Translatability Checker:

A Tool to Help Decide Whether to Use MT

Nancy L. Underwood, Bart Jongejan

Center for Sprogteknologi
Njalsgade 80, DK-2300 Copenhagen S
Denmark
nancy@cst.ku.dk, bart@cst.ku.dk

Abstract

This paper describes a tool designed to assess the machine translatability of English source texts by assigning a translatability index to both individual sentences and the text as a whole. The tool is designed to be both stand-alone and integratable into a suite of other tools which together help to improve the quality of professional translation in the preparatory phase of the translation workflow. Assessing translatability is an important element in ensuring the most efficient and cost effective use of current translation technology, and the tool must be able to quickly determine the translatability of a text without itself using too many resources. It is therefore based on rather simple tagging and pattern matching technologies which bring with them a certain level of indeterminacy. This potential disadvantage can, however, be offset by the fact that an annotated version of the text is simultaneously produced to allow the user to interpret the results of the checker.

Keywords

Translatability, MT

Introduction

Although the use of MT in busy translation departments and agencies can lead to significant benefits in terms of improved productivity, increased profits and worker satisfaction (see e.g. Bech, 1997), machine translation quality can be, to say the least, variable and dependent upon both the text to be translated and the MT system used. Using an MT system to translate certain texts can result in such bad translations that it becomes more time-consuming and costly to post-edit the raw output than to translate the text by more conventional means. So, in order to avoid such wastage it is important for a translation manager to be able to identify those texts which are most suitable for MT.

This paper presents a tool designed to assess the machine-translatability of English source texts which has been developed under the auspices of the TQPro (Translation Quality for Professionals) project[1] (Thurmair, 2000). The current version of the tool is stand-alone, and called from the command line, however it is designed ultimately also to be integrated into the overall TQPro architecture as part of a suite of preparatory tools in the translation workflow.

Thus the intention behind the Translatability Checker is not to invoke the full machinery of an MT system (such as its parser, and lexicon or terminology look-up) but rather to carry out a rapid and somewhat shallow analysis of the source text. As will be seen this approach involves a constant trade-off between speed and robustness on the one hand and accuracy on the other. Such a trade-off, though, is mitigated by the fact that the tool outputs not only a numeric score for the translatability of the text, the "Translatability Index" (TI), but also an annotated version of the text displaying the analysis of each sentence to enable the user to interpret and evaluate the TI.

The notion of translatability is based on so-called "translatability indicators" where the occurrence of such an indicator in the text is considered to have a negative effect on the quality of machine translation. The fewer translatability indicators, the better suited the text is to translation using MT. Once such indicators have been identified in the source text a set of weights associated with the indicators is used to calculate the translatability indices. Thus the tool performs 3 basic steps: Analysing the text to identify the translatability indicators - Calculating the translatability indices on the basis of these indicators - Generating the report and annotated text. These will be described in the following sections.

Identifying Translatability Indicators

In this implementation, there are two sets of translatability indicators. Firstly a set of phenomena identified in others' work on translatability as causing problems for MT systems (see e.g. Bernth & Gdaniec, 2000 and Gdaniec, 1994) have been defined as potential translatability indicators for the use of MT in general. These cover a) structural ambiguity caused by: PP-attachment; relative and other sub-clause attachment and multiple coordination b) compounds comprising 3 or more nouns, c) "sentences" without (finite) verbs, d) lexical ambiguity and e) sentence length (both very long and very short sentences). However, it is important to be able to assess translatability wrt specific MT systems and text-types and a set of indicators for a specific English-Danish MT system, PaTrans, (Povlsen et al. 1998*)* has also been developed. A number of the general indicators of course also apply to the PaTrans system, but in addition to these, sentence initial PPs, adverbs and subclauses often cause serious word-order problems in the Danish translation, as do long sentences which contain adverbs. On the other hand, PPs headed by "of" are much less problematic, and due to the

[1] The project is supported by the European Commission in the Fifth Framework Programme.

pre-editing/term identification phase in the use of PaTrans, nominal compounds do not pose a problem.

Given the requirement that the tool must be fast and robust, the linguistic analysis is restricted to POS tagging using a well established generally available POS tagger for syntactic phenomena and word lists for lexically ambiguous words.

POS-based analysis

The stand-alone version of the checker operates on flat text files which are first tokenised and segmented into sentences before being tagged by the tagger.

Tagging the text

The text is tagged using the Brill Tagger[2] (Brill, 1994) taking as a starting point the data (lexicon and rules) distributed with the software and based on the Brown Corpus of American English from 1961 (Francis & Kucera, 1979). It is clear that the quality of the final analysis depends in large part on the quality of the output of the tagger. In order to achieve the most satisfactory results from the tagger it is necessary to tune it to specific text types both in terms of their subject domain and the linguistic constructions they contain. However, notwithstanding the age of the data on which the tagger was trained, very good results were obtained after some fine-tuning of the program and data, for example to deal with capitalised words in headings: so that words in a heading that are not found in the dictionary but start with an upper case letter are looked up once again after being converted to lower case. Results of around 97% accuracy were achieved (see below). It would be expected that training the tagger on modern texts could increase the accuracy further and by training and/or fine-tuning the tagger on different text-types, in future versions of the tool it would be possible to activate different lexica and rule sets depending on the text-type specified.

Pattern matching

In order to identify potential translatability indicators pattern matching rules are applied to the tagged text. The rules are generally simple and rely on determining the presence or absence of a particular tag (or number of instances of that tag) in a sentence and the length of the sentence. For example, the presence of more than one coordinating conjunction (*and, or*) reflects the potential for structural ambiguity due to complex coordination. If the first word in a sentence is a gerund then this would indicate a sentence initial subordinate clause. The absence of a verb tag reflects the fact that the "sentence" does not include a verb and is thus probably a heading or list item, which can also lead to translation problems. In addition to the more straightforward verbless sentences, there is a pattern matching rule to identify "sentences" which do not include finite verbs, which is also a characteristic of headings and list items.

The quality of the final analysis produced by the checker is not only dependent upon the accuracy of the tagger itself but also on the usefulness of the tags applied and this of course affects the definition of the pattern matching rules. The tag set is the same as that used in the Penn Tree Bank Project (Santorini, 1990) and does not distinguish between prepositions (e.g. "in", "over", "for", "under") and subordinating conjunctions (e.g. "if", "whether", "that", "whereas") and in this version we have implemented the simplest analysis and do not distinguish between PPs and subclauses either. However, for specific MT systems it may make sense to distinguish between prepositions proper and subordinating conjunctions in order to be able to weight their occurrences differently. For example over and above the problems caused by attachment ambiguity, prepositions are notoriously difficult to translate. It would be relatively simple to augment the current pattern matching rules with a list of subordinating conjunctions.

The POS-based analysis comprises rules to identify the following phenomena on a sentence by sentence basis:

General indicators

- No verb present
- No finite verb present
- Multiple coordination
- Long sentence (>25 words)
- Short sentence (< 3 words)
- One or more nominal compounds (>2 nouns)
- PPs and/or subclauses

System-specific indicators

- Sentence over 25 words with at least one adverb
- Sentence-initial adverbs or subclauses
- Sentence-initial PPs and/or subclauses
- Non-sentence-initial PPs and/or subclauses
- PP headed by "of"

Specifying the MT system in question will have the effect of ensuring that only the relevant rules are activated. Thus specifying PaTrans will have the effect that the system-specific rules and the general rules for the first five indicators will be activated.

Word-based analysis

In English there is widespread ambiguity between verbs and nouns (e.g. *report, set, order, record)* and between verbs and adjectives (e.g. *correct, appropriate*) whilst many words can function as all three word classes (e.g. *light, cross, present, split*). Such lexical ambiguities are identified by means of word lists classifying the different types of homographs. Whilst for MT in general these have been derived from a large general lexicon for a specific MT system such lists are compiled from the system's own lexica. The translatability indicators identified via word-based analysis are:

- noun-verb homographs
- adjective-verb homographs
- adjective-noun-verb homographs

Translatability Indicator Values

From the above it might seem that the identification of a translatability indicator would result in a value of 1 or 0 depending on the presence or absence of a particular indicator in a sentence. However, it is not always sufficient to simply count the presence or absence of an

[2] The Brill tagger is available from http://www.cs.jhu.edu/~brill/

indicator since many of them can occur several times in one sentence and the more occurrences, the worse will be the predicted outcome of the translation. For example, the presence of six conjunctions in a sentence would be more detrimental to translation than only two conjunctions and this must be reflected.

If there are m_{ik} occurrences of an indicator i in sentence k, then the fractional indicator value I_{ik} is computed as follows:

$$I_{ik} = \frac{m_{ik}}{1 + m_{ik}}$$

In the example above, six conjunctions give a value of 6/7 (or 0.857) and two conjunctions give a value of 2/3 (0.667). Even a few instances result in very substantial values. One can reduce the result for only a few instances while at the same time ensuring a higher penalty for many instances, for example by replacing the term '1' in the denominator by a higher number. A number of 10, for example, would give values 6/16 (0.375) and 1/11 (0.091).

Thus the indicator value for such translatability indicators present in a sentence will be a number between 0 and 1, as opposed to those indicators which by definition can only occur once in a sentence (e.g., "No verb present" or "Long Sentence"), whose value will be 1. It is these indicator values which are used in calculating the translatability index.

Calculating Translatability

Once all the translatability indicators have been identified and their values calculated where this is relevant, the weightings defined for each indicator are applied and the TI for each sentence and the text as a whole is calculated.

Weighting the translatability indicators

For each translatability indicator a weight is defined, to indicate the relative effect of the indicator on the translation process. Weights are defined in a separate file which in this stand-alone version can be directly accessed and amended according to the relative importance of indicator. Weights can have values between 0 and 100, but their sum must not be greater than 100. Setting a weight at 0 means that the indicator is not considered relevant and thus has the effect of disabling the pattern matching rule for that indicator. In the first instance, the weights were assigned more or less intuitively based on the perceived relative importance of each indicator i.e. how badly a phenomenon affects the translation of a text. They are then adjusted after experimentation. A different weight file will be maintained for each MT system.

Calculating the Translatability Index

The TI for a sentence is computed as the weighted sum of all the values of the translatability indicators for that sentence. The TI for the text as a whole is the average of the TIs of all the sentences. The translatability index is a number between 0 and 100. A translatability index of 100 would mean that no translatability indicators are present that would have an adverse effect on the quality of the translation and thus the text is extremely well-suited for MT.

Generating Results

Since the translatability checker is designed to be integrated into the TQPro architecture which uses the Lotus DTO (Domino Translation) Environment and is web-based, providing a front-end interface to the web-server which hosts the TQPro toolbox, the results of the analysis are output as two .html files: a report file and the annotated text file.

The Report File

The report file contains both the overall translatability index of the text and the total number of instances of each translatability indicator found in the text. This is given in the form of a table with two columns in which the first column indicates how many sentences in the input text have been found to contain the translatability indicator mentioned in the second column. After the table, the overall translatability index of the text is given as a whole number between 0 and 100.

#	TI		annotations
3	81	However, in practice, this is not the way that such containers are filled.	initial subclause/adverb prep/subord. conj noun-verb homograph
5	95	The two halves are then welded together enclosing the normal atmospheric gases inside the secondary chamber.	prep/subord. conj
16	56	naturally the ratio of sizes **or** numbers of vents 5 **and** 11 are arranged to provide generation of the required amount of foam in the shell 10 as the stay-on tab 3 is opened.	**conjunction** initial subclause/adverb prep/subord. conj long & 1 adverbs (34 words) adj-verb homograph
76	89	Claims 1:	missing verb (2 words)

Figure 1: An Annotated Text File

Annotated text

Given the nature of the translatability checker, relying as it does on a shallow analysis, the analyser/report generator also creates an annotated version of the text, which allows users to interpret the translatability indices calculated. Figure 1 shows an example of an extract from an annotated text file using the system-specific analysis rules.

The annotated text file consists of a table with four columns. The first column contains the sentence number. The second column contains the translatability index of the sentence. The third column contains the sentence itself. Phenomena that can be localised somewhere in the sentence are highlighted. The fourth column lists all found phenomena, using the same highlighting as used in the sentence, where appropriate. Highlighting in the text is as follows: prepositions and subordinating conjunctions are underlined, compounds of 3 or more nouns are in italics, conjunctions (if there is more than 1 present) are in bold font, whilst the different types of homograph are in different coloured fonts (shown as shading in this figure). In the case of long sentences containing adverbs the adverbs are also highlighted in the text. Sentence-initial phenomena can easily be located and are not highlighted. Other phenomena which cannot be localised nevertheless generate annotations: For very short (under 3 words) or long sentences (over 25 words) the number of words is given in parentheses and if a sentence lacks a verb or finite verb, this is also noted in the annotations column.

Evaluation

Evaluation of the translatability checker encompasses both system internal and external evaluation. Internal evaluation has focussed on the tagger and the pattern matching analysis rules. Having chosen the Brill tagger the first job was to analyse its output both in terms of its accuracy and the types of tags used in order to build the pattern matching rules on top of these. Evaluation of analysis rules for identifying translatability indicators concerns not only how well the rules function according to their specifications but, more importantly, how well they identify the textual phenomena which have been defined as translatability indicators. External evaluation, on the other hand, concerns how well the TIs actually predict the translatability of text and the general utility of the checker in deciding whether to translate a text using a specific MT system or not.

Evaluating the Tagger Output

The first task was to evaluate the tagger as is, by running the tagger over a number of texts and checking by hand the tag assignments. After this first pass and the modifications to the tagger described above, the tagger was run again with improved results.

Two different corpora (representing different text-types) were analysed in detail: a corpus of 8 different patent application texts and a corpus derived from an on-line manual for commercial computer programs. Table 1 shows the results for the two text types. Figures in column A represent the results before the modifications to the tagger whilst column B contains the results obtained afterwards.

Test Corpus	% correctly tagged words	
	A	B
Computer Manual (6002 words)	93.52	96.45
Patent Applications (6139 words)	96.11	97.28

Table 1: Tagger Accuracy

When a word in a text does not appear in the tagger's lexicon, the tagger "guesses" its part of speech, and one can imagine that training the tagger on specific text types would improve its accuracy. However, due to the inherent imprecision of tagging, perfect performance can never be expected.

It is not surprising, given the age and nature of the tagger's data (based on a large general language corpus) that it performed better on the more discursive patent texts than on the computer manuals which are characterised by commands, headings and lists. It is not always easy, however, to judge the correctness of an assigned tag when looking at a sentence in isolation. For example in the string "Set domain name", "set" can either be a noun or a verb. In such cases, the larger context of the text was taken into account in order to try and determine what the correct tag would be. Also, some patent texts were found to contain both lexical and grammatical errors which made it difficult to judge the accuracy of the tagging.

Evaluating the Analysis Rules

To evaluate the accuracy of the analysis rules in detecting the translatability indicators in a text, all instances of such indicators in the test corpus were first identified and then compared with the output of the analysis rule module. Each of the analysis rules were evaluated in terms of both precision and recall of the translatability indicators. That is to say: how many of the actual instances of an indicator were correctly identified? (recall) and of those identified by the checker, how many were actually correct? (precision).

Table 2 shows the recall and precision results obtained for the general translatability indicators:

Translatability Indicator	recall (%)	precision (%)
Missing (finite) verb	90.47	95.45
Sentence length	100.00	100.00
Multiple coordination	100.00	100.00
PPs/subclauses	93.56	99.77
Homographs	100.00	100.00
Nominal compounds	78.12	91.46
Overall Scores	93.69	97.78

Table 2: Recall and Precision: General Translatability Indicators

The figures above are somewhat skewed, in that the number of actual instances of each indicator in the test corpus varied. So that whilst there were 1103 instances of PPs and subclauses there were only 96 instances of nominal compounds and 77 verbless sentences. The most straightforward of the analysis rules (for calculating sentence length, multiple coordination and homographs), not surprisingly, performed with perfect recall and precision. In the case of both nominal compounds and missing (finite) verbs the main stumbling block for the rules was due to the widespread verb-noun ambiguity in English. So, for example, the compound "data extract period" was not recognised because "extract" was tagged as a verb, whilst the clause "change master record" was incorrectly identified as a compound.

More interesting are the results for prepositional phrases and subclauses. The analysis rules for identifying these are (like the others) based on identifying tags assigned and thus generally they are unable to identify reduced relative clauses (e.g. "oxygen remaining in the sealed container"). Such clauses accounted for around 4% of those not correctly identified. In addition a number of prepositions and subordinating conjunctions are ambiguous in that they can also function as determiners or pronouns (e.g. "that", "which") and this also affected precision and recall.

The results obtained for the system-specific translatability indicators are shown in Table 3

Translatability Indicator	**recall (%)**	**precision (%)**
Long sentence with adverbs	100.00	100.00
(identification of adverbs)	100.00	99.14
Sentence initial adverbs/clauses	100.00	100.00
Sentence initial PPs/clauses	100.00	100.00
Non-initial PPs/clauses	92.61	99.78
"of"-PPs	100.00	100.00
Overall Scores	98.77	99.82

Table3: Recall and Precision: System-Specific Translatability Indicators

The results for the system specific indicators are generally better than the general results and this is in large part due to the fact that sentence initial PPs and subclauses and "of" are easier to identify. The results in the second row, (identification of adverbs), indicate that although the checker correctly identified all and only the sentences of more than 25 words containing at least one adverb, in two cases specific words were incorrectly marked as being adverbs.

The results for both the tagger and the analysis rules appear to be remarkably good. However, this may in part be due to the fact that the tokenisation and segmentation into sentences has been fine-tuned to each of the specific text types, in their flat text versions. In the integrated version, the tool will operate on texts marked up in the latest version of OTEXT[3]. It remains to be seen what, if any, further errors may be generated when this extra layer of processing (with its associated indeterminacy) is introduced and whether/how this will affect the overall performance of the checker.

External Evaluation

Evaluation of the predictive accuracy of the translatability indices generated by the checker has begun. A corpus of parallel texts (English patent documents and the corresponding raw MT output) have been used in the first instance to experiment with fine-tuning the system-specific weights for this text type. For each sentence, the TI calculated by the checker is compared with the quality of the raw MT output.

How to determine the quality of a translation has long been a vexed question and can maybe best be determined wrt the set-up in which this system is used. The output of PaTrans is always post-edited and so when assessing the quality of the translations the likely amount of post-editing necessary was an important factor. The fidelity of a the text to its original source is of course crucial, but cases in which only minor corrections are necessary (e.g. the inflectional form of a word needs to be changed) were considered to be less bad than cases where the word order of the translation would have to be substantially revised. This somewhat informal evaluation has resulted in the adjustment of the relative weights of the different indicators so that sentence initial subclauses, PPs and adverbs weigh much heavier than homographs and complex coordination.

However, to fully evaluate the checker, it is necessary to both have data on the actual post-editing effort required to transform the raw MT output into publishable quality, and over and above that, to evaluate its usability and usefulness in other set-ups and as an integrated part of the TQPro architecture.

Conclusions and Future Work

Assessing how suitable a text is for machine translation is an important element in the pursuit of the most productive and cost-effective use of current translation technology and the translatability checker described in this paper is a step in the right direction. The issue of translatability and the notion of calculating a Translatability Index wrt a specific MT system are not new. The Logos Translatability Index which assesses the translatability of texts wrt the Logos system, is described in Gdaniec (1994) and it shares a number of similarities with the translatability checker. Rather than parsing individual sentences it relies on identifying gross statistical properties of a document, but unlike the translatability checker it apparently does not produce an annotated version of the text. In addition to a TI for the text as a whole it produces a description of the ways in which a document is unsuitable for MT which is then used as a basis for improving the source document. There is, of course a range of measures which can also be employed to

[3] The latest version of OTEXT (the text handling format, originally developed during the OTELO project) can be found on the TQPro website: http://www.tqpro.de.

improve the translatability of source texts, ranging from the simple use of spelling checkers to prescribing controlled language use. However, as we have seen in the case of patent texts, users of translation technologies do not always have control over the nature and quality of the texts they must translate.

Since the purpose of the translatability checker is to avoid wasting effort and resources by mistakenly using MT, the checker must necessarily employ a speedy and rather shallow analysis with all its potential associated imperfections. In fact given the nature of natural language, and the state-of-the-art in natural processing it is unlikely that even using a full-blown parser will produce perfect results. There will, therefore, always be a trade-off between speed and accuracy in such a tool. However, the current evaluation results suggest that the level of accuracy is rather high and inadequacies may be offset by the simultaneous production of the annotated text, which explains the numerical results produced by the checker.

The current version of the checker depends on rather outdated linguistic data in the tagger (albeit delivering reasonable results for certain text types) and allows the user to choose between an analysis in terms of MT in general and a single specific MT system and text type. In future versions we would like to extend the checker to apply to other specific systems and maybe even train the tagger for more specific text types so that when the user specifies a particular text type or subject domain the tagger will access the relevant lexicon and rule set. Another interesting long-term possibility would be to extend the tool to treat texts in other source languages, which would involve incorporating not only a different tagger, but also language-specific rules for identifying the relevant translatability indicators.

References

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

Bernth, A. & Gdaniec, C. (2000) MTranslatability AMTA-2000 Tutorial.

Brill, E. (1994). Some advances in rule-based part of speech tagging. In Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94), Seattle, Wa.,

Francis, W. N., & Kucera, H. (1979). Brown Corpus Manual. Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. Department of Linguistics, Brown University.

Gdaniec, C. (1994). The Logos Translatability Index. In Technology Partnerships for Crossing the Language Barrier. Proceedings of the First Conference of the Association for Machine Translation in the Americas (pp. 97-105), AMTA.

Povlsen C., Underwood, N., Music, B., Neville, A. (1998). Evaluating Text-type Suitability for Machine Translation a Case Study on an English-Danish MT System. In A. Rubio, N. Gallardo, R. Castro & A. Tejada (Eds) Proceedings of the First International Conference on Language Resources and Evaluation, Vol 1, (pp. 27-31), Granada, Spain.

Santorini, B. (1990). Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd revision) (http://www.cis.upenn.edu/~treebank/home.html)

Thurmair Gr. (2000). TQPro: Quality Tools for the Translation Process. In Proceedings of the Twenty-Second International Conference on Translating and the Computer, London, November 2000 ASLIB-2000.