Rapid deployment of a new METIS language pair: Catalan-English

Toni Badia, Maite Melero and Oriol Valentín
Grup de Lingüística Computacional (Barcelona Media-UPF), Barcelona, Spain
Ocata, 1 08003 Barcelona, Spain
E-mail: {toni.badia, maite.melero, oriol.valentin}@upf.edu

Abstract

We show here the viability of a rapid deployment of a new language pair within the METIS architecture. Contrarily to other SMT or EBMT systems, the METIS architecture allows us to forgo parallel texts, which for many language pairs, such as Catalan-English are hard to obtain. In this experiment, we have successfully built a Catalan-English prototype by simply plugging a POS tagger for Catalan and a bilingual Catalan-English dictionary to the English generation part of the system already developed for other language pairs.

1. Introduction

During the past three years, the METIS-II project (Vandeghinste et al., 2006) has explored the feasibility of translating using a monolingual target corpus and a bilingual dictionary only. The bilingual dictionary functions as a flat translation model that provides a number of translations for each source word. The statistical ngram models, which have been built off the target language corpus [1], are then used to score and validate the most probable translation candidate. Vandeghinste et al. (2007) presents a detailed evaluation of the outcome of this project.

The METIS strategy, compared to other statistical MT systems, implies a shift of emphasis from the translation model to the target language generation. In the particular case of the Spanish-English METIS system, the shift is even more pronounced, since we do not use any kind of mapping rules between source and target language structures (see Section 2).

This strategy favours modularity and language independence and, thus, should be easily translatable to new language pairs, requiring only very basic linguistic resources. In the present experiment we put to test the portability of the METIS ideas.

Rapid deployment of a new language pair has been approached by corpus-based systems in the past. Among the most recent attempts in the SMT community, we find Abdelali et al. (2006), Engelbrecht & Schultz (2005), Majithia et al. (2005) and Lavie et al. (2004).

Pinkham & Smets (2002) describe the same thing for a hybrid EBMT system.

Both data driven approaches (SMT and EBMT) require large parallel corpora. Parallel corpora simply do not exist for many language combinations, and are scarce even for 'bigger' languages. They are an expensive resource that low-density languages such as Catalan cannot afford.

To overcome this problem, Gispert & Mariño (2006) present a Catalan-English SMT system, which does not use a Catalan-English parallel corpus. What they actually do is use Spanish as a bridge language. They are able to do without a parallel Catalan-English corpus, only by using two other parallel corpora: Catalan-Spanish and Spanish-English. Along the same lines, Pytlik & Yarowsky (2006) use French-English and Italian-English bitexts to train their Spanish-English system.

On the other hand, the METIS approach allows us to build a translation system between Catalan and English without resorting to any bilingual corpus at all. In the experiment that we have envisaged, since we keep English as the target, we can optimally reuse our existing Spanish-English system by simply plugging in a Catalan low-level preprocessing system and a Catalan-English dictionary.

2. Handling translation divergences in Generation

Surface divergences between the source and target languages are among the major challenges that our minimalist translation strategy faces.

Statistical MT systems try hard to learn these divergences from large parallel corpora. Linguistic based MT systems devise data representations that minimize translation divergences. Predicate-argument structure representations are able to neutralize different syntactic frames or cases. For example, in Spanish, human Direct Objects take the form of a Prepositional Phrase, with preposition "a" (e.g. *a los niños*). In a linguistic-based system, they would be represented as a flat NP (e.g. *los niños*), which could be simply translated into English just by translating the individual words (*the kids*). Remaining divergences, impossible to neutralize, would need to be solved in the translation module, either by:

- Hand written bilingual mapping rules (Transfer MT), or
- Mappings automatically extracted from parallel corpora (Example Based MT).

In the Spanish-English METIS system, we have approached the translation divergences problem with the following constraints in mind:

- Require very basic resources, both for source and target languages: only lemmatizer and POS tagger. Therefore:

[1] The English corpus is a lemmatized version of the British National Corpus tagged using the CLAWS5 tagset. It contains over 6 million sentences.

- No deep linguistic analysis to minimize divergences.
- No parallel corpus to learn mappings from.

- Keep translation model very simple. Therefore no mapping rules, either hand-written, or automatically learned.

Given these constraints, our solution goes in the line of handling translation divergences later in the translation process, namely in the generation component. Instead of bilingual structure modification rules, the Spanish-English system relies on the target language ngram models, as a basis both for lexical selection and for structure construction (Melero et al., 2007).
By pushing the treatment of translation mismatches to the TL end component of the system, we make the treatment independent of the source language and consequently much more general and reusable, as we will show with this experiment.
Our solution is in line with other Generation intensive systems such as Habash & Dorr (2002) and Carbonell et al. (2006). Like us, Habash and Dorr are able to dispense with expensive sophisticated resources for the SL, however, unlike us, they need rich TL resources, such as lexical semantics, categorial variation and subcategorisation frames. In the case of Carbonell, the output of the bilingual dictionary is decoded via long overlapping n-grams, built over full-form words; while we use non-overlapping n-grams over lemma-morphological tag pairs. Also, in their system, in order to account for translation divergences, words and phrases in the SL and TL are substituted by synonyms and near-synonyms, which have been previously learned from TL and SL monolingual corpora.

3. Catalan-English MT System

The choice of this particular language pair has been motivated by several factors:

- There are very few Catalan-English systems available. There is a commercial rule-based system (Translendium) and a couple of incipient research systems: the aforementioned Gispert & Mariño's and OpenMT, which is based on open source technologies (Alegria et al. 2005).
- Given our "Generation intensive approach", keeping English as target, already gives us a head start.
- Basic processing tools for Catalan are easily available to us.

3.1. What do we need?

The figure below shows the architecture of the METIS system. The greyed areas correspond to the parts of the system that are SL related and thus are needed to build a system for a new source language. The modules that can be reused from our existing Spanish-English system, as they are, are left uncoloured.

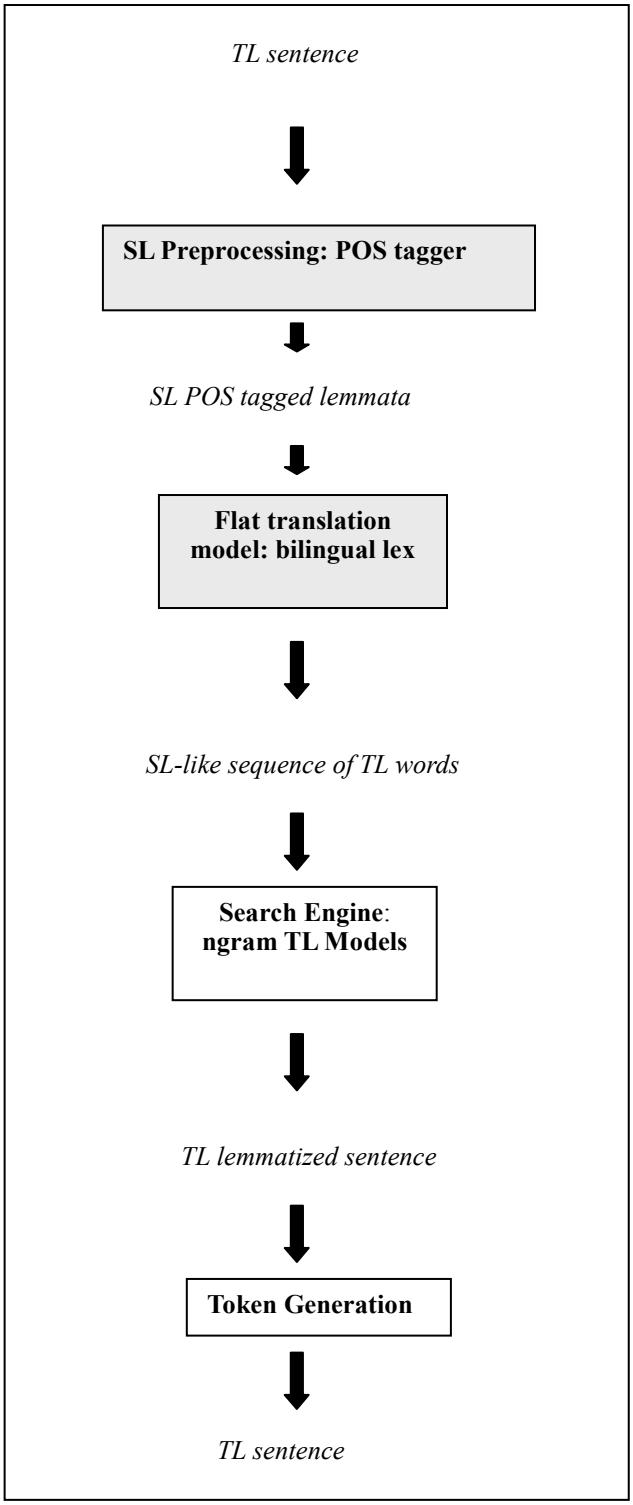

Figure 1: Spanish-English METIS architecture

3.2. Preprocessing of the Source Language

Preprocessing of the SL input requires, in our strategy, very basic resources: only a POS tagger and lemmatizer.

The tagger and lemmatizer of our choice is CatCG (Alsina et al., 2002), a shallow morphosyntactic parser for Catalan, based on the Constraint Grammar formalism. It has been built on the Machinese Phrase Tagger from Connexor[2] . The output of the tagger is a string of Catalan lemmas or base forms, with disambiguated POS tags and inflectional information. Morphological disambiguation is performed by selecting the most plausible reading for each word given the context, expressed in linear terms.

After lemmatizing and tagging has taken place, morphological tags are mapped into the Parole/EAGLES[3] tagset used by the dictionary. In this mapping step, information about POS, which will be used during dictionary look-up is separated from inflectional information which will be used only later, in token generation. Since the POS tagger used for Spanish is also based on the Machinese Phrase Tagger and uses the same tagset, the module that performs the mapping can be reused from the Spanish preprocessing module.

3.3. Translation model: Bilingual Lexicon

The Spanish-English dictionary that performs lexical translation is a lemma-to-lemma dictionary, which has information about the POS of both the source word and the target word. Mapping from source to target is always one-to-one, meaning that entries with more than one translation, or words with more than one POS, become different entries. The POS tags used come from the Parole/ EAGLES tagset, while for English, we use CLAWS5, which is the same tagset used to tag the BNC. The Spanish-English dictionary contains 66,000 entries, and has been automatically extracted from a machine readable dictionary, the Spanish-English Oxford Concise. To be able to reuse as much as possible our extraction scripts, we keep the same format for the Catalan-English METIS dictionary. We also use the Parole/ EAGLES tagset for Catalan[4] . As an initial source lexicon we have chosen to use DACCO[5] , an open-source, good-quality, but not very big Catalan-English dictionary that, at the moment, has 13,384 entries and 16,909 translations.

4. Experiment and evaluation

In order to test our rapidly assembled Catalan-English system, we use a test set of 200 sentences with a balanced distribution of four different text types (Grammar, Newspaper, Technical and Scientific).

The resulting translations have been evaluated using three automatic metrics: BLEU, NIST and the more recently proposed TER (Snover et al., 2006). The first two metrics measure edit distance between the machine-translated sentence and three human created references, while the TER measures the amount of editing that a human would have to perform to convert the MT output into the reference translation.

We compare these numbers with:

- The results obtained by the Spanish-English METIS system, on a similar test set (Vandeghinste et al., 2007);
- The results obtained on the same test set by the only existing commercial rule-based system for Catalan-English (Translendium).

	Grammar	News	Science	Tech	All
Cat-Eng METIS	0.2059	0.2533	0.2070	0.2365	0.2342
Sp-Eng METIS	0.2241	0.3273	0.2876	0.2633	0.2941
Cat-Eng Translendium	0.3334	0.4406	0.4226	0.4264	0.4250

Table 1: Evaluation results (BLEU), compared to Spanish-English METIS and commercial Catalan-System Translendium.

	Grammar	News	Science	Tech	All
Cat-Eng METIS	4.4252	5.6894	4.7482	4.8189	5.7543
Sp-Eng METIS	4.9688	6.3122	5.9071	5.7074	6.7779
Cat-Eng Translendium	4.3013	7.2239	7.1815	7.0841	8.0044

Table 2: Evaluation results (NIST), compared to Spanish-English METIS and commercial Catalan-System Translendium

	Grammar	News	Science	Tech	All
Cat-Eng METIS	49.238	56.372	63.358	60.447	51.894
Sp-Eng METIS	41.890	47.834	50.754	52.960	49.759
Cat-Eng Translendium	60.085	39.310	41.541	40.689	41.790

Table 2: Evaluation results (TER), compared to Spanish-English METIS and commercial Catalan-System Translendium

The results that we have obtained are, as was our expectation, not far from the results obtained by the Spanish-English pair, although they are not as good as the well-established, rule-based Translendium, which has years of development behind.

In our case, the METIS architecture has allowed us to assembly a new language pair, which compares well with the original system, in a very short time. Development time, which has been less than one person month has been employed mostly in obtaining and converting the dictionary and adapting the output of the tagger.

[2] http://www.connexor.com

[3] http://garraf.epsevg.upc.es/freeling/doc/userman/parole-es.html

[4] http://garraf.epsevg.upc.es/freeling/doc/userman/parole-ca.html

[5] http://www.catalandictionary.org/eng/

5. Conclusions

We show here the viability of a rapid deployment of a new language pair within the METIS architecture. In order to do it, we have benefited from the approach of our existing Spanish-English system, which is particularly generation intensive. Contrarily to other SMT or EBMT systems, the METIS architecture allows us to forgo parallel texts, which for many language pairs, such as Catalan-English are hard to obtain.

In this experiment, we have successfully built a Catalan-English prototype by simply plugging a POS tagger for Catalan and a bilingual Catalan-English dictionary to the English generation part of the system. The results of the evaluation are very encouraging.

Although Catalan and Spanish are closely related, we plan, as future research, to prove that the same strategy could be applied to other, more typologically distant source languages, provided that the required minimal processing resources are supplied.

6. References

Abdelali, Ahmed, James Cowie, Steve Helmreich, Wanying Jin, Maria Pilar Milagros, Bill Ogden, Hamid Mansouri Rad & Ron Zacharski (2006) Guarani: a case study in resource development for quick ramp-up MT. Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, "Visions for the Future of Machine Translation", August 8-12, 2006, Cambridge, Massa-chusetts, USA; pp.1- 9

Alegria, Iñaki and Arantza Diaz de Ilarraza and Gorka Labaka and Mikel Lersundi and Aingeru Mayor and Kepa Sarasola and Mikel L. Forcada and Sergio Ortiz-Rojas and Lluís Padró (2005) An open architecture for transfer-based machine translation between Spanish and Basque. Proceedings of the X Machine Translation Summit workshop OSMaTran: Open-Source Machine Translation X, Phuket, Tailandia.

Alsina, A., Badia, T., Boleda, G., Bott, S., Gil, A., Quixal, M. and Valentí, O. (2002) CATCG: a general purpose parsing tool applied. In Proceedings of Third International Conference on Language Resources and Evaluation. Vol. III, pages 1130–1134, Las Palmas, Spain.

Carbonell, J., Klein, S., Miller, D., Steinbaum, M., Grassiany, T. and Frei, J. (2006) Context-based machine translation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Visions for the Future of Machine Translation, pages 19–28, Cambridge, Massachussetts, USA.

Engelbrecht, Herman & Tanja Schultz (2005) Rapid development of an Afrikaans English speech-to-speech translator. International Workshop on Spoken Language Translation: Evaluation Campaign on Spo-ken Language Translation, 24-25 Pittsburgh, PA, US.

Gispert Adrià & José B. Mariño (2006). Catalan-English statistical machine translation without parallel corpus: bridging through Spanish. LREC-2006: Fifth International Conference on Language Resources and Evaluation. 5th SALTMIL Workshop on Minority Languages: "Strategies for developing machine translation for minority languages", Genoa, Italy, 23 May 2006; pp.65-68.

Habash, N. and Dorr, B. (2002) Handling translation divergences: Combining statistical and symbolic techniques in generation-heavy machine translation. In Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Ma-chine Translation: From Research to Real Users, London, UK. Springer-Verlag.

Lavie, Alon, Erik Peterson, Katharina Probst, Shuly Wintner & Yaniv Eytani (2004) Rapid prototyping of a transfer-based Hebrew-to-English machine translation system. In proceedings of the Tenth Conference on Theoretical and Methodological Issues in Machine Translation, Baltimore, USA; pp.1-10

Majithia, Hemali Philip Rennart & Evelyne Tzoukermann (2005) Rapid ramp-up for statistical machine translation: minimal training for maximal coverage . In Conference Proceedings: the tenth Machine Translation Summit; pp.438-444.

Melero, Maite, Oliver, Antoni, Badia, Toni and Suñol, Teresa (2007) Dealing with Bilingual Divergences in MT using Target language N-gram Models. In Proceedings of the METIS-II Workshop: New Approaches to Machine Translation. CLIN 17 - Computational Linguistics in the Netherlands. (pp. 19-26) Leuven, Belgium

Pinkham, Jessie & Martine Smets (2002) Modular MT with a learned bilingual dictionary: rapid deployment of a new language pair. Coling 2002, Taipei, Taiwan.

Pytlik Brock & David Yarowsky (2006). Machine translation for languages lacking bitext via multilingual gloss transduction. AMTA 2006: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, "Visions for the Future of Machine Translation", August 8-12, 2006, Cambridge, Massachusetts, USA; pp.156-165.

Snover, M., Dorr, B., Schwartz, R., Micciula, L., Makhoul, .(2006). A study of translation edit rate with targeted human annotation. In Proceedings of AMTA. pp. 223-231.

Vandeghinste, V., Schuurman, I., Carl, M., Markantonatou, S. and Badia, T. (2006) METIS-II: machine-translation for low-resource languages. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, pages 1284–1289, Genoa, Italy.