Morpho-syntactic Information for Automatic Error Analysis of Statistical Machine Translation Output

Maja Popović[⊥] **Hermann Ney**[⊥]
Adrià de Gispert[†] **José B. Mariño**[†]
Deepa Gupta[⋆] **Marcello Federico**[⋆]
Patrik Lambert[†] **Rafael Banchs**[†]

[⊥] Lehrstuhl für Informatik VI - Computer Science Department, RWTH Aachen University, Aachen, Germany
[†] TALP Research Center, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
[⋆] ITC-irst, Centro per la Ricerca Scientifica e Tecnologica, Trento, Italy

{popovic|ney}@informatik.rwth-aachen.de {agispert|canton}@gps.tsc.upc.es
{gupta|federico}@itc.it {lambert|banchs}@gps.tsc.upc.es

Abstract

Evaluation of machine translation output is an important but difficult task. Over the last years, a variety of automatic evaluation measures have been studied, some of them like Word Error Rate (WER), Position Indepedent Word Error Rate (PER) and BLEU and NIST scores became widely used tools for comparing different systems as well as for evaluating improvements within one system. However, these measures do not give any details about the nature of translation errors. Therefore some analysis of the generated output is needed in order to identify the main problems and to focus the research efforts. On the other hand, human evaluation is a time consuming and expensive task. In this paper, we investigate methods for getting use of morpho-syntactic information for automatic evaluation: standard error measures WER and PER are calculated on distinct word classes and types in order to get better idea of the nature of errors and possibilities for improvements.

1 Introduction

Evaluation of the generated output is an important issue for all natural language processing (NLP) tasks, especially for machine translation (MT). Human evaluation of machine translation system is a time consuming and expensive task. This is the reason why automatic evaluation is preferred in the research community. A variety of automatic evaluation measures have been proposed and studied over the last years, some of them are shown to be a very useful tool for comparing different systems as well as for evaluating improvements within one system. The most widely used are Word Error Rate (WER), Position Independent Word Error Rate (PER), BLEU score (Papineniet al., 2002) and NIST score (Doddington, 2000). However, none of these measures give any details about the nature of translation errors. A relationship between the error measures and the actual errors found in the translation outputs is not easy to find. Therefore some analysis of the translation errors is necessary in order to define the main problems and to focus the research efforst. A framework for human error analysis and error classification has been proposed in (Vilar et al., 2006), but as well as human evaluation, this is also a time consuming task.

The goal of this work is to present a framework for automatic error analysis of machine translation output using morpho-syntactic information and to analyse the obtained results.

2 Related Work

There are many publications dealing with various automatic evaluation measures for machine translation output, some of them proposing new measures, some proposing improvements and extendings of the existing ones (Doddington, 2000; Papineniet al., 2002; Babych and Hartley, 2004; Matusov et al., 2005). Semi-automatic evaluation measures have been also investigated, for example in (Niessen et al., 2000). An automatic metric which uses base forms and synonims of the words in order to correlate better to human judgements has been proposed in (Banerjee and Lavie, 2005). However, error analysis is still rather unexplored area. A framework for human error analysis and error classification has been proposed in (Vilar et al., 2006) and a detailed analysis of the obtained results has been carried out. To the best of our knowledge, automatic error analysis based on morpho-syntactic information have not been studied so far.

3 Morpho-syntactic Information and Automatic Evaluation

As already pointed out, automatic evaluation measures such as WER, PER, BLEU and NIST are widely spread and useful tools in the machine translation research community, but they do not give any idea about the nature of actual translation errors.

We propose the use of morpho-syntactic information in combination with automatic evaluation measures WER and PER in order to get more details about the translation errors. As any other automatic evaluation measures, these novel measures will be far from perfection. Possible POS-tagging errors may introduce additional noise in the measure. However, we expect this noise to be sufficiently small and the new measures to be able to give sufficiently clear ideas about particular errors.

We investigate two types of potential problems which might occur for the translation with Spanish-English language pair:

- local reorderings involving nouns and adjectives
- inflections involving mainly verbs, adjectives and nouns

These two points can be easily identified as potential sources of errors even with a moderate linguistic knowledge about Spanish and English. Additionaly, human error analysis (Vilar et al., 2006) also identified these problems as most common ones for this language pair and corpus.

3.1 Reordering

Adjectives in the Spanish language are usually placed after the corresponding noun, whereas in English is the other way round. Although in most cases the phrase based translation system handles these local permuations correctly, some deviations are still possible. In order to investigate this type of errors, we extract only nouns and adjectives from both the reference translations and the system output, and then calculate WER and PER. If the absolute or relative difference between WER and PER is large, this indicates reordering errors.

Additionally, we also calculate noun-adjective WER and PER for the output of the translation system based on POS-based word reordering method proposed in (Popović and Ney, 2006). The difference between WER and PER for this system is expected to be smaller than for the baseline system.

3.2 Inflection

Spanish has a rich inflectional morphology, especially for verbs. Person and tense are expressed by the suffix so that many different full forms of one verb exist. Additionaly, in contrast to English, Spanish adjectives have gender and number which have to match to those of the corresponding noun. Therefore the error rates for those word classes are expected to be higher for Spanish than for English. Additionaly, the error rates for the Spanish base forms are expected to be lower than for the full forms. In order to investigate potential inflection errors, we compare the error rates for verbs, adjectives and nouns for both languages. For the Spanish language, we also investigate differences between full form PER and base form PER. The larger these differences, more inflection errors are present.

4 Experimental Settings

4.1 Task and Corpus

The corpus analysed in this work is built in the framework of the TC-Star project. It contains more than one million sentences and about 35 million running words of the Spanish and English European Parliament Plenary Sessions (EPPS). A description of the EPPS data can be found in (Vilar et al., 2005). In order to investigate effects of sparse training data, we have randomly extracted a small subset containing about 13k sentences and 370k running words (about 1% of the original corpus) The statistics of the corpora can be seen in Table 1.

Training:		Spanish	English
full	Sentences	1281427	
	Running Words	36578514	34918192
	Running Words-Punctuation	32877238	31479559
	Vocabulary	153124	106496
	Singletons [%]	35.2	36.2
reduced	Sentences	13360	
	Running Words	385198	366055
	Running Words-Punctuation	346776	330264
	Vocabulary	22425	16326
	Singletons [%]	47.6	43.7
Develop:	Sentences	1008	
	Running Words	25778	26070
	Running Words-Punctuation	22948	23171
	Distinct Words	3895	3173
	OOVs (full) [%]	0.15	0.09
	OOVs (reduced) [%]	2.7	1.7
Test:	Sentences	840	1094
	Running Words	22774	26917
	Running Words-Punctuation	20427	24104
	Distinct Words	4081	3958
	OOVs (full) [%]	0.14	0.25
	OOVs (running words) [%]	2.8	2.6

Table 1: Corpus statistics

4.2 Translation System

The statistical machine translation system used in this work is based on a log-linear combination of seven different models. The most important ones are phrase based model in both directions, additionaly the IBM1 models at phrase level in both directions as well as phrase and length penalty are used. A more detailed description of the system can be found in (Vilar et al., 2005; Zens et al., 2005).

4.3 Experiments

The translation experiments have been done in both translation directions on both corpora. In order to analyse potential improvements of the baseline system, additional experiments with POS-based word reorderings of nouns and adjectives as proposed in (Popović and Ney, 2006) have been done as well. The translation results for Spanish→English are presented in Table 2 and those for English→Spanish in Table 3.

5 Error Analysis

5.1 Reordering errors

As explained earlier, the reordering errors have been measured by the relative difference between WER and PER calculated on nouns and adjectives. In order to get a better idea about what actually happens, we have calculated these differences also for verbs as well as adjectives and nouns separately.

Table 4 presents the relative differences for the English output and Table 5 for the Spanish output. It can be seen that the WER/PER differences for nouns and adjectives are relatively high for both language pairs (more than 20%), and those for the English output are higher than for the Spanish one. This corresponds to the fact that in the Spanish language, although the adjective *usually* is placed behind the noun, this is not *always* the case. On the other side, adjectives in English are always placed before the corresponding noun. Additionally, it can be seen that the differences are higher for the reduced corpus for both outpus, indicating that the local reordering problem is more important when only small amount of training data is available, as stated in (Popović and Ney, 2006).

Furthermore, the results show that the POS-based reordering of adjectives and nouns leads to decreasing of WER/PER difference for both outputs and both corpora. Like for the overall translation errors, relative decreasing of the WER/PER difference is larger for the small corpus than for the full corpus. It can be also noted that the relative decreasing for both corpora is larger for the English output than for the Spanish one.

For the verbs, WER/PER differences are less than 5% for both outputs and both training corpora, which indicates that for this language pair the word order of verbs is not an important issue. For the adjectives and nouns separately, the WER/PER differences are higher than those for verbs, those for the nouns being significantly higher than those for adjectives. The reason for this is probably the fact that in both languages some nouns actually have adjective function, for example “export control = control de exportación”.

5.2 Inflectional errors

Tables 6 and 7 represent the PERs of different word classes for English and Spanish output respectively. It can be seen that all PERs are higher for the Spanish output than for the English one due to the rich inflectional morphology of the Spanish language. Furthermore, it can be seen that the Spanish verbs are especially problematic (as stated in (Vilar et al., 2006)) reaching 60% of PER for the full corpus and more than 70% for the reduced corpus. Spanish adjectives also have a significantly higher PER than the English ones, whereas for the nouns this difference is not so high.

Results of the further analysis of inflectional errors are presented in Table 8. Relative differences between full form PER and base form PER also show that the verb inflections are the main source of translation errors into the Spanish language. The differences for adjectives and nouns are significantly lower than those for verbs.

For the small training corpus, there is basically no difference for verbs and nouns. Since nouns in Spanish only have singular and plural form as in English, the number of unseen forms is not particulary large even for the small training corpus. On the other side, the full/base relative difference of adjectives is significantly higher for the small corpus than for the full corpus due to increased number of unseen adjective full forms.

full corpus	dev			test		
	WER	PER	BLEU	WER	PER	BLEU
baseline	33.0	24.2	57.5	34.5	25.5	54.7
reord-adj	32.4	23.9	58.3	33.5	25.2	56.4
reduced corpus	**dev**			**test**		
	WER	PER	BLEU	WER	PER	BLEU
baseline	39.2	28.4	48.7	41.8	30.7	43.2
reord-adj	37.9	28.3	50.7	38.9	29.5	48.5

Table 2: Translation Results for Spanish→English (%)

full corpus	dev			test		
	WER	PER	BLEU	WER	PER	BLEU
baseline	39.8	30.2	50.5	39.7	30.6	47.8
reord-adj	39.7	30.2	50.9	39.6	30.5	48.3
reduced corpus	**dev**			**test**		
	WER	PER	BLEU	WER	PER	BLEU
baseline	48.3	35.8	40.6	49.6	37.4	36.2
reord-adj	47.4	35.6	41.7	48.1	36.5	37.7

Table 3: Translation Results for English→Spanish (%)

$1 - \frac{PER}{WER}$	full corpus		red. corpus	
	dev	test	dev	test
nouns+adjectives	24.7	24.7	27.8	25.7
+reordering	21.6	20.8	21.2	20.1
verbs	4.9	4.1	4.9	4.6
adjectives	8.4	10.2	6.8	8.4
nouns	19.8	20.1	19.8	19.1

Table 4: Relative differences between WER and PER (%) - English output

$1 - \frac{PER}{WER}$	full corpus		red. corpus	
	dev	test	dev	test
nouns+adjectives	20.5	21.5	22.9	22.9
+reordering	18.9	20.3	20.6	19.8
verbs	3.2	3.3	3.4	3.9
adjectives	6.0	5.6	6.0	5.4
nouns	18.2	16.9	21.2	19.3

Table 5: Relative differences between WER and PER (%) - Spanish output

As for verbs, intuitively it might be expected that the number of inflectional errors for this word class also increases with reducing of the training corpus, even more than for adjectives. However, the problem of choosing the right inflection of a Spanish verb apparently is not related to the number of the unseen full forms since the number of inflectional errors is very high even when the translation system is trained on a very large corpus.

6 Conclusion

In this work, we presented a framework for the automatic analysis of translation errors for the output of machine translation systems based on the use of morpho-syntactic information. We carried out a detailed analysis of the results which has shown that the results obtained by our method correspond to those obtained by human error analysis in (Vilar et al., 2006). Additionaly, it has been shown that the improvements of the baseline system can be adequately measured as well.

This work is just a first step towards the development of linguistically-informed evaluation measures which provide partial and more specific information of certain translation problems. These measures are very important to understand what weaknesses statistical machine translation systems have, and how future models can improve on that. In our future work, we plan to investigate aspects for more detailed error analysis, for example examination of exact inflection errors for Spanish verbs. We also plan to investigate possibilities for automatic morpho-syntax based error analysis for other types of translation errors and other language pairs.

Acknowledgement

This work was partly supported by the TC-STAR project by the European Community (FP6-506738) and partially funded by Generalitat de Catalunya and European Social Fund.

References

B. Babych and A. Hartley. 2004. Extending BLEU MT Evaluation Method with Frequency Weighting. In *Proc. of the 42th Annual Meeting of the Association for Computational Linguistics (ACL)*, Barcelona, Spain, July.

G. Doddington. 2002. Automatic Evaluation of Machine Translation Quality using n-gram Co-occurence Statistics. In *Proc. of ARPA Workshop on Human Language Technology*

S. Banerjee and A. Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments In *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June.

E. Matusov, G. Leusch, O. Bender, and H. Ney. 2005. Evaluating Machine Translation Output with Automatic Sentence Segmentation. In *Proc. Int. Workshop on Spoken Language Translation (IWSLT)*, pages 148–154, Pittsburgh, PA, October.

S. Niessen, F. J. Och, G. Leusch, and H. Ney. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proc. 2nd Int. Conf. on Language Resources and Evaluation (LREC)*, pages 39–45, Athens, Greece, May.

D. Vilar, E. Matusov, S. Hasan, R. Zens, and H. Ney. 2005. Statistical Machine Translation of European Parliamentary Speeches. In *Proc. MT Summit X*, pages 259–266, Phuket, Thailand, September.

D. Vilar, J. Xu, L. F. D'Haro, and H. Ney. 2006. Error Analysis of Statistical Machine Translation Output. To appear in *Proc. 5th Int. Conf. on Language Resources and Evaluation (LREC)*, Genova, Italy, May.

K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation.. *Proc. 40th Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, July.

M. Popovi´c and H. Ney. 2006. POS-based Word Reorderings for Statistical Machine Translation. To appear in *Proc. 5th Int. Conf. on Language Resources and Evaluation (LREC)*, Genova, Italy, May.

R. Zens, O. Bender, S. Hasan, S. Khadivi, E. Matusov, J. Xu, Y. Zhang, and H. Ney. 2005. The RWTH Phrase-based Statistical Machine Translation System. In *Proc. Int. Workshop on Spoken Language Translation (IWSLT)*, pages 155-162, Pittsburgh, PA, October.

PER	full corpus		red. corpus	
	dev	test	dev	test
verbs	41.0	44.8	51.8	56.1
adjectives	28.2	27.3	38.2	38.1
nouns	22.6	23.0	39.2	31.7

Table 6: PER for different word classes - English output

PER	full corpus		red. corpus	
	dev	test	dev	test
verbs	59.5	61.4	70.4	73.0
adjectives	40.4	41.8	50.0	50.9
nouns	27.8	28.5	35.0	37.0

Table 7: PER for different word classes - Spanish output

$1 - \frac{basePER}{fullPER}$	full corpus		red. corpus	
	dev	test	dev	test
verbs	25.9	26.9	25.9	23.7
adjectives	6.2	9.3	12.8	15.1
nouns	7.5	8.4	7.4	6.5

Table 8: Relative differences between PER of base forms and PER of full forms (%) - Spanish output