Experiments on Chinese Text Indexing
—CLARIT TREC-5 Chinese Track Report

Xiang Tong[1], ChengXiang Zhai[1], Nataša Milić-Frayling[2], and David A. Evans[2]

[1] Laboratory for Computational Linguistics
Carnegie Mellon University
Pittsburgh, PA 15213
{xt22@andrew.cmu.edu, cz25@andrew.cmu.edu}

[2] CLARITECH Corporation
5301 Fifth Ave.
Pittsburgh, PA 15232-2124
{natasa@clarit.com, dae@clarit.com}

1 Introduction

The focus of the CLARIT^{TM1} Chinese Track Experiments is on investigating the effectiveness of different automatic indexing methods for retrieval over Chinese texts. In particular, we explored indexing using linguistic units (words, compound words, and phrases), single Chinese characters, and overlapping character bigrams. In addition to fully automatic processing of queries, we ran experiments with manually constructed term vector queries supplemented by Boolean type constraints. The constraints were used for selecting documents for CLARIT automatic feedback or as a mean of refining the final set of retrieved documents [Milić-Frayling et al. 1997].

All the experiments were conducted using the CLARIT retrieval system [Evans & Lefferts 1995]. Since its current NLP component does not support the parsing of Chinese texts, we designed an appropriate parsing module and pre-processed the documents before submitting them for indexing and retrieval by the CLARIT system.

2 Chinese Text Indexing and Retrieval Experiments

In this section, we describe the two main indexing methods we explored: lexical term indexing and character N-gram indexing. The lexical term indexing method represents a linguistic approach while the N-gram indexing method is a purely statistical approach to indexing of Chinese texts. We also briefly discuss the general procedure and the design of our automatic and manual experiments.

2.1 Indexing by Lexical Terms

In contrast to English texts, Chinese texts have no obvious word boundaries. The text consists of a sequence of non-spaced ideographic characters, whose closest equivalents in English are morphemes. These morphemes normally are meaningful and represent words. But they can also be grouped together into multi-character units forming non-decomposable compound words. The problem of identifying words in Chinese text is very

[1] CLARIT is a registered trademark of CLARITECH Corporation

similar to that of identifying multi-word lexical items (or non-decomposable phrases) in English text, since the criteria for determining such English phrases is also not easy to specify.

We segment the Chinese text via a two-phase procedure using various lexical resources and heuristics. In the first phase, we segment the text into lexical units by applying

- lexical lookup using a Chinese lexicon that contains about 100,000 entries including words, compounds, and phrases extracted from several Chinese on-line dictionaries,
- a dynamic programming algorithm that optimizes the segmentation of a sentence into lexical words by searching for the segmentation with the minimum number of words,
- simple morphological rules, and
- heuristics about Chinese names, transliterations of foreign names, and other new compound words.

Portions of the segmented text that contain consecutive characters that are not part of any compounds or phrases are analyzed by applying heuristics for discovering new words. These items are collected into a list of unknown terms for further analysis and clean-up. After the whole corpus is processed, an item in the list is removed if it has a frequency lower than a pre-specified threshold or contains very frequent function words. This procedure when applied to the TREC Chinese data leads to a new word list of about 76,000 entries that are then incorporated into the original lexicon to form a new augmented lexicon with a total of approximately 176,000 entries.

In the second phase, the same dynamic algorithm and morphological rules are used to segment the texts into terms in the augmented lexicon. Moreover, all text segments with more than two characters are further decomposed into subcomponents that are valid words in the dictionary. For example, if $A, ..., G$ are Chinese characters and $ABCD$, AB, CD, DEF, DE, and G are valid words in the augmented lexicon, the sentence $ABCDEFG$ may be converted into $ABCD$<*space*>AB<*space*>CD<*space*>DEF<*space*>DE <*space*>G. All text segments and their valid subcomponents are used to represent the content of the text. One-character function words are not used for indexing, because they are highly frequent in the text and thus have low discrimination value.

After the text has been pre-processed by the text segmentation module and converted into lexical items separated by spaces, the converted text is submitted to the CLARIT indexer, which treats these lexical items as single terms.

In the Chinese automatic experiments, we constructed queries automatically, using all the parts of the topic description, after they were preprocessed by the same segmentation module that processed documents. The task involved 28 Chinese topics, each of which contains three parts: the title, topic description, and the narrative. The terms in each section were assigned a specific weight: terms in the title a weight of 2, terms in the description a wight of 4, and terms in the narrative a weight of 1.

In the manual runs, the automatically created queries were manually modified by adjusting the term weights and adding new terms. We also specified Boolean type constraints to select documents for automatic feedback. Documents that were retrieved in response to the initial query and satisfied the constraints were considered useful for automatic feedback. The constraints were written as a set of Boolean expressions using AND, OR, and NOT operators [Evans et al. 1996] [Milić-Frayling et al. 1997].

Manual modification of the queries took about 4 hours (5 min/query) and it was performed by a native speaker of Chinese. On average, 2–3 terms were added to a query.

2.2 Indexing with Overlapping Character N-grams

An alternative way to index Chinese text is to use character N-grams as indexing units. In our TREC-5 experiments we implemented and evaluated indexes based on unigrams (i.e., single characters) and overlapping character bigrams.

In our N-gram indexing approach, each sentence in the text is first converted into overlapping N-grams that are separated by spaces. For instance, indexing based on overlapping bigrams would convert a sequence of Chinese characters $ABCDEFG$ into a sequence of bigrams AB<*space*>BC<*space*>CD<*space*> DE<*space*>EF<*space*>FG. Text in this form is then processed by the CLARIT indexer.

Similar to the document text, queries are converted automatically into overlapping N-grams. The resulting character N-grams are weighted based on the section of the topic they originate from (see Section

2.1). We did not attempt any manual modification of the N-gram queries. Thus, all the runs that use this indexing method are considered TREC automatic runs.

3 Results and Analysis

The document collection used for the TREC-5 Chinese track consists of news articles from *People's Daily* and *Xinhua News Agency*. The size of the corpus is about 170 MB. We present the evaluation of our results based on the 19 topics that had been assessed at the time of TREC-5 meeting (November 1996).

The two official runs we submitted, CLCHNA and CLCHNM, use lexical term indexing. The difference between the two runs is in the construction of queries: CLCHNA uses automatically created queries, while CLCHNM uses manually modified queries.

We performed a number of experiments in our evaluation of different indexing methods. All the experiments with automatic feedback involve the query expansion based on the CLARIT thesaurus extraction technique [Evans & Lefferts 1995]. The results from the automatic runs using different indexing terms (characters, character bigrams, and lexical units) are shown in Table 1.

All manual runs were conducted over the database indexed by lexical terms. The results are presented in Table 2. In Figure 1 we show the precision-recall curves for all the runs.

Table 1: Automatic Run Results for Different Indexing Methods

Runs	Rcl(Rel Docs)	Ave Pre	Init Prec	R-Prec
Character-NoFeedback(baseline)	0.5318(744)	0.1235	0.3588	0.1502
Character-Feedback	0.3345(468)	0.0080	0.3133	0.1100
increase over baseline	-37.1%	-93%	-12.7%	-26.8%
Bigram-NoFeedback	0.9078(1270)	0.2541	0.6060	0.2877
increase over baseline	+70.7%	+100.5%	+68.9%	+91.5%
Bigram-Feedback	0.8549(1196)	0.2479	0.6535	0.2963
increase over baseline	+60.8%	+100.7%	+82.1%	+97.3%
Lexicon-NoFeedback	0.8670(1213)	0.2586	0.6674	0.2788
increase over baseline	+63.0%	+109.4%	+86.0%	+85.6%
Lexicon-Feedback(CLCHNA)	0.8449(1182)	0.2677	0.6223	0.2998
increase over baseline	+59.8%	+116.8%	+73.4%	+99.6%

Table 2: Manual Run Results

Runs	Rcl(Rel Docs)	Ave Prec	Init Prec	R-Prec
Manual-NoConstraint-NoFeedback	0.8863(1240)	0.3079	0.7758	0.3296
increase over baseline	+66.7%	+149.3%	+116.2%	+119.4%
Manual-NoConstraint-Feedback	0.84(1175)	0.2921	0.7613	0.3371
increase over baseline	+58.0%	+136.5%	+112.2%	+124.4%
Manual-Constraint-NoFeedback	0.8956(1253)	0.3855	0.8383	0.4122
increase over baseline	+68.4%	+212.1%	+133.6%	+174.4%
increase over NoConstraint-NoFeedback	+1.0%	+25.2%	+8.0%	+25.1%
Manual-Constraint-Feedback(CLCHNM)	0.8956(1253)	0.3583	0.7534	0.3872
increase over baseline	+68.4%	+190.0%	+110.0%	+157.8%
increase over NoConstraint-Feedback	+6.6%	+22.7%	-1.0%	+14.9%

From the collected performance statistics we can conclude that the results from the single character indexing are significantly worse than those from the experiments with other indexing methods. This indicates that Chinese characters have low discrimination value since they tend to be too general and ambiguous.

On the other hand, the retrieval performance based on overlapping character bigram indexing is comparable to that based on lexical term indexing. This suggests that the character bigrams are useful features in

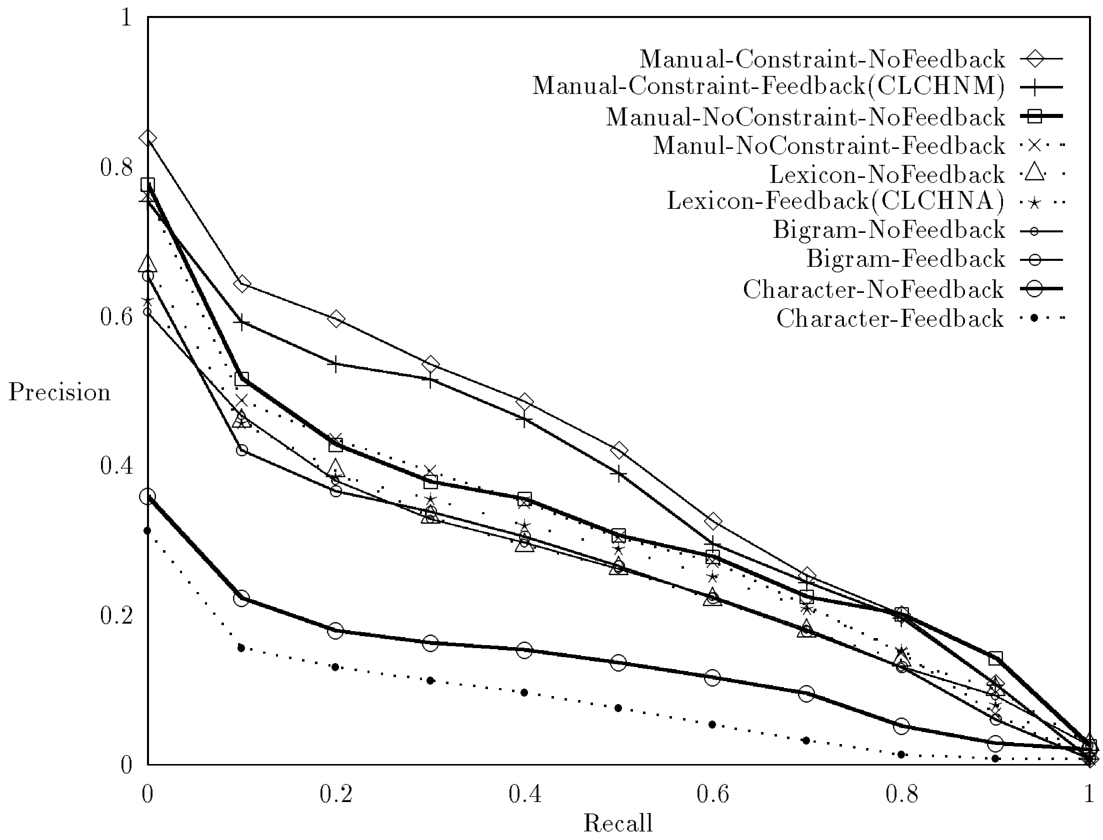

Figure 1: Precision-Recall Curves for All the Runs

capturing the patterns of Chinese texts, even though a large number of bigrams are not meaningful linguistic units.

Among the four manual runs, those that used Boolean constraints to control feedback achieved more than 20% improvement in average precision over the corresponding runs without constraints. Thus, Boolean constraints seem to be very effective in searching over Chinese texts.

The effect of automatic query augmentation based on the CLARIT thesaurus extraction technique is unclear at this point. Contrary to our past experience with retrieval over English texts, the feedback applied to manual runs hurts both the average and the initial precision. In fact, the best average and initial precision are achieved in the manual run using constraints and no feedback. We suspect that the present CLARIT thesaurus extraction technique is not quite appropriate for Chinese text. However, this issue needs to be explored further.

4 Conclusions

In the CLARIT TREC-5 Chinese experiments, we evaluated the effectiveness of different indexing methods for Chinese text retrieval. We found that indexing based only on single Chinese characters gives poor retrieval performance because the meaning of most Chinese characters tends to be general and ambiguous. On the other hand, indexing on overlapping character bigrams leads to results comparable to the retrieval based on lexical term indexing.

In the manual runs using lexical term indexing, Boolean type constraints proved to be very useful in specifying user information needs and therefore improve the retrieval performance.

In our future work we will continue investigating existing and new methods for automatic expansion of Chinese queries. Furthermore, we will explore the effectiveness of additional character N-gram indexing techniques.

References

[Evans & Lefferts 1995] Evans, D. A. and Lefferts, R. G. 1995. CLARIT-TREC Experiments. *Information Processing and Management*, Vol. 31, No. 3, 385-395.

[Evans et al. 1996] Evans, D. A., Milić-Frayling, N., Lefferts, R. G. 1996. CLARIT TREC-4 Experiments. In: Harman, D.(Ed.), *The Fourth Text REtrieval Conference (TREC-4).* NIST Special Publication 500-236, 1996, 305-321.

[Milić-Frayling et al. 1997] Milić-Frayling, N., Tong, X., Zhai, C., and Evans, D. A. 1997 CLARIT Compound Queries and Constraint-controlled Feedback in TREC-5 Ad Hoc Experiments. In: Harman, D.(Ed.), *The Fifth Text REtrieval Conference (TREC-5).* NIST Special Publication, 1997. (This volume.)