Semantic Re-ranking in Ad-hoc Robust Retrieval

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro

Department of Computer Science, University of Bari "Aldo Moro"
Via Orabona, I-70125, Bari, Italy
{basilepp,acaputo,semeraro}@di.uniba.it

Abstract. This paper proposes an investigation about a re-ranking strategy presented at SIGIR 2010. In that work we describe a re-ranking strategy in which the output of a semantic based IR system is used to re-weigh documents by exploiting inter-document similarities computed on a vector space. The space is built using the Random Indexing technique. The effectiveness of the strategy has been evaluated in the context of the CLEF Ad-Hoc Robust-WSD Task, while in this paper we propose new experiments in the TREC Ad-Hoc Robust Track 2004.

1 Background and Motivation

A general approach to overcome the word ambiguity problem in IR involves the representation of documents by word meanings. Among the most investigated techniques are those that rely on WordNet[1] synsets through which groups of synonym words are uniquely identified and linked to each other by semantic relations. The Robust-WSD task at Cross Language Evaluation Forum (CLEF) [1] has shown that results improve when aggregation strategies are exploited. The method proposed in [6] describes a different approach to document aggregation based on a variation of the "*inter-document similarities*" [8] idea. The method combines two retrieval strategies that work at two different representation levels: *keyword* and *synset*. The ranked list of documents retrieved using the synset-based representation (synset list) is exploited to re-rank the list of documents retrieved using the keyword-based one (keyword list). The insight of this method is that documents in the keyword list with the highest number of similar documents in the synset list should climb in the result set. The approach tries to re-weigh documents in response to a query by promoting those documents with the highest number of *supporters*. In this context, a *supporter* is a document with content similar to the target one. Inter-document similarities is computed relying on the Random Index technique to build a vector space in which similar documents are represented close.

Let us denote by L_k and L_s the ranked lists of documents retrieved using keywords and synsets representation, respectively. The idea behind our re-ranking method is to give more evidence to the documents in L_k that are widely supported by similar documents occurring in both lists.
The method requires the following steps:

[1] A semantic lexicon for the English language.

1. For each document $d_i \in L_k$ compute the $supporters(d_i, \alpha)$, which is the set of α documents $\{d_1, ...d_\alpha\} \subset L_k$ with the highest inter-document similarity to d_i.
2. Get the $overlap\_supporters = \{d_j \in L_s : d_j \in supporters(d_i, \alpha)\}$ which is the set of documents occurring in both L_s and $supporters$.
3. Assign to d_i a new score $S(d_i)$ taking into account *supporting* documents computed in the step 2. Formally:

$$S(d_i) = \theta * S_{supporters} + (1 - \theta) * S_k(d_i) \quad (1)$$

where

$$S_{supporters} = \sum_{d_j \in overlap\_supporters} S_k(d_j) * S_s(d_j) \quad (2)$$

and $S_k(d_j)$ is the score of d_j in L_k, while $S_s(d_j)$ is the score of d_j in L_s, and θ is a free parameter used to smooth $S_{supporters}$, which denotes the scores combination of *supporting* documents.

2 The new setting

The proposed approach involves two retrieval strategies which work at two different representation levels: *keyword* and *synset*. The *synset* level requires the disambiguation of the whole collections: CLEF 2009 Ad-hoc Robust Task and TREC Ad-Hoc Robust Track 2004. The TREC collection counts 528,155 documents, while the CLEF 2009 collection consists of 166,717 documents disambiguated by the task organizers. The Word Sense Disambiguation (WSD) algorithm [7] used by the CLEF organizers is not available, for this reason we adopt our WSD strategy to disambiguate TREC documents. Our WSD method is based on [5]. It is important to underline that our WSD strategy obtains similar results wrt [7] in terms of precision when the two WSD algorithms are evaluated "in vitro". The WSD method used by the CLEF organizers obtains 0.578 as precision in SemEval-2007 All-Words Task, while our system obtains 0.59 in terms of precision using the dataset of Senseval-3 All-Words Task. The two datasets are not directly comparable, but the results give an idea of the effectiveness of both WSD strategies. To perform the WSD algorithm, several text processing operations are required such as tokenization, part-of-speech tagging and lemmatization. We adopt META [2], a text processing tool able to perform all the necessary natural language processing steps.

Moreover, to build the vector space in which similarity between documents is computed, we adopt a strategy based on Random Indexing using a modified version of Semantic Vectors package [9] able to work with large collections as TREC. Our modified version works on computational aspects to improve performance related to space and time.

Finally, we built a retrieval system [3] based on Lucene and the Okapi BM25 model for both levels of representation: keyword and synset. Stemming and stop word removal are applied to the keyword-based representation of documents and

topics. To evaluate the performance we executed several runs using the topics provided in each track. In detail, the CLEF 2009 collection has 160 topics, while the TREC collection has 259 topics. We used TITLE and DESCRIPTION topic fields adopting two different boosting factors (TITLE=4, DESCRIPTION=1) to highlight terms in the TITLE.

More details about the adopted IR system are in [3], while the Random Indexing strategy exploited in this work is thoroughly described in [4].

3 Evaluation and Remarks

The goal of the evaluation is to prove that the re-ranking method proposed in [6] is able to obtain good performance when both a different collection and a different WSD algorithm are involved.

The proposed approach requires disambiguated documents. As well known, WSD is a time consuming task. The disambiguation of the whole TREC collection has required about 6 days using a Linux-based PC with Intel Core2Quad processor having 6 GB of RAM, while, in CLEF collection, we rely on disambiguated documents provided by the organizers. Comparing different WSD algorithms is out of the scope of this work, while we want to evaluate the contribution of synset-based document representation in our re-ranking approach. We plan to perform CLEF disambiguation using our WSD method in future evaluations.

The evaluation was performed using the MAP and GMAP measures. Table 1 summarizes the main results. Foremost, we evaluated each system alone (*Keyword* and *Synset*). *Keyword* was used as baseline of the evaluation. Then, we evaluated an aggregation strategy, *CombSUM*. In particular we adopted a modified version of that strategy to assign different weights to each list during aggregation. Finally, the result of the proposed method has been denoted by *ReRank*. After a tuning step, we set the weights for L_k and L_s to 0.8 and 0.2, respectively. Moreover, we tested several values of $\theta \in \{0.1, 0.2, \ldots, 0.5\}$ and $\alpha \in \{5, 10, 20, 40\}$. Table 1 reports only the best results and the involved parameters α and θ.

Table 1. Experimental Results

Collection	**Exp**	**MAP**	**GMAP**
CLEF	*Keyword*	*0.4205*	*0.1900*
	Synset	0.3201	0.1242
	CombSUM	0.4252	0.1972
	ReRank($\theta = 0.3$ $\alpha = 20$)	**0.4332**	**0.1989**
TREC	*Keyword*	*0.2745*	*0.1692*
	Synset	0.1360	0.0381
	CombSUM	0.2729	0.1739
	ReRank($\theta = 0.2$ $\alpha = 10$)	**0.2784**	**0.1754**

The *ReRank* method achieves the best results in terms of MAP and GMAP in both the collections. These improvements are significant with respect to the baseline *Keyword*; we validated our experiments using the non parametric Randomization test, setting ρ to 5%. Results confirm our hypothesis: the ranking provided by synsets (L_s) contributes significantly to the final document score. Moreover, it is important to underline that, despite using a different WSD algorithm to disambiguate the TREC collection, the performance are not affected.

References

1. Agirre, E., Di Nunzio, G.M., Mandl, T., Otegi, A.: CLEF 2009 Ad Hoc Track Overview: Robust-WSD Task. In: Peters, C., Di Nunzio, G.M., Kurimo, M., Mostefa, D., Peñas, A., Roda, G. (eds.) Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers. pp. 36–49. Lecture Notes in Computer Science, Springer (2009)
2. Basile, P., de Gemmis, M., Gentile, A., Iaquinta, L., Lops, P., Semeraro, G.: META-MultilanguagE Text Analyzer. In: Proceedings of the Language and Speech Technnology Conference-LangTech. pp. 28–29 (2008)
3. Basile, P., Caputo, A., Semeraro, G.: UNIBA-SENSE @ CLEF 2009: Robust WSD Task. In: Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers - CLEF (1). Lecture Notes in Computer Science, vol. 6241, pp. 150–157. Springer (2010)
4. Basile, P., Caputo, A., Semeraro, G.: Integrating Sense Discrimination in a Semantic Information Retrieval System. In: Soro, A., Vargiu, E., Armano, G., Paddeu, G. (eds.) Information Retrieval and Mining in Distributed Environments. Studies in Computational Intelligence, vol. 324, pp. 249–256. Springer (2011)
5. Basile, P., de Gemmis, M., Lops, P., Semeraro, G.: Combining knowledge-based methods and supervised learning for effective italian word sense disambiguation. In: Proceedings of the 2008 Conference on Semantics in Text Processing. pp. 5–16. STEP '08, Association for Computational Linguistics, Morristown, NJ, USA (2008)
6. Caputo, A., Basile, P., Semeraro, G.: From fusion to re-ranking: a semantic approach. In: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval. pp. 815–816. SIGIR '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1835449.1835630
7. Chan, Y.S., Ng, H.T., Zhong, Z.: Nus-pt: Exploiting parallel texts for word sense disambiguation in the english all-words tasks. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007). pp. 253–256. Association for Computational Linguistics, Prague, Czech Republic (June 2007), http://www.aclweb.org/anthology/S/S07/S07-1054
8. Kozorovitzky, A.K., Kurland, O.: From "identical" to "similar": Fusing retrieved lists based on inter-document similarities. In: Azzopardi, L., Kazai, G., Robertson, S.E., Rüger, S.M., Shokouhi, M., Song, D., Yilmaz, E. (eds.) ICTIR. LNCS, vol. 5766, pp. 212–223. Springer (2009)
9. Widdows, D., Ferraro, K.: Semantic Vectors: A Scalable Open Source Package and Online Technology Management Application. In: Proc. of the 6th Int. Conf. on Language Resources and Evaluation (LREC 2008) (2008)