A Hybrid Word Alignment Approach to Improve Translation Lexicons with Compound Words and Idiomatic Expressions

Nasredine Semmar (1), Christophe Servan (1), Gaël de Chalendar (1), Benoît Le Ny (2), Jean-Jacques Bouzaglou (2)

(1) CEA, LIST, Vision and Content Engineering Laboratory, 18 route du Panorama, Fontenay-aux-Roses, F-92265, France

(2) Softissimo, 5 rue Soyer, Neuilly-sur-Seine, F-92200, France

Abstract

In this paper, we present a hybrid approach to align single words, compound words and idiomatic expressions from bilingual parallel corpora. The objective is to develop, improve and maintain automatically translation lexicons. This approach combines linguistic and statistical information in order to improve word alignment results. The linguistic improvements taken into account refer to the use of an existing bilingual lexicon, named entities recognition, grammatical tags matching and detection of syntactic dependency relations between words. Statistical information refer to the number of occurrences of repeated words, their positions in the parallel corpus and their lengths in terms of number of characters. Single-word alignment uses an existing bilingual lexicon, named entities and cognates detection and grammatical tags matching. Compound-word alignment consists in establishing correspondences between the compound words of the source sentence and the compound words of the target sentences. A syntactic analysis is applied on the source and target sentences in order to extract dependency relations between words and to recognize compound words. Idiomatic expressions alignment starts with a monolingual term extraction for each of the source and target languages, which provides a list of sequences of repeated words and a list of potential translations. These sequences are represented with vectors which indicate their numbers of occurrences and the numbers of segments in which they appear. Then, the translation relation between the source and target expressions are evaluated with a distance metric. The single and compound word aligners have been evaluated on a subset of 1103 sentences in English and French of the JOC (Official Journal of the European Community) corpus . The obtained results showed that these aligners generate a translation lexicon with 90 % of precision for single words and 84 % of precision for compound words. We evaluated the idiomatic expressions aligner on a subset of the Canadian Parliament Hansard corpus and we obtained a precision of 81%.

1 Introduction

Bilingual lexicons play a vital role in machine translation (MT) and cross-language information retrieval (CLIR). Word alignment approaches are generally used to construct bilingual lexicons [1]. Existing word alignment tools such as Giza++ [2] are efficient only for aligning single words. Approaches and tools for aligning multi-word units such as compound words, terms and idiomatic expressions are at experimental stage [3].

This paper aims to describe a hybrid approach combining linguistic and statistical methods to align simple and complex words from parallel texts.

We present in section 2 the state of the art of aligning words from parallel text corpora. In section 3, the main steps for automatic construction of translation lexicons are described; we will focus, in particular, on the word alignment process. We discuss in section 4 results obtained after aligning simple and complex words from parallel corpora. Section 5 concludes our study and presents our future work.

2 Previous work

There are mainly three approaches for word alignment using parallel corpora:

- Statistical approaches are generally based on IBM models [4].
- Linguistic approaches for single words and compound words alignment use bilingual lexicons and morpho-syntactic analysis on source and target sentences in order to obtain grammatical tags of words and syntactic dependency relations [5].
- A combination of the two previous approaches [6, 7, 8, 9]. Gaussier's approach [7] is based on a statistical model to establish the French and English word associations. It uses the dependence properties between words and their translations. Ozdowska's approach [10] consists in matching words regards to the whole corpus, using the co-occurrence frequencies in aligned sentences. These words are used to create couples which are starting points for the propagation of matching links by using dependency relations identified by syntactic analysis in the source and target languages.

The most popular word alignment tool is Giza++. This tool implements statistical approaches based on IBM models but its performance is proved only for aligning single words.

3 Steps for automatic construction of translation lexicons

Automatic building of bilingual lexicons using word alignment approaches is generally composed of the following steps:

- Sentence alignment;
- Word alignment;
- Cleaning and validating the generated bilingual lexicon.

This paper addresses only the first two steps.

3.1 Pre-processing the bilingual parallel corpus

A bilingual parallel corpus is an association of two texts in two languages, which represent translations of each other. In order to use this corpus in word alignment, two pre-processing tasks are involved on the two texts: sentence alignment and linguistic analysis.

3.1.1 Sentence alignment

Sentence alignment consists in mapping sentences of the source language with their translations in the target language. Our approach to align the sentences of the bilingual parallel corpus combines different information sources (bilingual lexicon, sentence length and sentence position) and is based on cross-language information retrieval which consists in building a database of sentences of the target text and considering each sentence of the source text as a "query" to that database [11]. This approach uses a similarity value to evaluate whether the two sentences are translations of each other. This similarity is computed by the comparator of the cross-language search engine and consists in identifying common words between source and target sentences. This search engine is composed of a deep linguistic analysis, a statistical analysis to attribute a weight to each word of the sentence, a comparator and a reformulator to translate the words of the source sentence in the target language by using a bilingual lexicon.

3.1.2 Linguistic analysis

Linguistic analysis consists in producing for a given text a set of normalized lemmas, a set of named entities and a set of compound words with their grammatical tags [12]. We used the CEA LIST Multilingual Analysis platform (LIMA) which is composed of a tokenizer, a morphological analyzer, a part-of-speech tagger, a named entity recognizer and a syntactic analyzer.

3.2 Word alignment

Word alignment consists of finding correspondences between single words, compound words and idiomatic expressions in a sentence aligned bilingual corpus. Our word alignment approach uses:

- an existing bilingual lexicon, linguistic properties of named entities and cognates to align single words,
- syntactic dependency relations to align compound words,
- sequences of words repeated in the bilingual corpora and their occurrences to align idiomatic expressions.

3.2.1 Single-word alignment

The single-word alignment is composed of the following steps:

- Alignment using the existing bilingual lexicon;
- Alignment using the detection of cognates;
- Alignment using the detection of named entities;
- Alignment using grammatical tags of words.

3.2.1.1 Bilingual lexicon look-up

Alignment using the existing bilingual lexicon consists in extracting for each word of the source sentence the appropriate translation in the bilingual lexicon. The result of this step is a list of lemmas of source words for which one or more translations were found in the bilingual lexicon. For example, Table 1 shows the result of this step for the English sentence "*Social security funds in Greece are calling for independence with regard to the investment of capital.*" and its French translation "*Les caisses de sécurité sociale de Grèce revendiquent l'indépendance en matière d'investissements.*".

Lemmas of the words of the source sentence	**Translations found in the bilingual lexicon**
security	sécurité
fund	caisse
Greece	Grèce
independence	indépendance
investment	investissement

Table 1: Single-word alignment with the existing bilingual lexicon.

3.2.1.2 Cognate detection

For those words that are not found in the bilingual lexicon, the single-word aligner searches cognates (pairs of words which share the first four characters) among not assigned target words. The result of this step is a one-to-one word mapping. For example, for the previous English sentence and its French translation, the single-word aligner detects that the lemma of the English word "*social*" is a cognate of the lemma of the French word "*social*".

3.2.1.3 Named entities detection

This step consists in searching named entities present in the source and target sentences. For example, for the previous English sentence and its French translation, the single-word aligner detects that the English word "*Greece*" and the French word "*Grèce*" are named entities. However, this step can produce alignment errors in the case the source and target sentences contain several named entities. To avoid these errors, we added a criterion related to the position of the named entity in the sentence.

3.2.1.4 Grammatical tags matching

If for a given word no translation is found in the bilingual lexicon and no named entities are present in the source and target sentences, the single-word aligner tries to use grammatical tags of source and target words. This is especially the case when the word to align is surrounded with some words already aligned. For example, because the grammatical tags of the words "*calling for*" and "*revendiquent*" are the same (Verb) and "*calling for*" is surrounded with the words "*Greece*" and "*independence*" which are already aligned in the previous steps, the single-word aligner considers that the lemma "*revendiquer*" is the translation of the lemma "*call for*".

3.2.2 Compound-word alignment

Compound-word alignment consists in establishing correspondences between the compound words of the source sentence and the compound words of the target sentences. First, a syntactic analysis is applied on the source and target sentences in order to extract dependency

relations between words and to recognize compound words structures. Then, reformulation rules are applied on these structures to establish correspondences between the compound words of the source sentence and the compound words of the target sentence.

For example, the rule Translation(A.B) = Translation(B).Translation(A) allows to align the English compound word "*social security*" with the French compound word "*sécurité sociale*" as follows:

Translation(social.security) = Translation(security).Translation(social) = sécurité.sociale.

Table 2 presents results after running single-word and compound-word alignment processes on the previous example.

Lemmas of the words of the source sentence	**Translations of the lemmas in of the target language**
social	social
security	sécurité
fund	caisse
Greece	Grèce
call_for	revendiquer
independence	indépendance
investment	investissement
capital	capital
security_social	sécurité_social
security_social_Greek	sécurité_social_Grèce
fund_security_social	caisse_sécurité_social
fund_security_social_Greek	caisse_sécurité_social_Grèce

Table 2: Results after running single-word and compound-word alignment processes.

3.2.3 Idiomatic expressions alignment

The approaches used for single and compound word alignment were not developed for the alignment of more general collocations. However, properly aligning relevant multi-word units is necessary for the construction of bilingual lexicons. The algorithm we use is based on a statistical approach that requires only shallow parsing, and is mostly language independent contrary to the techniques used for single-word alignment. Moreover, the collocation alignment approach only requires "similar" corpora, as it is very tolerant on the original text/sentence alignment.

3.2.3.1 Collocation alignment algorithm

The collocation alignment algorithm is actually part of a larger framework developed by Softissimo to automatically create bilingual lexicons. This implies first the identification of relevant "terms" to add to the lexicon, and then finding its proper translation. Our approach can be summarized as follows: First, we identify the relevant word groups through the use of n-gram statistics in both the source and target languages. Then for each source "term" extracted we compile a list of candidate translations through the use of two distance metrics. The list of candidates is then pruned through the use of heuristics like the length of each collocation, and a translation is "found" if it satisfies confidence threshold on the distance metric and the heuristics.

The alignment process has the following five steps:

1. Aligning the corpus: The algorithm works on aligned “segments” of text, although it is not required that they be the exact translation of each other. The alignment still works on noisy data although a sentence aligned corpus is of course preferable.

2. Monolingual extraction of collocations: Identify all the n-grams [up to 6-grams under certain conditions] that may represent a collocation. This is done through frequency analysis and heuristic scoring. This step outputs two lists of terms, which we will refer to as SC (collocations in the source language) and TC (collocations in the target language).

3. Frequency distance calculation: For all source collocations in SC, we calculate the distance to each of the target collocations in TC. The main idea of this metric is that if two collocations are translations of each other then they must appear together in the corpus segments, and only together. Their frequency distance is then calculated as follows:

   S: a SL (source language) collocation

   T: a TL (target language) collocation

   f(s): the frequency of the source collocation

   f(t): frequency of the target collocation under consideration

   $$Fd = \frac{|f(s) - f(t)|}{\max(f(s), f(t))}$$

   We see that if T is the translation of S, f(s)=f(t) and we have 0 distance. Also, if two collocations always occur together but one is much more frequent than the other, the distance reaches 1 and they are not considered translations of each other. Here we chose to apply a threshold of 0.25 as the maximum allowable distance. This threshold can be tuned to achieve better precision

4. Co-occurrence distance: The previous step only considers frequencies so it may be possible for two completely unrelated terms to achieve a low distance score. However we also check for a co-occurrence score as follows:

   T: a TL collocation

   Xi: number of occurrences of S in the ith segment of the SL

   Yi: number of occurrences of T in the ith segment of the TL

   N: number of segments

   $$Cd = \frac{\sqrt{\Sigma(X_i - Y_i)^2}}{N}$$

This check allows the rejection of the terms that fortuitously have similar frequency. Since they would not appear in the same segments, the terms Xi-Yi would increase. The candidate list can be ordered through Cd.

5. Pruning last candidates: Once we have an ordered list of target candidates, we remove:

    - The candidates whose length is too different from the source collocation;
    - The candidates who have been previously aligned with another source collocation and where the co-occurrence score was better.

4 Experimental results

The single and compound word aligners have been evaluated on the corpus of the Official Journal of the European Community of the ARCADE II project [13]. This corpus contains written questions asked by members of the European Parliament on a variety of topics and the corresponding answers from the European Commission. The part of the corpus used to evaluate the performance of these aligners is composed of a set of 1103 English sentences aligned to their French counterparts.

Table 3 presents the performance of these two aligners:

Type of the aligner	**Precision**	**Recall**	**F-measure**
Single-word aligner	0.90	0.81	0.85
Compound-word aligner	0.84	0.55	0.66

Table 3: Single and compound word aligners performance.

Analysis of these results shows that 54% of words are aligned with the bilingual lexicon, 8% are aligned with cognates detection and 26% are aligned by using grammatical tags. Consequently, the single-word aligner has added to the bilingual lexicon translations of about 34% of the words of the source sentences. In addition, new compound words and their translations are added to this lexicon by the compound-word aligner.

The collocation aligner has been evaluated on a subset of the Canadian Parliament Hansard. The Hansard Corpus consists of parallel texts in English and Canadian French, drawn from official records of the proceedings of the Canadian Parliament. While the content is therefore limited to legislative discourse, it spans a broad assortment of topics and the stylistic range includes spontaneous discussion and written correspondence along with legislative propositions and prepared speeches. Being one the few freely available French-English corpora, the Hansard has been widely used for language processing evaluations [8]. The sub-corpus we used comprises the first 100 files of the training data as distributed online by the USC Science Institute, for a total of 302 000 aligned sentences.

As a result of our testing, we extracted eight hundred terms with their corresponding translation. This result set was examined by trained linguists with more than 10 years experience in dictionary creation, and especially for machine translation. Result alignments were evaluated in a binary fashion, as either valid of invalid. A very rough estimation of recall was conducted by sampling 10 random pages in the corpus and manually extracting relevant terms. Because of the skilled manpower involved in such evaluation it was not possible to

examine a larger sample. Table 4 shows a sample of the results; Table 5 summarizes the performance of the algorithm in terms of precision and recall.

Source expression	Target expression	Frequency
opposition officielle	official opposition	high
taux de intérêt	interest rates	high
vache à lait	cash cow	low

Table 4: Alignment result analysis.

Precision	Recall	F-measure
0.81	0.38	0.52

Table 5: Collocation aligner performance.

Because of the statistical nature of our algorithm, it tends to perform much better for terms that occur often in the corpus. Therefore it is interesting to see if there happens to be a clear frequency threshold below which aligned collocations should be rejected. To achieve this, the alignments were ordered by frequency and the precision rate was plotted versus the number of collocations considered (Figure 1). The least frequent collocations extracted have only a few occurrences in the corpus so we are testing the full range of possible frequencies.



Figure 1: Precision rate against number of alignments. Aligned terms are ranked from most frequent (item 1) to least frequent.

The plot shows no clear threshold that would allow us to improve precision dramatically without losing a lot of recall performance.

Table 6 shows some incorrect correspondences produced by the algorithm.

Source collocation	Target collocation	Proper alignment	Error type
mouvement coopératif	operative movement	Co - operative movement	Parsing-missing word
étudiants à temps	part time students	Etudiant à temps partiel	Missing word
jeter le bébé	bath water	throw out the baby with the bath water (for “jeter le bébé avec l’eau du bain”)	N-gram size limit
sénateurs non élus	unaccountable senators	unelected senators	Loose translation

Table 6: Some incorrect correspondences produced by the collocation aligner.

The examples chosen here reflect most of the alignment errors. In the first case our tokenizer mistakenly chose to split “Co” and “operative” because of the spaces surrounding the hyphen. Then a second class of error came into play because the algorithm favors similar size collocation. When some collocations are much longer in one language than in the other, we have a truncated alignment as in the first two examples. This might be mitigated by a strategy to “extend” the source or target collocation, whenever it is always accompanied by another word. The third example shows a very long collocation that was split due to the limit in n-gram size. The subsequent alignment therefore aligns one part of the French expression with another of the English one. We would expect this kind of issue to disappear when using longer n-grams. In the last case, the alignment is “right” however because the text is loosely translated we would not want to add such an entry into a bilingual lexicon.

5 Conclusion and future work

In this paper, we have presented a hybrid approach to align simple and complex words from parallel corpora. The results we obtained showed, on the one hand, that around 28 % of the single words of the source sentence and their translations are added to the bilingual lexicon, and, on the other hand, the statistical algorithm for aligning collocations is robust, requires no linguistic knowledge, and can be easily adapted to many language pairs. In future work, we plan to develop strategies and techniques to filter word alignment results in order to clean the bilingual lexicons built automatically and to extend the collocation aligner to deal with the remaining issues limiting precision.

6 Acknowledgements

This research work is supported by the ANR WEBCROSSLING project (ANR - Programme Technologies Logicielles - 2007).

References

[1] I. D. Melamed. Empirical Methods for Exploiting Parallel Texts. MIT Press, 2001.

[2] F. J. Och. GIZA++: Training of statistical translation models. MIT Press http://www.fjoch.com/GIZA++.htm, 2003.

[3] J. DeNero, D. Klein. The Complexity of Phrase Alignment Problems. In Proceedings of the of ACL-2008, 2008.

[4] P. F. Brown, S. A. D. Pietra, , V. J. D. Pietra, R. L. Mercer. The mathematics of statistical machine translation : parameter estimation. Computational Linguistics 19(3), 1993.

[5] F. Bisson. Méthodes et outils pour l'appariement de textes bilingues. PhD Thesis of the Unveristy Paris VII, 2001.

[6] B. Daille, E. Gaussier, J. M. Langé. Towards automatic extraction of monolingual and bilingual terminology. In Proceedings of the 15th International Conference on Computational Linguistics, 1994.

[7] E. Gaussier, J. M. Langé. Modèles statistiques pour l'extraction de lexiques bilingues. Traitement Automatique des Langues, Volume 36. ATALA, 1995.

[8] F. Smadja, K. Mckeown, V. Hatzivassiloglou. Translation Collocations for Bilingual Lexicons: A Statistical Approach. Computational Linguistics 22(1), 1996.

[9] A. M. Barbu. Simple linguistic methods for improving a word alignment. In Proceedings of the 7th International Conference on the Statistical Analysis of Textual, 2004.

[10] S. Ozdowska. Appariement bilingue de mots par propagation syntaxique à partir de corpus français/anglais alignés. In Proceedings of 11ème conférence TALN-RECITAL, 2004.

[11] N. Semmar, C. Fluhr. Arabic to French Sentence Alignment: Exploration of A Cross-language Information Retrieval Approach. In Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, 2007.

[12] R. Besançon, G. de Chalendar, O. Ferret, F. Gara, M. Laïb, O. Mesnard, N. Semmar. LIMA: A Multilingual Framework for Linguistic Analysis and Linguistic Resources Development and Evaluation. In Proceedings of LREC-2010, 2010.

[13] J. Veronis, O. Hamon, C. Ayache, R. Belmouhoub, O. Kraif, D. Laurent, T. M. H. Nguyen, N. Semmar, F. Stuck, W. Zaghouani. Arcade II Action de recherche concertée sur l'alignement de documents et son évaluation. Chapitre 2, Editions Hermès, 2008.