Catalan-English Statistical Machine Translation without Parallel Corpus: Bridging through Spanish

Adrià de Gispert, José B. Mariño

TALP Research Center
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain
{agispert|canton}@gps.tsc.upc.edu

Abstract

This paper presents a full experiment on large-vocabulary Catalan-English statistical machine translation without an English-Catalan parallel corpus, in the context of the debates of the European Parliament. For this, we make use of an English-Spanish European Parliament Proceedings parallel corpus and a Spanish-Catalan general newspaper parallel corpus, both of which of more than 30 M words. Given the language proximity between Spanish and Catalan languages, we investigate the cost of using Spanish as a bridge towards large-vocabulary Catalan-English translation in a wholly automatical statistical machine translation framework. Experimental results are promising, as the achieved translation quality is nearly equivalent to that of the Spanish-English language pair, practically carrying SMT research for the Catalan language to the level of more prominent language, in terms of data availability.

1. Motivation

Catalan is a Romance language spoken or understood by as many as 12 million people who live mostly in Spain, where it is co-official in several regions, but also in Andorra (where it is the national language), and some parts of France and Italy.

Despite this, when it comes to the parallel corpora necessary to build statistical machine translation systems, it becomes nearly impossible to find freely-available large-vocabulary data in Catalan and other languages. Thanks to the bilingual nature of most of the Catalan society, one exception can be found in the Spanish language. In fact, much Catalan-Spanish parallel content is being generated on a regular basis (institutions, companies, newspapers, etc.), thus producing training material of which statistical machine translation models can make good use.

Belonging to the same family of languages, being very much influenced and sharing in many cases the same speakers, Spanish and Catalan languages exhibit a morphological and grammatical similarity favouring the quick deployment of high-quality machine translation tools. Since many more parallel corpora in Spanish are available, one can reasonably think of using Spanish as a bridge towards statistical machine translation from Catalan to other languages, and viceversa.

Under this particular situation, the question of whether the Spanish language can be used as a feasible bridge for building state-of-the-art SMT systems between Catalan and any other language is raised.

In this direction, this paper presents a full experiment on large-vocabulary Catalan-English statistical machine translation without an English-Catalan parallel corpus, in the context of the proceedings of the European Parliament debates[1]. For this, we make use of an English-Spanish European Parliament Proceedings parallel corpus and a Spanish-Catalan general newspaper parallel corpus, both of which of more than 30 M words, and we implement two strategies, namely the straight catenation of systems and a direct training between Catalan and English.

The organization of the paper is as follows. Section 2 details the experimental setup, presenting the copora we worked with, as well as the two approaches followed to produce English-Catalan translations, including the evaluation procedure. Section 3 introduces the statistical machine translation system used in all experiments, whereas section 4 reports results for all language pairs. Finally, conclusions are presented in section 5.

2. Experimental Setup

2.1. Parallel Corpora

In order to carry out the experiments, two parallel corpora have been used. On the one hand, a general newspaper Catalan-Spanish corpus has been used, whose main statistics are shown in Table 1, including number of sentences, running words, vocabulary size and average sentence length.

	General Newspaper	
	Catalan	Spanish
Sentences	2.18 M	
Running Words	43.28 M	41.51 M
Vocabulary	390.2 k	397.4 k
Avg. Sent. Length	19.86	19.05

Table 1: Catalan-Spanish corpus statistics

On the other hand, for English-Spanish a parallel corpus containing the proceedings of the European Parliament from 1996 to September 2004 has been used (see Table 2 for main statistics).

	European Parliament proceedings	
	English	Spanish
Sentences	1.22 M	
Running Words	33.37 M	34.96 M
Vocabulary	104.8 k	151.5 k
Avg. Sent. Length	27.28	28.60

Table 2: English-Spanish corpus statistics

[1] So far European Parliament debates are not being manually translated into Catalan.

As it can be seen, even though this is a large-vocabulary task, it proves much more domain-limited as the newspaper task, which includes politics, society, international and sports sections. Note that, whereas in the newspaper corpus a new English word occurs every 111 running words on average, this happens every 318 words for the EU Parliament corpus.

2.2. Bridging strategies

In order to carry out Catalan-English translation, we have implemented two strategies, namely sequential and direct.

Regarding the sequential strategy, it simply consists of catenating two independent statistical machine translation systems, one between Catalan and Spanish, and the other between English and Spanish. Therefore, this is an additive error approach, as errors from one system propagate to the input of the following system.

On the other hand, the direct approach consists of translating the whole Spanish side of the English-Spanish parallel text into Catalan by using the Spanish-to-Catalan system, which is of a more general domain. Then, an English-Catalan system is directly trained by using this automatically produced *noisy* Catalan text. With this we expect that some translation errors of the Spanish-Catalan system will not correlate with English text and may get very low probabilities when training English-Catalan translation models.

2.3. Task evaluation

For evaluating the Catalan-to-English task, we require clean Catalan source development and test data, as well as references for the English-to-Catalan direction. With the aim of minimising the human cost necessary to obtain these data, we automatically translated source and references from the Spanish-English task, and a human reviewer made the minimum corrections necessary to obtain a correct Catalan sentence carrying the same message as the Spanish and English sentences.

Apart from an effort reduction motivation, this has two additional advantages. Firstly, we end up with exactly the same development and test sets for Spanish-English and Catalan-English, which lets as compare both tasks and evaluate the quality loss from one task to the other. And furthermore, the corrected Catalan data can be used as a very accurate evaluation reference for the Spanish-Catalan system.

	Sent	Wrds	Vocab	AvLen	Refs
Cat **à** Eng					
Dev	504	15646	2627	31.0	3
Test	840	23140	3914	27.6	2
Eng **à** Cat/Spa					
Dev	504	15331	2300	30.4	3
Test	1094	26876	3975	24.6	2
Spa **à** Eng					
Dev	504	15415	2735	30.6	3
Test	840	22753	4085	30.4	2

Table 3: Statistics of European Parliament develop and test sets for each translation direction.

Table 3 shows the main statistics of the resultant Catalan develop and test sets, including number of sentences, runnings words, vocabulary size, average sentence length and number of reference translations. As this sets are produced from correcting the automatic translations of Spanish text, English references are identical both for Cat **à** Eng and Spa **à** Eng directions.

3. Baseline SMT system

The SMT system used for these experiments follows the maximum entropy framework (Berger, 1996), where we can define the translation hypothesis *t* given a source sentence *s* as the target sentence maximizing a log-linear combination of feature functions, as described in the following equation:

$$\hat{t}_1^I = \arg\max_{t_1^I} \left\{ \sum_{m=1}^{M} l_m h_m(s_1^J, t_1^I) \right\} \quad (1)$$

where λ_m correspond to the weighting coefficients of the log-linear combination, and the feature functions h_m(s,t) to a logarithmic scaling of the probabilities of each model.

Following this approach, the translation system described in this paper implements a log-linear combination of one translation model and four additional feature models. In contrast with standard phrase-based approaches, our translation model is a bilingual Ngram model expressed in tuples as bilingual units (Mariño et al. 2005). These units are extracted from the automatical word alignment of a given parallel text generated by IBM Models (Brown et al., 1993) with the GIZA++ Toolkit (Och and Ney, 2003).

The tuple N-gram translation model is a language model of a particular language composed by bilingual units which are referred to as tuples. This model approximates the joint probability between source and target languages by using N-grams as described by the following equation:

$$h_m(s_1^J, t_1^I) = \prod_{i=1}^{K} p((s,t)_i \mid (s,t)_{i-N+1}, ..., (s,t)_{i-1}) \quad (2)$$

where $(s,t)_i$ refers to the i^{th} tuple of a given bilingual sentence pair which is segmented into *K* units. It is important to notice that, since both languages are linked up in tuples, the context information provided by this translation model is bilingual.

As additional feature functions, the system includes the following models:

- a target language model
- a word bonus model
- a source-to-target lexicon model
- a target-to-source lexicon model

The first of these feature functions is a language model of the target language, estimated as an standard N-gram over the target words, as expressed by equation 3:

$$h_{LM}(t_k) = \prod_{n=1}^{k} p(w_n \mid w_{n-N+1}, ..., w_{ni-1}) \quad (3)$$

where t_k refers to a partial hypothesis containing k target words, and w_n to the nth target word in it.

Usually, this feature function is accompanied by a word bonus model in order to compensate the system preference for short target sentences caused by the presence of the previous target language model. This bonus depends on the total number of words contained in the partial translation hypothesis, and it is computed as follows:

$$h_{WB}(t_k) = e^{number\ of\ words\ in\ tk} \quad (4)$$

Finally, the third and fourth feature functions correspond to source-to-target and target-to-source lexicon models. These models use IBM model 1 translation probabilities to compute a lexical weight for each tuple, which accounts for the statistical consistency of the pairs of words inside the tuple. These lexicon models are computed according to equation 5, where word-to-word probabilities are obtained from IBM model 1:

$$h_{IBM1}((s,t)_n) = \frac{1}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} p(t_n^i \mid s_n^j) \quad (5)$$

Once the models were computed, sets of optimal log-linear coefficients are estimated on the development set for each translation direction using an in-house implementation of the widely-used **simplex** algorithm (Nelder and Mead, 1965). This system is proved to achieve state-of-the-art performance (Koehn and Monz, 2005; Eck and Hori, 2005).

As decoder, we used the freely-available Ngram-based decoder MARIE (Crego, 2005).

4. Results

In this section the translation results are presented and discussed. First, we evaluate separately the Catalan-Spanish and English-Spanish tasks. Then, translations for the Catalan-English task are evaluated. As automatic evaluation measures, we use Word Error Rate (WER), Position-independent word Error Rate (PER) and BLEU scores (Papineni et at., 2001).

	BLEU	WER	PER
Cat **à** Eng sequential	0.5147	36.31	27.08
Cat **à** Eng direct	**0.5217**	**35.79**	**26.79**
Spa **à** Eng	0.5470	34.41	25.45
Eng **à** Cat sequential	**0.4680**	40.66	32.24
Eng **à** Cat direct	0.4672	**40.50**	**32.11**
Eng **à** Spa	0.4714	40.22	31.41
Spa **à** Cat	0.8421	9.88	8.74
Cat **à** Spa	0.8334	10.08	8.86

Table 4: Translation results for all translation directions. Whereas all directions to and from English belong to the domain of EU Parlamentary Sessions, Spanish-Catalan results are in general newspaper task.

4.1. English-Spanish and Catalan-Spanish tasks

Results for the Spanish-to-English and English-to-Spanish translation directions (belonging to European Parliament task) are shown in the 3rd and 6th rows of Table 4, respectively. Results for the Catalan-to-English and English-to-Catalan translation directions are shown in the last two rows of Table 4.

Although these results correspond to a different task (general newspaper domain, evaluated on a 2000 sentences test with a single reference translation), the quality increase due to proximity between Spanish and Catalan is remarkable.

Interestingly, whereas the Ngram-based translation model takes advantage of the additional features in the Spanish-English tasks (obtaining a performance increase in development of several BLEU absolute points), this behaviour is not observed in the Catalan-Spanish tasks, where simply the ngram translation model suffices to generate high-quality translations. This behaviour tells about the grammatical similarity between Spanish and Catalan, which allows for a less-sparse estimation of the model even with much larger vocabulary sizes.

As mentioned in section 2.3., when generating development and test Catalan texts through correction of automatically-translated texts from Spanish to Catalan, we obtain references, which are *adapted* to the Spanish-to-Catalan European Parliament task. That is, references which evaluate only those mistakes committed by the Spanish-to-Catalan automatic translator (without including alternative translations which do not match the produced translation, if this is correct). When evaluating this task, we obtain the following results:

	BLEU	WER	PER
Spa **à** Cat *adapted ref*	0.9345	3.79	3.49

which reflect once again the high quality obtained with the Catalan-Spanish Ngram-based translation model trained on broad-domain newspaper data[2].

4.2. Catalan-English task

As it can be seen in Table 4, results show in general that the automatic evaluation measures achieved in the Catalan-English task are pretty similar to those of the Spanish-English task, meaning that nearly no loss is found when bridging through Spanish to obtain a Catalan-English system. This is especially remarkable when English is the source language, which turns to be the most challenging case, basically due to the morphological richness of Romance languages in contrast to English.

At a more specific level, in Catalan-to-English, the performance loss due to the bridging through Spanish is higher (comparing the scores of the Catalan-to-English with the scores of the Spanish-to-English) than in the opposite direction. However, in this case the direct training with a noisy Catalan European Parliament corpus achieves slightly yet significantly improved scores in contrast to the sequential strategy.

This behaviour is not so clearly observed in the more difficult opposite direction, where both strategies achieve qualitatively the same performance, as automatic measures show discrepancies (best result is marked in bold). Remarkably, the scores are nearly as good as the English-to-Spanish experiment, although in this case the references are in different languages and therefore the results not strictly comparable.

[2] A demo of this Catalan-to-Spanish and Spanish-to-Catalan ngram-based SMT system can be found at http://www.n-ii.org

5. Conclusions

All in all, we believe that this is a very positive experience, whose implications are relevant, as one can conclude that it is indeed possible to build a large-scale statistical machine translation system between Catalan and any other language, so long as a large-vocabulary parallel corpus between the selected language and Spanish is found.

The achieved translation quality is nearly equal to that of the statistical translation system between Spanish and the given language, practically putting research in Catalan machine translation at the level of a major language, more powerful in terms of data availability.

On the other hand, and contrary to the catenation strategy, the approach implemented here has the advantage of producing a new corpus that can be improved and reused in further research.

Even though this corpus contains the translation mistakes generated by the Spanish-Catalan statistical system, this experiment proves that it is useful enough for training a state-of-the-art Catalan system. Besides, as these mistakes do not necessarily correlate with English data, the direct training achieves a small cancellation of some errors, slightly improving performance in one translation direction.

6. Further research

Unfortunately, at the moment the presented strategy seems to be valid only for those minority languages that are very closely related to other better-represented languages (for example, Galician). Therefore, it remains as a further research to investigate whether similar strategies could be devised for those minority languages which do not relate closely to any major language (for example, Basque).

Additionally, it will be interesting to investigate ways to detect consistent errors and clean them from the new noisy Catalan corpus, improving both the training of the Catalan-English models and the post-processing of the Spanish-Catalan translation system.

Finally, another further research line points towards multi-lingual machine translation, trying to take advantage of the information from both Spanish and Catalan data to improve translation into and from English.

7. Acknowledgements

The authors want to thank Prof. Climent Nadeu for his kind support. This work has been partly funded by the European Union under the integrated project TC-STAR – Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738), by Generalitat de Catalunya and by the European Social Fund.

8. References

Berger, A., Della Pietra, S.A. and Della Pietra, V.J. (1996) A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39-72, March.

Brown, P.F., Della Pietra, V.J., Della Pietra, S.A. and Mercer, R.L. (1993). The Mathematics of Statistical Machine Translation: parameter estimation. *Computational Linguistics*, 19(2):263-312, June.

Crego, J.M., Mariño, J.B. and de Gispert. A. (2005). An Ngram-based Statistical Machine Translation Decoder. In *Proceedings of the 9th European Conf. on Speech Communication and Technology*, Lisboa, Portugal, pp. 3193-96. September.

Eck, M. and Hori, Ch. (2005). Overview of the IWSLT 2005 Evaluation Campaign. In *Proceedings of the 2nd Int. Workshop on Spoken Language Translation*. Pittsburgh, Pennsylvania, pp. 11-32. October.

Koehn, P. and Monz, Ch. (2005). Shared Task: Statistical Machine Translation between European Languages. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*. Ann Arbor, Michigan, pp. 119-124, June.

Mariño, J.B., Banchs, R., Crego, J.M., de Gispert, A., Lambert, P., Costa-jussà, M.R. and Fonollosa, J.A.R. (2005). Bilingual N-gram statistical machine translation. In *Proceedings of the MT Summit X*. Pukhet, Thailand, pp. 275-282. September.

Nelder, J.A. and Mead, R. (1965) A simplex method for function minimization. *The Computer Journal*, 7:308-313.

Papineni, K., Roukos, S. Ward, T. and Zhu, W. (2001) Bleu: a method for automatic evaluation of machine translation. Tech. Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center.