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

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

Abstract

- a rich flora of word space models have proven their efficiency in many different applications (information retrieval [Dumais 1988], word sense disambiguation [Schutze, 1992], various semantic knowledge tests [lund, 1995; Karlgren, 2001], and text categorization [SahlgrenK, 2005], etc.)
- we assume that each model captures some aspects of word meanings and provides its own empirical evidence
- we present a systematic exploration of the principal corpus-based word space models for bilingual terminology extraction from comparable corpora
- once we have identified the best procedures, a very simple combination approach leads to significant improvements compared to individual models

Approach

Intuition:

- each word space model (WSM) provides its own empirical evidence
- properties of mathematical transforms ensure better data representation
- we aim at taking advantage of each technique to yield better performance

Steps:

- build each word space model separately (we use mathematical transforms such as LSA [Deerweste 1990], PCA and ICA [Jutten 1991; Comon, 1994; Hyvarinen, 2001])
- project words in each model and compute vector similarity
- apply a simple combination technique based on scores and ranks (as it is naturally used in information retrieval) to re-rank translation candidates

Subspace Representation:

- for each method we use the same matrix representation
- data is represented as an ***n x (m + r)*** matrix in which rows correspond to translation pairs and 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*** included 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

Discussion and Conclusion

- in theory, unsupervised word space models constitute an appropriate framework for data representation
- in a practical case, these models rely greatly on the initial data from which they build the new sub-space (in a bilingual scenario there is an additional noise introduced by the translation phase)
- for WSMs, the number of dimensions needs to be set (depends on data and can affect the performance)
- 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
- 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
- combining different word space models improves bilingual terminology extraction from comparable corpora
- appropriate models combination leads to significant improvements as shown in the experiments
- the main question not solved in this study is: how to choose the appropriate variables and samples?
- our findings lend support for the hypothesis that combining multiple WSMs is an appropriate way to improve significantly bilingual terminology extraction from comparable corpora

Experiments and Results

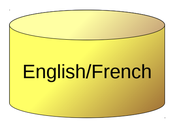
Resources

Breast cancer corpus
- 530,000 words
- reference list of 321 words

Wind-energy corpus
- 300,000 words
- reference list of 150 words

Volcano corpus
- 400,000 words
- reference list of 158 words

ELRA M0033 dictionary



240 000 entries

Word Space Models Comparison

- the results differ according to each association measure and word space model
- OCC ---> ICA
- PMI ---> variable results
- ODDS ---> SA and LSA
- LL ---> SA
- according to Table 1, the best configurations are OCC-ICA, PMI-PCA, ODDS-LSA and LL-SA

Word Space Models Combination

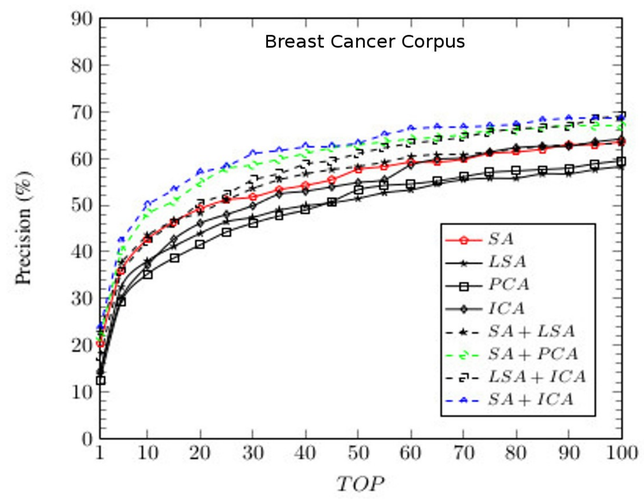
Breast Cancer
- SA performs better than LSA, PCA and ICA
- LSA+ICA model outperforms SA
- The best performance is obtained with the SA+ICA model closely followed by the SA+PCA model

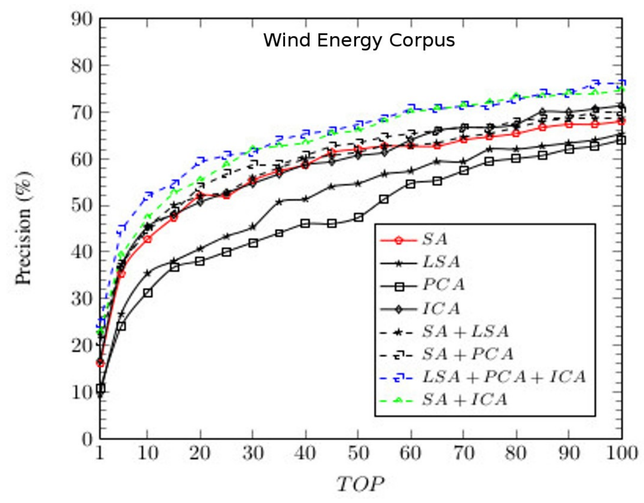
Wind energy
- SA and ICA obtain comparative results with a global advantage for ICA
- the best model is LSA+PCA+ICA
- SA+ICA model are close to those of LSA+PCA+ICA model

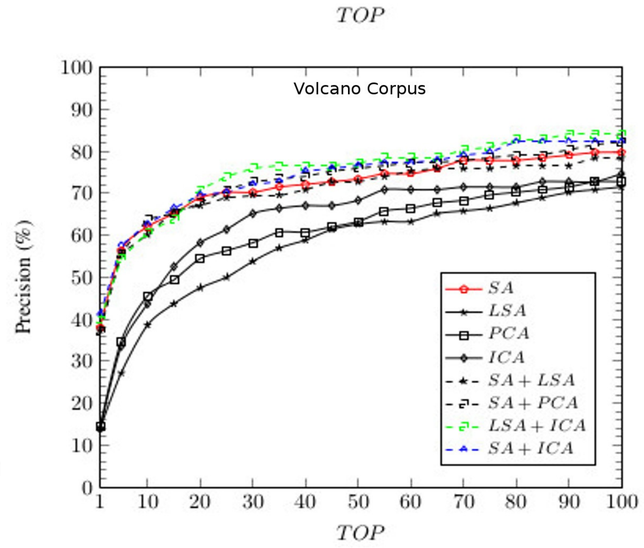
Volcano
- SA+ICA and LSA+ICA outperform significantly after the top 25

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

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







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

Acknowledgment

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