Comparing Word Sense Disambiguation and Distributional Models for Cross-Language Information Filtering

Cataldo Musto, Fedelucio Narducci, Pierpaolo Basile, Pasquale Lops, Marco de Gemmis, Giovanni Semeraro

Department of Computer Science - University of Bari "Aldo Moro", Italy
{cataldomusto,narducci,basilepp,lops,degemmis,semeraro}@di.uniba.it

Abstract. In this paper we deal with the problem of providing users with cross-language recommendations by comparing two different content-based techniques: the first one relies on a knowledge-based word sense disambiguation algorithm that uses MultiWordNet as sense inventory, while the latter is based on the so-called *distributional hypothesis* and exploits a dimensionality reduction technique called Random Indexing in order to build language-independent user profiles.
This paper summarizes the results already presented within the conference *AI*IA 2011* [1].

Keywords: Cross-language Information Filtering, Word Sense Disambiguation, Distributional Models

1 Introduction

Nowadays the amount of information we have to deal with is usually greater than the amount of information we can process in an effective way. In this context Information Filtering (IF) systems are rapidly emerging since they can adapt their behavior to individual users by learning their preferences and performing a progressive removal of non-relevant content. Specifically, the content-based filtering approach analyzes a set of documents (usually textual descriptions of items) and builds a model of user interests based on the features (usually keywords) that describe the items previously rated as relevant by an individual user. One relevant problem related to content-based approaches is the strict connection with the user language, since the information already stored in the user profile cannot be exploited to provide suggestions for items whose description is provided in other languages. In this paper we investigated whether it is possible to represent user profiles in order to create a mapping between preferences expressed in different languages. Specifically, we compared two approaches: the first one exploits a Word Sense Disambiguation (WSD) technique based on MultiWordnet, while the second one is based on the *distributional models*. It assumes that in every language each term often co-occurs with the same other terms (expressed in different languages, of course) thus, by representing a content-based user profile in terms of the co-occurences of its terms, user preferences become inerently independent from the language. The paper is organized as follows. Section 2 analyzes related works in the area of cross-language filtering and retrieval. An overview of the approaches is provided in Section 3. Experiments carried out in a movie recommendation scenario are described in Section 4. Conclusions and future work are drawn in the last section.

2 Related Work

The Multilingual Information Filtering task at CLEF 2009[1] has introduced the issues related to the cross-language representation in the area of Information Filtering. The use of distributional models [2] in the area of monolingual and multilingual Information Filtering is a relatively new topic. Recently the research about semantic vector space models gained more and more attention: Semantic Vectors (SV)[2] package implements a Random Indexing algorithm and defines a negation operator based on quantum logic. Some initial investigations about the effectiveness of the SV for retrieval and filtering tasks is reported in [3].

3 Description of the approaches

Learning profiles through MultiWordnet. In this approach we can imagine a general architecture composed by three main components: the *Content Analyzer* allows to obtain a language-independent document representation by using a Word Sense Disambiguation algorithm based on MultiWordnet [4]. Similary to WordNet, the basic building block of MultiWordNet is the synset (SYNonym SET), a structure containing sets of words with synonymous meanings, which represents a specific meaning of a word. In MultiWordNet, for example the Italian WordNet is aligned with the English one, so by processing textual descriptions of items in both the languages, a language-independent representation in terms of MultiWordNet synsets is obtained. The generation of the cross-language user profile is performed by the *Profile Learner*, using a naïve Bayes text classifier, since each document has to be classified as interesting or not with respect to the user preferences. Finally the *Recommender* exploits the cross-language user profiles to suggest relevant items by matching concepts contained in the semantic profile against those contained in the disambiguated documents.
Distributional Models. The second strategy used to represent items content in a semantic space relies on the distributional approach. This approach represents documents as vectors in a high dimensional space, such as `WordSpace` [2]. The core idea behind `WordSpace` is that words and concepts (and documents, as well) are represented by points in a mathematical space, and this representation is learned from text in such a way that concepts with similar or related meanings

[1] http://www.clef-campaign.org/2009.html
[2] http://code.google.com/p/semanticvectors/

are near to one another in that space (geometric metaphor of meaning). Therefore, semantic similarity between documents can be represented as proximity in a n-dimensional space. Since these techniques are expected to efficiently handle high dimensional vectors, a common choice is to adopt *dimensionality reduction* that allows for representing high-dimensional data in a lower-dimensional space without losing information. *Random Indexing* (RI) [2] targets the problem of dimensionality reduction by removing the need for the matrix decomposition or factorization since it is based on the concept of Random Projection: the idea is that high dimensional vectors randomly chosen are "nearly orthogonal". This yields a result that is comparable to orthogonalization methods, but saving computational resources. Given two corpus (one for language *L1* and another one for *L2*) we build two monolingual spaces S_{L1} and S_{L2} that share the same *random base* by following the procedure introduced in [3]. Since both spaces share the same random base it is possible to compare elements belonging to different spaces: for example we can compute how a user profile in S_{L1} is similar to an item in S_{L2} (or viceversa). This property is used to provide recommendations.

4 Experimental evaluation

The goal of the experimental evaluation was to measure the predictive accuracy of both the content-based multilingual recommendation approaches. We compared the language-independent user profiles represented through MultiWordNet sysnsets and the approaches based on distributional hypothesis (W-SV) and Random Indexing (W-RI), already presented in [3].

The experimental work has been performed on a subset of the MovieLens dataset[3] containing 40,717 ratings provided by 613 different users on 520 movies. The content information for each movie was crawled from both the English and Italian version of Wikipedia. User profiles are learned by analyzing the ratings stored in the MovieLens dataset while the effectiveness of the recommendation approaches has been evaluated by means of *Precision@n* ($n = 5, 10$). We designed four different experiments: In Exp#1 and Exp#2 we learned user profiles on movies with English (respectively, Italian) description and recommended movies with Italian (respectively English) description and we compared their accuracy with the classical monolingual baselines calculated in Exp#3 and Exp#4. Results of the experiments are reported in Table 1, averaged over all the users.

In general, the main outcome of the experimental session is that the strategy implemented for providing cross-language recommendations is quite effective for both the approaches. Specifically, the approach based on the bayesian classifier gained the best results in the *Precision@5.* This means that model has a higher capacity to rank the best items at the top of the recommendation list. On the other side, the absence of a linguistic pre-processing is one of the strongest point of the approaches based on the distributional model and the results gained by the W-SV and W-RI models in the *Precision@10* further underlined the effectiveness of this model. In conclusion, both the approaches gained good results. Even

[3] http://www.grouplens.org

Table 1. Precision@5 and Precision@10

	Precision@5			Precision@10		
Experiment	**W-SV**	**W-RI**	**Bayes**	**W-SV**	**W-RI**	**Bayes**
EXP#1 – ENG-ITA	84,65	84,65	85,61	84,73	84,43	84,60
EXP#2 – ITA-ENG	84,85	84,63	85,20	84,77	84,54	84,56
EXP#3 – ENG-ENG	85,23	85,29	85,23	85,10	84,86	84,89
EXP#4 – ITA-ITA	85,27	84,84	85,71	85,11	84,86	84,93

though in most of the experiments the cross-lingua recommendation approaches get worse results w.r.t. the mono-lingual ones, the difference in the predictive accuracy does not appear statistically significant. In general the bayesian approach fits better in scenarios where the number of items to be represented is not too high, and this can justify the application of the pre-processing steps required for building the MultiWordNet synset representation, while the distributional models, thanks to their simplicity and effectiveness, fit better in scenarios where real-time recommendations need to be provided.

5 Conclusions

This paper compared two approaches for providing cross-language recommendations. The key idea is to provide a bridge among different languages by exploiting a language-independent representation of documents and user profiles based on word meanings. Experiments were carried out in a movie recommendation scenario, and the main outcome is that the accuracy of cross-language recommmendations is comparable to that of classical (monolingual) content-based recommendations.

References

1. C. Musto, F. Narducci, P. Basile, P. Lops, M. de Gemmis, and G. Semeraro, "Cross-language information filtering: Word sense disambiguation vs. distributional models," in *AI*IA*, 2011, pp. 250–261.
2. M. Sahlgren, "The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces," Ph.D. dissertation, Stockholm University, Department of Linguistics, 2006.
3. C. Musto, "Enhanced vector space models for content-based recommender systems," in *Proceedings of the fourth ACM conference on Recommender systems*, ser. RecSys '10. New York, NY, USA: ACM, 2010, pp. 361–364.
4. E. Pianta, L. Bentivogli, and C. Girardi, "MultiwordNet: developing an aligned multilingual database," in *Proc. of the 1st Int. WordNet Conference, Mysore, India*, 2002, pp. 293–302.