Computational Linguistics and Chinese Language Processing
Vol. 10, No. 2, June 2005, pp. 251-276

Design and Development of a Bilingual Reading Comprehension Corpus

Kui Xu* **and Helen Meng***

Abstract

This paper describes our initial attempt to design and develop a bilingual reading comprehension corpus (BRCC). RC is a task that conventionally evaluates the reading ability of an individual. An RC system can automatically analyze a passage of natural language text and generate an answer for each question based on information in the passage. The RC task can be used to drive advancements of natural language processing (NLP) technologies imparted in automatic RC systems. Furthermore, an RC system presents a novel paradigm of information search, when compared to the predominant paradigm of text retrieval in search engines on the Web. Previous works on automatic RC typically involved English-only language learning materials (Remedia and CBC4Kids) designed for children/students, which included stories, human-authored questions, and answer keys. These corpora are important for supporting empirical evaluation of RC performance. In the present work, we attempted to utilize RC as a driver for NLP techniques in both English and Chinese. We sought parallel English, and Chinese learning materials and incorporated annotations deemed relevant to the RC task. We measured the comparative levels of difficulty among the three corpora by means of the baseline bag-of-words (BOW) approach. Our results show that the BOW approach achieves better RC performance in BRCC (67%) when compared to Remedia (29%) and CBC4Kids (63%). This reveals that BRCC has the highest degree of word overlap between questions and passages among the three corpora, which artificially simplifies the RC task. This result suggests that additional effort should be devoted to authoring questions with a various grades of difficulty in order for BRCC to better support RC research across the English and Chinese languages.

Keywords: bilingual, reading comprehension, corpus.

* Human-Computer Communications Laboratory, Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong SAR, China
E-Mail: {kxu, hmmeng}@se.cuhk.edu.hk

[Received February 2, 2005; Revised May 16, 2005; Accepted May 16, 2005]

1. Introduction

RC is a task that conventionally evaluates the reading ability of an individual, especially during language learning. Typically, the subject is presented with a passage of natural language text and asked to read and comprehend the passage. He or she is then presented with a series of questions about the passage and asked to answer each question based on information understood from the passage.

Recently, research efforts have been devoted to the development of *automatic* RC systems [Anand *et al*. 2000; Charniak *et al*. 2000; Hirschman *et al.* 1999; Ng *et al.* 2000; Riloff and Thelen 2000]. An RC system can automatically analyze a passage of natural language text. When the system is then presented with a series of (human-generated) questions, it is expected to automatically generate an answer for each question, based on information it extracted or retrieved from the passage. While the conventional RC task can evaluate the reading capability of a human, the task can also be used to drive advancements in natural language processing (NLP) technologies incorporated into automatic RC systems. Furthermore, an RC system presents a novel paradigm of information search, when compared to the predominant paradigm of text retrieval in search engines on the Web. Several comparisons are made below:

1. An RC system has only a *limited amount* of direct context upon which to draw in order to answer questions, whereas a Web-based search engine has *huge amounts* of direct context available on the Internet.
2. An RC system aims to generate *specific, precise answers* to user-posed questions based on the given passage, thus eliminating the need for the user to read the entire passage, whereas a Web-based search engine presents a list of text documents that closely match the user's query so that the user to *further browse for potential answers*.
3. A cross-language RC system is analogous to a cross-language text retrieval system, in that a user-posed question/query may be expressed in a different language from that used for passages or archived documents.

The first two points suggest that in-depth syntactic and semantic analyses are needed to facilitate the automatic RC task. Hence, the RC task has been a driver of NLP development.

Previous works on automatic RC, as cited above, typically involved *language learning materials* designed for children/students, which included stories, human-authored questions, and answer keys. These materials included the Remedia corpus [Hirschman *et al.* 1999] and the CBC4Kids corpus [Anand *et al*. 2000; Dalmas *et al.* 2003]. These corpora are important because they support empirical evaluation of RC performance. An analytical review of the passages, questions, and answer keys in the Remedia and CBC4Kids training sets reveals that a suite of natural language processing and information extraction technologies are

indispensable for achieving comprehension. Hence, an RC corpus needs to be annotated to support the development of such technologies, as explained below:

1. If questions and answers have equivalent meanings in term of the same frame structures, such as the same predicates, logical subjects, and logical objects, then RC systems should be able to identify the same frame structures.
2. Inference may be based on a common ontology (e.g., the synonymy/antinomy, is-a, part-of, causality, and entailment relations in WordNet), human feelings (e.g., what is strange, happy, sad, etc.) or semantically equivalent descriptions (e.g., "*the man is a farmer*" means "*the man makes a living by farming*").
3. Inference may be based on the context knowledge in a passage. For example, the two sentences "*A merry-go-round has wooden animals on it*" and "*The weather damages the animals*" imply that "*The weather damages the merry-go-round.*" An RC system should be able to find inferences by using context knowledge to identify equivalent meanings.
4. An RC system should be able to perform summarization to answer such questions as "*What can we draw from this story?*"
5. An RC system should be able to perform calculations in order to answer such questions as "*How many boroughs are there in New York City?*"
6. An RC system should be able to resolve anaphora in documents in order to identify equivalent meanings.
7. When questions ask for specific persons, times, locations, or numbers, an RC system should be able to identify different named entity types in order to identify equivalent meanings.
8. For definition questions, the answers follow some patterns. An RC system should be able to perform pattern matching to identify equivalent meanings. An example question-answer pair is "*Who is Christopher Robin?*" and "*He is the same person that you read about in the book, Winnie the Pooh.*"

In addition, previous corpora involved *English-only language learning materials*. This paper describes our initial attempt to design and develop a bilingual corpus for reading comprehension (RC). In the current work, we attempted to utilize RC as a driver for NLP techniques in both English and Chinese. As an initial step, we sought parallel English and Chinese learning materials for language learning and incorporated annotations deemed relevant to the RC task. We refer to this corpus as the bilingual reading comprehension corpus (BRCC).

The rest of this paper is organized as follows. Section 2 reviews related works.

Section 3 presents considerations for designing a bilingual corpus. Section 4 describes the development of the BRCC. Section 5 discusses BOW matching results on the BRCC, and section 6 draws conclusions.

2. Related Work

Remedia is the first corpus developed for the evaluation of automated RC systems [Hirschman *et al*. 1999]. This corpus consists of remedial reading materials for grades 3 to 6 and was annotated by the MITRE Corporation [Hirschman *et al*. 1999]. An example passage with questions and answer keys is shown in Table 1. An answer key is the answer to a given question as provided by the publisher. It may not be an extract from the passage itself. In each story, as exemplified in Table 1, the first line is the title; the second line is the dateline; the others are story sentences. This corpus was used as a test-bed in 2000 in an ANLP-NAACL workshop on "Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems" [Charniak *et al.* 2000; Hirschman *et al.* 1999; Ng *et al.* 2000; Riloff and Thelen 2000].

Table 1. Sample story and questions in the Remedia corpus

Passage	Pony Express Makes Final Run (JOPLIN, MISSOURI, October 26, 1861) From now on, mail will be sent a new, faster way. It is called the telegraph. It uses wires to send messages. Now there will be no need for the Pony Express. Since April, 1860, mail has been sent this way. The Last Pony Express rider leaves town today.…
Questions	Who left Joplin on October 26, 1861? What did the Pony Express riders do? When did the Pony Express start?
Answer keys	the last Pony Express rider they carried mail April, 1860

To evaluate RC systems on Remedia, three evaluation metrics were proposed in [Hirschman *et al.* 1999], namely, P&R, HumSent, and AutSent. P&R directly uses answer keys to compute precision and recall, while both HumSent and AutSent use answer sentences which are obtained according to answer keys. The difference between HumSent and AutSent is that answer sentences in HumSent are marked by humans, while those in AutSent are generated by an automated routine based on which sentence has the highest recall compared with the answer key [Hirschman *et al.* 1999]. HumSent accuracy is calculated by comparing the system answers with the human-marked answers, scoring one point to an answer from the system that is identical to the human marked answer and zero points otherwise. The average score across all the questions is the HumSent accuracy. Riloff and Thelen [2000] believed that

HumSent was more reliable than P&R and AutSent, since human-marked answer sentences were involved.

In 2000, the CBC4Kids corpus was developed based on the Canadian Broadcasting Corporation web page[1] for kids [Anand *et al.* 2000; Dalmas et al. 2003]. Stories in CBC4Kids are all news articles and cover 12 domains: politics, health, education, science, human interest, disaster, sports, business, crime, war, entertainment, and the environment [Anand *et al.* 2000]. For each story, Ferro and Bevins from the MITRE Corporation added between eight and twelve questions and answer keys [Anand *et al.* 2000]. According to the answer keys, the answer sentences were also annotated. In 2000, CBC4Kids was used to develop and evaluate reading comprehension technologies in a summer workshop on reading comprehension held at Johns Hopkins University.[2]

Details about Remedia and CBC4Kids[3] are given in Table 2. The distributions of different types of questions[4] in the Remedia training set and CBC4Kids training set are shown in Table 3 and Table 4, respectively. In Table 3 and Table 4, we divide *what* questions into four sub-types, since this question type asks for a variety of information, ranging from definitions to reasons, numbers, events, time, locations, person names, etc. *What–DEF* questions ask for definitions; *What–VP* questions ask about actions that the question subjects performed; *What–NP* questions ask for noun phrases which are subjects or objects of question predicates; *What–OTH* questions ask for reasons, numbers, times, locations, person names, organizations, etc.

Table 2. Details of Remedia and CBC4Kids

	Remedia	**CBC4Kids**
Publisher	Remedia Publications	Canadian Broadcasting Corporation
Training set	55 stories	73 stories
Test set	60 stories	52 stories
Corpus size	20K words	35K words
# questions	575	1232
Annotated information	named entities, anaphora co-references, and answer sentences	part-of-speech tags, base forms of words, named entity tags, anaphora co-references, parse trees, and answer sentences

[1] http://www.cbc4kids.ca

[2] http://www.clsp.jhu.edu/ws2000/groups/reading/

[3] We obtained CBC4Kids from Lisa Ferro.

[4] The question type "others" in Tables 3, Tables 4, 5 and 7 refers to *how many*, *how much*, *how long*, *how often*, *how far*, *how tall*, etc.

Table 3. Question distributions in the Remedia training set

Question type	# questions	Example
Who	43	Who is Christopher Robin?
What-DEF	13	What is the stock market?
What-VP	8	What did Jackie Cochran do?
What-NP	15	What did Alex eat on the island?
What-OTH	8	What causes these lights? What is the baby's name? What is the name of our national library?
When	43	When was Winnie the Pooh written?
Where	42	Where did young Chris live?
Why	42	Why did Chris write two books of his own?
Other	1	How high did she take her plane?
Overall	**215**	

Table 4. Question distributions in the CBC4Kids training set

Question type	# questions	Example
Who	100	Who runs the club?
What-DEF	37	What is Bill 101?
What-VP	39	What did Meiorin do to get her job black?
What-NP	69	What does the round goby eat?
What-OTH	29	What causes a solar eclipse? What company runs YNN? What is the capital of Turkey?
Which	10	Which leader is the premier of Ontario?
When	68	When did the metal shop close?
Where	79	Where is Brasilia?
Why	92	Why is the school opening the club?
How	53	How can the satellite help farmers?
Others	75	How big is the club? How many people live in La Ronge? How much was Babe paid to play basketball?
Overall	**651**	

In the above, we have summarized the common characteristics of two English corpora, Remedia and CBC4Kids, which can be used to guide the design of a bilingual reading comprehension corpus. Passages in these corpora are taken from language learning materials designed for children. They consist of stories that cover a variety of domains, which we refer to as open domains. As shown in Table 2, the two corpora contain on the order of a hundred passages, several tens of thousands of words, and several hundred questions. In addition, these two corpora provide experimental materials to support the development of automatic syntactic and semantic analysis techniques that can contribute to reading comprehension. Hence, they include annotations such as named entities, anaphora co-references, part-of-speech tags, and parse trees in order to support automatic named entity filtering, pronoun resolution, and syntactic and semantic analysis.

To estimate the reading difficulty of the passages in Remedia and CBC4Kids, we use the Dale-Chall readability formula[5] [Dale and Chall 1948]. It is believed that this formula is more accurate than other existing formulas when it is applied to the passages for grades four and above [Klare 1963]. To measure the readability of a passage, we first compute a raw score according to the following formula:

Raw score = (0.0496 × average sentence length) + (0.1579 × percent of words in passage not found on Dale Word List[6]*) + 3.6365*

We then consult the Grade Equivalent Conversion Chart [Dale and Chall 1948] with the raw score to obtain a grade equivalent reading score. For example, a raw score in the range from 7.0 to 7.9 implies suitability for grades 9–10. According to the Dale-Chall formula, the maximum readability grade of Remedia is 8; the minimum grade of Remedia is below grade 4. The Dale-Chall formula cannot determine the exact minimum readability score when the grade is below 4. CBC4Kids is at readability grades 7–15.

3. Design Considerations for the Bilingual Reading Comprehension Corpus (BRCC)

We referenced the English corpora, Remedia and CBC4Kids, in formulating the design considerations for the BRCC. We searched for language learning materials that have both English and Chinese versions. These materials needed to contain sufficient numbers of passages and corresponding English and Chinese questions as well as answer keys. The passages also needed to cover a range of readability, as measured using methods such as Dale Chall's formula. We summarize the design considerations for passages, questions, and answer keys, respectively, in the following:

- Parallel Chinese-English passages should be provided for cross-lingual evaluation.
- Passages should be open domain.
- Passages with different grades of difficulty should be selected in order to challenge RC systems to achieve higher levels of performance.

Our design considerations for *questions* were:

- Parallel Chinese-English questions should be provided.
- Questions should cover different types (see Table 4) that correspond to key information in the passages, and they should be authored with various levels of difficulty.

---

[5] http://www.interventioncentral.org/htmdocs/tools/okapi/okapi.shtml

[6] http://www.interventioncentral.org/htmdocs/tools/okapi/okapimanual/dalechalllist.shtml

Our design considerations for *answers* were:

- Parallel Chinese-English answer keys should be provided.
- Sentences that contain the same meanings as answer keys should be provided for the HumSent evaluation metric.

Annotations in the corpora should be based on solid principles. Inter-annotator agreement should also be enforced whenever possible. In order to evaluate different NLP technologies used in RC tasks, we propose to annotate linguistic knowledge about the structure and the meaning of language at different levels [Allen 1995]. In the initial version of the BRCC, we annotated such linguistic knowledge as the boundaries of noun phrases (e.g., "*the English language*" is a noun phrase), several types of named entities (e.g., person names, locations, etc.), and anaphoric references (e.g., "*Edison*" may be the referent of the pronoun "*he*").

4. Development of the BRCC

According to our design considerations discussed in the above section, we selected a bilingual RC book as raw data to develop the BRCC. The book "英語閱讀100天," published by the Chung Hwa Book Company (Hong Kong) Limited, supports English learning by Chinese readers. In total, there are 100 parallel Chinese-English reading comprehension passages. The corpus size is about 18K English words and 17K Chinese characters. There are 414 questions with corresponding answer keys in English. We manually translate the English questions and answers into Chinese. In addition, the passages cover the following domains: the English language, tourism, culture, society, sports, history, geography, arts, literature, economy, business, science, and technology. These attributes of the bilingual book are comparable to those of Remedia and CBC4Kids. We reserved 50 passages as the training set and the other 50 passages as the test set.

According to the Dale-Chall formula, the readability levels of the English training passages in the BRCC range from grades 7 to 12. Hence, compared to the passages in Remedia, those in the BRCC are more difficult. Compared to the passages in CBC4Kids, those in the BRCC are easier.

There are four questions on average for each passage. The distribution of different types of questions in the training set is listed in Table 5. We found that the corpus provides twelve types of questions that ask for key information from the passages, namely, *Who, What–DEF, What–VP, What–NP, What–OTH, Which, When, Where, Why, Yes/No, How* and *Others* (see Table 5). Table 6 shows a sample passage with questions and answer keys from the BRCC in both English and Chinese. In this table, the Chinese questions and answer keys are translated from English.

Table 5. Question type distribution in the BRCC training set

Question type	# questions	Example
Who	23	Who was William the Conqueror?
What-DEF	37	What is shocking story?
What-VP	4	What was Abraham Lincoln determined to do?
What-NP	35	What were well developed?
What-OTH	14	What is the size of Disney World? What is the nickname of Franz Beckenbauer? What is the main cause for those tragedies in America at schools?
Which	14	Which language was the other choice?
When	9	When did Disney World come into being?
Where	16	Where is New York City mainly located?
Why	15	Why do people have such a worry?
Yes/No	16	Did Thomas Edison like talking much?
How	15	How is an American nicknamed today?
Others	22	How many World Cups have there been? How long will the improvement take? How often do people use “hello”?
Overall	**220**	

Table 6. A sample passage with questions and answer keys in both Chinese and English in the BRCC

English passage	Image this: you have just won a competition, and the prize is an English language course at a famous school in Britain or the United States. You can either take a 30-week course for four hours a week, or a four-week course for 30 hours a week. Which one should you choose?…
English questions	1. If you win a competition, what may be the prize? 2. What may be the two kinds of courses?
English answer keys	1. The prize may be an English language course at a famous school in Britain or the United States. 2. They are either a 30-week course for four hours a week or a four-week course for 30 hours a week.
Chinese passage	想像一下：你刚赢得一场比赛，其奖赏是在英国或美国的一所名牌大学学习一门英语语言课程。你可以选一门为30周的课程，每周学习4小时，或者选一门为期4周的课程，每周30小时。你将作何选择？…
Chinese questions	1. 如果你赢得一场比赛，奖赏或许是什么？ 2. 两类课程会是什么？
Chinese answer keys	1. 奖赏会是在英国或美国的一所名牌大学学习一门英语语言课程。 2. 它们分别是一门为期 30 周的课程，每周学习 4 小时，或者选一门为期 4 周的课程，每周 30 小时。

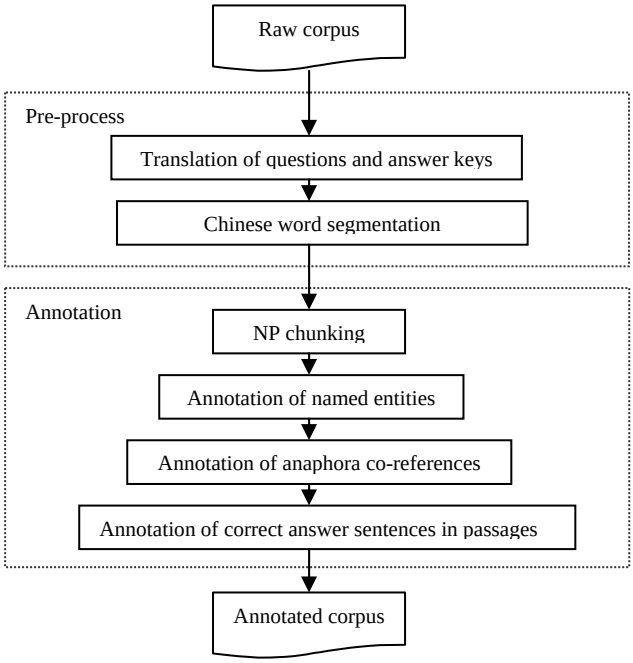

Figure 1. The development flow of the BRCC

The development flow is shown in Figure 1. We pre-processed the raw data in two steps:

1. We translated English questions and answer keys into Chinese manually.
2. We segmented Chinese words in Chinese passages and questions using the Chinese segmenter[7] provided by the LDC. The Chinese language is written without word delimiters. In order to process the text based on words, we needed to tokenize the Chinese character strings into Chinese words. In addition, we annotated noun phrases, named entities, anaphora co-references, and correct answer sentences for passages and questions in both languages.

4.1 Translation of Questions and Answer Keys

The original book contains English passages, English questions, English answer keys, and Chinese passages, but no Chinese questions or answer keys. We arranged for two individual translators with high proficiency in English to translate the English questions and answer keys into Chinese. They worked separately without any communication between them. The two translators then cross-checked the translations and labelled identical translations (exact string matching) as correct. Disagreements in translations were reviewed individually and resolved

[7] http://www.ldc.upenn.edu/Projects/Chinese/LDC_ch.htm

by both translators together. In total, they found that 48.1% of the translations were exact matches. The distribution of miss-matches is explained in the following.

1. Using synonymous words (36.2%): For example, the translator either used "喝" (meaning: drink) or "饮" (meaning: drink) in the translation. Either translation is acceptable.
2. Using semantic equivalent descriptions (8.2%): The two translators express the same meaning in different ways. For example, they used "北方军" (meaning: northern army) or "北方的士兵" (meaning: northern soldiers) in their translation. The meaning of the second translation is equivalent to the English question, "*How were the* ***northern soldiers*** *called by the southern army during the Civil War?*" Therefore, the second translation was adopted.
3. Using active or passive voice (2.7%): For example, the two translators used "什么被播出" (in passive voice) or "播出什么" (in active voice) to translate "*what is shown*." Since the English question uses the passive voice, the Chinese translation with passive voice was adopted.
4. Using anaphora versus direct reference (1.9%): The two translators represented the same entity using its name or its anaphor. For example, they used "他" (meaning: he) or "李克" (meaning: Rick) to refer to "*Rick*" in the question. Since "*Rick*" (i.e., direct reference) was used in the question, the second translation was adopted.
5. Different meanings (2.9%): For example, the English question was "*What did the music center on?*" The two translations were "该音乐用在什么地方" (meaning: what did the music use for) and "该音乐把什么作为中心" (meaning: what did the music center on). The first translation is wrong. The second translation is correct. The translation with the equivalent meaning to the English question was adopted.

4.2 Chinese Word Segmentation

Chinese word segmentation is a process in which word boundaries are identified. Typically, a space is inserted as a delimiter between Chinese words during segmentation. In this process, we manually corrected the outputs of an automated segmenter to produce correct Chinese words boundaries. We began by using the LDC Chinese segmenter to process all the passages, questions, and answer keys in order to maintain consistent segmentation across the board. This segmenter applies a dynamic programming approach to find the word segmentation path which has the highest multiple of word probabilities. In addition, this segmenter includes a lexicon in GB code, which contains 44,405 entries. Each entry of the lexicon contains a word, the occurrence frequency of the word, and the pinyin spelling (indicating the Mandarin pronunciation) of the word. Then, two linguistic annotators corrected and cross-checked the output of the Chinese segmenter so that the meanings of the Chinese sentences agreed with

those of the corresponding English sentences.

For example, "*Hingis*" is a person name in the English sentence, "*Hingis was so frustrated after the game.*" The corresponding Chinese translation is "赛后，轩芝丝感到灰心丧气。""轩芝丝" is the Chinese transliteration of "*Hingis*." It should be a Chinese word. However, the segmenter did not recognize the person name and mistakenly segmented it into three single-character words: "轩" (meaning: room), "芝" (meaning: a kind of herb), and "丝" (meaning: silk). These meanings are not found in the corresponding English sentence. Therefore, "轩", "芝", and "丝" was merged into a transliterated name, "轩芝丝".

As another example, consider the English sentence, "*the USA's population increased 3.3%.*" The segments in the output, "美国人 口," included "美国人" (meaning: American) and "口" (meaning: mouth), which do not agree with the meaning of the original meaning of the English sentence. Hence, the correct segments should be "美国(meaning: USA) 人口 (meaning: population)."

4.3 Noun Phrase Chunking

This section describes the annotation of noun phrase (NP) boundaries for all the passages as well questions in both English and Chinese. This step was important because we assumed that the named entities, anaphors, and antecedent boundaries were consistent with the NP boundaries. Named entities and anaphora co-references were annotated based on the NP boundaries.

Noun phrase chunking is the process that segments sentences into non-recursive portions of noun phrases. After NP chunking, the NP boundaries (denoted by square brackets) are marked in the sentence. For example, note the following sentence before and after NP chunking:

The government has other agencies and instruments for pursuing these other objectives.

[*The government*] *has* [*other agencies and instruments*] *for pursuing* [*these other objectives*] .

We first annotated English NP boundaries and referenced these to annotate the Chinese NP boundaries. Hence, the annotations of the NP boundaries in both English and Chinese were consistent. The annotation procedure is described below.

1. We applied Brill's part-of-speech transformational tagger[8] to annotate each word in the English text (passages, questions, and answer keys) with its part-of-speech tag [Brill 1994].
2. We applied the BaseNP Chunker[9] to identify English noun phrases. The input of the BaseNP Chunker was the output of the Brill's part-of-speech transformational tagger. With heuristic transformational rules trained on Wall Street Journal text from the UPenn Treebank, the BaseNP Chunker inserted square brackets marking the contained baseNP structures [Ramshaw and Marcus 1995].
3. We automatically replaced the left bracket "[" with "＜NP＞" and right bracket "]" with "＜/NP＞". Thus, the annotation tags of named entities and anaphora co-references could be inserted within the brackets of the noun phrase, "＜NP＞".
4. Two human annotators corrected and cross-checked the outputs of the BaseNP Chunker. Each noun phrase should have had a noun or pronoun as the head. Other words in the noun phrase were modifiers of the head. Normally, a noun phrase should have ended with a noun/pronoun/adjective. In addition, a noun phrase should not have overlapped other noun phrases since we only considered non-recursive baseNP structures.

For example, consider the input sentence:

Some say it came from the French, "ho" and "la" – "Ho, there!"

The output of the BaseNP chunker was:

[*Some say*] [*it*] *came from* [*the French*] *, "ho" and "la" – "Ho, there!"*

After replacing the left bracket "[" with "＜NP＞" and the right bracket "]" with "＜/NP＞", the output became:

＜NP＞ *Some say* ＜/NP＞ ＜NP＞ *it* ＜/NP＞ *came from* ＜NP＞ *the French* ＜/NP＞ *, "ho" and "la" – "Ho, there!"*

[8] http://www.cs.jhu.edu/～brill/RBT1_14.tar.Z

[9] ftp://ftp.cis.upenn.edu/pub/chunker/basenp_chunker_1/

Actually, the NP boundary behind “Some say” is wrong because “say” is a verb in the sentence. The valid boundary should be inserted behind “Some”. Therefore, after the annotator corrected the error, the annotated sentence became:

> ＜NP＞ *Some* ＜/NP＞ *say* ＜NP＞ *it* ＜/NP＞ *came from* ＜NP＞ *the French* ＜/NP＞ *, “ho” and “la” – “Ho, there!”*

In the process of annotating Chinese NP boundaries, we referenced the annotations of English NP and attempted to follow these to achieve cross-language consistency. For example, the corresponding Chinese sentence with word segmentations for the previous example is:

> 有人 认为 他 来自 於 法语 的 *“ho”* 和 *“la”* – 意思 是：「喂，那边的！」

The corresponding annotation of Chinese NP chunks is:

> ＜NP＞有人＜/NP＞认为＜NP＞他＜/NP＞来自於＜NP＞法语＜/NP＞的 *“ho”* 和 *“la”* – 意思 是：「喂，那边的！」

In this example, the English counterpart of “有人” (meaning: somebody) is “*some*”; the English counterpart of “他” (meaning: he or it) is “*it*”; the English counterpart of “法语” (meaning: French) is “*the French*.”

However, due to differences between the two languages, occasions may arise in which a strict correspondence between the NP in English and the NP in Chinese cannot be maintained. Consider the following example:

> English sentence: *It was once predicted that British and American English would draw so far apart that eventually they would become separate languages.*
>
> Chinese sentence: 曾经 有人 预言 英国 英语 和 美国 英语 的 区别 会 越来越大 直至 成为 两种 不同 的 语言。

“*It was predicted*” is translated as “有人 (meaning: somebody) 预言 (meaning: predict)” in the Chinese sentence. “*Somebody*” (有人) cannot match “*it*” because the subject of “*predict*” is not “*somebody*” in the English sentence but in the Chinese sentence. In addition, the translation of the Chinese word “区别” (meaning: difference) “*draw so far apart*,” which is

not a noun phrase. For Chinese NPs whose corresponding English NPs could not be found, the annotators still annotated them in the Chinese sentences. Therefore, "有人," "区别," "英国 英语," "美国 英语," and "两种 不同 的 语言" in the previous example were annotated as noun phrases. The annotated NP boundaries in English and Chinese were:

English sentence: ＜NP＞ *It* ＜/NP＞ *was once predicted that* ＜NP＞ *British* ＜/NP＞ *and* ＜NP＞ *American English* ＜/NP＞ *would draw so far apart that eventually* ＜NP＞ *they* ＜/NP＞ *would become* ＜NP＞ *separate languages* ＜/NP＞.

Chinese sentence: 曾经＜NP＞有人＜/NP＞预言＜NP＞英国 英语＜/NP＞和＜NP＞美国 英语＜/NP＞的＜NP＞区别＜/NP＞会 越来越大 直至 成为＜NP＞两种 不同 的 语言＜/NP>。

4.4 Annotation of Named Entities

This section describes the annotation of named entities. Knowledge about named entities has been applied to develop RC systems [Hirschman *et al*. 1999]. Here, we used a filtering module to rank sentences higher if they contained appropriate named entities. Positive results were achieved by applying named entity filtering.

Table 7. Named entity types for different question types

Question type	**Named entity type**	**examples**
Who	PERSON	Thomas Alva Edison, soldier, etc.
What	-	
Which	-	
When	TIME	100 years, weekend, etc.
Where	LOCATION	the US, England, etc.
Why	-	
Yes/No	-	
How	-	
Others	NUM	Millions, 20 stories, etc.

According to the question types listed in Table 5, we annotated four types[10] of named entities: PERSON, TIME, LOCATION, and NUM (see Table 7). Each annotator examined all the noun phrases and marked named entity tags in the left NP boundaries, "<NP>". The format used was "<NP NE=value>". The expected named entity type was assigned as the value to

[10] ORGANIZATION was not annotated because there were no questions that asked about organizations, even though organizations (e.g., government names, company names, etc.) do appear in the BRCC passages. We will include ORGANIZATION in the BRCC in the future if necessary.

"NE". The guidelines for identifying the four types of name entities were as follows:

- If the NP contains a person name or an occupation of person, the value should be PERSON.
- If the NP contains a year, month, week, date, duration, specific hour, or minute, the value should be TIME.
- If the NP contains a street name, a park name, a building name, a city name, or a country name, the value should be LOCATION.
- If the NP contains a specific number, e.g., for money, frequency, length, height, distance, width, area, weight, or age, the value should be NUM.

For example, after named entity annotation, the English sentence

This greeting may have arrived in England during the Norman Conquest in the year 1066

became

This greeting may have arrived in <NP NE=LOCATION> *England* </NP> *during the Norman Conquest in* <NP NE=TIME> *the year 1066* </NP>.

The corresponding Chinese sentence

这个 打招呼 的 用语 也许 是 在 1066年 诺曼第人 征服 英国 时 带去 的，

became

这个 打招呼 的 用语 也许 是 在 <NP NE=TIME> 1066 年</NP> 诺曼第人 征服 <NP NE=LOCATION> 英国 </NP> 时 带去 的。

In this example, the noun phrase "*England*" ("英国" in Chinese) is tagged LOCATION; "*the year 1066*" ("1066年" in Chinese) is tagged TIME.

4.5 Annotation of Anaphora Co-references

Anaphora co-references show the relationships between anaphors and their antecedents. Both an anaphor and the corresponding antecedent refer to the same referent. We believe

anaphora co-references are important in RC tasks because anaphors can be matched with their antecedents when anaphora resolution is applied.

In this process, noun phrases that contain pronouns are annotated as anaphors; their corresponding antecedents are noun phrases that contain the same entities to which the anaphors refer. If multiple antecedents exist, the nearest prior one is used. For example, consider the following sentences:

The American inventor, Thomas Alva Edison, is believed to be the first person to use "hello" in the late 1800's, soon after the invention of the telephone.

…

Thomas Edison was a man of few words.

He wasted no time.

"*He*" is the anaphor and refers to "*Thomas Edison*" and "*Thomas Alva Edison*." The annotator chooses "*Thomas Edison*" as the antecedent because it is the nearest one.

We annotate anaphora co-references in the left NP tag "<NP>". The annotation format of the anaphor is "< NP REF=value >". The format of the antecedent is "< NP REFID=value >". Each antecedent has a unique value, which is based on a counter that starts counting at the beginning of the passage. The value of an anaphor is identical to that of its antecedent. In other words, the annotator marks the co-reference relationship between an anaphor and its antecedent by assigning them the same value.

Refer again to the previous example; the last two sentences (in both English and Chinese) have the following co-reference annotations:

<NP NE=PERSON REFID=7> *Thomas Edison* </NP> *was a man of few words.*

<NP REF=7> *He* </NP> *wasted no time.*

<NP NE=PERSON REFID=7> 托马斯・爱迪生</NP> 是 个 沉默寡言 的 人。

<NP REF=7>他</NP> 从 不 浪费 时间。

In this example, "*Thomas Edison*" (托马斯・爱迪生 in Chinese) is assigned a "REFID" of 7. "*He*" (他 in Chinese) refers to "*Thomas Edison*," and its "REF" is assigned a value of 7.

4.6 Annotation of Correct Answer Sentences

In order to evaluate RC systems with the HumSent evaluation metric, we annotate answer sentences according to published answer keys, which are written by hand. Answer sentences are passage sentences that contain published answer keys or have the same meaning of published answer keys. Consider the following example:

> Question 1: *What word is used most often in the world?*
>
> Published answer key: *The word "hello" is used most often.*
>
> Answer sentence: *The word "hello" is used more often than any other one in the English language.*

In this example, the answer key is an "extract" from the passage. Consider another example:

> Question 2: *Did Thomas Edison like talking much?*
>
> Published answer key: *He didn't.*
>
> Answer sentence: *Thomas Edison was a man of few words.*

In this example, the answer key is not an "extract" from the passage. However, the answer sentence is equivalent to the published answer key, because the statement that "Thomas Edison" did not like talking much means he was a man of few words.

We mark each answer sentence with left and right boundaries: "<ANSQ*i*>" and "</ANSQ*i*>", where *i* is the sequence of the question. After the previous example is annotated, the correct answer sentences in English and Chinese are as follows:

> <ANSQ2> <NP NE=PERSON REFID=7> Thomas Edison </NP> was a man of few words. </ANSQ2>
>
> <ANSQ2> <NP NE=PERSON REFID=7> 托马斯·爱迪生</NP> 是 个 沉默寡言 的 人。 </ANSQ2>

We list the distributions of different annotations among the 100 passages in Table 8. After noun phrases, named entities, anaphora co-references, and correct answer sentences are annotated, the annotated passage and questions in Table 6 are as shown in Table 9.

Table 8. The distribution of different annotations among the 100 passages

Annotation type	# annotations
Noun phrase	6877
Named entity-PERSON	1504
Named entity-LOCATION	558
Named entity-TIME	307
Named entity-NUM	416
Anaphora co-reference	379

Table 9. The annotated sample passage and questions from Table 6. This sample includes annotation tags for NP boundaries, named entities, anaphora co-references, and answer sentences.

English passage	<ANSQ1>Image <NP>this</NP>: you have just won <NP>a competition</NP>, and <NP>the prize</NP> is <NP REFID=1>an English language course</NP> at <NP>a famous school</NP> in <NP NE=LOCATION>Britain</NP> or <NP NE=LOCATION>the United States</NP>. </ANSQ1> <ANSQ2>You can either take <NP NE=TIME>a 30-week course</NP> for <NP NE=TIME>four hours a week</NP>, or <NP NE=TIME>a four-week course</NP> for <NP NE=TIME>30 hours a week</NP>. </ANSQ2> <NP REF=1>Which one</NP> should you choose?…
English questions	1. If you win <NP>a competition</NP>, what may be <NP>the prize</NP>? 2. What may be <NP>the two kinds</NP> of <NP>courses</NP>?
Chinese passage	<ANSQ1>想像一下：你 刚 赢得 <NP>一场 比赛</NP>，其 <NP>奖 赏 </NP> 是 在 <NP NE=LOCATION> 英 国 </NP> 或 <NP NE=LOCATION>美国</NP> 的<NP>一所 名牌 大学</NP> 学习 <NP REFID=1>一门 英语 语言 课程</NP>。</ANSQ1> <ANSQ2>你 可以 选 <NP>一门 为期 30周 的 课程</NP> ，<NP>每周</NP> 学习 <NP>4小时</NP>，或者 选 <NP>一门 为期4周 的 课程 </NP>，<NP>每周</NP> <NP>30小时</NP>。</ANSQ2>你 将 作 <NP REF=1>何 选择</NP>？…
Chinese questions	1. 如果 你 赢得 <NP>一场 比赛</NP>，<NP>奖赏</NP> 或许 是 什么？ 2. <NP>两类</NP> <NP>课程</NP> 会 是 什么？

5. Benchmark Experiments and Discussion

In order to measure the comparative levels of difficulty among the BRCC, Remedia, and CBC4Kids, we applied the baseline bag-of-words (BOW) approach in our experiments. The same baseline has been previously applied to both Remedia and CBC4Kids. The RC system applied to Remedia is called Deep Read [Hirschman *et al.* 1999]. The input sentence of the

BOW matching approach is represented by a set of words, and the output is the first occurrence of the sentence that has the maximum number of matching words between the word set of the sentence and that of the question. The answer sentences are used to obtain HumSent results with the BOW matching approach.

Three pre-processes are performed prior to BOW matching:

1. The stemmed nouns and verbs are used to replace the original words[11].
2. English stop-word removal: We use the same stop-words list used in the Deep Read system [Hieschman *et al*. 1999]. They are forms of *be, have, do*, personal and possessive pronouns, *and, or, to, in, at, of, a, the, this, that*, and *which*.
3. Chinese stopword removal: The stop-words are the Chinese translations of the English personal/possessive pronouns, 和, 或, 到, 在, 中, 的, 这, and 那. We use the Chinese word segmentations in the BRCC directly.

In addition, we used named entity filtering (NEF) and pronoun resolution (PR) [Hirschman *et al*. 1999] to investigate the annotations of named entities and anaphora co-references. Both approaches have been applied in Deep Read [Hirschman *et al.* 1999]. The results obtained with both approaches showed significant improvements [Hirschman *et al.* 1999]. We applied these two approaches and repeated the experiments on the BRCC. For NEF, three named entity types (PERSON, TIME, and LOCATION) were used to perform answer filtering for three types of questions (*who, when and where*). The relationships are listed in the following [Hirschman *et al.* 1999]:

- For *who* questions, a candidate sentence that contains PERSON is assigned higher priority.
- For *where* questions, a candidate sentence that contains LOCATION is assigned higher priority.
- For *when* questions, a candidate sentence that contains TIME is assigned higher priority.

For Chinese questions, the question types refer to the corresponding English questions.

The Deep Read system uses a very simplistic approach to match five pronouns (*he, him, his, she* and *her*) to the nearest prior person name [Hirschman *et al.* 1999]. In addition, a different module uses the hand-tagged reference resolution of these five pronouns. In our experiment, we automatically resolved these five pronouns based on our hand-tagged references. For Chinese passages, we replaced the four pronouns 他, 她, 他的 and 她的 with their hand-tagged references. The detailed results obtained by applying BOW, NEF, and PR to the BRCC test set are listed in Table 10.

---

[11] A C function (morphstr) provided by WordNet is used to obtain the base forms of words [Miller *et al.* 1990].

Table 10. The detailed results obtained by applying bag-of-words (BOW), named entity filtering (NEF), and pronoun resolution (PR) to the BRCC test set

Corpus	BOW	BOW+NEF	BOW+PR
BRCC English test set	67%	68%	68%
BRCC Chinese test set	68%	69%	69%

The BOW approach achieved 29% HumSent accuracy when applied to the Remedia test set and 63% HumSent accuracy when applied to the CBC4Kids test set. As shown in Table 10, the BOW matching approach seemed to achieve especially good results when applied to the BRCC in comparison with the other corpora, as the improvement achieved by applying NEF and PR were not significant. A possible reason is that the questions tended to use the same words that were used in their answers in the BRCC. In the following example, we list a question and its correct answer in the BRCC training set. The corresponding word sets (i.e., bags of words) were obtained following stemming and stop-word removal:

Question: *Where did many sports played all over the world grow up to their present-day form?*

BOW: {*where many sport play all over world grow up present-day form*}

Correct answer: *Many sports which nowadays are played all over the world grew up to their present-day form in Britain.*

BOW: {*many sport nowadays play all over world grow up present-day form Britain*}

In this example, the intersection between the two word sets contains 10 words, 91% of which are in the question. We further calculated the overlap ratios for all the question-answer pairs in the English parts of the BRCC, Remedia, and CBC4Kids, and show the results in Table 11. We used the following formula:

$$\frac{\#\text{matching words between a question and its correct answer sentence}}{\#\text{ words in the question}}.$$

Table 11. The word overlap ratios for the English parts of the BRCC, Remedia, and CBC4Kids[12]

	BRCC(English)	Remedia	CBC4kids
Training set	71.7%	39.3%	46.3%
Test set	62.1%	37.8%	-

[12] Since the human-marked answers were not provided in the test set of our CBC4Kids copy, we were not able to compute the overlap ratio for the test set.

A high degree of word overlap between questions and correct answers could result in good BOW matching performance, which may mislead us to think that BOW is a sufficient approach for RC. Such overlap will artificially ease the task of RC. The difficulty levels of RC tests depend not only on the overlap between questions and correct answers but also on the world knowledge, domain ontology, etc. Questions may ask for information that is not provided in the passage, or for information that resides in different parts of the passage. Human beings can perform reasoning based on their world knowledge and domain ontology, but this process is really a challenge for machine performing automatic reading comprehension. For example, consider the following question and candidate answers:

Question: *Who owned the Negroes in the Southern States?*

Candidate sentence 1: *The blacks were brought to the Southern States as slaves.*

Candidate sentence 2: *They were sold to the plantation owners and forced to work long hours in the cotton and tobacco fields.*

If an RC system can infer that "*Negroes are sold to the plantation owners*" means "*they own Negroes*," then it will be easy to know that candidate sentence 2 is the correct answer. In this paper, we present RC performance measured using the bag-of-words (BOW) approach in order to use it as a baseline performance benchmark. The BOW approach relies heavily upon the degree of word overlap between questions and their corresponding answer sentences. Improvement beyond this benchmark requires the use of more sophisticated techniques for passage analysis, question understanding, and answer generation. It also requires further work in authoring questions that cover various grades difficulty in order to challenge techniques used in automatic natural language processing.

In addition, it is insufficient to only consider the word overlap in the BOW matching approach. The inter-word relationships, such as lexical dependencies among concepts in syntactic parsing, are also important for developing RC systems. In the following example, we list a question, two candidate sentences, and their corresponding word stems:

Question: *What is the new machine called?*

BOW: {*what new machine call*}

Candidate sentence 1: *A new machine has been made.*

BOW: {*new machine make*}

Candidate sentence 2: *The machine is called a typewriter.*

BOW: {*machine call typewriter*}

In this example, both candidate sentences have two matching words. The BOW matching approach cannot distinguish between them. The matching words between candidate sentence 1 and the question are: "*new*" and "*machine*," where "*new*" is the modifier of "*machine*." The matching words between candidate sentence 2 and the question are: "*machine*" and "*call*," where "*machine*" is the object of "*call*." Candidate sentence 2 and the question share a dependency with respect to the verb "*call*." But candidate sentence 1 and the question do not share any dependency with respect to any verb. Actually, candidate sentence 2 is the correct answer sentence. Based on this example, we believe that syntactic structures, such as verb dependencies between words, can be applied to improve the performance of RC systems.

6. Conclusions

In this paper, we have presented the design and development of a bilingual reading comprehension corpus (BRCC). The reading comprehension (RC) task has been widely used to evaluate human reading ability. Recently, this task has also been used to evaluate automatic RC systems [Anand *et al.* 2000; Charniak *et al.* 2000; Hirschman *et al.* 1999; Ng *et al.* 2000; Riloff and Thelen 2000]. An RC system can automatically analyze a passage and generate an answer for each question from the given passage. Hence, the RC task can be used to assess the state of the art of natural language understanding. Furthermore, an RC system presents a novel paradigm of information retrieval and complements existing search engines used on the Web.

So far, two English RC corpora, Remedia and CBC4Kids, have been developed. These corpora include stories, human-authored questions, answer keys, and linguistic annotations, which provide important support for the empirical evaluation of RC performance. In the current work, we developed an RC corpus to drive research of NLP techniques in both English and Chinese. As an initial step, we selected a bilingual RC book as the raw data, which contained English passages, questions, answer keys, and Chinese passages. We then manually translated the English questions and answer keys into Chinese and segmented the Chinese words. We also annotated the noun phrases, named entities, anaphora co-references, and correct answer sentences for the passages.

We gauged the comparative readability levels of the English passages by applying the Dale-Chall formula to the BRCC, Remedia, and CBC4Kids. We also measured the comparative levels of difficulty among the three corpora in terms of question answering using the baseline bag-of-words (BOW) approach. Our results show that the readability level of the BRCC is higher than that of Remeida and lower than that of CBC4Kids. We also observed that the BOW approach attains a better RC performance when applied to the BRCC (67%) than that it does when applied to Remedia (29%) and CBC4Kids (63%). The measured overlap values were 71.7% (training set) and 62.1% (test set) for the BRCC, compared with 39.3% (training set) and 37.8% (test set) for Remedia. This indicates that there is a higher degree of

word overlap which artificially simplifies the RC task with the BRCC. This strongly suggests that more effort must be made to author questions at various difficulty levels in order for the BRCC to better support RC research across the English and Chinese languages.

Acknowledgements

This project has been partially supported by a grant from the Area of Excellence in Information Technology of the Hong Kong SAR Government. In addition, we would like to thank the anonymous reviewers for their comments.

References

Allen, J., *Natural Language Understanding*, The Benjamin/Cummings Publishing Company, Menlo Park, CA, 1995.

Anand, P., E. Breck, B. Brown, M. Light, G. Mann, E. Riloff, M. Rooth, and M. Thelen, "Fun with Reading Comprehension," Final Report of the Workshop 2000 of Language Engneering for Students and Professonals Integrating Research and Education, Reading Comprehension, in Johns Hopkins University, 2000.

Brill, E., "Some advances in rule-based part of speech tagging," In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, 1994, pp. 722–727.

Buchholz, S., "Using Grammatical Relations, Answer Frequencies and the World Wide Web for TREC Question Answering," In *Proceedings of the tenth Text Retrieval Conference (TREC 10)*, 2001, pp. 502–509.

Chall, J. S. and E. Dale, *Readability revisited: The new Dale-Chall readability formula*, Cambridge, MA: Brookline Books, 1995.

Charniak, E., Towards a Model of Children's Story Comprehension, Ph.D. thesis, Massachusetts Institute of Technology, 1972.

Charniak, E., Y. Altun, R. D. S. Braz, B. Garrett, M. Kosmala, T. Moscovich, L. Pang, C. Pyo, Y. Sun, W. Wy, Z. Yang, S. Zeller, and L. Zorn, "Reading Comprehension Programs In a Statistical-Language-Processing Class," In *ANLP-NAACL 2000 Workshop: Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*, 2000, pp. 1–5.

Collins, M., Head-Driven Statistical Models for Natural Language Parsing, PhD thesis, University of Pennsylvania, 1999.

Dale, E. and J. S. Chall, "A Formula for Predicting Readability: Instructions," *Educational Research Bulletin*, 1948, pp. 37–54.

Dalmas, T., J. L. Leidner, B. Webber, C. Grover, and J. Bos, "Generating Annotated Corpora for Reading Comprehension and Question Answering Evaluation," In *Proceedings of the Workshop on Question Answering held at the Tenth Annual Meeting of the European Chapter of the Association for Computational Linguistics 2003 (EACL'03)*, 2003, pp. 13–19.

Hirschman, L., M. Light, E. Breck, and J. Burger, "Deep Read: A Reading Comprehension System," In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999, pp. 325–332.

Klare, G., "The Measurement of Readability," In *Iowa State University Press, Ames, Iowa*, 1963.

Light, M., G. S. Mann, E. Riloff, and E. Breck, "Analyses for Elucidating Current Question Answering Technology," In *Journal of Natural Language Engineering*, 4(7), 2001, pp. 1351–3249.

Miller, G. A., R. Beckwith, C Fellbaum, D Gross, and K. J. Miller, "Introduction to WordNet: An on-line lexical database," In *International Journal of Lexicography*, 1990, pp. 235–312.

Ng, H. T., L. H. Teo, and L. P. Kwan, "A Machine Learning Approach to Answering Questions for Reading Comprehension Tests," In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 2000, pp. 124–132.

Ramshaw, L. and M. Marcus,"Text Chunking Using Transformation-Based Learning," In *Proceedings of the Third ACL Workshop on Very Large Corpora*, 1995, pp. 82–94.

Riloff, E. and M. Thelen, "A Rule-based Question Answering System for Reading Comprehension Test," In *ANLP/NAACL-2000 Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*, 2000, pp. 13–19.

Voorhees, E. M., "Overview of the TREC 2001 Question Answering Track," In *Proceeding of the NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001)*, 2001, pp. 1–15.