Demonstration of the Spanish to English METIS-II MT system

Maite Melero and Toni Badia
Grup de Lingüística Computacional (Barcelona Media), Barcelona, Spain
{toni.badia, maite.melero}@upf.edu

We present an experimental Machine Translation prototype system that is able to translate between Spanish and English, using very basic linguistic resources. In our approach, no structural transfer rules are used to deal with structural divergences between the two languages: the target corpus is the basis both for lexical selection and for structure construction. Our strategy emphasises modularity and language independence and, thus, is translatable to languages with very little NLP development.

Our system is currently being developed in the framework of Metis-II (Vandeghinste et al., 2006). The goal of the Metis project is to achieve corpus-based translation on the basis of a monolingual target corpus and a bilingual dictionary only. The bilingual dictionary functions as a flat translation model that provides n translations for each source word. The most probable translation given the context is then selected by consulting the statistical models built off the TL corpus[1].

Clearly, syntactic divergences between the source and target languages are among the major challenges that this minimalist translation strategy faces. Transfer systems typically address structural translation divergences via explicit bilingual mapping rules, either handwritten or example-based. In the Spanish-English prototype, we are able to do without a rule-based structural transfer component by handling translation divergences in the TL generation component.

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. This solution is in line with other Generation intensive systems such as (Habash & Dorr, 2002). Like us, they are able to dispense with expensive sophisticated resources for the Source Language, however, unlike us, they need rich Target Language resources, such as lexical semantics, categorial variation and subcategorisation frames.

Our approach is also close to the work presented by (Carbonell et al., 2006). In their case, 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-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.

For the preprocessing of the Spanish input, only very basic linguistic resources are needed, namely only a POS tagger and lemmatiser[2], whose output is a string of Spanish 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. At a subsequent step, morphological tags are mapped into the Parole/EAGLES tagset[3] used by the bilingual 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.

Lexical translation is performed by a lemma-to-lemma dictionary, which contains information about the POS of both the source and the target word. The bilingual dictionary has been automatically extracted from a commercial machine readable dictionary, the Spanish-English Concise Oxford Dictionary (Rollin, 1998).

The output of the SL preprocessing and dictionary look-up is a set of translation candidates in form of strings of English lemmas and POS tags, ordered according to Spanish-like syntax.

As mentioned, translations that imply changes of structure are among the main difficulties of using a bilingual lexicon instead of a true translation model. These structure changes can ultimately be reduced to:

- local movement of Content Words (CW),
- deletion and insertion of Function Words (FW)[4], and

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

[2] Our current tagger-lemmatiser is CastCG (Alsina et al., 2002), a shallow morphosyntactic parser for Spanish, based on the Constraint Grammar formalism.

[3] http://www.lsi.upc.es/$\sim$nlp/freeling/parole-es.html

[4] The following parts-of-speech are typically considered to be function words: articles, conjunctions, determiners, pronouns,

- movement of sentence constituents.

Our strategy, which makes crucial use of the distinction between function and content words, provided by the POS tagger, is based on the use of the target-language model to validate any change of structure occurring between SL and TL, instead of writing source-language dependent mapping rules.

A series of target language models are built by indexing all the n-grams for $1 \leq n \leq 5$[5]. An n-gram can belong to one of the following types:

- a sequence of lemma/tag (e.g. always/ADV + wear/VV + a/AT + hat/NN)
- a sequence of lemma/tag except for one position of tag alone(e.g. ADV + wear/VV + a/AT + hat/NN)

During the indexing process, tokens are usually indexed as either lemma/tag or tag alone. Exceptions are:

- personal pronouns (PNP) which are always lemma/tag
- cardinals (CRD), ordinals (ORD) and unknown words (UNC) which are always indexed as tag alone.

To account for structure modifications, we allow permutation of CWs between two consecutive boundaries[6], as well as insertion and deletion of a predefined set of FWs.

In the experiment described in (Melero et al. 2007), we compared the effect of each structure modifying operation in isolation and combined (see results in Table 1). It was run on a test corpus of 227 sentences, for which a set of 3 translation references per sentence was manually created by three independent translators.

Test set	Base	Ins	Del	Perm	All
Grammar	0.4698	0.4518	0.4746	0.4818	0.4658
News	0.3473	0.3358	0.3475	0.3687	0.3516
Technic	0.3072	0.2928	0.3085	0.3205	0.3038
Wiki	0.2720	0.2585	0.2720	0.2960	0.2789

Table 1: BLEU scores for the different settings

In this experiment, we chose as baseline the results of the search on the TL corpus with no structure changing operations. This baseline turned out to be quite high, probably because the word orders of the two languages involved are not extremely different. The variations of the different settings on this baseline are consequently small. The experiment shows the potential of the approach although also brings to light aspects that need to be addressed, such as optimization of weighs and scoring.

prepositions and, specific to English, the existential (*there*) and the infinitive marker (*to*).

[5] The 5-gram model is used only to build the Insertion and Deletion models.

[6] Boundary detection is performed on the basis of the POS information at hand. A boundary is defined by a pair of adjacent POS tags (e.g. NounArticle), which are considered to unambiguously indicate a transition between two consecutive constituents.

References

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, Massachusetts, USA.

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 Machine Translation: From Research to Real Users*, London, UK. Springer-Verlag.

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

Rollin, N. (1998) The Concise Oxford Spanish Dictionary. Oxford University Press.

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.