Information Extraction Technology in Machine Translation

IE methods for improving and evaluating MT quality

by
Bogdan Babych

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

University of Leeds
Centre for Translation Studies

March, 2005

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Abstract

The thesis analyses the impact of Information Extraction technology on MT quality. Corpus-based experiments described in the dissertation show how IE can meet the demands of MT in two aspects – improving the accuracy of evaluating MT output and improving the adequacy of translation on lexical and morphosyntactic levels. These results also suggest that the IE technology models certain natural phenomena that are fundamental for the process of translation, but until now have been overlooked by MT researches: ranking relative relevance of translation equivalents, avoiding translation of specific items, etc. In this respect the ability of IE to concentrate on the most relevant information and to ignore irrelevant bits exactly meets this demand of MT technology and allows MT to overcome some of its fundamental limits. Improvements in MT quality via Named Entity recognition and higher correlation between IE-oriented MT evaluation metrics and human scores illustrate this suggestion. Therefore, IE technology has a potential to improve MT quality if it is properly integrated into MT architecture. IE methods can also point to some previously unknown limits of MT technology if they are used for MT evaluation.

Acknowledgements

Work on this thesis was supported by the White Rose Studentship (Universities of Leeds and Sheffield).

I would like to give special thanks to my supervisors – Prof. Tony Hartley, Dr. Eric Atwell and Prof. Yorick Wilks for their constant encouragement during my work on this thesis.

I am very grateful to my Examiners Prof. Harold Somers, Prof. David Hogg and Dr. Serge Sharoff for their insightful comments and suggestions.

Contents

List of Abbreviations

ADE - adequacy
AI - Artificial Intelligence
ARP - assumption of reference proximity
ASL - average sentence length
ASW - average syllables per word
BLEU - bilingual evaluation understudy metric
BLEUrXnY - BLEU score which uses X reference translations and Y maximal N-gram size
CLEF - Cross Language Evaluation Forum
DA - dictionary adaptation
DNT - "do not translate"
DOT - data-oriented translation
EBMT - example-based machine translation
EM - emails
EXP - expert
FAHQT - fully automatic high quality translations
FEM - feminine
FKGL - Flesch-Kincaid Grade Level score
FLU - fluency
FR - Flesch Reading Ease score
GATE - general architecture for text engineering
GPL - general public licence
HT - human translation
IDF - inverse document frequency
IE - information extraction
KF-IDF - category frequency / inverse document frequency
LFG - lexical function grammar
LTV - legitimate translation variation
MT - Machine Translation
MUC - Message Understanding Conference
MUMIS - Multimedia Indexing and Search Environment
NA - not available
NE - named entity
NER - named entity recognition
NIST - National Institute of Standards and Technology
NL - natural language
NLP - natural language processing
NLTK - natural language toolkit
NP - noun phrase
ORI - original
PCD - proper / common disambiguation
PLUR - plural
REF - reference
SING - singular
SL - source language
SMT - statistical machine translation

ST - source text
STDEV - standard deviation
STRAND - Structural Translation Recognition for Acquiring Natural Data
TDS - String Translated with Do-not-translate list
TF - text frequency
TF-IDF - text frequency / inverse document frequency
TL - target language
TT - target text
TWS - String Translated Without the do-not-translate list
USL - usability
WMN - weighted N-gram model
WP - whitepaper
WSD - word sence disambiguation

Preface

The output quality of Machine Translation systems has been always the central issue in MT research and development. However, it is not equally handled in different research paradigms. In recent decades Machine Translation became a commercially viable technology with a growing number of industrial applications. This was motivated by increasing usefulness of state-of-the-art MT systems in the workflow of professional translators due to the following developments: much wider lexical and grammatical coverage, integration of pre/post-editing and translation memory tools, the use of controlled language approaches, domain-specific terminological databases and disambiguation strategies, user dictionaries, etc. The "usefulness" of MT is no longer associated exclusively with the "quality" of raw MT output: there is recognition that even imperfect text produced by the systems can find its applications.

As a result there appeared two separate directions in MT research. The first one – the "perfectionist" direction – is treating MT as a "*venerable scientific enterprise*" and/or a "*technological challenge*" (Nirenburg and Wilks, 2000); to a large extent this direction is motivated by the idea of achieving "fully automatic high quality translation" (FAHQT). The second direction is "pragmatic"; it views MT as an "*economic necessity*" and is concerned primarily with the usefulness of existing systems and techniques, having conceded that very limited progress in MT quality is achievable. Still there is a gradual progress in text quality, and the advances in the "pragmatic" direction build up on extended capabilities of MT provided by the "perfectionist" route. However, the MT quality comparable to the quality of professional human translation (HT) has not been reached. In terms of MT evaluation scores there is still a huge gap between the quality of HT and MT. The disagreement between the "perfectionist" and "pragmatic" directions concerns the question whether MT quality can be made comparable to HT in a foreseeable future, or whether there is a ceiling for output quality, which MT may have already reached or is about to reach.

There is an intuitive recognition that such limits exist, so new conceptual models are needed (Kay, 2003/1980), (Kettunen, 1986: 37). Therefore, there is a need to identify limits of current MT architectures which could point to some new productive lines of research in linguistics, Artificial Intelligence (AI) or translation studies, which may yield improved MT quality. In the history of MT constructive attempts to identify limits of current approaches to MT often inspired new research

directions. Many claims have been made that solution to some particular problems are crucial and even play a vital role for MT quality.

However, most demonstrations of MT limits give only a local picture of a particular problem based on abstract theoretical lines of reasoning, on isolated and artificially constructed examples rather than on empirical corpus-based data. Little effort has been made to rank the importance of the problems and to identify which difficulties are most typical for the state-of-the-art MT systems. However, yet there is no empirical assessment of what impact particular aspects of NLP could have on MT quality. Without such assessments it is hard to identify scientific problems that are likely to provide best engineering solutions for MT: we just don't know what linguistic or cognitive issues need to be resolved in the first place to ensure considerable improvements in MT quality. There even is a suggestion by M. Kay that "even if all problems of syntax, morphology, and computational semantics had been individually solved, it might not improve MT" (quoted in Wilks, 2003: 203). Such suggestion calls for empirical verification on corpus data, so there is a need:

- to systematically asses the exact impact of different NLP and AI technologies on MT;
- to identify theoretical and technological problems whose solutions will be the most important for improving MT quality, i.e., those problems that need to be solved in the first place.

These two tasks are related: corpus-based evaluation of the impact, which has a particular NLP technology on MT, also highlights its limitations and makes it comparable within a bigger picture, where it becomes possible to identify the relative importance of individual solutions and discover some missing, perhaps previously unknown technologies, which might appear essential for improving MT quality.

This thesis concentrates on the first of the two tasks mentioned above. I have chosen to investigate the impact of some aspects of *Information Extraction* technology on MT quality, particularly those aspects which were found relevant for MT: Named Entity Recognition (NER) and identification of salient terms (which typically are Named Entities (NEs), names of events, etc.) in text. The thesis aims at setting an example of how a systematic corpus-based evaluation of the impact on MT quality can be carried out for other NLP and AI technologies.

There are two aspects how a particular NLP/AI technology can be useful for MT. On the one hand it may be integrated into analysis and transfer modules of some MT system and *directly* contribute to improvements in MT quality. On the other hand it may become a part of some evaluation metric for MT, and highlight

different problems which need to be solved by MT researchers. I will show that this *indirect* contribution of an NLP/AI technology to MT quality even more important than its direct impact on MT quality, because in this case we may discover some new, previously unknown facts about language and translation process and arrive at some unexpected empirical results, while with the direct impact we just measure the effect of some known approach. The thesis:

- proposes an MT evaluation framework based on ideas from IE;
- identifies technological limits for MT that can be revealed by the IE-based MT evaluation;
- suggests ways of improving MT quality with IE techniques, such as Named Entity Recognition, and outlines a wider IE-guided MT architecture.

To conclude, the thesis is an attempt to systematically identify some typical needs and perspective lines of improvement for the state-of-the-art MT systems, and to evaluate the effect of the proposed solutions on MT quality, which can be done using IE technology and corpus-based MT evaluation techniques. The thesis tries to show how it is possible to arrive at a bigger picture of limitations of MT quality from the perspective of IE and to empirically assess the effect of the suggested technological improvements using corpus data.

Part I

Information Extraction and Technological limits on MT quality

Chapter 1
Data acquisition and data processing problems in MT

> The *prima face* case against operational machine translation from the linguistic point of view will be to the effect that there is unlikely to be adequate engineering where we know there is no adequate science. A parallel case can be made from the point of view of computer science, especially that part of it called *artificial intelligence*. (Kay, 2003/1980: 222).
>
> … If we are doing something we understand weakly, we cannot hope for good results. And language, including translation, is still rather weakly understood. (Kettunen, 1986: 37)

Several NLP technologies (such as Information Retrieval, topic detection, Information Extraction for the most part) have reached a level of quality that is comparable to human performance on similar tasks. However, the quality of MT is still far behind the quality of professional human translation. The problem how to bridge the gap between the human quality and MT quality is central to research and development efforts in the field. There is recognition that on this stage we don't adequately understand cognitive processes involved in human translation, such as language comprehension, production, application of translation strategies and procedures, so we don't have appropriate models to implement in MT systems.

A legitimate question to ask is whether research and development experience in MT can be used as a "probing action" to systematically investigate what exactly is missing from our theoretical picture of the processes involved in translation and how important the discovered problems are for improving existing technologies. There have been a number of suggestions how MT quality can be improved, e.g., using anaphora resolution (Mitkov, 2002: xii; 2003: 257), disambiguation, including word sense and syntactic disambiguation (e.g., McEnery, 2003: 459), term extraction (Jacquemin and Bourigault, 2003: 604), representations of the rhetorical structure of texts, of common-sense knowledge, etc. (Wilks, 2003: 203). The majority of these suggestions were based on abstract theoretical lines of reasoning, on isolated and artificially constructed examples rather than on empirical corpus-based data. Little effort has been made to rank the importance of the problems and to identify which difficulties are most typical for the state-of-the-art MT systems. Inventories of possible improvements of MT often look like unstructured wish lists, so it is hard to justify the claims that a particular technology, e.g., anaphora

resolution, plays “a vital role” in machine translation, if there is no corpus-based analysis of its significance and no comparison to other problems.

According to (Hutchins and Somers, 1992: 161) translation inadequacies produced by systems have been the primary motivation behind MT evaluation efforts, which aimed at establishing how good raw MT output is, what is its potential for improvement and what is the best way of using imperfect MT output in practice. However, nowadays corpus-based MT evaluation can address not only such “post factum” problems (the problems outside the stage of fundamental MT research), but also some major pivotal problems of MT technology, which may determine ideology and structure of models for translation process behind MT, and point in the right direction for the mainstream MT development effort by assessing the potential of improvability and the limits of particular approaches and their possible “ripple effects” – cases when the desired effect is achieved but problems are created elsewhere (Hutchins and Somers, 1992: 169)

There is a need for a systematic corpus-based evaluation framework to asses the impact of a particular technology on MT quality and to compare it with the impact of other potentially useful technologies, and possibly – with baseline performance of alternative MT architectures on similar tasks. This evaluation would give a realistic view on what can and cannot be achieved by such technological amendments. Also we will be able to establish technological boundaries of different MT architectures and to identify what is still missing in MT.

The task of identifying limits of MT architectures is closely related to the task of discovering solutions to the problems identified in this way: some technological developments in MT may be viewed as responses to such limits (Wilks, 2003: 204), e.g., recognition of the role of knowledge in Ontological Semantics (Nirenburg and Raskin, 2004) responds to Bar-Hillel’s “demonstration of the nonfeasibility of FAHQT” (Bar-Hillel, 1960), i.e., accessing ontological knowledge can successfully disambiguate word senses in Bar-Hillel’s example: *Little John was looking for his toy box…The box was in a pen*. Similarly, development of data-driven approaches such as statistical MT (SMT) and Example-Based MT (EBMT) try to overcome the data-acquisition bottleneck in MT technology. However, at the moment there is no bird's-eye view which limits on MT are more serious and which are less serious; there is no empirical assessment of what impact particular aspects of NLP could have on MT quality. Without such assessments it is hard to identify scientific problems that are likely to provide best engineering solutions for MT: we just don’t know what linguistic or AI issues need to be resolved in the first place to ensure considerable improvements in MT quality.

Therefore, systematic identification of current limits on MT is essential for achieving progress in output quality. The progress could be achieved via theoretical analysis of boundaries of existing approaches. This analysis can be done on parallel corpora, which contain original texts, human translations and MT output, by comparing translation operations performed by human translators and by MT systems.

The goal is to push the state-of-the-art MT architectures and suggested amendments to their limits (at least theoretically), and either to show that every operation that we find in human translation can be systematically covered by a given approach, or that some classes of operations are not possible, or will always require unsystematic ad-hoc solutions within a particular framework.

Let us imagine the following "thought experiment": if we had an ideal dictionary, a full-coverage contrastive grammar, a very large and clean aligned parallel corpus and if we could apply efficient word sense disambiguation methods – would this be sufficient to reach the quality of human translation in MT? We can continue to add items to our "wish list", and to examine (at least in theory) whether they are sufficient to cover everything which happens in human translations. The experiment tests if models, architectural and methodological features suggested for MT (e.g., the noisy channel model, the direct or transfer architecture, statistical or example-based approaches) could in all cases be linked to translation strategies and procedures found in human texts. The experiment may highlight the issues that are essential for the process of translation, but are left behind by certain MT architectures or approaches.

The idea of this experiment may be attributed to M.Kay, who argued that "even if all problems of syntax, morphology and computational semantics had been individually solved, it might not improve MT" (quoted in (Wilks, 2003: 203)). A constructive part of Kay's argument is that the "wish list" of required MT features should be created systematically and motivated by empirical evidence from parallel corpora and defined in terms of concrete translation operations and procedures, not just in terms of abstract items of research agenda. The suggested "wish list" should be ranked according to frequency and seriousness of particular MT problems in the parallel corpus.

In this thesis we concentrated on the first of the two tasks described above. The thesis aims at systematic corpus-based identification of some typical needs and perspective lines of improvement for the state-of-the-art MT systems, and at evaluation the effect of the proposed solutions on MT quality, which can be done using different aspects of IE technology and IE-oriented MT evaluation techniques.

1.1. MT paradigms, evaluation and state-of-the-art of MT technology

The opening paragraph of Warren Weaver's Memorandum formulated the problem of Machine Translation as a practical task, aimed at "…contributing at least something to the solution of the world-wide translation problem through the use of electronic computers…" (Weaver, 2003/1949: 13). However, this task evolved into a wider set of research paradigms. Indications of this wider agenda for MT can be found in Weaver's text, e.g., a suggestion that "... in the manifold instances in which man has invented and developed languages, there are certain invariant properties which are, again not precisely but to some statistically useful degree, common to all languages. This may be, for all I know, a famous theorem of philology." (Weaver, 2003/1949: 13). This shows that in W.Weaver's view MT is not just the way of developing useful translation tools, it also aimed at discovering new facts about human cognition, structure of natural languages and translation process, and surely these discoveries are important beyond the practical task of automated translation of a text from one language into the other. However, different researchers give priority to practical and theoretical goals of MT. In this respect at least tree different research paradigms in MT can be identified – "pragmatic", "perfectionist" and "theoretical". The pragmatic paradigm has a greater engineering emphasis, the perfectionist paradigm emphasises the discovery of natural phenomena involved in translation, and the theoretical paradigm is concerned with developing and testing theoretical models of such phenomena.

The pragmatic paradigm views MT as a commercially viable technology with important industrial applications, bringing MT into the context of its users – translators, localisation developers, home users, etc. The *usefulness* of MT is no longer associated exclusively with the *quality* of raw MT output: there is recognition that even imperfect text produced by the systems can find its applications, in the first place – for *assimilation* purposes (i.e., for comprehension).

The followers of pragmatic MT work not only on fundamental issues of text quality, but also on extension of known methodologies and improving usefulness of existing systems and techniques, e.g., on increasing lexical and grammatical coverage, automatic creation of large-scale dictionaries, integration of pre/post-editing and translation memory tools, the use of controlled language approaches, domain-specific terminological databases and disambiguation strategies, user dictionaries, etc. An important advantage of the pragmatic paradigm is that MT is viewed in a wider framework of industrial and personal use of this technology.

The pragmatic paradigm in MT is completely justified by the existence of a gap between MT *quality* and *usefulness*. However, some researchers support the pragmatic paradigm with an additional argument about unattainability of "fully automatic high quality translation" (FAHQT). There are serious problems with this argument; nevertheless this argument is not necessary to support the case for pragmatic paradigm in MT, which is sufficiently strong without it.

The perfectionist paradigm in MT, as it is traditionally perceived, aims at FAHQT. However, this definition of its goal is not entirely accurate, rather this goal was attributed to the followers of the perfectionist paradigm by critics in the early days of MT: "Many groups engaged in MT research still regard fully automatic, high quality translation (FAHQT) as an aim towards which it is reasonable to work. […]. I believe to be in possession of an argument which amounts to an almost full-fledged demonstration of the unattainability of FAHQT, not only in the near future but altogether" (Bar-Hillel, 2003/1960: 45).

The person who formulated a concept of FAHQT and developed an argument against it was Y.Bar-Hillel – "an eminent philosopher of language and mathematical logician", who "has never written or designed an MT system" (Nirenburg, 2003: 7). The "perfectionist" pioneers (such as Erwin Reifler, who worked on developing MT systems) and later MT developers usually didn't state their goals this simplistic way. If we turn to their original papers, we get an impression that their thoughts were misunderstood or misinterpreted by Bar-Hillel, and that diverse lines of their research were labelled as "FAHQT" and to some extent "demonised". This is easy to see in the following quote from Reifler: "My research in comparative semantics, my experience in translation, and my teaching of foreign languages made me first relegate the MT to the realm of the impossible. In the course of further research, however, I began to see certain limited possibilities." (Reifler, 2003/1955: 21).

The real goal of MT perfectionists is to create "awareness of the obstacles that lie in the way of a complete mechanization of a translation process" (Reifler, 2003/1955: 21). Some of such obstacles were first exhibited in W.Weaver's memorandum, but "perfectionists" investigate these obstacles in a principled way. FAHQT may still be an ultimate goal, but it is not the area of everyday research. Creating awareness of fundamental MT problems is different from the long-term FAHQT goal. In reality the "perfectionist" problems are:

1. Are there any fundamental limits on particular approaches and methods used in MT (e.g., word-for-word translation, statistical MT, example-based MT, etc.), are there methods that in principle can solve all known problems, or can we prove that certain phenomena cannot be covered systematically?

2. What MT quality is achievable if a particular approach is pushed to its limits?

3. What are new lines of attack to deal with the phenomena that supposedly are not covered by current approaches in a systematic way?

This formulation of perfectionist goals is more accurate in characterising the paradigm in the past and nowadays. It highlights the fact that MT perfectionists study the limits of current approaches to MT in their everyday work, and are not pursuing unrealistic targets, so FAHQT is not the core business of MT perfectionists.

The theoretical paradigm uses MT as a test-bed for linguistic or translation theories and uses these theories systematically as a foundation for MT development (e.g., Rosetta system (Rosetta, 1994)). The distinctive feature of the theoretical paradigm is its links with full-scale theories (as opposed to methodology-based approaches), its recognition that ad-hoc methodology alone isn't sufficient for solving problems of MT.

The major difference between the theoretical paradigm and the perfectionist paradigm is that "perfectionists" adopt "*bottom-up*" approach – from open-ended MT problems and limitations on MT quality towards ways of addressing them, without committing themselves to any particular theory or model that is external to the research material. "Theoreticians" explore a "*top-down*" line of research, assuming that a particular "external" theory, model, some general framework, or even some interdisciplinary approach will be beneficial for MT and trying to apply this assumption for solving particular MT problems.

However, within the theoretical paradigm an MT-external theoretical interest of researchers often competes with the core practical goals of MT; therefore many of the developed systems remain experimental. The theoretical paradigm has been less prominent than the other two, e.g., according to S.Nirenburg, "it is, indeed, remarkable how little impact theoretical linguistics had on the early machine translation"; also nowadays MT "…can hardly be considered a direct application of theoretical or descriptive linguistics" (Nirenburg, 2003: 3, 4). Although in the 1960s many projects, "while paying lip service to the practical needs of MT, would concentrate much more on applying and testing a variety of linguistic (e.g., syntactic) and computational linguistic (e.g., parsing) theories within the framework of MT" (Nirenburg, 2003: 4). The theoretical paradigm generated interesting ideas, applicable in other NLP areas, such as Ontological Semantics (Nirenburg, 2004),

Preference Semantics (Wilks, 1975), Conceptual Dependence theory (Schank, 1972, 1975).

An idea about the development of a theoretical paradigm for MT in future can be also found in the paper "The Mechanical Determination of Meaning" by the "perfectionist" E.Reifler, who suggests that a separate discipline – "MT Linguistics" – should be a theoretical ground for MT. (Reifler, 2003/1955). Attempts to develop such discipline resulted in mainstream efforts of 1960ies, mentioned by Nirenburg, and in numerous all-out theoretical attacks later (e.g., syntactic translation (Yngve, 2003/1957), compositional translation (Landsbergen, 2003/1987), use of Esperanto as interlingua (Witkam, 1988), statistical MT (Brown et al., 2003/1990), etc.). However, these overreaching models don't solve MT problems once and for all – they have limits, which need to be studied from the "perfectionist" perspective, i.e., in a principled bottom-up direction, starting from systematic coverage of the material. In practice, motivation for theoretical MT models very often doesn't come from systematic corpus-based analysis of applicability of the model and its potential to improve MT quality. Instead, this motivation is often illustrated by a few simple artificially constructed examples. However, understanding model's limits within a general picture of MT problems, highlighted by corpus-based evaluation, gives the model its proper place in MT technology: it can be accommodated with other methodologies, which address different kinds of particular MT problems better.

In modern terms Reifler's "MT Linguistics" can be defined as an area of translation studies focused on the tasks of MT. Both in MT and in translation studies the general agreement is that traditional linguistics is not sufficient, e.g., for accounting for the phenomena which are found in professional human translations. The theoretical basis of translation studies is much wider, although less formal in this respect.

Until now Reifler's idea hasn't been fully exploited (perhaps due to much later appearance of the field of translation studies on the MT scene). MT and translation studies may generate mutually useful theoretical conceptions by formalising concepts of *translation shifts* (Catford, 1965), *translation transformations* (Shveitser, 1988), *translation strategies* (Vinay and Darbelnet, 1995), *relevance theory* (Sperber and Wilson, 1986), (Gutt, 1991), e.g., for modelling transfer rules, lexical and syntactic disambiguation, etc. Such models may be productive for learning higher order translation equivalents in data-driven approaches to MT. I'll discuss some of these suggestions in the following sections. Unfortunately, nowadays these sources are not extensively used in MT circles, however these descriptions may be adequately formalised and eventually form a theoretical basis for improving MT.

The point of disagreement between the MT paradigms discussed above is the issue of MT quality: whether MT can reach the quality of human translation in the foreseeable future and what are the productive ways to ensure the progress. Interestingly, MT quality may be the point of interaction between the paradigms (even though they have apparently different goals), which may yield interesting results, e.g., the limits of a particular approach to MT (the perfectionist paradigm) may be described via referring to the types of translation strategies that are systematically covered by this approach (the theoretical paradigm); some of these strategies may be learnt from the corpus, reducing the efforts in developing of large-scale MT systems (the pragmatic paradigm).

In this respect, MT evaluation is in fact a potential point of co-operation between different paradigms, since in this area the most "controversial" questions of whether, when and how the MT technology could reach the quality of human translation – are empirically testable. On the other hand, MT research and development could contribute to the solution of a serious theoretical problem in MT evaluation – objective definition of the concept of MT quality. Nowadays this concept is derived from human intuitive judgements about such parameters as adequacy and fluency of translation measured under specific experimental conditions (White et al., 1994). Automated measures of MT quality (e.g., Rajman and Hartley, 2001; Papineni et al., 2002) are calibrated with respect to these intuitive human criteria. Without appropriate analysis of the structure of the MT quality concept we cannot be completely sure whether we are measuring the right parameters, i.e., the quality itself, not usefulness of MT for a particular task.

Evaluation efforts in NLP in general and in MT in particular are traditionally aimed at monitoring the progress in quality achieved by large-scale systems, i.e., at quality assurance on their ***development cycle***. However, evaluation has a wider impact, which may allow researchers to achieve better understanding of the modelled phenomena, to discover new fundamental knowledge about natural phenomena that underlie linguistic functions modelled by a particular technology, e.g., to discover technological limits of current approaches and to suggest new research agenda for how to move the technology beyond such limits.

New facts and models revealed or developed through interpretation of the evaluation results can be reused for improving the quality of applications and put into the evaluation paradigm again and again. From this perspective corpus-based evaluation is part of the ***research cycle*** for linguistic engineering tasks. It is a

powerful tool for developing adequate models for modelled natural phenomena in a systematic way.

In this sense MT evaluation is an instrument of doing systematic research of fundamental MT problems. MT evaluation includes parameters of different stages in the development, installation and operation of MT systems, however testing of the "raw" MT output is common to all stages (Hutchins and Somers, 1992: 162-163). I will concentrate on this aspect and will sometimes refer to it by the general term "MT evaluation".

The need for evaluation of raw MT output was created when large-scale MT systems with sufficient coverage for real-world subject domains and for general-purpose texts were developed. The interest in evaluation was demonstrated in the late 1970s, when evaluation of Systran for European Communities received much attention. The principal motivation for the evaluation methodologies in MT comes from the MT development perspective: "A major question to be asked about any MT system is, therefore, how good are its raw translations, what is the potential for improvement, and how may it be best and most cost-effectively used in practice?" (Hutchins and Somers, 1992: 161). The central concern of MT evaluation is comparison of some large-scale MT systems and monitoring the progress in their development (or the effect which any changes may have on the quality of their output).

However, it is surprising that the significance of MT evaluation for fundamental research in MT hasn't been fully appreciated. The first suggestion to use MT evaluation in the "perfectionist" way was made by J.Hutchins and H.Somers (1992: 161): "… one role of evaluation must be to introduce realism in public discussions of what MT systems can and cannot do and what they may be able to do in the future". However the idea that MT evaluation can outline the limits of the current technology and suggest ways to overcome these limits on the basis of large-scale corpus-based experiments – hasn't been the core of MT evaluation field so far.

The advantage of current "numeric" approaches to MT evaluation is that MT quality parameters, such as adequacy and fluency, are comparable across different experiments. In this way we can compare relative performance of different systems or establish an absolute level of performance of a particular system in relation to some gold standard or a human level of performance. We can also assess an impact of a particular technology or solution on MT quality.

There are two aspects how a particular NLP/AI technology can be useful for MT. On the one hand it may be integrated into analysis and transfer modules of

some MT system and *directly* contribute to improvements in MT quality. On the other hand it may become a part of some evaluation metric for MT, and highlight different problems which need to be solved by MT researchers. I will show that this *indirect* contribution of an NLP/AI technology to MT quality is exactly the way to broaden *constructive* research horizons of MT technology, to search for new models, approaches and architectures in MT, to discover new, previously unknown facts and arrive at some unexpected empirical results, while with the direct impact we just measure the effect of some known approach and get an *indication* where it brings MT in terms of quality.

In this thesis we have chosen to investigate the impact of some aspects of IE technology on MT quality, particularly those aspects which were found relevant for MT: Named Entity Recognition (NER) and identification of salient terms (which typically are Named Entities (NEs), names of events, etc.) in text. The thesis aims at setting an example of how a systematic corpus-based evaluation of the impact on MT quality can be carried out for other NLP and AI technologies. we report on a series of experiments on both indicative and constructive MT evaluation related to IE technology.

The starting point in this discussion is a preliminary assessment of the state-of-the-art quality of several commercial MT systems in comparison to the quality of human translation. We need to know (at least approximately) where in absolute terms MT technology stands, and how much room it has for improvement. Even though researchers agree that MT quality is far behind the quality of human translation, we may wish to know how far, and what is the difference between the best and the average systems in absolute terms.

It is difficult to answer this question precisely, but automated corpus-based MT evaluation scores, such as BLEU (Papineni et al., 2002) may give a rough indication of the gap between the quality of human translation and MT. For this comparison we used what has now become a standard benchmark for MT quality – the DARPA 1994 MT evaluation corpus (White et al., 1994), which consists of 100 French newswire texts, each about 350 words long, 2 independent human translations into English – (called the “Expert” and the “Reference”) and the output of 5 different MT systems for each text. In DARPA 94 experiments both “Expert” and “Reference” translations were done by human translators, they were used differently: the “Reference” translation was used as a gold-standard translation, against which other translations – MT output and the human “Expert” translation were compared. Therefore, there are no human scores for the “Reference” translation – only the “Expert” translation was evaluated by human judges, while “Reference” was used only as a basis for comparison. Quick informal comparison of

the two groups of human translations suggests that the “Expert” texts are a bit more professional – overall they are more idiomatic and less literal than the “Reference” texts. However, for the purposes of my experiment such difference in the quality of human translations is negligible, on a larger scale these two groups of texts may be regarded as top-standard translations.

BLEU method uses one or more human translations as a reference and calculates the distance between the human reference(s) and the evaluated text (which may be an MT output or some other human translation) by computing precision of N-gram matches in these 2 translations. For this experiment the BLEU script was run two times: each time one of the two human translations was used as a reference; the other human translation was evaluated alongside with the 5 MT systems. BLEU scores are presented in Table 1 and in Figure 1:

	BLEU-r	BLEU-e
MT-ms	0.2037	0.2048
MT-globalink	0.2197	0.2207
MT-candide	0.2348	0.2387
MT-reverso	0.2724	0.2742
MT-systran	0.2771	0.2831
HT-expert/ref	0.4303	0.4304

Table 1.1. BLEU evaluation of MT and HT in DARPA-94 corpus

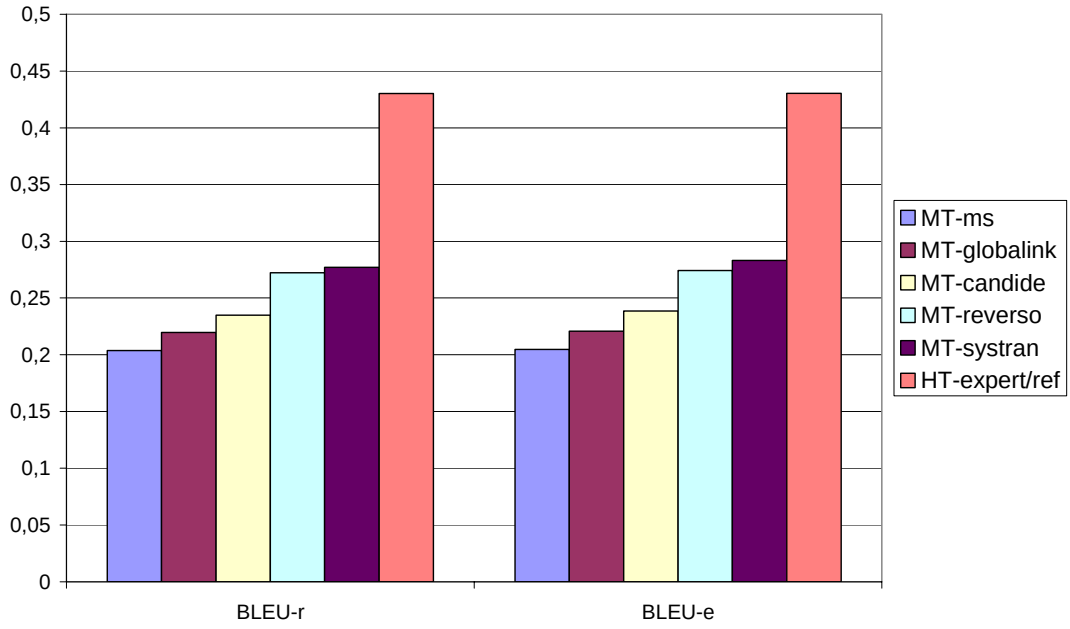


Figure 1.1. BLEU evaluation of MT and HT in DARPA-94 corpus

It can be seen from the charts that the results for the 2 runs of the experiment are very close (which is an indication that evaluation scores are accurate, and the size of the evaluation corpus is sufficient).

The most interesting information in these charts is the extent by which human translations are ahead of MT output (even for the best systems) – human translations

are 1.5 – 2 times "better" in terms of BLEU scores. Clearly, the range of BLEU scores may give an immediate indication (at least for the French-into-English direction) whether the translation was done by a human native speaker or an MT system.

However, the size of the gap in terms of BLEU scores cannot be directly interpreted as the size of the quality gap between human translation and MT. If we want to interpret this gap as an amount of effort that should be put into MT system in order to arrive at a certain level of BLEU scores, we should note that the BLEU scale is not homogeneous: it becomes harder and harder to achieve improvement in terms of the BLEU scores the further up the scale we get. The matter is that the BLEU method counts lexical matches (matches of individual words and their sequences) for calculating the score. The distribution of lexical items in text follows Zipf's law: a small number of frequent lexical sequences cover a relatively large proportion of text, but the number of lexical sequences needed to cover the remainder of the text grows exponentially. Therefore, in order to have a rough estimation about the amount of effort and time, which is still needed to reach the quality comparable to the quality of human translation, and in order to compare it to the amount of time and effort already spent, we need to transpose the numbers in Table 1 onto an exponential scale, which will match the Zipf's law distribution of lexical items in text and somehow "model" the efforts to achieve appropriate lexical coverage.

Of course, MT development is not just about ensuring appropriate dictionary coverage or an appropriate set of translation equivalents, which are known to follow Zipf's distribution. There may be other phenomena involved which will require even greater efforts and possibly – certain non-incremental "revolutionary" theoretical developments in MT, AI, etc. But if there is any substance behind the automated MT evaluation methods such as BLEU (at least on a larger corpus-level scale), all such developments will come down to finding appropriate words and word sequences as translation equivalents. Steepening of the scale of efforts will be at least exponential compared to the absolute number of lexical matches, if we try to achieve similar improvements measured in terms of BLEU scores further up the scale.

For that reason to get a rough estimation where the state-of-the-art MT systems stand after several decades of research and development, and how much effort is still needed as compared to the effort already put into the systems, we need to use an exponential scale or to rise the scores in Table 1 to some constant power, e.g., to the power e=2.718282. (The question which constant to use is an open issue, but here we are looking at the scale of the problem rather than on absolute values). Table 2 and Figure 2 give these modified scores.

	blue-r^E	blue-e^E
MT-ms	0.01323	0.01343
MT-globalink	0.01625	0.01645
MT-candide	0.01947	0.02036
MT-reverso	0.02916	0.02968
MT-systran	0.03054	0.03238
HT-expert/ref	0.10104	0.10110

Table 2. BLEU scaled exponentially

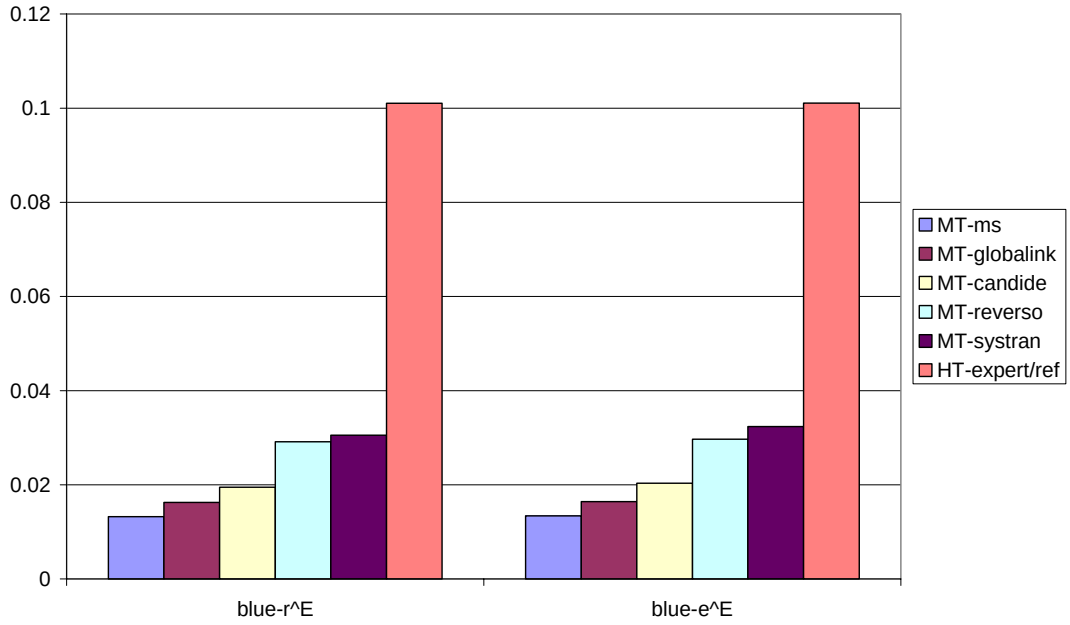


Figure 2. BLEU scaled exponentially

Once again, these figures are useful only as approximate guidelines, which give additional credibility to our intuitive feeling about how far we need to go in order to get to the level of human quality, e.g., one could safely suggest that reaching the quality of human translation in MT is at least several decades away, given the fact that in 1994, after 40 years of research and development, even the best MT systems covered only one third of the way.

However, from a certain perspective the results of this simple corpus-based evaluation experiment are astonishing: in any case the gap between the best and the worst MT systems is much smaller than the gap between the quality of human translation and MT. The magnitude of this difference shows that even having used corpora, wide-coverage grammars and dictionaries for the mainstream translation direction, the best commercial MT systems didn't achieve a breakthrough in MT quality. These results prove the suggestion that there are limits on what is achievable via an extensive way of developing MT (e.g., extending databases of translation equivalents with rule-based or data-driven methods), and that there is a need for

discovering fundamental limits on MT and looking for intensive solutions to these problems.

1.2. The Data Acquisition and the Data Processing bottlenecks

> Not the power to remember, but its very opposite, the power to forget, is a necessary condition for our existence. (Saint Basil, quoted in Barrow, 2003: vii)

Such an empirical verification of M.Kay's thesis implies that MT problems go beyond ***data acquisition***, so approaches based on the maxim "*there is no data like more data*" even at their limits don't guarantee "human" quality of translation. ***Data processing*** problems in MT are equally complicated and cannot be solved only by using current data-driven approaches, so we could say that "*there is no processing like intelligent processing*".

The processing side of MT is not specifically addressed by data-driven approaches, which primarily aim at acquiring databases of translation equivalents, and therefore remain relatively unsophisticated on the processing side: they generally use more advanced versions of the plain "dictionary lookup" strategy, which supplies equivalents for the maximal number of constructs successfully identified in the source text. However, it can be shown that on the processing side we need a number of diverse "artificial intelligence" tools for MT in order to systematically model even most common phenomena which we find in human translations.

A typical situation when MT problems rest on the "processing side" rather than on the "data side" is inability of MT to correctly *avoid* translation of certain less relevant information that is expressed in the source text (ST). Such information filtering is a core mechanism in human translation. The need for it arises from a well known fact in translation studies: generally it is not possible to preserve all information expressed in the source text, since the usage of translated units is different across languages. However, information expressed by ST has different degrees of relevance for communication, usually – according to its place on semiotic hierarchy: pragmatic functions are on top, followed by semantic (reference, then sense) functions, and finally syntactic functions.

Human translators rank information by its relevance and generate a target text (TT), preserving most relevant functions – as many as possible, starting from the top of their "relevance list". In case when lower level functions are preserved, but higher level functions are lost, the translation is "literal", and vice versa – if lower level functions are lost without clear motivation (e.g., to preserve higher level

functions), the translation is "free" (Shveitser, 1988: 87-88), (Barhudarov, 1975: 186). C.f.: "[…] a German maxim "so treu wie möglich, so frei wie nötig" (true, when possible, free, when necessary) reflects the logic of translator's decisions well: aiming at precision when this is possible, the translation allows liberty only if necessary […] The decisions taken by a translator often have a nature of a compromise, […] in the process of translation a translator often has to take certain losses. […] It follows that the requirement of adequacy has not a maximal, but an optimal nature." (Shveitser, 1988: 88, 96, quotes translated from Russian).

A professional translation is restricted from the two sides in this sense, so human translators operate within a relatively narrow "corridor" of possibilities. Within a particular language pair, there are requirements not only for what has to be preserved, but also – for what is *not* intended for translation and has to be filtered out and lost (!).

An example of such an obligatory loss is the following English sentence (from a text on history of football):

(1) "*The Danish flair and verve saw them beat France twice in 1908*",

which was translated by a human translator into French as:

(2) "*Le sens du jeu et la créativité des Danois a raison des Français à deux reprises en 1908.*" (lit.: The feeling of the play and the creativity of the Danes are right for the French twice in 1908).

The equivalent for "*The flair and verve saw…*" ("*le sens du jeu et la créativité a raison…*") in French relates to the English phrase on semantic "reference" level, but not on the lower levels of semantic "sense"– the mode of presentation of its reference (Richard, 2003: 2), and not on the level of syntactic structure. Both phrases have the same referent, but differ in lexical and syntactic means of "presenting" this referent, because collocation restrictions for lower level equivalents are incompatible.

A Russian professional translation of the phrase "*verve saw…*" in another sentence contains transformations on even deeper level:

(3) "*Bayern began with the verve which saw them come from behind to defeat Celtic FC a fortnight ago.*"

(4) *Гости, две недели назад одержавшие волевую победу над "Селтиком", с первых минут завладели инициативой*. (lit.: *Guests, who two weeks ago gained a strong-willed victory over "Celtic", from the first minutes took the initiative*.)

– In (4) there is even no referential semantic equivalent for the phrase "*verve saw…*" in sentence (3). Only pragmatic equivalence is preserved in Russian, the pragmatic functions of the "*verve saw*" are conveyed by phrases "*strong-willed victory*" and "*took the initiative*".

The metaphor "…verve saw…" is linguistically and culturally acceptable in English, but its literal translation does not make sense in Russian. In order to convey a relevant meaning, a human translator sensibly applies transformations on the referential semantic level, distributing the meaning of "verve saw" across two phrases, which also gives a different syntactic perspective to the sentence.

There are linguistic reasons for the differential cultural acceptability of such metaphors: as compared to Russian, English is characterised by a broader semantic range of verbs being able to take inanimate subjects, which gives rise to "personification" metaphors. This property is often interpreted as a compensation mechanism for the relatively fixed word order of English (Shveitser 1988:143).

As it could be expected, modern English-French and English-Russian MT systems produced literal translation for sentences (1) and (3), close to the point where the translations become unintelligible in TL: *Le flair et le verve danois les ont vus battre la France deux fois en 1908. (Systran); Bayern начался с воодушевления, которое видело, что они прибыли из-за нанести поражение Кельтскому FC две недели назад. (ProMT).* Dealing with such complex compensation strategies – i.e., diagnosing that there is a problem with a metaphor that does not have a literal translation, and finding a semantically equivalent sentence with a different syntactic perspective – is today clearly beyond the state-of-the-art of equivalent-oriented MT architectures. Enumerating all possible metaphors in a dictionary cannot solve the problem, since new metaphors are productively created in everyday speech within the limits of a particular language and culture. How can processing-oriented approaches address such problems?

Note that in both English sentences human translators *avoided* translating the main verb "*saw*". Lexical equivalents for such verbs in the context of metaphorically used subjects are very likely to fall beyond the "relevance" threshold (at least for translation directions English-into-French and English-into-Russian). It is difficult to see how MT system, which rely on simple "lookup" in databases of translation equivalents as their processing strategy, can systematically account for such regular losses of relevance by content words and trigger necessary translation transformations. This phenomenon requires more sophisticated processing strategies for MT, something like Saint Basil's "power to forget". In this particular case we need at least the following:

(a) A possibility for an MT system to process data annotations that could be called ***negative equivalents*** (for units like the verb "*see*" in (1) and (3)), which explicitly specify that by default a particular unit shouldn't be looked up in available databases of equivalents, and the system should try to use one of available compensation strategies instead to get out of a difficulty (e.g., to translate syntactically related units and then to supply collocations from monolingual TL corpus for the unit annotated as a negative equivalent). Further I argue that negative-equivalent-type annotations play essential role in finding proper translation strategies for Named Entities (NEs). In this respect the problem of translating NEs is similar to the "*verve saw*" case: translation is better if we restrict information used by MT for looking up equivalents, instead of extending it, and shift attention to the processing side.

(b) The ability of a system to assess relevance of translation units on-the-fly in unrestricted text (e.g., by using some rule-based approach or statistical measures which correctly approximates human intuition about those values, like tf.idf scores[1]), to produce appropriate annotation for potential negative equivalents, and to give priority to more relevant equivalents which could fire over the same segments.

A legitimate (and a very interesting) question is whether it is possible to correct the problem of the "verve saw" type with a dictionary update. The answer should be obvious if one tries to translate into some other language a reasonable number of different contexts, where such phrases can be used. Here are examples how they may be translated into Russian:

(4a) His pace and attacking <u>verve saw</u> him impress in England's World Cup game against Samoa.

– Его темп и <u>атакующая мощь</u> впечатляли во время игры Англии с Самоа на чемпионате мира.

Lit.*: His pace and <u>attacking power</u> impressed during the game of England with Samoa at the World Cup*

(4b) Legout's <u>verve saw</u> him past world No 9 Kim Taek-Soo in his first match.

---

[1] tf.idf scores will be formally introduced in Chapter 2 together with their IE-oriented modifications – S-scores

– Настойчивость Легу позволила ему в первом матче обойти Кима Таек-Соо, занимающего 9-ю позицию в мировом рейтинге.

Lit.*: Legout's persistency allowed him in the first match to get round Kim Taek-Soo*

(4c) The Pericos' extra verve saw them go a goal up in the first half.

– Перикос приложил все усилия, что позволило им повести в счете в первой половине встречи.

Lit.*: Pericos made every effort, which allowed them to lead in score in the first half of the game*

These examples show that it is not possible to create any reasonable dictionary entry for "verve saw" for an English-Russian MT system. However there is something common to all these examples: the verb "saw" in all cases is not translated (it is a true "negative equivalent"); translation of "verve" usually depends on "smoothest collocational choices" of neighbouring lexical items (the term from Wilks, 2003: 203). The argument is that a problem of distinguishing where to give preference to rules or collocational smoothness – is not the problem of a relevant dictionary entry or a grammar rule, or that it is tricky to represent a negatively-formulated rule for the verb "saw" in a grammar or a dictionary (obligatory drop it from the sentence structure) – it is not the problem of data acquisition at all. But this is a problem of correctly identifying relative relevance of units and focusing available lexical and syntactic resources on salvaging what is the top priority.

Therefore, the problem of the "verve saw" type phrases so far is completely outside the realm of approaches, which address the problem of facilitating data acquisition (like SMT or EBMT). The applicability of these approaches alone for dealing with this problem will be very limited, in the sense that correct translation can be produced only for the contexts (usually clauses) seen in the training corpus. The real difficulty is that even though all individual examples of translating "verve saw" could be learnt automatically from aligned parallel corpus (with an appropriate level of sophistication of an algorithm), but every new context of the "verve saw" type phrases will most certainly require inventing something new for the sentence structure as a whole (given that the main verb "saw" is gone), and for the noun "verve" (depending on the smoothest collocational choices of words which happen to be around and which trigger relevant lexical functions). The real problem is that nowadays data-driven MT doesn't provide a general model for translating (or avoiding translation) of such items. It is very likely that for some examples of the "verve saw" type there will be no two cases in a training corpus, where they are translated in the same way, no matter how many occurrences of such phrases are

there now or will be added in future. It may be also possible, that none of the solutions will work for new occurrences of such phrases in the test corpus, which will require something new still. Even different human translators when asked to translate the same sentence, may find different solutions for "verve saw" type phrases much more often than for other phrases, on which they normally agree with each other. (It is possible to empirically verify this suggestion). Even in an ideal situation for data-driven training of an MT system, if the corpus gets larger and larger, translation solutions for "verve saw" type phrases may keep approaching "infinite entropy", which means that no ready solutions can be found for new cases. The reason for this is that finding such solutions is not the problem of data acquisition – it is a problem of intelligent data processing.

1.3. New data sources and relevance of translation equivalents

It would be interesting to test the following conjecture: beyond certain threshold, which is measured in terms of to coverage for an MT system, addition of new knowledge sources will not improve MT quality unless there is a possibility to rank relative relevance of competing translation equivalents coming from different knowledge sources.

Intuition behind this conjecture could be illustrated by the following example: it may be argued that annotation of information structure (e.g., *theme / rheme* annotation) would be useful for MT (very much like anaphora resolution or word sense disambiguation, which were put forward by many researchers as a way of improving MT).

Indeed it is possible to find examples where information about *theme* / *rheme* distinction motivates translation transformations; such cases have been informally described in philological literature on translation studies. For instance information structure can motivate changes in word order or active / passive transformations. It has been noted that English and Russian are characterised by "gradual increase in communicative load towards the end of an utterance", which may collide with constraints on word order – normally it should be direct in English and is often indirect in Russian (Breus, 2002: 22), (Chernyahovskaya: 1976: 14-24) (theme is underlined with a single line, rheme – with a double line):

(5) *Иную позицию заняли Франция и Германия*. (lit.: A different stand (Acc.) took France and Germany (Nom.));

There are at least 2 ways of translating (5) into English: we could either apply transfer rules to restore a neutral English word order, which will preserve morphosyntactic structure and active voice of the sentence, as in (6), or we could

preserve the order of information structure (the “gradual increase in communicative load”), turning the sentence into passive, as in (7).

(6) **? France and Germany took a different stand.*

(7) *A different stand was taken by France and Germany.*

Professional translators disapprove of (*6) (this is indicated by the star), and prefer translation (7): (Breus, 2002: 23). Russian sentence (8) poses similar problems (if “*the room*” is its theme), but here preferred transformations turn an adjunct into a subject, so such shifts are not only about active / passive transformations:

(8) *В комнате установилась мертвая тишина*. (lit.: In the room established itself deathly silence).

(9) **? A deathly silence descended upon the room. – (this translation is fine only if the whole sentence expresses rheme, e.g., (8) is the first sentence in a text).*

(10) The room turned deathly silent. – (preferred translation)

On the other hand, note that adding such information about theme / rheme distinction into MT introduces competition between new and existing translation equivalents, which could fire on the same segment. In this particular case the equivalents exist on different levels: syntactic and information structure, so their competition shouldn’t be a problem if there exists a fixed hierarchical list of precedence, e.g., that information structure (as a pragmatic phenomenon) always takes precedence over syntactic or semantic equivalents and always triggers syntactic and lexical transformations. However, it is easy to show that this is not the case, since interaction of other phenomena, e.g., anaphora resolution, could reverse the order of relevance:

(11) *В комнате установилась мертвая тишина. Она была вызывающей.* – (lit.: In the room established itself deathly silence. It/[she]$_{=\text{the silence}}$ was defiant.)

(12) *A deathly silence descended upon the room. It was defiant.*

(13) **? The room turned deathly silent. It was defiant.*

Translation (10), which is preferred for sentence (8) in isolation, doesn’t allow us to establish correct co-reference relations: in (13) pronoun *it* cannot co-refer with an adjective phrase *deathly silent*; there is a need to restore *deathly silence* as a noun phrase, as in (12). In this context lower level syntactic functions become more relevant than pragmatic functions of information structure.

In general, it is an empirical question whether introduction of a certain knowledge source into state-of-the-art MT systems causes improvements or

deteriorations on average. However, the benefits will always be smaller than the full potential of added data, unless the problem of balancing relevance of translation equivalents is addressed systematically. This problem cannot be solved within a rigid equivalent-lookup strategy – no matter how much new knowledge and data is added to the system and how carefully ordered is application of equivalents or their levels. Systematic solution requires more "intelligent" flexible processing strategies, where relative relevance of conflicting translation equivalents (which could fire over the same segments) could be weighted, compared across different levels or within the same level and correctly balanced for all related segments. Without such strategies MT quality figures would probably flatten upon approaching a certain level, and certainly would not reflect the amount of efforts put into acquiring new data.

To summarise, the discussed examples point to a potentially important problem in MT, – the problem of balancing relative relevance of equivalents dynamically. Such relevance balancing problem is clearly beyond the task of data acquisition. In my dissertation I examine corpus-based evidence for this problem and ways of addressing it with Information Extraction technology.

1.4. Information Extraction and MT

> The meaning that a word, a phrase, or a sentence conveys is determined not just by itself, but by other parts of the text, both preceding and following… The meaning of a text as a whole is not determined by the words, phrases and sentences that make it up, but by the situation in which it is used". (M.Kay et. al.: 1994: 11)

Information Extraction is a technology for "…automatically extracting pre-specified sorts of information from short, natural language texts" (Gaizauskas and Wilks, 1998: 17), and includes such sub-tasks as Named Entity recognition and classification, Template element filling, Scenario template filling, (as defined for MUC-6) etc. (Gaizauskas and Wilks, 1998: 29-30) (Grishman, 2003). IE can be viewed as a step-wise normalisation of unrestricted text, usually based on cascaded shallow analysis of functions performed by its elements; this includes part-of-speech tagging, shallow parsing, identification of special phrases (e.g., date and time expressions, numbers, proper names) and general features of a text, like names of the described events, their structure and relations, co-reference between phrases describing participants of these events, etc. IE can be referred to as an intermediate-level technology, since it is used for a variety of higher-level tasks, such as Text

Mining, Question Answering, structuring information for Information Retrieval, engineering of ontologies, etc.

Can IE be useful for MT? Even though MT and IE have different tasks, IE encouraged NLP researchers to move from small-scale systems and artificial examples to real natural language data used on a large scale (Cowie and Lehnert, 1996). Therefore IE development experience could be useful for a parallel tendency in MT, which also evolved from experimental to large-scale systems.

There are at least two major aspects of potential interaction between these two technologies: IE can support MT on the data side and on the processing side. On the data side IE shares with MT several initial stages of the ST analysis – part-of-speech tagging, anaphora resolution, word-sense disambiguation. The experience in development, integration and testing these modules could be useful for MT. For instance, IE community has developed an efficient evaluation framework during MUC competitions, where each individual stage of IE was systematically benchmarked. This experience could be useful for quality assurance in MT development – for evaluating each individual stage of the ST analysis. The IE evaluation framework allows developers to compare different modules which perform similar tasks and to choose the best one. In this way IE modules themselves could be used in MT, even if they haven't been initially developed for MT. Some IE modules now can perform clearly defined tasks which provide additional data that could be used by state-of-the-art or future MT systems, e.g., term extraction, detection of phrase boundaries or even populating an ontology.

On the processing side interaction between IE and MT is even more interesting. Firstly, a classical definition of IE goals is "to find and link *relevant* information from NL text ignoring irrelevant information" (Neuman and Xu, 2004). This goal connects with the task to fill the processing gap in MT technology discussed earlier – to assess and compare relevance of translation equivalents across different levels.

IE specifications could be defined by users: the technology provides a general framework for identifying interesting information, writing rules, supervised and unsupervised learning of annotation patterns, but it leaves the possibility for the user to describe what information is "interesting" and should be extracted (Gaizauskas and Wilks, 1998: 49), (Wilks and Catizone, 1999). If we can define a model or some approximation for measuring relative relevance of translation equivalents (something similar to computing tf.idf scores), then user-defined IE systems can meet the demands of MT by providing necessary processing components.

Secondly, Template Element filling and Scenario Template filling tasks could be useful for domain-specific MT in domains typically processed by IE systems, such as newswires or football match reports. Note that templates are usually filled on the basis of information expressed by the whole text, i.e., information which spreads across several sentences. It could be possible to ensure consistency of translation of individual sentences by comparing templates filled from the original text and MT output, to resolve ambiguities on the level of individual sentences by boosting relevance of coherent translation equivalents which fit into the whole picture suggested by the text-level template, to avoid some obvious mistranslations and contrary-to-the-fact translations, etc. Further I give examples, where translation could be fixed with IE-guided processing component linked to MT architecture. English sentence (14) is taken from a paper on IE (Hobbs et al., 1997) and was translated into Russian by one of the best commercial MT systems:

(14) *Salvadoran President-elect Alfredo Christiani condemned the* ***terrorist killing*** *of Attorney General Roberto Garcia Alvarado.*

(15) *Сальвадорский Избранный президент Алфредо Христиани осудил* ***убийство террориста*** *Генерального прокурора Роберто Garcia Alvarado.*

*(*Lit.*: Salvadoran elected president Alfredo Christiani condemned the* ***killing of a terrorist*** *Attorney General Roberto Garcia Alvarado)*

Sentence (15) suggests that the Attorney General was a terrorist himself, not that he was killed by terrorists. Here again we have a competition of translation equivalents and a wrong equivalent gets through: *terrorist killing = killing of a terrorist* (presumably, by analogy to "*tourist killing*" or "*farmer killing*")*;* not *killing by terrorists*. From a compositional point of view such incorrect Russian translation of "*terrorist killing*" is perfect: it preserves semantically interpretable feature – the number of a noun "*terrorist*" and the default way of translating attributive use of English noun phrases into Russian (NP_1+NP_2 → $NP_{2(nom.)}$+$NP_{1(gen.)}$.). Here such compositional account is less relevant.

At the first site the problem requires an introduction of a non-compositional translation equivalent into the system or developing something like preference semantics for MT (Wilks, 1975), (Fass and Wilks, 1983), which could express something like: "terrorists more often kill people, but farmers and tourists are more likely to become victims than perpetrators".

Note, however, that the phrase "*terrorist killing*" can still be used compositionally (possibly with a different phrase structure), e.g.: "*…just pretending to be a terrorist killing war machine…*", "*... who is working for the police on a terrorist killing mission…*", "*...merged into the "TKA" (Terrorist Killing Agency),*

they would … proceed to wherever terrorists operate and kill them…", so setting a higher priority for the non-compositional equivalent would be noisy and will not solve the problem – the cases like these would be difficult for the preference semantics approach.

In the general case relative relevance of translation equivalents couldn't be determined on the level of those equivalents themselves and requires processing on a higher level. (A similar observation for semantics of language units was made in (Kay et al, 1994: 12)). In our example a systematic solution to interpretation of the phrase in question requires processing the information on the text level, which could be done by an IE system. Note, for example, that correct IE template, like (16) could be extracted from sentence (14), it couldn't be inferred from Russian sentence (15):

(16) … `Perpetrator`: ***terrorist***

`Human target:` ***Attorney General Roberto Garcia Alvarado***…

A template, which could be filled a Russian IE system run on incorrect MT output (15), would look like (17):

(17) … `Perpetrator`: [UNKNOWN]

`Human target:`

террорист Генеральный прокурор Роберто Garcia Alvarado

Lit: *terrorist Attorney General Roberto Garcia Alvarado*

It is easier to spot differences between structured templates (16) and (17), than between unstructured texts in different languages: the fact that one slot in (17) is empty and the other contains additional material could be noticed automatically even without translation of template elements, so an MT system could be alerted. If a system has several variants for translating the same segment, the one which ensures the best match in IE templates could be selected. In this way most relevant translation equivalents could be identified and their consistency with the general text meaning could be checked.

It is not necessarily the case that MUC-type IE systems will fill templates correctly in all cases and that in reality IE-guided MT will be able to use correct IE annotation efficiently in all cases for text-level disambiguation, performance figures will certainly be lower than 100%. Here we are talking about the difference of main principles in data processing in modern MT (equivalent-based approaches) and in proposed extension to MT (IE-guided approach). The main point is that equivalent based approaches are inherently limited by lack of flexibility in applying databases of equivalents (no matter, how large the databases could become), while IE guided MT paves the way to flexibly changing the order of application and even changing

the set of conflicting equivalents applied to the same fragment, depending on relative relevance of units in the ST, information coming from text-level templates (a kind of cross-sentence consistency check), etc., which enables MT to go beyond the data processing bottleneck of the "equivalent-based" approaches. The improvements in real performance will be dependent on future improvements of IE technology on template-filling tasks.

A legitimate objection to rise is that it may be not be necessary to complicate MT algorithms with IE processing, since for all disambiguation tasks it is possible to provide correct translations using only existing equivalent-based MT techniques. In fact it was pointed out that another English-Russian MT system (ProMT) correctly translates the example (14) into Russian:

(18a) Сальвадорский Избранный президент Альфредо Христиани осудил террористическое убийство Генерального прокурора Роберто Гарси Альварадо.

However, the argument in this thesis is not about individual examples. Indeed any of them can be correctly translated with equivalent-based MT techniques. My argument is about conflicting translation equivalents, which could fire on the same segments, i.e., about situations where exactly the same segment needs to be translated differently depending on dynamically changing relevance of its units within a larger context. Equivalent-based MT systems don't have capabilities for changing fixed order of equivalent application, since they don't have appropriate models for any possible "reasons" to do so. Such models can be provided only by intelligent data processing techniques, including IE template filling.

If the precedence of translation equivalents in any particular example is "guessed" correctly by an equivalent-based MT, any alternative way of prioritising equivalents will almost certainly be blocked (even if they are present in the database), by a kind of "ripple effect", so the contexts where this alternative order of relevance should be used will almost certainly be wrong. In other words, nothing can be inferred about the ability of MT architecture to deal with this kind of dynamic relevance problems from individual examples – whenever they are translated correctly or not. What really matters is system's ability to deal with all possible ways of prioritising application of conflicting equivalents for any given fragment, i.e., system's performance on the whole sets of examples, where the same fragments have to be translated differently depending on general consistency of information across different levels in text. For example, translations of other contexts for "terrorist killing" by the English-Russian system ProMT confirms this: even though the order of equivalent application is guessed correctly for (14), the

system doesn't change it for other contexts, failing to translate contexts which require alternative ordering:

(18b) I am just pretending to be a terrorist killing war machine.

Я только симулирую быть террористом, убивающим военную машину.

Lit.: I just simulate to be terrorist, killing a military machine.

(18c) Who is working for the police on a terrorist killing mission?

Кто работает для полиции на террористе, убивающем миссию?

Lit.: Who works for police on a terrorist, killing the mission?

It is possible that some other equivalent-based MT system guesses contexts (18b) and (18c) correctly, but will fail on (14). The problem is really the lack of flexibility to counter such "ripple effects" in equivalent databases, but not the lack of ability to translate individual examples.

Consistency check between ST and TT templates could also spot contrary-to-the-fact translations, like sentence (19), taken from the domain of football match reports.

(19)*Swedish playmaker scored a hat-trick in the 4-2 defeat of Heusden-Zolder*.

English-into-Russian MT:

(20) *Шведский плеймейкер выиграл хет-трик в этом поражении 4-2 Heusden-Zolder*. (lit.: Swedish playmaker won hat-trick in this defeat 4-2 Heusden-Zolder).

The name of a team – "Heusden-Zolder" isn't transliterated, so it cannot have a necessary morphological marker of the genitive case. This fact and also an unusual position of this Named Entity after the score for the game (the score in Russian cannot be used attributively) – impede its integration into the general syntactic structure of a sentence. With "Heusden-Zolder" left out, sentence (20) is contrary to the fact: the side of the Swedish playmaker wasn't defeated, in fact it won the game. The mistranslation is caused by enantiosemy of the noun "*defeat*" (Novikov, 1989: 229) which could be used with two opposite meanings in English: "*X's defeat*" means that 'X lost', "*X's defeat of Y*" means that 'X won'.

Yet again, correct interpretation could be enforced by a text-level template, which "knows" who won on the basis of processing the whole article, or even multiple knowledge sources describing the same event, as described e.g., (Saggion

et al., 2004). For sentence (19) mistranslation could be avoided by an *antonymic translation* (Shveitser 1988:141), which uses the word "*победа*" ('victory'): *"... в этой победе над ...*" ('…in this victory over…'). Such departure from the default translation "*defeat → поражение (loss)*" to "*defeat → победа (victory)*" is quite unsafe for an MT system; it really needs sound motivation for applying such radical translation transformation. Motivation on the level of translation equivalents themselves would be insufficient and still risky: cf.:

(21) "*its defeat of last night;*

their FA Cup defeat of last season;

their defeat of last season's Cup winners;

last season's defeat of Durham".

For addressing this problem an integrated IE module can check consistency between a global template, extracted from the ST, and different variants of translation.

Suggested methods for IE-guided MT make use of the fact that imperfect MT output often destroys necessary conditions for identification of relevant information by an IE system. It interferes with rules for filling scenario templates, finding Template Elements, Named Entities, co-reference relations. Firstly, these rules are written or trained on natural language texts produced by native speakers, so any ***fluency*** errors in MT could be punished – the relevant rules will not fire. Secondly, the relevant factual content which is extracted from the ST may no longer be present in the TT, so any ***adequacy*** errors which concerns relevant information will certainly be punished by an IE check-up. Therefore, IE is a powerful tool for spotting inconsistencies in restricted-domain MT.

Experiments presented in the remaining chapters of the dissertation illustrate how IE opens possibilities to spot inconsistencies in MT output and can improve the quality of MT. This suggests a novel perspective for thinking about IE technology. Besides important practical applications, IE also has a deeper theoretical significance: it cannot be viewed only as a shallow ad-hoc substitute for full-scale natural language understanding. Instead it touches some fundamental cognitive processes which are essential components of human understanding and other cognitive procedures involved in human translation, such as the ability to rank the relevance of information in the ST and to check consistency of translated information in the TT on the global level beyond individual sentences and to motivate appropriate translation transformation.

A more coherent viewpoint for IE would be that it establishes a link between language, knowledge and the structure of a domain (Neuman and Xu, 2004) – the

link, which is still beyond the capabilities of modern large-scale commercial MT systems. This gives an explanation for the potential of IE to check the consistency and to improve the quality of MT. Experiments on assessing the impact of IE on MT quality, presented further in this thesis, identify and address some fundamental limits of the state-of-the-art MT, which so far don't receive sufficient attention in the literature. These limitations come on the processing side of MT and go beyond data acquisition problems. In particular, IE can successfully deal with a problem which could be called "*a generalised Wilks's limit*" on MT. This limit is related to a suggestion made in (Wilks, 1994: 113) that the quality of statistical MT is inherently limited by redundancy of natural languages.

This suggestion can be generalised for other equivalent-oriented MT architectures (data-driven and rule-based): the quality of MT architectures that use lookup of translation equivalents as a single processing strategy is inherently limited by the level of information redundancy from the translation point of view: not all information in text is equally relevant and intended for translation. Serious errors in MT will be inevitable (no mater how large is a dictionary or training corpora for SMT or EBMT system) unless the relevance of equivalents is ranked and the number of cases when some of the equivalents "jump the relevance queue" is minimised.

There are two aspects of the *generalised Wilks's limit* on MT. On the one hand, less relevant (in many cases – redundant) information is always present in the ST and it needs to be ranked and filtered out before the transfer stage. IE which operates on the ST is capable of approximating relevance ranking of equivalents by rule-based approaches in some limited domain, or by some statistical measures such as tf.idf scores, that can operate more or less domain-independently. In this way it can support transfer operations on the level of individual sentences by establishing a dynamic relevance threshold for identified ST units, where a model of generation will give priority either to transfer rules or to "the smoothest collocational choices" (Wilks, 2003: 203).

On the other hand, TT should also contain some less relevant information which hasn't been present in the ST. This is related to widely-known phenomenon that the information in the ST is often insufficient for generating the TT. For domain-specific MT the IE templates can support transfer operations on the macro level by supplying part of such missing information, which is the result of processing the whole text (e.g., what are agents and patients of the identified events, etc.) or identify gaps in templates filled from the TT (as shown in example 17 above).

The experiments presented in this thesis illustrate how IE can identify such problems and successfully deal with them. The results indicate what the potential effect of IE integration into MT architecture might be. However, the extent of my experiments was limited by non-availability of the source code for the state-of-the-art commercial MT systems and for template-filling IE modules in public domain. The experiments were carried out using only publicly available tools, such as open-source IE modules for Named Entity (NE) recognition, the BLEU MT evaluation script, and control mechanisms in commercial MT systems open to end-users, such as do-not-translate lists and user dictionaries. This fact restricted the number of IE techniques whose impact on MT was actually evaluated: in particular we tested the impact of NE recognition on improving and evaluating MT and the effect of ranking relative relevance of translation equivalents with statistical salience scores for MT evaluation. Evaluating the exact impact of the other IE procedures, such as template element filling, scenario template filling and co-reference resolution will be possible when greater control over the processing side of large-scale MT systems becomes available. Nevertheless, the results of the evaluated aspects of IE are consistent with the general idea that IE is capable of overcoming some aspects of data-processing bottleneck in MT, in particular – it may successfully deal with many phenomena related to the generalised Wilks's limit on state-of-the-art MT architectures.

The remainder of the thesis is organised as follows: Part II deals with the problems of the role of Named Entity Recognition for MT. Chapter 2 discusses performance-based methods of MT evaluation, in particular – evaluating MT by running IE on degraded MT output. Chapter 3 presents the results of improving morphosyntactic and lexical quality in MT with NE recognition. Some problematic cases are pointed out, where IE has to meet the demands of MT for annotation of translation strategies. Part III examines the use of statistical IE-oriented techniques for improving the accuracy of MT evaluation. Chapter 4 presents the results of extending reference-proximity MT evaluation metric with salience scores. Chapter 5 describes the experiments on extending flexibility of these metrics for several related MT evaluation and MT development tasks. The main experimental results and implications for future work are discussed in the Conclusion section.

Part II
Evaluating and improving MT quality with Named Entity recognition

The two chapters in Part II describe the use of a particular IE sub-task – Named Entity recognition – for MT. This sub-task is particularly important, because unlike template-element filling and scenario template filling tasks, it is domain-independent, and its performance is much higher, so current NE recognition technology is ready for the real-world MT market. The performance of automatic IE systems on other tasks is still significantly lower than 85-95% figures on Precision and Recall achieved by NE recognition modules. Making template-element filling and scenario-template filling modules domain-independent is still an experimental issue (Etzioni et al., 2004). Finally, only NE recognition modules are available open-source in the time of writing. Therefore, even though all aspects of IE are potentially useful for MT, at present only the effects of NER, as the most reliable technique, can be reliably evaluated. However, the results of the presented experiments with NER indicate the extent of possible improvement of MT with other techniques as well, when these techniques become more reliable and domain-independent. This suggestion is also supported by further experiments (presented in Part III) on other domain-independent techniques used in IE, such as calculating salience weights for terms in text.

Chapter 2
IE for Performance-based methods of MT evaluation

In this thesis the term *MT evaluation* is used in the narrow the sense, it means "text quality evaluation", although this is only one of the possible aspects, e.g., identified in (Hutchins and Somers, 1992: 161-174). Here I am not dealing with the problems of evaluating extendibility, operational capabilities or the efficiency of use of MT systems, and concentrate only on two parameters of text quality evaluation: adequacy (fidelity) and fluency (intelligibility, clarity).

MT evaluation can be done by human judges, as described in (White et al., 1994) or using automated methods. Automated scores for MT evaluation are expected to correlate with these intuitive human judgements on the same texts.

We can identify two major groups of automated MT evaluation methods – *performance-based* and *reference proximity* methods. The *performance-based* MT evaluation adopts a "pragmatic" approach to MT, which is similar to human evaluation from a pragmatic point of view, as described in (Hutchins & Somers, 1992: 163): "… can someone using the translation carry out the instructions as well as someone using the original?" The difference with human evaluation is that the tasks are carried out by some automated system, e.g., a parser (Rajman and Hartley, 2001), a grammar correction system, an IE system which performs scenario template filling, co-reference resolution or NE recognition tasks. Parameters of the system's performance may be automatically computed, e.g., the average depth of syntactic trees, the number of syntactic relations of a particular kind, the number of extracted NEs, the ratio of filled template fields.

An advantage of the performance-based methods is that they do not require a human reference translation to compute the scores. Their disadvantage is that there is a potential mismatch between system's performance on some task and other aspects of MT quality, so the performance doesn't necessarily reflect what is considered to be "translation quality" by human evaluators. In general, performance-based methods are built on some prior assumption about the properties of natural language, e.g., that sentence structure should be always connected, that automatic tools built for the analysis of human texts will encounter difficulties in processing computer-generated texts, proportional to the relative amount of quality "degradation" in MT output. Therefore, one needs to be careful and explicit about these prior assumptions, because something which is a bad output for human users may be fine for an automatic NLP system on some task. Making the assumptions

explicit can justify the choice of the performance parameters which have the best correlation with human intuitions about MT quality. For example, in the experiment with a dependency parser (Rajman and Hartley, 2001) the best correlation for *fluency* was found for X-scores, which are computed as (#RELSUBJ + #RELSUBJPASS –#PADJ – #ADVADJ). Note that the parameters which have "plus" sign in this formula are high-level dependencies – subjects of active and passive relative clauses found within the main clause. On the other hand the parameters with the "minus" sign are low-level dependencies – attributive adjectives and adverbs, which are usually found between adjacent words. Therefore the assumption behind the X-score is that low-quality MT usually misses high-level dependencies, these dependencies are much harder to produce spuriously and in the degraded MT output they are usually under-generated; on the other hand, low-quality MT gives rise to spurious low-level dependencies, so there is a negative correlation between their number and MT fluency, such dependencies are usually over-generated.

The second group of MT evaluation methods are reference proximity approaches. These approaches are based on the assumption of reference proximity (ARP), made in (Papineni et al., 2002: 311): "…the closer the machine translation is to a professional human translation, the better it is". The way of computing such closeness can be different: it may be computed as a minimal edit distance between the two texts (Akiba et al., 2001), as modified precision of balanced N-gram matches between the two texts (Papineni et al., 2002), etc.

IE techniques can be used with both groups of MT evaluation methods. Chapter 2 describes experiments on performance-based MT evaluation with IE via measuring the performance of a NE recognition module on degraded MT output of different quality and on assessing usefulness of MT output for IE tasks. Chapter 4 in Part III will describe experiments on modifying the concept of reference proximity with statistical IE techniques via assigning salience weights to matched terms.

2.1. Comparative Evaluation of Automatic Named Entity Recognition from Machine Translation Output

This section reports on the results of an experiment on automatic NE recognition from Machine Translations produced by five different MT systems. NE annotations are compared with the results obtained from two high-quality human translations. The experiment shows that for recognition of a large class of NEs (Person Names, Locations, Dates, etc.) MT output is almost as useful as a human

translation. For other types of NEs (Organisation Names) Precision figures are close to the results for human annotation, although Recall is seriously distorted by the degraded quality of MT. The success rate of NE recognition doesn't strongly correlate with human or automatic MT evaluation scores, which suggests that the quality criteria needed for measuring MT usability for dissemination purposes are not pertinent for assimilation tasks such as Information Extraction (Babych and Hartley, 2004d).

2.1.1.Dissemination *vs* assimilation

Since the 1960's the 'Holy Grail' of Machine Translation technology has been Fully Automatic High Quality Translation, which aims at creating accurate and fluent texts in a target language suitable for dissemination (i.e. publication) purposes – a goal which has yet to be achieved.

However, there are successful attempts and suggestions to use 'crummy' MT output (Church and Hovy, 1993) for assimilation (i.e. comprehension) tasks: text classification, relevance rating, information extraction (White et al., 2000), for NLP tasks such as Cross-Language Information Retrieval (Gachot et al., 1998), and Multilingual Question Answering (a new task set up for CLEF 2003).

Multilingual Information Extraction is one such assimilation task and consequently an area where imperfect MT output is potentially useful. On the one hand MT can extend the reach of existing monolingual IE systems by translating a text before running IE; on the other hand, results of IE (identified Named Entities, template elements or scenario templates) can be translated into a foreign language after IE processing (Wilks, 1997: 7-8). The first scenario is more demanding for MT, because the performance of an automatic IE system may be influenced by MT quality.

There is an open question: Which aspects of MT quality are important for different IE tasks and may substantially influence the performance of IE?

MT quality is often benchmarked from the viewpoint of human users (White et al., 1994), focusing still on the goal of FAHQT for dissemination. As a result, automatic evaluation scores, such as BLEU (Papineni et al., 2002), are validated according to how well they correlate with human intuitive judgements of translation quality. Using edit distances between MT output and a human reference translation to evaluate MT (Akiba et al., 2001) also makes an implicit assumption that MT should be suitable for dissemination purposes.

However, MT has created its own demand precisely in the area where otherwise there would be no translation at all. Where it is primarily used for

assimilation purposes, the evaluation of NLP performance on MT output might give a better indication of its usefulness than dissemination criteria. Therefore there is a need for: (1) systematically benchmarking NLP technologies, such as IE (and its sub-tasks, e.g., NE recognition), on MT output; (2) developing and calibrating automatic MT evaluation scores for these *primary* uses of 'crummy' MT; (3) assessing quantitatively the extent to which certain human and automatic MT evaluation scores predict the performance of automatic systems on different NLP tasks.

2.1.2. Set-up of the experiment

We addressed some of the above issues by conducting a comparative evaluation of the performance of the ANNIE NE recognition module of Sheffield's GATE IE system (Gaizauskas et al., 1995; Cunningham et al., 1996, 2002). We used the DARPA-94 corpus of French-English MT and human translations (White et al., 1994). The MT systems were Candide, Globalink, Metal, and Systran (participants in DARPA), plus Reverso. Specifically, we focused on whether there is a significant divergence between NE recognition performance and the results of human and automatic evaluation of the MT systems. This indicates to what extent MT quality criteria may differ for human use and for the needs of NLP systems.

In the first stage NEs were annotated in translations of 100 news reports (each text is about 350 words), produced by each MT system.

NEs were also annotated in the two independent human translations of the same 100 texts: the Reference and the Expert translations.

Comparative evaluation of this NE annotation is different from standard evaluation procedure for NE recognition in two respects. The first difference is that in our experiment there is no gold standard NE annotation for any of the human translations or MT outputs. The second difference is that the annotated text is no longer constant.

2.1.2.1 Absence of a gold standard

As it was mentioned above, seven sets of texts in the DARPA 94 corpus were used: 5 sets of translations produced by different MT systems (Candide, Globalink, Metal, Systran, and Reverso) and 2 sets of human translations (Expert and Reference), each containing 100 texts translated from French.

Since all seven sets of texts are different, it would be too expensive to produce a gold standard annotation for each of them. However, all these texts have the same origin: all are translations of the same collection of French source texts, so it can be expected that there will be a great overlap between extracted NEs, namely for those

typical cases when French NEs have a standard translation into English. While most types of NEs are expected to stay the same across different translations, there is also a need to account for possible variations.

Two main things can go wrong when NEs are extracted from MT output (which is generally regarded to be of lower quality than a human translation):

– NE recognition often relies on certain contextual conditions being met, so if a lexical or morpho-syntactic context is distorted in MT output, NEs will be not extracted, resulting in NE 'undergeneration'; likewise the distorted context may give rise to false NEs, leading to NE 'overgeneration'.

– If NEs are wrongly translated despite the context meeting the requirements of the NE recognition system, they are of no use in any other NLP tasks.

The goal of our comparative evaluation is to estimate to what extent the output of different MT systems and the alternative human translation are 'robust' against these two pitfalls, i.e., to what extent they may be useful for the IE purposes. This means that we are less interested in absolute performance figures for the NE recognition system, than in the comparison between its runs on the output of different MT systems.

Furthermore, the accuracy scores for leading NE recognition systems are relatively high. The default settings of ANNIE NE modules produce between 80-90% Precision & Recall on news texts originally written in English (Cunningham et al., 2002). We assume that for comparable texts – human translations of news reports into English – NE recognition performance is similar.

Therefore, for our purposes it is possible to use the NE annotation in one of the human translations as a reference, which will serve as a 'silver standard' for benchmarking NE recognition performance from 'low quality' MT texts. The baseline for such comparisons will be the NE annotation in the other human translation: it will indicate what difference in accuracy may be expected if an alternative high-quality translation is used. This allows us to: (1) estimate the relative performance of the NE recognition system on texts with variable quality; (2) compare these relative figures with human and automatic MT evaluation scores; (3) answer the question whether usefulness of MT for IE should be characterised by criteria other than Adequacy and Fluency, or whether these correctly predict the potential performance of NE recognition.

2.1.2.2 Legitimate variation in translation

MT output and human translations available in DARPA 94 corpus were annotated with NEs. As it was mentioned above, each collection of texts has the

same origin – the same set of French text, but is different from other collections since it was generated by a different MT system or a human translator. This situation differs from standard set-up for NER evaluation experiments, where evaluated set of texts should be constant, and we compare annotation of these texts produced by different NER systems. In our research we compare annotation sets produced from different texts: annotated texts are no longer constant. On the contrary, we don't use different NER systems for annotation: the same system produces all compared sets of NEs.

This requires a different interpretation of the figures for Precision, Recall and F-score: strictly speaking they only characterise *differences* rather than the degree of *perfection*. Annotation mismatches do not necessarily mean deterioration; they may be also due to the improved performance of NE recognition on the test file, or due to choosing a legitimate alternative translation.

For example, we expect that NEs normally have a standard translation and will not vary across different human translations; therefore the quality of MT systems depends on how well this standard is followed. The only exceptions to this rule should be less well known organisations which do not have an established translation. But surprisingly, some degree of legitimate variation was found in human translations for well-known institutions also:

ORI: *De son côté, le département d'Etat américain, dans un communiqué, a déclaré: 'Nous ne comprenons pas la décision' de Paris.*

HT-Expert: *For its part, the* <Organization> *American Department of State* </Organization> *said in a communique that 'We do not understand the decision' made by* <Location> *Paris* </Location>.

HT-Reference: *For its part, the* <Organization> *American State Department* </Organization> *stated in a press release: We do not understand the decision of* <Location> *Paris* </Location>.

MT-Systran: *On its side, the* <Organization> *American State Department* </Organization>, *in an official statement, declared: 'We do not include/understand the decision' of* <Location> *Paris* </Location>.

This indicates the need to identify classes of NEs which may undergo legitimate translation variation, similarly to other words or phrases in language, and to account for the legitimate translation variation in our experiment.

2.1.2.3 Evaluation parameters and procedure

Note that the results of NE annotation from human translations and from MT output are rather distant for some types of NEs. For the two human translations there is a norm for the number of differences in annotation, but MT output sometimes goes far beyond this norm. This suggests that the extent of deviation from these baseline norms characterises the usefulness of MT systems for IE. Parameters can be computed which are interpretable from this point of view and account for the problems of the standard accuracy measures.

– *Counts of different types of annotated NEs*

This parameter is very robust against legitimate translation variation. It shows in how many cases *any* NE has been identified in a particular context, i.e., whether MT output preserves the contextual conditions for identifying an NE. On the other hand, this parameter does not take into account cases where conditions for identification of new (either spurious or genuine) NEs are created in tested MT output. It characterises only the 'upper bound' of cases where conditions for identification of an NE have been met.

– *Precision on the* union *of NE annotations for two human translations*

This parameter is sensitive to legitimate translation variation: it rewards annotations that match at least one of the alternatives found in two independent human translations. If no match is found for a particular NE, the case is treated as 'over-generation'. For a given MT output, this parameter shows how successfully over-generation of NEs may be avoided. This parameter uses a similar approach to the BLEU method for MT evaluation, which computes precision on the union of n-gram units from several human translations.

– *Recall on the intersection of NE annotations for two human translations*

This parameter rewards annotations that match a set of NEs, which are constant across different human translations. The intuition is that, if a given NE is present in both human translations, it is very likely to have some 'standard', obligatory translation, which it is necessary to preserve in MT. Such NE needs to be extracted exactly in the form used in both human translations. This parameter shows how successfully 'under-generation' for the set of most 'standard', uniformly translated NEs has been avoided.

In order to determine whether any human or automatic MT evaluation scores could predict the performance of NE recognition on MT output, the performance figures for correlation with each of the evaluation scores were tested, as follows.

The corpus was divided into 10 chunks each containing 10 texts. Human evaluation scores for Adequacy, Fluency and Informativeness are available for each machine-translated text in the DARPA corpus (not including Reverso, therefore). Automatic BLEU scores for each text were also generated. Average scores for chunks of 10 texts in the corpus were computed. The resulting sets of scores contained 40 samples each (10 for each MT system).

For corresponding sets we computed Pearson's correlation coefficient *r*. The statistical significance of this correlation was tested using *t* distribution.

2.1.3. Results of NE recognition on MT output

The counts of extracted NEs are summarised in Table 1 and Figure 1. It can be seen that for Organisation Names there is a significant difference in the number of extracted NEs for texts produced by humans and for MT output: many more Organisation Names are extracted from human translations. The difference is much smaller for Titles, but the tendency is similar. However, this tendency reverses for Job Titles: MT output tends to give rise to a greater number of this type of NEs [2].

Results for other types of NEs for human translations and MT output come very close together. This gives an indication that distorted MT quality seriously affects the results of NE recognition for a specific type of Named Entities – Organisation Names, which are more context-dependent and less distinguishable from other types of words than other NE types. The latter may have some explicit mark-up or clearly defined boundaries, so they tend to be less affected by MT and may be extracted from MT output more successfully.

---

[2] These clashing tendencies for Titles and Job Titles have a simple technical explanation: cases where human translators capitalise the initial letter (e.g. 'Colonel') normally go into the Title category; where MT renders them in lower case (e.g. 'colonel'), they are often annotated as Job Titles. The category {Job}Title joins these two categories and shows no significant differences between human translations and MT output.

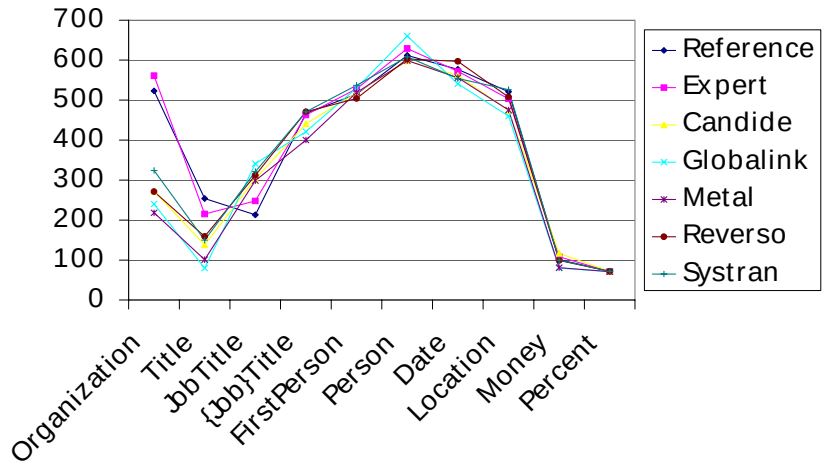


Figure 1. Number of extracted NEs

	Reference HT	Expert HT	Candide MT	Globalink MT	Metal MT	Reverso MT	Systran MT
Paragraph	*826*	*802*	*813*	*804*	*805*	*798*	*804*
Organization	*523*	*561*	*272*	*240*	*218*	*271*	*324*
Title	*254*	*215*	*138*	*80*	*101*	*159*	*150*
Job-Title	*213*	*248*	*303*	*341*	*299*	*312*	*321*
{Job}Title	*467*	*463*	*441*	*421*	*400*	*471*	*471*
First Pers.	*515*	*528*	*518*	*530*	*519*	*504*	*537*
Per-son	*612*	*629*	*598*	*660*	*599*	*603*	*608*
Date	*577*	*572*	*562*	*541*	*556*	*597*	*554*
Location	*521*	*503*	*474*	*460*	*475*	*508*	*526*
Money	*101*	*108*	*117*	*80*	*81*	*99*	*100*
Percent	*72*	*71*	*72*	*71*	*71*	*72*	*72*

Table 1. Number of extracted NEs

The Precision (P) and Recall (R) figures for each type of NE are summarised in the following tables and figures.

2.1.3.1 Organisation names

Figures for Organisation Names, taking as reference the NE annotations from the Reference and the Expert human translations and the union / intersection of these sets, are presented in Table 2 and Figure 2.

In all cases scores for annotations of human translations are the highest, but the contrast between human translations and MT is the largest for Recall, and there is a very little difference for Precision. The improvement in Precision when the union of both translations is used as a reference is very moderate. The improvement in Recall when the intersection of the two translations is used is much higher.

	HT-Ref	HT-Exp.	U/I
P.HT-exp.	0.5745	1	1
P.HT-ref	1	0.6172	1
P.candide	0.4924	0.5229	0.5763
P.globalink	0.4979	0.5319	0.5745
P.ms	0.5423	0.5672	0.6070
P.reverso	0.5709	0.5709	0.6552
P.systran	0.5096	0.5223	0.5892
R.HT-exp.	0.6172	1	1
R.HT-ref	1	0.5745	1
R.candide	0.252	0.2491	0.3639
R.globalink	0.2285	0.2273	0.3386
R.ms	0.2129	0.2073	0.3196
R.reverso	0.2910	0.2709	0.4019
R.systran	0.3125	0.2982	0.4399

Table 2. Precision, Recall – Organisations

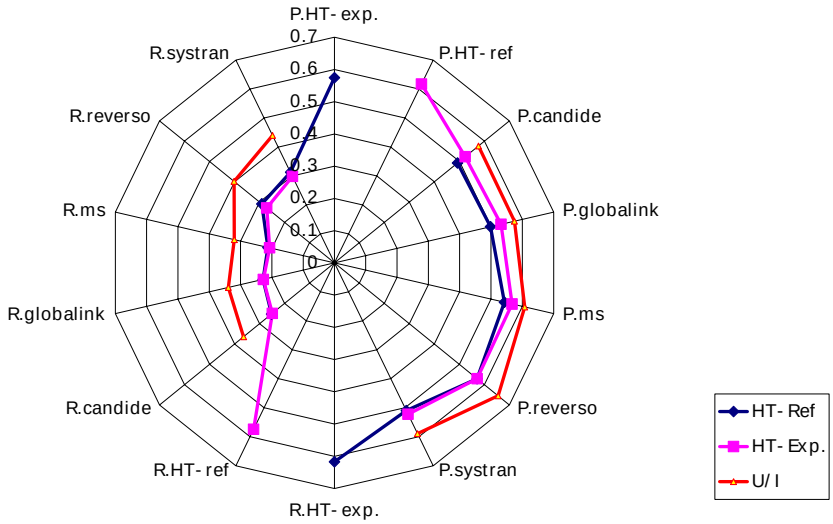


Figure 2. Precision, Recall – Organisations

This shows that in MT output over-generation of Organisation Names is very limited; the main problems are related to under-generation. Recall is the major

aspect affected by the degraded quality of MT output; Precision results for NE recognition from MT output are almost unaffected.

The results demonstrate that Organisation Names constitute a broad and highly dynamic class of NE whose identification is very sensitive to MT quality, hence the low Recall figures. In this typical example, ANNIE fails to identify the string 'Egyptian Diplomacy' in the MT output as an Organisation Name, since this is not an expected way of expressing this concept in English.

ORI: … *le chef de la diplomatie égyptienne*

HT: *the* <Title>Chief</Title> *of the* <Organization>*Egyptian Diplomatic Corps* </Organization>

MT-Systran: *the* <JobTitle> *chief* </JobTitle> *of the Egyptian diplomacy*

Such occurrences are frequent, so generally far fewer Organisation Names are identified in MT output as compared to a human translation.

2.1.3.2 Person names

In general, the accuracy for Person Names (Table 3 and Figure 3) is much higher than for Organisation Names. However, the figures are more dependant on a particular MT system and do not characterise any general tendency for MT output as compared to human translations. The results of leading MT systems for this type of NE are practically undistinguishable from the results obtained from human translations, in terms of both Precision and Recall.

But there is no correlation between human evaluation scores for the performance of an MT system and the accuracy figures for recognition of Person Names. This indicates that recognition of these NEs is not directly influenced by other translation problems, such as ambiguity between Person Names and common nouns (e.g. *Bill Fisher*). These cases are relatively rare and easily identifiable; current MT technology has successfully solved this problem.

	HT-Ref	HT-Exp.	U/I
P.HT-exp.	0.7850	1	1
P.HT-ref	1	0.8056	1
P.candide	0.7525	0.7425	0.8161
P.globalink	0.4932	0.5099	0.5478
P.ms	0.6868	0.6834	0.7437
P.reverso	0.6083	0.6333	0.6783
P.systran	0.7169	0.7318	0.7897
R.HT-exp.	0.8056	1	1
R.HT-ref	1	0.7850	1
R.candide	0.7353	0.7070	0.8235
R.globalink	0.5310	0.5350	0.6085
R.ms	0.6699	0.6497	0.7586
R.reverso	0.5964	0.6051	0.6856
R.systran	0.7075	0.7038	0.8073

Table 3. Precision, Recall – Person

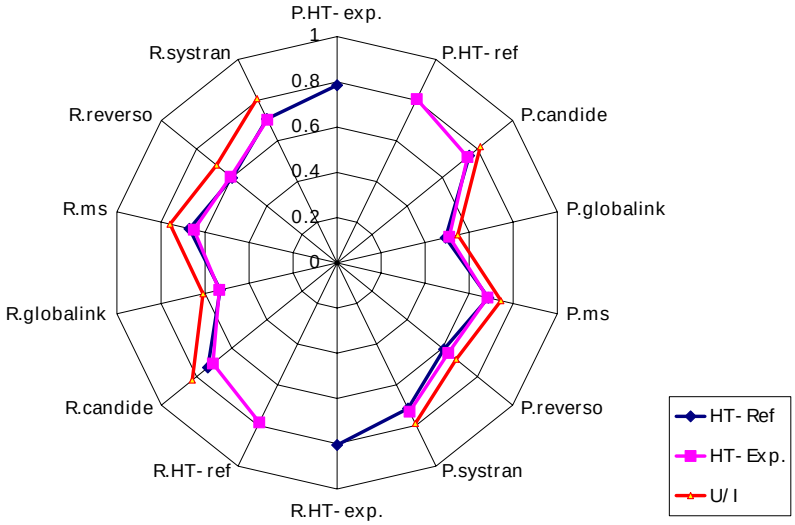


Figure 3. Precision, Recall – Person

2.1.3.3 Location names

The data for Location Names (Table 4 and Figure 4) shows a more even performance across different MT systems as compared to other types. This may be due to the fact that this type of NE is less ambiguous than the other types of NE. Once again, NE recognition from MT output is practically undistinguishable from NE recognition from a human translation, in terms of both Precision and Recall.

	HT-Ref	HT-Exp.	U/I
P.HT-exp.	0.8608	1	1
P.HT-ref	1	0.8311	1
P.candide	0.8481	0.8397	0.8840
P.globalink	0.8196	0.8087	0.8543
P.ms	0.8439	0.8312	0.8861
P.reverso	0.8185	0.8185	0.8679
P.systran	0.8042	0.8061	0.8574
R.HT-exp.	0.8311	1	1
R.HT-ref	1	0.8608	1
R.candide	0.7716	0.7913	0.8799
R.globalink	0.7236	0.7396	0.8222
R.ms	0.7678	0.7833	0.8637
R.reverso	0.7965	0.825	0.9007
R.systran	0.8119	0.8429	0.9145

Table 4. Precision, Recall – Location

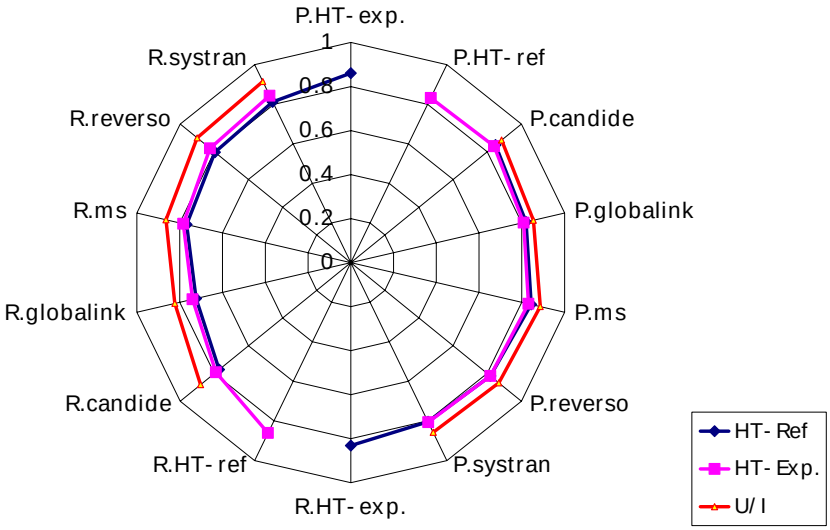


Figure 4. Precision, Recall – Location

2.1.3.4 Overgeneration

The results for other types of NE lie within the range of scores described above. Each type of NE behaves differently across MT systems and human reference translations. Nevertheless, for all these types of NEs, the best MT systems give results comparable with the accuracy of NE annotation from human translations.

Of course, some additional (spurious or genuine) NEs may appear in MT output despite this tendency. The rarity of such an event is because distorted MT output makes it much harder to create new conditions for identifying an NE than to reconstruct necessary conditions similar to those which often are created in an alternative human translation.

Nevertheless, in a few cases genuine new NEs are found in MT output. In these cases, MT is more favourable for IE tasks: it is more literal (therefore less misleading) than human translation, e.g.:

ORI : *Il a été fait chevalier dans l'ordre national du Mérite en mai 1991*

HT: *He was made a Chevalier in the National Order of Merit in May, 1991.*

MT-Systran: *It was made* `<JobTitle>` *knight* `</JobTitle>` *in the national order of the Merit in May 1991.*

MT-Candide: *He was knighted in the national command at Merite in May, 1991.*

The human translator used the borrowed French word *Chevalier*, which 'distracted' ANNIE. Although this translation might be more adequate from the human point of view, it is less useful for the NE recognition system. Interestingly, the more idiomatic translation of this sentence produced by the statistical MT system Candide removed this NE from the sentence. The literal output of rule-based MT systems, such as Systran, proves most favourable for identifying the NE.

For MT output, Recall is highest for 'constant' NEs which have a standard translation (and are identified in both human translations). The figures for Recall correlate highly with the counts of extracted Organisation Names (Table 1): Pearson's correlation coefficient is 0.9561 (there is no correlation for Precision). This confirms the suggestion that over-generation does not affect counts of NEs.

2.1.4. Correlation with MT evaluation scores

The second problem addressed in our experiment is determining whether human or automatic MT evaluation scores correlate with accuracy of NE recognition.

r(38)=	**ADE**	**FLU**	**INF**
ref=Exp ; Precision	-0.0047	-0.0232	0.1100
ref=Exp ; Recall	**0.2558** *p>0.05*	0.0671	-0.0128
ref=Ref ; Precision	0.1887	-0.0994	0.0011
ref=Ref ; Recall	**0.3997** *p<0.01*	0.0084	-0.0633
ref=u(ER); Precision	0.1804	-0.1284	-0.0071
ref=i(ER); Recall	**0.3465** *p<0.05*	-0.1070	-0.0458

Table 5. Pearson's *r* coefficient: (Organisations *vs* Adequacy,Fluency,Informativeness)

Only weak or moderate correlation was found in a number of cases, which suggests that human judgements or automatic MT evaluation scores is not deterministically linked to usefulness of MT output for assimilation purposes, e.g., for NLP tasks such as NE recognition, however, it is interesting to contrast cases where such correlation exists and where it doesn't exist.

For Organisation Names the highest correlation figures were between Recall and human scores for Adequacy – in cases when the Reference human translation or the intersection between the two human annotations were used (Table 5).

This weak correlation is also statistically significant (at the levels $p<0.01$ and $p<0.05$) for the Reference human translation and the intersection between the two references.

Other cases of moderately strong positive correlation were also found, although it is difficult to give linguistically meaningful interpretation to these correlations. This suggests that they may be due to indirect links between the overall quality of an MT system and the attention that particular groups of developers pay to specific NE problems.

The highest positive correlation for human evaluation scores was found between Recall on Date NEs and human scores for Fluency for the case when the Expert human translation was used as a reference:

$r(38)=0.4847, p<0.001; (t=3.8392)$.

With the other human reference translation the correlation is much weaker:

$r(38)=0.3559$; $p<0.05$; $(t=2.6388)$

The highest positive correlation for BLEU scores is close to these figures. Again, it is difficult to give any meaningful linguistic interpretation to this correlation. The correlation of BLEU scores is more consistent across different references, e.g. the correlation between BLEU and Recall figures for Title NEs with Expert and Reference human translations and the intersection of these annotations:

ref=Exp: $r(48)= 0.4844; p<0.001; (t=3.8364)$

ref=Ref: $r(48)= 0.4025; p<0.01; (t=3.0467)$

ref=i(ER): $r(48)= 0.3251; p<0.05; (t=2.3819)$

2.1.5. Conclusions of the experiment

It is possible to conclude that human evaluation scores and automatic BLEU scores do not reliably predict the performance of NE recognition for most of the NE types. Still, in a few cases the Pearson's *r* coefficient is significantly different from

zero, which indicates a positive link between better performance in some aspects of NE recognition (e.g., boosting recall for Organisation Names) and better quality (e.g., higher Adequacy scores) of MT from the point of view of human evaluators. Nevertheless, for many other aspects of NE recognition there is no such link, so the usability of MT output cannot be judged just by intuitive human criteria, which usually assume dissemination purposes for MT output.

Thus the results confirm the idea that the criteria most widely used for assessing MT quality for most NER tasks fail to reflect the needs of subsequent NLP processing in general and of NE recognition in particular; these criteria tend to underestimate the usefulness of MT for automated assimilation tasks, for which MT in most aspects may be as useful as a human translation.

There remain open questions why just for one particular type of NEs – Organisation Names –Recall of NE recognition of the rule-based ANNIE system is substantially distorted by the degraded MT output, and why the Recall weakly correlates with human scores for MT Adequacy. On the one hand, the existence of this link may suggest that the Recall figures give some indirect indication of the translation Adequacy, so technological improvements in recognition of Organisation Names in MT systems will boost translation quality (in eyes of human evaluators) and, therefore, will enhance the usability of MT for dissemination purposes as well.

On the other hand it may be the case that for the other type of NE recognition systems, namely the ones based on statistical Machine Learning, the MT output would appear much more useful (e.g., the distortion of Recall for Organisation Names would be significantly smaller), since the recognition could then adapt to any regular patterns produced by the MT system, even if they differed from the natural form. Whether or not any NE error patterns in MT output could be learnt by statistical IE systems is an interesting problem, which has important implications for MT and IE technology.

Future work in this direction may include research on usability of MT for other IE tasks, such as scenario template filling, co-reference resolution, automatic summarisation, etc. Also the suitability of the MT output for a range of learning IE systems will be investigated, and typologically different language pairs will be involved in the evaluation. Further direction of research may also include moving from the “black-box” to the “glass-box” evaluation and examining the ways in which particular MT systems treat different kinds of NE.

Another direction of research is investigating robustness of the suggested performance-based MT evaluation method (e.g., whether results can be replicated with another NER system built on different principles). At present the hypothesis is

that the method will be sufficiently robust and it will produce comparable results with all NER systems that have comparable performance figures on human translations. This hypothesis is based on observation that degraded MT output in most cases destroys the very basis for identification of Organisation Names – NEs themselves and their natural contexts, not just superficial ad-hoc triggers, on which NER system relies.

This research may lead to theoretical generalisations about the nature of dynamic quality criteria for translation, which correctly predict the usability of human translations and MT for different purposes.

2.2. Statistical modelling of MT output corpora for Information Extraction

The previous experiment demonstrated that the output of state-of-the-art MT systems could be useful for certain NLP tasks, such as IE. However, some unresolved problems in MT technology could seriously limit the usability of such systems. For example robust and accurate word sense disambiguation, which is essential for the performance of IE systems, is not yet achieved by commercial MT applications. In this section we try to develop an evaluation measure for MT systems which could be applicable for a wider variety of problems, not just NE recognition. This evaluation measure is designed to predict possible usability of MT output for some IE tasks, such as scenario template filling, or automatic acquisition of templates from texts. We focus on tf.idf-type scores which measure statistical salience of terms in a given text. This type of scores was developed for Information Retrieval, but different modifications of ft.idf are widely used in statistical Information Extraction, e.g., for automatic acquisition of domain relevant terms and their relations – KF-IDF scores (Xu et al., 2002), for automatic pattern acquisition (Sudo et al., 2001), automatic template creation (Collier, 1996), etc. In this section we propose a variant of statistical salience scores, the S-scores, which were found to have a closer link with translation *adequacy* (the experiment of NE recognition from MT output described in the previous section suggests that MT adequacy has the closest relation to this particular problem of IE). In Chapters 3 and 4 the properties and distribution of the S-score are compared with the standard tf.idf measures, which were found to have a closer link with translation *fluency*.

General importance of the salience scores for IE is also substantiated by the material, where highly salient words often include name entities and other important candidates for filling IE templates. I suggest MT evaluation metrics which are based on comparing the distribution of statistically significant words in corpora of MT

output and in human reference translation corpora. We show that there are substantial differences in such distributions between human translations and MT output, which could seriously distort IE performance. We compare different MT systems with respect to the proposed evaluation measures and look into their relation to other MT evaluation metrics. We also show that the statistical model suggested could highlight specific problems in MT output that are related to conveying factual information. Dealing with such problems systematically could considerably improve the performance of MT systems and their usability for IE tasks (Babych et al., 2003).

2.2.1. Overview of the experiment

Modern commercial MT systems do not yet achieve fully automatic high quality MT, but their output can still be used as input to some NLP tasks, such as IE. IE systems, such as GATE (Cunningham et al., 1996), are mainly used for "scenario template filling": processing texts in a specific subject domain (such as management succession events, satellite launches, or football match reports) and filling a predefined template for each text with strings taken from it. On the one hand, IE systems usually do local analysis of the input text and it is reasonable to assume that they tolerate low scores for MT fluency (besides it is the most difficult aspect to achieve in MT output). But in certain cases mistranslation could inhibit IE performance. In this section we develop an MT evaluation metrics that capture this aspect of MT quality, and relate them to other evaluation measures, such as MT adequacy scores.

On the other hand, some aspects of IE technology impose a specific set of requirements on MT output. These requirements are important for the general performance of IE systems. For example, NEs have to be accurately identified by MT systems: an IE system for Russian will not be able to correctly fill the template if a person name like "Bill Fisher" had been translated from English into Russian as "*выставить счет рыбаку*" ('to send a bill to a fisher'). Moreover, IE requires adequate translation of specific words which are significant for template filling tasks. These words are usually not highly frequent and have a very precise meaning. Therefore it is difficult to substitute such words with synonymous words. For example, the French phrase (1) was translated into English by one of our MT systems:

(1)	French original:	*un montant global de 30 milliards de francs*
	Human translation:	*a total amount of 30 billion francs*
	Machine translation:	*a global 30 billion franc amount*

The correct meaning of the word '*global*' could be guessed by a human post-editor, but the phrase could be misinterpreted by a template-filling module of an IE system, e.g, as an 'amount related to company's global operations', etc. Similarly in the translation of the French sentence (2):

(2)	French original:	*La reprise, de l'ordre de 8%, n'a pas été suffisante pour compenser la chute européenne*.
	Human translation:	*The recovery, about 8%, was not enough to offset the European decline.*
	Machine translation:	*The resumption, of the order of 8 %, was not sufficient to compensate for the European fall.*

The word '*order*' could be misinterpreted by a template-filling IE module as related to ordering of products, but not to uncertainty of information.

Developers of commercial MT systems often do not have sufficient resources to properly disambiguate such words, partly because they rarely occur in corpora that are used for the development and testing of MT systems, and partly because it is difficult to distinguish these problems from other types of issues in MT development. Therefore, it would be useful to have a reliable statistical criterion to highlight MT problems that are related to mismatches in factual information between human translation and MT output. This could be essential for improving the performance of IE systems that run on MT output.

Another important problem for present-day IE research is automatic acquisition of templates, which is aimed to making IE technology more adaptive (Wilks and Catizone, 1999). There have been suggestions to use lexical statistical models of a corpus and a text for IE to automatically acquire templates: statistically significant words (i.e., words in a text that have considerably higher frequencies than expected from their frequencies in a reference corpus) could be found in the text; templates could be built around sentences where these words are used (Collier, 1998).

However, it is not clear whether this method would be effective if applied to a corpus of MT output texts. On the one hand, the output of traditional knowledge-based MT systems produces significantly different statistical models from the models built on "natural" English texts (either original texts or human translations of texts, done by native speakers). In Chapter 1 it has been shown that N-gram precision of MT output text (in relation to a human reference translation), measured by the BLEU score (Papineni et al., 2001), is significantly lower then the N-gram precision of some other human translation (in relation to the same reference). This is due to the fact that translation equivalence in MT output texts is triggered primarily

by source-language structures, not by balancing the adequacy of the target text on the pragmatic level with its fluency, which depends on statistical laws in target language – as is the case for professional human translation. Structures that are treated by knowledge-based MT systems as translation equivalents could have a different distribution in "natural" source and target corpora. As a result, many words that are not statistically significant in "natural" English texts become significant in MT output, and vice versa. Subsequently, different sentences may be selected as candidates for a template pattern based on MT output and one based on human translation.

On the other hand, even if corresponding sentences are selected, the value of template patterns could be diminished by errors in word sense disambiguation, made by MT systems, e.g.:

(3)	French original:	*la reddition des armées allemandes*
	Human translation:	*the surrender of the German armed forces*
	Machine translation:	*the rendering of the German armies*

Words '*surrender*' and '*rendering*' could induce different IE templates, even if corresponding sentences in MT output have been correctly identified as statistically significant. Therefore the requirement of proper word sense disambiguation of statistically significant words is central to usability of MT output corpora for IE tasks.

High quality word sense disambiguation for large vocabulary systems is a complex task, which requires interaction of different knowledge sources and where "best results are to be obtained from optimisation of a combination of types of lexical knowledge" (Stevenson and Wilks, 2001). However, it is also important to find out to what extent the output of different state-of-the-art MT systems is now usable for IE tasks.

In this section we report on the results of an experiment for establishing an evaluation measure for MT systems which contrasts the distribution of statistically significant words in MT output and in human translation and gives an indication of how usable the output of particular MT systems could be for IE tasks. We also discuss linguistic intuitions behind this measure.

2.2.2. Experiment set-up and evaluation metrics

Statistical models were developed for the DARPA94 MT evaluation corpus (White et al., 1994). As it was mentioned before, this corpus contains 100 human reference translations of newspaper articles, alternative human "expert" translations, and the output of 5 French-English MT systems for each of these texts. The length

of each original French text is 300–420 words, with an average length of 370 words. For 4 of these systems scores of "fluency", "adequacy" and "informativeness" are also available.

The following method was used for measuring MT quality for IE tasks.

1. In the first stage a statistical model was developed for the corpus of MT output and for a parallel corpus of human translations. These models highlight statistically significant words for each text in the corpus and give a certain score of statistical significance for each highlighted word.

2. In the second stage these statistical models for MT output and for human translation corpora were compared. In particular,

- 2.a - We established which words in the MT output are "over-generated" – are marked as statistically significant, even though they are absent or not marked as significant in human translation – and what is the overall score of "statistical significance" for such words;

- 2.b – We established which words in MT output are "under-generated" – are absent or not marked as statistically significant, even though they are significant in human translation of the same text – and what is the overall score of "statistical significance" of these words;

- 2.c- We established which words are marked as significant both in MT and human translation, but which have different scores of statistical significance. Then I calculated the overall difference in the score for each pair of texts in the corpora;

- 2.d - We computed 3 measures that characterise differences in statistical models for MT and human translation of each text: a measure of "avoiding over-generation" (which is linked to the standard "precision" measure); a measure of "avoiding under-generation" (which is linked to the "recall" measure); and finally – a combined score based on these two measures (calculated similarly to the F-measure).

- 2.e - We computed the average scores for each MT system.

The scores for each system are different and the significance of the difference can be tested by contrasting it with standard deviation of scores for each of the compared systems (z-test), which will allow us either to accept or to reject the null-hypothesis that the difference is caused by some “noise” in data. This issue is not central for the discussion here, and it is properly addressed in Chapter 5. However, it worth mentioning here that the bigger the evaluated corpus, the smaller the “noise” – standard deviation of the scores for a particular system, and differences in evaluation data become more reliable. It will be shown that for the corpus of the

DARPA size standard deviation of the scores is usually smaller than 0.3%, so in our experiment any reported differences for MT systems which are greater than 1% will be statistically significant with a confidence level over 99.9%. Smaller test sets will introduce more noise and will require the difference to be greater in order to draw any conclusions about relative quality of the compared MT systems.

Besides general scores of translation quality, this method allows us to automatically generate lists of statistically significant words which have a problematic translation in MT output. Such lists could be directly useful for MT development and tuning MT systems for a particular subject domain. Further we present formulae used to compute the scores and illustrate this process with examples from the DARPA94 corpus.

It is possible to compute salience scores as standard tf.idf scores:

$$tf.idf_{(i,j)} = (1 + \log(tf_{i,j}))\log(N / df_i); and (tf_{i,j} \geq 1),$$

where:

$tf_{i,j}$ is the number of occurrences of the word w_i in the document d_j;

df_i is the number of documents in the corpus where the word w_i occurs;

N is the total number of documents in the corpus.

However, from the point of view of IE the problem with the standard tf.idf measure is that it uses *absolute* term frequencies, so highly frequent words have much better chance of getting higher tf.idf scores. But as it was mentioned before, low frequent words with precise meanings may be much more important for IE. There is a need to give an equal chance for highly-frequent words and for low-frequent words to be scored as salient within a given text.

There is a number of alternative salience scores, e.g., proposed in (Church, 2000), (Church and Gale, 1995), (Rayson and Garside, 2000), (Everitt, 1992). However, many of them characterise salience of words in the whole corpus rather than in individual text, and very few were tested across different technologies. Scores which are typically employed for statistical Information Extraction tasks usually use tf.idf as a baseline model. However, it will be an interesting problem for future research to adapt some of the most promising alternative salience scores, such as log-likelihood, and to test their applicability for IE or MT evaluation tasks.

The score proposed in this section uses a fundamental assumption behind tf.idf scores that distribution of words in text is very different from their distribution in the entire corpus; therefore a corpus is more than just one very large text, so there is a need for separate scores for each term in each individual text. At the same time unlike tf.idf, the proposed score isolates the issue of word frequency from a different

issue of word's salience within a given text. In Chapter 5 we show that tf.idf scores in text follow bell-shaped distribution (which is close to normal). If the issues of absolute frequencies were appropriately isolated from the issues of term's salience, the distribution would approach Zipfian shape. We need a score which would have this property, but preserve the ability to capture term's salience.

A possible way to do it is to use relative frequencies (which approximate probabilities) of terms in a document or in a corpus. In particular, we need to contrast the probability of a $word_i$ in a particular $text_j$ and its probability in the rest of the corpus: $P_{i,j}$–$P_{i(\text{rest-of-the-corpus})}$ and normalize this value by the word's probability in the whole corpus: $P_{i(\text{all-corpus})}$. A modified IDF measure will be another factor for the S-score. To keep it within the range [0…1], it may be computed as $1–df_{(i)}/N$ (or equivalently $(N\text{-}df_{(i)})/N$), instead of $N/df_{(i)}$. In this way the IDF factor has a clear intuitive interpretation: it describes the proportion of texts, where the word was *not* found in the corpus: $1–df_{(i)}/N = (N\text{-}df_{(i)})/N$.

The S-scores approximate statistical salience of words within a given text. The formula is:

$$S(i,j) = \log \frac{\left(P_{doc(i,j)} - P_{corp-doc(i,i)}\right) \times (N - df_{(i)}) / N}{P_{corp(i)}}$$

where:

$P_{doc(i,j)}$ is the relative frequency of the word in the text; ("relative frequency" is the number of tokens of this word-type divided by the total number of tokens).

$P_{corp\text{-}doc(i,j)}$ is the relative frequency of the same word in the rest of the corpus, without this text;

$(N - df_{(i)}) / N$ is the proportion of texts in the corpus, where this word does not occur (number of texts, where it is not found, divided by number of texts in the corpus);

$P_{corp(i)}$ is the relative frequency of the word in the whole corpus, including this particular text.

Intuitively, the first factor $(P_{doc(i,j)} - P_{corp\text{-}doc(i,j)})$ in this formula is the difference of relative frequencies in a particular text and in the rest of the corpus. Its value is very high for proper names, which tend to re-occur in one text, but have a very low (often 0) frequency in the rest of the corpus. The higher the difference, the more significant is the word for this text.

The second factor $(N\text{-}df_{(i)})/N$, is an alternative to IDF score: it describes how evenly the word is distributed across the corpus. If a word is concentrated in a small

number of texts, the value is high and the word has more chances of becoming statistically significant for this particular text.

The third factor ($1 / P_{corp(i)}$) boosts statistical significance of low-frequent words. The intuition behind it is that if a word occurs in a particular text more then 2 times (and we consider only words with absolute frequency in the text ≥ 2), it becomes more significant if its general relative frequency in the corpus is low.

We use the natural logarithm of the computed score to scale down the range of its values. In subsequent chapters we will show how S-scores could be useful for other IE-related tasks in MT, in particular – for reference proximity evaluation of MT output.

For the purposes of the current experiment the S-scores were used in the following way:

1. The S-scores were computed for each word with absolute frequency ≥ 2 in the particular text for each text in the corpus.

Here is an example of words ranked according to S-scores in Text 1 of the DARPA94 corpus:

Word	$S_{(i,j)}$	$(N-df_i) / N$	$P_{doc(i,j)} - P_{corp-doc(i,j)} * 100\%$	$P_{corp(i)} * 100\%$
urba-gracco	4.620857	0.99	1.098901	0.010710
Pezet	4.620857	0.99	0.824176	0.008032
sanmarco	4.620857	0.99	0.549451	0.005355
laignel	4.620857	0.99	0.549451	0.005355
hearing	4.620857	0.99	0.549451	0.005355
facet	4.620857	0.99	0.549451	0.005355
emmanuelli	4.620857	0.99	0.549451	0.005355
presiding	4.200307	0.98	0.546747	0.008032
marseille	4.190050	0.97	1.093494	0.016065
deputies	3.907667	0.98	0.544043	0.010710
lyon	3.897411	0.97	0.544043	0.010710
directors	3.897411	0.97	0.544043	0.010710

Expert translation, text 1:

In the ***Marseille Facet*** of the ***Urba-Gracco Affair***, Messrs. ***Emmanuelli***, ***Laignel***, ***Pezet***, and ***Sanmarco Confronted*** by the ***Former Officials*** of the ***SP Research Department***

On ***Wednesday***, February 9, the ***presiding judge*** of the ***Court*** of ***Criminal Appeals*** of ***Lyon***, ***Henri*** Blondet, charged with investigating the ***Marseille facet*** of the ***Urba-Gracco affair***, proceeded with an extensive ***confrontation*** among several ***Socialist deputies*** and ***former directors*** of ***Urba-Gracco***. Ten persons, including ***Henri Emmanuelli*** and Andre ***Laignel***, ***former*** treasurers of the ***SP***, Michel ***Pezet***, and

confrontation	3.897411	0.97	0.544043	0.010710
appeals	3.729578	0.96	0.813361	0.018742
forgeries	3.679541	0.98	0.541339	0.013387
sp	3.592717	0.96	0.810657	0.021420
henri	3.481956	0.97	0.538635	0.016065
questioned	3.301939	0.95	0.535932	0.018742
confronted	3.301939	0.95	0.535932	0.018742
research	3.019206	0.93	0.530524	0.024097
affair	3.019206	0.93	0.530524	0.024097
former	2.714896	0.82	1.578053	0.085678
director	2.647501	0.83	1.047529	0.061581
socialist	2.641580	0.94	0.519709	0.034807
brought	2.575622	0.88	0.519709	0.034807
criminal	2.529820	0.91	0.517005	0.037484
department	2.444534	0.90	0.514301	0.040162
judge	2.418210	0.94	0.511597	0.042839
companies	2.396704	0.92	0.511597	0.042839
officials	2.340823	0.87	0.511597	0.042839
wednesday	2.263339	0.86	0.508894	0.045517
political	2.261380	0.84	0.764692	0.066936
case	2.206641	0.83	0.761988	0.069614
court	2.110550	0.85	0.753877	0.077646
together	1.970650	0.81	0.498078	0.056226
part	1.736603	0.78	0.487263	0.066936
three	0.837934	0.68	0.427780	0.125840
were	0.800100	0.59	0.656540	0.174034
also	0.658376	0.60	0.422372	0.131195
these	0.525725	0.66	0.398038	0.155292
but	-0.478429	0.47	0.314220	0.238293

Philippe ***Sanmarco***, ***former deputies*** (***SP***) from the Bouches-du-Rhône, took ***part*** in a ***hearing*** which lasted more than seven hours.

Besides these ***political*** personalities, three ***former*** Urba ***directors***, Gérard Monate, chairman and managing ***director*** of Urbatechnic, Joseph Delcroix (editor of the "journals" detailing the internal operation of this exceptional ***research department***), and Bruno Desjoberts, ***director*** of the ***Marseille*** regional delegation, participated in this confrontational ***hearing***, which also ***brought together*** Bernard Pigamo, ***former*** campaign ***director*** for Mr. ***Pezet*** and ***director*** for "supporting associations" and a company head. All were ***questioned*** as ***part*** of a ***case*** bearing on acts of bribery, influence peddling, ***forgeries*** and the use of ***forgeries***, and complicity in, or concealment of, these major crimes.

Questions and answers turned mainly on the relationship and the operating methods implemented between ***Urba-Gracco*** and the ***Socialist*** Party. It was an opportunity for the examining magistrate to go further toward illuminating an organized financing system, since local decision makers and national ***political officials***, but also beneficiaries and intermediaries for sums paid by many ***companies*** were ***confronted*** with each other. The thirty-eight heads of ***companies questioned*** in the ***case*** had already been heard, but three of them were ***brought together Wednesday*** following the "***political***" ***confrontation***.

an	-0.766620	0.30	0.671701	0.433747
from	-1.536841	0.18	0.601402	0.503360
by	-2.715982	0.10	0.548968	0.830009
which	-3.039982	0.14	0.210413	0.615813
it	-3.216353	0.23	0.081693	0.468553
with	-3.230189	0.11	0.218525	0.607781
for	-3.839087	0.03	0.691207	0.963881
and	–	0.0	2.259603	2.158023
Of	–	0.0	2.210549	4.404402
A	–	0.0	0.183472	2.016118

The ***presiding judge*** of the ***Court*** of ***Criminal Appeals*** is to render a closing opinion, thus establishing a twenty-day deadline for requests from the various parties, followed by a "may it be communicated" order for settlement of the ***case*** by the ***Lyon*** public prosecutor's office. Considering the thickness of the file, which results from a long procedural battle in the ***Court*** of ***Appeals*** and the Council of State, initiated by an ecologist deputy from ***Marseille***, a trial is not foreseen before 1995.

Table 1: expert translation of Text 1 and word list

$S_{(i,j)}$ is computed for all words with a positive difference $P_{word[text]} - P_{word[rest\text{-}corp]}$. However, many function words also receive this score simply due to the fact that their frequency in a particular text happened to be somewhat higher than their general frequency in the rest of the corpus. So, for comparing statistical models of different MT systems, a threshold – $S_{(i,j)} > 1$ was established. This threshold separates content words and function words rather accurately, and words just above the threshold ("part" and "together" in the above example) are general "low-content" open-class words. The words with $S_{(i,j)} > 1$ are highlighted in the text.

2. In the second stage, the lists of statistically significant words for corresponding texts together with their $S_{(i,j)}$ scores are compared across different MT systems. Comparison is done in the following way:

For all words which are present in lists of statistically significant words both in the human reference translation and in the MT output, the sum of changes of their $S_{(i,j)}$ scores was computed:

$$S_{text.diff} = \sum \left| \left(S_{word-i,text.reference.j} - S_{word.i,text.MT.j} \right) \right|$$

The score $S_{text.diff}$ is added to the scores of all "over-generated" words (words that do not appear in the list of statistically significant words for human reference translation, but are present in such list for MT output). The resulting score becomes the general "over-generation" score for this particular text:

$$S_{over-generation.text} = S_{text.diff} + \sum_{words.text} S_{word.over-generated[text.j]}$$

The opposite "under-generation" score for each text in the corpus is computed by adding $S_{text.dif}$ and all $S_{(i,j)}$ scores of "under-generated" words – words present in the human reference translation, but absent from the MT output.

$$S_{under-generation.text} = S_{text.diff} + \sum_{words.text} S_{word.undergenerated[text.j]}$$

It is more convenient to use inverted scores, which increases as the MT system improves. These scores, $S_{o.text}$ and $S_{u.text}$, could be interpreted as scores for ability to avoid "over-generation" and "under-generation" of statistically significant words. The combined (o&u) score is computed similarly to the F-measure, where Precision and Recall are equally important:

$$S_{o.text} = \frac{1}{S_{over-generation.text}};\ S_{u.text} = \frac{1}{S_{under-generation.text}};\ S_{o\&u.text} = \frac{2S_{o.text}S_{u.text}}{S_{o.text} + S_{u.text}}$$

The number of statistically significant words could be different in each text, so in order to make the scores compatible across texts the average over-generation and under-generation scores per statistically significant word in a given text were computed. For the o_{text} score we divide $S_{o.text}$ by the number of statistically significant words in the MT text, for the u_{text} score we divide $S_{u.text}$ by the number of statistically significant words in the human (reference) translation:

$$o_{text} = \frac{S_{o.text}}{n_{statSignWordsInMT}};\ u_{text} = \frac{S_{u.text}}{n_{statSignWordsInHT}};\ u\ \&\ o_{text} = \frac{2o_{text}u_{text}}{o_{text} + u_{text}}$$

The general performance of an MT system for IE tasks could be characterised by the average o-score, u-score and u&o-score for all texts in the corpus.

The use of contrasting statistical models for human translation and MT output is illustrated by the following example in Table 2:

MT Reverso; Overgenerated words: motor, 4,565274; obligation, 4,565274; tires, 4,565274; debts, 3,841254; global, 3,404379; 12th, 3,255370; actions, 3,234316; franc, 2,839973; order, 2,829043; first, 1,042027	**"Expert"human translation** Undergenerated words (i.e., absent from MT): tire, 4,564768; automobile, 4,143929; fiscal, 4,143929; bonds, 3,840742; stock, 3,612322; reduce, 3,601959; debt, 3,403861; six, 2,839444; 12; 2,817465; amount, 2,716706; per, 2,657005; rates, 2,448991; itself, 2,128073; total, 2,068308; months, 1,956732; beginning, 1,745085; any, 1,297940; can, 1,294282
To reduce the cost of its debt Michelin throws a bond issue for 3,5 billion francs	To Reduce The Cost of Its Debt, Michelin Is Launching a Bond Issue for 3.5 Billion Francs
Michelin decided to proceed, from Wednesday,	Michelin has decided to begin issuing, ***beginning***

January ***12th***, to a bond issue convertible into 3,5 billion ***franc actions***. The ***first*** world manufacturer of tyres so intends to relieve his short-term ***debts***, while bringing him capital necessary for his recovery in the middle of a crisis of the European ***motor*** market. This broadcast will be opened to the public on January ***12th*** at the 255-***franc*** price the ***obligation*** and will concern 9 445 700 titles. His annual interest rate will be 2,5 % and its rate of return actuariel raw product of 5,03 % in case of non-conversion. Of a duration of six years, eleven months and a day, he will be quoted in the Paris Stock Exchange.	Wednesday, January ***12***, an issue ***bonds*** convertible into ***stock*** in the ***amount*** of 3.5 billion francs. In this way, the world's leading ***tire*** manufacturer wants to ***reduce*** its short-term ***debt*** while bringing in the capital needed to recover from the full-blown European ***automobile*** market crisis. This issue will be open to the public on January ***12*** at the price of 255 francs ***per*** bond, and will involve 9,445,700 ***bonds***. Its annual interest rate will be 2.5% and its gross actuarial yield rate will be 5.03% in the event of non-conversion. The issue will have a maturity period of six years, eleven months and one day and will be quoted on the Paris Stock Exchange.
According to Michelin, the conversion, at the rate of an action for an ***obligation***, can be made at any time from February 2nd, 1994. The loan will be altogether paid off itself on January 1st, 2001 at the 307-***franc*** price. A priority period of signature will be reserved for the shareholders, inclusive from 12 till 21 January, at the rate of an ***obligation*** for fifteen ***actions***.	According to Michelin, the conversion, at a rate of one share ***per*** bond can be made at any time ***beginning*** February 2, 1994. The loan itself will be repaid in full as of January 1, 2001 at the price of 307 francs. A subscription-priority period will be reserved for shareholders from January 12 through January 21, at the rate of one bond for fifteen shares.
This operation is going to allow Michelin not to weigh down too much its interest charges in this period of high interest rates, from which particularly suffered the clermontoise firm. A strong part of its debts, a ***global*** 30 billion ***franc*** amount, was it indeed with loans with floating interest rate.	This operation will enable Michelin to avoid burdening ***itself*** with finance costs during this period of high interest rates, which have hit the Clermont firm particularly hard. A large proportion of debt, in the ***total*** amount of 30 billion francs, was in fact borrowed at floating interest ***rates***.
Especially since Michelin can hardly count on the European ***motor*** market to raise its accounts. His losses amounted to 3,45 billion francs in the ***first*** half of the year and should border the 4 billion francs for the fiscal year 1993, according to certain analysts. This result succeeds three negative exercises (11 million from francs to 1992, 1 billion in 1991 and 5,3 billion francs in 1990), in spite of two recovery packages ending in more than 30 000 abolitions of employments on a ***global*** strength of the ***order*** of 125 000 persons.	Especially since Michelin can no hardly count ***any*** longer on the European ***automobile*** market to rehabilitate its books. Its losses rose to 3.45 billion francs for the ***first six months*** and should approach 4 billion francs for fiscal year 1993, according to some analysts. This result follows three negative ***fiscal*** years (11 million francs in 1992, 1 billion in 1991, and 5.3 billion in 1990), despite two recovery plans ending with the elimination of 30,000 jobs cut out of a ***total*** work force of approximately 125,000 persons.
In 1993, both the market of the ***tires*** of ***first*** horsemanship (for the new cars) and that of the ***tires*** of	In 1993, both the new car ***tire*** and the ***tire*** replacement markets collapsed in Europe. In the United States,

replacement collapsed in Europe. In the United States, where Michelin is very present thanks to the acquisition in April, 1990 of Uniroyal-Goodrich, the resumption, of the ***order*** of 8 %, was not sufficient to compensate for the European fall.	where Michelin has a strong presence because of its acquisition of Uniroyal-Goodrich in April 1990, the recovery, about 8%, was not enough to offset the European decline.
o_{text} = 0.612915	u_{text}=0.585990; u&o_{text} = 0.599452

Table 2:Overgenerated and undergenerated statistically significant words

The words highlighted in Table 2 are different for MT output and for human translation. In many cases these differences signal important problems in lexical well-formedness of the MT output which are related to word sense disambiguation or to necessary lexical transformations in the target text, e.g.:

(4)	French original:	*marché automobile européen*
	Human translation:	"*European automobile market*"
	Machine translation:	"*European motor market*"
(5)	French original:	*une obligation pour quinze actions*
	Human translation:	"*one bond for fifteen shares*"
	Machine translation:	"*an obligation for fifteen actions*"
(6)	French original:	*Ce résultat succède a trois exercices négatifs*
	Human translation:	"*This result follows three negative fiscal years* "
	Machine translation:	"*This result succeeds three negative exercises*"
(7)	French original:	*sur un effectif global*
	Human translation:	"*out of a total work force*"
	Machine translation:	"*on a global strength* "
(8)	French original:	*le marché des pneus de première monte (pour les voitures neuves) que celui des pneus de remplacement*
	Human translation:	"*the new car tire and the tire replacement markets* "
	Machine translation:	"*the market of the tires of first horsemanship (for the new cars) and that of the tires of replacement*"

(Only statistically significant words are underlined). Differences in the statistical models of aligned MT output and human translation allow us to spot most

serious factual mistakes automatically, and so improve an aspect of MT that is crucial for the performance of IE systems.

Note however, that the proposed scores could go beyond the range [0…1], which makes them different from precision/ recall scores.

2.2.3. Results of MT evaluation based on statistical modelling

MT evaluation was performed using one human reference at a time. Table 3 presents the results for both runs of the experiment, when the comparison was made to statistical models of each of the human translations – the "Expert" and the "Reference". Rows "DARPA-ade/flu" present human evaluation scores for the given texts, and the columns "CORREL" – Pearson's correlation coefficient *r(4)* for correlation between the human scores and the scores *o, u* and *ou*.

	HT ref/exp	MT systran	MT reverso	MT candide	MT ms	MT globalink	CORREL ade	CORREL flu
HT ref								
o~P	0.951	0.786	0.727	0.800	0.715	0.675	0.823680	0.955498
u~R	0.957	0.763	0.714	0.629	0.699	0.651	0.993564	0.941516
uo~F	0.954	0.774	0.721	0.714	0.707	0.663	0.956092	0.986828
HT expert								
o~P	0.957	0.776	0.719	0.811	0.693	0.677	0.790250	0.946763
u~R	0.951	0.752	0.707	0.634	0.677	0.651	0.994730	0.962116
uo~F	0.954	0.764	0.713	0.723	0.685	0.664	0.938578	0.996678
DARPAade	0.920	0.789		0.677	0.718	0.710		
DARPA-flu	0.850	0.508		0.454	0.382	0.381		

Table 3: MT evaluation scores for statistically significant words

A correlation could be found between some of the computed scores and human MT evaluation measures. The best match has been found between our u-score (the score for avoiding lexical under-generation) and the adequacy scores in DARPA94 MT evaluation. Correlation coefficient *r* for these series of data is around 0.993 – 0.994. Note, that the absolute values for the adequacy scores and u-scores are also very close (Figure 4):

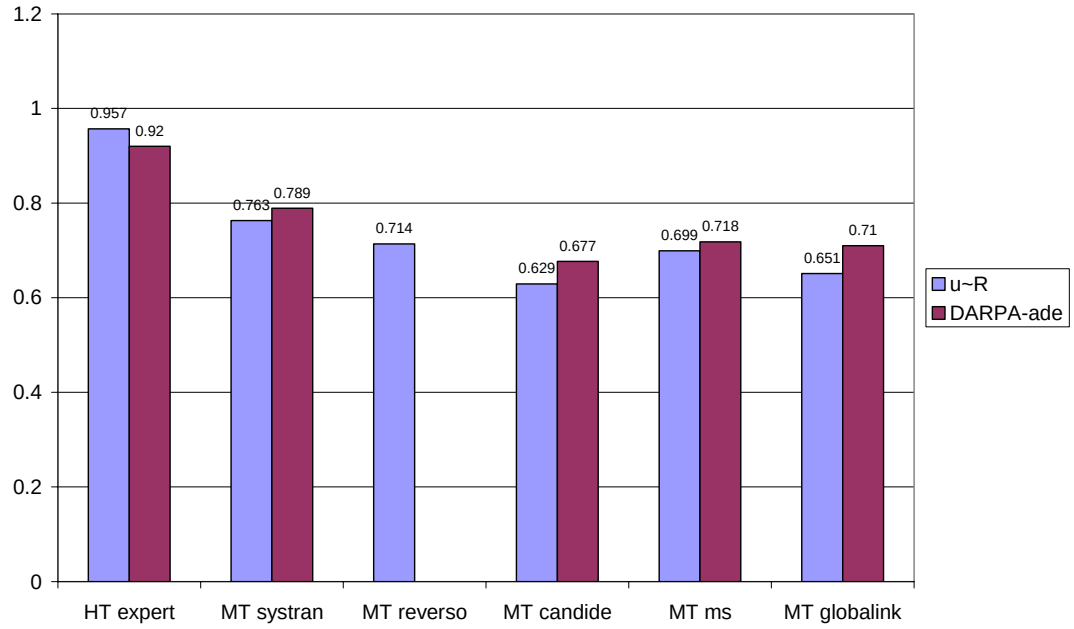


Figure 4: u-scores and DARPA 94 adequacy scores

This close match could be interpreted as a fact that translation adequacy always involves avoiding under-generation of salient lexical items: it demands that no important terms were missing from the translation. Remember that IE performance on the task of NE recognition for Organisation Names had a closest link with the translation *adequacy* parameter (Section 2.1). The results of the current experiment show what the material manifestation of this parameter is: it is the ability of MT not to miss important terms, to avoid their under-generation. In this way it provides corpus-based evidence how MT adequacy is materially realised and what is the mechanism of its influence on IE tasks.

Interestingly, the results also suggest that the link between MT adequacy and IE can also work the other way around: statistical or knowledge-based IE techniques can boost adequacy of MT by improving translation of highly salient terms.

There is also high correlation between "u&o" combined score, and the DARPA94 fluency measures. The correlation coefficient *r* for these series is 0.987–0.997. This may be interpreted as a more complex, combined nature of the fluency parameter, which may be presupposing adequacy as its necessary condition. This interpretation is intuitively plausible, but the data gives only an initial insight for more systematic research in this direction.

Note, that the proposed metrics measure only one aspect of MT, which is considered important for IE purposes, in particular – semantic appropriateness in translations of statistically significant words. It does not measure any other aspects, e.g, syntactic well-formedness.

O-scores and any of the DARPA94 human evaluation scores do not have strong correlation. DARPA94 "informativeness" scores do not have strong correlation with any of automatic evaluation scores.

Let us have a look at the absolute values of the scores. The scores for both runs of the experiment with each of the independent human translations are very similar. Several systems have a better "u&o" combined scores in relation to "reference" translation than in relation to "expert" translation. This might be due to the fact that the quality of the human "reference" translation is lower than that of the "expert" translation, so "reference" contains more cases of literal translation that better match MT output.

The exception to this rule is "Candide", which has a better u&o combined score for the "expert" translation. It also for some reason has a very high u-score, and considerably lower o-score. The exceptionality of “Candide” becomes obvious if we compare its scores for avoiding over-generation and under-generation, the configuration is the same in both runs of the experiment so let us see the run where the model for “Reference” translation is compared to other models (Figure 5):

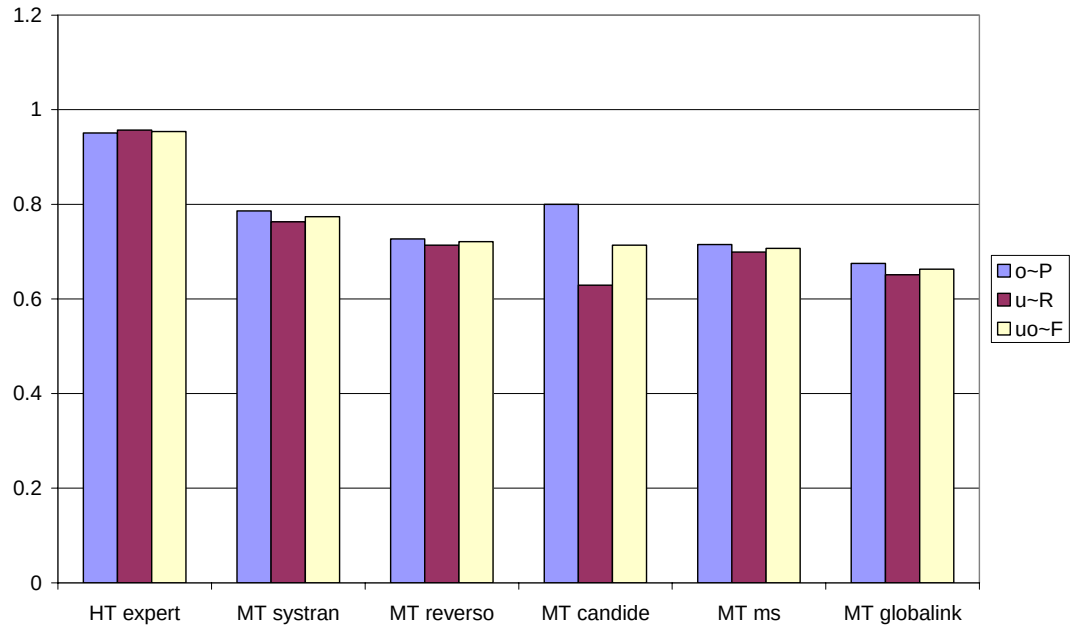


Figure 5: Contrastive scores o, u, u&o for HT Expert and MT systems

It can be seen from the table that scores for human "expert" translation are the best in relation to the other human translation – the "reference" translation. Scores for MT systems are substantially lower, which reflects the fact that they produce many more cases of lexical "under-generation" and "over-generation" of statistically significant words. But note, that the *o* and *u* scores are close for HT and MT, but not for “Candide” MT system: it is much better in avoiding over-generation of salient

lexical items (roughly speaking – it has much better lexical precision) than in avoiding their under-generation (roughly – recall of important terms is much worse).

Such exceptionality of "Candide" can be explained by the fact that this system implements the IBM statistical approach to MT (Berger et al., 1994), and (as it might be expected) produces a substantially different output, partially determined by the statistical structure of the target language. Our analysis allows us to see that the IBM statistical approach does not really improve the score for "avoiding under-generation", which has been found to closely match the DARPA "Adequacy" score. Instead, it considerably improves the score for "avoiding over-generation", which does not directly correspond to any of the DARPA evaluation scores (it influences the combined u&o score, which has been found to match (to some extent) the DARPA "Fluency" score, but more work needs to be done to determine if it really correspond to any important aspect in the quality of MT).

The same results were obtained for the output of a more recent statistical MT system built for Fr-En and for several other translation directions (En-Fr, Es-En, En-Es). Automated MT evaluation scores also over-estimated its adequacy, and human judges ranked it much lower. These results are presented in Chapter 5, Section 5.3 on replicating MT evaluation results on other language pairs. The problem of low adequacy seems to be an inherent feature of statistical MT systems.

These observations provide additional corpus-based evidence for the suggestion made in (Wilks, 1994) that there are fundamental limits for improving pure statistically-based systems (*Wilks's limit on MT*, discussed in Chapter 1): "Candide" showed lowest scores for "avoiding under-generation of statistically significant words" among all tested MT systems. Under-generation and possibly other "recall-based" measures seem to be the weakest point for statistical MT. At the same time the measure of translation adequacy (which is found to be related to our "prevent-under-generation" scores) is considered to be the most important aspect of the translation quality for IE tasks.

The goal of the current experiment is to find a measure of the usefulness of MT output for IE tasks, i.e., to find a performance-oriented measure for IE. However, the results link with the other type – reference proximity measures, which will be discussed in Chapter 4. In this respect it is interesting to compare the proposed *u* and *o&u* scores with BLEU scores.

2.2.4. Comparison with BLEU evaluation measure

BLEU evaluation measure proposed in (Papineni et al., 2001) was applied to the DARPA evaluation data, and the results were compared with the scores based on measuring statistical salience. BLEU scores are computed with 2 human references:

available for the DARPA corpus: “reference” and “human”, with N-gram size =4. Each of the 100 texts in the corpus was treated as a single segment.

The BLEU results and *r* correlation coefficients are presented in the table 6, (it also contains information from table 3 for better comparison with IE-oriented scores):

	HT ref/exp	MT systran	MT reverso	MT candide	MT ms	MT globalink	*r* ade	*r* flu
HT ref								
o~P	0.951	0.786	0.727	0.800	0.715	0.675	0.8237	0.9555
u~R	0.957	0.763	0.714	0.629	0.699	0.651	0.9936	0.9415
uo~F	0.954	0.774	0.721	0.714	0.707	0.663	0.9561	0.9868
HT expert								
o~P	0.957	0.776	0.719	0.811	0.693	0.677	0.7902	0.9468
u~R	0.951	0.752	0.707	0.634	0.677	0.651	0.9947	0.9621
uo~F	0.954	0.764	0.713	0.723	0.685	0.664	0.9386	0.9967
DARPAade	0.920	0.789		0.677	0.718	0.710		
DARPAflu	0.850	0.508		0.454	0.382	0.381		
bleu,N=1		0.7705	0.7650	0.7725	0.7007	0.7306	0.1701	0.862
bleu,N=2		0.4846	0.4653	0.4541	0.3824	0.4031	0.4664	0.9781
bleu,N=3		0.3171	0.2950	0.2797	0.2212	0.2376	0.5469	0.9869
bleu,N=4		0.2168	0.1971	0.1831	0.1373	0.1497	0.5759	0.9884
BLEU		0.4003	0.3793	0.3561	0.3004	0.3199	0.5936	0.9802

Table 6: BLEU scores and IE-oriented scores for the DARPA94 corpus

The BLEU scores strongly correlate with DARPA *fluency* scores, but correlation with other measures for adequacy is much weaker. The main reason for this is consistent overestimation of adequacy for the statistical MT system “Candide”. “Candide” and the BLUE evaluation measure were developed within the same paradigm of ideas, which could influence their close interpretation and formalisation of the “adequacy” concept.

In general, the IE-oriented evaluation measures give comparable results to BLUE scores for the MT systems developed within the same MT architecture (e.g., rule-based). However, BLEU has problems with correlation with *adequacy* if the evaluated set of MT systems is heterogeneous, (if a statistical system is included into the set of rule-based systems). IE-oriented scores in this case predict translation *adequacy* much more accurately than the BLEU method.

However, the IE-oriented scores cannot directly substitute BLEU scores. The problem is that they require the whole MT output corpus for evaluation. Their correlation is much higher than BLEU on the corpus level, but drops down on the

level of individual texts (which is due to the fact that it uses lexical material very selectively, so larger material is needed to ensure their required performance).

In Chapter 4 I examine the problem how IE-oriented MT evaluation metrics can be naturally integrated into reference proximity evaluation tools, such as BLEU.

2.2.5. Conclusion of the experiment

This experiment has investigated a word-salience measure S which compares word frequency within the current text against frequency across the rest of the corpus; by setting an experimentally established threshold, S>1, it is possible to eliminate high-frequency function words, leaving significant content words which characterise the text. (Further experiments presented in Chapter 3 show that this threshold also distinguishes content words and functional words in other languages, such as French and Russian). A comparison of words flagged by this S metric in MT output and human translation highlights factual mistakes. Statistical modelling of MT output corpora has shown substantial differences in distribution of significant words with respect to human translation, which implies that the usability of MT systems for IE technology is still substantially limited. However, the suggested evaluation methodology also allows us to highlight the problems of MT which might be important for the IE task, if MT output is to be used for template filling or acquiring templates automatically. It might also help developers of the state-of-the-art MT systems to identify specific problems relevant for preserving factual information in MT. Proposed measures of lexical match for statistically significant words were found to correlate with DARPA MT evaluation measure of "adequacy". This should allow prediction of the degree to which particular MT systems might be usable for IE tasks.

Future research in the suggested direction could look at the problem of stochastic models for the output of example-based MT systems, and comparing them with models for traditional knowledge-based applications and statistical MT. This could provide insights to establishing the formal properties of intuitive judgements about translation equivalence, adequacy and fluency both for human translation and for MT, and to investigating possible limits on improving MT quality with certain methodologies.

This experiment generalises the results of the Section 2.1. about the performance of one of IE modules (the NE recognition module) on degraded MT output and shows that adequacy of MT is linked to the performance of IE systems in general, so the performance of other IE modules (such as template element filling and scenario template filling, summary generation, co-reference resolution) can

measure the adequacy of MT output; on the other hand, MT adequacy could be improved if these modules are properly integrated into MT architecture.

2.2.6. Further possible applications of IE-based salience scores in MT

2.2.6.1 Application to automatic MT evaluation

Current automatic evaluation methods, such as BLEU (Papineni et al., 2001), do not make a distinction between lexical and morpho-syntactic differences, but distinguishing them and controlling the quality of MT on several separate levels might be useful to for the evaluation of MT systems under development (especially for target languages with a rich morphology, where these two types of differences clearly characterise different aspects of quality).

Another important problem for further research is establishing whether different degrees of legitimate variation in translation are allowed for items with different tf.idf and S-scores. One of the most serious problems for the BLEU method is related to legitimate variability in the reference translation. In order not to penalise acceptable MT that is different from human translation, the metric uses several reference translations of the same text. These resources can be expensive to create. However, if terms with different significance scores show different levels of legitimate variation, then the metric could rely on potentially more stable terms, so fewer reference texts would be needed to produce consistent evaluation scores for MT systems. This problem is partially addressed in Chapter 5: it is shown that there is an interesting and to some extent counter-intuitive link between salience and stability of terms in translation: most salient words appear to be also least stable across independent translations.

Yet another problem for the BLEU metric is high data scarcity of N-grams in languages with complex synthetic morphology, such as Slavonic languages. In order to achieve evaluation scores comparable with scores for English or other analytical languages, we need to use much larger reference corpora of human translations. An alternative solution to this problem could be to make automatically a rough distinction between lexical and morphological differences and to concentrate on the lexical differences that are expected to be less sparse across human translations and MT output.

2.2.6.2 Application to automatic alignment of parallel texts

An analysis of S-scores of lexical differences in the compared translations also gives interesting results. It can be noted that words which are translations of the same word in the DNT-processed and the baseline target texts (see Chapter 3 for

details) have very close scores. Ranked lists of differences for Russian MT are presented in Table 7:

DNT-processed translation	**Baseline translation**
1:NBC:3.939817	1:ЭН-БИ-СИ:3.906120
1:***Техники***:3.416626 technicians$_{(NOM.PLUR)}$	1:***Техников***:3.382496 (of) technicians$_{(GEN.PLUR)}$
1:***Electric***:3.416626	1:***Электрическая***:3.382496 electric$_{(NOM.SING.FEM)}$
1:***Broadcast***:3.416626	1:***Радиопередачи***:3.382496 of broadcast$_{(GEN.SING)}$
2:***Служащие***:2.959119 employees$_{(NOM.PLUR)}$	2:***Служащих***:2.924432 of employees$_{(GEN.PLUR)}$
2:***General***:2.959119	2:***Общая***:2.924432 general$_{(NOM.PLUR.FEM)}$
3:***Association***:1.886203	3:***Ассоциации***:2.303370 of association$_{(GEN.SING)}$

Table 7. Scores for corresponding words

The match between S-scores is closer for words with a unique translation, which implies that they have similar distribution in the text and in the corpus.

Another interesting property of the statistical significance measure is that different word forms which are translations of the same word (e.g., an English NE) often have very close S-scores, which are also close to the score of the original word. For example, S-scores for the first word in the NE "Pan Am" and for three morphological variants of its wrong translation into Russian are presented in Table 8. All are variants of the lexeme "кастрюля" – 'saucepan', and also have different frequencies in the text. This effect is also the strongest for words which have a unique translation in the corpus.

DNT-NE / S-score	*Abs. frq. in DNT text / in the rest of corpus*	*Baseline transl. of NE*	*Abs. frq. in baseline text / in the rest of corp.*
Pan 3.087052	14 / 0	Кастрюля$_{(NOM)}$ 3.112597	8 / 0
		Кастрюлю$_{(ACC)}$ 3.112597	2 / 0
		Кастрюли$_{(GEN)}$ 3.112597	2 / 0

Table 8. Scoring results

This property of the S-score may be useful in MT evaluation for highly inflected languages.

Chapter 3
Improving Machine Translation Quality with Automatic Named Entity Recognition

Named entities create serious problems for state-of-the-art commercial MT systems and often cause translation failures beyond the local context, affecting both the overall morphosyntactic well-formedness of sentences and word sense disambiguation in the source text. We report on the results of an experiment in which MT input was processed using output from the named entity recognition module of Sheffield's GATE IE system. The gain in MT quality indicates that specific components of IE technology could boost the performance of current MT systems. Experiments presented in Chapter 2 showed that Organisation Names have a special place among other NEs, they have the strongest link with MT quality. The idea is that if NE recognition of Organisation Names can characterise *adequacy* of translation, then this link should also work the other way around, so improving NE recognition of organisation names within MT systems should boost general quality of translation. This part tests this assumption and gives corpus-based results which illustrate such improvement of MT quality.

It is interesting to analyse the reasons why the quality of MT can be improved with NE recognition (especially – with identification of Organisation Names). Improvements affect not only NEs themselves, but also their lexical and morphosyntactic context. Even more surprisingly, this improvement is achieved not via a classical path – extension of knowledge sources, available for MT systems, but with quite an opposite method: the use of "do-not-translate" lists, i.e., via restricting the information which is given to the system (Babych and Hartley, 2003, 2004c).

Explanation of this fact could be similar to the suggestion about negative translation equivalents made in Chapter 1. Here again, St Basil's "power to forget" (or the power to rank relative relevance of information) is an essential component of human understanding and translation. In the case of NEs, the peek of relevance is on the boundaries and category of a NE, not on its internal structure. This is because NEs functions in text very differently from common words, and there is an extensive philosophical literature discussing this special function, starting with Russel's theory of descriptions and to Kripke's ideas of the "rigid designators". However, such function is a reflection of even deeper phenomenon of unequal relevance of information units in text and the need to rank this relevance in our models of understanding or translation. Such ranking can be efficiently modelled with IE tools.

3.1. Improving morphosyntactic quality with NE recognition

3.1.1 Motivation for the experiment

Correct identification of named entities (NEs) is an important problem for MT research and for the development of commercial MT systems (e.g., Somers, 2003). In the first place, translation of proper names often requires different approaches and methods than translation of other types of words (Newmark, 1982: 70-83). Mistakenly translating NEs as common nouns often leads to incomprehensibility or necessitates extensive post-editing. In many cases failure to correctly identify NEs has an effect not only on a local and immediate context, but also on the global syntactic and lexical structure of the translation, since proper segmentation of a source text might be seriously distorted.

However, the developers of commercial MT systems often pay insufficient attention to correct automatic identification of certain types of NE, e.g., organisation names. This is due partly to the greater complexity of this problem (the set of proper names is open and highly dynamic), and partly to the lack of time and other development resources.

On the other hand, the problem of correct identification of NE is specifically addressed and benchmarked by the developers of IE systems, such as the GATE system, created at the University of Sheffield and distributed under GPL (Cunningham et al., 1996, 2002). The quality of automatic NE identification has been evaluated at several message-understanding conferences sponsored by DARPA. Accuracy scores for leading systems are relatively high (in comparison to other IE tasks, such as co-reference resolution, template element filling or scenario template filling). The default settings of NE recognition module of the GATE system produces between 80-90% Precision & Recall on news texts (Cunningham et al., 2002).

This section describes the effect of using the GATE NE recognition module as a pre-processor for commercial state-of-the-art MT systems. The idea of our experiment is that high-quality automatic NE recognition, produced by GATE, could be used to create do-not-translate (DNT) lists of organisation names, a specific type of NE which in human translation practice is often left untranslated. (Newmark, 1982: 70-83).

In this experiment the effect of incorrect NE recognition on the surrounding lexical and morphosyntactic context in MT output was systematically analysed in order to establish how far NE recognition (specifically recognition of organisation names) influences grammatical well-formedness and word sense choices in the

context of NEs. The baseline translations (produced without NE DNT-processing) have been compared with translations produced using DNT lists (created with the GATE-1 NE recognition system), by systematically scoring cases of improvement and decline in lexical and morphosyntactic well-formedness. Texts with NE DNT-processing showed consistent improvement for all systems in comparison with baseline translations. The improvement was not lower than 20%.

This indicates that combining present-day MT systems with specific IE modules (where certain NLP problems are treated systematically) has beneficial effect on the overall MT quality.

3.1.2. Problems of NEs for MT

NEs usually require different approaches to translation than do other types of words. For example, foreign person names in Russian should be transcribed and written in Cyrillic; names that coincide with common nouns should not be looked up in the general dictionary. In some cases NEs (mostly organisation names) are not translated and preserve Roman orthography within Russian Cyrillic text. For example, in a 1000-word selection of 4 articles about the international economy on the Russian BBC World Service site, Roman-script NEs within the Cyrillic text covered 6% of the selection. The following NEs were neither translated, nor transliterated into Cyrillic: ‘Nestle’ (9 occurrences), ‘AOL’ (8); ‘Buffalo Grill’ (7); ‘Burger King’ (7); ‘Diageo’ (7); ‘Schweisfurth (Group)’ (2). In general, the practice not to translate organisation names is very common for translations into Slavic languages.

Mistakes related to the failure to distinguish between common nouns and proper nouns in MT can be very serious. For example, in our experiments an MT system translated the person name *Ray* as *Луч* ('beam of light'). Translating parts of compound NEs is also detrimental to MT quality, since it often involves incorrect segmentation of NEs: *American Telephone and Telegraph Corp.* was translated as *Американский Телефон и Компания Телеграфа* ('an American telephone and a company of a telegraph'). Yet another problem for MT systems is that failure to recognise NEs often has a negative effect on well-formedness of morphosyntactic and lexical context beyond the NEs themselves. Certain morphological features of neighbouring and syntactically related words, word order, a choice of word senses in MT output could be distorted if a NE is not correctly recognised. For example, an English phrase (1) was translated into Russian as (2):

Original: *Eastern Airlines executives notified union leaders …*

MT output: *Восточные исполнители Авиалиний уведомили профсоюзных руководителей*

(Lit.: *Oriental executives of the Airlines notified ...*')

This happened because the failure to identify *Eastern Airlines* as a NE led to incorrect syntactic segmentation of the sentence.

However, current MT systems allow the processing of MT input with DNT lists. Making a DNT of organisation names from the text in most cases improves not only the acceptability of NE translation, but also the overall well-formedness of the morphosyntactic and lexical context. For example, after the string *Eastern Airlines* was entered into a DNT list for the English-Russian MT system, the translation of (1) was morphologically and syntactically well-formed:

DNT-processed MT output: *Исполнители Eastern Airlines уведомили профсоюзных руководителей ...*

Creating DNT lists manually requires much effort from the user of an MT system. However, the high accuracy in NE tagging of current IE systems, including GATE, means that DNT lists for MT can be created automatically.

The performance results reported here are based entirely on automatically created DNT lists used to process NEs.

3.1.3. Description of the experiment

In order to measure the effect of NE recognition on MT quality, we took 30 texts (news articles) from the DARPA MUC-6 evaluation set. These texts were selected because they are relatively rich in NEs, and because clean NE annotation is available for them. We used the following linguistic resources of the Sheffield NLP group:

DARPA 'keys' – texts manually annotated with NEs;

GATE 'responses' – the output of the automatic NE annotation of the GATE-1 system, which participated in MUC-6.

Table 1 summarises statistical parameters of this corpus. The table indicates how frequently NEs (organisation names) occur and shows that GATE 'response' figures are very close to the DARPA "key" figures.

Number of:	*For the corpus*	*Av. per doc.*	*Av. per para.*	*Av. per sent.*
Paragraphs	283	9.4	–	–
Sentences	565	18.8	2.0	–
Word occurrences	11975	399.2	42.3	21.2
Different words	3944	235.7	36.3	19.7

NE occurrences keys/ ***GATE***	544/ **510**	18.1/ **17.0**	1.9/ **1.8**	1.0/ **0.9**
Different NEs: keys/ ***GATE***	201/ **174**	7.6/ **6.7**	1.5/ **1.4**	0.9/ **0.8**

Table 1: Statistical parameters of the corpus

The density of NEs in the DARPA corpus is also characterised by Table 2:

	Manual keys	*GATE*
Paragraphs with NEs	228 (80.6%)	**218 (77.0%)**
Sentences with NEs	329 (58.2%)	**315 (55.8%)**

Table 2: NE density in the corpus

The accuracy of GATE-1 in the NE recognition task at MUC-6 (Recall – 84%, Precision – 94%, Precision & Recall – 89.06 % (Gaizauskas et al., 1995)) is such that we used the GATE output for our MT experiment, rather than the cleaner manually annotated data. Moreover, the advantage of using automatic NE recognition is that the results of the experiment should be consistent with the results for other corpora on which the NE recognition task has been performed.

Having automatically generated DNT lists of organisation names from GATE 'response' annotation, we translated the texts using three commercial MT systems:

English-Russian 'ProMT 98' v4.0, released in 1998 (Softissimo)

English-French 'ProMT', (*Reverso*) v5.01, released in 2001 (Softissimo)

English-French 'Systran Professional Premium' v3.0b, released in 2000 (Systran)

Two translations were generated by each MT system:

– a baseline translation without a DNT list;

– a DNT-processed translation with the automatically created DNT list of organisation names;

The baseline translations were then compared with DNT-processed translations, with respect to the morphosyntactic well-formedness of the context surrounding the NEs.

3.1.3.1.Segmentation

To speed-up the process of finding contextual differences, we developed automatic tools, which allowed us to make a formal distinction between NE-internal and NE-external issues in MT. Whereas Al-Onaizan and Knight (2002) focus on the

former issue, our primary interest is in NE-external differences in context caused by improved NE recognition after DNT-processing. Thus, we automatically selected paragraphs with contextual differences and highlighted different strings in these paragraphs.

The example below illustrates the output of these annotation tools:

Different strings found in two translations are indicated by '---->'

'ORI' indicates the original English string in the DARPA corpus;

'TWS' (baseline translation) indicates a String Translated Without the do-not-translate list;

'TDS' (DNT-processed translation) indicates a String Translated with Do-not-translate list.

---->40;TDSnotInTWS: 40# Отдельно, в его регистрации

---->40;TDSnotInTWS: раскрыл детали его планов финансирования приобретения

40;ORI=40#<s> Separately, in its <ENAMEX>SEC</ENAMEX> filing, <ENAMEX>USAir</ENAMEX> disclosed details of its plans for financing the <ENAMEX>Piedmont</ENAMEX> acquisition.

40;TWS= 40# Отдельно, в ее регистрации СЕКУНДЫ, USAir раскрытые детали ее планов финансирования Предгорного приобретения.

40;TDS= ***40# Отдельно, в его регистрации*** *SEC, USAir* ***раскрыл детали его планов финансирования приобретения*** *Piedmont.*

Since the amount of manual annotation was relatively small, no complex alignment for the two translated texts was implemented. Instead, we implemented a simple segmentation algorithm for paragraphs, using NE annotation in the corpus.

The segmentation was done in two stages. First, tagged NEs from the 'ORI' paragraph were identified and searched for in the 'TDS' paragraph. Then they were used as separators for the TDS: parts of the TDS between (untranslated) NEs were identified and searched for in the 'TWS' paragraph. If any sub-string was not found in TWS, it was printed and also highlighted in bold in TDS. This shows that strings in the context of the NE are different in the DNT-processed translation and in the baseline translation. This difference was then manually scored.

3.1.3.2. Scoring

Contextual differences between the baseline translation and the DNT-processed translation were manually scored using the scale in Table 3.

The terms 'well-formed' and 'not well-formed' refer to the local morphosyntactic or lexical context within a segment where differences occur. It remains possible that well-formed structures require post-editing at a higher level in the translated text.

The term 'features' refers to morphosyntactic or lexical features of certain words in the context of the NE. By 'more correct', we mean that the features considered in the context are correct, but the corresponding features in the compared text are wrong.

Score	*Baseline translation*	*DNT-processed translation*
+ 1	not well-formed	well-formed
+ 0.5	not well-formed;	not well-formed; some features are more correct
= 0	equally (not) well-formed	
– 0.5	not well-formed; some features are more correct	not well-formed
– 1	well-formed	not well-formed

Table 3: Scoring scheme

Here are some example strings to illustrate each score:

+1	**Original:** (It) represents 4,400 ***Western Union*** **employees** around the country.
	Baseline translation: (Он) представляет 4,400 **Западных служащих Союза** по всей стране. ('It represents 4,400 **Western employees of the Union** around the country')
	DNT-processed translation: (Он) представляет 4,400 **служащих Western Union** по всей стране. ('(It) represents 4,400 **employees of Western Union** around the country')

+0.5	**Original:** Western Union Corp. **said its subsidiary**, Western Union Telegraph Co.…
	Baseline translation: Западная Корпорация Союза **сказала ее вспомогательную**, Западную Компанию Телеграфа Союза… ('Western Corporation of a Union said **its auxiliary (case.acc.)**, Western Company of Telegraph of a Union …')
	DNT-processed translation:

Western Union Corp. **Сказанный его филиал**, Western Union Telegraph Co. …
('Western Union Corp. **Its branch (case.nom) is said**, Western Union Telegraph Co.…')

=0	**Original:** *American Airlines* **Calls** for Mediation
	Baseline translation: Американские Авиалинии **Призывают** К посредничеству (American Airlines **Call(num.plur.)** for Mediation)
	DNT-processed translation: American Airlines **Призывает** К посредничеству (American Airlines **Calls(num.sing.)** for Mediation)

–0.5	**Original:** *USAir* said **that** William R. Howard, chairman and chief executive of *Piedmont*, will be elected president of *USAir*
	Baseline translation: USAir сказал **тот** Уильям Р. Говард, председатель и руководитель Предгорных, будут избраны президентом USAIR *USAir* said **that (particular)** (demonstr.pron,**nom.**) William R. Howard, chairman and chief executive of piedmont people, will be elected president of *USAir*
	DNT-processed translation: USAir сказал **того** Уильяма Ра. Говард, председатель и руководитель Piedmont, будут избраны президентом USAir *USAir* said **of that (particular)** (demonstr.pron,**gen.**) William Ra. Howard, chairman and chief executive of *Piedmont*, will be elected president of *USAir*

–1	**Original:** to discuss the benefits of **combining** *TWA* and *USAir*
	Baseline translation: чтобы обсудить выгоды от **объединения** TWA и USAIR ('to discuss the benefits of **the merge (noun) (of)** *TWA* and *USAir*')
	DNT-processed translation: чтобы обсудить выгоды от **объединяющегося** TWA и USAir ('to discuss the benefits of **the combining (participle, sing.)** *TWA* and **(of)** *USAir*')

For each MT system, we scored 50 strings showing differences. Table 4 summarises the number of paragraphs with contextual differences between the baseline and DNT-processed translations.

The figures in row 2 – *Paragraphs with contextual differences* – show to what extent DNT-processing affects the NE context for each system, showing also the

percentage of these paragraphs in relation to the corresponding figure in row 1. Row 3 represents the percentage of manually scored paragraphs in relation to the figure in row 2. These figures show the likely reliability of the results for manual scoring presented in the next section.

Number of:	*Original – GATE*	*MT E-R ProMT*	*MT E-F ProMT*	*MT E-F Systran*
Paras. with NE	218	225	225	239
Paras. with contextual differences		139 (61.8%)	132 (58.7%)	207 (86.6%)
Paras. manually scored		31 (22.3%)	28 (21.2%)	30 (14.5%)
Strings with differences		211	212	411
Strings scored		50 (23.7%)	50 (23.6%)	50 (12.2%)
Diff. strings per text		7.0	7.0	13.7
Diff. paras. per text		4.6	4.4	6.9

Table 4: Paragraphs with contextual differences

Note that in row 1 there is a mismatch between the number of paragraphs with NEs in the original GATE-annotated English texts (218) and in the translations produced by the three MT systems (225, 225 and 239 paragraphs with NEs). This is because the results of NE pre-processing could be submitted to the proprietary MT systems only in the form of a DNT list, which has its limitations. The most serious potential problem is over-generation: ambiguous items, which could be either NEs or common words in different contexts, are treated as NEs in *every* context, once they are written to the DNT list. For example, the word *Labour* could be either an organisation name ('the party'), a part of a larger NE, often of a type other than organisation name (*Federal Railway Labour Act*), or a common noun ('work', as in the phrase: *rise in labour costs*). As a result, in the translated corpus there are more NEs than in the original English corpus, annotated with GATE. This is reflected in the figures presented in row 1 of Table 2. Nevertheless, the difference is relatively low (less then 10% for the worst case). Given that there are (on average) only about 2 NE occurrences per paragraph in the corpus, over-generation does not greatly affect our evaluation results. (Some MT systems accept special "do-not-translate" annotation in the source text, which would be much more preferable. This option, however, was available neither in ProMT, nor in Systran at the time of experiment).

The segmentation method described above provided us with a clear formal distinction between NE-internal and NE-external problems for MT. However, we made one exception to this distinction: in the DNT-processed English-French, Systran often incorrectly inserts definite articles for organisation names which are present in DNT list, but does not do so in the baseline translation. Our segmentation method treats these articles as part of the morphosyntactic context of NEs, and

considerably increases the contextual degradation figures for Systran. But, linguistically, it is more correct to treat French articles as inner parts of NEs. Therefore, for the evaluation of contextual changes for Systran, we ignored strings where the inserted article was the only difference. As a result, Systran showed a net contextual improvement.

3.1.4. Results of the experiment

Table 5 summarises the results of the manual annotation of 50 strings containing differences for each MT system. (There are 61 scored differences for Systran, because in some strings there was more then one morphosyntactic or lexical difference).

	ProMT 1998 E-R		*ProMT 2001 E-F*		*Systran 2000 E-F*	
Mark	*N*	*Score*	*N*	*Score*	*N*	*Score*
+1*	28 =	+28.0	23 =	+ 23.0	18 =	+ 18.0
+0.5*	2 =	+1.0	5 =	+ 2.5	24 =	+ 12.0
0*	4 =	0	7 =	0	8 =	0
–0.5*	3 =	–1.5	1 =	– 0.5	1 =	– 0.5
–1*	13 =	–13.0	14 =	– 14.0	10 =	– 10.0
∑	50	**+14.5**	50	**+ 11.0**	61	**+ 19.5**
Gain	**+29%**		**+22%**		**+32%**	

Table 5: Manual annotation results

N is the number of differences, annotated with that particular score. To compute the overall score for the system we multiplied the scores by the number of strings with this particular score, and added the results. The improvement was then computed by dividing the overall score by the number of scored differences:

∑score / ∑ N.

In order to see how the resulting scores change when more data is analysed, scoring the English Russian ProMT 98 system continued, until 100 paragraphs with differences had been annotated. The results are presented in Table 6.

	ProMT 1998 E-R	
Mark	*N*	*Score*
+1*	59 =	+59.0
+0.5*	8 =	+4.0
0*	6 =	0
–0.5*	7 =	–3.5
–1*	31 =	–31.0
∑	111	**+28.5**
Gain	**+26%**	

Table 6: Results for additional E-R data

Here is an example of a sentence where improvement has been achieved in the DNT-processed translation for all three MT systems on several levels: morphological, syntactic and lexical.

	Original: The agreement was reached by a coalition of four of *Pan Am*'s five unions.
E-R ProMT	**Baseline translation**: Соглашение было достигнуто коалицией четырех Кастрюли пять союзов Ама. ('The agreement was reached by a coalition of four of a Saucepan five unions of Am.')
	DNT-processed translation: Соглашение было достигнуто коалицией четырех из пяти союзов Pan Am. ('The agreement was reached by a coalition of four out of five unions of *Pan Am* ')
E-F ProMT	**Baseline translation**: L'accord a été atteint par une coalition de quatre de casserole cinq unions d'Am. ('The agreement was reached by a coalition of four of saucepan five unions of Am.')
	DNT-processed translation: L'accord a été atteint par une coalition de quatre de cinq unions de Pan Am. ('The agreement was reached by a coalition of four of five unions of Pan Am.')
E-F Systran	**Baseline translation**: L'accord a été conclu par une coalition de quatre de la casserole étais cinq syndicats. ('The agreement was reached by a coalition of four of the saucepan was five trades-unions.')
	DNT-processed translation: L'accord a été conclu par une coalition de quatre de Pan Am's cinq syndicats.

('The agreement was reached by a coalition of four of Pan Am's five trades-unions.')

Here are further typical cases of morphosyntactic improvement in the translated material:

Improved syntactic segmentation:

Original:
Representatives for the 5,400-member *Allied Pilots Association* didn't return phone calls.

E-R ProMT

Baseline translation:
Представители для *Союзнических Пилотов с 5,400 членами Ассоциация* не возвращали обращения по телефону.
('Representatives for the *Allied Pilots with 5,400 members Association* didn't return phone calls.')

DNT-processed translation:
Представители для Allied Pilots Association с 5,400 членами не возвращали обращения по телефону.
Representatives for the *Allied Pilots Association* with 5,400-members didn't return phone calls.

Improved proper / common disambiguation:

Original:
A spokesman for the company said *American* officials 'felt that …'

E-F ProMT

Baseline translation:
Un porte-parole de la société a dit que les fonctionnaires *américains* 'ont estimé que …'
('A spokesman for the company said that the American [US] officials 'felt that …'')

DNT-processed translation:
Un porte-parole de la société a dit que les fonctionnaires *d'Américan* 'ont estimé que …'
('A spokesman for the company said that the officials of American 'felt that …'')

Improved word order:

Original:
USAir disclosed details of its plans for financing …

E-F ProMT

Baseline translation:
USAir les détails révélés de ses plans pour financer …
('USAir the details revealed (*Past participle*) of its plans for financing …')

DNT-processed translation:
USAir a révélé les détails de ses plans pour financer …
('USAir revealed (*Verb*) the details of its plans for financing …')

Improved lexical or syntactic disambiguation:

	Original: TWA *stock closed* at $28 …
E-F Systran	**Baseline translation**: *Fermé courant* de TWA à $28 … (‘Closed (*Past participle*) current (*Noun/Present participle*) of TWA at $28 …’)
	DNT-processed translation: *L’action* de TWA *s’est fermée* à $28 … ('The stock of TWA closed (*Verb*) at $28 …')
	Original: National Mediation Board is expected to release Pan Am Corp. and its Teamsters union from their long-stalled contract negotiations.
E-R ProMT	**Baseline translation**: Национальное Правление Посредничества, как ожидается, *выпустит* Кастрюлю - Корпорация и ее союз Водителей от их долго-остановленных переговоров контракта. ('National Mediation Board is expected to release [put on the market] a Saucepan - Corporation and its union of drivers from their long-stalled contract negotiations.')
	DNT-processed translation: National Mediation Board, как ожидается, *освободит* Pan Am Corp. И его союз Teamsters от их долго-остановленных переговоров контракта. ‘National Mediation Board is expected to release [make free] Pan Am Corp. and its Teamsters union from their long-stalled contract negotiations.’)

3.1.5. Conclusions for the experiment

The results indicate that combining IE technology with MT has a great potential for improving the state-of-the art in output quality. Taking advantage of efforts to resolve specific linguistic problems – as has happened with NE recognition within the IE framework – improves not only the treatment of that phenomenon by MT, but also morphosyntactic and lexical well-formedness more generally in the wider context of the target, thus boosting the overall quality of MT. The results show that modern MT systems still leave room to achieve a considerable improvement. Further gains in performance may be anticipated by harnessing other focussed technologies, such as word sense disambiguation, to MT.

Future research could look at the sensitivity of the performance gain to corpus size and variation. Table 6 shows that the difference in the score for 50 annotated paragraphs and the score for 100 paragraphs for E-R ProMT98 is 3%. In general, different occurrences of the same NE tend to have a similar morphosyntactic context, so they constantly tend to either improve or worsen the quality. In a particular text, the same NEs tend to re-occur. As a result, an improvement or a decline in quality is usually not homogeneous across corpora, but is more constant for a particular text. The score changes in more or less homogeneous chunks of text.

For E-R ProMT 98 MT system the average size of such chunks is about 7 differences (See Table 3, row 6 'Different strings per text'). For E-R ProMT 98, the value of each '+1' or '–1' score after 50 annotated differences is ±2%, so one text can potentially change the score by about ±14%. After checking 100 differences, the value of each '+1' or '–1' score becomes ±1%, so a new text could change the score by ±7% on average. In the case of E-R ProMT 98, scoring 50 additional new strings (about 7 new texts) changed the overall score by –3%. This indicates that, for our corpus, there is a reliable improvement after NE DNT-processing.

3.2. Selecting Lexical Translation Strategies in MT using Automatic Named Entity Recognition

The previous section looked into the quality of MT in the context of NEs, essentially *outside* the NEs themselves. In this section a different, but a related problem is addressed – whether the chosen translation strategies are optimal for the segments *inside* the boundaries of NEs

This section reports on the results of an experiment aimed at enabling a machine translation system to select the appropriate lexical strategy for dealing with words and phrases which have different translations depending on whether they are used as proper names or common nouns in the source text. ANNIE system was used to identify named entities in the source text and pass them to MT systems in the form of DNT lists. A consistent gain of about 20% in translation accuracy was achieved for all tested systems. The results suggest that successful translation strategy selection is dependent on accurate segmentation and disambiguation of the source text – aspects which could be significantly improved by named entity recognition. Further an automatic method for distinguishing and lexical differences in MT output is suggested that could have applications in automated MT evaluation for morphologically rich languages.

3.2.1. Motivation for the experiment

Language communities develop certain acceptable practices and norms for translating different types of concepts, expressions and texts from other languages and cultures. These practices are described as *translation methods*, *translation strategies* and *translation procedures*. (Vinay and Darbelnet, 1958, 1995). Translation methods relate to whole texts, while strategies and (finer-grained) procedures relate to sentences and smaller units (Newmark, 1988:81). The choice of a translation strategy often depends on the type of a translated unit. For example, for certain types of proper names the optimal translation strategy is *transference*, i.e., a "*do-not-translate*" or "*transliterate*" strategy, while the majority of common nouns

are translated with other strategies: *literal translation, transposition, modulation,* etc. (Newmark, 1988: 81-88). This implies that recognising different types of units in the source text is a necessary condition for optimising the choice of translation strategy and, ultimately, for improving the quality of the target text.

The problem of selecting translation strategies for words that may be used as proper names or common nouns in the source language is related to a more general problem of word sense disambiguation (WSD) – one of the most serious problems for Machine Translation technology. Dealing with "proper *vs* common disambiguation" (PCD) often requires combining different knowledge sources, in a similar way to WSD (Stevenson and Wilks, 2001). But the cross-level nature of this problem also suggests that improvement in MT quality could be achieved through improving related aspects of the source-text analysis, such as Named Entity recognition (Babych and Hartley, 2003; Somers, 2003:524). For the purposes of this discussion, we assimilate proper nouns to NEs and investigate NE recognition as a possible solution to the PCD problem insofar as it might enable the selection of the correct strategy.

Accurate NE recognition is important for the general quality of MT for the following reasons:

1. The translation of the same token may be different depending on whether the token is a common noun or part of an NE, e.g. in Russian if a common name is a part of an organization name, a "do-not-translate" or "transliterate" strategy should be used instead of a default translation strategy:

(1) **Original**: *…the Los Angeles office of the Hay Group, a management consulting firm.*

MT output[3]: *...Лос-Анджелесский офис Группы Сена, управление консультантская фирма.*

Lit.*:... the Los Angeles office of the group of the hay [i.e., the grass, cut and dried for fodder], management consulting firm*

Human translation: *Лос-Анджелесский офис Hay Group, управленческой консультантской фирмы.*

In this case NE recognition is directly linked to the PCD problem: we need to disambiguate between "common" and "NE" readings of the same string.

---

[3] The examples are taken from the output of MT systems that translated 30 texts of MUC-6 data, which was originally used for evaluating NE recognition.

2. Failure to recognise NEs as single syntactic units or to determine their correct morpho-syntactic category in the source text may cause segmentation errors, which lead to the wrong morpho-syntactic structure in the target text, e.g.:

(2) **Original***: a Big Board spokesman couldn't comment on the talks.*

MT output*: Большой представитель Правления не мог комментировать переговоры.*

Lit.*: A big spokesman of the Board [management] couldn't comment on the talks.*

In this case, NE recognition affects mainly morpho-syntactic segmentation, but individual words normally have correct translation strategies. However, a different morpho-syntactic context often requires the selection of a different translation strategy (either within or outside NEs), which may cause PCD errors in MT output, so there is an indirect link between morpho-syntactic disambiguation and PCD e.g.:

(3) **Original***: Moody's Investors Service Inc. placed the long-term debt under review.*

MT output*: Инвесторы Муди Обслуживают компанию, поместил долгосрочный долг под обзором.*

Lit*.: Investors of Moody serve the company, he placed the long-term debt under review.*

Here the NE *Investors Service Inc*. is not treated as a single segment, which causes a combined morpho-syntactic and PCD error: the system translates the word *service* as a verb that means ‘to serve’ instead of using the correct “do-not-translate” strategy.

Thus NE recognition could be beneficial both for morpho-syntactic well-formedness and for correct PCD in MT output. In (Babych and Hartley, 2003) we addressed the first of these two problems. In this section, we concentrate on the second problem and show how PCD can be improved using existing NE recognition modules.

Certain types of NEs, such as organisation names, appear to be a weak point even for some leading-edge MT systems, such as Systran and Reverso. At the same time, the problem of accurate NE recognition has been specifically addressed and benchmarked by the developers of IE systems. For example, the NE recognition module of the ANNIE IE system achieves a combined Precision & Recall score of 80-90% on news texts (Cunningham et al., 2002). Our suggestion is that combining this highly accurate NE recognition module with state-of-the-art MT systems would

be beneficial for MT output, even if we do not change any of the other MT components.

The source code for commercial MT systems is not publicly available, so for our experiment we used one of the pre-processing tools of these systems – DNT lists. These lists were created from NE annotation produced by the ANNIE NE recognition module. For each of the three available MT systems we generated two different translations: a baseline translation and the DNT-processed translation. An approximate distinction was made between PCD and morpho-syntactic differences automatically using statistical frequency weights similar to tf.idf scores. The improvement in PCD was evaluated by manually annotating the PCD differences in the baseline and NE-processed MT output.

3.2.2. Distinguishing lexical and morpho-syntactic differences in MT output

DNT-processing causes both morpho-syntactic and lexical differences in compared translations. In example (4) we annotate lexical (L) and morpho-syntactic (M) differences in the reference and DNT-processed translations. These differences are due to the fact that the company name "Eastern (Airlines)" received a correct morpho-syntactic category as a result of DNT-processing (Noun, not Adjective). Moreover, not translating this company name is the correct option for Russian target text.

(4)	**Original:** By proposing a meeting date, **Eastern** moved one step closer toward reopening current high-cost contract agreements
	Baseline translation: Предлагая дату встречи, **Восточный**-*(L)* перемещенный-*(M)* один шаг ближе к повторному открытию высокой стоимости-*(M)* потока-*(L)* заключают-*(L)* соглашения-*(M)* ('By proposing a meeting date, **Eastern** *(Adj.)* moved *(Participle)* one step closer toward reopening the high-cost$_{(ACC)}$ of a current *(Noun: 'the stream [of water, etc.]')* (they) conclude *(Verb)* agreements$_{(ACC)}$')
	DNT-processed translation: Предлагая дату встречи, **Eastern-***(L)* переместил-*(M)* один шаг ближе к повторному открытию текущих-*(L)* соглашений-*(M)* контракта-*(L)* с высокой стоимостью-*(M)* ('By proposing a meeting date, **Eastern** *(Noun)* moved *(Verb)* one step closer toward reopening of current *(Adj.)* agreements$_{(GEN)}$ of a contract *(Noun)* with high cost$_{(INST)}$')

	Original	**Baseline**	**DNT-proc.**
L	Eastern	Восточный ('Eastern$_{(ADJ)}$')	Eastern (not translated)
L	Current	потока (stream$_{(NOUN)}$')	текущих ('current$_{(ADJ)}$')
L	Contract	заключают ('conclude$_{(VERB)}$')	контракта ('contract$_{(NOUN)}$')
M	Moved	перемещенный (PARTICIPLE)	переместил$_{(VERB)}$
M	Cost	стоимости$_{(GEN)}$	стоимостью$_{(INST)}$
M	Agreements	соглашения$_{(ACC)}$	соглашений$_{(GEN)}$

Table 1. Examples of translation differences

In this example, all six variants in the DNT-processed translation are better than their counterparts in the baseline translation.

Note that a correct PCD choice for *lexical* differences is determined by the senses of the words in the source text, and there is no way of correctly using lexical items from the baseline translation as alternative translations. In contrast, the source text does not require particular values of *morpho-syntactic* categories in the target text. These values are determined by the rules of the target language and by the morpho-syntactic structure of a sentence, chosen by a translator. In many cases these values can be subject to greater variation then the lexical choices. For example, there is a legitimate way of using the last two words in the Table 1 in the genitive and accusative case, as in the baseline translation shown in example (5), if these values are required by their morpho-syntactic position:

(5) Предлагая дату встречи, Eastern переместился на один шаг ближе к тому, чтобы повторно открыть текущие контрактные соглашения$_{(ACC)}$ высокой стоимости$_{(GEN)}$.

> **Lit**.*: By proposing a meeting date, Eastern moved one step closer toward that [situation], to reopen current agreements$_{(ACC)}$ of high cost$_{(GEN)}$*

A rough distinction between morpho-syntactic and lexical differences in the compared output texts can be drawn automatically using S-scores for term frequency weights proposed in Chapter 2 for evaluating MT for Information Extraction purposes. These weights were found to make an accurate distinction between content and function words. With a varying degree of accuracy (depending on how analytic the grammar of a given language is) this distinction also separates lexical and morpho-syntactic differences in compared texts. For Russian (which has a not

highly analytic grammar) it achieves 88.4% Precision for lexical items, while for French the Precision is 98%.

The S-scores are computed for each word in each text using the formula given in Chapter 2:

$$S(i,j) = \log \frac{\left(P_{doc(i,j)} - P_{corp-doc(i)]}\right) \times (N - df_{(i)}) / N}{P_{corp(i)}}$$

S-scores were computed for words with:

$(P_{doc(i,j)} - P_{corp\text{-}doc(i)}) > 0$; $AbsFrq_i > 1$, where $AbsFrq_i$ is the number of occurrences of the word w_i in the corpus.

Table 2 illustrates the ranking of words according to their S-score for one of the English texts from MUC6 NE corpus, for which $tf_{i,j} > 1$ ($tf_{i,j}$ is the number of occurrences of the word w_i in the document d_j).

r	S	word	r	S	Word	r	S	word
1	2,918	OPEC	2	2,719	output	18	0,621	also
1	2,918	Emirates	3	2,449	others	19	0,527	much
1	2,918	barrels	3	2,449	manager	20	0,331	but
1	2,918	oil	3	2,449	government	21	0,291	over
1	2,918	quota	3	2,449	dropped	22	0,007	from
1	2,918	Subroto	3	2,449	declines	23	-0,079	there
1	2,918	world	3	2,449	agency	24	-0,126	after
1	2,918	cartel	4	2,375	day	25	-0,233	their
1	2,918	war	5	2,305	production	26	-0,244	new
1	2,918	ruler	6	2,096	well	27	-0,284	had
1	2,918	petroleum	6	2,096	demand	28	-0,411	as
1	2,918	markets	7	1,880	concern	29	-1,225	talks
1	2,918	gestures	8	1,844	total	30	-1,388	been
1	2,918	estimates	8	1,844	report	31	-1,594	at
1	2,918	conciliatory	9	1,692	current	33	-1,844	on
1	2,918	Zayed	10	1,593	price	34	-2,214	its
1	2,918	UAE	10	1,593	news	35	-3,411	for
1	2,918	Szabo	11	1,470	recent	36	-3,707	with
1	2,918	Sheik	12	1,270	month	38	-4,238	the
1	2,918	Saudi	13	1,161	officials	39	-4,319	by
1	2,918	Petroleum	14	0,972	because	40	-4,458	Mr
1	2,918	Dhabi	15	0,805	million	41	-5,323	the
1	2,918	Arabia	16	0,781	yesterday	42	–	a
1	2,918	Abu	17	0,651	that	42	–	of

Table 2. Ranking of words by the S-score

A threshold for distinguishing content words and functional words was experimentally established, it is: *S-score = 1*

This threshold gives good results for text in all analysed languages: English, French and Russian. Our assumption implies that for comparing lexical differences in two variants of translation we need to compare for each text sets of words with an S-score above the threshold.

Accordingly, all words that were different in each set were automatically highlighted in their respective texts and presented for manual scoring. In the examples of MT in the following sections, words with $tf_{i,j} > 1$ are bold, words with $tf_{i,j} = 1$ are bold and italic. In the original English sentences, the NEs used for the DNT lists are highlighted in bold.

3.2.3. Resources and scoring method

For the experiment the same linguistic resources were used as for the experiment presented in Section 4.1: 30 texts (news articles) which were processed with the NE recognition module of the GATE-1 IE system in the DARPA MUC6 competition and translated with the 3 commercial MT systems.

The baseline and the DNT-processed translation were automatically compared using the method presented in Section 4.2.2. Lexical differences were highlighted and scored according to the following criterion:

+1	– PCD is correct in the DNT-processed translation and is wrong in the baseline translation
0	– PCD in both translations is equally (not) correct
–1	– PCD is wrong in the DNT-processed translation, or DNT-processing is not acceptable translation strategy for the NE; PCD is correct in the baseline translation

Further examples illustrate these scores:

+1	**Original:** A week earlier, **Eastern** sued the Machinist and pilot unions
	Baseline translation: Неделей ранее, **Восточный~+1** ***преследуемый~+1*** перед Машинистом и экспериментальными союзами. ('A week earlier, Eastern$_{(ADJ)}$ (was) chased$_{(Participle)}$ before the Machinist and experimental unions')
	DNT-processed translation: Неделей ранее, **Eastern~+1** предъявил ***иск~+1*** Машинисту и экспериментальным союзам ('A week earlier, **Eastern**$_{(NOUN)}$ brought **suit**$_{(NOUN)}$ against the Machinist and experimental unions')

+0	**Original:** About 6,000 salaried workers are currently represented by the **United Auto Workers** union.
	Baseline translation: Приблизительно 6,000 оплачиваемых рабочих в настоящее время представлены Объединенным союзом Работников автомобильной промышленности***~0***. ('About 6,000 salaried workers are currently represented by the United union of Workers of automobile industry.')
	DNT-processed translation: Приблизительно 6,000 оплачиваемых рабочих в настоящее время представлены союзом United ***Auto~0*** Workers. ('About 6,000 salaried workers are currently represented by the union "United Auto Workers".')

–1	**Original:** **Treasury** Secretary James Baker held a 7 1/2-hour negotiating session with top Canadian officials.
	Baseline translation: **Министр~-1** финансов Джеймс Бакер проводил 7 1/2-часовых сессии ведения переговоров с высшими Канадскими должностными лицами ('The minister of finances James Baker held a 7 1/2-hour negotiating session with top Canadian officials..') – correct translation equivalent chosen
	DNT-processed translation: Секретарь**~-1** Treasury, Джеймс Бакер проводил 7 1/2-часовых сессии ведения переговоров с высшими Канадскими должностными лицами ('Secretary of "Treasury" James Baker held a 7 1/2-hour negotiating session with top Canadian officials.') – incorrect translation equivalent
	Original: The **Labour Department** has collected the statistics.
	Baseline translation: **Министерство~-1** труда**~-1**, собрало статистику. ('The Ministry of Labour has collected the statistics.')
	DNT-processed translation: **Labor~-1 Department~-1**, собрало статистику. ('The **Labor Department** has collected the statistics.') – unacceptable translation strategy

All differences highlighted in the whole MUC-6 NE corpus were manually annotated for each of the MT systems under consideration. Cases of morpho-syntactic differences were also annotated and excluded from the scored set of differences. The number of annotated differences is presented in Table 4:

	ProMT 1998 E-R	*ProMT 2001 E-F*	*Systran 2000 E-F*
Highlighted differences;	528	161	176
Including:			
diff.	61	3	2
scored lexical diff./Precision	467 (88.4%)	158 (98.1%)	174 (98.9%)

Table 4. Number of annotated differences

The larger number of differences and the lower Precision for the Russian system can be attributed to the largely synthetic morphology of Russian.

The overall score for improvement / decline in PCD for each MT system was calculated as a sum of all scores of lexical differences divided by the number of lexical differences for the particular system.

3.2.4. Results of the experiment for PCD

The set-up of this experiment gives a reasonable estimate of the influence of NE recognition on MT quality, and suggests that if improvement in MT can be achieved via pre-processing tools, then we can expect even greater improvement when an NE recognition module is properly integrated into MT systems (e.g., types of NEs requiring non-transference translation strategies are also distinguished). The improvement achieved for the MT systems under consideration was around 20%.

The results of manual annotation are summarised in Table 5:

	ProMT 1998 E-R		*ProMT 2001 E-F*		*Systran 2000 E-F*	
Mark	*N*	*Score*	*N*	*Score*	*N*	*Score*
+1*	154	+154	62	+62	77	+77
0*	239	0	66	0	61	0
–1*	74	–74	30	–30	36	–36
∑	467	**+ 80**	158	**+ 32**	174	**+ 41**
Gain	**+17.1%**		**+20.2%**		**+23.6%**	

Table 5 Scoring results

All systems showed consistent improvement in PCD tasks after NE recognition. The results indicate that systematic NE recognition has great potential for improving the quality of MT, and that successful PCD depends on appropriate analysis of other aspects in the source text, such as determining correct values for

morphological categories and correct syntactic segmentation. These aspects could be substantially improved via NE recognition.

However, finding appropriate segmentation and morpho-syntactic disambiguation is a necessary but not a sufficient condition for achieving improvement in MT: most cases of decline in MT quality after DNT-processing are due to the lack of flexibility in determining the optimal translation strategy for NEs. In our experiment, the overall improvement in the quality of PCD is due to the fact that the *transference* ("*do-not-translate*") strategy is optimal, or it is an acceptable translation strategy for the majority of NE that occurred in our corpus (Newmark, 1982). But many NEs might need to be translated by specific translation equivalents that are normally recognised by the state-of-the-art MT systems. This is especially important for names of well-known organisations, such as 'The Treasury', 'The Army', 'The Navy' 'Labour', which are often part of more complex NEs: 'The Treasury Secretary', 'The Labour Government', 'The Army Chief' – in all these cases a "do-not-translate" strategy could cause a serious decline in MT quality.

It is important to point out that this kind of problems usually doesn't require any additional dictionary update instead of DNT-processing. In most cases the necessary terms are already in system's dictionary. The great advantage of DNT processing against dictionary update is that it can be done completely automatically, using results of automatic NE annotation. DNT strategy for annotated items overrides existing strategies for lexical items in the ST, and more often happens to be right than wrong. However, still in many cases a general-purpose NE recognition system may over-generate DNT items for MT.

Therefore, this is not a problem of missing data, but the problem of competition of multiple equivalents for (which possibly implement different translation strategies) for the same segments in text.

What is needed is fine-grained distinction between different kinds of organisation names which require either the existing strategy and the use of existing system dictionaries, or the DNT strategy. In future MT performance can be improved even further by implementing new strategies, such as "transliterate" or "add a generic term" (e.g., "[+company] North Airlines", which may be a more natural way of translating such items into some TLs). IE techniques may generate annotation for items in the ST, which require indirect strategies, some kinds of translation transformations, etc. IE-type annotation may even be used to give different priorities to different user dictionaries "on-the-fly", depending on the required translation strategy for the same item in different contexts, or on automatically detected topic of the text, etc. However, this will require proper integration of IE modules into the source code of an MT system.

IE-type annotation gives MT much greater flexibility for changing translation strategies dynamically and resolving the issue of competition of equivalents for the same segments. It is introduced into the ST before translation is done, without making any prior translation commitments, therefore – without blocking the application of possibly useful equivalents, which happen to have lower lookup priorities. Another advantage is that unlike user-dictionary update, IE annotation is done completely automatically and triggers some universal strategies for items which are not in system dictionaries, (like the DNT or "transliterate" strategies for Organisation Names) greatly improving the coverage of some highly dynamic and open sets of such items. Therefore, IE-type annotation may increase flexibility of MT beyond dictionary update.

Our analysis suggests that targeting specific needs of MT could be a way of improving MT quality with IE technology: the NE recognition stage could meet the needs of MT systems by distinguishing different classes of NEs which require different translation strategies. Appropriate annotation of these NEs in the source text could then guide the MT system at the transfer stage.

3.2.5. Conclusions for the experiment

The potential improvement in PCD for MT systems has been characterised, which is achievable with accurate NE recognition. The results indicate that PCD is very sensitive to those aspects of MT quality which can be improved with NE recognition: finding appropriate morpho-syntactic categories and correct segmentation for NEs often influences the correctness of the general analysis of the source sentence. But some aspects of PCD cannot be improved with existing NE recognition and need to be addressed by the IE and MT communities jointly. NE recognition modules can be extended to distinguish between types of NEs that require different translation strategies; and MT systems can be adapted to deal more flexibly with user input, by using NE annotation designed specifically for MT purposes.

The proposed method of making a rough automatic distinction between lexical and morpho-syntactic differences allowed us to annotate important features in a relatively large corpus within a reasonable amount of time. It can be suggested that this method could have applications in other domains of NLP, in particular – in automated MT evaluation and in automatic alignment of parallel texts.

Future work in this direction could involve measuring the accuracy of the suggested method of distinguishing morpho-syntactic and lexical differences in MT output for typologically different languages and evaluating the degree of legitimate variation in translation at different levels of the significance scores.

Part III

Statistical IE techniques for automatic MT evaluation

Chapter 4
Improving the accuracy of reference-proximity methods of MT evaluation

4.1. Extending the BLEU MT Evaluation Method with Frequency Weightings

This section presents the results of an experiment on extending the automatic method of MT evaluation BLUE with statistical weights for lexical items, such as tf.idf and S-scores. We show that this extension gives additional information about evaluated texts; in particular it allows us to measure translation adequacy, which, for statistical MT systems, is often overestimated by the baseline BLEU method. The model suggests a linguistic interpretation which relates frequency weights and human intuition about translation *adequacy* and *fluency* (Babych and Hartley, 2004a).

4.1.1. Motivation for the experiment

Automatic methods for evaluating different aspects of MT quality – such as Adequacy, Fluency and Informativeness – provide an alternative to an expensive and time-consuming process of human MT evaluation. They are intended to yield scores that correlate with human judgments of translation quality and enable systems (machine or human) to be ranked on this basis. Several such automatic methods have been proposed in recent years since the seminal paper by (Brew and Thompson, 1994).

Among these methods the BLEU approach (Papineni et al., 2002) and its different modifications, e.g., NIST metric, have been the most popular. In contrast to the performance-based methods, it doesn't require expensive linguistic or processing resources (PoS-tagged or parsed data). The usability of other methods is seriously limited by the need to produce and maintain such resources, e.g., the RED method (Akiba et al., 2003) ranks texts and takes an edit-distance approach over PoS-tagged data which, it is claimed, handles long-distance co-occurrence and is less sensitive than BLEU to the choice of reference translations. However, the test suite and training data are expensive to produce. (Rajman and Hartley, 2001; 2002) propose a method combining syntactic relations and semantic vectors that dispenses with the need for reference translations but which requires parsed data and a large aligned training corpus.

The core assumption behind BLEU metric is the assumption of "reference proximity", that a good quality translation should be "close" (in some sense) to a professional human translation used as a reference. BLEU method computes this "closeness" as a match between N-gram models in MT output and in a set of human reference translations.

However, a serious problem for the BLEU method is the lack of a model for relative importance of matched and mismatched items. Words in text usually carry an unequal informational load, and as a result are of differing importance for translation. It is reasonable to expect that the choices of right translation equivalents for certain key items, such as expressions denoting principal events, event participants and relations in a text are more important in the eyes of human evaluators then choices of function words and a syntactic perspective for sentences. Accurate rendering of these key items by an MT system boosts the quality of translation. Therefore, at least for evaluation of translation Adequacy (Fidelity), the proper choice of translation equivalents for important pieces of information should count more than the choice of words which are used for structural purposes and without a clear translation equivalent in the source text. (The latter may be more important for Fluency evaluation).

It is sensible to suggest that a model of statistical salience of terms in text, which was developed in Chapter 2 to characterise usability of MT output for IE, could be used to account for such variable information load of terms in text and will boost the accuracy of the Adequacy evaluation

The problem of different significance of N-gram matches is related to the issue of legitimate variation in human translations, when certain words are less stable than others across independently produced human translations. BLEU accounts for legitimate translation variation by using a set of several human reference translations, which are believed to be representative of several equally acceptable ways of translating any source segment. This is motivated by the need not to penalise deviations from the set of N-grams in a single reference, although the requirement of multiple human references makes automatic evaluation more expensive. I discuss the problem of the legitimate translation variation in Chapter 5.

However, the "significance" problem is not directly addressed by the BLEU method. On the one hand, the matched items that are present in several human references receive the same weights as items found in just one of the references. On the other hand the model of legitimate translation variation cannot fully accommodate the issue of varying degrees of "salience" for matched lexical items, since alternative synonymic translation equivalents may also be highly significant for an adequate translation from the human perspective (Babych and Hartley,

2004b). Therefore it is reasonable to suggest that introduction of a model which approximates intuitions about the significance of the matched N-grams will improve the correlation between automatically computed MT evaluation scores and human evaluation scores for translation Adequacy.

In this section we present the result of an experiment on augmenting BLEU N-gram comparison with statistical weight coefficients which capture a word's salience within a given document: the standard tf.idf measure used in the vector-space model for Information Retrieval (Salton and Leck, 1968) and the S-score proposed in Chapter 2 for evaluating MT output corpora for the purposes of Information Extraction (see also Babych et al., 2003). Both scores are computed for each term in each of the 100 human reference translations from French into English available in DARPA-94 MT evaluation corpus (White et al., 1994).

The proposed weighted N-gram model for MT evaluation is tested on a set of translations by four different MT systems available in the DARPA corpus, and is compared with the results of the baseline BLEU method with respect to their correlation with human evaluation scores.

The scores produced by the N-gram model with tf.idf and S-Score weights are shown to be consistent with baseline BLEU evaluation results for Fluency and outperform the BLEU scores for Adequacy (where the correlation for the S-score weighting is higher). We also show that the weighted model may still be reliably used if there is only one human reference translation for an evaluated text.

Besides saving cost, the ability to dependably work with a single human translation has an additional advantage: it is now possible to create Recall-based evaluation measures for MT, which has been problematic for evaluation with multiple reference translations, since only one of the choices from the reference set is used in translation (Papineni et al. 2002:314). Notably, Recall of weighted N-grams is found to be a good estimation of human judgements about translation Adequacy. Using *weighted* N-grams is essential for predicting Adequacy, since correlation of Recall for non-weighted N-grams is much lower.

It is possible that other automatic methods which use human translations as a reference may also benefit from an introduction of an explicit model for term significance, since so far these methods also implicitly assume that all words are equally important in human translation, and use all of them, e.g., for measuring edit distances (Akiba et al, 2001; 2003).

The weighted N-gram model (WNM) has been implemented as an MT evaluation toolkit (which includes a Perl script, example files and documentation). It computes evaluation scores with tf.idf and S-score weights for translation Adequacy

and Fluency. The toolkit is available at http://www.comp.leeds.ac.uk/bogdan/evalMT.html (Babych and Hartley, 2004e).

4.1.2. Set-up of the experiment

The experiment used French–English translations available in the DARPA-94 MT evaluation corpus. The corpus contains 100 French news texts (each text is about 350 words long) translated into English by 5 different MT systems: "Systran", "Reverso", "Globalink", "Metal", "Candide" and scored by human evaluators; there are no human scores for "Reverso", which was added to the corpus on a later stage. The corpus also contains 2 independent human translations of each text. Human evaluation scores are available for each of the 400 texts translated by the 4 MT systems for 3 parameters of translation quality: "Adequacy", "Fluency" and "Informativeness". The *Adequacy* (Fidelity) scores are given on a 5-point scale by comparing MT with a human reference translation. The Adequacy parameter captures how much of the original content of a text is conveyed, regardless of how grammatically imperfect the output might be. The *Fluency* scores (also given on a 5-point scale) determine intelligibility of MT without reference to the source text, i.e., how grammatical and stylistically natural the translation appears to be. The *Informativeness* scores (which wasn't used for the current experiment) determine whether there is enough information in MT output to enable evaluators to answer multiple-choice questions on its content (White, 2003:237)

In the first stage of the experiment, each of the two sets of human translations was used to compute tf.idf and S-scores for each word in each of the 100 texts. The tf.idf score was calculated as described in Chapter 2.

In the second stage we carried out N-gram based MT evaluation, measuring Precision and Recall of N-grams in MT output using a single human reference translation. N-gram counts were adjusted with the tf.idf weights and S-scores for every matched word. The following procedure was used to integrate the S-scores / tf.idf scores for a lexical item into N-gram counts. For every word in a given text which received an S-score and tf.idf score on the basis of the human reference corpus, all counts for the N-grams containing this word are increased by the value of the respective score (not just by 1, as in the baseline BLEU approach). The original matches used for BLEU and the weighted matches are both calculated.

The weighted N-gram evaluation scores of Precision, Recall and F-measure may be produced for a segment, for a text or for a corpus of translations generated by an MT system.

In the third stage of the experiment the weighted Precision and Recall scores were tested for correlation with human scores for the same texts and compared to the results of similar tests for standard BLEU evaluation.

Finally we addressed the question whether the proposed MT evaluation method allows us to use a single human reference translation reliably. In order to assess the stability of the weighted evaluation scores with a single reference, two runs of the experiment were carried out. The first run used the "Reference" human translation, while the second run used the "Expert" human translation (each time a single reference translation was used). The scores for both runs were compared using a standard deviation measure.

4.1.3. The results of the MT evaluation with frequency weights

With respect to evaluating MT systems, the correlation for the weighted N-gram model was found to be stronger, for both Adequacy and Fluency, the improvement being highest for Adequacy. These results are due to the fact that the weighted N-gram model gives much more accurate predictions about the statistical MT system "Candide", whereas the standard BLEU approach tends to over-estimate its performance for translation Adequacy.

Table 1 present the baseline results for non-weighted Precision, Recall and F-score. It shows the following figures:

– Human evaluation scores for Adequacy and Fluency (the mean scores for all texts produced by each MT system);

– BLEU scores produced using 2 human reference translations and the default script settings (N-gram size = 4);

– Precision, Recall and F-score for the weighted N-gram model produced with 1 human reference translation and N-gram size = 4.

– Pearson's correlation coefficient *r* for Precision, Recall and F-score correlated with human scores for Adequacy and Fluency ***r(2)*** (with 2 degrees of freedom) for the sets which include scores for the 4 MT systems.

The scores at the top of each cell show the results for the first run of the experiment, which used the "Reference" human translation; the scores at the bottom of the cells represent the results for the second run with the "Expert" human translation.

System **[ade] / [flu]**	**BLEU** **[1&2]**	**Prec.** **1/2**	**Recall** **1/2**	**Fscore** **1/2**
CANDIDE 0.677 / 0.455	0.3561	0.4068 0.4012	0.3806 0.3790	0.3933 0.3898
GLOBALINK 0.710 / 0.381	0.3199	0.3429 0.3414	0.3465 0.3484	0.3447 0.3449
MS 0.718 / 0.382	0.3003	0.3289 0.3286	0.3650 0.3682	0.3460 0.3473
REVERSO *NA / NA*	0.3823	0.3948 0.3923	0.4012 0.4025	0.3980 0.3973
SYSTRAN 0.789 / 0.508	0.4002	0.4029 0.3981	0.4129 0.4118	0.4078 0.4049
*Corr **r(2)** with **[ade]** – MT*	0.5918	0.1809 0.1871	0.6691 0.6988	0.4063 0.4270
*Corr **r(2)** with **[flu]** – MT*	0.9807	0.9096 0.9124	0.9540 0.9353	**0.9836** **0.9869**

Table 1. Baseline non-weighted scores.

Table 2 summarises the evaluation scores for BLEU as compared to tf.idf weighted scores, and Table 3 summarises the same scores as compared to S-score weighed evaluation.

System **[ade] / [flu]**	**BLEU** **[1&2]**	**Prec.** **(w) 1/2**	**Recall** **(w) 1/2**	**Fscore** **(w) 1/2**
CANDIDE 0.677 / 0.455	0.3561	0.5242 0.5176	0.3094 0.3051	0.3892 0.3839
GLOBALINK 0.710 / 0.381	0.3199	0.4905 0.4890	0.2919 0.2911	0.3660 0.3650
MS 0.718 / 0.382	0.3003	0.4919 0.4902	0.3083 0.3100	0.3791 0.3798
REVERSO *NA / NA*	0.3823	0.5336 0.5342	0.3400 0.3413	0.4154 0.4165
SYSTRAN 0.789 / 0.508	0.4002	0.5442 0.5375	0.3521 0.3491	0.4276 0.4233
*Corr **r(2)** with **[ade]** – MT*	0.5918	0.5248 0.5561	0.8354 0.8667	0.7691 0.8119
*Corr **r(2)** with **[flu]** – MT*	0.9807	**0.9987** **0.9998**	0.8849 0.8350	0.9408 0.9070

Table 2. BLEU *vs* tf.idf weighted scores.

System **[ade] / [flu]**	**BLEU** **[1&2]**	**Prec.** **(w) 1/2**	**Recall** **(w) 1/2**	**Fscore** **(w) 1/2**
CANDIDE 0.677 / 0.455	0.3561	0.5034 0.4982	0.2553 0.2554	0.3388 0.3377
GLOBALINK 0.710 / 0.381	0.3199	0.4677 0.4672	0.2464 0.2493	0.3228 0.3252
MS 0.718 / 0.382	0.3003	0.4766 0.4793	0.2635 0.2679	0.3394 0.3437
REVERSO *NA / NA*	0.3823	0.5204 0.5214	0.2930 0.2967	0.3749 0.3782
SYSTRAN 0.789 / 0.508	0.4002	0.5314 0.5218	0.3034 0.3022	0.3863 0.3828
Corr ***r(2)*** *with* ***[ade]*** **–** *MT*	0.5918	0.6055 0.6137	**0.9069** **0.9215**	0.8574 0.8792
Corr ***r(2)*** *with* ***[flu]*** **–** *MT*	0.9807	0.9912 0.9769	0.8022 0.7499	0.8715 0.8247

Table 3. BLEU *vs* S-score weights.

It can be seen from the table that there is a strong positive correlation between the baseline BLEU scores and human scores for Fluency: $r(2)=0.9807,\ p<0.05$. However, the correlation with Adequacy is much weaker and is not statistically significant: $r(2)=0.5918,\ p>0.05$. The most serious problem for BLEU is predicting scores for the statistical MT system Candide, which was judged to produce relatively fluent, but largely inadequate translation. For other MT systems (developed with the knowledge-based MT architecture) the scores for Adequacy and Fluency are consistent with each other: more fluent translations are also more adequate. BLEU scores go in line with Candide's Fluency scores, but do not account for its Adequacy scores. When Candide is excluded from the evaluation set, *r* correlation goes up, but it is still lower than the correlation for Fluency and remains statistically insignificant: $r(1)=0.9608,\ p>0.05$. Therefore, the baseline BLEU approach fails to consistently predict scores for Adequacy.

Correlation figures between non-weighted N-gram counts and human scores are similar to the results for BLEU: the highest and statistically significant correlation is between the F-score and Fluency: $r(2)=0.9836,\ p<0.05$, $r(2)=0.9869,\ p<0.01$, and there is somewhat smaller and statistically significant correlation with Precision. This confirms the need to use *modified* Precision in the BLEU method that also in certain respect integrates Recall.

The proposed weighted N-gram model outperforms BLEU and non-weighted N-gram evaluation in its ability to predict Adequacy scores: weighted Recall scores have much stronger correlation with Adequacy (which for MT-only evaluation is

still statistically insignificant at the level $p<0.05$, but come very close to that point: $t=3.729$ and $t=4.108$; the required value for $p<0.05$ is $t=4.303$).

Correlation figures for S-score-based weights are higher than for tf.idf weights (*S-score*: $r(2)= 0.9069$, $p > 0.05$; $r(2)= 0.9215$, $p > 0.05$, *tf.idf score*: $r(2)= 0.8354$, $p >0.05$; $r(2)= 0.8667$, $p >0.05$).

The improvement in the accuracy of evaluation for the weighted N-gram model can be illustrated by the following example of translating the French sentence:

> **ORI-French:** *Les trente-huit chefs d'entreprise mis en examen dans le dossier ont déjà fait l'objet d'auditions, mais trois d'entre eux ont été confrontés, mercredi, dans la foulée de la confrontation "politique".*

English translations of this sentence by the knowledge-based system Systran and statistical MT system Candide have an equal number of matched unigrams (highlighted in italic), therefore conventional unigram Precision and Recall scores are the same for both systems. However, for each translation two of the matched unigrams are different (underlined) and receive different frequency weights (shown in brackets):

MT "Systran":

> *The thirty-eight* heads (tf.idf=**4.605**; S=**4.614**) *of* undertaking put in examination *in the* file *already* were the subject of hearings, *but three of them were* confronted, *Wednesday*, in *the* tread of "*political*" confrontation (tf.idf=**5.937**; S=**3.890**).

Human translation "Expert":

> ***The thirty-eight*** *heads* ***of*** companies questioned ***in the*** **case had** ***already*** been heard, ***but three of them were*** brought together ***Wednesday*** following ***the*** "***political***" *confrontation*.

MT "Candide":

> *The thirty-eight* counts *of* company put into consideration *in the* case (tf.idf=**3.719**; S=**2.199**) *already* had (tf.idf=**0.562**; S=**0.000**) the object of hearings, *but three of them were* checked, *Wednesday*, in *the* path of confrontal "*political*."

(In the human translation the unigrams matched by the Systran output sentence are in italic, those matched by the Candide sentence are in bold).

It can be seen from this example that the unigrams matched by Systran have higher term frequency weights (both tf.idf and S-scores):

heads (tf.idf=**4.605**;S=**4.614**)

confrontation (tf.idf=**5.937**;S=**3.890**)

The output sentence of Candide instead matched less salient unigrams:

case (tf.idf=**3.719**;S=**2.199**)

had (tf.idf=**0.562**;S=**0.000**)

Therefore for the given sentence weighted unigram Recall (i.e., the ability to avoid under-generation of salient unigrams) is higher for Systran than for Candide (Table 4):

	Systran	Candide
R	0.6538	0.6538
R * tf.idf	0.5332	0.4211
R * S-score	0.5517	0.3697
P	0.5484	0.5484
P * tf.idf	0.7402	0.9277
P * S-score	0.7166	0.9573

Table 4. Recall, Precision, and weighted scores

Weighted Recall scores capture the intuition that the translation generated by Systran is more adequate than the one generated by Candide, since it preserves more important pieces of information.

On the other hand, weighted Precision scores are higher for Candide. This is due to the fact that Systran over-generates (doesn't match in the human translation) much more "exotic", unordinary words, which on average have higher cumulative salience scores, e.g., *undertaking, examination, confronted, tread* – vs. the corresponding words "over-generated" by Candide: *company, consideration, checked, path*. In some respect higher weighted precision can be interpreted as higher Fluency of the Candide's output sentence, which intuitively is perceived as sounding more naturally (although not making much sense).

On the level of corpus statistics the weighted Recall scores go in line with Adequacy, and weighted Precision scores (as well as the Precision-based BLEU scores) – with Fluency, which confirms such interpretation of weighted Precision and Recall scores in the example above. On the other hand, Precision-based scores and non-weighted Recall scores fail to capture Adequacy.

The improvement in correlation for weighted Recall scores with Adequacy is achieved by reducing overestimation for the Candide system, moving its scores closer to human judgements about its quality in this respect. However, this is not

completely achieved: although in terms of Recall weighted by the S-scores Candide is correctly ranked below MS (and not ahead of it, as with the BLEU scores), it is still slightly ahead of Globalink, contrary to human evaluation results.

For both methods – BLEU and the Weighted N-gram evaluation – Adequacy is found to be harder to predict than Fluency. This is due to the fact that there is no good linguistic model of translation adequacy which can be easily formalised. The introduction of S-score weights may be a useful step towards developing such a model, since correlation scores with Adequacy are much better for the Weighted N-gram approach than for BLEU.

Also from the linguistic point of view, S-score weights and N-grams may only be reasonably good approximations of Adequacy, which involves a wide range of factors, like syntactic and semantic issues that cannot be captured by N-gram matches and require a thesaurus and other knowledge-based extensions. Accurate formal models of translation variation may also be useful for improving automatic evaluation of Adequacy.

The proposed evaluation method also preserves the ability of BLEU to consistently predict scores for Fluency: Precision weighted by tf.idf scores has the strongest positive correlation with this aspect of MT quality, which is slightly better than the values for BLEU; (*S-score: $r(2)= 0.9912$, $p<0.01$; $r(2)= 0.9769$, $p<0.05$; tf.idf score: $r(2)= 0.9987$, $p<0.001$; $r(2)= 0.9998$, $p<0.001$*).

The results suggest that weighted Precision gives a good approximation of Fluency. Similar results with non-weighted approach are only achieved if some aspect of Recall is integrated into the evaluation metric (either as *modified* precision, as in BLEU, or as an aspect of the F-score). Weighted Recall (especially with S-scores) gives a reasonably good approximation of Adequacy.

On the one hand using 1 human reference with uniform results is essential for our methodology, since it means that there is no more "trouble with Recall" (Papineni et al., 2002:314) – a system's ability to avoid under-generation of N-grams can now be reliably measured. On the other hand, using a single human reference translation instead of multiple translations will certainly increase the usability of N-gram based MT evaluation tools.

The fact that non-weighted F-scores also have high correlation with Fluency suggests a new linguistic interpretation of the nature of these two quality criteria: it is intuitively plausible that Fluency subsumes, i.e. presupposes Adequacy (similarly to the way the F-score subsumes Recall, which among all other scores gives the best correlation with Adequacy). The non-weighted F-score correlates more strongly with Fluency than either of its components: Precision and Recall; similarly

Adequacy might make a contribution to Fluency together with some other factors. It is conceivable that people need adequate translations (or at least translations that make sense) in order to be able to make judgments about naturalness, or Fluency.

Being able to make some sense out of a text could be the major ground for judging Adequacy: sensible mistranslations in MT are relatively rare events. This may be the consequence of a principle similar to the "second law of thermodynamics" applied to text structure, – in practice it is much rarer to some alternative sense to be created (even if the number of possible error types could be significant), than to destroy the existing sense in translation, so the majority of inadequate translations are just nonsense. However, in contrast to human translation, fluent mistranslations in MT are even rarer than disfluent ones, according to the same principle. A real difference in scores is made by segments which make sense and may or may not be fluent, and things which do not make any sense and about which it is hard to tell whether they are fluent.

This suggestion may be empirically tested: if Adequacy is a necessary precondition for Fluency, there should be a greater inter-annotator disagreement in Fluency scores on texts or segments which have lower Adequacy scores. This will be a topic of future research.

We note that for the DARPA corpus the correlation scores presented are highest if the evaluation unit is an entire corpus of translations produced by an MT system, and for text-level evaluation, correlation is much lower. A similar observation was made in (Papineni et al., 2002: 313). This may be due to the fact that human judges are less consistent, especially for puzzling segments that do not fit the scoring guidelines, like nonsense segments for which it is hard to decide whether they are fluent or even adequate. However, this randomness is leveled out if the evaluation unit increases in size – from the text level to the corpus level.

Automatic evaluation methods such as BLEU (Papineni et al., 2002), RED (Akiba et al., 2001), or the weighted N-gram model proposed here may be more consistent in judging quality as compared to human evaluators, but human judgments remain the only criteria for meta-evaluating the automatic methods.

4.1.4. Stability of weighted evaluation scores

In this section we investigate how reliable is the use of a single human reference translation. The stability of the scores is central to the issue of computing Recall and reducing the cost of automatic evaluation. We also compare the stability of our results with the stability of the baseline non-weighted N-gram model using a single reference.

It is essential to control the stability of the scores, since for pure IE-based MT evaluation the scores tend to become less stable, since only statistically salient lexical material is involved into evaluation. In this experiment IE-based evaluation is combined with the standard reference proximity model, which is intended to ensure that the stability of the weighted scores is comparable to the stability of the baseline reference proximity measures such as BLEU.

In this stage of the experiment we measured the changes that occur for the scores of MT systems if an alternative reference translation is used – both for the baseline N-gram counts and for the weighted N-gram model. Standard deviation was computed for each pair of evaluation scores produced by the two runs of the system with alternative human references. An average of these standard deviations is the measure of stability for a given score. The results of these calculations are presented in Table 5.

	systems	StDev-basln	StDev-tf.idf	StDev-S-score
P	**candide**	0.0040	0.0047	0.0037
	globalink	0.0011	0.0011	0.0004
	ms	0.0002	0.0012	0.0019
	reverso	0.0018	0.0004	0.0007
	systran	0.0034	0.0047	0.0068
	AVE SDEV	0.0021	0.0024	0.0027
R	**candide**	0.0011	0.003	0.0001
	globalink	0.0013	0.0006	0.0021
	ms	0.0023	0.0012	0.0031
	reverso	0.0009	0.0009	0.0026
	systran	0.0008	0.0021	0.0008
	AVE SDEV	0.0013	0.0016	0.0017
F	**candide**	0.0025	0.0037	0.0008
	globalink	0.0001	0.0007	0.0017
	ms	0.0009	0.0005	0.0030
	reverso	0.0005	0.0008	0.0023
	systran	0.0021	0.0030	0.0025
	AVE SDEV	0.0012	0.0018	0.0021

Table 5. Stability of scores

As it was expected, standard deviation for weighted scores is generally slightly higher in comparison to the baseline BLEU scores, but both the baseline and the weighted N-gram approaches give relatively stable results: the average standard deviation was not greater than 0.0027, which means that both will produce reliable

figures with just a single human reference translation (although interpretation of the score with a single reference should be different than with multiple references).

Somewhat higher standard deviation figures for the weighted N-gram model also confirm the suggestion that a word's importance for translation cannot be straightforwardly derived from the model of the legitimate translation variation implemented in BLEU and needs the salience weights, such as tf.idf or S-scores.

4.1.5. Interpretation of significance weights for MT evaluation

The model of weighted N-grams distinguishes different possible "angles" of matches between the evaluated text and the human reference, which can be visualised by the two diagrams in Figure 6. On the diagrams the set of reference N-grams is divided into the following subsets:

H – N-grams that have high salience scores ("high" N-grams);

L – N-grams that have low salience scores ("low" N-grams);

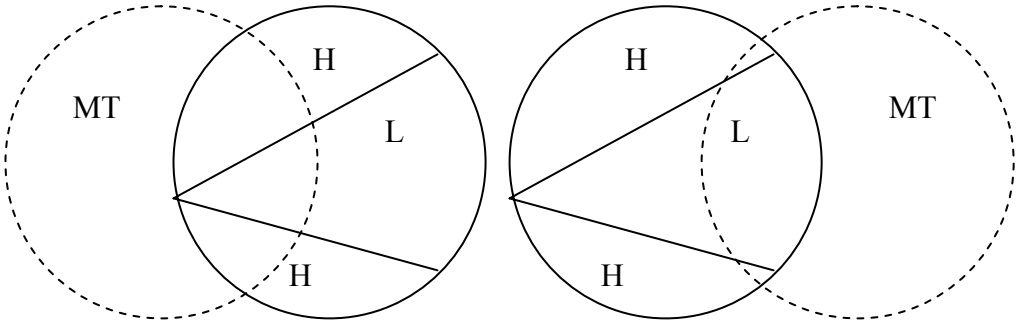


Figure 6. Distinguishing different types of matches

N-grams from the MT output intersect with the reference set at different angles. For instance, even if total count of the matched N-grams is the same, the output of knowledge-based MT systems probably intersects with more important, i.e., higher scored N-grams, while the output of statistical MT intersects with the set of reference N-grams "from the other side", where mostly low score N-grams are matched.

The proposed model provides a framework for fine-tuning MT evaluation tools with respect to different quality criteria, allowing the user to modify the range and magnitude of the significance scores involved in MT evaluation.

4.1.6. Conclusion of the experiment

The results for weighted N-gram models have a significantly higher correlation with human intuitive judgements about translation Adequacy and Fluency than the baseline N-gram evaluation measures which are used in the BLEU MT evaluation toolkit. This shows that they are a promising direction of research. Future work will apply our approach to evaluating MT into languages other than English, extending the experiment to a larger number of MT systems built on different architectures and to larger corpora.

However, the results of the experiment may also have implications for MT development: significance weights may be used to rank the relative "importance" of translation equivalents. At present all MT architectures (knowledge-based, example-based, and statistical) treat all translation equivalents equally, so MT systems cannot dynamically prioritise rule applications, and translations of the central concepts in texts are often lost among excessively literal translations of less important concepts and function words. For example, for statistical MT significance weights of lexical items may indicate which words have to be introduced into the target text using the *translation model* for source and target languages, and which need to be brought there by the *language model* for the target corpora. Similar ideas may be useful for the Example-based and Rule-based MT architectures. The general idea is that different pieces of information expressed in the source text are not equally important for translation: MT systems that have no means for prioritising this information often introduce excessive information noise into the target text by literally translating structural information, etymology of proper names, collocations that are unacceptable in the target language, etc. This information noise often obscures important translation equivalents and prevents the users from focusing on the relevant bits. MT quality may benefit from filtering out this excessive information as much as from frequently recommended extension of knowledge sources for MT systems. The significance weights may schedule the priority for retrieving translation equivalents and motivate application of compensation strategies in translation, e.g., adding or deleting implicitly inferable information in the target text, using non-literal strategies, such as transposition or modulation (Vinay and Darbelnet, 1995). Such weights may allow MT systems to make an approximate distinction between salient words which require proper translation equivalents and structural material both in the source and in the target texts. Exploring applicability of this idea to various MT architectures is another direction for future research.

4.2. Calibrating resource-light automatic MT evaluation metrics

In the previous section the proposed WNM metric was developed and tested on DARPA 94 MT evaluation corpus. In this section it is applied to a new material and tested for correlation with an additional quality parameter – usability of business email texts. The problems of capturing improvements in MT quality after the dictionary update and of estimating the quality of a non-native human translation with the automated MT evaluation metric are also addressed in this section.

The WNM metric is based on IE techniques, It has shown high correlation with Adequacy on DARPA94 MT evaluation corpus. However, to prove its usability we need to calibrate it on some different material, possibly on some smaller text. Calibration will involve collecting human evaluation scores for different quality parameters and measuring correlation between these scores and automated metrics – WNM and BLEU (Babych et al., 2004a).

In this calibration experiment a novel parameter – usability (or utility) of output was tested. It was found to integrate both fluency and adequacy. two automated metrics, BLEU and WNM, with new data for which human evaluation scores were also produced; we then measured the agreement between the automated and human evaluation scores. The resources produced in the experiment are available on the same website (Babych and Hartley, 2004e).

4.2.1. Motivation for the experiment

A comparative evaluation of two mature knowledge-based MT systems was performed, based on human judgments of three quality attributes, designed to calibrate the two automatic methods – BLEU and WNM. Both the BLEU and the WNM metrics were applied to a corpus of business texts translated from French into English by two mature knowledge-based MT systems, with a view to scoring the systems. The experiment also tested whether the quality of the translations was judged by humans to be improved by updating the dictionaries of each system in line with a benchmark provided by a human translation, and whether the automated metrics would capture this perception.

4.2.1.1. Automatic evaluation – BLEU method

The BLEU automatic evaluation metric has been shown to strongly correlate with human judgements about fluency of knowledge-based MT systems, which is also confirmed by the results presented here. The BLEU method is based on matches of N-grams (individual words or sequences of several words, usually up to 4) in MT and in one or more human “gold standard” reference translations. More specifically,

BLEU measures N-gram precision (the proportion of N-grams found both in MT output and in any of the "gold standard" human reference translations).

The rationale of using BLEU is to explore objective properties of the evaluated texts as compared to a gold standard human reference translation. This gives an "absolute" measure for comparison across different evaluation attributes, e.g. fluency, adequacy and usability, which are not directly comparable through human scoring. The BLEU scores are in the range [0…1].

4.2.1.2. Automatic evaluation – WNM method

The WNM method as described in Section 3.1 is based on BLEU, but the matched words in the tested MT output and the "gold standard" translation have unequal weight when they are matched. More weight is given to statistically salient words in the evaluated text. Statistical significance weights, suggested in Chapter 2 are computed by contrasting the word's frequency in a text and in the rest of the corpus

Usually the content words, names of events, event participants, and terminology happen to be more statistically significant. The intuition is that matches of the "significant" words should count more, when the MT output is evaluated, which is captured in WNM method by assigning greater weights to words whose statistical salience S-score is >1.

WNM computes three scores for each evaluated document: Precision (or degree of avoiding "over-generation" of "significant" words), recall (or degree of avoiding "under-generation") and F-score, where precision and recall are weighted equally. In our previous experiments with the DARPA corpus, *recall* was found to be the best match for *adequacy*, and the *F-score* for *fluency*.

4.2.2. Calibrating BLEU and WNM

4.2.2.1. Set up of the experiment

The French-to-English versions of two leading commercial MT systems – System 1 and System 2 were evaluated in order to assess the quality of their output and to determine whether updating the system dictionaries brought about an improvement in performance.

The input for the evaluation were a whitepaper (3,334 words in 120 segments) from the European Commission and a collection of 36 business and personal emails (average length 107 words). Translations of all the texts by a professional translator were also available. These were used as a gold standard reference for creating new dictionary entries. These human translations also figured in the evaluation exercise.

For the emails, translations were produced by a non-professional, French-speaking translator. This was intended to simulate a situation where, in the absence of MT, the author of the email would have to write in a foreign language (here English). We anticipated that the quality would be judged lower than the professional, native speaker translations.

The evaluations were performed by 70 judges – 42 business people and 28 postgraduate students who knew very little or no French.

Using a five-point scale in each case, judgments were solicited on three attributes of text quality by means of the following questions:

Usability – "Using each reference email on the left, rate the three alternative versions on the right according to **how usable you consider them to be for getting business done**." The non-native translations were dispersed anonymously in the data set and so were also judged.

Fluency – "Look carefully at each segment of text and give each one a score according to how much you think the text reads like fluent English written by a native speaker." No reference text was seen.

Adequacy – "For each segment, read carefully the reference text on the left. Then judge how much of the same content you can find in the candidate text."

Five independent judgments were collected for each segment and for each email.

4.2.2.2. Human evaluation results

Figure 1 and Table 1 summarise the results of human evaluation for 3 different evaluation tasks:

1. Fluency of the whitepaper translations (the 2 MT systems before and after dictionary update), judged by students (40%) and business users (60%) – FLU.

2. Adequacy of the whitepaper translation (the 2 MT systems before and after dictionary update), judged by students – ADE.

3. Usability of the email translations (the 2 MT systems before and after dictionary update and a non-native speaker translation), judged by business users – USL.

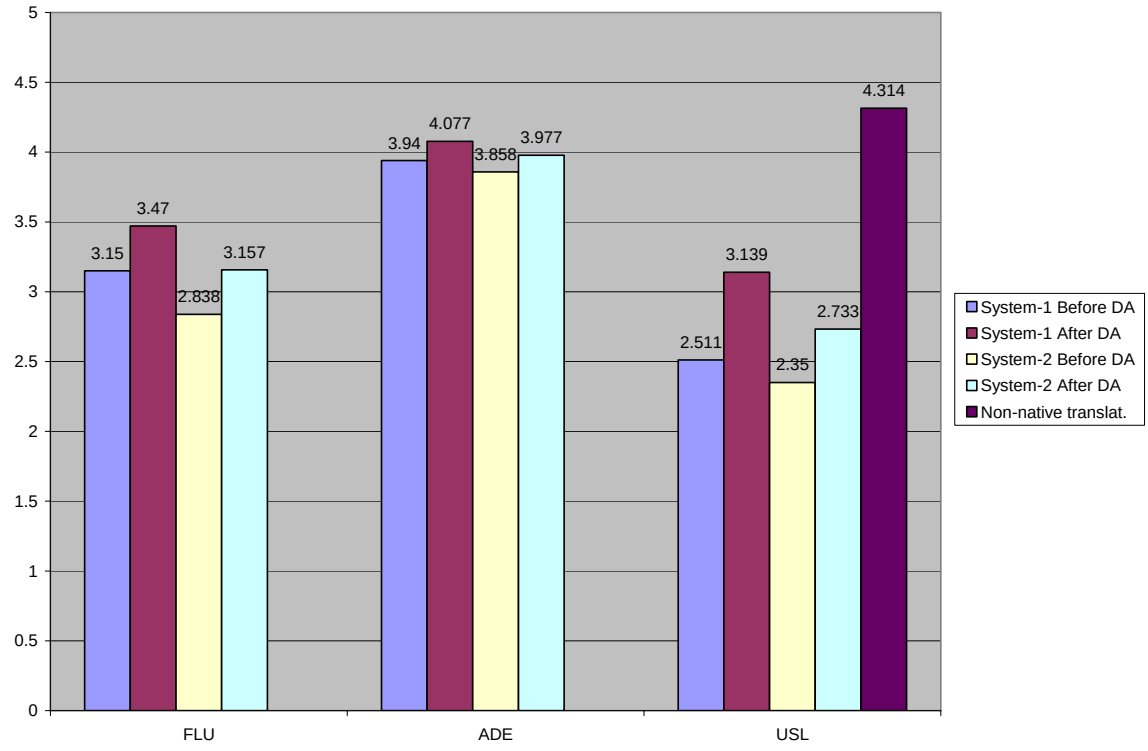


Figure 1. Human evaluation results

	FLU	ADE	USL
System-1 Before DA	3.150	3.94	2.511
System-1 After DA	3.470	4.077	3.139
System-2 Before DA	2.838	3.858	2.350
System-2 After DA	3.157	3.977	2.733
Non-native translation			4.314

Table 1. Human evaluation results

It can be seen from the figures that the results for adequacy are very high: on average MT systems scored "four" on the five-point scale. The results for fluency are worse: "three" on the five-point scale is the most likely score for MT systems. This shows that MT is useful primarily for "assimilation", i.e., "understanding" purposes, where the users try to grasp the meaning, and are less interested in getting well-formed, i.e., grammatically and lexically impeccable and stylistically natural sentences (which might be important for "dissemination", e.g., publication purposes – for these tasks MT is still not so good).

On the other hand, usability most probably has integrated "fluency" and "adequacy" aspects of the text quality (and perhaps has been influenced by the presence of the non-native human translation). It is natural to suggest that the text

which is easier to read requires less effort on the part of the user to reconstruct the meaning. From the point of view of usability, fluency and adequacy MT errors aggravate each other, so the scores for usability are lower than for the other two attributes.

All human scores for texts *after dictionary update* are consistently higher both for System 1 and for System 2, but the degree of improvement is different: it is the biggest for usability of the e-mail translations (25% for System 1 and 16% for System 2), and the smallest for adequacy of the whitepaper translation (3.5% for System 1 and 3.1% for System 2).

4.2.2.3. Automatic evaluation results

The results of BLEU evaluation for the whitepaper document and for emails are summarised in Figure 2. BLEU used a single human reference translation and counted N-grams up to N=4.

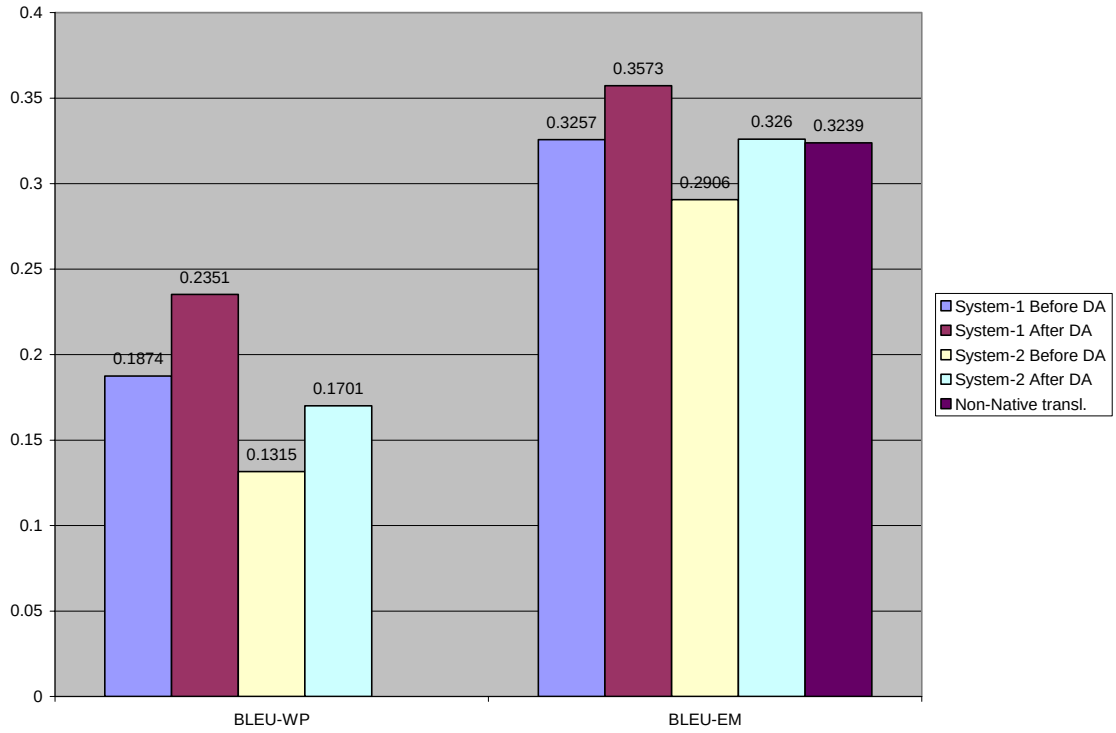


Figure 2. BLEU evaluation: whitepaper and emails

	BLEU-WP	BLEU-EM
System-1 Before DA	0.1874	0.3257
System-1 After DA	0.2351	0.3573
System-2 Before DA	0.1315	0.2906
System-2 After DA	0.1701	0.3260
Non-Native transl.		0.3239

Table 2. BLEU evaluation: whitepaper and emails

Another aspect of the BLEU evaluation is a possible comparison between the whitepaper text and in the business emails. There are many more matches of N-grams in the emails as compared to the whitepaper. Table 3 summarises the growth of matches between these two types of documents.

System-1 Before DA	0.737994
System-1 After DA	0.519779
System 2 Before DA	1.209886
System 2 After DA	0.916520

Table 3. Percentage growth of N-gram matches in the emails over the whitepaper

The table shows that translating emails is objectively easier for MT systems than translating the legal documents (this may be also true for human translators, although it is hard to find formal parameters to empirically test this suggestion). However, human judges adjust the scores according to the evaluation task, so the difference becomes apparent only with automatic evaluation. In this experiment, since the human non-native translation was involved in usability evaluation of the emails, a kind of "masking effect" was introduced, so the scores for usability were lower than for adequacy or fluency (where there was no comparison with the human translation). Therefore the BLEU score allows us to make comparison between different types of texts, which were not directly compared in our evaluation and shows that translating emails is easier for MT systems and much better results are objectively achievable, in comparison to the legal documents.

Also in Table 3 the difference between the whitepaper and the email matches for System 1 is lower than for System 2 (74% and 52% vs. 121% and 91%). This shows that System 1 translation gives more stable quality across genres, and the quality for System 2 is more dependent on the genre of the translated text: it achieves its quality is greatly improved for "easier" texts, such as emails as compared to the "hard" texts.

Figure 3 and Table 4 summarise the WNM evaluation results.

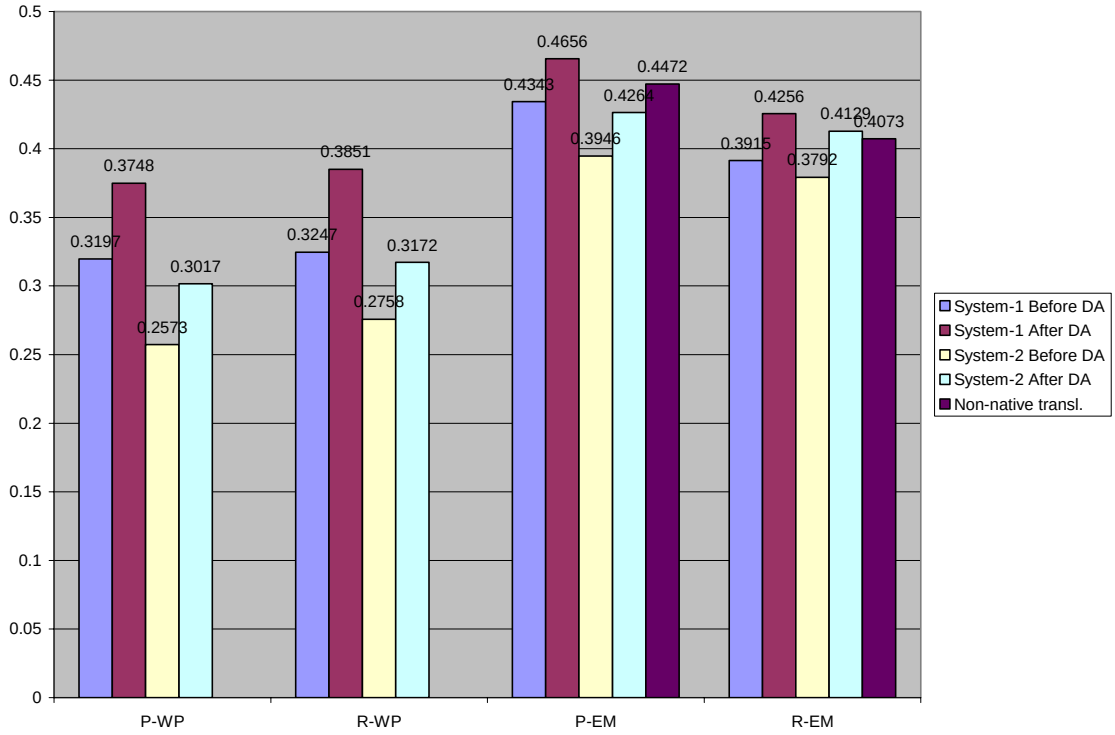


Figure 3. WNM scores – Precision and Recall

	P-WP	**R-WP**	**F-WP**
System-1 Before DA	0.3197	0.3247	0.3222
System-1 After DA	0.3748	0.3851	0.3799
System-2 Before DA	0.2573	0.2758	0.2663
System-2 After DA	0.3017	0.3172	0.3093
	P-EM	**R-EM**	**F-EM**
System-1 Before DA	0.4343	0.3915	0.4118
System-1 After DA	0.4656	0.4256	0.4447
System-2 Before DA	0.3946	0.3792	0.3868
System-2 After DA	0.4264	0.4129	0.4196
Non-native transl.	0.4472	0.4073	0.4263

Table 4. WNM scores

BLEU and WNM agree with human judgments with respect to ranking the two systems, although they differ in their precise scores. Results after dictionary update are better than before the update, and scores for System 1 are somewhat higher than for System 2; however, System 2 is shown to be capable of reaching System 1's baseline quality (the quality "before update") after its dictionary has been updated. The ratios of improvement and ratios of differences between systems are close to the ratios for human evaluation. This is an indication that human intuitive judgments

about fluency, adequacy and usability of MT quality across systems and before and after the dictionary update are confirmed by the objective criteria: precision of N-gram matches in MT and the "gold standard" translation.

An important difference between the two automated metrics and the human evaluation results is the score for the non-native translation: BLEU seriously underestimates the quality of the human translation, WNM slightly less so. The explanation for this fact could be that for knowledge-based MT and for native-speaker human translations there is a close match between the adequacy and fluency of translation, but this is not the case for non-native translation (as well as for the output of statistical MT systems, see (Babych, et al., 2003)). Therefore, the N-gram precision is not a good model for usability of Non-native human translations, which doesn't use similar words that are required in "natural" English, and doesn't sufficiently match the N-grams in the "gold standard" translation, but nevertheless "makes sense" for the readers of the text. The second aspect of the explanation could be that BLEU is a much better measure for fluency than for adequacy; usability of emails, supposedly, has stronger links with the latter than with the former.

BLEU and WNM also indicate that emails are easier for MT than the whitepaper text: the absolute evaluation scores of both automated methods are higher for the emails.

WNM measures both precision and recall, so we may see that the recall measure is more stable across "easy" and "hard" texts, while precision changes much more if the type of the text changes. "Harder" texts, such as the whitepaper legal documents usually cause much greater over-generation of N-grams, but the under-generation of N-grams changes to a much smaller extent.

4.2.2.4. Correlation between automatic and human evaluation scores

Table 5 summarises correlation between automatic scores – BLEU and WNM and the human evaluation scores. The WNM and BLEU scores which previously have been found to closely correlate with corresponding human evaluation measures are underlined.

	WNM-P-WP	WNM - R-WP	WNM - F-WP
cFLU	0.984809	0.989558	0.988328
cADE	0.949595	0.970463	0.960599
	WNM - P-EM	WNM - R-EM	WNM - F-EM
cUSL/MT	0.905698	0.967349	0.969011
cUSL/MT+HT	0.593061	0.475047	0.562204
	BLEU-WP	BLEU-EM	
cFLU	0.982683		
cADE	0.945306		
cUSL/MT		0.933908	
cUSL/MT+HT		0.333796	

Table 5. Correlation between automatic and human evaluation scores

The chart shows that although BLEU provides scores which correlate closely with human judgments, especially for fluency, WNM outperforms BLEU for all the measured scores. The greatest advantage of the WNM is for adequacy and usability. Usability scores were not part of previous experiments, but the closest match for it is the WNM F-score, and the WNM Recall comes very close behind it.

4.2.3. Conclusions from the experiment

The following conclusions can be drawn:

1. Both automatic methods capture quality increase after dictionary update and rank systems correctly, in line with human judgments about MT quality.

2. The WNM method measures both precision and recall of N-gram matches, which allows flexible evaluation of different aspects of MT quality, such as adequacy and usability.

3. The usability metric integrates elements of adequacy and fluency, as is reflected in both human and automatic evaluation scores.

Chapter 5
Extending flexibility of MT evaluation techniques

This chapter deals with several open research questions in MT evaluation, which are relevant for the topics discussed in Part III. Section 5.1 addresses the question of evaluating texts of different genres and designing a knowledge-light metric for estimating translation complexity and the issues of evaluating homogenous vs. heterogeneous text collections. Section 5.2 examines the problem of how big a text collection has to be before BLEU or WNM evaluation scores become informative. Section 5.3 discusses the relation between statistical salience of terms and their legitimate variation.

The results of the presented experiments have interesting theoretical implications for IE-guided MT, since they point out to some non-obvious phenomena in translation, which can be used to improve MT quality.

5.1. Extending MT evaluation tools with translation complexity metrics

This section reports on the results of an experiment in designing resource-light metrics that predict the potential translation complexity of a text or a corpus of homogenous texts for state-of-the-art MT systems. We show that the best prediction of translation complexity is given by the average number of syllables per word (ASW). The translation complexity metrics based on this parameter are used to normalise automated MT evaluation scores such as BLEU and the IE-oriented metric based on the WNM, which otherwise are variable across texts of different types. The suggested approach makes a fairer comparison between the MT systems evaluated on different corpora (Babych et al., 2004b).

5.1.1. Motivation for the experiment

Automated evaluation is much quicker and cheaper than human evaluation. Another advantage of the scores produced by automated MT evaluation tools is that intuitive human scores depend on the exact formulation of an evaluation task, on the granularity of the measuring scale and on the relative quality of the presented translation variants: human judges may adjust their evaluation scale in order to discriminate between slightly better and slightly worse variants – but only those variants which are present in the evaluation set. For example, absolute figures for a human evaluation of a set which includes MT output only are not directly

comparable with the figures for another evaluation which might include MT plus a non-native human translation, or several human translations of different quality. Because of the instability of this intuitive scale, human evaluation figures should be treated as relative rather than absolute. They capture only a local picture within an evaluated set, but not the quality of the presented texts in a larger context. Although automated evaluation scores are always calibrated with respect to human evaluation results, only the relative performance of MT systems within one particular evaluation exercise provide meaningful information for such calibration.

In this respect, automated MT evaluation scores have some added value: they rely on objective parameters in the evaluated texts, so their results are comparable across different evaluations.

Furthermore, they are also comparable for different types of texts translated by the same MT system, which is not the case for human scores. For example, automated scores are capable of distinguishing improved MT performance on easier texts or degraded performance on harder texts, so the automated scores also give information on whether one collection of texts is easier or harder than the other for an MT system: the complexity of the evaluation task is directly reflected in the evaluation scores.

However, there may be a need to avoid such sensitivity. MT developers and users are often more interested in scores that would be stable across different types of texts for the same MT system, i.e., would reliably characterise a system's performance irrespective of the material used for evaluation. Such characterisation is especially important for state-of-the-art commercial MT systems, which typically target a wide range of general-purpose text types and are not specifically tuned to any particular genre, like weather reports or aircraft maintenance manuals.

The typical problem of having "task-dependent" evaluation scores (which change according to the complexity of the evaluated texts) is that the reported scores for different MT systems are not directly comparable. Since there is no standard collection of texts used for benchmarking all MT systems, it is not clear how a system that achieves, e.g., BLEUr4n4[4] score 0.556 tested on "490 utterances selected from the WSJ" (Cmejrek et al, 2003:89) may be compared to another system which achieves, e.g., the BLEUr1n4 score 0.240 tested on 10,150 sentences from the "Basic Travel Expression Corpus" (Imamura et al., 2003:161).

---

[4] BLEUrXnY means the BLEU score with produced with X reference translations and the maximum size of compared N-grams = Y.

Moreover, even if there is no comparison involved, there is a great degree of uncertainty in how to interpret the reported automated scores. For example, BLEUr2n4 0.3668 is the highest score for a top MT system if MT performance is measured on news reports, but it is a relatively poor score for a corpus of e-mails, and a score that is still beyond the state-of-the-art for a corpus of legal documents. These levels of perfection have to be established experimentally for each type of text, and there is no way of knowing whether some reported automated score is better or worse if a new type of text is involved in the evaluation.

The need to use stable evaluation scores, normalised by the complexity of the evaluated task, has been recognised in other NLP areas, such as anaphora resolution, where the results may be relative with regard to a specific evaluation set. So "more absolute" figures are obtained if we use some measure which quantifies the complexity of anaphors to be resolved (Mitkov, 2002).

MT evaluation is harder than evaluation of other NLP tasks, which makes it partially dependent on intuitive human judgements about text quality. However, automated tools are capable of capturing and representing the "absolute" level of performance for MT systems, and this level could then be projected into task-dependent figures for harder or easier texts. In this respect, there is another "added value" in using automated scores for MT evaluation.

Stable evaluation scores could be achieved if a formal measure of a text's complexity for translation could be cheaply computed for a source text. Firstly, the score for translation complexity allows the user to predict "*absolute*" performance figures of an MT system on harder or easier texts, by computing the "absolute" evaluation figures and the complexity scores for just one type of text. Secondly, it lets the user compute "*standardised*" performance figures for an MT system that do not depend on the complexity of a text (they are reliably within some relatively small range for any type of evaluated texts).

Designing such standardised evaluation scores requires choosing a point of reference for the complexity measure: e.g., one may choose an average complexity of texts usually translated by MT as the reference point. Then the absolute scores for harder or easier texts will be corrected to fit the region of absolute scores for texts of average complexity.

This section reports on the results of an experiment in measuring the complexity of translation tasks using resource-light parameters such as the average number of syllables per word (ASW), which is also used for computing the readability of a text. On the basis of these parameters normalised BLEU and WNM scores are computed, which are relatively stable across translations produced by the

same general-purpose MT systems for texts of varying difficulty. This suggests that further testing and fine-tuning of the proposed approach on larger corpora of different text types and using additional source text parameters and normalisation techniques can give better prediction of translation complexity and increase the stability of the normalised MT evaluation scores.

5.1.2. Set-up of the experiment

We compared the results of the human and automated evaluation of translations from French into English of three different types of texts which vary in size and style: an EU whitepaper on child and youth policy (120 sentences), a collection of 36 business and private e-mails and 100 news texts from the DARPA 94 MT evaluation corpus (White et al., 1994). The translations were produced by two leading commercial MT systems. Human evaluation results are available for all of the texts, with the exception of the news reports translated by System-2, which was not part of the DARPA 94 evaluation. However, the human evaluation scores were collected at different times under different experimental conditions using different formulations of the evaluation tasks, which leads to substantial differences between human scores across different evaluations, even if the evaluations were done at the same time.

Further, two sets of automated scores were produced: BLEUr1n4, which have a high correlation with human scores for fluency, and WNM Recall, which strongly correlate with human scores for adequacy. These scores were produced under the same experimental conditions, but they uniformly differ for both evaluated systems: BLEU and WNM scores were relatively higher for e-mails and relatively low for the whitepaper, with the news texts coming in between. These differences could be interpreted as reflecting the relative complexity of texts for translation.

For the French originals of all three sets of texts resource-light parameters are computed, which are used in standard readability measures (Flesch Reading Ease score or Flesch-Kincaid Grade Level score), i.e. average sentence length (ASL – the number of words divided by the number of sentences) and average number of syllables per word (ASW – the number of syllables divided by the number of words).

Pearson's correlation coefficient *r* was computed between the automated MT evaluation scores and each of the two readability parameters. Differences in the ASL parameter were not strongly linked to the differences in automated scores, but for the ASW parameter a strong negative correlation was found.

Finally, normalised ("absolute") BLEU and WNM scores were computed using the automated evaluation results for the DARPA news texts (the medium

complexity texts) as a reference point. We compared the stability of these scores with the stability of the standard automated scores by computing standard deviations for the different types of text. The absolute automated scores can be computed on any type of text and they will indicate what score is achievable if the same MT system runs on DARPA news reports. The normalised scores allow the user to make comparisons between different MT systems evaluated on different texts at different times. In most cases the accuracy of the comparison is currently limited to the first rounded decimal point of the automated score.

5.1.3. Results of human evaluations

The human evaluation results were produced under different experimental conditions. The output of the compared systems was evaluated each time within a different evaluation set, in some cases together with different MT systems, or native or non-native human translations. As a result human evaluation scores are not comparable across different evaluations.

Human scores available from the DARPA 94 MT corpus of news reports were the result of a comparison of five MT systems (one of which was a statistical MT system) and a professional ("expert") human translation. For the current experiment DARPA scores for adequacy and fluency for one of the participating systems were used.

Human scores for translations of the whitepaper and the e-mails were obtained from one of our MT evaluation projects at the University of Leeds. This had involved the evaluation of French-to-English versions of two leading commercial MT systems – System 1 and System 2 – in order to assess the quality of their output and to determine whether updating the system dictionaries brought about an improvement in performance. (An earlier version of System 1 also participated in the DARPA evaluation.) Although the human evaluations of both texts were carried out at the same time, the experimental set-up was different in each case.

The evaluation of the whitepaper for *adequacy* was performed by 20 postgraduate students who knew very little or no French. A professional human translation of each segment was available to the judges as a gold standard reference. Using a five-point scale in each case, judgments were solicited on adequacy by means of the following question:

"*For each segment, read carefully the reference text on the left. Then judge how much of the same content you can find in the candidate text*."

Five independent judgments were collected for each segment.

The whitepaper *fluency* evaluation was performed by 8 postgraduate students and 16 business users under similar experimental conditions with the exception that the gold standard reference text was not available to the judges. The following question was asked:

"*Look carefully at each segment of text and give each one a score according to how much you think the text reads like fluent English written by a native speaker*."

For e-mails a different quality evaluation parameter was used: 26 human judges (business users) evaluated the *usability* (or *utility*) of the translations. Also translations produced by a non-professional, French-speaking translator were included in the evaluation set for e-mails. (This was intended to simulate a situation where, in the absence of MT, the author of the e-mail would have to write in a foreign language (here English); we anticipated that the quality would be judged lower than the professional, native speaker translations.) The non-native translations were dispersed anonymously in the data set and so were also judged. The following question was asked:

"*Using each reference e-mail on the left, rate the three alternative versions on the right according to* ***how usable you consider them to be for getting business done***."

Figure 1 and Table 1 summarise the human evaluation scores for the two compared MT systems. The judges had scored versions of the e-mails ("em") and whitepaper ("wp") produced both before and after dictionary update ("DA"), although no judge saw the before and after variants of the same text. (The scores for the DARPA news texts are converted from [0…1] into [0…5] scale).

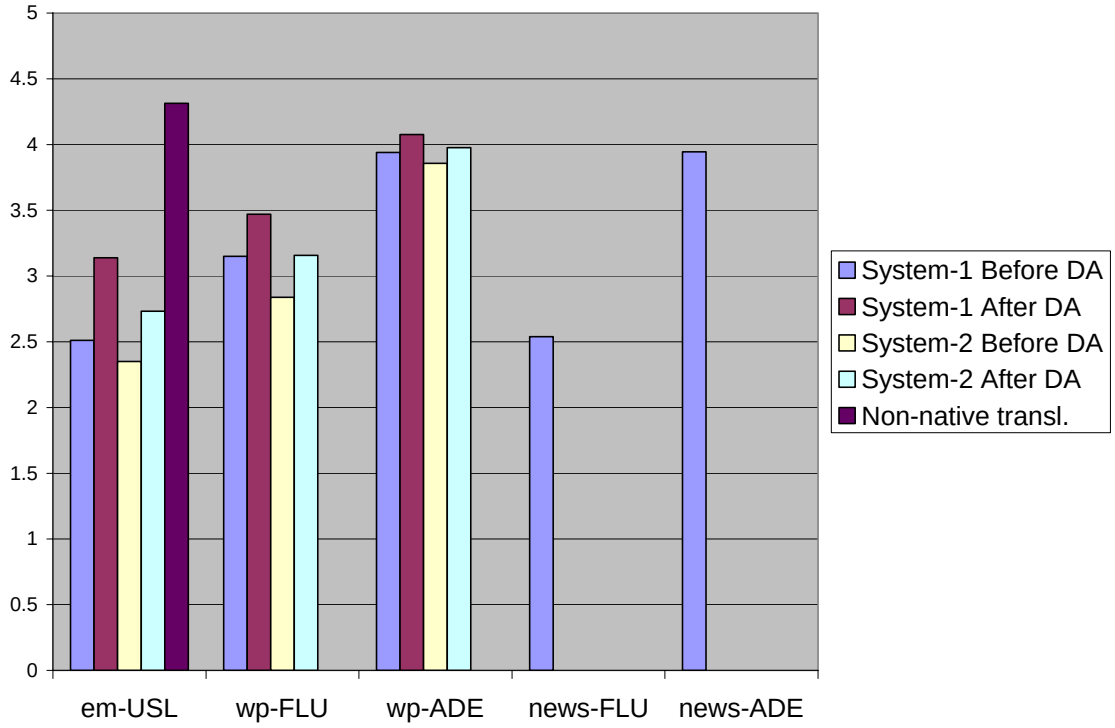


Figure 1. Human evaluation results

	S1	S1da	S2	S2da	NN
em [usl]	2.511	3.139	2.35	2.733	4.314
wp [flu]	3.150	3.47	2.838	3.157	
wp [ade]	3.940	4.077	3.858	3.977	
news [flu]	2.540				
news [ade]	3.945				

Table 1. Human evaluation scores

It can be inferred from the data that human evaluation scores do not allow us to make any meaningful comparison of the scores outside a particular evaluation experiment, which necessarily must be interpreted as relative rather than absolute.

We can see that dictionary update consistently improves the performance of both systems, that System 1 is slightly better than System 2 in all cases, although after dictionary update System 2 is capable of reaching the baseline quality of System 1. However, the usability scores for supposedly easier texts (e-mails) are considerably lower than the adequacy scores for harder texts (the whitepaper), although the experimental set-up for adequacy and usability is very similar: both used a gold-standard human reference translation. This suggests that the presence of a higher quality translation done by a human non-native speaker of the target language "over-shadowed" lower quality MT output, which dragged down

evaluation scores for e-mail usability. No such higher quality translation was present in the evaluation set for the whitepaper adequacy, so the scores went up.

Therefore, no meaning can be given to any absolute value of the evaluation scores across different experiments involving intuitive human judgements. Only a relative comparison of these evaluation scores produced within the same experiment is possible.

5.1.4. Results of automated evaluations

Automated evaluation scores use objective parameters, such the number of N-gram matches in the evaluated text and in a gold standard reference translation. Therefore, these scores are more consistent and comparable across different evaluation experiments. The comparison of the scores indicates the relative complexity of the texts for translation. For the output of both MT systems under consideration two sets of automated evaluation scores were generated: BLEUr1n4 and WNM Recall (as described in Section 3.1)

Figures 2 and 3 and Table 2 summarise the automated evaluation scores for the two MT systems.

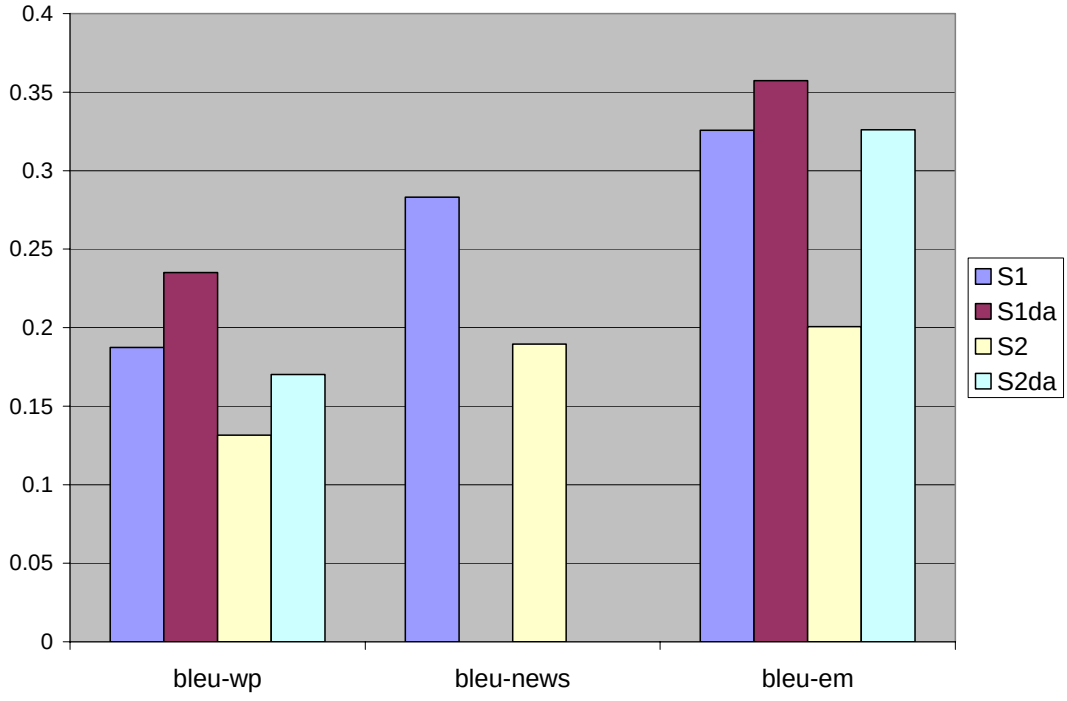


Figure 2. Automated BLEUr1n4 scores

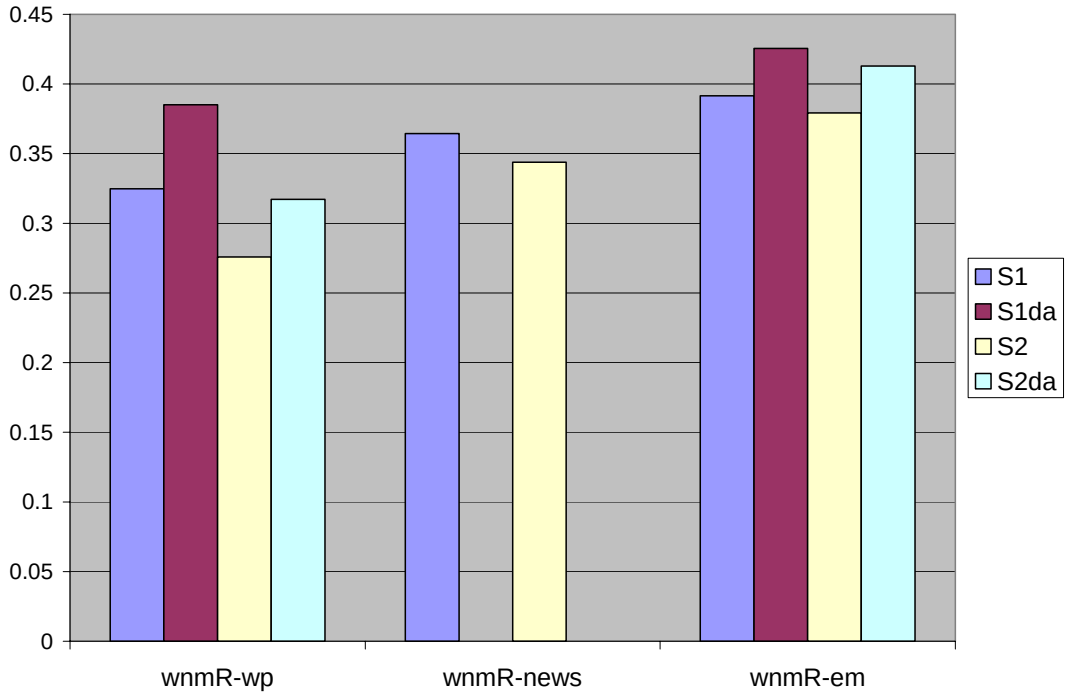


Figure 3. Automated WMN Recall scores

scores	S1	S1da	S2	S2da
bleu-wp	0.1874	0.2351	0.1315	0.1701
bleu-news	0.2831		0.1896	
bleu-em	0.3257	0.3573	0.2006	0.326
wnmR-wp	0.3247	0.3851	0.2758	0.3172
wnmR-news	0.3644		0.3439	
wnmR-em	0.3915	0.4256	0.3792	0.4129
r correlation	[flu]	[ade/usl]		
bleu-wp	0.9827	0.9453		
bleu-em		0.7872		
wnmR-wp	0.9896	0.9705		
wnmR-em		0.9673		

Table 2. Automated evaluation scores

It can be seen from the charts that automated scores consistently change according to the type of the evaluated text: for both evaluated systems BLEU and WNM are the lowest for the whitepaper texts, which emerge as most complex to translate, the news reports are in the middle and the highest scores are given to the e-mails, which appear to be relatively easy. A similar tendency also holds for the system after dictionary update. However, technically speaking the compared systems are no longer the same, because the dictionary update was done individually for each system, so the quality of the update is an additional factor in the system's performance – in addition to the complexity of the translated texts.

The complexity of the translation task is integrated into the automated MT evaluation scores, but for the same type of texts the scores are perfectly comparable. For example, for the DARPA news texts, newly generated BLEU and WNM scores confirm the observation made, on the basis of comparison of the whitepaper and the e-mail texts, that S1 produces higher translation quality than S2, although there is no human evaluation experiment where such translations are directly compared.

Thus the automated MT evaluation scores derive from both the "absolute" output quality of an evaluated general-purpose MT system and the complexity of the translated text.

5.1.5. Readability parameters

In order to isolate the "absolute" MT quality and to filter out the contribution of the complexity of the evaluated text from automated scores, we need to find a formal parameter of translation complexity which should preferably be resource-light, so as to be easily computed for any source text in any language submitted to an MT system.

Since automated scores already integrate the translation complexity of the evaluated text, we can validate such a parameter by its correlation with automated MT evaluation scores computed on the same set that includes different text types.

In this experiment, the following resource-light parameters were examined for their correlation with both automated scores:

Flesch Reading Ease score, which rates text on a 100-point scale according to how easy it is to understand; the score is computed as follows:

$$FR = 206.835 – (1.015 * ASL) – (84.6 * ASW),$$

where:

– ASL is the average sentence length (the number of words divided by the number of sentences);

– ASW is the average number of syllables per word (the number of syllables divided by the number of words)

Flesch-Kincaid Grade Level score which rates texts on US grade-school level and is computed as:

$$FKGL = (0.39 * ASL) + (11.8 * ASW) – 15.59$$

each of the ASL and ASW parameters individually.

Table 3 presents the averaged readability parameters for all French original texts used in our evaluation experiment and the *r* correlation between these parameters and the corresponding automated MT evaluation scores.

	FR	FKGL	ASL	ASW
wp	17.30	15.70	19.65	2.00
news	27.80	14.70	21.40	1.86
em	61.44	6.98	9.22	1.608
r/bleu-S1	0.872	-0.804	-0.641	-0.928
r/bleu-S2	0.785	-0.701	-0.513	-0.859
r/wnm-S1	0.920	-0.864	-0.721	-0.963
r/wnm-S2	0.889	-0.825	-0.669	-0.941
r *Average*	0.866	-0.799	-0.636	-0.923

Table 3. Readability of French originals

Table 3 shows that the strongest negative correlation exists between ASW (average number of syllables per word) and the automated evaluation scores. Therefore the ASW parameter can be used to normalise MT evaluation scores. Therefore translation complexity is highly dependent on the complexity of the lexicon, which is approximated by the ASW parameter.

The other parameter used to compute readability – ASL (average sentence length in words) – has a much weaker influence on the quality of MT, which may be due to the fact that local context is in many cases sufficient to produce accurate translation and the use of the global sentence structure in MT analysis is limited.

5.1.6. Normalised evaluation scores

The ASW parameter was used to normalise the automated evaluation scores in order to obtain absolute figures for MT performance, where the influence of translation complexity is neutralised.

Normalisation requires choosing some reference point – some average level of translation complexity – to which all other scores for the same MT system will be scaled. The difficulty of the news texts in the DARPA 94 MT evaluation corpus can be used as one such "absolute" reference point. Normalised figures obtained on other types of texts will mean that if the same general-purpose MT system is run on the DARPA news texts, it will produce raw BLEU or WNM scores approximately equal to the normalised scores. This allows users to make a fairer comparison between MT systems evaluated on different types of texts.

For the WNM scores the best normalisation can be achieved by multiplying the score by the complexity normalisation coefficient *C*, which is the ratio:

$$C = \mathrm{ASW}_{\mathrm{evalText}} / \mathrm{ASW}_{\mathrm{DARPAnews}}.$$

For BLEU the best normalisation is achieved by multiplying the score by C^2 (the squared value of $ASW_{evalText}$/ $ASW_{DARPAnews}$).

Optimal values of normalisation coefficients were established experimentally, although there is a need to find some theoretically grounded method of normalising evaluation scores.

Normalisation makes the evaluation relatively stable – in general, the scores for the same system are the same up to the first rounded decimal point. Table 4 summarises the normalised automated scores for the evaluated systems.

	C	S1	S1da	S2	S2da
bleu-wp	1.156	0.217	0.272	0.152	0.197
bleu-news	1.	0.283		0.190	
bleu-em	0.747	0.243	0.267	0.150	0.244
wnmR-wp	1.075	0.349	0.414	0.297	0.341
wnmR-news	1.	0.364		0.344	
wnmR-em	0.865	0.338	0.368	0.328	0.357

Table 4. Normalised BLEU and WNM scores

The accuracy of the normalisation can be measured by standard deviations of the normalised scores across texts of different types. The improvement in stability of the normalised scores was also measured and compared to the stability of the raw scores generated on different text types. Standard deviation was computed using the formula:

$$STDEV = \sqrt{\frac{n\sum x^2 - (\sum x)^2}{n(n-1)}}$$

Table 5 summarises standard deviations of the raw and normalised automated scores for the e-mails, whitepaper and news texts.

	S1	S1da	S2	S2da	Ave-rage
bleu-stdev	0.071	0.086	0.037	0.11	0.076
N-bleu-stdev	0.033	0.003	0.022	0.033	0.023
improved *X					**3.299**
wnm-stdev	0.034	0.029	0.053	0.068	0.046
N-wnm-stdev	0.013	0.033	0.024	0.011	0.02
improved *X					**2.253**

Table 5. Standard deviation of BLEU and WNM

It can be seen from the table that the standard deviation of the normalised BLEU scores across different text types is 3.3 times smaller; and the deviation of the normalised WNM scores is 2.25 times smaller than for the corresponding raw

scores. So the normalised scores are much more stable than the raw scores across different evaluated text types.

5.1.7. Conclusion from the experiment

In this section we presented empirical evidence for the observation that the complexity of an MT task influences automated evaluation scores, and proposed a method for normalising the automated scores by using a resource-light parameter of the average number of syllables per word (ASW), which relatively accurately approximates the complexity of the particular text type for translation.

The fact that the potential complexity of a particular text type for translation can be accurately approximated by the ASW parameter can have an interesting linguistic interpretation. The relation between the length of the word and the number of its meanings in a dictionary is governed by the Menzerath's law (Koehler, 1993: 49), which in its most general formulation states that there is a negative correlation between the length of a language construct and the size of its "components" (Menzerath, 1954; Hubey, 1999: 239). In this particular case the size of a word's components can be interpreted as the number of its possible word senses. It can be suggested that the link between ASW and translation difficulty can be explained by the fact that the presence of longer words with a smaller number of senses requires a more precise word sense disambiguation for shorter polysemantic words, so the task of word sense disambiguation becomes more demanding: the choice of very specific senses and the use of more precise (often terminological translation equivalents) is required.

Further research could empirically test this suggestion as well as look further into improving the stability of the normalised scores by developing better normalisation methods. The proposed approach could be evaluated on larger corpora containing different genres, and will investigate other possible resource-light parameters, such as type/token ratio of the source text or unigram entropy, which can predict the complexity of the translated text more accurately. Different formal parameters for syntactic and semantic complexity of a text could be also tested, such as average tree depth, an average number of word senses in monolingual and bilingual dictionaries for words in text, etc. Another direction of prospective research is comparison of stability of evaluation scores on subsets of the evaluated data within one particular text type and across different text types.

5.2. Determining minimal size of MT evaluation corpus

This section reports the results of an experiment on measuring stability and the accuracy of MT evaluation scores produced for text collections of different length. It is generally accepted that the longer the evaluation corpus, the more reliable are the evaluation scores, since in smaller collections the differences may be due to noisy data. However, many MT evaluation experiments are built around relatively small amounts of data, which may be expensive and time-consuming to acquire. In this respect it is important to establish, at what point the results of MT evaluation experiments clearly stand out of the noise and yield meaningful comparison between MT systems and reliable correlation figures with human judgements. The goal of this experiment is practical, but the results have an interesting theoretical interpretation, which is important for the topic of the previous section: at what level we can view an evaluation corpus as homogenous.

5.2.1. Motivation for the experiment

There are two aspects in which we can view the results of MT evaluation as reliable. The first aspect is internal to automated scores. We can ask how stable the scores are across comparable collections of data, so the results of an MT evaluation experiment for same MT systems can be replicated using a different data set. If the evaluation results for different collections of comparable texts are similar, the size of text collections should be sufficient. This aspect is a standard method of evaluating "reliability" of MT evaluation scores (BLEU computes confidence intervals for the scores using a similar method). A possible way to assess stability of the evaluation scores is to compute standard deviation for different runs of the experiment using different sizes of MT evaluation corpus.

However, this aspect doesn't link automated scores to any external parameters, so it doesn't give a direct answer to the question what are possible distortions in correlation between automated and human scores. MT evaluators are mostly interested in these correlation figures. Therefore, a second aspect of this problem is determining which minimal size of the corpus gives sufficiently high correlation figures and also – whether we can expect that these correlation figures will be stable across comparable text collections of the same size. A way to assess these parameters would be to compute correlation figures for different sizes of evaluation corpus as well as standard deviation of these figures for different runs of the experiment on different comparable collections.

The results of the two experiments are be complementary and provide independent sources of evidence for determining a minimal size of MT evaluation corpus.

5.2.2. Set-up of the experiment

This experiment was carried out using the DARPA 94 evaluation corpus. On the first stage the experiment BLEU and WNM scores were generated for each text in the output of the 5 MT systems available in the corpus. Both metrics were computed using a single human reference, and the scores for each text were computed twice, with each of the two human translations as a reference (this gives additional comparable data for the experiment, since BLEU and WNM scores computed for the same text with an alternative human reference are independent). Therefore the scores were generated for all 100 texts for each of the 5 MT systems available in the corpus and twice for each text, using the "Reference" and the "Expert" translation.

On the second stage of the experiment the scores were grouped into larger and larger sets – starting from 1 text and finishing with a group covering the whole 100 texts in a corpus. Sets of 1, 5, 10, 20, 33, 50 and 100 texts were created. There were 20 sets of 5 texts, 10 sets of 10 texts, 5 sets of 20 texts, and so on. For each set 2 average scores were computed – using each of the two human references.

The third and the forth stages of the experiment are independent, they address the two complementary questions formulated above: (1) what is the dynamics of standard deviation figures for automated scores if we increase the size of evaluation corpus from 360 words to 36000 words; (2) what is the dynamics of correlation figures and what will be the standard deviation for correlation in this case.

Therefore on the third stage of the experiment standard deviation was computed for each collection of sets of the same size: i.e., across all 100 sets of 1 text, across 20 sets of 5 texts, across 10 sets of 10 texts, etc. Since there are two scores for each set, standard deviation can be computed even for the set containing all 100 texts.

On the forth stage the r correlation was found for each of the sets, and an average and standard deviation for correlation figures were computed across all sets of the same size., i.e., an average correlation of a set containing 1 text, containing 5 texts, etc., and similarly – standard deviation across all sets of the given size.

5.2.3. Results of the experiment

The results of the third stage of the experiment, which characterise internal stability of automated evaluation scores, are presented in Table 6 / Figure 6 and Table 7 / Figure 7 – for the BLEU and WNM scores respectively. Columns show the size of the chunk: from 1 text long to 100 texts long.

BLUE-stdev	L1	L5	L10	L20	L33	L50	L100
candide	0.0519	0.0236	0.0142	0.0124	0.0094	0.0080	0.0030
globalink	0.0487	0.0239	0.0121	0.0102	0.0043	0.0045	0.0006
Ms	0.0435	0.0180	0.0095	0.0051	0.0025	0.0037	0.0007
systran	0.0513	0.0249	0.0113	0.0088	0.0043	0.0048	0.0040
reverso	0.0524	0.0272	0.0149	0.0099	0.0046	0.0050	0.0011

Table 6. Standard deviation of BLEU evaluation scores

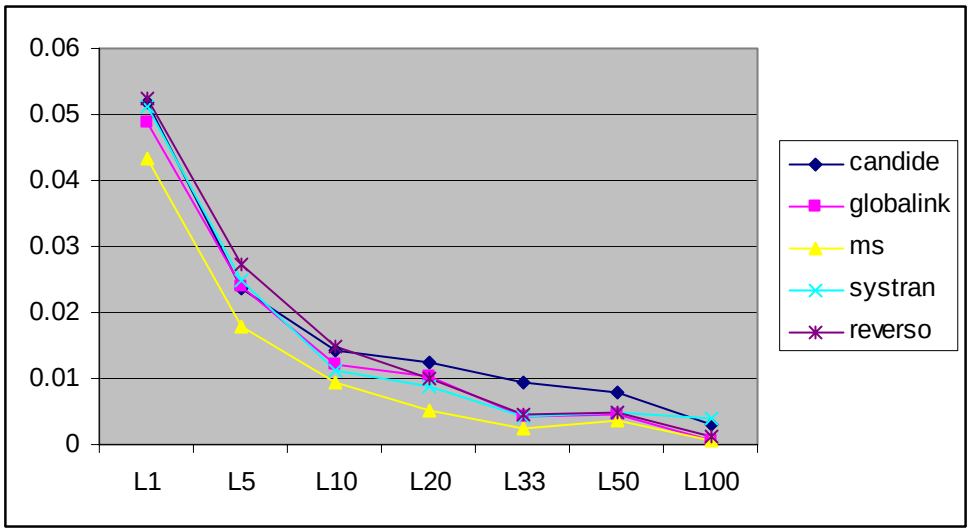


Figure 6. Standard deviation of BLEU evaluation scores

WNM-stdev	L1	L5	L10	L20	L33	L50	L100
candide	0.0548	0.0258	0.0156	0.0127	0.0109	0.0089	0.0069
globalink	0.0529	0.0248	0.0129	0.0098	0.0067	0.0044	0.0044
ms	0.0508	0.0192	0.0118	0.0063	0.0057	0.0021	0.0024
systran	0.0549	0.0289	0.0141	0.0112	0.0074	0.0068	0.0075
reverso	0.0548	0.0261	0.0164	0.0095	0.0077	0.0061	0.0042

Table 7. Standard deviation of WNM evaluation scores

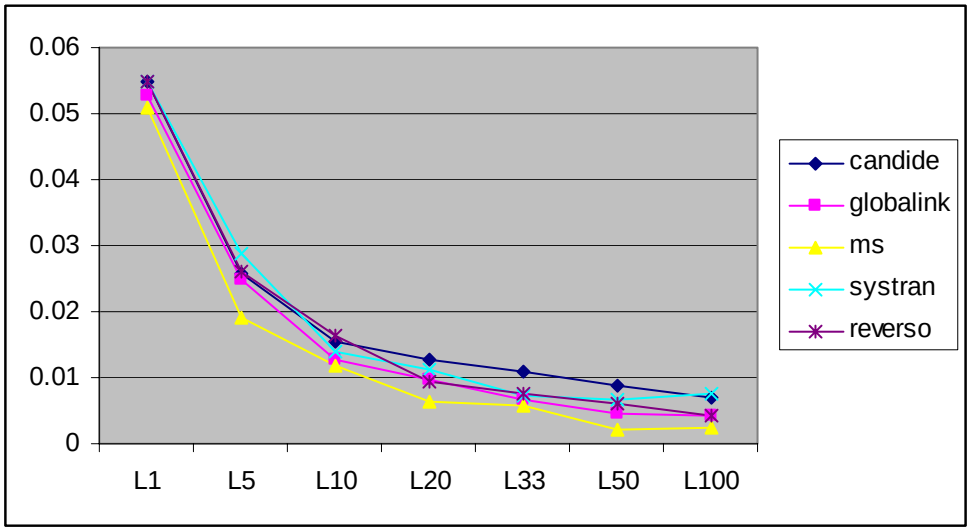


Figure 7. Standard deviation of WNM evaluation scores

It can be seen from the charts that standard deviation for both scores gradually decreases if the size of evaluated corpus gets larger, starting from more than 5% of the scale and approaching figures of less than 1%. It is interesting, however, that the improvement is the largest on the initial stages, and gradually flattens, especially after the size of the corpus has approached 33 texts (around 12000 words).

The presented results can be used to assess whether the difference between the systems is statistically significant using a standard z-test: if it is greater than 1.96 standard deviations, we may be 95% confident that the difference is not by chance. If the difference increases to 2.576 standard deviations, or to 3.09 standard deviations, the confidence increases respectively to 99% and to 99.9%. We see that if the size of the evaluated corpus is 12000 words, standard deviation for both scores is usually smaller than 1%, so if the difference between automated scores is more than 0.025, there is a 99% chance that this reflects differences in number of N-gram matches in a greater "text population".

However, greater number of matches doesn't necessarily mean higher quality in eyes of human evaluators. The results of the stage 4 of the experiment show the dynamics of correlation between automated and human scores, if we gradually move to greater evaluation corpus. Table 8 and Figure 8 present the results for correlation of the two metrics with adequacy, and Table 9 and Figure 9 show the results for fluency.

ade	L1	L5	L10	L20	L33	L50	L100
r-correl-BLUE	0.2166	0.4907	0.6709	0.7444	0.7191	0.7406	0.7418
stdev-BLUE	0.5438	0.4322	0.1845	0.1335	0.2189	0.1197	0.0101
r-correl-WNM	0.2402	0.5235	0.7665	0.8476	0.8376	0.8561	0.8648
stdev-WNM	0.5433	0.4989	0.1765	0.057	0.0655	0.0358	0.0209

Table 8. *r* correlation figures and standard deviation of *r* – adequacy

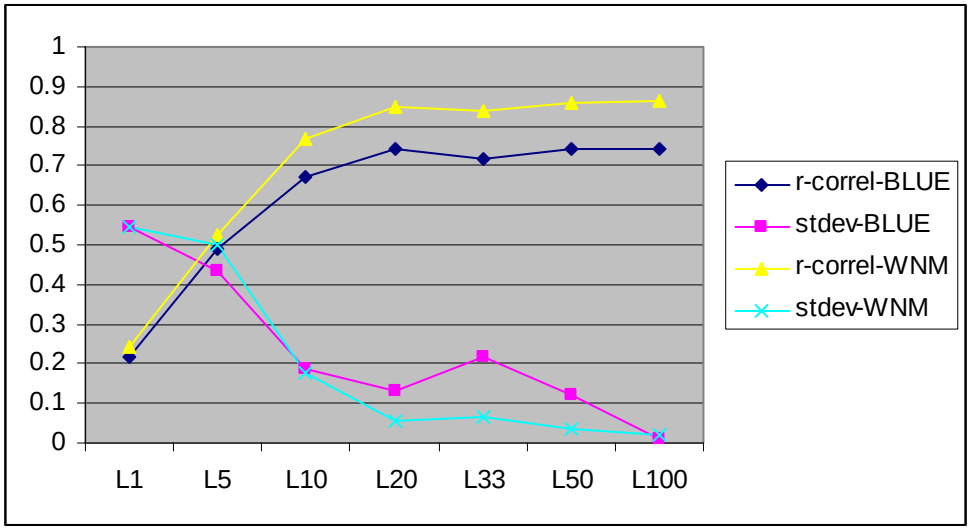


Figure 8. *r* correlation figures and standard deviation of *r* – adequacy

flu	L1	L5	L10	L20	L33	L50	L100
r-correl-BLUE	0.3466	0.5755	0.7758	0.8977	0.9167	0.9293	0.9509
stdev-BLUE	0.5059	0.4555	0.1905	0.0997	0.088	0.0604	0.0062
r-correl-WNM	0.3504	0.5589	0.7446	0.844	0.8804	0.8836	0.8887
stdev-WNM	0.4889	0.4039	0.1468	0.1532	0.0444	0.0296	0.0296

Table 9. *r* correlation figures and standard deviation of *r* – fluency

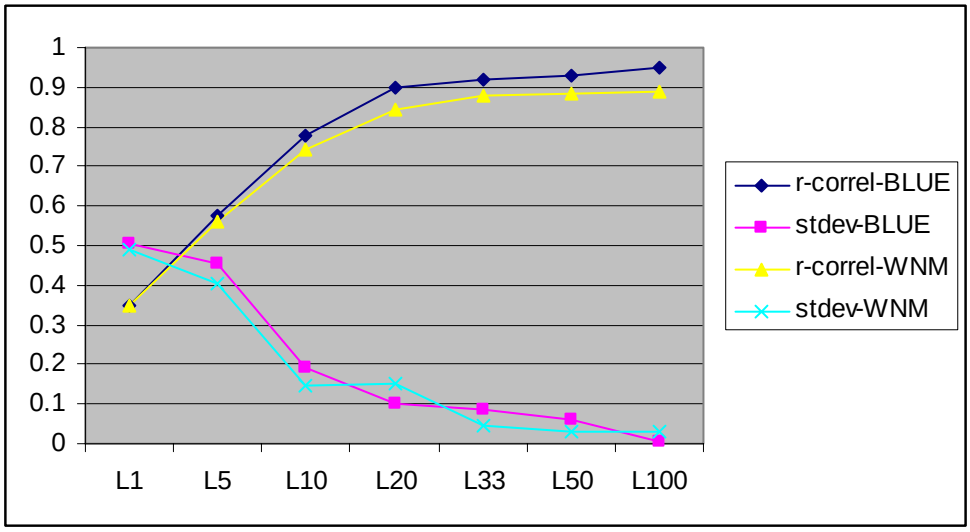


Figure 9. *r* correlation figures and standard deviation of *r* – fluency

It can be seen from the charts that correlation between automated scores and human judgements gradually increases with the size of evaluated corpus. However, again correlation flattens for both scores after size has approached 20 texts (7200 words). In this experiment, whoever, we can also see the relation between the size of

correlation coefficient *r* and its standard deviation. If the corpus is too small, the standard deviation is even greater than the value of *r* (therefore, no meaningful conclusions about correlation can be drawn for corpora of 5 texts, or 1800 words, and even the corpora of 10 texts, or 3600 words will still be too noisy in terms of correlation). Once again, standard deviation of *r* flattens after 20 texts. Note however, that WNM gives better correlation with adequacy, and BLEU is more accurate for evaluating fluency. Interestingly, standard deviation of *r* is also in line with this tendency: higher *r* is more stable in WNM evaluation of adequacy and in BLEU evaluation of fluency.

Therefore, an evaluated corpus needs to be at least 7200 words long to provide reliable correlation figures. Increasing the size of the corpus after this point only slightly improves correlation figures and the stability of *r* coefficient.

If we compare the results with the stage 3, we see that the potential of improvement in correlation figures is exhausted even before the stability of automated scores finally flattens after 12000 words.

5.2.4. Interpretation of the results

It is interesting to point out that improvement in stability of the automated scores and in the values of correlation can be explained by two reasons. Firstly, greater corpus compensates for legitimate translation variation, so it becomes less relevant whether still a great proportion of correct terms in translation didn't match anything in human translation, since the number of matches becomes sufficient to characterise the performance of an individual MT system in general.

Secondly, differences in scores for individual texts can be due to the fact that the corpus is not homogeneous on the text level: there are more difficult and less difficult texts, which are translated better or worse, depending on their difficulty for MT (similarly as sentences can be translated better or worse). So even if we finally solve the problem of legitimate translation variation, still we will need larger corpus, because the objective quality of translation of any individual text and even any individual sentence is different. Therefore, if we are interested not in quality of translation of sentences and texts, but in the level of performance of an MT system in general, we need to account for such non-homogeneity of a corpus on the micro-level (which is different from non-homogeneity on the macro-level that is due to differences in genres, as discussed in the previous Section 5.1).

Such instability of the performance on the micro-level reflects the general problem of MT systems, once pointed out by M.Kay: MT can translate some sentences and some easier texts without any major errors, but we can never be sure that this level will be the same for all sentences and for all texts in a corpus. Even

worse, if we don't know where there is a potentially serious error, we still need to check the entire translation. Such instability of the performance on the micro-level is yet another parameter of MT quality, which needs to be improved, apart from general level of MT quality. Ideally, MT output should not only show high general figures of the performance on the corpus level, but also this performance should be stable on the micro-level across individual sentences and texts. The next step is achieving the stability of MT performance across different genres and text types, i.e., macro-level stability on heterogeneous corpus.

5.3. Replicating evaluation results to other language pairs

This section briefly describes results of on-going research on replicating MT evaluation results for other target languages. This experiment is part of a larger MT evaluation project with a commercial company with a goal to select an MT system for on-line translation software solutions. The experiment compares MT output of 4 commercial MT systems – abbreviated as *s03, s04, s05* and *s06* (one of them is a statistical system – *s04*), which translated into 23 translation directions groped by 6 target languages: German, English, Spanish, French, Italian and Portuguese. For 2 of the knowledge-based systems, versions with updated dictionaries were also included into the evaluated set – abbreviated as *u03* and *u05*.

The assumption behind grouping by target language is that human and automated evaluation scores will be comparable for the same target language, so correlation figures will be meaningful even if source languages are different, but the translated documents are the same. The size and text types in evaluated corpus is the same as reported in Section 5.1: 36 email texts (3800 words) translated into 6 languages and EU whitepaper (3200 words). Each segment in this corpus was evaluated by three human judges and by the two automated metrics – BLEU and WNM. As it follows from the results of the previous section, the correlation figures for such a small corpus will be noisy, so commercial users relied on human scores for their business decisions. Automated scores were used as a preliminary estimate of MT quality.

However, interesting results were obtained by comparing correlation figures from the present experiment with average correlation figures of similarly small subsections of the DARPA 94 corpus. For all target languages and for both text types the correlation figures were found to stay reliably within the limits of expected variance. However other parameters – such as the slope and intercept of the regression line – were found to be specific for a target language and a text type. Therefore, the experimental results suggest that BLEU and WNM automated scores

may work similarly well for other target languages and text genres in terms of *correlation* with human scores – the results don't disprove the null-hypothesis that all differences with DARPA corpus figures are due to some accidental factors. On the other hand, null-hypothesis was rejected for *regression* parameters for certain combinations of the target language and text type, which means that it will be necessary to use language-specific and genre-specific correction coefficients, if we want to predict acceptability level and possible range of human scores of an MT system that works on certain combinations of a target language and text type.

5.3.1. Multilingual MT evaluation experiment

Human and automated MT evaluation scores for all translation directions and text types are presented in Table 1.

Sys	SL	TL		hEm	wnmEm	BleuEm		hWp	wnmWp	BleuWp
s06	fr	de	1	3,665	0,2653	0,1496	1	3,818	0,2061	0,1314
s05	en	de	2	3,602	0,3029	0,236	3	3,342	0,1386	0,0762
s06	en	de	3	3,503	0,2759	0,1969	2	3,469	0,1441	0,0581
s03	it	de	4	3,184	0,1901	0,0644	4	2,707	0,0977	0,0224
				corr	**0,88238**	**0,7694**		**corr**	**0,9487**	**0,9168**
Sys	SL	TL		hEm	wnmEm	BleuEm		hWp	wnmWp	BleuWp
s05	de	en	1	4,383	0,4213	0,3207	8	4,071	0,2055	0,1258
u05	fr	en	2	4,247	0,4446	0,3339	1	4,589	0,4330	0,3354
s06	de	en	3	4,194	0,4005	0,2951	7	4,153	0,2320	0,1374
s05	fr	en	5	4,151	0,3513	0,2475	4	4,273	0,2565	0,2002
s05	es	en	4	4,151	0,3473	0,218	5	4,242	0,2721	0,1702
s06	fr	en	6	4,080	0,3920	0,2862	2	4,347	0,3091	0,2242
s06	es	en	7	3,902	0,3196	0,1959	9	4,018	0,2528	0,1585
u03	fr	en	8	3,845	0,3880	0,2659	3	4,338	0,3636	0,2370
s03	it	en	9	3,746	0,2716	0,1320	15	2,907	0,2024	0,0974
s04	fr	en	10	3,689	0,3294	0,2259	6	4,224	0,3674	0,2746
s04	es	en	11	3,447	0,2612	0,1712	11	3,927	0,3308	0,2237
s03	fr	en	12	3,423	0,2982	0,1740	14	3,131	0,2026	0,1253
s03	es	en	13	3,294	0,2518	0,1432	13	3,147	0,2231	0,1278
s06	it	en	14	3,250	0,2856	0,1746	10	3,971	0,2339	0,1296
s06	pt	en	15	3,124	0,3075	0,2051	12	3,711	0,2256	0,1216
				corr	**0,82152**	**0,7699**		**corr**	**0,6742**	**0,7086**
Sys	SL	TL		hEm	wnmEm	BleuEm		hWp	wnmWp	BleuWp
s06	fr	es	1	3,618	0,2460	0,1720	1	4,456	0,5771	0,5570
s05	en	es	2	3,379	0,2600	0,1988	2	3,696	0,2785	0,2158
s03	en	es	3	3,149	0,2136	0,1412	3	3,498	0,2704	0,2058
s06	en	es	4	3,126	0,2410	0,1925	4	3,460	0,2539	0,1987
s04	en	es	5	2,490	0,2250	0,1592	5	3,171	0,3269	0,2414
				corr	**0,56742**	**0,3539**		**corr**	**0,8487**	**0,8909**
Sys	SL	TL		hEm	wnmEm	BleuEm		hWp	wnmWp	BleuWp
s06	en	fr	1	3,974	0,3079	0,2414	5	3,924	0,3285	0,2551
u05	en	fr	2	3,846	0,3024	0,2498	2	4,298	0,4276	0,3321
u03	en	fr	3	3,811	0,2909	0,2228	4	4,118	0,3814	0,2617
s06	de	fr	4	3,654	0,2573	0,1656	7	3,882	0,3294	0,2328
s05	en	fr	5	3,649	0,291	0,237	6	3,902	0,3247	0,2460
s03	it	fr	6	3,450	0,2075	0,1398	10	3,436	0,3851	0,3381
s06	it	fr	7	3,446	0,2513	0,1695	8	3,880	0,4351	0,3721

s06	es	fr		8	3,377	0,2474	0,1613		1	4,562	0,5360	0,4748
s04	en	fr		9	3,351	0,2523	0,201		9	3,620	0,3990	0,2998
s06	pt	fr		10	3,303	0,2508	0,178		3	4,204	0,4991	0,4236
s03	en	fr		11	2,854	0,2204	0,139		11	2,647	0,2215	0,1341
					corr	**0,82017**	**0,7732**			**corr**	**0,7883**	**0,7182**
Sys	SL	TL			hEm	wnmEm	BleuEm			hWp	wnmWp	BleuWp
s06	fr	it		1	3,743	0,253	0,1799		1	4,500	0,4848	0,4348
s03	es	it		2	3,705	0,2049	0,1166		3	3,667	0,2761	0,2028
s03	fr	it		3	3,598	0,232	0,1536		2	3,902	0,3738	0,3172
s06	en	it		4	3,333	0,2206	0,1544		4	3,611	0,2362	0,1397
s03	en	it		5	3,282	0,2018	0,1170		5	2,964	0,1800	0,0898
s03	de	it		6	2,551	0,1869	0,0858		6	2,287	0,1837	0,0648
					corr	**0,73446**	**0,74**			**corr**	**0,8873**	**0,9064**
Sys	SL	TL			hEm	wnmEm	BleuEm			hWp	wnmWp	BleuWp
s05	en	pt		1	3,409	0,2703	0,2076		2	3,771	0,2010	0,1341
s06	fr	pt		2	3,377	0,2353	0,1616		1	4,262	0,4512	0,4214
s06	en	pt		3	3,114	0,2232	0,1435		3	3,196	0,1497	0,0901
					corr	**0,76596**	**0,7833**			**corr**	**0,9174**	**0,902**

Table 1. Scores for multilingual MT evaluation experiment

Note that the output of a commercial statistical MT system *s04* is among the lowest according to human ranking, but it is one of the highest according to automated scores, which holds for all target languages, for which *s04* is available. This observation confirms the results presented in Chapter 2 about a similar phenomenon for an earlier statistical MT system Candide: automated scores also over-estimated the quality of its output, and human evaluation figures for adequacy ranked it much lower than the automated scores. This observation confirms an earlier suggestion that reference proximity evaluation methods work best with homogeneous MT architectures, and overestimate the adequacy of statistical MT.

In the next stage of the experiment we computed Pearson's correlation coefficient *r* between automated N-gram metrics (BLEU and WNM) and the human evaluation scores. We also computed the two parameters of the regression line (the slope and the intercept), which allow us to predict human scores, given automated scores for some new system:

HumanSc= ***Slope****AutomatedSc+***Intercept***

All coefficients were computed individually for each target language and for each evaluated text type.

The resulting figures are given in Table 2.

TL/Text Type	*r* corr BLEU/ WNM	Slope BLEU/ WNM	Intercept BLEU/ WNM
DE/em	0.7694 0.8824	1.3301 0.9973	-0.7663 -0.4373
DE/wp	0.9168 0.9487	0.0898 0.0915	-0.2275 -0.1583

EN/em	0.7699 0.8215	0.5996 0.6096	-0.2291 -0.1247
EN/wp	0.7086 0.6742	0.0961 0.0957	-0.1992 -0.1026
ES/em	0.3539 0.5674	0.0997 0.1224	0.1099 0.1599
ES/wp	0.8909 0.8487	0.2823 0.2355	-0.7484 -0.5198
FR/em	0.7732 0.8202	0.5043 0.4278	-0.1636 -0.0394
FR/wp	0.7182 0.7883	0.1351 0.136	-0.2153 -0.1371
IT/em	0.74 0.7345	0.2847 0.1965	-0.0573 0.0841
IT/wp	0.9064 0.8873	0.1687 0.1379	-0.3803 -0.1918
PT/em	0.7833 0.766	0.7996 0.5787	-0.3568 -0.139
PT/wp	0.902 0.9174	0.3042 0.2774	-0.9233 -0.7709

Table 2. Correlation and regression coefficients

It can be seen from the table that there are differences in terms of absolute values for correlation, slope and intercept across languages and text types.

In the second stage of the experiment we addressed the problem whether the differences in values of these coefficients are statistically significant or whether they can be attributed to chance and random error, that may be due to the relatively small size of the evaluated text.

In order to answer this question we contrasted the computed coefficients with the gold standard MT evaluation benchmark – the DARPA 94 MT evaluation corpus (White et al., 1994). We used the French-into-English part of the corpus which contains 100 news texts; each text being approximately 360 words long. For a corpus of this size high correlation figures are reported for both BLEU and WNM with human evaluation scores (Babych and Hartley, 2004a).

In the current experiment we divided the DARPA corpus into 10 chunks; each chunk contains 10 texts and is about the same size as our new texts – approximately 3,600 words.

We generated BLEU and WNM scores for each text in the corpus using a single human reference. Since two independent human translations are available for each text in the DARPA corpus, the scores for each text were generated twice – using both the “expert” and the “reference” human translations. We then computed the average human and automated scores for the 10 texts in each of the 10 chunks.

These scores became the basis for making comparisons with the corresponding scores in our new texts.

The comparison was done in the following way: first we examined the variation of the correlation and regression parameters across chunks in the DARPA corpus; second, we established whether the same parameters in our new texts were within the limits of such variation or whether they stood significantly beyond the outer limits of such variation "noise". If so, they could be said to carry some "signal" about the evaluated target language or the text type.

We computed the same set of parameters for each chunk: the *r* correlation coefficient, and the slope and the intercept of the regression line.

We assessed the variation of these parameters in the DARPA corpus by computing the averages and standard deviation figures for the 10 chunks. These figures are presented in Table 3.

TL/Text Type	*r* corr BLEU/ WNM	Slope BLEU/ WNM	Intercept BLEU/ WNM
EN/news/ AVERAGE	0.6709 0.7666	0.4611 0.404	-0.096 -0.0009
EN/news/ STDEV	0.1873 0.1799	0.2479 0.203	0.1676 0.1376

Table 3. Average and StDev: DARPA

It can be seen from the table that, on average, WNM scores have a higher correlation with adequacy than BLEU (r = 0.767 vs. 0.671), which confirms previous results obtained on complete DARPA corpus and on other texts (Babych and Hartley, 2004a; Babych and Hartley, 2004b). However, since the size of the evaluated texts is smaller, the standard deviation figures are also high (about 25%–30% of the mean) and, again, slightly higher for BLEU.

On the one hand, such a high level of variation "noise" on smaller corpora makes any predictions about MT evaluation scores more risky; on the other hand, for the purposes of the current experiment we are not interested in specific predictions per se, rather we want to know if the accuracy of such predictions depends on the target language or text type. For this purpose having a smaller corpus with "noisier" variation is even beneficial, because only the parameters that carry the strongest "signal" will stand out from the noise.

For each of the correlation and regression parameters in our new texts we computed the z-score (the standard score which tells how far the tested score is from the expected average in terms of standard deviations):

$$z = \frac{TestedSc - ExpectedMean}{STDEV}$$

ExpectedMean and *STDEV* are taken from the Table 3, while *TestedSc* comes from Table 2.

We assume that variations in the DARPA scores fit a Gaussian distribution, so 95% of the points are within the limit of 1.96 standard deviations from the mean, and 99% are within the limit of 2.576 standard deviations. Therefore, if the z-score for a particular parameter is outside the range ±2.576, we can be 99% confident that the difference between the tested parameter and the corresponding parameter in the DARPA corpus can be attributed to some features in the target language and the text type, and did not happen by chance, e.g., was not influenced by the size of the evaluated text.

5.3.2. Results of the comparison of correlation and regression parameters

The z-scores for each of the tested correlation and regression parameters are presented in Table 4.

TL/Text Type	*z* – *r* corr BLEU/ WNM	*z* – Slope BLEU/ WNM	*z*–Intercpt BLEU/ WNM
DE/em	0.5259 0.6434	**3.506** **2.9228**	**-3.999** **-3.1717**
DE/wp	1.3132 1.012	-1.498 -1.54	-0.785 -1.144
EN/em	0.5284 0.3051	0.5587 1.0127	-0.794 -0.8998
EN/wp	0.2015 -0.513	-1.472 -1.519	-0.616 -0.739
ES/em	-1.693 -1.1072	-1.458 -1.3871	1.2276 1.1687
ES/wp	1.1749 0.4559	-0.721 -0.83	**-3.892** **-3.771**
FR/em	0.5462 0.2976	0.1745 0.1173	-0.404 -0.28
FR/wp	0.2523 0.1205	-1.315 -1.32	-0.712 -0.99

IT/em	0.3691 -0.1788	-0.712 -1.022	0.2308 0.6177
IT/wp	1.2575 0.6707	-1.18 -1.311	-1.696 -1.388
PT/em	0.6003 -0.0037	1.3657 0.8606	-1.556 -1.004
PT/wp	1.2338 0.8378	-0.633 -0.624	**-4.935** **-5.596**

Table 4. z-scores for correlation/regression

It can be seen from the table that for most parameters across the target languages and text types the z-scores are smaller that 1.96, therefore the differences in such parameters can be attributed to variation that is typical for the evaluation corpus of a size of around 3,600 words. However, several parameters have z-scores higher than 2.576 (even higher than the next convenient "confidence threshold" of 99.9% – 3.09). For these parameters the null-hypothesis should be rejected: their values are influenced by the target language and text type.

First, note that for the Pearson's correlation coefficient *r* the z-scores for all target languages and text types are contained within the limits of the variation present in the French-English part of the DARPA corpus. The null-hypothesis for the *r* coefficient always holds, which confirms that for all evaluated target languages and text types the n-gram MT evaluation metrics can be used reliably, if the user is only interested in correlation between the human scores and automated scores, e.g., for internal development purposes. Correlation is not influenced by the these "external" factors, so higher automated n-gram scores will always mean better quality in the eyes of human evaluators for all evaluated target languages and text types.

Second, however, having reliable correlation figures is not enough for making predictions about human scores on the basis of automated scores (as well as about the level of "acceptability" of the output of a particular MT system for end-users). The additional parameters needed to make these predictions (such as the slope and the intercept of the regression line) are not stable across the target languages and text types, and are influenced by these "real-world", evaluation-external factors.

Note that for the target language German for emails the z-scores for both parameters of the regression line "stand out" from the variation "noise" for both n-gram metrics. The regression line for German emails is much steeper – higher slope – and is moved down the *y* axis (the axis of human scores) – lower intercept. This means that "better quality" for human evaluators here needs a smaller number of n-gram matches, and that the improvement in human scores involves a much greater

increase in the number of n-gram matches than is the case for the news texts in the French-English part of the DARPA corpus.

This does not hold for the whitepaper texts translated into German: here all the differences in the slope and the intercept of the regression line are within the variation limits of the French-English DARPA corpus.

Also note that for the whitepaper texts in the target languages Spanish and Portuguese the intercept parameter of the regression line is also much lower than expected: here higher "human" quality again relies on smaller number of N-gram matches. But the slope of the regression line is within the variation limits both for Spanish and Portuguese.

A surprising fact about these results is that regression parameters can be changed by the target language (possibly influenced by some language-specific features) or by text type, which from the point of view of MT evaluation may behave as a different language (or sub-language). The mechanism whereby such a language/sub-language influences the regression parameters is not clear, but it can be suggested that typological features (rather than genealogical factors) play an important role, since genealogically related languages (such as English and German or French/Italian and Spanish/Portuguese) often show differences in the parameters. An important factor could be the degree to which the target language is "analytic" (relies on the use of free functional morphemes and syntactic means to express concepts) or "synthetic" (more often uses fused functional morphemes and word formation for concepts). The difference in the degree of "analytism" may explain the differences in the parameters for French and Spanish whitepaper texts.

It should be also noted that within a particular language "typological distance" between sub-languages (or text-types) could be different: it is intuitively plausible that the colloquial style of emails in German is very different from the style of legal documents, such as the whitepaper – in terms of lexicon and syntax – and such a distance is possibly greater than between English or French emails and the whitepaper texts (cf. Kittredge, 1982). This could provide a clue as to why there is a difference in regression parameters across text types in German, but there is no such difference in English, French or Italian.

However, the most important and interesting result of our experiment is the very fact that the regression parameters do vary across text types and target languages (TLs), so they cannot be re-used for previously untested combinations of TLs/text-types. This means that knowing the regression line parameters for a certain combination of these evaluation-external factors is not helpful for predicting human evaluation scores or the acceptability of an MT system for some other combination.

In order to predict these values, one needs to carry out expensive human evaluations for every TL/text-type combination for which there is a demand to predict human evaluation scores from automated n-gram-based scores.

There is still an open question whether the TL and the text-type are the only factors which influence the parameters of the regression line. If this is the case, "calibration" of human scores needs to be done only once for each TL/text-type combination by computing the parameters of slope and intercept on a larger corpus. Furthermore, these parameters can be re-used for the reliable prediction of human evaluation scores within the same TL/text-type combinations.

However, other external factors may also influence the regression parameters, e.g., the architecture of the evaluated MT system (statistical, example-based, rule-based, etc.), the source language. Further experiments are needed to estimate their effect on the prediction of human scores.

5.3.3. Conclusions of the experiment

We carried out a large-scale MT evaluation experiment for a number of languages and text types, which had not been the subject of automated MT evaluation. The experiment involved generating human scores for adequacy and two sets of automated evaluation scores (BLEU and WNM), computing correlation and regression parameters between human and automated scores, and predicting the acceptability of the output of particular MT systems for different target languages and text types. We established experimentally the acceptability threshold at 3.5 on the 5 point scale used by the human judges and mapped this threshold to the automated scores for each combination of text type and target language.

The analysis of this data involved measuring the difference between the correlation/regression parameters in our newly evaluated texts and in the gold standard DARPA 94 MT evaluation corpus. The principal findings are that the correlation figures for all target languages and text types are always reliably within the expected variation limits, so it can be expected that the correlation between human and automated n-gram metrics for all the evaluated target languages and sub-languages will be equally high. So the metrics can be reliably used for internal system development for all evaluated target languages.

However, end users of MT systems often need to estimate the level of acceptability of a particular MT system on the basis of automated MT evaluation scores, i.e., to predict human evaluation scores for the system on the basis of the automated scores. This task requires estimating regression parameters – the slope and the intercept of the regression line. Our results suggest that, unlike the correlation coefficient, these regression parameters may be specific to some

languages and text types. Consequently, human evaluation scores for each TL/text-type combination will still be necessary for making reliable predictions about human evaluation scores for new texts and MT systems. Absolute values of BLEU and WNM (which eventually come down to the number of n-gram matches) are influenced by such evaluation-external factors and therefore their predictive power is "local" to a particular language or a sub-language (text type). In the general case, the number of n-gram matches cannot give a "universal" prediction of "human" quality.

Future work will involve accounting for the influence of other possible factors on the regression parameters (e.g., source language) and extending the number of evaluated target languages.

5.4. Modelling legitimate translation variation for automatic evaluation of MT quality

This section discusses a potential problem for MT, which can be successfully identified with IE techniques. The experiment presented in this section examines the link between salience of terms in text and their stability and their legitimate variation (LTV) across several independent human translations of the same text. Information about stability of lexical items across human translations is important for MT evaluation and MT development.

Automated MT evaluation metrics need to take this phenomenon into account in order to reduce the minimal necessary size of MT evaluation corpus. Several widely used automatic methods for MT evaluation are based on the assumption of reference proximity – the assumption that MT quality is related to some kind of distance between the evaluated text and a professional human translation (e.g., an edit distance or the precision of matched N-grams). However, independently produced human translations are necessarily different, conveying the same content by dissimilar means. Such legitimate translation variation is a serious problem for distance-based evaluation methods, because mismatches do not necessarily mean degradation in MT quality.

Also the developers of data-driven MT systems may exploit the information about the degree of word's stability in order to improving retrieval of translation equivalents from parallel corpus. Stable words are more likely to represent "normative" equivalents, i.e., some obligatory conventionalised ways of translation, and therefore will exemplify more "adaptable" examples (Collins and Somers, 2003: 141), (Collins, 1998), which is in general a much "cleaner" material for automatic acquisition of equivalents (the intuition is that if an equivalent is stable across

independent translations, it can be more safely reused in new contexts). Variable words very often represent solutions generated "on-the-fly" and usually will be less adaptive, so an MT system should take extra caution while reusing such equivalents in a different context. Therefore, having some kind of "LTV weighting" in aligned parallel corpora used for the development of data-driven MT can be beneficial for MT quality.

Ideally such weighting should be generated from a parallel corpus that contains multiple translations of the same text. However, such data will be even more scarce and expensive than usual "single translation" corpora.

The idea of the experiment presented in this section is to test whether salience weightings of terms in text, such as tf.idf or S-scores scores, can provide indirect evidence whether a given term is likely to be stable or variable – without using multiple independent translations. An initial assumption that salience weighting and LTV will not be orthogonal is related to what can be called a "*morpho-syntactic variation conjecture*" – an intuitively plausible suggestion that the primary source of LTV is a morphological and syntactic domain: greater variation is expected for syntactic frames of sentences, function words and function morphemes, but content words are expected to be more stable across independent translations of the same texts. Since content words and function words differ in their salience scores, such scores can be used as indirect "LTV weighting" coefficients.

The presented experiment is inspired by the "morpho-syntactic variation conjecture" and examines the relation between LTV and salience scores. However, since no prior assumptions have been made on the basis of the conjecture, the side-effect of the experiment is that it tests whether this conjecture is true or not.

The results of the experiment have been unexpected and counter-intuitive: indeed there is a link between LTV and salience of terms, but the "morpho-syntactic variation conjecture" is false: the greatest degree of variation was found for most salient words. These results have interesting implications for equivalent-based approaches to MT in general and to data-driven MT in particular. The problem is that most unsafe, inadaptable translations appear at the very top of the salience hierarchy in text. The model, where most variation is centred around functional material and where content words are mostly stable, is far too simplistic. The problem is that greatest number of examples of LTV comes from the most salient content words, not from functional words. Such words are organising centres of the text and their choice often determines other lexical choices, syntactic frames, sentence structure, etc. It appears that most important items in text are inherently inadaptable, and adaptable examples are relatively marginal with respect to their salience.

A possible interpretation of this fact is still very speculative at present, and needs to be tested experimentally, but it gives a natural interpretation to the discovered relation between LTV and statistical salience. The suggestion is that such counter-intuitive finding presents equivalent-based MT with a fundamental problem of "marginality" of "safe examples" – examples, which may be "canned" in constant databases of translation equivalents and safely learnt and reused in new contexts. At present, this suggestion is still.

On the learning stage, such adaptable examples will be found in some local syntactic pockets inside sentences, and will rarely spread across the entire sentence structure. In runtime the ready translation equivalents retrieved from databases will reliably capture only marginal information from text, and may still miss the central, the most salient points, which in general don't have ready translation solutions and require some general modifications in the sentence structure, indirect translations and global adjustments in the network of related translation equivalents. However, this effect may be smaller in certain genres and subject domains which developed standardised translation practices and use a smaller number of "imaginative" elements that don't have ready translation solutions. Indeed, this is the area, where equivalent-based MT so far has been most successful.

IE-oriented salience scores can highlight this potentially important problem for equivalent-based MT, that most variation happens at the top level of the salience hierarchy and in general the central points in the text and sentence structures may be least adaptable, which might degrade even usefulness of adaptable equivalents retrieved from the text. It will be much harder to find a systematic solution to it. In this respect, IE-guided architecture for MT could be a step in the right direction. A more general solution might require developing yet more advanced intelligent processing strategies for MT than IE-based processing (Babych and Hartley, 2004b).

5.4.1. Motivation for the experiment

Automatic evaluation tools enable MT developers and users to make quick judgements about MT quality without going through a lengthy and expensive process of human evaluation. Reference-proximity automatic MT evaluation methods need to account for LTV – the fact that independent human translations of the same text often use different words and structures to convey the same content, which results in different sets of N-grams for such translations. In two independent human translations available in the DARPA-94 corpus the average overlap of unigrams in a text (about 350 words long) is approximately 72% for tokens and 68% for types. This means that while translating the same text independently two human translators use roughly around 70 % of the same lexicon, so 30% of words in their

translation will be different. However, if we count sequences of words (e.g., bi-grams, tri-grams, N-grams), then the proportion of common items in two independent translation decreases. If $N_{max} = 4$ (i.e., if we take into account the set of all N-grams starting from individual words and up to 4-grams), the overlap decreases to 46% for tokens and 44% for types.

In this section we link LTV with the phenomenon of variable importance of translated units for MT evaluation. We put forward the suggestion that such differences in information load may be approximated by statistical weighting of words in reference translations with tf.idf scores (Salton and Lesk, 1968) or S-scores, proposed in Chapter 2, Section 2.2 and also in (Babych, et al., 2003). These scores capture the "salience" of lexical items within a given document. There is a noticeable difference in distribution of such significance weights for the words that are, respectively, variable or stable in independent human reference translations, although the relation between the significance weights and the LTV factor is not straightforward.

5.4.2. Assumption of Reference Proximity

The key hypothesis behind methods that compute different kinds of distances between the human translations and MT output is the assumption of reference proximity (or ARP, mentioned in Chapter 2), which states that "*the closer the machine translation is to a professional human translation, the better it is*" (Papineni et al., 2002: 311). Strictly speaking LTV undermines this assumption, since there is an ambiguity in interpreting deviations from the reference in the evaluated text. On the one hand these deviations may be the result of mistranslations, inadequate or nonsense translations or degraded fluency of MT. On the other hand they may be the result of choosing a legitimate alternative construction, which could be equally fluent, adequate and comprehensible.

A possible way of accounting for LTV is suggested by the BLEU method, which allows users to employ several human references. This reduces the ambiguity in interpreting deviations from the single human reference (there is a greater chance that a legitimate alternative will be found at least in one translation), but there is no guarantee that such set of references exhausts legitimate translation variants, or that deviations from an N-gram set for all available references necessarily mean deterioration in MT quality. In addition this clearly increases the cost of MT evaluation, since multiple reference translations of the same text may be expensive to obtain.

Another disadvantage of using multiple human references is the so called "trouble with Recall" (Papineni et al., 2002:314): it only makes sense to compute

Precision on a *union* of reference N-grams, because a good translation will use only one of the possible translation choices for a given unit, but not all of them. Still, Recall may contain important information about some aspects of MT quality. Intuitively this disadvantage means that, despite there being no proper translation equivalent for a certain concept which might be central for a given text, the MT system may still somehow "get away with it", without being directly punished for this omission.

Yet in practical terms, despite the above theoretical drawbacks, the ARP has been found to give good estimation of translation quality for mainstream rule-based commercial MT systems (even with a single reference). The relative number of legitimate and erroneous deviations from the reference appears to be relatively stable for MT systems built with the same architecture. If human translations produced by a native-speaker are included in the evaluation, the ARP approach still correctly ranks human translations higher than MT, although the difference in scores becomes much smaller.

Problems with ARP become more visible for "non-classic" types of translations, i.e., if we include data-driven MT systems, such as statistical MT, or non-native human translations into the evaluation set. In these cases the absence of a proper model for LTV cannot be compensated by other factors. With non-native human translation, a much greater proportion of mismatches "makes sense" and is judged useful by human evaluators. With statistical MT the situation is the opposite: relatively fewer mismatches actually "make sense" for human evaluators (and possibly the proportion of "spurious" matches is also relatively higher). Thus, as it was discussed in Chapter 2, when human evaluators compare the output of systems based on different architectures, the statistical MT system Candide is ranked higher with respect to its translation fluency than with respect to its adequacy.

As a result present-day ARP-based methods consistently underestimate the usefulness of non-native human translation and overestimate the adequacy scores for statistical MT. Moreover, such "non-classic" texts cannot be judged using a single quality criterion. Different aspects of translation quality (such as adequacy and fluency) do not necessarily match up in the same translation. Therefore, a further challenge for ARP evaluation is the need to account for different quality criteria that may produce different rankings for evaluated systems.

The fact that statistical MT produces more fluent, but still not highly adequate translation indicates the need for ARP-based evaluation tools to account for different aspects of MT quality. An ARP-based model should predict which terms are more important for evaluation and which terms might be subject to greater LTV.

5.4.3. LTV and frequency weighting scores

To account for LTV within ARP-based MT evaluation models we used extensions of reference-proximity MT evaluation scores with weights of term "salience" within a text, such as tf.idf and S-scores. This extension is based on the assumption that these measures approximate the relative importance of lexical items for human translators and evaluators, and this will be necessarily reflected in LTV across different human translations. In this experiment tf.idf and S-scores were computed for each lexical type in each text in the DARPA corpus as described in Chapter 2. The tf.idf and S-scores were computed on the basis of both reference translations.

To establish the impact of these salience scores on LTV, we divided the unigrams from the two human translations into three classes: those found in both translations (the intersection set of unigrams) and those found in only one of the translations (two differences sets of unigrams). The distribution of tokens with different significance scores was examined in each of these classes. For the intersection class we did two calculations on the basis of each of the human translations. For the difference class we did the calculation only on the basis of the "native" reference translation.

Since the intersection set (IS) of unigrams is larger than the difference sets, the average tf.idf and S-scores were compared as well as frequency polygons of scores normalised by the size of each set. Table 1 presents the average scores for the sets:

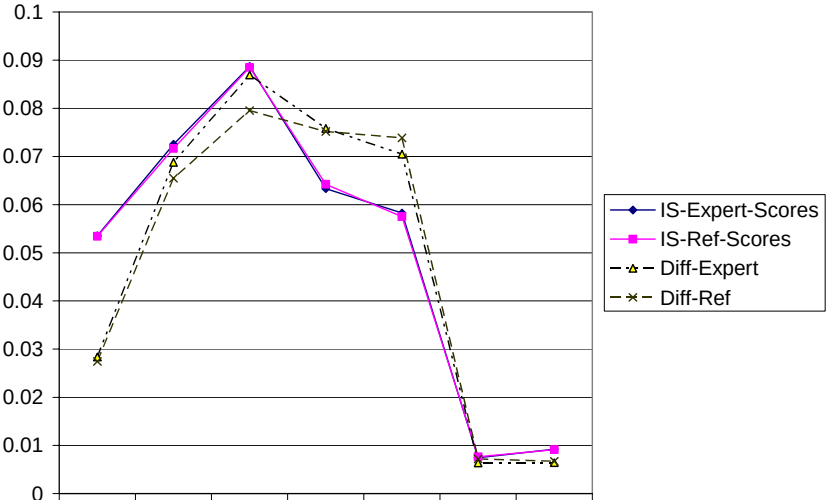
	tf.idf score	S-score
IS-Expert-Scores	2.6057	1.9825
IS-Ref-Scores	2.6146	2.0011
Diff-Expert	2.8290	2.2206
Diff-Ref	2.9200	2.3046

Table 1. Average scores: Intersection and Difference sets

These results are surprising because terms in the difference sets (those which were found to undergo LTV) have somewhat higher average significance scores that the supposedly more stable terms in the intersection sets. (Intuitively one may be inclined to believe that more significant words, such as content words, should be also more stable, and translation variation may be mostly due to the choice of low-salience functional words or different morpho-syntactic perspectives for a sentence).

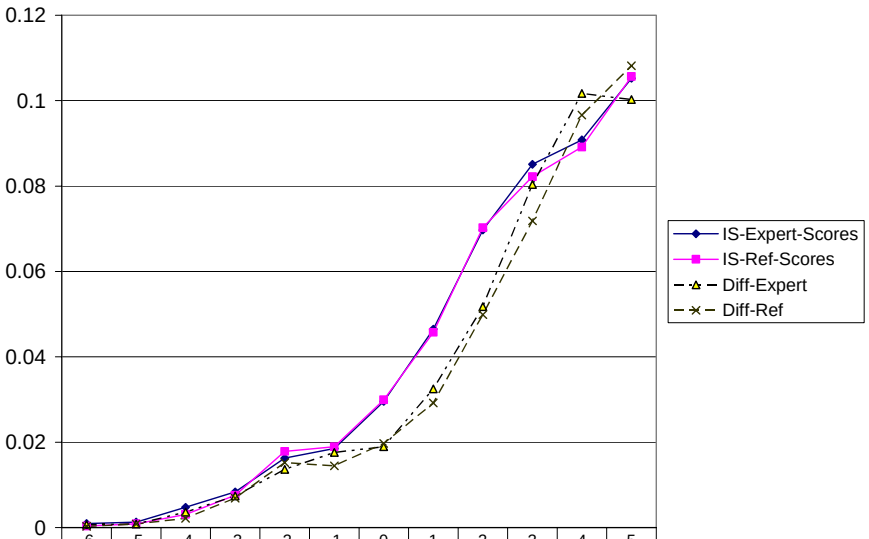
This means that stable words across human translations are somewhat less "salient" than the words with variable translation equivalents. Frequency polygons for each of these scores describe the distribution of significance scores for the

intersection and difference N-gram sets. Figures 3 and 4 compare the frequency polygons (normalised by the size of each N-gram set) for each set weighted by tf.idf or by S-scores.



	0	1.5	3	4.5	6	7.5	9
IS-Expert-Scores	0.0536	0.0725	0.0887	0.0633	0.0583	0.0074	0.0092
IS-Ref-Scores	0.0534	0.0717	0.0885	0.0642	0.0575	0.0076	0.0091
Diff-Expert	0.0284	0.0688	0.0869	0.0758	0.0705	0.0064	0.0064
Diff-Ref	0.0274	0.0655	0.0795	0.0752	0.0739	0.0072	0.0067

Figure 3. Frequency polygons weighted by tf.idf



	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5
IS-Expert-Scores	1E-03	0.001	0.005	0.008	0.016	0.019	0.03	0.046	0.07	0.085	0.091	0.105
IS-Ref-Scores	4E-04	9E-04	0.003	0.008	0.018	0.019	0.03	0.046	0.07	0.082	0.089	0.106
Diff-Expert	7E-04	8E-04	0.004	0.007	0.014	0.018	0.019	0.032	0.052	0.08	0.102	0.1
Diff-Ref	3E-04	8E-04	0.002	0.007	0.015	0.014	0.02	0.029	0.05	0.072	0.097	0.108

Figure 4. Frequency polygons weighted by the S-score

It can be seen from the charts that terms with different salience scores vary in their stability across independent human translations. In the first place there is no

significant difference for “under-represented” words (S-score < 0). The words with low and average salience scores (0 < tf.idf < 3; 0 < S-score <3) constitute the majority of the words used in texts. They tend to be much more frequent in the intersection set, i.e., they are more stable across independent human translations. On the contrary, a relatively small number of highly salient words (S-score > 3; if.idf > 3) become more frequent in the difference set, therefore being subject to the greater translation variation.

The threshold of S-score = 1 accurately distinguishes function words from content words, and it can be seen that the majority of function words show clear stability, which is not substantially different from the stability of content words that are highly frequent in a corpus.

These results suggest that the words which are not salient within a given text usually have some optimal translation equivalents. Different human translators usually agree on these equivalents.

However, individual human translators are consistent in using words which are subject to great translation variation, which makes these words statistically salient. (Infrequent words can also become highly salient according to S-score, which includes normalisation by a relative frequency of a word in corpus, giving all words an equal chance to become statistically salient). This means that highly salient units typically do not have ready translation solutions and require some “artistic creativity” on the part of human translators. Such words also give the translators a degree of freedom, making translation to some extent a creative process, even supposedly “non-computable” or “non-algorithmic” (cf. Penrose, 1990), which involves creative invention of translation strategies.

Finding out a proper translation strategy for such unstable words is very important for the general quality of the text, since highly salient words make the biggest contribution to the texts general content, and matter most of all in evaluation of the text quality by human judges.

The following sentence is an example of LTV related to the absence of a clear-cut translation strategy; transformations are applied differently by the two human translators:

> **ORI** : *Le président de la chambre d'accusation doit **rendre** un **avis** de clôture, ouvrant un **délai** de vingt jours pour les **requêtes** des diverses parties, suivi d'un arrêt de "soit communiqué" pour le **règlement** du dossier par la parquet général de Lyon.*
>
> **REF** : *The Director of the Public Prosecutor's Office must **give** a closing **decision**, which will open a 20-day **period** for the various parties to file*

petitions, *after which no papers may be sent to the public prosecutor so that the Office of the Public Prosecutor of Lyon can* ***prepare*** *the case.*

EXP : *The presiding judge of the Court of Criminal Appeals is to* ***render*** *a closing* ***opinion***, *thus establishing a twenty-day* ***deadline*** *for* ***requests*** *from the various parties, followed by a "may it be communicated" order for* ***settlement*** *of the case by the Lyon public prosecutor's office.*

The results presented suggest that there is a potentially serious problem for ARP-based approaches to MT evaluation: the most important terms in translation are the most unstable ones, which may not be necessarily present in any number of human reference translations. However, this problem may be partly solved by assigning different weights to highly salient and low significant N-gram matches in a reference translation and the evaluated text.

5.4.4. Conclusions for the experiment

The discovered difference in the tf.idf and S-scores for terms that are subject to various degrees of translation variation indicates that there is a link between the potential stability of units across independent human translations and their "salience" within a given text. Highly significant words, which are consistently used within a single translation, were found to be the most unstable across different translations. The possible reason for this fact could be that translation of significant units typically requires invention of some novel translation strategy.

The results also indicate that there exist fundamental limits on using data-driven approaches to MT, since the proper translation for the most important units in text may be not present in the corpus of available translations. Discovering the necessary translation equivalent might involve a degree of inventiveness and genuine intelligence, because the set of translation equivalents for most salient items is open, so the solutions will not be found in any pre-arranged data source.

A systematic solution to this problem would require generation of translation equivalents on-the-fly from pre-defined or learnt translation strategies. Data-driven approaches to MT will have to move from learning translation equivalents to learning translation strategies and procedures and to the ability to generate novel translation equivalents in the process of translation.

Future research in this direction could involve testing the applicability of the proposed method for highly-inflected languages, where N-gram scarcity is higher, finding a linguistic interpretation of the significance weights, and establishing the potential limits of legitimate variation across multiple human translations of a single text.

Conclusions

The main suggestion put forward in the theses follows from the presented experimental results: there are fundamental limits on MT quality, achievable by existing equivalent-based approaches to MT (either rule-based or data-driven). Currently equivalent-based approaches do not generalise translation strategies (in particular, non-literal, or oblique, strategies), which are used by human translators for inventing novel non-compositional translation equivalents. Equivalent-based approaches are inherently limited to translating either previously "seen" units in the ST, or constructs, which can be derived from such units in a compositional way, and where this compositionality can be paralleled in the TT. These approaches do not provide adequate models for "processing" phenomena that frequently occur in human translation: inventing novel translation equivalents along the lines of known translation strategies, changing the "perspective", or the point of view on a situation, dynamically assessing relevance of information and preserving only the most relevant levels, harmonising translation equivalents which may compete between each other on different linguistic levels, identifying levels which have to be lost according to translation norms for a particular genre, e.g., etymology of proper names or certain metaphorical uses of common words.

In the thesis I tried to show that such cases cannot be systematically covered simply by extending knowledge sources available for MT systems: "there is no data like more data" cannot be a solution here. The central point is that even though equivalent-based MT is powerful enough to account for all individual examples (including all examples presented in this thesis) – via extending an equivalent-based MT system and modifying the structure of a database of translation equivalents, but a system can never acquire a complete set of such examples: there will always be the need to generate them "on the fly" along the lines of different types of known translation strategies. MT should move from learning individual equivalents to generalising and learning such strategies, move from data-oriented to intelligence-oriented processing.

Contrary to common-sense intuition, novel translation equivalents often have to be created in a non-compositional way (either because compositionality cannot be mirrored in a TL, or because a novel unit in the SL is created in a non-compositional way). Therefore, no corpus will be large enough and will ever be able to cover such units, unless there is a way to learn not individual equivalents, but more general strategies for dealing with these units and to "dynamically" generate novel equivalents.

For data-driven methods this means that whenever such dynamic strategies were applied by human translators in the parallel training corpus, the fragments will become "poorly-adaptable", and therefore very risky for training an equivalent-based system. Such equivalents will almost certainly remain idiosyncratic; it would be very hard to "decompose" them safely. Only a model for generating new equivalents dynamically within general translation strategies will allow us to take "oblique" examples from parallel corpora safely onboard – obviously with a purpose of identifying and learning these general oblique translation strategies.

These limitations have parallels in the early years of MT, when it became evident that having an automated dictionary is not enough for creating a good translation (e.g., Kay, 1979). Now we see that even having a very flexible database of translation equivalents, a wide-coverage grammar and robust parsing algorithms is not sufficient: there is a need to apply translation strategies (generalised from these individual equivalents) in an intelligent, and to a great extent – "creative", dynamic way. It appears that even until now a "naïve" model of word-by-word translation which has its roots in an early "dictionary" metaphor wasn't abandoned completely. It still has its impact on the current research paradigm and obscures a more general view on the translation process.

This suggestion could be an explanation for the surprising finding presented in Chapter 1: in general, the quality achieved by state-of-the-art MT systems in terms of objective quality parameters (such as the number of N-grams matches found in a corpus) is still far behind the quality of human translation, and the difference between the best and the worst MT systems is much smaller that the gap between even the best system and a human translation. This finding justifies intuitive feeling that at present MT quality is not fit for the market of texts translated for publication, and is not really in competition with human translators, despite several decades of research.

However, the presented experiments also suggest that Information Extraction offers a way to overcome some of these fundamental limits by adding more sophisticated "intelligence" to the process of looking up databases of translation equivalents. In particular, NE recognition gives a clear example how MT quality can be improved not via extending databases of equivalents, but via "cleverer" processing of the ST (since the set of organisation names is open and highly dynamic it cannot be stored in a gazetteer of an IE system, which needs to use some "intelligent" processing techniques. Surprisingly, even though a module for processing organisation names is a feasible and useful extension for MT, it hasn't been implemented so far in any of the tested state-of-the-art MT systems, possibly because such an extension has been not in line with predominant MT "philosophy",

primarily concerned with extending systems' coverage, not systems' "intelligence". Supposedly, it is contrary to expectations of the mainstream MT thinking that the quality can be improved without extension of system dictionary or grammar, but with the very opposite – restricting the application of the dictionary and grammar rules for units, which can be identified in the ST by some intelligent processing techniques and whose translation strategy is clearly different.

The results noticeably call for changing the philosophy of MT from coverage-oriented towards intelligence-oriented processing, which so far have been marginal in MT. The results also suggest that the intelligent processing techniques can be applied outside the realm of NE, to a much wider variety of units.

Another finding comes from comparison of experimental results presented in Chapter 2 and Chapter 3: a particular type of NEs – organisation names –can be efficiently used both for "performance-based" MT evaluation and for improving MT quality, which suggests that significance of MT evaluation goes far beyond its usual role in quality assurance and monitoring the progress of MT development. The search for more adequate formal evaluation criteria, which more closely correlate with human intuitive judgements about MT quality, can pave the way for discovering new approaches and techniques capable of boosting MT quality. These findings suggest that in the general case, things which work for evaluation will also work for improving MT.

Such a technique, which improves the accuracy of MT evaluation, is salience weightings of terms in text. The results presented in Chapter 4 show that weighting terms with statistical "salience" scores (the standard tf.idf scores and S-scores, specifically designed for IE purposes) improves correlation between automated and human evaluation both for fluency and adequacy evaluation. It is also reasonable to interpret these results within the suggested "intelligent processing" approach to MT: the improvement is achieved via letting the evaluation tool to concentrate on more important bits of information in text, and to ignore less important items. This approach resembles the experiment with NEs in that restricting some types of items from being taken into account, instead of adding any new knowledge sources, is found to be beneficial.

Salience weighting can be viewed as a rough approximation of identification of most important lexical items in text, and it is also domain-independent. In the thesis I put forward a suggestion that similar salience weighting approach may improve the quality of data-driven MT systems (EBMT and SMT) by allowing the system to make a clearer distinction between important items which need to be translated and less important structures and units, whose translation is either optional or shouldn't be done at all (e.g. function words, or occasional metaphorical

usage of content words). An alternative ways of grading relevance of items in the ST for translation and ensuring consistency between individual sentences and the whole text structure could be to use full-scale rule-based IE in a specific subject domain. At this stage the technology can make a step from roughly approximating the importance of ST units to accurate analysis of their relevance in the ST. In this respect IE systems can be specifically tuned to meet specific the demands of MT technology. Testing this suggestion will also require an access to the source code of MT systems and scenario template filling modules of IE systems and will be the matter of my future work.

Experiments on extending flexibility of MT evaluation metrics presented in Chapter 5 examine the problem of increasing usability of automated evaluation scores beyond correlation issues. The results of the first experiment show that intelligent "knowledge-light" processing techniques can be used to evaluate text difficulty for translation. This measure of difficulty can further be used to normalise MT evaluation scores of large general-purpose MT systems. For evaluated language pairs and genres it was found that difficulty of a particular text type for translation is most closely related to the issue of word-sense disambiguation (an average number of possible word senses per word, which is closely correlated with the number of syllables per word). Syntactic complexity or sentence length plays much smaller role in determining the translation complexity. However, the most interesting and counter-intuitive fact is that the relation goes in opposite direction from an expectation that greater number of word senses results in a more difficult text. On the contrary, texts where the average number of word senses per token was smaller (EU Whitepaper) were found to be more difficult than texts with more word senses per token (emails). A possible explanation of this result could be that WSD in translation doesn't follow a naïve word-for-word or fixed-number-of-senses model, and the picture should be more complex. Possibly, it is inappropriate to think of word senses as unordered collections of some fixed possible meanings, and the need for disambiguation really arises in context of some unambiguous (and long) words with very precise meanings.

It is possible that in easier texts, like emails, word senses can exist in some highly abstract state without the need for disambiguation, and therefore translation requires much smaller disambiguation effort – default translation supplied by an MT systems are fine. On the contrary, harder texts, like legal documents, require greater disambiguation effort for potentially ambiguous words, because there is a need to synchronise them with exact meaning of (usually unambiguous) terminology. Therefore, default translations for such polysemous words more frequently fail, their meaning has to be much more precise. The text itself may have generated WSD

problems “on-the-fly”, creating “unexpected”, context-dependent, and out-of-dictionary word senses.

The results of this experiment give yet another example how MT evaluation can go beyond its usual function of quality assurance or monitoring the development progress and identify potential gaps and problems in our knowledge of language and translation process, as well as point to possible solutions and bring in new knowledge, which can boost MT quality beyond the limits of current MT technology.

Problems of designing an optimal MT evaluation set-up were addressed by the experiments on determining minimal size of MT evaluation corpus and on applying automated MT evaluation methods to other languages. The results suggest that reliable correlation figures between automated and human scores can be obtained with a corpus containing at least 7200 words, but the effects related to the lack of text-level “homogeneity” are filtered out when the size approaches 12000 words – at this stage automated scores become stable, and start to reflect the general level of performance of an MT system on a particular text type or genre, not the translation difficulty of individual sentences and texts within this genre.

Correlation between automated MT evaluation scores and human judgements was found to be similarly high for all evaluated target languages. However, in order to predict acceptability levels of MT output or the absolute values of human judgements, we will need to know the two parameters of regression line: the slope and intercept. These parameters were found to be specific to particular combinations of evaluated text types and target languages.

The experiment on relating words’ salience and their stability across several independent human translations points out to a potentially important limit of current MT technology. The experiment shows that the relation between salience weighting of terms in text and legitimate variation of these terms also goes in the opposite direction to naïve expectations. One might expect that more salient terms, e.g., content words, should be more stable across different independently created human translations than less salient terms, e.g., function words. It is plausible to believe that different people may choose different prepositions, phrasing of sentences or syntactic frameworks while translating the same sentence, but they should agree in important words, which describe events, event participants and relations. Once again, the results of the experiment show that contrary to such expectations the most salient words in text are also the most unstable across independently created human translations.

Qualitative analysis of such cases indicates that these are the words which don't have a readily available translation equivalent, and require some "inventiveness" or "creativity" on the part of human translator. There is no single straightforward and "correct" way to translate these items. Implication for data-drive MT could be that the most central lexical items are also least "adaptable" (in terms of EBMT). Being on the top of the relevance hierarchy, they supposedly act as organising centres for the entire sentence structure, so the resulting parallel sentences became also inadaptable. As a result, not much can be learned from aligning these sentences, since the solution to the translation problem makes sense only in its entirety, so it is hard to identify safe points of decomposing such asynchronous structure into smaller reusable aligned fragments.

The results suggest that there is supposedly an open set of possible solutions to central translation problems of the text, and the solutions therefore are non-local – they spread out into the level of the entire sentences or even the entire text, so there are very few safe points for decomposition of such solutions into smaller reusable fragments, which are independent of each other. For EBMT this means that the boundary friction problem will never be solved if we try to learn individual translation equivalents. For SMT these results indicate that it is not possible to avoid learning noise together with learning translations. Larger amounts of text will never be able to filter out that noise – it will be constantly introduced, because inadaptable fragments are central for the text, not marginal, and will always pose a problem if we try to translate any seriously challenging text, which goes beyond some simple artificially constructed or carefully chosen examples, which nicely fit into classical "word-for-word" model.

On the other hand, the amount of these "unstable" solutions has to be smaller for certain less imaginative genres (e.g., for maintenance manuals), and greater for more creative writing styles (like fiction). This fact goes in line with empirical observations that MT is much more useful for "mundane" text types, and becomes practically useless if translation problems become more and more creative.

A possible way of addressing this problem has been already mentioned: we should try to develop (in a rule-based framework) a model of translation strategies and transformations, or (in the case of data-driven MT) try to learn such strategies and rules of their application from observing individual translation examples. Ideally, a data-driven MT system should be able to fit any cases of indirect translation (found in parallel training corpus of human translation) in its evolving set of possible translation strategies and then to apply these strategies for new, previously unobserved structures. It is too early to speculate whether this approach will provide a solution to most serious and truly "creative" translation problems,

which are often heartily discussed by human translators, but I hope this “generalised translation strategy” approach would be a step in the right direction.

To summarise, the central points defended in my thesis are the following:

1. Automatic identification of essential bits of information in text with Information Extraction methods allows us to overcome limitations for MT that are related to unequal relevance and unequal information load of translation equivalents potentially found in the source text. IE methods can efficiently deal with the intelligent processing bottleneck of current MT architectures.

2. Filtering out redundant information is as much important for MT as traditionally emphasised extension of knowledge sources available for MT systems; since it meets the need of ranking relative relevance of translation equivalents which could fire over the same segment. This ranking can be efficiently modelled with IE techniques, e.g., IE-oriented statistical term salience scores, proposed in the thesis. Improvements of MT quality with NE recognition and higher correlation for IE-oriented MT evaluation measures, described in the thesis, reflect this phenomenon.

3. MT evaluation is an integral a part of a research cycle in MT, it allows researchers to discover new knowledge about natural phenomena (e.g., cognitive or social facts) related to translation process. There is a bi-directional link between the performance of MT evaluation metrics and the ways of improving MT: the ideas which work for MT evaluation may improve MT quality on the development stage and vice versa. Therefore, new knowledge discovered on the evaluation stage can be reused for linguistic engineering tasks and improve the quality of applications (e.g., new knowledge about language comprehension / generation and about selection of optimal translation strategies can be used in improving source language analysis, transfer and target language generation algorithms in MT).

4. During the project the following tools and resources were developed:

– open-source MT evaluation toolkit, based on the proposed weighted N-gram model;

– a multilingual MT evaluation workspace for calibration of automated MT evaluation scores on texts of several genres.

These resources have been put into the public domain, so the results of the research can make contribution to the development of an open-source framework for MT development and evaluation, which may integrate into MT design some other open-source NLP technologies, such as IE methods described above, and may test some of the hypothesis and ideas presented in this thesis.

A possible direction of future work will be developing an open-source example-based MT system, which will implement some of the suggested principles. For instance, it and can be guided by IE in generating and selecting translation equivalents which will be consistent with the text-level information provided by IE templates, make full use of NE annotation of the source text and the target text. The IE algorithms for this system could also be adapted to meet the demands of EBMT, for instance for identifying and annotating different translation strategies in the training corpora. The development of an open-source EBMT system will solve the problem which seriously limited the extent of experiments in my dissertation: no access to MT source code. Further there is a brief outline of a proposal for developing such a system is presented.

Recent advances in MT technology made it much more useful for professional translators. MT considerably increases the productivity of translators' work by automating the search for direct translation equivalents and previously translated fragments and allowing the translators to concentrate on more creative tasks, which go beyond the direct translation strategy, e.g., on translating items that do not have available translation equivalents or cannot be decomposed into sequences of previously translated fragments. Data-driven approaches to MT, such as SMT and EBMT, as well as the use of large corpora for the development of Rule-Based translation systems allowed the developers to overcome the most serious technological bottleneck – the problem of data acquisition. Modern MT systems may rely on large collections of aligned human translations, semi-automatically constructed dictionaries, terminological databases, wide-coverage grammars, ontologies and other extensive knowledge sources, which increase their coverage and make them scalable to new subject domains.

However, these developments also reveal limitations of the current approaches to MT (limitations which were less visible before, since the data acquisition problem used to be much more serious). In the first place, the limitations now come on the processing side, hindering MT from taking full advantage of the available data sources. For example, MT systems often need to filter out redundant translation equivalents (items not intended for translation) (Babych and Hartley, 2004b: 629), which may be as beneficial for MT quality as the widely suggested use of extensive knowledge sources. Strategies of handling the databases of translation equivalents need to become more flexible, they need to be extended beyond the direct "look-up" procedures and systematically cover also oblique translation procedures used by human translators, such as "transposition" or "modulation" (Vinay and Darbelnet, 1995). To a certain extent these cases require simulation of advanced aspects of

human intelligence, since many translation problems are non-trivial and require creative and flexible solutions.

It is likely that no single approach and no single system architecture will be able to solve all theoretical and technological problems in MT. The experiments that properly accommodate different "ideologies" are much more promising (e.g., Imamura et al., 2004: 99-105). The lack of a wider technological environment for translation models, where more general models from the field of Artificial Intelligence (AI) can be implemented, seriously impedes the progress in achieving better MT quality (c.f. Key, 2003: xix). Better results could be attained if MT developers concentrate on combining specific NLP and AI techniques in the process of translation, which addresses a diverse variety of particular translation problems, instead of looking for a single overreaching MT model, methodology or architecture (e.g., statistical, syntax-driven, compositional, etc.). This requires a common experimental ground for evaluating NLP modules, where the developers can test usability of their technologies for MT quality (and more generally – for natural language understanding and/or natural language generation quality), and not just claim in their papers that a particular technology can be useful or even plays a "vital role" in MT applications (e.g., Mitkov, 2002: xii).

Open-source software systems proved to be very successful as co-operative experimental frameworks for NLP research groups. Examples of such systems that are being developed jointly by the NLP community include GATE, which is used for Information Extraction and Text Mining tasks by academic, governmental and commercial organisations, and NLTK (http://nltk.sourceforge.net/) used for teaching Computational Linguistics. These systems adopt modular architecture and supply a framework for interaction between different modules, which can be implemented and tested independently of each other. The systems already include many modules, which are usable for MT technology, e.g., the Named Entity recognition module in GATE or the Word Sense Disambiguation module in NLTK. However, at the moment there is no transparent open-source MT system, which may implement different MT architectures, integrate a variety of potentially useful open-source NLP modules in a flexible way and measure their influence on MT quality against its baseline performance.

The goal of the proposed project is to create an open-source MT development environment able to integrate independently developed NLP and AI modules and flexibly build alternative system architectures from such modules. Such environment may function as a common experimental ground, where research groups could test applicability of their technologies to MT and compare the impact

of their modules on MT quality with the system's baseline performance or with the impact of modules developed by other research groups.

Monolingual analysis and synthesis modules can annotate features that may become a basis for formal models of human translation. For example, such models can condition application of indirect translation procedures, which are used by human translators, on sets of the identified features. Therefore the MT development environment can help to validate theoretical models of some phenomena found in parallel texts, so it could be useful for researchers in Translation Studies.

Such environment can be made practical for teaching courses on MT in the curriculum of Computational Linguistics (following the examples of GATE and NLTK, which are now extensively used for teaching Information Extraction, Machine Learning, Parsing, etc.). Example-Based MT architecture is most suitable as the baseline system implementation for the proposed MT development environment. EBMT integrates both data-driven and rule-based techniques (Carl and Way, 2003: xix), it can be built around a reasonably small aligned corpus in a given subject domain and may naturally incorporate statistical techniques as well as linguistic knowledge and be transparent for the developers of the system. EBMT architecture can store examples as annotated linguistic structures (Way, 2003: 444), making use of arbitrary sophisticated linguistic representations, e.g., part-of-speech annotation, lemmatisation, automatically aligned syntactic trees (Groves et al., 2004), semantic representations, e.g., preference semantics formulas (Wilks, 1975), qualia structures (Pustejovsky, 1995), annotation of semantic classes, synsets, etc. On the other hand EBMT architecture does not necessarily require the use of such resources; it may be implemented with a resource-light approach, although availability of additional linguistic resources may substantially boost the MT quality.

Implementation of particular language directions for EBMT depends on availability of parallel texts for a given language pair. The size of the parallel corpus influences the quality of EBMT, but relatively small (preferably word aligned) corpus could be a good starting point (c.f. Lavie et al., 2004: 116). Parallel resources in the STRAND database (Resnik and Smith, 2003) can be also used for the development. The core system can implement an algorithm of run-time retrieval of translation examples from sentences with morphological annotation, as described in (Andriamanankasina et al., 2003). Similarly, inductive learning can be used to predict word alignment in new translation examples to be included in the corpus.

In the framework of the proposed project the core EBMT engine later can be extended with MT-oriented Information Extraction module which can be based on

Sheffield's NE recognisers available in GATE for English (ANNIE) and for Russian (RusIE) – (Popov et al., 2004).

MT-oriented IE module can be implemented as an example of possible extension of the core EBMT system. IE technology can target specific MT problems such as annotating correct translation strategies for proper nouns using existing NER modules. Extended NER may distinguish different subclasses of Named Entities that require different translation strategies for a given language pair, such as transference ("do-not-translate" or "transliterate"), literal translation, etc. In our experiments NER module interacted with MT systems via "do-not-translate" lists, but proper integration of the modules into MT architecture will give further improvement, since cases of ambiguity between proper vs. common nouns within the same text could be properly addressed in order to avoid potential over-generation for the ambiguous items. Strings that require oblique translation may be annotated in advance and sent to specific MT modules which will apply appropriate translation transformations.

Other aspects of IE technology may be also useful for MT. Properly defined IE templates may be used as a kind of "Interlingua", which gives proper structure of recognised events and annotates roles of event participants. Such annotation can guide the core MT system in specifying the event structures in the target language, which will allow it to avoid typical mistranslations caused by the wrong event structure, especially when the events in the source and target texts are seen from different perspectives.

IE may provide annotation of relative salience of the lexical items in the source text, identifying lexical items and constructions, which are central for the meaning of the text, using some salience scores. This annotation allows the core MT system to prioritise lookup of translation equivalents in relation to the salience of terms. Salience of lexical items can be also used in other external modules, e.g., as a feature for Word Sense Disambiguation algorithms.

If the core EBMT system comes up with several translation variants for a source segment, IE modules could be run on these alternatives, doing performance-based evaluation via IE (Babych and Hartley, 2004c), e.g., they may attempt to identify Named Entities or event participants. The preferred candidate would have the closest number of identified items of a particular type to the number identified by IE in the corresponding source segment, or more generally – the structure of IE templates generated on the preferred target would be maximally isomorphic to the structure of templates filled from the source segment.

The suggested MT-oriented IE module can test the assumption that IE technology may provide the necessary flexibility for core MT systems, properly

addressing the problem of interaction between language, knowledge and the structure of the subject domain.

Interestingly, the proposed environment may allow the developers to extend the system in a principled way, providing a general structure for the extendibility of the system and ensure that transfer algorithms take full advantage of monolingual processing of the source and target texts. Richer linguistic annotations and new modules, such as Anaphora Resolution, Word Sense Disambiguation, etc. can be run on the aligned development corpus and can provide new features which accommodate translation shifts and transformations done by human translators. E.g., availability of Word Net hierarchies for source and target languages will allow the system to annotate the cases of using hyponyms and hyperonyms as aligned translation equivalents (Shveytser, 1988: 131), to generalise such cases (e.g., with supervised Machine Learning techniques) and to apply dynamically these translation transformations to new sentences. Similarly, an AI module that implements an automated reasoning system in a given subject domain may introduce appropriate annotation for the changes of the point of view on certain events, so the "modulation" translation procedures can be learnt, e.g.: "En.: *it is not difficult to show*" – Fr.: "*il est facile de démontrer*" [lit.: it is easy to show] (Munday, 2001: 57). Annotation of information structure will accommodate the procedures where professional translators change sentence's syntactic perspective in order to convey an appropriate order of presenting given and new information, e.g.: Rus.: *Иную позицию заняли Франция и Германия* [*lit.: "Different*$_{case.acc}$ *position*$_{case.acc}$ *took France*$_{case.nom}$ *and Germany*$_{case.nom}$] – Eng: *A different stand was taken by France and Germany* (An active sentence with the inverse word order was translated by a passive sentence, in order to preserve the information structure of the original) – (Breus, 2003: 23). Richer monolingual annotations and more sophisticated monolingual processing modules can give deeper insights into human approaches to translation, letting the system to simulate more complex indirect translation strategies.

An MT evaluation framework for such system may allow the developers to do qualitative error analysis and annotate the cases of improvement and deterioration, caused by introduction of the new modules. Thus it might be possible to make a direct comparison of competing approaches to different NLP problems from the point of view of their usability for MT and to evaluate viability of different MT architectures for different language pairs given linguistic resources available for those languages.

Bibliography

Akiba, Y., Imamura, K., & Sumita, E. (2001). *Using multiple edit distances to automatically rank machine translation output.* MT Summit VIII.

Akiba, Y., Sumita, E., Nakaiwa, H., Yamamoto, S., & Okuno, H. G. (2003). *Experimental Comparison of MT Evaluation Methods: RED vs. BLEU.* MT Summit IX.

Al-Onaizan, Y., & Knight, K. (2002). *Translating Named Entities Using Monolingual and Bilingual Resources.* 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia.

Andriamanankasina, T., Araki, K., & Tochinai, K. (2003). EBMT of POS-Tagged Sentences via Inductive Learning. In M. Carl & A. Way (Eds.), *Recent Advances in Example-Based Machine Translation* (pp. 225-252). Dordrecht: Kluwer Academic Publishers.

Babych, B. (2004). *Weighted N-gram model for evaluating Machine Translation output.* 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics, University of Birmingham.

Babych, B., Elliott, D., & Hartley, A. (2004a). *Calibrating resource-light automatic MT evaluation : a cheap approach to ranking MT systems by the usability of their output.* 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal.

Babych, B., Elliott, D., & Hartley, A. (2004b). *Extending MT evaluation tools with translation complexity metrics.* 20th International Conference on Computational Linguistics (COLING 2004), University of Geneva, Switzerland.

Babych, B., & Hartley, A. (2003). *Improving Machine Translation Quality with Automatic Named Entity Recognition.* 7th International EAMT workshop on MT and other language technology tools. Improving MT throught other language technology tools. Recourses and tools for building MT, Budapest, Hungary.

Babych, B., & Hartley, A. (2004a). *Extending the BLEU MT Evaluation Method with Frequency Weightings.* 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), Barcelona, Spain, 21-26 July, 2004.

Babych, B., & Hartley, A. (2004b). *Modelling legitimate translation variation for automatic evaluation of MT quality.* 4th International Conference on Language Resources and Evaluation (LREC 2004).

Babych, B., & Hartley, A. (2004c). *Selecting Translation Strategies in MT using Automatic Named Entity Recognition.* EAMT 2004 Workshop, Malta.

Babych, B., & Hartley, A. (2004d). *Comparative Evaluation of Automatic Named Entity Recognition from Machine Translation Output.* IJCNLP Workshop on Named Entity Recognition for Natural Language Processing Applications, Sanya, Hainan Island, China.

Babych, B., & Hartley, A. (2004e). Open Source MT evaluation toolkit (Version 01.1). Leeds. http://www.comp.leeds.ac.uk/bogdan/evalMT.html

Babych, B., Hartley, A., & Atwell, E. (2003). *Statistical Modelling of MT output corpora for Information Extraction.* Corpus Linguistics 2003 conference, Lancaster University (UK).

Bar-Hillel, Y. (2003/1960). The Present Status of Automatic Translation of Languages. In F. L. Alt (Ed.), *Advances in Computers* (Vol. 1, pp. 91-63). New York: Academic Press.

Barhudarov, L. S. (1975). *Язык и перевод (Language and Translation)*. Москва: Международные отношения.

Barrow, J. D. (2003). *The Constants of Nature. From Alpha to Omega*. London: Vintage.

Berger, A., Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Gillett, J., Lafferty, J., Mercer, R., Printz, H., & Ures, L. (1994). *The Candide system for Machine Translation.* ARPA workshop on Human Language Technology, San Mateo.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., & Roossin, P. S. (2003/1990). A statistical approach to Machine Translation. In S. Nirenburg & H. Somers & Y. Wilks (Eds.), *Readings in machine translation* (pp. 355-362). Cambridge, MA: MIT Press.

Breus, E. V. (2002). *Основы теории и практики перевода с русского языка на английский (Foundations of Theory and Practice of Translation from Russian into English).* Москва: Издательство УРАО.

Brew, C., & Thompson, H. (1994). *Automatic Evaluation of Computer Generated Text.* ARPA/ISTO Workshop on Human Language Technology.

Catford, J. C. (1965). *A Linguistic Theory of Translation*. London: Oxford University Press.

Chernyahovskaya, L. A. (1976). *Перевод и смысловая структура (Translation and Sense Structure)*. Москва: Международные отношения.

Church, K. (2000). *Empirical estimates of adaptation: The chance of two Noriega's is closer to p/2 than p²*. The 18th International Conference on Computational Linguistics.

Church, K., & Gale, W. (1995). Poisson mixtures. *Journal of Natural Language Engineering, 1*(2), 163-190.

Church, K. W., & Hovy, E. H. (1993). Good applications for crummy machine translation. *Machine Translation, 8*, 239-258.

Čmejrek, M., Cuřín, J., & Havelka, J. (2003). *Czech-English Dependency-based Machine Translation.* 10th Conference of the European Chapter of Association for Computational Linguistics (EACL 2003), Budapest, Hungary.

Collier, R. (1996). *Automatic template creation for information extraction, an overview* (Technical report CS-96-07). Sheffield: University of Sheffield.

Collins, B., & Somers, H. (2003). EBMT seen as case-based reasoning. In M. Carl & A. Way (Eds.), *Recent advances in Example-Based Machine Translation* (pp. 115-153). Dordrecht: Kluwer Academic Publishers.

Collins, B. (1998). *Example-based Machine Translation: an adaptation-guided retrieval approach.* PhD thesis, Trinity College, Dublin.

Cowie, J., & Lehnert, W. (1996). Information Extraction. *Communications of the ACM, 39*(1), 80-91.

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). *GATE: A Framework and Graphical Development Environment for robust NLP Tools and Applications.* ACL'02, Philadelphia.

Cunningham, H., Wilks, Y., & Gaizauskas, R. (1996). *GATE - a General Architecture for Text Engineering.* COLING-96, Copenhagen.

Etzioni, O., Cafarella, M., Downey, D., Popesku, A.-M., Shaked, T., Soderland, S., Weld, D. S., & Yates, A. (2004). *Methods for Domain-Independent Information Extraction from the Web: An experimental Comparison.* AAAI 2004.

Everitt, B. S. (1977). *The analysis of contingency tables*. London: Chapman and Hall.

Fass, D., & Wilks, Y. (1883). Preference Semantics, Ill-Formedness, and Metaphor. *American Journal of Computational Linguistics, 9*(3-4), 178-187.

Gachot, D., Lange, E., & Yang, J. (1998). The Systran NLP browser: an application of Machine Translation technology in Cross-Language Information Retrieval. In G. Grefenstette (Ed.), *Cross-Language Information Retrieval* (pp. 105-118): Kluwer.

Gaizauskas, R., Wakao, T., Humphreys, K., Cunningham, H., & Wilks, Y. (1995). *University of Sheffield: Description of the LaSIE system as used for MUC-6.* MUC-6.

Gaizauskas, R., & Wilks, Y. (1998). Information Extraction: Beyond Document Retrieval. *Computational Linguistics and Chinese Language Processing, 3*(2), 17-60.

Grishman, R. (2003). Information Extraction. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (pp. 545-560). Oxford: Oxford University Press.

Groves, D., Hearne, M., & Way, A. (2004). *Robust Sub-Sentential Alignment of Phrase-Structure Trees.* COLING 2004.

Gutt, E. (1991). *Translation and Relevance: Cognition and Context*. Oxford: Blackwell.

Hobbs, J. R., Appelt, D. E., Bear, J., Israel, D., Kameyama, M., Stickel, M., & Tyson, M. (1997). FASTUS: a cascaded finite-state transducer for extracting information from natural-language text, In E. Roche and Y. Schabes (Eds.), *Finite State Devices for Natural Language Processing.* (pp. 381-406). Cambridge, MA: MIT Press.

Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Hubey, M. (1999). *Mathematical Foundations of Linguistics*. München: Lincom Europa.

Hutchins, W. J., & Somers, H. L. (1992). *An Introduction to Machine Translation*. London: Academic Press.

Imamura, K., Okuma, H., Watanabe, T., & Sumita, E. (2004). *Example-based Machine Translation Based on Syntactic Transfer with Statistical Models.* COLING-2004.

Imamura, K., Sumita, E., & Matsumoto, Y. (2003). *Automatic Construction of Machine Translation Knowledge Using Translation Literalness.* 10th

Conference of the European Chapter of Association for Computational Linguistics (EACL 2003), Budapest, Hungary.

Jacquemin, C., & Bourigault, D. (2003). Term extraction and automatic indexing. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (pp. 599-615). Oxford: Oxford University Press.

Kay, M. (1979). *Syntactic Process.* 17th Annual Meeting of the Association for Computational Linguistics.

Kay, M. (2003/1980). *The Proper Place of Men and Machines in Language Translation* In S. Nirenburg & H. Somers & Y. Wilks (Eds.), *Readings in machine translation* (pp. 221-232). Cambridge, MA: MIT Press.

Kay, M. (2003). Introduction. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (pp. xvii-xx): Oxford University Press.

Kay, M., Gawron, M., & Norvig, P. (1994). *Verbmobil: A Translation System for Face-to-Face Dialog* (Vol. Lecture Note No. 33). Stanford: Center for the Study of Language and Information.

Kettunen, K. (1986). Letter to the Editor. Is MT Linguistics? *Computational Linguistics, 12*(1), 37-38.

Kittredge, R. (1982). Variation and homogeneity of sublanguages. In R. Kittredge & J. Lehrberger (Eds.), *Sublanguages: studies of language in restricted semantic domains* (pp. 107-137). New York: Walter de Gruyter.

Köhler, R. (1993). Synergetic Linguistics. In R. Köhler & B. B. Rieger (Eds.), *Contributions to Quantitative Linguistics* (pp. 41-51). Dordrecht: Kluwer Academic Publishers.

Landsbergen, J. (2003/1987). Montague Grammar and Machine Translation. In S. Nirenburg & H. Somers & Y. Wilks (Eds.), *Readings in machine translation* (pp. 239-254). Cambridge, MA: MIT Press.

Lavie, A., Probst, K., Peterson, E., Vogel, S., Levin, L., Font-Llitjos, A., & Carbonell, J. (2004). *A Trainable Transfer-Based Machine Translation Approach for Languages with Limited Resources.* Ninth EAMT Workshop, Valetta, Malta.

McEnery, T. (2003). Corpus Linguistics. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (pp. 448-465). Oxford: Oxford University Press.

Menzerath, P. (1954). *Die Architektonik des deutchen Wortschatzes (Architectonics of the German Lexicon)*. Bonn: Dummler.

Mitkov, R. (2002). *Anaphora Resolution*. London: Pearson Education.

Mitkov, R. (2003). Anaphora Resolution. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (pp. 266-283). Oxford: Oxford University Press.

Munday, J. (2001). *Introducing Translation Studies: Theories and Applications*: London: Routledge.

Neuman, G., & Xu, F. (2004). *Intelligent Information Extraction. Part 1: Introduction.* ESSLLI 2004 - The 16th European Summer School in Logic, Language and Information. Available: http://www.dfki.de/~neumann/esslli04/reader/ie-lec1.pdf.

Newmark, P. (1982). *Approaches to translation*. Oxford: Pergamon Press.

Newmark, P. (1988). *A textbook of translation*. London: Longman.

Nirenburg, S. (2003). Introduction. In S. Nirenburg & H. Somers & Y. Wilks (Eds.), *Readings in machine translation* (pp. 3-12). Cambridge, MA: MIT Press.

Nirenburg, S., & Raskin, V. (2004). *Ontological Semantics*. Cambridge, MA: MIT Press.

Nirenburg, S., Somers, H., & Wilks, Y. (Eds.). (2003). *Readings in machine translation*. Cambridge, MA: MIT Press.

Nirenburg, S., & Wilks, Y. (2000). Machine Translation at 50. In M. Zelkowitz (Ed.), *Advances in Computers*. New York: Academic Press.

Novikov, L. A. (1989). Лексикология (Lexicology) . In V. A. Beloshapkova (Ed.), *Современный русский язык (Modern Russian Language)* (pp. 165-236). Москва: Высшая школа.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). *Bleu: a method for automatic evaluation of machine translation* (research report RC22176 (W0109-022)): IBM.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *BLEU: a method for automatic evaluation of machine translation.* 40th Annual Meeting of the Association for the Computational Linguistics (ACL). Philadelphia, July 2002.

Penrose, R. (1989). *The Emperors New Mind*: Oxford University Press.

Popov, B., Kirilov, A., Maynard, D., & Manov, D. (2004). *Creation of reusable components and language resources for Named Entity Recognition in Russian.* LREC-2004.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge: MIT Press.

Rajman, M., & Hartley, A. (2001). *Automatically predicting MT systems ranking compatible with Fluency, Adequacy and Informativeness scores.* 4th ISLE Workshop on MT Evaluation, MT Summit VIII, Santiago de Compostela.

Rayson, P., & Garside, R. (2000). *Comparing corpora using frequency profiling.* Workshop on Comparing Corpora, 38th ACL

Reifler, E. (2003/1955). The Mechanical Determination of Meaning. In S. Nirenburg & H. Somers & Y. Wilks (Eds.), *Readings in machine translation* (pp. 21-36). Cambridge, MA: MIT Press.

Resnik, P., & Smith, N. A. (2003). Web as a parallel corpus. *Computational Linguistics, 29*(3), 349-380.

Richard, M. (2003). Introduction: Conceptions of Meaning. In M. Richard (Ed.), *Meaning* (pp. 1-35). Oxford: Blackwell.

Rosetta, M. T. (Ed.). (1994). *Compositional Translation*. Dordrecht: Kluwer Academic Publishers.

Saggion, H., Cunningham, H., Bontcheva, K., Maynard, D., Hamza, O., & Wilks, Y. (2004). Multimedia indexing through multi-source and multi-language information extraction: the MUMIS project. *Data & Knowledge Engineering, 48*(2), 247-264.

Salton, G., & Lesk, M. E. (1968). Computer evaluation of indexing and text processing. *Journal of the ACM, 15*(1), 8-36.

Schank, R. C. (1972). Conceptual Dependency: A Theory of Natural Language Understanding. *Cognitive Psychology, 3*(4), 532--631.

Schank, R. C. (1975). *Conceptual Information Processing*. Amsterdam: Elsevier.

Shveitser, A. D. (1988). *Теория перевода: статус, проблемы, аспекты (Theory of Translation: Status, Problems, Aspects).* Москва: Наука.

Somers, H. (2003). Machine Translation: latest developments. In R. Mitkov (Ed.), *The Oxford handbook of Computational Linguistics* (pp. 512-528). Oxford, NY: Oxford University Press.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and Cognition*. Oxford: Blackwell.

Stevenson, M., & Wilks, Y. (2001). The integration of knowledge sources in word sence disambiguation. *Computational Linguistics, 27*(3), 321-349.

Sudo, K., Sekine, S., & Grishman, R. (2001). *Automatic Pattern Acquisition for Japanese Information Extraction.* First International Conference on Human Language Technology Research.

Van Eynde, F. (1993). Machine Translation and Linguistic Motivation. In F. Van Eynde (Ed.), *Linguistic Issues in Machine Translation* (pp. 1-43). London and New York: Pinter Publishers.

Vinay, J.-P., & Darbelnet, J. (1958). *Stylistique comparée de l'anglais et du français: Méthode de traduction*, Paris: Didier, translated and edited by J. C. Sager and M.-J. Hamel (1995) as *Comparative Stylistics of French and English: A methodology for Translation.* Amsterdam and Philadelphia, PA: John Benjamins.

Way, A. (2003). Translating with Examples: the LFG-DOT Models of Translation. In M. Carl & A. Way (Eds.), *Recent Advances in Example-Based Machine Translation* (pp. 443-472). Dordrecht: Kluwer Academic Publishers.

Weaver, W. (2003/1949). Translation. In S. Nirenburg & H. Somers & Y. Wilks (Eds.), *Readings in machine translation* (pp. 13-17). Cambridge, MA: MIT Press.

White, J. (2003). How to evaluate machine translation. In H. Somers (Ed.), *Computers and Translation: a translator's guide* (pp. 211-244). Amsterdam and Philadelphia: J. Benjamins B.V.

White, J., Doyon, J., & Talbott, S. (2000). *Determining the tolerance of text-handling tasks for MT output.* LREC-2000, Athens.

White, J., OConnell, T., & OMara, F. (1994). *The ARPA MT evaluation methodologies: evolution, lessons and future approaches.* 1st Conference of the Association for Machine Translation in the Americas, Columbia, MD.

Wilks, Y. (1975). A Preferential Pattern Seeking Semantics for Natural Language Inference. *Artificial Intelligence, 6*, 53-74.

Wilks, Y. (1994). Development in MT research in the US. *Aslib Proceedings, 46*(4), 111-116.

Wilks, Y. (1997). Information Extraction as a core language technology. In M. T. Pazienza (Ed.), *Information Extraction. A multidisciplinary approach to an emerging technology* (pp. 1-9). Berlin: Springer.

Wilks, Y. (2003). Theoretical and Methodological Issues. Introduction. In S. Nirenburg & H. Somers & Y. Wilks (Eds.), *Readings in machine translation* (pp. 203-205). Cambridge, MA: MIT Press.

Wilks, Y., & Catizone, R. (1999). Can We Make Information Extraction More Adaptive?, *Information Extraction: Towards Scalable, Adaptable Systems. Proceedings of the SCIE99 Workshop* (pp. 1-16). Berlin and Rome: Springer-Verlag.

Witkam, T. (1988). *DLT - an industrial R&D project for multilingual MT.* 12th International Conference on Computational Linguistics (COLING 1988).

Xu, F., Kurz, D., Piskorski, J., & Schmeier, S. (2002). *A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and their Relations with Bootstrapping.* LREC 2002, the third international conference on language resources and evaluation, Las Palmas, Canary Island, Spain.

Yngve, V. H. (2003/1957). A framework for syntactic translation. In S. Nirenburg & H. Somers & Y. Wilks (Eds.), *Readings in machine translation* (pp. 39-44). Cambridge, MA: MIT Press.