

Sense-based Blind Relevance Feedback for Question Answering

Matteo Negri

ITC-irst, Centro per la Ricerca Scientifica e Tecnologica
Via Sommarive 18, 38050
Povo (TN), Italy
negri@itc.it

ABSTRACT

This paper addresses the problem of enhancing document retrieval under the specific restrictions posed by the Question Answering scenario. In particular, given an input question, we aim at defining a reliable method for expanding its keywords with semantic information extracted from WORDNET (*e. g.* synonyms or hypernyms). This is a challenging task, since it is intrinsically dependent on high quality disambiguation of natural language questions which so far has been out of the reach of state-of-the-art Word Sense Disambiguation tools. The proposed solution relies on a two-step access to the target document collection, and can be seen as a “*sense-based*” relevance feedback. According to this technique, once the top d_1, d_2, \dots, d_n documents have been retrieved using the question keywords, the most frequent *senses* of the question terms are considered instead of drawing for expansion the most relevant *words* that appear within d_1, d_2, \dots, d_n . Query enrichment is then carried out adding terms semantically related to these senses. Our experiments, carried out using part of the TREC-2003 factoid questions set and the AQUAINT corpus for document retrieval, demonstrate the viability of this approach. Preliminary results show that the application of Sense-based Relevance Feedback to the QA scenario can improve retrieval up to 7% in terms of answer-bearing documents obtained with the best performing expansion strategy.

Keywords

Question Answering, Word Sense Disambiguation, Query Expansion, WordNet

1. INTRODUCTION

Dealing with the general problem of finding textual information that is relevant with respect to a particular user’s information need, Information Retrieval (IR) and Question Answering (QA) systems are faced with the challenges posed by language variability and word ambiguity. In both these frameworks, one of the major needs is to bridge the gap between the query terms (the query space) and the actual form in which the sought-after information is stored in the target collection (the document space). Often, in fact, queries do not contain the same words that are used in the document space to represent concepts [18]. The impact of the mismatch between the query and the document space becomes particularly evident in the QA scenario, where (*i*) actual answers are required instead of relevant documents; and (*ii*) the target collection often contains a limited number of rele-

vant text passages from which answers can be mined ([14] reports an average of 7.0/5.0/3.0 correct answer-bearing documents per question at TREC-9/10/11 respectively).

Since it is likely that correct answers will appear within documents which have little or no similarity to the question keywords, point (*i*) makes a big difference between IR (whose purpose is just finding relevant documents with respect to a given topic statement) and the more refined QA task. As a consequence, while IR techniques proved to be effective at locating within large collections of documents those relevant to a user’s query, more fine-grained techniques are required in QA. At the same time, point (*ii*) demands that these techniques not be too restrictive, since the performance of the retrieval module clearly represents an upper bound for the overall system’s performance (if a relevant document is missed, there is no way to finally return a correct answer at the end of the whole process).

The modalities for query expansion with terms related (either semantically, as it is discussed in this paper, or in virtue of their co-occurrence tendency) to the words of an input query represent one of the many facets of the problem, and have become crucial issues both for the IR and QA communities. However, while in IR the research in this direction has reached a certain degree of maturation, little has been done with respect to the specificities of the QA benchmark, where it is still not clear what kinds of expansion techniques are best suited for the task.

Focusing on the specific issues raised by query expansion with semantically related terms in the QA scenario, our work aims at filling this lack. In this direction, we analyze different WORDNET-based query enrichment techniques whose common underlying approach is a two-step access to the local document collection. This approach can be seen as a semantically oriented variant of Blind Relevance Feedback (BRF). Like in BRF ([1], [12]), terms for expansion are drawn from the top d_1, d_2, \dots, d_n documents obtained after a first search with the original query keywords. The novelty here introduced is that instead of taking as additional expansion terms the *most relevant words* contained in d_1, d_2, \dots, d_n , we consider the *most frequent senses* of the original question’s keywords. Experimenting with an unsupervised Word Sense Disambiguation (WSD) tool, we aim to demonstrate that the recognition of these senses within d_1, d_2, \dots, d_n is a more feasible task than the disambiguation

of the input question.

The paper is organized as follows. Section 2 will briefly overview the background related to this work, motivating our effort on query expansion for QA. Section 3 will propose “Sense-based” Blind Relevance Feedback (S-BRF) as a solution to the problem of reliably disambiguating words occurring in a natural language question. Section 4 will present some experiments carried out to determine what kind of information can be mined from WORDNET to improve document retrieval in a QA system. Our results, discussed in Section 5, show that S-BRF can improve retrieval performance up to 7% in terms of answer-bearing documents obtained with the best performing expansion strategy.

2. BACKGROUND

In the last few years WORDNET [10] has been widely adopted as a core resource in the construction of QA systems. The literature (see for example [5], [4], and [24]) presents a number of effective variants in using the semantic information encoded in WORDNET to construct most of the basic components involved in the QA process. Among these, the modules in charge of answer type recognition, named entity recognition, and answer justification often implement techniques based on taxonomic information. However, its actual contribution to improve document retrieval still represents a debatable issue. Even though WORDNET seems to have great potential in providing additional terms to improve search results, only a limited number of participants at the TREC and CLEF QA tasks¹ have documented its effective exploitation for this purpose. For instance, only 6 out of 25 participants to the last TREC edition describe some kind of query enrichment technique using WORDNET. These techniques vary from query expansion with hypernyms of the question focus [16] or synonyms of the question terms [23], to the extraction from WORDNET of multiword expressions that should not be separated in the query process [6]. An interesting approach is presented in [24], which uses Google to extract from the Web relevant terms with respect to an input question, and WordNet to adjust their weights and perform query expansion. However, since none of these systems seems to use WSD at any phase of the process, the actual impact of WORDNET-based expansions to the overall systems’ performance is rather unclear.

The limited use of WORDNET for query expansion in QA is probably motivated by the skepticism, developed within the IR community, towards dealing explicitly with semantic information to improve document retrieval (*e.g.* by using hand-made thesauri). This general feeling originates from a number of experiments targeted to assess the impact on the search results of applying WSD and different strategies of query enrichment with synonyms. Part of these works agreed on the conclusion that while recall is actually improved by query enrichment, precision can be dramatically affected by poor WSD performance. Discouraging results are presented in [17], which concludes that a WSD system performing with less than 90% of accuracy decreases docu-

¹Both the Text REtrieval Conference (TREC, see <http://trec.nist.gov>) and the Cross Language Evaluation Forum (CLEF, see <http://www.clef-campaign.org>) are annual evaluation exercises designed to boost research in a number of IR related areas.

ment retrieval results. Accordingly, it seems better to avoid any expansion attempt since state-of-the-art WSD systems are still far from this level of accuracy. The best performing system in the English all-words task at Senseval-2², in fact, reached a 69% precision and recall performance.

Independently from the quality of WSD, other work also puts in doubt the usefulness of WORDNET-based enrichments. Among these, [21], [20], and [19] did not notice any improvement when expanding queries (even manually) with different combinations of synonyms, hypernyms, and hyponyms, unless such expansion is limited to short, incomplete queries. For longer topic statements (in this case the *Narrative field* of a TREC topic statement), [20] reports that none of the expansion methods produced more than a 2% improvement with respect to unexpanded queries.

Nevertheless, a number of works criticized part of these conclusions. Among these, [13] introduces the idea of using WORDNET to extend Web searches based on semantic similarity. Their results show that WSD-based query enrichment actually improves retrieval performance in the Web scenario. [2] shows that using WSD to expand TREC-6 query titles with synonyms and hyponyms extracted from WORDNET 1.5 does not necessarily decrease performance, even at degrees of accuracy lower than the 90% threshold set by [17] (the WSD system they adopt performs at 72%). Similar conclusions have been reached in [3], which experimented with the effect of searching SemCor [11] with manually disambiguated and expanded queries. Their results show that, combined with synonymy enrichment, a WSD system performing at 70% still increases retrieval performance, and even with 40% proper sense recognition there are not significant losses. [15] describes a simple WSD algorithm, based on structured pattern recognition used to disambiguate short queries. Using the Google search engine to access the Web and WORDNET 1.6 as a word sense repository, they report limited effects on retrieval performance when expanding with synonyms (2.4%) and hypernyms (1.6%), but a considerable improvement (26.8%) when words extracted from WORDNET glosses are used. [25] proposes a technique which effectively substitutes WSD with corpus-based statistics for ranking lexical paraphrases of the input queries. The reported results show how these paraphrases, generated using WORDNET synonyms and part-of-speech information, can improve overall retrieval performance up to 8% for queries with less than 5 words.

3. SENSE-BASED BLIND RELEVANCE FEEDBACK

3.1 The Problem

From the QA perspective, since the length of a query (after stop words removal from the input questions) is usually small³, the last results reported in the previous section are encouraging. At the same time, it must be kept in mind that the performance of state-of-the-art WSD algorithms is

²Senseval (<http://www.senseval.org>) is a periodic competition aiming at the evaluation of WSD systems. In the all-words task, the evaluation is on the disambiguation of all the content words contained in a sample text.

³The average length of a question at TREC-2003 was 7.7 words, ranging from a minimum of 3, to a maximum of 18.

highly sensitive to the dimension of the context surrounding a candidate target word. In practice, the restrictions posed by the QA benchmark (*i.e.* dealing with short natural language questions) may determine a loss in accuracy with respect to WSD over longer texts, where larger pieces of context are available for disambiguation. Given this situation, on one hand the retrieval phase of the QA process represents the ideal “short-query scenario” which proved to benefit from WORDNET-based query expansion techniques. On the other hand, it is affected by the possible degradation of WSD accuracy caused by the limited context provided by questions.

3.2 Our Solution

To address this problem, we propose Sense-based Blind Relevance Feedback (S-BRF) as a method to enhance a WSD system’s reliability in the QA scenario. Geared to query expansion, S-BRF builds on the assumptions that: (i) the limited context provided by short natural language questions can be expanded in order to perform a reliable disambiguation of the words they contain, and (ii) this expansion can be profitably accomplished by accessing the local document collection available for the final answers’ extraction. S-BRF is similar to BRF in that it performs query expansion through a two-step access to the target collection. At the first step, the top n documents returned are used to mine relevant “knowledge” with respect to the input query. At the second step such “knowledge” is combined into an expanded query which is supposed to return more relevant documents. The difference is that while BRF considers as relevant “knowledge” the most relevant *words* found within the documents retrieved at the first step, S-BRF builds expanded queries using semantic information related to the most frequent *senses* of the query keywords in the retrieved documents.

Given $Q(q_1, q_2, \dots, q_n)$, a question Q containing the keywords q_1, q_2, \dots, q_n , the S-BRF process can be summarized in the following steps.

1. **Context enlargement.** Q is part-of-speech tagged, and subject to multiwords and named entities recognition. This processing will serve to select good candidates for disambiguation in the following phase of the process. The first retrieval step is carried out querying the collection with:

$[q_1 \text{ AND } q_2 \text{ AND } \dots \text{ AND } q_n]$

in order to produce the necessary context for a reliable disambiguation of q_1, q_2, \dots, q_n .

Since this strict query may return empty results, relaxation heuristics (see Section 5.1) are recursively applied until at least 10 paragraphs⁴ are retrieved. Even though in the TREC competition the average number of answer-bearing documents per question is usually less than 10 (see Section 1), we settled to 10 the number of retrieved paragraphs under the hypotheses that one answer-bearing document may contain more than

⁴In our experiments, the document collection has been indexed at the paragraph level. For this reason, from now on we will refer to the output of document retrieval as a set of *paragraphs*, instead of *documents*.

one passage in which q_1, q_2, \dots, q_n appear close to each other, and that the co-occurrence of these keywords is always a good clue for disambiguation, also within a non answer-bearing document. In further developments of the approach, we plan to assess the impact of varying the number of documents used in the relevance feedback set.

At the end of this first processing phase, the resulting extended context is represented by the set of paragraphs:

$$P_Q = \{p_{(q_1, q_2, \dots, q_n)}^1, p_{(q_1, q_2, \dots, q_n)}^2, \dots, p_{(q_1, q_2, \dots, q_n)}^k\}$$

in which q_1, q_2, \dots, q_n co-occur.

2. **Word Sense Disambiguation.** Once P_Q has been extracted from the local collection and duplicates have been removed, each of its elements is subject to part-of-speech tagging, multiword recognition and named entities recognition. WSD is then carried out trying to assign a WORDNET 1.6 sense to each of the occurrences of q_1, q_2, \dots, q_n in P_Q , which have the same part of speech in Q .

Term occurrences with different parts of speech are discarded in order to reduce the risk of introducing noise by disambiguating wrong elements. For instance, given the input question Q_{1934} : “*What is the play “West Side Story” based on?*”, and the retrieved paragraph “... *everything from “Maria” (from “WEST SIDE STORY”) to “Take me Out to the Ball Game” will PLAY 28 performances*”, the verb “*play*” will not be marked as a candidate term for WSD, as the occurrence of “*play*” in the part-of-speech-tagged question is a noun. Named entities *e.g.* “*Notre Dame*” not present in WORDNET will be discarded as well, in order to avoid wrong expansions with synonyms of their component words (*e. g.* “*doll*”, “*wench*”, “*skirt*”, “*chick*”, and the other synonyms of “*Dame*”). The same holds for multiword expressions present in the question (*e. g.* “*square miles*” in Q_{1977} : “*How large in square miles is North Carolina?*”). If their components appear scattered through a retrieved paragraph (as in “... *about four MILES west of the Colonial SQUARE ...*”) they will not be disambiguated.

3. **Query Expansion.** Finally, the most frequent *senses* of q_1, q_2, \dots, q_n within P_Q are collected and used for query expansion. The new enriched query will be of the form

$$[(q_1 \text{ OR } q_1' \text{ OR } q_1'' \dots q_1^w) \text{ AND } (q_2 \text{ OR } q_2' \text{ OR } q_2'' \dots q_2^y) \text{ AND } \dots \text{ AND } (q_n \text{ OR } q_n' \text{ OR } q_n'' \dots q_n^z)]$$

which indicates a boolean combination containing the original terms $q_1, q_2, \dots, q_n \in Q$, plus additional words q_i^j semantically related to these terms mined from the WORDNET hierarchy (*e. g.* synonyms, hypernyms, holonyms, etc.).

S-BRF presents a twofold advantage with respect to directly applying WSD to the input question. First, the enlarged context we create is “question-specific”. The retrieval of paragraphs where q_1, q_2, \dots, q_n co-occur, in fact, actually acts as an implicit mutual disambiguation. This is likely to filter out paragraphs which contain occurrences of these words with different senses with respect to the ones intended by the question Q .

Second, by including a large number of occurrences of q_1, q_2, \dots, q_n , the new enlarged context P_Q contributes to coping with WSD deficiencies both in terms of recall and precision. Recall is improved by the fact that, instead of a single occurrence of q_i within Q , the disambiguation tool will have many more chances to resolve its sense. Precision in the recognition of more ambiguous words benefits from the selection of the most frequent sense, among the ones recognized by the WSD system, of a word q_i within P_Q .

S-BRF also allows us to overcome the known problems related to BRF when applied to the QA scenario. BRF is not used in QA, probably due to the small number of relevant documents. The fact that under those circumstances BRF performs rather poorly, and that its application in the context of QA is not appropriate, has been empirically confirmed by [14]. Results reported show a performance drop with respect to plain retrieval when BRF is applied. However, by providing a more restrictive and controlled way for adding new terms to the original queries, S-BRF allows us to reduce the impact of the noise introduced by expanding with too many terms. This is confirmed by the preliminary results of our experiments carried out using S-BRF to find the WORDNET-based query expansion technique best suited for QA.

4. EXPERIMENTAL SETTING

4.1 Dataset

The experiments described in this section have been carried out using the MG search engine [22] to access the AQUAINT Corpus of English News Text used in the TREC QA track. The collection, which was indexed at the paragraph level, contains approximately 1,033,000 newswire documents for a total of 3.0 Gigabytes.

The test set has been extracted from the 2003 Main Task, and is composed of all the 380 factoid questions (out of a total of 500 questions, subdivided into the three categories *factoid*, *definition*, and *list*) for which a correct answer can be found within the collection. Factoid questions typically require a single named entity as an answer (e.g. “How far is it from Earth to Mars?” or “What city is Disneyland in?”). Answers to these questions are represented by [*document-id*, *answer-string*] pairs, where *answer-string* is an “exact answer”⁵, and *document-id* indicates a document supporting this answer. In our test set, an average of 3.47 words per question are ambiguous (out of an average length of 7.7 words per question), and the average ambiguity factor for each term in WORDNET 1.6 is 1.75.

The results here reported were calculated using the perl answer patterns together with the documents in which they appear, which have been kindly made available by Ken Litkowski through the TREC website. Manual checking was adopted when a correct answer string returned by the search engine could not be found within the judgment files.

⁵At TREC, an “exact answer” is roughly defined as a text portion containing nothing more and nothing less than the correct answer. For instance, “Mississippi” and “Mississippi River” are exact answers to the question “What is the longest river in the United States?”, while “the Mississippi river is the” is wrong.

4.2 Word Sense Disambiguation

In our implementation of S-BRF, the recognition of the proper sense of a term in P_Q is in charge of a Domain Driven Disambiguation (DDD) tool. DDD [9] is an unsupervised WSD technique that makes use of only domain information in order to solve lexical ambiguity. The basic idea is that the disambiguation of a word in its context is mainly a process of comparison between the domain of the context and the domains of the word’s senses. The system uses the semantic domains associated with WORDNET synsets available in WORDNET DOMAINS [8], an extension of WORDNET 1.6 where each synset has been annotated with at least one domain label, selected from a set of about 200 labels hierarchically organized.

The methodology requires three steps:

1. Compute $DV(c)$ for each sense c of the word w to be disambiguated.
2. Compute $DV(t)$ for the context t of the word w to be disambiguated.
3. Choose the sense:

$$\hat{c} = \operatorname{argmax}_{c \in \text{senses}(w)} \{ \text{sim}(DV(c), DV(t)) \}$$

where a Domain Vector (DV) assigns a weight to each considered domain⁶ representing the strength of the association between a domain and either a word’s sense c or a text t .

Step 1 is based on domain annotations available in WORDNET DOMAINS and does not require any training.

Step 2 considers a 50 word text window (i.e. the context) around the target word and builds a domain vector considering the contribution of all the content words in this window belonging to WORDNET. In this phase the system makes use of a *domain relevance threshold* which gives an estimation of how many instances (i.e. words) in a text are necessary in order to select a word sense belonging to a certain domain. This threshold has been estimated, for each domain, over the BNC Corpus, and it takes into account the fact that different domains have different coverage (e.g. ECONOMY is more frequent than VETERINARY) and, as a consequence, their relevance is based on a different amount of words.

Finally, step 3 compares each sense vector of the target word with the context vector, and the most similar one is selected. The system performance can be tuned using a parameter that allows us to tune the tradeoff between precision and recall.

State-of-the-art performance emerged from the assessment of this disambiguation technique in different evaluation settings. In the framework of the SENSEVAL-2 competition, the system achieved 74.8% Precision and 35.7% Recall values in the English all-words task. Further tests on the same scenario, carried out using SemCor as a test collection, resulted in 84% Precision, 53% Recall, and 65% F1 scores.

⁶For simplicity, a set of 43 disjoint domain labels (e.g. SPORT in place of VOLLEY or TENNIS) is used, instead of 200, which allows for a good level of abstraction without losing relevant information.

Table 1: Baseline results. Precision and answer-bearing documents retrieved

	p@5	a@5	p@10	a@10	p@20	a@20	p@40	a@40	p@50	a@50	# a
BASILINE	11.52	29.21	10.38	40.00	9.37	50.52	8.25	61.57	7.94	65.52	249

Table 2: Setting1 results. Precision and answer-bearing documents retrieved without S-BRF

	p@5	a@5	p@10	a@10	p@20	a@20	p@40	a@40	p@50	a@50	# a
Monosemous Synonyms	17.25	43.70	14.35	51.32	11.42	59.74	8.90	66.84	8.22	69.21	263
1 POS Synonyms	16.80	43.16	14.22	51.05	11.25	59.74	8.75	66.60	8.10	69.21	263
All Synonyms	16.72	42.63	14.14	50.53	11.23	57.63	8.71	63.70	8.12	67.90	258

4.3 Evaluation Measures

As for the evaluation measures, we adopt the same metrics proposed by [14] to assess the impact of different retrieval strategies for QA. However, since the number of relevant documents in the collection is unknown, recall values cannot be calculated. Thus, only two of these metrics are used, namely **p@n** (*i.e.* precision at n documents retrieved, with $n=5, 10, 20, 40, 50$), and **a@n** (*i.e.* the percentage of questions for which at least one correct answer-bearing document was retrieved at the same five cut-off levels). As suggested by [14], we did not consider higher cut-offs since most QA systems seldomly consider more than the top 50 documents returned by the retrieval component. The total number of questions for which at least one correct answer-bearing document is obtained (**# a**, also called *productivity* by [13]) is also reported as a compact and intuitive way to represent the effectiveness of each retrieval strategy.

5. RESULTS AND DISCUSSION

A baseline (see 5.1) was calculated for our evaluation purposes. Such a baseline is used as a term of comparison to estimate the impact of different WORDNET-based query enrichment strategies within two experimental settings. The first of these settings (5.2) aims at measuring the potentialities of query expansion with synonyms in the framework of a less precise WSD (*i.e.* when disambiguation is carried out without recourse to the local document collection). The second setting (5.3) is intended to measure the improvements on document retrieval given by the S-BRF method described in the previous section. To this end, we experiment combining S-BRF with different expansion strategies using information mined from WORDNET 1.6.

5.1 Baseline

Baseline figures have been calculated by processing an input TREC question (*e.g.* Q2104: “When was the submarine invented?”), and searching AQUAINT with a slight variant of the retrieval module actually used by DIOGENE, the QA system developed at ITC-irst (see [7] for details). In the variant adopted, while any WORDNET-based query enrichment was disabled, query expansion with morphological derivations has been left active (*e.g.* allowing the addition of the query terms “invention”, “inventing”, and “invent” derived from the verb “invented”). This actually produced quite high results for a “baseline” (see Table1), but our intention was to compare the improvement given by WORDNET-based query expansions with realistic performance results.

Search queries are boolean combinations obtained from the question keywords and their morphological derivations. As

the boolean model frequently returns either too many or too few documents to the user, query relaxation and paragraph ranking techniques (not available when using MG in the boolean mode) had to be implemented.

Query relaxation obeys several heuristics, which are applied in the following order. Firstly the leftmost common name (*e.g.* “country” in Q2387: “What country celebrates Guy Fawkes Day?”) is removed, as it often indicates the semantic type of the answer which usually does not appear within relevant passages (*e.g.* “Plenty of similar celebrations can be found worldwide. In England, for example, Nov. 5 is Guy Fawkes Day”). At the second step, lower case adverbs and several frequent less important verbs (*e.g.* “call”, “name”, “abbreviate”) are deleted. Then, lower case verbs with more than 20, 12, and 5 WORDNET senses, lower case hyponyms of “abstract synsets” (*i.e.* **abstraction#6**, **grouping#1**, and **psychological_feature#1**), and lower case adjectives are iteratively removed right to left.

Paragraph ranking takes into account the number and the relative distance of the query keywords, pushing paragraphs maximizing these factors higher in the final output. At the end, the top 50 paragraphs are returned and inspected to check for the presence of correct answers. Baseline results are reported in Table 1.

5.2 Setting 1

Using the same retrieval module previously described, *Setting 1* results have been calculated activating wordnet-based query expansion *without* S-BRF. The main goal of this evaluation setting is to measure the impact of different expansion strategies under less precise WSD performance conditions. These conditions are created by disambiguating the input questions without recourse to the local document collection. Once the input query keywords have been disambiguated, actual query expansion is carried out by considering three types of WORDNET synonyms. This process can be summarized as follows:

- Step1: Keyword Extraction.** The input question Q is subject to part-of-speech tagging, multiwords and named entities recognition. Then, stop words removal is applied to select the most relevant keywords q_1, q_2, \dots, q_n , which are marked as candidates for disambiguation.
- Step2: WSD.** Disambiguation is carried out over Q , trying to determine the proper senses of q_1, q_2, \dots, q_n .

Table 3: Setting2 results. Precision and answer-bearing documents retrieved using S-BRF

	p@5	a@5	p@10	a@10	p@20	a@20	p@40	a@40	p@50	a@50	# a
Monosemous Synonyms	17.66	45.26	14.59	53.42	11.45	61.32	8.85	68.94	8.18	71.31	271
1 POS Synonyms	17.46	44.74	14.48	52.63	11.34	61.58	8.78	68.42	8.10	70.79	269
All Synonyms.	17.61	44.74	14.56	51.84	11.36	60.53	8.94	68.68	8.27	71.05	270
Mon. Syn.+ADJ2N	17.67	45.53	14.67	54.21	11.58	62.10	9.15	70.00	8.44	72.63	276
Mon. Syn.+Monos. Glos.	17.54	42.63	14.25	48.95	11.52	55.26	9.23	61.58	8.60	63.68	242
Mon. Syn.+1 POS. Glos.	16.54	40.26	13.54	46.84	11.08	52.89	9.15	60.00	8.61	62.10	236
Mon. Syn.+All Glos.	16.54	40.26	13.54	46.84	11.08	52.89	9.15	60.00	8.61	62.10	236
Mon. Syn.+All Glos.+ADJ2N	17.28	42.37	14.20	48.95	11.57	55.53	9.35	61.84	8.71	63.95	243

3. **Step3: WordNet mining.** Word senses are used to mine WORDNET terms that are semantically related to q_1, q_2, \dots, q_n . In particular, for each $q_i \in q_1, q_2, \dots, q_n$, we collect:

- all synonyms (if any)
- synonyms which have only one part-of-speech (if any)
- monosemous synonyms (if any)

4. **Step4: Query Expansion.** Boolean queries such as:

$[(q_1 \text{ OR } q_1' \text{ OR } q_1'' \dots q_1^w) \text{ AND } (q_2 \text{ OR } q_2' \text{ OR } q_2'' \dots q_2^y) \text{ AND } \dots \text{ AND } (q_n \text{ OR } q_n' \text{ OR } q_n'' \dots q_n^z)]$

are made combining the original question terms with words taken from one of the three lists.

Table 2 reports the results obtained by the three types of query expansion with WORDNET synonyms. As it can be seen, retrieval performance is improved by all the strategies, at all the cut-off levels, both in terms of precision and of answer-bearing documents retrieved. The best results are given by query enrichment with monosemous synonyms, which improves over the baseline up to 14.5% at a@5. It’s noteworthy that, in general, the best improvements with respect to the baseline are achieved at the lowest cut-off level. This is particularly relevant under the QA perspective, where the computational cost of analyzing a large number of documents is very high.

5.3 Setting 2

Once demonstrated that query expansion with synonyms improves document retrieval, Setting 2 experiments aim at verifying: (i) the extent to which the retrieval phase in a QA scenario can be improved by a more precise WSD, and (ii) if other kinds of information extracted from WORDNET can be profitably used.

As for (i), the disambiguation of the input questions is carried out by activating the S-BRF module. The process involves the four steps used for Setting 1, *i.e.* Keyword Extraction, WSD, WORDNET mining, and Query Expansion. Moreover the **Context enlargement** described in Section 3.2 is added as an intermediate step between Keyword Extraction and WSD.

As for (ii), additional information is searched within the WORDNET mining phase, which is in charge of providing terms semantically related to the input question’s keywords.

In particular, besides synonyms, for each $q_i \in q_1, q_2, \dots, q_n$ we collect:

- the noun to which q_i is connected through the PERTAINS-TO or IS-VALUE-OF relation (if exists) when q_i is an adjective (*e.g.* “old” → “age”, “British” → “Great Britain”)
- all the content words contained in the gloss of q_i
- the monosemous content words contained in the gloss of q_i
- all the content words contained in the gloss of q_i which have only one part-of-speech

Summing up, first S-BRF is used to enhance the WSD precision; then the various kinds of information mined from WORDNET are combined in different ways during the query expansion phase.

The results obtained are reported in Table 3. A first conclusion that can be drawn is that under the same query expansion configuration used in Setting 1, *i.e.* when only synonyms are added, S-BRF improves the performance at all cut-off levels. This confirms our initial working hypothesis, even if the overall improvement of Setting 2 with respect to Setting 1 (an average of about 2% in terms of answer-bearing documents retrieved) is lower than what we expected. The best results are obtained when S-BRF exploits the semantic information conveyed by a combination of monosemous synonyms and the PERTAINS-TO and IS-VALUE-OF WORDNET relations (Mon. Syn.+ADJ2N). With this configuration, performance improves at all cut-off levels, reaching its maximum at a@50, obtaining an increase of 3.42% with respect to the best result of Setting 1, and of 7.11% with respect to the baseline in terms of answer-bearing documents. As regards the additional information acquired from WORDNET glosses, we must notice a considerable performance drop, especially at higher cut-off levels. This is due to the fact that expansion with glosses is a less restrictive strategy introducing more noise than semantically relevant terms, as it happens when applying BRF to the QA benchmark.

Even though the overall improvement achieved by S-BRF with respect to Setting 1 is an indirect indication of the suitability of the approach, a preliminary evaluation of the intermediate steps of the algorithm has been carried out

to throw light on why such improvement is lower than expected. The analysis of the WSD process revealed that the less precise disambiguation carried out in Setting 1 returns at least one synonym for at least one word for 353 questions (93% of the dataset). In contrast, the more restrictive S-BRF technique produces at least one synonym for at least one word for 212 questions (55.7% of the dataset). In particular, as regards the Monosemous Synonyms, we observed a drop from 321 questions (84.47% of the dataset) in Setting 1 (with an average of 2.46 Monosemous Synonyms added per question), to 177 in Setting 2 (46.57% of the dataset, with an average of 1.31 Monosemous Synonyms added per question). A similar trend can be observed also for the other classes of terms mined from WORDNET (*i.e.* 1POS Synonyms, ADJ2Ns, Gloss words). Summing up, S-BRF leads to the expansion of less questions (*e.g.* 177 vs. 321 in the case of Monosemous Synonyms) with a smaller average number of synonyms (1.31 vs. 2.46). On one side, the reduction of the noise produced by imprecise query enrichments gives better results than those achieved in Setting 1. However, on the other side, the concrete realization of S-BRF experimented so far is still too restrictive: the expansion of a too limited number of questions does not allow for a significant performance improvement. Before considering our results conclusive, we think worthwhile exploring more relaxed criteria both in the selection of the top n paragraphs used to perform S-BRF, and in the selection, within these paragraphs, of the words to be disambiguated.

6. CONCLUSIONS

In this paper we proposed a method which aims at enhancing document retrieval under the specific restrictions posed by the QA scenario. Geared to query expansion, Semantic-based Blind Relevance Feedback (S-BRF) builds on the assumptions that: (*i*) the limited context provided by short natural language questions can be expanded in order to perform a reliable disambiguation of the words they contain, and (*ii*) this expansion can be profitably accomplished by accessing the local document collection available for the final answers' extraction. S-BRF is similar to Blind Relevance Feedback (BRF) in that it performs query expansion through a two-step access to the target collection. The difference is that while BRF considers as relevant "knowledge" for query expansion the most relevant *words* found within the documents retrieved at the first step, S-BRF builds these queries using semantic information related to the most frequent *senses* of the query keywords in the retrieved documents. In order to evaluate the proposed methodology, a number of experiments has been carried out, using different combinations of semantic information mined from WORDNET. Preliminary results confirm the viability of S-BRF, showing an improvement up to 7% over the baseline. Since the overall improvement on document retrieval is lower than expected, we have started a first qualitative analysis of the intermediate steps of the process, in order to better understand the extent to which this is due to disambiguation errors or to unsuitability of WORDNET-based expansions. Building on our preliminary observations we are going to investigate new methods to relax the S-BRF algorithm.

7. REFERENCES

- [1] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic Query Expansion Using SMART. In D. Harman, editor, *The Third Text REtrieval Conference (TREC3) NIST Special Publication*, 1995.
- [2] C. de Loupy and M. El-Beze. Managing Synonymy and Polysemy in a Document Retrieval System Using WordNet. In *LREC-2002 Workshop on "Creating and Using Semantics for Information Retrieval and Filtering*, Las Palmas, Canary Islands, Spain, June 2002.
- [3] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran. Indexing with WordNet Synsets can Improve Text Retrieval. In *Proceedings of the COLING/ACL'98 Workshop on Usage of WordNet for NLP Proceedings of the COLING/ACL'98 Workshop on Usage of WordNet for NLP*, Montreal, 1998, Montreal, Canada, 1998.
- [4] S. Harabagiu, D. Moldovan, C. Clark, M. Bowden, J. Williams, and J. Bensley. Answer Mining by Combining Extraction Techniques with Abductive Reasoning. In *TREC-2003 Notebook Papers*, pages 46–53, Gaithersburg, MD, USA, November 18-21 2003.
- [5] S. Harabagiu, D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus, and P. Morarescu. The Role of Lexico-Semantic Feedback in Open Domain Textual Question Answering. In *ACL01 - Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 274–281, Toulouse, France, 2001.
- [6] B. Katz, J. Lin, D. Loreto, W. Hildebrandt, M. Bilotti, S. Felshin, A. Fernandes, G. Marton, and F. Mora. Integrating Web-Based and Corpus-Based Techniques for Question Answering. In *TREC-2003 Notebook Papers*, Gaithersburg, MD, USA, November 18-21 2003.
- [7] M. Kouylekov, B. Magnini, M. Negri, and H. Tanev. ITC-irst at TREC-2003: the DIOGENE QA System. In *TREC-2003 Notebook Papers*, pages 425–433, Gaithersburg, MD, USA, November 18-21 2003.
- [8] B. Magnini and G. Cavaglià. Integrating Subject Field Codes into WordNet. In *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*, Athens, Greece, June 2000.
- [9] B. Magnini, C. Strapparava, G. Pezzulo, and A. Gliozzo. The Role of Domain Information in Word Sense Disambiguation. *Natural Language Engineering, special issue on Word Sense Disambiguation*, 8(4):359–373, 2002.
- [10] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Introduction to WordNet: an On-Line Lexical Database. *Journal of Lexicography*, 3(4):235–244, 1990.
- [11] G. Miller, C. Leacock, R. Teng, and R. Bunker. A Semantic Concordance. In *Proceedings of the DARPA Workshop on Human Language Technology*, pages 303–308, Plainsboro, New Jersey, USA, 1993.

- [12] M. Mitra, A. Shinghal, and C. Buckley. Improving Automatic Query Expansion. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development of Information Retrieval*, pages 206–214, 1998.
- [13] D. Moldovan and R. Mihalcea. Using WordNet and Lexical Operators to Improve Internet Searches. *IEEE Internet Computing*, 2000.
- [14] C. Monz. *From Document Retrieval to Question Answering*. PhD thesis, Institute for Logic, Language and Computation (ILLC), 2003.
- [15] R. Navigli and P. Velardi. An Analysis of Ontology-Based Query Expansion Strategies. In *Workshop on Adaptive Text Extraction and Mining (ATEM 2003), in the 14th European Conference on Machine Learning*, Cavtat-Dubrovnik, Croatia, September 22-26 2003.
- [16] J. Prager, J. Chu-Carroll, K. Czuba, C. Welty, A. Ittycheriah, and R. Mahindru. IBM’s PIQUANT in TREC-2003. In *TREC-2003 Notebook Papers*, pages 36–45, Gaithersburg, MD. USA, November 18-21 2003.
- [17] M. Sanderson. Word Sense Disambiguation and Information Retrieval. In *Proceedings of the 17th annual international ACM-SIGIR conference on Research and development in information retrieval*, pages 142–151, Dublin, Ireland, June 1994.
- [18] A. F. Smeaton, F. Kellely, and R. O’Donnel. TREC-4 Experiments at Dublin City University: Thresholding Posting Lists, Query Expansion with WordNet and POS Tagging of Spanish. In *Proceedings of the 4th Text Retrieval Conference (TREC-4)*, pages 373–390, Gaithersburg, MD. USA, 1995.
- [19] A. F. Smeaton and A. Quigley. Experiments on Using Semantic Distances Between Words in Image Caption Retrieval. In *Proceedings of the 19th International Conference on Research and Development in IR*, 1996.
- [20] E. Voorhees. *WordNet: An Electronic Lexical Database*, chapter Using Wordnet for Text Retrieval, pages 285–303. The MIT Press, Cambridge, Massachusetts, USA, 1998.
- [21] E. M. Voorhees. Query Expansion Using Lexical-Semantic Relations. In *Proceedings of the 17th annual international ACM-SIGIR conference on Research and development in information retrieval*, pages 61–69, Dublin, Ireland, June 1994.
- [22] I. Witten, A. Moffatt, and T. Bell. *Compressing and Indexing Documents and Images*. Morgan Kaufmann, 2nd edition, 1999.
- [23] M. Wu, X. Zheng, M. Duan, T. Liu, and T. Strzalkowski. Question Answering by Pattern Matching, Web Proofing, Semantic Form Proofing. In *TREC-2003 Notebook Papers*, Gaithersburg, MD. USA, November 18-21 2003.
- [24] H. Yang, H. Chui, M. Kan, M. Maslennikov, L. Qiu, and T. Chua. QUALIFIER in TREC-12 QA Main Task. In *TREC-2003 Notebook Papers*, pages 54–63, Gaithersburg, MD. USA, November 18-21 2003.
- [25] I. Zukermann and B. Raskutti. Lexical Query Paraphrasing for Document Retrieval. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan, 2002.