Sara Stymne

Using a grammar checker and its error typology for annotation of statistical machine translation errors

ABSTRACT

In this study I extend an error typology for machine translation based on an error typology for a grammar checker. I use this typology to classify errors in a Swedish text machine translated from English by a statistical system. I also investigate if the grammar checker can be used to automatically identify machine translation errors, and how its performance differs from that on several human text types. I found that the extended error typology was useful for annotating the machine translation output.

1 INTRODUCTION

Texts that are produced by machine translation (MT) systems are becoming increasingly more common in today’s society, mainly because of the availability of free online machine translation systems such as Google Translate1. Google Translate, and many current state-of-the-art research systems, are phrase-based statistical machine translation (SMT) systems, that are based on statistical models trained on human translations. One problem with such systems is that the output often is ungrammatical, since there generally is no linguistic knowledge in the systems.

Machine translation systems are often only evaluated using automatic metrics, such as Bleu (Papineni et al., 2002), which compare the machine translation output to one or several human reference translations. Automatic metrics are a quick way to judge the performance of a system and they usually perform well for small changes in one system, or for comparisons of similar systems, while not always being fair for systems of different kinds. An alternative is human evaluation, which is often done to compare systems, either by ranking sentences, or by assigning numerical values to dimensions such as adequacy and fluency (see e.g. Callison-Burch et al., 2007). Adequacy is used to describe the level of semantic equivalence between the source

1 http://translate.google.com/
and target texts, and fluency is used to describe how fluent the source text is on its own, i.e. if it is a good source language text. Human evaluation is a good complement to automatic metrics, but it is very time consuming. A problem with both these evaluation types is that they do not give much information of what the particular problems in a system are. To address this issue, error analysis is needed.

Error analysis has been attempted before for MT. Vilar et al. (2006) described a general error typology that has also been used by several other authors. Their typology has five top categories of errors: Missing words, Word order, Incorrect words, Unknown words, and Punctuation, most of which are further subdivided. I find their suggestion problematic in several respects. First of all the typology mixes identification of errors with finding the cause for certain errors, e.g. by distinguishing unknown words by words that have a lemma that is known to the system, and words that are completely unknown. While such an analysis certainly can be useful, I think it is outside the scope of a general error typology. On the other hand the typology is rather shallow, and has several very large categories. The authors suggest that it can be extended for specific languages, and do so for Chinese, but I would prefer a more thorough base typology, even though I agree with the need for adding language specific elements, since target languages differ, for instance in the type of agreement they have. Moreover, there is an error type for Wrong form of words, but no further division as to the function of the word with wrong form, such as agreement errors or co-reference errors, both of which result in a word with the wrong form. Other authors have used their own error typologies in single studies, but to the best of our knowledge, there are no other generally used error typologies for machine translation.

Error analysis of machine translation can be useful in many respects. For machine translation system developers or researchers it is useful to know what the most common and/or serious errors are, in order to focus development and research on issues that are likely to really improve translation performance. For users of MT systems it can also be useful to know what the common error types are, in order to be able to easier understand the MT output. For buyers of MT systems it can be advantageous to know the weaknesses and strengths of different systems. The possibility to automatically identify, and possibly correct MT errors, is also useful both for system developers, who can use it to improve their systems, and for users, who could possibly use reading tools which highlight errors.

In this study I define an error typology for machine translation based on a previous error typology for a Swedish grammar checker, Granska (Knutsson, 2001). An error typology for a grammar checker, however, is purely monolingual, concerning the fluency of a monolingual text. For machine translation there are also adequacy errors, where the monolingual target text does not reflect the source text semantics, due to mistranslations. This means that I have to extend the typology to cope with adequacy errors. This study is an extension of Stymne and Ahrenberg (2010), where the grammar checker was used for automatic identification and post-editing of SMT output. In this study I focus on human error analysis using the extended Granska error typology. I also include an enriched discussion of the possibilities of automatic error annotation by Granska, based on a comparison with the human annotations. The Granska error typology is based on human texts and the
Granska tool is developed for human texts. SMT output is different from human texts in many respects, so this is also an investigation of how well typologies and tools that were developed for human texts can be applied to a completely different text type, SMT output, compared to different types of human texts.

2 TOOLS AND CORPUS

A statistical machine translation system is created by collecting statistics from a large corpus of texts translated by humans. It contains a number of statistical models that accounts for different aspects of the translation process. The two most important are the translation model, which accounts for adequacy, and contains probabilities for translations of short text segments, and the language model, which accounts for fluency, and contains probabilities for sequences of source words. I trained a standard phrase-based machine translation system for translation from English to Swedish, with the commonly used Moses decoder (Koehn et al., 2007). The system is further described in Stymne & Ahrenberg (2010)². I automatically check the machine translation output with the Swedish grammar checker Granska (Knutsson, 2001), which is a hybrid, mainly rule-based grammar checker. Granska identifies errors and generates error descriptions and correction suggestions.

For both training and testing of the machine translation system I use texts from the Europarl corpus (Koehn, 2005), which is a collection of European parliament speeches, available in 11 languages. I use the English-Swedish portion of Europarl. The system was trained on 701,157 sentences, and the error analysis was performed on 500 sentences, with a total of 11,574 Swedish words.

3 ERROR TYPOLOGY

In this study I base an MT error typology on the error typology of Knutsson (2001), who investigated errors in Swedish human texts in order to develop and evaluate the Swedish grammar checker Granska (Knutsson, 2001). Knutsson performed a large scale investigation of errors in five types of Swedish texts: sport news, foreign news, popular science, government texts and student essays. Table 1 shows the error categories identified by Knutsson, their overall frequency, and their implementation status in Granska. Table 1 also shows the number of errors in each type in SMT output. As can be seen there are many more errors in the SMT output, despite the fact that the text is only 5.8% of the length of the human texts.

All error categories in Table 1 are errors that are found in monolingual texts produced by humans. Thus they only account for the fluency in a text. In machine translation, the target text should also correspond to the semantics of the source sentence. The MT output should not omit, change or add anything to the content of the source text. To be able to use the grammar checker based typology of Knutsson, I thus need to extend it with categories that account for adequacy. These categories are shown in Table 2. Overall fluency errors are much more common than adequacy

---

² The system was trained on 701,157 sentences, and used a part-of-speech sequence model.
errors, which is positive, since they generally are less problematic for understanding the target text, and they are also, in principle, possible to correct automatically.

Table 1. Error types in human texts, their frequency in human texts, their implementation status in Granska (based on Knutsson, 2001), their total counts in SMT output, and the number of errors found by Granska in SMT output. These error types are all fluency errors.

<table>
<thead>
<tr>
<th>Type of error</th>
<th>Count, human texts (of 201,019 words)</th>
<th>Can be identified by Granska</th>
<th>Count, MT output (of 11,574 words)</th>
<th>Found by Granska in MT output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun phrase agreement</td>
<td>69</td>
<td>Yes</td>
<td>126</td>
<td>63</td>
</tr>
<tr>
<td>Predicative agreement</td>
<td>16</td>
<td>Yes</td>
<td>34</td>
<td>18</td>
</tr>
<tr>
<td>Verb form errors</td>
<td>89</td>
<td>Yes</td>
<td>148</td>
<td>41</td>
</tr>
<tr>
<td>Word order</td>
<td>8</td>
<td>To some extent</td>
<td>134</td>
<td>9</td>
</tr>
<tr>
<td>Split compounds</td>
<td>74</td>
<td>Yes</td>
<td>65</td>
<td>11</td>
</tr>
<tr>
<td>Wrong pronoun form</td>
<td>14</td>
<td>To some extent</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Missing function word</td>
<td>56</td>
<td>To some extent</td>
<td>31</td>
<td>-</td>
</tr>
<tr>
<td>Coreference</td>
<td>1</td>
<td>No</td>
<td>26</td>
<td>-</td>
</tr>
<tr>
<td>Extra words</td>
<td>4</td>
<td>No</td>
<td>12</td>
<td>-</td>
</tr>
<tr>
<td>Wrong proposition</td>
<td>11</td>
<td>To some extent</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>Definiteness</td>
<td>4</td>
<td>No</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>Punctuation</td>
<td>2</td>
<td>No</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>Spelling error with gram-sem consequence</td>
<td>55</td>
<td>No</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Contamination of expressions</td>
<td>4</td>
<td>To some extent</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Upper/lower case</td>
<td>11</td>
<td>No</td>
<td>141</td>
<td>-</td>
</tr>
<tr>
<td>Foreign words</td>
<td>-</td>
<td>-</td>
<td>69</td>
<td>-</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>418</strong></td>
<td></td>
<td><strong>810</strong></td>
<td><strong>142</strong></td>
</tr>
</tbody>
</table>

Table 2: Adequacy error types and their counts in SMT output

<table>
<thead>
<tr>
<th>Type of error</th>
<th>Count, MT output (of 11,574 words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mistranslated words</td>
<td>199</td>
</tr>
<tr>
<td>Missing words</td>
<td>88</td>
</tr>
<tr>
<td>Extra words</td>
<td>36</td>
</tr>
<tr>
<td>Wrong form</td>
<td>30</td>
</tr>
<tr>
<td>Wrong order</td>
<td>3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>356</strong></td>
</tr>
</tbody>
</table>
4 ERROR ANALYSIS USING THE TYPOLOGY

In this section I will discuss the results of a manual annotation of 500 machine translated Swedish sentences, annotated for errors using the error typology described in Section 3. The annotation was performed by the author, who is a native Swedish speaker. The number of errors found of each type are shown in Table 1 and 2. In the following I will discuss and exemplify the error types that occurred in the machine translation output. The error types described in section 4.1-4.11 are fluency errors, i.e. they are problematic on the target side, while basically carrying the source semantics, whereas the last five categories, described in 4.12-4.16 are adequacy errors that are problematic with respect to the source side, while they can possibly be grammatical considering only the target. Three of the error types from Table 1 never occur in the MT output; these and the rare punctuation errors will not be discussed.

4.1 Noun phrase agreement errors

In Swedish there is noun phrase agreement between the article, noun and adjectives, on gender, number, and definiteness. Noun phrase agreement errors are very common in the SMT output. The majority of the erroneous NPs contain one adjectival modifier, errors on NPs without modifiers are very rare. There are also a number of more difficult cases, that are not identified by the grammar checker, including NPs with several adjectival modifiers, adjectives with adverbial modifiers, or NPs with conjunctions, exemplified in (1).

(1) There needs to be a sensible and sensitive partnership in the external field
Det behövs en förnuftig och känslig partnerskap på det utrikespolitiska området
It needs a-UTR sensible-UTR and sensitive-NEU partnership-NEU on the
external field-DEF
Det behövs ett förnuftigt och känsligt partnerskap på det utrikespolitiska området

4.2 Predicative agreement errors

There is also agreement in Swedish predicatives between the subject and predicative on gender, number, and definiteness. Just like for NP agreement the simple cases are often found by the grammar checker, but there are many complex cases with conjunction or adverbial modifiers that are problematic, as in (2).

(2) it must be clear and resolute
det måste vara tydlig och resolut
it-SG-NEU must be clear-PL and resolute-SG-NEU
det måste vara tydligt och resolut

---

3 Most examples are snippets of full sentences. All translation examples contain the English source text, the Swedish machine translation and its gloss, and a (close) correct Swedish translation. Italics are used to highlight relevant parts in the snippet. The following abbreviations are used for grammatical functions, when needed for clarity: SG: singular, PL: plural, DEF: definite, INDEF: indefinite, NEU: neuter gender, UTR: utter gender.
4.3 Verb form errors

This is a relatively diverse error type, with many subcategories, such as missing (finite) verb, form after aux/modal/cop, double supine/passive. It includes two main error types, either it detects that a verb, or infinitive marker, is missing, or that a verb has the wrong form. That a verb is missing does not always constitute a real error, since the test text contains many independent snippets and headlines, as (applause from the UEN group), which should not contain a verb. Many of the verb errors in the SMT output cannot be attributed to any of the verb error subcategories, showing that the types of verb errors produced by SMT, not surprisingly, are different from the verb errors in human texts. In many of these cases the target text erroneously keeps the structure of the English source text, as in (3), where the copula of the English progressive form is kept instead of using only the single main verb ignorerar in Swedish.

(3) but that we in the committee are ignoring the fact
men att vi i utskottet är att ignorerar det faktum
but that we in committee-DEF are to ignore that fact
men att vi i utskottet ignorera det faktumet

4.4 Word order errors

Word order errors belong to the most common in the output, and they are hard to find for the grammar checker, which only has partial coverage of word order errors. The most common word order error is related to the fact that Swedish is a V2 language, and English is not, which leads to many cases where the verb and subject should be reversed in Swedish, but are not, as in (4). Another common word order error is misplacement of adverb, which is sometimes different in the two languages.

(4) As a member of the Convention, I have to admit that
Som ledamot av konventet, jag tillstår att
As member of convention-DEF, I admit that
Som ledamot av konventet tillstår jag att

4.5 Split compounds

Splitting compounds into several words is a common writing error in Swedish and many other languages where compounds generally are closed, i.e. written as one word without spaces or other word boundaries between the parts. Writing them as two words changes the stress of the words, and often changes the meaning of the words, which is the basis of many jokes, e.g. kycklinglever (chicken liver) versus kyckling lever (chicken is alive). The grammar checker often finds these when the words have the correct form, but it is also common that words have the incorrect form, which makes them harder to identify, as in (5). This is also the category where the grammar checker has the highest number of false alarms.
4.6 Spelling-related errors

Granska contains a spell checker that marks spelling errors. The SMT output does not have any spelling errors as such, since it can only produce words that are found in the parallel text it is trained on, or words that are present in the input to the system. There are, however, two types of errors that the spell checker can manage to find, problems with true casing and foreign words in the output. Examples of both these types of errors are shown in (6).

(6) listened very carefully to Commissioner Verheugen's words about 'selling' the EU to the candidate countries
lyssnat mycket noga till detta påpekande från kommissionär verheugen ord om 'selling' eu att kandidatländerna
listened very carefully to this point from commissioner verheugen about 'selling' eu that
candidate-countries-DEF
lyssnade därför mycket noga till kommissionär Verheugens ord om 'att sälja' EU till
kandidatländerna

4.7 Missing function words

Missing function words are considered fluency errors when they do not obscure the meaning of the sentence. These errors are quite rare, and generally concern either a missing article in noun phrases (7) or a missing preposition. These errors are only classified as fluency errors when the meaning of the sentence is clear even with the function word missing. If the loss of a function word makes the meaning unclear it has been classified as a missing word (see subsection 4.13).

(7) for it to comply with the very specific decisions adopted in Helsinki
föra det till överensstämma med mycket konkreta beslut som antogs i helsingfors
for that it shall comply with very concrete decisions which were adopted in helsinki
för att det skall överensstämma med de mycket konkreta beslut som antogs i Helsingfors

4.8 Coreference errors

Coreference is when a word, typically a pronoun, refers back to a previous expression. In the coreference errors the form of the coreferent is wrong. These errors are often third person pronouns with the wrong gender, as in (8), or the wrong choice of reflexive pronoun. The anaphora of the coreferent can generally be found in the same sentence, close to the referent, as in (8), but in a few cases it crosses sentence boundaries. These errors mainly occur when Swedish has distinctions that English does not have, such as the neuter and uter gender, in which case the SMT
system tends to choose the most common translation of a word, unless the exact context has been seen before.

(8) so that the Commission does not feel it is engaged on a major task
så att kommissionen inte anser att det är engagerad i en stor uppgift
so that commission-DEF-UTR not feel that it-NEU is engaged in a large task
så att kommissionen inte anser att den är engagerad i en stor uppgift

4.9 Extra words
This category includes extra function words that do not change the main meaning of the sentence, such as doubled prepositions (9) or other words, and spurious determiners or infinitive markers.

(9) where marked differences of opinion were evident in the ministerial discussions
där stora skillnader var påtagligt i på minister-nivå
there big differences-PL were evident-SG in on minister-level
där stora åsiktsskillnader var påtagliga på minister-nivå

4.10 Preposition errors
Preposition errors are those errors where the wrong preposition is used, but where this does not obscure the meaning of the sentence, as in (10). Changing a preposition can also result in changed meaning, for instance if replacing to by from, in which case it has been classified as a mistranslated word (see subsection 4.12). Generally the erroneous preposition is a common translation of the English preposition, as in (10).

(10) What has not been achieved in eleven months
Vad som inte har uppnåtts i elva månader
What which not has been achieved in eleven months
Vad som inte har uppnåtts under elva månader

4.11 Definiteness errors
In Swedish, definiteness is mostly expressed by a definite suffix on the noun, but are in some cases also marked by a definite article. This is different from English, where only the definite article is used for marking definiteness. This difference between the languages cause some definiteness errors, where the wrong way of expressing definiteness is chosen, as in (11), where an article is used instead of a suffix. In a few cases definiteness errors are also due to a different distribution of definiteness in English and Swedish.

(11) because of the fragmentation of our decision-making processes
på grund av den uppdelnings av våra beslutsprocesser
because of the fragmentation-INDEF of our decision-making-processes
på grund av uppdelningen av våra beslutsprocesser
4.12 Mistranslated words
This is the most common of the adequacy errors, and occurs when the wrong translation of a source word is chosen, given the context. Most of these errors are due to the ambiguity of words, and the translation of a word is often correct in a difference context. But there are also some completely wrong translations. These errors occur for most parts-of-speech, and sometimes also change the part-of-speech, as in (12), where aim is translated with a noun instead of a verb.

(12) If you aim for these targets, there is a curious range of percentages
Om ni målsättning för dessa mål, det är ett märkligt rad procentsats
If you target for these goals, it is a curious row percentages
Om ni satsar på dessa mål, finns det en märklig skala av procentsatser

4.13 Missing words
These errors are also relatively common, and occur when a content word from the source is missing in the translation, as in (13). Like the previous category, this is really problematic for the understanding of the target text.

(13) Will that research be incorporated in the sixth framework programme?
Kommer att införlivas i det sjätte ramprogrammet?
Will that be-incorporated in the sixth framework-programme-DEF?
Kommer den forskningen att införlivas i det sjätte ramprogrammet?

4.14 Extra words
In these errors a word that is not present in the source is inserted in the target, changing the meaning of the sentence. An example is shown in (14), which also has a missing word, perspectives.

(14) In tandem with integrating equality perspectives into all policy areas
Tillsammans med integreringen av lika i alla politikområden budgetplanen
Together with integration-DEF of equal in all policy-areas budget-plan-DEF
Tillsammans med integreringen av jämställdhetsperspektiv i alla politikområden budgetplanen

4.15 Wrong form of words, considering the source
These errors differ from the fluency form errors in that they might be grammatically correct, but that they do not reflect the intended meaning of the source, i.e., they can be in indefinite instead of definite, as in (15), which also has several other problems.

(15) weaknesses in the systems of control
svagheter i system för kontroll
weaknesses in system-INDEF for control
svagheter i kontrollsystemen
4.16 Wrong order of words, considering the source

Like the previous category, these differ from the fluency word order errors in that they might be grammatically correct, but the word order is changed with respect to the source, rendering a different meaning than in the source, as in (16). There are only three such errors in the analysed text.

(16) in which the money of poor and persecuted people ends up in the clutches of the mafia där pengarna i fattiga människor slutar i och förföljs försämringen av maffian
where money-DEF of poor people ends up and persecuted degradation of mafia-DEF där pengarna för fattiga och förföljda människor hamnar i maffians nypor

5 AUTOMATIC ERROR ANALYSIS

I compared the automatic identification performance of Granska on the five grammar checker categories that occurred in the statistical machine translation output, with two earlier evaluations on different types of text. One evaluation was on the 201,019 words of human adult texts in five text categories (sport news, foreign news, popular science, government texts and student essays) that the Granska error classification is based on (Knutsson, 2001). The other evaluation was performed on 29,812 words of text written by primary school children (Sofkova Hashemi, 2003, who does not report any results on split compounds).

The results are shown in Figure 1 and 2. Figure 1 shows precision, i.e. the proportion of the identified errors that really are errors, and Figure 2 shows recall, i.e. the proportion of the actual errors in the text that were identified. For all categories except verb form errors, the grammar checker has the highest precision. It has much lower recall than on adult texts, but approximately the same recall as on children's texts. The performance for the grammar checker varies with text type also for the adult texts, however, for instance the precision on NP agreement varies between 0.11-0.72 for the five genres. When the grammar checker does find an error in the SMT output, it tends to be correctly identified. The problem is that it only finds a low proportion of all errors, especially considering that it does not find any other error types, except some spelling related errors, than these five categories.

I also found an evaluation of Granska on texts written by learners of Swedish as a second language (Öhrman, 2000). This study uses different error categories and does not report precision on the individual categories, so it is not included in Figures 1 and 2. The overall precision of error identification in this study is 0.85, and overall recall is 0.35. The recall is very varied between categories, however, with a recall of 0.56 on NP agreement and of only 0.05 on word order. These results are actually quite similar to that of the SMT output. As for the SMT output, the spell checker was able to identify some foreign words and casing errors. Overall, the grammar checker only has a good recall on adult texts. The performance on SMT output is actually similar or sometimes better than for human texts with a high number of errors, as those written by children or learners.
Figure 1: Precision for Granska error identification on three different text types

Figure 2: Recall for Granska error identification on three different text types

6 CONCLUSION

The error typology of Granska, extended with MT adequacy categories, was useful for SMT error annotation. I could classify all errors into the categories, except that I needed additional subcategories for verb form errors. I also think that the results give quite a good picture of the strengths of our SMT system. As system developers it will certainly help us to focus our efforts on common problems. In some cases a further sub-classification would be desirable, for instance by describing which class of function words is missing for Missing function word errors. I plan to develop a
general error typology for machine translation, both based on the lessons learned from using this typology, and from comparisons with previous work on error typology.

There is some promise in automatic identification of SMT errors, since the precision of the grammar checker is very high, it tends to only flag real errors. For most practical applications, however, it is problematic that only a small proportion of all errors is found, which is similar to the grammar checker behavior on non-adult texts. The human error analysis showed that the majority of the SMT errors were fluency errors, for which automatic error identification, and possibly also correction, at least in principle is possible. Using Granska in its current form is not enough, but I believe that it would be useful to investigate other postprocessing strategies.

**BIBLIOGRAPHY**


