ICA for Bilingual Lexicon Extraction from Comparable Corpora

Amir HAZEM and Emmanuel MORIN

Laboratoire d'Informatique de Nantes-Atlantique (LINA)
Université de Nantes, 44322 Nantes Cedex 3, France
Amir.Hazem@univ-nantes.fr, Emmanuel.Morin@univ-nantes.fr

Abstract

Independent component analysis (ICA) is a statistical method used to discover hidden features from a set of measurements or observed data so that the sources are maximally independent. This paper reports the first results on using ICA for the task of bilingual lexicon extraction from comparable corpora. We introduce two representations of data using ICA. The first one is called global ICA (GICA) used to design a global representation of a context according to all the target entries of the bilingual lexicon, the second one is called local ICA (LICA) and is used to capture local information according to target bilingual lexicon entries that only appear in the context vector of the candidate to translate. Then, we merge both GICA and LICA to obtain our final model (GLICA). The experiments are conducted on two different corpora. The French-English specialised corpus 'breast cancer' of 1 million words and the French-English general corpus 'Le Monde / New-York Times' of 10 million words. We show that the empirical results obtained with GLICA are competitive with the standard approach traditionally dedicated to this task.

1. Introduction

The use of comparable corpora for the task of bilingual lexicon extraction has received great interest since the beginning of 1990. It was introduced by Rapp (1995) as an alternative to the inconvenience of parallel corpora, which are not always available and are also difficult to collect especially for language pairs not involving English and for specific domains, despite many previous efforts in compiling parallel corpora (Church and Mercer, 1993). According to Rapp (1995, p320): *<...The availability of a large enough parallel corpus in a specific field and for a given pair of languages will always be the exception, not the rule.>*

The standard approach proposed by Rapp (1995) for aligning words from comparable corpora is, without doubt, the gold standard and the main state of the art in this domain based on a word space model. Words are represented by context vectors in high dimensional vector spaces by using distributional statistics. Contextual information has been widely used in statistical analysis of natural language corpora (Deerwester et al., 1990), (Honkela et al., 1995), (Ritter and Kohonen, 1989). Words are represented by the contexts in which they occur. This representation is motivated by the distributional hypothesis, which states that words with similar meanings tend to occur in similar contexts. Many investigations and a number of studies have emerged, (Fung, 1995; Fung, 1998; Fung and Lo, 1998; Peters and Picchi, 1998; Rapp, 1999; Chiao and Zweigenbaum, 2002; Déjean et al., 2002; Gaussier et al., 2004; Morin et al., 2007; Laroche and Langlais, 2010, among others).

Word space models, are not specific to bilingual lexicon extraction. Considerable attention is given to it in current research on semantic indexing (Sahlgren and Karlgren, 2005). Many different applications use word space models, including information retrieval (Dumais et al., 1988), word sense disambiguation (Schütze, 1992), (Hanson et al., 1993), various semantic knowledge tests (Lund et al., 1995), (Karlgren and Sahlgren, 2001), and text categorisation (Sahlgren and Coster, 2004).

In the standard word space methodology, for bilingual lexicon extraction from comparable corpora, each word is represented by its context vector for both source and target languages. For a word to be translated in the source language, its context vector is first translated using a bilingual lexicon, then, a similarity measure is used between the translated context vector and all the target context vectors. Finally, The target words are ranked according to their similarity scores. It is worth noticing that context vectors which are the basis of the word space model, may contain information redundancy, and suffer from data sparseness. We believe that a better representation of context vectors, by using a subspace in which vectors are orthogonal and data is maximally independent, should provide a better representation of data and thus reach a better accuracy for word alignment. In this paper, we propose to apply the independent component analysis (ICA) transform, which is basically an extension of the principal component analysis (PCA) transform. Both have proven their efficiency in data representation in many fields such as face recognition, data compression, etc. The remainder of this paper is organised as follows: Section 2. presents the standard approach based on lexical context vectors dedicated to word alignment from comparable corpora. Section 3. describes ICA technique. Section 4. describes our approach. Section 5. describes the different linguistic resources used in our experiments. Section 6. evaluates the contribution of the standard and ICA approaches to the quality of bilingual terminology extraction through different experiments. Section 7. presents our discussion and finally, Section 8. presents our conclusion and some perspectives.

2. Standard Approach

The main work in bilingual lexicon extraction from comparable corpora is based on lexical context analysis and relies on the simple observation that a word and its translation tend to appear in the same lexical contexts. The basis of this observation consists in the identification of first-order affinities for each source and target language: "*First-order affinities describe what other words are likely to be*

found in the immediate vicinity of a given word" (Grefenstette, 1994a, p. 279). These affinities can be represented by context vectors, and each vector element represents a word which occurs within the window of the word to be translated (for instance a seven-word window approximates syntactical dependencies).
The implementation of this approach can be carried out by applying the following four steps (Rapp, 1995; Fung and McKeown, 1997):

Context Characterisation

Let us denote, by $\mathbf{i}$ the context vector of the word i [1]. All the words in the context of each word i are collected, and their frequency in a window of n words around i extracted. For each word i of the source and the target languages, we obtain a context vector $\mathbf{i}$ where each entry $\mathbf{i}_j$, of the vector is given by a function of the co-occurrences of words j and i. Usually, association measures such as mutual information (Fano, 1961) or the log-likelihood (Dunning, 1993) are used to define vector entries.

Vector Transfer

The words of the context vector $\mathbf{i}$ are translated using a bilingual dictionary. Whenever the bilingual dictionary provides several translations for a word, all the entries are considered but weighted according to their frequency in the target language. Words with no entry in the dictionary are discarded.

Target Language Vector Matching

A similarity measure, $\mathrm{sim}(\bar{\mathbf{i}}, \mathbf{t})$, is used to score each word, t, in the target language with respect to the translated context vector, $\bar{\mathbf{i}}$. Usual measures of vector similarity include the cosine similarity (Salton and Lesk, 1968) or the weighted Jaccard index (WJ) (Grefenstette, 1994b) for instance.

Candidate Translation

The candidate translations of a word are the target words ranked following the similarity score.

The translation of the words of the context vectors, which depends on the coverage of the bilingual dictionary vis-à-vis the corpus, is an important step of the standard approach; as more elements of the context vector are translated, the context vector will be more discriminating in selecting translations in the target language. This drawback can be partially circumvented by combining a general bilingual dictionary with a specialised bilingual dictionary or a multilingual thesaurus (Chiao and Zweigenbaum, 2003; Déjean et al., 2002). Moreover, this approach is sensitive to the choice of parameters such as the size of the context, the choice of the association and similarity measures. The most complete study about the influence of these parameters on the quality of bilingual alignment has been carried out by Laroche and Langlais (2010).

[1] Generally, bold lower case letters indicate vectors and bold upper case letters indicate matrices.

3. Independent Component Analysis

In the classic version of the linear ICA model (Jutten and Hérault, 1991), (Comon, 1994), (Hyvarinen et al., 2001), each observed random $x = (x_1, x_2, ..., x_n)^T$ is represented as a weighted sum of independent random variables $s = (s_1, ..., s_k, ..., x_n)^T$, such as:

$$x = As \tag{1}$$

where A is the mixing matrix that contains the weights which are assumed to be different for each observed variable and s is the vector of the independent components. If we denote the columns of matrix A by a_i the model can be written as:

$$x = \sum_{i=1}^{D} a_i s_i \tag{2}$$

The statistical model in equation 1 is called the ICA model which describes how the observed data are generated by a process of mixing the components s_i. Both the mixing matrix A and the independent components s are learned in an unsupervised manner from the observed data x.
The starting point for ICA is the assumption that the components s_i are statistically independent. ICA can be seen as an extension to principal component analysis (PCA) and factor analysis. The main difference between ICA and PCA is, while PCA finds projections which have maximum variance, ICA finds projections which are maximally non-Gaussian. PCA is useful as a pre-processing technique that can reduce the dimension of the data with minimum mean-squares error. In contrast, the purpose of ICA is not dimension reduction. For our analysis we applied the FastICA (Hyvarinen, 1999) algorithm where the data matrix X is considered to be a linear combination of independent components:

$$X = AS \tag{3}$$

where columns of S contain the independent components and A is a linear mixing matrix. The dimension of the data was first reduced by PCA in order to decorrelate the data, to reduce over-learning and to get the square mixing matrix A. After variance normalisation (the whitened data), n independent components which create a feature representation in the component space were extracted with ICA.

4. Method

Our method consists in building a discriminating subspace using ICA which represents a double interest. Indeed, the mathematical properties of ICA ensure a better data representation, and using PCA as a pre- processing step, provides a dimension reduction which can be very useful when using large comparable corpora.

Data Representation

In our case, the observed data x is an N-by-N word-word matrix where columns represent contexts and rows represent words. The N words of the target language that appear in the bilingual dictionary are retained for constructing matrix X. Each column of X represents a context vector of a word i with $i \in N$. For a given element X_{cr} of matrix X, X_{cr} denotes the association measure of the r:th analysed word with the c:th context word. The chosen association measures are mutual information and the log likelihood.

GICA Representation

Data representation in GICA consists in building a whole component space s that represents a global view of words in the target corpus. Each component s_k encodes some interesting features extracted from the N target words. Here, we can analyse how the positions of the words in the target language are related according to the general representation of data which gives a global view of the distribution of words by considering contexts of all the words of the corpus that appear in the bilingual lexicon.

LICA Representation

Data representation in LICA consists in building a partial component space s that represents a local view of words in the target corpus according to the translated context vector of the candidate. Each component s_k encodes some interesting features extracted from the M target words that are part of the translated context vector of the candidate. Here, we can analyse how the positions of the words in the target language are related according to the partial representation of data by considering only the contexts of the candidate. The aim of this specific representation is to capture information related to the candidate only. This can be seen as a local or a specific representation.

For each method GICA and LICA, we use the same context characterisation and vector transfer in the same way that the standard approach. Context vectors of source and target words are computed and the words of the context vector of the candidate are translated using a bilingual dictionary. The main difference of our method resides in building a new vector space using ICA that transforms matrix X into a new component space $s = (s_1, ..., s_k, ..., x_n)^T$. Matrix X can be seen as the concatenation of N context vectors of the target words that appear in the bilingual lexicon.

4.1. Words Projection

Once the new component space s is built, The translated context vector of the candidate and all the context vectors of the target words are projected into the new subspace.

Let us denote $\mathbf{i}$ a context vector of a given word i. The projection of the context vector of i in the new subspace and noted $\mathbf{i}_p$ is shown in equation 4.

$$\mathbf{i}_p = \mathbf{i}^T \times S \tag{4}$$

4.2. Distance Measure

As in the standard approach, the candidate translations of a word are the target words ranked following the similarity score or dissimilarities (proximities). Here we only deal with dissimilarity that can often be understood as distance. We use a normalised Euclidean distance also called Chord distance (Korenius et al., 2006) as shown in equation 5.

$$d(\mathbf{i}, \mathbf{j}) = \sqrt{\sum_{k=1}^{n} (\frac{\mathbf{i_k}}{||\mathbf{i}||} - \frac{\mathbf{j_k}}{||\mathbf{j}||})^2} \tag{5}$$

4.3. GLICA Model

Let us denote $d_{GL}(i,j)$, ($d_G(i,j)$ and $d_L(i,j)$), the GLICA, GICA and LICA distances. GLICA is merely a weighted sum of GICA and LICA as given by the following equation:

$$d_{GL}(i,j) = \lambda \times d_G(i,j) + (1-\lambda) \times d_L(i,j) \tag{6}$$

Although the representation of GLICA is simple, it is important to highlight the fact that this model retains only candidates that appear in both GICA and LICA. That is to say, all the target words that are not present in the local or the global independent component space are discarded.

5. Linguistic Resources

The experiments have been carried out on two different French-English corpora: a specialised corpus from the medical domain within the sub-domain of 'breast cancer' and a general corpus from newspapers 'LeMonde/New-York Times '. Due to the small size of the specialised corpus we wanted to conduct additional experiments on a large corpus to have a better idea of the behaviour of our approach. Both corpora have been normalised through the following linguistic pre-processing steps: tokenization, part-of-speech tagging, and lemmatisation. The function words have been removed and the words occurring less than twice (i.e. hapax) in the French and the English parts have been discarded.

5.1. Specialised Corpus

We have selected the documents from the Elsevier website[2] in order to obtain a French-English specialised comparable corpus. We have automatically selected the documents published between 2001 and 2008 where the title or the keywords contain the term 'cancer du sein' in French and 'breast cancer' in English. We collected 130 documents in French and 118 in English and about 530,000 words for each language. The comparable corpus comprised about 7,400 distinct words in French and 8,200 in English.

In bilingual terminology extraction from specialised comparable corpora, the terminology reference list required to evaluate the performance of the alignment programs is often composed of 100 single-word terms (SWTs) (180 SWTs in (Déjean and Gaussier, 2002), 95 SWTs in (Chiao and Zweigenbaum, 2002), and 100 SWTs in (Daille and Morin, 2005)). To build our reference list, we selected 400 French/English SWTs from the UMLS[3] meta-thesaurus and the *Grand dictionnaire terminologique*[4]. We kept only

[2] www.elsevier.com

[3] www.nlm.nih.gov/research/umls

[4] www.granddictionnaire.com/

the French/English pair of SWTs which occur more than five times in each part of the comparable corpus. As a result of filtering, 122 French/English SWTs were extracted.

5.2. General Corpus

We chose newspapers as they offer a large amount of data. We selected the documents from the French newspaper 'Le Monde' and the English newspaper 'The New-York Times '. We automatically selected the documents published between 2004 and 2007 and obtained 5 million words for each language. The comparable corpus comprised about 41,390 distinct words in French and 44,311 in English.
The terminology reference list is much more consequential and contains 500 SWTs. It has been extracted from ELRA-M0033 randomly.

5.3. Bilingual Dictionary

The French-English bilingual dictionary required for the translation phase was the ELRA-M0033 dictionary. It contains, after projection in the 'breast cancer 'corpus and linguistic pre-processing steps, 3600 English single words and 3550 french single. And contains after projection in the corpus 'Le Monde/New-York Times' and linguistic pre-processing steps, 17.100 English single words and 16600 french single words belonging to the general language.

6. Experiments and Results

In this section, we first give the parameters of the standard and ICA based approaches, than we present the results conducted on the two corpora presented above: 'Breast cancer' and 'LeMonde/New-YorkTimes'.

6.1. Experimental Setup

Three major parameters need to be set to the standard approach and the ICA based approaches (LICA, GICA and GLICA), namely the similarity measure, the association measure defining the entry vectors and the size of the window used to build the context vectors. Laroche and Langlais (2010) carried out a complete study of the influence of these parameters on the quality of bilingual alignment. As a similarity measure, we chose to use the Cosine (Salton and Lesk, 1968) and the Weighted Jaccard Index (Grefenstette, 1994b) for the standard approach, while for ICA approaches, we chose the Euclidean distance which is the standard measure for PCA and ICA transforms. The entries of the context vectors were determined by the mutual information (Fano, 1961) and the log-likelihood (Dunning, 1993), and we used a seven-word window since it approximates syntactic dependencies. Other combinations of parameters were assessed but the previous parameters turned out to give the best performance.

6.2. Evaluation on the Breast Cancer Corpus

We investigated the performance of the standard approach (SA) and ICA based approaches (GICA, LICA and GLICA) on the 'Breast Cancer' corpus, using the evaluation list of 122 words.

We evaluate the accuracy by using the term : "top k" which means that the correct translation was found in the first k words presented by a given approach.

Evaluation Using Mutual Information

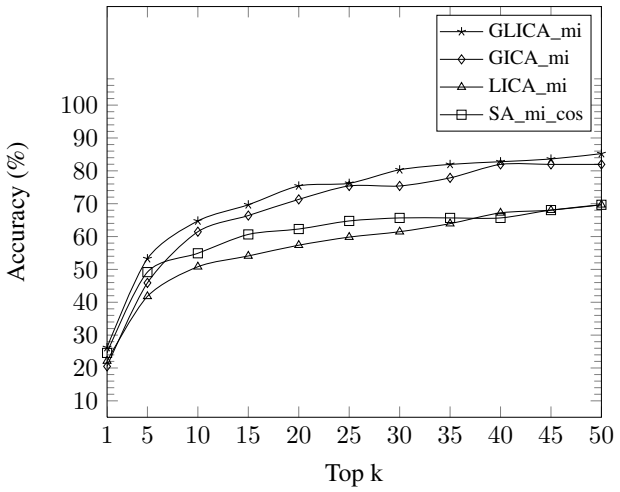

Figure 1: Accuracy at top k for the breast cancer corpus using mutual information.

We can see in Figure 1 that GLICA_mi approach always outperforms the standard approach for all values of k. The accuracy at the top 20 for SA_mi_cos is 62.29% while GLICA_mi approach gives 75.40%. We can also notice that GICA_mi outperforms SA_mi_cos from $k = 5$. Even if LICA_mi is almost always under the other approaches, according to Figure 1, it remains useful for GLICA.

Evaluation Using Log-Likelihood

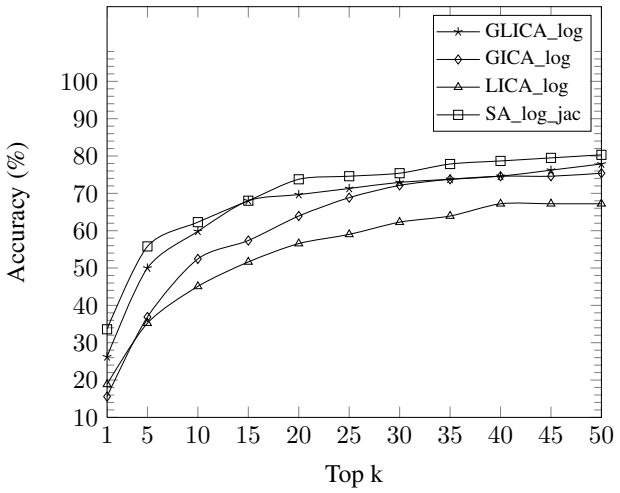

Figure 2: Accuracy at top k for the breast cancer corpus using log-likelihood.

We can see in Figure 2 that GLICA_log approach is under the standard approach for almost all values of k (except at $k = 15$). The accuracy at the top 20 for SA_log_jac is 73.77% while GLICA_mi approach gives 69.67%. Both, GICA_log and LICA_log are also under the baseline.

According to Figure 1 and Figure 2, we can notice that the best configuration for the standard approach is SA_log_jac with an accuracy of 73.77% for the top 20, while for our approach, the best configuration is GLICA_mi with an accuracy of 75.40% for the top 20. It is worth to notice that the merging process of the local and the global ICA plays an important role for improving the accuracy of our final model GLICA.

Evaluation on the best configuration of the Standard and GLICA approaches

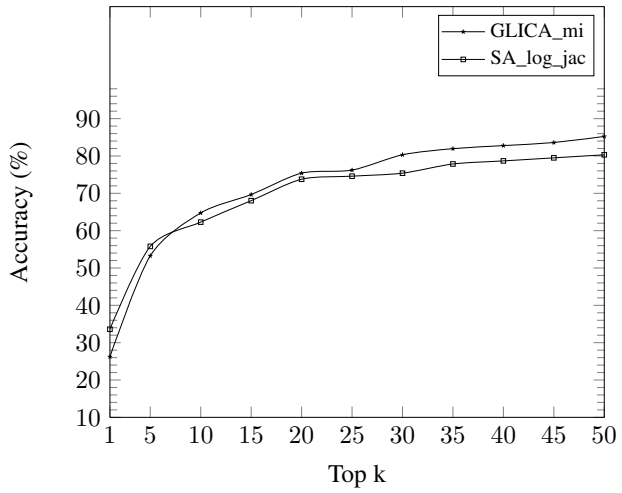

Figure 3: Accuracy at top k for the breast cancer corpus using the best parameters configuration of the standard and GLICA approaches.

Figure 3 presents the best performance of the standard and GLICA approaches. We can see that our approach outperforms the standard approach from $k > 5$. GLICA_mi reaches an accuracy of 64.75% at $k = 10$ and 75.40% at $k = 20$ while the standard approaches reaches an accuracy of 62.29% at $k = 10$ and 73.77% at $k = 20$. We can also notice that the standard approach outperforms our approach for both $k = 1$ and $k = 5$. GLICA_mi reaches an accuracy of 26.22% at $k = 1$ and 53.27% at $k = 5$ while the standard approach reaches an accuracy of 33.60% at $k = 1$ and 55.79% at $k = 5$.

Evaluation of the GLICA approach according to λ

Figure 4 shows how the GLICA (GLICA_mi) approach can be sensitive to the variations of the parameter λ. It seems that our approach is more accurate for $0.5 < \lambda < 0.9$ which means that the merging process gives more importance to the global ICA (GICA) than to the local ICA (LICA).

Evaluation on the LeMonde/New-YorkTimes Corpus

We then investigate the performance of the standard approach (SA) and ICA based approaches (GICA, LICA and GLICA) on 'LeMonde/New-YorkTimes' corpus, using an evaluation list of 500 words.

Evaluation Using Mutual Information

We can see in Figure 5 that GICA_mi LICA_mi and GLICA_mi approaches always outperform the standard approach for all values of k. The accuracy for the top 20

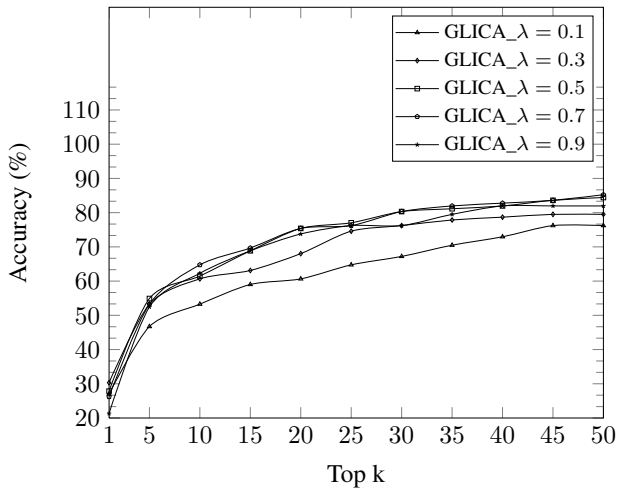

Figure 4: Accuracy at top k for the breast cancer corpus according to λ.

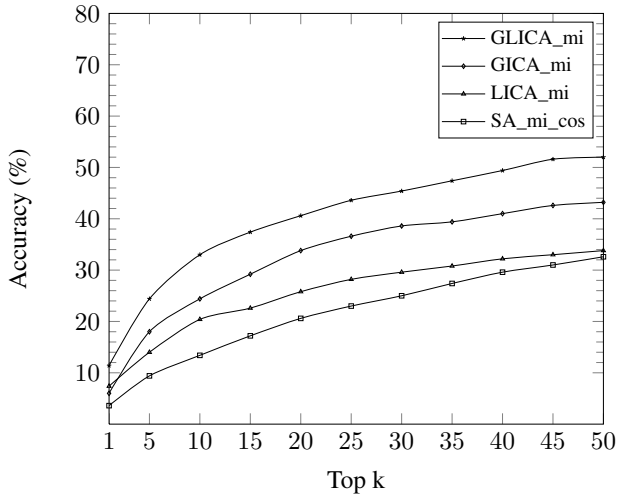

Figure 5: Accuracy at top k for LeMonde/NewYorkTimes using mutual information

for SA_mi_cos is 20.6% while GICA_mi approach gives 33.8%, LICA_mi approach gives 25.8% and GLICA_mi approach gives 40.6%. According to Figure 5 All the ICA models outperform the standard approach for this configuration (using mutual information as the association measure).

Evaluation Using Log-Likelihood

We can see in Figure 6 that the GLICA_log is slightly better than the standard approach. The accuracy for the top 20 for SA_log_jac is 38.8% while GLICA_mi approach gives 39.4%. Both, GICA_log and LICA_log are under the baseline.

According to Figure 5 and Figure 6, we can notice that the best configuration for the standard approach is SA_log_jac with an accuracy of 38.8% at the top 20, while for our approach, the best configuration is GLICA_mi with an accuracy of 40.6% at the top 20. It is also interesting to notice that GLICA_log outperforms SA_log_jac with an accuracy of 39.4% for $k = 20$.

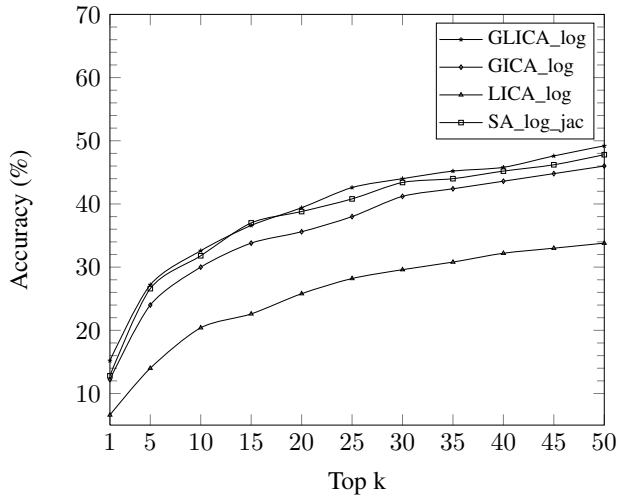

Figure 6: Accuracy at top k for LeMonde/NewYorkTimes using log-likelihood

Evaluation on the best configuration of the Standard and GLICA approachs

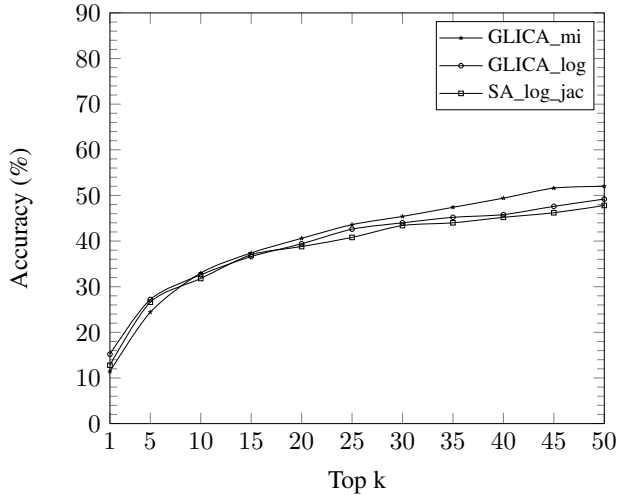

Figure 7: Accuracy at top k for LeMonde/NewYorkTimes corpus using the best parameters configuration of the standard and GLICA approaches.

Figure 7 presents the best performance of the standard and GLICA approaches. We can see that our approach outperforms the standard approach from $k > 5$. GLICA_mi reaches an accuracy of 33% at $k = 10$ and 40.6% at $k = 20$ while the standard approach reaches an accuracy of 31.8% at $k = 10$ and 38.8% at $k = 20$. We can also notice that the standard approach outperforms our approach for both $k = 1$ and $k = 5$. GLICA_mi reaches an accuracy of 11.4% at $k = 1$ and 24.4% at $k = 5$ while the standard approach reaches an accuracy of 12.8% at $k = 1$ and 26.6% at $k = 5$. On the contrary, GLICA_log outperforms the standard approach for both $k = 1$ with an accuracy of 15.2% and $k = 5$ with an accuracy of 27.2%.

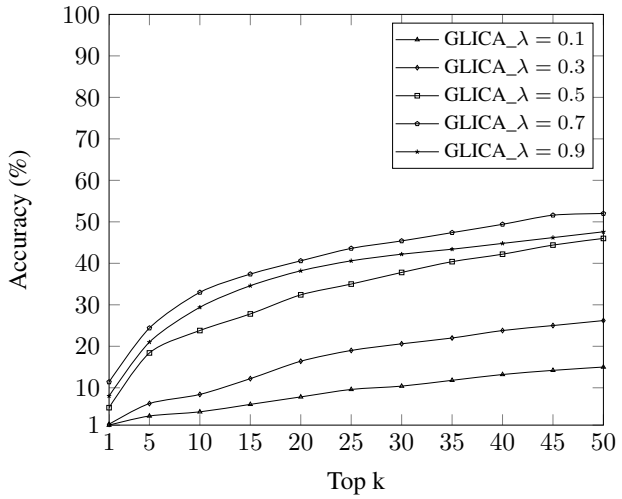

Figure 8: Accuracy at top k for LeMonde/NewYorkTimes according to λ.

Evaluation of the GLICA approach according to λ

Figure 8 shows how the GLICA (GLICA_mi) approach can be sensitive to the variations of the parameter λ. It seems that our approach is more accurate for $0.7 < \lambda < 0.9$ which means that the merging process gives more importance to the global ICA (GICA) than to the local ICA (LICA).

7. Discussion

The purpose of our experiments was to compare the proposed method with the baseline not only according to the best parameters configuration of each method, but also, in terms of behaviour according to the two main association measures that have proven their efficiency in thise domain (Rapp, 1999), and by choosing two different comparable corpora, a domain specific and a general one. The main interest of using two different comparable corpora is to test and validate our method according to the size and the type of the corpus.

For the 'breast cancer' corpus, the experiments based on mutual information, have shown that GLICA_mi and GICA_mi outperform SA_mi_cos while LICA_mi is slightly under SA_mi_cos. On the contrary, the use of the log-likelihood on the same corpus have shown that SA_log_jac outperforms LICA_log , GICA_log and GLICA_log. For the best configuration of each method, GLICA_mi shows better results than SA_log_jac. We can conclude from this first set of experiments on the breast cancer corpus that the standard approach reaches its best accuracy with log-likelihood while GLICA reaches its best performance with mutual information and for the best configuration of each method, GLICA_mi outperforms Sa_log_jac (except for $k = 1$ and $k = 5$).

For the 'LeMonde/New-YorkTimes' corpus, the results have also shown that GLICA_mi, GICA_mi and LICA_mi outperform SA_mi_cos. And that GLICA_log outperforms SA_log_jac while LICA_log, GICA_log were under the baseline (SA_log_jac). For the best configuration, GLICA_mi outperforms SA_log_jac (except for $k = 1$ and $k = 5$). This second set of experiments allows us to confirm

that both ICA-based methods and the standard method have the same behaviour on two different comparable corpora, and that the best association measure for the standard approach is the log-likelihood while for the ICA-based methods mutual information performs better.

According to the results stated previously, it is rightful to try to understand the reasons why GLICA accuracy is better using mutual information than log-likelihood on the 'Breast Cancer' corpus, while conversely, GLICA_log performs better than GLICA_mi on the 'LeMonde/New-YorkTimes' corpus for $k = 1$ and $k = 5$. Is it a matter of corpus size? or is it a matter of data representation? Further experiments need to be conducted in this direction.

In the GLICA approach, the parameter λ was fixed at 0.7, which means that we gave an advantage to GICA in the merging process. In fact, it was not our aim in this paper to deal with the parameter λ. We believe that in an appropriate environment, with an optimal data representation for both local and global component spaces, λ should be fixed at 0.5, so we consider GICA and LICA with the same importance. It is our hope for future work to carry out an in-depth study on this parameter, in addition to other merging techniques other than the one used for GLICA.

The GLICA method shows two advantages : (1) it is a merger of GICA which captures global context information of words, and LICA which captures local context information. Thus, GLICA has both global and local views on context representation. (2) Thanks to PCA pre-processing, GLICA offers a dimension reduction which enables a faster computation. As a comparison, the context vector size of a given word in the standard approach varies between the frequency of the word to its frequency multiplied by the size of the context window, which can easily reach thousands of words for frequent words and hundreds for less frequent words. For GLICA, the size of the context vectors in the ICA subspace is fixed to one hundred, it is independent from word frequency.

Finally, GLICA can be considered as promising for future work. The GLICA model does not take into account any linguistic or semantic information, it is just based on bag of words context. Many improvements need to be done especially for context representation.

8. Conclusion

In this paper, we have described and compared two techniques which focus on bilingual lexicon extraction from comparable corpora. The standard method considered as the state of the art and our method based on independent component analysis transform. This work represents, to the best of our knowledge, the first application of ICA to the task of bilingual lexicon extraction from comparable corpora. We have shown that a GLICA-based model can significantly outperform the standard approach model, for both the specialised and the general comparable corpora. The fact that our GLICA-based model outperforms the standard approach indicates that independent component analysis deserves more attention and can be considered as an alternative to the standard approach. It is our hope that this work will encourage further exploration of the potential of ICA modeling within alignment based on comparable corpora.

9. Acknowledgement

The research leading to these results has received funding from the French National Research Agency under grant ANR-08-CORD-013 and from the European Communitys Seventh Framework Programme (*/*FP7/2007-2013*/*) under Grant Agreement no 248005.

10. References

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1208–1212, Tapei, Taiwan.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2003. The Effect of a General Lexicon in Corpus-Based Identification of French-English Medical Word Translations. In Robert Baud, Marius Fieschi, Pierre Le Beux, and Patrick Ruch, editors, *The New Navigators: from Professionals to Patients, Actes Medical Informatics Europe*, volume 95 of *Studies in Health Technology and Informatics*, pages 397–402, Amsterdam. IOS Press.

Kenneth Ward Church and Robert L. Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1):1–24.

P. Comon. 1994. Independent component analysisa new concept? *Signal Processing*, 36:287314.

Béatrice Daille and Emmanuel Morin. 2005. French-English Terminology Extraction from Comparable Corpora. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCLNP'05)*, pages 707–718, Jeju Island, Korea.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

Hervé Déjean and Éric Gaussier. 2002. Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*, pages 1–22.

Hervé Déjean, Fatia Sadat, and Éric Gaussier. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 218–224, Tapei, Taiwan.

Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Scott Deerwester, and Richard Harshman. 1988. Using latent semantic analysis to improve access to textual information. In *SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS*, pages 281–285. ACM.

Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.

Robert M. Fano. 1961. *Transmission of Information: A Statistical Theory of Communications*. MIT Press, Cambridge, MA, USA.

Pascale Fung and Yuen Yee Lo. 1998. An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th international conference on Computational linguistics (COLING'98)*, pages 414–420.

Pascale Fung and Kathleen McKeown. 1997. Finding Terminology Translations from Non-parallel Corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora (VLC'97)*, pages 192–202, Hong Kong.

Pascale Fung. 1995. Compiling Bilingual Lexicon Entries From a non-Parallel English-Chinese Corpus. In David Farwell, Laurie Gerber, and Eduard Hovy, editors, *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'95)*, pages 1–16, Langhorne, PA, USA.

Pascale Fung. 1998. A Statistical View on Bilingual Lexicon Extraction: From ParallelCorpora to Non-parallel Corpora. In David Farwell, Laurie Gerber, and Eduard Hovy, editors, *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, pages 1–16, Langhorne, PA, USA.

Éric Gaussier, Jean-Michel Renders, Irena Matveeva, Cyril Goutte, and Hervé Déjean. 2004. A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, pages 526–533, Barcelona, Spain.

Gregory Grefenstette. 1994a. Corpus-Derived First, Second and Third-Order Word Affinities. In *Proceedings of the 6th Congress of the European Association for Lexicography (EURALEX'94)*, pages 279–290, Amsterdam, The Netherlands.

Gregory Grefenstette. 1994b. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publisher, Boston, MA, USA.

Stephen Jose Hanson, Jack D. Cowan, and C. Lee Giles, editors. 1993. *Advances in Neural Information Processing Systems 5, [NIPS Conference, Denver, Colorado, USA, November 30 - December 3, 1992]*. Morgan Kaufmann.

Timo Honkela, Ville Pulkki, and Teuvo Kohonen. 1995. Contextual relations of words in grimm tales analyzed by self-organizing map. In *ICANN*, pages 3–7.

A. Hyvarinen, J. Karhunen, and E Oja. 2001. Independent component analysis. *New York:a John Wiley Sons.*

A. Hyvarinen. 1999. Fast and robust fixed-point algorithms for independent component a analysis. *IEEE Transactions on Neural Networks*, 10(3):626634.

C Jutten and J. Hérault. 1991. Blind separation of sources. part i. an adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:110.

J. Karlgren and M. Sahlgren. 2001. From words to understanding. In *Foundations of Real-World Intelligence*, pages 294–308.

Tuomo Korenius, Jorma Laurikkala, Martti Juhola, and Kalervo Järvelin. 2006. Hierarchical clustering of a finnish newspaper article collection with graded relevance assessments. *Inf. Retr.*, 9(1):33–53.

Audrey Laroche and Philippe Langlais. 2010. Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 617–625, Beijing, China.

Kevin Lund, Curt Burgess, and Ruth Ann Atchley. 1995. Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pages 660–665.

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. 2007. Bilingual Terminology Mining – Using Brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 664–671, Prague, Czech Republic.

Carol Peters and Eugenio Picchi. 1998. Cross-language information retrieval: A system for comparable corpus querying. In Gregory Grefenstette, editor, *Cross-language information retrieval*, chapter 7, pages 81–90. Kluwer Academic Publishers.

Reinhard Rapp. 1995. Identify Word Translations in Non-Parallel Texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'95)*, pages 320–322, Boston, MA, USA.

Reinhard Rapp. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 519–526, College Park, MD, USA.

Helge Ritter and Teuvo Kohonen. 1989. Self-organizing semantic maps. *biological Cybernetics*, 4(64):241–254.

M. Sahlgren and R Coster. 2004. Using bag-of-concepts to improve the performance of support vector machines in text categorization. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING, August 23-27, Geneva, Switzerland*, pages 487–493.

Magnus Sahlgren and Jussi Karlgren. 2005. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering*, 11(3):327–341.

Gerard Salton and Michael E. Lesk. 1968. Computer evaluation of indexing and text processing. *Journal of the Association for Computational Machinery*, 15(1):8–36.

Hinrich Schütze. 1992. Word space. In *NIPS*, pages 895–902.