An approach based on multilingual thesauri and model combination for bilingual lexicon extraction

Hervé Déjean, Éric Gaussier
* Xerox Research Centre Europe
6, Chemin de Maupertuis,
38240 Meylan, France
firstname.lastname@xrce.xerox.com

Fatia Sadat
Graduate School of Information Science
Nara Institute of Science and Technology,
Nara, Japan
fatia-s@is.aist-nara.ac.jp

Abstract

This paper focuses on exploiting different models and methods in bilingual lexicon extraction, either from parallel or comparable corpora, in specialized domains. First, a special attention is given to the use of multilingual thesauri, and different search strategies based on such thesauri are investigated. Then, a method to combine the different models for bilingual lexicon extraction is presented. Our results show that the combination of the models significantly improves results, and that the use of the hierarchical information contained in our thesaurus, UMLS/MeSH, is of primary importance. Lastly, methods for bilingual terminology extraction and thesaurus enrichment are discussed.

Introduction

The growing availability of comparable corpora, through the Internet or via distribution agencies providing newspapers articles in different languages, has led researchers to develop methods to extract bilingual lexicons from such corpora, in order to enrich existing bilingual dictionaries, and help cross the language barrier for cross-language information retrieval. The results obtained thus far on comparable corpora, even though encouraging, are not completely satisfactory yet. (Fung, 2000) reports, for the Chinese-English language pair an accuracy of 76% to find the correct translation in the top 20 candidates, a figure we do not believe to be good enough to consider manual revision. Furthermore, the evaluation is carried out on 40 English words only. (Rapp, 1999) reaches 89% on the German-English language pair, when considering the top 10 candidates. If this figure is rather high, it was obtained on a set of 100 German words, which, even though not explicit in Rapp's paper, seem to be high frequency words, for which accurate and reliable statistics can be obtained.

We want to show in this paper how previously proposed methods can be extended to and improved for specialized domains. In particular we will focus on the use and enrichment of multilingual thesauri, which, even though partially related they may be to the texts under consideration, are nonetheless an available and valuable resource for the task. We rely in this work on two main linguistic resources: a general bilingual dictionary (available through the ELRA consortium[1]) and a specialized multilingual thesaurus (the Medical Subject Headings, MeSH, provided through the metathesaurus Unified Medical Language System, UMLS[2]). Without anticipating too much on the linguistic preprocessing we use, it has to be noted that, unless otherwise stated, when we speak of a "word" we refer to a single (as opposed to compound), lexical word (as opposed to stop word). All our examples and experiments use the (German, English) language pair.

1 Context vectors: a basic building block

Bilingual lexicon extraction from non-parallel but comparable corpora has been studied by a number of researchers, (Peters, 1995; Tanaka, 1996; Shahzad 1999; Rapp, 1999; Fung, 2000) among others. Their work relies on the assumption that if two words are mutual

[1] http://www.icp.grenet.fr/ELRA/home.html

[2] http://www.nlm.nih.gov/mesh/meshhome.html

translations, then their more frequent collocates (taken here in a very broad sense) are likely to be mutual translations as well. Based on this assumption, a standard approach consists in building context vectors, for each source and target word, which aim at capturing the most significant collocates. The target context vectors are then translated using a general bilingual dictionary, and compared with the source context vectors.

Our implementation of this strategy relies on the following steps, and follows the one given in (Rapp, 1999):

- for each word *w*, build a context vector by considering all the words occurring in a window encompassing several sentences that is run through the corpus. Each word *i* in the context vector of *w* is then weighted with a measure of its association with *w*. We chose the log-likelihood ratio test, (Dunning, 1993), to measure this association
- the context vectors of the target words are then translated with our general bilingual dictionary, leaving the weights unchanged (when several translations are proposed by the dictionary, we consider all of them with the same weight)
- the similarity of each source word *s*, for each target word *t*, is computed on the basis of the cosine measure
- the similarities are then normalized to yield a probabilistic translation lexicon, *P(t|s)*.

To illustrate the above steps, we give here the first 5 words of the context vector of the German word *Leber* (*liver*), together with their associated score: (Transplantation 138, Resektion 53, Metastase 41, Arterie 38, cirrhose 26). Once this context vector translated, the English top five becomes: (transplant 138, tumour 48, secondary 42, metastatis 41, artery 38). One can note that the German term *Resektion* was not found in our bilingual dictionary, and thus not translated. However, the translated context vector contains English terms characteristic of the co-occurrence pattern for *liver*, allowing one to associate the two words *Leber* and *liver*. We refer to the above method as the **standard method.**

2 Lexical translation model based on a multilingual thesaurus

A multilingual thesaurus bridges several languages through cross-language correspondences between concept classes (a concept class in the thesaurus links alternative names and views of the same concept together. For example, concept class C0751521, for which the main entry is *splenic neoplasms*, also contains *cancer of spleen*, *splenic cancer*, *spleen neoplasms*). The correspondence can be one-to-one, i.e. the same concept classes are used in the different languages, or many-to-many, i.e. different concept classes are used in different languages, and a given concept class in a given language corresponds to zero, one or more concept classes in the other languages. The correspondence between concept classes across languages helps us write the probability *P(t|s)* of selecting word *t* as a translation of word *s* in the following general way, where *C* represents a multilingual concept class in MeSH (we omit the derivation, which is mainly technical, and uses the fact that the correspondence between concept classes in MeSH is one-to-one):

$$P(t|s) = \Sigma_C P(C|s) P(t|C,s) \quad (1)$$

a formula which can be interpreted as follows: from a source word *s* of the source corpus, select a (interlingual) concept class in the thesaurus, according to *P(C|s)*, then generate a target word *t* of the target corpus from the concept class and the source word, according to *P(t|C,s)*. The dependence on *s* in the last probability distribution (*P(t|C,s)*) allows one to privilege one possible lexicalization of a given concept class. It could be used, for example, to choose *spleen neoplasms* from concept class C0751521 as the translation of *Milztumoren*. However, since such a distinction between the different lexicalizations of a given concept is beyond the scope of the current paper, we make the additional simplifying assumption that, given a concept class, the target word *t* is independent of the source word *s*, which leads to the simplified formula:

$$P(t|s) \approx \Sigma_C P(C|s) P(t|C) \quad (2)$$

The above equation views the thesaurus as a trellis linking source and target words. As such, given probabilities *P(C|s)* and *P(t|C)* (see section 2.2 for the way we estimate these probabilities), there are several ways to compute an association score between source and target words. The most obvious one is to carry the sum over all concept classes, or a large subset of them, as indicated by the formula. We refer to

this method as the **complete search**. However, if the relation between a word and a concept class is not significant, the complete search has the disadvantage of bringing noisy data in the estimation of *P(t|s)*. An alternate solution is to select just the concept class which maximizes the association between *s* and *t*. Because of its analogy with the Viterbi algorithm, we refer to this method as the **Viterbi search**.

Nevertheless, neither the complete nor the Viterbi search makes use of the hierarchical information contained in the thesaurus, which is, in the above formulations, mainly viewed as a specialized lexicon. We present below a third search strategy which directly makes use of the structure of the thesaurus. For reasons that will become clear, we call this strategy the **subtree search**.

2.1 The subtree search

Complete search and the Viterbi search represent two extreme ways of making use of the thesaurus since they consider either all or only one of the concept classes it contains. In order to find a way in-between and to focus on a subset of interesting concept classes, we first select for each source word *s* the *n* best concept classes in the thesaurus, i.e. the first *n* concept classes according to the probability distribution *P(C|s)*. We then extend this set of classes by adding new classes using the hierarchy in the thesaurus.

Intuitively, if two or more classes in the selected subset have the same parent class, then the source word is likely to be related to this parent as well as to the classes themselves, since the parent is the direct node "conceptually" linking the classes. For example, if a source word *s* selects the two classes *Hepatitis* and *Cirrhosis*, then *s* is likely to be related to *Liver Diseases*, the parent class. We make use of this intuition in the following way: for each pair of classes from the set of the *n* best classes associated with source word *s*, select the subtree formed by the classes, their common ancestor, and all the nodes that appear between the classes and their ancestor.

This algorithm provides a set of subtrees from the 15 sub-thesauri corresponding to the 15 main categories of the MESH classification (MeSH, rather than being a single thesaurus, contains 15 different sub-thesauri, artificially related through a common root node in UMLS. We do not make use of this distinction in the complete and Viterbi methods, but use it for the subtree search to avoid linking classes via the artificial root concept). One can also note that the above algorithm suggests a way to identify polysemous words, or words used through different points of view, via the different sub-thesauri they select subtrees from. This refinement, which should lead to more fine-grained bilingual lexicons, will be the focus of future research.

The set of classes contained in the subtrees is then used in equation (2) to derive associations between source and target words.

2.2 Linking words and concept classes

The estimation of the probability distributions *P(C|s)* and *P(t|C)* used in equation (2) can be easily carried out by resorting once again to context vectors. Indeed, if a word of the corpus is similar to a term present in a concept class, then they are likely to share similar contexts and have similar context vectors. We thus extend the notion of context vectors to concept classes, and rely again on the cosine measure to compute similarities between words and concept classes. The probability distributions *P(C|s)* and *P(t|C)* are finally derived through normalization.

To build a context vector for a concept class, we first build the context vector of each term the class contains. For single-word units, we directly rely on the context vectors extracted in section 1. If the term is a multi-word unit, as *liver disease*, we consider the conjunction of the context vectors of each word in the unit, normalizing the weights by the number of words in the unit. For example, the context vector for *liver disease* will contain only those words that appear in the context of both *liver* and *disease*, since the whole unit is a narrower concept than its constituents. We then take the disjunction of all context vectors of each entry term in the class, normalizing the weights by the number of terms in the class, to build the context vector of each concept class.

The following example illustrates the complete process: the German spelling variant *Actinomykose* is used in our corpus in addition to *Aktinomykose*, which is the only form listed in the UMLS class C0001261; nevertheless, our

process associates C0001261 as the closest class to *Actinomykose* and *actinomycosis* (English), and retain them as translation candidates.

3 Combining different models

The previous section provides us with two different probabilistic lexical translation models: one derived from the standard method, and one based on the bilingual thesaurus. A third lexical translation model can be directly derived from the bilingual dictionary by considering the different translations of a given entry as equiprobable. For example, our dictionary associates *abbilden* with the two words *depict* and *portray*, thus *P(depict|abbilden) = P(portray|abbilden) = 0.5*. Note that these three models are not independent of each other, since the corpus is used, through the estimation of *P(C|s)* and *P(t|C)*, in the thesaurus-based model, and the bilingual dictionary is used for translating context vectors in the corpus-based model. The final estimate of *P(t|s)* is then based on the following mixture of models:

$$P(t|s) = \Sigma_i P(i|f(s)) Pi(t|s) \quad (3)$$

where *i* is an integer used to index the different models (here $1 \leq i \leq 3$), and *P(i|f(s))* denotes the probability of selecting model *i* based on characteristics of *s* (f is a function mapping the source word to a set of relevant features) . The problem is now one of estimating the mixture weights *P(i|f(s))*, which can be done by maximizing the likelihood of some held-out data. To this end, we manually created a reference bilingual lexicon, part of which is reserved for estimating the mixture weights. Let *l* denote the part of the reference lexicon we use for estimation purposes, and *l(s)* the set of translations of *s* in *l*. The mixture weights are obtained through a standard constrained optimization problem, and are given by:

$$P(i \mid f(s) = g) = \frac{\sum_{t \in l(s), f(s)=g} Pi(t \mid s)}{\sum_k \sum_{t \in l(s), f(s)=g} Pk(t \mid s)} \quad (4)$$

The set of features we retained aim at capturing the reliability of each model for a given source word. The reliability of the standard method can be indirectly measured through the frequency of *s*, the more frequent *s* is, the more reliable the information available to this method is. We capture this with a binary valued attribute, being 1 if *s* occurs at least 5 times in our corpus, and 0 otherwise. Similarly, the reliability of the thesaurus-based model uses a binary valued attribute which is 1 if *s* is close to the thesaurus (i.e. if $\arg\max_C P(C \mid s)$ is greater than 0.5) and 0 otherwise. For the dictionary-based model, the reliability is directly computed in terms of presence/absence of *s* in the dictionary. The above thresholds were empirically tuned, and constitute what we believe to be a good compromise between fine-grained mixtures and data sparseness problems.

Nevertheless, despite this tuning, some configurations of the above attributes still suffer estimation problems. Starting with a reference lexicon containing 1,800 translation pairs, we used 10 different splits into estimation and evaluation lexicons (two third of the data are reserved for estimation, one third for evaluation), and then estimated the mixture weights on each split. The results show that the variance for the configuration “low frequency, not in thesaurus, not in dictionary" is 10 times larger than the variance obtained for the other 7 configurations. Unfortunately, many source words fall into this configuration. We thus decided to fall back on a simplified version of equation 3 in which the dependence of *i* on f(s) is dropped (the adaptation of equation 4 is straightforward). This time, the variance is around 10^{-4}, 5 times lower than the lowest value previously obtained, thus rendering the estimation of the mixture weights more reliable. Table 1 below presents the mixture weights finally obtained for the different search methods.

	Viterbi	Complete	Subtree
corpus	0.59	0.45	0.33
thesaurus	0.1	0.24	0.37
dictionary	0.31	0.31	0.29

Table 1: Mixture weights for the 3 models

4 Linguistic preprocessing

As a preprocessing step, we tag and lemmatize texts in both languages. This step allows us to focus on content words only (nouns, verbs, adjectives and adverbs), and reduces the noise in our model (content words are the primary focus for thesaurus enrichment and cross-language

information retrieval). Nevertheless, since we use the (German, English) language pair for all our experiments, a major problem still resides in the difference in the word definition between the two languages, mainly due to the particular usage of compounding the German language has. Two alternatives are offered: either use a direct phrasal alignment, or decompose the German compounds into smaller units. Inasmuch as the models presented in the preceding sections implicitly assume a one-to-one correspondence between words in the two languages, we rely on the second strategy. However, an additional complication is introduced by the fact that our corpora belong to the medical domain, thus leaving our German lemmatizer clueless when it comes to decomposing medical compounds. We thus used two additional heuristics, recursively applied on all German words:

- some sequences, e.g. -ungs-, -heits-, -keits-, -schafts, -aets- and -ions-, as well as their plural forms, are considered as boundaries between two words in a compound, and break a word into two parts
- if a word is composed of the sequence AB, and if A and B are both longer than 3 characters and both occur in the corpus, then the sequence AB is decomposed into A and B.

The above heuristics reduce the number of different lemmas in the German vocabulary by 28% (from 14,700 to 10,500), while not hurting too much the quality of the vocabulary since their precision is estimated to be above 90%. For example, they allow us to accurately decompose the compound *Adhaesionsileusbehandlung* into the three parts *Adhaesion*, *Ileus* and *Behandlung*.

5 Experiments and results

To test the above models and their combination, we used roughly 700 abstracts from MEDLINE[3], in German and English (each portion, German and English, contains approximately 100,000 words). These abstracts are "partial" translations of each other, because in some cases the English writer directly summarizes the articles in English, rather than translating the German abstracts. That set of abstracts is used both as our comparable corpus, in which case we do not make use of alignment information, and as our parallel corpus (see section 6). There is a continuum from parallel corpora to fully unrelated texts, going through comparable corpora. The comparable corpus we use is in a way "ideal" and is biased in the sense that we know the translation of a German word of the German corpus to be, almost certainly, present in the English corpus. However, this bias, already present in previous works, does not impact the comparison of the methods we are interested in, all methods being equally affected. Indeed, the results we obtain with the standard method (see below) are in the range of those reported in previous works.

As already mentioned, we manually extracted a reference lexicon comprising 1,800 translation pairs from our comparable corpus. From this, we reserve approximately 1,200 pairs for estimating the mixture weights, and 600 for the evaluation proper. All our results are averaged over 10 different such splits. Since the models we rely on yield a ranked set of translation candidates for each source word, and since one cannot expect the right translation to be *the* first candidate, we compute precision and recall of each method in the following way: for each pair *(s,t)* in the evaluation lexicon, we consider the first *p* candidates provided for *s* by the method under evaluation, and judge the set as correct if it contains *t*, as incorrect otherwise; precision is then obtained by dividing the number of correct sets by the number of sets proposed by the method for the words in the evaluation lexicon, whereas recall is obtained by dividing the number of correct sets by the number of pairs in the evaluation lexicon. In addition, we evaluate the average rank of the first correct translation in the proposed list of translations, for each method.

Table 2 shows the results we obtained on our comparable corpora, for *p=10*, without combining the different models. ST50 refers to the subtree search strategy within the thesaurus, with *n=50*. The precision of the dictionary-based model is around 78%, which is not that bad considering the domain we focused on, but, as one can expect, its recall reaches only 48%.

[3] http://www4.ncbi.nlm.nih.gov/PubMed/

The F1-score, which combines precision and recall, obtained for the corpus-based model is similar to the ones obtained in previous works.

Model	Dictionary	Corpus	Thesaurus
F1-score	56.16	62.04	51.34

Table 2: Results for separate models

Table 3 presents the results (F1-score) we obtained with the different search strategies for the thesaurus-based model: the Viterbi search , the complete one considering the first 100 and first 200 concept classes for each source word, and the subtree search with different values of *n*), and two different values for *p*, 5 and 10. The average rank is given next to each F1-score. As one can see, the combination significantly improves the results over the models alone, since the F1-score goes from 62% to 84%, a score that may be good enough to consider manual revisions.

	P=5	p=10
Viterbi	71.3/14.7	79.7/14.7
Complete(100)	75.4/14.1	80.3/14.1
Complete(200)	75.4/12.3	83.2/12.3
SubTree 10	75.8/11	82.4/11
SubTree 20	76.4/11.7	84.1/11.7
SubTree 50	77.3/11.2	83.6/11.2
SubTree 100	76.9/11.8	83/11.8

Table 3: Evaluation of search strategies.

Furthermore, the best results are obtained with the subtree search, with *n=20*, thus validating our hypothesis that using the structure of the thesaurus is beneficial. One can note however that the results obtained with the complete search using 200 classes are close to the best results. Nevertheless, the optimal subtree search (ST20) uses 7.5 times less classes than the complete search, and is also two times faster. This proves that the subtree search is able to focus on accurate concept classes in the thesaurus, whereas the complete search needs considering more classes to reach a comparable level of performance. Interestingly, it also seems that the candidates provided by the subtree search closely correspond to a semantic field, whereas the ones given by the complete search are more varied. Where this to be the case, the subtree search would also certainly outperform the other methods when used for cross-language information retrieval. We will try to validate this hypothesis in future work.

6 Bilingual terminology extraction

Bilingual terminology extraction is based on three steps: word alignment, term extraction term alignment.

In this section, we rely on the word to word translation lexicon obtained from the parallel corpus, following the method described in (Gaussier *et al*., 2000).

6.1 Term extraction

For identifying German and English candidate terms we use the following patterns, similar to those proposed by (Heid, 1999) and (Blanck, 2000):

1. single words which appear in the thesaurus (for alignment purposes) or which contain English morphemes extracted from *The Specialized Lexicon* found in UMLS and translated in German .

2. syntactic patterns: [(ADJ)+ NOUN GEN+] and [ADJ+ NOUN (GEN)+] for German, and all non recursive noun phrases for English.

Our morpheme list contains 40 elements, some of which are general, *-ion*, *-ung*, but the majority of which is specific to the medical domain, *-ektomie*, *-itis*. The syntactic patterns match nouns which occur with a complement (adjective and/or genitive structures). The German sequence *problematösen Gebieten der Chirurgie* is then defined as a candidate term, when the English translation *problematic fields of surgery* is composed of two candidate terms: *problematic fields* and *surgery*.

6.2 Term alignment

Our algorithm allows alignment of a sequence of candidate terms, and follows the one proposed in (Hull, 1997). We first try to align candidate terms, and then test if a longer unit, composed of several candidate terms, improve the alignment score. A unit is extended if and only if the next contiguous candidate term is a prepositional phrase, the relaxation of this constraint introducing too much noise. The extension stops when the score is lower than the score of the "non-extended term". For instance, an alignment score is computed for [*problematösen Gebieten der Chirurgie*] and [*problematic fields*]. Then the English term is extended to [[*problematic fields*] of [*surgery*]], which provides a better

alignment score, and is then kept. In this particular example, neither the German nor the English units can be further extended, since the German term occurs at the end of a sentence and the English unit is not followed by a prepositional phrase. The German candidate *problematösen Gebieten der Chirurgie* is thus finally aligned with the English candidate *problematic fields of surgery*.

Most German compounds, decomposed for word alignment purposes, are aligned with English terms corresponding to a sequence *adjective+noun* (*Nierenfunktion*/*renal function*) or *noun+of+noun* (*Lebensqualitaet*/*quality of live*). Correspondences between acronyms and translated developed forms can also be found (*Nierenzellcarcinom*/*RCC*). In practice, no unit composed of three candidate terms is found. The longest units are generated by German candidate term with a genitive structure (*Plattenepithelcarcinom des Oesophagus*/*squamous cell esophageal cancer*).

We manually extracted 150 candidate terms with their translation for evaluating our procedure. Table 4 shows precision and recall for our method. If the first 5 candidates are retained, the F1-score reaches 80%. Precision is always higher than recall, which can be explained by the fact that the reference terms were extracted manually when the automatic extraction can propose incorrect units due to chunking errors.

	precision	Recall
1	56.52	50.98
2	71.01	64.05
5	84.78	76.47
10	89.85	81.04

Table 4: Evaluation of term alignment

Conclusion

We have shown in this paper how to optimally combine different models derived from different resources for bilingual lexicon extraction from comparable corpora. We have proven that such a combination significantly (by 30%) improves the results over the models alone. We have also presented different models based on a multilingual thesaurus, and have obtained the best results with the model integrating hierarchical information. Lastly we have proposed different ways to enrich existing thesauri with new terms discovered in parallel corpora. Future work should focus on terminology extraction from comparable corpora.

Acknowledgements

We wish to thank anonymous reviewers for useful comments on the first version of this paper. The research herein described has in part been supported by the EC/NSF grant IST-1999-11438 for the MUCHMORE project.

References

Blank, I., 2000. Terminology extraction from parallel technical texts. In J. Véronis (Ed.), Parallel Text Processing - Alignment and Use of Translation Corpora. Kluwer Academic Publishers.

Dunning, T., 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):64-74.

Fung, P., 2000. A statistical view on bilingual lexicon extraction - From parallel corpora to non-parallel corpora. In J. Véronis (Ed.), Parallel Text Processing - Alignment and Use of Translation Corpora. Kluwer Academic Publishers.

Gaussier, E., Hull, D., Ait-Mokhtar, S., 2000. Term alignment in use: Machine-aided human translation. In J. Véronis (Ed.), Parallel Text Processing - Alignment and Use of Translation Corpora. Kluwer Academic Publishers.

Heid, U., 1999. A linguistic bootstrapping approach to the extraction of term candidates from German text. *Terminology*, 5(2).

Hull, D., 1997. Automating the construction of bilingual terminology lexicons. *Terminology*, 4(2).

Peters, C., Picchi, E., 1995. Capturing the comparable: A system for querying comparable text corpora. *JADT Proceedings.*

Rapp, R., 1999. Automatic identification of word translations from unrelated English and German corpora, ACL *Proceedings.*

Shahzad, I., Ohtake, K., Masuyama, S. Yamamoto, K., 1999. Identifying translations of compound nouns using non-aligned corpora. *Workshop MAL Proceedings.*

Tanaka, K., Iwasaki, H., 1996. Extraction of lexical translations from non-aligned corpora. *COLING Proceedings.*

Vivaldi, J., Rodriguez, H., 2001. Improving term extraction by combining different techniques. *Terminology*, 7(1).