Automatic Processing of Multilingual Medical Terminology: Applications to Thesaurus Enrichment and Cross-Language Information Retrieval

H. Déjean[†], E. Gaussier[†], J.-M. Renders[†], F. Sadat[‡]

† Xerox Research Centre Europe

6, chemin de Maupertuis

F-38240 Meylan, France

‡ Nara Institute of Science and Technology,

Ikoma, Nara, Japan

Corresponding author: Hervé Déjean

Tel: 33 4 76 61 50 81 – Fax: 33 4 76 61 50 99 – Email: Herve.Dejean@xrce.xerox.com

Abstract

We present in this article experiments on Multi-Language Information Extraction and Access in the medical domain. Methods for extracting bilingual lexicons from parallel and comparable corpora are described and their use in Multi-Language Information Access is illustrated. Our experiments show that these automatically extracted bilingual lexicons are accurate enough for semi-automatically enriching mono- or bilingual thesauri (such as UMLS), and that their use in Cross-language Information Retrieval (CLIR) significantly improves the retrieval performance and clearly outperforms existing bilingual lexicon resources (both general lexicons and specialized ones).

Key words: *Bilingual Lexicon Extraction from Parallel and Comparable Corpora, Thesaurus Enrichment, Corpus Linguistics, Cross-language Medical Information Retrieval*

1. Introduction

Recent advances in Natural Language Processing (NLP) and Information Retrieval (IR) owe a great deal to the corpus-based approaches. The use of large text corpora has become a standard in language processing and text retrieval: they

represent resources for automatic lexical acquisition and multilingual lexical resource enrichment, they provide indispensable training data for statistical translation models and they can provide the connection between vocabularies in Cross-Language Information Retrieval (CLIR), which consists of retrieving documents written in one language using queries written in another language.

When working on specialized languages and specific domains, terminology plays a crucial role, inasmuch as it aims at describing and organizing the knowledge of the domain through the concepts, and their lexical realizations, that are used. Enriching terminological resources like thesauri thus requires first to identify the different terms of a domain, and then to relate them to existing concepts in the thesaurus. When one is interested in multilingual thesauri, the additional step of linking terms across languages is required. In specialized domains, parallel texts, i.e. texts in two or more languages which are translation of each other, provide an ideal material to follow, from a multilingual point of view, the evolution of the terminology of a domain, and to update existing resources. Research on bilingual lexicon extraction from parallel corpora has mainly focused on single word lexicons (as opposed to terms, usually made up of several words), with few works ([11], [16]) explicitly addressing the problem of extracting multilingual term lexicons. In the present paper, we first focus on the task of enriching a domain-specific thesaurus through bilingual lexicon extraction from parallel corpora. The thesaurus we are interested in is the Medical Subject Headings (MeSH), and its German version DMD, available through the Unified Medical Language System (UMLS). We also show how the extracted lexicons can be useful for CLIR in the medical domain.

However, with the explosive growth of the World Wide Web, it has become obvious that comparable corpora, i.e. multilingual texts which are not direct translations of each other, but cover similar topics, could be collected more easily from downloading electronic copies of newspapers, articles and other resources for specific domains and languages. This has led researchers to develop methods for bilingual lexicon extraction from comparable corpora, in order to enrich existing bilingual dictionaries, and help bridge the language barrier for cross-language information retrieval. The results obtained thus far on comparable corpora, even though encouraging, are not completely satisfactory yet. In [9] it is reported, for the Chinese-English language pairs an accuracy of 76% in finding the correct translation in the top 20 candidates, a figure we do not believe to be good enough to deliver for manual revision. Furthermore, the evaluation is carried out on 40 English words only. In [21] the evaluation score 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. In the second part

of the paper, we propose a new method for extracting bilingual lexicons from comparable corpora, and show that the combination of this new method with previous ones yields a significant improvement in the accuracy of the lexicons obtained. We also show results obtained in CLIR with this lexicon. Finally, the main goal of our research is to show that one can develop methods for bilingual lexicon extraction, from parallel and comparable corpora, and tools, based on the extracted lexicons, first to help terminologists enrich existing thesauri, and second to help users find relevant information in different languages.

The remainder of the present paper is organized as follows: section 2 presents the experimental framework used in this work; section 3 describes the linguistic processing applied on this data; bilingual lexical extraction from parallel corpora and its use in CLIR are presented in section 4; bilingual lexical extraction from comparable corpora, the integration of multilingual thesauri with hierarchical search strategies as well as a combination of different models and their use in CLIR are described in section 5; finally, section 6 concludes the paper.

2. Experimental background

2.1. Description of the Data

For our experiments, we use several resources: a corpus, a general bilingual lexicon and a specialized thesaurus. We focus in this article on two languages, German and English, and on one specific domain: the medical domain. Our corpus is based on the bilingual collection provided by the MuchMore project (http://muchmore.dfki.de). The MuchMore corpus is composed of 9,300 medical abstracts extracted from the Springer Verlag website, corresponding to about 1,200,000 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. This corpus is used as a comparable corpus, in which case we do not make use of alignment information. In order to rely on a parallel corpus for which sentences are well aligned, only a subpart (5,500 abstracts, 900,000 words) of the original corpus is used to form our parallel corpus (some German abstracts contain only a title where the English ones contain both title and body).. 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 section 5) are in the range of those reported in previous works.

As a general bilingual resource, we use the German/English ELRA dictionary (http://icp.grenet.de/ELRA/home.html), which contains about 50,000 bilingual entries. The medical thesaurus used is MeSH (Medical Subject Headings), and its German version, DMD, provided by DIMDI (http://dimdi.de). Both thesauri were extracted from UMLS (Universal Medical Language System: http://www.nlm.nih.gov/mesh/meshhome.html). Through UMLS the MeSH English entries and the DMD German entries are aligned, so we can extract a bilingual thesaurus. Since DMD is smaller than MeSH, the resulting bilingual thesaurus contains only 15,000 bilingual entries, while MeSH contains 200,000 entries.

2.2. Evaluation measures

To evaluate the results of the different methods for bilingual lexicon extraction, we first manually extracted reference lexicons for single words and terms from our parallel corpus (1000 single words and 150 multi-words terms). Since the models we rely on yield a ranked set of translation candidates for each source word, we compute precision and recall of each method in the following way: for each pair *(s,t)* in the reference 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*, otherwise it is marked incorrect. 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 reference lexicon, whereas recall is obtained by dividing the number of correct sets by the number of pairs in the reference lexicon. In addition to precision and recall, we also compute the average rank of the first correct translation in the proposed sets, as well as the F1-score, a synthetic measure based on precision and recall (the F1-score is the harmonic mean of precision and recall);

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

To evaluate the results for cross-language information retrieval, bilingual queries and associated relevance judgements were established by American and German medical students within the MuchMore project. We rely on the standard synthetic measure used in IR, namely macro-averaged precision and break-even point (macro-average gives equal importance to all queries; see [26] for more details).

3. Linguistic Processing

In order to extract bilingual terminology from corpora, we first need to normalize the different variants of an item (such as the different conjugations of a verb) to a given form, which will be used as the entry in the lexicons. Such a normalization corresponds to an automatic procedure, called **lemmatization**, which is based on a) a morphological analysis step, which assigns to each word its different parts-of-speech, as well as associated information on number, gender, declensions and conjugations, if any; and b) part-of-speech tagging, which consists in selecting, according to the context of each word, the correct part-of-speech and associated information. For example, the word *rate* can be a noun or a verb, but in the sequence "*The hepatocyte respiratory rate …*", it is most certainly a noun. These steps allow us to focus on content words (nouns, verbs, adjectives and adverbs), which will constitute the main entries of the lexicons we are going to consider. The other words, called "empty words", correspond to grammatical words which do not bear meaning. They are of less interest for the tasks we target, and are usually filtered out.

Nevertheless, since we work on German and English, a major problem still resides in the difference in the word definition between the two languages, mainly due to productive compounding in German. For example, the German compound *Kreuzband* is translated in English as *cruciate ligament*, a 2-word sequence. 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 following 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, developed for general language, clueless when it comes to decomposing medical compounds. Several methods have been proposed for decomposition of German compounds ([4], [13], [20], [26], [27]). The method we propose relies on the following two heuristics, recursively applied on all German words composed of more than 7 letters:

- Some join sequences e.g., -ungs- -heits-, -keits-, -schafts-, -ets- and –ions-, as well as plural endings are considered as morpheme boundaries between two words in a compound and break a word into two units. This heuristic decomposes the word Adhaesionsileusbehandlung into Adhaesion and ileusbehandlung. The second part is also a compound, but it can be decomposed only by applying additional heuristics.

- If a word is composed of the sequence AB and if A and B are both longer than three characters and both occur stand-alone in the corpus, then the sequence AB is decomposed into two components A and B. If we come back to the preceding example (*Adhaesionsileusbehandlung*), this heuristic is applied on the second part (*ileusbehandlung*), and since the word *Ileus* and *Behandlung* occur in our corpus, this compound is finally correctly decomposed thanks to both heuristics.

Together these heuristics reduce the number of lemmas in the German vocabulary by 28% (from 14,700 to 10,500); though not perfect, precision is estimated to be above 90%.

4. Bilingual Lexicon Extraction from Parallel Corpora

We present in this section the method we follow to extract bilingual lexicons from parallel corpora. A parallel corpus is a bilingual corpus the elements of which are translations of each other, and which is aligned, usually through an automatic procedure, at the sentence level (which means that each sentence in one language is associated with its translation in the other language). The methodology we follow to extract bilingual lexicons made of single and multi-word units is based on three steps:

1. Word alignment across languages,
2. Term extraction in both languages,
3. Term alignment across languages, based on the outputs of steps 1 and 2.

The multi-word units we will consider are terms, part of the medical terminology. We first describe the method we use for step 1, focusing on single word units, and then address the problem of finding translation equivalents for terms.

4.1. Single Word Lexicons

Recent research has demonstrated that statistical alignment models can be highly successful at extracting word correspondences from parallel corpora; see for example [3], [5], [10], [11], [15], [16]. All these works are based on the assumption that, once documents have been aligned at the sentence level, the more two words from different languages co-occur in aligned sentences, the more likely they are translations of each other. In the present paper, we rely on the word-to-word translation lexicon obtained from parallel corpora, following the method described in [12], which can be summarized as follows.

We first represent co-occurrences between words across translations by a matrix, the rows of which represent the source language words[1], the columns the target language words, and the elements of the matrix the *expected alignment frequencies* (*EAFs*) for the words appearing in the corresponding row and column. Empty words are added in both languages in order to deal with words with no equivalent in the other language.

The estimation of the expected alignment frequency is based on the Iterative Proportional Fitting Procedure (IPFP) presented in [1]. This iterative procedure updates, at the k^{th} iteration, the current estimate $n_{ij}^{(k)}$ of the EAF of source word *i* with target word *j*, using the following two-stage equations:

$$n_{ij}^{(k,1)} = \sum_{s,(i,j)\in s} n_{ij}^{(k-1,2)} \times \frac{s_i}{n_{i.}^{(k-1,2)}}$$
$$n_{ij}^{(k,2)} = \sum_{s,(i,j)\in s} n_{ij}^{(k,1)} \times \frac{s_j}{n_{.j}^{(k,1)}} \qquad (1)$$

where $n_{i.}$ and $n_{.j}$ are the current estimates of the row and column marginals, *s* is a pair of aligned sentences containing words *i* and *j*, and s_i and s_j are the observed frequencies of words *i* and *j* in *s*. The initial estimates $n_{ij}^{(0,2)}$ are the observed frequencies of co-occurrences, obtained by considering each pair of aligned sentences and by incrementing the alignment frequencies accordingly. The sequence of updates will eventually converge and the EAFs are then normalized (by dividing each element n_{ij} by the row marginal $n_{i.}$), so as to yield probabilistic translation lexicons, in which each source word is associated with a target word through a score. In the remainder of the paper, we will use *P(t|s)* to denote the probability of selecting target word *t* as translation for source word *s*.

4.2. Term Lexicons

In order to illustrate the term alignment process, we will use the following aligned sentences as basic examples, and will describe each step with them:

Das hintere Kreuzband ist das kraeftigste Band des menschlichen Kniegelenks. (1)

The posterior cruciate ligament (PCL) is the strongest ligament of the human knee joint. (2)

Term extraction

[1] We use source and target to refer to elements of different languages, which does not imply, in our case, any privileged direction.

For identifying German and English candidate terms we rely on the following characteristics, similar to those proposed by Heid [14] and Blank [2]:

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 into German. Since the *Specialized Lexicon* provides only English morphemes, we created a list of German morphemes by translating the English ones (thus -ectomy is translated into -ektomie),

2. Syntactic patterns corresponding to [ADJ* NOUN GEN*] for German, and all non recursive noun phrases for English, i.e. noun phrases of the form [(ADJ* NOUN*)* NOUN], where ADJ stands for an adjective, NOUN for a noun, GEN for a genitive noun phrase (corresponding to the pattern [DET* ADJ* NOUN] with the noun tagged genitive), and * for the Kleene star.

Our morpheme list contains 40 elements, some of which are general, *-ion*, *-ung*, but the majority of which are specific to the medical domain, *-ektomie*, *-itis*. The above morpho-syntactic patterns match nouns which occur with a complement (adjective and/or genitive structures), and cover the vast majority of multi-word terms present in the UMLS thesaurus (ca. 85% for both German and English). The German and English terms not covered by these morpho-syntactic patterns correspond to noun phrases including coordination or prepositions (as *strongest ligament of the human knee joint*, based on the composition of two sub-terms *strongest ligament* and *human knee joint* with the preposition *of*). However, most of the noun phrases including coordination or prepositions in German and English are not actual terms, and extending the above patterns to deal with coordination and prepositions would lead to the extraction of too many irrelevant candidates. Furthermore, actual terms including prepositions can be decomposed, as in the above example, into simpler terms, captured by the above patterns. This suggests that it should be posible to retrieve alignments between complex terms from the alignments between simpler ones, an extension that we describe below.

From the above patterns, we extract the following German and English terms from sentences (1) and (2):

hintere Kreuzband (ADJ NOUN); *kraeftigste Band des menschlichen Kniegelenks (*ADJ NOUN GEN)

posterior cruciate ligament (ADJ ADJ NOUN); *strongest ligament* (ADJ NOUN); *human knee joint* (ADJ NOUN NOUN)

The German sequence *kraeftigste Band des menschlichen Kniegelenks* is then defined as a candidate term, when the English translation *strongest ligament of the human knee joint* is composed of two candidate terms: *strongest ligament* and

human knee joint. This type of mismatch is inherent to the strategy we have adopted, since it is very unlikely that, due to peculiarities of each language, the German and the English term grammars extract exactly the same units. The patterns we considered focus on minimal units for which the terminological nature is almost always certain.

Term alignment

However, the above considerations show that it may be necessary to extend some units so as to get translation equivalents. Indeed, some terms comprise prepositional phrases that need be recovered if one wants to provide adequate translations for most German genitive structures. This extension has to be conditioned on an optimization criterion, in order to control its scope. To this end, we define a measure of association between candidate terms, based on single word associations. Lastly, because in a pair of aligned sentences, a given source term is usually translated by a single target term, we make use of a one-to-one alignment constraint to filter the list of candidate translations. This leads us to the following strategy:

1. Compute an association score between candidate terms,
2. Extend candidate terms whenever this improves the alignment score,
3. Filter the set of candidate translations.

The association score for step 1 is based on the probabilistic lexicon extracted for single words. Let *s* be a source candidate term composed of *n* single words and *t* a target candidate term composed of *l* words; the association score between *s* and *t* is given by:

$$score(s,t) = \prod_{i=1}^{n} \sum_{j=1}^{l} \frac{p(t_j \mid s_i)}{l} = \frac{1}{l^n} \prod_{i=1}^{n} \sum_{j=1}^{l} p(t_j \mid s_i)$$

where $P(t_j|s_i)$ is the probability of translating the single word s_i by the single-word t_j, as obtained from section 4.1. The sum in the right-hand side of the first equation corresponds to the probability of translating s_i by one of the elements of *t*, normalized by the length of t^2. It thus corresponds to the average probability of translating s_i in *t*. Collecting those contributions in a product over *i* amounts to computing the probability that *s* are translates as *t*.

For step 2, not all extensions are considered. An English term is extended if and only if the next contiguous candidate term is a prepositional phrase introduced by the preposition *of*, the relaxation of this constraint introducing too much noise (in addition to the disputable terminological status of sequences containing prepositions different from *of*, we face the prepositional-phrase attachment problem – where to attach a prepositional phrase in a sentence-, which has not received

satisfactory solutions yet, except for the preposition *of*, for which the attachment to the preceding noun phrase is a valid heuristics).

For German, we allow as extension all prepositional phrases. One important grammatical construction we do not cover is the coordination of nouns (i.e. *Academies and Institutes*), present in approximately 5% of the English and German terms of UMLS, since such coordinated structures are difficult to precisely extract.

The extension stops when the score is lower than the score of the "non-extended term". For instance, from the list of candidate terms in our example, an alignment score is computed between each source candidate term and each target candidate term:

(*hintere Kreuzband*; *posterior cruciate ligament:* 0.0009),

(*hintere Kreuzband; strongest ligament:*3.17e-07),

(*hintere Kreuzband; human knee joint:*6.17e-16)

(Kraeftigste Band des menschlichen Kniegelenks; posterior cruciate ligament: 3.45e-20*)*

(Kraeftigste Band des menschlichen Kniegelenks; strongest ligament: 1.36e-12*)*

(Kraeftigste Band des menschlichen Kniegelenks; human knee joint: 2.37e-07*)*

From this, only [*strongest ligament]* can be extended with *human knee* (only contiguous prepositional terms are considered). Then the English term is extended to [*strongest ligament of the human knee joint*], and its score with the German terms is computed. In our case, the score between this extended term and the German one [*kraeftigste Band des menschlichen Kniegelenks*] provides a better alignment score (8.26e-05) than with the other basic terms, and is then kept. In this particular example, neither the German nor the English units can be further extended, since the term occurs at the end of a sentence. The new list after term extension is then:

(*hintere Kreuzband*; *posterior cruciate ligament:* 0.0009),

(*hintere Kreuzband*; *strongest ligament:* 3.17e-07),

(*hintere Kreuzband*; *human knee joint:* 6.17e-16),

(Kraeftigste Band des menschlichen Kniegelenks; strongest ligament of the human knee joint: 2.87e-05),

(Kraeftigste Band des menschlichen Kniegelenks; posterior cruciate ligament: 3.45e-20*),*

---

[2] Note that without this normalization, longer target units yield higher probabilities.

(Kraeftigste Band des menschlichen Kniegelenks: strongest ligament: 1.36e-12*)*

(Kraeftigste Band des menschlichen Kniegelenks: human knee joint: 2.37e-07*)*

At the end, the best-match criterion is applied to this resulting list: the list is first sorted by decreasing score; then a term occurring in a pair (*s*,*t*) is kept if and only if neither *s* nor *t* occurs in a previous pair (with a higher score). After the best match criterion the final list is:

(*hintere Kreuzband*; *posterior cruciate ligament:* 0.0009),

(Kraeftigste Band des menschlichen Kniegelenks; strongest ligament of the human knee joint: 2.87e-05)

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 life*). Since our method allows an n-m aligment, and since acronyms are simply processed like standard words, correspondences between acronyms and translated spelled out forms can also be found (*Nierenzellcarcinom*/*RCC*). With the MuchMore corpus, 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*).

4.3. Lexicon evaluation

From the parallel corpus (comprising 5,500 abstracts), we manually extracted 150 candidate terms with their translation for evaluating our complete procedure. Table 1 shows precision and recall for terms. 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 while the automatic extraction can propose incorrect units due to chunking errors.

These results show that in almost 90% of the cases, a correct translation can be found in the first 10 propositions made by our tools. This means that a terminologist can very easily review the lists proposed by the system, and select the translations judged appropriate for enriching existing resources. We are now going to see how our tools can be used for such an enrichment, when the resources under consideration are thesauri.

Number of target candidates	Precision	Recall
1	56.52	50.98
2	71.01	64.05
5	84.78	76.47
10	89.85	81.04

Table 1: Evaluation of term alignment according to the number of target candidates considered.

4.4. Thesaurus enrichment

We propose in this section some solutions for enriching monolingual as well as bilingual thesauri. Our goal is to propose tools to the terminologist, in order to enrich semi-automatically existing thesauri. We focus on a thematic subpart of the corpus, comprising abstracts from the journal DerChirurg (700 abstracts). Since the English language is predominant in UMLS, enrichment in other languages via term alignment is possible. Indeed, the German thesaurus we use, DMD, is a partial translation of MeSH.

The first extension we address is the introduction of new strings (following UMLS terminology) associated with a concept in the thesaurus. If one element of the bilingual extracted lexicon is in the thesaurus, the translated candidate can be directly proposed as a possible addition to the part of the thesaurus corresponding to its language. Such new strings usually correspond to synonyms as well as spelling or term variants. For example, the German string *Karzinom* is associated with the UMLS concept C0007097. The corresponding English string in UMLS is *carcinoma*. Through the term alignment process, we can propose a new German string: *Carcinom*. Note that this spelling difference is in fact due to two different German spellings being used in medical texts. New strings corresponding to morpho-syntactic variations can also be detected. Thus, our alignment provides a new string for the entry *Lebertransplantation*: *Transplantion der Leber*, which was not part of the original entry which contains: *Lebertransplantation*, *Hepar-Transplantation*, *Transplantation*, *Hepar-*, *Leber-*. A second kind of enrichment is the addition of new concepts in one language. In some cases no German string is proposed for a given concept class. For example, the German thesaurus has no strings associated with C0334281 (*malignant insulinoma*) or with C0406864 (*flap loss*). Our alignment tool allows us to propose the following candidates:

malignen Insulinom for C0334281, and *Lappenverlust* for C0406864, propositions that a terminologist can review before deciding whether to enter them in the thesaurus. From the 700 medical abstracts that our corpus contains, about 1400 new German terms can be proposed in such a way to the terminologist.

The most difficult situation is the addition of new concepts and associated strings in the two languages. In some cases, it can be done through the combination of hierarchical information with particular morpho-syntactic patterns. For example, words with the suffix *-ectomy* tend to occur under the concept C0543467 (*Surgical Procedures*) and words with the suffix *-graphy* tend to occur under *Diagnosis*. Through this information, we can propose new concepts to be added to the thesaurus, as well as targeting the subpart of the thesaurus closest to this new concept. However, we do not have this information in most cases, and need to rely on a more general strategy: for a given pair of aligned terms, propose to the terminologist in charge of updating the thesaurus the set of concept classes which are closest to the pair of terms under consideration. The terminologist will then decide whether or not to create new concept classes in the thesaurus. This strategy is most adequate when the concept described by the pair of terms is close to or a refinement of existing concepts. For instance, no concept exists in UMLS for the German string *chronische Pankreatitis* or its English translation, *chronic pancreatitis*. Computing a similarity between concepts and these terms in the way described in section 5.2. below yields C0030305 (*pancreatitis*) as the closest concept class to both candidate terms. In general, if the term is composed with some words present in the thesaurus, the list of concepts proposed to the terminologist is relevant.

4.5. Application to CLIR

In this section, we consider the following CLIR experiment: the task consists in retrieving documents (actually abstracts) in a English corpus dedicated to Medical Science (the MuchMore corpus, as described in section 2), starting from German queries. For this cross-lingual task, resources such as the ELRA bilingual dictionary and the Mesh/DMD bilingual thesaurus are available.

When dealing with German queries on this corpus, the most obvious approach is to use the bilingual resources to translate the queries. However, it is well known that using these resources as such can raise coverage problems (entries are missing; translations are not domain-specific, etc.). It is therefore necessary to extend these resources by extracting specialized bilingual lexicons for the corpus. This can be done by alignment from parallel or comparable corpora. In this section, we investigate the benefits of using the bilingual lexicon extracted from the Parallel corpus as “translator” in the CLIR task, while the added value of using the Comparable corpus is examined in Section 5.5.

To be more precise, the whole Corpus[3] (9,300 documents) is used in this experiment, but no information about document structure (title, body, …) is available. A set of 25 German short but realistic queries was provided by medical experts, as well as corresponding relevance assessments on our collection. As an example, here are the first 5 queries we used:

1. Arthroskopische Behandlung bei Kreuzbandverletzungen
2. Komplikationen bei arthroskopischen Eingriffe
3. Pathophysiologie und Prävention der Arthrofibrose
4. Ätiologie und Therapie der Hämarthrose
5. Arthroskopische Behandlung der Meniskusläsionen

Queries and assessments can be found under: http://muchmore.dfki.de/resources2.htm.

As this corpus is bilingual and parallel, and as we used this parallelism (at least partially) in the lexicon extraction, we will compare our approach with the baseline method, namely the monolingual IR task (retrieving documents in the German part of the corpus which best match the German queries).

The coverage ratio (proportion of normalized German terms in the queries that are present in the German corpus or as entry of the extracted bilingual lexicon) is about 92%, for both approaches. This can be explained mainly by spelling errors in the query and, more marginally, by a normalization/decomposition stage which is not 100% reliable. The candidate translations considered are words occurring in the corpus. Two extra factors were analyzed in these experiments: the weighting schemes applied to the documents/queries and the number of retained translation candidates when using the bilingual lexicon (the *n* translations which correspond to the *n* highest translation probabilities are chosen). The following weighting schemes were examined: *ltc*, *ttc* and *tnc* for the documents and *ttx* and *tnx* for queries (SMART notation[4]). Other schemes were tried, but with poorer performance. Table 2 gives the experimental results of the CLIR task.

---

[3] Note that the bilingual lexicon extraction was based only on a subset of this corpus (5500 abstracts). This development subset was chosen in order to ensure the alignment quality.

[4] The SMART notation describes the weighting schemes in two triples, XYZ.UVW. The first triple (XYZ) refers to the document, the second (UVW) to the query. The first letter --X or U-- is a function of term frequency (for instance, *t* is the pure term frequency, *l* is a logarithmic transformation of the term frequency); the second letter --Y or V-- is a function of collection frequency (*n* means 1, *t* designated the "Inverse Document Frequency" value), while the third one is a normalization factor (*x* means 1, *c* is the cosine normalization). The final weight of an indexing term is then given by X*Y/Z for documents and U*V/W for queries.

Method	Weighting Scheme	Number of retained translations	Macro-averaged Precision	Macro-averaged Break-Even Point
Monolingual _German	*ltc-tnx*	--	0.424	0.414
Monolingual _German	*ltc-ttx*	--	0.475	0.443
Monolingual _German	*ttc-ttx*	--	0.449	0.428
Monolingual _German	*tnc-ttx*	--	0.386	0.380
Parallel Bilingual Lex.	*ltc-ttx*	100	0.542	0.510
Parallel Bilingual Lex.	*ltc-ttx*	70	0.543	0.511
Parallel Bilingual Lex.	*ltc-ttx*	40	0.543	0.510
Parallel Bilingual Lex.	*ltc-ttx*	30	0.543	0.509
Parallel Bilingual Lex.	*ltc-ttx*	20	0.543	0.510
Parallel Bilingual Lex.	*ltc-ttx*	10	0.540	0.509
Parallel Bilingual Lex.	*ltc-ttx*	5	0.537	0.507

Table 2 : Comparison of performance (average precision and break-even point) between different approaches using the parallelism of the corpus.

The main observation is that the bilingual lexicon approach -translating and enriching the queries through the lexicon extracted from the parallel corpus - significantly enhances the performance measures of the retrieval task (average precision increased from 47% to 54%). This clearly results from the efficient enrichment provided by the bilingual lexicon. It turns out that "*ltc-ttx*" is the weighting scheme that provides the best retrieval performance, and this is the case for all approaches we adopted in this study (Figure 1 gives the evolution of the average precision for the monolingual approach, but quite similar trends were observed with the bilingual lexicon methods). Finally, as illustrated by Figure 2, the number of retained translation candidates seems to have little influence, provided this number is high enough (above 10); indeed, these candidates do not bring significant extra noise.

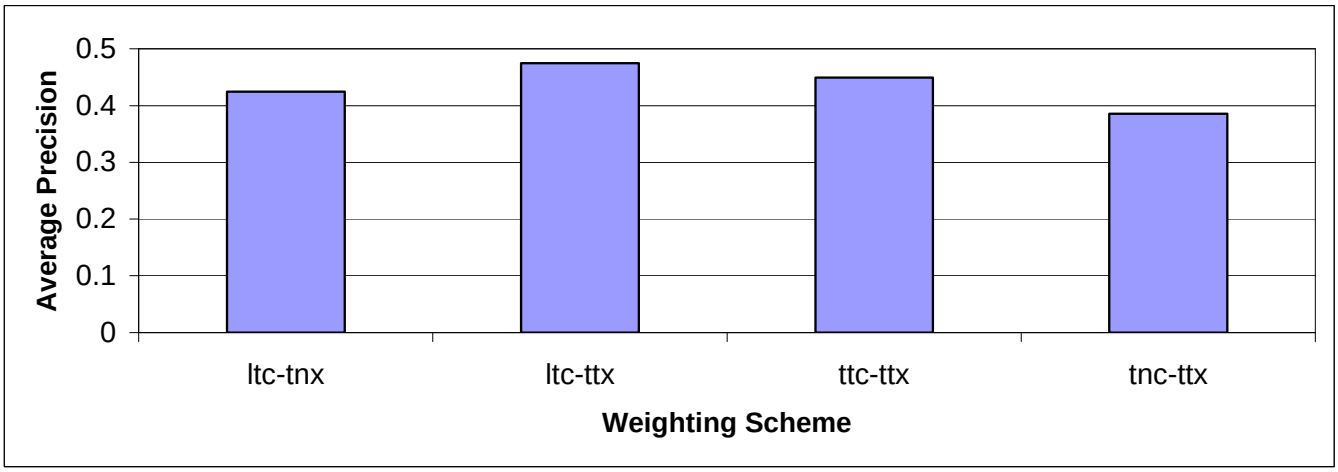

Figure 1. Comparison of different weighting schemes on the Monolingual (German) approach performance in CLIR.

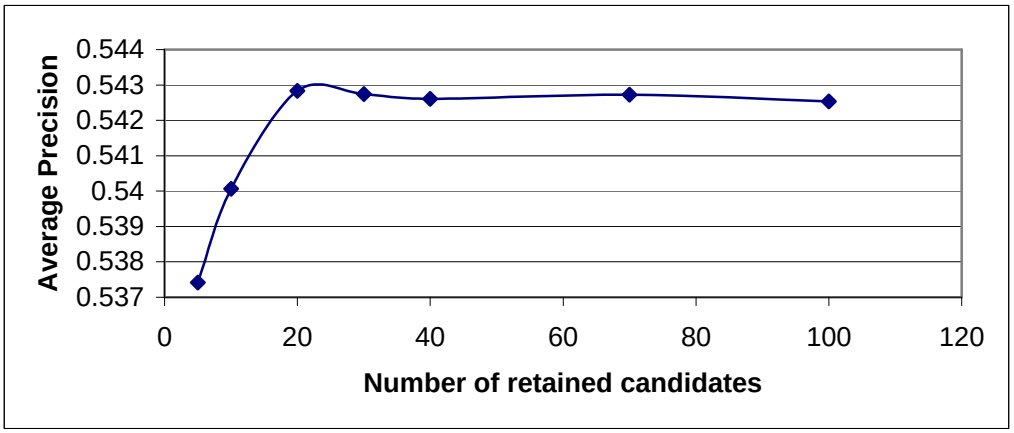

Figure 2 : Influence of the number of retained translation candidates.

5. Bilingual Lexicon Extraction from Comparable Corpora

In many cases, however, the bilingual collection under consideration does not constitute a parallel corpus, but rather a comparable corpus. In this latter case, we cannot make use of the sentence alignment information underlying parallel corpora. Searching for the translation of a given source word amounts to searching the entire target corpus, whereas in the case of parallel corpora the search is limited to the sentence pairs containing the source word. Because of this increase in the dimension of the search space, one cannot expect to reach results equivalent to the ones obtained with parallel corpora. This problem is even worse for terms, since good clues for term alignments are provided by word alignments. This has led researchers to focus on the extraction of bilingual lexicons of single words in the case of comparable corpora.

We will present in this section the standard method developed for extracting bilingual lexicons from comparable corpora. We will then present a new, complementary method that uses a bilingual thesaurus as a pivot between languages, and which is partially described in [6]. We will then show how to combine these two methods, and will present results obtained for bilingual lexicon extraction from comparable corpora, and for their use in cross-language retrieval.

5.1. Standard approach

Bilingual lexicon extraction from non-parallel but comparable corpora has been studied by a number of researchers, [7], [9], [19], [21], [23], [24] among others. Their work relies on the assumption that if two words are mutual 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.

The use of a general bilingual dictionary to translate context vectors is justified by the fact that if the context vectors are sufficiently large, then some of their elements are likely to belong to the general language and to the bilingual dictionary, and we can thus expect the translated context vector of word *t* to be, in average, closer to the context vector of the translation *s* of *t*. It has to be noted that the above strategy makes sense even when *t* is present in the bilingual dictionary, since the corpus may display a particular, technical usage of *t*.

Our implementation of this strategy relies on the following standard algorithm:

1. 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, [8], to measure this association,
2. 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; if different words lead to the same translation, their weights is added),
3. The similarity of each source word *s*, to each target word *t*, is computed on the basis of the cosine measure,
4. The similarities are then normalized to yield a probabilistic translation lexicon, $P_1(t|s)$.

To illustrate the above steps, we give below the first 5 words of the context vector of the German word *Leber* (*liver*), together with their associated score:

transplantation 138.897; resektion 53.501; metastase 41.668; arterie 38.519; cirrhose 26.302

which, when translated, become:

transplant 138.897; tumour 48.654; secondary 42.552; metastasis 41.668; artery 38.519

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.**

5.2. Lexical translation models based on a multilingual thesaurus

A multilingual thesaurus bridges several languages through cross-language correspondences between concept classes[5]. 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. Based on the fact that the thesaurus we use (MeSH-DMD) relies on a one-to-one correspondence between concept classes, we can write the probability $P(t|s)$ of selecting word t as a translation of word s as follows:

$$P_2(t|s) = \Sigma_C P(C,t|s) = \Sigma_C P(C|s)\, P(t|C,s) = \Sigma_C P(C|s)\, P(t|C,s) \approx \Sigma_C P(C|s)\, P(t|C) \quad (2)$$

where C represents a multilingual concept class in MeSH-DMD. The last approximation is based on the fact that we do not want to privilege any of the possible lexicalizations of a given concept class, which could be done from the probability distribution $P(t|C,s)$.

The above formula suggests the use of all the classes of the thesaurus to compute the association between source and target words. However, if the relation between a word and a concept class is not significant this has the disadvantage of bringing

[5] 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*. In many cases, it corresponds to a synonym class.

in noisy data in the estimation of $P_2(t|s)$. We thus propose to rely on the following algorithm, which slightly generalizes equation (2):

1. For each source word *s*, for each target word *t*, compute $P(C|s)$, $P(t|C)$,
2. For each pair (*s*,*t*)
    a. Select the concept classes to be used,
    b. Compute $P_2(t|s)$ based on the selected classes and the quantities computed in step 1.

The computation of $P(C|s)$ and $P(t|C)$ is 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. 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 5.1. 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 sum 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. Once context vectors have been computed for concept classes, we rely again on the cosine measure to compute similarities between words and concept classes. The probability $P(C|s)$ and $P(t|C)$ are finally derived through normalization.

There are several ways to select concept classes in step 2. a. of the above algorithm. The most obvious one consists in retaining all the concept classes as indicated by equation 2. A direct extension is to retain only the best n concept classes, i.e. the set *E* of *n* concept classes such that $\Sigma_{C \in E} P(C|s) P(t|C)$ is maximal. We refer to this method as the **complete-*n* search**. Nevertheless, the complete *n* search does not make use of the hierarchical information contained in the thesaurus. We thus devised another method which makes use of the structure underlying a thesaurus and selects concept classes from the thesaurus in the following way: a) form the set *F* of the *p* best concept classes associated with source word *s*; b) for each pair of concept classes $(C1,C2)$ from *F*, add to *F* all the classes, not already in *F*, that appear in the path between *C1*, *C2* and their common ancestor. This procedure corresponds to the following intuition: if two or more concept classes associated with a source word *s* have the same parent in the thesaurus, then this parent is likely to be related to *s* as well. 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 refer to this method as the **subtree search**.

The subtree search provides a set of subtrees from the 15 sub-thesauri corresponding to the 15 main categories of the MeSH classification. Table 3 shows the different behaviors of the complete method, using the 200 concept classes closest to the source word, and the subtree method with $p = 20$ on the source word *Leber*. Even if some candidates are not actual translations of *Leber*, the subtree search provides *liver* (the correct translation), as the first translation candidate. Note that the subtree search, with $n = 20$, yields in average 4 subtrees per source word.

Subtree	Complete-200
Liver	Hepatocyte
Orthotopic	Neoplasm
Hepatic	Enormous
Survival	Orthotopic
Metastasis	Inherit

Table 3: First 5 candidates for *Leber*

5.3. Combining Different Models

The previous section provides us with two different probabilistic lexical translation models: one derived with the standard method, and one based on alignment through a bilingual thesaurus. A third model, which we will denote P_3, can be directly derived from our general bilingual dictionary by considering the different translations of a given entry as equiprobable. For example, our dictionary associates *abbilden* with the words *depict* and *portray*. Thus, $P_3(depict|abbilden) = P_3(depict|abbilden) = 0.5$.

The final estimate of the probability of translating source word *s* by target word *t* is then given by the following mixture of models:

$$P(t|s) = \Sigma_i P(i) P_i(t|s)$$

where $P(i)$ denotes the probability of selecting model *i* and corresponds to the mixture weights. A standard way to estimate mixture weights is to rely on the EM algorithm by maximizing the likelihood of some held-out data. In order to do so, we

manually created, from our corpus, a "held-out reference lexicon", denoted *l*, containing ca. 1,200 pairs. The E- and M-step formulas of the EM algorithm are then defined as follows:

$$< Iist >= \frac{P(i)P_i(t \mid s)}{\sum_k P(k)P_k(t \mid s)}$$

$$P(i) = \frac{\sum_{s,t} < Iist >}{\sum_k \sum_{s,t} < Ikst >}$$

where <*Iist*> denotes the probability of choosing model *i* from the pair *(s,t)*. Table 4 below presents the mixture weights we obtained when model 2 is based on the complete n search (with n = 200 and n = 1) and the subtree search (with p = 20).

	Complete-1	Complete-200	Subtree 20
Model 1	0.59	0.45	0.33
Model 2	0.1	0.24	0.37
Model 3	0.31	0.31	0.29

Table 4: Mixture weights for the 3 models

As one can note, these results suggest that the thesaurus, heavily used in model 2, is less reliable then the other resources when used with the complete-1 search, whereas it becomes the most important resource with the subtree search. This is not surprising if one relates these results with the way the thesaurus is exploited in each case: precise (the best class) but incomplete (only one class) information is used in the complete-1 search, more complete (200 classes) but less precise information in the complete 200 search, and more complete and more precise information in the subtree search, since only accurate and relevant concept classes are selected through the thesaurus hierarchy.

5.4. Lexicon evaluation

As already mentioned, we manually extracted a reference lexicon comprising 1,800 translation pairs from our 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.

Table 5 shows the results we obtained, retaining only the first 10 candidates for each source word, without combining the different models.

	Model 1	Model 2 (subtree-50)	Model 3
F1-score	62.04	51.34	56.16

Table 5: Results for separate models.

The precision obtained with the general bilingual dictionary only (model 3) reaches 78%, which is respectable considering the fact that we focused on a highly specialized domain. However, as one could expect, the recall reaches only 48%.

Table 6 shows the results obtained by combining the different models, using different search strategies for model 2, and considering the first 5 and 10 candidates for each source word (the first number corresponds to the F1-score, whereas the second one is the average rank Comp. Stands for complete and Subt. for subtree).

	Comp.-1	Comp.-100	Comp.-200	Subt. 10	Subt. 20	Subt. 50	Subt. 100
p=5	71.3/14.7	85.4/14.1	75.4/12.3	75.8/11	76.4/11.7	77.3/11.2	76.9/11.8
p=10	79.7/14.7	80.3/14.1	83.2/12.3	82.4/11	84.1/11.7	83.6/11.2	83/11.8

Table 6: Evaluation of combined models with different search strategies

As one can see, the combination significantly improves the results over the individual models alone, since the F1-score goes from 62% to 84%, a score that we believe is good enough to forward for manual revision. The best results obtained with the subtree and the complete searches are comparable, with a slight advantage to the subtree search. However, the optimal subtree search uses 7.5 times fewer classes than the optimal complete search, and is 2 times faster. This proves that the subtree search is able to focus on accurate concept classes while the complete search need to consider more classes to reach a similar level of performance. Lastly, since the average rank is computed on the entire list, and not on the first *p* elements, it is possible that the evolution of this rank does not parallel the one of the F1-score, as one can note for the subtree search.

5.5. Application to CLIR

In this section, we consider the same CLIR task as in section 4.5., now using the bilingual lexicon extracted from the comparable corpus as the translation resource. As we did not use the link (more precisely the parallelism) between the German and the English corpus, the baseline will now be the translation of German queries by the usual resources (the ELRA bilingual dictionary and the Mesh/DMD specialized bilingual thesaurus), followed by a monolingual (English) standard Information Retrieval task.

The coverage ratio (proportion of normalized German terms in the queries that are present as entry in the translation resource) is less than 50% for the baseline approach, and about 92% for the bilingual lexicon approach based on the comparable corpus. Moreover, 10% of the translation candidates given by the usual resources are not present in the English corpus. This clearly emphasizes some reasons for the weakness of the baseline method. We applied several weighting schemes, but once again the '*ltc-ttx*" scheme appeared to be the best one, for both approaches. Table 7 gives the experimental results of the CLIR task.

Method	**Weighting Scheme**	**Number of retained translations**	**Macro-averaged Precision**	**Macro-averaged Break-Even Point**
Monolingual _English	*ltc-ttx*	--	0.223	0.258
Comparable Bilingual Lex.	*ltc-ttx*	100	0.346	0.361
Comparable Bilingual Lex.	*ltc-ttx*	40	0.348	0.359
Comparable Bilingual Lex.	*ltc-ttx*	20	0.360	0.367
Comparable Bilingual Lex.	*ltc-ttx*	10	0.387	0.385
Comparable Bilingual Lex.	*ltc-ttx*	8	0.397	0.386
Comparable Bilingual Lex.	*ltc-ttx*	6	0.397	0.389
Comparable Bilingual Lex.	*ltc-ttx*	4	0.379	0.338

Table 7: Comparison of performance (average precision and break-even point) between different approaches not using the parallelism of the corpus.

The main observation is that the bilingual lexicon approach --translating and enriching the queries through the lexicon extracted from the comparable corpus-- significantly enhances the performance measures of the retrieval task with respect to the baseline (average precision increased from 22% to 40%). However, it can be argued that the bilingual lexicon approach was favorably biased by using the parallel corpus, even if it did not directly use the parallelism. Further experiments will be conducted in the near future to see if the performance trend is similar with a less favorable comparable corpus. Still the fact remains that the level of precision achieved is lower than the level realized when the parallelism is directly exploited. Finally, as illustrated by Figure 3, the number of retained translation candidates has now a strong impact on performance: there is an optimal number of translation candidates (between 8 and 10); below that, the enrichment is not sufficient to give an accurate translation, while higher numbers introduce useless noise.

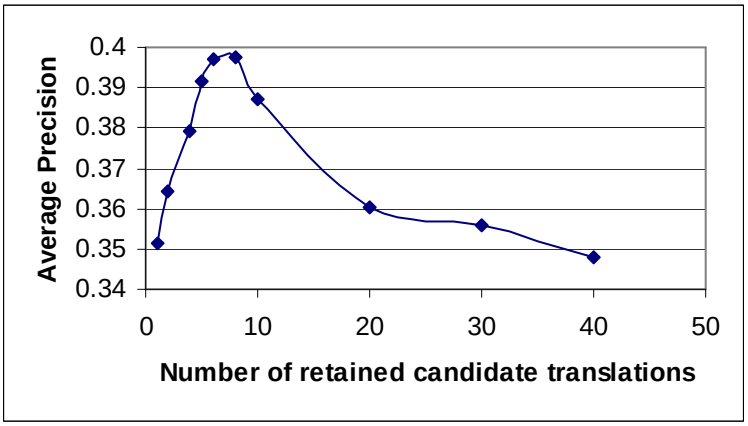

Figure 3 : Influence of the number of retained translation candidates (comparable corpus approach).

6. Conclusion and Future Work

We have shown in this paper that bilingual lexicon extraction from parallel corpora in the medical domain could lead to accurate, specialized lexicons, which can be used to help enrich existing thesauri. To this end, we have proposed several methods, relying on the tools we described. We have also shown that such an enrichment can be substantial, even when one starts with a “small” collection, since we were able to propose about 1,400 new German terms from 700 medical abstracts for direct enrichment, i.e. enrichment of existing concept classes. We have also shown how to use the extracted lexicons for cross-language information retrieval in the medical domain. Not surprisingly, using lexicons extracted from, and thus

adapted to, the collection under consideration significantly improves the retrieval performance. Because more and more scientific articles contain bilingual abstracts, as is the case for the medical articles on Springer Verlag, bilingual lexicons can be extracted, articles can be indexed, and accurate information can be made available to medical doctors and researchers, in the way described in the first part of this paper.

However, many bilingual collections are not parallel but only comparable. In such cases, the methods developed for parallel corpora cannot be used directly, and one needs to resort to different approaches. We propose in this paper a new approach for bilingual lexicon extraction from comparable corpora, which uses a bilingual thesaurus as a pivot. Because of the availability of bilingual thesauri in the medical domain, we believe this method to be well adapted to this domain. Since our approach is complementary to previous ones, we propose a way to combine the different approaches, and show that such a combination significantly (by more than 20%) improves the accuracy and coverage of the lexicons obtained. Furthermore, we have shown that using lexicons extracted in such a way outperforms once again general bilingual resources for cross-language information retrieval. As shown in [22], one can gain from the combination of the two, since in many cases only part of the collection to be searched is fully parallel. We plan in the future to have a closer look at this problem. Lastly, we have restricted ourselves to bilingual lexicons of single words, when dealing with comparable corpora. The natural follow-up of our work will be to extend our approach to bilingual lexicons of terms.

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

[1] Bishop, S.Fiendbeg and P.Holland, “Discrete Multivariate Analysis”, In MIT Press, 1975.

[2] I.Blank, “Terminology Extraction from Parallel technical Texts”, In J. Véronis, Ed. Parallel Text Processing – Alignment and Use of Translation Corpora, Kluwer Academic Publishers, 2000.

[3] P. Brown, S. Della Pietra, V. Della Pietra and R. L. Mercer, “The Mathematics of Statistical Machine Learning Translation: Parameter Estimation”, Proc. Computational Linguistics, vol. 19, no. 2, pp. 263-311, 1993.

[4] P. Buitelaar and B. Sacaleanu, “Extending Synsets with Medical Terms”, Proc. First International WordNet Conference, Mysore, India, Jan. 21-25, 2002.

[5] I. Dagan and I. ITAI, “Word Sense Disambiguation using a Second Language Monolingual Corpus”, Proc. Computational Linguistics, vol. 20, no. 4, pp. 563-596, 1994.

[6] H. Déjean, É.Gaussier and F. Sadat, “An Approach based on Multilingual Thesauri and Model Combination for Bilingual Lexicon Extraction”, Proc. 19th International Conference on Computational Linguistics COLING, Taipei, Taiwan, pp.218-224, Aug. 2002.

[7] M. Diab and S. Finch, “A Statistical Word-Level Translation Model for Comparable Corpora”, Proceedings of the Conference on Content-based Multimedia Information Access, RIAO, 2000.

[8] T. Dunning, “Accurate Methods for the Statistics of Surprise and Coincidence”, Proc. Computational linguistics, vol. 19, no. 1, pp. 61-74, 1993.

[9] P. Fung, “ A Statistical View of Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora”, In Jean Véronis , Ed. Parallel Text Processing, 2000.

[10] W.A. Gale and K. W. Church, “A Program for Aligning Sentences in Bilingual Corpora”, Proc. Annual Meeting of the Association of Computational Linguistics, pp. 177-184, 1991.

[11] E. Gaussier, “Flow Network Models for Word alignment and Terminology Extraction from Bilingual Corpora”, In Proc. Of the joint 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics, 1998.

[12] E. Gaussier, D. Hull and S. Ait-Mokhtar, “Term Alignment in Use: Machine-Aided Human Translation”, In J. Véronis, Editor. Parallel Text Processing – Alignment and Use of Translation Corpora, Kluwer Academic Publishers, 2000.

[13] T. Hedlund, "Compounds in Dictionary-based Cross-Language Information Retrieval" In. Information Research, vol. 7, no. 2, 2002.

[14] U. Heid, “A Linguistic Bootstrapping Approach to the Extraction of Term Candidates from German Text”, Proc. Terminology, vol. 5, no.2, 1999.

[15] D. Hiemstra, “Using Statistical Methods to Create a Bilingual Dictionary”, In Master Thesis, Universiteit Twente, 1996.

[16] D. Hull, “Automating the construction of bilingual terminology lexicons”, In Terminology, 4(2), 1999.

[17] M.Kay and M. Röscheisen, “Text Translation Alignment”, Proc. Computational linguistics, vol. 19, no. 1, pp. 121-142, 1993.

[18] I. D. Melamed, “A Word-to-word Model of Translation Equivalence”, Proc. 35th Annual Meeting of the Association for Computational Linguistics, pp. 490-497, 1997.

[19] C. Peters and E. Picchi, “Capturing the Comparable: A System for Querying Comparable Text Corpora”, Proc. The 3rd International Conference on Statistical Analysis of Textual Data, pp. 255-262, 1995.

[20] A. Pirkola, H. Keskustalo and K. Järvelin, “The effects of Conjunction, Facet Structure, and Dictionary Combinations in Concept-based Cross-Language Retrieval”. In. Information Retrieval, vol. 1, no. 3, pp. 217-250, 1999.

[21] R. Rapp, “Automatic Identification of Word Translations from Unrelated English and German Corpora”, Proceedings of the European Association for Computational Linguistics Conference, EACL, 1999.

[22] J.-M. Renders, H. Déjean, E. Gaussier, “Assessing Automatically Extracted Bilingual Lexicons for CLIR in Vertical Domains: XRCE participation the GIRT track of CLEF 2002”, to appear in Lecture Notes in Computer Science, C. Peters, M. Braschler, J. Gonzalo and M. Kluck Editors, Springer Verlag, 2003.

[23] I. Shahzad, K. Ohtake, S. Masuyama and K. Yamamoto, “Identifying Translations of Compound Using Non-aligned Corpora”, Proc. of the Workshop MAL, pp. 108-113, 1999.

[24] K. Tanaka and H. Iwasaki, “Extraction of Lexical Translations from Non-Aligned Corpora”, Proceedings of The 13th International Conference on Computational Linguistics, COLING, 1996.

[25] C.J. Van Rijsbergen Information Retrieval, Buttersworth, 1975.

[26] M. Volk, B. Ripplinger, S. Vintar, P. Buitelaar, D. Raileanu and B. Sacaleanu, "Semantic Annotation for Concept-based Cross-Language Medical Information Retrieval", International Journal of medical Informatics, 67 (1-3), 2002.

[27] D. Widdows, B. Dorrow, and C. K. Chan, "Using Parallel Corpora to Enrich Multilingual Lexical Resources", Proc. Third International Conference on Language Resources and Evaluation LREC, Spain, pp. 240-244, May, 2002.