Dealing with Bilingual Divergences in MT using Target Language N-gram Models

Maite Melero
GLiCom
U. Pompeu Fabra
maite.melero@upf.edu

Antoni Oliver
U. Oberta de Catalunya
aoliverg@uoc.edu

Toni Badia
GLiCom
U. Pompeu Fabra
toni.badia@upf.edu

Teresa Suñol
GLiCom
U. Pompeu Fabra
teresa.sunol@upf.edu

Abstract

In this paper we present a prototype translation system that uses only a source-language (SL) tagger, a bilingual dictionary and a lemmatised target-language (TL) corpus. In our approach, the TL corpus is innovatively exploited both for lexical selection (selecting among the different translations proposed by the dictionary) and for structure building of the output. To that end a series of n-gram model over lemmas and POS tags are built from the TL corpus, which are then searched at run-time. The system presented here uses Spanish as SL and English as TL but the architecture is language independent and translatable to languages with very little NLP development.

1 Introduction

The prototype translation system that we present in this paper is currently being developed in the framework of Metis-II (Vandeghinste et al., 2006). The goal of this 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. The English corpus is a lemmatized version of the British National Corpus [1] tagged using the CLAWS5 tagset. 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. Example-based systems learn those mappings from big parallel corpora (e.g. MSR-MT (Richardson et al., 2001)). However, we are keen not to use an expensive resource such as a parallel corpus, essential to EBMT, and we also want to avoid the laborious and time-consuming task of manually encoding all structure changes between SL and TL in a set of hand-written mapping rules.

As we will see, in our approach, we are able to do without a complex transfer component by handling translation divergences in the TL generation component. This solution is in line with other Generation intensive systems such as the Generation-Heavy MT approach of (Habash and Dorr, 2002). 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. Like us, Habash and Dorr 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 subcategorization frames. Our approach requires only a TL corpus, tagged for part-of-speech and very basically chunked. Habash and Dorr's approach is accomplished by employing their complex symbolic TL resources to overgenerate multiple lexico-structural variations from a target-glossed syntactic dependency of the source-language sentence. This overgeneration is constrained at a later stage by a statistical TL model. By contrast, in

[1] The British National Corpus, version 2 (BNC World). 2001. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

our approach it is the statistical TL model itself which accounts for the different possible lexico-structural variations.

Our approach is also in line with the work presented by (Carbonell et al., 2006). The output of the bilingual dictionary, in their case, is decoded via long overlapping, built over full-form words, and not lemma-tag pairs like in our case. 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.

2 Organization of this paper

This paper is structured as follows. In section 3, we describe the basic resources employed by the system, i.e. POS taggers, bilingual dictionary and monolingual corpus. In section 4, we present the basic structure changing operations needed to deal with the main translation divergences, illustrated for the pair Spanish-English. In section 5, we give an account of the implementation of our strategy using n-gram models. Finally in section 6, we present a small evaluation exercise of the current state of the system.

3 SL Preprocessing and translation dictionary

For the preprocessing of the Spanish input, only very basic linguistic resources are needed, namely only a POS tagger and lemmatizer. Our current tagger and lemmatizer is CastCG (Alsina et al., 2002), a shallow morphosyntactic parser for Spanish, based on the Constraint Grammar formalism. The output of the tagger 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[2] 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.

Lexical translation is performed by a lemma-to-lemma dictionary, which has information about the POS of both the source word and the target word. The bilingual dictionary has been automatically extracted from a commercial machine readable dictionary, the Spanish-English Concise Oxford (Rollin, 1998). The bilingual dictionary functions as a flat translation model and no complex operations can take place in it. However, some translation divergences between SL and TL are actually dealt with in the dictionary, such as the following:

1. Category change, e.g. *mundial* (ADJ) translated by *world* (N) (as in *economia mundial* (*world economy*));
2. A single SL word is translated into a fixed TL multi-word expression, e.g.: *acequia* translates into *irrigation ditch* or *muchos* translates into *a lot of*;
3. A SL compound has a single-word translation, e.g.: *abeto falso* translates into *spruce*

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.

4 Addressing translation divergences using TL n-gram models

4.1 Basic operations to handle structure modification

Translations that imply changes of structure going from source to target, are among the main difficulties of using a bilingual lexicon instead of a true translation model. These structure changes can ultimately be reduced to:

1. local movement of content words,
2. deletion and insertion of function words[3], and
3. 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 occuring between SL and TL, instead of writing source-language dependent mapping rules.

[2] http://www.lsi.upc.es/∼nlp/freeling/parole-es.html

[3] The following parts-of-speech are typically considered to be *function words*: articles, conjunctions, determiners, pronouns, prepositions and, specific to English, the existential *there* and the infinitive marker *to*.

By allowing reordering of elements, plus deletions and insertions, the combination of possibilities in the search algorithm explodes. In order to limit the search space in a linguistically principled way, we use a sort of pseudo-chunking strategy by identifying constituents' boundaries on the strings of English lemmas, which constitute the output of the dictionary. 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 unambiguosly indicate a transition between two consecutive constituents.

Boundaries are used twofold:

1. On the one hand, two consecutive boundaries mark the limits within which content words are allowed to move;

2. On the other hand, boundaries are used to build a second-level language model (or syntactic model) needed to handle non-local order changes, such as movement of constituents.This is an n-gram model over sequences of POS tags. The tags in this model are complex tags of the type AT0-AJ0-NN, limited by boundaries. In this way, this model yields a representation of the syntactic patterns of the target language which is then used to rank all possible permutations of the input tag sequences.

4.2 Spanish-English structural divergences

We talk about a translation divergence when a literal, or word-by-word translation (even if the word senses are correct) is incapable of yielding a correct translation of the original sentence. (Habash and Dorr, 2002) classifies divergences between Spanish and English in five types:

1. Categorial divergence: e.g. *tener hambre* (*have hunger*) translates into *be hungry*;

2. Conflational: e.g. *dar un beso* (*give a kiss*) translates into *kiss*;

3. Structural: e.g. *ver a Juan* (*see to Juan*) translates into *see Juan*;

4. Head swapping: e.g. *cruzar nadando* (*cross swimming*) translates into *swim across*;

5. Thematic: e.g. *X gusta a Y* (*X pleases to Y*) translates into *Y likes X*.

In our experiment, we adress divergences of the structural type, which are among the most common ones. As Table 1 summarizes, most structural divergences between English and Spanish may be reduced to a combination of the basic operations exposed in Section 4.1.

4.3 Creation of the TL Models

A target language model is built by indexing all the n-grams for

$$1 \leq n \leq 4$$

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.

In order to optimize the indexing and the search process, only n-grams with words appearing among the 30K more frequent are indexed. The model is stored in 4 Berkeley databases, one for each n. An extra table for n=5 is also calculated, which will be used for the calculation of the deletion and insertion models (see below).

In order to deal with structure changes between SL and TL, a Deletion and an Insertion models are created. To build the Deletion Model, for every n-gram (for n between 3 and 5) containing functional words in any position, excluding the first and the last, the n-gram resulting from deleting the functional word(s) is looked-up in the TL model. If the resulting n-gram has a frequency a fixed times greater than the original n-gram frequency, then the original (longer) n-gram is linked to the new (shorter) n-gram in the Deletion Model, which is also implemented as a Berkeley hash table. The Insertion Model is built in much the same way.

Spanish - English contrast	Deletion	Insertion	Local movement	Global movement
Extraposition and clitic duplication (El coche se lo ha comprado José / José bought the car)	Pron	-	-	+
Human Direct Object (PP→NP) (Contesta al profesor / Answer the teacher)	Prep	-	-	-
DO + IO → Hum DO + Prep O (Yo le pediré dinero a mi amigo / I will ask my friend for money)	Prep Pron	Prep	-	+
Se pronoun → Passive (Se construyó otro puente / Another bridge was built)	Pron	Aux verb	-	+
N + de N → N + prep N (la chica de la perla / the girl with the pearl)	Prep	Prep	-	-
N + de N → N + N (jugo de naranja / orange juice)	Prep	-	+	-
Noun + Adjective (una mesa redonda / a round table)	-	-	+	-
Generic NP Subject (Los hombres son mortales / Men are mortal)	Art	-	-	-
Apposition (Vi al profesor Chomsky / I saw Professor Chomsky)	Art	-	-	-
Temporal NPs (Terminamos el lunes / We will finish on Monday)	Art	Prep	-	-
Generic NP Object (Luisa siempre lleva sombrero / Louise always wears a hat)	-	Art	-	-
Reflexive pron → possessive (Tomás se puso la chaqueta / Thomas put on his jacket)	Pron	Poss	-	-
Decausative se → inchoative constr (El barco se hundió / The boat sank)	Pron	-	-	-
Transposed DO and Subject (Compró mi padre una casa / My father bought a house)	-	-	-	+
Do / Aux Inversion (¿Quieres café? / Do you want coffee?)	-	Do-particle	-	-

Table 1: Spanish-English structural divergences in terms of the basic operations

5 Search against the TL n-gram model

5.1 Lexical Selection

The set of translation candidates obtained form the dictionary is validated against the TL model. In order to reduce the search space, an initial preselection of candidates is performed, based on the co-apparition probability of the content words appearing in the candidate in question. Lexical co-apparition probability is calculated using a table where TL words (i.e. lemma-tag pairs) are stored together with the sentences' identifiers where they appear.

5.2 Candidate expansion

The initial set of candidates, after the initial preselection phase, is then expanded as new candidates result from the application of the structure modification operations explained above: deletion and insertion of functional words; permutation of content words between boundaries; and permutation of whole constituents according to the second level (or syntactic) model.

5.3 Ranking of the candidates

At this point, the set of candidates is ranked according to the evidence found in the target corpus. Scoring follows a logarithmic progression based on lenght and frequency of the n-grams, complemented with a negative scoring on the pieces that remain untranslated. The rationale of the negative scoring is that even if a long n-gram has been identified as a good candidate for translation, if the remaining pieces are unfrequent POS tag combinations, there is a penalizing score that counterbalances the positive scoring.

5.4 Token Generation

The highest ranking translation is still in lemma form and full-form words need to be generated. This is done in two successive steps. In a first step, the extended CLAWS5 tag for each lemma is calculated, by combining the reduced tag (used in the lemma-to-lemma dictionary and lemmatized TL corpus) with part of the SL morphological information. In a second step, the reversible tokenizer, previously used in the other direction to lemmatise the TL corpus, is applied.

6 Experiment

We want to test our experimental approach in which translation divergences between Spanish and English can be solved by means of target language models, instead of being dealt with by explicit rules, either hand-written or learnt from parallel corpus. To that end we have created a test bench of Spanish sentences, which have been translated to English by three different human translators.

6.1 Test corpus

The test corpus is composed of a set of 227 sentences which are distributed evenly among four different categories:

- 46 sentences extracted from Spanish grammar books and books about contrastive English-Spanish phenomena (Whitley, 2002).
- 60 sentences of journalistic text type extracted from an online newspaper [4].
- 62 sentences of encyclopedia extracted from Wikipedia articles in Spanish [5]
- 60 sentences of technical text extracted from an online technical manual for Open office [6].

6.2 Restrictions on the size of the search space

To limit the computational complexity in this experiment we have chosen to work with a subset of the total BNC corpus and limit the number of translations provided by the dictionary. Therefore, we have built our language models following the procedure exposed in section 4.3 over a target corpus of 1M sentences and we have restricted to three the number of possible translations for each source word. Since at the time of writing this paper, token generation has not yet been implemented, evaluation has been carried out on the lemmatized output.

6.3 Scoring weights

For this experiment we have manually assigned weights to score the translation candidates, based on the n-grams found in the corpus and on penalizations for n-grams not found in the corpus. These values, which are shown in the tables below, need to be seen as a first attempt to scoring potential candidates.

[4] http://www.lavanguardia.es/
[5] http://es.wikipedia.org/wiki/Wiki
[6] http://www.linex.org/

	Lemma|Tag	Lemma|Tag + 1 Tag
4-gram	1.000.000 x freq	10.000 x freq
3-gram	10.000 x freq	100 x freq
2-gram	100 x freq	10 x freq
1-gram	1 x freq	–

Table 2: Accumulated scores for n-grams found in the TL corpus.

	Lemma|Tag	Lemma|Tag + 1 Tag
4-gram	1	10
3-gram	100	1000
2-gram	10.000	100.000
1-gram	1.000.000	–

Table 3: Penalties for n-grams not found in the TL corpus.

6.4 Description of the experiment

We have evaluated the following scenarios, using the standard measure BLEU with three reference human translations:

- A baseline translation where the incoming Spanish word order has not been altered.
- Search against the TL n-gram model with the insertion model activated.
- Search against the TL n-gram model with the deletion model activated.
- Search against the TL n-gram model, allowing permutations to happen between content words inside boundaries.
- Search against the TL n-gram model allowing all operations to happen: insertion, deletion and permutation.

The effect of the second-language or syntactic model has been excluded from this experiment.

6.5 Experiment results and error analysis

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 4: BLEU scores for the different settings

The results of our experiment are shown in Table 4. The following facts become apparent:

- The baseline is quite high, particularly for the Grammar set, with shorter sentences, and the variations on this baseline are small.
- The insertion operation performs consistently slightly under the baseline.
- The permutation operation yields the best results in all test sets.
- The results of the combination of all the operations are a combination of the results of each operation tested independently; in two cases they are slightly over the baseline and in two cases they fall slightly under.

6.5.1 Analysis of the results for the Insertion operation

The functional words that have been inserted in this experiment are:

- Articles (*the*, *a/an*).
- Prepositions.
- *To* particle

Insertion of prepositions yields a higher error rate than that of articles. Often it does not take into account enough context, as in the following example:

interpret a song → interpret *as* a song

In certain cases, it seems apparent that the choice of the preposition should belong to the translation model and not to the generation model:

guess the riddle → guess *at* the riddle

There are several cases where a preposition has been inserted between two words that should otherwise undergo permutation:

change climatic → change *in* climatic
optimism prudent → optimism *for* prudent
diversity notable → diversity *in* notable

Insertion of articles and the *to* particle works much better, as illustrated by the following examples:

say me → say *to* me
leave house → leave *the* house
call to united states → call *to* the united states
without need → without *the* need

can be object of help → can be *the* object of help
in few hour → in *a* few hour
be conceived as holy drink → be conceived *as* a holy drink

The fact that we are using lemmas and not forms sometimes affects the resulting translation negatively. In the following example, an article is incorrectly inserted in the absence of information about number.

attractive for company → attractive for *the* company (ref. attractive for companies)

6.5.2 Analysis of the results for the Deletion operation

The functional words that have been experimentaly deleted in this experiment are:

- Conjunction *that*.
- Preposition *of*.

The small impact of deletion on the baseline is clearly related to the reduced set of words that have been allowed to disappear for this experiment. The impact of removing the preposition *of* is clearly positive and is illustrated by this example:

he leave *of* the meeting → he leave the meeting

On the other hand, all examples of deletion of the *that*-conjunction happen in complex sentences that have a Bleu score of 0.0000 and do not seem to have an overall positive effect.

6.5.3 Analysis of the results for the Permutation operation

The most frequent permutation takes place in the ”Noun + Adj” group , as in the examples:

nation arab → arab nation
engagement official → official engagement
that man tall → that tall man
beach catalan → catalan beach

Particularly interesting are the results obtained in groups with more than one adjective, such as:

ceremony traditional hindu → traditional hindu ceremony
system operating favorite → favorite operating system
attack aerial american → aerial american attack

These examples, which would be very hard to re-order with traditional hand-written rules based only on SL surface order, have been satisfactorily solved by the TL n-gram models[7].

The catch of our approach to permutations, where we do not specify which categories may permute, but only that they need to be major categories is that other, more difficult, permutations can also take place, such as ”Noun + Noun”:

menu format → format menu

It is also interesting to note that a deterministic rule to permute ”Noun + Adj” would not always be right, such as in the following example where the operation has -correctly- not happened:

a center of *culture famous* in all the world

6.5.4 Analysis of the results for the combination of all the operations

The combination of all the operations too often gives more credit to insertion than permutation, as in the following example:

attack aerial in the gaza strip →
attack *on* aerial in the gaza strip
(instead of *aerial attack*...)

This is probably due to the fact that in our current weighing system longer n-grams are preferred over shorter n-grams and is an evidence that the insertion operation should be more penalized than it currently is.

There are less examples where permutation is -correctly- preferred over insertion:

that conflict armed → that armed conflict
(and not *that conflict with armed*)

[7] Except perhaps for the last example, which could be better expressed as: american aerial attack

7 Conclusions

In this paper we have presented an experimental Machine Translation prototype 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. For this purpose, we build a series of TL n-gram models which are searched to validate the set of candidates proposed by the dictionary and extended by the structure changing operations, reduced to insertion and deletion of function words and movement of content words. Our strategy emphasises modularity and language independence and, thus, is translatable to languages with very little NLP development.

The experiment described here shows the potential of the approach and highlights some interesting aspects that will be addressed in further development of the system.

8 Future work

As a first step we aim at optimizing the tuning of the several parameters intervening in the searching process against the TL models, such as scoring of n-grams in the ranking phase on the basis of lenght, type of n-gram, frequency, etc. as well as negative scoring of n-grams not present in the corpus and n-grams that have undergone certain structure changes, such as insertion. Currently these paramenters are manually fixed and have been set in a quite straightforward way as a first approach, but we plan to use Machine Learning techniques to tune them more optimally.

We plan to test the effect of expanding or reducing the set of functional words used in the structure changing models. We also plan to experiment with the syntactic model which will allow for reordering of constituents.

Lastly, we also plan to optimize execution times by applying dynamic programming to the search algorithms. In this way we expect to be able to increase the size of the target models by using the whole BNC (6M sentences) which will probably have a boosting effect on the overall results as well.

References

Àlex Alsina, Toni Badia, Gemma Boleda, Stefan Bott, Àngel Gil, Martí Quixal, and O. Valentín. 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.

J. Carbonell, S. Klein, D. Miller, M. Steinbaum, T. Grassiany, and J. Frei. 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.

Nizar Habash and Bonnie Dorr. 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.

S. Richardson, W. Dolan, A. Menezes, and J. Pinkham. 2001. Achieving commercial-quality translation with example-based methods. In *Proceedings of the Machine Translation Summit VIII*, pages 293–298, Santiago de Compostela, Spain.

Nicholas Rollin. 1998. *The Concise Oxford Spanish Dictionary*. Oxford University Press.

V. Vandeghinste, I. Schuurman, M. Carl, S. Markantonatou, and T. Badia. 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.

M. Stanley Whitley. 2002. *Spanish/English Contrasts: A Course in Spanish Linguistics, 2nd edition*. Georgetown University Press, Washington, USA.