Knowl Inf Syst (2015) 44:689–718
DOI 10.1007/s10115-014-0785-4

REGULAR PAPER

Using knowledge-based relatedness for information retrieval

Arantxa Otegi · Xabier Arregi · Olatz Ansa · Eneko Agirre

Received: 18 January 2013 / Revised: 10 February 2014 / Accepted: 7 September 2014 /
Published online: 20 September 2014

Abstract Traditional information retrieval (IR) systems use keywords to index and retrieve documents. The limitations of keywords were recognized since the early days, specially when different but closely related words are used in the query and the relevant document. Query expansion techniques like pseudo-relevance feedback (PRF) and document clustering techniques rely on the target document set in order to bridge the gap between those words. This paper explores the use of knowledge-based semantic relatedness techniques to overcome the vocabulary mismatch between the query and documents, both on IR and Passage Retrieval for question answering. We performed query expansion and document expansion using WordNet, with positive effects over a language modeling baseline on three datasets, and over PRF on two of those datasets. Our analysis shows that our models and PRF are complementary; in that, PRF is better for easy queries, and our models are stronger for difficult queries and that our models generalize better to other collections, being more robust to parameter adjustments. In addition, we show that our method has a positive impact in an end-to-end question answering system for Basque and that it can be readily applied to other knowledge bases, as our good results using Wikipedia show, paving the way for the use of other knowledge structures such as medical ontologies and linked data repositories.

Keywords Knowledge-based systems · Semantic similarity · Semantic relatedness · Information retrieval · Query and document expansion

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

A. Otegi (✉) · X. Arregi · O. Ansa · E. Agirre
IXA Group, University of the Basque Country UPV/EHU, Manuel Lardizabal 1,
20018 Donostia, Basque Country
e-mail: arantza.otegi@ehu.es

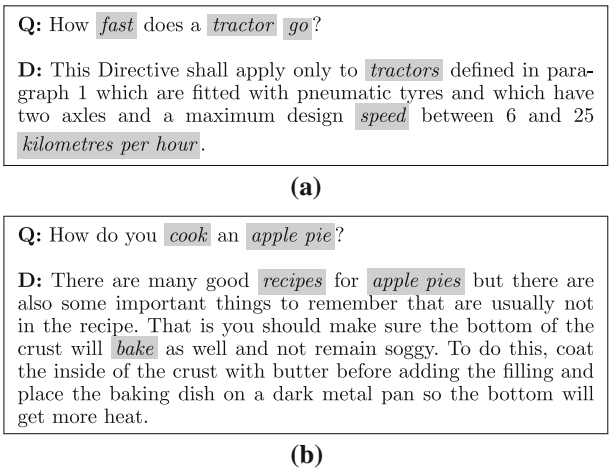


Fig. 1 Examples of lexical gaps (ResPubliQA and Yahoo! datasets, respectively)

lexical gap between the query and the document. Lexical gaps are also a problem for question answering, as well as related passage-retrieval and answer-finding systems [38,43]. Those systems face the task of recovering short pieces of information that satisfy the users' needs, passages or exact answers, respectively, instead of whole documents. They also differ in the nature of the queries: While IR queries are usually composed by a few keywords, the queries on question answering scenarios are formulated as natural language questions. Furthermore, as noted in [10], users who submit questions to a question answering system can not be expected to anticipate the lexical content of an optimal response, and there is often little overlap between the terms in the question and those appearing in its answer. The datasets we have used in this work have been selected to test the effects of the lexical gap problem in ad-hoc IR, passage-retrieval and answer-finding tasks.

Figure 1 shows two examples from two of the datasets used in this article exemplifying lexical gaps. In each example, there is a query (Q) and a relevant document (D), which answers the question using different but related words. For example, the question in Fig. 1a contains *fast*, *tractor* and *go*. Only one of these words appears in the document (*tractor*), but other words related to the query are also present, such as *speed* and *kilometers per hour*. Something similar happens on Fig. 1b, where the query keyword *cook* does not occur on the document, which does contain related words such as *recipes* or *bake*.

In order to bridge the gap, IR has resorted to distributional models. Most research concentrated on Query Expansion (QE) methods, which typically analyze term co-occurrence statistics in the corpus and/or in the highest scoring documents in order to select terms for expanding the query [32]. PRF is one of the most notorious techniques in this area. Document expansion (DE) is a natural alternative to QE. Several researchers have used distributional methods from similar documents in the collection in order to expand the documents with related terms that do not actually occur in the document [22,26,31,33,53]. The work presented here is complementary to those works; in that, we explore QE and DE, but use relatedness-based methods (on WordNet) instead of distributional ones.

As an alternative to distributional methods, WordNet has been used with great success in psycholinguistic datasets of word similarity and relatedness, where it surpasses distributional methods based on keyword matches [3,6]. Table 1 shows the relatedness between some words

Table 1 Relatedness between sample words in queries and documents in Fig. 1a, b

	Tractor	Speed	kmh	Recipe	Apple pie	Bake
Fast	**1.34**	**8.09**	1.11	**1.24**	0.78	1.17
Cook	1.68	1.90	0.87	**5.93**	**2.57**	**4.31**

The numbers are produced by a WordNet-based relatedness software [6]. Numbers are scaled by 10^3. Three highest relatedness numbers in each row in bold

Table 2 Scores for selected words in documents from Fig. 1, returned by random walks initialized with the queries [6]

	Tractor	Speed	kmh	Recipe	Apple pie	Bake
How fast does a tractor go?	**408.59**	**7.06**	**−0.02**	−0.04	−0.03	−0.18
How do you cook an apple pie?	−0.03	−0.11	−0.03	**0.48**	**13.28**	**14.14**

Higher scores indicate higher relatedness to the query words. Numbers are scaled by 10^3. Three highest scores in each row in bold

in the queries and documents in Fig. 1, as returned by the WordNet relatedness software proposed in [6]. As the table shows, the word which is most related to *fast* among the highlighted words in the documents is *speed* and the word most related to *cook* is *recipe*. Given these relatedness scores, each query could be paired with the corresponding document automatically. This example shows the motivation of our approach, where we want to use WordNet-based relatedness to bridge the lexical gap.

WordNet has been applied to IR before. Some authors extended the query with synonyms from WordNet [30,54], while others have explicitly represented and indexed word senses after performing word sense disambiguation (WSD) [19,25,50]. More recently, the WSD-Robust task at CLEF[1] provided queries and documents with automatically disambiguated word senses, where some high-scoring participants reported significant improvements when using WordNet information [4,17].

The work reported here is novel in that we use WordNet-based relatedness beyond synonymy for query and document expansion. As computing and using word-by-word relatedness as in Table 1 is a costly process, we compute the related words for whole queries or documents instead. Given a query (or full document), a relatedness algorithm using random walks over the WordNet graph [6] returns the concepts, which are closely related to the words in the query (or document). This is in contrast to previous WordNet-based works, which focused on WSD to replace or supplement words with their senses. Our method discovers important concepts, even if they are not explicitly mentioned in the query or document. Table 2 shows that using this technique for the two queries in Fig. 1, the words in the respective documents get the highest scores.

In this work, we adopt a language modeling framework to implement the query likelihood (QL) and PRF baselines, as well as our relatedness-based query expansion and document expansion methods. In order to test the performance of our method, we selected several datasets with different domains, topic typologies and document lengths, including ad-hoc IR, passage retrieval and answer finding. Given the relevance among the community using WordNet-related methods, we selected the Robust-WSD dataset from CLEF [4], which is a typical ad-hoc dataset on news.

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

We think that our method is specially relevant for question answering on specific document collections,[2] as failing to retrieve the passage which contains the answer negatively affects the performance of the whole system. We thus evaluated our system on widely used answer-finding and passage-retrieval datasets. The first is the Yahoo! Answers dataset, which contains questions and answers by real users on diverse topics [52]. The second is ResPubliQA, a passage-retrieval task on European Union laws organized at CLEF [41]. In addition, we applied our method to an in-house question answering system on Science and Technology documents in Basque, showing significant improvement in the end-to-end results.

The results show that our methods provide improvements in all datasets when compared to the QL baseline and that they compare favorably to PRF in most datasets. The analysis suggests that our models and PRF are complementary, in that PRF improves results for easy queries and our models are stronger for difficult queries. We also show that our models are more robust in face of sub-optimal parameters.

Finally, we will show that our method can be applied successfully to other knowledge bases, as exemplified by a knowledge base extracted from Wikipedia, paving the way for the use of medical ontologies and arbitrary linked data repositories in query expansion.

The work presented in this article follows [1], which used the same WordNet-based relatedness algorithm for document expansion but in a probabilistic setting, and [39], where we explored query expansion. In the present work, we subsume both works providing an implementation on a language modeling framework for IR and provide additional analysis, including the factors that affect the performance of the algorithm, qualitative analysis of the concepts produced, an application to question answering, and applying the method to Wikipedia.

The article is structured as follows. After the introduction, Sect. 2 introduces the random-walk model and the relatedness-based models for query and document expansion. Section 3 presents the experimental setup. Section 4 shows our main results, followed by a section on the analysis of performance factors. In Sect. 6, we show how our method can make use of Wikipedia instead of WordNet. Section 7 does some qualitative analysis of the expansion terms produced by our system. Section 8 presents the application to an end-to-end question answering system. Section 9 reviews related work. Finally, the conclusions and future work are mentioned.

2 Relatedness-based expansion models

In this section, we describe the relatedness-based method to expand queries and documents, followed by the expansion models we propose for IR.

2.1 Obtaining expansion terms

The key insight of our model is to expand the query or the document with related words according to the background information in WordNet [16], which provides generic information about general vocabulary terms. WordNet groups nouns, verbs, adjectives and adverbs into sets of synonyms (synsets), each expressing a distinct concept. Synsets are interlinked with conceptual semantic and lexical relations, including hypernymy, meronymy and causality.

[2] As opposed to open question answering, which typically searches the web.

In contrast with previous work using WordNet, we select those concepts that are most closely related to the text as a whole. As we will see in the following sections, this text could be a query or a document. For that, we use a technique based on random walks over the graph representation of WordNet concepts and relations [23], which has been successfully used in word similarity [6] and WSD [5], and made publicly available by the authors.[3]

We represent WordNet as a graph as follows: Graph nodes represent WordNet concepts (synsets) and dictionary words; relations among synsets are represented by undirected edges; and dictionary words are linked to the synsets associated to them by directed edges. We use version 3.0, with all relations provided, including the gloss relations. This was the setting obtaining the best results in a word similarity dataset as reported by Agirre [6].

Given a text and the graph-based representation of WordNet, we obtain a ranked list of WordNet concepts as follows: (1) We first preprocess the text to obtain the lemmas and parts of speech of the open category words. (2) We then assign a uniform probability distribution to the terms found in the text. The rest of nodes are initialized to zero. (3) We compute Personalized PageRank [21] over the graph, using the previous distribution as the reset distribution, and producing a probability distribution over WordNet concepts. The higher the probability for a concept, the more related it is to the given text. (4) Given the topology of the graph, some concepts from very dense areas receive high probabilities, regardless of the words used to initialize the random walk. In order to avoid this effect, we run PageRank over the whole graph, which produces a probability independent of the specific target words, and subtracted the resulting probability from each concept. That is, the score of each concept is obtained subtracting the PageRank from the probability returned by Personalized PageRank. Table 2 shows the scores attained by Personalized PageRank when initialized with each of the two queries. The positive scores show that the Personalized PageRank value is higher than that of the PageRank value, indicating high relevance to the query, while negative scores show the contrary.

Basically, Personalized PageRank is computed by modifying the random jump distribution vector in the traditional PageRank equation. In our case, we concentrate all probability mass in the concepts corresponding to the words in the text. Let G be a graph with N vertices $v_1, \ldots, v_N$ and d_i be the outdegree of node i; let M be a $N \times N$ transition probability matrix, where $M_{ji} = \frac{1}{d_i}$ if a link from i to j exists, and zero otherwise. Then, the calculation of the *PageRank vector* $\mathbf{Pr}$ over G is equivalent to resolving Eq. (1).

$$\mathbf{Pr} = cM\mathbf{Pr} + (1 - c)\mathbf{v} \tag{1}$$

In the equation, $\mathbf{v}$ is a $N \times 1$ vector and c is the so-called *damping factor*, a scalar value between 0 and 1. The first term of the sum on the equation models the voting scheme described in the beginning of the section. The second term represents, loosely speaking, the probability of a surfer randomly jumping to any node, e.g., without following any paths on the graph. The damping factor, usually set in the [0.85..0.95] range, models the way in which these two terms are combined at each step.

The second term on Eq. (1) can also be seen as a smoothing factor that makes any graph fulfill the property of being aperiodic and irreducible and thus guarantees that PageRank calculation converges to a unique stationary distribution.

In the traditional PageRank formulation, the vector $\mathbf{v}$ is a stochastic normalized vector whose element values are all $\frac{1}{N}$, thus assigning equal probabilities to all nodes in the graph in case of random jumps. In the case of Personalized PageRank as used here, $\mathbf{v}$ is initialized with uniform probabilities for the terms in the document, and 0 for the rest of terms.

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

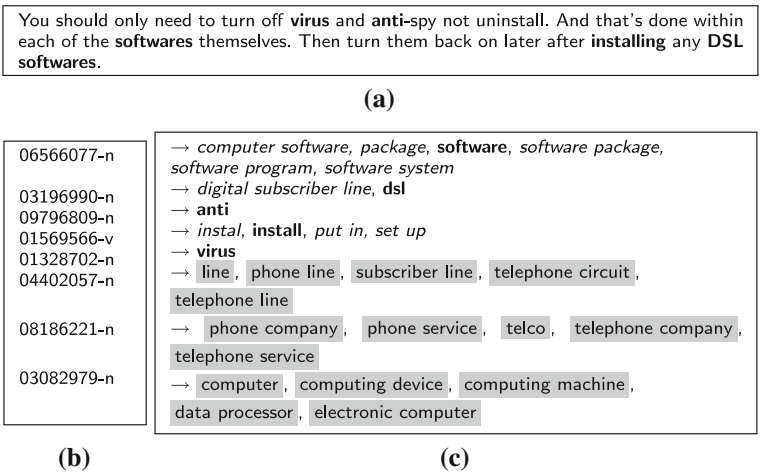


Fig. 2 Example of the expansion for document 1005121303076 of the Yahoo! dataset: **a** original document, **b** *synset* numbers for some of the concepts to be expanded, **c** words obtained from the expansion

PageRank is actually calculated by applying an iterative algorithm, which computes Eq. (1) successively until a fixed number of iterations are executed. In our case, we used a publicly available implementation[4] with the default values provided by the software, i.e., a damping value of 0.85, and 30 iterations.

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. When expanding the documents (cf. Sect. 2.2), we follow the work in [1] and fix N to 100. When expanding the queries (cf. Sect. 2.3), we explore several values of N and tune it in order to get the optimum value, as discussed in Sect. 3.

Figure 2 shows the expansion process for a document. After applying the graph algorithm to the document in 2a, we obtain the concepts with the *synset* numbers, as partially shown in 2b, sorted by relatedness to the document in decreasing order. The words that lexicalize these concepts are shown in Fig. 2c. The words that are in the original document are in **bold**; their synonyms are in *italic*; and other related words are highlighted. In addition to synonyms, words that are not in the document but are related to related concepts are suggested for expansion, as for instance, *phone company* and *computer*.

Similarly, Fig. 3 illustrates query expansion. After applying the graph algorithm to the query in 3a, we obtain the concepts with the *synset* numbers, as partially shown in 3b, sorted by relatedness to the query in decreasing order. The words that lexicalize these concepts are shown in Fig. 3c. We can see that words such as *vehicle* and *distance*, which are not in the query but are related to it, are suggested for expansion.

2.2 Relatedness-based document expansion (RDE)

The RDE approach requires the document collection to be preprocessed to obtain a list of most related terms for each document, following the method explained in Sect. 2.1. These related terms are indexed separately. Documents are ranked by their probability of generating the query [42], where this probability is estimated as a weighted combination of query likelihoods from the different document representations:

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

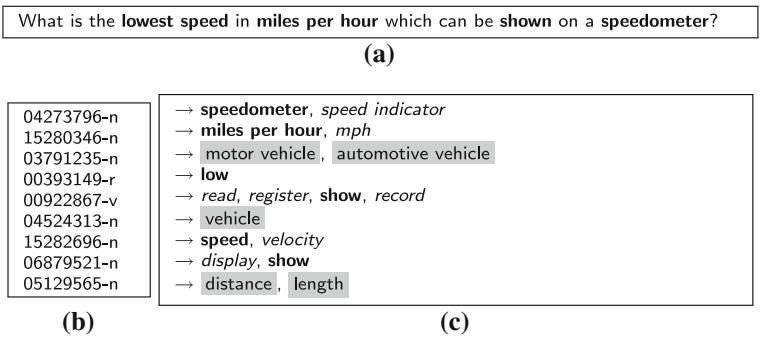


Fig. 3 Example of the expansion for query 91 of the ResPubliQA dataset: **a** original query, **b** *synset* numbers for some of the concepts to be expanded, **c** words obtained from the expansion

$$P_{\text{RDE}}(Q \mid \Theta_{\text{RDE}}) = P(Q \mid \Theta_D)^w P(Q \mid \Theta_E)^{1-w} \tag{2}$$

where Θ_D and Θ_E are the language models estimated from the original document representation and the expanded document representation, respectively, and w is the weight given to the original document language model set in the [0..1] range. Query likelihood is estimated following the multinomial distribution (we show the document model, but the expansion model is analogous):

$$P(Q \mid \Theta_D) = \prod_{i=1}^{|Q|} P(q_i \mid \Theta_D)^{\frac{1}{|Q|}} \tag{3}$$

where q_i is a query term of query Q and $|Q|$ is the length of Q. And following the Dirichlet smoothing [56], we have

$$P(q_i \mid \Theta_D) = \frac{tf_{q_i D} + \mu \frac{tf_{q_i C}}{|C|}}{|D| + \mu} \tag{4}$$

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.

2.3 Relatedness-based query expansion (RQE)

In this approach, we expand each query with the terms obtained following the expansion technique described in Sect. 2.1. Thus, we retrieve documents based on the expanded query, which contains 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:

$$P_{\text{RQE}}(Q_{\text{RQE}} \mid \Theta_D) = P(Q \mid \Theta_D)^w P(Q' \mid \Theta_D)^{1-w} \tag{5}$$

where w is the weight given to the original query and Q' is the expansion of query Q. The query likelihood probability $P(Q \mid \Theta_D)$ is again calculated following a multinomial distribution and Dirichlet smoothing, as specified in Eqs. 3 and 4. The probability of generating the expansion terms is defined as

$$P(Q' \mid \Theta_D) = \prod_{q_i'}^{|Q'|} P(q_i' \mid \Theta_D)^{\frac{w_i}{W}} \tag{6}$$

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

$$w_i = P(q' \mid Q) = \sum_{j=1}^{N} P(q' \mid c_j) P(c_j \mid Q) \tag{7}$$

where c is a concept returned by the expansion algorithm (cf. Sect. 2.1), 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.

3 Experimental setup

In order to test the performance of our method, we selected several datasets from different domains, topics, typologies and document lengths, including ad-hoc IR, passage retrieval and answer finding. Table 3 shows some statistics for the three selected datasets. Note that all three datasets have a separate subset for developing and training the systems.

The first is the English dataset of the **Robust-WSD** task at CLEF 2008 and 2009 [4], a typical ad-hoc dataset on news. This dataset has been widely used among the community interested on WSD and WordNet-related methods for IR, as the organizers run state-of-the-art WSD software on all questions and documents, making it easy to experiment with IR method. Note that we need to reuse existing relevance judgments (customary on standard datasets), which were pooled among participants of the task, and thus, systems that are based on different expansion strategies (e.g., WSD or WordNet) might return relevant documents, which were not available in the pool that was manually judged at competition time. For this reason, the organizers of the Robust-WSD dataset used relevance judgments obtained pooling both monolingual and multilingual runs. The organizers of the exercise hoped that the inclusion of multilingual runs, with a larger variability due to translation strategies, would include relevance judgments for query-document pairs where different wording had been used [4].

The documents in the Robust-WSD comprise news collections from LA Times 94 and Glasgow Herald 95. The topics are statements representing information needs, consisting of three parts: a brief title statement; a one-sentence description; a more complex narrative describing the relevance assessment criteria. Following the rules of the Robust-WSD task, we use the title and the description parts of the topics in our experiments.

As we think that our method is specially relevant for question answering, we also evaluated our methods on an answer-finding dataset, Yahoo! Answers, which contains questions and answers as phrased by real users on diverse topics [52], and a paragraph retrieval task, ResPubliQA, which is related to European Union laws, organized at CLEF [41].

The **Yahoo! Answers** corpus[5] Surdeanu et al. [52] is a subset of a dump of the Yahoo! Answers web site,[6] where people post questions and answers, all of which are available for browsing. The task is to find the document, which contains the answer. Before releasing the dataset, the Yahoo team filtered the dataset as follows: (1) It comprised a subset of the

[5] Yahoo! Webscope dataset: L4—Yahoo! Answers Manner Questions, version 1.0 http://webscope.sandbox.yahoo.com/catalog.php?datatype=l.

[6] http://answers.yahoo.com/.

Table 3 Statistics for each of the datasets: number of documents, average document length, number of queries and average query length

Dataset	Documents		Training queries		Test queries	
	#	Length	#	Length	#	Length
Robust	166,754	532	150	8.37	160	8.64
Yahoo!	89,610	104	1,000	11.32	30,000	11.25
ResPubliQA	1,379,011	20	100	10.22	500	10.71

Table 4 Optimal values in each dataset for free parameters

Dataset	QL	PRF				RDE		RQE		
	μ	μ	d	t	w	μ	w	μ	N	w
Robust	1,000	1,000	10	50	0.3	1,200	0.8	2,000	100	0.5
Yahoo!	200	200	2	20	0.8	200	0.8	200	50	0.7
ResPubliQA	100	100	10	30	0.8	100	0.7	100	125	0.7

questions, selected for their linguistic properties (for example, they all start with "how {to | do | did | does | can | would | could | should}"). (2) Questions and answers of obvious low quality were removed. (3) 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.

The other collection is the English dataset of the **ResPubliQA** exercise at the Multilingual Question Answering Track at CLEF 2009 [41]. The exercise is aimed at retrieving paragraphs that contain answers to a set of 500 natural language questions. The document collection is a subset of the JRC-Acquis Multilingual Parallel Corpus and consists of 21,426 documents in English, which are aligned to a similar number of documents in other languages.[7] For evaluation, we used the gold standard released by the organizers, which contains a single correct passage for each query.

Documents and queries have been lemmatized and tagged with parts of speech. The Robust dataset is already tagged with lemmas and parts of speech. We have used OpenNLP[8] for the other two datasets.

Our experiments were performed using the Indri search engine [51], which is a part of the open-source Lemur toolkit.[9] In order to determine whether the two expansion models we developed are useful to improve retrieval performance, we set up a number of experiments in which we compared our expansion models with other retrieval approaches. We used two baseline retrieval approaches for comparison purposes. One of the baselines is the default **QL** language modeling method implemented in the Indri search engine. The other one is **PRF** using a modified version of Lavrenko's relevance model [27], where the final query is a weighted combination of the original and expanded queries, analogous to Eq. 5. As in our own model presented in the previous sections, 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 parameters: number of documents (d) and terms (t), and w (cf. Eq. 5). The RDE model also has w (cf. Eq. 2).

[7] Note that Table 3 shows the number of paragraphs, which conform the units we indexed.

[8] http://incubator.apache.org/opennlp/.

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

Table 5 Results of all methods

Dataset	Measure	QL	PRF		RDE		RQE	
		Result	Result	Δ QL (%)	Result	Δ QL (%)	Result	Δ QL (%)
Robust	MAP	0.3322	**0.3669*****	10.44	**0.3387****	1.95	**0.3367**	1.36
	GMAP	0.1321	**0.1438*****	8.90	**0.1351**	2.26	**0.1434****	8.59
	P@5	0.4250	**0.4363**	2.65	**0.4300**	1.18	0.4225	−0.59
	P@10	0.3531	**0.3738*****	5.84	**0.3556**	0.71	**0.3581**	1.42
Yahoo!	MRR	0.2636	**0.2640**	0.15	**0.2752*****	4.42	**0.2722*****	3.26
	P@5	0.0667	0.0663**	−0.56	**0.0691*****	3.64	**0.0688*****	3.21
	P@10	0.0395	**0.0396**	0.25	**0.0412*****	4.29	**0.0410*****	3.91
ResPubl.	MRR	0.4877	0.4633***	−5.00	**0.4926**	1.02	**0.4978**	2.07
	P@5	0.1244	0.1200*	−3.54	0.1236	−0.64	**0.1268**	1.93
	P@10	0.0680	0.0678	−0.29	**0.0694**	2.06	0.0678	−0.29

Δ columns show relative improvement with respect to QL. Bold means better than QL

The RQE model has two parameters: w (cf. Eq. 5) and N the number of concepts for the expansion (Eq. 7). 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, 1,200] range for ResPubliQA and Yahoo! and [100, 2,000] 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 4.

4 Results

In this section, we present the results for the baseline QL model, PRF and our relatedness-based query and document expansion models. 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). Note that in this setting, MAP is identical to MRR. We also report Mean Precision at ranks 5 and 10 (P@5 and P@10). GMAP is also included, and we will introduce and mention it afterward. Statistical significance was computed using Paired Randomization Test [49]. In the tables throughout the paper, we use * to indicate statistical significance for 90 % confidence level, ** for 95 % and *** for 99 %.

4.1 Comparison with respect to QL

Our main results are shown in Table 5. The first three columns of results in Table 5 show the results for QL and PRF, and the performance difference between them. The results for PRF are mixed. In Yahoo!, the improvements are 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. In contrast, it is very effective in the Robust dataset, with dramatic improvements, specially in MAP. This finding is common for relevance feedback algorithms, which is a recall-enhancing technique at the cost of precision [32,47]. The results for PRF in Robust are partly consistent with this statement, as apart from improving recall (5.81 %

Table 6 Results of PRF, RDE and RQE

Dataset	Measure	PRF	RDE		RQE	
		Result	Result	Δ PRF (%)	Result	Δ PRF (%)
Robust	MAP	0.3669	0.3387***	−7.69	0.3367***	−8.22
	GMAP	0.1438	0.1351**	−6.10	0.1434	−0.29
	P@5	0.4363	0.4300	−1.43	0.4225	−3.15
	P@10	0.3738	0.3556***	−4.85	0.3581*	−4.18
Yahoo!	MRR	0.2640	**0.2752****	4.26	**0.2722****	3.11
	P@5	0.0663	**0.0691****	4.22	**0.0688****	3.79
	P@10	0.0396	**0.0412****	4.03	**0.0410****	3.65
ResPubliQA	MRR	0.4633	**0.4926****	6.33	**0.4978****	7.44
	P@5	0.1200	**0.1236**	3.00	**0.1268****	5.67
	P@10	0.0678	**0.0694**	2.36	0.0678	0.00

Δ columns show relative improvement with respect to PRF. Bold means better than PRF

not shown in table), PRF also improves precision at early rank (in a less degree, but still significant for P@10). Note that all differences for PRF at Robust are statistically significant, except for P@5. As MAP encapsulates both precision and recall aspects, it is the one with largest improvement. In the other two datasets, there is one relevant document for each query, and recall is thus irrelevant.

Continuing rightwards on Table 5, the last columns show the results for RDE and RQE, together with their difference with respect to QL. RDE and RQE improve QL in nearly all datasets and measures. The strongest improvements are in Yahoo!. For Robust, the improvements in precision are not so substantial, but the recall improvements are significant, 1.36 % for RDE and 4.67 % for RQE (not shown in table).

4.2 Comparison with respect to PRF

Results of PRF, RDE and RQE are repeated in Table 6 to better compare results with respect to PRF. Note that figures in bold mean better performance than PRF. We can see that the best results vary across datasets, with PRF yielding the best results for Robust, RDE for Yahoo! and RQE for ResPubliQA. Both RDE and RQE improve over PRF in Yahoo! and ResPubliQA, with mostly statistically significant differences.

PRF is known to perform well for some topics and datasets but not for others [47]. We have included results for the GMAP in the Robust dataset (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 [44]. The figures show that RQE gets worse results for MAP but approximates the performance of PRF for GMAP. In the next section, we will analyze the results per query, and we will see that RDE and RQE perform better for some queries, concretely, for difficult queries.

5 Performance factors

In order to understand the behavior of our method, we performed some detailed analysis. First of all, we analyze the performance of each technique on a query by query basis. Next, we study the intercollection generalization of each technique, followed by their sensitivity

to model parameters. We will also check the associated computational costs. Finally, we will summarize the main factors.

5.1 Performance by query

We first compare the performance of RDE and RQE with respect to PRF, calculating, for each query, the difference with respect to PRF in terms of average precision (ΔAP). We sorted the queries by decreasing ΔAP, grouped the queries according to ΔAP ranges, and plotted the number of queries falling into each bucket, as shown in Fig. 4. A positive difference indicates an improvement over PRF for those queries.

The plot for Robust confirms that PRF performs better than RDE and RQE in this dataset, with more queries with negative ΔAP, but note that our expansion models outperform PRF for some of the queries. The situation is reversed for Yahoo! and ResPubliQA, with more queries getting worse results with PRF. In addition, the plots show that for ResPubliQA the majority of queries get the same performance with either method (Δ AP equals 0). In Yahoo! the trend is similar, but less steep. Surprisingly, in the Robust dataset, the number of queries getting the same performance is very low, showing that PRF and our methods are complementary. The plots of our methods versus the QL baseline show the same trends. We have omitted them for the sake of brevity.

In order to study the behavior of our expansion models with respect to easy and hard topics, Fig. 5 shows the performance of each query according to MAP (MRR for Yahoo! and ResPubliQA) obtained by our expansion methods (vertical axis) and PRF (horizontal axis). Hard queries are those which get low performance and are located close to the origin, on the bottom-left quadrant. The best fitting line for the Robust plots shows that PRF does better than RDE and RQE on easy queries (i.e., those with high performance, on the right), but the performance on difficult queries is better for RQE and specially RDE. The best fitting lines for Yahoo! and ResPubliQA also show that the RDE and RQE are performing better on difficult queries. It thus seems that PRF and our expansion techniques are complementary, with one doing better on easy queries and the others doing better on hard queries. The plots with respect to the QL baseline (not shown for the sake of brevity) are very similar, with RQE and RDE doing specially better for queries with low performance.

Figure 6a shows an example of a difficult query from ResPubliQA. Both QL and PRF obtain a low MRR for this query (0.33), while RQE gets a perfect score of 1. Figure 6b shows some of the words proposed by RQE for query expansion. The expansion words include *vehicle*, *distance* and *mph*, which are contained in the relevant document (cf. Fig. 6c).

5.2 Intercollection generalization

In Table 4, we showed the optimum parameters for each technique and dataset, developed according to cross-validation results on the training subset of each dataset. In most practical situations, though, there are no training data to adjust the parameters, and parameters estimated on other scenarios are used, with some performance loss.

In this section, we analyze the behavior of the methods when parameters adjusted on other datasets are used. This analysis was named *intercollections generalization* in [35]. Metzler proposed to measure generalization properties of a model by computing the effectiveness ratio, which is the ratio of the observed effectiveness of a target model with respect to the optimal effectiveness (when optimal values in train are used). Thus, an effectiveness ratio of 100 % represents a model that generalizes optimally. We take a simpler approach and apply the idea directly to MAP (or MRR) values, obtaining a MAP (or MRR) ratio

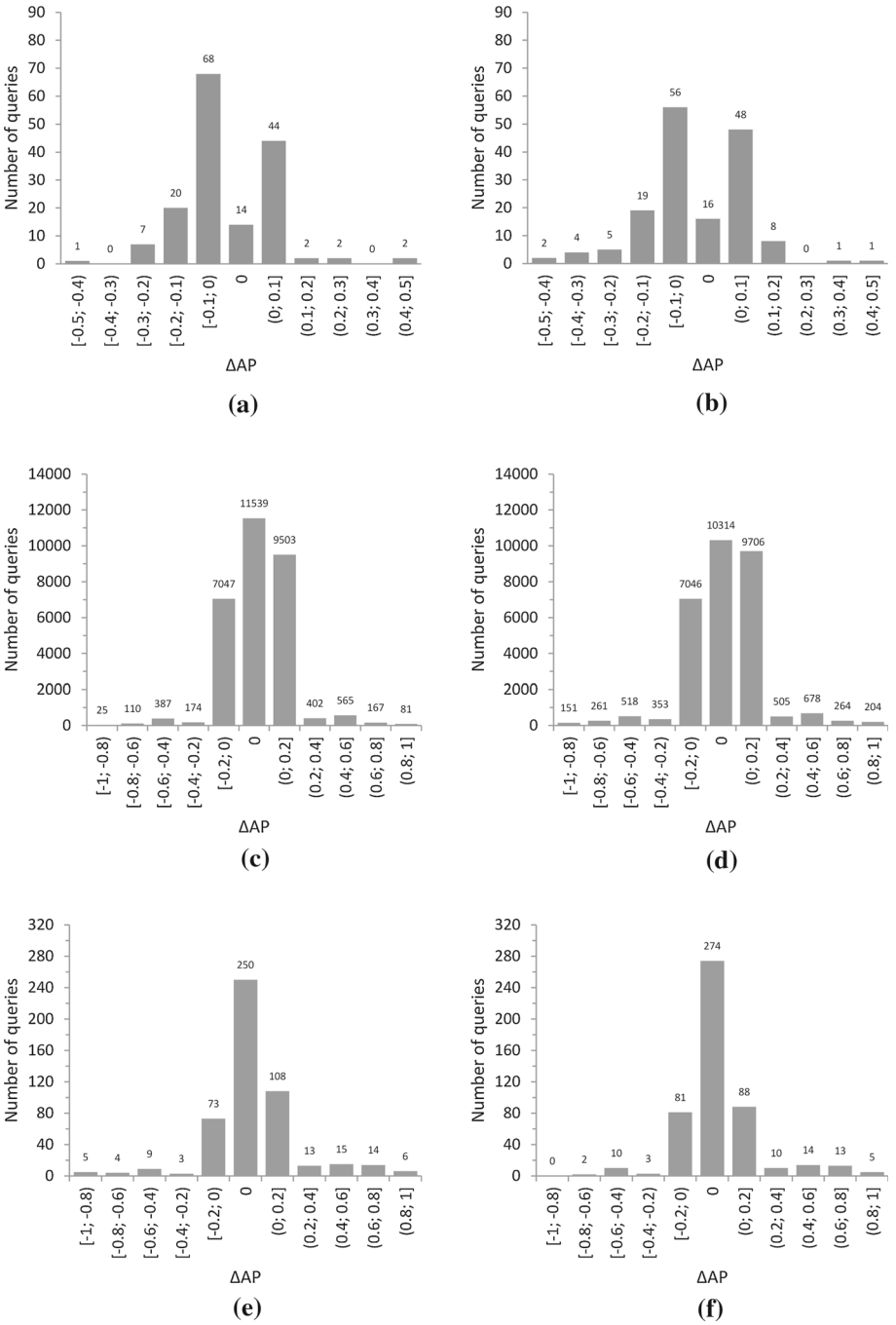


Fig. 4 Queries grouped by differences in improvement over PRF for all datasets. **a** RDE over PRF in Robust, **b** RQE over PRF in Robust, **c** RDE over PRF in Yahoo!, **d** RQE over PRF in Yahoo!, **e** RDE over PRF in ResPubliQA, **f** RQE over PRF in ResPubliQA

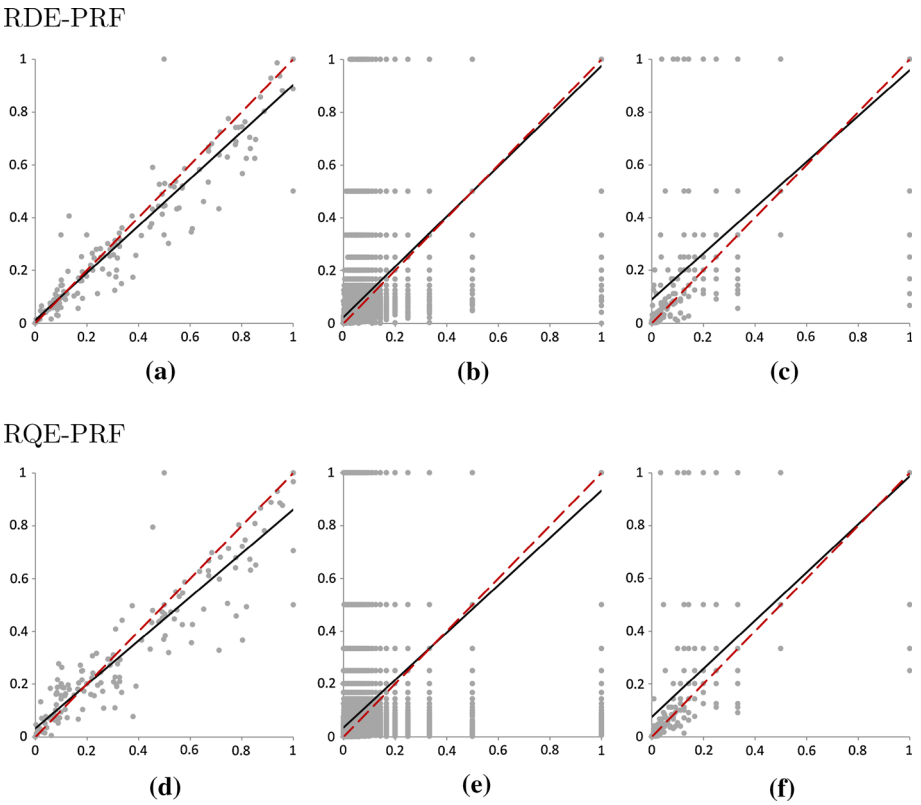


Fig. 5 MAP plots (MRR for Yahoo! and ResPubliQA) of all queries, comparing RDE and RQE (y axis) to PRF (x axis). RDE plots on the *top row*, RQE on the *bottom*. Best fitting linear trend (*solid*) and equality ($y = x$, *dashed*) *lines* are also shown. **a, d** Robust, **b, e** Yahoo!, **c, f** ResPubliQA

What is the lowest speed in miles per hour which can be shown on a speedometer?

(a)

speedometer speed_indicator miles_per_hour **mph** motor_vehicle automotive_vehicle low read register show record **vehicle** speed velocity display show **distance** length

(b)

where a **vehicle** is intended for sale in a Member State where imperial **distances** are used, the speedometer must also be graduated in **mph** (miles per hour), with subdivisions of 1, 2, 5 or 10 **mph**. Marked numerical speed value intervals must not exceed 20 **mph** and must begin at either 10 **mph** or 20 **mph**;

(c)

Fig. 6 A difficult query from ResPubliQA which has been correctly answered with query expansion, including expansion terms proposed by relatedness and the relevant document. **a** The English query (number 91), **b** some of the words obtained by query expansion, **c** a relevant document for the query (jrc32000L0007-en/92)

Table 7 Effectiveness ratios for *intercollections generalization* (based on MAP or MRR)

	PRF			RDE			RQE		
	Rob	Yah	Res	Rob	Yah	Res	Rob	Yah	Res
Robust	–	−9.7	−18.8	–	0.0	1.0	–	−4.3	−7.6
Yahoo!	−6.3	–	1.7	0.0	–	1.0	0.5	–	−0.4
ResPubliQA	−7.3	−0.9	–	−0.7	−0.7	–	0.9	1.3	–
Average	−6.9 %			0.11 %			−1.60 %		

The first column specifies the training dataset for the respective row and the columns the test dataset. Empty slots correspond to the reference (0.0 %). The average row shows the macro-average of all differences above it

for each combination of training/testing datasets, and macro-averaging across all possible combinations (cf. Table 7). Note that, in order to keep the analysis simpler, we kept μ fixed at the optimal values. The smoothing parameter μ has a direct relation with document length and can be thus adjusted according to past experiences easily.

For instance, the *Rob* column for PRF shows a negative ratio of − 6.3 when Yahoo! is used to estimate the parameters and the system is tested on Robust, meaning that the performance is 6.3 % less than when using parameters estimated on the training subset of Robust. The figures in the table show that RDE is the least sensitive to optimization (it actually improves performance), with RQE losing some performance and PRF with the largest losses, −6.9 %.

5.3 Sensitivity to model parameters

We will now explore the sensitivity of the results to changes in the parameters. For that purpose, we will display the effects of the results for different models varying one parameter each time, maintaining the other parameters in their optimal values. This analysis has been performed on the training subsets of the dataset, and thus, the figures reported here are not directly comparable to those obtained on the test datasets.

5.3.1 The number of terms

The main parameter when expanding queries is the number of terms that are added to the query. Figure 7 shows the behavior of PRF and RQE with respect of the number of query terms, when keeping the other parameters fixed. Figure 7a shows that PRF behaves differently on each datasets, with maximum performances at different points. Figure 7b shows that, for RQE, all datasets respond similarly to each newly added term, growing steadily until they plateau at around 20–75 concepts.

5.3.2 The weight of the original query or document

Figure 8 displays the effect of varying the weight of the original query or document. For PRF, we observe that the best value is very different for each dataset, with approximately 0.3 for Robust and 0.8 for the other two. RDE obtains the best result for similar values, around 0.7 or 0.8. For RQE, the best results range from 0.5 for Robust to 0.7 for the other two. RDE shows the most consistent behavior from all three methods, with PRF behaving worst.

Note that when the weight for the original query or the original document language model is 0, the results show the performance of using PRF or expansions alone. PRF terms seem

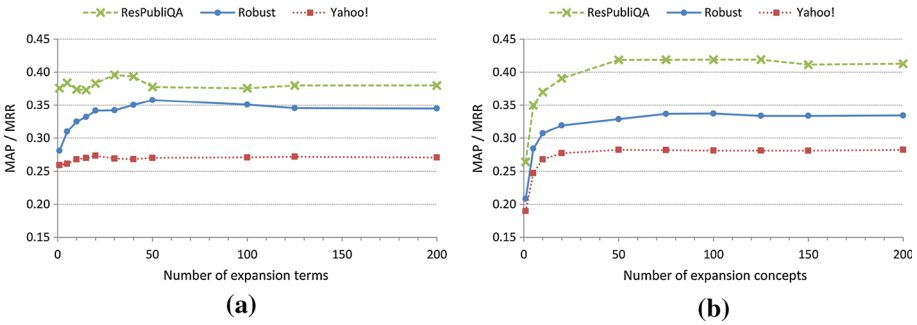
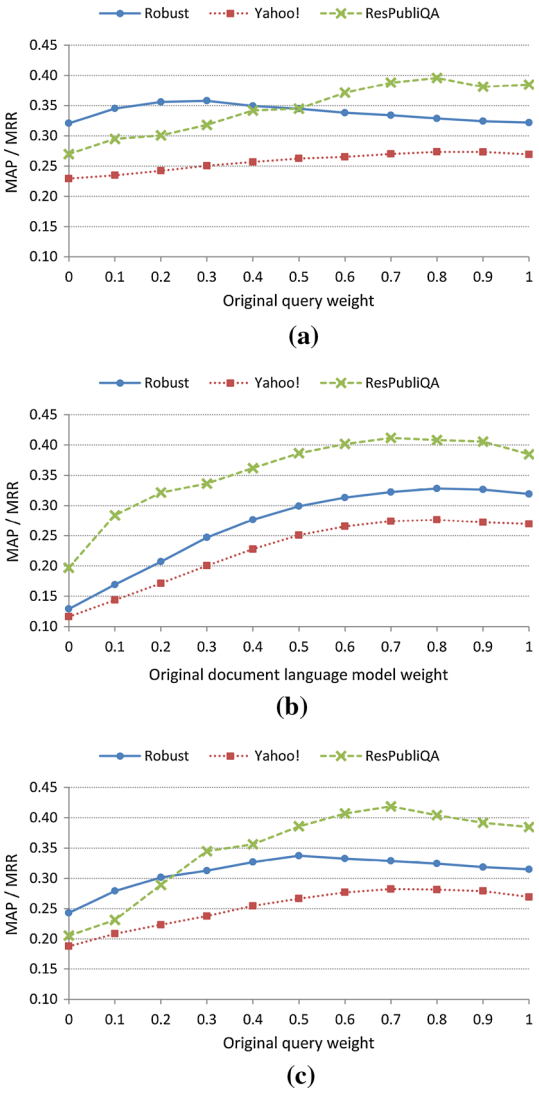


Fig. 7 Results for varying the number of expansion terms for each of the models. **a** PRF, **b** RQE

Fig. 8 Plots of the results when varying the weight of the original query for each of the models. **a** PRF, **b** RDE, **c** RQE



to yield good results on their own, with RQE terms performing slightly worse. Regarding RDE, when the weight is 0, the query is processed using the terms that expand the document alone, and the results are very low.

5.4 Computational cost

Improved performance comes at a computational cost. A query of the Robust test set (160 queries) takes 0.14 s on average for the QL baseline on a server with two Intel QuadCore Xeon X5460 processors at 3,160 MHz with 32 GB of memory. PRF takes 2.75 s, RDE 0.37 s, and RQE 8.53 s per query on average. The larger cost for PRF and RQE at query time comes from the added complexity of examining additional terms in the expanded query. Given that RQE is using more terms than PRF, the cost is higher. The added cost for RDE is the overhead of searching in two indexes and merging the results from both indexes.

In addition, running the random walk on one query or document takes approximately 6 s. In the case of RDE, the process can be easily parallelized and done in batch in advance. In the case of RQE, query time computations could be sped up using less iterations in the random-walk algorithm, or we could have precomputed the random walks for each word in advance. In the later case, at query time, one would just need to do a linear combination of the probability vectors of the words in the query. For the future, we would like to check whether there is any performance loss involved in these computational improvements.

5.5 Summary

We have shown that our two methods provide improvements in all three datasets when compared to the QL baseline. PRF is beneficial in two datasets, but degrades performance in ResPubliQA. RDE and RQE compare favorably to PRF in two datasets, but perform worse in Robust. Our analysis shows that our models and PRF are complementary; in that, PRF is better for easy queries and our models are stronger for difficult queries. Note also that, in the robust dataset, there are very few queries where PRF and our models perform equally, underscoring the possibilities for future combinations. RQE and specially RDE generalize very well across collections, with PRF suffering 7 % on average. The analysis of each individual parameter also shows that RQE and RDE behave nicely regarding the number of terms and the weight of the original query or document. The analysis of performance shows that, at query time, RDE is the most efficient. Finally, we have shown that our method is implicitly doing WSD and that it could possibly be improved using other WSD methods.

6 Using Wikipedia as a knowledge base

Our method is generic in that it can be applied to any graph that links together related words. We are specifically interested in using the knowledge stored in knowledge bases like medical ontologies [24] or those derived from Wikipedia (e.g., Freebase,[10] DBpedia,[11]). As an illustration, we will apply our RQE method to a knowledge base extracted from Wikipedia articles and hyperlinks. We selected RQE because it is less computationally demanding than RDE, which requires processing the whole document collection.

[10] http://www.freebase.com.

[11] http://dbpedia.org.

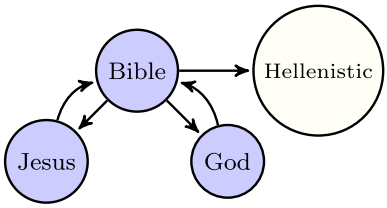


Fig. 9 Simplified example of reciprocal and unidirectional hyperlinks between Wikipedia articles. We create a relation between articles when there exists a reciprocal hyperlink

6.1 Building the hyperlink graph

Wikipedia can be used to derive a knowledge base, as exemplified by several resources, including DBpedia and Freebase. We follow their approach, where articles in Wikipedia correspond to concepts and the titles correspond to the lexicalization of those concepts. Regarding the relations between concepts, we use a simpler setup, where the hyperlinks between articles are interpreted to indicate a generic "related" relation between the source and target articles. Freebase and DbPedia, in comparison, use a richer set of relations, derived using several heuristics from InfoBoxes, which we plan to use in future work. We will now explain how we extract the concepts (articles), the relations between concepts and the dictionary listing the lexicalization of the concepts.

We start from a Wikipedia dump (5th April, 2011), retaining article pages and discarding redirect, disambiguation and category pages, and mining hyperlinks between articles. Given the large number of hyperlinks, we say that there is a relation between the two concepts when there are hyperlinks from one article to the other and back. The resulting knowledge base contains 2,325,876 concepts (articles) and 5,549,696 relations. Figure 9 shows an example where reciprocal hyperlinks to Bible denote closely related articles, and the hyperlink to Hellenistic does not. Given the knowledge base, it is straightforward to represent it as a graph, with concepts as vertices and relations as undirected edges.

Regarding the dictionary, in addition to the article title, we also mined lexicalizations (i.e., different ways of referring to the article) from redirection and disambiguation pages, as well as the anchor text in hyperlinks that point to the article. All those strings are lower-cased, and all texts between parenthesis are removed. When an hyperlink points to a disambiguation page, its anchor text is associated with all articles the disambiguation page points to. The anchors in the text are used to gather prior probabilities. The prior for a string is estimated as the number of times that the string occurs in the anchor text pointing to an article divided by the total number of occurrences of the string as anchor text.

6.2 RQE experiments

In order to see if Wikipedia is a good external resource for query expansion, we apply our RQE method using Wikipedia instead of WordNet. We followed the same experimental design (cf. Sect. 3) and use the train part of the collections to tune the parameters of the RQE method. Table 8 shows the results on the three datasets, compared to those of the QL baseline and of the RQE method on WordNet (these results are the ones shown in Table 5, column RQE). RQE on Wikipedia is always better than the baseline and attains better results than when using WordNet on Robust and Yahoo!, but worse on ResPubliQA. All differences with respect to the baseline and RQE on WordNet are statistically significant, except for ResPubliQA. Furthermore, Wikipedia RQE beats PRF on two datasets and gets even on Robust (cf. Table 5), as the difference is not statistically significant.

Table 8 Results of RQE on Wikipedia, compared to QL and RQE on WordNet

Dataset	Measure	QL	RQE-WN		RQE-Wiki	
		Result	Result	Δ QL (%)	Result	Δ QL (%)
Robust	MAP	0.3322	0.3367	1.36	**0.3568*****	7.40
Yahoo!	MRR	0.2636	0.2722	3.26	**0.2789*****	5.81
ResPubl.	MRR	0.4877	**0.4978**	2.07	0.4929	1.08

Δ columns show relative improvement with respect to QL. Bold shows best in row

7 Qualitative analysis

We have already mentioned that our expansion methods compare well to PRF, as they generalize well to other collections and are more robust to parameter adjustments. We also mentioned that they are complementary, as PRF is better for easy queries and our models perform better on difficult queries. In this section, we will examine the expansion terms proposed by PRF and RQE (both on WordNet and Wikipedia). We will then comment analyze the relation between our method and other WSD methods.

7.1 Expansion terms

Both PRF and RQE propose terms to be added to the query for more effective retrieval. PRF runs the query on the document collection and selects those terms which occur distinctively on highly ranked documents. When the top-ranked documents are closely related to the query, this is known to introduce good quality terms and improve performance. On the contrary, when the top-ranked documents are not related to the query, unrelated terms are introduced and topic drift occurs [37]. This would explain the poorer results of PRF on difficult queries.

In the case of our expansion methods, the document collection is not searched for, and the expansion terms are selected from the respective knowledge base (WordNet or Wikipedia in our case) and are thus specially useful on difficult queries. RQE also exhibits topic drift, but, contrary to PRF, the topic drift is independent of the target collection, and it thus produces terms which are not necessarily found in the collection in the context of the query terms, limiting the negative influence.

For instance, in one query about the Dayton agreement involving countries in the Balcan war, PRF produces terms such as "*Sarajevo*" and "*Milosevic*", which, although related, retrieve documents which have nothing to do with the Dayton agreement. In another query about a woman's rights conference in Beijing, Wikipedia brings in terms such as "*blood*" and "*donor*", which, although unrelated to the query, do not affect negatively the results, as those terms do not alter the ranking of the relevant documents. These examples illustrate the fact that the errors made by PRF when expanding the query can produce a negative impact on final performance, as they refer to terms which do occur frequently with the query terms in the document collection and may return wrong documents. Our method, on contrast, produces errors, which are not correlated with the frequencies and cooccurrences in the document collection, and are thus less harmful.

7.2 Relation to WSD

As mentioned in the introduction, most of the techniques based on WordNet have focused on performing explicit WSD before doing expansion. Our method initializes the random walk

Title: Computer Mouse RSI
Desc: Find documents that report on computer mouse repetitive strain injuries (RSI).
Narr: Relevant documents report injuries that are caused by the continuous use of a computer mouse. Documents proposing ways to avoid repetitive strain injuries (RSI) when using the computer are also relevant.

(a)

computer mouse rsi repetitive strain injuries

(b)

Fig. 10 A query example from the Robust dataset to illustrate relation to WSD and polysemy. **a** Topic 10.2452/064-AH from Robust dataset, **b** The formulated query using the *title* and *desc* fields of the topic

mouse-1: any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails.
mouse-4, computer mouse: a hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move it around on a pad; on the bottom of the device is a ball that rolls on the surface of the pad.

(a)

strain-3, tune, melody, air, melodic line, line, melodic phrase: a succession of notes forming a distinctive sequence; *she was humming an air from Beethoven.*
strain-5, breed, stock: a special variety of domesticated animals within a species; *he experimented on a particular breed of white rats.*
strain-7: injury to a muscle (often caused by overuse), results in swelling and pain.

(b)

Fig. 11 Some nominal senses for the words *mouse* and *strain*, as given by WordNet

using a set of words. If the words are polysemous, the random walk follows all senses of the words, but the probabilities of the senses which are close to other senses in the input set raise, and the probabilities of unrelated senses decrease. When selecting the concepts for expansion, those concepts which are close to the intended senses of the input words will get higher scores than the rest. Table 2 shows this effect, with *tractor*, *speed* and *kmh* getting high scores for the first query ("*How fast does a tractor go*") and *recipe*, *apple pie* and *bake* getting higher scores for the second query ("*How do you cook an apple pie*"). We thus interpret that our algorithm does implicit WSD. In fact, an algorithm based on random walks has been successfully used to perform WSD [5].

For instance, Fig. 10 shows an information need and respective query from the Robust dataset. Some of the terms in this query are polysemous. Figure 11 shows a subset of the senses of the words *mouse* and *strain* in WordNet.[12] Given this query, the IR baseline system retrieves, among others, the documents displayed in Fig. 12, which are not relevant to the given query. These documents are considered to be relevant by the system because they contain two of the query words (highlighted as gray), but used with a different meaning. For instance, the word *mouse* in the query is used in the computer mouse sense, whereas in the document it refers to a kind of animal. In the case of the word *strain*, the query refers to an injury in the muscle, while the document in Fig. 12a refers to a variety of an animal, and the document in Fig. 12b refers to a tune or melody of music.

For the future, we would like to check whether a state-of-the-art dedicated system doing WSD prior to running the random walks would improve the performance of our expansion methods.

[12] http://wordnetweb.princeton.edu/perl/webwn.

RESEARCHER ACCUSED OF FAKING DATA; HER STUDY PURPORTED TO USE GENES TO TRANSFER DISEASE RESISTANCE.
(...) Her results were published in the April 25, 1986, issue of the journal Cell in an article co-authored by Nobel laureate David Baltimore. The article "purposed to show that a gene from one **strain** of **mouse** had been transferred to another **strain** of **mouse**, resulting in the latter's production of high levels of antibody molecules it would not normally produce – antibody molecules mimicking the antibody molecules produced by the original **strain**," investigators said in a written statement. (...) after reviewing scientific evidence and performing a **computerized** statistical analysis that showed the false data was not made up of chance errors (...)

(a)

SOUNDS: LATEST WORK IS BOWEN'S MOST HIGH-PROFILE; COMPOSER AND PERFORMER OF NEW MUSIC SPENT YEARS WORKING ON THE FRINGES.
Listening to the lilting **strains** of Gene Bowen's new album "The Vermilion Sea" (...) the Nordic-looking Bowen has a few guitars, a synthesizer and the all-important **computer** – his main composing tool – and piles of records and CDs. (...) Three years ago, Bowen began his work-in-progress, creating the raw material on synthesizers and **computers**. (...) "My interests came through guitar music and songwriting coupled with interest in folk and ethnic music, where **repetition** is always so important. **Repetition** and texture are almost more important than (...)

(b)

Fig. 12 Some nonrelevant documents retrieved for the given query in the previous example, due to polysemy. **a** Document LA112694-0025 from the Robust dataset, **b** Document LA063094-0099 from the Robust dataset

8 Application to question answering

We consider our method particularly relevant for question answering tasks. In order to test this hypothesis, we have set a radically different experimental context, where the contribution of our method to a full-fledged question answering system is evaluated. We have taken a Basque corpus of science and technology, and we have integrated the expansion techniques in an in-house question answering system for Basque [7]. Being Basque an agglutinative language with quite different features with respect to English, this experimental setup implies a test for the robustness of the method we propose. Note, furthermore, that Basque is a less-resourced language.

The ZT Corpus (Basque Corpus of Science and Technology) is a tagged collection of specialized texts in Basque, which aims to be a major resource in research and development with respect to written technical Basque [8]. It was released in December 2006 by ELDA and can be also queried online.[13] This corpus has been promoted, released and maintained by Elhuyar, a foundation dedicated to the dissemination of scientific materials. It is composed of two parts, a 2 million-word balanced part, whose annotation has been revised by hand, and another automatically tagged 6.6 million-word part. In terms of contents, ZT is a balanced corpus that comprises a wide range of areas (e.g., exact sciences, technology, physics or biology) and genres (from dissemination materials to highly specialized articles).

The evaluation dataset comprises 100 questions and the respective answers, which where compiled based on the queries that Elhuyar had logged in a related science and technology website.[14] This website is popular among interested persons and secondary school teachers and students alike. We selected one hundred queries and edited them for grammaticality. They are factoid questions whose answers refer to named entities related to science and technology, for instance, "Zein mendetan argitaratu zen Philosofiae Naturalis Principia Mathematica libu-

[13] http://www.ztcorpusa.net.

[14] http://zientzia.net/.

Table 9 Results of the Basque Q&A system

Dataset	Measure	Baseline	RDE	
		Result	Result	Δ baseline (%)
ZT corpus	MRR on paragraphs	0.7978	**0.8275****	3.72
	MRR on exact answers	0.4587	**0.4732**	3.16

Δ columns show relative improvement with respect to the baseline without expansion. Bold means better than baseline

rua?",[15] or "Nork asmatu zuen erresonantzia magnetiko nuklearra (EMN) molekula biologiko handiekin erabiltzeko metodoa?".[16]

Due to the nature and dimension of the text collection, the presence of correct answers for each question is expected to be scarce. We estimate that most of the questions have one or two instances of the correct answers in the corpus. We do not use any external resource to find hypotheses or to validate answer candidates.

Questions and documents have been lemmatized, tagged with parts of speech, and named entities have been recognized and classified, using in-house linguistic tools. The question answering system uses a paragraph retrieval method based on the MG4J search engine [11], as tested in CLEF competitions, attaining a performance comparable to other systems for less-resourced languages [7,17]. Passage retrieval is a key module of the system, as in most conventional question answering systems.

In our experiments, we try to improve paragraph retrieval and overall performance of the system introducing the use of document expansion (RDE) on top of the MG4J search engine, using the Basque WordNet as knowledge base. No specific parameter optimization was done, and default parameters were used in all experiments.

Table 9 reports the MRR over the top five paragraphs and exact answers, showing that RDE overcomes significantly the baseline system both in the retrieval of the paragraphs and in finding the exact answers. The results also, which show that paragraph retrieval is quite effective, in contrast to the answer extraction module often misses the correct answer. Despite of this, the improvement in the paragraph retrieval module translates to improved answer retrieval.

9 Related work

Our work stems from the use of random walks over the WordNet graph to compute the similarity and relatedness between pairs of words [23]. In this approach, WordNet is represented as a graph, with word senses and concepts as vertices, and relations between concepts as edges (cf. Sect. 2.1 for more details). The method first computes a random walk over the graph for a single word, obtaining the probability distribution over all WordNet concepts. The probability distribution represents the meaning of the word in the concept space. To judge the degree of similarity between any two words, it suffices to compute the similarity of the probability distributions of each word. In later work, different configurations of the graph were tested [3,6], obtaining the best results on a word similarity benchmark among WordNet-based systems to date. Note that the results were comparable to the results of a

[15] In which century was the Philosofiae Naturalis Principia Mathematica published?

[16] Who developed the method to apply nuclear magnetic resonance (NMR) to large biological molecules?

distributional similarity method, which used a crawl of the entire web [6]. The same method also ranks highest among WordNet-based methods for relatedness [6], where the task is to judge the degree to which the words are related to each other. The random-walk software is open-source,[17] and it is the same as we use in this work.

Other authors have proposed alternative similarity algorithms, including corpus-based and WordNet-based methods. Li et al. [28] presents a semantic similarity measure based on WordNet and corpora. They use WordNet to compute the path length between the synsets containing the two target words, as well as the depth most specific synset subsuming the two corresponding synsets corresponding. They also use statistical information for concepts estimated using a large corpus. The proposed similarity measure combines these information sources nonlinearly. Mihalcea et al. [36] defines similarity between pairs of texts, extending the notion of word similarity. These and other similarity methods have been designed to return a score for pairs of texts, while query (and document) expansion requires that, given a set of terms, a list of similar or related words are generated in decreasing order of similarity. It is not clear that the similarity techniques just mentioned can be applied to the generation setting, but we have shown that our random-walk method can generate those related words efficiently (cf. Tables 1 and 2 in Sect. 1).

Although we will focus on applications of similarity to IR, knowledge-based similarity has also been found to be useful in other applications. For instance, similarity among ontology items is used to support communication between agents in multi-agent systems [34,46]. Those works are complementary to ours; in that, we use our technique to propose expansion terms and improve IR and show a general trend where the information in knowledge structures enables practical applications via similarity measures.

As mentioned in the introduction, IR relies heavily on keyword match. As an alternative to bridge lexical mismatches between query and documents, QE and DE methods have been proposed. QE methods analyze user query terms and incorporate related terms automatically [54] and are usually divided into local and global methods. Local methods adjust a query relative to the documents that initially appear to match the query [32]. PRF is one of the most widely used expansion methods [45,55]. This method assumes that the top-ranked documents returned by the original query are relevant (and in some cases, that low-ranked documents are irrelevant) and selects additional query terms from the top-ranked documents. Since Rocchio presented an algorithm for relevance feedback [45], lots of variations have been developed. The TREC 2008 Relevance Feedback Track results confirmed that relevance feedback consistently improves different kinds of retrieval models, but the amount of relevance information needed to improve results and the use or not of nonrelevant information varied among systems [12].

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 [32]. As an example of the former, Bai et al. [9] proposed a language modeling approach that integrates term relationships mined from documents in a query expansion model. They considered two specific types of term relationship: co-occurrence relationships and inferential relationships extracted from documents. As examples of the later, several researchers have expanded queries with synonyms from WordNet after performing WSD with some success [13,15, 29,30,54,57]. For instance, Zhong and Ng [57] use a combination of PRF, WSD and query expansion using WordNet relations. They use the top documents returned by the query to provide a context for disambiguating the queries, in a way reminiscent of PRF. The senses

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

and the synonyms of the senses are then used to smooth term probabilities in a language modeling approach to IR. They show very strong results, with significant improvements and state-of-the-art results, but their expansion system might suffer on datasets where PRF is not effective.

The query expansion method proposed here is also a global expansion technique based on WordNet, but in contrast to the references just cited, it does not require explicit WSD and uses related words beyond synonyms for expansion. As mentioned above, the use of explicit WSD could further improve our technique to suggest expansion techniques.

An alternative to QE is to perform the expansion in the document. DE was first proposed in the speech retrieval community [48], where the task is to retrieve speech transcriptions, which are quite noisy. Singhal and Pereira proposed to enhance the representation of a noisy document by adding to the document vector a linearly weighted mixture of related documents. In order to determine related documents, the original document is used as a query into the collection, and the ten most relevant documents are selected. Two related papers [26,31] followed a similar approach on the TREC ad-hoc document retrieval task. They use document clustering to determine similar documents, and document expansion is carried out with respect to these. Both papers report significant improvements over nonexpanded baselines. Instead of clustering, more recent work [22,33,53] uses language models and graph representations of the similarity between documents in the collection to smooth language models with some success.

The document expansion method presented here is complementary to those methods; in that, we also explore DE, but use WordNet instead of distributional methods. The comparison with respect to other DE techniques and the exploration of potential combinations will be the focus of future research.

Another strand of WordNet-based IR work has explicitly represented and indexed word senses after performing WSD, without performing any expansion proper [19,25,50]. Word senses conform a different space for document representation, but contrary to us, these works incorporate concepts for all words in the documents and are not able to incorporate concepts that are not explicitly mentioned in the document. Stokoe et al. [50] performed WSD on WordNet senses for both documents and queries and achieved significant improvements over a vector-space model baseline. Unfortunately, the baseline was very weak, making it difficult to judge whether the word senses would be helpful in a stronger IR system. Kim et al. [25] tagged nouns with 25 semantic tags from WordNet and adjusted term weights in the baseline IR system according to the sense matches between query and document, improving over a strong system. More recently, a CLEF task was organized [4] where terms were semantically disambiguated to see the improvement that this would have on retrieval. Several teams participated, exploring different ways to index word senses. The conclusions were mixed, with some participants slightly improving results over baselines with information from WordNet. Our method to find related concepts both for queries and documents is complementary to those methods; in that, we could have used an index of concepts and word senses in addition to the additional index in RDE. We would like to explore these possibilities in the future.

As an alternative to WordNet, other authors have used Wikipedia as the word sense or concept repository. For instance, Egozi et al. [14] use a method to augment text with concepts from Wikipedia, based on Explicit Semantic Analysis [18]. In order to improve over the baseline, they need to use feature selection methods to prune the concept representation and combine concept and bag-of-words retrieval. In contrast, we show that using our graph-based method using Wikipedia links for query expansion leads to improvements with no need for feature selection. Although we use WordNet and Wikipedia separately, heterogeneous information sources could be combined following the method laid out in [40].

In previous work [1], we used 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. Subsequently, we moved to a language modeling approach, experimenting with query expansion and comparing the performance with PRF [39]. The work presented here extends [39] with an implementation of RDE in a language modeling framework and provides more extensive analysis and experimentation.

Finally, we would like to mention the performance of other systems on the same datasets. The systems which performed best in the Robust evaluation campaign [4] report 0.4509 MAP, but note that they deployed a complex system combining probabilistic and monolingual translation-based models. In ResPubliQA [41], the official evaluation included manual assessment, and we cannot therefore reproduce those results exactly. As an alternative, the organizers released all runs, but only the first ranked document for each query was included, so we could only compute P@1. The P@1 of the best run was 0.40, which is not so far from our best P@1 result, as we obtain 0.3940 P@1 for RDE. Regarding Yahoo!, Surdeanu et al. [52] report an MRR of around 0.68. This number in an overestimation of the real performance, as they evaluate only in the questions where the correct answer is retrieved by their document retrieval engine in the top 50 answers, and it is thus not directly comparable to our setting.

10 Conclusions

In this paper, we explore a generic method to improve IR results using structured knowledge for both query and document expansion. Our work has been motivated by the success of knowledge-based methods in word similarity and relatedness tasks [6], where it outperformed distributional similarity methods. Note that distributional methods are closely related to query expansion and document clustering techniques for IR. In query expansion, techniques such as PRF expand the query with terms which are deemed to be related to the query according to the retrieved documents [55]. In document clustering, terms from documents in the same cluster are used to re-estimate counts and to expand the documents with new terms [48]. Our research is complementary to the aforementioned techniques; in that, we also experimented with question and document expansion (RQE and RDE), but our core technique is based on knowledge-based similarity.

Our expansion method is based on random walks over a graph representation of a knowledge base. The random walk returns sets of concepts, which are related to the input query (or document), even if those concepts are not explicitly mentioned in the texts. The query (or document) is then expanded using the terms lexicalizing the related concepts. In this work, we mainly focused on WordNet, but any other knowledge structure could be used, as shown by our successful results using Wikipedia.

We adopted a language modeling framework to implement the QL and PRF baselines, as well as our RQE and RDE methods, where the expansion terms for documents are indexed separately. We wanted to check the performance on a diverse typology of document collections, ranging from ad-hoc IR, answer finding and passage retrieval, as follows: Robust-WSD dataset from CLEF (ad-hoc dataset on news which got the attention of the WSD community), Yahoo! Answers (answer-finding collection, including questions and answers by real users on diverse topics) and ResPubliQA (a passage-retrieval task on European Union laws in the context of a question answering exercise).

Our two methods provide improvements in all three datasets, when compared to the QL baseline. PRF is beneficial in two datasets, but degrades performance in ResPubliQA. RDE and RQE compare favorably to PRF in two datasets, but perform worse in Robust. Our analysis shows that our models and PRF are complementary; in that, PRF is better for easy queries and our models are stronger for difficult queries. We also show that our models generalize better to other collections and are more robust to parameter adjustments. In addition, we tested the contribution of our method to a Basque question answering system on Science and Technology with positive results on the final system.

Given the very positive results obtained with WordNet, we also experimented with Wikipedia, further improving the results of RQE with WordNet on two out of the three datasets. In related work, we used random walks over Wikipedia for semantic enrichment in a task held at CHiC 2012 (Cultural Heritage in CLEF), obtaining good results [2]. On the other hand, we have also tested our query expansion method on the 2012 and 2013 TREC Medical Tracks,[18] where we used the UMLS metathesaurus [24] as the knowledge base. The results where positive in both years.

In the future, we would like to combine several knowledge sources when doing the expansion. We would also like to combine our relatedness method with other WSD-based techniques and to explore the ability of RQE and RDE to perform well on difficult queries, perhaps combining them with PRF and document clustering techniques.

A limitation of our method is that it would suffer in the case of noisier datasets such as blogs or tweets, where informal language abounds. Recent work on normalization [20] would be very helpful, as the text could be normalized prior to checking the lexical resources, making our method amenable to the task.

Acknowledgments This work was partially funded by MINECO in Projects READERS and SKATER (PCIN-2013-002-C02-01, TIN2012-38584-C06-02) and by the European Commission in Project NEWSREADER (ICT FP7-ICT-2011-8-316404).

References

1. Agirre E, Arregi X, Otegi A (2010) Document expansion based on WordNet for robust IR. In: Proceedings of the 23rd international conference on computational linguistics: posters, COLING '10, Association for Computational Linguistics, pp 9–17
2. Agirre E, Clough P, Fernando S, Hall M, Otegi A, Stevenson M (2012) The Sheffield and Basque Country Universities Entry to CHiC: using random walks and similarity to access cultural heritage. In: CLEF (Online Working Notes/Labs/Workshop)'12
3. Agirre E, Cuadros M, Rigau G, Soroa A (2010) Exploring knowledge bases for similarity. In: Proceedings of the seventh international conference on language resources and evaluation (LREC '10), European Language Resources Association (ELRA), pp 373–377
4. Agirre E, Di Nunzio GM, Mandl T, Otegi A (2010) CLEF 2009 ad hoc track overview: robust-WSD task. In: Multilingual information access evaluation I. Text retrieval experiments, Vol. 6241 of Lecture Notes in Computer Science. Springer, Berlin, pp 36–49
5. Agirre E, Soroa A (2009) Personalizing pagerank for word sense disambiguation. In: Proceedings of the 12th conference of the European chapter of the the association for computational linguistics, EACL '09, Association for Computational Linguistics, pp 33–41
6. Agirre E, Soroa A, Alfonseca E, Hall K, Kravalova J, Paşca M (2009) A study on similarity and relatedness using distributional and WordNet-based approaches. In: Proceedings of human language technologies: the 2009 annual conference of the North American Chapter of the Association for Computational Linguistics, NAACL '09, Association for Computational Linguistics, pp 19–27

[18] http://trec.nist.gov/pubs/call2012.html.

7. Ansa O, Arregi X, Otegi A, Soraluze A (2009) Ihardetsi: a basque question answering system at QA@CLEF 2008. In: Evaluating systems for multilingual and multimodal information access, Vol. 5706 of Lecture Notes in Computer Science. Springer, Berlin, pp 369–376
8. Areta N, Gurrutxaga A, Leturia I, Polin Z, Saiz R, Alegria I, Artola X, de Ilarraza AD, Ezeiza N, Sologaistoa A, Soroa A, Valverde A (2006) Structure, annotation and tools in the basque ZT corpus. In: International conference on language resources and evaluations (LREC 2006), pp 1406–1411
9. Bai J, Song D, Bruza P, Nie JY, Cao G (2005) Query expansion using term relationships in language models for information retrieval. In: Proceedings of the 14th ACM international conference on Information and knowledge management, CIKM '05, ACM, pp 688–695
10. Berger A, Caruana R, Cohn D, Freitag D, Mittal V (2000) Bridging the lexical chasm: statistical approaches to answer-finding. In: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval. ACM, pp 192–199
11. Boldi P, Vigna S (2005) MG4J at TREC 2005. In: The fourteenth text retrieval conference (TREC 2005) proceedings, number SP 500–266 in 'Special Publications', National Institute of Standards and Technology (NIST)
12. Buckley C, Sanderson M (2008) Relevance feedback track overview: TREC 2008. In: Proceedings of The seventeenth text retrieval conference, TREC 2008, Vol. Special Publication 500-277, National Institute of Standards and Technology (NIST)
13. Cao G, Nie J, Bai J (2005) Integrating word relationships into language models. In: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '05. ACM, pp 298–305
14. Egozi O, Markovitch S, Gabrilovich E (2011) Concept-based information retrieval using explicit semantic analysis. ACM Trans Inf Syst 29(2):8:1–8:34
15. Fang H (2008) A re-examination of query expansion using lexical resources. In: Proceedings of the 46th annual meeting of the association for computational linguistics. Human language technologies. Association for Computational Linguistics, pp 139–147. http://www.aclweb.org/anthology/P/P08/P08-1017
16. Fellbaum C (1998) WordNet: an electronic lexical database and some of its applications. MIT Press, Cambridge
17. Forner P, Penas A, Agirre E, Alegria I, Forăscu C, Moreau N, Osenova P, Prokopidis P, Rocha P, Sacaleanu B, Sutcliffe R, Sang E (2009) Overview of the CLEF 2008 multilingual question answering track. In: Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access, CLEF'08. Springer, Berlin, pp 262–295
18. Gabrilovich E, Markovitch S (2007) Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th international joint conference on artificial intelligence. Morgan Kaufmann Publishers Inc., pp 1606–1611
19. Gonzalo J, Verdejo F, Chugur I, Cigarran J (1998) Indexing with WordNet synsets can improve text retrieval. In: Proceedings of the COLING/ACL workshop on usage of wordnet in natural language processing systems, pp 38–44
20. Han B, Baldwin T (2011) Lexical normalisation of short text messages: Makn Sens a #twitter. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. Association for Computational Linguistics, pp 368–378
21. Haveliwala TH (2002) Topic-sensitive PageRank. In: Proceedings of the 11th international conference on world wide web, WWW '02. ACM, pp 517–526
22. Huang Y, Sun L, Nie J (2009) Smoothing document language model with local word graph. In: Proceeding of the 18th ACM conference on Information and knowledge management, CIKM '09, ACM, pp 1943–1946
23. Hughes T, Ramage D (2007) Lexical semantic relatedness with random graph walks. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), pp 581–589
24. Humphreys L, Lindberg D, Schoolman H, Barnett G (1998) The unified medical language system: an informatics research collaboration. J Am Med Inf Assoc 1(5):1–11
25. Kim S, Seo H, Rim H (2004) Information retrieval using word senses: root sense tagging approach. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '04. ACM, pp 258–265
26. Kurland O, Lee L (2004) Corpus structure, language models, and ad hoc information retrieval. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '04. ACM, pp 194–201
27. Lavrenko V, Croft WB (2001) Relevance based language models. In: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '01. ACM, pp 120–127

28. Li Y, Bandar Z, McLean D (2003) An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans Knowl Data Eng 15(4):871–882
29. Liu S, Liu F, Yu C, Meng W (2004) An effective approach to document retrieval via utilizing WordNet and recognizing phrases. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '04. ACM, pp 266–272
30. Liu S, Yu C, Meng W (2005) Word sense disambiguation in queries. In: Proceedings of the 14th ACM international conference on Information and knowledge management, CIKM '05. ACM, pp 525–532
31. Liu X, Croft WB, Bruce W (2004) Cluster-based retrieval using language models. In: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '04. ACM, pp 186–193
32. Manning CD, Raghavan P, Schütze H (2009) An introduction to information retrieval. Cambridge University Press, Cambridge
33. Mei Q, Zhang D, Zhai C (2008) A general optimization framework for smoothing language models on graph structures. In: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '08. ACM, pp 611–618
34. Meo PD, Quattrone G, Rosaci D, Ursino D (2012) Bilateral semantic negotiation: a decentralised approach to ontology enrichment in open multi-agent systems. Int J Data Mining Model Manag (IJDMMM) 4(1):1–38
35. Metzler D (2006) Estimation, sensitivity, and generalization in parameterized retrieval models. In: Proceedings of the 15th ACM international conference on Information and knowledge management, CIKM '06. ACM, pp 812–813
36. Mihalcea R, Corley C, Strapparava C (2006) Corpus-based and knowledge-based measures of text semantic similarity. In: Proceedings of the 21st national conference on artificial intelligence—Volume 1', AAAI '06. AAAI Press, pp 775–780
37. Mitra M, Singhal A, Buckley C (1998) Improving automatic query expansion. In: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '98. ACM, pp 206–214
38. Moldovan D, Surdeanu M (2003) On the role of information retrieval and information extraction in question answering systems. Information Extraction in the Web Era, pp 129–147
39. Otegi A, Arregi X, Agirre E (2011) Query expansion for IR using knowledge-based relatedness. In: Proceedings of 5th international joint conference on natural language processing, Asian Federation of Natural Language Processing, pp 1467–1471
40. Palopoli L, Rosaci D, Terracina G, Ursino D (2005) A graph-based approach for extracting terminological properties from information sources with heterogeneous formats. Knowl Inf Syst 8(4):462–497
41. Peñas A, Forner P, Sutcliffe R, Rodrigo A, Forăscu C, Alegria I, Giampiccolo D, Moreau N, Osenova P (2009) Overview of ResPubliQA 2009: question answering evaluation over European legislation. In: Proceedings of the 10th cross-language evaluation forum conference on multilingual information access evaluation: text retrieval experiments, CLEF '09. Springer, Berlin, pp 174–196
42. Ponte JM, Croft WB (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. ACM, pp 275–281
43. Riezler S, Vasserman A, Tsochantaridis I, Mittal V, Liu Y (2007) Statistical machine translation for query expansion in answer retrieval. In: Proceedings of the 45th annual meeting of the association of computational linguistics. Association for Computational Linguistics, pp 464–471
44. Robertson S (2006) On GMAP: and other transformations. In: Proceedings of the 15th ACM international conference on Information and knowledge management, CIKM '06. ACM, pp 78–83
45. Rocchio JJ (1971) Relevance feedback in information retrieval. In: Salton G (ed) The smart retrieval system: experiments in automatic document processing. Prentice-Hall, Englewood Cliffs, pp 313–323
46. Rosaci D (2007) CILIOS: connectionist inductive learning and inter-ontology similarities for recommending information agents. Inf. Syst. 32(6):793–825
47. Ruthven I, Lalmas M (2003) A survey on the use of relevance feedback for information access systems. Knowl Eng Rev 18(2):95–145
48. Singhal A, Pereira F (1999) Document expansion for speech retrieval. In: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '99. ACM, pp 34–41
49. Smucker MD, Allan J, Carterette B (2007) A comparison of statistical significance tests for information retrieval evaluation. In: Proceedings of the sixteenth ACM conference on information and knowledge management, CIKM '07. ACM, pp 623–632

50. Stokoe C, Oakes MP, Tait J (2003) Word sense disambiguation in information retrieval revisited. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '03. ACM, pp 159–166
51. Strohman T, Metzler D, Turtle H, Croft WB (2005) Indri: a language-model based search engine for complex queries. In: Technical report, Proceedings of the international conference on intelligent analysis
52. Surdeanu M, Ciaramita M, Zaragoza H (2008) Learning to rank answers on large online QA collections. In: Proceedings of the 46th annual meeting of the association for computational linguistics. The Association for Computer Linguistics, pp 719–727
53. Tao T, Wang X, Mei Q, Zhai C (2006) Language model information retrieval with document expansion. In: Proceedings of the main conference on human language technology conference of the North American Chapter of the Association of Computational Linguistics, HLT-NAACL '06. Association for Computational Linguistics, pp 407–414
54. Voorhees EM (1994) Query expansion using lexical-semantic relations. In: Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '94. Springer, New York, pp 61–69
55. Xu J, Croft WB (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. ACM, pp 4–11
56. Zhai C, Lafferty J (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. ACM, pp 334–342
57. Zhong Z, Ng HT (2012) Word sense disambiguation improves information retrieval. In: Proceedings of the 50th annual meeting of the Association for Computational Linguistics: Long Papers—Volume 1. Association for Computational Linguistics, pp 273–282

Arantxa Otegi received her Ph.D. in 2012 and B.A. degree in 2006 in Computer Engineering, both from the University of the Basque Country. She is currently a researcher of the natural language processing research group of the University of the Basque Country. She has been a member of that group since 2003, working mainly on the fields of information retrieval and lexical semantics. She has participated in several projects and has been involved in the organization of several workshops in some of the international conferences like ACL and CLEF.

Xabier Arregi is Associate Professor in the Department of Computer Languages and Systems at the University of the Basque Country. He received his Ph.D. in Computer Engineering from the same university in 1995. Since 1990, he has been working in the IXA natural language processing group, where he has coordinated several projects. His research interests include information retrieval, question answering and text understanding.

Olatz Ansa is Lecturer Professor in the Department of Computer Languages and Systems at the University of the Basque Country. Since 1996, she has been working in the IXA natural language processing group, participating in several projects. Her research interests include question answering, information retrieval and lexical semantics.

Eneko Agirre received his Ph.D. (1999) and B.A. (1990) in Computer Science by the University of the Basque Country, and is Associate Professor in the Computer Science Faculty there since 2005. He received his M.Sc. in Knowledge-Based Systems by Edinburgh University (1991). He has published over 150 international peer-reviewed articles and conference papers in Natural Language Processing, mainly in the areas of Lexical Knowledge Acquisition, Word Sense Disambiguation and Semantic Processing. He has been secretary and president of the ACL SIGLEX and member of the editorial board of Computational Linguistics. He is a usual reviewer for top international journals including LRE, NLE, JSL, ACM CS, ACM TOIS and IEEE IS. He is a regular area chair and member of the program committees of international conferences like GWC, ACL, EACL, NAACL, IJCNLP, HLT, CONLL, TSD, RANLP, SEPLN, EMNLP, AAAI and IJCAI. He has lead the local team in several European projects.