Spoken versus Written Queries for Mobile Information Access: an Experiment on Mandarin Chinese

Heather Du
Department of Computer and Information Sciences
University of Strathclyde
Glasgow
heather@cis.strath.ac.uk

Fabio Crestani
Department of Computer and Information Sciences
University of Strathclyde
Glasgow
fabioc@cis.strath.ac.uk

Abstract

As Chinese is not alphabetic and the input of Chinese characters into computer is still a difficult and unsolved problem, voice retrieval of information becomes apparently an important application area of mobile information retrieval (IR). It is intuitive to think that users would speak more words and require less time when issuing queries vocally to an IR system than forming queries in writing. This paper presents some new findings derived from an experimental study on Mandarin Chinese to test this hypothesis and assesses the feasibility of spoken queries for search purposes.

1 Introduction

There is an increasing demand for mobile access to online information today. Mobile phone subscriptions are increasing faster than Internet connection rates. According to official figures released by The Chinese Ministry of Information Industries, the number of mobile phone subscribers in China reached 200 million at the end of November 2002, even though mobile phone calls cost about three times as much as calls made with fixed or "wired" telephones. The number of Chinese web pages is increasing rapidly. With the upcoming 3G wireless networks, it will enable this huge mobile user community to access information anywhere and anytime. Currently, the means of input user's information needs available are very much limited in keypad capability by either keying in or using a stylus on the mobile phone screen, but the speed is far from satisfaction. Moreover, such input style does not work well for those users who are moving around, using their hands or eyes for something else, or interacting with another person, not to mention those with visual impairment such as blindness or difficulty in reading words in ordinary newsprint. In all those cases, given the difficulty of Chinese character input and the ubiquity of mobile phone access, speech enabled interface has come to the lime light of today's mobile IR research community which lets users access information solely via voice.

The transformation of user's information needs into a search expression, or query, is known as query formulation. It is widely regarded as one of the most challenging activities in information seeking (Cool et al., 1996). Research on query formulation with speech is denoted as spoken query processing (SQP), which is the use of spoken queries to retrieve textual or spoken documents. While Chinese spoken document retrieval has been extensively studied over the years, especially supported by TREC-6 (Text REtrieval Conference) (Wilkinson, 1998), little work has been devoted to Chinese SQP. Two groups of researchers have investigated the level of degradation of retrieval performance due to errors in the query terms introduced by the automatic speech recognition (ASR) system by carrying out experimental studies (Chang et al., 2002; Chen et al., 2000). These experiments claimed that despite the current limitations of the accuracy of speech

recognition software, it is feasible to use speech as a means of posing questions to an IR system which will be able to maintain considerable effectiveness in performance.

However, the query sets created in these experiments were dictated from existing queries in textual forms. Will people use same words, phrases or sentences when formulating their information needs via voice as typing onto a screen? If not, how different are their queries in written form from spoken form? Dictated speech is considerably different from spontaneous speech and easier to recognise (Keller, 1994). It would be expected that spontaneous spoken queries to have higher levels of word error rate (WER) and different kinds of errors. Thus, the claim will not be valid until further empirical work to clarify the ways in which spontaneous queries differ in length and nature from dictated ones.

We have carried out an experiment in English languages previously and derived some interesting findings which are reported in (Du and Crestani, 2003). We are motivated by the comments from one of the reviewers for that paper to conduct a similar experiment in Mandarin language to see if we could obtain similar findings as the ones with English queries. From the experiment we conducted in English language, the results showed that using speech to formulate one's information needs not only provides a way to express naturally, but also encourages one to speak more semantically which resulted in longer queries than written text. Thus, we would expect, in the case of Mandarin Chinese, that spoken queries would be longer in length than written ones. Furthermore, the translation of thoughts to speech is faster than the transition of thoughts to writing. In order to test these two hypotheses on Mandarin queries, we repeated the same experiment with just some minor setup changes.

In this paper we present the results of an experimental study on the differences between Mandarin written and spoken queries. The paper is structured as follows. Section 2 describes our experimental environment of the study. The results of this study are reported in section 3. In section 4, the results are compared to the ones derived from the experiment on English queries. Conclusion with some remarks on the potential significance of the study is presented in section 5.

2 Experimental study

Our view is that the best way to assess the differentiations in query formulation between spoken form and written form is to conduct an experimental analysis with a group of potential users in a setting as close as possible to a real world application (Miller, 1984). We used a within-subjects experimental design and in total 10 subjects participated.

2.1 Subjects

As retrieving information via voice is still relatively in its infancy, it would be difficult to identify participants for our study. We therefore decided to recruit 10 native Mandarin speakers who are currently studying different subjects for a Master degree who all have good experience in using Chinese web search engines, but few have prior experience with vocal information retrieval. In the previous experiment, we recruited 12 native English speakers.

2.2 Test collection

The topics we used for this experimental study was a subset of 10 topics extracted from TREC-5 Mandarin topic collection (topics 19-28). Each topic consists of four parts: id, title, description and narrative. The English parts of these topics were ignored and only the parts in Mandarin were presented to the participants. We used 10 TREC topics (topics 151-160) in the previous experiment on English language.

2.3 Experimental procedure

After we have experimented in English language, our aim now is to find out if we can derive the same results from an identical experiment in Mandarin Chinese. Therefore, the main experimental procedure remained the same as the English experiment and a detailed description of the procedural can be found in (Du and Crestani, 2003). The experiment consisted of two sessions. Each participant was given the same descriptions of 10 topics in text form for both sessions. These topics were in a predetermined order and each had a unique ID. For session 1, each participant was asked to form his/her queries in written form for

the first 5 queries and in spoken form for the second 5 queries via a graphic user interface (GUI) on a desktop screen. For session 2, the order was reversed, that was each participant presented his/her queries in spoken form for the first half topics and in written form for the second half topics via the GUI. A maximum of 5 minutes time constraint was also imposed on each topic. We deliberate run session 2 one week later than session 1 in attempt to minise the risk of participants' familiarities with the data.

2.4 Data capture

We utilised three different methods of collecting data for post-experimental analysis: background system loggings, interviews and questionnaires. By these means we could collect data that would allow us to analyse and test the experimental hypotheses.

During the course of the experiment, the written queries were collected and saved in text format along with the duration of each query. The duration of each written query was counted as the total time a participant spent to comprehend a topic and formulate his/her query in the query field and submit it. The spoken ones were recorded using a close-talk microphone and saved in audio format automatically for each participant along with the duration for each query which was calculated in a similar way.

The interviews sought to solicit participants' comments after each session and they were also asked to point out the easiest and most difficult topics in written and spoken form and the reasons for their judgments.

The same questionnaires were handed out after the completion of both sessions to gather participants' assessment on the complexity of the tasks. By comparing their answers, we could see how their ratings on the difficulty of the tasks would vary from session 1 to session 2.

3 Experimental results and analysis on Mandarin queries

From this experiment, we collected 100 written queries and 100 spoken queries. We manually transcribed spoken queries into textual form. Each Chinese character has its own meanings, and it can form a compound word to give more complete and richer meaning with neighboring characters. In the context of a Chinese sentence, there are no spaces, or explicit separators between words to indicate boundaries, appearing as a linear sequence of equally spaced characters (Oard and Wang, 1999). With the fact that a character may form a meaningful word with its previous or next neighboring character(s) in the context of a sentence, many different word combinations are possible. One mistaken word combination would lead to an incorrect segmentation for the rest of the sentence. Therefore, Chinese word segmentation has been a popular topic in the IR community. There are a number of segmentation approaches which have been proposed to tackle this problem (Sun et al., 1998). The main approaches are statistical-based and dictionary-based methods. With the dictionary-based segmentation, words in Chinese text are segmented by comparing and matching the words in a dictionary. Such methods have a stronger accordance to the semantics and sentence structure of the Mandarin Chinese. As Chinese words can vary in length, the dictionary is flexible to change as the vocabulary can be updated and new words can be inserted manually. Therefore, we have chosen a dictionary-based method developed at the in2in research laboratory to segment the transcriptions of the spoken queries (Chen, 2001). There were more than 58,000 entries in the dictionary adopted by this method. Some of the characteristics of written and spoken queries are reported in Table 1. This table pictures clearly that the average length of spoken queries is longer than written queries with a ratio rounded at 2.38 as we have hypothesised. After stopwords removal, the average length of spoken queries is reduced from 22.66 to 17.61with a 22% reduction rate and the average length of written queries is reduced from 9.51 to 8.29 with a reduction rate at 13%. These numbers indicate that spoken queries contain more stopwords than written ones. There is also a significant difference on durations for formulating queries in spoken and written forms.

3.1 Length of queries across topics

The average length of spoken and written queries for each topic across all 10 participants is calculated and presented in Fig. 1. In Fig. 1, the scat-

Table 1: Characteristics of written and spoken queries: + indicates presence of stopwords, - indicates absence of stopwords

Data set	Written	Spoken
Queries	100	100
Average length +	9.51	22.66
Average length -	8.29	17.61
Median length +	7.9	17.4
Average duration +	2m 33s	1m 36s

tered points for spoken queries are always above the ones for written queries, which suggests the spoken queries are lengthier than the written ones. This is the case for every topic persistently. This is exactly what we expected to see. When formulating queries verbally, the ease of speech encourages participates to speak more words. A typical user spoken query looks like the following:

"有关苏联在海湾战争中如何担任调停的角色，具体包括苏联与伊拉克的沟通，以及提出的停火协议还有要求多国部队从伊拉克撤出的和平建议"

Whereas its textual counterpart is much shorter:

"海湾战争，苏联，调停，沟通，停火，撤军"

3.2 Length of queries across participants

We also summarise the average length of queries for all 10 topics across all participants and presented it in Fig. 2. We could observe from Fig. 2 that for 9 out of 10 participants his/her spoken queries are longer than written ones. There is only one participant whose written queries are just slightly lengthier than spoken queries. However, the variations of the length between spoken and written queries for some participants are very timid. In fact, after we have studied the transcriptions of spoken queries, we observed that the spoken queries generated by these participants are very much identical to their written ones. The discrepancies of length within written queries are very insignificant and relatively stable. All participants used similar approach to formulate their written queries by specifying only keywords. The experience of using textual search engines influenced the participants' process of query formulations. For most popular textual search engines, the stopwords would be removed from a query before creating the query representation. Conversely, the length fluctuates rapidly within spoken queries among participants.

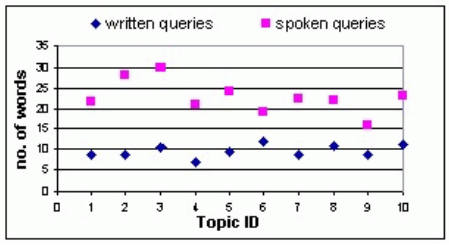

Figure 1: Avg length of queries per topic.

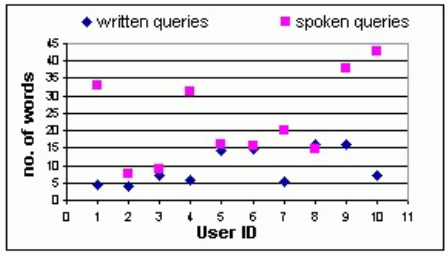

Figure 2: Avg length of queries per user.

We didn't run a practice session prior to the experiment such as to give an example of how to formulate a written query and a spoken query for a topic, because we felt this would set up a template for participants to mimic later on during the course of experiment and we wouldn't be able to find out how participants would go about formulating their queries. Existing research also shows that people do not use the same language when talking to a computer as when talking to another human (William, 1998). In this experiment, we observed that half of the participants adopted natural language to formulate their queries which were very much like conversational talk and the other half participants stuck to the traditional approach by only speaking keywords and/or broken phrases. They said they didn't "talk"to the computer was because they felt strange and uncomfortable to speak to a machine. This suggests that participants own personalities played a key role in the query formulation process.

3.3 Duration of queries across topics

The time spent to formulate each query was measured. A maximum of 5 minutes was imposed on each topic and participants were not allowed to work past this. All participants felt that the time given was sufficient. The average time participants spent on each topic is shown in Fig. 3. Just as we would expect, it is the same case for every topic consistently that less time is required to form a spoken query than a written one. However, the discrepancies between the spoken queries and written queries for the first half topics are quite obvious whereas much less differences exist for the second half topics.

The variations of durations within written queries and within spoken queries were quite irregular, but nevertheless, we could observe that on average the written queries for the first half topics need more time to formulate than the second half topics, and for spoken queries, the second half topics require more time than the first half topics. This could be due to the way in which the tasks were assigned to the participants. They formulated written queries for the first half topics in session1 and the second half topics in session 2, while spoken queries were formulated for the second half topics in session 1 and the first half topics in session 2. Although we intentionally carried out the session 2 at least one week after the completion of session 1 to minimise the participants' memory jogging on these topics, some participants commented that they could still recall the topics very well, therefore, they could save time from comprehending the topics again in session 2. This trend is also reflected on the shifting of their ratings on task complexities which are reduced from session 1 to session 2. However, we couldn't neglect the fact that the cognitive load of participant to speak out their thoughts was also high. Some of them commented that they had to well-formulate their queries in head before speaking aloud with no mistakes. One could revise one's textual queries easily in a query field, but it would be difficult for the computer to understand if one alters one's words while speaking. Information retrieval via voice is a relatively new research area and there aren't many working systems available currently. Lacking of experience also pressurised the spoken query formulation process. We could claim that the complexity of the topics and the participants' knowledge about the topics also played a part in the query formulation process. We asked the participants to point out the most difficult and easiest topics during the interview. Their judgments conformed to Fig. 3, that is they spent longest time on the topics thought to be most difficult and least time on the ones deemed to be simple and straightforward.

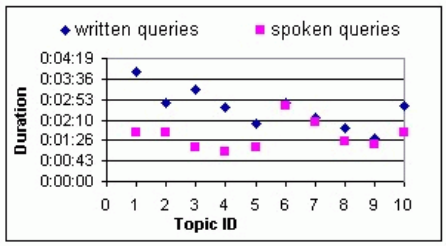

Figure 3: Avg duration of queries per topic.

3.4 Duration of queries across participants

The duration of queries per participant is shown in Fig. 4. Every participant spent less time on spoken queries than written ones with only one exception. This strongly supports our claim that people would require less time to form spoken queries since that is the way they communicate to each other, furthermore, this means that it would be quicker to specify one's information needs via voice than key in or hand write into a mobile device. Since the mobile communication cost is still relatively expensive, the less the time needed to access information via a mobile device, the cheaper the cost incurs. The discrepancies between the written and spoken queries for some participants are significant and for other participants are minimal. The variations within spoken queries are relatively steady, but change rapidly among the written queries. Statistically, the quickest participant required only 01:44 on average for a written query whereas the slowest participant doubled that amount of time. This is because of the difficulty arisen from the input of Chinese using keyboards. In Chinese, there are two kinds of basic units: PinYin and Chinese character. PinYin represents the pronunciation of the character and usually consists of several alphabets. Chinese character is for expressing Chinese sentences and has several thousands

kinds. Therefore, there are in general two kinds of technique for input Chinese into a computer: PinYin input method and character input method. The PinYin input method is most commonly used. The character input method is quite demanding as it requires memorising the keyboard distribution for the radicals of characters and it is difficult for a normal user. Therefore we have chosen the Microsoft Pin Yin input method in this experiment for participants to enter their written queries. To enter a Chinese character using a Pin Yin method generally consists of following steps:

1. A user inputs a PinYin string corresponding to twenty-six alphabets.
2. System presents the user a list of characters which have the same pronunciation.
3. The user selects the desired character from that list.

This process is considerably slow as a Pin Yin for a Chinese character is often spelled by several alphabets and multiple characters have the same pronunciation, users need to browse through the list of candidate characters linearly to locate the correct one. Pin Yin as the standard pronunciation for Chinese characters is spoken among the population in the north of China, whereas people from different regions speak dialects which would pronounce same character differently. For those people, on top of the tedious character input process, it would be more difficult as they often make Pin Yin spelling mistakes. The participants we recruited for our experiment were from different parts of China, therefore their Pin Yin spelling skills were at different levels. Hence, whilst we can claim that people formulate their spoken queries more quickly we can't ignore the fact that the difficulty of Chinese character input also contributes to the extended written query formulation process.

3.5 Length of spoken and written queries without stopwords across topics

From the previous analysis, we know that spoken queries as a whole are definitely lengthier than written queries. One would argue that people with natural tendency would speak more conversationally which results in lengthy sentences containing

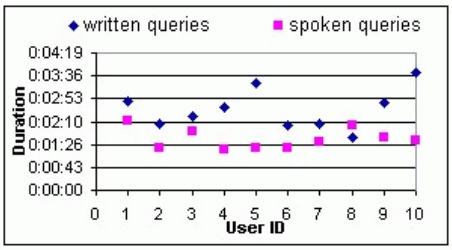

Figure 4: Avg duration of queries per user.

a great deal of words such as conjunctions, measurements or pronouns, that need to be linked to other words to form a complete expression, which have been referred as stopwords in information retrieval community, whereas the written queries are much terser but mainly contain content words such as nouns, adjectives and verbs, therefore, spoken queries would not contribute much than written queries semantically. However, after we removed the stopwords within both the spoken and written queries and plotted the average length of spoken and written queries against their original length in one graph, as shown in Fig. 5, which depicts a very different picture. We adopted a standard Chinese stopword list for our experiment which was compiled at the in2in research laboratory.

As we can see from Fig. 5, the scattered points for spoken queries are consistently on top of the ones for the written queries; after stopword removal, each of them is also undoubtedly becoming shorter. Moreover, the points for spoken queries without stopwords stay above the ones for written queries without stopwords consistently across every topic. Statistically, the average spoken query length without stopwords is 17.61 and for written query, that is 8.29, which shows the spoken queries have almost doubled the length of the written ones. This significant improvement in length indicates that the ease of speaking encourages people to express not only more conversationally, but also more semantically. From information retrieval point of view, more search words would improve the retrieval results (Yang and Ma, 2001). The accuracy of ASR systems is critical in SQP because IR engines try to find documents that contain words that match those in the query; therefore any errors in the query have the potential

for derailing the retrieval of relevant documents. But the longer the query is, the higher the likelihood of an important word is repeated. If this word is not recognised one instance, it will probably be recognised from other occurrence. This redundancy provides some resilience to recognition errors. In the case in our experiment, even the recognition accuracy is only 50the meanings for spoken queries than written queries, in other word, the spoken information clearly has the potential to be at least as valuable as written material.

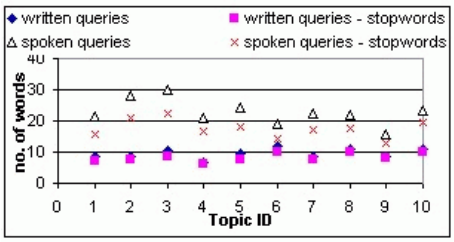

Figure 5: Avg length of queries across topics.

3.6 Length of spoken and written queries without stopwords across participants

The average length of spoken and written queries with and without stopwords across all 10 participants is shown in Fig. 6. This graph shows that very few stopwords are specified in the written queries. Statistically, on average a written query contains 1.5 stopword. This figure also demonstrates that half of our participants issued their spoken queries using natural language whereas the other half felt uncomfortable "talking" to a machine and only specified keywords in their spoken queries. This further sheds light on our claim that people's individual personality also influenced the spoken query formulation process.

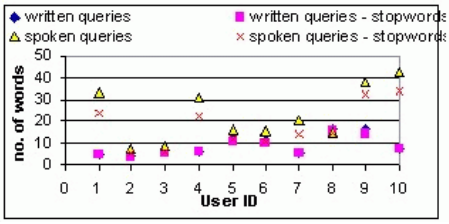

Figure 6: Avg length of queries per user.

4 Comparison with the experimental results of English queries

We have conducted two experiments consecutively to analyze the differences between written and spoken queries in two different languages. The results of the first experiment in English language can be found in our paper (Du and Crestani, 2003) and the average length of English queries across topics and across participants are reproduced here in Fig. 7 and Fig. 8 respectively. The second one in Mandarin Chinese is presented in the previous section in this paper. We could observe some similarities between the results of these two experiments by comparing the data collected and figures generated. In both languages, people tend to speak more words for spoken queries than written ones. Spoken queries contain not only more stopwords but also more content words than written ones for both languages. Despite these two languages with completely different semantic structures, we have found that the number of words used to specify both written and spoken queries were extremely close. The discrepancy in the two experiment results exists in the durations of query formulation process. For the experiment in English language, we were unable to establish any strong claim because no significant differences existed between the two query forms in terms of duration. However, in this experiment reported here, we have observed that considerably less time is required to formulate spoken queries than written ones in Mandarin Chinese.

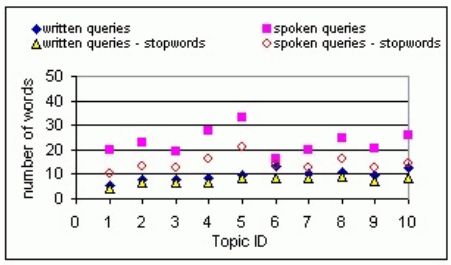

Figure 7: Avg length of English queries across topics.

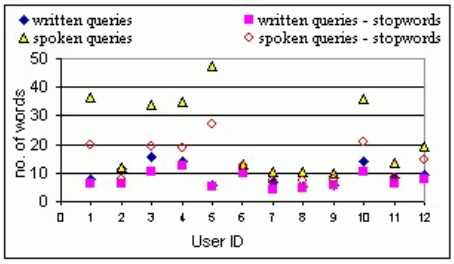

Figure 8: Avg length of English queries per user.

5 Conclusions

This paper reports on an experimental study on the differentiations between Mandarin written and spoken queries in terms of length and durations of the query formulation process, which can also contribute to the speech dialogue design for Chinese mobile information access systems. To the best of our knowledge there is no other similar study carried out in Mandarin Chinese for mobile IR. The results show that using speech to formulate one's information needs not only provides a way to express naturally, but also encourages one to speak more semantically. This indicates that spoken queries will have better quality than written ones in terms of information seeking since more search words will lead to better IR performance. Furthermore, as far as ASR systems are concerned, longer queries will have more context and redundancy which would potentially decease the impact of the ASR errors on retrieval effectiveness. One's individual personality is another factor influencing query-issuing approaches. In spite of the cognitive load, one can translate one's thoughts via voice more quickly than write them down in text form. Consequently, accessing information verbally on mobile devices will be more cost-effective and more affordable than conventional text input. This means that we can reach the conclusion that spoken queries as a means of formulating and inputting information needs on mobile devices are utterly feasible. Nevertheless, this empirical study was carried out with a small number of participants, further studies are required with larger user population to underpin these results.

References

Eric Chang, Frank Seide, Helen M. Meng, Zhouran Chen, Yu Shi, and Yuk-Chi Li. 2002. A System for Spoken Query Information Retrieval on Mobile Devices. *IEEE Transactions on Speech and Audio Processing*, VOL. 10, Pp 531-541.

Berlin Chen, Hsin-min Wang, and Lin-shan Lee. 2000. Retrieval of Mandarin Broadcast News Using Spoken Queries. In *Proc. International Conference on Spoken Language Processing*, (ICSLP2000), Beijing.

Hongbiao Chen. 2001. CLAS: a general purpose system for Chinese corpus processing. *First International Conference on Formal Linguistics*, Changsha, China.

Colleen Cool, Soyeon Park, Nicholas Belkin, Jurgen Koenemann, and Kwong Bor Ng. 1996. Information seeking behavior in new searching environment. *CoLIS 2*, Copenhagen. Pp 403-416.

Heather Du and Fabio Crestani. 2003. Spoken versus Written Queries for Mobile Information Access. *Proceedings of the MobileHCI03 workshop on Mobile and Ubiquitous Information Access*, Udine, Italy.

Eric Keller (Ed.). 1994. Fundamentals of Speech Synthesis and Speech Recognition, *John Wiley and Sons*, Chichester, UK.

Doug Oard and Jianqiang Wang. 1999. Effects of Term Segmentation on Chinese/English Cross-Language Information Retrieval. In *Proceedings of the Symposium on String Processing and Information Retrieval*, Cancun, Mexico. Pp 149-157.

Steve Miller. 1984. Experimental design and statistics. *Routledge*, London, UK, second edition.

Maosong Sun, Dayang Shen, and Benjamin K. Tsou. 1998 Chinese word segmentation without using lexicon and hand-crafted training data. In *Proceedings of COLING-ACL-98*, Pp 1265-1271. Montreal, Canada.

Ross Wilkinson. 1998. Chinese document retrieval at TREC-6, in *Proc. 6th Text Retrieval Conference*, (TREC-6), Pp. 25-30.

D William and Christine Cheepen. 1998 "Just speak naturally": Designing for naturalness in automated spoken dialogues. In *Proceedings of Conference on Human Factors in Computing Systems*, Pp 243–244, Los Angeles.

Yiming Yang and Nianli Ma. 2001. CMU cross-language information retrieval at NTCIR-3, in *Proc. Of the Third NTCIR Workshop on research in information Retrieval, Automatic Text Summarization and Question Answering*.