# Application of Axiomatic Approaches to Crosslanguage Retrieval

Roman Kern[1], Andreas Juffinger[1] and Michael Granitzer[1,2]
Know-Center, Graz[1]
Graz University of Technology[2]
rkern,ajuffinger,mgrani@know-center.at

**Abstract**

Natural languages contain many ambiguous words. Detecting the correct sense of words within documents and queries could potentially improve the performance of an information retrieval system. This is the major motivation for the Robust WSD tasks of the Ad-Hoc Track of the CLEF 2009 campaign. For these tasks we have build a customizable and flexible retrieval system. The best performing configuration of this system is based on research in the area of axiomatic information retrieval approaches. Further, our experiments show that configurations that incorporate WSD information into the retrieval process did outperform those without. For the monolingual task the performance difference is more pronounced than for the bilingual task. Finally we are able to show that our query translation approach does work effectively, even if applied in the monolingual task.

## Categories and Subject Descriptors

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

## General Terms

Measurement, Performance, Experimentation

## Keywords

Information Retrieval, Co-occurrene Statistics, Word Sense Disambiguation, Crosslanguage Retrieval

# 1 Introduction

The intuition that determining the correct sense of ambiguous words could improve the performance of information retrieval systems has generated a lot of research in the last couple of years. Results in the area of monolingual retrieval could not life up to the expectations, see for example [16] and [20]. Short queries and the skewed distribution of senses are among the explanations for the observed results.

Despite moderate results for monolingual retrieval tasks, the question is still open whether WSD also has no impact on other areas of information retrieval, like for example Question Answering (QA) and Cross-language Information Retrieval (CLIR). In [12] the authors indicate that word sense disambiguation could help in multilingual retrieval.

For the CLEF2009 challenge we customized our retrieval system which has been developed for the CLEF2008 tasks, see [7]. This system is based on the open-source retrieval library Lucene[1], and has been modified to integrate different types of retrieval and ranking functions. The system contains *TFIDF* weighting schemes as provided by Lucene, the BM25 [14] weighting function and finally retrieval function utilizing axiomatic retrieval approaches [4]. In our experiments we evaluated those different retrieval functions with and without incorporating WSD information.

Results show, that the best performing runs are based on axiomatic retrieval approaches. Further, runs incorporating WSD information did outperform those without, whereas for the monolingual task the performance difference is more pronounced than for the bilingual task. Finally, we are able to show that our query translation approach does work effectively, even if applied in the monolingual task.

The paper is structuered as follows: The next section provides a detailed description of our system. In section 3 their results of the various evaluation runs are presented and the main observations are discussed. Finally section 4 concludes our findings.

# 2 Indexing & Retrieval System

Our information retrieval system consists of multiple separate components, that can be split into two groups. The first group of component processes and parses the input sources - articles and additional resources - and build the retrieval indices. The second group of components takes these indices together with the queries as input to retrieve and rank relevant documents.

## 2.1 CLEF Article Index

The document index is build using the collection of articles from the Los Angeles Times (1994) and the Glasgow Herald (1995) supplied by the organizers of the Robust WSD Task. These articles have already been tokenized, lemmatized, contain POS tags and are annotated with senses using WordNet synsets. These senses are computed using two different word sense disambiguation systems - labeled UBC [1] and NUS [2]. We will report our results for both WSD information separately in the evaluation section. For all terms that are associated with multiple senses, we took the sense with the highest score.

For indexing we used Lucene, which is an open-source search engine library implemented in Java. A single Lucene index can consist of multiple fields, that can be seen as separate indices, each with its own dictionary and statistics. We exploited this feature and for each article we created a single document that contains multiple fields. From the articles we only took the article body. The headline of the articles were not processed as they did not appear to contribute to the relevance of the articles judging by results of the experiments made with our CLEF2008 system. No stop word removal was applied in the indexing stage.

### 2.1.1 Co-occurrence Term Statistics

Using WordNet and the annotated sense of ambiguous terms it is possible to determine the synonyms for a specific sense. The relation between synonymous word are one of many semantic relatedness relationship types between words. Statistical methods provide unsupervised means to detect word pairs with a high semantic relatedness, without restriction to a specific relationship type. One of these methods is based on the co-occurrence statistics of words within a corpus. Many algorithms have been proposed to accomplish this task, using different weighting functions to measure the relationship between words. The Pointwise Mutual Information (PMI) has been been found to provide good performance in this regard [19].

The calculated similarity between words - estimated through their distribution in a corpus - can be used to enrich the retrieval approach. In our system, we utilized a query expansion technique based on the findings in [18]. Calculated on the CLEF2009 article corpus the utilized

[1] http://lucene.apache.org/java/docs/

| Field Name | Number of Terms |
|---|---|
| Word-Form | 512725 |
| Lemma | 459326 |
| Stems | 403759 |
| Synonyms (NUS) | 57840 |
| Synonyms (UBC) | 56013 |
| Synset IDs (NUS) | 55279 |
| Synset IDs (UBC) | 53292 |
| Cooccurrence Terms | 256306 |

Table 1: Number of distinct terms in each of the index field of the CLEF article index.

co-occurrence statistics uses a modified PMI measure for the similarity between two word based on the occurrence probability $P_{w_i}$ of word $w_i$:

$$S_{CondPMI}(w_i, w_j) = \frac{log_2 \frac{P(w_i|w_j)}{P(w_j)}}{log_2(\frac{1}{P(w_j)})}$$

#### 2.1.2 CLEF Article Index Fields

Each article is represented in the retrieval system using the following different fields.

**Word-Form** From the articles the word form for each token was taken as indexing term. The tokens were only marginally processed - non-letter characters were ignored and diacritical signs were removed from the letters.

**Lemma** The lemmas for each token were taken to build this field. No further processing has been applied to these terms.

**Stems** The word form tokens were stemmed using the snowball stemmer[2] and indexed in its own field.

**Synonyms** For each token of the articles the synset with the highest score was selected. For this synset all synonyms were listed using the MIT Java Wordnet Interface[3]. These synonyms were added to a dedicated field. Thus, this field contains all synonyms of the most probable sense according to the WSD annotations.

**Synset IDs** This field was filled similar to the Synonyms field, but using only the Synset ID of the highest ranked sense of each token. This can be seen as representation of the article in the WordNet Synset ID feature space.

**Co-occurrence Terms** For each term in an article the terms with the highest co-occurrence weight were selected and indexed in an separate field. The number of the associated term was limited by twice the number of terms within the article itself. For each article this field contains semantically related terms based on their distributions.

Table 1 lists the number of term in the different index fields for the 166717 CLEF articles.

### 2.2 Multilingual Index

The multilingual index is used to translate individual terms from one language to another. Again each entry in the index is made up of multiple fields. Each of these fields corresponds to a single

[2] http://snowball.tartarus.org/
[3] http://projects.csail.mit.edu/jwi/

| | Entries | English Terms | Spanish Terms |
|---|---|---|---|
| Wikipedia | 2896802 | 5139238 | 1365908 |
| Europarl | 1304243 | 88370 | 146537 |

Table 2: Statistics of the Wikipedia and Europarl multilingual indices.

language. The multilingual index can be created using various multilingual resources. We used two resources in our system, the *Wikipedia*[4] and the *Europarl* corpus[5]. Both differ largely in their characteristics, such as domain and number of distinct terms.

### 2.2.1 Wikipedia Multilingual Index

The free encyclopedia *Wikipedia* is an effort of many voluntary contributors and is continuously growing. There exist various editions in different languages that also contain links between corresponding articles. We exploited this link infrastructure to automatically build a multilingual index for all query languages, namely English and Spanish. The articles contained in the XML dumps[6] provided by *Wikimedia* organization were parsed the *Wikipedia Java API*[7]. The *Wikipedia* multilingual index thus finally contains aligned articles that are available in the two target languages.

### 2.2.2 Europarl Multilingual Index

Additionally to the Wikipedia another multilingual resource was used. The *Europarl* corpus[10] is created using the proceedings of the European parliament taken from the years 1996-2006. This resource again offers the possibility to build a sentences aligned multilingual index. We accomplish this by using the Church and Gale algorithm[5]. The Europarl corpus contains versions in 11 European languages, but for our system we used only the English and Spanish versions.

Table 2 gives an overview of the two multilingual indices. The Wikipedia index consists of whole articles whereas the Europarl index is build out of sentences. One can observe that there is huge gap in the number of terms between the two resources.

### 2.2.3 Multilingual Index Translation

The goal of the multilingual index is to find the best matching terms in a language that is different to the original language of a input term. This is achieved using information retrieval techniques. For each term to be translated, which can either be a singe word or a phrase, a query is build. This query is then used to search for relevant documents in the source language. From this result set the unique identifiers and the score of the top hits are collected. Using the identifier information the version of the hits in the target language are retrieved. From these documents the translated terms are extracted. The intuition behind this procedure is similar to selecting terms for query expansion using the top ranked documents in pseudo relevance feedback methods[11].

We implemented two scoring algorithms for selecting the best translation for the input term. The first is a simple heuristic based on the well known *TF IDF* weighting scheme. For each term the weight $w_i$ is calculated using the score of the most relevant documents $D$:

$$w_i^{TFIDF} = log(\frac{N}{docFreq_i + 1} + 1) * \sum_j^D score_j$$

The intuition behind the second scoring algorithm is to maximize the likelihood that a term has caused the document to be relevant. To accomplish this the same formula that is used to calculate

[4] http://en.wikipedia.org/wiki/Main_Page
[5] http://www.statmt.org/europarl/
[6] http://download.wikimedia.org/backup-index.html
[7] http://matheclipse.org/en/Java_Wikipedia_API

the score of a document in the source language is applied on all target language terms found in the most relevant hits. The aggregated difference between the actual score and the reconstructed score serves as base for the weight of a single term:

$$w_i^{reconstruction} = \frac{1}{\sum_j^D |tf_{i,j} * log(\frac{N}{docFreq_i+1} + 1) - score_j| + 1}$$

## 2.3 Query Processing

The first step of the query processing is the selection which parts of the topics are used for the queries. In all our experiments we used the title and description part. The narrative section of the topics was not included in the query generation process.

### 2.3.1 Query Types

Both the title and the description part of the topics do not only contain the word form of the tokens, but also offer a lemmatized version and annotations for the sense of the terms. As with the articles the sense information is also available from two different algorithms for the English topics. For the Spanish topics a first sense heuristic was applied by the organizers. Using the available features of the topics our system can be configured to generate different types of queries. Each of these query types are generated to search in the according fields of the CLEF article index. For example the synonyms for the query terms are searched in the synonyms field of the articles.

**Word-Form** The word form of the tokens in the topics and description elements of the topics are used to create the query

**Lemma** The lemmatized version of the tokens are taken to build the query terms

**Stems** The word form of the tokens were processed using the same stemming algorithm using for the articles (the Spanish version of the Snowball stemmer was used for the Spanish topics)

**Synonyms** From the top scored synset of each token in the topics the synonyms according to the English WordNet were selected.

**Synset IDs** The identifier of the synset with the highest score was used as query term

**Cooccurrence Terms** For all stems in the query the terms with the highest co-occurrence weight are selected for query expansion

### 2.3.2 Query Translation

If the language of the topic differs from the languages of the articles, the query terms are individually translated. This is done using the Wikipedia and Europarl multilingual indices. For each of the two indices a weighted list of translated terms was generated and then normalized between 0 and 1 using the highest score as denominator. The sum of the two normalized scores for each term was then used as final weight for the translation candidates. The top n candidates were then added to the query as translation for a single query term. Using the training topics and relevance judgments we found that using only the two highest scoring translation terms to offer the best overall performance.

## 2.4 Document Ranking

The result of the query generation is an unordered list of terms extracted out of a topic definition. In the next step relevant documents are retrieved and ranked. The *TFIDF* [15] weighting scheme and the *BM25* [14] approach are textbook methods to this problem and demonstrated robust and reliable performance in the past. A variant of the *TFIDF* retrieval model did provide good, but not state-of-the-art performance in the CLEF2008 Robust WSD task [7]. Many of the CLEF2008 participants incorporated the BM25 approach into their retrieval systems with great success (for example [3] and [6]). We therefore also report the performance our system using an implementation of the BM25 weighting scheme[8]:

$$S_{BM25}(Q,D) = \sum_{t \in Q \cap D} \frac{tf_{t,D}}{k_1((1-b) + b * \frac{docLength_D}{averageDocLength}) + tf_{t,D}} * log \frac{N - docFreq_t + 0.5}{docFreq_t + 0.5}$$

For our main experiments we have chosen to apply findings in the area of axiomatic approaches to information retrieval. Fang and Zhai present in [4] several variations of weighting functions build using a set of axioms that constrain the properties of a weighting function. The authors did recommend one of their derived retrieval functions which has shown promising performance in their evaluation. We did adapt this function for our retrieval system. The score of a document $D$ out of $N$ documents given a set of query terms $Q$ is build using the tuning parameter $\alpha$ and $\beta$:

$$S_{Axiomatic}(Q,D) = \sum_{t \in Q \cap D} (\frac{N}{docFreq_t})^{\alpha} * \frac{tf_{t,D}}{tf_{t,D} + 0.5 + \beta \frac{docLength_D}{averageDocLength}}$$

Using the training topics we found the setting of 0.25 for $\alpha$ and 0.75 for $\beta$ to provide a satisfying performance.

### 2.5 Question Answering

Due to the fact that the Robust WSD Task is not only an information retrieval task but also a question answering (QA) task we experimented also with methods from that field [9, 17]. In question answering passage retrieval algorithms are used to find the answering passage to a question. In Tellex et al. the authors report well performing algorithms based on varios statistics of term and sentence overlap. In this work we claim that our retrieval system provides already a good ranking with the best answering documents at the top. Due to the fact that our retrieval system performs the task based on term, co-occurrence and sentence statistics we aimed to exploit a different feature - the part-of-speach (POS) graph spectrum.

The rationale to use the POS graph is that we experienced a stylistic similarity between the answering documents and the questions. The POS graph was thereby constructed for each document by a fixed number of nodes (17 POS-Tags). For each co-occurring POS tag within a sentence an edge was introduced or the appropriate edge weight was increased by one. The same procedure was also applied to each query. Based on the trainingset and the spectral difference as defined in [8] we trained a Support Vector Machine (SVM) with a linear kernel. The trained SVM was then used to rerank the result documents similar to the methods in [13]. Unfortunately none of our experiments showed a significant improvement when applying this method. A reason for this might be the type of data and the homogene document set.

## 3 Results & Discussion

The main motivation for the Robust WSD task is to measure the performance impact of using word sense disambiguation as part of a information retrieval system. A first step to determine the influence of WSD information is the creation of a state-of-the-art retrieval system that does not incorporate a disambiguation process. We tried to build such a system and then use the WSD information as an optional processing step using query expansion. The results of these two system configurations should provide insights into the influence of word sense disambiguation. To further increase the validity of the observed behavior we also report the performance of our system using query expansion based on co-occurrence term statistics. All reported performance figures were calculated using 160 test topics and relevance assessments.

[8] http://nlp.uned.es/ jperezi/Lucene-BM25/

| Token Feature | MAP | GMAP |
|---|---|---|
| Word-Form | 0.3510 | 0.1471 |
| Lemma | 0.3911 | 0.1771 |
| Stems | 0.4022 | 0.1805 |

Table 3: Baseline performance of the monolingual system using no query expansion.

| Retrieval Function | MAP | GMAP | Notes |
|---|---|---|---|
| TFIDF1 | 0.3083 | 0.1182 | *Default Lucene Boolean Query* |
| TFIDF2 | 0.3313 | 0.1331 | *Lucene Disjunction Max Query* |
| BM25 | 0.3889 | 0.1566 | *Using* $k_1 = 0.8$ *and* $b = 0.5$ |
| Axiomatic | 0.4022 | 0.1805 | |

Table 4: Baseline performance of the monolingual system using no query expansion.

### 3.1 Monolingual Performance

Table 3 gives an overview of the results of the baseline system using the different token features of the topics. The best performance is achieved using the stemmed version of the word forms. Therefore in all following evaluation runs we report only the performance of the configuration that is based on the stemmed tokens of the topics.

In table 4 different retrieval functions are compared using the CLEF2009 test collection, using the stemmed tokens of the title and the description of the topics without query expansion. Although this comparison gives no insights into the question whether WSD information could improve the performance, it demonstrates that the results of the axiomatic approach is indeed a valuable contribution to the arsenal of information retrieval techniques. The according GMAP metric is improved over the BM25 run, which indicates that especially low performing topics did improve using the axiomatic approach.

For the comparison with the configurations that utilize the WSD information we only report the performance figures achieved using the axiomatic retrieval function. Table 5 lists the performance metrics of the various query expansion configurations. The best performing configuration combines the synonym, synset and term co-occurrence information. The performance figures do show that integrating the words sense disambiguation data into the retrieval process of our system does improve performance. Not only does the baseline configuration benefit from the sense annotations, but also the configuration that already uses a (successful) query expansion technique is improved further. The difference between the two WSD data sets (NUS and UBC) and between the Synonym and the Synset features are too small to allow any conclusions. The p-values are calculated via a Wilcoxon signed rank test using R[9] and reflect whether the improvement over the baseline (or the query expansion using co-occurrence statistics for the last two runs) is statistically significant.

### 3.2 Bilingual Performance

For the Spanish topics of the Robust WSD task we added the translation step into the query processing as described in section 2.2.3. This processing step is executed prior to the query expansion step. The baseline performance of our system using this configuration is listed in table 6. Using another language for the queries than the languages used for the documents has clearly a negative effect on the performance of our system[10]. Using the stemmed version yields the best performance.

As in the monolingual task the axiomatic retrieval function outperforms the other retrieval

[9]http://www.r-project.org/

[10]We had considerable trouble processing the Spanish topics as their did contain numerous encoding errors, leading to a worse performance

| Query Expansion | MAP | GMAP | p-value |
|---|---|---|---|
| Synonyms (NUS) | 0.4061 | 0.1849 | 0.9624 |
| Synonyms (UBC) | 0.4036 | 0.1837 | *0.6714* |
| Synset IDs (NUS) | 0.4047 | 0.1856 | 0.9697 |
| Synset IDs (UBC) | 0.4070 | 0.1869 | 0.9881 |
| Cooccurrence Terms | 0.4170 | 0.1864 | 0.9999 |
| Cooccurrence Terms + Synonyms + Synset IDs (NUS) | 0.4222 | 0.1947 | 0.9826 |
| Cooccurrence Terms + Synonyms + Synset IDs (UBC) | 0.4212 | 0.1942 | *0.8397* |

Table 5: Performance of the monolingual system using a combination of features.

| Token Feature | MAP | GMAP |
|---|---|---|
| Word-Form | 0.2619 | 0.0629 |
| Lemma | 0.2702 | 0.0570 |
| Stems | 0.2885 | 0.0746 |

Table 6: Baseline performance of the bilingual system without query expansion.

functions. The results for the Spanish topics are listed in table 7 using the same parameters as for the monolingual runs.

For the next evaluation runs we added the WSD information to our retrieval system, which again resulted in a performance improvement, see table 8. The gap between the best configuration and the baseline is just about 1%. The difference between the configuration that incorporate the WSD information are not statistically significant better than their respective baseline.

### 3.3 Translation Impact

For our final evaluation runs we investigated the impact of the query translation step. The motivation for this are the findings in [7] where the authors state that the query translation did not cause serious performance deterioration even if both the query and the documents are of the same language. Table 9 summarizes the performance of our system using different languages and query translation functions. The results demonstrate that using our approach to translate a query does not have a pronounced negative effect on the retrieval performance when using the best performing translation strategy.

## 4 Conclusion

In order to investigate the influence of words sense disambiguation in the area of cross language retrieval we built a system that can be operated in a number of configurations. This system was designed in a way to also study the performance of different retrieval functions. Additionally to the well known *TFIDF* weighting scheme and the *BM25* ranking function we adapted a retrieval function that has been developed using an axiomatic approach to information retrieval. This

| Retrieval Function | MAP | GMAP |
|---|---|---|
| TFIDF1 | 0.1992 | 0.0472 |
| TFIDF2 | 0.2445 | 0.0665 |
| BM25 | 0.2713 | 0.0658 |
| Axiomatic | 0.2885 | 0.0746 |

Table 7: Performance of the different retrieval functions for the bilingual task.

| Query Expansion | MAP | GMAP | p-value |
|---|---|---|---|
| Synonyms (1st) | 0.2923 | 0.0762 | *0.9090* |
| Synset IDs (1st) | 0.2933 | 0.0773 | *0.7813* |
| Cooccurrence Terms | 0.2917 | 0.0718 | 0.9910 |
| Cooccurrence Terms + Synonyms + Synset IDs (1st) | 0.2982 | 0.0746 | *0.7141* |

Table 8: Performance of the bilingual system using a combination of features.

| Language & Translation Function | MAP | GMAP |
|---|---|---|
| English TFIDF | 0.3979 | 0.1570 |
| Spanish TFIDF | 0.2885 | 0.0746 |
| English Reconstruction | 0.3942 | 0.1618 |
| Spanish Reconstruction | 0.2086 | 0.0379 |

Table 9: Performance of the different translation functions.

method did provide the best performance, not only for the monolingual task, but also for topics that are formulated in a language other than the language of the documents. For the bilingual retrieval task we developed a translation mechanism based on the freely available Wikipedia and the Europarl corpus.

In our evaluation runs we have found that incorporating the word sense disambiguation information does indeed improve the performance of our system by a small margin. This was the case for the monolingual and the bilingual task, although for the bilingual task the improvements are not statistically significant. Also when using the WSD information additionally to an existing query expansion technique the performance was further improved. In none of our tests we observed that the performance did decrease when applying the word sense disambiguation information. Just on a few queries there has been a negative impact. The reason for this and possible means to detect and to avoid poor performing queries are still open questions and require further research.

# Acknowledgements

The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

# References

[1] E. Agirre and O.L. de Lacalle. UBC-ALM: Combining k-nn with SVD for WSD. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval, Prague, Czech Republic)*, 2007.

[2] Y.S. Chan, H.T. Ng, and Z. Zhong. NUS-PT: Exploiting parallel texts for word sense disambiguation in the english all-words tasks. *Proceedings of Proceedings of SemEval*, pages 253–256, 2007.

[3] L. Dolamic, C. Fautsch, and J. Savoy. UniNE at CLEF 2008: TEL, Persian and Robust IR. 2008.

[4] H. Fang and C.X. Zhai. An exploration of axiomatic approaches to information retrieval. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 480–487. ACM New York, NY, USA, 2005.

[5] W.A. Gale and K.W. Church. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102, 1994.

[6] J. Guyot, G. Falquet, S. Radhouani, and K. Benzineb. UNIGE Experiments on Robust Word Sense Disambiguation. 2008.

[7] A. Juffinger, R. Kern, and M. Granitzer. Exploiting Cooccurrence on Corpus and Document Level for Fair Crosslanguage Retrieval. 2008.

[8] A. Juffinger, R. Willfort, and M. Granitzer. Spectral web content trend analysis. In *Proc. of IADIS International Conference WWW/Internet*, 2009.

[9] B. Katz, G. Marton, G. Borchardt, A. Brownell, S. Felshin, D. Loreto, J. Louis-Rosenberg, B. Lu, F. Mora, S. Stiller, O. Uzuner, and A. Wilcox. External knowledge sources for question answering. In *Proceedings of the 14th Annual Text REtrieval Conference (TREC2005), November 2005, Gaithersburg, MD.*, 2005.

[10] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, 2005.

[11] C.D. Manning, P. Raghavan, and H. Schtze. *Introduction to information retrieval*. Cambridge University Press New York, NY, USA, 2008.

[12] D.W. Oard and B.J. Dorr. A survey of multilingual text retrieval. 1998.

[13] T. Pahikkala, E. Tsivtsivadze, A. Airola, J. Boberg, and T. Salakoski. Learning to rank documents for ad-hoc retrieval with regularized models. In *Proceedings of SIGIR 2007 Workshop Learning to Rank for Information Retrieval*, 2007.

[14] S.E. Robertson, S. Walker, S. Jones, MM Hancock-Beaulieu, and M. Gatford. Okapi at TREC-4. In *Proceedings of the Fourth Text Retrieval Conference*, pages 73–97, 1996.

[15] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. 1987.

[16] M. Sanderson. Word sense disambiguation and information retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 142–151. Springer-Verlag New York, Inc. New York, NY, USA, 1994.

[17] S. Tellex, B. Katz, J. Lin, G. Marton, and A. Fernandes. Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003.

[18] E. Terra and C.L.A. Clarke. Scoring missing terms in information retrieval tasks. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 50–58. ACM New York, NY, USA, 2004.

[19] P.D. Turney. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. *Lecture Notes in Computer Science*, pages 491–502, 2001.

[20] E.M. Voorhees. Natural language processing and information retrieval. *Lecture notes in computer science*, 1714:32–48, 1999.