A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora

E. Gaussier[†], J.-M. Renders[†], I. Matveeva[*], C. Goutte[†], H. Déjean[†]
[†]Xerox Research Centre Europe
6, Chemin de Maupertuis — 38320 Meylan, France
Eric.Gaussier@xrce.xerox.com
[*]Dept of Computer Science, University of Chicago
1100 E. 58th St. Chicago, IL 60637 USA
matveeva@cs.uchicago.edu

Abstract

We present a geometric view on bilingual lexicon extraction from comparable corpora, which allows to re-interpret the methods proposed so far and identify unresolved problems. This motivates three new methods that aim at solving these problems. Empirical evaluation shows the strengths and weaknesses of these methods, as well as a significant gain in the accuracy of extracted lexicons.

1 Introduction

Comparable corpora contain texts written in different languages that, roughly speaking, "talk about the same thing". In comparison to parallel corpora, ie corpora which are mutual translations, comparable corpora have not received much attention from the research community, and very few methods have been proposed to extract bilingual lexicons from such corpora. However, except for those found in translation services or in a few international organisations, which, by essence, produce parallel documentations, most existing multilingual corpora are not parallel, but comparable. This concern is reflected in major evaluation conferences on cross-language information retrieval (CLIR), e.g. CLEF[1], which only use comparable corpora for their multilingual tracks.

We adopt here a geometric view on bilingual lexicon extraction from comparable corpora which allows one to re-interpret the methods proposed thus far and formulate new ones inspired by latent semantic analysis (LSA), which was developed within the information retrieval (IR) community to treat synonymous and polysemous terms (Deerwester et al., 1990). We will explain in this paper the motivations behind the use of such methods for bilingual lexicon extraction from comparable corpora, and show how to apply them. Section 2 is devoted to the presentation of the standard approach, ie the approach adopted by most researchers so far, its geometric interpretation, and the unresolved synonymy and polysemy problems. Sections 3 to 4 then describe three new methods aiming at addressing the issues raised by synonymy and polysemy: in section 3 we introduce an extension of the standard approach, and show in appendix A how this approach relates to the probabilistic method proposed in (Dejean et al., 2002); in section 4, we present a bilingual extension to LSA, namely canonical correlation analysis and its kernel version; lastly, in section 5, we formulate the problem in terms of probabilistic LSA and review different associated similarities. Section 6 is then devoted to a large-scale evaluation of the different methods proposed. Open issues are then discussed in section 7.

[1]http://clef.iei.pi.cnr.it:2002/

2 Standard approach

Bilingual lexicon extraction from comparable corpora has been studied by a number of researchers, (Rapp, 1995; Peters and Picchi, 1995; Tanaka and Iwasaki, 1996; Shahzad et al., 1999; Fung, 2000, 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, the standard approach builds context vectors for each source and target word, translates the target context vectors using a general bilingual dictionary, and compares the translation with the source context vector:

1. For each source word v (resp. target word w), build a context vector $\overrightarrow{v}$ (resp. $\overrightarrow{w}$) consisting in the measure of association of each word e (resp. f) in the context of v (resp. w), $a(v, e)$.
2. Translate the context vectors with a general bilingual dictionary $\mathcal{D}$, accumulating the contributions from words that yield identical translations.
3. Compute the similarity between source word v and target word w using a similarity measures, such as the Dice or Jaccard coefficients, or the cosine measure.

As the dot-product plays a central role in all these measures, we consider, without loss of generality, the similarity given by the dot-product between $\overrightarrow{v}$ and the translation of $\overrightarrow{w}$:

$$\begin{aligned}\langle \overrightarrow{v}, \overrightarrow{tr(w)} \rangle &= \sum_e a(v,e) \sum_{f,(e,f) in \mathcal{D}} a(w,f) \\ &= \sum_{(e,f)\in\mathcal{D}} a(v,e)\, a(w,f) \qquad (1)\end{aligned}$$

Because of the translation step, only the pairs (e, f) that are present in the dictionary contribute to the dot-product.

Note that this approach requires some general bilingual dictionary as initial seed. One way to circumvent this requirement consists in automatically building a seed lexicon based on spelling and cognates clues (Koehn and Knight, 2002). Another approach directly tackles the problem from scratch by searching for a translation mapping which optimally preserves the intralingual association measure between words (Diab and Finch, 2000): the underlying assumption is that pairs of words which are highly associated in one language should have translations that are highly associated in the other language. In this latter case, the association measure is defined as the Spearman rank order correlation between their context vectors restricted to "peripheral tokens" (highly frequent words). The search method is based on a gradient descent algorithm, by iteratively changing the mapping of a single word until (locally) minimizing the sum of squared differences between the association measure of all pairs of words in one language and the association measure of the pairs of translated words obtained by the current mapping.

2.1 Geometric presentation

We denote by $s_i, 1 \leq i \leq p$ and $t_j, 1 \leq j \leq q$ the source and target words in the bilingual dictionary $\mathcal{D}$. $\mathcal{D}$ is a set of n *translation pairs* (s_i, t_j), and may be represented as a $p \times q$ matrix M, such that $M_{ij} = 1$ iff $(s_i, t_j) \in \mathcal{D}$ (and 0 otherwise).[2]

Assuming there are m distinct source words $e_1, \cdots, e_m$ and r distinct target words $f_1, \cdots, f_r$ in the corpus, figure 1 illustrates the geometric view of the standard method.

The association measure $a(v, e)$ may be viewed as the coordinates of the m-dimensional context vector $\overrightarrow{v}$ in the vector space formed by the orthogonal basis $(e_1, \cdots, e_m)$. The dot-product in (1) only involves source dictionary entries. The corresponding dimensions are selected by an orthogonal projection on the sub-space formed by $(s_1, \cdots, s_p)$, using a $p \times m$ projection matrix P_s. Note that $(s_1, \cdots, s_p)$, being a sub-family of $(e_1, \cdots, e_m)$, is an orthogonal basis of the new sub-space. Similarly, $\overrightarrow{w}$ is projected on the dictionary entries $(t_1, \cdots, t_q)$ using a $q \times r$ orthogonal projection matrix P_t. As M encodes the relationship between the source and target entries of the dictionary, equation 1 may be rewritten as:

$$\mathcal{S}(v,w) = \langle \overrightarrow{v}, \overrightarrow{tr(w)} \rangle = (P_s \overrightarrow{v})^\top M (P_t \overrightarrow{w}) \quad (2)$$

where $^\top$ denotes transpose. In addition, notice that M can be rewritten as $S^\top T$, with S an $n \times p$ and T an $n \times q$ matrix encoding the relations between words and pairs in the bilingual dictionary (e.g. S_{ki} is 1 iff s_i is in the k^{th} translation pair). Hence:

$$\mathcal{S}(v,w) = \overrightarrow{v}^\top P_s^\top S^\top T P_t \overrightarrow{w} = \langle S P_s \overrightarrow{v}, T P_t \overrightarrow{w} \rangle \quad (3)$$

which shows that the standard approach amounts to performing a dot-product in the vector space formed by the n pairs $((s_1, t_l), \cdots, (s_p, t_k))$, which are assumed to be orthogonal, and correspond to translation pairs.

2.2 Problems with the standard approach

There are two main potential problems associated with the use of a bilingual dictionary.

Coverage. This is a problem if too few corpus words are covered by the dictionary. However, if the context is large enough, some context words are bound to belong to the general language, so a general bilingual dictionary should be suitable. We thus expect the standard approach to cope well with the coverage problem, at least for frequent words. For rarer words, we can bootstrap the bilingual dictionary by iteratively augmenting it with the most probable translations found in the corpus.

Polysemy/synonymy. Because all entries on either side of the bilingual dictionary are treated as orthogonal dimensions in the standard methods, problems may arise when several entries have the same meaning (synonymy), or when an entry has several meanings (polysemy), especially when only one meaning is represented in the corpus.

Ideally, the similarities wrt synonyms should not be independent, but the standard method fails to account for that. The axes corresponding to synonyms s_i and s_j are orthogonal, so that projections of a context vector on s_i and s_j will in general be uncorrelated. Therefore, a context vector that is similar to s_i may not necessarily be similar to s_j.

A similar situation arises for polysemous entries. Suppose the word *bank* appears as both *financial institution* (French: *banque*) and *ground near a river*

[2] The extension to weighted dictionary entries $M_{ij} \in [0, 1]$ is straightforward but not considered here for clarity.

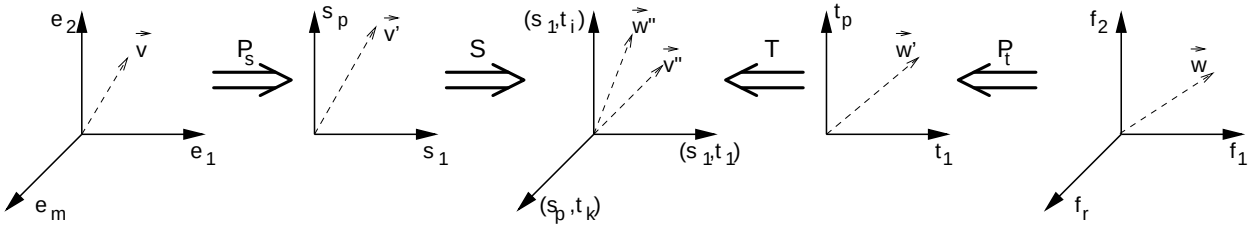

Figure 1: Geometric view of the standard approach

(French: *berge*), but only the pair *(banque, bank)* is in the bilingual dictionary. The standard method will deem similar *river*, which co-occurs with *bank*, and *argent* (*money*), which co-occurs with *banque*.

In both situations, however, the context vectors of the dictionary entries provide some additional information: for synonyms s_i and s_j, it is likely that $\overrightarrow{s_i}$ and $\overrightarrow{s_j}$ are similar; for polysemy, if the context vectors $\overrightarrow{banque}$ and $\overrightarrow{bank}$ have few translations pairs in common, it is likely that *banque* and *bank* are used with somewhat different meanings. The following methods try to leverage this additional information.

3 Extension of the standard approach

The fact that synonyms may be captured through similarity of context vectors[3] leads us to question the projection that is made in the standard method, and to replace it with a mapping into the sub-space formed by the context vectors of the dictionary entries, that is, instead of projecting $\overrightarrow{v}$ on the sub-space formed by $(s_1, \cdots, s_p)$, we now map it onto the sub-space generated by $(\overrightarrow{s_1}, \cdots, \overrightarrow{s_p})$. With this mapping, we try to find a vector space in which synonymous dictionary entries are close to each other, while polysemous ones still select different neighbors. This time, if $\overrightarrow{v}$ is close to $\overrightarrow{s_i}$ and $\overrightarrow{s_j}$, s_i and s_j being synonyms, the translations of *both* s_i and s_j will be used to find those words w close to v. Figure 2 illustrates this process. By denoting Q_s, respectively Q_t, such a mapping in the source (resp. target) side, and using the same translation mapping (S, T) as above, the similarity between source and target words becomes:

$$\mathcal{S}(v, w) = \langle SQ_s\overrightarrow{v}, TQ_t\overrightarrow{w}\rangle = \overrightarrow{v}^\top Q_s^\top S^\top T Q_t \overrightarrow{w} \quad (4)$$

A natural choice for Q_s (and similarly for Q_t) is the following $m \times p$ matrix:

$$Q_s = R_s^\top = \begin{pmatrix} a(s_1, e_1) & \cdots & a(s_p, e_1) \\ \vdots & \vdots & \vdots \\ a(s_1, e_m) & \cdots & a(s_p, e_m) \end{pmatrix}$$

[3]This assumption has been experimentally validated in several studies, e.g. (Grefenstette, 1994; Lewis et al., 1967).

but other choices, such as a pseudo-inverse of R_s, are possible. Note however that computing the pseudo-inverse of R_s is a complex operation, while the above projection is straightforward (the columns of Q correspond to the context vectors of the dictionary words). In appendix A we show how this method generalizes over the probabilistic approach presented in (Dejean et al., 2002). The above method bears similarities with the one described in (Besançon et al., 1999), where a matrix similar to Q_s is used to build a new term-document matrix. However, the motivations behind their work and ours differ, as do the derivations and the general framework, which justifies e.g. the choice of the pseudo-inverse of R_s in our case.

4 Canonical correlation analysis

The data we have at our disposal can naturally be represented as an $n \times (m + r)$ matrix in which the rows correspond to translation pairs, and the columns to source and target vocabularies:

$$\mathcal{C} = \begin{array}{c|ccc|ccc|l} & e_1 & \cdots & e_m & f_1 & \cdots & f_r & \\ \hline & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & (s^{(1)}, t^{(1)}) \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & (s^{(n)}, t^{(n)}) \end{array}$$

where $(s^{(k)}, t^{(k)})$ is just a renumbering of the translation pairs (s_i, t_j).

Matrix $\mathcal{C}$ shows that each translation pair supports two views, provided by the context vectors in the source and target languages. Each view is connected to the other by the translation pair it represents. The statistical technique of *canonical correlation analysis* (CCA) can be used to identify directions in the source view (first m columns of $\mathcal{C}$) and target view (last r columns of $\mathcal{C}$) that are *maximally correlated*, ie "behave in the same way" wrt the translation pairs. We are thus looking for directions in the source and target vector spaces (defined by the orthogonal bases $(e_1, \cdots, e_m)$ and $(f_1, \cdots, f_r)$) such that the projections of the translation pairs on these directions are maximally correlated. Intuitively, those directions define latent semantic axes

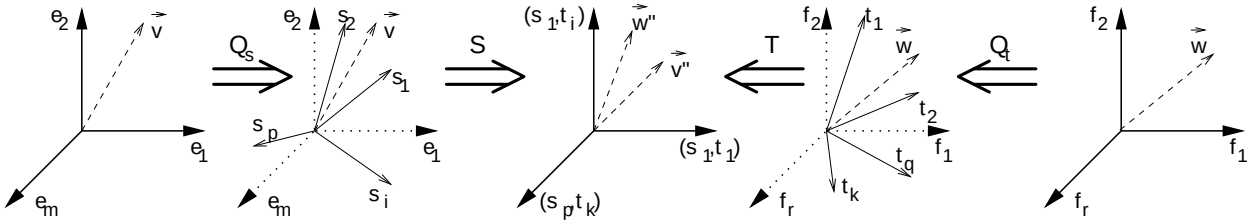

Figure 2: Geometric view of the extended approach

that capture the implicit relations between translation pairs, and induce a natural mapping across languages. Denoting by ξ_s and ξ_t the directions in the source and target spaces, respectively, this may be formulated as:

$$\rho = \max_{\xi_s,\xi_t} \frac{\sum_i \langle \xi_s, \overrightarrow{s}^{(i)} \rangle \langle \xi_t, \overrightarrow{t}^{(i)} \rangle}{\sqrt{\sum_i \langle \xi_s, \overrightarrow{s}^{(i)} \rangle \sum_j \langle \xi_t, \overrightarrow{t}^{(j)} \rangle}}$$

As in principal component analysis, once the first two directions (ξ_s^1, ξ_t^1) have been identified, the process can be repeated in the sub-space orthogonal to the one formed by the already identified directions. However, a general solution based on a set of eigenvalues can be proposed. Following e.g. (Bach and Jordan, 2001), the above problem can be reformulated as the following generalized eigenvalue problem:

$$\mathbf{B}\,\xi = \rho \mathbf{D}\,\xi \tag{5}$$

where, denoting again R_s and R_t the first m and last r (respectively) columns of $\mathcal{C}$, we define:

$$\mathbf{B} = \begin{pmatrix} 0 & R_t R_t^\top R_s R_s^\top \\ R_s R_s^\top R_t R_t^\top & 0 \end{pmatrix},$$

$$\mathbf{D} = \begin{pmatrix} (R_s R_s^\top)^2 & 0 \\ 0 & (R_t R_t^\top)^2 \end{pmatrix},\ \xi = \begin{pmatrix} \xi_s \\ \xi_t \end{pmatrix}$$

The standard approach to solve eq. 5 is to perform an incomplete Cholesky decomposition of a regularized form of $\mathbf{D}$ (Bach and Jordan, 2001). This yields pairs of source and target directions $(\xi_s^1, \xi_t^1), \cdots, (\xi_s^l, \xi_t^l)$ that define a new sub-space in which to project words from each language. This sub-space plays the same role as the sub-space defined by translation pairs in the standard method, although with CCA, it is derived from the corpus via the context vectors of the translation pairs. Once projected, words from different languages can be compared through their dot-product or cosine. Denoting $\Xi_s = \left[\xi_s^1, \ldots \xi_s^l\right]^\top$, and $\Xi_t = \left[\xi_t^1, \ldots \xi_t^l\right]^\top$, the similarity becomes (figure 3):

$$\mathcal{S}(v, w) = \langle \Xi_s \overrightarrow{v}, \Xi_t \overrightarrow{w} \rangle = \overrightarrow{v}^\top \Xi_s^\top \Xi_t \overrightarrow{w} \tag{6}$$

The number l of vectors retained in each language directly defines the dimensions of the final subspace used for comparing words across languages.

CCA and its kernelised version were used in (Vinokourov et al., 2002) as a way to build a cross-lingual information retrieval system from parallel corpora. We show here that it can be used to infer language-independent semantic representations from comparable corpora, which induce a similarity between words in the source and target languages.

5 Multilingual probabilistic latent semantic analysis

The matrix $\mathcal{C}$ described above encodes in each row k the context vectors of the source (first m columns) and target (last r columns) of each translation pair. Ideally, we would like to cluster this matrix such that translation pairs with synonymous words appear in the same cluster, while translation pairs with polysemous words appear in different clusters (soft clustering). Furthermore, because of the symmetry between the roles played by translation pairs and vocabulary words (synonymous and polysemous vocabulary words should also behave as described above), we want the clustering to behave symmetrically with respect to translation pairs and vocabulary words. One well-motivated method that fulfills all the above criteria is Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999).

Assuming that $\mathcal{C}$ encodes the co-occurrences between vocabulary words w and translation pairs d, PLSA models the probability of co-occurrence w and d via *latent classes* α:

$$P(w, d) = \sum_\alpha P(\alpha)\, P(w|\alpha)\, P(d|\alpha) \tag{7}$$

where, for a given class, words and translation pairs are assumed to be independently generated from class-conditional probabilities $P(w|\alpha)$ and $P(d|\alpha)$. Note here that the latter distribution is language-independent, and that the same latent classes are used for the two languages. The parameters of the model are obtained by maximizing the likelihood of the observed data (matrix $\mathcal{C}$) through Expectation-Maximisation algorithm (Dempster et al., 1977). In

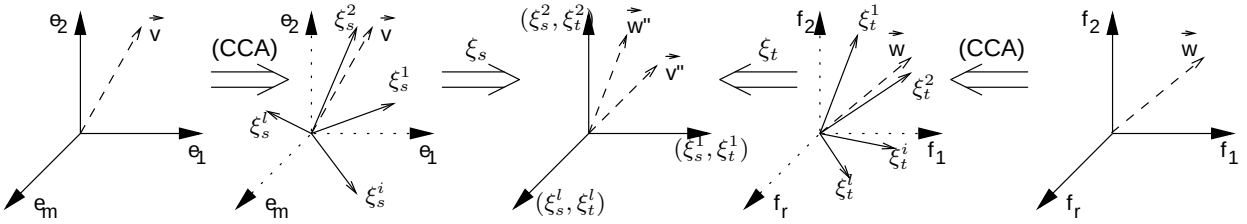

Figure 3: Geometric view of the Canonical Correlation Analysis approach

addition, in order to reduce the sensitivity to initial conditions, we use a deterministic annealing scheme (Ueda and Nakano, 1995). The update formulas for the EM algorithm are given in appendix B.

This model can identify relevant bilingual latent classes, but does not directly define a similarity between words across languages. That may be done by using Fisher kernels as described below.

Associated similarities: Fisher kernels

Fisher kernels (Jaakkola and Haussler, 1999) derive a similarity measure from a probabilistic model. They are useful whenever a direct similarity between observed feature is hard to define or insufficient. Denoting $\ell(w) = \ln P(w|\theta)$ the log-likelihood for example w, the Fisher kernel is:

$$K(w_1, w_2) = \nabla\ell(w_1)^\top \mathbf{I_F}^{-1} \nabla\ell(w_2) \qquad (8)$$

The Fisher information matrix $\mathbf{I_F} = \mathsf{E}\left(\nabla\ell(x)\nabla\ell(x)^\top\right)$ keeps the kernel independent of reparameterisation. With a suitable parameterisation, we assume $\mathbf{I_F} \approx \mathbf{1}$. For PLSA (Hofmann, 2000), the Fisher kernel between two words w_1 and w_2 becomes:

$$\begin{aligned} K(w_1, w_2) &= \sum_\alpha \frac{P(\alpha|w_1)P(\alpha|w_2)}{P(\alpha)} \qquad (9) \\ &+ \sum_d \widehat{P}(d|w_1)\widehat{P}(d|w_2) \sum_\alpha \frac{P(\alpha|d,w_1)P(\alpha|d,w_2)}{P(d|\alpha)} \end{aligned}$$

where d ranges over the translation pairs. The Fisher kernel performs a dot-product in a vector space defined by the parameters of the model. With only one class, the expression of the Fisher kernel (9) reduces to:

$$K(w_1, w_2) = 1 + \sum_d \frac{\widehat{P}(d|w_1)\widehat{P}(d|w_2)}{P(d)}$$

Apart from the additional intercept ('1'), this is exactly the similarity provided by the standard method, with associations given by scaled empirical frequencies $a(w, d) = \widehat{P}(d|w)/\sqrt{P(d)}$. Accordingly, we expect that the standard method and the Fisher kernel with one class should have similar behaviors. In addition to the above kernel, we consider two additional versions, obtained:through normalisation (NFK) and exponentiation (EFK):

$$\begin{aligned} NFK(w_1, w_2) &= \frac{K(w_1, w_2)}{\sqrt{K(w_1)K(w_2)}} \qquad (10) \\ EFK(w_1, w_2) &= e^{-\frac{1}{2}(K(w_1)+K(w_2)-2K(w_1,w_2))} \end{aligned}$$

where $K(w)$ stands for $K(w, w)$.

6 Experiments and results

We conducted experiments on an English-French corpus derived from the data used in the multilingual track of CLEF2003, corresponding to the newswire of months May 1994 and December 1994 of the *Los Angeles Times* (1994, English) and *Le Monde* (1994, French). As our bilingual dictionary, we used the ELRA multilingual dictionary,[4] which contains ca. 13,500 entries with at least one match in our corpus. In addition, the following linguistic preprocessing steps were performed on both the corpus and the dictionary: tokenisation, lemmatisation and POS-tagging. Only lexical words (nouns, verbs, adverbs, adjectives) were indexed and only single word entries in the dicitonary were retained. Infrequent words (occurring less than 5 times) were discarded when building the indexing terms and the dictionary entries. After these steps our corpus contains 34,966 distinct English words, and 21,140 distinct French words, leading to ca. 25,000 English and 13,000 French words not present in the dictionary.

To evaluate the performance of our extraction methods, we randomly split the dictionaries into a training set with 12,255 entries, and a test set with 1,245 entries. The split is designed in such a way that all pairs corresponding to the same source word are in the same set (training or test). All methods use the training set as the sole available resource and predict the most likely translations of the terms in the source language (English) belonging to the

[4] Available through www.elra.info

test set. The context vectors were defined by computing the mutual information association measure between terms occurring in the same context window of size 5 (ie. by considering a neighborhood of +/- 2 words around the current word), and summing it over all contexts of the corpora. Different association measures and context sizes were assessed and the above settings turned out to give the best performance even if the optimum is relatively flat. For memory space and computational efficiency reasons, context vectors were pruned so that, for each term, the remaining components represented at least 90 percent of the total mutual information. After pruning, the context vectors were normalised so that their Euclidean norm is equal to 1. The PLSA-based methods used the raw co-occurrence counts as association measure, to be consistent with the underlying generative model. In addition, for the extended method, we retained only the N ($N = 200$ is the value which yielded the best results in our experiments) dictionary entries closest to source and target words when doing the projection with Q. As discussed below, this allows us to get rid of spurious relationships.

The upper part of table 1 summarizes the results we obtained, measured in terms of F-1 score for different lengths of the candidate list, from 20 to 500. For each length, precision is based on the number of lists that contain an actual translation of the source word, whereas recall is based on the number of translations provided in the reference set and found in the list. Note that our results differ from the ones previously published, which can be explained by the fact that first our corpus is relatively small compared to others, second that our evaluation relies on a large number of candidates, which can occur as few as 5 times in the corpus, whereas previous evaluations were based on few, high frequent terms, and third that we do not use the same bilingual dictionary, the coverage of which being an important factor in the quality of the results obtained. Long candidate lists are justified by CLIR considerations, where longer lists might be preferred over shorter ones for query expansion purposes. For PLSA, the normalised Fisher kernels provided the best results, and increasing the number of latent classes did not lead in our case to improved results. We thus display here the results obtained with the normalised version of the Fisher kernel, using only one component. For CCA, we empirically optimised the number of dimensions to be used, and display the results obtained with the optimal value ($l = 300$).

As one can note, the extended approach yields the best results in terms of F1-score. However, its performance for the first 20 candidates are below the standard approach and comparable to the PLSA-based method. Indeed, the standard approach leads to higher precision at the top of the list, but lower recall overall. This suggests that we could gain in performance by re-ranking the candidates of the extended approach with the standard and PLSA methods. The lower part of table 1 shows that this is indeed the case. The average precision goes up from 0.4 to 0.44 through this combination, and the F1-score is significantly improved for all the length ranges we considered (bold line in table 1).

7 Discussion

Extended method As one could expect, the extended approach improves the recall of our bilingual lexicon extraction system. Contrary to the standard approach, in the extended approach, all the dictionary words, present or not in the context vector of a given word, can be used to translate it. This leads to a noise problem since spurious relations are bound to be detected. The restriction we impose on the translation pairs to be used (N nearest neighbors) directly aims at selecting only the translation pairs which are in true relation with the word to be translated.

Multilingual PLSA Even though theoretically well-founded, PLSA does not lead to improved performance. When used alone, it performs slightly below the standard method, for different numbers of components, and performs similarly to the standard method when used in combination with the extended method. We believe the use of mere co-occurrence counts gives a disadvantage to PLSA over other methods, which can rely on more sophisticated measures. Furthermore, the complexity of the final vector space (several millions of dimensions) in which the comparison is done entails a longer processing time, which renders this method less attractive than the standard or extended ones.

Canonical correlation analysis The results we obtain with CCA and its kernel version are disappointing. As already noted, CCA does not directly solve the problems we mentioned, and our results show that CCA does not provide a good alternative to the standard method. Here again, we may suffer from a noise problem, since each canonical direction is defined by a linear combination that can involve many different vocabulary words.

Overall, starting with an average precision of 0.35 as provided by the standard approach, we were able to increase it to 0.44 with the methods we consider. Furthermore, we have shown here that such an improvement could be achieved with relatively simple

	20	60	100	160	200	260	300	400	500	Avg. Prec.
standard	0.14	0.20	0.24	0.29	0.30	0.33	0.35	0.38	0.40	0.35
Ext (N=500)	0.11	0.21	0.27	0.32	0.34	0.38	0.41	0.45	0.50	0.40
CCA (l=300)	0.04	0.10	0.14	0.20	0.22	0.26	0.29	0.35	0.41	0.25
NFK(k=1)	0.10	0.15	0.20	0.23	0.26	0.27	0.28	0.32	0.34	0.30
Ext + standard	**0.16**	**0.26**	**0.32**	**0.37**	**0.40**	**0.44**	**0.45**	**0.47**	**0.50**	**0.44**
Ext + NFK(k=1)	0.13	0.23	0.28	0.33	0.38	0.42	0.44	0.48	0.50	0.42
Ext + NFK(k=4)	0.13	0.22	0.26	0.33	0.37	0.40	0.42	0.47	0.50	0.41
Ext + NFK (k=16)	0.12	0.20	0.25	0.32	0.36	0.40	0.42	0.47	0.50	0.40

Table 1: Results of the different methods; F-1 score at different number of candidate translations. *Ext* refers to the extended approach, whereas *NFK* stands for normalised Fisher kernel.

methods. Nevertheless, there are still a number of issues that need be addressed. The most important one concerns the combination of the different methods, which could be optimised on a validation set. Such a combination could involve Fisher kernels with different latent classes in a first step, and a final combination of the different methods. However, the results we obtained so far suggest that the rank of the candidates is an important feature. It is thus not guaranteed that we can gain over the combination we used here.

8 Conclusion

We have shown in this paper how the problem of bilingual lexicon extraction from comparable corpora could be interpreted in geometric terms, and how this view led to the formulation of new solutions. We have evaluated the methods we propose on a comparable corpus extracted from the CLEF colection, and shown the strengths and weaknesses of each method. Our final results show that the combination of relatively simple methods helps improve the average precision of bilingual lexicon extraction methods from comparale corpora by 10 points. We hope this work will help pave the way towards a new generation of cross-lingual information retrieval systems.

Acknowledgements

We thank J.-C. Chappelier and M. Rajman who pointed to us the similarity between our extended method and the model DSIR (distributional semantics information retrieval), and provided us with useful comments on a first draft of this paper. We also want to thank three anonymous reviewers for useful comments on a first version of this paper.

References

F. R. Bach and M. I. Jordan. 2001. Kernel independent component analysis. *Journal of Machine Learning Research*.

R. Besançon, M. Rajman, and J.-C. Chappelier. 1999. Textual similarities based on a distributional approach. In *Proceedings of the Tenth International Workshop on Database and Expert Systems Applications (DEX'99)*, Florence, Italy.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

H. Dejean, E. Gaussier, and F. Sadat. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *International Conference on Computational Linguistics, COLING'02*.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Mona Diab and Steve Finch. 2000. A statistical word-level translation model for comparable corpora. In *Proceeding of the Conference on Content-Based Multimedia Information Access (RIAO)*.

Pascale Fung. 2000. A statistical view on bilingual lexicon extraction - from parallel corpora to non-parallel corpora. In J. Véronis, editor, *Parallel Text Processing*. Kluwer Academic Publishers.

G. Grefenstette. 1994. *Explorations in Automatic Thesaurus Construction*. Kluwer Academic Publishers.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 289–296. Morgan Kaufmann.

Thomas Hofmann. 2000. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In *Advances in Neural Information Processing Systems 12*, page 914. MIT Press.

Tommi S. Jaakkola and David Haussler. 1999. Ex-

ploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems 11*, pages 487–493.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *ACL 2002 Workshop on Unsupervised Lexical Acquisition*.

P.A.W. Lewis, P.B. Baxendale, and J.L. Bennet. 1967. Statistical discrimination of the synonym/antonym relationship between words. *Journal of the ACM*.

C. Peters and E. Picchi. 1995. Capturing the comparable: A system for querying comparable text corpora. In *JADT'95 - 3rd International Conference on Statistical Analysis of Textual Data*, pages 255–262.

R. Rapp. 1995. Identifying word translations in nonparallel texts. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

I. Shahzad, K. Ohtake, S. Masuyama, and K. Yamamoto. 1999. Identifying translations of compound nouns using non-aligned corpora. In *Proceedings of the Workshop MAL'99*, pages 108–113.

K. Tanaka and Hideya Iwasaki. 1996. Extraction of lexical translations from non-aligned corpora. In *International Conference on Computational Linguistics, COLING'96*.

Naonori Ueda and Ryohei Nakano. 1995. Deterministic annealing variant of the EM algorithm. In *Advances in Neural Information Processing Systems 7*, pages 545–552.

A. Vinokourov, J. Shawe-Taylor, and N. Cristianini. 2002. Finding language-independent semantic representation of text using kernel canonical correlation analysis. In *Advances in Neural Information Processing Systems 12*.

Appendix A: probabilistic interpretation of the extension of standard approach

As in section 3, $SQ_s\overrightarrow{v}$ is an n-dimensional vector, defined over $((s_1,t_l),\cdots,(s_p,t_k))$. The coordinate of $SQ_s\overrightarrow{v}$ on the axis corresponding to the translation pair (s_i,t_j) is $\langle\overrightarrow{s_i},\overrightarrow{v}\rangle$ (the one for $TQ_t\overrightarrow{w}$ on the same axis being $\langle\overrightarrow{t_j},\overrightarrow{w}\rangle$). Thus, equation 4 can be rewritten as:

$$\mathcal{S}(v,w) = \sum_{(s_i,t_j)} \langle\overrightarrow{s_i},\overrightarrow{v}\rangle\langle\overrightarrow{t_j},\overrightarrow{w}\rangle$$

which we can normalised in order to get a probability distribution, leading to:

$$\mathcal{S}(v,w) = \sum_{(s_i,t_j)} P(v)P(s_i|v)P(w|t_j)P(t_j)$$

By imposing $P(t_j)$ to be uniform, and by denoting C a translation pair, one arrives at:

$$\mathcal{S}(v,w) \propto \sum_C P(v)P(C|v)P(w|C)$$

with the interpretation that only the source, resp. target, word in C is relevant for $P(C|v)$, resp. $P(w|C)$. Now, if we are looking for those ws closest to a given v, we rely on:

$$\mathcal{S}(w|v) \propto \sum_C P(C|v)P(w|C)$$

which is the probabilistic model adopted in (Dejean et al., 2002). This latter model is thus a special case of the extension we propose.

Appendix B: update formulas for PLSA

The deterministic annealing EM algorithm for PLSA (Hofmann, 1999) leads to the following equations for iteration t and temperature β:

$$P(\alpha|w,d) = \frac{P(\alpha)^\beta P(w|\alpha)^\beta P(d|\alpha)^\beta}{\sum_\alpha P(\alpha)^\beta P(w|\alpha)^\beta P(d|\alpha)^\beta}$$

$$P^{(t+1)}(\alpha) = \frac{1}{\sum_{(w,d)} n(w,d)} \sum_{(w,d)} n(w,d)P(\alpha|w,d)$$

$$P^{(t+1)}(w|\alpha) = \frac{\sum_d n(w,d)P(\alpha|w,d)}{\sum_{(w,d)} n(w,d)P(\alpha|w,d)}$$

$$P^{(t+1)}(d|\alpha) = \frac{\sum_w n(w,d)P(\alpha|w,d)}{\sum_{(w,d)} n(w,d)P(\alpha|w,d)}$$

where $n(w,d)$ is the number of co-occurrences between w and d. Parameters are obtained by iterating eqs 11–11 for each β, $0 < \beta \leq 1$.