In Rubio, A., N. Gallardo, R. Castro & A. Tejada (eds.) *Proceedings of the First International Conference on Language Resources and Evaluation, Granada, Spain 28-30 May 1998*, Vol. 1, 27-31.

Evaluating Text-type Suitability for Machine Translation a Case Study on an English-Danish MT System

Claus Povlsen, Nancy Underwood, Bradley Music, Anne Neville
Center for Sprogteknologi
DK-2300 Copenhagen, DENMARK
[{claus|nancy|music|anne}@cst.ku.dk]

Abstract

This paper reports on an evaluation of how well a specific MT system would perform in translating new text-types including an assessment of in what ways the system itself could be extended to deal with new text-types.
The overall evaluation and quality criterion is defined in terms of how much effort it takes to post-edit the text after having been translated by the MT system. A structured questionnaire rating different error types was given to the post-editors involved. The results were then "translated" into a number of mainly linguistic phenomena occurring in the input text causing these errors.
In order to achieve consistency and reliability the analysis of the new text-types was automated as far as possible. A suite of programs was developed, each of which identifies a particular phenomenon and assigns scores for each occurrence.
A reference text, known as being a good text, was first analysed using the procedure in order to provide a benchmark against which to assess the results from analysing the new text-types. After running the evaluation, a representative subset of the new text-types were then selected and translated by a slightly revised version of the MT system and assessed by the post-editors (using the same questionnaire).

Introduction

Over the last seven years CST has developed and maintained an English-to-Danish machine-translation system (PaTrans) under contract to the Danish translation company Lingtech A/S, which translates more than 3 million words a year using PaTrans (Bech 1997). The domain covered by the PaTrans system is petro-chemical and mechanical patent documents.
PaTrans is a production MT system directly descended from the Eurotra MT prototype (EUROTRA 1991). The commercialisation process included extensions for optimisation, syntactic error recovery, grammatical coverage of text-type-specific phenomena, conversion to/from WordPerfect, document handling (with preservation of layout information), pre- and post-editing tools, term-coding tools and lookup in multiple term dictionaries, i.a. (See relevant articles in (Bits & Bytes 1993)). Recent extensions include the integration of a part-of-speech tagger in pre-processing, and an experimental automatic post-editing tool.
At Lingtech, documents are first analysed for new terms, which are then coded in term dictionaries. Some mark-up is done manually in the pre-editing tool, such as identification of untranslatable text segments. A raw translation is then generated by PaTrans, which is post-edited by experts within the domain. This process compared to entirely manual translation saves Lingtech an estimated 60-75% of translation costs (Maegaard & Hansen 1995).
In 1996, Lingtech together with CST, initiated a project the purpose of which was to find out whether it would be profitable for Lingtech to extend the commercial area of machine-translation with new text-types and to find out whether it would be beneficial for Lingtech A/S to adapt the present PaTrans system to new text-types. This paper describes the overall methodology used concentrating on the methods and programs developed at CST for the analysis of source texts to ascertain their suitability for translation with PaTrans.

The Overall Method

The first step in the project was a market analysis instigated by Lingtech to assess the commercial potential for offering machine-translation services treating different text-types to new customers. Having, in this way, acquired knowledge of potential new customers and the text-types they would want translated, the next step in the investigation was to assess whether it would make economic sense to use PaTrans for these text-types. Whether it is worthwhile to translate a particular text using PaTrans depends on the quality of the translation produced. The higher the quality, the less post-editing is necessary and thus the more cost-effective it is to use the system. In order to find out drawbacks and advantages of the PaTrans system, a post-editor survey was carried out. Based on a questionnaire, the post-editors described their experiences with editing the machine-translation results of petro-chemical and mechanical patent documents. The next phase was to interpret these results from the post-editor survey into measurable criteria and assign the criteria scores in accordance with the replies given by the post-editors. The measurable criteria were then formalised into a set of specifications which formed the basis for (semi) automatic methods for measuring the suitability of new text-types to be translated by the PaTrans system.
In parallel with implementing the measuring methodology, work was carried out to collect representative corpora of text samples from the potential customers identified in the market analysis. Representative subsets of the collected corpora were then defined and (semi)automatically tested. In order to assess the scores for the new text-types, a benchmark in terms of a petro-chemical patent document was processed as well for the purposes of comparison. The PaTrans system was known to produce optimal translation quality in processing this document.

The final part of the investigation consisted of testing or verifying the automatic scoring of the new text-types. This was done in the following way: a subset of the tested corpora was defined and the lexical coverage (both terminology and general words) of the PaTrans system was extended to cover the texts in the subset. These texts were then translated by the revised PaTrans system and finally the translation results were evaluated via a new post-editor survey.

First Post-editor Survey

The aim of the survey, designed and carried out by Lingtech, (Post-editor Survey 1, 1996) was to obtain information from Lingtech's post-editors regarding their experience with editing the translation results of PaTrans. The information elicited from the survey was to be used to prioritise various improvements to the existing system as well as forming the basis for assessing the suitability of new text-types. 16 post-editors were asked to participate in the survey of whom 11 answered the questionnaire on which the survey was based. The questions were answered for both petro-chemical and mechanical patent documents.

The survey consisted of four parts. The first part of the survey was concerned with the general quality of the translations from PaTrans, the second part aimed to determine the characteristics of a translation that make it difficult or easy to post-edit, and the fourth dealt with tools facilitating the post-editing work. The third part of the survey which is the focus of this paper was concerned with the frequency and disturbing effect of 12 pre-defined error types. The post-editors were also asked to list frequent and disturbing error types themselves. The 12 errors types were selected by Lingtech on the basis of their experience with the system. The post-editors were asked to score the 12 error types with respect to their frequency and disturbing effect on a rating scale from 1 to 5. This was again done for both the mechanical and petro-chemical texts. The error types ranged from *totally incomprehensible and messy word order at sentence level* to *missing words*. The scores for frequency and disturbing effect for both text-types were combined to achieve a greater level of generalisation. The conclusion of this part of the survey was that incorrect word order caused the most problems for post-editing. Another serious error type was the wrong translation of prepositions.

From Error Types to Measurable Criteria

The error types identified in the post-editor survey are necessarily described in informal and non-technical language, which must be converted into criteria which can be measured. In order to use the translation error types in evaluating the suitability of a source text, the error types had to be translated into phenomena occurring in the source text which give rise to these translation problems. This was done by determining for each error type which observable phenomena in a text could cause this error to occur. The conversion from error types to phenomena in a text was based on the CST's expert knowledge as developers of the PaTrans system. The following list shows a selection of the error types and some of the phenomena which have been identified as causing them.

A. *totally incomprehensible and messy word order at sentence level*
  A1 fronted adverbial subordinate clauses and prepositional phrases
  A2 compound elisions with a hyphen
  A3 translation units containing numbered lists
B. *incorrect translations of general words*
  B1 homographs (with different target translations)
  B2 deverbal nouns and adjectives
  B3 tokens split by a hyphen
C. *missing words*
  C1 valency bound prepositions
D. *wrong prepositions*
  D1 non-valency bound prepositions
E. *incorrect placement of words*
  E1 words which translate into nexus adverbs

Thus if *totally incomprehensible and messy word order at sentence level* occurs in the translation then it is an indication that the sentence in the source text has contained either fronted adverbial subordinate clauses and prepositional phrases or compound elisions with a hyphen or translation units containing numbered lists. Similarly, if *incorrect translations of general words* occurs then it is a sign that a sentence has contained a homograph, a deverbal noun or adjective or tokens split by a hyphen. The list is far from exhaustive, and is only meant to illustrate the step from translation error types to phenomena detectable in the source text. In the rest of the description of the evaluation we will concentrate on the examples given here.

Based on the post-editors' scores, the error types were divided into three groups, ranging from seriously disturbing error types to less seriously disturbing error types. The three groups were scored on a scale from 1 to 3, where 1 indicates the least disturbing and 3 indicates the most seriously disturbing. These scores were used in the following way. Each occurrence of phenomenon X that caused error type Y triggered the score associated with error type Y.

Specification of Methods for Measuring Criteria

Having interpreted the post-editors' error types as identifiable phenomena in a given text, the next step was to specify the methods by which each of these criteria could be measured. At the most abstract level each occurrence of a given phenomenon in the text is identified and the relevant score is recorded. In order to achieve consistency in the analyses of texts and in line with other work in evaluation (EAGLES 1996; Paggio & Underwood, *in press*) our aim has been to automate the process as much as possible. However, as will be shown below, some of the phenomena must be identified by hand. Thus a suite of programs which also allow for human input has been developed.

The different phenomena can be broadly classified into three types: *layout issues*, *lexical items*, and *syntactic constructions*. Such a typology proved useful in specifying the methods for identifying different phenomena, since it also reflects the different knowledge resources which programs must access. It is interesting to note that this typology cuts right across the translation error typology.

Layout issues

Three phenomena fall under this type: B3 *(tokens split by a hyphen*); A2 (*compound elisions with a hyphen*); and A3 (*translation units containing numbered lists*). Measuring the first two phenomena is arguably the simplest to carry out, since it is based on identifying layout features in the original input file. To identify numbered lists occurring within a translation unit the relevant program operates on a version of the original text which has been segmented into PaTrans translation units. Thus these measures do not rely on any linguistic knowledge resources.

Lexical items

A number of phenomena are concerned with individual lexical items. Here the lexical resources of PaTrans itself play a vital role. Identifying B2 *(deverbal nouns and adjectives*) relies on information in the English lexical databases. For example *deverbal nouns* cause problems when a lexical item is ambiguous with respect to its part of speech (e.g. "report") and the PaTrans lexical lookup procedure must be run on the text and then all lexical items which have been assigned both the categories verb and noun are identified as deverbal nouns.

Identifying B1 (*homographs (with different target translations)*) and E1 (*words which translate into nexus adverbs*) on the other hand relies on the bilingual English-Danish lexicon. For example, Danish nexus adverbs produce word order problems in the Danish translation. Therefore it is necessary to identify which words in the text will be translated into nexus adverbs by looking them up in the bilingual lexicon.

Syntactic constructions

Other phenomena involve syntactic constructions of various kinds and therefore require the use of more complex linguistic knowledge and analyses than that needed for individual lexical items. The methods for identifying different phenomena also vary in their complexity and the resources they must access.

For example, to identify C1 (*valency bound prepositions*) and D1 (*non-valency bound prepositions*) the relevant programs must make use of subcategorisation information available in the lexical databases in order to analyse and mark-up which prepositional phrases in the text are valency bound. Thus these phenomena are tightly bound to the specific analysis of subcategorisation in PaTrans, and can be identified reasonably straightforwardly using heuristic rules.

In order to identify more complex constructions we made use of a shallow analysis provided by the commercial constraint-grammar parser ENGCG, (Voutilainen *et al*, 1992) which provides part-of-speech disambiguation augmented with functional information, e.g. identification of the main verb.

The phenomenon A1 (*fronted adverbial subordinate clauses and prepositional phrases*) is probably one of the most complex to measure. It encompasses a number of different construction types and to facilitate their identification it has been divided up into the following five separate phenomena ,

A1.1 Fronted prepositional phrases
e.g. "In the next few weeks we will see an upturn in the economy"

A1.2 Fronted subordinate clauses
e.g. "If the compressor is not used the procedure will take three hours"

A.1.3 Fronted past participle clauses
e.g. "Given the current situation, we advise against a merger"

A.1.4 Fronted subordinate clauses within a subordinate clause
e.g. "It is important to remember that whenever the process is running all other windows become inactive".

A.1.5 Fronted infinitives or present participles
e.g. "To restart the process depress the control key"

For the first four of these phenomena the heuristic rules for identifying these proved satisfactory. However in the case of A.1.5 (*fronted infinitives or present participles*) it was not possible to automatically distinguish reliably between cases where for example a present participle functions as an adverbial and where it is a subject

e.g. "Turning to the next paragraph we see..."
"Centrifuging the compound produces a white residue".

So for this phenomenon (as for a number of others), human judgement must also be involved.

Implementation

A program was developed ("paca" for "PaTrans corpus analysis") comprising a suite of modules counting occurrences of the phenomena and calculating an overall score for the text. The platform used was an HP workstation running HP-UX. Standard unix tools were utilised for the most part, including Bourne shell scripts, awk scripts and a make file. The result of applying ENGCG was used as input for analysing several of the syntactic phenomena. Text analysis is initiated by running paca from the unix command-line with a filename argument. The file must be in the PaTrans document input format (SGML-based). The output is a list of phenomena with their scores. In addition, a number of files are created containing occurrences of certain phenomena that have to be double checked manually (see below). The rest of this section describes some of the different programs for the different types of phenomena, and implementational considerations associated with them.

In the identification of layout phenomena, it was important to distinguish between tokens split with a hyphen and compound elision with hyphen, as these received different scores. A line ending with a token with a final hyphen wasn't enough to qualify as a split token, since it could be the initial member of a compound with elision occurring over a line break, e.g. "... is done by single- <linebreak> or double-clicking...". The following line thus had to be checked for coordinating conjunctions.

The individual lexical phenomena were even easier to identify. Deverbal nouns for instance were simply entries which could be either verbs or nouns. Nexus adverbs

were identified by comparing input tokens with a list of English words which can be translated into Danish nexus adverbs.

The syntactic (or collocational) phenomena presented more of a challenge, and in some cases could only be tallied using heuristics. Fronted adverbials is one particularly broad class of phenomenon to be identified, and was broken down into five sub-types, as mentioned above. The first three, viz. fronted PPs, subordinate clauses and past participles, are easily identified by checking the initial element of each sentence in the (disambiguated) tagger output.

Fronted elements within subordinate clauses were found by checking the tagger output for comma-separated elements immediately following an appropriate coordinating or subordinating conjunction (e.g. "because", "however" and non-demonstrative "that"). Again, these are heuristics whose only function is to identify the presence of the phenomena, not to give them any kind of analysis, nor even to delimit them. Nonetheless, their performance for this evaluation was satisfactory.

Fronted infinitives and present participles were more difficult, since, as noted, these constructions sentence-initially can also be the subject of the main verb and not a fronted adverbial. The heuristic used here is to count sentences with a gerund or an infinitive marker as the first element with a comma before the main verb. Obviously for this type of analysis, relying on identification of infinitive markers and main verbs, the input must first be disambiguated by the part-of-speech tagger.

A rating based on a part-of-speech tagger and punctuation is somewhat unreliable. The strategy chosen here (and with respect to other phenomena) was to identify the most difficult phenomena in a relatively unconstrained manner, then explicitly tell the evaluator which phenomena within which sentences to double-check. A file is generated for each phenomenon to be double-checked, based on which a list of score adjustments is then created manually. For each phenomenon, the automatically generated score is then toted with its score adjustment to arrive at a final adjusted list of scores for the input text, from which a total document score and an average per word is derived.

Running the Evaluation

A test corpus collected by Lingtech was analysed using paca. The test corpus consists of three text-type-specific corpora each representing sample texts from text-types Lingtech had identified as being potentially commercially viable. The size of the total corpus is about 210 pages (56390 tokens). The overall scores for each of the three test corpora were 0.63, 0.68 and 0.64. These figures themselves do not say anything about their suitability to be translated by the PaTrans system. Since an average means expressing the threshold of a good/bad text-type to be translated by the PaTrans system does not exist and is not to be theoretically deduced, it was decided to run a petro-chemical patent document through paca as a benchmark or reference text with which to compare the results for the new text-types. This text, upon which the linguistic coverage of the overall PaTrans system to a large extent has been based, represents a text which the PaTrans system translates very well. The overall score for this reference text was 0.76. Bearing in mind that the higher the score the less suitable a text is, the three text-types represented by the three test corpora at first glance seemed to be better suited to the PaTrans system than the reference text.

In order to get a more precise picture of the overall results, the scores for each individual phenomenon were examined separately. On the whole, the scores for each phenomenon in the new text-types and the benchmark were very similar. For example the scores for fronted adverbial subordinate clauses and prepositional phrases leading to *totally incomprehensible and messy word order at sentence level* yielded approximately the same results, i.e. the same number of hits per token. In other cases, differences in the average occurrence of the various phenomena could be observed between the new text-types and the reference text. Especially the occurrence of words which would translate into nexus adverbs deviated radically. The occurrences of this group of adverbs in the new text-types were from 2,5 to 5 times as frequent as in the reference text. Another example is deverbal nouns which in new text-types occurred, on average, approximately half as frequently again as in the reference text.

Results from the Second Post-editor Survey

In order to survey post-editors' reactions to the new text-types and compare these with the results of the paca analysis, a subset of the test corpora (hereafter sub-corpus) was defined and translated by the PaTrans system. In order to ensure that the various collections of texts in the sub-corpus did represent the text-type domain, the selections were based on the results from the evaluation of the new text-types. This was done in the following way: first a selection was done manually based primarily on quantitative criteria then the reduced text corpora were run through the evaluation to examine whether the ratio per token of phenomena found reflected the ratio in the three test corpora. This was done iteratively until an approximately similar occurrence of phenomena was achieved.

In order that the post-editors could reasonably judge the quality of the new text-types, the lexical coverage of PaTrans was extended. Lingtech provided a list of uncovered general words that they had collected translating the sub-corpus. This was done with a pre-and post-editing tool one of whose facilities is the generation of a list including the general words and terms that the system did not recognise during translation. General words were then coded at CST and Lingtech coded the unknown terms themselves using the term-coding tool developed. As unknown words are of critical importance to the translation quality, this exercise was deemed necessary in order to allow a proper comparison with the post-editors' normal experience with PaTrans output, where unknown general words and terms are coded before the translation is run.

The translations produced were then given to the post-editors in order make a new survey of the output from the revised PaTrans system. Letting the post-editors evaluate the translations of the text corpus selected can be said to be a test of the semi(automatic) method described above. The crucial question will of course be: will the post-editors consider the quality of the translations of the new text-types to be as good as for the reference text?

According to the second post-editor survey (Post-editor Survey 2, 1997) the post-editors did not arrive at the same conclusions as the automatic evaluation did. In general, they assessed the quality of the translations of the test corpora to be somewhat worse compared to the reference text. Especially the errors categorised as *totally incomprehensible and messy word order at sentence level* were emphasized as being more resource demanding in the new text-types compared to the reference text. This discrepancy was found to be due to various factors.

One influential factor is the difference in language usage between the new text-types and the patent documents. The post-editors being used to edit very technical and precise patent documents may find it difficult to edit translations of the new text-types, which are characterised by having a wide target group and thus being written in a much more informal style.

The most important factor, however, is the following. The list of phenomena is based on the post-editors' assessment of how well the PaTrans system translates patent documents and that the design of the PaTrans system is tuned to analyse petrochemical patent documents. The consequence of these facts is that linguistic phenomena in the new types which lie beyond the linguistic coverage of the PaTrans system are not accounted for. Interrogatives and imperative forms of verbs, for instance, do not occur in patent documents and are therefore not included in the linguistic coverage. This means that the PaTrans system performed badly in cases where imperatives and interrogatives occurred in the sub-corpus, leading to poor translation quality.

Conclusions and Future Work

This paper has described the development and performance of a concrete evaluation of the suitability of certain text-types for translation by a specific MT system. In addition our remit was to assess in what ways the linguistic coverage of the system should be extended to deal with new text-types. The discrepancy between the evaluation results and the post-editors' feedback on the translations of the new text-types can to a large extent be explained by coverage gaps in the current system.

It is thus clear that the evaluation of new text-types cannot rely solely on criteria developed for assessing the translation quality of an existing text-type. The evaluation described above should therefore be considered as being a first stage in an iterative process, in which the suite of programs is extended to account for the newly identified gaps in coverage and the evaluation of the text-type carried out again. An alternative would be to implement the linguistic coverage gaps found in the PaTrans system and then make a third post-editor survey in order to test the evaluation results.

The implications of the evaluation and its results are two-fold. Despite the fact that this evaluation was devoted to a very specific context, the lessons learned and the general approach, taking into account properties of the system and texts as well as user requirements, contains a number of elements re-usable in similar evaluation tasks. For example in the Transrouter project in which CST is involved, a similar methodology for analysing properties of texts will be used to route texts to the most suitable translation solutions. In addition, based on the evaluation results Lingtech has already ordered various changes to the PaTrans system in order to improve the translation quality. Initially the changes focus on improvements that triggered the errors categorised as *totally incomprehensible and messy word order at sentence level*. Furthermore it is foreseen that Lingtech later on will order improvements to the MT system that will bridge the coverage gaps identified during the evaluation.

Acknowledgements

The authors would like to thank the Lingtech staff and especially its managing director, Annelise Bech, for a very fruitful collaboration and permission to use the results from the two post-editor surveys in this paper.

.

References

Bech, A. (1997). MT from an Everyday User's Point of View. In *Proceedings of MT Summit VI* (pp. 98--105) San Diego.

Bits & Bytes (1993). *Datalingvistisk Forenings årsmøde*. ISSN 0109-5501. Odense: Institut for sprog og kommunikation.

EAGLES (1996). *EAGLES Evaluation of natural language processing systems*. Final Report. ISBN 87-90708-00-8. Copenhagen: CST.

EUROTRA (1991). Copeland, C., Durand, J., Krauwer, S. & Maegaard, B. (Eds.), *Studies in Machine Translation and Natural Language Processing*, Vols. 1 and 2. Luxembourg: CEC.

Maegaard, B. & Hansen, V. (1995). PaTrans: Machine Translation of Patent Texts, from Research to Practical Applications. In *Engineering Proceedings of the Second Language Convention* (pp.1--8).

Paggio P. & Underwood, N.L. (in press). Validating the TEMAA LE evaluation methodology: a case study on Danish spelling checkers. *Journal of Natural Language Engineering*, 4(3). Cambridge University Press.

Post-editor Survey 1 (1996). *Analyse og sammenfatning af sprogrevisorsvar. Lingtechs spørgeskemaundersøgelse 1996.*

Post-editor Survey 2 (1997). *Evaluering af oversættelser af tekstkorpusser K; edb og kontormaskiner. Baseret på input fra Lingtechs sprogrevisorer og datalingvister.*

Voutilainen, A., Heikkilä, J. & Anttila, A. (1992). *Constraint Grammar of English. A Performance-Oriented Introduction*. University of Helsinki: Department of General Linguistics.