Comparison of Word-Based and Syllable-Based Retrieval for Tibetan

Paul G. Hackett and Douglas W. Oard

College of Information Studies, University of Maryland, College Park, MD 20742 U.S.A
Email: hackettp@cstone.net, oard@glue.umd.edu

Abstract

Tibetan retrieval based on automatically segmented words is compared with the use of overlapping syllable *n*-grams using a known-item retrieval evaluation. The optimal span of fixed-length *n*-grams is found to be 2 syllables, and indexing words is found to be as effective as indexing syllable bigrams.

Keywords: Tibetan; ranked retrieval; word segmentation; n-gram indexing.

1 Introduction

Research on information retrieval in Asian languages has increased dramatically in recent years, spurred in part by the increased ease with which text in electronic form can be generated, and shared. Widely spoken languages, such as Japanese, Mandarin Chinese, and Korean have received a great deal of attention, reflecting their important role in international commerce. Development of effective information retrieval systems for less widely spoken languages is also important, both as a way of improving access to existing information within each language and as a foundation for cross-language information retrieval systems that can provide global access to information regardless of language. This paper presents what we believe to be the first reported work on Tibetan information retrieval using automatic indexing.

Tibetan is a member of the Tibeto-Burmese language family. It is an alphabetic language, with phonetic characteristics that mirror those of Sanskrit. Tibetan has fixed rules for combining letters into syllables, resulting is approximately 81,000 possible syllables. The written form of the language contains explicit syllable delimiters and phrase delimiters, but no word delimiters; in this manner it is similar to Vietnamese. There are two broad approaches to retrieval for languages that lack word delimiters: techniques based on automatic segmentation, and techniques based on overlapping *n*-grams [Wilkenson, 1996]. In this paper we compare those approaches for Tibetan using a known-item retrieval methodology.

2 Experiment

We conducted a series of retrieval experiments to compare the performance of three variants of *n*-gram indexing (syllable unigrams, syllable bigrams, and syllable trigrams) and three variants of word-based indexing (automatically segmented words, automatically segmented word stems, and automatically segmented word stems with stopwords removed). Syllable *n*-grams spanning the boundaries of unambiguously marked Tibetan phrases were suppressed. Tibetan words were automatically segmented using a shallow parser that we built for Tibetan based on Wilson's formulation of Tibetan grammar [Wilson, 1992] to bracket morphemes, words, and phrases. We used a verb lexicon and a noun/adjective dictionary to guide the parsing process. The primary source for the verb lexicon was an augmented version of Wilson's verb classification tables, supplemented with other verb tables and glossaries. The resulting lexicon contained over 4,800 verbs. The noun/adjective dictionary was formed by combining two dictionaries and three glossaries, collapsing duplicate entries, and removing all collocations containing verbs. The parser's design hinges on sentence boundary detection and verb identification. Tibetan is a verb-final language, so a verb lexicon is used to recognize sentence boundaries. We used this as the starting point for shallow parsing of the input text, and then exploited syntactic constraints to guide the segmentation process. Greedy longest substring matching using a noun/adjective dictionary was applied to regions in which shallow parsing failed to produce a unique segmentation. Optional stemming was implemented by automatically removing embedded case marking particles from declined Tibetan words, and optional stopword removal was accomplished by deleting all case marking, lexical, and syntactic particles.

We used Release 4 of the Asian Classics Input Project (ACIP) Tibetan collection for our experiments [ACIP, 1998]. The collection consists entirely of religious literature dating from 200 CE ("Common Era") to 1990 CE,

and contains 2,120 Tibetan documents (41 million syllables, totaling approx. 200 MB of text in a standard ASCII transliteration). The document collection was randomly sampled to select candidate "known-item" documents. The title of each target document was then used as a basis for forming the associated query; synonyms and narrower terms were added to the query from a Tibetan presentation of the standard philosophical concept hierarchy. Details can be found in [Hackett, 2000].

We examined two aspects of system performance, retrieval effectiveness and index size, using a known-item retrieval evaluation paradigm that assessed the ability of the system to locate a specific document based on an incomplete description. A known-item retrieval evaluation simulates a user seeking a particular half-remembered document. Each query is associated with a single document, and retrieval effectiveness is assessed on the basis of the rank (first, second, third, …) assigned to that known item by the system when the query is processed. We constructed our queries in a manner similar to that used for the TREC Confusion and SDR tracks [Garofolo *et al.* 1998]. Inquery (3.1p1) was used for retrieval. The rank assigned to the target document of each of 39 queries was used to compute the mean inverse rank for each configuration. We chose the sign test as a measure of statistical significance because inverse rank produces unevenly quantized values (e.g., 1.0, 0.5, 0.33, …). We chose a 0.05 significance level as the criterion for rejection of the null hypothesis that the observed inverse rank values for a pair of runs were drawn from the same distribution.

The use of syllable bigrams (n=2) resulted in a mean inverse rank of 0.63, which can be thought of as typically placing the desired document in the first or second position of a ranked list. This result was statistically significantly better than that achieved using either syllable unigrams (0.24) or syllable trigrams (0.47). The mean inverse rank achieved using the most effective word-based technique (0.68, using words without stopword removal) was found to be statistically indistinguishable from that achieved using bigrams. Stemming (0.64) and stopword removal (0.62) seemed to produce small negative effects, but neither effect was statistically significant. No segmentation errors from ambiguous word boundaries were observed, and incorrect (premature) segmentations yielded results comparable to syllable bigrams. As has been observed in other languages [Miller *et al*., 2000], *n*-gram indexing resulted in explosive growth in the number of terms with increasing *n*. The index size for word-based indexing was less than one quarter of that of syllable bigrams.

3 Conclusions and Future Work

We have shown that large-scale Tibetan text retrieval is possible using fully automatic indexing. It is important to caveat our results by noting that they were obtained using a known-item retrieval evaluation. Carbonell et al. have shown (in the related context of "mate retrieval") that although poor known-item retrieval effectiveness implies poor effectiveness at more general *ad hoc* search tasks, the converse does not hold [Carbonell et al., 1997]. In other words, known-item retrieval evaluation is an inexpensive means for rejecting bad systems, but it does not discriminate well between good systems. Since no *ad hoc* evaluation collection exists for Tibetan, known-item retrieval evaluation provides a cost-effective way of exploring a broad range of candidate techniques. It is presently impractical to assemble TREC-style pooled relevance assessments for Tibetan because many systems must contribute to the pools before reliable results can be obtained. Precision-oriented evaluation metrics such as precision at 10 documents offer one alternative, but the resulting requirement to perform new relevance assessments for each run limits the opportunity for parameter optimization. Exhaustive relevance assessment does not suffer from such a limitation, but such an approach would be affordable only for relatively small collections. The nature of this evaluation challenge is, of course, not specific to Tibetan. Ultimately, we must reach agreement as a research community on widely accepted evaluation techniques for information retrieval in minority languages that are both insightful and affordable if we are to make progress on this important research challenge.

4 Acknowledgements

This work has been supported in part by DARPA contract N6600197C8540. The authors wish to thank Rebecca Green and Philip Resnik for their helpful comments.

5 References

1. ACIP, Asian Classics Input Project, *Release 4*, 1998.
2. Carbonell, J., Y. Yang, R. Frederking, R.D. Brown, Y. Geng, and D. Lee, Translingual Information Retrieval: A comparative evaluation. In *International Joint Conference on Artificial Intelligence*, 1997.
3. Garofolo, J., E.M. Voorhees, V.M. Stanford, and K. Sparck-Jones, TREC-6 1997 Spoken Document Retrieval Track Overview and Results. In *Proceedings of the Sixth Text Retrieval Conference*, Gaithersburg, 1998, pp.83-91. http://trec.nist.gov
4. Hackett, P.G., *Approaches to Tibetan Information Retrieval: Segmentation vs. n-grams.* Master's Thesis. College of Library and Information Services, University of Maryland, College Park, 2000. http://www.glue.umd.edu/~oard
5. Miller, E., D. Shen, J. Liu, and C. Nicholas, Performance and Scalability of a Large-scale N-gram Based Information Retrieval System. *Journal of Digital Information*, January 2000.

6. Wilkenson, R., Chinese Document Retrieval at TREC-6. In *Proceedings of the Sixth Text Retrieval Conference, Gaithersburg*, 1998, pp.25-30.

7. Wilson, Joe, *Translating Buddhism from Tibetan*. Ithaca: Snow Lion Publ. 1992.