Pivot strategies as an alternative for statistical machine translation tasks involving iberian languages*

Estrategias pivote como alternativa a las tareas de traducción automática estadística entre idiomas ibéricos

Carlos Henríquez†, Marta R. Costa-jussà⋆, Rafael E. Banchs‡, Lluis Formiga† and José B. Mariño†
† Universitat Politècnica de Catalunya-TALP
C/Jordi Girona, 08034, Barcelona
{carlos.henriquez,lluis.formiga,jose.marino}@upc.edu
⋆Barcelona Media Innovation Center
Av Diagonal, 177, 9th floor, 08018 Barcelona, Spain
marta.ruiz@barcelonamedia.org
‡ Institute for Infocomm Research
1 Fusionopolis Way 21-01, Singapore 138632
rembanchs@i2r.a-star.edu.sg

Resumen: Este artículo describe diferentes aproximaciones para construir sistemas de traducción automática estadísticas (SMT por sus siglas en inglés) entre idiomas de escasos recursos paralelos. La estrategia es especialmente interesante para España, un país con tres idiomas oficiales (catalán, vasco y gallego) aparte del castellano, en donde es difícil conseguir corpus paralelo entre cualquiera de los tres primeros pero es comparativamente fácil hacerlo entre castellano y cualquiera de ellos. Tal particularidad nos permite aprovechar el castellano como puente o pivote para construir sistemas que traduzcan entre catalán e inglés, por ejemplo. Estos sistemas son de gran utilidad para los idiomas minoritarios pues ayudan a darles una presencia global y a promover su uso. Como caso de uso, se describe un sistema catalán-inglés siguiendo la estrategia pivote de corpus sintético, la comparamos con una aproximación de cascada y comentamos sobre mejoras adicionales que pudieran implementarse para este par de idiomas en particular.
Palabras clave: idioma pivote, traducción automática estadística, corpus paralelo escaso, cascada, pseudo-corpus, modelos de traducción, frases, n-gramas

Abstract: This paper describes different pivot approaches to built SMT systems for language pairs with scarce parallel resources. The strategy is particularly interesting for Spain, a country with three official languages (Catalan, Basque, and Galician) besides Spanish, where it is difficult to find parallel corpora between two of the first three mentioned languages but it is relatively easy to collect it between Spanish and any of them. This characteristic, however, allow us to develop machine translation systems from major languages like English, to Catalan for instance, using Spanish as pivot. Such systems help these minority languages giving them global presence and promoting their use in content collaboration. We describe a English-Catalan baseline system built following the synthetic approach, we compare it with the transfer approach and comment about future enhancement that could be implemented for this language pair.
Keywords: pivot language, statistical machine translation, scarse parallel corpora, cascade, pseudo-corpus, phrase-based, ngram-based, translation models

1. Motivation

Spain is a multilingual country with four official languages: Catalan, Euskera, Galician and Spanish. Catalan is spoken by 11.5 million people, Euskera by 1.2 million people, Galician by 3.2 million people and Spanish by 400 million people. Given the high number of Spanish speakers compared to the other languages, Spanish has much more linguistic and data resources.

The quantity of resources is relevant in statistical machine translation. The more parallel text we have, the better the translation quality. In order to face the lack of resources in translation, there are many research works on pivot approaches which consist on using a pivot language to perform a source to target translation (Bertoldi et al., 2008a) (Costa-jussà, Henríquez, y Banchs, 2011). For example, in order to translate from Galician to Catalan, we could use Spanish as pivot language. There are much more resources in Galician-Spanish and Spanish-Catalan than between Galician and Catalan directly. The same could happen when interested in translating Catalan, Euskera or Galician into English. In this work, we introduce a state-of-the-art English-Catalan translation system recently built for the free web translator N-II[1].

The main differences with the Catalan-English SMT system presented in (de Gispert y Mariño, 2006) are that in this paper we use an extended corpus and we propose to build a hybrid system which uses an Ngram-based system for Catalan-Spanish and a phrase-based system for Spanish-English. The Ngram-based system outperforms the phrase-based system in Catalan-Spanish (Farrús et al., 2009) while the opposite occurs for the case of Spanish-English (Costa-Jussà y Fonollosa, 2009). Additionally, for the Catalan-Spanish system we are using a further competitive system using rules and statistical features (Farrús et al., 2011).

The remainder of this paper is organized as follows. Section 2 reports a brief description of the phrase-based and Ngram-based translation approaches. Section 3 presents the pivot approaches used in this paper. Section 4 describes the English-Catalan SMT system. Section 5 compares the pivot strategies in terms of translation quality and Section 6 presents the most relevant conclusions.

2. Statistical Machine Translation approaches

As mentioned in the previous section, we are working with two SMT systems: the phrase-based (Koehn, Och, y Marcu, 2003) and Ngram-based systems (Mariño et al., 2006; Casacuberta y Vidal, 2004), which are briefly described as follows.

2.1. Phrase-based

This approach to SMT performs the translation splitting the source sentence in segments and assigning to each segment a bilingual phrase from a phrase-table. Bilingual phrases are translation units that contain source words and target words, e.g. $<$ *unidad de traducción* | *translation unit* $>$, and have different scores associated to them. These bilingual phrases are then selected to maximize a linear combination of feature functions. Such strategy is known as the log-linear model (Och y Ney, 2002) and it is formally defined as:

$$\hat{e} = \arg\max_{e} \left[\sum_{m=1}^{M} \lambda_m h_m (e, f) \right] \qquad (1)$$

where h_m are different feature functions with weights λ_m. The two main feature functions are the translation model (TM) and the target language model (LM). Additional models include POS target language models, lexical weights, word penalty and reordering models among others.

Moses (Koehn et al., 2007) was used to build the phrase-based system.

2.2. Ngram-based

The base of the Ngram approach is the concept of tuple. Tuples are bilingual units with consecutive words both on the source and target side that are consistent with the word alignment. They must provide a unique monotonic segmentation of the sentence pair and they cannot be inside another tuple

* The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 247762 (FAUST) and from the Spanish Ministry of Science and Innovation through the Juan de la Cierva research program and the Buceador project (TEC2009-14094-C04-01).

[1] available at http://www.n-ii.org

in the same sentence. This unique segmentation allows us to see the translation model as a language model, where the language is composed of tuples instead of words. That way, the context used in the translation model is bilingual and implicitly works as a language model with bilingual context as well. In fact, while a language model is required in phrase-based and hierarchical phrase-based systems, in Ngram-based systems it is considered just an additional feature.

This alternative approach to a translation model defines the probability as:

$$P(f,e) = \prod_{n=1}^{N} P\left((f,e)_n \,|\, (f,e)_{n-1}, \ldots, (f,e)_1\right) \quad (2)$$

where $(f,e)_n$ is the n-th tuple of hypothesis e for the source sentence f.

As additional features, we used a Part-Of-Speech (POS) language model for the target side and a target word bonus model.

We used the open source decoder MARIE (Crego, de Gispert, y Mariño, 2005) to build the Ngram-based system.

3. Pivot Approaches

The best approaches to build a SMT system through a pivot language are: the cascade system, also known as the transfer approach and the pseudo-corpus or synthetic approach. Other pivot approaches do not outperform these two (Wu y Wang, 2007) (Cohn y Lapata, 2007). The cascade and the pseudo-corpus approaches have been evaluated and compared in works such as (de Gispert y Mariño, 2006; Bertoldi et al., 2008a; Bertoldi et al., 2008b). Consistently, both works have shown that the pseudo-corpus approach is the best performing strategy.

3.1. Cascade or transfer method

This approach considers the language pairs source-pivot and pivot-target independently. It consists in training and tuning two different SMT systems and combine them in a two-step process: first, we translate a source sentence using the source-pivot system; then, we use the resulting sentence as input for the pivot-target translation. A common variation for this strategy presented in (Khalilov et al., 2008) considers a n-best output instead of the single-best during the first translation and then produce a m-best translation in the last step. At the end, mn-best hypotheses are produced, which are reranked by using Minimum Bayes Risk (MBR) (Kumar y Byrne, 2004), allowing the introduction of additional features such as new language models.

3.2. Pseudo-corpus or synthetic approach

Instead of considering the two language pairs independently, this approach produces a single source-target SMT system. Assuming we have a source-pivot and a pivot-target parallel corpus, we build and tuned a pivot-target SMT system and we use it to translate the pivot part from the source-pivot corpus. This results in a source-target synthetic corpus (hence the name) which is finally used to build the source-target SMT system. For the tuning process, we could also use a synthetic development corpus but an actual source-target corpus is prefered, if possible. A simple variation for this approach is to build a pivot-source SMT system in order to translate the pivot part of the pivot-target corpus, and use the resulting source-target synthetic corpus to build the final system.

4. Building an English-Catalan SMT using Spanish as pivot

We present an English-Catalan SMT baseline system, using Spanish as the pivot language. In this case, the parallel corpus available for the Catalan-Spanish language pair was provided by the bilingual newspaper "El Periódico"[2] and the English-Spanish corresponds to the train corpora provided during the 2010 WMT's translation task[3], i.e. Europarl and News Commentary. We followed the synthetic approach described before to build the final system. Therefore, the Spanish part from the WMT Corpus was translated into Catalan and a English-Catalan phrase-based SMT system was built using the resulting synthetic corpus. Table 1 shows a summary of the statistics of both corpora. We also used the Catalan-Spanish baseline together with the Spanish-English baseline system presented in the 2010's WMT (Henríquez Q. et al., 2010) to build the other direction and compare the different approaches in it.

[2] http://www.elperiodico.es

[3] http://www.statmt.org/wmt10/translation-task.html

Corpora	Catalan	Spanish
Training sents.	4,6M	4,6M
Running words	96,94M	96,86M
Vocabulary	1,28M	1,23M
Development sents.	1966	1966
Running words	46765	44667
Vocabulary	9132	9426
Corpora	**Spanish**	**English**
Training sents.	1,18M	1,18M
Running words	26,45M	25,29M
Vocabulary	118073	89248
Development sents.	1729	1729
Running words	37092	34774
Vocabulary	7025	6199
Test sents.	2525	2525
Running words	69565	65595
Vocabulary	10539	8907

Cuadro 1: Catalan-Spanish and Spanish-English corpora (M stands for Millions)

4.1. Spanish-Catalan baseline system

As mentioned before, the Spanish-Catalan SMT system (named N-II) is based on the corpus provided by the bilingual newspaper "El Periódico". It is a Ngram-based SMT system that includes several improvements specific to the language pair: a homonym disambiguation for the Catalan verb 'soler' and Catalan possessives, special consideration for pronominal clitics, upper-case words and the Catalan apostrophe, gender concordance, numbers and time categorization and text processing for common mistakes found when writing in Catalan. The full description can be found in (Farrús et al., 2011).

4.2. English-Catalan system description

Once obtained the Catalan translation from the Spanish section of the WMT corpus, a phrase-based SMT system was built using Moses as the decoder. Apart from the baseline pipeline, the system also includes a POS target language model computed with TnT (Brants, 2000), numbers and time categorization similar to N-II and the parallel corpus was aligned considering the Catalan lemmas computed with Freeling (Padró et al., 2010) and the English stems of words obtained with Snowball[4].

[4]http://snowball.tartarus.org

Pivot approach	Direction	BLEU
Cascade	cat-eng	21,63
Cascade	eng-cat	24,29
Pseudo-corpus	cat-eng	23,19
Pseudo-corpus	eng-cat	26,97

Cuadro 2: English-Catalan results

5. Results

Table 2 shows the BLEU score of the cascade and pseudo-corpus approaches in both directions. The test set was the one provided as internal test set during the WMT translation task. It is also important to mention that the score was computed using one reference.

The final quality of the Catalan-English system is determined by the quality of the Spanish-English corpus, whose baseline has a BLEU around 24 (Henríquez Q. et al., 2010). The Catalan-Spanish baseline has a BLEU around 80 (Farrús et al., 2009). Also there is a negative effect given the difference in domain between the Catalan-Spanish corpus (a regional newspaper) and Spanish-English corpus (Europarl).

Using paired bootstrap resampling (Koehn, 2004), we can see that for these systems, the Pseudo-corpus approach is better than Cascade with 95 % statistical significance.

6. Conclusions and further work

We have presented an English-Catalan SMT system built using Spanish as pivot language, given the scarce resources for English-Catalan.

Similarly to previous research work, we have seen here that, in the particular translation task under consideration, the pseudo-corpus approach constitutes the best strategy for pivot translation. Although the cascade approach clearly performs worse than the pseudo-corpus approach, it could be also beneficial to consider a system combination between these two strategies to further boost the quality of the translations.

Further work should focus on building Spanish-pivot systems between all the official languages and English, as well as among them. The similarities between the languages (except Basque) and the availability of parallel corpora between Spanish and the others encourage the approach.

Bibliografía

Bertoldi, N., R. Cattoni, M. Federico, y M. Barbaiani. 2008a. FBK @ IWSLT-2008. En *Proc. of the International Workshop on Spoken Language Translation*, páginas 34–38, Hawaii, USA.

Bertoldi, Nicola, Madalina Barbaiani, Marcello Federico, y Roldano Cattoni. 2008b. Phrase-Based Statistical Machine Translation with Pivot Languages. En *Proceedings of IWSLT*.

Brants, T. 2000. TnT – a statistical part-of-speech tagger. En *Proc. of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, WA.

Casacuberta, F. y E. Vidal. 2004. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(2):205–225.

Cohn, T. y M. Lapata. 2007. Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. En *Proc. of the ACL*.

Costa-Jussà, M. R. y J. A. R. Fonollosa. 2009. Phrase and ngram-based statistical machine translation system combination. *Applied Artificial Intelligence: An International Journal*, 23(7):694–711, August.

Costa-jussà, M.R., C. Henríquez, y R. Banchs. 2011. Evaluación de estrategias para la traducción automática estadística de chino a castellano con el inglés como lengua pivote. En *Proc. of the SEPLN*, Huelva.

Crego, J.M., A. de Gispert, y J.B. Mariño. 2005. An Ngram-based Statistical Machine Translation Decoder. En *Proceedings of 9th European Conference on Speech Communication and Technology (Interspeech)*.

de Gispert, A. y J.B. Mariño. 2006. Catalan-English Statistical Machine Translation without Parallel Corpus: Bridging through Spanish. En *Proc. of LREC 5th Workshop on Strategies for developing Machine Translation for Minority Languages (SALTMIL'06)*, páginas 65–68, Genova.

Farrús, M., M. R. Costa-jussà, J. B. Mariño, M. Poch, A. Hernández, C. Henríquez, y J. A. R. Fonollosa. 2011. Overcoming statistical machine translation limitations: error analysis and proposed solutions for the catalan-spanish language pair. *Language Resoures and Evaluation*, 45(2):181–208.

Farrús, M., M. R. Costa-jussà, M. Poch, A. Hernández, y J. B. Mariño. 2009. Improving a catalan-spanish statistical translation system using morphosyntactic knowledge. En *Proceedings of European Association for Machine Translation 2009*.

Henríquez Q., C. A., M.R. Costa-jussà, V. Daudaravicius, R. E. Banchs, y J. B. Mariño. 2010. Using collocation segmentation to augment the phrase table. En *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, páginas 104–108, Uppsala, Sweden, July.

Khalilov, M., M. R. Costa-Jussà, C. A. Henríquez, J. A. R. Fonollosa, A. Hernández, J. B. Mariño, R. E. Banchs, B. Chen, M. Zhang, A. Aw, y H. Li. 2008. The TALP & I2R SMT Systems for IWSLT 2008. En *Proc. of the International Workshop on Spoken Language Translation*, páginas 116–123, Hawaii, USA.

Koehn, P. 2004. Statistical significance tests for machine translation evaluation. En *Proceedings of EMNLP*, volumen 4, páginas 388–395.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, y E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. En *ACL '07: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, páginas 177–180, Morristown, NJ, USA.

Koehn, P., F.J. Och, y D. Marcu. 2003. Statistical phrase-based translation. En *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics*.

Kumar, S. y W. Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. En *Proceedings of the Human Language Technology and North American*

Association for Computational Linguistics Conference (HLT/NAACL'04), páginas 169–176, Boston, USA, May.

Mariño, José B., Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert, José A. R. Fonollosa, y Marta R. Costa-jussà. 2006. Ngram-based Machine Translation. *Computational Linguistics*, 32(4):527–549.

Och, F. J. y H. Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. En *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Padró, Ll., M. Collado, S. Reese, M. Lloberes, y I. Castellón. 2010. FreeLing 2.1: Five Years of Open-Source Language Processing Tools. En *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*, La Valleta, Malta, May.

Wu, H. y H. Wang. 2007. Pivot Language Approach for Phrase-Based Statistical Machine Translation. En *Proc. of the ACL*, páginas 856–863, Prague.