Pseudo Relevance Feedback Using Named Entities for Question Answering

Luiz Augusto Pizzato and **Diego Mollá**
CLT - Macquarie University
Sydney, Australia
{pizzato, diego}@ics.mq.edu.au

Cécile Paris
ICT - CSIRO
Sydney, Australia
Cecile.Paris@csiro.au

Abstract

Relevance feedback has already proven its usefulness in probabilistic information retrieval (IR). In this research we explore whether a pseudo relevance feedback technique on IR can improve the Question Answering task (QA). The basis of our exploration is the use of relevant named entities from the top retrieved documents as clues of relevance. We discuss two interesting findings from these experiments: the reasons the results were not improved, and the fact that today's metrics of IR evaluation on QA do not reflect the results obtained by a QA system.

1 Introduction

Probabilistic Information Retrieval estimates the documents' probability of relevance using a small set of keywords provided by a user. The estimation of these probabilities is often assisted by the information contained in documents that are known to be relevant for every specific query. The technique of informing the IR system which documents or information are relevant to a specific query is known as relevance feedback. As reported by Ruthven and Lalmas (2003), relevance feedback techniques have been used for many years, and they have been shown to improve most probabilistic models of Information Retrieval (IR).

Relevance feedback is considered as pseudo (or blind) relevance feedback when there is an assumption that the top documents retrieved have a higher precision and that their terms represent the subject expected to be retrieved. In other words, it is assumed that the documents on the top of the retrieval list are relevant to the query, and information from these documents is extracted to generate a new retrieval set.

In this paper we explore the use of a pseudo relevance feedback technique for the IR stage of a Question Answering (QA) system. It is our understanding that most questions can be answered using an arbitrary number of documents when querying an IR system using the words from the topic and the question. Because QA is normally a computationally demanding task, mostly due to the online natural language processing tools used, we believe that IR can help QA by providing a small list of high-quality documents, i.e. documents from where the QA system would be able to find the answer. In this sense, documents containing answers for a question in a sentence structure that can be easily processed by a QA system will be highly relevant.

In this work, we describe an experiment using a pseudo relevance feedback technique applied over a probabilistic IR system to try to improve the performance of QA systems. Since most types of factoid questions are answered by named entities, we assume that documents addressing the correct topic but not containing any named entity of the expected answer class would have a low probability of relevance regarding QA. Therefore, we hypothesise that documents containing named entities of the correct class have higher probability of relevance than those not containing them.

The relevance feedback applied to QA differs from the one applied to general IR in the sense that QA deals more with the presence of a passage that can answer a certain question than with the presence of its topic. In this sense, our technique focuses on feeding terms into the IR engine that could represent an answer for the questions.

Despite the fact that it is possible to apply the

Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006), pages 83–90.

technique to most question types, in this work we only report the results of questions regarding people's names. We understand that the other types of questions are as important as this and may generate different results due to the different frequency of their appearance in the documents; however people's names can provide us with concrete results since it is a type of named entity that has been widely experimented on recognisers and is likely to be present in most types of newswire texts, as those in Aquaint corpus (Graff, 2002). We performed our experiments using the Aquaint corpus and the set of question from the QA track of TREC'2004 (Voorhees, 2005).

The next section provides some background information on IR techniques applied to QA. Section 3 explains the principles behind the named-entity relevancy feedback technique and how we implemented it. Section 4 focuses on the evaluation of the technique regarding its use as an IR tool and as a module of a QA system. Section 5 presents the concluding remarks and future work.

2 Document Retrieval for Question Answering

Question Answering is the field of research that focuses on finding answers for natural language questions. Today, QA focuses on finding answers using textual documents, as in the TREC QA Tracks (Voorhees and Tice, 2000). Although finding answers by first populating a database with likely answers to common questions and then consulting the database at question time is still an interesting task from an information extraction point of view, most research in this area is focusing on online open domain question answering using large collections of documents.

Research on offline/database and online/textual QA styles has shown that using offline/database information is possible to achieve a higher precision with the cost of a lower recall comparing with online/textual information (Mur, 2004). Even though methods using textual corpora have not yet obtained a precision high enough for practical applications, a large amount of question types can hypothetically be answered.

Most QA systems follow a framework that involves processing the question, finding relevant documents and extracting the required answer. The majority of QA systems tend to apply their complex methodologies on both ends of this framework (on question analysis and on answer extraction), but in order to extract the correct answer for a question, a QA system needs to find a document that contains the answer and some supporting evidence for it. Therefore, one of the fundamental stages of a QA system is the document retrieval phase. It does not matter how advanced the techniques used by the QA are if the retrieved set of documents does not include the answer it requires.

In Section 4 we show that, using just the question topic as the IR query, it is possible to obtain reasonable results on a QA system. However, depending on the complexity of the techniques used by the QA system, it is necessary to reduce the retrieval set to a minimal and optimal number of documents in order to allow the completion of the task in a reasonable time.

Some work has been done on specific IR models for aiding the QA task. The work of Monz (2004) defines a weighting scheme that takes into consideration the distance of the query terms. Murdock and Croft (2004) propose a translation language model that defines the likelihood of the question being the translation of a certain document. Tiedemann (2005) uses a multi-layer index containing more linguistic oriented information and a genetic learning algorithm to determine the best parameters for querying those indexes when applied for the QA task. In other words, Tiedemann argues that since question answering is an all-natural language task, linguistic oriented IR will help finding better documents for QA. However, just the use of extra information may not necessarily improve QA when it is important to know the right configuration of your modules for the task.

Another way of limiting the amount of information sent to the QA system is by selecting the best passages or sentences that a QA system will analyse. Some IR work focuses on improving QA by passage retrieval re-ranking using word overlap measures. For instance, Tellex et al. (2003) compared a group of passage retrieval techniques and concluded that those that apply density-based metrics are the most suitable to be used on QA.

Since most IR is treated as a blackbox by the QA community, manipulation of the IR results is normally performed by query modification. The most common query modifications are lexical substitution and query expansion. These techniques seem to be obligatory for most QA systems when

the original retrieval set has a small recall. However, it tends to reduce the precision in a way that harms QA by introducing documents of unrelated subjects. On the other hand, White and Sutcliffe (2004) have shown that since only a small amount of terms from questions match the supporting answer sentence, it is important for QA systems that rely on word overlap to apply some semantic or morphological expansion.

Another way of modifying the IR phrase is by performing passive to active voice transformation of the question, as in Dumais et al. (2002). This has been shown to work well since some IR systems give preference to the distance and order of terms in the query by making the affirmative voice of the answers preferable over the passive one of the questions.

Most IR research applied to QA use similar metrics to Roberts and Gaizauskas (2004) to evaluate their systems. These metrics, defined by the authors as coverage and redundancy, evaluate respectively the percentage of a set of questions that could be answered using the top-N documents of a retrieval set, and how many answers on average it finds.

It is understandable that this metrics are closely related to the needs of QA systems, and we will show that even though they provide us with the information of how likely we are of finding the answer in the retrieval set, they do not guarantee a better QA performance. This and other issues are addressed in the following sections.

3 Relevance Feedback Using Named Entities

Because named entities are required as answers for most fact-based questions, we are hypothesising that a relevance feedback mechanism that focuses on this kind of information will be useful. Therefore, we are focusing on the QA concept of relevance by trying to reduce the number of documents that would not be able to answer a factoid question. By doing this, not only the process will guide the document retrieval towards documents relevant to the question topic (general IR relevancy) but also towards those containing entities that could answer the question (QA relevancy).

Let us say that we have a question *Q* of a topic T[1] and a probabilistic IR engine using the combination *Q+T* to obtain *R1* as a set of documents. Our process applies a named entity recogniser over the top-N ranked documents of *R1*, thus obtaining a set of named entities *E*. The feedback process consists of enriching the previous query as *Q+T+E* in order to obtain a new set of documents *R2*.

Our expectation on this technique is that not only documents containing the correct answer in *R1* will be boosted in ranking on *R2*, but also that documents that have a high ranking in *R1* and do not contain any name entity of the expected answer type will then be demoted in *R2*. Therefore, documents that theoretically would not contribute to the QA performance will not take part on the answer extraction phase, allowing their slots of processing to be occupied by other more relevant documents.

In order to exemplify this process, consider the TREC 2005 QA Track question 95.3 regarding *the return of Hong Kong to Chinese sovereignty*: "*Who was the Chinese President at the time of the return?*"

The first phase of the process is the question analysis that defines what the expected answer type is and what the question main words are. Then the question and its topic define an IR query that generates the retrieval set R1.

The next process extracts the named entities of the expected answer type of the first *N* documents in the R1 set of documents. For the example, fifteen names of people were extracted, mostly Chinese and all of them related with politics. A new IR query is built using these fifteen names and the final set *R2* of documents is retrieved.

The list of names found for this query is listed on Table 1. We can observe that, among those names there is the correct answer for the question (*President Jiang Zemin*), which helped generating a better retrieval for this question with the pseudo relevance feedback mechanism.

Table 1: Extracted Named Entities

President Mario Alberto N. L. Soares	President Jiang Zemin
General Secretary Aleksandr Zharikov	Minister Qian Qichen
Minister Sabah Al- Ahmad Al-Jaber	Minister Zhou Nan
Prime Minister Mahmoud Zouebi	Mr. Deng Xiaoping
President Maumoon Abdul Gayoom	Premier Li Peng
President Ugo Mifsud Bonnici	Liu Huaqiu
President Meets Chinese	laws Will
President Leonid Kuchma	

[1]The distinction between topic (or target) and question is made clear on recent TREC QA Tracks (Voorhees, 2005). However, most cases of question answering systems would have the topic extracted from the question itself.

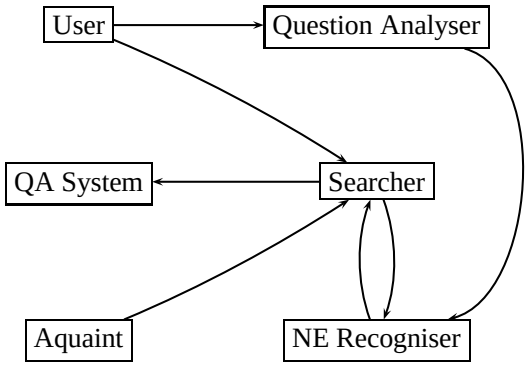

Figure 1: System overview for the the relevance feedback process.

Our hypothesis is that the named-entity feedback technique improves the overall document retrieval for QA by providing a retrieval set of documents that facilitates the extraction of the correct answer by a QA system. The technique should theoretically improve good questions (where a correct feedback is obtained) and not deteriorate bad ones[2]. Next section describes the experiments we performed and the results.

3.1 Implementation

The technique consists of posting the original question to a probabilistic IR engine, extracting the named entities of the expected answer type from the top-N results, and re-feeding the IR engine with an expanded query. By doing this, we are telling the IR system that documents containing those named entities are relevant to the question. Several implementations and set-ups can be tested using this approach, but the basic framework we implemented is shown on Figure 1.

We developed our IR system using C++ and the XAPIAN[3] Toolkit for Probabilistic IR. The Aquaint Corpus (Graff, 2002) was indexed using full text but stopwords, and it was searched using Xapian Probabilistic Methods (it uses Robertson's BM25 (Robertson et al., 1992) for ranking).

As can be seen in Figure 1, the user poses a question to the system. It is simultaneously processed by the question analyser and the searcher. The question analyser returns the expected answer type (a named-entity class for factoid questions), while the searcher returns a list of documents or snippets of text from the Aquaint corpus ranked by Xapian BM25 implementation. The named-entity recogniser receives the output of these two processes and extracts the corresponding named entities from the received files. Once this is done, it re-feeds the query to the searcher with the additional named entities. The searcher then feeds the results into the QA system.

4 Experiments and Evaluation

We use in our experiments the data collection made available by NIST on the TREC QA Tracks[4]. All the questions and judgement files of TREC 2003 QA Track were used on a preliminary evaluation of this process. Because this experiment required that all components shown on Figure 1 be fully functional, several setups were implemented, including a manual question classification (to ensure 100% correctness) and the implementation of a simple passage retrieval algorithm.

In our evaluation, we labelled documents as relevant or not relevant by assuming that relevant documents are those containing the required answer string. These early tests showed us that using the set of 500 TREC 2003 questions with our pseudo-relevance feedback technique improved the results over the initial retrieval. The improvement, however, was small and not statistically relevant.

On our system architecture, the question classification was performed using the Trie-based technique (Zaanen et al., 2005) which has a performance of around 85% accuracy when trained with the set of questions made available by Li and Roth (2002). This means that in 15% of the cases, we might have an immediate degradation of the results (by adding the wrong named-entities to the query). Because of this, we trained the classification with the same questions as the verification set. This was done to ensure complete correctness on this part of the module. However, because of the large amount of expected answer types present in the classification we used, named entity recognition proved to be a particularly complex task.

Since many questions required numbers as their answers and most documents contain some kind of number, defining a document relevant and using numbers as indication of relevancy does not work well. This demonstrated that even though we obtained better overall results using all categories

[2] A question is bad when used as a query on an IR system, it is unable to retrieve any document containing an answer.

[3] `http://www.xapian.org/`

[4] `http://trec.nist.gov/data/qa.html`

available, some of them were a real challenge for the evaluation.

We also observed that some named-entity classes could not be properly identified by our named-entity recogniser. Therefore we shifted our attention to only people's names, as we understood them to be less likely to suffer from the issues above reported. We also started to use two well known named entity recognisers: Lingpipe[5] and ANNIE[6] on Gate.

The evaluation was performed intrinsically and extrinsically in the same sense as Spärck Jones and Galliers (1996). Intrinsic and extrinsic evaluations differ because the former evaluates a system according to its primary function, while the latter evaluates a system according to its function or its setup purpose. In our study, the evaluation was performed using the combined set of questions and topics of the TREC 2004 and 2005 along with their respective judgement sets. Different setups were experimented, but mainly variations of passage window, the number of top documents used and the weights assigned to the different components (*T*, *Q* and *E*) of the query. We extrinsically evaluated the effectiveness of the retrieval sets by the percentage of correct answers the AnswerFinder(Molla and van Zaanen, 2006) system generated, and intrinsically evaluated the same sets of documents using the standard precision metric for IR and other metrics defined by Roberts and Gaizauskas (2004) for IR on QA:

- **Precision**: percentage of related documents over all questions;
- **Coverage**: percentage of questions that potentially could be answered using the top-*N* documents; this means that at least one of the top-*N* documents potentially answers the question; and
- **Redundancy**: average of how many answers can be found using the top-*N* documents;

We applied the retrieved document set on AnswerFinder and measured the exact results using the patterns made available by Litkowski on the TREC QA Data Webpage.

[5] `http://www.alias-i.com/lingpipe/`
[6] `http://gate.ac.uk/ie/annie.html`

4.1 Results

Our evaluation focused on using pseudo relevance feedback to enrich the IR query used by QA systems to find some documents that could answer natural language questions. We performed an intrinsic evaluation using some standard metrics for IR on QA, and, at the same time, we also performed an extrinsic evaluation by using the retrieval set on the QA system.

Sets of documents were retrieved using a combination of Topics (*T*), Questions (*Q*), Entities (*E*) and Answers (*A*). The following combinations were tested:

- *T*: Only the topic is sent as a query. This set of queries evaluates the potentiality of improving the retrieval set that NIST provides for every topic.
- *TQ*: The queries are made of Topic and Question. This is the current retrieval set used by the AnswerFinder system.
- *TQE*: This is the feedback technique, where Topic, Question and the Named Entities extracted from top-N documents are combined;
- *TQA*: This is the optimal feedback technique, where Topic, Question and Answers are combined. This set evaluated how far from the optimal retrieval we are;
- *TQEA*: These queries combine the feedback technique with the answers, so we can measure the amount of noise introduced by adding bad named entities. We made sure that a named entity that was also the answer was not introduced twice so its score would not be erroneously duplicated on the query.

Different combinations could also be tested, for instance *TA*, *TE* or just *A*, *E* and *Q*. We understand that those and other combinations could provide some insight on certain matters, but we believe that they would not represent a realistic retrieval set. It is a fact that the terms from *T* must be present in the retrieval set, since all documents must address the correct topic. For instance, including *Q* without having *T* will not generate a relevant retrieval because the subject of the question is not present. Also, including *A* or *E* without *Q* and *T* may represent a totally different retrieval that is not desired in this study.

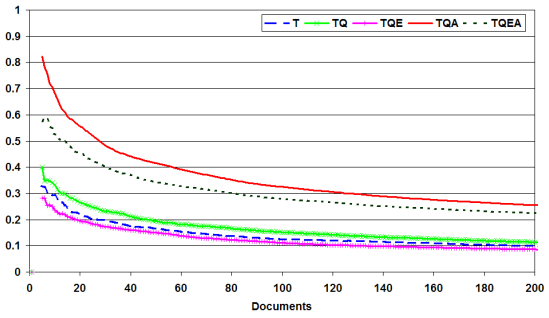

Figure 2: Precision

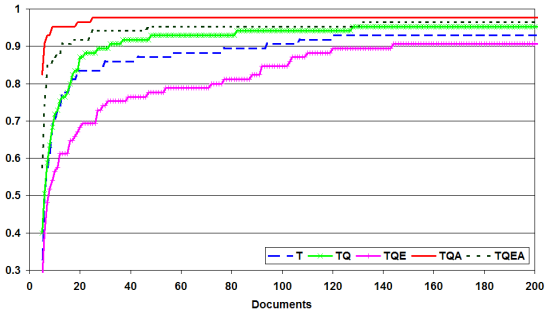

Figure 3: Coverage

The precision, coverage and redundancy obtained for the TREC 2004 and 2005 questions regarding people's name are respectively shown in Figures 2, 3 and 4. We note that the results for the feedback technique do not improve the results on neither *T* nor *TQ* on any of the measures we obtained. As expected, the addition of the answer on *TQA* represents the optimal retrieval set, obtaining the coverage of 86% on the first document per question and over 90% on the second.

The noise introduced on TQEA is not a major concern when the answers are involved in the

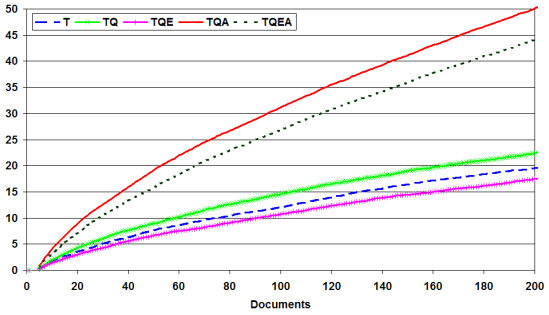

Figure 4: Redundancy

query. This is an indication that most entities found by the feedback mechanism do not represent an answer. This raises two issues: how to improve the technique so that the answers are included in the feedback; and how to minimise the noise so that a potential good feedback is not worsened.

To address the first problem we can foresee two solutions: one is improving the accuracy of the named-entity recogniser, something we cannot address in this study. The other is increasing the search space without adding more noise in the query. This is a difficult task and it could be achieved by finding the smallest possible windows of text containing the answer on several documents. We performed some experiments using different numbers of documents and variable passage size, at the moment fewer documents and smaller passages provide our best results.

We understand that documents in the first retrieval set *R1* will contain named-entities of the same type, but not necessarily the correct one (the answer), thus creating some noise in the query. We believed that a certain degree of noise would not hurt the retrieval performance. However, our experiments, as shown, demonstrate otherwise. The noise created by erroneous entities affects the performance once the elements in *E* become more important than the elements in *Q*. Because we cannot guarantee the correctness of any of the named-entities included in *E*, the resulting retrieval set *R2* might represent a worse retrieval set than *R1*. However, these cases may not influence the results in a QA system since *R1* would also not lead to the correct result.

This shows that our feedback technique suffers from the same flaws most pseudo-feedback techniques have. For instance Ruthven and Lalmas (2003) show that when the initial retrieval set is not good, the pseudo-feedback techniques is likely to worsen the results because, instead of bringing the query closer to the topic at hand, it will take it further away (a phenomenon called query drift). We hypothesise that since our technique is meant to be applicable over a QA system, if the initial set of results is bad (i.e. it does not contain the answer), there is not much that can be worsened. To confirm this hypothesis, it is necessary to perform an evaluation over a QA system. Table 2 shows the runs of QA performed using the same set of questions of the intrinsic evaluation and the documents retrieved by the retrieval sets

Table 2: Correct Answers on AnswerFinder

Run	Exact
T	19.6%
TQ	28.6%
TQE	23.2%
TQA	28.6%
TQEA	32.1%

shown before.

What can be observed here is that the feedback technique (*TQE*) offers a better set of documents than the one using only the topics (*T*). However, they are still worse than the topic and question ones (*TQ*). An interesting result is that *TQEA* is the best run, which may show that the inclusion of entities can help improve QA. We have not yet performed a deep analysis of this case to verify its cause. Even though our process did not show improvements over the baseline techniques, it was very important to find that the results of the (intrinsic) evaluations of the IR component do not parallel the results of the (extrinsic) evaluation of the QA system. In spite of the fact that high precision, coverage and redundancy represent a better chance of finding answers, we show that they do not guarantee a better performance over a QA system.

Comparing the results of *T* and *TQ* it is possible to observe that they are very similar on the intrinsic evaluation and quite different on the QA system. Therefore, what appears to help question answering is the presence of more context words so that the answers not only appear in the document but are also present in the context of the questions. This is mostly due to the fact that most QA systems tend to work with full discourse units, such as sentences and paragraphs, and the selection of those are normally based on words from the topic and the question.

In summary our experiments did not confirm the hypothesis that named-entities feedback would help improving QA. But, in the ideal situations where the answers are identified and included in the queries, the improvements are clear under an intrinsic evaluation. The differences between the intrinsic evaluation and extrinsic one point out that there are many issues that IR metrics are not currently covering.

5 Concluding Remarks and Future Work

In this paper, we have looked at whether a pseudo relevance feedback mechanism could help the QA process, on the assumption that a good indication of a document relevancy for its usage on a QA system is the presence of named entities of the same class required as the answer for a certain question. Our assumption was based on the fact that documents not containing those entities are less likely to help provide the correct answer and every entity of the right type has a probability of being the answer.

We have described our evaluation of the hypothesis using known IR metrics and a QA system. Our main conclusions are:

- Because we have not yet reported satisfactory results, we believe that even though the method is conceptually sound, it will not produce good results unless a more sophisticated control over the introduced noise is achieved; and

- The evaluation of the technique brought to our attention the fact that it is not possible to state that a retrieval technique is better just by relying on conventional IR evaluation metrics. The differences on the intrinsic and extrinsic evaluations demonstrate that there are many hidden variables that are not taken into account in metrics such as precision, coverage and redundancy.

As further work, we plan to repeat our evaluation using different QA systems, since other QA systems may give preference to different text features offering a better insight on how the IR evaluation metrics correlate with the QA results. We are also planning to use the named-entity recogniser that is being developed by our research group and to extend the system to use a more advanced passage retrieval algorithm.

References

Susan Dumais, Michele Banko, Eric Brill, Jimmy Lin, and Andrew Ng. 2002. Web question answering: is more always better? In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 291–298, Tampere, Finland. ACM Press.

David Graff. 2002. The AQUAINT corpus of english news text. CDROM. ISBN: 1-58563-240-6.

Karen Spärck Jones and Julia R. Galliers. 1996. *Evaluating Natural Language Processing Systems: An*

Analysis and Review. Springer-Verlag New York, Inc.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics*, pages 556–562, Taipei, Taiwan. Association for Computational Linguistics.

Diego Molla and Menno van Zaanen. 2006. Answerfinder at TREC 2005. In *The Fourteenth Text REtrieval Conference (TREC 2005)*, Gaithersburg, Maryland. National Institute of Standards and Technology.

Christof Monz. 2004. Minimal span weighting retrieval for question answering. In *Proceedings of the SIGIR-2004 Workshop on Information Retrieval For Question Answering (IR4QA)*, Sheffield, UK, July.

Jori Mur. 2004. Off-line answer extraction for dutch qa. In *Computational Linguistics in the Netherlands 2004: Selected papers from the fifteenth CLIN meeting*, pages 161–171, Leiden University.

Vanessa Murdock and W. Bruce Croft. 2004. Simple translation models for sentence retrieval in factoid question answering. In *Proceedings of the SIGIR-2004 Workshop on Information Retrieval For Question Answering (IR4QA)*, Sheffield, UK, July.

Ian Roberts and Robert J. Gaizauskas. 2004. Evaluating passage retrieval approaches for question answering. In Sharon McDonald and John Tait, editors, *ECIR*, volume 2997 of *Lecture Notes in Computer Science*, pages 72–84. Springer.

Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aarron Gull, and Marianna Lau. 1992. Okapi at TREC. In *Text REtrieval Conference*, pages 21–30.

I. Ruthven and M. Lalmas. 2003. A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review*, 18(2):95–145.

Stefanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes, and Gregory Marton. 2003. Quantitative evaluation of passage retrieval algorithms for question answering. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 41–47, Toronto, Canada. ACM Press.

Jorg Tiedemann. 2005. Optimizing information retrieval in question answering using syntactic annotation. In *Proceedings of RANLP 2005*, pages 540–546, Borovets, Bulgaria.

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 200–207, Athens, Greece. ACM Press.

Ellen M. Voorhees. 2005. Overview of the TREC 2004 question answering track. In *Text REtrieval Conference*.

Kieran White and Richard F. E. Sutcliffe. 2004. Seeking an upper bound to sentence level retrieval in question answering. In *Proceedings of the SIGIR-2004 Workshop on Information Retrieval For Question Answering (IR4QA)*, Sheffield, UK, July.

Menno Van Zaanen, Luiz Augusto Pizzato, and Diego Molla. 2005. Classifying sentences using induced structure. In *Proceedings of the Twelfth Symposium on String Processing and Information Retrieval (SPIRE-2005).*, Buenos Aires, Argentina, November.