Revising the Compositional Method for Terminology Acquisition from Comparable Corpora

Emmanuel Morin   Béatrice Daille
Université de Nantes, LINA UMR CNRS 6241
2, rue de la Houssinière, BP 92208
F-44322 Nantes cedex 03
`{emmanuel.morin,beatrice.daille}@univ-nantes.fr`

ABSTRACT
In this paper, we present a new method that improves the alignment of equivalent terms monolingually acquired from bilingual comparable corpora: the Compositional Method with Context-Based Projection (CMCBP). Our overall objective is to identify and to translate high specialized terminology made up of multi-word terms acquired from comparable corpora. Our evaluation in the medical domain and for two pairs of languages demonstrates that CMCBP outperforms the state-of-art compositional approach commonly used for translationally equivalent multi-word term discovery from comparable corpora.

KEYWORDS: Comparable corpora, bilingual lexicon, compositionality, multi-word term, context information.

1 Introduction

The automatic compilation of bilingual dictionaries has received considerable attention in recent years for *language for special purposes (LSP)* (especially coming from scientific domains). LSP is characterised by the small amount of available textual data compared with general language, and a high proportion of specialised terms which are not be found in general language monolingual or bilingual dictionaries. For LSP, a specialised term could be either a single-word term (SWT) or a multi-word term (MWT), the latter being highly productive (Sag et al., 2002). A term is a lexical unit which represents a concept within a domain. As an example, in the medical domain, *cancer* is an SWT, *breast cancer* is an MWT.

Comparable corpora that are sets of texts in two or more languages without being translations of each other, seem to be the right solution to solve the textual scarcity of LSP: as monolingual productions, they are authentic texts, and the babel web ensures that there is a sufficient number of multilingual documents. The comparability of the corpus should be ensured by using various shared characteristics across languages that are checked during the compilation phase (McEnery, 2007). For LSP, the domain and sub-domain are requested as well as the communicative settings and the textual genre to identify reliable translations (Bowker and Pearson, 2002).

To build highly-specialised terminologies, the terms are first of all extracted monolingually from the comparable corpus. To collect close candidate terms across languages, it is necessary to use a term extraction program that applies the same method in the source and in the target languages. The translation of MWTs is the main need as they constitute around 80% of the domain-specific terms. See Nakagawa and Mori (2003) for the Japanese language.

Proceedings of COLING 2012: Technical Papers, pages 1797–1810,
COLING 2012, Mumbai, December 2012.

Our goal is to find the right translation for a source MWT in the set of MWT candidates in the target language. The simplest method assumes that the right translation of an MWT could be obtained by translating each component individually thanks to a general dictionary, by generating all the combinations of word positions, and then filtering the translated expressions using either the list of target MWTs (Morin and Daille, 2010), the target corpus (Robitaille et al., 2006) or the web (Grefenstette, 1999). This method is limited to the subset of MWTs that share the same compositional property - 48.7% were reported by Baldwin and Tanaka (2004) for English/Japanese N N compounds. However, even if the MWT is characterised by the compositional property, the translation is not found when some words, which are part of the MWTs, do not belong to the general bilingual dictionary or when the translated combinations do not exist or have not been extracted by the term extraction program in the target language.

Within this context, we propose to improve the compositional approach by using context information collected from LSP comparable corpora when one or several of the components, part of the MWTs, are not found in the dictionary. We demonstrate that the use of context information when performing terminology translation helped us to learn a significant number of additional correct lexical entries that could not be identified by the compositional method.

The remainder of this paper is organised as follows: Section 2 shows the intricate problems with the translations of MWTs. Section 3 presents the compositional method used for the automatic translation of MWTs. Section 4 introduces CMCBP that takes advantage of the context information to improve the compositional method. Section 5 describes the linguistic resources and the open-terminology extraction tool used for our experiments. Section 6 evaluates the influence of CMCBP on the quality of bilingual terminology extraction through experiments involving French as a source language, and English and German as target languages. Section 7 discusses works related to this study. Finally, Section 8 presents our conclusions.

2 Translation of MWTs

If MWTs are less polysemous (Savary and Jacquemin, 2003) and more representative (Nomura and M., 1989; Nakagawa and Mori, 2003) of domain specialities than SWTs, pinpointing their translations poses specific problems that are well-known, such as fertility, non-compositionality, or term variation[1]:

Fertility is known as a problem of difference of length between the source and the target MWT (Brown et al., 1993): for instance, the German SWT *axilladissektion* (1 content word) is translated into English by the MWT *axillary dissection* (two content words); the French MWT *dépistage du cancer du sein* (three content words) is translated into English by the MWT *breast screening* (two content words).

Non-compositionality is illustrated when the target MWT is not typically composed of the translation of its parts (Melamed, 2001). For instance the French MWT *curage axillaire* is translated into the English language as *axillary dissection* whereas the English word *dissection* is not the translation of the French word *curage*. Baldwin and Tanaka (2004) report that at least 50% of the Japanese N N compounds are not translated through a compositional strategy into English.

[1] The French/English/German examples in this paper are extracted from the specialized medical comparable corpus described in Section 5.

Term variation refers to an MWT that appears in texts in different forms reflecting either graphical, syntactic, morphological or semantic differences: for example, the French MWTs *cancer du sein* and *cancer mammaire* are both translated by the same English MWT *breast cancer*. Source and target MWTs can appear in different syntactic structures. For example, the French MWT *prolifération tumorale* of N A pattern is translated by the English MWT *tumour proliferation* of N N pattern, where the French adjective *tumorale* is linked through morphological derivation to the English noun *tumour*. The term variations could also involve paradigmatic variation when one element of the MWT is substituted by a synonym or a hypernym such as *tumour size* → *diameter tumour* in the source language and not in the target language such as *taille tumorale* (lit. 'tumour size').

It is quite difficult to design a general framework that can address all these problems simultaneously (Robitaille et al., 2006) and for any language. The non-compositionality has to be solved during the translation process as it involves an MWT and its translation. The term variant problem is generally handled at the monolingual level during the term extraction task. This is done in two steps: term variant extraction and term variant grouping. A sophisticated variant recognition and conflation program will handle several types of variants: graphical, but also morphological and syntactic variation, ideally paradigmatic variants and acronyms. Using a term variant program allows us to cluster a set of term-like sequences reflecting base or variant forms. This clustering could be interpreted as a terminology normalization in the same way as lemmatisation at the morphological level. Handling term variation could indirectly solve part of the fertility problem using the syntactic variant of MWTs: as an example, in French, the term *dépistage du cancer du sein* (lit. *breast cancer screening*) could be collected as a syntactic variant of the term *dépistage du sein* (lit. *breast screening*) and thus provides a word-to-word translation. In German, the fertility problem could be solved by establishing an equivalence relation between a morphological compound of the type $N_1|N_2$ where | is the concatenation operator, and a syntagmatic compound of N_1N_2 pattern: the noun *axilladissektion* that is morphologically analyzed as *axilla|dissektion* will be a variant of the MWT *axilläre dissektion*.

The compositional method with context-based projection that we introduce in Section 4 will take into account the non-compositionality and term variation problems and indirectly the fertility problem through the term variant and the German compound splitting treatments.

3 Compositional Approach

Compositionality is defined as the property where *"the meaning of the whole is a function of the meaning of the parts"* (Keenan and Faltz, 1985, p. 24-25): a *frying pan* is indeed a *pan* used for *frying*. The implementation of the principle of translation compositionality from a comparable corpus relies on the following steps (Grefenstette, 1999; Tanaka, 2002; Robitaille et al., 2006):

Translation of the source MWT For an MWT of the source language to be translated, each component of the MWT is translated by looking it up in a dictionary. The lexical form is examined without checking the part-of-speech (POS). For example, for the French MWT *examen clinique* (*clinical examination*), there are six English translations for *examen* (*consideration*/N, *examen*/N, *examination*/N, *inspection*/N, *review*/N, *test*/N) and two translations for *clinique* (*clinic*/N, and *clinical*/A).

Generation of the candidate translations All possible mappings are constructed regardless of word order with a total of $O(\prod_{i=1}^{p} t_i n!)$ possible mappings (where t_i is the number of translations of the content word i, and n the number of content words). In the above example, 24 combinations are obtained. The number of generated translations can be reduced using MWT POS patterns in the source and the target languages. For instance, Tanaka and Baldwin (2003) defined the following templates to filter translation candidates: N_1 N_2 Japanese structure is translated by N_1 N_2 (33.2% of the cases), A_1 N_2 (28.4%), N_2 `of (the)` N_1 (4.4%) English structures.

Selection of the candidate translations From the set of translation candidates, the most likely translations are selected according to term frequency in the target language. In the above example, the translations are MWTs of the target language identified by the terminology extraction system.

4 Compositional Method with Context-Based Projection (CMCBP)

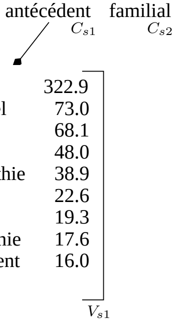
The compositional approach that finds translations of multi-word terms is easy to implement, but it fails when:

1. At least one element of an MWT is not found in the bilingual dictionary and thus cannot be translated.
2. The translated combination is valid but is not provided by the term extraction program for the target language. One explanation could be that the target MWT does not occur in the comparable corpus, or the source concept occurs in the target corpora but under a non terminology-like form, or an error during the preprocessing of the corpora induces that the terminology extraction program misses the MWT.
3. The translated combination is not valid. One of the MWT translation problems has been encountered (see Section 2).

When there is no translation candidates for an MWT, a first solution would be to find its synonyms in the source language. Similar words are predicted by Pekar et al. (2006) for low-frequency words and by Sharoff et al. (2009) for wrong translations. CMCBP that deals with term variants performs a clustering of synonymic terms. The translations that are proposed are for the set of synonymic term variants.

CMCBP is designed to identify MWT translations in a comparable corpus on a large scale and is able to solve points 1 and 3. It includes the use of the context of the words (which are parts of the MWT to be translated) when the compositional approach fails. We refer to the two monolingual parts of the comparable corpus as the source and target corpus. CMCBP uses four steps:

Computing the context of the MWT For an MWT or a morphological compound in the source corpus defined as $C_{s1}C_{s2}\cdots C_{sk}$ to be translated (where k is the number of content words or autonomous morphemes), we look up each component C_{si} in the bilingual dictionary. When a component is not found in the bilingual dictionary, we replace it by co-occurrence information. We compute the co-occurrence information between a component C_{si} and the words that co-occur in a window of w words around C_{si} from the source corpus. Mutual information or Likelihood-ratio are good measures of the co-occurrence relationship between 2 words. The co-occurrence information is expressed with a vector representation called context vector (V_{si}). As an example, let us consider the French MWT *antécédent familial* ($C_{s1}C_{s2}$). If the first component *antécédent* (C_{s1}) is not found in the bilingual dictionary then this component is replaced by its vector context: V_{s1} (see Figure 1).



familial	322.9
personnel	73.0
cancer	68.1
sein	48.0
mastopathie	38.9
degré	22.6
patient	19.3
mastodynie	17.6
saignement	16.0
...	

Figure 1: Computing context in the source corpus

Transfer of the MWT At this level, we are able to identify two situations depending on whether or not the components of the MWT are translated:

1. If the component C_{si} is found in the dictionary, we compute the co-occurrence information of each translation in the target corpus and store it in a context vector: V'_{si}.
2. If the component C_{si} is not found in the dictionary, we use the context vector of the source corpus V_{si}. The elements of V_{si} are projected into the target corpus using the bilingual dictionary and the transferred context vector becomes: $V^{'}_{si}$. If the bilingual dictionary provides several translations for an element, all of them are used but the different translations are weighted according to their frequency in the target language. If an element is not found in the bilingual dictionary it is discarded.

In the previous example, if we find two English translations for the component *familial* (C_{s2}) such as *familial* and *family* in the target language then we obtain two context vectors: $V^{'}_{s2_1}$ and V'_{s2_2} (see Figure 2).

Generation of candidate translations Each MWT of the target language, for which each component C_{ti} is described by its context vector V_{ti}, is then compared to the transferred MWT through a similarity measure such as Cosine or Weighted Jaccard. For an MWT composed of two context vectors V_{t1} and V_{t2} in the target language and a transferred MWT composed of two context vectors $V^{'}_{s1}$ and $V^{'}_{s2}$, two pairs of similarity scores corresponding to the possible mappings are computed: $sim(V_{t1}, V^{'}_{s1})$ with $sim(V_{t2}, V^{'}_{s2})$,

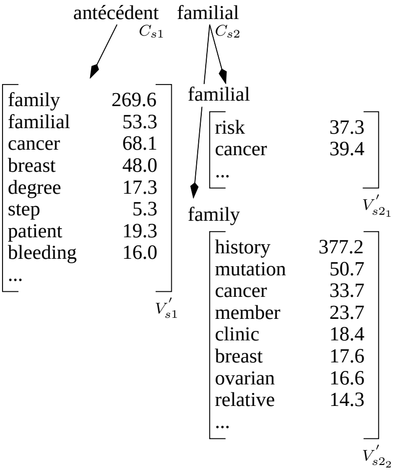


Figure 2: Projection in the target corpus

and $sim(V_{t1}, V'_{s2})$ with $sim(V_{t2}, V'_{s1})$. The combination score for each pair is then defined as the geometric mean of each similarity score: $\sqrt{sim(V_{t1}, V'_{s1}).sim(V_{t2}, V'_{s2})}$ and $\sqrt{sim(V_{t1}, V'_{s2}).sim(V_{t2}, V'_{s1})}$. Figure 3 illustrates this comparison for the previous example.

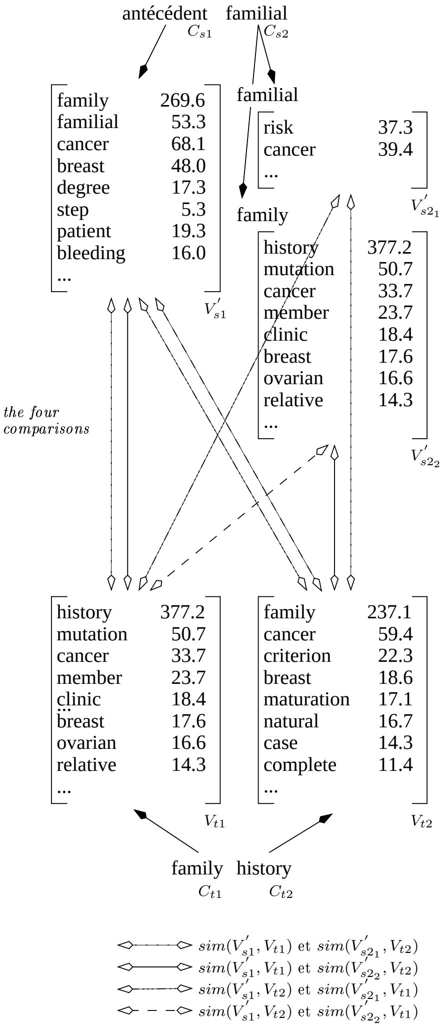
Ranking of candidate translations We rank the candidate translations in decreasing order of their combination score (see Figure 4).

5 Resources

In this section, we describe the different resources used for our experiments: the comparable corpus, the bilingual dictionary, and the multi-word term test set.

5.1 Comparable Corpora

The documents comprising the specialised comparable corpora were taken from the medical domain within the sub-domain of 'breast cancer'. These documents have been automatically selected from scientific paper websites where the title or the keywords of the articles contain the MWT 'breast cancer' in English, 'cancer du sein' in French and 'brustkrebs' in German. The compilation of the comparable corpus fulfils the requirements of an LSP comparable corpus: domain, sub-domain, communicative settings (experts-to-experts) and textual genre are common characteristics across languages. In this way, we collected 118 documents in English, 130 in French and 103 in German (about 530,000 words for English and French languages and 220,000 words for German language).



antécédent
familial
C_{s1}
C_{s2}
familial
family
the four comparisons
family
history
C_{t1}
C_{t2}
$sim(V^{'}_{s1}, V_{t1})$ et $sim(V^{'}_{s2_1}, V_{t2})$
$sim(V^{'}_{s1}, V_{t1})$ et $sim(V^{'}_{s2_2}, V_{t2})$
$sim(V^{'}_{s1}, V_{t2})$ et $sim(V^{'}_{s2_1}, V_{t1})$
$sim(V^{'}_{s1}, V_{t2})$ et $sim(V^{'}_{s2_2}, V_{t1})$

$V^{'}_{s1}$

family	269.6
familial	53.3
cancer	68.1
breast	48.0
degree	17.3
step	5.3
patient	19.3
bleeding	16.0
...	

$V^{'}_{s2_1}$

risk	37.3
cancer	39.4
...	

$V^{'}_{s2_2}$

history	377.2
mutation	50.7
cancer	33.7
member	23.7
clinic	18.4
breast	17.6
ovarian	16.6
relative	14.3
...	

V_{t1}

history	377.2
mutation	50.7
cancer	33.7
member	23.7
clinic ...	18.4
breast	17.6
ovarian	16.6
relative	14.3
...	

V_{t2}

family	237.1
cancer	59.4
criterion	22.3
breast	18.6
maturation	17.1
natural	16.7
case	14.3
complete	11.4
...	

Figure 3: Comparison between the MWT to be translated and an MWT of the target corpus

antécédent familial

↓

family history	0.75
cancer family	0.57
family member	0.22
high−risk family	0.18
familial risk	0.06
...	

Figure 4: Rank list of candidate translations

5.2 Bilingual Dictionary

The bilingual dictionaries used in our experiments are the French/English dictionary ELRA-M0033 and the French/German dictionary ELRA-M0034 available from the ELRA catalogue[2]. The French/English dictionary contains 243,539 translations and the French/German dictionary 170,967 translations. These are two general language dictionaries which contain only a few terms related to the medical domain.

5.3 Multi-Word Term Test Set

Terms are extracted monolingually from the comparable corpora. To collect close candidate terms across languages, it is necessary to use a term extraction program that is multilingually designed. We choose the TTC TermSuite (Rocheteau and Daille, 2011)[3] that applies the same term extraction method to several languages including French, German and English. TermSuite first normalises the texts through the following linguistic pre-processing steps: tokenisation, part-of-speech tagging and lemmatisation using TreeTagger (Schmid, 1995). TermSuite then extracts SWTs and MWTs whose syntactic patterns correspond either to a canonical or a variation structure. The patterns are expressed using MULTEXT part-of-speech tags and are provided for each language. The main patterns, whatever the language is, are N and A for SWTs. The main patterns of MWTs are for:

- **French** N N: *ganglion sentinelle* (*sentinel lymph node*); N Sp N: *cancer du sein* (*breast cancer*); N A: *curage axillaire* (*axillary dissection*);
- **English** N N: *breast cancer*; A N: *far therapy*; N Sp N: ;
- **German** A N: *thromboembolischer vorfall* (*thromboembolic incident*); N Sp N: *patientin mit mammakarzinom* (*patient with breast cancer*); N D:g N: *erfahrung der früherkennung* (*experience of early detection*).

The variants handled for MWTs are graphical, morphological, and syntactic. Both SWTs and MWTs accept variants but some are more likely to concern one main type such syntactic variants for MWTs. TermSuite defines a morphological variant as a morphological modification of one of the components of the MWT, and a syntactical variant as the adding of another word at the frontier or inside the MWT. For example, in the French part of the comparable corpus, the MWT candidate *cancer du sein* (*breast cancer*) appears in the following forms where shared items are numbered with the same values.

[2] http://www.elra.info/

[3] http://code.google.com/p/ttc-project

- **base form** of N_1 S_1p N_2 pattern: *cancer du sein* (*breast cancer*);
- **inflexional variant**: *cancers du sein* (*breast cancers*);
- **syntactic variant** (insertion inside the base form of a modifier): N_1 A S_1p N_2 *cancer primitif du sein* (*primary breast cancer*);
- **syntactic variant** (expansion coordination of base form): N_1 Sp N S_1p N_2 *cancer des ovaires et du sein* (*ovarian and breast cancer*).

In German, it is necessary to reconsider the rough distinction between single- multi-word terms in order to take morphological compounds into account. The common German compounds of the type $N_1|N_2$ ("|" is the concatenation operation) are often translated by N Sp N patterns in French: *Produktionsstandort* ⟷ *site de production* or by the N N patterns in English. Morphological compounds are identified by tokenisation programs as single-word terms but they look quite similar to multi-word terms. We use the morphological splitter which is combined with a dictionary look-up developed by Weller and Heid (2012) in order to get the MWT syntagmatic equivalence of a German morphological compound.

In order to build the test set, we have selected the French MWTs extracted by TermSuite for which the number of occurrences is greater than or equal to 5. The test set is composed of 976 French MWTs for which 90% of the base forms are only composed of two content words.

6 Experiments

In this section, we evaluate the performance of the dictionary look-up, the compositional method and CMCBP on the quality of bilingual terminology extraction.

6.1 Dictionary Look-up

First of all, we count the number of terms of the test set directly translated by looking them up in the bilingual dictionaries. From the 976 French MWTs to be translated, 51 are recorded in the French/English dictionary and 12 in the French/German dictionary. Here, the MWTs correctly translated are mainly generic terms that are not specific to the thematic of breast cancer such as *traitement médical/medical treatment* and *acide aminé/amino acid:* in French/English and *analyse statistique/statistische untersuchung* (*statistical analysis*) and *effet secondaire/begleiterscheinung* (*side effect*) in French/German. In this instance we were unable to generate any translations for 836 French MWTs in English and for 964 French MWTs in German.

6.2 Compositional Method

We then evaluate the quality of the translations provided by the compositional method (the MWTs found in the dictionary are not used). Table 1 shows the results obtained for the translation from French/English and German/English. The first column indicates the number of French MWTs that are translated. Since the compositional approach can give several target translations for one French MWT, the last two columns indicate the Top_1 and Top_5 accuracy. To evaluate the Top_n accuracy, we first keep for each French word to be translated its n first candidate translations and then measure the accuracy of the ranked lists obtained, *i.e.* the proportion of lists comprising the expected translation. Here, the candidate translations are ranked according to their frequency in the target part of the comparable corpus. The results of this experiment show that 140 of the 836 French MWTs are translated into English for the Top_5 with a high level of accuracy: 79.1%, and 87 of the 964 French MWTs are translated into German for the Top_5 with a high level of accuracy: 95.7%. Here, we were unable to generate any translations for 785 French MWTs in English and 877 French MWTs in German.

	# trans.	Top_1	Top_5
French/English	140	73.2%	79.1%
French/German	87	88.8%	95.7%

Table 1: Results for the compositional method

6.3 Compositional Method with Context-Based Projection

We now apply CMCBP (here again the MWTs found in the dictionary are not used). In this experiment, the parameters required for our approach are as follows: the size of the context window w is up to 3 (i.e. a seven-word window), the association measure is Mutual Information, and the distance measure is Cosine. Other combinations of parameters were assessed but the previous parameters gave the best performance. Table 2 presents the percentage of French terms for which the correct translation is obtained among the Top $_1$, $_5$, $_{10}$, and $_{20}$ candidates translations from French to English and German. Table 2 shows that 514 of the 836 French MWTs are translated into English with the CMCBP with an accuracy of 42.1% for the Top_1 and 57.1% for the Top_{20} and 510 of the 964 French MWTs are translated into German with an accuracy of 44.3% for the Top_1 and 51.2% for the Top_{20}. These results indicate that the majority of the correct translated MWTs are in fact obtained from the Top_5. Moreover, the CMCBP retains the advantages of the compositional method. All translations obtained with the compositional method are found in the same rank with the CMCBP.

	# trans.	Top_1	Top_5	Top_{10}	Top_{20}
French/English	514	42.1%	55.4%	56.8%	57.1%
French/German	510	44.3%	49.4%	51.2%	51.2%

Table 2: Results for the compositional method enhanced with context alignments

From the MWTs that are correctly translated and not found by the compositional approach, we found a large majority of French MWTs involving a relational adjective. However, the French MWT *dépistage mammographique* is not translated by the compositional approach since the French relational adjective *mammographique* is not found in the dictionaries. In contrast, the correct English translation *mammographic screening* is found in the Top_3 with the CMCBP because we have associated the French context vector of *mammographique* with the English context vector of *mammographic* and the French/English pair *dépistage/screening* is found in the dictionary. The other French MWTs correctly translated are mainly MWTs with a compositional structure for which one element is not found in the dictionary such as: *amélioration significative/significant benefit* (Top_1), and *caractéristique tumoral/tumor charakteristik* (tumor characteristic) (Top_1) or without a compositional structure such as: *bras témoin/control arm* (Top_1), and *curage axillaire/axillary dissection* (Top_{11}). From the MWTs incorrectly translated, we can point out two main cases. First, we find target MWTs semantically close to the French MWTs to be translated such as: *postmenopausalen frau* (*postmenopausal women*) (Top_5) and *prämenopausalen frau* (*premenopausal women*) (Top_7) for *femme ménopausé* (*menopausal women*). *Postmenopausalen frau* and *prämenopausalen frau* are morphological variants of *menopausalen frau* and should have been identified as thus by the term extraction program, but unfortunately the canonical form *amenopausalen frau* does not occur in the comparable corpora. Secondly, we found only a sub-part of the English MWTs such as: *node dissection* for *curage ganglionnaire* (*lymph node dissection*). This case needs further work as it deals with a fertility case that is not able to be solved with the term variant recognition program.

7 Related Work

The principle of translation compositionality is restrictive. Several studies have concentrated on enhancing the compositional approach: Robitaille et al. (2006) proposed a backing-off method: if there is insufficient data in the dictionary to translate an MWT of *n* content words, a scaled MWT with a length less than, or equal to, *n* is used instead. Morin and Daille (2010) proposed an extended compositional method that bridges the gap between MWTs of different syntactic structures through morphological links. The compositional approach is also called the "bag-of-equivalents" approach (Vintar, 2010) when the bilingual dictionary is built from a parallel corpus and contains all words that occur in the corpus and their suggested translation equivalents, together with a probability score. The "bag-of-equivalents" approach has been used for SMT to build a word-level translation lexicon from parallel corpora (Munteanu and Marcu 2006) and a cognate lexicon from comparable corpora (Koehn and Knight 2002).

Much of the work involving general or LSP comparable corpora has focused on extracting SWT translations using only contextual information (Fung, 1998; Rapp, 1999; Chiao and Zweigenbaum, 2002; Gaussier et al., 2004; Morin et al., 2007; Laroche and Langlais, 2010, among others). The contextual information method as defined by Fung (1998) gives very low results for MWTs: from the 785 non translated French MWTs, 483 French MTWS with an accuracy of 15.6% (Top$_{10}$) were found.

CMCBP is a new method dedicated to MWT that combines both the compositional and the contextual information. CMCBP significantly improves both the compositional method commonly used for bilingual alignment of MWTs extracted from comparable corpora, and the contextual information method we obtained from English: 514 French translations of MWTS with a precision of 56.8% and 510 German translations of MWTS with a precision of 51.2% for Top$_{10}$.

8 Conclusion

In this study, we have investigated the compilation of bilingual terminologies from a specialized comparable corpus and show how to push back the limits of the compositional approach used in alignment programs to translate MWTs. We have proposed CMCBP: a compositional method enhanced with pre-processed context information. The experiments that we carried out have shown that we increase the results of the compositional approach by providing a significant number of additional correct lexical entries that could not be identified either by the dictionary look-up or by compositional methods.

In future work, we will generalise this method to obtain an homogeneous modular design for all languages by reconsidering the rough distinction between simple and complex terms and applying the CMCBP both at the morphological and the lexical levels. We will investigate how to improve the solving of the fertility problem for MWTs which produces incomplete translations. Fertility was only partially solved thanks to the term variation treatment associated to the CMCBP. We aim to modify the evaluation protocol by accepting one-to-many translations in the case of synonym or semantically-related translation candidates that are not handled through term variation processing.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement no 248005.

References

Baldwin, T. and Tanaka, T. (2004). Translation by Machine of Complex Nominals: Getting it Right. In *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, pages 24–31, Barcelona, Spain.

Bowker, L. and Pearson, J. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. Routledge, London/New York.

Brown, P., Della Pietra, S., Della Pietra, V., and Mercer, R. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Chiao, Y.-C. and Zweigenbaum, P. (2002). Looking for Candidate Translational Equivalents in Specialized, Comparable Corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1208–1212.

Fung, P. (1998). A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora. In Farwell, D., Gerber, L., and Hovy, E., editors, *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, pages 1–16, Langhorne, PA, USA.

Gaussier, E., Renders, J.-M., Matveeva, I., Goutte, C., and Déjean, H. (2004). A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL'04)*, pages 526–533, Barcelona, Spain.

Grefenstette, G. (1999). The World Wide Web as a Resource for Example-Based Machine Translation Tasks. In *ASLIB'99 Translating and the Computer 21*, London, UK.

Keenan, E. L. and Faltz, L. M. (1985). *Boolean Semantics for Natural Language*. D. Reidel, Dordrecht, Holland.

Koehn, P. and Knight, K. (2002). Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*, pages 9–16, Philadelphia, Pennsylvania, USA.

Laroche, A. and Langlais, P. (2010). Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 617–625, Beijing, China.

McEnery, A.and Xiao, Z. (2007). Parallel and comparable corpora: What is happening? In Anderman, G. and Rogers, M., editors, *Incorporating Corpora: The Linguist and the Translator*, Multilingual Matters. Clevedon.

Melamed, I. D. (2001). *Empirical Methods for Exploiting Parallel Texts*. MIT Press, Cambridge, MA, USA.

Morin, E. and Daille, B. (2010). Compositionality and Lexical Alignment of Multi-word terms. In *Language Resources and Evaluation*, volume 44, pages 79–95. Springer.

Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (2007). Bilingual Terminology Mining – Using Brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 664–671, Prague, Czech Republic.

Munteanu, D. S. and Marcu, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'06)*, Sydney, Australia.

Nakagawa, H. and Mori, T. (2003). Automatic term recognition based on statistics of compound nouns and their components. *Terminology*, 9(2):201–219.

Nomura, M. and M., I. (1989). *Gakujutu Yogo Goki-Hyo*. National Language Research Institute, Tokyo.

Pekar, V., Mitkov, R., Blagoev, D., and Mulloni, A. (2006). Finding translations for low-frequency words in comparable corpora. *Machine Translation*, 20(4):247–266.

Rapp, R. (1999). 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)*, pages 519–526, College Park, MD, USA.

Robitaille, X., Sasaki, X., Tonoike, M., Sato, S., and Utsuro, S. (2006). Compiling French-Japanese Terminologies from the Web. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 225–232, Trento, Italy.

Rocheteau, J. and Daille, B. (2011). TTC TermSuite - A UIMA Application for Multilingual Terminology Extraction from Comparable Corpora. In *Proceedings of the IJCNLP 2011 System Demonstrations*, pages 9–12, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'02)*, pages 1–15, Mexico City, Mexico.

Savary, A. and Jacquemin, C. (2003). Reducing Information Variation in Text. In Grefenstette, G., editor, *Text- and Speech-Triggered Information Access*, Lecture Notes in Computer Science, pages 141–181. Springer Verlag.

Schmid, H. (1995). Improvements In Part-of-Speech Tagging With an Application To German. In *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

Sharoff, S., Babych, B., and Hartley, A. (2009). 'irrefragable answers' using comparable corpora to retrieve translation equivalents. *Language Resources and Evaluation*, 43(1):15–25.

Tanaka, T. (2002). Measuring the Similarity between Compound Nouns in Different Languages Using Non-parallel Corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1–7, Taipei, Taiwan.

Tanaka, T. and Baldwin, T. (2003). Noun-Noun Compound Machine Translation: A Feasibility Study on Shallow Processing. In *Proceedings of the ACL 2003 workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 17–24, Sapporo, Japan.

Vintar, S. (2010). Bilingual term recognition revisited: The bag-of-equivalents term alignment approach and its evaluation. *Terminology*, 16(2):141–158.

Weller, M. and Heid, U. (2012). Simple methods for dealing with term variation and term alignment. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. ELDA.