Extending MT evaluation tools with translation complexity metrics

Bogdan BABYCH
Centre for Translation Studies, University of Leeds
Leeds, UK, LS2 9JT
bogdan@comp.leeds.ac.uk

Debbie ELLIOTT
School of Computing
University of Leeds
Leeds, UK, LS2 9JT
debe@comp.leeds.ac.uk

Anthony HARTLEY
Centre for Translation Studies, University of Leeds
Leeds, UK, LS2 9JT
a.hartley@leeds.ac.uk

Abstract

In this paper we report 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, which otherwise are variable across texts of different types. The suggested approach makes a fairer comparison between the MT systems evaluated on different corpora. The translation complexity metric was integrated into two automated MT evaluation packages – BLEU and the Weighted N-gram model. The extended MT evaluation tools are available from the first author's web site: http://www.comp.leeds.ac.uk/bogdan/evalMT.html

1 Introduction

Automated evaluation tools for MT systems aim at producing scores that are consistent with the results of human assessment of translation quality parameters, such as adequacy and fluency. Automated metrics such as BLEU (Papineni et al., 2002), RED (Akiba et al, 2001), Weighted N-gram model (WNM) (Babych, 2004), syntactic relation / semantic vector model (Rajman and Hartley, 2001) have been shown to correlate closely with scoring or ranking by different human evaluation parameters. 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 [1] 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).

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.

In this paper we report 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 we compute normalised BLEU and WNM scores which are relatively stable across translations produced by the same general-purpose MT systems for texts of varying difficulty. We suggest 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.

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, we produced two sets of automated scores: 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. We interpreted these differences as reflecting the relative complexity of texts for translation.

For the French originals of all three sets of texts we computed resource-light parameters used in

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

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).

We computed Pearson's correlation coefficient *r* 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, we computed normalised ("absolute") BLEU and WNM scores 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.

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 our experiment we used DARPA scores for adequacy and fluency for one of the participating systems.

We obtained human scores for translations of the whitepaper and the e-mails 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. We also included translations produced by a non-professional, French-speaking translator 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] to [0, 5] scale).

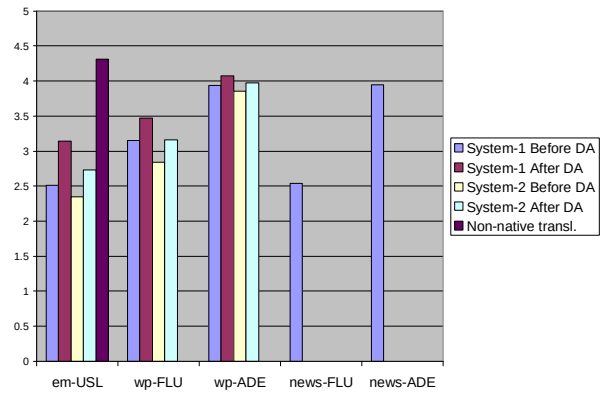

Figure 1. Human evaluation results

	S1	S1da	S2	S2da	NN
em [usl]	2.511	3.139	2.35	2.733	4.314
wp [flu]	3.15	3.47	2.838	3.157	
wp [ade]	3.94	4.077	3.858	3.977	
news [flu]	2.54				
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. We suggest 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.

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 we generated two sets of automated evaluation scores: BLEUr1n4 and WNM Recall.

BLEU computes the modified precision of N-gram matches between the evaluated text and a professional human reference translation. It was found to produce automated scores, which strongly correlate with human judgements about translation fluency (Papineni et al., 2002).

WNM is an extension of BLEU with weights of a term's salience within a given text. As compared to BLEU, the WNM recall-based evaluation score was found to produce a higher correlation with human judgements about adequacy (Babych, 2004). The salience weights are similar to standard tf.idf scores and are computed as follows:

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

where:

- $P_{doc(i,j)}$ is the relative frequency of the word w_i in the text j; ("Relative frequency" is the number of tokens of this word-type divided by the total number of tokens).
- $P_{corp\text{-}doc(i)}$ is the relative frequency of the same word w_i in the rest of the corpus, without this text;
- 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.
- $P_{corp(i)}$ is the relative frequency of the word w_i in the whole corpus, including this particular text.

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

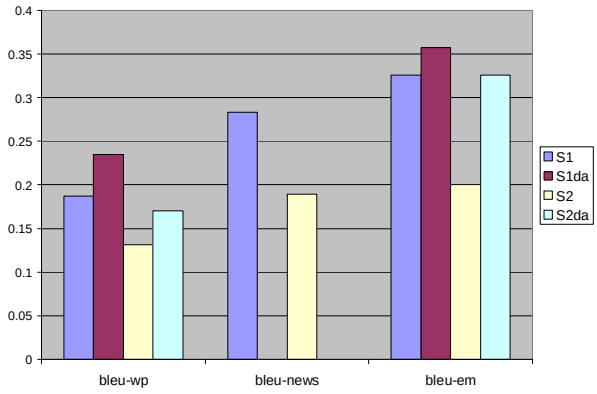

Figure 2. Automated BLEUr1n4 scores

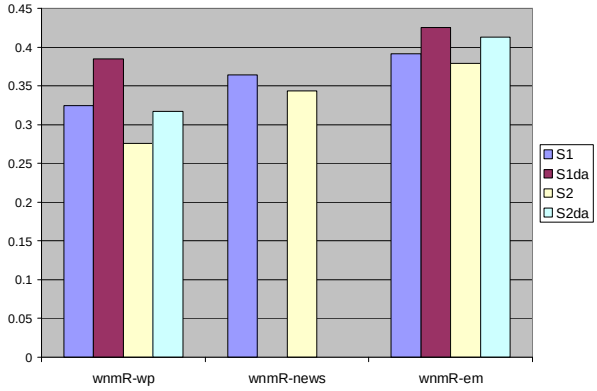

Figure 3. Automated WMN Recall scores

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 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 our experiment, we examined the following resource-light parameters 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	17.3	15.7	19.65	2
news	27.8	14.7	21.4	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	0.92	-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.

6 Normalised evaluation scores

We used the ASW parameter 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. We suggest using the difficulty of the news texts in the DARPA 94 MT evaluation corpus 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.

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

$C = ASW_{evalText} / ASW_{DARPAnews}$.

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

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		0.19	
bleu-em	0.747	0.243	0.267	0.15	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. We also measured the improvement in stability of the normalised scores as 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.

7 Conclusion and future work

In this paper, we presented empirical evidence for the observation that the complexity of an MT task influences automated evaluation scores. We 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. We suggest 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.

Future work will involve empirical testing of this suggestion as well as further experiments on improving the stability of the normalised scores by developing better normalisation methods. We will evaluate the proposed approach 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. Another direction of future research is comparison of stability of evaluation scores on subsets of the evaluated data within one particular text type and across different text types.

Acknowledgments

We are very grateful for the insightful comments of the three anonymous reviewers.

References

Y. Akiba, K. Imamura and E. Sumita. 2001. U*sing multiple edit distances to automatically rank machine translation output.* In "Proc. MT Summit VIII". pages 15–20.

B. Babych. 2004. *Weighted N-gram model for evaluating Machine Translation output.* In "Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics". M. Lee, ed., University of Birmingham, 6-7 January, 2004. pages 15-22.

M. Cmejrek, J. Curin and J. Havelka. 2003. *Czech-English Dependency-based Machine Translation.* In "Proceedings of the 10th Conference of the European Chapter of Association for Computational Linguistics (EACL 2003)". April 12th-17th 2003, Budapest, Hungary.

K. Imamura, E. Sumita and Y. Matsumoto. 2003. *Automatic Construction of Machine Translation Knowledge Using Translation Literalness.* In "Proceedings of the 10th Conference of the European Chapter of Association for Computational Linguistics (EACL 2003)". April 12th-17th 2003, Budapest, Hungary.

M. Hubey. 1999. *Mathematical Foundations of Linguistics*. Lincom Europa, Muenchen.

R. Koehler. 1993. *Synergetic Linguistics.* In "Contributions to Quantitative Linguistics", R. Koehler and B.B. Rieger (eds.), pages 41-51.

P. Menzerath. 1954. *Die Architektonik des deutchen Wortschatzes.* Dummler, Bonn.

R. Mitkov. 2002. *Anaphora Resolution.* Longman, Harlow, UK.

K. Papineni, S. Roukos, T. Ward, W-J Zhu. 2002 *BLEU: a method for automatic evaluation of machine translation.* In "Proceedings of the 40th Annual Meeting of the Association for the Computational Linguistics (ACL)", Philadelphia, July 2002, pages 311-318.

M. Rajman and T. Hartley. 2001. *Automatically predicting MT systems ranking compatible with Fluency, Adequacy and Informativeness scores.* In "Proceedings of the 4th ISLE Workshop on MT Evaluation, MT Summit VIII". Santiago de Compostela, September 2001. pages. 29-34.

J. White, T. O'Connell and F. O'Mara. 1994. The ARPA MT evaluation methodologies: evolution, lessons and future approaches. Procs. 1st Conference of the Association for Machine Translation in the Americas. Columbia, MD, October 1994. 193-205.