Technical Issues in Building An Information Retrieval System for Chinese[1]

John Broglio, Jamie Callan and W. Bruce Croft

Center for Intelligent Information Retrieval

Amherst, Massachusetts

[1] This material is based on work supported in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

Building a Chinese Retrieval System

1.1.1 General issues

Information retrieval in a foreign language requires modification to text and user interfaces. Stemming, word boundary identification, punctuation and stopword identificdation must all be modified; appropriate input and presentation methods must be provided. But once these interface issues are resolved the retrieval model and enhancement techniques operate equally effectively in all the languages we have worked with.

Text and user interface issues:

- Writing direction: right-left, left-right, or top to bottom starting on the left.
- The fundamental concept of what is an indexable word (or *term*) changes from language to language, as does the concept of a word stem or root. Some languages, like Chinese and Japanese, are written continuously, with no spaces between words.

  In Chinese, *artificial intelligence* becomes a four character phrase, which could be translated literally as *man-made-cognition-able*. The first two characters are purely semantic, like the compound word *man-made* in English. The third character is semantic and the fourth operates much like the suffix *-able* in English. How many words do we have here? one, two or three?

  In Semitic languages, words are classically viewed as consonant stems with the addition of prefixes and suffixes. The stems undergo changes in vowels or doubling of consonants. Finding a term stem for indexing can generate a lot of false relations between words.

  In agglutinative languages, a single printed word can express the lexical semantics of a complex noun phrase or even a whole sentence.

- A given language may have multiple non-ASCII character encodings. Occasionally, as in SJIS, JIS and EUC for Japanese, there is a comparatively simple algorithmic mapping from one encoding to another. More typically, as in Chinese, conversion tables are required because the code for a character in one encoding will bear no relation to the code for the same character in a different encoding.
- Users have different expectations. For example, although research indicates that a bigram model for languages like Chinese may be very effective, a user may be disconcerted to see the second character of one word juxtaposed with the first character of the following word as a search item (compare *ificial intel*).

1.1.2 Character Encodings

The ASCII encoding was evolved and standardized on the English language, so input and display of any other language presents problems for ASCII-oriented display technology and languages such as C, where even the datatype *char* is ambigous and not guaranteed to support more than 7-bit ASCII. Because of this, many foreign languages have alternative character encodings. Languages that do not use the Roman alphabet may have any number of competing encodings in use by different agencies or in different countries or on different platforms.

Modern Chinese has two graphic character sets: Beijing (*simplified*) characters and Taiwan (*traditional*) characters. (Classical Chinese has additional graphic character styles.). In the unix world, these two display sets are encoded by the GB (GuoBiao - Beijing) and Big5 (Taiwan) two-byte eight-bit encodings, although on other systems (e.g., DOS) there are seven-bit encodings of these display character sets.

One problem that must be handled is a query in one character set retrieving documents encoded in different character sets. The query must be transcoded for retrieval from each database and the documents retrieved must then be transcoded into the input set for display. There will be information loss due to incomplete conversion tables or due to local expressions that have no equivalent in the another writing.

1.1.3 Indexing and Segmentation.

Our experience with both Japanese and Chinese has shown that character-based indexing is the most flexible approach to take for Chinese. Indexing each Chinese character as a term ensures that no information is lost.

Although, it is possible to segment the documents into words automatically and index each word as a term, this can cause well-posed queries to fail for two reasons:

- words improperly joined by the automatic segmentation.
- Different understandings of the definition of a word. For example, the Chinese expression for *Beijing Institute of Physics* may be legimately represented in a Chinese lexicon as a single word and a Chinese-speaking user may also perceive it as a word. But if this expression is stored as a single term, then perfectly reasonable queries such as *Physics institutes in China* or *Beijing technical institutes* would fail to match that term. For the same reason, query-time segmentation should include the raw characters or at least the bigrams in the query.

When the document index is character-based, then query-processing can determine proximity constraints based on word and phrase formation. A user may hand-segment the query, the query may be segmented automatically or adjacent bigrams from the query may be used.

Automatic segmentation in Chinese raises the special problem of name recognition. Foreign names are represented phonetically in Chinese by a small set of Chinese characters. These characters may appear individually in Chinese words, but when they are combined to sound out non-Chinese names they form sequences that are not otherwise part of a Chinese lexicon.

Chinese names present a different problem. There is a relatively small number of traditional Chinese surnames, but given names are essentially unrestricted combinations of two-character sequences. A Chinese name recognizer must look for sequences of unsegmented (or poorly segmented) characters, and try to identify a traditional family name, followed by two characters that could be a given name (i.e., not otherwise segmentable as a word or part of a word.)

Ideally name recognition should be efficiently interleaved with segmentation, so that when segmentation fails on a short sequence, the name recognizers can be called. A makeshift substitute for this interaction is to run the name segmenter to identify guaranteed names (from name lexicon), run the segmenter, and then run the name recognizer again, this time to identify possible names from the still unsegmented characters.

1.1.4 Query Processing

There are several issues in query processing besides those encountered by the user interface.

- The input character representation must be matched to the document collection representation, and converted if necessary.
- Characters which carry no meaning, such as punctuation or grammatical particles, should be discarded.
- Groupings of characters that represent words should be identified, either manually or automatically. This may include the special problem of Chinese and foreign name recognition.
- Query expansion methods. We relied on an automatic collection-driven concept-relation technology called InFinder described below and in [Jing].

In our first version of the Chinese IR system, we convert between whatever character sets are represented in our document database. We relied on user hand-segmentation to identify words. Our second version of the system has an automatic segmentation component. Where the user has indicated a preferred segmentation, however, it will be respected by the automatic component.

A future modification will be to combine the segmented query with the raw-character query, and possibly to break long words into their bigram subcomponents.

1.1.4.1 Query Expansion with Related Terms

One of the objectives of information retrieval with respect to the user is to render the technology more accessible by diminishing the gap between the retrieval performance of an expert or trained user and that of a novice or casual user. The InFinder technology shows a lot of promise in this area. The goal was to offer automatic or user-guided query expansion by supplying terms which are related in meaning to the query terms.

In the past, this has been attempted with a general-purpose thesaurus or with a keyword list or topic navigation outline. The general purpose thesaurus fails by bringing in terms which are unrelated to the usage or the context at hand, and by neglecting other terms which are germane to a query term in context. The topic navigation and keyword lists are very expensive to construct and fail in heterogenous collections or in domains which change rapidly

The InFinder technology constructs an automatic related-term database which attacks the two problems of currency relevance with the same mechanism. An automatic catalog is constructed from a collection based on word co-occurrence. Taking any word or phrase as a *concept*, the InFinder program collects and filters frequency information on the words that are most frequently found within two or three sentences of the concept of interest. Since all the information is gathered from the text collection at hand, the term relations are relevant to the text. The resulting database is an INQUERY database which can be updated as desired, so that as new usages appear in the text, they can be added automatically to the InFinder database.[i]

When a query is submitted to the InFinder database for expansion, concepts which are contextually related to the query terms will be retrieved. Some number of the top terms can be automatically added to the original query to add coverage and specificity, or the user can be prompted to select which terms to add to the original query. In the user-guided approach, the user gets the added benefit of immediate feedback as to which concepts in the collection are related to the query. This information can lead to selection of a different collection, or modification of the original query to alter a term that has a domain-specific meaning not intended by the user. For the JINGTIAN project, user-guided expansion was supported.

1.1.5 Relevance Feedback

In relevance feedback, selected documents are processed by the system, and terms which are suggested by those documents are added to the original query. Since the Chinese indexing is character-based, the relevance feedback approach treated characters as query enhancement terms. Since this did not produce good results, we modified the feedback selection techniques to select significant pairs of adjacent characters from the relevant documents (bigram model). This model appears to produce very good results, although the terms added are occasionally disconcerting for the user, since they represent parts of words, or characters from two different words that commonly appear together in a phrase.

We could segment the relevant documents so that we can use actual words in the feedback query. This will produce a more "readable" query, but ongoing research suggests that the results may be the same or worse than those produced by the bigram model.

1.1.6 User Interface

To enable query input to the Chinese language version of INQUERY, it was desirable to have a graphical user interface platform that would allow the input and display of Chinese characters. While there is a great deal of grassroots support in the unix world for display of Chinese and Japanese (*kterm, cxterm*), documentation and stability are unreliable and they do not support sophisticated pointer-driven or menu-based interaction. The best candidate for a platform for a user interface was the New Mexico State University Computing Research Laboratory XAT library of *widgets* based on the Motif library for the X Window System. The XAT library supports display of several different languages, and two important characters encodings for Chinese: the *traditional* or *Big5* encoding, and the *simplified* or *GuoBiao* (GB) encoding. In addition, the XAT library supports several different input methods for both character sets, including both Beijing and Cantonese *pinyin* and the *Standard Telegraphic Code* (STC) 4-digit numeric representation.

The XAT library would allow input of Chinese text, which could then be comunicated to a program. It permits the program to display Chinese text by including an opening and closing annotation which indicated which character-encoding the text was using. It was often the case that documents, most of which came from Beijing, arrived in the simplified character set, while the client users might be more familiar with the STC input method and/or the traditional character encoding and display. Therefore it was necessary to have the XAT library receive STC or Big5 encodings and display traditional characters, and to have INQUERY translate the traditional encodings into simplified characters to retrieve documents from a text collection. For this purpose, we used tables supplied by the client, and conversion programs provided freely on the network (GB-BIG5) or created at CIIR (STC).

1.2 Evaluation of the prototype system

1.2.1 Evaluation Methodology

The purpose of evaluation is to assess retrieval effectiveness against some standards of expected performance. For information retrieval evaluations, a reasonably large set of documents is collected, a set of queries is prepared by domain experts, or collected from users, and the relevance of each document to each query is judged. In practice, the thoroughness of relevance judgments will vary. Only an extremely small collection of documents can be judged completely. For reasonably large sets, a subset of documents is identified and judged for each query. Then the performance of a system can be evaluated based on the subset of judged documents. This is an expensive and time-consuming procedure when done properly, requiring many months of work assembling queries and judging retrieved documents by domain experts.

A given system's performance will be reported in terms of *recall* and *precision*: recall indicates what percentage of all the relevant documents were retrieved at a given point; precision indicates what percentage of the documents retrieved were relevant. As recall increases to 100%, precision will decrease correspondingly.

The INQUERY technology has been formally evaluated in TIPSTER and TREC trials in English, Spanish and Japanese with outstanding results and comparable performance in each language. Since there is as yet no TREC track for a complete evaluation of Chinese IR systems, we have conducted an in-house evaluation with limited resources to determine if the quality of retrieval appeared to be in line with our performance in other languages.

We assembled thirty "natural language" queries, modeled on a current set of TREC queries, a typical query being: "Investment prospects in China for American companies". For each query we had a Chinese language expert examine and judge the ten documents ranked most highly by Chinese INQUERY. The queries were submitted in three different experimental sets: raw characters and two sets of word-based queries: hand segmented and automatically segmented.

The database used was the Chinese *Peoples Daily* collection containing more than 100 megabytes of text.

A second stage of the experiment tested relevance feedback on the same queries. Relevant documents were selected and two-character sequences common to the relevant documents were automatically added to the original query. The modified query was resubmitted to the system and the first ten documents returned were evaluated for relevance.

1.2.2 Evaluation Results

As the precision figures for the thirty queries in Table 1 show, even the unsegmented character-based queries give respectable results. On the average six out of the first ten documents will be relevant to a given query. Interpreted another way, the first document listed will be relevant in eight queries out of ten.

Hand segmentation requires the user to insert spaces between the Chinese words when entering the text of the query. As the table shows, this gives an average improvement in performance of about 10% over the unsegmented query. Automatic segmentation gives a similar increase in performance. The difference between the two segmentation methods is largely due the presence of proper names in the queries. Although we have developed a Chinese and foreign name recognizer, it was not used in the segmentation for this experiment. As a result names were interpreted as a series of characters.

The relevance feedback stage of the experiment was based on a bigram model, which means that a number of two-character sequences from the relevant documents were selected for query expansion. We have previously observed that two-character sequences perform much better than single-character selection in relevance feedback. It would also be possible to automatically segment the relevant documents for feedback analysis, but it is not clear that this method would produce a measurable difference within the parameters of this experiment.

As the table shows, relevance feedback gives a performance increase of 10-20%. Relevance feedback expands the original query, so the difference observed in the feedback experiment are due to the influence of the original segmented or unsegmented query terms.

1.2.3 Evaluation conclusions.

Within the limitation of the evaluation methods, we can conclude that the performance of Chinese INQUERY is quite satisfactory and conforms to that of INQUERY in other languages.

Based on work in English and Japanese, it is expected that a combination method, combining a word-based query with its character-based raw text, would perform best. Based on the quality of our bigram-based relevance feedback, we also intend to experiment with a bigram method of segmentation. This would be faster and simpler than lexicon-based segmentation.. If used in a combination query, it is possible that the results would equal or surpass the more expensive automatic segmentation performance.

.

.[Jing] Jing, Y.; Croft, W.B. (1994) An association thesaurus for information retrieval. *Proceedings of RIAO 94*, 146-160.

.

.

.