Error Analysis of Statistical Machine Translation Output

David Vilar*, Jia Xu*, Luis Fernando D'Haro†, Hermann Ney*

*Lehrstuhl für Informatik VI – Computer Science Department
RWTH Aachen University
52056 Aachen, Germany
{vilar,xujia,ney}@cs.rwth-aachen.de
†Speech Technology Group – ETSI de Telecomunicación
Dpto. Ingeniería Electrónica
Universidad Politécnica de Madrid
28040 Madrid, Spain
lfdharo@die.upm.es

Abstract

Evaluation of automatic translation output is a difficult task. Several performance measures like Word Error Rate, Position Independent Word Error Rate and the BLEU and NIST scores are widely use and provide a useful tool for comparing different systems and to evaluate improvements within a system. However the interpretation of all of these measures is not at all clear, and the identification of the most prominent source of errors in a given system using these measures alone is not possible. Therefore some analysis of the generated translations is needed in order to identify the main problems and to focus the research efforts. This area is however mostly unexplored and few works have dealt with it until now. In this paper we will present a framework for classification of the errors of a machine translation system and we will carry out an error analysis of the system used by the RWTH in the first TC-STAR evaluation.

1. Introduction

Evaluation of machine translation (MT) ouput is a controversial task in the MT community. Several automatic measures have been proposed the Word Error Rate (WER), the Position independent word Error Rate (PER), the BLEU (Papineni et al., 2002) and the NIST (Doddington, 2002) measures being the most widely used ones. A relationship between these error measures and the actual errors found in the translations is however not easy to find. The identification of the most prominent problems of a translation system is important in order to focus research efforts. The goal of this work is to present a framework for (human) error analysis of machine translation output and analyse the results obtained by our group in the first TC-STAR evaluation campaign.

The goal of the TC-STAR project[1] is to build a speech-to-speech translation system that can deal with real life data. We concentrate on three translation directions: Spanish to English, English to Spanish and Chinese to English.

For the Spanish-English language pair we have collected data from speeches held in the European Parliament Plenary Sessions to build an open domain corpus. There are three different versions of the data, the official version of the speeches as available on the web page of the European Parliament, the actual exact transcription of the speeches produced by human transcribers and the output of an automatic speech recognition system. This provides an useful framework for testing various translation technologies. The first version of the data, called Final Text Edition (FTE), consists of written text and text-to-text translation methods can be used. Using the verbatim human produced transcription we can investigate the impact that spontaneous speech effects (ungrammaticality, false starts, hesitations, etc.) have on the translation quality. Lastly, in the third condition (ASR), the integration of speech recognition and translation systems is investigated.

For Chinese to English translation we do not have such appropriate data available. We use broadcast news as provided by the Linguistic Data Consortium (LDC), but in this case the distinction between the FTE and verbatim data is somewhat artificial.

[1] http://www.tc-star.org/

2. The RWTH Statistical Machine Translation System

Our statistical machine translation system is based on a log-linear combination of seven different models, the most important ones being phrase based models in both source-to-target and target-to-source directions and a target language model. Additionally we use IBM1 models at phrase level, also in source-to-target and target-to-source directions; and phrase and length penalties. We then proceed to generate n-best lists and rescore them with IBM1 models at sentence level and additional clustered language models. A more detailed description of the system can be found in (Vilar et al., 2005).

3. Error Classification

In order to find the errors in a translation, it is useful to have one or more reference translations in order to contrast the output of the MT system with a correct text[2]. However, as it is well known in the machine translation community, there are several correct translations for a given source sentence, which poses a difficult problem for automatic evaluation and comparison of machine translation systems. Therefore the use of this reference translations must be done with care.

[2] And a tool for highlighting the differences also proved to be quite useful.

The classification of the errors of a machine translation system is by no means unambiguous. The classification scheme we propose in this work is an extension of the error typology presented in (Llitjós et al., 2005). It has a hierarchical structure as shown in Figure 1. In the first level we have split the errors in five big classes: "Missing Words", "Word Order", "Incorrect Words", "Unknown Words" and "Punctuation" errors.

A "Missing Word" error is produced when some word in the generated sentence is missing. We can distinguish two types of errors, when the missing words is essential for expressing the meaning of the sentence, and when the missing word is only necessary in order to form a grammatically correct sentence, but the meaning is preserved. Normally the first type of errors are caused by missing "main words" like nouns or verbs, but this not always the case, as for example a missing preposition can alter the meaning of the sentence significantly. This first type of errors is of course more important and should be addressed first. For each of these divisions one could further distinguish which lexical category ("Part of Speech") is missing, as different word types may have different treatments. For simplicity these subclasses are not included in Figure 1.

The next category concerns the word order of the generated sentence. Here we can distinguish between word or phrase based reorderings, and within each of these categories between local or long range reorderings. In the case of word based reorderings, we can generate a correct sentence by moving individual words, independently of each other, whereas when a phrase based reordering is needed, blocks of consecutive words should be moved together to form a right translation out of the generated hypothesis. The distinction between local or long range is difficult to define in absolute terms, but it tries to express the difference between having to reorder the words only in a local context (within the same syntactic chunk) or having to move the words into another chunk. For the Chinese-English language pair, a more refined classification scheme, dependent on the sentence type has been carried out, see Section 5.3 for more details.

The widest category of error are the "Incorrect Words" errors. These are found when the system is unable to find the correct translation of a given word. Here we distinguish five subcategories. In the first one, the incorrect word disrupts the meaning of the sentence. Here we could further distinguish two additional subclasses, when the system chooses an incorrect translation and when the system was not able to disambiguate the correct meaning of a source word in a given context, although the distinction between them is certainly fuzzy.

The next subcategory within the "Incorrect Words" errors is caused when the system was not able to produce the correct form of a word, although the translation of the base form was correct. This is specially important for inflected languages, where the big variability of the open word classes poses a difficult problem for machine translation. How to further analyze the errors that fall into this category is very much dependent of the language pair we are considering. For example, for the Spanish language, being a highly inflected language, it is useful to distinguish between bad verb tenses and concordance problems between nouns and adjectives or articles.

Another class of errors is produced by extra words in the generated sentence. This kind of error was introduced mainly when investigating the translation of speech input, as artifacts of spoken language may produce additional words in the generated sentence.

The last two classes are less important. The first one ("Style Errors") concerns a bad choice of words when translating a sentence, but the meaning is preserved, although it can not be considered completely correct. A typical example is the repetition of a word in a near context. In this case a human translator would choose a synonym and avoid word repetition. The second one concerns idiomatic expressions[3] that the system does not know and tries to translate as normal text. Normally these expressions can not be translated in this way, which causes some additional errors in the translation.

Unknown words are also a source of errors. Here we can further distinguish between truly unknown words (or stems) and unseen forms of known stems.

A variation of this category has a special importance for the Chinese-English language pair. For the majority of European languages, or even languages that share the same alphabet, unknown proper names can be "translated" simply by copying the input word to the generated sentence, without further processing. Chinese characters, however, can not be translated into English by itself, and a conversion, sometimes guided by the pronunciation, is required. Therefore for this language pair we also distinguish between unknown person, location, organization and other proper names.

Lastly there can also be punctuation errors, but, for the current machine translation output quality, these represent only minor disturbances for languages without fixed punctuation rules, and are not further considered in this work.

Of course, the error types so defined are not mutually exclusive. In fact it is not infrequent that one kind of error causes also another one to occur. So for example, a bad word translation can also cause a bad ordering of the words in the generated sentence.

4. Corpora

The corpora considered in this analysis are the corpora used in the TC-STAR evaluation: the European Parliament Plenary Sessions (EPPS) corpora for the English-Spanish language pair, and broadcast news for the Chinese-English language pair.

A description of the EPPS data can be found in (Vilar et al., 2005). The statistics of the corpora can be found in Table 1 and the results in Table 2. Note that for all the EPPS tasks the same training corpus was used, consisting of the Final Text Editions data, only the preprocessing of the corpus was different. This produces a slight mismatch between training and testing data, which contributes in increasing the error rates for the Verbatim and ASR conditions.

For the Chinese to English translation, the training corpora are provided by the Linguistic Data Consortium (LDC), the

[3] As an example: "It's raining cats and dogs".

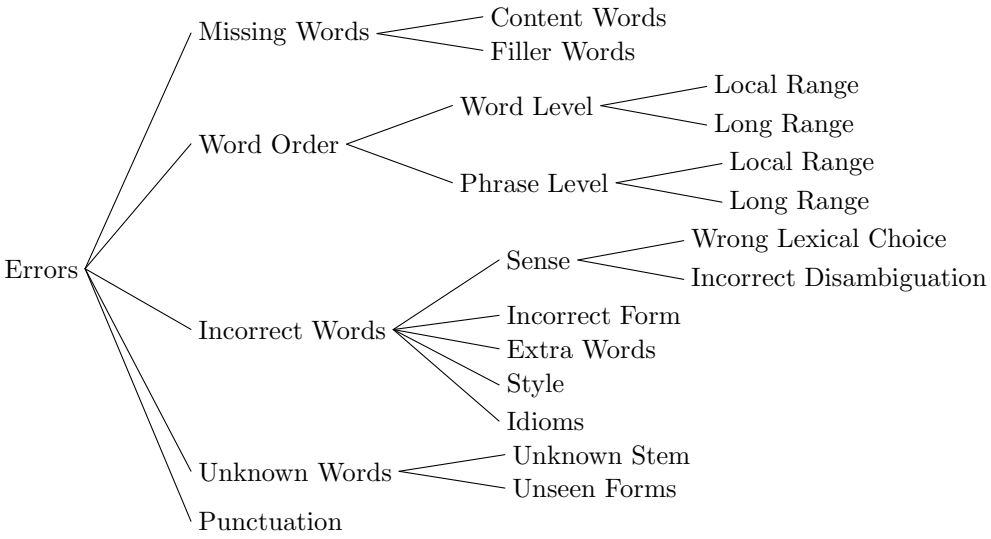

Figure 1: Classification of translation errors.

domain being news articles. A list of these corpora can be found at the LDC web pages (LDC, 2005) under the "Large Data Condition". The evaluation data is selected from the manual transcription of the "Voice of America".

As shown in Table 1, the whole training corpus contains more than seven million sentences after the filtering. Each of the evaluation data has 494 sentences. After preprocessing, such as Chinese word segmentation and the number-, hour- and date-categorization, we obtained nearly 200 million Chinese running words for training. The evaluation data were also preprocessed. Because of the large amount of training data, there were very few Chines unknown words. The translation results for the Chinese-English tasks are presented in Table 2.

5. Error Statistics

In this section we will analyze in more detail which are the most prominent source of errors in each of the tasks within the TC-STAR project.

5.1. English to Spanish

5.1.1. EPPS FTE data

As stated earlier, Spanish is a highly inflected language, havig for example 17 different verb tenses (not counting impersonal forms like gerundium). It is often the case that the correct verb gets chosen, but the tense is incorrect. This is epecially true for past tenses, as Spanish differentiates several tenses depending if the action was terminated or not, and the subjunctive tenses, which have no direct correspondence into English. The errors due to bad tense amount to 15.1% of the total. There are also cases where the tense is correctly generated, but the person is not correct. This is mainly motivated by the relatively long sentences present in the corpus, as the verb and the corresponding subject information necessary for generating the correct form of the verb are relatively far apart. Neither the translation nor the language models are able to handle so long range context information.

Incorrect lexical choice is also an important problem. Especially there is an important disambiguation problem, namely the pair of Spanish verbs "ser" and "estar". Both are translations of the English verb "to be", the first one being used for permanent properties of objects or persons, and the second one is used for expressing temporary qualities or locations[4]. In many cases the system is not able to distinguish between these two verbs.

The next most frequent errors are caused by missing words, 7.9% of the total errors caused by missing content words. Another important source of errors concerns the generation of the correct order of the sentence. Although English and Spanish have a very similar word order, there are some deviations. The most frequent ones are the adjective-noun pairs, English uses the form "adjective-noun" while in Spanish it is more common to use "noun-adjective". In most cases this permutations are correctly handled by the phrase based translation model, as they occur only in a local context, but for some longer ranging reorderings or for unseen adjective noun pairs, the system is not able to handle them correctly. 11.6% of the errors are caused by local range word based reorderings.

There are also problems with the concordance between names, adjectives and articles. In contrast with English, Spanish articles and adjectives must match the gender and number of the noun. As was the case when handling reordering, in most of the cases this gets modelled by the phrase based translation model, but there are still some errors left. The complete error statistics for this task can be found in the column of the FTE Spanish-English in Table 3.

5.1.2. EPPS Verbatim data

The errors found for the verbatim data condition are quite similar to those found for the FTE condition. However, the input in this condition has some ungrammatical constructions which constitute an additional source of errors, as discussed in Section 4. The statistics are shown in Table 3.

[4]This is a rough simplification and the exact use is more refined than that.

		EPPS		BN	
		Spanish	English	Chinese	English
TRAINING DATA	Sentence pairs	1 207 740		7 082 390	
	Running Words	34 851 423	33 335 048	198 867 499	212 674 144
	Vocabulary	139 587	93 995	223 258	351 198
	Singletons	48 631	33 891	99 937	162 240
FTE TEST DATA	Sentences	840	1094	494	
	Running Words	22 756	26 885	13 852	
	Vocabulary	3644	3744	2 585	
	OOVs (running words)	40	102	1	
VERBATIM & ASR TEST DATA	Sentences	1073	792	494	
	Running Words	18 896	19 306	12 508	
	Vocabulary	3302	2772	2 586	
	OOVs (running words)	145	44	1	
	Input WER (ASR only)	10.1%	9.5%	13.7%	
	Number of Politicians*	36	11	—	

* Unknown number of interpreters.

Table 1: Statistics of the TC-STAR corpora.

		WER [%]	PER [%]	BLEU [%]	NIST
ENGLISH TO SPANISH	FTE	39.9	30.6	48.6	9.95
	Verbatim	46.1	35.4	42.5	9.33
	ASR	49.8	38.6	38.7	8.73
SPANISH TO ENGLISH	FTE	34.3	25.9	55.0	10.68
	Verbatim	42.5	31.7	45.9	9.75
	ASR	46.6	35.4	41.5	9.12
CHINESE TO ENGLISH	FTE	75.8	55.4	16.5	5.95
	Verbatim	78.6	58.0	16.8	5.99
	ASR	78.1	57.8	16.2	5.87

Table 2: TC-STAR evaluation results.

When comparing the error statistics with the FTE data, the most prominent difference is an increase in the number of missing words. This can be explained by the ungrammatical constructions of the input text. If we decompose this kind of errors into missing context words and missing filler words, the increase is mainly due to this last kind of errors. That means that the ungrammaticality of the input sentence is somewhat transfered to the generated sentence.

5.1.3. EPPS ASR data

The analysis carried out for the Verbatim data is also applicable to the ASR data. In this condition, however, we have an additional source of errors, namely the errors due to the speech recognizer. The input data has a 9.5% word error rate. If we decompose these errors into insertion, deletion and substitution errors we see that the most important errors are substitution errors amounting to a total of 54.7% of the errors (deletions amount to 25.0% and insertions to 20.3%). This trend gets transfered to the translations. If we compare the output of the verbatim system with the output of the ASR system, we find that 62.8% of the differences correspond to substitutions. This increase is easily explained if we consider that a change in a word (a substitution) also changes the surrounding words in the translation, as the context changes and another phrase gets selected in the translation process. The deletion and substitution errors are not so important, as they affect normally articles or prepositions that are not essential for the translation process.

5.2. Spanish to English

For the reverse direction, namely translating from Spanish to English, we have observed similar problems. However, as English is a language with nearly no inflections, the error rates achieved by the systems are better than for Spanish. The main problem in this direction for each of the conditions are presented in this subsection.

5.2.1. EPPS FTE data

When generating English, the most prominent source of errors is a bad lexical choice. The amount of errors due to incorrect translations and bad disambiguation together amount to 28.2% of the total errors. However, more than an increase in the absolute number of errors when comparing to the opposite direction, the higher percentage is motivated by the decrease in the number of other errors.

Missing words are the second most important source of errors. However, most missing words are simply filler words, 18.8% of the total errors, that is, the meaning of the sentence is still preserved. It is often the case, for example,

Type	Sub-type	E-S [%]		S-E [%]	
		FTE	Verbatim	FTE	Verbatim
Missing Words		19.9	26.4	26.0	19.6
	Content Words	7.9	9.9	7.2	4.4
	Filler Word	12.0	16.5	18.8	15.2
Word Order		15.4	11.5	20.4	21.1
	Local Word Order	11.6	4.8	12.7	13.2
	Local Phrase Order	2.1	5.5	6.0	6.9
	Long Range Word Order	1.7	1.1	0.6	1.0
	Long Range Phrase Order	0.0	0.0	1.1	0.0
Incorrect Words		64.4	61.0	50.8	57.3
	Sense	21.9	24.6	28.2	36.8
	Wrong Lexical Choice	13.0	15.4	15.5	21.1
	Disambiguation	8.9	9.2	12.7	15.7
	Incorrect Form	33.9	30.2	9.9	11.7
	Verbs				
	Incorrect Tense	15.1	13.2	7.7	7.8
	Incorrect Person	8.2	8.5	2.2	3.9
	Concordance				
	Incorrect Gender	7.5	4.8	0.0	0.0
	Incorrect Number	3.1	3.7	0.0	3.9
	Extra Words	0.0	2.9	1.1	3.9
	Style	7.9	3.3	9.9	3.9
	Idioms	0.7	0.0	1.7	0.0
Unknown Words		0.3	1.1	2.8	2.0
	Unknown Words	0.3	1.1	1.1	1.5
	Unseen Forms	0.0	0.0	1.6	0.5

Table 3: Error statistic for the English–Spanish EPPS FTE Task.

that the "to" particle of English infinitives is missing. Only in 7.2% of the cases, a content word, essential for the meaning of the sentence is missing. The complete statistics can be found in the column of FTE S-E in Table 3.

5.2.2. EPPS Verbatim data

As was the case for the English to Spanish direction, if we switch to verbatim data, the input looses in grammatical correctness. In this translation direction this is even more important, as the number of interpreters increases. This produces a distorted input and the system is not always able to produce suitable translations. We can observe this effect in the increased number of bad lexical choice errors with respect to the final text editions. We also encounter an increase in the number of extra words, which originate from the spontaneous speech effects of the input text. The other errors are quite similar to the FTE condition. The statistics can be found in the column of the Verbatim S-E in Table 3.

5.2.3. EPPS ASR data

As was the case for the reverse direction, the most important source of errors of the speech recognizer are substitution errors, amounting to a total of 58.3% of the total errors (with 25.5% deletions and 16.2% insertions). This also has the effect that the most significant differences between the output of the Verbatim and the ASR conditions are substitution errors, amounting to 57.9% of the differences. In this case however, most of the substitution errors of the recognizer are due to changes in the morphology of the words, but the base form remains the same. It is not unusual that plural and singular or masculine and feminine forms of the words are exchanged. In these cases the contextual information is not lost in the same way as for the English to Spanish direction and the proportion of errors remains nearly the same.

5.3. Chinese to English

For the Chinese to English direction, we introduce new types of reordering errors. The main difference between the two languages is the position of the modifiers, and so we distinguish three major categories related to the sentence construction. In Chinese declarative sentences, the modifiers are usually located before the predicates, and the modifier of the place/time can also be at the beginning of a sentence. In interrogative sentences the word order is generally the same as in declarative sentences, but in the Is-Question sentence, a Chinese key word "Ma" is appended to show the tone, and in the Wh-Question sentence, the question part is substituted by a word "Shenme". Lastly subordinate/infinitive sentences are placed after the main sentence in English but before the main sentence in Chinese.

5.3.1. CE FTE data

The statistics of the errors are presented in Table 4. The main source of errors are the "Wrong Lexical Choice". For the FTE translation, it contributes 27%-33% of the total errors. The second biggest error type are the "Missing Words" with 19%-28% of the total errors. The following

Type Sub-Type	FTE [%]	Verb [%]
Missing Words	27.5	20.9
Content Words	22.1	14.2
Filler Words	5.4	6.7
Word Order	17.8	17.3
Declare	10.1	10.3
Question	0.2	0
Sub-ordinate	0.7	0
Infinitive	6.8	5.9
Long Range	10.6	11.1
Local Range	7.3	6.2
Incorrect Words	27.9	32.0
Wrong Lexical Choice	18.5	25
Incorrect Form	9.4	7.0
Named Entity	8.9	10.1
Person	5.4	7.0
Location	2.6	2.1
Organization	0.7	0.8
Others	0.2	0.3
Extra Words	17.8	19.8
Content Words	5.4	10.3
Filler Words	12.4	9.5
Unknown Words	0	0

Table 4: Error statistics for the Chinese–English Tasks.

are the "Extra Words" and the "Word Order", which contribute 17.8% of the errors respectively. At last the "Named Entity Words" amount to 8.9% of the total errors. Because of the very low OOV rate as shown in Table 1, no "Unknown Words" were found in the analysis.

Unlike the EPPS translation systems, the CE translation system produces numerous "Extra Words", most of them filler words like prepositions. One reason could be that the translations are in fact very short because of the missing words, and then the system inserts filler words to make the sentence longer. Therefore we expect a reduction of "Extra Words" as the "Missing Words" problem is suitably handled.

The next error class are the errors caused by the reordering. If we calculate the translation mistakes both at the word and phrase levels, in the FTE translation 17.8% of the total errors are caused by bad reorderings, and if we only count at the word level, 20.4% words are taken as incorrect because of the wrong positions in the sentence. This is related to the difference between the WER and PER in Table 2.

We categorize the reordering types with respect to three criteria: the sentence type, the local/long range reordering and the reordering at the word/phrase level. In the statistics of Table 4, 10.6% reorderings take place in long ranges, i.e. across more than two positions, and 7.3% reorderings are in short ranges. From the statistics and the linguistic view as presented in the beginning of Section 5.3, we see that the CE translation system requires a phrase based reordering and the reordering may have different lengths of ranges.

As described in Section 3, for the Chinese-English translation we distinguish the named entity words with four categories: the person name, the location name, e.g. the city, country, the organization name and other names. Here the person name is the biggest problem, which contains 5.4% of the total errors. The translation of the location names is also an error source.

5.3.2. CE Verbatim & ASR data

In the Chinese English translation task, the verbatim data is the same as the FTE data without punctuation marks. As shown in Table 4, the order of the error classes according to their number of the errors has not changed, but in the translation of the verbatim data, the percentage of the "Missing Words" decreases from 27.5% to 20.9% and the number of the "Incorrect Words" and "Extra Words" increase. The conclusions for the analysis of the CE ASR data are similar to the conclusions presented for the EPPS task.

6. Conclusion

In this paper we presented a framework for the analysis of errors for the output of machine translation systems, and carried out a detailed analysis of the results presented by our group in the first TC-STAR evaluation. The most important class of errors is language-pair dependent, e.g. the verb tense generation for translation from English into Spanish or the word order for translation from Chinese to English.
Future work will study in more detail the relationship between the automatic evaluation measures (maybe on the level of word classes) and the error classes used in this work.

7. Acknowledgments

This work has been partly funded by the integrated project TC-STAR – Technology and Corpora for Speech-to-Speech Translation – (IST-2002-FP6-506738) and partly sponsored by the National Office of Universities and Research - Regional Education Ministry - Community of Madrid, Spain.

8. References

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. ARPA Workshop on Human Language Technology*.

LDC. 2005. Linguistic data consortium chinese training data resources. http://www.ldc.upenn.edu/Projects/TIDES/mt2005cn.htm.

Ariadna Font Llitjós, Jaime G. Carbonell, and Alon Lavie. 2005. A framework for interactive and automatic refinement of transfer-based machine translation. In *Proc. of the 10th Annual Conf. of the European Association for Machine Translation (EAMT)*, Budapest, Hungary, May.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, July.

David Vilar, Evgeny Matusov, Saša Hasan, Richard Zens, and Hermann Ney. 2005. Statistical Machine Translation of European Parliamentary Speeches. In *Proceedings of MT Summit X*, pages 259–266, Phuket, Thailand, September. Asia-Pacific Association for Machine Translation (AAMT).