UNIBA-SENSE @ CLEF 2009: Robust WSD task

Pierpaolo Basile, Annalina Caputo, Giovanni Semeraro
Dept. of Computer Science - University of Bari
Via E. Orabona, 4 - 70125 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. During the participation at the same task of CLEF 2008, SENSE showed that WSD can be helpful to improve retrieval, even though the overall performance was not exciting mainly due to the adoption of a pure Vector Space Model with no heuristics. In this edition, 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 new model, based on Okapi BM25, was adopted at each level. Moreover, we improved the word stemming algorithm and we normalized words removing some characters that made more evident the word mismatch problem. The use of these simple heuristics allowed us to increases of 106% the MAP value, compared to our best result obtained at CLEF 2008. 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 Mono-lingual task MAP increased of about 2% exploiting disambiguation, while GMAP increased from 4% to 9% when we used WSD in both Mono- and Cross- lingual tasks.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Management**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Information Retrieval, Word Sense Disambiguation

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.

The main motivation behind our model is that the presence of multiple meanings for one word (polysemy), together with synonymy (occurring when different words have the same meaning), negatively affects the retrieval performance. Generally, the result is that, due to synonymy, relevant documents can be missed if they do not contain the exact query keywords, while wrong documents are deemed as relevant due to polysemy. These problems call for alternative methods that work not only at the lexical level of the documents, but also at the *meaning* level.

Therefore, in our interpretation semantic information could be captured from a text by looking at *word meanings*, as they are described in a reference dictionary (e.g. WordNet [5]). SENSE is an IR system which manages documents indexed at multiple separate levels: keywords and senses (word meanings). The system is able to combine keyword search with semantic information provided by the word meaning level.

The main idea underlying the definition of an open framework to model different semantic aspects (or levels) pertaining document content is that there are several ways to describe the semantics of a document. Each semantic facet needs specific techniques and ad-hoc similarity functions. To address this problem we propose a framework in which a different IR model is defined for each level in the document representation. Each level corresponds to a *logical view* that aims to describe one of the possible semantic spaces in which documents can be represented. The adoption of different levels is intended to guarantee acceptable system performance even when not all semantic representations are available for a document.

We suppose that the keyword level is always present and, when other levels are available too, they are exploited to enhance retrieval capabilities. Furthermore, our framework allows to associate each level with the appropriate representation and similarity measure. The following semantic levels are currently available in the framework:

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). Word Sense Disambiguation (WSD) algorithms are adopted to assign synsets to words.

SENSE is able to manage different models for each level. In CLEF 2008 edition 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 adopting a more powerful model, called Okapi BM25, and the introduction of a pseudo-relevance feedback mechanism based on Local Context Analysis.

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. Moreover, Section 3 contains details about the Okapi BM25 model implemented in SENSE and the 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. Moreover, 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 the case in point, 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.

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

[2] Stemming is performed by Snowball library.

An intermediate format is necessary because SENSE supports an indefinite number of levels, not restricted to keyword and meaning ones as in CLEF Ad-Hoc WSD Robust track. For that reason we developed a flexible indexing mechanism able to support further levels.

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, such as: *find, report, information, provide, describe, include, discuss, specific, interest, concern.* 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 is also performed for queries during the search process. In that way the match between documents and query is not compromised.

As regards 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. Hence, features at the word meaning level are *synsets* obtained from WordNet, a semantic lexicon for the English language. 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 model described in Section 3.1. 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 Okapi BM25 model in SENSE

We employed Lucene API to build the SENSE search engine. An important change we made concerns the adoption of a new model, based on Okapi BM25 [7], to implement a new weighting scheme and local similarity function at each level. In order to implement BM25 in SENSE we exploited the technique described in [6]. 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.

First of all, in the multi-field representation the weight of each term is computed taking into account the aggregate amount of the term weights for all fields, as follows:

$$weight(t,d) = \sum_{c \in d} \frac{occurs^d_{t,c} * boost_c}{((1-b_c) + b_c * \frac{l_c}{avl_c})} \quad (1)$$

where $occurs^d_{t,c}$ is the occurrence of the term t in the field c, l_c is the field length and avl_c is the average length for the field c. b_c is a constant related to the field length, similar to b constant in classical BM25 formula, while $boost_c$ is the boost factor applied to field c.

Then, the similarity between query and document is computed exploiting the accumulated weight for each term t that occurs both in the query q and in the document d.

$$R(q,d) = \sum_{t \in q} idf(t) * \frac{weight(t,d)}{k_1 + weight(t,d)} \quad (2)$$

Inverse document frequency is computed according to the classical BM25 model:

$$idf(t) = log \frac{N - df(t) + 0.5}{df(t) + 0.5} \quad (3)$$

where N is the number of documents in the collection and $df(t)$ is the number of documents where the term t appears. Table 1 shows the BM25 parameters used in SENSE. Parameters are different for keyword level ($HEADLINE_k$, $TEXT_k$) and meaning level ($HEADLINE_s$, $TEXT_s$)

Field	k_1	N	avl_c	b_c	$boost_c$
$HEADLINE_k$	3.25	166,726	7.96	0.70	2.00
$TEXT_k$	3.25	166,726	295.05	0.70	1.00
$HEADLINE_s$	3.50	166,726	5.94	0.70	2.00
$TEXT_s$	3.50	166,726	230.54	0.70	1.00

Table 1: BM25 parameters used in SENSE.

3.2 Query Expansion and Term Reweighting

During 2008 edition of CLEF, SENSE showed promising results, although its overall performance was not exciting. This deterred us from using query expansion techniques. Indeed, a preliminary condition to avoid the query drift problem, an intrinsic problem for automatic query expansion methods, is to have a system with good precision in the first retrieved documents. The performance improvement expected as a consequence of the adoption of BM25 weighting scheme, made it possible the use of these techniques into our system. 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* techniques 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:

$$codegree(f, q_i) = \frac{log_{10}(co(f, q_i) + 1) * idf(f)}{log_{10}(n)} \quad (4)$$

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.

$$co(f, q_i) = \sum_{d \in S} tf(f, d) * tf(q_i, d) \quad (5)$$

$$idf(f) = min(1.0, \frac{log_{10} \frac{N}{N_f}}{5.0}) \quad (6)$$

where $tf(f, d)$ and $tf(q_i, d)$ are the frequency of f and q_i in d respectively, S is the set of top-ranked documents, N is the number of documents in the collections 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:

$$lca(f, q) = \prod_{q_i \in q} (\delta + codegree(f, q_i))^{idf(q_i)} \quad (7)$$

This formula is used to rank the list of features that occur in the top-ranked documents; δ is a smoothing factor and the exponent is used to give an higher impact to rare features. A new query q' is created by adding the k top ranked features to the original query, 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 Mono-language and Cross-language. 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. More details on the track are reported in the track overview paper [1].

In particular for the Ad-Hoc Mono-language track we performed the following runs:

1. **unibaKTD**: the query is built using word stems in the fields TITLE and DESCRIPTION of the topics. All query terms are joined adopting the OR boolean operator. The terms in the TITLE field are boosted using a factor 8.

2. **unibaKTDN**: similar to the previous run, but in this case we add the NARRATIVE field and we adopt different term boosting values: 8 for TITLE, 1 for DESCRIPTION and 1 for NARRATIVE.

3. **unibaKRF**: we used the query produced in unibaKTDN adding a pseudo-relevance feedback mechanism which implements LCA.

4. **unibaWsdTD**: in this experiment we exploited only the word meaning level. The query is built using the synset with the highest score for each token. The synset score is also used to give a weight to the synset into the query. Synset boosting values are: 8 for TITLE and 2 for DESCRIPTION.

5. **unibaWsdTDN**: similar to the previous run, but in this case we add the NARRATIVE field. Synset boosting values are: 8 for TITLE, 2 for DESCRIPTION and 1 for NARRATIVE.

6. **unibaWsdNL0802**: in this experiment we exploit the N-level architecture of SENSE. For the keyword level we adopt the query method described in unibaKTDN and for the word meaning level that in unibaWsdTDN. The two levels are combined using a factor of 0.8 for keyword and 0.2 for meaning.

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

7. **unibaWsdNL0901**: similar to the previous run, but using different combination factors: 0.9 for keyword and 0.1 for meaning.

8. **unibaKeySynRF**: in this experiment we exploit both the N-level architecture of SENSE and LCA. For the keyword level we adopt the query method described in unibaKRF and for the word meaning level the unibaWsdTDN applying pseudo-relevance feedback. The two levels are combined using a factor of 0.8 for keyword and 0.2 for meaning.

For the Ad-Hoc Cross-language track we performed the following runs:

1. **unibaCrossTD**: the query is built using word stems in the TITLE and DESCRIPTION fields of the topics. In the Cross-language track the topics are in Spanish, thus a translation of terms in English is required. We adopt Google Translation API to translate queries from Spanish to English. Term boosting values are: 8 for TITLE and 1 for DESCRIPTION.

2. **unibaCrossTDN**: similar to the previous run, adding the NARRATIVE field. Term boosting values are: 8 for TITLE, 1 for DESCRIPTION and 1 for NARRATIVE.

3. **unibaCrossKeyRF**: queries are built using the method described in unibaCrossTDN and pseudo-relevance feedback is applied using LCA.

4. **unibaCrossWsdTD**: the query is built using for each token the synset with the highest score. Synset boosting values are: 8 for TITLE and 2 for DESCRIPTION. It is important to note that in this case the synset with the highest score is always the first synset in Spanish WordNet because word sense disambiguation is not applied to Spanish topics.

5. **unibaCrossWsdTDN**: similar to the previous run, but in this case we add the NARRATIVE field.

6. **unibaCrossWsdNL0802**: in this experiment we exploit the N-level architecture of SENSE. For the keyword level we adopt the query method described in unibaCrossTDN and for the word meaning level the unibaCrossWsdTDN. The two levels are combined using a factor 0.8 for keyword and a factor 0.2 for meaning.

7. **unibaCrossWsdNL0901**: similar to the previous run, but using different combination factors: 0.9 for keyword and 0.1 for meaning.

8. **unibaCrossKeySynRF**: in this experiment we exploit both the N-level architecture of SENSE and relevance feedback in the context of cross-language retrieval. For the keyword level we adopt the query method described in unibaCrossKeyRF and for the word meaning level the unibaCrossWsdTDN applying pseudo-relevance feedback. The two levels are combined using a factor 0.8 for keyword and 0.2 for meaning.

For all the runs we removed the stop words from both the index and the topics.

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 R-precision where R is set to 5 and 10 respectively) for each run.

Run	**MONO**	**CROSS**	**N-levels**	**WSD**	**LCA**
unibaKTD	X	-	-	-	-
unibaKTDN	X	-	-	-	-
unibaKRF	X	-	-	-	X
unibaWsdTD	X	-	-	X	-
unibaWsdTDN	X	-	-	X	-
unibaWsdNL0802	X	-	X	X	-
unibaWsdNL0901	X	-	X	X	-
unibaKeySynRF	X	-	X	X	X
unibaCrossTD	-	X	-	-	-
unibaCrossTDN	-	X	-	-	-
unibaCrossKeyRF	-	X	-	-	X
unibaCrossWsdTD	-	X	-	X	-
unibaCrossWsdTDN	-	X	-	X	-
unibaCrossWsdNL0802	-	X	X	X	-
unibaCrossWsdNL0901	-	X	X	X	-
unibaCrossKeySynRF	-	X	X	X	X

Table 2: Overview of 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**

Table 3: Results of the performed experiments

Though a comparison with the CLEF 2008 results is not reported, we have to point out that the worst run without WSD (*unibaKTD*) registered a rise of 106% in MAP when compared to the best CLEF 2008 run. Analyzing the mono-lingual task, as expected 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 as 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 bilingual task, queries are disambiguated using the first sense heuristics. This clearly has an impact on the use of synsets in the query processing and pseudo-relevance feedback steps. Performance of the word meaning level are 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.

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.

Respect to the last participation of SENSE to CLEF, 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 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 cross-lingual 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.

As future research we plan to improve the pseudo-relevance feedback strategy. We can achieve this goal applying the Local Context Analysis to the merged list of documents provided by SENSE. Currently, the Local Context Analysis is applied separately to the top ranked documents present in each level: keyword and word meaning.

References

[1] E. Agirre, G. M. Di Nunzio, N. Ferro, T. Mandl, and C. Peters. CLEF 2009: Ad Hoc Track Overview. In *CLEF 2009 Workshop: Working notes*, 2009.

[2] P. Basile, A. Caputo, A. L. Gentile, M. Degemmis, P. Lops, and G. Semeraro. Enhancing semantic search using n-levels document representation. In S. Bloehdorn, M. Grobelnik, P. Mika, and D. T. Tran, editors, *SemSearch*, volume 334 of *CEUR Workshop Proceedings*, pages 29–43. CEUR-WS.org, 2008.

[3] E. A. Fox and J. A. Shaw. Combination of multiple searches. In *TREC*, pages 243–252, 1993.

[4] J.-H. Lee. Analyses of multiple evidence combination. In *SIGIR*, pages 267–276. ACM, 1997.

[5] G. A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[6] S. Robertson, H. Zaragoza, and M. Taylor. Simple bm25 extension to multiple weighted fields. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 42–49, New York, NY, USA, 2004. ACM.

[7] K. Sparck Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments. *Information Processing Management*, 36(6):779–808, 809–840, 2000.

[8] J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1):79–112, 2000.