Improving Bilingual Terminology Extraction from Comparable Corpora via Multiple Word-Space Models

Amir HAZEM, Emmanuel MORIN
LINA, Universite de Nantes.
2 rue de la houssiniere 44000 Nantes, France.
amir.hazem@univ-nantes.fr, emmanuel.morin@univ-nantes.fr

Abstract

There is a rich flora of word space models that have proven their efficiency in many different applications including information retrieval (Dumais et al., 1988), word sense disambiguation (Schütze, 1993), various semantic knowledge tests (Lund et al., 1995; Karlgren and Sahlgren, 2001), and text categorization (Sahlgren and Karlgren, 2005). Based on the assumption that each model captures some aspects of word meanings and provides its own empirical evidence, we present in this paper a systematic exploration of the principal corpus-based word space models for bilingual terminology extraction from comparable corpora. We find that, once we have identified the best procedures, a very simple combination approach leads to significant improvements compared to individual models.

Keywords: Comparable corpora, Bilingual lexicon extraction, word-space models

1. Introduction

The distributional hypothesis which states that words with similar meanings tend to occur in similar contexts (Harris, 1954), has been extended to the bilingual scenario (Fung, 1998; Rapp, 1999). Hence, using comparable corpora, a translation of a source word can be found by identifying a target word with the most similar context. A popular method often used as a baseline is the *Standard Approach* (Fung, 1998). It consists of using the bag-of-words paradigm to represent words of source and target language by their context vector. After word contexts have been weighted using an association measure (the point-wise mutual information (Fano, 1961), the log-likelihood (Dunning, 1993), the discounted odds-ratio (Laroche and Langlais, 2010)), the similarity between a source word's context vector and all the context vectors in the target language is computed using a similarity measure (cosine (Salton and Lesk, 1968), Jaccard (Grefenstette, 1994)...). Finally, the translation candidates are ranked according to their similarity score. Many variants of the *Standard Approach* have been proposed. They can differ in context representation (window-based, syntactic-based) (Gamallo, 2008), corpus characteristics (small, large, general or domain specific...)(Chiao and Zweigenbaum, 2002; Déjean and Éric Gaussier, 2002; Morin et al., 2007), type of words to translate (single word terms (SWTs) or multi-word terms (MWTs))(Rapp, 1999; Daille and Morin, 2005), words frequency (less frequent, rare...)(Pekar et al., 2006), etc. There exist other approaches for bilingual lexicon extraction. Déjean et al. (2002) introduce the *extended approach* to avoid the insufficient coverage of the bilingual dictionary required for the translation of source context vectors. A variation of the latter method based on centroid is proposed by (Daille and Morin, 2005). Haghighi et al. (2008) employ dimension reduction using canonical component analysis (CCA).

The majority of the proposed approaches rely on context similarity. The starting point of context characterization is word co-occurrence statistics. It can provide a natural basis for semantic representation. Corpus-based word space models allow to go from distributional statistics to a geometric representation that induce the semantic representation of words from their patterns of co-occurrence in text. While literature suggest numerous techniques that could be used for that purpose, it is not obvious which is the best and furthermore in bilingual configuration where few studies have been proposed so far (Gaussier et al., 2004; Haghighi et al., 2008). The *Standard Approach* can be seen as a raw word-space model where each word defines a new dimension. This latter suffers from data sparseness where a tiny amount of words in language are distributionally promiscuous, and the vast majority only occur in a very limited set of contexts. In order to counter problems with very high dimensionality and data sparseness, different unsupervised and popular models can be applied, such as Latent Semantic Analysis (LSA), Principal Component analysis (PCA), Independent component Analysis (ICA) and Canonical Correlation Analysis (CCA), etc.

In this paper, we propose to extend the work of (Gaussier et al., 2004) in which they compared different projection techniques such as LSA, PLSA, and CCA, etc. Our first contribution is to investigate other techniques such as PCA and ICA in addition to LSA[1] on three specialized comparable corpora. Our second contribution is to propose a streamlined approach that can be seen as a combination system of multiple word space models based on empirical evidence. We show that our approach leads to significant improvements compared to individual models.

The remainder of this paper is organized as follows. Section 2. presents our approach. Section 3. describes the different linguistic resources used in our experiments. Section 4. evaluates the contribution of all the approaches on the quality of bilingual terminology extraction through different experiments. Section 5. presents our conclusions.

[1] CCA was assessed but the results were disappointing

2. Method

Starting from the intuition that each word space model (WSM) provides its own empirical evidence, we aim at taking advantage of each technique to yield better performance. Our method is first to build each word space model separately, then project words in each model and finally apply a simple combination technique based on scores and ranks as it is naturally used in information retrieval to re-rank translation candidates. In order to build a discriminant subspace we use mathematical transforms such as LSA (Deerwester et al., 1990), PCA and ICA (Jutten and Hérault, 1991; Comon, 1994; Hyvarinen et al., 2001). The main interest of using mathematical transforms is that their properties ensure a better data representation. For each method we use the same matrix representation. Data is represented as an $n \times (m + r)$ matrix in which the rows correspond to translation pairs, and the columns to source and target vocabularies. The most frequent $m + r$ words of the source and target language that appear in the bilingual dictionary are retained for constructing the matrix x. Each column of x represents a context vector of a word i with $i \in m + r$. For a given element x_{cr} of the matrix x, x_{cr} denotes the association measure of the r:th analyzed word with the c:th context word.

3. Experimental Setup

3.1. Corpus Data

The experiments have been carried out on three English-French comparable corpora. A specialized corpus of 1 million words from the medical domain within the sub-domain of 'breast cancer'[2], a specialized corpus from the domain of 'wind-energy' of 600,000 words and a specialized corpus from the domain of geology within the sub-domain of Volcanoes of 800,000 words. The three bilingual corpora have been normalized through the following linguistic pre-processing steps: tokenization, part-of-speech tagging, and lemmatization. We also used the French-English bilingual dictionary ELRA-M0033 of about 200,000 entries[3].

3.2. Reference Lists

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 Éric Gaussier, 2002), 95 SWTs in (Chiao and Zweigenbaum, 2002), and 100 SWTs in (Daille and Morin, 2005)). To build our reference lists, we selected only the French/English pair of SWTs which occur more than five times in each part of the comparable corpus. As a result of filtering, 321 French/English SWTs were extracted (from the UMLS[4] meta-thesaurus.) for the breast cancer corpus, 150 pairs for the wind-energy corpus and 158 for the volcano corpus.

3.3. Evaluation Measure

Three major parameters need to be set, namely the association measure, the similarity measure and the size of the window used to build the context vectors (Laroche and Langlais, 2010). As association measure we use the point-wise mutual information (PMI) (Fano, 1961), the log-likelihood measure (LL) (Dunning, 1993), the discounted odds-ratio (ODDS) (Laroche and Langlais, 2010) and the raw co-occurrence value (OCC). As a similarity measure, we use weighted Jaccard index (Grefenstette, 1994) for the standard approach and the normalized Euclidean distance (Korenius et al., 2006) for LSA, PCA and ICA. We also chose a 7-window size. Other combinations of parameters were assessed but the previous parameters turned out to give the best performance.

3.4. Baseline

The baseline in our experiments is the standard approach (Fung, 1998) often used for comparison (Pekar et al., 2006; Gamallo, 2008; Prochasson and Morin, 2009), etc.

4. Experiments and Results

We note that 'Top k' means that the correct translation is present in the k first candidates of the list returned by a given method. We use also the mean average precision *MAP* (Manning and Schuze, 2008).

4.1. Word Space Models Comparison

We carry out a comparison between the standard approach (SA), LSA, PCA and ICA according to the main association measures (OCC, PMI, ODDS and LL).

Table 1 shows that the results differ according to each association measure and word space model. For the co-occurrence association measure (OCC) for instance, the most appropriate WSM is ICA for the three comparable corpora while the results are more variable for the PMI where all the WSM's obtain in general the same results. For the ODDS measure, SA and LSA obtain the best results with a slight advantage of SA using the breast cancer corpus and a slight advantage of LSA on the volcano corpus. Results on the wind energy corpus are quite similar for both SA and LSA. Finally, the SA method shows the best results using the LL measure. We can note that the best MAP is obtained using OCC-ICA for the wind energy corpus with a score of 27.1% and a score of 27.9% and 46.8% respectively for the breast cancer and the volcano corpus using SA. In summary, according to Table 1 the best configurations are OCC-ICA, PMI-PCA, ODDS-LSA and LL-SA.

4.2. Word Space Models Combination

In this experiment we compare different combinations of WSM's according to their best individual performance. We use a weighted arithmetic score combination for LSA, PCA or ICA as they are based on the same similarity measure (Normalized Euclidean distance) and a weighted arithmetic ranks combination while using SA in the combination process (SA is based on Jaccard similarity). Here also different combination measures were assessed but the chosen ones give the best results.

Figure 1 shows the results of word space models combinations and each model taken separately using the best association measure. We notice that SA performs better than

[2] www.elsevier.com

[3] ELRA dictionary has been done by Sciper in the Technolangue/Euradic project

[4] http://www.nlm.nih.gov/research/umls

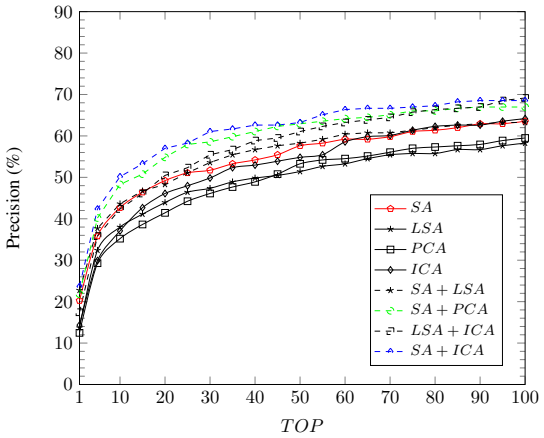
	SA	LSA	PCA	ICA	
OCC	16.9	14.4	06.7	**20.2**	Breast
PMI	**22.6**	21.1	18.5	21.1	
$ODDS$	**24.8**	22.6	18.1	19.2	
LL	**27.9**	10.0	09.7	14.8	
OCC	18.5	18.3	09.8	**27.1**	Wind
PMI	15.6	12.6	**17.6**	13.8	
$ODDS$	20.2	**21.3**	17.5	16.4	
LL	**24.2**	11.1	12.8	14.1	
OCC	30.1	26.0	16.6	**37.5**	Volcano
PMI	21.7	20.5	24.6	**26.8**	
$ODDS$	30.3	**33.9**	24.2	26.6	
LL	**46.8**	18.2	19.4	34.4	

Table 1: Mean average precision (MAP) of word space models using different association measures.

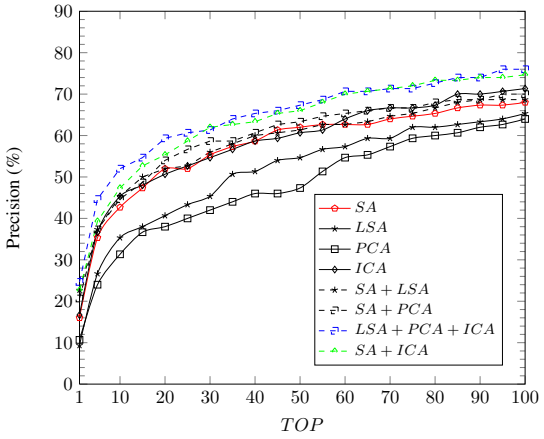
the three other models for the breast cancer corpus. We can also notice that LSA+ICA model outperforms SA. The best performance is obtained with the SA+ICA model closely followed by the SA+PCA model. For the wind energy corpus, SA and ICA obtain comparative results with a global advantage for ICA. Unlike the results of the breast cancer corpus, the best model is LSA+PCA+ICA. We notice that the scores of the SA+ICA model are close to those of LSA+PCA+ICA model. Finally, for the volcano corpus we notice that SA+ICA and LSA+ICA outperform SA significantly after the top 25.

5. Discussion and Conclusion

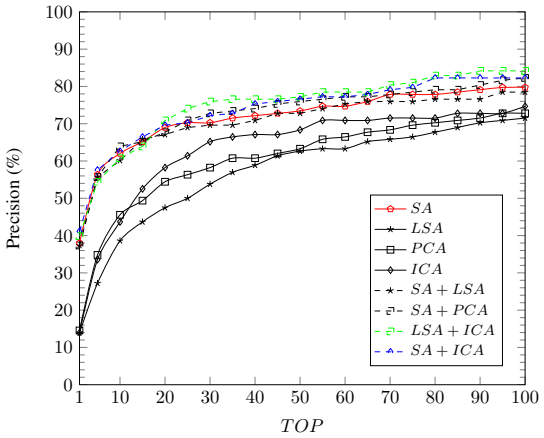
In this paper we have proposed the idea of combining different word space models to improve bilingual terminology extraction from comparable corpora. We notice that appropriate models combination leads to significant improvements as shown in the experiments. In theory, unsupervised word space models constitute an appropriate framework for data representation. However, in a practical case these models rely greatly on the initial data from which they build the new sub-space. Moreover, in a bilingual scenario there is an additional noise introduced by the translation phase. This can explain the results variability in some cases. For WSMs, the number of dimensions needs to be set. This parameter depends on data and can affect the performance. Nevertheless, we notice in our experiments that a number of 300 dimensions fixed empirically was an appropriate choice for the three corpora. For each WSM we need to select variables and samples from the corpus. In our case, variables are the words of the target language that appear in the bilingual dictionary and the samples are all the words of the target language. Variables allow a mapping between the source and the target language. The main question not solved in this study is: how to choose the appropriate variables and samples? Not all the words are of the same influence on a WSM as we notice in our experiments, so further investigation is certainly needed in this direction. Finally, our findings lend support for the hypothesis that combining multiple WSMs is an appropriate way to improve significantly bilingual terminology extraction from comparable corpora.



(a) Breast cancer corpus



(b) Wind energy corpus



(c) Volcano corpus

Figure 1: Comparison of different word space models combinations (the improvements indicate a significance at the 0.05 level using Student's t-test)

Acknowledgments

The research leading to these results has received funding from the French National Research Agency under grant ANR-12-CORD-0020 (CRISTAL project).

6. References

Chiao, Y.-C. and Zweigenbaum, P. (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.

Comon, P. (1994). Independent component analysis—a new concept? *Signal Processing*, 36:287–314.

Daille, B. and Morin, E. (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.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 281–285, New York, USA. ACM.

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

Déjean, H. 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.

Déjean, H., Sadat, F., and 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.

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

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

Gamallo, O. (2008). Evaluating two different methods for the task of extracting bilingual lexicons from comparable corpora. In *Proceedings of LREC 2008 Workshop on Comparable Corpora (LREC'08)*, pages 19–26, Marrakech, Marroco.

Gaussier, , Renders, J.-M., Matveeva, I., Goutte, C., and Déjean, H. (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.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publisher, Boston, MA, USA.

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., and Klein, D. (2008). Learning bilingual lexicons from monolingual corpora. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08)*, pages 771–779, Columbus, Ohio, USA.

Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

Hyvarinen, A., Karhunen, J., and Oja, E. (2001). Independent component analysis. *John Wiley Sons.*

Jutten, C. and Hérault, J. (1991). Blind separation of sources. part i. an adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10.

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

Korenius, T., Laurikkala, J., Juhola, M., and Järvelin, K. (2006). Hierarchical clustering of a finnish newspaper article collection with graded relevance assessments. *Inf. Retr.*, 9(1):33–53.

Laroche, A. and Langlais, P. (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.

Lund, K., Burgess, C., and Atchley, R. A. (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.

Manning, Christopher D.; Prabhakar, R. and Schuze, H. (2008). Introduction to information retrieval. Cambridge University Press.

Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (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.

Pekar, V., Mitkov, R., Blagoev, D., and Mulloni, A. (2006). Finding translations for low-frequency words in comparable corpora. *Machine Translation*, 20(4):247–266.

Prochasson, E. and Morin, E. (2009). Anchor points for bilingual extraction from small specialized comparable corpora. *TAL*, 50(1):283–304.

Rapp, R. (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.

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

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

Schütze, H. (1993). Word space. In Hanson, S. J., Cowan, J. D., and Giles, C. L., editors, *Advances in Neural Information Processing Systems 5*, pages 895–902. Morgan Kaufmann.