A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora

E. Gaussier, J.-M. Renders
Xerox Research Centre Europe
6, Chemin de Maupertuis
38320 Meylan, France
Firstname.Lastname@xrce.xerox.com

I. Matveeva
Dept of Computer Science
University of Chicago
58th St. Chicago, IL 60637 USA
matveeva@cs.uchicago.edu

C. Goutte, H. Déjean
Xerox Research Centre Europe
6, Chemin de Maupertuis
38320 Meylan, France
Firstname.Lastname@xrce.xerox.com

Abstract
We adopt in this study a geometric view on bilingual lexicon extraction from comparable corpora. This view makes it possible to re-interpret the methods proposed so far and identify unresolved problems. We then motivate and formulate two new methods, partly inspired by latent semantic analysis, that aim at solving these problems. We finally evaluate these methods, showing their strengths and weaknesses. Our final results show 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 translation centers and 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), eg CLEF (http://clef.iei.pi.cnr.it:2002/), which only use comparable corpora for their multilingual tracks.

We adopt here a geometric view on bilingual lexicon extraction from comparable corpora which allows us to re-interpret the methods proposed so far and formulate new ones inspired by latent semantic analysis (LSA), which was developed within the information retrieval 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 and 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 formulate the problem in terms of probabilistic LSA and review different associated similarities. Section 5 is then devoted to a large-scale evaluation of the different methods proposed. Open issues are discussed in section 6.

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, a standard approach consists in building context vectors, for each source and target word, in order to capture the most significant collocates. The target context vectors are then translated using a general bilingual dictionary, and compared with the source context vectors.

Denoting by $\vec{v}$ the context vector associated with word $v$, the above approach can be described through the following algorithm:

1. For each source word $v$ (resp. target 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 $e$ in $\vec{v}$ (resp. $f$ in $\vec{w}$) is then weighted with a measure of its association with $v$ (resp. $w$), denoted as $\alpha(v, e)$.

2. The context vectors of the target words are then translated with a general bilingual dictionary $D$; when several words lead to the same translation, the weight of the translation is the sum...
of the weights of the original words.

3. The similarity between each source word $v$ and each target word $w$ is computed through standard similarity measures, e.g., Dice, Jaccard or 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 $\mathbf{v}$ and the translated context vector $\mathbf{tr}(w)$:

$$S(v, w) = \langle \mathbf{v}, \mathbf{tr}(w) \rangle$$

$$= \sum_{e} a(\mathbf{v}, e) \sum_{f, (e, f) \in \mathcal{D}} a(\mathbf{w}, f)$$

$$= \sum_{(e, f) \in \mathcal{D}} a(\mathbf{v}, e) a(\mathbf{w}, f)$$

Because of the translation, only words $e$ that are present in the dictionary will contribute to the dot-product. The above sum can thus be restricted to source words $e$ present in the bilingual dictionary.

2.1 Geometric presentation

In the remainder of the paper, we will denote by $s_1, 1 \leq i \leq p$ (resp. $t_j, 1 \leq j \leq q$) a source (resp. target) word present in the bilingual dictionary. A pair of words, translation of each other in the bilingual dictionary, will be called a translation pair, and will be denoted $(s_i, t_j)$. The set of all translation pairs, the cardinality of which will be denoted by $n$, can be represented by a $p \times q$ matrix $M$, such that:

$$M_{ij} = \begin{cases} 1 & \text{if } (s_i, t_j) \text{ is present in the dictionary} \\ 0 & \text{otherwise} \end{cases}$$

$M$ directly encodes the translation relations given by the dictionary. Lastly, we will assume that there are $m$ (resp. $r$) distinct words in the source (resp. target) corpus, denoted by $e_1, \ldots, e_m$ (resp. $f_1, \ldots, f_r$). $v$ will be used to denote a source word, and $w$ a target word.

As illustrated on figure 1, the association $a(\mathbf{v}, e)$ may be viewed as the coordinate of $\mathbf{v}$ on $e$, and the context vector $\mathbf{v}$ of source word $v$ is naturally represented as a vector in the $m$-dimensional vector space formed by the orthogonal basis $(e_1, \ldots, e_m)$. The dot-product (1) only involves words that are present in the dictionary, the only dimensions of interest in this vector space are the ones associated with the dictionary entries, $(s_1, \ldots, s_p)$. Selecting those dimensions amounts to project $\mathbf{v}$ on the sub-space formed by $(s_1, \ldots, s_p)$, through a $p \times m$ orthogonal projector that will be denoted $P_s$. As $(s_1, \ldots, s_p)$ is a sub-family of $(e_1, \ldots, e_m)$, it is an orthogonal basis of the new sub-space. A similar projection $P_t$ is conducted on the target side, restricting $\mathbf{w}$ to the target dictionary entries $(t_1, \ldots, t_q)$. The translation step of the standard approach (step 2 above) amounts to mapping, through the matrix $M$, the projection of $\mathbf{w}$ on $(t_1, \ldots, t_q)$ into $(s_1, \ldots, s_p)$. Equation 1 can then be rewritten:

$$S(v, w) = \langle P_s \mathbf{v}, (P_t \mathbf{w}) \rangle$$

(2)

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

$$S(v, w) = \langle \mathbf{v}, TP_t \mathbf{w} \rangle = \mathbf{v}^t P_s S^t T P_t \mathbf{w}$$

(3)

which shows that the standard approach finally amounts to performing a dot-product in the vector space formed by the $n$ pairs $(s_i, t_j)$. In this vector space, the axes are assumed to be orthogonal, and correspond to translation pairs.

2.2 Problems with the standard approach

The above approach relies on the assumption that the bilingual dictionary is appropriate for the corpus under consideration. This assumption may be wrong on two counts: first, the dictionary may contain few entries found in the corpus (coverage problem), and second, dictionary entries may have several meanings, i.e., be polysemous, with only part of their meanings represented in the corpus. Because of the structure of the vector space used for the final comparison (eq. 3), the standard approach fails to account for either polysemous or synonymous dictionary entries (or, in a broader sense, semantically related entries). Using a general bilingual dictionary to translate context vectors ensures that, provided context vectors are large enough, some of their elements will belong to the general language and to the bilingual dictionary. We therefore expect the standard approach to cope well with the coverage problem, at least for frequent words. For less frequent ones, one might consider a bootstrapping approach in which the bilingual dictionary is gradually augmented with translations found in the corpus.

There is no direct answer however to the problems raised by synonymy and polysemy of dictionary entries, and that is what motivated the present work. Let us illustrate how the standard approach fails in that case, before we review potential solutions. If two words $s_i$ and $s_l$ are synonyms in the

---

1 The extension to weighted bilingual dictionary entries is straightforward. For clarity, we only consider the more common case of unweighted entries.
corpus, it is desirable that their context vectors be more similar than for unrelated words. However, since axes corresponding to $s_i$ and $s_j$ are orthogonal in the vector space used for calculating the similarity, the standard method fails to account for synonymous dictionary entries. A similar situation arises for polysemous dictionary entries: if $s_i$ is used in different, polysemous ways in the corpus, then it is desirable to treat differently context vectors in which $s_i$ is used with different meanings. However, because of the direct projection on the translation pairs, the standard method fails to account for synonyms, even if the nearest context vectors in the corpus contain the French word *banque* and the English word *bank*.

In both situations, however, the context vectors of the dictionary entries provide us with some elements to address these problems: if $s_i$ and $s_j$ are synonyms, then it is likely that $\overrightarrow{s_i}$ and $\overrightarrow{s_j}$ are close together; on the other hand, if the nearest context vectors of *banque* contain $(s_{i_1}, \ldots, s_{i_k})$, and the one of *bank* $(t_{j_1}, \ldots, t_{j_k})$, and if few translation pairs can be built from these two sets, then it is likely that *banque* and *bank* are used with somewhat different meanings in the source and target corpora. The methods we are now turning to make use of these facts.

### 3 Extension of the standard approach

The fact that synonyms may be captured through similarity of context vectors\(^2\) 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 $(s_1, \ldots, s_p)$, we now map it into the sub-space generated by $(\overrightarrow{s_1}, \ldots, \overrightarrow{s_p})$. 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 $\overrightarrow{s_j}$ being synonyms, both the translations of $s_i$ and $s_j$ will be used to find those words $w$ close to $v$. Figure 2 illustrates this process. Let $Q_s$ (resp. $Q_t$) be the corresponding mapping on the source (resp. target) side. The similarity between source and target words becomes:

$$ S(v, w) = (SQ_s \overrightarrow{v}, TQ_t \overrightarrow{w}) = \overrightarrow{v}^T Q_s S Q_t \overrightarrow{w} 
\tag{4} $$

The mapping from $(\overrightarrow{s_1}, \ldots, \overrightarrow{s_p})$ to $(e_1, \ldots, e_m)$ is given by the $p \times m$ context vector matrix:

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

A natural choice for the reverse mapping $Q_s$ (and similarly for $Q_t$) is to take $Q_s = R_s^+$. An obvious alternative would be the pseudo-inverse of $R_s$ instead, computing that pseudo-inverse is a complex operation, while the above projection is straightforward (the rows of $Q$ are the context vectors of the dictionary words). In appendix A we show how this method generalizes the probabilistic approach presented in (Dejean et al., 2002).

### 4 Multilingual probabilistic LSA

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:

$$ C = \begin{pmatrix} e_1 & \cdots & e_m & f_1 & \cdots & f_r \\ (s, t)_1 & \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ (s, t)_n & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix} $$

where for clarity we have re-indexed as $(s, t)_k$ all the translation pairs $(s_i, t_j)$ in the bilingual dictionary.

---

\(^2\)This assumption has been experimentally validated in different studies, eg (Grefenstette, 1994; Lewis et al., 1967).
Figure 2: Geometric view of the extended approach

The matrix $C$ encodes in each row $k$ the context vectors of $s_i$ (first $m$ columns) and $t_j$ (last $r$ columns). What we are indeed trying to achieve is a clustering of the above matrix such that translation pairs with synonymous words appear in the same cluster, and such that translation pairs with polysemous words appear in more than one cluster (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; Hofmann, 2001).

Assuming here that the $C$ matrix encodes the co-occurrences between vocabulary words $w$ and translation pairs $d$, PLSA models the probability of co-occurrence between a vocabulary word $w$ and a translation pair $d$ via latent classes $\alpha$, as:

$$P(w, d) = \sum_{\alpha} P(\alpha) P(w|\alpha) P(d|\alpha)$$  \hspace{1cm} (5)

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 observations $C$ through the Expectation-Maximisation algorithm (Dempster et al., 1977). In 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.

Even though the model above can identify relevant bilingual latent classes, it does not directly define a similarity between words across languages. That can be done however through the use of Fisher kernels.

**Associated similarities: Fisher kernels**

Fisher kernels (Jaakkola and Haussler, 1999) have been proposed as a way to derive a similarity measure from a probabilistic model. Fisher kernels may be useful whenever the direct similarity between observed feature is insufficient. For example, in the context of a mixture model, two examples generated from the same class could be more similar than their locations in feature space may suggest.

If we denote by $\ell(w) = \ln P(x|\theta)$ the log-likelihood of the model for example $w$, the Fisher kernel is derived as:

$$K(w_1, w_2) = \nabla \ell(w_1)^T \mathbf{I}_F^{-1} \nabla \ell(w_2)$$  \hspace{1cm} (6)

where $\mathbf{I}_F = E \left( \nabla \ell(x) \nabla \ell(x)^T \right)$ is the Fisher information matrix, which keeps the kernel independent of re-parameterisation. For a suitable choice of parameterisation, the Fisher information matrix is usually approximated by the identity matrix. Using this approximation on PLSA (Hofmann, 2000), the Fisher kernel between two words $w_1$ and $w_2$ is:

$$K(w_1, w_2) = \sum_{\alpha} \frac{P(\alpha|w_1)P(\alpha|w_2)}{P(\alpha)}$$  \hspace{1cm} (7)

$$+ \sum_d \hat{P}(d|w_1) \hat{P}(d|w_2) \sum_{\alpha} \frac{P(\alpha|d, w_1)P(\alpha|d, w_2)}{P(d|\alpha)}$$

where $w_1$ and $w_2$ range over the vocabularies of the two languages, and $d$ over the translation pairs. Note that the above kernel amounts to a dot-product in the vector space defined by the parameters of the model, the size of which is approximately $K \times (n + m + 1)$, where $K$ is the number of latent classes. Interestingly, with only one class, the expression of the Fisher kernel (7) reduces to:

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

This is exactly the similarity provided by the standard method, with one intercept (giving the front ‘1’) and associations given by empirical frequencies $\hat{P}(d|w)/\sqrt{\hat{P}(d)}$. We thus conjecture that, despite
the different weightings adopted in context vectors, the standard method and the Fisher kernel with one class should have similar behaviors.

In addition, we consider two other kernels, which directly derive from (7) either through normalisation or through exponentiation. In summary, we retain the following three similarities for PLSA:

\[
K_1(w_1, w_2) = K(w_1, w_2)
\]

\[
K_2(w_1, w_2) = \frac{K(w_1, w_2)}{\sqrt{K(w_1)K(w_2)}}
\]

\[
K_3(w_1, w_2) = e^{-\frac{1}{2}(K(w_1)+K(w_2)-2FK(w_1,w_2))}
\]

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

5 Experiments and results

We conducted experiments on an English-French corpus derived from the data used in the multilingual track of CLEF2003. We used the documents 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,\(^3\) 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 resource: tokenisation, lemmatisation and POS-tagging. Only lexical words (nouns, verbs, adverbs, adjectives) were indexed and only single word entries in our resource were used. Infrequent words (frequency less than 5) 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.

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. 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 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. This pruning appeared to give even better performance than using the complete context vector, probably because it removes noisy, non statistically significant components. 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 \( N \) 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 value of \( N \) was optimized on a validation set.

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. Our considering long candidate lists is motivated by CLIR considerations, where longer lists might be preferred over shorter ones for query expansions 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.

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).

6 Discussion

Extended method: As expected, the extended approach improves the recall of our bilingual lexicon extraction system. In the standard approach, only the dictionary words found in the context vector of a given word are used to translate it, whereas in the extended approach all the dictionary words can \textit{a priori} be used. 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 select-

\(^3\)Available through www.elra.info
### 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.

| Method                  | 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       |
| 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       |

Multilingual PLSA: Although PLSA is theoretically well-founded, it 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 approach. 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.

The low scores we obtain for the first 20 candidates are explained by two factors. First, contrary to previous works, we do not restrict ourselves to highly frequent words for evaluation, but randomly select from the dictionary a set of words that may appear as little as 5 times in the corpus. Second, we rely on a standard collection used for CLIR purposes, which is rather small compared to the corpora used in previous works. In some cases, it may be possible to extend a given comparable corpus, e.g. through web querying, but this introduces biases that can be harmful in restricted, specialized domains.

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 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 our validation set. Such a combination could involve Fisher kernels only in a first step, as proposed in (Hofmann, 2000), 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 sure that we can gain over the combination we used here. Other issues concern the tuning of the different methods, as the optimal value of $N$ for the extended approach. We plan to investigate these issues in future work.

### 7 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 collection, 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 comparable corpora by 10 points. We hope our work will pave the way toward the new generation of cross-lingual information retrieval systems.

### References
S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *J. American Soc. Inf. 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 *Proc. COLING’02*.

A.P. Dempster, N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society, B*, 39(1):1–38.
We can however use any other association measure
ity distribution through appropriate normalisation,
the same axis being
ation pair
be rewritten as:
\\( SQ_s(v, w) = \sum_{(s_i, t_j)} \langle s_i, v; w \rangle \langle t_j, w \rangle \quad (8) \)

We can however use any other association measure
between \( s_i \) and \( v' \), and transform it into a probability
distribution through appropriate normalisation,
leading to:
\\[ S(v, w) = \sum_{(s_i, t_j)} P(s_i, v)P(t_j, 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:
\\[ 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 instead of a symmetric similarity between \( v \) and \( w \), we are looking for those \( w_s \)
closest to a given \( v \), we get:
\\[ 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 version of the EM algorithm (Ueda and Nakano, 1995) for PLSA leads to the
following equations for iteration \( t \) and temperature \( \beta \):
\\[ P(\alpha|w^{(i)}, d^{(i)}) = \frac{P(\alpha)^\beta P(w^{(i)}|\alpha)^\beta P(d^{(i)}|\alpha)^\beta}{\sum_\alpha P(\alpha)^\beta P(w^{(i)}|\alpha)^\beta P(d^{(i)}|\alpha)^\beta} \]
\\[ P^{(t+1)}(\alpha) = \frac{1}{N} \left( \sum_{(w,d)} n(w^{(i)}, d^{(i)}) P(\alpha|w^{(i)}, d^{(i)}) \right) \]
\\[ P^{(t+1)}(w|\alpha) = \frac{\sum_{(w,d)} n(w^{(i)}, d^{(i)}) P(\alpha|w^{(i)}, d^{(i)})}{\sum_{(w,d)} n(w^{(i)}, d^{(i)}) P(\alpha|w^{(i)}, d^{(i)})} \]
\\[ P^{(t+1)}(d|\alpha) = \frac{\sum_{(w,d)} n(w^{(i)}, d^{(i)}) P(\alpha|w^{(i)}, d^{(i)})}{\sum_{(w,d)} n(w^{(i)}, d^{(i)}) P(\alpha|w^{(i)}, d^{(i)})} \]

where we let \( n(w^{(i)}, d^{(i)}) \) denote the number of co-
ocurrences between \( w \) and \( d \). Parameters are
obtained by iterating the above equations until convergence, for each \( \beta, 0 < \beta \leq 1 \).