UNIBA-SENSE @ CLEF 2009: Robust WSD Task

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro

Department of Computer Science
University of Bari
70126 Bari, Italy
{basilepp,acaputo,semeraro}@di.uniba.it

Abstract. This paper presents the participation of the semantic N-levels search engine SENSE at the CLEF 2009 Ad Hoc Robust-WSD Task. Our aim is to demonstrate that the combination of the N-levels model and WSD can improve the retrieval performance even when an effective retrieval model is adopted. To reach this aim, we worked on two different strategies. On one hand a model, based on Okapi BM25, was adopted at each level. On the other hand, we integrated a local relevance feedback technique, called Local Context Analysis, in both indexing levels of the system (keyword and word meaning). The hypothesis that Local Context Analysis can be effective even when it works on word meanings coming from a WSD algorithm is supported by experimental results. In monolingual task MAP increased of about 2% exploiting disambiguation, while GMAP increased from 4% to 9% when we used WSD in both mono- and bi- lingual tasks.

1 Introduction

In this paper we present our participation at the CLEF 2009 Ad Hoc Robust-WSD Task. Our retrieval system is based on SENSE [2], a semantic search engine which implements the N-levels model. For the CLEF 2009 experiments, the following levels were exploited:

Keyword level - the entry level in which a document is represented by the words occurring in the text.

Word meaning level - at this level a document is represented through *synsets* obtained by WordNet, a semantic lexicon for the English language. A synset is a set of synonym words (with the same meaning).

SENSE is able to manage different models for each level. In CLEF 2008 campaign we adopted the standard Vector Space Model implemented in Lucene for both the keyword and the word meaning level. For CLEF 2009 our goal is to improve the overall retrieval performance by adopting a more powerful model, called Okapi BM25, and a pseudo-relevance feedback mechanism based on Local Context Analysis.

C. Peters et al. (Eds.): CLEF 2009 Workshop, Part I, LNCS 6241, pp. 150–157, 2010.

The rest of the paper is structured as follows: the indexing step adopted in SENSE is described in Section 2, while Section 3 presents the searching step with details about Local Context Analysis strategy. The details of the system setup for the CLEF competition are provided in Section 4. Finally, the experiments are described in Section 5. Conclusions and future work close the paper.

2 Indexing

In CLEF Ad-Hoc WSD Robust track, documents and queries are provided in XML format. In order to index the documents and read the queries we developed an XML parser using the XMLBeans[1] tool. As SENSE supports an indefinite numbers of levels, we developed a flexible indexing mechanism. Hence, we produced an intermediate data format which contains all the data necessary to the N-levels model. For each token this format provides a set of features needful to build each level. In CLEF, for the keyword level the stemming of the word[2] is provided, for the meaning one we provided the list of all possible meanings with the corresponding score. During the indexing we performed several text operations. One is stop words elimination. We built two different stop words lists, one for documents and one for queries. In this way we removed irrelevant words from queries. Moreover, before storing each token in a document, we replaced all occurrences of not alphanumeric characters with a single underscore character "_". This text normalization operation was also performed for queries during the search process. With respect to the meaning level, we index for each token only the WordNet synset with the highest score. For each document a bag of synsets is built. Consequently, the vocabulary at this level is the set of distinct synsets recognized in the collection by the WSD procedure.

3 Searching

The local similarity functions for both the meaning and the keyword levels are computed using a modified version of the Lucene default document score, that implements the Okapi BM25 [7]. In order to implement BM25 in SENSE we exploited the technique described in [5]. In particular, we adopted the BM25-based strategy which takes into account multi-field documents. Indeed, in our collection each document is represented by two fields: HEADLINE and TEXT. The multi-field representation reflects the XML structure of documents provided by the organizers. Table 1 shows the BM25 parameters used in SENSE, where avl is the average length for each field. b is a constant related to the field length, similar to b constant in classical BM25 formula, k_1 is a free parameter, while $boost$ is the boost factor applied to that field. All parameters were tuned on the training data and are different for keyword and meaning level.

[1] `http://xmlbeans.apache.org/`

[2] Stemming is performed by the Snowball library.

Table 1. BM25 parameters used in SENSE

Level	Field	k_1	N	avl	b	$boost$
Keyword	$HEADLINE$	3.25	166,726	7.96	0.70	2.00
	$TEXT$	3.25	166,726	295.05	0.70	1.00
Word Meaning	$HEADLINE$	3.50	166,726	5.94	0.70	2.00
	$TEXT$	3.50	166,726	230.54	0.70	1.00

For the meaning level, both query and document vectors contain synsets instead of keywords.

In SENSE each level produces a list of documents ranked according to the similarity function defined for that level (*local similarity function*). Since the ultimate goal is to obtain a *single* list of documents ranked in decreasing order of relevance, a *global ranking function* is needed to merge all the result lists that come from each level. This function is independent of both the number of levels and the specific local scoring and similarity functions because it takes as input n ranked lists of documents and produces a unique merged list of the most relevant documents.

The aggregation of lists in a single one requires two steps: the first one produces the n normalized lists and the second one merges the n lists in a single one. The two steps are thoroughly described in [2]. In CLEF we adopt Z-Score normalization and CombSUM [3,4] as score normalization and rank aggregation function, respectively. Each level can be combined using a different weighting factor in order to give different relevance to each level.

3.1 Query Expansion and Term Reweighting

We extended the SENSE architecture by integrating a query expansion module, as well as a technique for term reweighting. We adopted the Local Context Analysis (LCA) [8], a strategy that proved its effectiveness on several test collections. LCA is a *local* technique as it analyzes only the first top-ranked documents that are assumed to be the relevant ones. LCA relies on the hypothesis that terms frequently occurring in the top-ranked documents frequently co-occur with all query terms in those documents too. We employed the LCA for both levels exploited in our experiments: keyword and word meaning. The underlying idea is that the LCA hypothesis could also be applied to the word meaning level, in which meanings are involved instead of terms. Therefore, we extended the original measure of co-occurrence degree in order to weigh a generic feature (keyword or word meaning) rather than just a term. According to the original formula, we define the following function:

$$codegree(f, q_i) = \frac{log_{10}(co(f, q_i) + 1) * idf(f)}{log_{10}(n)} \tag{1}$$

codegree measures the degree of co-occurrence between the feature f and the query feature q_i ($co(f, q_i)$), but it takes also into account the frequency of f in

the whole collection ($idf(f)$) and normalizes this value with respect to n, the number of documents in the top-ranked set.

$$co(f, q_i) = \sum_{d \in S} tf(f, d) * tf(q_i, d) \tag{2}$$

$$idf(f) = min(1.0, \frac{log_{10} \frac{N}{N_f}}{5.0}) \tag{3}$$

where $tf(f, d)$ and $tf(q_i, d)$ are the frequency in d of f and q_i respectively, S is the set of top-ranked documents, N is the number of documents in the collection and N_f is the number of documents containing the feature f. For each level, we retrieve the n top-ranked documents for a query q by computing a function lca for each feature in the results set, as follows:

$$lca(f, q) = \prod_{q_i \in q} (\delta + codegree(f, q_i))^{idf(q_i)} \tag{4}$$

This formula is used to rank the list of features that occur in the top-ranked documents; δ is a smoothing factor and the power is used to raise the impact of rare features. A new query q' is created by adding the k top ranked features to the original query, where each feature is weighed using the lca value. Hence, the new query is re-executed to obtain the final list of ranked documents for each level. Differently from the original work, we applied LCA to the top ranked documents rather than passages[3]. Moreover, no tuning is performed over the collection to set the parameters. For the CLEF experiments, we decided to get the first ten top-ranked documents and to expand the query using the first ten ranked features. Finally, we set up the smoothing factor to 0.1 in order to boost those concepts that co-occur with the highest number of query features.

4 System Setup

We exploited the SENSE framework to build our IR system for the CLEF evaluation. We used two different levels: keyword (using word stems) and word meaning (using WordNet synsets). All SENSE components involved in the experiments are implemented in Java using the version 2.3.2 of Lucene API. Experiments were run on an Intel Core 2 Quad processor at 2.6 GHz, operating in 64 bit mode, running Linux (UBUNTU 9.04), with 4 GB of main memory.

Following CLEF guidelines, we performed two different tracks of experiments: Ad Hoc Robust-WSD monolingual and bilingual. Each track required two different evaluations: with and without synsets. We exploited several combinations between levels and the query relevance feedback method, especially for the meaning level. All query building methods are automatic and do not require manual operations. Moreover, we used different boosting factors for each topic field and gave more importance to the terms in the fields TITLE and DESCRIPTION.

[3] In the original work, passages are parts of document text of about 300 words.

Table 2 shows all performed runs. In particular, the column *N-Levels* reports the different weighting factors used to merge each result list. The columns *WSD* and *LCA* denote in which runs the word meaning level and pseudo-relevance feedback technique were involved. Details about boosting factors assigned to each query field are reported in the *Boost* column (T=Title, D=Description, N=Narrative). More details on the track are reported in the track overview paper [1]. For all the runs we removed the stop words from both the index and the topics.

Table 2. Overview of experiments

RUN	**Mono**	**Bi**	**N-levels**		**WSD**	**LCA**	**Boost**		
			Key	Syn			T	D	N
unibaKTD	X	-	-	-	-	-	8	1	-
unibaKTDN	X	-	-	-	-	-	8	2	1
unibaKRF	X	-	-	-	-	X	8	2	1
unibaWsdTD	X	-	-	-	X	-	8	1	-
unibaWsdTDN	X	-	-	-	X	-	8	2	1
unibaWsdNL0802	X	-	0.8	0.2	X	-	8	2	1
unibaWsdNL0901	X	-	0.9	0.1	X	-	8	2	1
unibaKeySynRF	X	-	0.8	0.2	X	X	8	2	1
unibaCrossTD	-	X	-	-	-	-	8	1	-
unibaCrossTDN	-	X	-	-	-	-	8	2	1
unibaCrossKeyRF	-	X	-	-	-	X	8	2	1
unibaCrossWsdTD	-	X	-	-	X	-	8	1	-
unibaCrossWsdTDN	-	X	-	-	X	-	8	2	1
unibaCrossWsdNL0802	-	X	0.8	0.2	X	-	8	2	1
unibaCrossWsdNL0901	-	X	0.9	0.1	X	-	8	2	1
unibaCrossKeySynRF	-	X	0.8	0.2	X	X	8	2	1

5 Experimental Session

The experiments were carried out on the CLEF Ad Hoc WSD-Robust dataset derived from the English CLEF data, which comprises corpora from "Los Angeles Times" and "Glasgow Herald", amounting to 166,726 documents and 160 topics in English and Spanish. The relevance judgments were taken from CLEF. Our evaluation has two main goals:

1. to prove that the combination of two levels outperforms a single level. Specifically, we want to investigate whether the combination of keyword and meaning levels turns out to be more effective than the keyword level alone, and how the performance varies.
2. to prove that Local Context Analysis improves the system performance. We exploit pseudo-relevance feedback techniques in both levels, keyword and meaning. Our aim is to demonstrate the effectiveness of pseudo-relevance feedback when it is applied not only to a keyword but to a word meaning representation, too.

To measure retrieval performance, we adopted the Mean-Average-Precision (MAP) and the Geometric-Mean-Average-Precision (GMAP) calculated by CLEF organizers using the DIRECT system on the basis of the first 1,000 retrieved items per request. Table 2 summarizes the description of system setup for each run, while Table 3 shows the results of five metrics (Mean-Average-Precision, Geometric-Mean-Average-Precision, R-precision, P@5 and P@10 are the precision after 5 and 10 documents retrieved respectively) for each run.

Table 3. Results of the performed experiments

Run	***MAP***	***GMAP***	***R-PREC***	***P@5***	***P@10***
unibaKTD	.3962	.1684	.3940	.4563	.3888
unibaKTDN	.4150	.1744	.4082	.4713	.4019
unibaKRF	.4250	.1793	.4128	.4825	.4150
unibaWsdTD	.2930	.1010	.2854	.3838	.3256
unibaWsdTDN	.3238	.1234	.3077	.4038	.3544
unibaWsdNL0802	.4218	.1893	.4032	.4838	.4081
unibaWsdNL0901	.4222	.1864	.4019	.4750	.4088
unibaKeySynRF	**.4346**	**.1960**	**.4153**	**.4975**	**.4188**
unibaCrossTD	.3414	.1131	.3389	.4013	.3419
unibaCrossTDN	.3731	.1281	.3700	.4363	.3713
unibaCrossKeyRF	**.3809**	.1311	**.3755**	.4413	.3794
unibaCrossWsdTD	.0925	.0024	.1029	.1188	.1081
unibaCrossWsdTDN	.0960	.0050	.1029	.1425	.1188
unibaCrossWsdNL0802	.3675	.1349	.3655	.4455	.3750
unibaCrossWsdNL0901	.3731	.1339	.3635	.4475	.3769
unibaCrossKeySynRF	.3753	**.1382**	.3709	**.4513**	**.3850**

Analyzing the mono-lingual task, the word meaning level used alone is not enough to reach good performance (*unibaWsdTD*, *unibaWsdTDN*). However, an increase of 1,7% in MAP is obtained when word meanings are exploited in the N-levels model (*unibaWsdNL0901*) with respect to the keyword level alone (*unibaKTDN*). Looking at the N-levels results, we can notice the impact of word meanings on GMAP. In fact, as the weight of the word meaning level raises the MAP decreases while the GMAP increases. In both runs, with or without WSD, the adoption of pseudo-relevance feedback techniques increases the MAP: 2.9% with WSD (*unibaKeySynRF* vs. *unibaWsdNL0901*) and 2.4% without WSD (*unibaKRF* vs. *unibaKTDN*). Finally, LCA combined to WSD (*unibaKeySynRF*) works better than LCA without WSD (*unibaKRF*) with an increment in all measures (+2.3% MAP, +9.3% GMAP, +0.6% R-prec, +3.1% P@5, +0.9% P@10) and, in general, it shows the best results.

In the bilingual task, queries are disambiguated using the first sense heuristics. This has an impact on the use of synsets in the query processing and pseudo-relevance feedback steps. Word meaning level performance is very bad. Moreover, runs without WSD generally outperform those with WSD, with an increment of

1.5% in MAP (*unibaCrossKeyRF* vs. *unibaCrossKeySynRF*). As LCA has shown to be helpful, with or without WSD, a higher increment is obtained without WSD: 2.09% in MAP (*unibaCrossKeyRF* vs. *unibaCrossTDN*). Nevertheless, also in the bilingual task WSD has improved the GMAP with an increment of 5.42% (*unibaCrossKeySynRF* vs. *unibaCrossKeyRF*). The increment in GMAP emphasizes the improvement for poorly performing (low precision) topics. This suggests that WSD is especially useful for those topics with low scores in average precision.

However, there are poorly performant queries (Average Precision<0.1). A query-by-query analysis suggested the reasons of the system failures. Hard topics can be split in two macro-categories. On one hand, there are complex topics which require precise information similar to a question answering task. For example, requests for game winners[4], name of countries/cities, events in specific periods of time (Topic 171, 172, 258, 310, 313, 345, 346). On the other hand, there are topics which specify non-relevance constraints (Topics 160, 305, 309[5], 322). Obviously, in the Bag-of-Word representation this kind of information is lost.

We validate our experiments (with respect to MAP metric) using both the parametric Student paired t-test and the non parametric Randomization test, as suggested in [6] ($\alpha = 5$ %). For the Randomization test we use a Perl script supplied by the authors. Both tests give similar results: all improvements are significant with only two exceptions. In both, mono/bi-lingual tasks without WSD, the differences obtained using the *NARRATIVE* field of the query are not significant. We achieve significant improvements in WSD task using the *NARRATIVE* field: this field is helpful to recognize the proper word meanings belonging to the query. In the monolingual task, the improvement obtained using the combination of keyword and word meaning levels are generally significant, except for *unibaWsdNL0802*.

6 Conclusion and Future Work

We have described and tested SENSE, a semantic N-levels IR system which manages documents indexed at multiple separate levels: keywords and meanings. The system is able to combine keyword search with semantic information provided by the other indexing levels.

With respect to the last participation of SENSE to CLEF 2008, we introduce in this edition new features in order to improve the overall retrieval performance. In particular, we adopt the Okapi BM25 model for both keyword and word meaning levels. Moreover, we propose a pseudo-relevance feedback strategy based on

[4] Who won a tennis Grand Slam Tournament event in 1995?

[5] What are the dangers to health of illegal hard drugs, such as heroin and cocaine, as opposed to soft drugs? Relevant documents must provide information on the medical risks involved in the illegal use of and dependence on hard drugs. Information on problems resulting from the use of "soft" drugs is not relevant.

Local Context Analysis. This strategy is applied to keyword and word meaning levels.

The results of the evaluation prove that the combination of keyword and word meaning can improve the retrieval performance. Only in bilingual task the combination of levels is outperformed by the only keyword level. Probably this is due to WSD technique adopted for Spanish topics. In particular, no WSD algorithms for Spanish are available and the organizers assign the first synset in Spanish-WordNet to each keyword in a topic. Moreover, the results prove that the pseudo-relevance feedback based on Local Context Analysis improves the IR performance.

References

1. Agirre, E., Di Nunzio, G.M., Ferro, N., Mandl, T., Otegi, A.: CLEF 2009 Ad Hoc Track Overview: Robust-WSD Task. In: Peters, C., Nunzio, G.D., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., Roda, G. (eds.) Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments. LNCS (LNAI), Springer, Heidelberg (2009)
2. Basile, P., Caputo, A., Gentile, A.L., Degemmis, M., Lops, P., Semeraro, G.: Enhancing Semantic Search using N-Levels Document Representation. In: Bloehdorn, S., Grobelnik, M., Mika, P., Tran, D.T. (eds.) SemSearch. CEUR Workshop Proceedings, vol. 334, pp. 29–43. CEUR-WS.org. (2008)
3. Fox, E.A., Shaw, J.A.: Combination of Multiple Searches. In: TREC, pp. 243–252 (1993)
4. Lee, J.H.: Analyses of multiple evidence combination. In: SIGIR, pp. 267–276. ACM, New York (1997)
5. Robertson, S., Zaragoza, H., Taylor, M.: Simple BM25 extension to multiple weighted fields. In: CIKM 2004: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, pp. 42–49. ACM, New York (2004)
6. Smucker, M.D., Allan, J., Carterette, B.: A comparison of statistical significance tests for information retrieval evaluation. In: CIKM 2007: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, pp. 623–632. ACM, New York (2007)
7. Sparck Jones, K., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments. Information Processing Management 36(6), 779–808, 809–840 2000)
8. Xu, J., Croft, W.B.: Improving the effectiveness of information retrieval with local context analysis. ACM Trans. Inf. Syst. 18(1), 79–112 (2000)