An n-gram approach to exploiting a monolingual corpus for Machine Translation

Toni Badia, Gemma Boleda, Maite Melero, Antoni Oliver
Universitat Pompeu Fabra
Passeig de Circumval·lació, 8 - 08003 Barcelona
{toni.badia,gemma.boleda,maite.melero,antonio.oliver}@upf.edu

Abstract

In this paper we present an approach to Statistical Machine Translation that uses a bilingual dictionary and a target language model based on n-grams extracted from a monolingual corpus. This approach is still in an experimental stage and is being developed in the context of Metis-II, a UE project that aims at constructing free text translations by retrieving the basic stock for translations from large monolingual corpora. The architecture described in this paper is being applied to translation from Spanish to English and is designed so as to depend as little as possible on complex linguistic processing tools. The only required tools are a POS tagger and lemmatizer for the source language, and another for the target language.

1 Introduction

Corpus-Based Machine Translation (MT), including Statistical Machine Translation (SMT) and Example-Based Machine Translation (EBMT), use bilingual parallel corpora to train translation models. SMT is based on probability theory (Yamada and Knight, 2001); EBMT, on the other hand, is inspired by analogical reasoning: every new translation is computed in analogy to already known translations extracted from a bilingual corpus (Carl and Way, 2003). This approach basically relies on finding translated maximal-length phrases that combine to form a translation.

One basic pre-requisite for Corpus-Based Machine Translation is the existence of adequate bilingual parallel corpora, which may be difficult to acquire, even for widely spoken languages, let alone minority languages. Considering that for statistical systems one of the best ways to improve the results is by using a larger corpus (Banko and Brill, 2001), difficulty to acquire parallel corpora is a major drawback.

Another factor worth taking into consideration is the fact that the existing parallel corpora often belong to a very limited number of domains, such as parliamentary debates like the Hansards (debates of the Canadian Parliament) or Europarl (minutes from the European Parliament; Koehn (2002)).

On the other hand, the availability of monolingual corpora in digital format, belonging to a large variety of domains, keeps growing for all languages.

Given this scenario, the goal of the project Metis-II is to achieve corpus-based translation on the basis of a monolingual target language corpus and a bilingual dictionary only. The project aims at building a translation system for Dutch, German, Greek, and Spanish to English, using the British National Corpus (BNC; Burnard, (1995)) as the monolingual target language corpus.

Metis-II was preceded by Metis-I, which operated on a sentence-level base (Dologlou et al., 2003). Using a bilingual dictionary and some reordering rules, a near word-by-word translation was produced from the source sentence. The target corpus was then searched for the closest match, which was then proposed as the best known translation for the source sentence. Even though the performance of Metis-I was superior to that of a Translation Memory (built using a more expensive resource, namely a parallel corpus), it was clearly limited by the size of its base unit: the sentence.

Metis-II aims at improving on the results of the approach initiated by Metis-I by using segments below the sentence level. Since finding the exact match of a sentence is too strict a requirement, the sentence has to be decomposed into some kind of constituents, in order to perform a partial match. Proposals about how to decompose example sentences abound in the literature on EBMT (Turcato and Popowich, 2001). In most cases, some sort of linguistic analysis is used, from the most low-level to the most deep-level, e.g. clustering methods for chunking (Brown, 2003), shallow parsing for extraction of translation units (Carl, 2003), use of dependency trees (Watanabe et al., 2003), logical forms or predicate-argument structures in the Microsoft Research MT system (Richardson et al., 2001), etc. The idea behind all these proposals is that examples can be decomposed into smaller constituents to be processed independently. Every approach addresses in one way or other the two main problems of decomposition, namely "boundary definition", i.e. where to segment, and "boundary friction", i.e. how to stitch together the translated pieces.

Different approaches to decomposition and reuse of the material are currently being explored within the Metis-II Consortium. We next explain the approach that is being explored by GLiCom[1], which uses n-grams as base units.

2 The n-gram approach

Statistical MT systems typically consist of a translation model and a target language model (Brown et al., 1993). In our case, the bilingual dictionary functions as a lexical translation model and we only need to compute the target language model, out of the target language corpus.

2.1 Linguistic pre-processing

In our approach n-grams are not built out of words, as it is usually the case in SMT systems, but out of lemmas and/or morphological tags. This implies that both the target corpus and the input sentences have to be lemmatized and tagged. In addition to providing a more generalized representation of the corpus, to avoid data sparseness, this representation has the advantage that it can be directly used in the dictionary lookup: typical machine readable bilingual dictionaries are lemma-to-lemma, so that they need a lemmatized input and provide a lemmatized output.

In order to process the Spanish input, we use a morphological analyzer called KURD (Carl and Schmidt-Wigger, 1998). KURD is a constraint-based formalism that works on the basis of a pattern matching approach that is suitable for shallow or partial linguistic processing. It manipulates morphological analysis in order to kill, unify, replace or delete parts of the structure. The result of the pre-processing with KURD yields a disambiguated morphological analysis that can be fed into the lemma-to-lemma bilingual dictionary.

2.2 The bilingual dictionary

The bilingual lexicon that we use is based on a commercial machine-readable dictionary, the Concise Oxford Spanish (Rollin, 1998), which has 32,653 entries in the Spanish-English direction with an average of 4 translations per headword.

The coverage is being enlarged, using automatic procedures, with entries coming from the reverse direction (English-Spanish) as well as from terminological glossaries. Orthographic and regional variants, such as British and American spellings are also being added, as well as compounds, that appear in the original dictionary as secondary entries under the main headword.

Lemma to lemma translations are automatically extracted from the machine readable dictionary such that mapping from source to target is always one-to-one. Because of simplicity of design, identical headwords with different translations constitute different entries. Likewise, identical headwords with different parts of speech constitute different entries even if the translation is the same. The structure of the resulting entries looks as follows:

- Entry identifier
- Spanish headword (lemma)
- POS of the Spanish headword (PAROLE/EAGLES tag set)[2]
- English translation (lemma)
- POS of the English translation (CLAWS5 tag set)[3]

In order to build our dictionary, we need to calculate the POS of the translation, which is not present in the original Spanish-English dictionary. This POS is automatically assigned on the basis of the POS of the source word and is subsequently validated on the English-Spanish dictionary and other sources, like the target corpus itself. In the few cases where the POS of the target does not coincide with the POS of the source, the validation will overwrite the default. The value of the POS is expressed using the CLAWS5 tag set, which is the same tag set used to tag the BNC.

[1] Computational Linguistics Group of the Universitat Pompeu Fabra

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

[3] http://www.comp.lancs.ac.uk/computing/research/ucrel/

The machine readable dictionary from which our dictionary is extracted provides other types of lexical information as well, such as collocations, sense indicators[4], field labels, examples, etc. In future enhancements of the system, we are considering exploiting part of this information, particularly collocations, in the translation process.

2.3 The language model

The target corpus that we use to validate the translations coming from the dictionary is a lemmatized version of the BNC. In a first step all n-grams are sequences of lemmas. In a second step, one -and just one- of the lemmas of a given n-gram is substituted by its POS tag. This is done for every lemma in the n-gram, one lemma at a time[5]. Here is an example of a 4-gram with tag substitution: *inconvenient on those occasions*:

inconvenient on this occasion
AJ0 on this occasion
inconvenient PRP this occasion
inconvenient on DT0 occasion
inconvenient on this NN1

This is repeated for the tri-grams and bigrams contained in the 4-gram. The last model to be built is the unigram model. This model, which does not provide contextual information, is nevertheless used as a frequency measure for single words. If no other evidence is found, at least the most frequent word is chosen as translation. The purpose of the target language model within our architecture is twofold:

- Perform lexical selection: i.e. select one translation out of the possible candidates provided by the bilingual lexicon
- Build the sentence structure: i.e. select one of the possible orderings of the tokens within the n-grams, as well as among n-grams

2.4 The translation process

Once the source sentence has been tokenized, tagged and lemmatized, it goes through the following steps:

1. Every lemma in the source sentence is matched against the left side of the bilingual dictionary. Part-of-speech information obtained from the tagger is used to guide lexicon look-up in order to disambiguate between homonymous words, i.e. words with the same lemma but different category. Other morphosyntactic values such as tense or number are not used at this point but are stored and will be consulted at the end of the process in order to generate the right inflected form in the target language. If a source word (i.e. lemma) is not found in the bilingual lexicon, the word is left untranslated.
2. All possible n-grams are built out of the sequence of translated lemmas, starting with the highest value of n (i.e. n=4). A different n-gram is built for each translation possibility. For instance, in the sentence *el niño pequeño come carne* 'the little child eats meat', if *carne* is translated as *meat* or *flesh*, both *the child little eat meat* and *the child little eat flesh* are built.
3. At this point, a reduced set of hand-written mapping rules may need to apply in order to deal with specific phenomena. These rules are an ad-hoc mechanism apt to deal with hard translation problems, such as thematic role inversion (e.g like - gustar) and other structure changing issues. However, for alternatives to the use of mapping rules, see section 3.2.
4. Validation of the translated n-grams proceeds. Based on the frequency in the target language corpus and the length (i.e. the value of n) of the n-gram, a weight is assigned to each candidate. In the case no evidence is found for a given n-gram formed by lemmas, the process is repeated by successively substituting one lemma by its tag. This substitution affects negatively the weight of the resulting n-gram.
5. When all calculations have been done, the n-grams with the highest weights are kept as translation candidates.
6. The n-grams of the portions which have not yet been validated by the model are recalculated, and steps 2-5 are repeated with n = n-1, until all portions are calculated or n = 1. The portions that are validated at a particular stage of the process are not further taken into consideration.

[4] These may be near synonyms or guiding words or explanations.

[5] Except in the case when there is a proper noun or a cardinal number in the n-gram, in which case, we may find more than one tag: the one that is being substituted, plus the tag for the proper noun (NP0) or the tag for the number (CRD)

7. Any POS tag left in the final string, different from *cardinal* or *proper noun*, is replaced by the most frequent translation of the original word according to the unigram model. If none of the proposed translations appear in the target corpus, the first translation provided by the lexicon is then chosen. Tokens tagged as *cardinal* or *proper noun* are replaced by the original word.

3 Dealing with changes of structure

Translations that imply changes of structure, going from source to target, are among the main difficulties of using a bilingual lexicon, and not a true translation model. These changes of structure can be reduced to:

- Insertions.
- Deletions.
- Movements: local and non-local.

Although a small set of hand-written mapping rules can be advisable for some phenomena, and is indeed foreseen in the general Metis architecture, they cannot be the only device to deal with changes in structure, if the system is to be robust and scalable. More generally, we plan to use our target language model to perform these changes.

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 intend to use the information provided by the POS tagger to distinguish between *content words* and *grammatical words*. The idea is to limit (local) movement to content words, and possibility of insertion or deletion to grammatical words.

3.1 Insertions and deletions

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

We assume that insertion or deletion affect only grammatical words. These words function as true markers, in the same way as morphological inflection does and therefore, very often, only appear in the source or in the target, but not in both. The following are common examples of this phenomenon:

(1) Dormían en un coche
sleep-PAST-P3-PLR in a car
'THEY slept in a car'

(2) La policía detuvo a
the police arrest-PAST-P3-SNG TO
un sospechoso
a suspect
'The police arrested a suspect'

(3) Los perros ladran
THE dogs bark-PRES-P3-PLR
'Dogs bark'

Example 1 illustrates a case of pro-drop, i.e. absence of explicit subject. This is a common phenomenon in Spanish. The subject pronoun, on the other hand, is obligatory in English and needs to be inserted.

In example 2 the Spanish sentence contains an *a* preposition which functions as a Direct Object marker and must not appear in the English version. Likewise, example 3 illustrates differences in the use of articles in the two languages: generic sentences in English require bare plural nouns while in Spanish the definite article is obligatory.

The way the search algorithm described in 2.4 is intended to deal with insertion and deletion is that the presence or absence of grammatical words does not hinder *n*-gram matching. Grammatical words are part of the model but they function as if they were effectively *invisible* in the same way that inflection is generally not used when searching for an *n*-gram candidate.

3.2 Local movements

We distinguish between local and non-local movements. Local movements are changes in the order of individual words that occur inside a linguistic constituent, such as an NP. Non-local movements affect reordering of constituents in the sentence. We address non-local movements in 3.3.

As stated above, only content words are allowed to move. Major categories, such as nouns, verbs, adjectives and adverbs are considered to be *content words*. As a way of example, let us look at the reordering of adjectives inside an NP.

(4) la guerra civil española
the war civil Spanish
'the Spanish Civil war'

Reordering of translated adjectives in example 4 would require in traditional linguistic MT systems information about scope and/or type of the adjectives and position in the source sentence. In a statistical system such as ours, the adjectives, together with the noun, are freely allowed to move, thus expanding the *n*-gram set. The correct order is eventually validated by the target language model. In contrast, the determiner (being a grammatical word) is not allowed to move.

Certainly, we do not want the adjectives to move *outside* of the boundaries of the NP. How do we achieve such a restriction, considering that we are not using any kind of parser or chunker, but only a POS tagger?

In order to detect linguistically significant constituents, we mirror a chunking procedure in which we pre-define phrase boundary markers. For instance, *Det* is a boundary marker, and so is *Verb-FIN* and *Prep*. Content words are only allowed to move inside two consecutive boundaries.

Another example of what can be achieved by our approach is the translation of noun complements, which in Spanish tend to appear after the head noun, preceded by a *de* preposition, and in English appear as a noun pre-modifying the head. Example 5 is an illustration of this, which includes both reordering and deletion.

(5) un regalo de cumpleaños
a present of birthday
'a birthday present'

3.3 Non-local movements

The procedure described in the previous section is insufficient when changes in order are not local but affect sentence constituents. This happens only occasionally when translating from Spanish to English, e.g. different position of the adverb, subject inversion, etc., but is particularly frequent when going from German to English.

For instance, sentence 6 would not be correctly handled by our system, such as it has been described so far:

(6) In dem Garten isst der
In the-DAT garden eats the-NOM
junge Mann
young man
'The young man eats in the garden'

In German, the finite verb in main clauses must always be in second position, regardless of which kind of constituent occurs in the first position of the clause. In example 6, a locative adjunct (marked in dative case) occurs in first position, and the subject (marked in nominative case) after the verb. This order needs to be reversed in the translation (or at least the subject has to be placed before the verb).

To handle non-local order changes, we propose creating a "second-level language model" apart from the token level language model described in Section 2.3. This is an n-gram model over *sequences of tags*. The tags in this model are complex tags of the type: *DetAdjNoun*. Sequences of tags are limited by the same type of boundaries described in Section 3.2. In this way, the 'second-level' language model gives us a parser-free representation of the syntactic patterns of the target language.

Boundary detection is performed on the output of the lemma-to-lemma dictionary look-up. The result of boundary detection on the lemma-to-lemma translation of the original sentence in example 6 is shown in 7, where * marks the boundaries:

(7) In the garden eats the
*Prep Det Noun *Verb-FIN *Det
young man
Adj Noun

This sequence of complex tags would then be checked against the 'second-level' or 'syntactic model', yielding as a result the most frequent of all possible permutations, in our example, (8c).

(8) a. *PrepDetNoun *Verb-FIN *DetAdjNoun
b. *DetAdjNoun *PrepDetNoun *Verb-FIN
c. *DetAdjNoun *Verb-FIN *PrepDetNoun

As a further enhancement, lexical information could be introduced into this model, most prominently verbal lemmas. In this way, subcategorization information could be taken into account, defining syntactic patterns over verb types. Without this information, interference between different subcategorization frames could bring noise into the model, for example, intransitive structures would give wrong models for transitive verbs. If this solution makes the model too sparse, an alternative would be to build the model with clusters (Resnik, 1993).

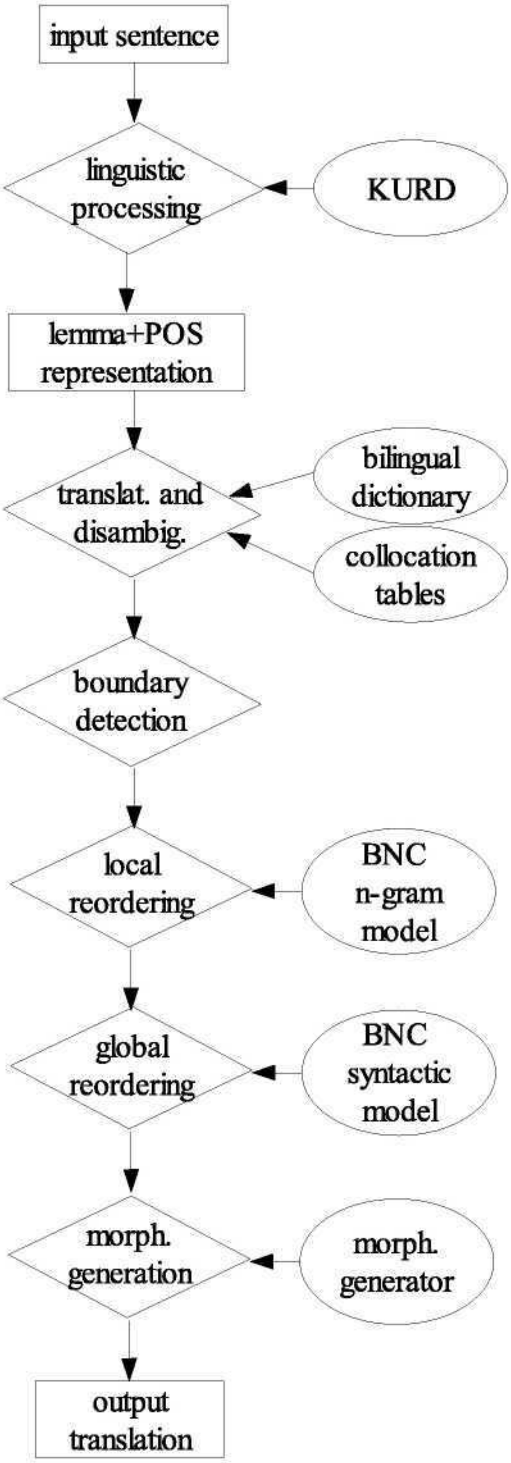

Figure 1: Proposed architecture of the Spanish-English n-gram based translation system

4 Use of a table of collocations to optimize lexical selection

To choose among different translations of a word, or at least to discard many translations, less information than a full language model is arguably needed, if a collocation module is built that exploits the information in the target corpus relative to selectional restrictions. A similar approach, using co-occurrence statistics in a target reference corpus, has been exploited in cross-language information retrieval by (Qu et al., 2002).

We plan to extract from the corpus a table of collocations for certain pairs of POS: Verb+Noun, Adj+Noun, Noun+Noun, and Verb+Adv.[6] The frequencies of the lemma pairs in an n-word window will be collected, associated with one of several possible measures for collocation detection (Evert and Krenn, 2001), and stored in a table.

Our goal is not to just store collocations, but to more generally model selectional restrictions. In cases such as the example *el niño pequeño come carne* cited above, the pair *eat-meat* will presumably have a higher score than *eat-flesh*, so that *flesh* does not need to enter into the n-gram building process. In this way we expect to help discard some of the lexical combinations resulting from the dictionary look-up, prior to actually doing the n-gram search on the model. In cases where the collocation table does not provide enough evidence, the remaining translations can still be validated with the general translation algorithm.

5 Conclusions

In this paper we have presented an experiment, which is being carried out in the context of Metis-II, to translate from Spanish to English using very basic linguistic resources, namely a POS tagger and lemmatizer for Spanish, a machine readable bilingual dictionary and the tagged and lemmatized version of the British National Corpus. Its architecture, as shown in fig 1, is thus translatable to languages with very little NLP development.

The target corpus is the basis both for lexical selection (selecting among the different translations found in the dictionary) and for structure construction (allowing for both local and global changes in structure). To that end, the building and exploitation of the following models will be explored:

- an n-gram language model over lemma and tag tokens. This model should allow for an efficient treatment of common structural changes in translation, involving insertion,

[6] For simplicity reasons we will only encode binary collocations, at least in a first step.

deletion and local movement. This treatment will make crucial use of the distinction between grammatical and content words, provided by the POS tagger;

- a syntactic model over sequences of tags within sentences, as a representation of the syntactic patterns of the target language, to deal with global movement;
- a collocation table to account for selectional restrictions.

The use of the models as explained in this paper has been designed to dispense with explicit mapping rules, or at least keep them to a really minimal set. If these models can be conveniently exploited, it would be an enormous boost to the scalability and robustness of the system.

The implemented version of the system is still too immature to perform a meaningful evaluation. However, we have discussed promising lines of research to build a full-fledged system which can eventually be evaluated analogously to other MT systems.

References

M. Banko and E. Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of 39th Annual Meeting of ACL*, pages 26–33, Toulouse, France.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter stimation. *Computational Linguistics*, 19(2):263–311.

Ralf Brown. 2003. Clustered transfer rule induction for example-based translation. In Michael Carl and Andy Way, editors, *Recent Advances in Example-Based Machine Translation*, pages 287–306. Kluwer Academic Publishers.

L. Burnard. 1995. *Users reference guide for the British National Corpus*. OUCS.

M. Carl and A. Schmidt-Wigger. 1998. Shallow post morphological processing with kurd. In *Proceedings of NeMLaP'98*, pages 5–11, Sydney, Australia.

M. Carl and A. Way. 2003. *Recent Advances in Example-Based Machine Translation*. Kluwer Academic Publishers.

Michael Carl. 2003. Inducing translation grammars from bracketed alignments. In Michael Carl and Andy Way, editors, *Recent Advances in Example-Based Machine Translation*, pages 339–364. Kluwer Academic Publishers.

Yannis Dologlou, Stella Markantonatou, George Tambouratzis, Olga Yannoutsou, Athanassia Fourla, and Nikos Ioannou. 2003. Using monolingual corpora for statistical machine translation: The metis system. In *Proceedings of the EAMT-CLAW 03:Controlled Language Translation*, pages 61–68, Dublin City University, Dublin, Ireland.

Stefan Evert and Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th ACL*, pages 188–195.

P. Koehn. 2002. *Europarl: a multilingual corpus for evaluation of machine translation*. Unpublished.

Y. Qu, G. Gefenstette, and D. A. Evans. 2002. Resolving translation ambiguity using monolingual corpora. In *Working Notes for the CLEF 2002 Workshop*, pages 115–126.

Philip Resnik. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania.

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.

D. Turcato and F. Popowich. 2001. What is example-based machine translation? In *Proceedings of the Machine Translation Summit VIII*, Santiago de Compostela, Spain.

Hideo Watanabe, Satoshi Kurohshi, and Eiji Aramaki. 2003. Finding translation patterns from dependency structures. In Michael Carl and Andy Way, editors, *Recent Advances in Example-Based Machine Translation*, pages 287–306. Kluwer Academic Publishers.

K. Yamada and K. Knight. 2001. A syntax-based statistical translation model. In *Proceedings of 39th Annual Meeting of ACL*, pages 5–11, Toulouse, France.