Comparing Rule-based and Statistical MT Output

Gregor Thurmair

linguatec
Gottfried Keller Str. 12
D 81245 Munich
g.thurmair@linguatec.de

Abstract

This paper describes a comparison between a statistical and a rule-based MT system. The first section describes the setup and the evaluation results; the second section analyses the strengths and weaknesses of the respective approaches, and the third tries to define an architecture for a hybrid system, based on a rule-based backbone and enhanced by statistical intelligence.

This contribution originated in a project called "Translation Quality for Professionals" (TQPro)[1] which aimed at developing translation tools for professional translators. One of the interests in this project was to find a baseline for machine translation quality, and to extend MT quality beyond it. The baseline should compare state-of-the-art techniques for both statistical packages and rule-based systems, and draw conclusions from the comparison. This paper presents some insights into the results of this work.

1 Baseline

The experiment was to compare the state-of-the-art quality of MT, and it used a current statistical MT package and a commercial rule-based MT system.
The material was provided by SAP; it consisted of Translation Memory material, German to English, more than 100.000 segments in the domain of the R/3 system, to have sufficient training data for a statistical package.

1.1 Statistical MT

The statistical analysis and translation was done by the team of RTH Aachen; this team had the best results in the Verbmobil project (Vogel et al. 2000) and is a leading center of statistical MT in Europe (Och et al. 2003).

Setup

The data were processed as follows: After a preprocessing step, the material was split into a training corpus (with 1.068 mio German and 1.128 mio English tokens, representing 44.400 German and 26.600 English types, respectively). This was used as input for the alignment template SMT system to train the MT.
A test corpus (5% of the corpus) was then analysed, of which all sentences of (randomly) of 14 tokens of length and containing no unknown words were selected. This resulted in 68 sentences.

Evaluation

These sentences were evaluated by splitting them into three categories:

- **grammatical**: This means the sentences are syntactically correct, and convey the content.
- **understandable**: This means the sentences are incorrect but still convey the content (without reference to the source text).
- **wrong**: This means that the sentences cannot be understood without reference to the source text.

Such an evaluation scheme is a common standard in commercial MT development, often used for quality assessment[2].
About 10% of the resulting 68 sentences contain ill-formed input (incorrect German sentences: segmentation, agreement, and syntactic errors), which is a realistic figure. With the translations, a reference human translation (resulting from the SAP memory production) is available.
The resulting translation quality is as follows:

grammatical	16	23,5%
understandable	31	45,6%
wrong	21	30,9%

It can be seen that there is a significant amount of understandable results, while the really good and really bad sentences are less frequent. This underlines the robustness of such an approach. Together the good + understandable sentences are close to 70%. It should be noted, however, that from a practical point of view, understandable sentences need to be post-edited, while for grammatical sentences this is not necessarily the case.

Improvements

The authors propose some improvements to these results like: morphological analysis of German noun compounds, special treatment of variable and product names, lookup of (manual) lexicon (cf. Nießen/Ney 2000).
Such improvements point into the direction of creating a hybrid system, with statistical basis and additional linguistic features to improve the statistical machinery.

[1] This project (IST-1999-11407) has as partners: SAP, Lotus Ireland, SailLabs, and CST on the development side, and CAT technologies and Logoscript on the user and testing side. Details are given in (Thurmair, 2000).

[2] Note that the notion of a "word error rate" as used in teh NIST evaluations (NIST 2001) is not a suitable evaluation concept for translation as there is not such a thing as a 'canonical' or 'reference translation' from which deviations could be computed: Three human translators produce four different versions of a text, all of which they claim to be correct.

1.2 Rule-based MT

In a second evaluation step, the output of the statistical MT was compared to a commercial rule-based MT system (linguatec's "Personal Translator" German-to-English).

Setup

The system was basically used as a raw MT system, with no specific tuning towards the domain.

The only action was to add some of the unknown words to the system dictionary. The 68 test sentences contained about 860 words, mainly very specialised database terminology. About 60 were not in the system dictionary. Of those, 20 were coded, using the system's coding tool. This was done to match the requirement that all words should be known (as it holds for the statistical MT).

Coding took less than 10 minutes as only 1:1 transfers were added to the dictionary. No further tuning was done.

Evaluation

The same evaluation measure was taken as for the statistical MT. The result can be given in the following table:

grammatical	30	44,1%
understandable	24	35,3%
wrong	14	20,6%

This result shows that the system is less strong in the middle category; either it finds a parse, and then produces good and grammatical results, or it fails. This fact shows that rule-based systems are less robust than alternative approaches.

However, the rule-based system produces significantly more grammatical results, and significantly better overall results (close to 80%) than the statistical MT system, under the same conditions (14 words sentences, no unknown words).

Improvements

Of course there is plenty of room to improve the translation quality of the rule-based system; mainly by tuning translation alternatives; this can easily be done, e.g. by assigning subject area codes to translations and choosing the right subject areas in translation. Recent studies (cf. Weber 2003) also underline a significant quality potential just using lexical measures. This was not done, however, as effects on the rest of the corpus could not be predicted, and it would have been an unfair tuning compared to the statistical package.

Also, recognition of named entities, proper names, product names etc. has been shown to improve the translation quality (Babych/Hartley 2003).

So there are significant tuning options just in the paradigm of rule-based systems; and there are customers which report error rates of only 3-4% for such systems.

2. Improvements

However, the question is not so much which approach is better; the more interesting question is what can be learned for the respective other approach, and how a hybrid system by which significant improvement in MT quality could be achieved should look like. To learn from the comparison, it is worthwhile to look at the translation results in more detail, and identify typical strengths and weaknesses of the respective approaches.

2.1 Statistical MT

This system basically works on chunks of input and assigns translations running a language model over the target words. Correlations of such chunks in source and target are learned, and used to translate the test corpus.

Quality

Translation quality is good if proper corresponding chunks can be identified in source and target language, like in (1)[3]; and fails if this is not the case, like in (29, 60). This counts for about 45% of the cases where translation quality is evaluated "wrong".

However, even if proper chunks are identified the translation fails in typical cases. Such failures can be described in linguistic terms, i.e. they can be generalised ("rule-based"). Typical failures are:

- German verb order and Satzklammer (split verbs) phenomena. Verbs in subordinate clauses must go from German last to English second position, and Satzklammer needs to be resolved.
  Here the system is not able to build a proper verb phrase (5, 27, 58), or drops one verb part altogether (31, 19).
- Constituent order: The system tends to keep the constituent order as in the source language (37, 68); cases where re-ordering is required (like in (63) where the German direct object is topicalised) tend to fail. Cf. also the wrong adverb placement in (57)
- Special constructions like German conditional clauses without subjunction.(47). The system translates plain indicative.
- Pronouns have several translations; the system tends to drop them altogether (22).

Such mis-handlings are systematic, they are responsible for about 55% of the 'wrong' evaluations, and it is hard to see how they could be overcome even if the training corpus could be extended significantly, because the "normal" material always outperforms the special cases.

Another systematic grammatical problem is to be mentioned, which is morphology. Statistical MT systems going from e.g. English into languages with richer morphology usually fail in assigning proper case information to their target output, in particular if the case indicates some functional relationship (like functional subject / object). This is less obvious in the current investigation as English does not use to many morphological markups.

On the lexical side, the statistical MT system performs quite well; so it is able to collect proper translation proposals from the training corpus. Sometimes wrong translations are given, however (4, 61, 64).

Usability

The crucial point is not that wrong lexical assignment can happen but that there is no possibility to control or influence the system behavior from a user's point of view. How can users add lexical items? How can they select a preferred translation in such a context? All this is crucial for a practical MT system.

[3] The numbers refer to the sentence numbers in the annex.

Another issue is domain-dependency. While statistical MT can be trained to a given domain with limited effort; this also means that it *has to be* trained to such domains every time anew. This is a never-ending task for a full-coverage MT system, and it is a severe problem in cases where no bilingual texts are available (which is nearly the majority of all cases). Even the best example-based systems (Richardson et al. 2001) have been tuned for one domain only (or one at a time).
From a practical and usability point of view, many questions remain to be solved before statistical MT systems can be considered to be operational.

2.2 Rule-based MT

These systems try to do a full parse on the input, and identify the basic syntactic functions in the sentence which are used for translation. Translation is done by looking up the words in the transfer dictionary and generating a proper word order and inflection.

Quality
The main sources of failure lie in the two main steps:

- **Parse failures** do not allow to identify the sentence parts; systems often use fall-back rules for those cases, but there will always be sentences which cannot be analysed properly. (cf. 25, 55)
- **Lexical failures** are the other main source of bad translations. This is not just that a word has no transfer entry in the dictionary; very often the problem is that there are *several* transfers in the dictionary and the system picks the wrong one.
  Examples are (10, 37, 57)
  In the tests mentioned above, two thirds of the "wrong" evaluation for the rule-based MT system are due to the problem of wrong lexical selection; so this seems to be more serious than the wrong-parse problem.
  A sub-section of this problem is translation of prepositions. They are notoriously difficult to translate, and there is much knowledge involved which is not rule-based but collocation-based; cf. (27, 56, 58).

In general, statistical MT performs better in these cases than rule-based MT. It is more robust than the fall-back strategies of rule-based systems, and it never picks translation readings which are outside of the domain (i.e. would simply not occur in a given corpus). Also, translation of prepositions contains less errors in statistical than in rule-based MT.

Usability
To select the right transfer from a set of options is a very difficult task, as current rule-based systems use systematic-linguistic features for disambiguation. They code in their transfer dictionaries under which conditions a term is transferred into a target term. Such conditions are mainly expressed in terms of features and values based on the conceptual model of underspecified morphosyntactic trees (good examples can be found in the OLIF (McCormick, 2001) and MILE (Calzolari et al, 2002) standardisation efforts for transfer entries). Examples are:

- Existence of certain **features** on the local node (e.g.: different transfers depending on gender),
- Existence of certain **syntactic functions** in a partial tree (e.g. different transfers of a verb depending on the presence of a direct object)
- Presence of certain surrounding **lexical material** (different transfer for adjective depending on the semantic type of the noun which it modifies; different transfer for nouns in compound specifier position vs. in head position)

and other such possibilities (more elaborate examples in (Thurmair 1990)).
Often however, either the text does not provide the required formal clues and neutralises readings, or the clues are more subtle to be detected by the current state of the art. Therefore it is not obvious how the selection process could be improved.
Of course, a rigid use of subject areas could prevent the system from picking out-of-area translations, but there are still sufficiently many cases of 1:n transfers left inside of such a subject area.

3. Conclusions

In the light of these discussions, the best way to proceed seems to be to create a hybrid system base a system on a rule-based architecture, and enrich it by features of statistical MT.

3.1 Rule-based backbone

The reasons to base it on a rule-based approach are the following:
1. It starts from a better quality baseline, and has already solved many of the usability and engineering problems which statistical MT still would have to overcome.
2. There are some ways how statistical MT can be improved:

- Preprocessing steps (better segmentation, morphological decomposition, name recognition etc.) definitely help to improve the MT quality by providing cleaner input to the statistical procedures.
- Replacing the (rather primitive) target language models by smarter linguistic-based generation components. Such components would use the lexical material produced by the statistical alignment, and try to 'make some sense' out of it, by putting them into the right constituent order and word formation. There have been related approaches in the paradigm of "shake and bake translation" in the early nineties (Whitelock 1992), however with limited success. But this approach would definitely improve results, and push some 'understandable' sentences into the 'grammatical' category.
- However, grammatical reference to the source sentence is still necessary, esp. in the area of grammatical functions (subject, object etc.). If this is not known, morphological case markings and/or word order cannot be stabilised. This kind of information requires significant linguistic analysis.

As a result, there are sources of knowledge which are indispensable for good MT, and it needs to be incorporated into a statistical backbone. A hybrid system based on such a statistical backbone is proposed in (Och et al. 2003), based on POS modeling, syntactic chunking probabilistic parsing and tree-tree alignment, with mixed quality results due to unreliable parses and the huge number of possible alternatives.

3. The main argument against rule-based MT is that it is costly to set up. There are three answers to this:

- There isn't such a thing as a free MT system. Building an MT system is always work.
- Cost is always relative, and is related to the savings which can be achieved, be it in productivity or in informativness. Examples show that investment into MT (esp. in the lexicon domain) pays off easily (Brundage 2001)
- Cost of a general-purpose MT system must not be compared to the cost of a special-purpose (one-domain) statistical system. Special purpose rule-based MT, with customised domain-specific dictionaries and grammars, can be set up in few months time. Cost for multi-domain general-purpose statistical MT is unknown as it does not exist.

For these reasons there is not really an alternative to a rule-based system backbone.

3.2 Statistical Enhancements

Assuming a decision in favor of a rule-based architecture, there are several ways how such systems could be improved by statistical means.

Robust Parsing

The idea is to improve rule-based parsing by statistical means. Instead of current approaches for probabilistic parsing only, the better strategy is to use probabilistic information to improve deep-linguistic analysis.

The side-effect of such a project would be to improve the analysis in robustness: In case of a parse failure, still the most probable analysis would be taken, just like in current statistical systems.

Transfer Selection

Instead of trying full statistical MT, the approach would be to find translation equivalents on word and phrase level for a given corpus / domain, and filter out all translation proposals which are not part of this corpus. After lexical transfer, standard target language generation components could be called.

This would reduce the hilarious results which MT is famous for, and leave only proposals which are valid for this domain.

Such an approach is promising also in cases of prepositions and other idiosyncratic translations, which make a good deal of the translation problems.

The challenge then would be to engineer such a solution: Create a special knowledge source for these cases, and have it interact with the current transfer components in a convincing way.

Productivity Tools

To increase productivity, statistical MT can be used as productivity tools in several respects:

- Pre-translation filter: Text analysis for the MT-translatability of a text. While most tools work on linguistic basis (Underwood/Jongejan 2001) (and repeat strengths and weaknesses of a rule-based MT system), a different technology may be better to detect such problems.
- Post-Translation filter: A statistical tool comparing MT output with 'standard' target text might help to locate problems which the MT system had: 'Strange' translations could be flagged, and postediting could focus on such segments first.

Dictionary work

Also in the preparation phase there are many options for statistical tools, mainly in the area to propose transfers from a given bilingual corpus. This is the intention of monolingual and bilingual terminology extraction tools (Thurmair 2003, Piperidis et al. 1997) which analyse corpus material to help to build linguistic resources.

Elaborate versions of such support users to create large bilingual linguistic dictionaries fast, and increase overall system productivity by shortening the coding phase. They assume, however, a rule-based type of MT system.

References

Babych, B., Hartley, A. (2003). Improving Machine Translation Quality with Automatic named Entity Recognition. Proc. EACL-EAMT, Budapest.

Brundage, J. (2001). Machine Translation – Evolution not Revolution. Proc. MT Summit VIII Santiago

Calzolari, N., Bertagna, F., Lenci, A., Monachini, M., ed., (2002). Standards and best practice for multilingual computational Lexicons and MILE (the Multilingual ISLE Lexical Entry). ISLE-Report 2002

Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics. Proc. AAAI.

McCormick, S. (2001). The structure and content of the body of an OLIF v.2 File. www.olif.net

Nießen, S., Ney, H. (2000): Improving SMT Quality with Morpho-syntactic analysis. Proc. COLING 2000

(NIST, 2001) Automatic Evaluation of machine Translation Quality Using N-gram Co-Occurrence Statistics. www.nist.gov/speech/tests/mt

Och, F., Gildea, D, Khudanpur, S., et al. (2003): Syntax for Statistical Machine Translation. J. Hopkins Summer Workshop. www.clsp.jhu.edu/ws03/groups/translate

Piperidis, St., Boutsis, S., Demiros, J. (1997). Automatic Translation Lexicon Generation from Multilingual texts.. Proc. AAAI 1997.

Richardson, St., Dolan, W., Menezes, A., Pinkham, J. (2001): Achieving Commercial-quality Translation with Example-based Methods. Proc. MT Summit VIII, Santiago

Thurmair, G. (1990). Complex lexical transfer in METAL. Proc. TMI 3, Austin, Tx.

Thurmair, G. (2000): TQPro, Quality Tools for the translation process. proc. ASLIB, London

Thurmair, G. (2003). Making Term Extraction Tools Usable. Proc EAMT-CLAW Dublin.

Underwood, N., Jongejan, B. (2001). Translatability Checker: A Tool to Help Decide Whether to Use MT. Proc. MT Summit VIII, Santiago.

Vogel, S., Och, F, Ney, H. (2000). The Statistical tramslation Module in the Verbmobil System. Proc. KONVENS Ilmenau.

Vogel, S., Och. F., Tillmann, Chr., Nießen, S., Sawaf, H., Ney, H. (2000). Statistical Methods for Machine Translation. In. Wahlster, W. (ed.): Verbmobil: Foundations of Speech-to-Speech Translation. Springer.

Weber, N. (2003): MÜ-Lexikografie. Proc. GLDV, Köthen

Whitelock, P. (1992) Shake-and-bake Translation. Proc. COLING Nantes.

Annex: Examples

The following list shows some examples of the output of the translation runs. Behind the sentence number there is

- the German source text (marked as “DE”)
- the output of the statistical machine translation (marked as “SMT”)
- the output of the rule-based translation (“RMT").

Important aspects or missing parts are underlined.

01 DE Mit dieser Methode ermitteln Sie den ersten Kindknoten eines bestimmten Knotens.
SMT use this method to find out the first child node of a given node .
RMT With this method you find the first child node of a certain node out.

04 DE Die folgende Abbildung zeigt die Benutzersicht auf das R/3-System:
SMT the following diagram shows the View Point to the R /3 System :
RMT The following illustration shows the user view on the R/3 system:

05 DE Es gibt zwei unterschiedliche Möglichkeiten, aus ABAP-Programmen auf Datenbanken zuzugreifen:
SMT there are two different ways from ABAP programs access databases:
RMT There are two different possibilities of accessing databases from ABAP programs:

10 DE Der zweite Teilbaum beschreibt die Struktur des zu lesenden Datenbestandes .
SMT to read data with data with data with the second sub – tree describes the structure of the .
RMT The second partial tree describes the structure of the data stock to be read.

19 DE Von ihrem Arbeitsplatz aus sollen Ihre Mitarbeiter auch vereinfachte Transaktionen ausführen können .
SMT from their workplace should your employee also simplified transactions . [__]
RMT Your employees also shall be able to execute simplified transactions from their workstation.

22 DE tragen Sie sich bitte als Pilot ein , und ordnen Sie sich einem Flug zu .
SMT enter this as a pilot , and assign a flight .
RMT Please put down your name as a pilot and assign yourself to a flight.

25 DE liefert Liste von Schlüsselfeldern der Objekte , die bestimmten Auswahl - Kriterien genügen [*incomplete*]
SMT returns a list of the key fields of the objects satisfying the selection criteria
RMT List of key fields of the objects which are enough for certain choice criteria delivers

27 DE Die Reihenfolge der Ausgabe dieser Textelemente wird dann über das Druckprogramm gesteuert .
SMT the sequence of the text elements output this is then using the print program controlled .
RMT The order of the version of these tokens is then controled about the print program.

29 DE Die geschützten Komponenten sind eine Schnittstelle zu den Unterklassen von C 1 .
SMT the protected components are an interface and of C 1 .
RMT The protected components are an interface to the subclasses of C 1.

31 DE Beim Anlegen einer Tabelle wird ein Schlüssel nach programmtechnischen Gesichtspunkten vergeben.
SMT when you create a table is a key after a a technical assign . [__]
RMT When creating a table a key is allocated after technical points of view.

37 DE Für die Verknüpfung mit Organisationsobjekten müssen Sie kein HR im Einsatz haben.
SMT for the link with organizational objects must use in HR you do not have .
RMT You don't have to have any HR for the bonding with organization objects in use.

47 DE Ist diese Berechtigung nicht vorhanden, so wird das Empfangs-programm abgebrochen
SMT this authorization is not available , the receiving program terminated
RMT If this right isn't existing, then the receiving program is broken off

55 DE Über die Job - Übersicht (Transaktion SM 37) erfahren Sie näheres .
SMT use the job overview (transaction SM 37) tell you for further information .
RMT Learn you about the job overview (transaction sm 37) near.

56 DE Der Datenbankmonitor zeigt zum Redo-Log-Puffer die folgenden Informationen an:
SMT the Database Monitor shows the redo log buffer the following information:
RMT The database monitor displays the following information to the redo log buffer:

57 DE Ein Job kann jedoch erst ausgeführt werden, wenn er freigegeben wurde.
SMT a job but can be executed if it was released .
RMT A job can, however, be executed only if it was unlocked.

58 DE Sie können die Tabellenpuffer über die Parameter im Profil der Instanz neu einstellen .
SMT you can the table buffers using the parameters in the instance profile set a new .
RMT You can adjust the table buffers newly about the parameters in the profile of the instance.

60 DE Die effizientesten Schreiboperationen werden am Checkpoint durchgeführt (“Chunk Writes”)
SMT the writes to the checkpoint to be performed (“chunk Writes”)
RMT The most efficient write operations are executed at the checkpoint (“Chunk Writes”)

61 DE Alle drei Workprozesse sind frei und die folgenden Jobs sind startfähig:
SMT all three work processes are empty and the following jobs are eligible to run:
RMT All three background processes are free and the following jobs are eligible to start:

63 DE Die Parameter der Datenbanksystemprüfung können Sie in der R/3- Tabelle DBCHECKORA konfigurieren.
SMT the parameters of the database system check [__] in R /3 configure table DBCHECKORA .
RMT The parameters of the database system test can configure you in the R/3 table DBCHECKORA.

64 DE Der Status der Aktionszeile im DBA – Einplanungskalender besitzt folgende Farbcodierung :
SMT the status of the Delete in the DBA Planning Calendar has the following color coding:
RMT The status of the action line in the DBA planning Calendar has the following color coding:

68 DE Um den Dialogmodus von saposcol zu verlassen, verwenden Sie den Befehl quit:
SMT dialog mode to the saposcol to leave , use the command quit:
RMT To exit the dialog mode of saposcol, you use the command quit: