French-English multi-word term alignment based on lexical context analysis

Béatrice Daille*, Samuel Dufour-Kowalski*, Emmanuel Morin*

*University of Nantes
LINA - FRE CNRS 2729
2, rue de la Houssinière - BP 92208
44322 Nantes Cedex 3, France
{daille, dufour, morin}@lina.univ-nantes.fr

Abstract

This article presents a method of extracting bilingual lexica composed of single-word terms (SWTs) and multi-word terms (MWTs) from comparable corpora of a technical domain. First, this method extracts MWTs in each language, and then uses statistical methods to align single words and MWTs by exploiting the term contexts. After explaining the difficulties involved in aligning MWTs and specifying our approach, we show the adopted process for bilingual terminology extraction and the resources used in our experiments. Finally, we evaluate our approach and demonstrate its significance, particularly in relation to non-compositional MWT alignment.

1. Introduction

Traditional research into the automatic compilation of bilingual dictionaries from corpora exploits parallel texts, i.e. a text and its translation (Veronis, 2000). From sentence-to-sentence aligned corpora, symbolic (Carl and Langlais, 2002), statistical (Gaussier and Langé, 1995), or combined (Daille et al., 1994) techniques are used for word and expression alignments.
The use of parallel corpora raises two problems:

- As a parallel corpus is a pair of translated texts, the vocabulary appearing in the translated text is highly influenced by the source text, especially for technical domains;
- such corpora are difficult to obtain for paired languages not involving English.

New methods try to exploit comparable corpora. The main studies concentrate on finding translation candidates for one-item words. The method is based on lexical context analysis and relies on the simple observation that a word and its translation tend to appear in the same lexical contexts. These contexts can be represented by vectors, and each vector element represents a word which occurs within the window of the word to be translated. Translation is obtained by comparing the source context vector to each translation candidate vector after having translated each element of the source vector with a general dictionary. This method is known as the "direct context-vector approach". Using this method, (Fung, 1998) extracts English-Chinese candidate translations from two years of English and Chinese newspaper articles by matching the context vector with 76% precision on the first 20 candidates. From English-German newspaper corpora of 85 million words, (Rapp, 1999) improves the precision to 89% on the first 10 candidates using the same techniques. Cao and Li (2002) adapted this approach to deal with many-to-many word translations. In extracting bilingual nominal phrases belonging to general domains from the web, they obtain a precision of 91% on the first 3 candidates.
Some improvements have been proposed by Déjean and Gaussier (2002) who avoided direct vector translation by using a similarity-vector approach. From English-German medical corpora of 8 million words, they obtain a precision of 84% on the first 10 candidates.

If the results obtained in the field of bilingual lexicon extraction from comparable corpora are promising, they only cover either bilingual single words from general or specialised corpora, or bilingual nominal phrases from general corpora. Our goal is to find translation for multi-word terms (MWTs) from specialised comparable corpora.

If MWTs are more representative of domain specialities than single-word terms (SWTs), pinpointing their translations poses specific problems:

- SWTs and MWTs are not always translated by a term of the same length. For example, the French MWT *peuplement forestier* (2 content words) is translated into English as the SWT *crop* and the French term *essence d'ombre* (2 content words) as *shade tolerant species* (3 content words). This well-known problem, referred to as "fertility", is seldom taken into account in bilingual lexicon extraction, a *word-to-word* assumption being generally adopted.
- When a MWT is translated into a MWT of the same length, the target sequence is not typically composed of the translation of its parts (Melamed, 2001). For example, the French term *plantation énergétique* is translated into English as *fuel plantation* where *fuel* is not the translation of *énergétique*. This property is referred to as "non-compositionality".
- A MWT could appear in texts under different forms reflecting either syntactic, morphological or semantic variations (Jacquemin, 2001; Daille, 2003). Term variations should be taken into account in the translation process. For example, the French sequences *aménagement de la forêt* and *aménagement forestier* refer to the same MWT and are both translated into the same English term: *forest management*.

We propose tackling these three problems, fertility, non-compositionality, and variations, by using both linguistic and statistical methods. First, MWTs are identified in both

the source and target language using a monolingual term extraction program. Second, a statistical alignment algorithm is used to link MWTs in the source language to single words and MWTs in the target language. Our alignment algorithm extracts the words and MWT contexts and proposes translations by comparing source and target words and MWT contexts.

2. Extraction process

We present in this section the bilingual extraction process which is composed of both linguistic and statistical steps:

2.1. Linguistic step

The goal of this step is to identify the set of candidate MWTs in our corpus. The corpus is cleaned and tokenized, then part-of-speech- and lemma-tagged. Then, MWTs are extracted using a terminology extraction program available for French and English: *ACABIT*[1] (Daille, 2003). This program implements shallow parsing and morphological conflating. The different occurrences referring to a MWT or one of its variants are grouped and constitute an unique candidate MWT. Second, morphological analysis is performed to conflate synonymic derivational variants of MWTs such as *acidité du sang* (acidity of the blood) ↔ *acidité sanguine* (blood acidity). Morphological and morphosyntactic variants which introduce a semantic distance are not grouped with the same candidate MWT. For example, French sequences such as *bois chauffé* and *bois non chauffé* reflect two different candidate MWTs linked by an antonymy variation.

In the following steps, we do not consider a unique sequence reflecting a candidate MWT but a set of sequences. We consider only term variants that are grouped under a unique MWT. This grouping of term variations could be interpreted as a terminology normalisation in the same way as lemmatisation at the morphological level.

2.2. Statistical step

The goal of this step, which adapts the similarity vector-based approach defined for single words by Déjean and Gaussier (2002) to MWTs, is to align source MWTs with target single words, SWTs or MWTs. From now on, we will refer to lexical units as words, SWTs or MWTs.

2.2.1. Context vectors

First, we collect all the lexical units in the context of each lexical unit and count their occurrence frequency. For each lexical unit of the source and the target language, we obtain a context vector which gathers the set of co-occurrence units associated with their frequency. We normalise context vectors using an association score such as Mutual Information. In order to reduce the arity of context vectors, we keep only the co-occurrences with the highest association scores.

2.2.2. Similarity vectors

Once context vectors have been built, it is possible to compare the lexical contexts of the set of the lexical units.

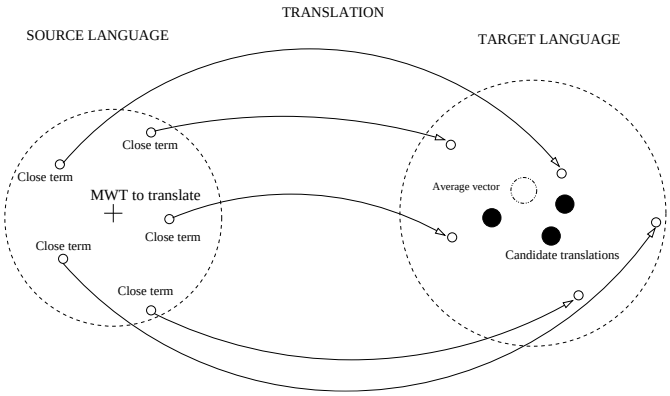

Figure 1: Transfer procedure of similarity vectors from source to target language.

This comparison is done using a vector distance measure such as Cosine Measure. It allows us to group lexical units which share the same lexical contexts and to associate with each lexical unit a similarity vector which gathers the close lexical units associated with their similarity score. In order to reduce the arity of similarity vectors, we keep only the lexical units with the highest similarity scores. Up to now, similarity vectors have only been built for the source language.

2.2.3. Translation of the similarity vectors

Using a bilingual dictionary, we translate the lexical units of the similarity vector and identify their context vectors in the target language. Figure 1 illustrates this translation process.

Depending the nature of the lexical unit, two different treatments are carried out:

Translation of a SWT If the bilingual dictionary provides several translations for a word belonging to the similarity vector, we generate as many target context vectors as possible translations. Then, we calculate the union of these vectors to obtain only one target context vector.

Translation of a MWT If the translation of the parts of the MWT are found in the bilingual dictionary, we generate as many target context vectors as translated combinations identified by *ACABIT* and calculate their union. When it is not possible to translate all the parts of a MWT, or when the translated combinations are not identified by *ACABIT*, the MWT is not taken into account in the translation process.

2.2.4. Finding the MWT translations

We calculate the barycentre of all the target context vectors obtained in the preceding step in order to propose a target average vector. The candidate translations of a lexical unit are the target lexical units closest to the target average vector according to vector distance.

3. Resources presentation

We present in this section the different resources used for our experiments:

[1] http://www.sciences.univ-nantes.fr/info/perso/permanents/daille/

3.1. Comparable corpus

Our comparable corpus has been built from the *Unasylva* electronic international journal published by FAO[2] and representing 4 million words. This journal deals with forests and forest industries and is available in English, French and Spanish. In order to constitute a comparable corpus, we only select texts which are not the translation of each other.

3.2. Bilingual dictionary

Our bilingual dictionary has been built from lexical resources on the Web. It contains 22,300 French single words belonging to the general language with an average of 1.6 translation per entry.

3.3. Reference bilingual terminology

The evaluation of our bilingual terminology extraction method has been done from a reference bilingual terminology. This reference list has been built from three different terminological resources:

1. a bilingual glossary of the terminology of silviculture[3]. It contains 700 terms of which 70% are MWTs.
2. the Eurosilvasur multilingual lexicon[4]. It contains 2,800 terms of which 66% are MWTs.
3. the multilingual AGROVOC thesaurus[5]. It contains 15,000 index terms of which 47% are MWTs.

These three terminological resources are complementary, the glossary being the most specialised, the thesaurus the least. From these resources, we automatically select 300 terms with the constraint that each French term should appear at least 5 times in our corpus. These terms are divided into three sub-lists:

- [list 1] 100 French SWTs of which the translation is an English SWT. Of course, this translation is not given by our bilingual dictionary.
- [list 2] 100 French MWTs of which the translation could be an English SWT or a MWT. In the case of MWTs, the translation could not be obtained by the translation of the MWT's parts.
- [list 3] 100 MWT of which the translation is an English MWT. The translation of these MWTs is obtained by the translation of their parts.

This reference list contains a majority of terms with low frequency (cf. Table 1). Two main reasons explain this fact: on the one hand, the different resources which have been used to build this reference list are either specific or generic; on the other hand, our corpus covers several domains linked to forestry and does not constitute a highly specialised resource.

[2]http://www.fao.org/forestry/foris/webview/forestry2/
[3]http://nfdp.ccfm.org/silviterm/silvi_f/silvitermintrof.htm
[4]http://www.eurosilvasur.net/francais/lexique.php
[5]http://www.fao.org/agrovoc/

# occ.	$<$ **50**	$\leq$ **100**	$\leq$ **1 000**	$>$ **1 000**
[list 1]	50	21	18	11
[list 2]	54	21	25	0
[list 3]	51	18	29	2

Table 1: Frequency in the corpus of the French terms belonging to the reference list

4. Evaluation

We present now the evaluation of the bilingual terminology extraction.

4.1. Parameter estimation

Several parameters appear in the extraction process presented in Section 2. These parameters interact in complicated ways. We summarise below those which arise in the statistical step and the most interesting values obtained after a few experiments:

1. **Context vectors**
   - context window size: 3 sentences;
   - retained type of lexical units (single words alone or single words and MWTs) appearing in the context vector: single words alone;
   - association score: MI or Loglike;
   - context vector size: 20 to 30 items.
2. **Similarity vectors**
   - distance measure: Cosine or Jaccard;
   - similarity vector size: 20 to 30 items.

4.2. Result analysis

Table 2 gives the results obtained with our experiments. For each sublist, we give the number of translations found (NB_{trans}), and the average and standard deviation position for the translations in the ranked list of candidate translations (AVG_{pos}, $STDDEV_{pos}$).

	NB_{trans}	AVG_{pos}	$STDDEV_{pos}$
[list 1]	60	35.1	38,5
[list 2]	65	37.5	45,9
[list 3]	90	3.4	15,2

Table 2: Bilingual terminology extraction results

We note that translations of MWTs belonging to [list 3] which are compositionally translated are well-identified and often appear in the first 20 candidate translations. The translations belonging to [lists 1 and 2] are not always found and, when they are, they seldom appear in the first 20 candidate translations.
The examination of the candidate translations of a MWT regardless of the list to which it belongs shows that they share the same semantic field. For example, the first 20 candidate

	NB_{trans}	AVG_{pos}	$STDDEV_{pos}$	**Top 10**	**Top 20**
[list 1]	60	22.7	26.1	37	47
[list 2]	65	21.3	25.0	41	51
[list 3]	90	1.5	11.3	88	89

Table 3: Bilingual MWT extraction with parameter combination

translations of *gaz à effet de serre* (greenhouse gas) are: *carbon, carbon cycle, atmosphere, greenhouse gas, greenhouse, global carbon, atmospheric carbon, emission, sink, carbon dioxide, fossil fuel, fossil, carbon pool, mitigate, global warming, climate change, atmospheric, dioxide, sequestration, quantity of carbon*.

As noted above, our results differ widely according the chosen parameter values. Because of time constraints, we cannot evaluate all the possible values of all the different parameters, but manual examination of the candidate translations for a few different configurations shows:

- Some good translations obtained for one parameter configuration are not found for another, and, inversely, some terms which are not translated in the first configuration could be correctly translated by another. So, it is difficult to choose the best configuration, especially for [lists 1 and 2].
- More precisely, for a given term, the first candidate translations are different for different configurations. For example, for the French MWT *pâte à papier* (*paper pulp*), the first 50 candidate translations of 20 different configurations have only 30 items in common.
- The right translation appears in different positions for different configurations.

In order to identify more correct translations, we decided to take into account the different results proposed by different configurations by fusing the first 20 candidate translations proposed by each configuration. The different configurations concern the size of the context and similarity vectors, and the association and similarity measures. The results obtained and presented in Table 3 show a slight improvement in bilingual extraction. The results for [list 3] are still very satisfactory. The results for [list 1] improve, 41% and 51% for the first 10 and 20 candidates, but remain a little below the results obtained by (Déjean et al., 2002) who obtained 43% and 51% for the first 10 and 20 candidates respectively for a 100,000-word medical corpus, and 79% and 84% for a multi-domain 8 million word corpus. This difference in results could be explained by the fact that we used automatic evaluation, which is more constrained than manual evaluation. For example, our reference list gives *haulage road* as the translation of *piste de débardage*. In our candidate translation list, *haulage road* is not present. We find an acceptable translation, *skid trail*, in the first 20 candidates, but this is never considered valid by our automatic evaluation.

Our results for MWTs are better than those for single words. The method seems promising, especially for MWTs for which translation is not compositional.

5. Conclusion

In this paper, we proposed and evaluated a combined method for bilingual MWT extraction from comparable corpora which takes into account three main characteristics of MWT translation: fertility, non-compositionality, and variation clustering. We first extracted monolingually MWTs and clustered synonymic variants. Secondly, we aligned them using a statistical method adapted from (Déjean et al., 2002) for single words which exploits the context of these MWTs. This combined approach for MWTs gives satisfactory results compared to those for single word. It also allows us to obtain non compositional translations of MWTs. Our further works will concentrate on the interaction parameters, the combining of the source-to-target and target-to-source alignment results, and the handling of non-synonymic term variations.

6. References

Cao, Yunbo and Hang Li, 2002. Base Noun Phrase Translation Using Web Data and the EM Algorithm. In *Actes, COLING'02*.

Carl, M. and P. Langlais, 2002. An intelligent Terminology Database as a pre-processor for Statistical Machine Translation. In *Actes, COMPUTERM 2002, COLING 2002 workshop*.

Daille, B., 2003. Terminology Mining. In M.T. Pazienza (ed.), *Information Extraction in the Web Era*. Springer, pages 29–44.

Daille, B., E. Gaussier, and J. Langé, 1994. Towards automatic extraction of monolingual and bilingual terminology. In *Actes, COLING 1994*.

Déjean, H., F. Sadat, and E. Gaussier, 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Actes, COLING 2002*.

Fung, Pascale, 1998. A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora. In *Actes, AMTA'98*. Springer.

Gaussier, Eric and Jean-Marc Langé, 1995. Modèles statistiques pour l'extraction de lexiques bilingues. *T.A.L, Vol. 36(1-2)*:133–155.

Jacquemin, C., 2001. *Spotting and Discovering Terms through Natural Language Processing*. MIT Press.

Melamed, I. Dan, 2001. *Empirical Methods for Exploiting Parallel Texts*. MIT Press.

Rapp, Reinhard, 1999. Automatic identification of word translations from unrelated English and German corpora. In *Actes, ACL'99*.

Veronis, Jean (ed.), 2000. *Parallel Text Processing*. Kluwer Academic Publishers.