Improving a Catalan-Spanish Statistical Translation System using Morphosyntactic Knowledge

Mireia Farrús, Marta R. Costa-jussà, Marc Poch, Adolfo Hernández, and José B. Mariño
TALP Research Centre, Department of Signal Theory and Communications
Universitat Politècnica de Catalunya
C/ Jordi Girona 1-3, 08034 Barcelona, Spain
{mfarrus,mruiz,mpoch,adolfohh,canton}@gps.tsc.upc.edu

Abstract

In this paper, a human evaluation of a Catalan-Spanish Ngram-based statistical machine translation system is used to develop specific techniques based on the use of grammatical categories, lexical categorisation and text processing, for the enhancement of the final translation. The system is successfully improved when testing with ad hoc and general corpora, as it is shown in the final automatic evaluation.

1 Introduction

Statistical Machine Translation (SMT) nowadays has become one of the most popular Machine Translation paradigms. The SMT approach allows to build a translator with open-source tools as long as a parallel corpus is available. If the languages involved in the translation belong to the same linguistic family, the translation quality can be surprisingly nice. Furthermore, one of the most attractive reasons to build an statistical system instead of an standard rule-based system is the little human effort required.

Theoretically, when using SMT, no linguistic knowledge is required. In practice, once the system is built and specially, if the translation quality is high, then the linguistic knowledge becomes necessary to make further improvements (Niessen and Ney, 2000; Popović and Ney, 2004; Popović et al., 2006). In fact, the main question that arose at the beginning of this work was: which are the steps to follow when the intention is to improve a high quality statistical translation?

Let's consider a high quality statistical translation defined as the system which has a BLEU around 75% with a single reference in an in-domain test. This is a relatively unusual situation as most of the statistical translation systems have much lower performance. This study is devoted to develop this stage in the Catalan-Spanish pair in both directions.

The study starts from a high quality Ngram-based statistical translation baseline system, trained with the aligned Spanish-Catalan parallel corpus taken from *El Periódico* newspaper, which contains 1.7 million sentences. A human error analysis of the translation is then performed and used to further improve the translation by introducing statistical techniques and linguistic rules.

This paper is organised as follows. Section 2 describes the Ngram-based statistical translation system used as baseline system. Section 3 reports the human error analysis and evaluation of the baseline system, whose solutions based on statistical techniques, linguistic rules and text processing are explained in section 4. In section 5, an automatic evaluation of the new system is performed and discussed. Finally, Section 6 sums up the conclusions.

2 Ngram-based statistical translation system

An Ngram-based SMT system regards translation as a stochastic process. In recent systems, such an approach is faced using a 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:

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

where the *argmax* operation denotes the search

problem, i.e. the generation of the output sentence in the target language, $h_m(t, s)$ are the feature functions and λ_m are their corresponding weights.

The main feature function (and the only one in our baseline system) is the Ngram-based translation model which is trained on bilingual n-grams. This model constitutes a language model of a particular bi-language composed of bilingual units (translation units) which are referred to as tuples. In this way, the translation model probabilities at the sentence level are approximated by using n-grams of tuples.

The Ngram-based approach is monotonic in that its model is based on the sequential order of tuples during training. Therefore, the baseline system with only one feature function may be specially appropriate for pairs of languages with relatively similar word order schemes. Further details can be found in Mariño (2006 et al.).

3 Linguistic error analysis

In this section we report the linguistic error analysis performed over the Ngram-based baseline output. The analysis was performed by a Catalan and Spanish native linguist at the level of syntaxis, semantics and morphology using out-of-domain text. The set of errors are listed and briefly described next.

Obligation The obligation Spanish expression *tener que* (*have to*) was literally translated as **tenir que* into Catalan, instead of *haver de*.

***Solo* confusion** The term *solo* in Spanish can be related to three distinct parts of speech (POS): adverb (*only*), adjective (*alone*) or noun (*solo*). Since the translation into Catalan depends on the POS, the translated term becomes erroneous when the Spanish POS is not well-recognised, which happens specially in this case between the adverb and the adjective.

Apostrophe In the Spanish-Catalan translation, the apostrophe rules for the Catalan articles *el*, *la* and the preposition *de* in front of vowels are not fulfilled.

Geminated *l* (*l·l*) Although the Catalan geminated *l* should be always written with a middle dot (·), it is very frequent to find it written with normal dot, which leads to erroneous translations into Spanish.

Omission of prepositions The preposition *de* is frequently omitted when translating the Spanish verb *deber* (*must*) into the phrasal verb *haver de*. On the other hand, Spanish normally uses the preposition *a* in front of a direct object while Catalan does not, so that such preposition is usually omitted in the Catalan-Spanish translation.

***a*, *en* prepositions** These prepositions are used in very distinct ways in both Catalan and Spanish languages, so that it becomes difficult to achieve correct translations in both directions.

Possessive pronouns and adjectives In Catalan, possessive pronouns and adjectives are expressed with the same term, whereas Spanish does not. This ambiguity in Catalan leads to confusion in the translation to Spanish.

Conjunction *perquè* This conjunction is ambiguous in the Catalan-Spanish translation since it depends on whether the conjunction is causal, in which case corresponds to *porque* (because), or final, where it corresponds to *para que* (in order to).

Verb *soler* The conjugated forms *sol* and *sols* of the verb *soler* (to use to) can be confused by the adjective meaning *alone* that uses the same term.

Conjunctions *i,o* These Catalan conjunctions must be translated into Spanish as *e* and o instead of *y* and *u* when the following word begins with *i* and *u*, respectively.

Numbers Many numeric expressions are not included in the training corpus, so that no translation can be generated in any of the target languages.

Hours Catalan and Spanish time expressions differ significantly, being usually impossible to use literal translations. The main difference is found in the use of the quarters: where Spanish hours express the quarters that pass from a specific hour, Catalan uses the following hour. E.g. *Las cuatro y cuarto* (four and a quarter) in Spanish would correspond to *Un quart de cinc* (a quarter of five) in Catalan.

Pronominal clitics Frequently, the translation fails in the combination of the pronominal clitic and the corresponding verb.

***Cuyo* relative pronoun** The relative constructions involving the Spanish pronoun *cuyo* are subject to a lexical reordering in the translation into Catalan and viceversa. E.g. the Spanish expression *la mesa cuyo propietario es* (the table whose owner is) would correspond to *la taula el propietari del qual és*.

Gender concordance A masculine Spanish term can correspond to a feminine Catalan term, and viceversa. E.g. *la señal* (Spanish fem., the signal) corresponds to *el senyal* (Catalan masc.).

Unknown words Apart from the numbers, there are other words that are not found in the training corpus due to the fact that they appear only at the beginning of the sentence in capital letters, so that the same words written in lower case letters are not translated.

4 Applying improvement techniques

In order to solve some of the problems described in the previous section, three different techniques have been applied, based on the use of the grammatical category of the words, lexical categorisation and direct text processing, respectively.

4.1 Grammatical category-based techniques

Grammatical categories have been successfully implemented in statistical machine translation in order to deal with some problems such as reordering (Crego and Mariño, 2007) and automatic error analysis (Popović and Ney, 2006). The aim is to add the grammatical category (tag) corresponding to the word we are dealing with, so that the statistical model will be able to distinguish the words according to its category and to learn from context.

Homonymy disambiguation

In the translation task, it is common to find two words in the source language with the same spelling and different meaning that correspond to two different words in the target language, which leads to incorrect translations. When equal words in the source language differ from each other by their grammatical category or associated tag (they are homonymous), such tag can be used for disambiguation.

In the case of the Catalan verb *soler*, instead of generating a series of rules to detect whether *sol* and *sols* are verbal conjugations of *soler*, the tag is directly taken from Freeling tool (Carreras et al., 2004).

However, in some cases, the tag information given by the FreeLing tool is not correct, and some additional processing is needed in order to perform the word disambiguation. In the *solo* case, a series of context-based rules have been designed to identify the *solo* adverb from the *solo* adjective in the doubtful cases. The rules are applied over the source language and the corresponding tag is added to the word in question. Thus, a source language sentence such as *venía solo* (he was coming alone) is transformed into *venía solo_<ADJ>*, so that the statistical model will be able to distinguish between both cases.

A similar process is performed in the Catalan possessives: a set of rules has been designed in order to assign a tag indicating the category of the word (adjective or pronoun), and the tags are then implemented in the source language. Some examples of the resulting translations after applying homonymy disambiguation can be found in Table 1.

Soler	(S) La CR **sol** disposar de quatre.
	(T1) La CR ***solo** disponer de cuatro.
	(T2) La CR **suele** disponer de cuatro.
Solo	(S) Era **solo** un niño.
	(T1) Era ***sol** un nen.
	(T2) Només era un nen.
Poss.	(S) Els meus amics no són **els teus**.
	(T1) Mis amigos no están ***tus**.
	(T2) Mis amigos no son **los tuyos**.

Table 1: Examples of correction after homonymy disambiguation.

Pronominal clitics

The pronominal clitics are initially detected and separated from the verb by using the Freeling tool. After translating them, they are combined again with the corresponding verb. In order to solve the errors in this combination process, a set of rules is defined, in which two grammatical aspects are considered: the Spanish accentuation rules and the pronoun-verb combination in Catalan. In Spanish, for instance, the stressed syllable position changes when adding an enclitic pronoun to the verb:

vende + ***lo*** → *véndelo* (sell **it**)

while in Catalan, the accentuation rules are not altered and the pronoun-verb combination is performed by using apostrophes or hyphens:

seguir + ***lo*** → *seguir-**lo*** (follow **it**)
compra + ***el*** → *compra'**l*** (buy **it**)
el + *aixecava* → ***l'***aixecava (lifted **it**)

Apostrophe

A series of rules have been applied in order to fulfil the Catalan apostrophe rules. The basic apostrophe rule states that the singular articles *el, la* and the preposition *de* must be apostrophised when preceding a word that begins with a vowel or an unsounded *h* (in Catalan language the letter *h* is not pronounced):

el + *arbre* → *l'arbre* (the tree)
la + *hora* → *l'hora* (the hour)
de + *eines* → *d'eines* (of tools)

Some exceptions to these rules have also been included:

- The articles and the preposition are not apostrophised when they precede terms beginning with semiconsonantic *i* or *u* (including *hi*, *hu*): *el uombat* (the wombat), *la hiena* (the hyena), *de iogurt* (of yoghurt).
- The feminine article is not apostrophised when precedes a word that begins with atonic *i* or *u* (including *hi* and *hu*): *la universitat* (the university), *la Irene*.
- The feminine article and the preposition are not apostrophised when preceding the negative prefix *a*: *la anormalitat* (the anormality), *de asimètric* (of asimètric).
- *La una* [hora](one o'clock), *la ira* (the wrath), *la host* (the host) and the names of letters(*la e*, *la hac*, *la erra*, etc.) are not apostrophised.

Some examples of clitics and apostrophe correction can be found in Table 2.

Clitics	(S) No quiero ver**te** más por aquí.
	(T1) No vull veure ***et** més per aquí.
	(T2) No vull veure'**t** més per aquí.
Apostr.	(S) **La acepta** hasta el final.
	(T1) ***La accepta** fins al final.
	(T2) **L'accepta** fins al final.

Table 2: Examples of clitics and apostrophe correction.

Capital letters at sentence beginning

It was also seen in section 3 that some of the unknown words appear in the training corpus only in capital letters, since they are found only at the beginning of sentences. In order to solve this problem, all those words that appear at the sentence beginning are changed to lower case words, except for proper nouns, common nouns and adjectives, since common nouns and adjectives could be also proper nouns, and they are usually not found at sentence beginnings. Therefore, those words that appeared only in capital letters will be translated when writing them in lower case. An example of this type of correction can be found in Table 3.

(S) No entenc per què no hi **assisteixes**.
(T1) No entiendo por qué no ***assisteixes**.
(T2) No entiendo por qué no **asistes**.

Table 3: Example of capital letter unknown word correction.

Gender concordance

In order to improve the translation of those words that change the gender between Catalan and Spanish, a tag containing the part-of-speech information has been used. This technique benefits those word sequences that maintain the gender coherence; for instance: *pilota*_FN *verda*_FAdj (where FN is feminine noun and FAdj feminine adjective) will have a higher probability that *pilota*_FN *verd*_MAdj (where MAdj is a masculine adjective), since the tags model will have seen more time the sequence FN-FAdj than the sequence FN-MAdj.

Nevertheless, the tags model will be useful only if the language model (i.e. the tuples included in the training corpus) allows it. Thus, the translation of *senyal*_MN *blanc*_MAdj will remain as *señal*_FN *blanco*_MAdj instead of *señal*_FN *blanca*_FAdj, since the tuple *blanc#blanca* is not contained in the translation model.

Cuyo

In order to solve the problem of the relative pronoun *cuyo*, a preprocessing rule was applied to transform the Spanish structure into a literal translation of the Catalan structure *del qual*; i.e. the sentences containing *cuyo* or some of its other forms (*cuya*, *cuyos*, *cuyas*), were transformed to

sentences containing *del cual* or its corresponding forms (*de la qual*, *de los cuales*, *de las cuales*), so that the alignment was easier, and some translation errors related to this pronoun were avoided.

Table 4 shows some examples of gender concordance and *cuyo* correction.

Gender	(S) Me encantan **las espinacas**.
	(T1) M'encanten ***les espinacs**.
	(T2) M'encanten **els espinacs**.
Cuyo	(S) Un pueblo **cuyo** nombre es largo.
	(T1) Un poble ***amb un nom** és llarg.
	(T2) Un poble el nom **del qual** és llarg.

Table 4: Examples of gender concordance and *cuyo* relative pronoun correction.

4.2 Numbers and time categorisation

As it was seen in section 3, many numeric expressions are not included in the training corpus and they appear as unknown words in the translation process. In order to solve this problem, the numeric expressions are detected in the source language, codified, and generated again in the target language.

In order to detect the numbers in the source language, two issues must be considered: the structure of the numeric expressions (compound words, use of dashes, etc.) and the gender of the number, if applicable. Then, a specific codification is defined in order to maintain the coherence of the detected expression. Numbers like *un/una* (one), *nou* (nine) and *deu* (ten) have not been categorised because they can be related to non numeric expressions.

On the other hand, it was also seen in section 3 that time expressions differ in Catalan and Spanish languages. Since the training corpus contains few examples related to time expressions, it is difficult to learn from context and to obtain correct translations. As in the numbers, time expressions are detected (considering three possible expression structures), codified and generated in the target language. In some cases, where a verb exists, this changes in the translation, so that it becomes necessary to include it in the detection step. In the following Catalan-Spanish example: *són dos quarts de dues* (it's half past one), which is translated into *es la una media*, the verb changes from plural to singular; thus, the verb must also be included in the detected structure.

Some examples of the correction after number and time categorisation can be found in Table 5.

Numbers	(S) L'alliberament de **quatre-cents quaranta-un** presoners.
	(T1) La liberación de ***quatre-cents *quaranta-un** prisioneros.
	(T2) La liberación de **cuatrocientos cuarenta y un** prisioneros.
Hours	(S) Són **tres quarts de vuit**.
	(T1) Son ***tres cuartos de ocho**.
	(T2) Son **las ocho menos cuarto**.

Table 5: Examples of correction after number and time categorisation.

4.3 Text processing

Some of the errors need to be solved by performing a text processing before or after the translation. The geminated *l*, for instance, have been treated before the translation, by normalising the writing of the middle dot. In other cases such as the obligation *tener que* and the conjunctions *y* and *o* have been treated as a postprocessing after the translation. Some examples correction by text processing can be found in Table 6.

Gemin.	(S) Reformat a **Brussel.les**.
l	(T1) Reformado en ***Bruselas. las**.
	(T2) Se ha reformado en **Bruselas**.
Obligat.	(S) Nos lo **tenemos que** creer.
	(T1) Ens ho ***tenim que** creure.
	(T2) Ens ho **hem de** creure.
y/o	(S) Com a Blanes **o** Olot.
	(T1) Como Blanes ***o** Olot.
	(T2) Como Blanes **u** Olot.

Table 6: Some examples of text processing correction.

5 Evaluation

In order to evaluate the final system after applying the grammatical rules and statistical techniques described in the current paper, a test corpus containing the above-mentioned problematic cases was developed. The built corpus contains 636 sentences for each of the source and target languages, where the problems to deal with can be found in a balanced proportion. In addition, an evaluation with a 2000-sentence test extracted from *El Per-*

ió́dico itself was also performed. The obtained results are shown in Table 7.

	Sent.	ES > CA	CA > ES
Baseline N-II	636	75.91	73.50
Improved N-II		81.35	76.12
Baseline N-II	2000	83.80	83.01
Improved N-II		83.91	83.23

Table 7: BLEU results in both directions of translation.

The results obtained with the 636-sentence test corpus show that the problems we were focusing on are being solved better than in the baseline system. A slight improvement is also observed when using the *El Periódico* test set, although the improvement is not so obvious since the corpus does not contain explicitly the error cases we were dealing with. Additionally, the following points could explain some reasons why the improvement was not higher:

1. The improved translation has an additional knowledge with respect to the corpus. Therefore, some translations from the improved system are correct but differ from the reference while the baseline system outputs the reference as it is. E.g. *EUA està (...)* instead of *Els EUA estan (...)*

2. The CA>ES translation from the improved system contains more words than the CA>ES translation from the baseline system. It must be taken into account that BLEU measures the precision and not the recall.

6 Conclusions

The initial aim of the current paper was to improve an Ngram-based statistical machine system. Once a set of common errors were detected through a human evaluation, a set of techniques based on the used of grammatical category, lexical categorisation and text processing have been applied.

When using an *ad hoc* built test corpus, the results show that the use of grammatical information and the correction of the text as a pre- and postprocessing are useful techniques in order to achieve this goal, as it has been shown in the automatic evaluation: the BLEU of the improved N-II is higher with respect to the baseline system.

A higher performance in terms of BLEU is also reflected in the improved N-II when using a general corpus extracted from *EL Periódico*, although the relative improvement is less than the previous one, since the corpus does not contain explicitly the problems we were tackling in the current paper. Additionally, possible causes for the less improvement observed have been analysed.

References

Carreras, Xavier , Chao, Isaac , Padró, Lluís, and Padró, Muntsa. 2004. FreeLing: An Open-Source Suite of Language Analyzers. *Proceedings of the Conference on Language Resources and Evaluation*, Lisboa.

Crego, Josep M. and Mariño, José B. 2007. Improving SMT by coupling reordering and decoding. *Machine Translation*, 20:3:199–215.

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

Niessen, S., Ney, H. 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*, Sapporo, Japan. 160–167.

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

Popović, M., de Gispert, A., Gupta, D., Lambert, P., Ney, H., Mariño, J.B. y Banchs, R. 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.