Processing Multilingual Collections for Text Mining Applications

Eric Gaussier

Xerox Research Centre Europe
6, Chemin de Maupertuis – 38240 Meylan France
Eric.Gaussier@xrce.xerox.com

Abstract: We address in this presentation the problem of processing multilingual collections, for such text mining applications as cross-language clustering, categorisation and information retrieval. We review different models proposed for this task, while focusing on the most important problems that need to be solved.

1 Introduction

Large international companies, translation centres, international organisations (e.g. the European Commission or international patent offices) make use of text collections written in more than one language. Despite this linguistic diversity, the operations one wants to perform on documents (searching, filtering, categorising, taxonomy induction) have to be consistent across languages. For example, when categorising documents within the IPC (International Patent Classification), one wants that related patents, written in different languages, end up in the same categories. To address this problem, several methods have been developed, ranging from dictionary-based methods, in which multilingual lexicons are automatically extracted from given collections, to latent semantics representation, in which documents are mapped to a language-independent space (as is done e.g. in Latent Semantic Indexing or Canonical Correlation Analysis, and their kernel versions). Underlying all these methods is the type of indexing required for a given application, indexing which needs to be consistent throughout languages.

In our presentation, we will first review methods developed so far to deal with multilingual collections. We will then present some of the most important problems which remain to be solved.

2 General Characteristics

Multilingual collections can be either parallel, i.e. documents are translations of each other, or comparable, i.e. documents cover the same topics in the same domains.

Translation memories constitute a well-known example of parallel collections, whereas newspapers articles, written in different languages, at the same moment are prototypical examples of comparable corpora.

For most text mining applications, the units to be considered range from words, to terms and entities, the latter covering standard named entities (proper names, place names, …) as well as domain specific entities (e.g. gene or protein names in biology). Even though those units are often considered independently of each other, semantic relations do exist between them, as synonymy, hyponymy and hyperonymy, and in many cases it is important to split a given unit according to the different meanings it has (polysemy).

For most text mining applications one might consider, it is important, in order not to privilege any language, to ensure both consistency and performance equivalence across languages. Consistency can be seen as a qualitative constraint, and corresponds to the example given in introduction: if two documents are translations of each other, then one wants the processings done on each document to be equivalent (e.g. they are categorised in the same language-independent classes). Performance equivalence is the quantitative counterpart of consistency, and requires that the performance obtained on each language to be equivalent. As we are going to see, consistency, hence performance equivalence, are difficult to ensure, and thus correspond to goals rather than constraints that one can impose.

3 Feature Extraction in Multilingual Collections

We want to show here that it is, in most cases, not possible to extract the same features, in terms of types and units, in different languages. To this end, we will consider the particular case of terminology, since terms are often used as indexes in text mining applications[1].

3.1 Terminology Extraction

There is no fully operational definition of terms. Similarly to compounds, a term represents a lexicalised entity, at least when it is used within its technical context. Classical definitions for compounds integrate syntactic, semantic and referential criteria. If a term univoquely refers to a concept (or class of concepts), in general, the sense of a multi-word term can be derived, through composition, from the sense of the words it contains. The first syntactic criteria used to determine whether an entity was a compound or not reflected the fact that the structure of a compound was believed to be frozen. Most recent studies, e.g. (Gross, 1988), have shown that the syntactic structure of compounds are nonetheless subject to variations, introducing the idea of a continuum between frozen and non-frozen entities. As we will see, most

[1] The following presentation is partly derived from *Gaussier E. General Considerations on Bilingual Lexicon Extraction. D. Bourigault, C. Jacquemin, M.-C. L'Homme Editors. Computational Terminology, 2000.*

terms undergo such variations, with insertions of adjectives and/or adverbs. Computational studies on compounds and terms have then tried to characterize elementary structures of compounds and terms, as well as the modifications such structures can support (see (Mathieu-Colas, 1988; Jacquemin, 2001) on French, and (Nkwenthi-Azeh, 1992; Justeson and Katz, 1995) on English). Such structures for characterizing French and English terms are the basis of the vast majority of monolingual term extractors (sometimes coupled with statistical information), and we first give an overview of these structures and their modifications, for terms of length 2, i.e. composed of two lexical (as opposed to grammatical) words. The restriction to terms of length 2 is justified by the central role they play in terminology: (a) they are by far the most frequent type of terms, (b) terms of length 3 and more are usually derived from terms of length 2 by various operations that we describe hereafter. Furthermore, all our examples are taken from a corpus on telecommunication satellites, provided by the EEC within the framework of the European Project ET-10/63. Lastly, the structures we are going to present are derived from the examination of terminology lists and dictionaries, as well as from corpus studies.

Syntactic Patterns for French Terms

All the syntactic patterns we are considering correspond to noun phrases (not all noun phrases are covered by these patterns). A short form, summarizing all the relevant patterns, is given below, where E corresponds to the empty string, N(1,2) to a noun, and Adj to an adjective:

N1 PREP DET E2

with:

PREP={de} DET={E,le,la,l',les}, E2=N2 (1)
PREP={à} DET={E,le,la,l'}, E2=N2 (2)
PREP={en,sur} DET={E}, E2=N2 (3)
PREP={dans,par} DET={le,la,l',les}, E2=N2 (4)
PREP={E} DET={E}, E2=N2 (5)
PREP={E} DET={E}, E2=Adj (6)

These patterns make use of definite articles, but not of indefinite ones. Even though some terms may integrate indefinite articles, as *mouvement d'une orbite*, most sequences containing indefinite articles are not terms. Moreover, the terms with indefinite articles often have a correspondent with definite articles. We face here the problem of recall versus precision. If recall is privileged, then one should consider both definite and indefinite articles. If, on the contrary, the emphasis is put on precision, then one should focus on candidates with definite articles. But, since the vast majority of candidates with indefinite articles are not valid terms, relying only on definite articles represents a good trade-off between recall and precision. The same remarks apply for the selection of the articles, among the set of definite articles, with respect to the preposition used.

The pattern **Adj N**, where the adjective appears before the noun, is not retained, since most candidates of this form are not terms. The criterion here is anyway indirect, since it is the type of adjective which mainly determines the terminological status of the unit. In French, only certain classes of adjectives can appear before the noun they modify. These adjectives are, in general, not used to form terms. However, most of these adjectives can also appear after the noun. We thus see that we could refine, if this information is present in our lexicons, the pattern **N Adj**, by **N Adjna**, where *Adjna* represents an adjective that cannot appear before the noun.

Here are examples of terms for each of the preceding patterns (we provide their English translation in parentheses):

(1) durée de vie (lifetime); vitesse du faisceau (beam velocity)
(2) trafic a l'émission (transmit traffic)
(3) répartition en fréquence (frequency division)
(4) répartition dans le temps (time division)
(5) diode tunnel (tunnel diode)
(6) lobe latéral (side lobe)

Syntactic Patterns for English Terms

The syntactic patterns for English are less numerous, since English relies on a composition of Germanic type, without prepositions, to produce compounds, and of Romance type, with prepositions, to produce free noun phrases, as in *examples of calculations*, whereas French relies on Romance type for both, as described in (Chuquet and Paillard, 1989). Only two patterns are retained: **N N** and **Adj N**. Examples of Adj N compounds are *hot stand-by* (*secours permanent*) and *orthogonal polarization* (*polarisation orthogonale*). Examples of N N compounds are *frequency band* (*bande de fréquence*) and *telephone channel* (*voie téléphonique*).

However, two remarks need to be done, the first one concerning the use of the Saxon genitive in certain compounds, the second one the modifiation of an N N sequence into an N of N sequence:

1. The Saxon genitive (N1's N2) can be used to specify to which category N2 belongs to, as in *a man's job* (*un metier d'homme*), which could be considered as a term. Nevertheless, most studies reject this pattern insofar as such a use of the Saxon genitive is rare. The only example of Saxon genitive we found in our corpus, occurring only once, is *earth's curvature* (*courbure de la terre*). Furthermore, all the dictionaries we looked at propose *curvature of the earth*, and not *earth's curvature*.
2. There are very few terms in English corresponding to the pattern **N PREP N**. Our corpus contains 4300 candidates of the form **N N** or **Adj N** occurring at least twice, whereas only 530 candidates of the form **N PREP N**, occurring at least twice, are encountered. Among these candidates, 360 contain the preposition *of*.

The construction **N1 of N2** is usually used when the two nouns are considered independently of one another, when it is not possible to form, through composition, a new concept. Furthermore, the sequence **N1 of N2** is also used to translate the French sequence **N1 de N2** when the first noun is a quantifier or a classifier, as *type d'antenne* (*type of antenna*). Such French candidates are not valid terms, and if the pattern **N1 of N2** is not retained, we might expect these French candidates to be eliminated during the alignment process since their translation is not taken into account.

However, there are cases where syntactic constraints may force the use of **N1 of N2** instead of **N2 N1**. This is the case when one wants to unambiguously qualify, with an adjective, the noun *N2*. A sequence **Adj N2 N1** is ambiguous with respect to which noun is qualified by the adjective, whereas the sequence **N1 of Adj N2** is not. Such a process is illustrated in the following example, where *N1* and *N2* usually appear as **N2 N1**, and where the qualification of *N2* by an adjective yields the sequence **N1 of Adj N2** (the French translations are given in parentheses):

interference levels (niveaux de brouillage)
levels of permissible interference (niveaux de brouillage admissible).

Thus, certain term variants are to be found in sequences with preposition *of*. However, such variants are of length greater than 2, and should be recovered from the underlying term of length 2, as is done in (Jacquemin, 2001). Furthermore, terms (of length 2 and more) with preposition *of* are not frequent. This explains why they are not retained in most studies. Lastly, in a bilingual environment, as argued before, not retaining English patterns with prepositions may act as an additional filter for French candidate terms.

Correspondences and Non-correspondences across Languages

In an ideal world, the English and French patterns cover exactly the same units, i.e. French terms following French patterns are translated into English terms following English patterns. (Maxwell, 1992) speaks of "regularities" in rendering certain English patterns into certain French patterns.

On a small portion of our corpus, we have looked at which structures were used in English (French) to translate French (English) candidates following the above patterns. Considering each candidate only once, we obtained the following results (we do not claim that the figures we obtained are representative of the phenomena taken into account. They mainly serve as an illustration of our discussion):

Table 1. Pattern correspondences and non-correspondences.

	N N	Adj N	N of N	N’s N	N
N de N	122	15	2	1	8
N prep N	28	9	-	-	2
N Adj	23	63	-	-	1
N N	11	-	-	-	-
N	1	1	-	-	-

The numbers indicate the correspondence between elements of English and French terms. The preceding results can also be seen in the following graphs, which show how the different patterns we have considered in one language are realized in the other one (**N1 PREP N2** represents all the French patterns formed with two nouns):

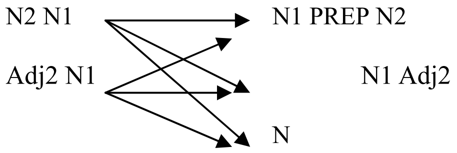

Fig. 1. Pattern alignments, from English to French

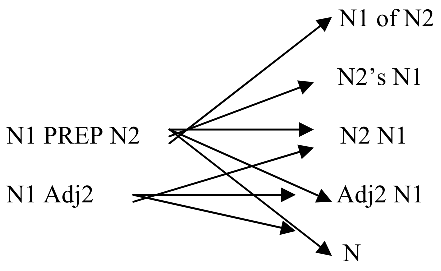

Fig. 2. Pattern alignments, from French to English

The differences between the preceding graphs show that we should extract more French candidates than English ones. It is the case since we obtain a set of 2 235 English candidates occurring at least twice, against 3 205 French candidates occurring at least twice. Furthermore, we see that relying on patterns in both English and French should act as a filter for French candidate terms. As we already mentioned, French candidate terms translated as an **N of N** sequence are not usually terms. Thus, the bilingual dimension could serve as a refinement for monolingual terminology extraction. However, if valid for French, this remark is not true for English, since most of the structures used to translate English candidate terms are retained in French.

The correspondences and non-correspondences between patterns across languages show that it is not possible to come up with patterns which cover exactly the same phenomena in both languages. We are thus bound to face non-correspondences between a set of candidate terms extracted following some patterns in one language and other patterns in another language. Different strategies can be envisaged to cope with this problem, as we see in next section.

3.2 Bilingual Term Alignment from Parallel Corpora

Several methods have been proposed to align noun phrases and/or terms within parallel corpora. These methods usually rely on the following steps: extraction of candidate terms (or NPs) in each language, and alignment of the extracted units, assuming that monolingually extracted units correpond to each other cross-lingually (see (Kupiec, 1993; Gaussier, 1995) for example). Unfortunately, this is not always the case, and the above methodology suffers from the weaknesses pointed out by (Wu, 1997) concerning parse-parse match procedures:

1. appropriate, robust, monolingual grammars may not be available,
2. grammars may be incompatible across languages,
3. selection between multiple possible arrangements may be arbitrary.

If the first point above concerns more particularly the grammatical analysis of complete sentences, the last two ones have a direct impact on bilingual terminology extraction, as we have seen in previous section. Three main solutions have been proposed to overcome these problems.

Extended Parse-parse Method

The first solution can be viewed as an extended parse-parse match procedure, and can be summarized as follows: starting with word alignments, use syntactic dependencies and contiguity constraints to derive unit alignment. As an example, consider the French and English terms *largeur de bande admissible* and *permissible bandwidth*, and let us assume that our word-to-word alignment produces a correspondence between *admissible* and *permissible*, *largeur* and *bandwidth*, and *bande* and *bandwidth*. Then, we can rely on a syntactic parser to extract the dependencies *permissible->bandwidth*, and *admissible*->(*largeur de bande*). In the latter case, a syntactic parser would most likely produce either *admissible->largeur* or *admissible->bande*. However, in both cases, we are able to derive the association *largeur de bande admissible* <-> *permissible bandwidth*, based on the correspondences across languages, the dependency relations between words and the contiguity of the sequences considered. Of course, for bilingual terminology alignment, only certain dependency relations have to be taken into account, namely the ones corresponding to the patterns given in previous section.

(Debili and Zribi, 1996) are the first ones, to our knowledge, to have propose this method. (Hull, 1998) uses a variant of this method, based on the following sequences: sort candidate term associations in descending order of a score based on the word alignments they contain (this step produces a sorted list of (Ts<->Tc) pairs, where Ts (Tc) is a candidate term in the source (target) language), take the largest association and align the associated terms, unless both are already aligned. If, for example, the source term Ts is already aligned with another term Tc1, then the target term Tc is concatenated to Tc1 if there is a dependency relation between head words of Tc and Tc1 or Tc and Tc1 are contiguous (i.e. there is no unit between Tc and Tc1 which is aligned elsewhere).

The method used by Hull differs from the method proposed by Debili, inasmuch as candidate terms are extracted in both languages, and so the first list of term associations built suffers from the problems associated with parse-parse match procedures. Furthermore, it is not possible, with this method, to recover the translation of a source term when this translation is a subpart of a target candidate term. Since there is no restriction on the length of candidate terms, such a case will happen.

Crucial for Debili's method is the word alignment algorithm used. If a type (1,1) alignment is used, that is each English (French) word is associated to one and only one French (English) word, then, in the preceding example, either *largeur* or *bande*, but not both, will be associated with *bandwidth*, and the correspondence between *largeur de bande admissible* and *permissible bandwidth* may not be recovered. Less restrictive alignment types can be used, but they may lead to less precise results, and thus endanger the whole procedure.

One Way Parsing

An alternative solution to the problems of parse-parse match methods for bilingual terminology extraction can be found in (Gaussier, 1998), where candidate terms are extracted in one language (English), and guessed, through the alignment process, in the other language (French). The method is based on flow network models for aligning units within aligned sentences. An English sentence is represented by a set of vertices corresponding to the different candidate terms and words in the English sentence. In order to take into account contiguity constraints, the French sentence is represented as a set of layers of vertices, the lowest layer being associated with the French words themselves, and each vertex in any upper layer being linked to two consecutive vertices of the layer below. The uppermost layer contains only one vertex and can be seen as representing the whole French sentence. Capacity values can also used to control the length of a French unit a given English unit can be aligned to. A minimum cost flow algorithm is then run between the English and French vertices to discover translations of English candidate terms.

Such a procedure allows one to discover translation of English terms which are subparts of French candidate terms, without suffering from parse-parse match procedure problems. Furthermore, the associations between words do not serve as a starting point from which larger associations are derived, but are rather used to define a score between English and French units. A possible extension of this approach is to replace the vertices corresponding to contiguous units with vertices corresponding to dependency relations between units. However, for terminology extraction purposes, since terms are usually made up of contiguous elements, it is not clear that there will be much difference between the two approaches, if one restricts oneself to dependency relations present in terms. Nevertheless, we can expect that with larger grammars, i.e. not restricted to term identification, we will have a better disambiguation of different relations, since the context is no more local to a term. We could thus take advantage of the dependency relations in this case.

Parallel Parsing

Lastly, a third solution can be envisaged along the lines given by (Wu, 1997). In this work, an inversion transduction bilingual grammar is built to parse in parallel aligned sentences, based on word alignments. Once the grammatical analyses have been built for each sentence, we can, for example, select the units corresponding to candidate terms in the English parse and associate them with their corresponding units in the French parse. This method is thus general enough to accommodate our needs for bilingual terminology extraction. Nevertheless, since this method aims at finding parallel parses for complete sentences, it is more subject to errors than a method restricted to term alignments, and may miss correct associations or yield incorrect ones that the methods restricted to terminology alignment will not miss or yield.

3.3 The Case of Comparable Corpora

In the case of parallel corpora, sentence alignments impose strict constraints on the set of possible translations for a given source word. For comparable copora, however, no such restrictions exist, and searching for the translations of a given source word amounts to searching the entire target corpus.

Bilingual lexicon extraction from comparable corpora has been studied by a number of researchers, (Fung 2000; Peters and Picchi 1995; Rapp 1999; Shahzad et al. 1999; Tanaka and Iwasaki 1996) among others. Their works rely on the assumption that if two words are mutual translations, then their more frequent collocates are likely to be mutual translations as well. Based on this assumption, the 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, through a similarity measure, as the standard cosine.

A different approach is presented in (Dejean et al. 2002), where translation equivalences are searched via similarities to dictionary entries. Even though this method, when combined to the previous one, improves the accuracy of the bilingual lexicon extracted, the level of performance reached is still below the one achieved with parallel corpora (this level obviously depends on the dictionary used as well as on the degree of comparability of the corpus under study), and does not allow one to consider correspondences between complex units, as terms.

4 Crossing the Language Barrier in Text Mining Applications

The algorithms we have presented in the previous section represent just one way of crossing the language barrier, namely through the building of a corpus specific translation lexicon. Other methods exist, most of which are built upon a vector space model, and the so-called bag-of-words representation of documents. Even though different attempts have been made to propose new models (as word sequences in Cancedda et al. 2003), most methods proposed so far are based on the vector space model (arguably, all kernel-based methods, as the one described in the above mentioned work, are implicitely based on vector spaces). We will review in this section the most common methods for crossing the language barrier in text mining applications, and more particularly for information retrieval purposes.

Dictionary-based Methods

A probabilistic bilingual lexicon, whether extracted from a parallel or comparable corpus, can naturally be represented as a translation matrix, P, the lines of which correspond to target words, the columns to source words and each element P_{ij} to the probability of translating source word j by target word i (i.e. the quantity $P(i|j)$ of previous section). Translating a source document then amounts to mapping the vector representing the document into target language through the matrix P, as: $d_t = P \cdot d_s$.

The above formulations can be extended to the case where the bilingual lexicon is directly derived from a bilingual dictionary, in which case the translation matrix P_d is a binary matrix defining the mapping between source and target words as found in the dictionary.

Latent Semantic Spaces

Unlike dictionary based methods which aim at establishing direct translation links between indexes representing documents, other techniques aim at capturing latent semantic spaces « shared » by the different documents of the mutilingual collection. Indeed, the main techniques (GVSM and LSI) used to build monolingual latent semantic spaces can be extended to the multilingual case to build interlingual concept spaces whenever a parallel corpus is available, i.e. whenever the collection itself or a significant part of it is a parallel corpus. We will here review some of the models previously presented in a monolingual setting and show their formulation when such a parallel corpus is available. We use, in the following, the decomposition of the term-document matrix *D* into language specific matrices *A* and *B*, such that $D = A^t \cdot B^t$, as well as the resulting singular value decomposition at the basis of LSI. The document vectors we consider are extended monolingual vectors, i.e. they contain 0's for all the term components in the other languages.

The multilingual version of GVSM is straightforwardly given by :

$$K(d_1, d_2) = d_1{}^t A \cdot B^t d_2$$

as established e.g. in (Brown et al., 1998).

For LSI, two different multilingual similarities can be derived. The first one directly relies on the singular value decomposition (Littman et al. 1998) and leads to:

$$K(d_1, d_2) = d_1{}^t U_c \cdot U_c^t d_2$$

where U_c represents the first *c* columns of *U*, obtained from *D* by singular value decomposition. The second multilingual LSI similarity is directly related to the formulation proposed in (Jiang and Littman 2000), which consists in decomposing, again through singular value decomposition, the two language specific matrices *A* and *B*, independently of each other. Using $U_{AB}{}^t = U_A{}^t \cdot U_B{}^t$, and denoting V_{AB} and V_{BA} the multilingual extensions of V_A and V_B (obtained by adding 0's on the components of the other language), the associated similarity takes the following form:

$$K(d_1, d_2) = d_1{}^t . U_{AB} . I_c . V_{BA}{}^t . V_{AB} . I_c U_{AB}{}^t . d_2$$

where *c* denotes once again the number of dimensions one wants to retain.

Comparing the forms of the two multilingual LSI similarities, one can note that they relate to the two traditional ways of establishing correspondences between concepts across languages: either through a set of interlingual concepts (the same set of concepts is used in both languages), or through a many-to-many mapping between sets of monolingual concepts. The first multilingual LSI similarity assumes a set of interlingual concepts, terms and documents being mapped into the vector space they induce, whereas the second similarity relies on two different monolingual sets of concepts, mapped across languages.

Remarks:

1. In the case when the collection is a comparable corpus rather than a parallel one (which is usually the case when e.g. the collection is made up from news articles), we can either try to extract a parallel corpus from the comparable one, and fall back on the above models, or extract a multilingual dictionary directly from the comparable corpus and rely on the dictionary-based method.
2. Other multilingual similarities can be envisaged, as the ones based on Fisher kernels derived from probabilistic models (see for example (Gaussier et al., 2001)), or canonical correlation analysis (Vinokourov et al. 2002). In this lat-

ter case, monolingual concepts, correlated across languages are searched for and use to relate documents in different languages.

4 Conclusion

We have addressed in this paper the problem of processing multilingual collections, for such text mining applications as cross-language clustering, categorisation and information retrieval. We have shown that in most cases it was not possible to guarantee equivalent processings of different languages. This implies that consistency and performance equivalence across languages are difficult to achieve, and should be viewed as objectives towards which we should tend. We have finally presented the conceptual differences between different methods used to cross the language barrier in text mining applications.

The growing availability of multilingual collections, both parallel and comparable, allows the extraction of corpus-specific resources, which are sufficient to meet most of the text mining needs. However, even though bilingual lexicon extraction from parallel corpora has been deeply investigated and has led to exploitable results, the situation is different with comparable corpora, for which state-of-art methods do not yield satisfactory results yet. Being able to fully exploit both parallel and comparable corpora is one of the major challenges for text mining on multilingual collections.

References

1. Bourigault, D. 1994. LEXTER, un Logiciel d'Extraction de TERminologie. Application à l'acquisition de connaissances à partir de textes. PhD Thesis. Paris: É cole des Hautes Études en Sciences Sociales.
2. Brown, P., Della Pietra, S., Della Pietra, V. and Mercer, R. 1993. "The Mathematics of Statistical Machine Translation: Parameter Estimation". Computational Linguistics, 19(2).
3. Brown, R.D., Carbonell, J.G., Yang, Y. Automatic Dictionary Extraction for Cross-Language Information Retrieval. In J. Véronis, editor, Parallel Text Processing, 2000.
4. Cancedda N., Gaussier E., Goutte C. and Renders J.-M. 2003. Word-Sequence kernels. In Journal of Machine Learning Research, Special Issue on Machine Learning Methods for Text.
5. Chuquet, H. and Paillard, M. 1989. Approche linguistique des problèmes de traduction anglais-français. Ophrys.
6. Debili, F. and Zribi, A. 1996. "Les dépendances syntaxiques au service de l'appariement des mots". In Proceedings of 10ième Congrès Reconnaissance des Formes et Intelligence Artificielle.
7. Déjean, H., Gaussier E., and Sadat F. 2002. An Approach based on Multilingual Thesauri and Model Combination for Bilingual Lexicon Extraction. Proceedings of the 19th International Conference on Computational Linguistics, COLING 2002.
8. Dempster, A., Laird, N. and Rubin, D. 1977. "Maximum likelihood from incomplete data via the EM algorithm". Journal of the Royal Statistical Society, 34(B).
9. Fung, P. 2000. A statistical View of Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora. In Jean Véronis (Ed.) Parallel Text Processing.

10. Gaussier, E. 1995. Modèles statistiques et patrons morphosyntaxiques pour l'extraction de lexiques bilingues de termes. PhD Thesis. Paris: Univ. Paris 7.
11. Gaussier, E. 1998. “Flow Network Model for Bilingual Lexicon Extraction”. In Proceedings of the joint COLING-ACL Conference.
12. Gaussier, E., Goutte, C., Popat, K., Chen, F. A Hierarchical Model for Clustering and Categorisaing Documents. In Advances in Information Retrieval, Lecture Notes in Computer Science, 2291. Springer-Verlag, 2002.
13. Gross, G. 1988. “Degré de figement des noms composés”. Langages, vol. 90.
14. Hull, D. 1998. “A practical approach to teminology alignment”. In Proceedings of the First Workshop on Computational Terminology. Montreal, 1998.
15. Jaakola, T.S., Haussler, D. Exploiting Generative Models in Discriminative Classifiers. In Advances in Neural Information Processing Systems11, 1999.
16. Jacquemin, C. 2001. Spotting and and discovering terms through NLP, MIT Press, Cambridge, MA.
17. Jiang, F., Littman, M. Approximate Dimension Equalization in Vector-Based Information Retrieval. In Proceedings of the 17th International Conference on Machine Learning. Morgan-Kauffman, 2000.
18. Justeson, J. and Katz, S. 1995. "Technical terminology: some linguistic properties and an algorithm for identification in text". Natural Language Engineering, 1(1).
19. Kupiec, J. 1993. “An algorithm for finding noun phrase correspondences in bilingual corpora". In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics.
20. Littman, M., Dumais, S., Landauer, K. Automatic Cross-Language Information Retrieval Using Latent Semantic Indexing. In G. Grefenstette, editor, Cross-Language Information Retrieval. Kluwer, 1998.
21. Mathieu-Colas, M. 1988. Typologie des noms composés. Rapport technique. Univ. Paris 13.
22. Maxwell, K. 1992. Automatic translation of English compounds: problems and prospects. Rapport technique Working Papers in Language Processing, 39, University of Essex.
23. Nkwenti-Azeh, B. 1992. Positional and Combinational characteristics of Satellite Communications terms. Technical Report, CCl-UMIST, Manchester.
24. Peters C. and Picchi E. 1995. Capturing the Comparable: A System for Querying Comparable Text Corpora, Proceedings of the 3rd International Conference on Statistical Analysis of Textual Data.
25. Rapp R. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. Proceedings of the European Association for Computational Linguistics.
26. Shahzad I., Ohtake K., Masuyama S. And Yamamoto K. 1999. Identifying Translations of Compound Using Non-aligned Corpora. Proceedings of the Workshop MAL.
27. Tanaka K. And Iwasaki H. 1996. Extraction of lexical translations from Non-Aligned Corpora. Proceedings of the 13th International Conference on Computational Linguistics, COLING’96.
28. Vinokourov, A., Shawe-Taylor, J., Cristianini, N. Inferring a semantic representation of text via cross-language correlation analysis, Advances in Neural Information Processing Systems 15, 2002.
29. Wu, D. 1997. "Stochastic inversion transduction grammars and bilingual parsing of parallel corpora". Computational Linguistics, 23(3).