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

Maja Popović[⋆] **Adrià de Gispert**[†] **Deepa Gupta**[⊥] **Patrik Lambert**[†]
Hermann Ney[⋆] **José B. Mariño**[†] **Marcello Federico**[⊥] **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 Independent Word Error Rate (PER) and BLEU and NIST scores have become 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 using of morpho-syntactic information for automatic evaluation: standard error measures WER and PER are calculated on distinct word classes and forms in order to get a better idea about the nature of translation errors and possibilities for improvements.

1 Introduction

The evaluation of the generated output is an important issue for all natural language processing (NLP) tasks, especially for machine translation (MT). Automatic evaluation is preferred because human evaluation is a time consuming and expensive task. 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), the BLEU score (Papineni et al., 2002) and the NIST score (Doddington, 2002). However, none of these measures give any details about the nature of translation errors. A relationship between these error measures and the actual errors 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 efforts. A framework for human error analysis and error classification has been proposed in (Vilar et al., 2006), but like 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 based on morpho-syntactic information.

2 Related Work

There is a number of publications dealing with various automatic evaluation measures for machine translation output, some of them proposing new measures, some proposing improvements and extensions of the existing ones (Doddington, 2002; Papineni et 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 synonyms of the words in order to correlate better to human judgements has been

Proceedings of the Workshop on Statistical Machine Translation, pages 1–6,
New York City, June 2006. ©2006 Association for Computational Linguistics

proposed in (Banerjee and Lavie, 2005). However, error analysis is still a 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. Automatic methods for error analysis to our knowledge have not been studied yet.

Many publications propose the use of morpho-syntactic information for improving the performance of a statistical machine translation system. Various methods for treating morphological and syntactical differences between German and English are investigated in (Nießen and Ney, 2000; Nießen and Ney, 2001a; Nießen and Ney, 2001b). Morphological analysis has been used for improving Arabic-English translation (Lee, 2004), for Serbian-English translation (Popović et al., 2005) as well as for Czech-English translation (Goldwater and McClosky, 2005). Inflectional morphology of Spanish verbs is dealt with in (Popović and Ney, 2004; de Gispert et al., 2005). To the best of our knowledge, the use of morpho-syntactic information for error analysis of translation output has not been investigated so far.

3 Morpho-syntactic Information and Automatic Evaluation

We propose the use of morpho-syntactic information in combination with the automatic evaluation measures WER and PER in order to get more details about the translation errors.

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

- syntactic differences between the two languages considering nouns and adjectives
- inflections in the Spanish language considering mainly verbs, adjectives and nouns

As any other automatic evaluation measures, these novel measures will be far from perfect. Possible POS-tagging errors may introduce additional noise. However, we expect this noise to be sufficiently small and the new measures to be able to give sufficiently clear ideas about particular errors.

3.1 Syntactic differences

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 is able to handle these local permutations correctly, some errors are still present, especially for unseen or rarely seen noun-adjective groups. In order to investigate this type of errors, we extract the nouns and adjectives from both the reference translations and the system output and then calculate WER and PER. If the difference between the obtained WER and PER is large, this indicates reordering errors: a number of nouns and adjectives is translated correctly but in the wrong order.

3.2 Spanish inflections

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. Spanish adjectives, in contrast to English, have four possible inflectional forms depending on gender and number. Therefore the error rates for those word classes are expected to be higher for Spanish than for English. Also, 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 PER 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 analyse effects of data sparseness, we have randomly extracted a small subset referred to as 13k containing about thirteen thousand sentences and 370k running words (about 1% of the original

Training corpus:		Spanish	English
full	Sentences	1281427	
	Running Words	36578514	34918192
	Vocabulary	153124	106496
	Singletons [%]	35.2	36.2
13k	Sentences	13360	
	Running Words	385198	366055
	Vocabulary	22425	16326
	Singletons [%]	47.6	43.7
Dev:	Sentences	1008	
	Running Words	25778	26070
	Distinct Words	3895	3173
	OOVs (full) [%]	0.15	0.09
	OOVs (13k) [%]	2.7	1.7
Test:	Sentences	840	1094
	Running Words	22774	26917
	Distinct Words	4081	3958
	OOVs (full) [%]	0.14	0.25
	OOVs (13k) [%]	2.8	2.6

Table 1: Corpus statistics for the Spanish-English EPPS task (running words include punctuation marks)

corpus). The statistics of the corpora can be seen in Table 1.

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 models in both directions, additionally IBM1 models at the 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 sizes of the corpus. In order to examine improvements of the baseline system, a new system with POS-based word reorderings of nouns and adjectives as proposed in (Popović and Ney, 2006) is also analysed. Adjectives in the Spanish language are usually placed after the corresponding noun, whereas for English it is the other way round. Therefore, local reorderings of nouns and adjective groups in the source language have been applied. If the source language is Spanish, each noun is moved behind the corresponding adjective group. If the source language is English, each adjective group is moved behind the corresponding noun. An adverb followed by an adjective (e.g. "more important") or two adjectives with a coordinate conjunction in between (e.g. "economic and political") are treated as an adjective group. Standard translation results are presented in Table 2.

Spanish→English		WER	PER	BLEU
full	baseline	34.5	25.5	54.7
	reorder	33.5	25.2	56.4
13k	baseline	41.8	30.7	43.2
	reorder	38.9	29.5	48.5

English→Spanish		WER	PER	BLEU
full	baseline	39.7	30.6	47.8
	reorder	39.6	30.5	48.3
13k	baseline	49.6	37.4	36.2
	reorder	48.1	36.5	37.7

Table 2: Translation Results [%]

5 Error Analysis

5.1 Syntactic errors

As explained in Section 3.1, reordering errors due to syntactic differences between two languages have been measured by the relative difference between WER and PER calculated on nouns and adjectives. Corresponding relative differences are calculated also for verbs as well as adjectives and nouns separately.

Table 3 presents the relative differences for the English and Spanish output. It can be seen that the PER/WER difference for nouns and adjectives is relatively high for both language pairs (more than 20%), and for the English output is higher than for the Spanish one. This corresponds to the fact that the Spanish language has a rather free word order: although the adjective usually is placed behind the noun, this is not always the case. On the other hand, adjectives in English are always placed before the corresponding noun. It can also be seen that the difference is higher for the reduced corpus for both outputs indicating that the local reordering problem

English output		$1 - \frac{PER}{WER}$
full	nouns+adjectives	24.7
	+reordering	20.8
	verbs	4.1
	adjectives	10.2
	nouns	20.1
13k	nouns+adjectives	25.7
	+reordering	20.1
	verbs	4.6
	adjectives	8.4
	nouns	19.1

Spanish output		$1 - \frac{PER}{WER}$
full	nouns+adjectives	21.5
	+reordering	20.3
	verbs	3.3
	adjectives	5.6
	nouns	16.9
13k	nouns+adjectives	22.9
	+reordering	19.8
	verbs	3.9
	adjectives	5.4
	nouns	19.3

Table 3: Relative difference between PER and WER [%] for different word classes

is more important when only small amount of training data is available. As mentioned in Section 3.1, the phrase based translation system is able to generate frequent noun-adjective groups in the correct word order, but unseen or rarely seen groups introduce difficulties.

Furthermore, the results show that the POS-based reordering of adjectives and nouns leads to a decrease of the PER/WER difference for both outputs and for both corpora. Relative decrease of the PER/WER difference is larger for the small corpus than for the full corpus. It can also be noted that the relative decrease for both corpora is larger for the English output than for the Spanish one due to free word order - since the Spanish adjective group is not always placed behind the noun, some reorderings in English are not really needed.

For the verbs, PER/WER difference is less than 5% for both outputs and both training corpora, indicating that the word order of verbs is not an important issue for the Spanish-English language pair. PER/WER difference for adjectives and nouns is higher than for verbs, for the nouns being significantly higher than for adjectives. The reason for this is probably the fact that word order differences involving only the nouns are also present, for example "export control = control de exportación".

English output		PER
full	verbs	44.8
	adjectives	27.3
	nouns	23.0
13k	verbs	56.1
	adjectives	38.1
	nouns	31.7

Spanish output		PER
full	verbs	61.4
	adjectives	41.8
	nouns	28.5
13k	verbs	73.0
	adjectives	50.9
	nouns	37.0

Table 4: PER [%] for different word classes

5.2 Inflectional errors

Table 4 presents the PER for different word classes for the 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. It can be also 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 5. Relative difference between full form PER and base form PER is significantly lower for adjectives and nouns than for verbs, thus showing that the verb inflections are the main source of translation errors into the Spanish language.

Furthermore, it can be seen that for the small cor-

Spanish output		$1 - \frac{PERb}{PERf}$
full	verbs	26.9
	adjectives	9.3
	nouns	8.4
13k	verbs	23.7
	adjectives	15.1
	nouns	6.5

Table 5: Relative difference between PER of base forms and PER of full forms [%] for the Spanish output

pus base/full PER difference for verbs and nouns is basically the same as for the full corpus. Since nouns in Spanish only have singular and plural form as in English, the number of unseen forms is not particularly enlarged by the reduction of the training corpus. On the other hand, base/full PER difference of adjectives is significantly higher for the small corpus due to an increased number of unseen adjective full forms.

As for verbs, intuitively it might be expected that the number of inflectional errors for this word class also increases by reducing the training corpus, even more than for adjectives. However, the base/full PER difference is not larger for the small corpus, but even smaller. This is indicating that the problem of choosing the right inflection of a Spanish verb apparently is not related to the number of 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 automatic analysis of translation errors based on the use of morpho-syntactic information. We carried out a detailed analysis which has shown that the results obtained by our method correspond to those obtained by human error analysis in (Vilar et al., 2006). Additionally, 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. Such measures are very important to understand what are the weaknesses of a statistical machine translation system, and what are the best ways and methods for improvements.

For our future work, we plan to extend the proposed measures in order to carry out a more detailed error analysis, for example examinating different types of inflection errors for Spanish verbs. We also plan to investigate other types of translation errors and other language pairs.

Acknowledgements

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

References

Bogdan Babych and Anthony Hartley. 2004. Extending bleu mt evaluation method with frequency weighting. In *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Barcelona, Spain, July.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgements. In *43rd Annual Meeting of the Assoc. for Computational Linguistics: Proc. Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72, Ann Arbor, MI, June.

Adrià de Gispert, José B. Mariño, and Josep M. Crego. 2005. Improving statistical machine translation by classifying and generalizing inflected verb forms. In *Proc. of the 9th European Conf. on Speech Communication and Technology (Interspeech)*, pages 3185–3188, Lisbon, Portugal, September.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. ARPA Workshop on Human Language Technology*, pages 128–132, San Diego.

Sharon Goldwater and David McClosky. 2005. Improving stastistical machine translation through morphological analysis. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, Vancouver, Canada, October.

Young-suk Lee. 2004. Morphological analysis for statistical machine translation. In *Proc. 2004 Meeting of the North American chapter of the Association for Computational Linguistics (HLT-NAACL)*, Boston, MA, May.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005. Evaluating machine translation output with automatic sentence segmentation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 148–154, Pittsburgh, PA, October.

Sonja Nießen and Hermann Ney. 2000. Improving SMT quality with morpho-syntactic analysis. In *COLING '00: The 18th Int. Conf. on Computational Linguistics*, pages 1081–1085, Saarbrücken, Germany, July.

Sonja Nießen and Hermann Ney. 2001a. Morpho-syntactic analysis for reordering in statistical machine translation. In *Proc. MT Summit VIII*, pages 247–252, Santiago de Compostela, Galicia, Spain, September.

Sonja Nießen and Hermann Ney. 2001b. Toward hierarchical models for statistical machine translation of inflected languages. In *Data-Driven Machine Translation Workshop*, pages 47–54, Toulouse, France, July.

Sonja Nießen, Franz J. Och, Gregor Leusch, and Hermann Ney. 2000. An evaluation tool for machine translation: Fast evaluation for mt research. In *Proc. Second Int. Conf. on Language Resources and Evaluation (LREC)*, pages 39–45, Athens, Greece, May.

Kishore Papineni, Salim Roukos, Todd Ward, and Wie-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.

Maja Popović and Hermann Ney. 2004. Towards the use of word stems & suffixes for statistical machine translation. In *Proc. 4th Int. Conf. on Language Resources and Evaluation (LREC)*, pages 1585–1588, Lissabon, Portugal, May.

Maja Popović and Hermann Ney. 2006. POS-based word reorderings for statistical machine translation. In *Proc. of the Fifth Int. Conf. on Language Resources and Evaluation (LREC)*, Genova, Italy, May.

Maja Popović, David Vilar, Hermann Ney, Slobodan Jovičić, and Zoran Šarić. 2005. Augmenting a small parallel text with morpho-syntactic language resources for Serbian–English statistical machine translation. In *43rd Annual Meeting of the Assoc. for Computational Linguistics: Proc. Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages 41–48, Ann Arbor, MI, June.

David Vilar, Evgeny Matusov, Saša Hasan, Richard Zens, and Hermann Ney. 2005. Statistical machine translation of european parliamentary speeches. In *Proc. MT Summit X*, pages 259–266, Phuket, Thailand, September.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *Proc. of the Fifth Int. Conf. on Language Resources and Evaluation (LREC)*, page to appear, Genova, Italy, May.

Richard Zens, Oliver Bender, Saša Hasan, Shahram Khadivi, Evgeny Matusov, Jia Xu, Yuqi Zhang, and Hermann Ney. 2005. The RWTH phrase-based statistical machine translation system. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 155–162, Pittsburgh, PA, October.