Linguistic-based Evaluation Criteria to identify Statistical Machine Translation Errors

Mireia Farrús*, Marta R. Costa-jussà, José B. Mariño* and José A.R. Fonollosa***
*Universitat Politècnica de Catalunya, TALP Research Center
C/Jordi Girona 1-3, 08034 Barcelona, Spain
{mfarrus,canton,adrian}@gps.tsc.upc.edu
** Barcelona Media Innovation Center
Av. Diagonal 177, 08018 Barcelona, Spain
marta.ruiz@barcelonamedia.org

Abstract

Machine translation evaluation methods are highly necessary in order to analyze the performance of translation systems. Up to now, the most traditional methods are the use of automatic measures such as BLEU or the quality perception performed by native human evaluations. In order to complement these traditional procedures, the current paper presents a new human evaluation based on the expert knowledge about the errors encountered at several linguistic levels: orthographic, morphological, lexical, semantic and syntactic. The results obtained in these experiments show that some linguistic errors could have more influence than other at the time of performing a perceptual evaluation.

1 Introduction

One of the aims in the research community is to find accurate evaluation methods that allow analyzing and comparing the performance of these translation systems. The most commonly used evaluation methods are the standard automatic measures such as BLEU (Papineni et al., 2002), NIST (Doddington, 2002), TER (Snover and Dorr, 2006) and WER (McCowan, 2004 et al.), as well as the use of human native evaluators that analyze and compare translated sentences according to a general perception of the linguistic quality.

In this paper, these evaluation methods are used to evaluate and compare two translation systems based on the statistical approaches in the Catalan-to-Spanish language pair: Google Translate and N-II; this one developed at the Universitat Politècnica de Catalunya (UPC).

In addition, a new human evaluation method is applied, based on an expert linguistic evaluation, which provides information about the errors classified according the level they are encountered: orthographic, morphological, lexical, semantic and syntactic. The number of errors found in each level is then used to compare both human evaluations: linguistic and perceptual. Since the aim is to achieve a good human perception in our final translation, one of the main points is to see which linguistic errors have more impact in the human evaluation.

The structure of this paper is as follows. Next section presents a brief summary of the related work. Section 3 presents an overview of the statistical machine translation approach. Section 4 includes the description of the systems and the human evaluations used in the experiments. Section 5 shows the results obtained in each of the evaluations, and finally, conclusions are presented in section 6.

2 Related work

Automatic and human evaluation has been widely investigated by the scientific community. Having an automatic evaluation is a must in order to optimize a MT system. Actually, there are many interesting measures, for example the ones which have been presented and evaluated in the Annual Workshop of Machine Translation (WMT) [1]. Some measures include linguistic knowledge and do correlate with human criteria. However, as mentioned in the introduction, in this area, BLEU is still the most widely used measure by most MT research

[1] http://www.statmt.org/wmt09/

groups. Some of the main problems in automatic evaluation are that: the measure depends on the quality of the references; and, the measure do not behave objectively among different types of MT translation systems. Given that a source sentence may have multiple correct target sentences, it is difficult to compose a test set which covers all of them.

Human evaluation is time consuming. One of the main problems here is that the criteria changes for each annotator. People do not have the same criteria when evaluating or ranking one translation. Recently, in the GALE project, one effective way to evaluate was asking annotators to edit the translation. In that sense, the less number of editions, the better the translation. In (Callison-Burch, 2009), they proposed to edit the translation output as fluent as possible which reflects the annotators' understanding of the sentence.

Apart from the inconveniences mentioned above, both automatic and human evaluation provide little information about the linguistic errors committed by the system, which would help further research. In this paper, we propose a linguistic evaluation which aims at being objective over any translation output and at specifying the type of errors committed by the system in order to help MT developers to improve it.

Some proposals regarding evaluation classification schemas can be found in the literature. (Vilar et al., 2006), for instance, propose a 5-category schema that does not use linguistic criteria. The classification presented in the current paper offers more linguistic information about the type of error; e.g. (Vilar et al., 2006) use the concept of *incorrect words* that can be related to multiple linguistic levels: lexical, semantic and morphological. On the other hand, Flanagan classification (Flanagan, 1994) lists a series of errors that are pair language-dependent. In the current paper, a similar list of subcategories for Catalan-Spanish is presented. However, these subcategories are included in a 5-category schema, which is language-independent.

3 Statistical Machine Translation

Nowadays, Statistical Machine Translation (SMT) has become one of the most popular machine translation paradigms. The SMT approach allows building a translation system by means of open-source tools as long as a parallel corpus is available. Moreover, one of the most attractive reasons to build a statistical system is that, unlike standard rule-based system, little human effort is required.

In SMT, statistical weights are used to decide the most likely translation of a word. Modern SMT systems are phrase-based rather than word-based, and assemble translations using the overlap in phrases. Thus, given a source string $s_1^J = s_1 \dots s_j \dots s_J$ to be translated into a target string $t_1^I = t_1 \dots t_i \dots t_I$, the aim is to choose, among all possible target strings, the string with the highest probability:

$$\tilde{t_1^I} = \underset{t_1^I}{argmax}\, P(t_1^I|s_1^J)$$

where I and J are the number of words of the target and source sentence, respectively.

The first SMT systems were reformulated using Bayes' rule. In recent systems, such an approach has been expanded to a more general maximum entropy approach in which a log-linear combination of multiple feature functions is implemented (Och, 2003). This approach leads to maximising a linear combination of feature functions:

$$\tilde{t} = \underset{t}{argmax} \left\{ \sum_{m=1}^{M} \lambda_m h_m(t,s) \right\}.$$

Given a target sentence and a foreign sentence, the translation model tries to assign a probability that t_1^I generates s_1^J. While these probabilities can be estimated by thinking about how each individual word is translated, modern statistical MT is based on the intuition that a better way to compute these probabilities is by considering the behavior of phrases (sequences of words). The intuition of phrase-based statistical MT is to use phrases as well as single words as the fundamental units of translation. Phrases are estimated from multiple segmentation of the aligned bilingual corpora by using relative frequencies.

The translation problem has also been approached from the finite-state perspective as the most natural way for integrating speech recognition and machine translation into a speech-to-speech translation system (Vidal, 1997; Bangalore and Riccardi, 2001; Casacuberta, 2001). The Ngram-based system implements a translation model based on this finite-state perspective (de Gispert and Mariño, 2002) which is used along with a log-linear combination of additional feature functions (Mariño, 2006 et al.).

In addition to the translation model, SMT systems use the language model, which is usually formulated as a probability distribution over strings

that attempts to reflect how likely a string occurs inside a language (Chen and Goodman, 1998). Statistical MT systems make use of the same n-gram language models as do speech recognition and other applications. The language model component is monolingual, so acquiring training data is relatively easy.

The lexical models allow the SMT systems to compute another probability to the translation units based on the probability of translating word per word of the unit. The probability estimated by lexical models tends to be in some situations less sparse than the probability given directly by the translation model. Many additional feature functions can also be introduced in the SMT framework to improve the translation, like the word or the phrase bonus.

Although SMT systems provide, in general, good performance, it has been demonstrated in recent papers that the addition of linguistic information can be highly useful in this kind of systems (Niessen and Ney, 2000; Popović and Ney, 2004; Popović and Ney, 2006; Popović et al., 2006).

4 Experimental Framewok

Machine translation systems can be evaluated by means of human judgments in many different ways. The main objective of this work is to utilize three kinds of evaluations (automatic, perceptual and linguistic) and see whether they are somehow correlated or not. The three evaluations have been performed over two SMT systems: Google and N-II. This section includes an overview of both systems and a brief description of the human evaluations used in the current work.

4.1 Systems Description

Google Translate[2] has been developed by Google's research group on multiple pairs of languages. This system feeds the computer with billions of text words, including monolingual text in the target language, as well as aligned text consisting of examples of human translations between the languages. Then, statistical learning techniques are applied in order to build a translation model. The accuracy of the automatic language detection increases with the amount of text entered.

Google is constantly working to support more language in order to introduce them as soon as the automatic translation meets their standards. Large amounts of bilingual texts are needed to further develop new systems.

N-II[3], developed at the UPC mainly for the Spanish-Catalan pair, is an engine based on an N-gram translation model integrated in an optimized log-linear combination of additional features. Although it is mainly statistical, additional linguistic rules are included in order to solve some errors caused by the statistical translation, such as ambiguity in adjective and possessive pronouns, orthographic errors or time expressions, among others.

Time expressions, which differ largely in both languages, are solved by detecting them, codifying them as numeric expressions, and generating them in the target language (Farrús, 2004 et al.). The same procedure is used in the numbers, since many of them were not included in the training corpus. Other unknown words apart from numbers are solved by including a dictionary as a post-process after the translation, and a spell checker in order to avoid wrong-written words in the input.

4.2 Perceptual and Linguistic Evaluations

Human evaluations of the systems can be performed in different ways. The most commonly used, is the one called *perceptual* in the current paper. It consists in selecting a reasonable number of evaluators, which are not necessary linguistic experts but having a good knowledge of the language in question. Such evaluators are then asked to compare translations output by two or more systems. In addition, another human evaluation is presented in this paper, consisting of a linguistic analysis made by an expert linguist. Next, both evaluations are briefly described.

4.2.1 Perceptual Evaluation

The comparison between different translation system outputs was performed by ten different human evaluators. All of them were bilingual in both Catalan and Spanish languages, therefore no reference of translation was shown to them, in order to avoid any bias in their evaluation.

Each evaluator was asked to make a system-to-system (pairwise) comparison, where the system pairs were randomized, so that the evaluator did not know which system was being judged. Each judge evaluated 100 randomly extracted translation pairs, and assessed, in each case, whether

[2]http://translate.google.com

[3]http://www.n-ii.org/

one system produced a better translation than the other one, or whether both outputs were equivalent. Therefore, a total number of 1000 judgements was collected. Next, an example of an output shown to the evaluators is presented:

> Source: Cal que hi hagi oferta per a tothom.
> (1): Hace falta que haya ofrecida para todo el mundo.
> (2): Es necesario que haya oferta para todos.
> Which translation was better? (Type 0 for same quality)

4.2.2 Linguistic Evaluation

In order to evaluate the translations by means of linguistic criteria, rather than using only the common knowledge of the language speakers, a linguistic error classification was proposed in order to linguistically evaluate the encountered errors. The error-annotation process was very time consuming. Since the linguistic evaluation guidelines were very specific, only one evaluator was required, and no inter-annotator agreement was needed. The system order was randomized, so that the annotator did not know which system was being judged.

The errors are reported according to the different linguistic levels involved: orthographic, morphological, lexical, semantic and syntactic, and according to the specific cases that can be found in a Catalan to Spanish (and vice versa) translation task.

The annotation guidelines are described in detail in a journal paper[4] submitted and pending of acceptance at the time of writting this paper. The guidelines include a detailed description of the linguistic levels, providing the kind of errors that could be encountered in each level, and giving examples of each of them.

Next, the annotation guidelines are summarized. For each linguistic level, the most common errors encountered for the Spanish-Catalan pair are briefly described.

- **Orthographic errors** include punctuation marks, erroneous accents, letter capitalisation, joined words, spare blanks coming from a wrong detokenisation, apostrophes, conjunctions and errors in foreign words.

  An apostrophe error, for instance, can be seen in the following example, where the pronoun in Spanish is not apostrophized in Catalan as it should be:

  Source (es): *la acepta*.
  Incorrect T (ca): **la acepta*.
  Correct T (ca): *l'accepta*.

- **Morphological errors** include lack of gender and number concordance, apocopes, errors in verbal morphology (inflection) and lexical morphology (derivation and compounding), and morphosyntactic changes due to changes in syntactic structures.

  The next example shows an error regarding lack of gender concordance. The gender of the feminine Spanish term *señal* (signal) must be translated into a masculine term in Catalan:

  Source (es): *la señal*.
  Incorrect T (ca): **la senyal*.
  Correct T (ca): *el senyal*.

- **Lexical errors** include no correspondence between source and target words, non-translated source words, missing target words, and non-translated proper nouns or translated when not necessary.

  The next example shows a non-translated word in the source language:

  Source (es): *el número dieciséis*.
  Incorrect T (ca): *el número *dieciséis*.
  Correct T (ca): *el número setze*.

- **Semantic errors** include polysemy, homonymy, and expressions used in a different way in the source and target languages.

  Next, an example of an homonymy problem is shown. The word *solo* in Spanish can be an adverb or an adjective. In the Catalan translation, the wrong category was chosen: it was translated as an adjective when in that context should have been taken as an adverb:

  Source (es): *era solo un niño*.
  Incorrect T (ca): *era *sol un nen.*.
  Correct T (ca): *era només un nen*.

- **Syntactic errors** include errors in prepositions, errors in relative clauses, verbal periphrasis, clitics, missing or spare article in front of proper nouns, and syntactic element reordering.

  Next, two examples regarding syntactic errors are presented. The first one shows a wrong combination of a pronominal clitic with the verb. The second one shows an error in the

[4] Overcoming statistical machine translation limitations: error analysis and proposed solutions for the Catalan-Spanish language pair.

translation of a relative clause involving the relative pronoun *cuyo*.

Source (es): *quiero verte*.
Incorrect T (ca): *vull veure *et*.
Correct T (ca): *vull veure't*.

Source (es): *un pueblo cuyo nombre es largo*.
Incorrect T (ca): *un poble *amb un nom és llarg*.
Correct T (ca): *un poble el nom del qual és llarg*.

5 Evaluation Results

This section shows the results obtained in the automatic evaluation and in both human evaluations described above: the perceptual-non-expert evaluation, and the linguistic-expert evaluation.

The test set selected for the current evaluation consists as follows. The Spanish source test corpus consists of 711 sentences extracted from *El País* and *La Vanguardia* newspapers, while the Catalan source test corpus consists of 813 sentences extracted from the *Avui* newspaper plus transcriptions from the TV program *Àgora*. For each set and each direction of translation, two manual references were provided. Table 1 shows the number of sentences, words and vocabulary used for each language.

	Spanish	**Catalan**
sentences	711	813
words	15974	17099
vocabulary	5702	5540

Table 1: Corpus statistics for the Catalan-Spanish test.

5.1 Automatic Evaluation Results

Table 2 presents the results obtained by using two standard measures: BLEU and TER, for both systems and both directions of translation: Spanish to Catalan (es2ca) and Catalan to Spanish (ca2es). BLEU (Bilingual Evaluation Understudy) computes lexical matching accumulated precision for n-gram up to length four, while TER (Translation Error Rate) measures the number of edits required to change a system output into one of the references.

5.2 Perceptual Evaluation Results

Table 3 presents the results obtained in the perceptual evaluation, for both systems and both directions of translation: Spanish to Catalan (es2ca) and Catalan to Spanish (ca2es).

	es2ca		**ca2es**	
Errors	**Google**	**N-II**	**Google**	**N-II**
BLEU	86.10	86.54	92.37	88.58
TER	11.32	10.76	5.70	7.80

Table 2: Automatic evaluation measures for both statistical systems and both directions of translation.

direction	**Google**	**N-II**
es2ca	48%	52%
ca2es	53%	47%

Table 3: Human judgments after the system-to-system comparison, showing in which percentage each system was found better than the other one.

The results show in which percentage each of the systems was perceived as better than the other one by the human evaluators. Thus, in the es2ca direction of translation, the performance of the N-II translation system was perceived as better than the performance of Google. On the other hand, opposite results were found in the ca2es direction of translation, where the Google system performed better than the N-II one in terms of the evaluators perception. All the results obtained in this evaluation seem to be consistent with the results obtained in the automatic evaluation.

5.3 Linguistic Evaluation Results

The results found in the linguistic evaluation are shown in Table 4. It can clearly be seen that, in the es2ca translation, the N-II system performance outperformed largely the Google performance: the latter doubled the N-II in the total number of errors. This is consistent with the results obtained in the perception evaluation, where N-II was found better than Google in 52% of the cases. The same consistency is found in the automatic evaluation, where the BLEU and TER in the NII sytem equal 86.54 and 10.76, respectively, slightly better than in the Google system, where BLEU and TER equal 86.10 and 11.32, respectively.

Nevertheless, and despite these consistencies, the difference of quality between both systems in the linguistic evaluation is not reflected neither in the perceptual evaluation, nor in the automatic evaluation. In both perceptual and automatic evaluations the difference of performance quality is smaller. They are mutually consistent and, in con-

sequence, they differ from the linguistic evaluation in the same way.

In the opposite direction of translation (ca2es), the evaluation results differ from the es2ca translation: N-II outperforms Google translator only in three linguistic levels: orthographic, morphological and syntactic. In the lexical and the semantic domains, the Google system outperforms the N-II translator.

	es2ca		ca2es	
Errors	**Google**	**N-II**	**Google**	**N-II**
orthographic	169	62	102	82
morphological	80	29	40	37
lexical	113	65	54	67
semantic	101	61	50	65
syntactic	183	79	111	87
total	646	295	357	338

Table 4: Number and type of errors encountered in both systems and both directions of translation.

The total number of errors is similar in both translation systems, although it is slightly lower in the N-II system (338 in front of 357 in the Google system). Nevertheless, the perceptual evaluation is not consistent with these results, since the evaluators judged the Google performance as better than the N-II performance in 53% of the cases.

The presented results can be interpreted in different ways. First, it seems that some linguistic errors have more influence than others at the time of performing a perceptual evaluation, and that the lexical and semantic errors (which are, in turn, highly related) could have a higher weight. Second, that human evaluations do not have a mutual and real consistency, and thus, they are highly independent from each other, since the evaluators may not rely on any specific linguistic error level when performing the evaluations.

Thus, it seems that further experiments by using other corpora, other languages and other translation approaches should be performed in order to see whether a real correlation exists between all the evaluation methods included. Nevertheless, the proposed linguistic human-expert evaluation gives more detailed information regarding the type of errors occurred. Therefore, a more specific starting point is provided in order to improve the translation system in the future.

6 Conclusions

In this paper, a new evaluation method has been proposed in order to evaluate two statistical machine translation systems. System evaluation is a decisive task when trying to improve a system of such characteristics. Therefore, a lot of effort has been put into trying to find the best or the most accurate and consistent evaluation method.

The evaluation procedure proposed in this paper takes into account the type of errors encountered in each system, by classifying them into different linguistic levels: orthographic, morphological, lexical, semantic and syntactic. When comparing the results obtained through this classification to the ones obtained by performing a traditional human evaluation, it could be stated that some levels (the lexical and the semantic levels) have more influence in the way how the human evaluators perceive the errors. In the same way, both lexical and semantic errors seem to be also consistent with the automatic evaluation measures BLEU and TER.

Nevertheless, the experiments in the current paper where only carried out within one pair of languages (Spanish-Catalan). Further experiments should be performed in order to analyze more accurately this possible correlation and whether exists or not a dependency with the languages used in the translation.

Acknowledgments

The N-II machine translation system developed at the UPC has been funded by the European Union under the integrated project TC-STAR: Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738), the Spanish Government under the BUCEADOR project (TEC2009-14094-C04-01), and partially funded by the Spanish Department of Education and Science through the *Juan de la Cierva* fellowship program.

References

Bangalore, Srinivas, and Giuseppe Riccardi. 2001. Finite-state models for lexical reordering in spoken language translation. *Proceedings of the ICSLP*, 4:422–425, Beijing, China.

Callison-Burch, Chris, Philipp Koehn, Christof Monz and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 1–28, Athens, Greece.

Casacuberta, Francisco. 2001. Finite-state transducers for speech-input translation. *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 375–380, Trento, Italy.

Chen, Stanley F., and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. *Technical Report TR-10-98*, Harvard University.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Proceedings of the HLT-NAACL*, 138–145, San Diego.

Farrús, Mireia, Marta R. Costa-jussà, Marc Poch, Adolfo Hernández, and José B. Mariño. 2009. Improving a Catalan-Spanish Statistical Translation System using *Proceedings of the EAMT*, 52–57, Barcelona.

Flanagan, Mary A. 2002. Error classification for MT evaluation. *Proceedings of the AMTA Conference*, 65–72, Columbia, Maryland.

de Gispert, Adrià, and José B. Mariño. 2002. Using X-grams for speech-to-speech translation. *Proceedings of the ICSLP*, 1885–1888, Denver, Colorado.

Mariño, José B., Rafael E. Banchs, Josep M. Crego, Josep M., Adrià de Gispert, Patrick Lambert, José A.R. Fonollosa and Marta R. Costa-jussà. 2006. N-gram Based Machine Translation. *Computational Linguistics*, 32:4:527–549.

McCowan, Iain, Darren Moore, John Dines, Daniel Gatica-Perez, Mike Flynn, Pierre Wellner and Hervé Bourlard. 2004. On the Use of Information Retrieval Measures for Speech Recognition Evaluation. *Technical Report of the IDIAP*, 73, Martigny, Switzerland.

Niessen, Sonja, and Hermann Ney. 2000. Improving SMT quality with morpho-syntactic analysis. *Proceedings of the International conference on Computational Linguistics*, Saarbrücken, Germany.

Och, Franz Josef. 2003. Minimum Error Rate Training in Statistical Machine Translation. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 160–167, Sapporo, Japan.

Papineni, Kishore, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Meeting of the Association for Computational Linguistics*, 311–318, Philadelphia, Pennsylvania.

Popović, Maja, and Hermann Ney. 2004. Towards the Use of Word Stems and Suffixes for Statistical Machine Translation. *Proceedings of International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Popović, Maja and Hermann Ney. 2006. POS-based Word Reorderings for Statistical Machine Translation. *Proceedings of International Conference on Language Resources and Evaluation*, Genoa, Italy.

Popović, Maja, Adrià de Gispert, Deepa Gupta, Patrick Lambert, Hermann Ney, José B. Mariño and Rafael E. Banchs. 2006. Morpho-syntactic Information for Automatic Error Analysis of Statistical Machine Translation Output. *Proceedings of the HLT/NAACL Workshop on Statistical Machine Translation*, New York.

Snover, Matthew, and Bonnie Dorr. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the AMTA*, Boston, USA.

Vidal, Enrique. 1997. Finite-state speech-to-speech translation. *Proceedings of the ICASSP*, 111–114, Munich, Germany.

Vilar, David, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of statistical machine translation output. *Proceedings of the LREC*, Genoa, Italy.