Query Expansion based on Pseudo Relevance Feedback from Definition Clusters

Delphine Bernhard
LIMSI-CNRS
Delphine.Bernhard@limsi.fr

Abstract

Query expansion consists in extending user queries with related terms in order to solve the lexical gap problem in Information Retrieval and Question Answering. The main difficulty lies in identifying relevant expansion terms in order to prevent query drift. We propose to use definition clusters built from a combination of English lexical resources for query expansion. We apply the technique of pseudo relevance feedback to obtain expansion terms from definition clusters. We show that this expansion method outperforms both local feedback, based on the document collection, and expansion with WordNet synonyms, for the task of document retrieval in Question Answering.

1 Introduction

Question Answering (QA) systems aim at providing precise answers to user questions. Most QA systems integrate a document retrieval component, which is in charge of retrieving the most relevant documents or passages for a given user question. Since document retrieval is performed in early stages of QA, it is of the uttermost importance that all relevant documents be retrieved, to limit the loss of relevant answers for further processing. However, document retrieval systems have to solve the lexical gap problem, which arises from alternative ways of conveying the same piece of information in questions and answers. One of the solutions proposed to deal with this issue is query expansion (QE), which consists in extending user queries with related terms.

This paper describes a new method for using lexical-semantic resources in query expansion with a focus on QA applications. While some research has been devoted to using explicit semantic relationships for QE, such as synonymy or hypernymy, with rather disappointing results (Voorhees, 1994), we focus on the usefulness of textual and unstructured dictionary definitions for question expansion. Definitions extracted from seven English lexical resources are first grouped to obtain definition clusters, which capture redundancies and sense mappings across resources. Expansion terms are extracted from these definition clusters using pseudo relevance feedback: we first retrieve the definition clusters which are most related to the user query, and then extract the most relevant terms from these definition clusters to expand the query.

The contributions of this work are as follows: (i) we build definition clusters across seven different lexical resources for English, (ii) we thoroughly compare different question expansion methods using local and global feedback, and (iii) we address both the lexical gap and question ambiguity problems by integrating expansion and disambiguation in one and the same step.

In the next section, we describe related work. In Section 3, we describe our method for acquiring definition clusters from seven English lexical resources. In Section 4, we detail query expansion methods. We present experimental results in Section 5 and conclude in Section 6.

2 Related Work

Query expansion attempts to solve the vocabulary mismatch problem by adding new semantically related terms to the query. The goal is to increase recall by retrieving more relevant documents. Two types of query expansion methods are usually distinguished (Manning et al., 2008): *global* techniques, which do not take the results obtained for the original query into account, and

Coling 2010: Poster Volume, pages 54–62,
Beijing, August 2010

local techniques, which expand the query based on an analysis of the documents returned. Local methods are also known as *relevance feedback*.

A first type of global QE methods relies on external hand-crafted lexical-semantic resources such as WordNet. While expansion based on external resources is deemed more efficient than expansion relying on relevance feedback, it also has to tackle problems of semantic ambiguity, which explains why local analysis has been shown to be generally more effective than global analysis (Xu and Croft, 1996). However, recent work by Fang (2008) has demonstrated that global expansion based on WordNet and co-occurrence based resources can lead to performance improvement in an axiomatic model of information retrieval.

Corpus-derived co-occurrence relationships are also exploited for query expansion. Qiu and Frei (1993) build a corpus-based similarity thesaurus using the method described in Schütze (1998) and expand a query with terms which are similar to the query concept based on the similarity thesaurus. Song and Bruza (2003) construct vector representations for terms from the target document collection using the Hyperspace Analogue to Language (HAL) model (Lund and Burgess, 1996). The representations for all the terms in the query are then combined by a restricted form of vector addition. Finally, expansion terms are derived from this combined vector by information flow.

Quasi-parallel monolingual corpora have been recently employed for query expansion, using statistical machine translation techniques. Expansion terms are acquired by training a translation model on question-answer pairs (Riezler et al., 2007) or query-snippets pairs (Riezler et al., 2008) and by extracting paraphrases from bilingual phrase tables (Riezler et al., 2007).

The main difficulty of QE methods lies in selecting the most relevant expansion terms, especially when the query contains ambiguous words. Moreover, even if the original query is not ambiguous, it might become so after expansion. Recent attempts at integrating word sense disambiguation (WSD) in IR within the CLEF Robust WSD track[1] have led to mixed results which show that in most cases WSD does not improve performance of monolingual and cross-lingual IR systems (Agirre et al., 2009). For query expansion based on translation models, ambiguity problems are solved by a language model trained on queries (Riezler et al., 2008), in order to select the most likely expansion terms in the context of a given query.

In this article, we propose to integrate disambiguation and expansion in one and the same step by retrieving expansion terms from definition clusters acquired by combining several English lexical resources.

3 Acquisition of Definition Clusters

Dictionary definitions constitute a formidable resource for Natural Language Processing. In contrast to explicit structural and semantic relations between word senses such as synonymy or hypernymy, definitions are readily available, even for less-resourced languages. Moreover, they can be used for a wide variety of tasks, ranging from word sense disambiguation (Lesk, 1986), to producing multiple-choice questions for educational applications (Kulkarni et al., 2007) or synonym discovery (Wang and Hirst, 2009). However, all resources differ in coverage and word sense granularity, which may lead to several shortcomings when using a single resource. For instance, the sense inventory in WordNet has been shown to be too fine-grained for efficient word sense disambiguation (Navigli, 2006; Snow et al., 2007). Moreover, gloss and definition-based measures of semantic relatedness which rely on the overlap between the definition of a target word and its distributional context (Lesk, 1986) or the definition of another concept (Banerjee and Pedersen, 2003) yield low results when the definitions provided are short and do not overlap sufficiently.

As a consequence, we propose combining lexical resources to alleviate the coverage and granularity problems. To this aim, we automatically build cross-resource sense clusters. The goal of our approach is to capture redundancy in several resources, while improving coverage over the use of a single resource.

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

3.1 Resources

In order to build definition clusters, we used the following seven English resources:

WordNet We used WordNet 3.0, which contains 117,659 synset definitions.[2]

GCIDE The GCIDE is the GNU version of the Collaborative International Dictionary of English, derived from Webster's 1913 Revised Unabridged Dictionary. We used a recent XML version of this resource,[3] from which we extracted 196,266 definitions.

English Wiktionary and Simple English Wiktionary Wiktionary is a collaborative online dictionary, which is also available in a simpler English version targeted at children or non-native speakers. We used the English Wiktionary dump dated August 16, 2009 and the Simple English Wiktionary dump dated December 9, 2009. The English Wiktionary comprises 245,078 definitions, while the Simple English Wiktionary totals 11,535 definitions.

English Wikipedia and Simple English Wikipedia Wikipedia is a collaborative online encyclopedia. As Wiktionary, it provides a Simple English version. We used the Mediawiki API to extract 152,923 definitions from the English Wikipedia[4] and 53,993 definitions from the Simple English Wikipedia. Since full Wikipedia articles can be very long in comparison to the other resources, we only retrieved the first sentence of each page to constitute the definition database, following (Kazama and Torisawa, 2007).

OmegaWiki OmegaWiki is a collaborative multilingual dictionary based on a relational database. We used the SQL database dated December 17, 2009,[5] comprising 29,179 definitions.

3.2 Definition Clustering

In order to cluster definitions, we first build a definition graph: each node in the graph corresponds to a definition in one of our input resources and two definition nodes are linked if they define the same term and their definitions are similar enough. Links are weighted by the cosine similarity of the definition nodes. To compute the cosine similarity, we stem the definition words with the Porter Stemmer and remove stop words. Moreover, we weigh words with their *tf.idf* value in the definitions. Document frequency (*df*) counts are derived from the definitions contained in all our resources.

Definition clusters are identified with a community detection algorithm applied to the definition graph. Communities correspond to groups of nodes with dense interconnections: in our case, we aim at retrieving groups of related definitions. We used the algorithm proposed by Blondel et al. (2008), based on modularity optimisation.[6] The modularity function measures the quality of a division of a graph into communities (Newman and Girvan, 2004).

In order to increase the precision of clustering, we remove edges from the graph whose cosine value is lower than a given threshold.

3.3 Evaluation of Definition Clusters

We built a gold-standard by manually grouping the definitions contained in our source resources for 20 terms from the Basic English Word List,[7] totalling 726 definitions, grouped in 321 classes. We evaluated the definition clusters in terms of clustering purity (Manning et al., 2008), which is a classical evaluation measure to measure clustering quality. Purity is defined as follows:

$$purity(\Omega, C) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j| \qquad (1)$$

where N is the number of clustered definitions, $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ is the set of definition

[2] Statistics obtained from http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html

[3] Retrieved from http://rali.iro.umontreal.ca/GCIDE/

[4] As we mainly aimed at capturing the redundancy across resources, we only extracted definitions for the Wikipedia terms which were also found in the GCIDE, Omegawiki, Wiktionary or Simple English Wikipedia.

[5] Retrieved from http://omegawiki.org/

[6] We used its Python implementation by Thomas Aynaud, available at http://perso.crans.org/aynaud/communities/community.py [Visited on October 26, 2009].

[7] http://en.wiktionary.org/wiki/Appendix:Basic_English_word_list

Resource	Definition
WordNet	an arc of colored light in the sky caused by refraction of the sun's rays by rain
Gcide	A bow or arch exhibiting, in concentric bands, the several colors of the spectrum, and formed in the part of the hemisphere opposite to the sun by the refraction and reflection of the sun's rays in drops of falling rain.
Simple Wikipedia	A rainbow is an arc of color in the sky that you can see when the sun shines through falling rain.
Simple Wiktionary	The arch of colours in the sky you can see in the rain when the sun is at your back.

Table 1: Excerpt from a definition cluster.

clusters obtained, w_k is the set of definitions in cluster k, $C = \{c_1, c_2, \ldots, c_J\}$ is the set of definition families expected and c_j is the set of definitions in family j.

We also report the amount of clusters obtained for each cosine threshold value. The evaluation results are detailed in Table 2.

Cosine threshold	Purity	# Clusters
0.0	0.363	73
0.1	0.464	135
0.2	0.644	234
0.3	0.848	384
0.4	0.923	458
0.5	0.957	515

Table 2: Evaluation results for definition clustering.

Overall, the results which account for the best compromise between purity and cluster count are obtained for a threshold of 0.3: for this threshold, we obtain 384 clusters, which is closest to the expected value of 321 classes. The purity obtained for this cosine threshold is very close to the values obtained by Kulkarni et al. (2007), who clustered definitions extracted from only two source dictionaries and report a purity of 0.88 for their best results. In total we obtain 307,570 definition clusters. Table 1 displays an excerpt from one of the definition clusters obtained.

4 Query Expansion Methods

In this section, we describe the methods used for performing query expansion. We first describe two simple baseline methods, one based on local feedback, the other based on WordNet. Then, we detail our method relying on the definition clusters previously described.

4.1 Query Expansion based on Local Feedback

In order to perform local feedback based on the document collection, we used the pseudo relevance feedback methods implemented in the Terrier information retrieval platform (Ounis et al., 2007): Bo1 (Bose-Einstein 1), Bo2 (Bose-Einstein 2) and KL (Kullback-Leibler). These methods extract informative terms from the top-ranked documents retrieved using the original query and use them for query expansion.

4.2 Query Expansion based on WordNet Synonyms

As a second baseline for query expansion, we expand the query terms with their synonyms extracted from WordNet. For each query term t, we retrieve its WordNet synsets and keep the corresponding synset members as expansion terms.[8] We weigh the expansion terms in each synset by the frequency score provided in WordNet, which indicates how often the query term t occurs with the corresponding sense. In the rest of the paper, this method is referred to as **WN-synonyms**.

The expansion terms obtained using WN-synonyms are further reweighted using Rocchio's *beta* formula which computes the weight qtw of

[8] We use NLTK (`http://www.nltk.org/`) to access WordNet.

query term t as follows (Rocchio, 1971; Macdonald et al., 2005):

$$qtw = \frac{qtf}{qtf_{max}} + \beta \frac{w(t)}{w_{max}(t)} \quad (2)$$

where qtf is the frequency of term t in the query, qtf_{max} is the maximum query term frequency among the query terms, $w(t)$ is the expansion weight of t, detailed in Equation 3, and $w_{max}(t)$ is the maximum $w(t)$ of the expansion terms. In all our experiments, β is set to 0.4, which is the default value used in Terrier.

Given this formula, if an original query term occurs among the expansion terms, its weight in the expanded query increases. For expansion terms which do not occur in the original query, $qtf = 0$.

This formula has been proposed in the setting of pseudo relevance feedback, where expansion terms are chosen based on the top documents retrieved for the original query. However, in our WN-synonyms setting, one and the same expansion term might be obtained from different original query terms with different weights. It is therefore necessary to obtain a global similarity weight for one expansion term with respect to the whole query. Following Qiu and Frei (1993), we define $w(t)$ as:

$$w(t) = \frac{\sum_{t_i \in q} qtf_i \cdot s(t, t_i)}{\sum_{t_i \in q} qtf_i} \quad (3)$$

where q is the original query and $s(t, t_i)$ is the similarity between expansion term t and query term t_i, i.e., the frequency score in WordNet.

For final expansion, we keep the top T terms with the highest expansion weight.

4.3 Query Expansion Based on Definition Clusters

In order to use definition clusters (DC) for query expansion, we first use Terrier to index the clusters which obtained the best overall results in our evaluation of definition clustering, corresponding to a cosine threshold of 0.3.[9] For each cluster, we index both the definitions and the list of terms they define, which makes it possible to include synonyms or Wikipedia redirects in the index.

For a given question, we retrieve the top D definition clusters: the retrieval of definition clusters is based on all the question terms, and thus enables indirect contextual word sense disambiguation. Then, we extract expansion terms from these clusters using pseudo relevance feedback (PRF) as implemented in Terrier. The top T most informative terms are retrieved from the top D definition clusters retrieved and used for expansion. The expansion terms are weighted using the KL (Kullback-Leibler) term weighting model in Terrier. We chose this particular weighting model, as it yielded the best results for local feedback (see Table 3).

We name this method **DC-PRF**.

5 Experiments

In this section, we describe the experimental results obtained for the query expansion methods presented in the previous section. We used the Microsoft Research Question-Answering Corpus[10] (MSRQA) as our evaluation dataset.

5.1 Microsoft Research Question-Answering Corpus (MSRQA)

MSRQA provides a fully annotated set of questions and answers retrieved from the Encarta 98 encyclopedia. The Encarta corpus contains 32,715 articles, ranging from very short (3 tokens) to very long (59,798 tokens). QA systems usually split documents into smaller passages. We have therefore segmented the Encarta articles into smaller parts representing subsections of the original article, using a regular expression for identifying section headers in the text. As a result, the dataset comprises 61,604 documents, with a maximum of 2,730 tokens. The relevance judgements provided comprise the document id as well as the sentences (usually one) containing the answer. We processed these sentence level relevance judgements to obtain judgements for documents: a document is considered as relevant if it contains an exact answer sentence. Overall, we obtained relevance judgements for 1,098 questions.

[9] We used the 2.2.1 version of Terrier, downloadable from `http://terrier.org/`

[10] Downloadable from `http://research.microsoft.com/en-us/downloads/88c0021c-328a-4148-a158-a42d7331c6cf/`

	All questions		**Easy questions**		**Medium questions**		**Hard questions**	
Expansion	**MAP**	**MRR**	**MAP**	**MRR**	**MAP**	**MRR**	**MAP**	**MRR**
none	0.2257	0.2681	0.2561	0.3125	0.1720	0.1965	0.1306	0.1392
Terrier-Bo1	0.2268	0.2674	0.2625	0.3157	0.1642	0.1903	0.1222	0.1240
Terrier-Bo2	0.2234	0.2602	0.2581	0.3077	0.1660	0.1872	0.1126	0.1146
Terrier-KL	0.2274	0.2684	0.2635	0.3167	0.1644	0.1915	0.1220	0.1236
WN-synonyms	0.2260	0.2687	0.2536	0.3098	0.1785	0.2055	0.1254	0.1260
DC-PRF	**0.2428**	**0.2929**	**0.2690**	**0.3361**	**0.2004**	**0.2294**	0.1385	0.1472
	+7.6%	+9.2%	+5.0%	+7.5%	+16.5%	+16.7%	+6.0%	+5.7%
DC-PRF + Terrier KL	0.2361	0.2796	0.2625	0.3184	0.1928	0.2213	**0.1389**	**0.1484**

Table 3: Experimental results. The performance gaps between the DC-PRF and the baseline retrieval models without expansion (none), Terrier-KL and WN-synonyms are statistically significant (two-tailed paired t-test, $p < 0.05$), except for hard questions and for the MAP comparison with Terrier-KL for easy questions. We also report the improvement percentage.

Based on the annotations available in the MSRQA dataset, we further distinguish three question types:

- *easy* questions, which have at least one answer with a strong match (two or more query terms in the answer).
- *medium* questions, which have no strong match answer, but at least an answer with a weak match (one query term in the answer).
- *hard* questions, which have neither a strong nor a weak match answer, but only answers which contain no query terms, and at the best synonyms and derivational morphological variants for query terms.

Overall, the evaluation dataset comprises 651 easy questions, 397 medium questions and 64 hard questions (some of these questions have no exact answer).

5.2 Results

As our baseline, we use the BB2 (Bose-Einstein model for randomness) retrieval model in Terrier. We varied the values for the parameters T (number of expansion terms) and D (number of expansion documents) and used the settings yielding the best evaluation results. For the PRF methods implemented in Terrier, the default settings (T=10, D=3) worked best; for DC-PRF, we used T=20 and D=40. Finally, for WN-synonyms we used T=10. We also combined both DC-PRF and Terrier-KL by first applying DC-PRF expansion and then using local Terrier-KL feedback on the retrieved documents (DC-PRF + Terrier KL). Prior to retrieval, all questions are tokenised and part-of-speech tagged using Xerox's Incremental Parser XIP (Aït-Mokhtar et al., 2002). Moreover, we retrieve 100 documents for each question and stem the Encarta document collection. The results shown in Table 3 are evaluated in terms of Mean-Average Precision (MAP) and Mean Reciprocal Rank (MRR). Table 4 provides examples of the top 5 expansion terms obtained for each expansion method.

The DC-PRF expansion method performs best overall, as well as for easy and medium question types. For medium questions, DC-PRF leads to an increase of 16.5% in MAP and 16.7% in MRR, with respect to the 'none' baseline. Local feedback methods, such as Terrier-KL, only bring minor improvements for easy questions, but lead to slightly lower results for medium and hard questions. This might be due to the small size of the document collection, which therefore lacks redundancy. The simple baseline expansion method based on WordNet leads to very slight improvements for medium questions over the setting without expansion. The combination of DC-PRF and Terrier-KL leads to lower results than using only

Terrier-KL	**WN-synonyms**	**DC-PRF**
12: *Are there UFOs?*		
sight – unidentifi – report – object – fly	flying – unidentified – object – UFO – saucer	unidentified – ufo – flying – ufology – objects
104: *What is the most deadly insect in the world?*		
speci – plant – feed – anim – liv	cosmos – creation – existence – macrocosm – universe	nightshade – belladonna – mortal – death – lethal
107: *When was the little ice age*		
drift – glacial – ago – sheet – million	small – slight – historic – period – water	floe – period – glacial – cold – interglacial
449: *How does a TV screen get a picture from the air waves?*		
light – beam – televi – electron – signal	moving – ridge – image – icon – ikon	television – movie – image – motion – door
810: *Do aliens really exist?*		
sedition – act – govern – deport – see	live – subsist – survive – alienate – extraterrestrial	alien – extraterrestrial – monsters – dreamworks – animation

Table 4: Expansion examples. The expansion terms produced by Terrier-KL are actually stemmed, as they are retrieved from a stemmed index.

DC-PRF, except for hard questions, for which the combination brings a very slight improvement.

The expansion samples provided in Table 4 exemplify the query drift problem of local feedback methods (Terrier-KL): for question 810, expansion terms focus on the “foreigner” sense of *alien* rather than on the “extraterrestrial” sense. The WN-synonyms method suffers from the problem of weighting synonyms, and mainly focuses on synonyms for the most frequent term of the question, e.g. “world” in question 104. Interestingly, the DC-PRF method accounts for neologisms, such as “ufology” which can be found in new collaboratively constructed resources such as Wikipedia or Wiktionary, but not in WordNet. This is made possible by the combination of diversified resources. It is also able to provide encyclopedic knowledge, such as “dreamworks” and “animation” in question 810, referring to the feature film “Monsters vs. Aliens”.

The DC-PRF method also has some limitations. Even though the expansion terms “dreamworks” and “animation” correspond to the intended meaning of the word “alien” in question 810, they nevertheless might introduce some noise in the retrieval. Some other cases exemplify slight drifts in meaning from the query: in question 104, the expansion terms “nightshade” and “belladonna” refer to poisonous plants and not insects; “*deadly* nightshade” is actually the other name of the “belladonna”. Similarly, in question 449, the expansion term “door” is obtained, in relation to the word “screen” in the question (as in “screen door”). This might be due to the fact that the terms defined by the definition clusters are indexed as well, leading to a high likelihood of retrieving syntagmatically related terms for multiword expressions. In future work, it might be relevant to experiment with different indexing schemes for definition clusters, e.g. indexing only the definitions, or adding the defined terms to the index only if they are not present in the definitions.

6 Conclusions and Future Work

In this paper, we presented a novel method for query expansion based on pseudo relevance feedback from definition clusters. The definition clusters are built across seven different English lexical resources, in order to capture redundancy while improving coverage over the use of a single resource. The expansions provided by feedback from definition clusters lead to a significant improvement of the retrieval results over a retrieval setting without expansion.

In the future, we would like to further ameliorate definition clustering and incorporate other resources, e.g. resources for specialised domains. Moreover, we have shown that query expansion based on definition clusters is most useful when applied to medium difficulty questions. We therefore consider integrating automatic prediction of query difficulty to select the best retrieval method. Finally, we would like to evaluate the method presented in this paper for larger datasets.

Acknowledgments

This work has been partially financed by OSEO under the Quæro program.

References

Agirre, Eneko, Giorgio M. Di Nunzio, Thomas Mandl, and Arantxa Otegi. 2009. CLEF 2009 Ad Hoc Track Overview: Robust - WSD Task. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece.

Aït-Mokhtar, Salah, Jean-Pierre Chanod, and Claude Roux. 2002. Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering*, 8(2-3):121–144.

Banerjee, Satanjeev and Ted Pedersen. 2003. Extended Gloss Overlaps as a Measure of Semantic Relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008+, October.

Fang, Hui. 2008. A Re-examination of Query Expansion Using Lexical Resources. In *Proceedings of ACL-08: HLT*, pages 139–147, Columbus, Ohio, June.

Kazama, Jun'ichi and Kentaro Torisawa. 2007. Exploiting Wikipedia as External Knowledge for Named Entity Recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 698–707.

Kulkarni, Anagha, Jamie Callan, and Maxine Eskenazi. 2007. Dictionary Definitions: The Likes and the Unlikes. In *Proceedings of Speech and Language Technology in Education (SLaTE2007)*, pages 73–76.

Lesk, Michael. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26.

Lund, Kevin and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical cooccurrence. *Behavior Research Methods, Instruments & Computers*, 28(2):203–208.

Macdonald, Craig, Ben He, Vassilis Plachouras, and Iadh Ounis. 2005. University of Glasgow at TREC 2005: Experiments in Terabyte and Enterprise Tracks with Terrier. In *Proceedings of the 14th Text REtrieval Conference (TREC 2005)*, Gaithersburg, MD, USA.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Navigli, Roberto. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 105–112.

Newman, M. E. J. and M. Girvan. 2004. Finding and evaluating community structure in networks. *Physical review E*, 69.

Ounis, Iadh, Christina Lioma, Craig Macdonald, and Vassilis Plachouras. 2007. Research Directions in Terrier: a Search Engine for Advanced Retrieval on the Web. *Novatica/UPGRADE Special Issue on Web Information Access, Ricardo Baeza-Yates et al. (Eds), Invited Paper*.

Qiu, Yonggang and Hans-Peter Frei. 1993. Concept based query expansion. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 160–169.

Riezler, Stefan, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu Mittal, and Yi Liu. 2007. Statistical Machine Translation for Query Expansion in Answer Retrieval. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 464–471, Prague, Czech Republic, June.

Riezler, Stefan, Yi Liu, and Alexander Vasserman. 2008. Translating Queries into Snippets for Improved Query Expansion. In *Proceedings of the*

22nd International Conference on Computational Linguistics (Coling 2008), pages 737–744, Manchester, UK, August.

Rocchio, J., 1971. *The SMART Retrieval System*, chapter Relevance Feedback in Information Retrieval, pages 313–323.

Schütze, Hinrich. 1998. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97–123.

Snow, Rion, Sushant Prakash, Daniel Jurafsky, and Andrew Y. Ng. 2007. Learning to Merge Word Senses. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1005–1014, Prague, Czech Republic, June.

Song, Dawei and Peter D. Bruza. 2003. Towards context sensitive information inference. *Journal of the American Society for Information Science and Technology (JASIST)*, 54(4):321–334.

Voorhees, Ellen M. 1994. Query expansion using lexical-semantic relations. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 61–69.

Wang, Tong and Graeme Hirst. 2009. Extracting Synonyms from Dictionary Definitions. In *Proceedings of RANLP 2009*.

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