Multi-grained alignment of parallel texts with endogenous resources

Emmanuel Giguet
Emmanuel.Giguet@info.unicaen.fr
GREYC UMR 6072
Université de Caen, France

Abstract

This paper deals with the spotting of multi-grained translation equivalents from parallel corpora. The idea is to contribute to the processing of languages for which few linguistic resources are available.
We especially pay attention to the handling of highly inflectional languages. Our approach is endogenous: it does not require external linguistic resources such as stemmers or taggers.

1. Introduction

1.1. Automatic processing of bilingual and multilingual corpora

Processing bilingual and multilingual corpora constitutes a major area of investigation in natural language processing. The linguistic and translational information that is available make them a valuable resource for translators, lexicographers as well as terminologists. They also constitute the nucleus of example-based machine translation and translation memory systems.

Corpora of this type are today freely available and their increasing size demands the exploration of automatic methods.

1.2. Multi-grained alignment

Alignment becomes an important issue for research on bilingual and multilingual corpora. Existing alignment methods define a continuum going from purely statistical methods to linguistic ones. A major point of divergence is the granularity of the proposed alignments (entire texts, paragraphs, sentences, clauses, words) which often depends on the application.

In a coarse-grained alignment task, punctuation or formatting can be sufficient. At finer-grained levels, methods are more sophisticated and combine linguistic clues with statistical ones. Statistical alignment methods at sentence level have been thoroughly investigated (Gale & Church, 1991b ; Brown et al., 1991 ; Kay & Röscheisen, 1993). Others use various linguistic information (Simard et al., 1992 ; Papageorgiou et al., 1994). Purely statistical alignment methods are proposed at word level (Gale & Church, 1991a ; Kitamura & Matsumoto, 1995). (Tiedemann, 1993 ; Boutsis & Piperidis, 1996 ; Piperidis et al., 1997) combine statistical and linguistic information for the same task. Some methods make alignment suggestions at an intermediate level between sentence and word (Smadja, 1992 ; Smadja et al., 1996 ; Kupiec, 1993 ; Kumano & Hirakawa, 1994 ; Boutsis & Piperidis, 1998).

A common problem is the delimitation and spotting of the units to be matched. This is not a real problem for methods aiming at alignments at a high level of granularity (paragraphs, sentences) where unit delimiters are clear. It becomes more difficult for lower levels of granularity (Simard, 2003), where correspondences between graphically delimited words are not always satisfactory.

2. The alignment approach

2.1. Endogenous approach

The approach proposed here deals with the spotting of multi-grained translation equivalents. We do not adopt very rigid constraints concerning the size of linguistic units involved, in order to account for the flexibility of language and translation divergences. Alignment links can then be established at various levels, from sentences to words and obeying no other constraints than the maximum size of candidate alignment sequences and their minimum frequency of occurrence.

At the preprocessing stage, input texts have been segmented and aligned at sentence level. They do not contain any syntactical annotation, and they have not been lemmatised. Inflectional divergencies of isolated words are taken into account without external linguistic information (lexicon) and without linguistic parsers (stemmer or tagger). The morphology is learnt automatically using an endogenous parsing module integrated in the alignment tool (Déjean, 1998). When occurring, wrong morphemes are filtered by the alignment algorithm.

2.2. The grammatical way

We strive to keep to a minimalist approach, in the line of GREYC. Therefore we avoid using a large amount of *a priori* knowledge on the languages of the texts to be aligned. In fact, many languages do not have available linguistic resources for automatic processing, neither inflectional or syntactical annotation, nor surface syntactic analysis or lexical resources (machine-readable dictionaries etc.). Moreover, we think that studying the contribution of grammatical structures is an attractive way, leading to multilingual processing.

3. Considerations on the Corpus

3.1. Corpus definition

The current version of our alignment system deals with one language pair at a time, whatever the languages are. It requires a corpus of bitexts aligned at sentence level. The bitext is a quadruple < *T1, T2, Fs, C*> where *T1* and *T2* are the two texts, *Fs* is the function that reduces *T1* to an element set *Fs(T1)* and *T2* to an element set *Fs(T2)*, and *C* is a subset of the Cartesian product of *Fs(T1) x Fs(T2)* (Harris, 1988).

Concretely, the texts constituting the input corpus are documents written in one source language (for instance, scientific papers or technical reports), and their translation, written in a target language. Thus, it is a bi-directional corpus.

3.2. Corpus preparation

The algorithm takes as input a corpus of bitexts aligned at sentence level. Usually, the alignment at this level outputs aligned windows containing from 0 to 2 segments. One-to-one mapping corresponds to a standard output. An empty window corresponds to a case of addition in the source language or to a case of omission in the target language. One-to-two mapping corresponds to split sentences.

The alignment at sentence level is obtained automatically by sentence a alignment tool. For instance, the translation memory system TrAid (Triantafyllou et al., 2000) includes an independent sentence aligner. The algorithm used in this system is based on the statistical model of character lengths proposed by (Gale & Church, 1991a). Sentence segmentation and sentence alignment are the only required preprocessing stages.

Our system natively handles TMX and XCES file format, with UTF-8 or UTF-16 encoding.

4. The Resolution Method

The resolution method is composed of two stages, based on two underlying hypotheses. The first stage handles the document grain. The second stage handles the corpus grain.

4.1. Hypothesis

hypothesis 1 : let's consider a bitext composed of the texts T_1 and T_2. If a sequence S_1 is repeated several times in T_1 and in well-defined sentences[1], there are many chances that a repeated sequence S_2 corresponding to the translation of S_1 occurs in the corresponding aligned sentences in T_2.

hypothesis 2 : let's consider a corpus of bitexts, composed of two languages L_1 and L_2. There is no guarantee for a sequence S_1 which is repeated in many texts of language L_1 to have a unique translation in the corresponding texts of language L_2.

4.2. Stage 1 : Bitext analysis

The first stage handles the document scale. Thus it is applied on each document, individually. There is no interaction at corpus level.

Multi-grained Alignment Method

In the beginning of the alignment procedure we check the internal coherence of every bitext, making sure that every textual segment (or sentence) in one language has its corresponding segment in the other language. Segments aligned with null segments are removed (in case of omission or addition in one language). The equality of the number of segments is important: it is necessary for the construction of an orthonormal space. Obviously, this verification is not useful if the alignment tools used at sentence level guarantee the equality of the number of segments in the two languages.

Determining the sequences to be aligned

Hence, we consider the two languages of the document independently, the source language L1 and the target language L2. For each language, we compute the repeated sequences as well as their frequency. The settings of the underlying algorithm are the minimum and maximum number of words forming the sequences, as well as the minimum frequency of the sequences that must be kept.

We use a greedy algorithm, similar to the algorithm used by (Vergne, 2005) for term extraction. The idea is to keep sequences of 1, 2, 3, … words, while the sequence frequency in the document is greater than a particular threshold (1 for instance).

The algorithm does not retain the sub-sequences of a repeated sequence if they are as frequent as the sequence itself. For instance, if "*subjects*" appears with the same frequency than "*healthy subject*s" we retain only the second sequence. On the contrary, if "*disease*" occurs more frequently than "*thyroid disease*" we retain both.

When computing the frequency of a repeated sequence, the offset of each occurrence is memorized. So the output of this processing stage is a list of sequences with their frequency and the offset list in the document.

Handling inflections

Inflectional divergencies of isolated words are taken into account without external linguistic information (lexicon) and without linguistic parsers (stemmer or tagger). The morphology is learnt automatically using an endogenous approach derived from (Déjean, 1998). The algorithm is reversible: it allows to compute prefixes the same way, with reversed word list as input.

The basic idea is to approximate the border between the nucleus and the suffixes. The border matches the position where the number of distinct letters preceding a suffix of length n is greater than the number of distinct letters preceding a suffix of length n-1.

For instance, in the first English document of our corpus, "*g*" is preceded by 4 distinct letters, "*ng*" by 2 and "*ing*" by 10: "*ing*" is probably a suffix. In the first Greek document, "*ά*" is preceded by 5 letters, "*κά*" by 1 and "*ικά*" by 10. "*ικά*" is probably a suffix.

[1] Here, « sentences » can be generalized as « textual segments »

The algorithm can generate some wrong morphemes, from a strictly linguistic point of view. But at this stage, no filtering is done in order to check their validity. We let the alignment algorithm do the job with the help of contextual information.

Vectorial representation of the sequences

As stated earlier, the equal number of segments in the two languages allows the construction of an orthonormal space. This space can be used in order to explore the existence of translation relations between the sequences and define translation couples. For the construction of this space we pick up the segment offset in the document for each occurrence.

"thyroid cancer": list of segments where the sequence appears
45, 46, 46, 48, 51, 51

Then we convert this list in a *n*-dimension vector (where *n* corresponds to the number of textual segments of the corpus). Each dimension contains the number of occurrences present in the segment.

"*thyroid cancer*" : *associated with a vector of 63 dimensions*.

1	2	…	45	46	47	48	49	50	51	…	63
0	0		1	2	0	1	0	0	2		0

Sequence alignment

For each sequence of L_1 to be aligned, we look for the existence of a translation relation between it and every L_2 sequence to be aligned. The existence of a translation relation between two sequences is approximated by the cosine of the vectors associated to them.

The cosine is a mathematical tool used in in Natural Language Processing for various purposes, e.g. (Roy & Beust, 2004) uses the cosine for thematic categorisation of texts. The cosine is obtained by dividing the scalar product of two vectors with the product of their norms.

$$\cos(x_i,y_i)=\frac{\sum x_i \cdot y_i}{\sqrt{\sum x_i^2} \times \sqrt{\sum y_i^2}}$$

We note that the cosine is never negative as vectors coordinates are always positive. The sequences proposed for the alignment are those that obtain the largest cosine. We do not propose an alignment if the best cosine is inferior to a certain threshold.

4.2. Stage 2 : Corpus management

The second stage handles the corpus grain and merges the information found at document grain, in the first stage.

Handling the Corpus Dimension

The bitext corpus is not a bag of aligned sentences and is not considered as if it were. It is a bag of bitexts, each bitext containing a bag of aligned sentences.

Considering the bitext level (or document grain) is useful for several reasons. First, for operational sake. The greedy algorithm for repeated sequence extraction has a cubic complexity. It is better to apply it on the document unit rather than on the corpus unit. But this is not the main reason.

Second, the alignment algorithm between sequences relies on the principle of translation coherence: a repeated sequence in L1 has many chances to be translated by the same sequence in L2 in the same text. This hypothesis holds inside the document but not in the corpus: a polysemic term can be translated in different ways according to the document genre or domain.

Third, the confidence in the generated alignments is improved if the results obtained by the execution of the process on several documents share compatible alignments.

Alignment Filtering and Ranking

The filtering process accepts terms which have been produced (1) by the execution on at least two documents, (2) by the execution on solely one document if the aligned terms correspond to the same character string or if the frequency of the terms is greater than an empirical threshold

function. This threshold is proportional to the inverse term length since there are fewer complex repeated terms than simple terms.

The ranking process sorts candidates using the product of the term frequency by the number of output agreements.

5. Results

This section is problematic. We carried an evaluation of our system but our rights to use the bitext corpus expired. Right now, we are not allowed to give any information about this corpus.

The evaluation concerns an alignment task between a rich inflectional language and a weak inflectional language, carried on scientific papers.

5.2. Quantitative study of proposed alignments

An evaluation has been performed on monograms:

- number of words of the sequences: 1
- minimum cosine : 0.9
- minimum frequency: 2

The system proposed 400 word alignments on the 19 documents of the corpus (Giguet & Apidianaki, 2005). It achieves 100% decision: 1 word from language L1 is always mapped to 1 word from language L2. It achieves 92% precision: number of correct alignments.

Among the correct alignments, we mainly find domain dependant lexical terms (*κορτιζόλης* / cortisol, *ομοκυστεΐνης* / homocysteine) and invariant terms (min / min, SH / SH, vitro / vitro).

The wrong alignments mainly come from candidates that have not been confirmed by running on several documents.

5.2. Discussion

The main drawback is silence. First, we note that possible alignments of grammatical words are not generated. The difference in linguistic roles carried by grammatical words is the problem. For instance, the English grammatical word “the” has 3 inflectional variations (gender and number) in French : “le” “la” et “les” (lemma = “le”). But the function carried by both words is not the same, “the” being used under more constraints than “le”.

Second, thematic terms of the corpus are not always aligned, since they are not repeated. Corefence is used instead, thanks to nominal anaphora, acronyms, and also lexical reductions. Accuracy depends on the document domain. In the medical domain, acronyms are aligned but not their expansion. However, we consider that this problem has to be solved by an anaphora resolution system, not by this alignment algorithm.

6. Conclusion

We showed that it is possible to contribute to the processing of languages for which few linguistic resources are available. We propose a solution to the spotting of multi-grained translation from parallel corpora.

We use an endogenous approach in order to handle inflectional variations. We also show the importance of using the proper knowledge at the proper level (sentence grain, document grain and corpus grain).

The results are surprisingly good, considering the precision rate.

The next improvement is to properly handle translations in multiple languages. An effort should be made to reduce silence. Another perspective is to integrate an endogenous coreference solver (Giguet & Lucas, 2004).

References

(Altenberg & Granger, 2002) Altenberg B. & Granger, S. *Recent trends in cross-linguistic lexical studies.* In *Lexis in Conrast,* Altenberg & Granger (eds.), 2002.

(Boutsis & Piperidis, 1998) Boutsis, S., & Piperidis, S. *Aligning clauses in parallel texts.* In *Third Conference on Empirical Methods in Natural Language Processing,* 2 June, Granada, Spain, p. 17-26, 1998.

(Brown et al., 1991b) Brown P., Lai J. & Mercer R. *Aligning sentences in parallel corpora*. In *Proc. 29th Annual Meeting of the Association for Computational Linguistics*, p. 169-176, 18-21 June, Berkley, California, 1991.

(Déjean, 1998) Déjean H. *Morphemes as Necessary Concept for Structures Discovery from Untagged Corpora.* In Workshop on Paradigms and Grounding in Natural Language Learning, pages 295-299, PaGNLL Adelaide.

(Gale & Church. 1991a) Gale W.A. & K.W. Church, *Identifying word correspondences in parallel texts*. In *Fourth DARPA Speech and Natural Language Workshop*, p. 152-157. San Mateo, California: Morgan Kaufmann, 1991.

(Gale & Church, 1991b) Gale W.A. & Church K. W. *A Program for Aligning Sentences in Bilingual Corpora.* In *Proc. 29th Annual Meeting of the Association for Computational Linguistics*, p. 177-184, 18-21 June, Berkley, California, 1991.

(Giguet & Apidianaki, 2005) Giguet E. & Apidianaki M. *Alignement d'unités textuelles de taille variable*. Journée Internationales de la Linguistique de Corpus. Lorient.

(Giguet & Lucas, 2004) Giguet E. & Lucas N. *La détection automatique des citations et des locuteurs dans les textes informatifs*. In *Le discours rapporté dans tous ses états : Question de frontières*, J. M. López-Muñoz S. Marnette, L. Rosier, (eds.). Paris, l'Harmattan, pp. 410-418. 2004

(Harris, 1998) Harris B. *Bi-text, a New Concept in Translation Theory,* Language Monthly (54), p. 8-10, 1998.

(Isabelle, 1993) Isabelle P. & Warwick-Armstrong S. *Les corpus bilingues: une nouvelle ressource pour le traducteur.* In Bouillon, P. & Clas A. (eds.), *La Traductique : études et recherches de traduction par ordinateur.* Montréal : Les Presses de l'Université de Montréal, 1993, p. 288-306.

(Kay & Röscheisen, 1993) Kay M. & Röscheisen M. *Text-translation alignment*. Computational Linguistics, p.121-142, March 1993.

(Kitamura & Matsumoto, 1996) Kitamura M. & Matsumoto Y. *Automatic extraction of word sequence correspondences in parallel corpora*. In *Proc. 4th Workshop on Very Large Corpora*, p. 79-87. Copenhagen, Denmark, 4 August 1996.

(Kupiec, 1993) Kupiec J. *An algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora*, *Proceedings of the 31st Annual Meeting of the Association of Computational Linguistics*, p. 23-30, 1993.

(Papageorgiou et al., 1994) Papageorgiou H., Cranias L. & Piperidis S. *Automatic alignment in parallel corpora*. In *Proceed. 32nd Annual Meeting of the Association for Computational Linguistics*, p. 334-336, 27-30 June, Las Cruses, New Mexico, 1994.

(Salkie, 2002) Salkie R. *How can linguists profit from parallel corpora?,* In *Parallel Corpora, Parallel Worlds: selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22-23 April, 1999,* Lars Borin (ed.), Amsterdam, New York: Rodopi, 2002, p. 93-109.

(Simard, 1992) Simard M., Foster G., & Isabelle P. *Using cognates to align sentences in bilingual corpora*. In *Proceedings of TMI-92*, Montréal, Québec, 1992.

(Simard, 2003) Simard M. *Mémoires de Traduction sous-phrastiques.* Thèse de l'Université de Montréal, 2003.

(Smadja, 1992) Smadja F. *How to compile a bilingual collocational lexicon automatically.* In *Proceedings of the AAAI-92 Workshop on Statistically -based NLP Techniques,* 1992.

(Smadja et al., 1996) Smadja F., McKeown K.R. & Hatzivassiloglou V. *Translating Collocations for Bilingual Lexicons: A Statistical Approach*, Computational Linguistics. March, p. 1-38, 1996.

(Tiedemann, 1993) Tiedemann J. *Combining clues for word alignment*. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, p. 339-346, Budapest, Hungary, April2003.

(Triantafyllou et al., 2000) Triantafyllou I., Demiros I., Malavazos C. & Piperidis S. *An alignment architecture for Translation Memory bootstrapping*. In *MT 2000*.

(Vergne, 2005) Vergne J. *Une méthode indépendante des langues pour indexer les documents de l'Internet par extraction de termes de structure contrôlée.* In Conférence Internationale sur le Document Electronique. Beyrouth, Liban. May, 2005