Error profiling for evaluation of machine-translated text: a Polish-English case study

Sandra Weiss[1], Lars Ahrenberg[2]
Department of Culture and Communication[1] / Department of Computer and Information Science[2]
Linköping University
E-mail: sandre17@gmail.com, lars.ahrenberg@liu.se

Abstract

We present a study of Polish-English machine translation, where the impact of various types of errors on cohesion and comprehensibility of the translations were investigated. The following phenomena are in focus: (i) The most common errors produced by current state-of-the-art MT systems for Polish-English MT. (ii) The effect of different types of errors on text cohesion. (iii) The effect of different types of errors on readers' understanding of the translation. We found that errors of incorrect and missing translations are the most common for current systems, while the category of non-translated words had the most negative impact on comprehension. All three of these categories contributed to the breaking of cohesive chains. The correlation between number of errors found in a translation and number of wrong answers in the comprehension tests was low. Another result was that non-native speakers of English performed at least as good as native speakers on the comprehension tests.

Keywords: Machine translation evaluation, Error analysis, Polish-English machine translation.

1. Introduction

Nowadays translation is not only a profession but an everyday activity. For our convenience, since quite a while now, there are many translating tools available which can be used instantly on the internet and help us get access to information written in a language that we do not understand. In this study we wished to gauge the performance of those systems, restricted to the language pair Polish-English. The focus of the study is on the text quality they produce and the effect of errors on text cohesion and readers' comprehension.

Automatic metrics for machine translation output such as BLEU, NIST and METEOR have benefitted the development and comparison of machine translation systems tremendously. They are not without drawbacks, however. They are hard to interpret in qualitative terms and they are not really fit for task-based evaluation, as they are defined and applied independently of the intended use of the system. While some of the metrics have parameters that can be set differently, e.g. giving different weights to different n-gram lengths, they are based on comparisons with reference translations, for which the purpose and quality characteristics are not usually known or seen as irrelevant.

For assimilative translation, where the goal is to provide a translation that is good enough to enable a user with little knowledge of the source language, to gain a correct understanding of the contents of the source text, it is hard to avoid using human subjects in the evaluations. It is also of interest, however, to know what features of a translation may cause comprehension problems. Therefore, occurrences of different types of error were investigated, and, as we hypothesized that comprehension problems may correlate with a lack of text cohesion, we also investigated the effect of errors on the cohesion of the translations and observed difficulties of comprehension.

More specifically, we were interested in the following questions:

- What are the most common errors produced by current state-of-the-art MT systems for Polish-English MT?
- What is the effect of various types of errors on text cohesion?
- What is the effect of various types of errors on readers' understanding of the translation?
- Are there differences between native and non-native speakers in their ability to comprehend machine-translated text?

Some recent studies indicate that qualitative evaluations that employ error categories can be at least partly automated (e.g. Xiong et al., 2010; Popović and Burchardt, 2011). In this study, however, the tasks of recognizing and categorizing errors have been performed by one of the authors.

The outline of the paper is as follows. In the next section we describe related work. In Section 3 we describe our method and the data used. In Section 4 we state the most important results which is followed by a discussion in Section 5. Finally, in Section 6 we state the conclusions.

2. Related work

Many different techniques have been available to evaluate MT output. Initially accepted measures of MT evaluation have included examination of MT system output by humans, who grade the correctness of the

translation in terms of conveyed message meaning from the source language and the fluency of expression of this meaning in the target language (White, et al. 1994). Other evaluation methods that have been proposed and used previously include subjective ratings and comprehension tests such as multiple choice questions (Hovy et al, 2003; Wojak and Graliński, 2010). Other than this and supposedly better are procedures that establish how well some human task can be performed when a human is supplied with machine-translated text as suggested in (Taylor & White, 1998) and implemented for various tasks by e.g. (Jones et al., 2005; Voss & Tate, 2006; Specia, 2011).

As regards machine translation between Polish and English an evaluation experiment at the Posen's Faculty of Mathematics and Computer Science is reported in (Wojak & Graliński, 2010). It involved asking subjects some questions based on the source text without supplying them with the source text but only the output text. The exercise requires an understanding of the target text on the part of participating subjects and based on the accuracy of the answers allowed the evaluator to measure the level of comprehensibility and adequacy of the conveyed information in the target text. The experiment successfully showed a correspondence between test results and the quality of translation. The number of subjects were unevenly distributed on the selected texts, however.

3. Method and data

Hovy et al. (2003) provides a general framework for machine translation evaluation that was applied in the planning of the study. The ***purpose*** of the evaluation was defined as identifying how comprehensibility and cohesion are maintained in Polish-English machine translated texts and consequently identifying how accessible the translation texts were to the reader. The ***object*** of evaluation considers MT systems as wholes, selecting from state-of-the-art online Polish-English systems. For ***usability*** the goal was to assess whether the selected machine translation engines can be used with success for the purpose of producing texts that give valuable information to the reader. The ***translation task*** is assimilation with the focus on the type of information which may be of interest to free movers to Poland without advanced knowledge of Polish and on the satisfactory level of translation quality needed for the readers to make use of the translated information. The ***user characteristics*** of the systems are English speakers with basic or zero knowledge of Polish, but familiar with MT engines. In the conducted comprehension tests they were both native and non-native English speakers with zero knowledge of Polish. The ***document types*** selected were organisational websites, local news sites, and information portals for Internet access.

3.1 Texts and systems

Five online Polish texts were chosen for the study. They were selected as likely texts to be searched for by English speakers residing in Poland without knowing Polish. The texts differ in length as well as in complexity. Some basic data on the texts can be found in Table 1. For Polish, a word is regarded as complex if it has four or more syllables (Gorczyca, 2010).

The titles of the texts were as follows:
Text1: O *Wojewódzkim Urzędzie Pracy* (About the Regional Labour Office). A promoting homepage of the labour office in Krakow.
Text 2: *Młodzieżowe Targi Pracy i Edukacji* (Work and Education Youth Fairs). A text containing information about the fairs including the date and place and application information.
Text3: *Zatrudnianie cudzoziemców - zmiana wzorów dokumentów* (Employing foreigners – change of document patterns). A webpage produced by the Poznan job centre.
Text 4: *Polonijne media w londyńskim ratuszu* (Polish Media in London's City Hall). National, Local News: onet.pl internet portal.
Text 5: *FAQ telewizja kablowa* (FAQ cable TV). Information Portal of Internet/TV provider Multimedia.

Each of the texts were translated by three online translation engines: Google Translate, Bing Translator and SYSTRANet Translator, giving a total of 15 target texts. While Bleu scores for the translations differed substantially between systems (lowest: 0.14, highest 0.38 for the same text) they had rather similar performance, when looking at error counts. For this reason we anonymize them in the rest of the paper.

Texts	W	S	ASL	ASW	Complex words/sent
Text1	231	14	16.5	2.6	5
Text 2	302	13	23.2	2.3	4.7
Text3	239	13	18.4	2.7	5.5
Text 4	629	35	18	2.1	2.3
Text 5	466	37	12.6	2.1	2

Table 1: Data on the selected texts. W – number of words; S – number of sentences; ASL – average sentence length; ASW – average syllables per word.

3.2 Error analysis

Each of the fifteen texts were analysed for errors according to a common multi-level error taxonomy, having at the top level the following categories:
Missing word (MW), a word that should have appeared in the translation does not do so,
Extra word (EW), a word appears in the translation which should not be there,
Not translated (NT), a word from the source text appears in the translation, instead of a proper target language counterpart,
Incorrect word (IW), a word can be explained from the source text, but is not an adequate translation
Incorrect form (IF), a word of the translation has a

proper stem but inaccurate inflection,
Word order (WO), a word or phrase is misplaced in the translation.

All the error categories were further subdivided according to the part-of-speech affected and some were also subdivided according to phenomena where Polish grammar differ from English. As can be seen there is a partial overlap with the commonly used taxonomy employed by Vilar et al. (2006) but we include no categories, such as Unknown words, that requires knowledge of the system dictionaries.

3.3 Analysis of cohesion

Source texts and translations were analysed for the occurrence of cohesive chains. The analysis of cohesion is based on Halliday & Hasan (1976). Four types of chains were recorded: **referential** chains, i.e., repeated reference to the same entity, **repetition**, i.e., use of the same word or phrase for a property or relation, **semantic relations** such as hyponymy, synonymy, partonomy, between lexical items, and **conjunction**, i.e., explicit signals that establish causal, temporal and other relations between clauses.

Particular attention was paid to chains that were broken in translation on the assumption that such broken chains could cause comprehension problems. As an example, consider a system translation such as *"Politely we inform, that submissive change the examples drukow of declaration about the intention of the entrusting of the execution of work by the foreigners without the necessity of the obtainment of permit on the work."*. Here, the word 'drukow' should have been translated to something like 'form' which has a semantic relation to the word 'document' in the previous sentence, and so the semantic cohesive chain is broken due to a Non-translated word.

3.4 The comprehension tests

For the comprehension tests, two texts were selected, Text1 and Text3, and the two translations that could be considered the best and the worst given the number of errors in them. Incidentally, the system that produced the "best" translation for Text1 was the same that produced the "worst" translation for Text3, and vice versa. This procedure gave the following four translations to be used in the tests:

Translation 1b *"O Wojewódzkim Urzędzie Pracy"* the translation with the least number of errors (13) for this text.
Translation 1w the same text translated by another system and having the most errors (15)..
Translation 3b *"Zatrudnianie cudzoziemców - zmiana wzorów dokumentów"* the translation with the least number of errors (37) for this text.
Translation 3w, the same text translated by another system, having the most errors (44).

The comprehension tests were composed of ten open questions each, which subjects had to answer with the information retrieved from the output translation given to them. The questions were based on the source text which was not accessible to the subjects. The subjects were informed that the translations presented to them were done by machine translating engines but they were not aware of which translation engines were used, avoiding any possible prejudice affecting their commitment to the task by making them less attentive in case of a translator they might dislike. There were twenty subjects and each of the twenty subjects read one version of each of the two texts and answered the questions. Thus, each version was read by ten different subjects, five of which were native English speakers and five who were not. All subjects were students in the age range 20-30 and having passed the level of English required to enter an international Master's programme at Linköping University. Native speakers were English, American as well as Irish, while non-native English speakers had Indian, Iranian, French, Chinese, Mexican, and Norwegian backgrounds.

The difficulty of questions ranged, according to a scale with three levels:
Level 1 questions – relatively easy - asking about short pieces of information, requiring one or two words to answer.
Level 2 questions – medium difficulty – requiring one sentence answers but offering word-clues. The key word in the question is present in the answer in the reference translation, and any corresponding synonym in a system translation would still give a clue.
Level 3 questions – questions which require as an answer more than one idea discussed in the text, sometimes ideas placed in different places in the text.

The answers of the subjects were analysed in relation to the reference answers that were established and then grouped into a) correct answer marked as '1' and b) incorrect answer marked as '0' in the statistical tables. In order to confirm that the source texts in question were perfectly understandable in the source language, Polish, twenty Polish native speakers (10 subjects to each of the 2 source texts) answered exactly the same questions translated from English to Polish, about the texts with a 100% correct answer rate.
All questions for which there were fewer than 60% correct answers were analyzed in detail. This happened for 6 question translations out of 20, or 30% of Text-1, and 11 question ranslations, or 55%, of Text-3 (see Tables 4 and 5). In this analysis the part of the translation needed to get the correct answer was identified as well as the errors occurring in that part. Errors were then categorized as critical for missing the correct answer, possibly contributing to the difficulty, or not really affecting the answer. As an example consider

Q7 of Text-3 as translated by one of the systems, its reference translation, and its annotation:

Q7: What happens with the copy of declaration?
Ref: The copy stays in the Labour Office.
System translation: "the original statement, the employer shall provide to an alien who must provide it to the consular establishments of the dyplomatyczno-in the place of residence, and a copy of the registered claim remains in Office work."
Corrected translation: "The original statement, the employer shall provide to a foreigner who must provide it to the consular-diplomatic establishment in the place of residence, and a copy of the registered claim remains in the work/Labour office."

With this analysis we find one instance of an incorrect word (IW[Noun]): the noun in dative case 'cudzoziemcowi' translates to 'foreigner' not 'alien'. There is also a word order problem: the adjective 'work/Labour' should precede the noun 'office' in the noun phrase providing the answer. These two errors are judged to be the critical ones.

	IW	EW	MW	NT	IF	WO	Total
Text1	13	5	10	5	1	8	42
Text2	44	6	33	18	2	26	129
Text3	40	12	31	14	2	22	121
Text4	43	23	39	21	14	21	161
Text5	59	17	42	5	8	26	157
Total	**199**	**63**	**155**	**63**	**27**	**103**	**610**
%	32.6	10.3	25.4	10.3	4.5	16.9	100

Table 2. The distribution of error types over texts.

4. Results

A summary of the number of errors in each text can be found in Table 2. It can be seen that the generally most common error is Incorrect Word (IW), that accounts for just about one third of all errors, followed by Missing Word (MW) and Word Order (WO). Very few errors are of the type Incorrect Form (IF) which may be attributed to the fact that translation is into English, which is morphologically poor. The most common types (IW, MW, WO) have a fairly even distribution over all texts, while the distribution for the other types is more varied. Also, the variation in the frequency of errors is noteworthy.

If we relate the number of errors to the number of words in the texts, we can see that Text3 has the highest relative frequency of errors and is also the text with highest value on most of the complexity metrics. For the other texts there are no clear correlations.

4.1 Effects of errors on cohesion

When looking at cohesive chains, the number of chains is about the same for each translation. Except for one case, the ratio of the text version with the smallest number of chains to the one with the largest number was 90% or above. The ratio of broken chains ranged from 1.5% to 15% with a mean of 8.5%. System averages were 6.3%, 10.6%, and 9.6%, respectively.

	Errors	Words	Errors/Words
Text1	42	231	0.18
Text2	129	302	0.43
Text3	121	239	0.51
Text4	161	629	0.26
Text5	157	466	0.33
Total	610	1874	0.33

Table 3. Average number of errors per word for each text.

The most frequently occurring cohesive chain is repetition that constitutes 38% of all cohesive chains. The second most common chains are referential chains and the chains of semantic relations. They both represent 23% of all cohesive chains. The chains of conjunctions were relatively few (16%).

The most interesting finding was that different types of chains were affected quite differently. Reference chains were affected the most with some 19% (out of 258) being broken, while conjunctions were only broken in one instance out of 60. Repetition was the second most affected with 7% (out of 451) being broken.

Referential chains represent 58% of all broken chains. The second most frequently broken chains are repetition chains (32%). From the analysis, it is apparent that most of those chains are broken due to Incorrect words, Not Translated words or Missing words in the translations. Extra words can also create unnecessary chains or elements in the chain which in turn, sometimes may confuse the reader as in the case of, for example, additional personal pronouns or prepositions.

4.2 Effects of errors on comprehension

Two texts were selected for the comprehension test. The chosen texts were Text1 and Text3 as these were of equal size but differed widely in text complexity and the number of errors in them (see Table 1). The translations with the highest and lowest number of errors were used. The questions were selected to be evenly spread over the texts and ranged from ones that could be answered by a single word or short phrase to questions that required the collection of information from different sentences (section 3.4). Somewhat contrary to expectations difficulties occurred also for questions considered to be at the lowest level. Results for translations of Text1 are shown in Table 4, while results for translations of Text3 are shown in Table 5.

Text1 had fewer errors than Text3 for both systems. This correlates well with the fact that, on average, there were more correct answers for Text1 than Text3. However, when one looks at individual questions, the

Qs	Translation 1b		Translation 1w	
	Native	Non-Nat.	Native	Non-nat.
Q1	4	4	5	5
Q2	5	5	5	5
Q3	5	4	3	3
Q4	3	3	5	4
Q5	1	1	1	1
Q6	5	5	5	4
Q7	3	3	4	5
Q8	4	5	3	5
Q9	4	5	5	5
Q10	0	2	4	4
Aver.	3.4	3.7	4.0	4.1

Table 4. Correct answers per question and subject category for the translations of Text1.

Qs	Translation 3b		Translation 3w	
	Native	Non-Nat.	Native	Non-nat.
Q1	3	4	0	1
Q2	5	5	5	5
Q3	5	5	3	2
Q4	1	4	3	2
Q5	0	1	3	3
Q6	5	4	5	4
Q7	0	4	2	2
Q8	5	3	5	3
Q9	2	2	5	5
Q10	2	4	1	2
Aver.	2.8	3.6	3.3	2.9

Table 5. Correct answers per question and subject category for the translations of Text3.

number of errors shows only a weak tendency to correlate negatively with the number of correct answers for that question. Correlations are in the interval -0.1 to -0.6.

The distribution of error types that have occurred in the relevant sections of the texts is shown in Table 6. As it is hard to judge with any certainty whether an error is critical or not for other readers, we have judged some errors as potentially critical and made counts both with and without those errors. This gives an interval rather than an exact figure for the critical instances. In any case, the results indicate that the kind of error is important and not just the number of errors. Close to half of the errors of type Not translated (NT) are judged as critical, while the ratio for Incorrect word (IW) is in the range 25-30%. For the other types, the results are more uncertain.

4.3 Non-native vs. native speakers

A rather surprising outcome was that non-native speakers of English did not perform worse than native speakers. In fact, for three out of four combinations used, the non-native English speakers as a group had better results than the native English speakers.

.

Error types	**#Instances in the texts**	**#Critical instances**
Incorrect word	36	9-11
Extra word	15	2-5
Missing word	18	3-3
Not translated	18	7-8
Incorrect form	3	0
Word order	19	2-5
All	109	23-32

Table 6. Distribution of errors that were judged to be critical for comprehension.

Unfortunately we did not do an independent rating of the language skills of the subjects, but it has been claimed that bi- or multiliterate readers can be more flexible in their reading (Singhal, 1998). It may be also conditioned in the subjects' educational background and generally in their proficiency in English. It is possible that the non-native speakers had a command of English close to native speakers. It may just as well depend on the fact that English non-native speakers in the process of learning the English language became familiar with many error types that are possible to be made in such translations and consequently, it was easier for them to fill in the gaps caused by the errors.

5. Discussion

According to statistical correlation analysis of percentages of correct answers and the amount of errors in the relevant parts of the text the questions were relating to, there is no linear relation. The correlation is very weak (36% at most) in case of both texts which suggests that the understanding of the translations by human subjects was not conditioned by the amount of the generated errors but rather the error type and its seriousness. There were questions where the correct answer rate was from 80% to 100% even though the number of errors was from 5 to 7 in one sentence or two sentences of a text. On the other hand, there was a question concerning a fragment of the translation that did not have any errors and the subjects answered it correctly only in 20%. That lack of correlation between incorrect answers and number of errors would also explain why the initially assumed best translation-1b of Text1 turned out to be the one with the least percentage of correct answers, while the high percentage of correct answers (over 80%) for translation-1w suggests that it was better understood by the subjects. Given that the difference in the number of errors was small for this text (only 2 instances), the difference in the number of correct answers is still striking. The plausible reason for that may be found in the seriousness of the mistakes found in the translations. Even though the translation-1w has more errors (15), they seem to be of less importance. The most frequent error for that translation is the WO (4) constituting nearly 1/3rd of all errors which we claim to be less influential. The second largest group is MW (4) which in 50% concerns only

prepositions. At this point, we can see that a little more than a half of the errors did not have a great influence on the general comprehensibility of this translation. On the other hand the assumed best translation-1b, had a smaller amount of errors but more errors of significant influence.

In the case of translations of Text3, they have, on average, very similar results. Translation-3b had an average of 64% of correct answers and translation-3w an average of 62%. Again we see that the difference in number of errors is not significant for comprehension.

The influence of broken chains on the comprehensibility of the translations is very closely related to the influence of significant errors. For translations of Text3, the number of broken chains in the near context of relevant answers to problematic questions is high, with referential chains and repetitive chains being the most affected (10, and, 8 respectively for 13 instances). For Text1, on the other hand, they were very few. If there are broken cohesive chains then the cause is often missing or incorrect words or non-translated words that become missing or affected elements of the chains. Therefore, we cannot separate the effects of broken chains from that of the three most influential general errors.

Originally,a goal of this work was to compare system performance and find out which system had the best performance for Polish – English translations. However, no such claim can be made as all the systems generated relatively similar translations. However, there was a difference in their error profiles with one system having a higher frequency of Incorrect word and Not translated than the others.

6. Conclusions

While this study is small, it lends support to the view that for assimilative translation some error types are more likely to cause difficulties than others; and in our case Incorrect words and Untranslated words are the most troublesome, while errors in word forms are negligible. The statistics show that on average, IW represents 32%, MW 26%, NT 10% of all errors in all fifteen translations. We can see that even though the error NT is less frequent it is more serious and still affects the comprehensibility greatly. It also depends very much on what part of speech was affected. Generally, the errors that affected verbs and nouns are the most serious ones and those that affect articles, noun genders and prepositions make the least impact on the comprehensibility of the text.

Somewhat to our surprise, non-native English speakers did not fare worse than native speakers, rather the opposite tendency was seen.

There are vast possibilities of improving this research and pushing it in other interesting directions. Firstly, and similarly to the conclusions of (Wojek & Galińsky, 2010) it would certainly be more appealing to carry out a similar study on a larger scale with more text samples and more subjects solving the tests, in order to be able to generalize and see how the claims and assessments would apply in relation to, for example, text types. It would allow us to see what the most common mistakes are, for texts written in formal or informal language, etc. It is also worth analyzing further which errors affect what parts of speech and look for correlations to general comprehensibility of the translations. This study also brings to mind a question: how do human subjects make up for the gaps those errors create? An answer to this question could probably explain why in the comprehension tests, questions concerning fragments of the text with many errors were answered correctly and vice versa. However, the greatest obstacle in carrying out such studies is the disadvantages of human MT evaluation which is inefficient time-wise, but for the purpose of an in-depth analysis of cohesive relations in the translations, there may not be other or better alternatives.

7. References

Halliday, M A K, & Hasan, R. (1976). *Cohesion in English*. (R. Quirk, Ed.) (Vol. 9, p. 374). London: Longman Group Ltd.

Hovy, E., King, M., & Popescu-Belis, A. (2003). Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, (17), 43-75.

Gorczyca, D. (2010). Tekst w Internecie. Zarządzanie Zmianami Biuletyn POU, 5.

Jones, D., Shen, W., Granoien, N., Herzog, M. and Weinstein, C. (2005). *Proceedings of the 2005 International Conference on Intelligence Analysis*, McLean, VA

Popović, M. and Burchardt, A. (2011). From Human to Automatic Error Classification for Machine Translation Output. *Proceedings of the 15th Conference of EAMT*, Leuven, Belgium, pp. 265-272.

Singhal, Meena (1998) A Comparison of L1 and L2 Reading: Cultural Differences and Schema. The Internet TESL Journal. Vol. IV, No. 10, October 1998

Specia, L. (2011). Exploiting objective annotations for measuring translation post-editing effort. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, pp. 73–80, Leuven, Belgium.

Taylor, K. and White, J. (1998). Predicting What MT is Good for: User Judgements and Task Performance. In *Proceedings of the Third Conference of AMTA*, Langhorne, PA.

Vilar, D., Xu, J., D'Haro, L. F., & Ney, H. (2006). Error Analysis of Statistical Machine Translation Output. *Proceedings of the Conference on Language Resources and Evaluation* (pp. 697-702).

Voss, C. R. and Tate, C. R. (2006). Task-based Evaluation of Machine Translation (MT) Engines: Measuring How Well People Extract Who, When, Where-Type Elements in MT Output. *Proceedings of the 11th Annual Conference of EAMT*, Oslo, Norway.

White, J. S., O'Conell, T., & O'Mara, F. (1994). The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. *Proc of the AMTA* (pp. 193-205).

Wojak, A., & Graliński, F. (2010). Matura Evaluation Experiment Based on Human Evaluation of Machine Translation. *Proceedings of the International Multiconference on Computer Science and Information Technology* (pp. 547-551). Poznań, Poland.

Xiong, Deyi, Zhang, Min, and Haizhou, Li (2010). Error detection for statistical machine translation using linguistic features. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 604-611.