Error Analysis of Verb Inflections in Spanish Translation Output

Maja Popović and Hermann Ney

Lehrstuhl für Informatik 6 – Computer Science Department
RWTH Aachen University, D-52056 Aachen, Germany
{popovic,ney}@informatik.rwth-aachen.de

Abstract

Evaluation of machine translation output is an important but difficult task. Over the last years, various automatic evaluation measures have been studied and have become widely used. 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 possibilities for improvements. In this work, we present the results of automatic error analysis of Spanish verbs in statistical machine translation output generated by RWTH in the second TC-STAR evaluation. Different types of verb inflections referring to the mode, tense and person are defined. For each inflection type, PER-based precision and recall measures are calculated as well as corresponding F-measure. Additionally, a ratio between relative frequency and PER-based F-measure is defined in order to estimate significance of each verb inflection. Analysis based on the F-measure and the relative frequency has shown which verb inflections are the most difficult to translate and which are the most important to improve. Analysis of the PER-based precision-recall graph has indicated which inflections are tending to be translated wrongly and which are tending to replace other inflections.

1. Introduction and Related Work

The evaluation of the generated output is an important issue for machine translation (MT). Automatic evaluation is preferred because human evaluation is a time consuming and expensive task. Various automatic evaluation measures have been proposed and studied over the last years, the most widely used being Word Error Rate (WER), Position-independent word Error Rate (PER), the BLEU score (Papineni et al., 2002), the NIST score (Doddington, 2002). Various publications are dealing with extensions and improvements of these measures, e.g. (Babych and Hartley, 2004; Matusov et al., 2005). An automatic metric which uses base forms and synonyms of the words in order to correlate better to human judgements has been proposed in (Banerjee and Lavie, 2005). Semi-automatic evaluation measures have been also investigated, for example in (Nießen et al., 2000).

However, none of these measures gives any details about the nature of translation errors. Therefore some analysis of the translation output 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. A framework for automatic error analysis based on morpho-syntactic information is proposed in (Popović et al., 2006), but this work is just a first step towards an automatic error analysis of distinct word classes.

This work presents a detailed automatic analysis of verb inflection errors in Spanish output. Precision and recall measures based on the PER are defined along with the corresponding PER-based F-measure and calculated for each type of verb inflection occurring in the text. Additionally, a ratio between relative frequency and PER-based F-measure is defined in order to estimate actual significance of each verb inflection.

2. Automatic Evaluation of Syntactic and Morphological Errors

Morpho-syntactic information can be used in combination with the automatic evaluation measures WER and PER in order to get more details about the translation errors. As any other automatic evaluation measures, these new measures are far from perfect. The obtained numbers should not be taken as absolutely correct and precise - possible POS-tagging errors may introduce additional noise. However, this noise seems to be sufficiently small for the new measures to give sufficiently clear information about particular errors.

Errors caused by syntactic differences between two languages can be measured by difference between WER and PER: the larger this difference, more reordering errors are present. For example, one of the problems for the translation between Spanish and English is a different word order of nouns and adjectives: in the Spanish language, adjectives are usually placed after the corresponding noun whereas in English is the other way round. Possible errors due to these differences can be measured by the relative difference between WER and PER of noun-adjective groups: if this difference is large, this indicates reordering errors - there is a number of nouns and adjectives translated correctly but in the wrong order.

Inflectional errors can be measured by difference between PER for full forms and PER for base forms: the larger this difference, more inflection errors are present. For the English translation output, this type of errors has little importance since English morphology is not particularly rich. For the outputs of morphologically rich languages as for example Spanish, this type of errors can be problematic. For example, Spanish adjectives, in contrast to English, have four possible inflectional forms depending on gender and number. The verbs have even richer inflectional morphology: one base form can have up to about fifty different inflected forms.

3. Automatic Evaluation of Spanish Verbs

As already pointed out, the Spanish language has a rich inflectional morphology, especially for verbs. Person and tense are expressed by the suffix, in some cases variations of the stem are also present, so that many different full forms of one verb exist. The pronoun is often omitted since this information is contained in the suffix. Apart from that, some modes and tenses in Spanish have no direct equivalent in English. For example, subjunctive mode does not exist at all in English, and different Spanish past tenses might correspond to different English tenses but without any rule. Therefore translation of the verb word class is difficult if Spanish is the target language. Table 1 shows examples of two Spanish translations of the English verb "to need" in the present tense of the first person plural: one should be translated as the indicative mode and the other one into the subjunctive mode.

English	Spanish
That principle is that **we need** to use...	Ese principio consiste en que **necesitamos** usar...
The need means that **we need** more...	La necesidad implica que **necesitemos** más...

Table 1: Examples of the indicative and subjunctive mode: the same form of the English verb might correspond to the two different Spanish forms, the second one (subjunctive mode) not having a direct English equivalent

Some examples of the differences between past tenses is shown in Table 2. The same past tense third singular form of the English verb "to say" can be translated as three different Spanish tenses: the first example is the past tense composed using the auxiliary verb and a past participle, the second one is the perfect tense in the indicative mode and the third one is the imperfect tense in the subjunctive mode.

English	Spanish
Mr Voggenhuber **said** that...	Sr. Voggenhuber **ha dicho** que...
As Mrs Roure **said**...	Como **dijo** la Sra. Roure...
...whatever Mr Barosso **said**...	independientemente de lo que **dijera** el Sr. Barroso...

Table 2: Examples of the different past tenses: the same form of the English verb might correspond to the three different Spanish forms, the third one (subjunctive mode) not having a direct English equivalent

3.1. Inflections

In order to examine details of the verb inflectional errors, PER-based precision and recall measure as well as PER-based F-measure are calculated for each type of verb inflection. Inflection types are defined by combination of mode and tense and by person. Ten types of mode-tense combinations which are occuring in the corpus are analysed and four types of person (second singular and second plural occur extrremely rarely in the corpus).

Additionally, for each inflection type a ratio between relative frequency and PER-based F-measure is calculated. The reason for this is the following: if the F-measure of certain inflection type is low, it means that it is difficult to be translated correctly; however, if this inflection type rarely occurs in the corpus, it is not very important to translate it correctly. For the frequent inflection types, the reasoning is the other way round: even if the F-measure is relatively high, small improvements might be significant.

PER-based recall (perR) for the verb inflection type Vx is defined as follows:

- reference: all verb forms of the type Vx are extracted from the translation reference;

- hypothesis: each verb whose base form occurs in the corresponding translation reference sentence is extracted from the translation hypothesis;

- PER is calculated and subtracted from 1: this is the PER-based recall perR;

The perR measure indicates the percentage of verb inflections Vx in the translation reference which are found.

PER-based precision (perP) for the verb inflection type Vx is defined as follows:

- reference: all verb forms of the type Vx are extracted from the translation hypothesis;

- hypothesis: each verb whose base form occurs in the corresponding translation hypothesis sentence is extracted from the translation reference;

- PER is calculated and subtracted from 1: this is the PER-based precision perP;

The perP measure indicates the percentage of inflections Vx in the translation hypothesis which are translated correctly.

PER-based F-measure (perF) is defined as the standard F-measure, i. e. as harmonic mean of precision and recall:

$$perF = HM(perR, perP) = \frac{2 \cdot perR \cdot perP}{perR + perP} \quad (1)$$

The perF measure indicates the difficulty of correct translation for certain inflection type Vx - the higher this measure, the less problematic is the translation.

It should be noted that these measures are not strictly equivalent to the standard precision, recall and F-measure since the translation process is not an 1:1 mapping - the number of words (verbs) in the hypothesis is [often] not equal to the number of words (verbs) in the reference.

4. Experimental Settings

4.1. Corpus

The EPPS training corpus used for this TC-STAR evaluation is the same we used for the previous evaluation, extended with the data corresponding to the period between December 2004 and May 2005. The text analysed in this work is the Final Text Edition (FTE) version of the test corpus. The statistics can be found in Table 3.

		Spanish	English
Train	Sentences	1 167 627	
	Running Words+PM	35 320 646	33 945 468
	Vocabulary	159 080	110 636
	Singletons [%]	39.6	41.7
Test	Sentences	1 782	1 117
	Running Words+PM	56 468	28 492
	Distinct Words	7 204	4 172
	OOV Words [%]	0.6	0.2

Table 3: EPPS corpus statistics (PM = punctuation marks)

4.2. Translation system

The statistical machine translation system used in this work models the translation probability directly using a log-linear model (Och and Ney, 2002) with seven different models and corresponding scaling factors. The most important models are phrase based models in both directions, and also 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). Additionally, POS-based reorderings (Popović and Ney, 2006) of the source languages are applied as a preprocessing step, both in the training phase before the alignment computation, and before the translation of the test corpus. Additional local reorderings are applied in the style of (Kanthak et al., 2005) during the search process.

The translation output analysed in this work is generated using two pass approach. First, lists of the n best translation candidates are generated, then additional rescoring models on these generated hypotheses are applied in order to extract the final translation. The most important models used for rescoring are the IBM1 model and additional language models.

4.3. Translation Results

Table 4 presents standard translation results, i.e. WER, PER, the BLEU and the NIST score of the analysed Spanish output obtained in the second TC-STAR evaluation.

WER	PER	BLEU	NIST
39.8	30.5	49.4	10.16

Table 4: Translation Results [%] for the Spanish output

5. Error Analysis

5.1. Comparison with other word classes

Table 5 presents the relative difference between full form PER and base form PER desribed in Section 2 and in (Popović et al., 2006) for the three main open word classes: verbs, adjectives and nouns. It can be seen that this difference is significantly lower for adjectives and nouns than for verbs, thus confirming that the verb inflections are the main source of translation errors into the Spanish language.

word class	$1 - \frac{PERb}{PERf}$
verb	21.0
adjective	3.7
noun	3.2

Table 5: Relative difference between PER of base forms and PER of full forms [%] for verbs, adjectives and nouns

5.2. Detailed error analysis of verb inflections

In Table 6 PER-based F-measure perF is presented together with the relative frequency and the ratio between the relative frequency and perF. It can be seen that the inflection types with the lowest perF are those corresponding to the subjunctive mode in both present and imperfect tense as well as the indicative mode in the imperfect tense. This indicates that these three verb inflection types are most difficult to be translated correctly. This could also be expected considering the diversity of past tenses in Spanish, lack of straightforward correspondence to the English tense, as well as absence of the subjunctive mode in English. However, the imperfect tense inflections occur very rarely in the corpus. Therefore they do not have significant importance even though their perF is very low.

type of verb inflection		$perF$	rel. freq.	$\frac{rel.freq.}{perF}$
mode and tense	indicative present	53.8	37.5	**69.7**
	infinitive	44.9	25.1	**55.9**
	past participle	35.1	12.5	35.6
	subjunctive present	**14.8**	6.9	**46.6**
	indicative perfect	35.7	5.6	15.7
	indicative future	34.7	3.8	10.9
	present participle	29.8	3.4	11.4
	conditional	40.2	2.7	6.7
	indicative imperfect	**21.1**	0.8	3.8
	subjunctive imperfect	**11.2**	0.7	6.2
person	first singular	53.7	5.2	9.6
	first plural	50.4	9.6	10.4
	third singular	45.5	29.4	64.6
	third plural	36.1	13.8	38.2

Table 6: PER-based F-measure (perF) [%], relative frequency [%] and relative frequency/perF ratio for different types of verb inflections

On the other hand, the indicative present tense and the infinitive form have relatively high perF which means that the

system is capable of translating these inflections correctly in most cases. Nevertheless, it would be reasonable to improve the translation quality of these categories since they are the most frequent types of verb inflections. Improving the subjunctive present tense also would be reasonable because it occurs sufficiently often and has a very low perF.

As for person-based inflections, third person plural seems to be most difficult for translation and the third person singular to be most significant. However, the latest requires some further analysis, because some modes and tenses in the third singular actually have the same inflection form as in the first singular.

It should be noted that a potential improvement for certain inflection type cannot be independent of the other types. For example, if the system can be improved so that the subjunctive present form is more often translated correctly thus becoming less often replaced with the indicative present form, the translation quality of both forms will be improved. In the similar way, if the infinitive form can be better translated so that some other inflection types are less often replaced with it, the quality of both the infinitive form and the inflections in question will be improved. The exact methods for improving the translation quality of verb inflections can be defined after a deeper analysis of the precision and recall measure, as well as after examination of confusion tables.

PER-based precision and recall graph for all the inflection types analysed in this work is showed in Figure 1. Precision and recall for the most frequent type of verb inflection i.e. the indicative present form are both more than 50%. The value of perR indicates that this verb inflection is translated correctly in almost 60% cases, and the perP that in almost 50% cases some other verb form is wrongly translated as this one. The infinitive form shows a similar tendence but with slightly lower precision. The subjunctive present form has both measures rather low, recall being significantly lower than precision. The numbers are indicating that this form is discovered correctly in only about 10% cases and it is very often replaced by some other form - in more than 80% cases. Intuitively it might be expected that it is most often translated as the indicative present form, but a confusion table analysis should be carried out before deriving any conclusion.

Another interesting point is that the difference between recall and precision for the inflection type with the lower F-measure, i.e. the subjunctive imperfect form is very high - very low recall indicates that this verb form is very difficult to be translated correctly. However, relatively high precision indicates that in a number of (rare) cases when the system decides that this type of verb inflection is the appropriate translation, the decision is correct.

6. Conclusions

In this work, a detailed analysis of translation errors of Spanish verbs based on automatic evaluation measures is presented. Analysis of the PER-based F-measure and the relative frequency of different inflection types based on the mode, tense and person has shown which verb inflections are the most difficult to translate and which are the most important to improve. Analysis of the PER-based precision-

recall graph has indicated which inflections are tending to be translated wrongly and which are tending to replace other inflections.

For our future work, we plan to introduce and investigate confusion tables in order to discover more details about particular verb inflection errors. We also plan to investigate errors of verb groups, like for example composed past tense being formed with the auxiliary verb and the past participle.

7. References

B. Babych and A. 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.

S. Banerjee and A. 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.

G. 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.

S. Kanthak, D. Vilar, E. Matusov, R. Zens, and H. Ney. 2005. Novel reordering approaches in phrase-based 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 167–174, Ann Arbor, MI, June.

E. Matusov, G. Leusch, O. Bender, and H. 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.

S. Nießen, F. J. Och, G. Leusch, and H. 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.

F. J. Och and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 295–302, Philadelphia, PA, July.

K. Papineni, S. Roukos, T. Ward, and W. J. 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.
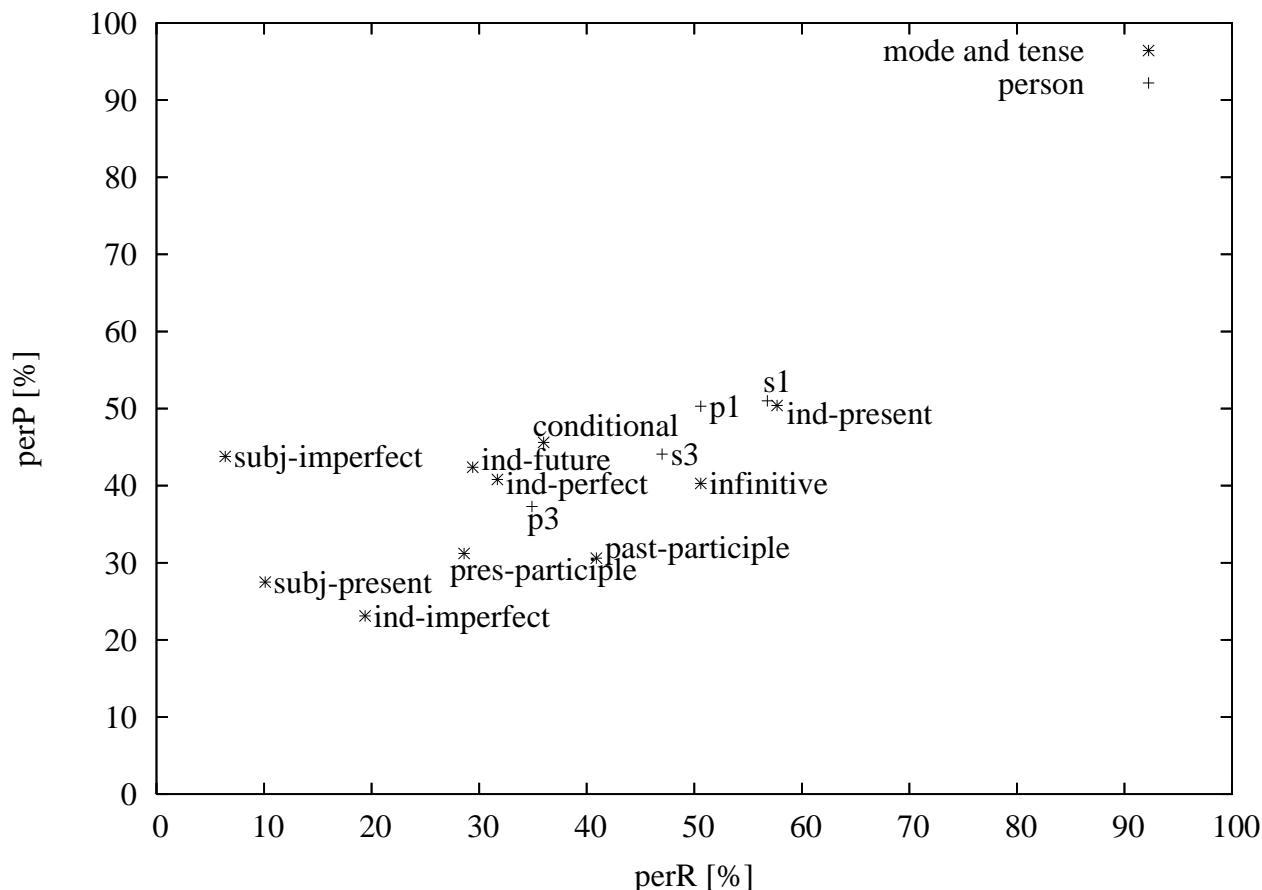
Figure 1: PER-based precision-recall graph for different types of verb inflections

M. Popović and H. Ney. 2006. POS-based word reorderings for statistical machine translation. In *Proc. of the Fifth Int. Conf. on Language Resources and Evaluation (LREC)*, pages 1278–1283, Genoa, Italy, May.

M. Popović, A. de Gispert, D. Gupta, P. Lambert, H. Ney, J. B. Mariño, M. Federico, and R. Banchs. 2006. Morpho-syntactic information for automatic error analysis of statistical machine translation output. To appear in *Proc. of the HLT-NAACL Workshop on Statistical Machine Translation*, New York, NY, June.

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. In *Proc. of the Fifth Int. Conf. on Language Resources and Evaluation (LREC)*, pages 697–702, Genoa, 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 *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 155–162, Pittsburgh, PA, October.