French-English Terminology Extraction from Comparable Corpora

Béatrice Daille and Emmanuel Morin

University of Nantes, LINA - FRE CNRS 2729,
2, rue de la Houssinière - BP 92208, 44322 Nantes Cedex 3, France
{beatrice.daille, emmanuel.morin}@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 [17]. From sentence-to-sentence aligned corpora, symbolic [2], statistical [11], or combined [7] 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: texts that are of the same text type and on the same subject without a source text-target text relationship. The main studies concentrate on finding in such corpora translation candidates for one-item words. For example, the French SWT *manteau* is translated in English by *mantle* in the domain of forestry, *shield* in the domain of marine activities, and by *coat* in the domain of clothing. 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. Thus, for our three possible translations of *manteau*, three different lexical contexts are encountered which are expressed below by English lexical units:

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 707–718, 2005.

- *manteau/mantle* : vegetation, forest, wood...
- *manteau/shield* : boat, sea, shipbuilding...
- *manteau/coat* : cloth, cold, wear...

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, [10] extracts English-Chinese one-item 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, [14] improves the precision to 89% on the first one-item 10 candidates using the same techniques. [4] obtain 50% precision on the first one-item 10 candidates from a French/English corpus of 1.2 million words. [1] adapted this approach to deal with many-to-many word translations. In extracting English-Chinese 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 [9] to avoid the insufficient coverage of bilingual dictionary and thus not to get context vectors with too many elements that are not translated. This method is called "similarity-vector approach": it associates to the word to be translated the context vectors of the nearest lexical units that are in the bilingual dictionary. With this method, they obtain for one-item French-English words 43% and 51% precision on the ten and twenty first candidates applied on a medical corpus of 100 000 words (respectively 44% and 57% with the direct method) and 79% and 84% precision on the ten and twenty first candidates applied on a social science corpus of 8 millions words (respectively 35% and 42% with the direct method).

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 [13]. 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 [12],[5]. 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 two steps:

1. Identification in source and target languages of MWTs and their variations;
2. Alignment of theses MWTs using a method close to the "similarity-vector approach".

2.1 MWT Identification

MWTs are extracted using a terminology extraction program available for French and English: *ACABIT*[1]. This program is open source and one of its characteristics is to take into account variants of MWTs (graphical, inflectional, syntactic, and morphosyntactic)[6]. It does not need any external linguistic resources and is domain-independent. ACABIT applies on a corpus with the following pre-processing:

– tokenisation and sentence segmentation;
– part-of-speech and lemma tagging.

First, ACABIT carries out shallow parsing: it scans the corpus, counts and extracts strings whose tag sequences characterise patterns of MWTs or one of their variants. The different occurrences referring to a MWT or one of its variants are grouped and constitute an unique candidate MWT. Thus the candidate MWT *produit forestier* 'forest product' appears under the following forms:

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

- **base form:** *produit forestier* ;
- **graphical variant:** *produit fo-restier*, *pro-duit forestier* ;
- **inflexional variant:** *produits forestiers* ;
- **syntatic variant: modification:** *produit non forestier*, *produit alimentaire forestier*, *produit fini d'origine forestiére*, *produit ligneux non forestier* ;
- **syntactic variant: coordination:** *produit halieutique et forestier*, *produit agricole ou forestier*, *le produit et le service forestier*.

The MWT candidates *produit de la forêt*, *produit agroforestier*, *non-produit agroforestier*, and *sous-produit forestier*, *sous-produit de la forêt* have also been identified.

Second, ACABIT performs semantic grouping thanks to the following operations:

Merging of two MWTs. Two MWT candidates are merged if they are synonymic variants obtained by derivation or conversion. Such variants include a relational adjective: either a denominal adjective, i.e. morphologically derived from a noun thanks to a suffix, such as *forêt/forestier* 'forest', or an adjective having a noun usage such as *mathématique* 'mathematical/mathematics'.

Dissociation of some MWT variants. Syntactical variants that induce semantic discrepancies are retrieved from the set of the candidate variants and new MWT candidates are created. Modification variants with the insertion of an adverb of negation denoting an antonymy link such as *produit non forestier* 'non forest product' and *produit forestier* 'forest product', or insertion of a relational adjectives denoting an hyperonymy link such as *produit alimentaire forestier* 'food forest product' with *produit forestier* 'forest product' [6].

Grouping of MWTs. All MWT candidates linked by derivational morphology or by variations inducing semantic variations are clustered. For example, the following MWT candidates constitutes a cluster of MWTs: *produit forestier/produit de la forêt*, *produit non forestier*, *non-produit agroforestier*, *produit agroforestier*, *sous-produit forestier/sous-produit de la forêt*, *produit alimentaire forestier* and*produit forestier*.

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 MWT Alignment

The goal of this step, which adapts the similarity vector-based approach defined for single words by [9] 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.

Context Vectors. First, we collect all the lexical units in the context of each lexical unit i and count their occurrence frequency in a window of n sentences around i. For each lexical unit i of the source and the target language, we obtain a context vector v_i which gathers the set of co-occurrence units j associated with the number of times that j and i occur together occ_j^i. We normalise context vectors using an association score such as Mutual Information or Log-likelihood. (cf. equations 1 and 2 and table 1). In order to reduce the arity of context vectors, we keep only the co-occurrences with the highest association scores.

Table 1. Contingency table

	j	$\neg j$
i	$a = occ(i, j)$	$b = occ(i, \neg j)$
$\neg i$	$c = occ(\neg i, j)$	$d = occ(\neg i, \neg j)$

$$MI(i,j) = log\frac{a}{(a+b)(a+c)} \tag{1}$$

$$\begin{aligned}\lambda(i,j) = {} & a\log(a) + b\log(b) + c\log(c) + d\log(d) \\ & +(a+b+c+d)\log(a+b+c+d) - (a+b)\log(a+b) \\ & -(a+c)\log(a+c) - (b+d)\log(b+d) - (c+d)\log(c+d)\end{aligned} \tag{2}$$

Similarity Vectors. For each lexical unit k to be translated, we identify the lexical units which the context vectors are similar to v_k thanks to a vector distance measure such as Cosine [15] or Jaccard [16] (cf. equations 3 and 4). From now, we call "similarity vector" of the unit k a vector that contains all the lexical units which the context vectors are similar to v_k. To each unit l of the similarity vector v_k, we associate a similarity score $simil_{v_l}^{v_k}$ between v_l and v_k. 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.

$$simil_{v_l}^{v_k} = \frac{\sum_t assoc_t^l \, assoc_t^k}{\sqrt{\sum_t {assoc_t^l}^2 \, {assoc_t^k}^2}} \tag{3}$$

$$simil_{v_l}^{v_k} = \frac{\sum_t min(assoc_t^l, assoc_t^k)}{\sum_t {assoc_t^l}^2 + \sum_t {assoc_t^k}^2 - \sum_t assoc_t^l \, assoc_t^k} \tag{4}$$

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:

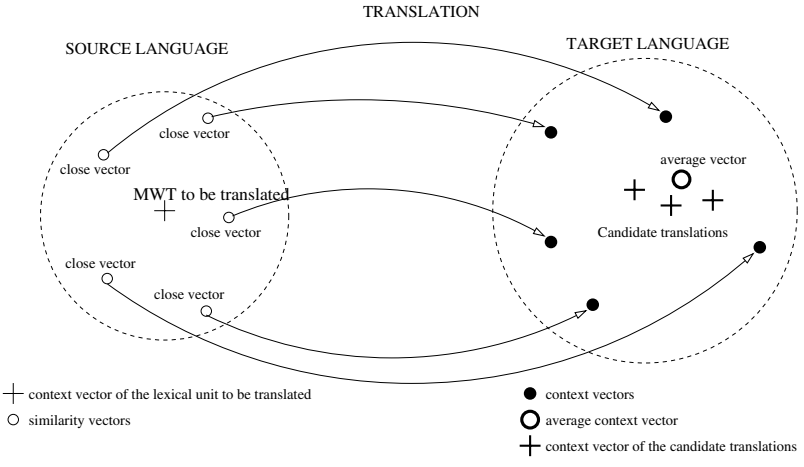

Fig. 1. Transfer procedure of similarity vectors from source to target language

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.

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:

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

[2] http://www.fao.org/forestry/foris/webview/forestry2/

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 2). 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.

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

Table 2. Frequency in the corpus of the French terms belonging to the reference list

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

4 Evaluation

We present now the evaluation of the bilingual terminology extraction. We have to deal with 55 013 SWTs and MWTs, but only 7 352 SWTs and 6 769 MWTs appear both in the reference bilingual terminology and in the corpus.

4.1 Parameter Estimation

Several parameters appear in the extraction process presented in Section 2. The most interesting results have been obtained with the following values:

- **Size of the context window** is 3 sentences around the lexical unit to be translated;
- **Context vectors** are built only with one-item words to increase representativity. For example, the context vector of the French term *débardage* 'hauling' includes the MWT *tracteur à chenille* 'crawler tractor' which is more discriminating than its parts, *tracteur* or *chenille*. But including MWTs into context vectors increases the vectorial space dimension and reduces the representativity of the terms appearing both in the corpus and the reference bilingual terminology. The term *débardage* 'hauling' has a frequency of 544 as a SWT and only a frequency of 144 as part of a MWT as it appears in several MWTs. The context vector size are limited to the first 100 values of the Log-likelihood association score.
- **Similarity vectors** are the first 30 values of Cosine distance measure.
- **Finding translations** is done with Cosine distance measure.

4.2 Result Analysis

Table 3 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}, $STDDEV_{pos}$).

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 (cf. table 5).

Table 3. Bilingual terminology extraction results

	NB_{trans}	AVG_{pos}	$STDDEV_{pos}$
|list 1]	56	32.9	23,7
[list 2]	63	30.7	26,7
[list 3]	89	3.8	7,9

Table 4. Bilingual MWT extraction with parameter combination

	NB_{trans}	AVG_{pos}	$STDDEV_{pos}$	**Top 10**	**Top 20**
|list 1]	59	16.2	15.9	41	51
[list 2]	63	14.8	22.3	45	55
[list 3]	89	2.4	3.7	87	88

Table 5. Exemples of candidate translations obtained for 3 terms belonging to [list 2]

degré de humidité (# occ. 41)	gaz à effet de serre (# occ. 33)	papeterie (# occ. 178)
humidity	carbon	newsprint
saturation	carbon cycle	paper production
aridity	atmosphere	raw material
evaporation	**greenhouse gas**	mill
saturation deficit	greenhouse	pulp mill
rate of evaporation	global carbon	raw
atmospheric humidity	atmospheric carbon	manufacture
water vapor	emission	**paper mill**
joint	sink	manufacturing
dry	carbon dioxide	capacity
hot	fossil fuel	printing
rainy	fossil	paper manufacture
temperature	carbon pool	factory
moisture control	mitigate	paperboard
meyer	global warming	fiberboard
party	climate change	bagasse
atmospheric	atmospheric	paper-making
dryness	dioxide	board
monsoon	sequestration	material supply
joint meeting	quantity of carbon	paper pulp

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 4 show a slight improvement in the position of the correct translations among the set of candidate translations.

The results for [list 3] are still very satisfactory. The results for [list 1] improve, but remain a little below the results obtained by [8] 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.

4.3 Comment

In a general way, it is difficult to compare our experiments to previous ones [3],[8] as the corpora are different. Indeed, our comparable corpus covers several domains belonging to forestry, and does not constitute a very specialised resource on the contrary of the medical corpus of [3] built thanks to the key words "symptoms, pathological status". Moreover, half of the terms of the reference bilingual terminological database have a frequency of less than 50 occurrences in the corpus that lead to non-discriminating context vectors. [8] use for their experiments a social sciences corpora of 8 millions words and a reference bilingual terminological database of 180 words with high frequencies in the corpus: from 100 to 1000. Our automatic evaluation is also 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 similarity-vector approach 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.

Acknowledgements

We are particularly grateful to Samuel Dufour-Kowalski, who undertook the computer programs. This work has also benefited from his comments.

References

1. Cao, Y., Li, H.: Base Noun Phrase Translation Using Web Data and the EM Algorithm. In: Proceeding of the 19th International Conference on Computational Linguistics (COLING'02), Tapei, Taiwan (2002) 127–133
2. Carl, M., Langlais, P.: An intelligent Terminology Database as a pre-processor for Statistical Machine Translation. In Chien, L.F., Daille, B., Kageura, L., Nakagawa, H., eds.: Proceeding of the COLING 2002 2nd International Workshop on Computational Terminology (COMPUTERM'02), Tapei, Taiwan (2002) 15–21
3. Chiao, Y.C.: Extraction lexicale bilingue à partir de textes médicaux comparables : application à la recherche d'information translangue. PhD thesis, Université Pierre et Marie Curie, Paris VI (2004)
4. Chiao, Y.C., Zweigenbaum, P.: Looking for candidate translational equivalents in specialized, comparable corpora. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING'02), Tapei, Taiwan (2002) 1208–1212
5. Daille, B.:. Conceptual Structuring through Term Variations. In Bond, F., Korhonen, A., MacCarthy, D., Villacicencio A., eds.: Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment (2003) 9–16
6. Daille, B.: Terminology Mining. In Pazienza, M., ed.: Information Extraction in the Web Era. Springer (2003) 29–44
7. Daille, B., Gaussier, E., Langé, J.-M..: Towards Automatic Extraction of Monolingual and Bilingual Terminology. Proceedings of the 15th International Conference on Computational Linguistics (COLING'94) **1** (1994) 515–521
8. Déjean, H., Sadat, F., Gaussier, E.: An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING'02). (2002) 218–224
9. Déjean, H., Gaussier, E.: Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. Lexicometrica, Alignement lexical dans les corpus multilingues (2002) 1–22
10. Fung, P.: A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora. In Farwell, D., Gerber, L., Hovy, E., eds.: Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98), Springer (1998) 1–16

11. Gaussier, E., Langé, J.M.: Modèles statistiques pour l'extraction de lexiques bilingues. Traitement Automatique des Langues (TAL) **36** (1995) 133–155
12. Jacquemin, C.: Spotting and Discovering Terms through Natural Language Processing. Cambridge: MIT Press (2001)
13. Melamed, I.D.: Empirical Methods for Exploiting Parallel Texts. MIT Press (2001)
14. Rapp, R.: Automatic Identification of Word Translations from Unrelated English and German Corpora. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99). (1999) 519–526
15. Salton, G., Lesk, M.E.: Computer Evaluation of Indexing and Text Processing. Journal of the Association for Computational Machinery **15** (1968) 8–36
16. Tanimoto, T.T.: An elementary mathematical theory of classification. Technical report, IBM Research (1958)
17. Veronis, J., ed.: Parallel Text Processing. Kluwer Academic Publishers (2000)