Prism: An Answer Projection System

by

Lynn Wu

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 2003
[September 2003]

Author ....
Department of Electrical Engineering and Computer Science
August 8, 2003

Certified by ....
Boris Katz
Principal Research Scientist
Thesis Supervisor

Accepted by ....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Prism: An Answer Projection System

by

Lynn Wu

Submitted to the Department of Electrical Engineering and Computer Science
on August 8, 2003, in partial fulfillment of the
requirements for the degree of
Master of Engineering

Abstract

Prism is an answer projection system that combines the best attributes of traditional and Web-based question answering systems. Traditional question answering systems retrieve answers with pinpoint accuracy but limited coverage, using a small and reliable data source, while Web-based systems retrieve answers with broad coverage but limited accuracy, using a large but noisy data source. By taking advantage of the strengths of each system, an answer projection system can answer as many questions as a Web-based system while still being as accurate as a traditional system. In fact, Prism improves the performance of a traditional question answering system by 25%. It improves the accuracy of answers retrieved from the World Wide Web by 10%, and more importantly, it verifies answers retrieved from the Web by providing reliable supporting documents. By combining Prism with traditional and Web-based systems, we obtain a question answering system with high accuracy and broad coverage.

Thesis Supervisor: Boris Katz
Title: Principal Research Scientist

Acknowledgments

I could not have completed this thesis without help. Here is an incomplete list of those to whom I owe thanks.

- Boris, for his guidance over the past two years.
- Greg and Jimmy for all the admonishments, discussions, and advice.
- Stefie, for patiently answering my questions even the stupid ones.
- Anant, Alexy, Lilyn, Sheetal and Sue who read drafts.
- Peter, for helping me days and nights during this time.
- My family for their love and support.

Dedication

To my family and my friends at MIT. And to Stefie's cat.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The amount of information available online is growing rapidly, and as Internet use becomes widespread, anyone can add to this massive information repository. However, information access technology is not yet mature enough for people to rapidly find specific information. Past research into finding specific information focused on document retrieval using keyword search, with Google being a very popular example. Unfortunately, keyword searches have serious drawbacks. For example, if a user asks a question such as "When did Alaska become a state?", the user must spend time manually searching for the result in the corpus of documents returned. A more efficient alternative is the question answer search. The goal of this type of search is to return a concise but complete response to a question, rather than an entire document or set of documents.

Answers may be easier to find in a large corpus of documents, such as the Web for the following reasons. First, more information is available. Second, the same information may be restated in a variety of ways. In a small corpus of information, an answer phrased in a different way from the question asked can be missed. However, if the large corpus is unreliable, it is important to have a way to check the reliability of the answer. One approach to improve reliability is a technique known as answer projection. Answer projection involves taking a question and a candidate answer to the question and attempting to verify the answer by searching for evidence in a corpus of documents that is considered to be more reliable.

This thesis presents Prism, a system that uses answer projection to perform question answer searches. Prism takes a question and searches for candidate answers from the largest corpus of documents available, the Web, using an existing Web-based system. Because the Web is so large, it is likely that we can find some answer to any question on the Web; by the same token, because there is no control over adding information to the Web. the answers may not be very reliable. Therefore, Prism attempts to verify the answer by feeding both the question and the candidate answer into a variant of a traditional question-answer search engine, which then tries to find a supporting paragraph for the answer from a smaller, more reliable corpus of documents.

1.1 Motivations

Below I describe two scenarios that motivate the answer projection system.

Tom, a software developer, experiences an obscure bug using Microsoft's Visual Studio .NET. He searches the official documentation published by Microsoft, but unfortunately, the bug was so obscure that he was unsuccessful in finding any information describing his exact situation. Elsie suggests to Tom that perhaps he should try to search the mailing list and discussion group archives for Visual Studio .NET. It is likely that someone else has already encountered a similar situation. Tom takes Elsie's suggestion and searches the discussion archive which is a much larger and noisier information source than Microsoft's official documentation. Indeed, Tom is able to find several answers that can potentially lead to the solution. However, answers provided by the discussion group are short and not very descriptive and sometimes inaccurate. One of the answers suggests that the bug is caused by an XML transformation syntax error and it can be fixed by using the correct XML transformation. Tom then searches through Microsoft's documentation again using the additional keyword "XML transformations." Because he knows the exact answer he is looking for, Tom easily finds the description of the functionality that solves his problem in the official documentation, although he failed to find it by searching the documentation initially using his own description of the bug. Prism, an answer projection system, automates this search process. It finds a large number of candidate answers using a larger database, and verifies each answer by looking for a paragraph that supports the answer from a more reliable data source.

In the second scenario, Jane, a middle-school student, wants to find out who the prime minister of Israel is for a school assignment. Her Web-based QA system responds with 'Steven Spielberg'. Although the system answers the question, it is wrong because of its unreliable data source. This happens because some humorous documents on the Web claim that President Bush thinks the prime minister of Israel is Steven Spielberg. In this case, Prism can help verify the answer by projecting it onto a reliable data source. Since it obviously can not find Steven Spielsberg as the prime minister of Israel, Prism returns no supporting document. This can be used as a verification system to indicate the accuracy of answers coming from the Web.

Another motivation for answer projection is to allow the user to see the answer extracted from the Web in a context that is considered to be reliable. Many current Web-based question-answering systems provide very short answers. However, as Lin *et al.* demonstrated, users tend to prefer paragraph-sized answers over short phrases because the large answers provide more context [14]. For example, if a user asks the question "Who was the first woman killed in the Vietnam War", the system responds with "Sharon Lane." The user may have no idea how reliable this information is. However, if the following paragraph is returned, the user can judge better for himself.

> A piece of steel from a rocket that landed between Ward 4A and Ward 4B of the 312th Evacuation Hospital had ripped through Sharon Lane's aorta. She bled to death less than a month before her 26th birthday, the first woman killed by hostile fire in Vietnam.

Context can be especially important in situations where the reliability of the answer is questionable, which is often the case with answers extracted from the Web. There are two sources of unreliabilities: answers retrieved can have nothing to do with the question asked, or the data itself is erroneous. By projecting the answer into a corpus of documents that is considered reliable, and showing the context of the answer from a supporting document, question-answering systems can increase the user's trust in the answers returned.

1.2 Advantages and Drawbacks of Current Approaches to Question-answering Search

As mentioned above, question-answering search programs tend to fall into two broad categories. The first, more traditional category involves restricting the search to the small corpus of documents provided. The second approach involves trying to find an answer from the World Wide Web. Although the traditional systems produce answers that are likely to be reliable, they may fail to return an answer because they rely on the phrasing, and especially the words of the query, closely matching the answer. On the other hand, Web-based systems will almost always produce an answer, although not necessarily a very reliable one. In a small corpus of documents, the correct answer for a given question may only be present in a very small number of documents. Furthermore, current passage retrieval algorithms are sufficiently unreliable that the handful of documents that have the answer may be missed entirely. For example, assume a system is given a highly reliable set of documents about flowers. A user may ask a question about the state of flower of Hawaii, and the information is phrased using different words from the question asked. Despite the reliability of the documents provided to the system, it will be unable to answer the question because it simply could not match keywords of the question asked in any of the documents. Conversely, because the World Wide Web is so vast, systems that search it are almost guaranteed to come up with some candidate answer to a question, simply because there is in all likelihood at least one document on the Web that contains at least one word from the question. Of course, the reliability of these documents is by no means guaranteed. If the same question about the flower is asked in a Web-based system, it may find a document mistakenly stating the state flower of Hawaii is camellias. Although it finds the answer on the Web, the information is incorrect.

1.3 Contribution

The major goal of Prism is to use projection to combine the strengths of the two approaches to question answering mentioned above. Prism takes the question from the user and performs a Web-based search using Aranea, a Web-based question answering system developed at MIT, to find candidate answers to this question. It then takes keywords from both the question and a candidate answer, weights them appropriately, and forms a new query, which it then passes into a more traditional, small corpus-based question-answer system, such as Pauchok [16] in order to extract a reliable final answer. By using the Web as the data source for the initial search, Prism takes advantage of the wealth of data available to try to ensure that some answer is available. By using the candidate answer as part of a query into a traditional question-answering system, Prism can determine the reliability of the answer provided. By using projection to combine the two traditional question-answering techniques, Prism can make question-answer searches on large data sets more reliable and more informative to the user.

1.4 Outline

This thesis is organized in the following way.

Chapter 2 reviews existing answer projection systems.

Chapter 3 describes Prism's architecture.

Chapter 4 details answer projection algorithms.

Chapter 5 outlines evaluation guidelines and techniques used for Prism.

Chapter 6 provides examples and performance improvements from various answer projection algorithms. Adjusting the weighting for question and answer keywords leads to significant improvements in passage retrieval. Furthermore, Prism improves the performance of Aranea, a Web-based system in two ways: it improves answer projection significantly, and it slightly improves the accuracy of its answers.

Chapter 7 suggests future directions for answer projection.

Chapter 8 summarizes Prism's contributions to question answering research.

Chapter 2

Related Work

To date, few studies have systematically examined the value of answer projection. This chapter describes traditional systems and Web-based systems and how they can use a good answer projection system to improve its performance. It then discusses two existing kinds of answer projection systems.

2.1 Traditional Question Answering Systems

Traditional generic question answering systems use a small but reliable corpus to find answers. Many systems in TREC [18, 19] are of this type. TREC is a competition sponsored by National Institute of Standards and Technology (NIST). It allows researchers to evaluate their question answering systems by running them on a common set of questions, using a specific corpus of newspaper articles. Questions in TREC are restricted to fact-based, short-answer questions. Most of the systems employed by TREC contestants find passages within the given corpus. Typically, these systems share four common components: question type classification, document retrieval, passage retrieval, and named entity matching [18, 19]. Question type classification identifies the expected answer type of the question. For example, the answer type of "Who shot Abraham Lincoln?" might be "person." There is no standard answer type ontology, so systems vary from having a very broad to very specific answer type classifications. Next, these systems use document and passage retrieval algorithms to find passages from the corpus that are likely to contain an answer. Finally the entity matching module searches the retrieved passage to find the entity that matches the expected answer type. Prism can improve the performance of traditional question answering systems by incorporating answers retrieved from Web-based systems into the search.

2.2 Web-based Question Answering Systems

Web-based question answering systems utilize the Web to retrieve answers. Because of its massive size and redundancy, they can answer many more questions than traditional systems. Web-based systems share some similar components with the traditional question answering systems, but many components are created specifically for the Web. Aranea [13] is a a question answering system that focuses on extracting answers from the World Wide Web. It uses two techniques: knowledge annotation and knowledge mining. Knowledge annotation [9, 10] allows heterogeneous sources on the Web to be accessed as if they were a uniform database. That database contains answers for certain classes of TREC questions. When these questions are asked, Aranea looks up one of the annotated pages stored in the database and uses the annotation to find the best answer. Knowledge mining uses statistical techniques to find answers by leveraging data redundancy in the Web [13]. However, Aranea needs to find supporting documents from a reliable corpus. A good answer projection system can help Aranea find better and more reliable supporting documents.

START [9], a question-answering system developed at MIT, also uses the Web. However, unlike Aranea, START focuses on giving more detailed answers, usually in paragraph format. START uses Omnibase [10, 11], a virtual database that integrates heterogeneous data sources using an object-property-value model. However, Omnibase requires manual indexing for different kinds of knowledge bases. Although it can answer questions it knows about very precisely, the range of questions it can answer is limited. If START can use the broad range of questions Aranea can answer and project them onto a reliable data source, the range of questions START can answer can be expanded. A good answer projection system such as Prism can help START to realize this goal.

2.3 Previous Answer Projection Systems

Microsoft's AskMSR [2] is one of the first systems to perform answer projection. It uses the World Wide Web to extract answers and project them onto the TREC corpus. In their projection phase, five possible supporting documents are found for each answer using the Okapi IR system [5, 15]. The query submitted to Okapi is just the list of query words along with the candidate answer. Documents are ranked using the standard best match rank function bm25 [8].

Aranea also attempts to perform answer projection using existing passage retrieval algorithms implemented in the Pauchok [16] framework. Pauchok uses the Lucene index for document retrieval and the MultiText algorithm for passage retrieval. However, Aranea's strategy is very primitive. It simply concatenates the answer keywords with the question keywords to find supporting passages, Prism is designed to be a better answer projection system that avoids many of the pitfalls inherent in Aranea.

Authors of both AskMSR and Aranea have suggested that answer projection module needs to be strengthened to improve their systems.

Chapter 3

Architectural Overview

Prism is built upon Pauchok's architecture [16]. It modifies the query generation, document retrieval, and passage retrieval components in Pauchok to support answer projection. It adds a new component–answer selection module. Figure 3-1 shows the data flow in Prism graphically. A query is generated using a question and answer pair. The answers used for projections are results returned using Aranea, a Web-based question answering system [13]. A document retriever takes the query and returns a list of documents, and a passage retriever finds the best passage within each document. Finally, the passage selection module takes the best passage from each of the documents and finds the best one for the query.

3.1 Query Generation

The original Pauchok framework only uses queries that are generated from questions. In Prism, answers are added to the query generation process. Instead of generating one query per question, Prism generates five queries from five question-answer pairs, using the top five answers from Aranea. The sixth query is the default query that contains the question only. These six queries are used in Lucene's document retriever [1], a boolean keyword search engine, to retrieve relevant documents. The Lucene document retriever requires that the documents it retrieves contain all the keywords in a query. Aranea may not always return the correct answer; in these

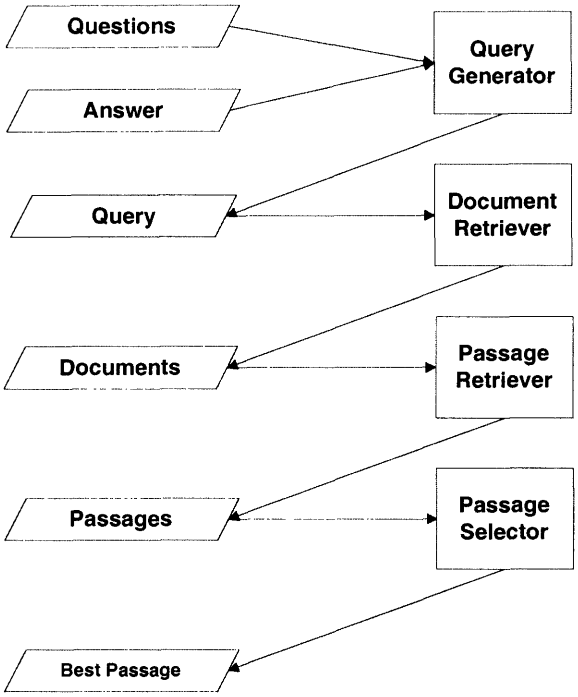

Figure 3-1: The graph above illustrates the Prism's architecture. It is a modification of the existing Pauchok architecture. Currently, both answer and questions can be used as input.

cases, the search engine may not be able to find a document containing all the question terms along with the incorrect answer terms. Even if it does, the documents may be erroneous. This limitation can be eliminated by dropping some keywords in the query or simply by using the default question. For the time being, the default question is used as a back-off, in case no appropriate document is returned using question-answer queries. Table 3.1 is an example of a typical question and its corresponding Aranea answers. From the five Aranea answers, six queries are generated. The query generator eliminates common words and punctuation.

Both document and passage retrieval algorithms use these queries to retrieve documents and extract passages. The top passage returned for each query is used as the final answer for the query. The rank of the passage is given by the rank of the Aranea answer.

What language is mostly spoken in Brazil		
	Aranea Answer	Query generated
Rank 1.	Portuguese	1. language AND mostly AND spoken AND Brazil AND **Portuguese**
Rank 2.	English	2. language AND mostly AND spoken AND Brazil AND **English**
Rank 3.	French	3. language AND mostly AND spoken AND Brazil AND **French**
Rank 4.	million people	4. language AND mostly AND spoken AND Brazil AND **million people**
Rank 5.	official	5. language AND mostly AND spoken AND Brazil AND **official**
Rank 6.		6. language AND mostly AND spoken AND Brazil

Table 3.1: An example of a typical TREC question and its corresponding Aranea answers. The rank 1 answer is the most probable answer and the rank 5 answer is the least probable answer. The rank 6 is the default query generated using only the question.

3.2 Document Retrieval

Prism uses the Lucene Indexer [1], a freely available open-source boolean keyword IR engine. It uses boolean queries to retrieve the initial set of documents and the inverse document frequency to sort the documents. In the example above, query keywords are connected using "AND", so in this case, Lucene finds documents only if they contain all the keywords in a query. Table 3.2 displays the document ID of the first five documents retrieved for each query.

Because Lucene is a boolean keyword search engine, it suffers from a common drawback of a boolean system: poor control over the size of the result set. If keywords are connected using "AND", no documents may be returned. However, if keywords are connected using "OR", documents with any of the query terms are returned, which is usually too many. This drawback can be alleviated by using a back-off: for example, dropping query terms incrementally until documents can be found. Empirically, studies by Tellex *et al.* [17] show that boolean IR systems can supply a reasonable set of documents for passage retrieval and answer extraction. In fact, many TREC systems employ simple boolean queries for this reason.

3.3 Passage Retrieval

Passage retrieval algorithms take a document and a query and find the best passage within the document. Pauchok supports many passage retrieval algorithms. At

Query	Document Number
1. language AND mostly AND spoken AND Brazil **Portuguese**	1. AP901216-0003
	2. AP890121-0027
	3. AP881107-0172
	4. SJMN91-06119067
	5. FBIS3-10134
2. language AND mostly AND spoken AND Brazil **English**	1. AP890121-0027
	2. SJMN91-06119067
	3. AP901225-0030
	4. WSJ900523-0141
3. language AND mostly AND spoken AND Brazil **French**	1. SJMN91-06119067
	2. AP901225-0030
	3. WSJ900523-0141
	4. LA040290-0137
4. language AND mostly AND spoken AND Brazil **million people**	1. AP890121-0027
	2. AP890522-0302
	3. AP880517-0204
	4. WSJ900523-0141
	5. FBIS3-10756
5. language AND mostly AND spoken AND Brazil **official**	1. AP890121-0027
	2. FBIS3-11105
	3. AP881107-0172
	4. FBIS3-32308
	5. AP890522-0302
6. language AND mostly AND spoken AND Brazil	1. AP901216-0003
	2. AP890121-0027
	3. FBIS3-11105
	4. AP881107-0172
	5. FBIS3-32308

Table 3.2: The first five queries are generated from question-pairs, and the sixth is from the default question. Aranea answer keywords are shown in bold font. Prism retrieves document using Lucene's search engine. Because query keywords are connected using "AND", only documents containing all the keywords are returned. The table displays the top five ranking documents Lucene retrieves. For query 2 and 3, only 4 documents are found by Lucene.

the time Prism was being developed, experiments with Pauchok showed that the MultiText algorithm yields the best performance using queries generated from the questions. Furthermore, in the course of this thesis research, it turned out that the MultiText algorithm is also the best performer among all the passage retrieval algorithms implemented in Pauchok for a query generated from questions and answers. Hence, Prism uses Pauchok's implementation of the MultiText algorithm for answer projection. The details of enhancements on passage retrieval can be found in the next section.

First, Prism converts the boolean query used for document retrieval into a "bag of words" query by stripping off its boolean connectors. The passage retrieval algorithm uses the Bag-of-Word query to retrieve the best passage. Prism calculates the *idf* value of each word in the "bag of word" query. For each document returned by Lucene, Prism extracts the best passage using a modified MultiText passage retrieval algorithm. The best passages found in each document generated from the first query are displayed in Table 3.3.

3.4 Passage Selection

Passage selection is the last step of the process. After the best passage is extracted from each document, Prism chooses the best passage as the candidate passage for the query. The final output for each of the six queries is displayed in Table 3.4.

1. language AND mostly AND spoken AND Brazil **Portuguese**	
Document number	Passage
1. AP901216-0003	77.900186948623366 and removes some accents and letters that are no longer spoken But critics say the accord could muddle the tongue When it takes full effect by the beginning of 1994 some Portuguese words with very different roots and meaning will be spelled the same
2. AP890121-0027	61.894195883344573 secretary of state for foreign affairs Antonio Materrula "We have to push our mutual language at an international level " Portuguese is not an official language of the United Nations but the countries that use it Angola Brazil Cape Verde Guinea-Bissau
3. SJMN91-06119067	43.694054393939965 Monsanto Co even wants to match dialects For its European customers it has customer training guides in French Spanish and Portuguese ; Now it is tailoring those guides to match the way Spanish and Portuguese are spoken respectively in Colombia and Brazil
4. AP881107-0172	43.3185723712253964 Paiakan wearing native headdresses and collars and jungle body paint addressed their U S hosts in Portuguese Brazil's official language and Posey translated In an inexplicable move the government also indicted the Indians who were born in Brazil and never
5. FBIS3-10134	29.225912742462157 Report 2 Mar 1994 Brazil Finance Minister Gives News Briefing PY0103014094 Sao Paulo Rede Bandeirantes Television in Portuguese 1822 GMT 28 Feb 94 PY0103014094 Sao Paulo Rede Bandeirantes Television Language: Portuguese Article Type:BFN [News briefing by

Table 3.3: This table shows the all passages retrieved for the first query, which was generated using the first ranked Aranea answer. There are five documents retrieved for the first query, so five passages are generated, one from each document.

Query	Passage
1. language mostly spoken Brazil **Portuguese**	1. AP901216-0003 77.90018694862366 removes some accents and letters that are no longer spoken But critics say the accord could muddle the tongue When it takes full effect by the beginning of 1994 some Portuguese words with very different roots and meaning will be spelled the same
2. language mostly spoken Brazil **English**	2. AP890121-0027 120.31122787189483 Portugal and Sao Tome e Principe count about 180 million inhabitants Only Mandarin Chinese Hindi English Spanish and Russian are spoken by more people worldwide Brazilian and Mozambican delegates at the chemical weapons conference spoke
3. language mostly spoken Brazil **French**	3. SJMN91-06119067 145.57490318625133 manuals were in English a language that few car wash operators could read So Ryko translated its manual into French and 1991 sales took off Professional Careers DEMAND FOR TRANSLATORS IS ON THE RISE "Operating in local languages has become a
4. language mostly spoken Brazil **million people**	4. AP890121-0027 98.30959759249924 count about 180 million inhabitants Only Mandarin Chinese Hindi English Spanish and Russian are spoken by more people worldwide Brazilian and Mozambican delegates at the chemical weapons conference spoke in Portuguese further enraging Portugal's
5. language mostly spoken Brazil **official**	5. AP890121-0027 181.75082506950966 Antonio Materrula "We have to push our mutual language at an international level " Portuguese is not an official language of the United Nations but the countries that use it Angola Brazil Cape Verde Guinea-Bissau Mozambique Portugal and Sao
6. language mostly spoken Brazil	6. SJMN91-06119067 115.76214157786963 and Portuguese ; Now it is tailoring those guides to match the way Spanish and Portuguese are spoken respectively in Colombia and Brazil "We want to accommodate any culture that uses our equipment " saidd Joseph M Morris a marketing manager in

Table 3.4: The first five queries are generated from question-pairs, and the sixth one is from the default question. Prism finds the best passage for each query, and the rank of the passage is the rank of the Aranea answer. The passage retrieved using the default query is placed last. The format of the passage is as follows: the rank of the passage, the supporting document number, the passage score, and the passage text.

Chapter 4

Modification to Passage Retrieval

Finding the best passage requires two steps. First, passage extraction uses a modified version of the MultiText passage retrieval algorithm to find the best passage for each document. Passage selection is then used to choose the best one from the group of best passages. In this chapter, I briefly describe the MultiText algorithm. Then I describe six types of modifications that can be applied to both passage extraction and selection.

4.1 The MultiText Algorithm

The MultiText algorithm [3, 4] scores passages based on the passage length and the weights assigned to the query terms they match. First, the algorithm finds the best cover within a document. A cover is a chunk of text that starts and ends with a term in the query. Short covers containing many high inverse document frequency (*idf*) terms are favored. The rarer the word, the higher the *idf* score. Once the best cover is identified, it is expanded to 250 bytes of text centered around the original cover. The score of the passage depends on the cover only. The MultiText algorithm uses an *idf*-like weight rather than the actual *idf*. Prism uses Pauchok's [16] implementation of MultiText algorithm, which uses the actual *idf*.

4.2 Modification to Passage Extraction and Selection

Six types of modifications are applied to passage extraction and selection. First, incorporating Lucene's document score into the passage selection process can potentially improve its accuracy. Second, ensuring proper relative weighting on question and answer keywords can help select the most accurate and relevant passages. Third, scoring all the keywords in the passage, not just the ones in the the scoring text, can help passage extraction. Fourth, answer type analysis can help narrowing down the number of candidate passages. Fifth, naively, the closer answer keywords are to the question keywords, the more likely the passage is to be correct. SiteQ's [12] distance measure, which rewards passages with keywords close together, could potentially help Prism locate the correct passage. Lastly, if more than one passage retrieval algorithm finds the same passage, it is likely that the passage is correct. Voting uses a combination of three distinct algorithms to find the best passage.

4.2.1 Incorporating the Document Score

In the original Pauchok question-answering framework, only the passage score is used. If the document is reliable, the passage from it is more likely to be correct. Incorporating the document score into the passage score could improve the overall performance of the system. Therefore, it is incorporated with the MultiText passage score in the following fashion:

$$score = \alpha \times LuceneDocScore + MultiTextPassageScore.$$

Training is performed on α by trying a range of values and calculating the global maximum. The best value for α is found to be 66. The reason for such a high weight on Lucene is that of different scoring scales are used in Lucene and MultiText. Lucene's score ranges from 0 to 1, whereas the MultiText score is based on the sum of *idf* scores, which could be any positive real number.

4.2.2 Ensuring Proper Relative Weighting On Question And Answer Keywords

In the third baseline experiment, question and answer keywords are weighted equally. The score of each query keyword is simply its *idf* score. However, this simple scheme is not sufficient to ensure the presence of the correct Aranea answer in the passage. Sometimes, the best passage returned from the third baseline experiment does not contain the Aranea answer at all. Sometimes, the passage contains the Aranea answer but does not answer the question. Prism uses four enhancements to ensure the proper weighting of question and answer keywords.

Normalizing Query Keywords

The number of words in the question and the answer are generally not the same. If we simply assume the weight of all keywords to be equal by adding individual *idf* scores, then we are treating short questions or short answers unfairly. Therefore, a normalization scheme should be used to adjust the *idf* score of each query keyword. The following formula is used:

$$\mathit{new\_idf} = \frac{q}{a} \times \mathit{old\_idf}.$$

new_idf is the new *idf* score of an answer keyword.
q is the number of keywords in the question.
a is the number of keywords in the answer.
old_idf is the original *idf* score of an answer keyword.

If a question has more keywords than its answer, the *idf* values of answer keywords are increased proportionally. For example, if a query has 3 keywords in the question and 2 keywords in the answer, the *idf* of each answer keyword is multiplied by (3/2), the ratio between the number of question keywords and answer keywords. This adjustment ensures that the question and the answer are treated fairly.

Checking If the Passage Contains the Answer Keywords

Many passages returned by the MultiText algorithm do not contain the Aranea answer in the passage. This is due to high *idf* scores from question keywords and low *idf* scores from answer keywords. In this case, the passage returned by the MultiText passage retrieval algorithm primarily contains question keywords. However, if the Aranea answer was correct, a passage without it would certainly be incorrect. Therefore, Prism modifies the MultiText passage retrieval algorithm to reward passages that contain answer keywords.

Several strategies for rewarding these passages are described below.

1. Boost the passage score by 50% if the cover found in the MultiText algorithm contains the exact answer phrase.

2. (a) Weight the score of the expanded passage by the number of answer keywords it contains.

$$\begin{aligned} inter\_score &= (1 + step\_size \times N) \times old\_score \\ step\_size &= \frac{\delta}{ta} \end{aligned}$$

   N is the number of answer keywords in the passage.

   inter_score is the new answer *idf* score before answer rank adjustment.

   old_score is the original *idf* score.

   ta is the total number of answer keywords in the Aranea answer.

   δ is a constant, trained to be 0.1.

   (b) Weight the passage score by the answer rank to favor low-ranked answers.

$$new\_score = inter\_score(\gamma + \frac{1}{answer\_rank})$$

γ is a constant greater than 1. In the current implementation, it is trained to be 1.2.

answer_rank is the Aranea answer rank. Rank 1 is most likely to be correct.

(c) The presence of answer keywords and question keywords in a passage should be rewarded differently. The passage score is modified by multiplying a constant, β, that represents the importance of answer keywords compare to question keywords. β is trained to be 0.63.

$$new\_score = inter\_score \times \beta$$

Similarly there is also a weight for questions keywords, which is defined to be $1 - \beta$. Details on question keywords are described below.

Checking If the Passage Contains Question Keywords

Conversely, it is possible for a passage to have answer keywords but not actually answer the question. This is due to the presence of very few question keywords in the passage. This occurs when the *idf* value for an answer keywords are much higher than question keywords. For example, one of the Aranea answers for the question "What is the name of the newspaper in the Seattle area?" is "Tacoma News Tribune." Many high score passages are found to contain Tacoma News Tribune, but none of them relate to the question asked, because many newspaper articles in TREC starts with "Tacoma News Tribune reports..." Thus, it is also necessary to make sure that some of the question keywords are also present in the passage. The procedure to boost the passage score based on the presence of question keywords is similar to answer keywords described above. The difference is that answer rank is not incorporated in question keyword search.

(a) For each expanded passage, its score is modified as follows.

$$\begin{aligned} inter\_score &= (1 + stepsize \times N) \times old\_score \\ stepsize &= \frac{\delta}{tq} \end{aligned}$$

N is the number of answer keywords in the passage.

inter_score is the new passage score before the weight adjustment.

old_score is the original *idf* score of the passage.

tq is the total number of question keywords in the Aranea answer.

δ is a constant, trained to be 0.1.

(b) As described above, question keywords are rewarded differently than answer keywords. The passage score is further modified by incorporating the question keyword weight, 1-β.

$$new\_score = inter\_score \times (1 - \beta)$$

β is what is defined above, and trained to be 0.63.

4.2.3 Scoring the Entire Passage

After finding the best cover, the MultiText algorithm expands the cover from the left and the right until it reaches a certain byte limit. Sometimes after expanding the cover, the passage may contain other important query keywords, and they are not counted into the final passage score. The passage score is based on the sum of *idf* scores of query keywords in the cover, not the entire passage. The passage score is thus modified by adding the *idf* values of all keywords that are in the passage but not in the cover. This number is then scaled and added to the original passage score.

$$score = ops \times \alpha + et\_idf \times (1 - \alpha)$$

ops is the original passage score, which is the score of the cover.

et_idf is the extra keywords *idf* value, which is defined to be the sum of *idf* values of words that are in the passage but not in the cover.

Again α is trained, and the best value for α is 0.67.

4.2.4 Analyzing Answer Type

Rudimentary answer type analysis is performed on questions. Although Aranea already uses answer type analysis, Prism still needs to perform similar analysis when the Aranea answer is not present in the passage. Ideally Prism should adopt Aranea's answer type analysis; however, it is nontrivial to incorporate it into Prism. Instead, a simple answer type analysis is implemented in Prism as a proof of concept. Specifically, Prism classifies questions into three categories: date, number and proper noun. Questions starting with keywords such as "when", "what date" or "what year", etc., are classified as date questions. When passages answering date questions contain dates, their scores are boosted. Questions that start with keywords such as "How many" are put into the number category. Like before, passages answering number question have their scores increased if they contains numbers. There is also a proper noun category with person and place sub-categories, corresponding to the "who" and "where" keywords respectively. A coarse way to identify a proper noun is to check if the word is capitalized and if the *idf* score of the word is above a certain threshold. In general, proper nouns should have higher *idf* weight than common words. Capitalized words with high *idf* values are more likely to be a proper noun. If the passage is found to contain proper nouns, its score is again boosted.

4.2.5 Incorporating Distance Between Question and Answer Keywords

How close answer keywords are to question keywords in the passage can be important. One might suppose that the closer the answer is to the question, the more likely the answer is to be correct. Prism uses two techniques to incorporate the closeness measure. One uses the minimum linear distance among pairs of question and answer keywords; one incorporates SiteQ's [12] distance calculation into Prism's passage retrieval algorithm. However, SiteQ's distance calculation is between any query keywords. I modified SiteQ's algorithm such that it only measures the distance between an answer keyword and a question keyword.

$$\begin{aligned} new\_score &= distance + old\_score \\ distance &= \frac{\sum_{j=1}^{k-1} \frac{wgt(dw_j)+wgt(dw_{j+1})}{\alpha \times dist(j,j+1)^2}}{k-1} \times matched\_cnt \end{aligned}$$

$wgt(dw_j)$ is the weight of question keyword j.

$wgt(aw_{j+1})$ is the weight of answer keyword $j+1$.

$dist(j, j+1)$ is the distance between document word j and $j+1$.

$matched\_cnt$ is the number of query words matched.

α is a constant, trained to be 1.0

4.2.6 Voting Using Multiple Passage Retrieval Algorithms

Voting is a passage retrieval meta-algorithm which combines the results from a collection of passage retrieval algorithms [17]. Prism uses the simple voting scheme in Pauchok that scored each passage based on its initial rank and also based on the number of answers the other algorithms returned from the same document. More precisely, given the results from various passage retrieval algorithms, the score for each passage is calculated as follows [17]:

$$
\begin{aligned}
A &= \text{number of algorithms} \\
R &= \text{number of passages returned} \\
\mathit{docids} &= A \times R \text{ matrix of document ids} \\
&\quad \text{returned by each algorithm} \\
\mathit{docscore}(\mathit{doc}) &= \sum_{a=1}^{A} \sum_{r=1}^{R} \begin{cases} 1/r & \text{if } \mathit{docids}[a,r] = \mathit{doc} \\ 0 & \text{otherwise} \end{cases} \\
\mathit{score}(a,r) &= \frac{1}{r} + \frac{1}{2}\mathit{docscore}(\mathit{docids}[a,r])
\end{aligned}
$$

Chapter 5

Evaluation

The performance of the Prism system is measured using mean reciprocal rank(MRR), which is a well-established way to measure a question answering system. The scores of the Prism system are compared to the scores for several baseline calculations to evaluate the performance of the system.

5.1 MRR Evaluation

For each question-answer pair, a system should return a supporting document. An individual question receives a score equal to the reciprocal of the rank at which the first correct response is returned, or 0, corresponding to infinite rank, if none of the responses contains the correct answer. An answer is correct if it matches one of the regular expressions in answer patterns obtained from NIST. The mean reciprocal rank (MRR) is the average score of all the answers. Two measures of MRR are adopted. Lenient MRR only requires the answer to be correct, whereas strict MRR also requires the correct supporting document. A document supports its answer if it appears in the list provided by NIST containing correct documents for that question.

5.2 Baseline Evaluations

Three relatively simple approaches described below are used as the baselines, by using questions only, using answers only and using both questions and answers.

5.2.1 Using Question Only

Using the question as the query provides a baseline that shows how the question alone can influence the accuracy of answer projection. If the document and the passage retrieval algorithms are accurate enough, providing the question alone is enough to give us sufficient information. The MRR score for this baseline analysis is 0.341.

5.2.2 Using Answer Only

Using the answer as the query provides a baseline that shows how the answer alone can affect the accuracy of the supporting documents retrieved. Although most document retrieval and passage retrieval algorithms are used with the question as the query only, having the answer as the input could improve the performance of passage retrieval, in some cases. If the answer given was specific enough, a simple regular expression matching the passage might give the supporting documents. However, if the answer given is too short or too general, such as a date or a number, then it would be difficult to determine which is the correct supporting document, since there would be too many documents that match the answer string. This baseline experiment could shed light on when having the answer would help find supporting documents and what characteristics in the algorithms make using the answer alone a good or bad answer projection. Aranea answers are categorized using a set of features: answer length, the number of capitalized words in the answer, the number of stop-words in the answer, and the length of the longest word. Experiments show that these characteristics of the answer are not the crucial reason for variation of MRR scores across different type of questions. The difference is mainly due to how accurately Aranea can answer questions. In some cases, Prism is able to correct Aranea answers, but the performance of Prism is largely proportional to the performance of Aranea.

The MRR score for using the answer alone is 0.095. The low MRR score indicates that although the answer is important, it must be used in conjunction with the question to retrieve passages. This low MRR score is partially due to low Aranea answer accuracy.

5.2.3 Using Concatenation of Questions and Answers

One simple algorithm uses the concatenation of the question and answer as the query. This third baseline can also be compared to the results from the question alone or the answer alone as the input. It can also be used to test the performance increase from applying algorithms in the previous section. The MRR score using this scheme is 0.351, which is better than the first baseline measures. This makes sense since there is more information used to find the supporting passage.

Chapter 6

Discussion and Results

This chapter presents the end-to-end performance of the algorithms described in Chapter 4 when applied to the TREC-9 data set. Various examples are included to show how these algorithms are used to improve the performance. The first section describes the theoretical upper bound on passage retrieval. The next several sections describe the impact of applying various strategies to the basic algorithms.

Table 6.9 shows the empirical results for individual algorithms, as measured in terms of performance improvement. From this table, we can see that adjusting the relative weights of question keywords and answer keywords improves the performance the most. More specifically, the largest improvement comes from checking for the presence of a partial answer. Also, scoring the whole passage, analyzing the answer type and including document score in the passage score give a reasonable improvement in performance. On the other hand, distance measures between question keywords and answer keywords and voting do not help the overall performance. The three algorithms that improve the performance the most are checking for the presence of answer and question keywords, scoring the entire passage, and analyzing the answer type. These three algorithms alone account for 80% of the overall improvement.

	strict	lenient	increase from original Clarke
MultiText Original	0.341	0.377	0%
Adding answer	0.351	0.475	3%
Containing the entire answer	0.378	0.504	10%
Containing partial answer	0.390	0.516	14%
Containing question keywords	0.376	0.504	9%
Incorporating Answer Rank	0.382	0.505	6%
Normalization	0.368	0.497	8%
All together	0.404	0.509	18%

Table 6.1: The MRR values for TREC-9 questions after applying each strategy for ensuring the proper relative weighting of query terms. Each strategy is built on top of the third baseline. The last row shows the final MRR score after incorporating all the strategies in Section 4.2.

6.1 Theoretical Upper Bound on MRR Scores For Passage Retrieval

Because the document retrieval algorithms do not find relevant documents for every question, there is an upper bound on their performance. Lucene finds documents for 380 questions out of 500. Out of 380 questions, 265 have the correct relevant document. Assuming these documents are all in the first rank, the maximum MRR score that passage retrieval can achieve is 265/380 = 0.69.

6.2 The Most Effective Approach: Ensuring Proper Relative Weighting On Query Keywords

Weighting question keywords properly with respect to answer keywords is a hard balance to strike. The following strategies are applied to achieve the proper weighting. Results and examples are reported below. Applying all the strategies together boosts the MRR score to 0.404, a 18% increase from the the first baseline experiment and 15% from the third baseline experiment. Table 6.1 summarizes the performance improvements of applying these algorithms.

Question	What continent is Bolivia on?
Aranea Answer	South American
Query	continent Bolivia **South American**
rewarding exact answer	FT944-10662 4.994507662448498 six rainy months each year Now that Bolivia's development plans hinge on becoming a strategic hub for the *South American* **continent** transport is a high priority The ministry's ambitious brief is to build 2 700km of roads within the next five years
ignoring exact answer	FT944-10662 4.9790795146907065 thousands of tones of soya beans from the Mato Grosso in south-west Brazil to cross the **continent** through **Bolivia** for eventual shipment to Asia through Peruvian or Chilean Pacific seaports Also arousing the enthusiasm of top-level politicians and FT944-10662 3.324507662448498 six rainy months each year Now that Bolivia's development plans hinge on becoming a strategic hub for the **South American continent** transport is a high priority The ministry's ambitious brief is to build 2 700km of roads within the next five years

Table 6.2: The Aranea answer is the correct answer for the question. Passage returned by Prism is in the following format: document ID, passage score, passage. The Aranea answer is in italic and query keywords are in boldface.

6.2.1 Checking If the Passage Contains the Answer Keywords

Ensuring that the Aranea answer is present in the passage is important, especially when it is correct. If the Aranea answer is wrong, it is still likely to be related to the actual answer.

Containing the Exact Answer

Passages containing the exact Aranea answer phrase are likely to be correct. Prism rewards passages containing the exact Aranea answer handsomely. The motivation for this is that the exact Aranea answer should not appear along with the question keywords by chance.

Table 6.2 provides an example where having the exact Aranea answer helps to select the correct passage. If the presence of the exact Aranea answer is not rewarded, then the wrong passage is selected, since the correct passage has a lower passage score. In this case, the cover containing the question keyword "Bolivia" has a higher *idf* score than the sum of *idf* scores for the answer keywords, "South" and "American." However, if the presence of the exact Aranea answer is rewarded by boosting the score of the passage by 50%, the correct passage becomes first ranked.

Containing a Partial Answer

If the entire Aranea answer is not included in the passage, having a part of the answer can also help in finding the correct passage. In many cases, Aranea answers are more specific than the question requires. Sometimes, an Aranea answer that is incorrect is fairly close to the required answer. Under these situations, if the passage contains some answer keywords, the required answer may have been embedded in the passage. Also, in some cases, answer projection can actually correct the Aranea answer. As Table 6.3 shows, the Aranea answer is July 1, 1981, but 1981 alone would have been a correct answer. By finding a part of the answer in the passage, Prism is able to find the required answer. However, if partial answers are not rewarded, the first-ranked passage does not contain the answer at all. Again, this is due to higher *idf* scores for question keywords than answer keywords.

Not only does this technique help in finding the correct passage, it also corrects the Aranea answer. As illustrated in Table 6.3, the correct answer is actually July 21, 1981, not the Aranea answer, July 1, 1981. By checking for partial answers in the passage, Prism is able to find that July and 1981 are inside the passage and boost the passage score. It turns out that the passage found the actual correct answer July 21, 1981, even though the Aranea answer was slightly different. This algorithm is the most effective algorithm in this category, improving the MRR score to 0.390, a 14% increase from the first baseline experiment, and a 11% from the third baseline experiment.

Question	When did Princess Diana and Prince Charles get married?
Correct Answer	1981
Aranea Answer	July 1 1981
Query	Princess Diana Prince Charles married **July 1 1981**.
with partial answer	LA053190-0005 10.979235857673952 Prince Charles and Princess Diana finally received a wedding gift nine years after they were married when they opened the Prince and Princess of Wales Hospice in Glasgow Scotland The royal couple married **July 21 1981** on Tuesday chatted with patients
without partial answer	AP900529-0130 19.391621203035832 them GLASGOW Scotland (AP) There's nothing fashionable about being this late but no one seemed offended Prince Charles and Princess Diana finally received a wedding gift from Glasgow nearly nine years after they were married The present: naming a

Table 6.3: The format of the passage returned by Prism is in the following format: document ID, passage score, passage. The correct answer is in boldface.

6.2.2 Checking If the Passage Contains Question Keywords

It is possible that a passage contains the answer, but does not answer the question. Sometimes, the selected passage may contain the Aranea answer, which could be incorrect. In these situations, it is important to check for not only the answer keywords, but the presence of question keywords as well. This scheme is used primarily to balance the previous answer-checking algorithms, to avoid situations where answer key terms are heavily rewarded, especially if they are wrong. Using this scheme, the MRR score increased to 0.376, a 9% increase from the first baseline and 6% increase from the third baseline. Using this strategy in conjunction with checking for partial answer, the MRR score increased to 0.394, a 16% increase from the first baseline and a 13% increase from the third baseline.

6.2.3 Incorporating Answer Rank

The rank of the Aranea answer determines how likely the Aranea answer is to be correct. A lower answer ranking indicates a more likely accurate answer. Therefore, answer rank is an important factor in determining the relative weights of question and answer keywords. The relative weight for low ranking Aranea answers should be higher than high ranking Aranea answers. After incorporating answer rank into the passage score, the MRR score improved to 0.362, a 6% increase from the first baseline experiment and a 3% increase from the third baseline experiment.

6.2.4 Normalizing Query Keywords

Normalizing the query keywords improves the MRR score by 6% compared to the first baseline. However, this is less effective than checking for question and answer keywords. In fact, normalizing query keywords after checking for both question and answer keywords leads to a negligible improvement. This makes sense since these algorithms are fundamentally very similar.

6.3 Other Effective Approaches

Expanding the cover and analyzing the answer type improves the performance of the overall score by 7% and 6% respectively. The performance analysis is detailed below. Results are shown in Table 6.9.

6.3.1 Scoring the Entire Passage

The MultiText algorithm uses a cover to score a passage. A cover string is an excerpt of text that starts and ends with query terms. Its length can vary from a word to a paragraph. The cover is expanded or contracted to a certain prescribed length, and the final result is the passage returned from the MultiText. Most of the time, a small cover is chosen, which is then expanded to a paragraph. In such cases, the expanded cover may contain keywords that are not considered for the passage score calculation.

As illustrated in Table 6.4, the cover found for the passage is simply "Ronald Reagan." But when scoring the entire passage, many keywords outside of the cover are found. These keywords are instrumental in determining the relevance and accuracy of the passage. The passage obtained by scoring the cover alone has one more keyword

Question	Who was the oldest U.S. president
Aranea Answer	rank 1 Ronald Reagan
Query	oldest U.S. president *Ronald Reagan*
with expanding cover	AP890611-0018 6.1042909026145935 and only a dozen were older at the end of their terms than Bush is now Forty men have been *president* The *oldest* on Inauguration Day were: **Ronald Reagan** 69; William Henry Harrison 68 James Buchanan 65 and Bush 64 The *oldest U S presidents* ever were:
without expanding cover	LA031290-0004 4.319760770411017 gum which they consumed at a prodigious rate In their 103rd year Guinness proclaimed them the world's oldest twins **President Ronald Reagan** telegraphed his greetings a Japanese TV crew showed up at the nursing home where they lived and the National

Table 6.4: The passage returned by Prism is in the following format: document ID, passage score, passage. The cover is in boldface, and extra keywords outside of the cover are in italic.

in the cover but many fewer keywords in the passage. Therefore, by scoring the entire passage, the score of the correct passage is increased above all others. The MRR score increased to 0.364 after scoring the entire passage, a 7% increase from the first baseline experiment and a 4% increase from the third baseline experiment.

6.3.2 Analyzing Answer Type

Answer type analysis makes it more likely for passages containing the correct answer to have high scores. The answer type analysis in Prism is rather simple, and hence, the expected performance increase is not as significant as expected. But with a better classification of questions and identification of answers, perhaps by incorporating Aranea's question and answer type analysis into Prism, the performance should increase significantly. Table 6.5 shows the performance of 3 different types of questions before and after answer type analysis.

Number Question Analysis

Question and answer type analysis improves the performance over the third baseline experiments, where queries are generated from questions and answers; however, this

Question Type	Baseline 1		Baseline 3		Type Analysis	
	strict	lenient	strict	lenient	strict	lenient
Number	0.278	0.278	0.218	0.266	0.220	0.266
Proper noun	0.423	0.442	0.479	0.657	0.483	0.650
When	0.135	1.171	0.250	0.381	0.337	0.456

Table 6.5: QA analysis by categories. "When" questions are questions that require a specific time or date. "Proper noun" questions are questions that starts with "where" or "who". "Number" questions usually start with keywords such as "how many", and require the answer to be a quantity.

performance is much worse than the first baseline where queries are generated using only the question. The main reason for such a poor result is that the Aranea answers are either incorrect or in the wrong format.

Table 6.6 presents an example where answer type analysis actually does worse than the first baseline, which uses the original MultiText algorithm. In the example, none of the Aranea answers are correct. Prism is therefore unable to find the relevant document using question-answer queries. At best, if the correct passage can be found using the default question query, the passage would be at rank 6, the last rank. In the first baseline experiment, only the question query is used, and six passages (not just one) are returned. As shown in the example, the third passage is the correct one in the first baseline experiment. In the third baseline experiment, the correct passage at the third rank was not found, because only the best passage for each query is returned. However, after boosting the passage score using answer type analysis, Prism is able to find the correct passage using the default query. This shows the need for researching how to rank passages that come from different queries. If all the Aranea answers are fairly unreliable, then perhaps answers coming from the default query should be placed first. If Prism can somehow know when it should use the default query first, then the passage found in the example could have been ranked first.

Out of 28 number type questions, 4 questions perform worse while 5 questions perform better compared to the third baseline result. Although 2 out of those 4 questions have the correct answer, they have the wrong format. For example, one question asks for the number of continents in the world. The answer in the document

Question	How many miles is it from London, England to Plymouth, England?
Aranea answers	30 miles 10 miles 15 miles 50 miles 22 miles
Correct Answer	200 miles
With Answer Type Analysis	AP890114-0067 4.451597669124604 Typhoon was expected to reach the Yarrowanga on Sunday and attempt to salvage it The Royal Air Force Rescue Center in Plymouth said the ship had lost 120 square yards of plating below the waterline on the right side and about 200 meters of plating LA061090-0021 5.944230185627816 parade will be led by a replica of an 18th-Century East Indiaman A tall ships race meanwhile starts in Plymouth England and will go via Le Coruna Spain; Bordeaux France and Zeebrugge the Netherlands to Amsterdam Visitors will be allowed on board LA061889-0051 4.451597669124604 " Massachusetts – Mrs Evelyn C Weber Los Alamitos: Enjoyed the Hawthorne Hill Bed amp; Breakfast 3 Wood St Plymouth Mass 02360 Rates: $40 single/$50 double Montana – June and Larry Pierce Desert Hot Springs: "Happy Landen Trailer Park P O AP880918-0002 3.679006338119507 attribution that the convoy was transporting four nuclear depth bombs to the Royal Navy's armament depot in Plymouth on the south coast Depth bombs are designed for use aircraft to attack submarines Colin Hines a spokesman for the environmentalist FT934-11521 4.451597669124604 this vale packed with history from the wild country of Exmoor The M5 Somerset's main artery goes south to Exeter Plymouth and France and north to the rest of England and its motorway network Caravans bound for Devon and Cornwall crowd it in *AP900516-0223 11.854143158164826* *late Mondayand early today in London the Welsh city of Cardiff and the southwestern town of Plymouth about **200 miles** from London No group claimed responsibilty in any of the incidents but the domestic news agency Press Association noted that*
Baseline 1 experiment (Using Question Only)	LA061090-0021 13.146522503971934 parade will be led by a replica of an 18th-Century East Indiaman A tall ships race meanwhile starts in Plymouth England and will go via Le Coruna Spain; Bordeaux France and Zeebrugge the Netherlands to Amsterdam Visitors will be allowed on board AP900516-0223 11.854143158164826 to close several popular tourist beaches Oil washed up on 12 miles of beaches and coves forming Bigbury Bay east of Plymouth in far western England Authorities closed beaches to permit an unhindered cleanup "We were fighting a battle against *AP881220-0059 11.800750436361918* *late Mondayand early today in London the Welsh city of Cardiff and the southwestern town of Plymouth about **200 miles** from London No group claimed responsibility in any of the incidents but the domestic news agency Press Association noted that* AP880720-0075 11.731439228293945 approach that he insisted on finishing a game of bowls on a lawn at the southwestern port city of Plymouth 80 miles from Kynance before sailing to battle As crowds watched a team of actors in Elizabeth costumes enacted Drake's game at Plymouth AP891229-0034 11.209296266793963 and left by the side of the road near Torquay 150 miles southwest of London Police on Thursday returned the taxi found 25 miles away in Plymouth "Finding the money was surprising " said a police spokesman "It must be the Christmas spirit " AP880829-0180 9.263881212226414 review the situation The Karin B was spotted Monday morning anchored eight miles off Eddystone lighthouse about 20 miles south of Plymouth said a spokesman for the Department of Environment who by custom was not identified "As yet it hasn't made

Table 6.6: The format of the passage returned by Prism is in the following format document ID, passage score, passage. The correct passage is italic and the correct answer is also in bold font.

is "seven", but the Aranea answer is "7." Another problem is that Prism eliminates punctuation within a word. For example, 500,000 is converted to 500 000. The Aranea answer is correct, but because of formatting issues, Prism is not able to recognize that "500,000" is the same as "500 000". These are the major drawbacks that make incorporating the Aranea answer problematic.

Time and Date Question Analysis

Answer type analysis on time questions improves the performance significantly: it increases the MRR score from the first baseline of 0.135 to 0.337, a 150% increase. The enhancement from the third baseline is 35%, corresponding to a baseline score of 0.250. Three reasons contribute to this big improvement: Aranea answers for time questions are more accurate; type checking on the presence of the answer type in a passage improves the accuracy especially if the *idf* values of question keywords are high; type checking sometimes corrects the wrong Aranea answer.

Having more accurate Aranea answers certainly helps in retrieving the correct passages, since answer type analysis favors passages containing the correct answer by boosting passages that contain the correct answer type. Therefore, a part of this huge increase in MRR can be reproduced by favoring passages containing the Aranea answer.

However, as explained earlier, passages without the Aranea answer can still be chosen because of higher *idf* scores of question keywords . In cases where the passages do not contain the Aranea answer, time question analysis is performed. Table 6.7 shows an example where time analysis improves the performance of the system. In the example, the question keywords are far away from the answer keywords, and the *idf* scores of the question keywords are much higher than those of the answer keywords. Even if Prism favors passages containing the Aranea answer, the paragraph containing the question keywords would still be chosen due to their higher *idf* values. The Aranea answer in this example is correct, so obviously any passage without it is incorrect. In this case, answer type analysis helps ensure that the passage contains the correct answer.

Question	When did Chernobyl nuclear accident occur?
Aranea Answer	26 April 1986
Correct Answer	1986
Query	Chernobyl nuclear accident occur **26 April 1986**
without time/date analysis (the third baseline)	
The Chosen Passage	FBIS3-60326 5.184221270557177 any connection Regardless–the national assembly Verhovna Rada certainly stands by what it said when it states that the *Chernobyl accident* ”will be seen as the worst tragedy in 20th century Ukrainian history and will affect the lives of several
The Passage Containing The Correct Answer	Six and one-half years after the explosion and the fire at the nuclear power plant–which began on **26 April 1986**–the effects are being observed on people, animals, and plants.
with time/date analysis	
Chosen Passage	LA010189-0001 7.465278629602334 (5.184221270557177) and has allowed substantial new Western insights into Soviet society David R Marples' new book his second on the *Chernobyl accident* of **April 26 1986** is a shining example of the best type of non-Soviet analysis into topics that only recently

Table 6.7: The format of the passage returned by Prism is in the following format document ID, passage score, passage. For the second chosen passage, its original passage score from the third baseline experiment is included in parenthesis.

With answer type checking, the correct passage is found using the same query. The two chosen passages in Table 6.7 have the same passage score in the third baseline experiment. Just by chance, the incorrect passage is chosen because its document number comes first alphabetically. However, answer type analysis can be used to boost the score of the correct passage to ensure that it is chosen.

Dates, times and numbers all suffer from formatting problems. As illustrated in Table 6.7, the format of the Aranea answer is different from the format in the correct passage. However, in this case, answer type analysis actually helps to correct the formatting problem. It recognizes that April 26 1986 is a date object. Although the format is not the same as the Aranea answer, it nonetheless finds the correct passage.

Proper Noun Analysis

Proper noun recognition does not improve the performance significantly. Only 3 out of 149 questions benefited from this strategy. It is not clear that these questions found the correct answer due to the proper noun boost. The techniques for recognizing proper nouns in Prism are very primitive. Better proper noun recognition such as using WordNet should help performance.

6.3.3 Incorporating the Document Score

It is possible to use the document retriever only to retrieve a list of documents, and use the score returned from the passage retrieval algorithm to rank passages. The motivation to use only the passage score might be that passage retrieval algorithms were studied in much detail in TREC, and the passage scoring scheme should be more reliable than Lucene, a free open-source software. However, as illustrated in Table 6.8, a document score indeed helps in retrieving the best passage by combining the document score to the passage score.

The first passage displayed in Table 6.8 has a document score of 1.0 and a MultiText passage score of 4.683. The second passage has a document score of 0.88, and a MultiText passage score of 5.146. Clearly, the second passage has a higher passage

Question	What is the most common cancer?
Aranea answer	skin
Query	most common cancer **skin**
with using document score	AP900419-0049 23.683565679069293 cause is less certain but is believed to be partly intermittent harsh sun exposure particularly during adolescence Koh said **Skin** cancer is the most common form of cancer in the United States with the number of cases of malignant melanoma doubling
without using document score	FT933-10912 5.146421477121814 quarter between 1980 and 1987 the most recent year for which statistics are available Sufferers of the most common skin cancer need an operation to remove the diseased area Half of those who develop the more rare malignant melanoma die from it

Table 6.8: The format of the passage returned by Prism is in the following format: document ID, passage score, passage. The Aranea answer for the question is the correct answer, and is shown in boldface.

score but a lower document score. If one were to rely on passage score alone, the second passage in the table, which is incorrect, would have been chosen. But by using the document score in conjunction with the passage score, the first and the correct passage is selected.

However, there is a trade-off. Sometimes, the document score is higher for wrong passages. But experiments show that combining the document score with the passage score in the right proportion will increase the MRR score. Indeed, the MRR score increases to 0.475, 6% increase from the first baseline, and 3% increase from the third baseline.

6.4 Ineffective Approaches

Two algorithms, voting and distance measure between question terms and answer terms, do not improve the performance of the system. The possible reasons for their failure are described below.

6.4.1 Incorporating Distance Between Question and Answer Keywords

Incorporating the minimum linear distance among pairs of question and answer keywords did not improve the system performance. This is mainly due to the fact that passages are sufficiently short such that proximity is apparently not important. Incorporating the SiteQ [12] distance score does not help find more accurate passages either, because the MRR increase from the third baseline is insignificant. This is probably due to the fact that both the SiteQ and the MultiText algorithms are density-based scoring systems, and fundamentally do the same thing.

6.4.2 Voting Using Multiple Passage Retrieval Algorithms

Unfortunately, Prism only supports three passage retrieval algorithms for answer projection. They have been found to be insufficient to significantly improve passage retrieval performance compared to just using the MultiText algorithm. In the future, Prism will be able to implement more passage retrieval algorithms for answer projection. With more algorithms, voting may be able to improve the performance.

6.5 Error Analysis for Missed Questions

The strict MRR score presented is lower than the actual strict MRR score. This is because the relevant document list provided in TREC is incomplete. This leads to much incorrect scoring for correct passages, when their supporting documents are not recognized. This is illustrated by looking at the difference between the lenient and the strict MRR for the first baseline experiment, where only the question is used to find passages. In this case, the strict MRR and lenient MRR should be very similar to each other if not exactly the same. However as Table 6.9 illustrates, the lenient MRR is 10% higher than the strict MRR. So the actual improvements may be higher than what is reported here, but although only up to the percentage increase of lenient MRR scores, which is 43%.

	strict	lenient	increase from original MultiText
MultiText Original	0.341	0.377	0%
Adding answer	0.351	0.475	3%
Adding Doc score	0.363	0.480	6%
Answer Type Analysis	0.361	0.483	6%
Weighting answer question terms	0.394	0.516	16%
Expanding cover	0.364	0.490	7%
Distance	0.375	0.506	4%
All together	0.427	0.550	25%

Table 6.9: The final MRR values for TREC-9 questions that Aranea tries to answer. This includes questions that Aranea could not answer. Each row shows the new MRR score from applying each strategy described in algorithm section. The last row shows the final MRR score after incorporating all the strategies. Strict MRR is calculated from a correct answer and supporting document pair, whereas lenient MRR ignores the supporting document.

6.6 Overall Results

Table 6.9 reports the results of applying these algorithms to all questions in TREC-9 excluding definition questions, and Table 6.10 reports the results from TREC-9 questions which Aranea has answers to. The performance improvements of the original Pauchok is 25% higher. Prism is able to improve Aranea as well. As shown in the table, the lenient MRR score is 0.55, 10% higher than Aranea's lenient score, which is 0.50.

Results show that using one algorithm from each category can account for most of the MRR increase. Containing answer keywords, expanding the cover and answer type analysis accounts for 80% of the increase in MRR.

	strict	lenient	increase from original MultiText
MultiText Original	0.352	0.388	0%
Adding answer	0.365	0.487	3%
Adding Doc score	0.369	0.478	5%
Answer analysis	0.376	0.558	7%
Weighting question and answer	0.410	0.540	16%
Expanding cover	0.416	0.529	7%
Distance	0.366	0.506	4%
All together	0.437	0.558	24%

Table 6.10: The final value MRR values for TREC-9 questions that Aranea has answers to. Each row shows the new MRR score from applying each strategy described in algorithm section. The last row shows the final MRR score after incorporating all the strategies. Strict MRR is calculated from a correct answer and supporting document pair the passage and the supporting document is correct, whereas lenient MRR ignores the supporting document.

Chapter 7

Future Work

Prism is the first step in building a good answer projection system. There is much more work to be done, including exploration of several new ideas. Below I outline some future directions.

7.1 Combining Passages From Different Queries

Currently, each question generates multiple queries. For each one, passages are chosen, and the best one is selected. As a result, each question generates multiple passages—one from each query. Ranking these passages can be a challenge. Simply comparing them using the raw passage scores is inappropriate, since the passage scores are calculated using different queries. Currently, Prism ranks the passage by the rank of the Aranea answer. However, a better system for performing this task should be devised and implemented.

7.2 Generating More Queries

Aranea answers are sometimes too specific, such that the documents retrieved are not relevant to the question asked. This happens because the Lucene document retriever needs all the keywords in the query to be present in the document. If the Aranea answer is more specific than the answer in the document, the correct document is not selected because it lacks terms in the Aranea answer. In the future, query generation could be modified to include partial answer queries. Instead of just one query per Aranea answer, a variable number of queries could be generated, depending on the number of important keywords in the Aranea answer.

7.3 Exploring A Better Answer Type Analysis Module

The current answer type analysis used in Prism is very rudimentary. It only recognizes three categories of answers: times and dates, numbers, and proper nouns. As shown in Chapter 6, good answer type analysis can help the performance significantly. Therefore, it is worth the effort to explore a better answer type analysis module. Aranea already has a good answer type system. The first step in realizing this goal is perhaps incorporate Aranea's techniques into Prism.

7.4 Exploring Different Passage Retrieval Algorithms For Answer Projection

At the time Prism was first implemented, the MultiText was the best-performing algorithm in Pauchok. Since then, more passage retrieval algorithms have been incorporated into its general framework. Tellex *et al.* [17] has shown that the best passage retrieval algorithms employ density-based measures for scoring query term. Although the MultiText algorithm uses a density-based scoring, it does not perform as well as IBM [8, 7] and ISI [6] passage retrieval algorithms. In the future, these algorithms could be modified for use in answer projection, much like the MultiText, using techniques outlined in this thesis. As more passage retrieval algorithms are adopted for answer projection, voting could potentially help the performance by rewarding cases where multiple passage retrieval algorithms point to the same passage.

Chapter 8

Contributions

Answer projection is starting to receive more attention from the IR community because of its potential to combine the best of both worlds: finding some answer within a massive and redundant data source and then verifying it in a small and more reliable data source. The goal of Prism is to design an answer projection system that can answer as many questions as the Web and yet retain the accuracy of traditional systems. It combines and extends existing document and passage retrieval algorithms, and includes new algorithms specific to answer projection. By combining Prism with traditional and Web-based systems, we obtain a question answering system that improves the performance of a traditional question answering system by 25%. It improves the accuracy of answers retrieved from the World Wide Web using Aranea by 10% and more importantly, it verifies answers retrieved from the Web by providing reliable supporting documents. Prism improves upon existing question answering systems by providing answers with high accuracy and broad coverage.

Bibliography

[1] Lucene. http://jakarta.apache.org/lucene/docs/index.html.

[2] Eric Brill, Susan Dumais, and Michele Banko. An analysis of the AskMSR question-answering system. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, 2002.

[3] Charles Clarke, Gordon Cormack, Derek Kisman, and Thomas Lynam. Question answering by passage selection (Multitext experiments for TREC-9). In *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*, 2000.

[4] Charles Clarke, Gordon Cormack, and Thomas Lynam. Exploiting redundancy in question answering. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-01)*, 2001.

[5] S. E. Robertson et al. Okapi at TREC-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, 1995.

[6] Eduard Hovy, Ulf Hermjakob, and Chin-Yew Lin. The use of external knowledge in factoid QA. In *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*, 2001.

[7] Abraham Ittycheriah, Martin Franz, and Salim Roukos. IBM's statistical question answering system—TREC-10. In *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*, 2001.

[8] Abraham Ittycheriah, Martin Franz, Wei-Jing Zhu, and Adwait Ratnaparkhi. IBM's statistical question answering system. In *Proceedings of the 8th Text REtrieval Conference (TREC-9)*, 2000.

[9] Boris Katz. Annotating the World Wide Web using natural language. In *Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet (RIAO '97)*, 1997.

[10] Boris Katz, Sue Felshin, Deniz Yuret, Ali Ibrahim, Jimm y Lin, Gregory Marton, Alton Jerome McFarland, and Baris Temelkuran. Omnibase: Uniform access to heterogeneous data for question answerin g. In *Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems (NLDB 2002)*, 2002.

[11] Boris Katz, Jimmy Lin, and Sue Felshin. The START multimedia information system: Current technology and future directions. In *Proceedings of the International Workshop on Multimedia Information Systems (MIS 2002)*, 2002.

[12] Gary Geunbae Lee, Jungyun Seo, Seungwoo Lee, Hanmin Jung, Bong-Hyun Cho, Changki Lee, Byung-Kwan Kwak, Jeongwon Cha, Dongseok Kim, JooHui An, Harksoo Kim, and Kyungsun Kim. SiteQ: Engineering high performance QA system using lexico-semantic pattern matching and shallow NLP. In *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*, 2001.

[13] Jimmy Lin, Aaron Fernandes, Boris Katz, Gregory Marton, and Stefanie Tellex. Extracting answers from the Web using knowledge annotation and knowledge mining techniques. In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, 2002.

[14] Jimmy Lin, Dennis Quan, Vineet Sinha, Karun Bakshi, David Huynh, Boris Katz, and David R. Karger. The role of context in question answering systems. In *Proceedings of the 2003 Conference on Human Factors in Computing Systems (CHI 2003)*, 2003.

[15] Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Mike Gatford, and A. Payne. Okapi at TREC-4. In *Proceedings of the 4th Text REtrieval Conference (TREC-4)*, 1995.

[16] Stefanie Tellex. Pauchok: A modular framework for question answering. Master of engineering thesis, MIT AI Lab, 2003.

[17] Stefanie Tellex, Boris Katz, Jimmy Lin, Gregory Marton, and Aaron 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 (SIGIR-2003)*, 2003.

[18] Ellen M. Voorhees. Overview of the TREC 2001 question answering track. In *Proceedings of the 2001 Text REtrieval Conference (TREC 2001)*, 2001.

[19] Ellen M. Voorhees and Dawn M. Tice. Overview of the TREC-9 question answering track. In *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*, 2000.