English-Chinese Information Retrieval at IBM

Martin Franz, J. Scott McCarley, Wei-Jing Zhu
IBM T.J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Abstract

We describe TREC-9 experiments with an IR system that incorporates statistical machine translation trained on sentence-aligned parallel corpora for both query translation (English⇒Chinese) and document translation (Chinese⇒English .) These systems are contrasted with monolingual Chinese retrieval and with query translation based on a widely available commercial machine translation package. These systems incorporate both words and characters as features for the retrieval. Comparisons with a baseline from TREC-5/6 enable our experiments to address issues related to the differences between Beijing and Hong Kong dialects.

1 Chinese preprocessing

The TREC-5/6 corpus is in the Taiwanese dialect of Chinese, and is encoded in the GB-2312 character set. The TREC-9 corpus consists of news stories from Hong Kong, and is encoded in the Big-5 character set. In order to perform comparable experiments on both corpora, we adopt UTF-8 encoded unicode as our internal representation of Chinese characters. In order to study baseline retrieval performance, we converted the TREC-5/6 Chinese track corpus from GB to unicode. We converted the TREC-9 corpus from Big-5 to unicode (ignoring the "extra" HKSAR hanzi.) We note that unicode often contains at different code points both the simplified and traditional forms of the same hanzi; the mappings relating the simplified and traditional forms, as well as other semantic variants within unicode are well-documented [1]. Any character that could be linked to a simplified Chinese character (including indirect linkings) was mapped to that character; simplified characters linked to each other were mapped to the smaller unicode number.

2 Chinese IR System Description

The Chinese IR track in TREC-5/6 triggered extensive experimentation on whether Chinese characters should be automatically tokenized ("segmented") into words to use as features for IR, or whether the characters themselves (and n-grams of characters) should be used as tokens for IR. No clear consensus has emerged [2, 3], see also [4, 5], although there are good reasons to prefer shorter words (limited to less than about 4 characters) [6] as well as to incorporate both types of features. Our approach to incorporating both words and characters is to build two separate systems, closely modeled on our English IR system [7], and to merge the results by linear combination of scores.

Both corpora were segmented with a statistical segmenter similar to the one discussed in [8]. The corpus-based iterative approach to Chinese segmentation allowed us to customize the segmenter's language model probabilities to each corpus. The segmenter's vocabulary consisted mostly of two-character words, with no words exceeding 5 characters, since there is evidence that short words are preferable (longer words often fail to match any terms in queries) for information retrieval purposes.

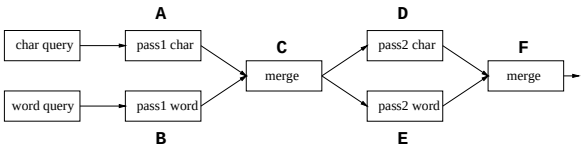

Figure 1: Diagram of system

Our Chinese (monolingual) IR system is a two-pass system, in which the results of the initial retrieval are used to construct an expanded query, which is then used for a second pass retrieval. The outline of the system is indicated in Fig. (1). For generality of explanation, we assume that the query has already been preprocessed into two forms, one in which it has been automatically been tokenized into short words for use as IR features, and one in which each character is a separate token. The first pass scoring is based on the Okapi formula [9], using the characters as features in (A), and using short-word tokens as features in (B). The results are merged at (C) by linear combination of scores. Our query expansion is based on LCA [10], which selects features from the top-ranked documents (output at (C)) which frequently cooccur with query features in these documents. At (D) we query-expand the character-based representation, i.e., we look for characters that frequently cooccur with query characters in the top-ranked documents. At (E) we query-expand the short-word-based representation of the corpus. Both query expansions are merged at (F) to yield the final results.

3 Crosslingual IR Experiments

3.1 Query translation with a statistical model

We used two parallel corpora (Hong Kong Laws and Hong Kong News, available from the Linguistic Data Consortium as part of the Topic Detection and Tracking (TDT) project [12]), and a smaller amount of material from the FBIS, to build a *character-based* statistical translation model. Because the majority of parallel text was from Hong Kong, we expect this translation model to be particularly well-matched to the TREC-9 test set, and less well suited for the TREC-5/6 baseline (in contrast to the commercial translation package described above.) We built a model of the probability $p(c|E, E_+, E_-)$ where c is a Chinese character, E is an English word, and E_+ and E_- are the nearest following and preceding content words. Models of this structure have previously been described in [13]. These models are trained from a sentence-aligned parallel corpus, together with a word alignment. The word alignment is constructed automatically from the parallel corpus using Poisson-fertility model, as described in [14, 15] This model predicts only characters, not the order of characters. The ordering was imposed when possible by dictionary lookup, using a Chinese-English dictionary made available for the TDT project. We note that ordering the characters was not necessary for the character-based aspect of IR, but was necessary in order to segment the query into words (and hence for the word-based aspect of IR.) Experiments with this model are denoted QT in the results.

3.2 Document Translation with a Statistical Model

Because of our prior success mixing query translation and document translation [16], we also built a Chinese⇒English translation model from the same parallel corpora as above. This model is not directly comparable to statistical query translation model - it is a word-based model for $p(E|C)$ the probability of an English (morphological root) word E given a Chinese word C (determined from an automatic segmentation of the corpus. This model is also a Poisson- fertility model. When the corpus was translated, the translation model was supplemented by the LDC dictionary. Since the resulting translation of the corpus is English, there is no character/word distinction in the IR system associated with this retrieval. Results with this model are denoted DT in the results.

3.3 Query translation using commercial software

Another set of experiments involved translating the English version of the query using a widely available commercial machine translatation package [11]. These experiments will be denoted TW in the results. Since this software was developed in Taiwan, we expected that it would be more closely matched to the TREC

5/6 baseline than to the the TREC-9 test set. The output of the translation package, which consists of unsegmented characters, is automatically segmented as into words and characters and then used in our Chinese IR system.

4 Discussion of Results

The character-based half of the system generally outperformed the word-based half of the system, across both types of Chinese⇒English translation, and monolingually, especially on the first pass of scoring. Query expansion made the differences between character- and word-based retrieval less clear. The gain from mixing character-based and word-based results was only slight. This result seems to be true for both the TREC-5/6 set and the TREC-9 set, and so is probably independent of dialect. On the other hand, dialect strongly influenced the relative behavior of the two query translation systems. The Taiwan-built commercial system, as expected, performed better on the TREC 5/6 task (Beijing data), whereas the statistical system, trained on Hong Kong data, performed better on the TREC-9 task (Hong Kong corpus.)

Our submission system was a merging of the TW, QT, and DT systems. However, the relative ranks of the TW, QT, and DT systems are completely reversed between TREC-5/6 and TREC-9, presumably mostly as a result of dialect differences. Thus TREC-5/6 was could not be used as a training set to predicting merging weights, etc. for TREC-9. However, in both sets (unlike many other IR tasks) the value of merging the results of different systems questionable.

5 Acknowledgements

This work is supported by DARPA under SPAWAR contract number N66001-99-2-8916. We would like to thank Salim Roukos and Todd Ward for valuable discussions.

References

[1] www.unicode.org/Public/MAPPINGS/EASTASIA/UNIHAN.TXT

[2] "Spanish and Chinese Document Retrieval in TREC-5", A.Smeaton and R. Wilkinson, in *Proceedings of the Fifth Text REtrieval Conference (TREC-5)* ed. by E.M. Voorhees and D.K. Harman. NIST Special Publication 500-238, 1997

[3] "Chinese Document Retrieval at TREC-6", R. Wilkinson, in *Proceedings of the Sixth Text REtrieval Conference (TREC-6)* ed. by E.M. Voorhees and D.K. Harman. NIST Special Publication 500-240, 1998

system	aveP TREC 5-6	aveP TREC 9
ML (c)	0.4783	0.2887
ML (w)	0.4813	0.3112
ML (c+w)	0.4904	0.2973
TW (c)	0.3152	0.1865
TW (w)	0.2689	0.2153
TW (c+w)	0.3193	0.2030
QT (c)	0.2778	0.2402
QT (w)	0.2228	0.1805
QT (c+w)	0.2810	0.2420
DT (w)	0.2889	0.2181
QT+DT	0.2979	0.2203
all 3	**0.3054**	**0.2258**

Table 1: Final scores by subsystem: ML = monolingual baseline, TW = query translation with commercial software QT = statistical query translation, DT = statistical document translation

[4] "On Chinese Text Retrieval" J.-Y. Nie, M. Brisebois, and X.Ren, in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Zurich, Switzerland, pp. 4-11.

[5] K.L. Kwok "Comparing Representations in Chinese Information Retrieval", in SIGIR 1997.

[6] K.L. Kwok "Lexicon Effects on Chinese Information Retrieval", in EMNLP, 1997

[7] M. Franz, J.S.McCarley, S.Roukos, "Ad hoc and Multilingual Information Retrieval at IBM", in *Proceedings of the Seventh Text REtrieval Conference (TREC-7)* ed. by E.M. Voorhees and D.K. Harman, 1999

[8] X.Luo and S. Roukos, "An Iterative Algorithm to Build Chinese Language Models", in *34th Annual Meeting of the Association for Computational Linguistics* Santa Cruz, CA, 1996.

[9] S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, M. Gatford, "Okapi at TREC-3" in *Proceedings of the Third Text REtrieval Conference (TREC-3)* ed. by D.K. Harman. NIST Special Publication 500-225, 1995.

[10] J. Xu and W. B. Croft 1996 Query Expansion Using Local and Global Document Analysis. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Zurich, Switzerland, pp. 4-11.

[11] TransWhiz C-E Translation PRO Version 3.0 package, by American Insight, Inc. (39-02 Main Street, 3F, Flushing, NY 11354)

[12] http://morph.ldc.upenn.edu/Projects/Chinese/

[13] J.S. McCarley and S.Roukos, "Fast Document Translation for Cross-Language Information Retrieval", in *Machine Translation and the Information Soup* ed. by D.Farwell, L.Gerber, and E.Hovy. (1998)

[14] P. F. Brown et al. "The mathematics of statistical machine translation: Parameter estimation", *Computational Lingustics*, 19 (2), 263-311, June 1993.

[15] S. Della Pietra, M. Epstein, S. Roukos, T. Ward "Fertility Models for Statistical Natural Language Understanding" *35th Annual Meeting of the Association for Computational Linguistics* Madrid, Spain, 1996.

[16] J.S. McCarley, "Should we Translate the Documents or the Queries in Cross-Language Information Retrieval?", in *37th Annual Meeting of the Association for Computational Linguistics* College Park, MD, 1999.