Query Expansion for IR using Knowledge-Based Relatedness

Arantxa Otegi
IXA NLP Group
Univ. of the Basque Country
arantza.otegi@ehu.es

Xabier Arregi
IXA NLP Group
Univ. of the Basque Country
xabier.arregi@ehu.es

Eneko Agirre
IXA NLP Group
Univ. of the Basque Country
e.agirre@ehu.es

Abstract

The limitations of keyword-only approaches to information retrieval were recognized since the early days, specially in cases where different but closely-related words are used in the query and the relevant document. Query expansion techniques like pseudo-relevance feedback rely on the target document set in order to bridge the gap between those words, but they might suffer from topic drift. This paper explores the use of knowledge-based semantic relatedness in order to bridge the gap between query and documents. We performed query expansion, with positive effects over some language modeling baselines.

1 Introduction

The potential pitfalls of keyword retrieval have been noted since the earliest days of Information Retrieval (IR). Keyword retrieval proves ineffective when different but closely-related words are used in the query and the relevant document. The use of different words creates a lexical gap between the query and the document.

In order to bridge the gap, IR has resorted to distributional semantic models. Most research concentrated on Query Expansion (QE) methods, which typically analyze term co-occurrence statistics in the corpus and/or in the highest scored documents in order to select terms for expanding the query terms (Manning et al., 2009). The work presented here is complementary, in that we explore QE, but we use an approach based on semantic relatedness instead of distributional methods.

In a closely related work, (Agirre et al., 2010) proposed a WordNet-based document expansion method using random walks: given a document, a random walk algorithm over the WordNet graph, inspired in (Agirre et al., 2009b), ranks concepts closely related to the words in the document. Note that the method can return concepts which are not explicitly mentioned in the document. The highest ranking concepts were then selected to expand the document.

In this work, we explore an alternative method to exploit relatedness, query expansion, so we thus run the relatedness algorithm over the queries and we expand the queries. We adopt a language modeling framework to implement the query likelihood and pseudo-relevance feedback baselines, as well as our relatedness-based query expansion method.

In order to test the performance of our method we selected several datasets with different domains, topic typologies and document lengths. Given the relevance among the community using WordNet-related methods, we selected the Robust-WSD dataset from CLEF (Agirre et al., 2009a), which is a typical ad-hoc dataset on news. As we think that our method is specially relevant for short queries and/or short documents, we also selected the Yahoo! Answers dataset, which contains questions and answers as phrased by real users on diverse topics (Surdeanu et al., 2008), and ResPubliQA, a paragraph retrieval task on European Union laws organized at CLEF (Peñas et al., 2009).

The results show that our method provide improvements in all three datasets, when compared to the query likelihood baseline, and that they compare favorably to pseudo-relevance feedback in two datasets.

The paper is structured as follows. We first briefly introduce related work. We then mention the random walk model for query expansion. The design of the experiments is presented in Section 4. Section 5 shows our results, and, finally, Section 6 presents the conclusions.

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 1467–1471,
Chiang Mai, Thailand, November 8 – 13, 2011.

2 Related Work

Query expansion methods analyze user query terms and incorporate related terms automatically (Voorhees, 1994). They are usually divided into local and global methods.

Local methods adjust a query relative to the documents that initially appear to match the query (Manning et al., 2009). Pseudo-relevance Feedback (PRF) is one of the most widely used expansion methods (Rocchio, 1971; Xu and Croft, 1996). This method assumes top-ranked documents to be relevant (and sometimes, also that low-ranked documents are irrelevant), and selects additional query terms from the top-ranked documents.

Global methods are techniques for expanding query terms without checking the results returned by the query. These methods analyze term co-occurrence statistics in the entire corpus or use external knowledge sources to select terms for expansion (Manning et al., 2009). For example, techniques using Word Sense Disambiguation (WSD) techniques and synonyms from WordNet have been used for query expansion with some success (Voorhees, 1994; Liu et al., 2005).

The query expansion method proposed in this paper is a global expansion technique based on WordNet, but in contrast to the previous work based on WordNet, it does not perform WSD and adds related words beyond synonyms.

(Agirre et al., 2010) is the work which is closest to ours. They use the same WordNet-based relatedness method in order to expand documents, following the BM25 probabilistic method for IR, obtaining some improvements, specially when parameters had not been optimized. In contrast to their work, we investigate methods to apply relatedness to query expansion, and we compare the results with pseudo-relevance feedback. Besides, we found that a language modeling (Ponte and Croft, 1998) approach to IR combined with inference networks (Turtle and Croft, 1991) offered more flexibility for query expansion.

Our work stems from the use of random walks over the WordNet graph to compute the relatedness between pairs of words (Hughes and Ramage, 2007). In this work a single word was input to the random walk algorithm, obtaining the probability distribution over all WordNet synsets. The similarity of two words was computed as the similarity of the distributions of each word. In later work, (Agirre et al., 2009b) tested different configurations of the graph, and obtained the best results for a WordNet-based system, comparable to the results of a distributional similarity method which used a crawl of the entire web. The same authors later released their UKB software, which is the one we use here.

3 Relatedness-based Query Expansion (RQE)

The key insight of our model is to expand the query with related words according to the background information in WordNet (Fellbaum, 1998), which provides generic information about general vocabulary terms.

In contrast with previous work using WordNet, we select those concepts that are most closely related to the query as a whole. To this end, we follow the approach in (Agirre et al., 2010), which, based on random walks over the graph representation of WordNet concepts and relations, obtains concepts related to the documents. We use the same settings and implementation for the graph algorithm, which is publicly available[1]. Details are omitted here due to lack of space, please refer to (Agirre et al., 2010).

In order to select the expansion terms, we choose the top N highest scoring concepts, and get all the words that lexicalize the given concept. We explored several values of N, and tune it in order to get the optimum value, as discussed in Section 4. For instance, given a query like "*What is the lowest speed in miles per hour which can be shown on a speedometer?*", our method suggests related terms like *vehicle*, *distance* and *mph*.

Our retrieval model runs queries which contain the original terms of the query and the expansion terms. Documents are ranked by their probability of generating the whole expanded query (Q_{RQE}), which is given by:

$$P_{RQE}(Q_{RQE} \mid \Theta_D) = P(Q \mid \Theta_D)^w P(Q' \mid \Theta_D)^{1-w} \quad (1)$$

where w is the weight given to the original query and Q' is the expansion of query Q.

The query likelihood probability is estimated following the multinomial distribution:

$$P(Q \mid \Theta_D) = \prod_{i=1}^{|Q|} P(q_i \mid \Theta_D)^{\frac{1}{|Q|}} \quad (2)$$

[1] http://ixa2.si.ehu.es/ukb/

where q_i is a query term of query Q and $|Q|$ is the length of Q. And following the Dirichlet smoothing (Zhai and Lafferty, 2001) we have

$$P(q_i \mid \Theta_D) = \frac{tf_{q_i D} + \mu \frac{tf_{q_i C}}{|C|}}{|D| + \mu} \quad (3)$$

where $tf_{q_i D}$ and $tf_{q_i C}$ are the frequency of the query term q_i in the document D and the entire collection, respectively, and μ is the smoothing free parameter.

The probability of generating the expansion terms is defined as

$$P(Q' \mid \Theta_D) = \prod_{q'_i}^{|Q'|} P(q'_i \mid \Theta_D)^{\frac{w_i}{W}} \quad (4)$$

where q'_i is a expansion term, $W = \sum_{i=1}^{|Q'|} w_i$ and w_i is the weight we give to a expansion term, which we can see as the relatedness between the original query Q and the expansion term, and is computed as

$$w_i = P(q' \mid Q) = \sum_{j=1}^{N} P(q' \mid c_j) P(c_j \mid Q) \quad (5)$$

where c is a concept returned by the expansion algorithm, N is the number of concepts we chose for the expansion, $P(q' \mid c_j)$ is estimated using the sense probabilities estimated from Semcor (i.e. how often the query term q' occurs with sense c_j), and $P(c_j \mid Q)$ is the similarity weight that the mentioned expansion algorithm assigned to c_j concept.

4 Experiments

In order to test the performance of our method we selected several datasets with different domains, topic typologies and document lengths. Table 1 shows some statistics for each.

The first is the English dataset of the **Robust-WSD** task at CLEF 2009 (Agirre et al., 2009a), a typical ad-hoc dataset on news. This dataset has been widely used among the community interested on WSD and WordNet-related methods. The documents in the Robust-WSD comprise news collections from LA Times 94 and Glasgow Herald 95.

The **Yahoo! Answers** corpus is a subset of a dump of the Yahoo! Answers web site, where people post questions and answers, all of which are public to any web user willing to browse them

	docs	length	q. train	q. test	length
Robust	166,754	532	150	160	8.6
Yahoo!	89,610	104	1,000	30,000	11.7
ResPubliQA	1,379,011	20	100	500	12.2

Table 1: Number of documents, average document length, number of queries for train and test in each collection, and average query length.

	QL	PRF				RQE		
	μ	μ	d	t	w	μ	N	w
Rob	1000	1000	10	50	0.3	2000	100	0.5
Yah	200	200	2	20	0.8	200	50	0.7
Res	100	100	10	30	0.8	100	125	0.7

Table 2: Optimal values in each dataset for free parameters.

(Surdeanu et al., 2008). The document set was created with the best answer of each question (only one for each question). We use the dataset as released by its authors[2].

The other collection is the English dataset of **ResPubliQA** exercise at the Multilingual Question Answering Track at CLEF 2009 (Peñas et al., 2009). The exercise is aimed at retrieving paragraphs that contain answers to a set of 500 natural language questions.

Our experiments were performed using the Indri search engine (Strohman et al., 2005), which is a part of the open-source Lemur toolkit[3].

To determine whether the query expansion model we developed is useful to improve retrieval performance, we set up a number of experiments in which we compared our expansion model with other retrieval approaches. We used two baseline retrieval approaches for comparison purposes. One of the baselines is the default query likelihood (**QL**) language modeling method implemented in the Indri search engine. The other one is pseudo-relevance feedback (**PRF**) using a modified version of Lavrenko's relevance model (Lavrenko and Croft, 2001), where the final query is a weighted combination of the original and expanded queries, analogous to Eq. 1. As in our own model presented in the previous section, we chose the Dirichlet smoothing method for the baselines. We consider **QL** and **PRF** to be strong, reasonable baselines.

All the methods have several free parameters. The PRF model has three: number of documents (d) and terms (t), and w (cf. Eq. 1). The RQE

[2]Check the features of the dataset at Yahoo! Webscope dataset: http://webscope.sandbox.yahoo.com/ ("ydata-yanswers-manner-questions-v1_0")

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

		QL	PRF	Δ QL	RQE	Δ QL	Δ PRF
Rob	MAP	33.22	**36.69**	10.44% ***	33.67	1.36%	-8.22% ***
	GMAP	13.21	**14.38**	8.90% ***	14.34	8.59% **	-0.29%
	P@5	42.50	**43.63**	2.65%	42.25	-0.59%	-3.15%
	P@10	35.31	**37.38**	5.84% ***	35.81	1.42%	-4.18% *
Yah	MRR	26.36	26.40	0.15%	**27.22**	3.26% ***	3.11% ***
	P@5	6.67	6.63	-0.56% **	**6.88**	3.21% ***	3.79% ***
	P@10	3.95	3.96	0.25%	**4.10**	3.91% ***	3.65% ***
Res	MRR	48.77	46.33	-5.00% ***	**49.78**	2.07%	7.44% ***
	P@5	12.44	12.00	-3.54% *	**12.68**	1.93%	5.67% ***
	P@10	**6.80**	6.78	-0.29%	6.78	-0.29%	0.00%

Table 3: Results of all methods. Δ columns show relative improvement with respect to QL or PRF.

model has two parameters: w (cf. Eq.. 1) and N the number of concepts for the expansion (Eq. 5). In addition, all methods use Dirichlet smoothing, which has a smoothing parameter μ. We used the train part of each dataset to tune all these parameters via a simple grid-search. The μ parameter was tested on the [100,1200] range for ResPubliQA and Yahoo! and [100,2000] for Robust, with increments of 100. The w parameter ranged over [0,1] with 0.1 increments. The d parameter ranged over [2,50] and the t and N in the range [1,200] (we tested 10 different values in the respective ranges). The parameter settings that maximized mean average precision for each model and each collection are shown in Table 2.

5 Results

Our main results are shown in Table 3. The main evaluation measure for Robust is Mean Average Precision (MAP), as customary. In two of the datasets (Yahoo! and ResPubliQA), there is a single correct answer per topic, and therefore we use Mean Reciprocal Rank (MRR). We also report Mean Precision at ranks 5 and 10 (P@5 and P@10). GMAP is also included (we will introduce and mention it afterwards). Statistical significance was computed using Paired Randomization Test (Smucker et al., 2007). In the tables throughout the paper, we use * to indicate statistical significance at 90% confidence level, ** for 95% and *** for 99%.

QL and PRF. The first two columns in Table 3 shows the results for QL and PRF and the performance difference between them. The results for PRF are mixed. It is very effective in the Robust dataset, with dramatic improvements, specially in MAP. All differences are statistical significant, except for P@5. In Yahoo! the improvement is small in MRR and P@10, without statistical significance, but P@5 is lower. In ResPubliQA the results are bad, with statistical significant degradation in MRR.

RQE. Continuing rightwards with Table 3, the following columns show the results for RQE, together with its difference with respect to QL and PRF. Note that figures in bold mean the best performance for each metric. It can be seen that, although RQE is not effective for Robust, it is the best method for Yahoo! and ResPubliQA. Moreover, the improvements over QL, and also over PRF, for Yahoo! are all statistical significant.

PRF is known to perform well for some topics and datasets but not for others. Table 3 includes results for the geometrical mean, GMAP (Robertson, 2006), in the Robust dataset, as it is not relevant in the other datasets. GMAP tries to promote systems which are able to perform well for all topics, in contrast to systems that perform better in some but worse in others. The figures show that RQE approximate the performance of PRF, showing that it perform better for difficult topics.

Combining PRF and RQE. In a preliminary experiment, we added the expansion terms produced both by RQE and PRF, obtaining a **MAP of 37.67** in the Robust collection, the best result. We would like to explore the potential for combination further in the future.

6 Conclusions

Motivated by the recent success of knowledge-based methods in word similarity and relatedness tasks (Agirre et al., 2009b), we explored a generic method to improve IR results using WordNet-based query expansion, and compared it to baseline query likelihood and pseudo-relevance feedback methods.

Our results on a diverse range of ad-hoc datasets with different domains, topic typologies and document lengths show that our method improves over a query likelihood baseline in all three datasets, while Pseudo Relevance Feedback is beneficial in only two datasets. Our method compares favorably to PRF in two datasets, and, in a preliminary experiment, the combination of PRF and our method yielded the best results in the third dataset.

In the future, we would like to analyze the differences between PRF and our method, and explore further combinations. We would also like to use our method on domains where large lexical resources are available, such as UMLS (Humphreys et al., 1998) and linked data repositories

Acknowledgments

This work has been supported by KNOW2 (TIN2009-14715-C04-01). Arantxa Otegi's work is funded by a PhD grant from the Basque Government. Part of this work was done while Arantxa Otegi was visiting ILPS group of the University of Amsterdam.

References

E. Agirre, G. M. Di Nunzio, T. Mandl, and A. Otegi. 2009a. CLEF 2009 Ad Hoc Track Overview: Robust - WSD Task. In *Working Notes of the Cross-Lingual Evaluation Forum*.

E. Agirre, A. Soroa, E. Alfonseca, K. Hall, J. Kravalova, and M. Pasca. 2009b. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In *Proc. of NAACL*, Boulder, USA.

E. Agirre, X. Arregi, and A. Otegi. 2010. Document expansion based on WordNet for robust IR. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 9–17, Stroudsburg, PA, USA. Association for Computational Linguistics.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database and Some of its Applications*. MIT Press, Cambridge, Mass.

T. Hughes and D. Ramage. 2007. Lexical semantic relatedness with random graph walks. In *Proceedings of EMNLP-CoNLL-2007*, pages 581–589.

L. Humphreys, D. Lindberg, H. Schoolman, and G. Barnett. 1998. The Unified Medical Language System: An Informatics Research Collaboration. *Journal of the American Medical Informatics Association*, 1(5):1–11.

V. Lavrenko and W. B. Croft. 2001. Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 120–127, New York, NY, USA. ACM.

S. Liu, C. Yu, and W. Meng. 2005. Word sense disambiguation in queries. In *Proceedings of CIKM '05*, pages 525–532.

C. D. Manning, P. Raghavan, and H. Schütze. 2009. *An introduction to information retrieval*. Cambridge University Press, UK.

A. Peñas, P. Forner, R. Sutcliffe, A. Rodrigo, C. Forăscu, I. Alegria, D. Giampiccolo, N. Moreau, and P. Osenova. 2009. Overview of ResPubliQA 2009: Question Answering Evaluation over European Legislation. In *Working Notes of the Cross-Lingual Evaluation Forum*.

J. M. Ponte and W. B. Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281, New York, NY, USA. ACM.

S. Robertson. 2006. On GMAP: and other transformations. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 78–83, New York, NY, USA. ACM.

J. J. Rocchio. 1971. Relevance feedback in information retrieval. In G. Salton, editor, *The Smart retrieval system - experiments in automatic document processing*, pages 313–323. Englewood Cliffs, NJ: Prentice-Hall.

M. D. Smucker, J. Allan, and B. Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proc. of CIKM 2007*, Lisboa, Portugal.

T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. 2005. Indri: a language-model based search engine for complex queries. Technical report, in Proceedings of the International Conference on Intelligent Analysis.

M. Surdeanu, M. Ciaramita, and H. Zaragoza. 2008. Learning to Rank Answers on Large Online QA Collections. In *Proceedings of ACL 2008*.

H. Turtle and W. B. Croft. 1991. Evaluation of an inference network-based retrieval model. *ACM Trans. Inf. Syst.*, 9:187–222, July.

E. M. Voorhees. 1994. Query expansion using lexical-semantic relations. In *Proceedings of SIGIR '94*, page 69.

J. Xu and W. B. Croft. 1996. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pages 4–11, New York, NY, USA. ACM.

C. Zhai and J. Lafferty. 2001. A study of smoothing methods for language models applied to Ad Hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 334–342, New York, NY, USA. ACM.