IMPROVING CHINESE TOKENIZATION WITH LINGUISTIC FILTERS ON STATISTICAL LEXICAL ACQUISITION

Dekai Wu
Department of Computer Science
University of Science & Technology (*HKUST*)
Clear Water Bay, Hong Kong
dekai@cs.ust.hk

Pascale Fung
Computer Science Department
Columbia University
New York, NY 10027
pascale@cs.columbia.edu

Abstract

The first step in Chinese NLP is to tokenize or segment character sequences into words, since the text contains no word delimiters. Recent heavy activity in this area has shown the biggest stumbling block to be words that are absent from the lexicon, since successful tokenizers to date have been based on dictionary lookup (e.g., Chang & Chen 1993; Chiang *et al.* 1992; Lin *et al.* 1993; Wu & Tseng 1993; Sproat *et al.* 1994).

We present empirical evidence for four points concerning tokenization of Chinese text: (1) More rigorous "blind" evaluation methodology is needed to avoid inflated accuracy measurements; we introduce the *nk-blind* method. (2) The extent of the unknown-word problem is far more serious than generally thought, when tokenizing unrestricted texts in realistic domains. (3) Statistical lexical acquisition is a practical means to greatly improve tokenization accuracy with unknown words, reducing error rates as much as 32.0%. (4) When augmenting the lexicon, linguistic constraints can provide simple inexpensive filters yielding significantly better precision, reducing error rates as much as 49.4%.

HOW TO HANDLE DOUBLE STANDARDS

Current evaluation practice favors overly optimistic accuracy estimates. Because partially-tokenized words are usually evaluated as being correctly tokenized, failures to tokenize unknown words can be overlooked. For example, what makes 援助金 (yuán zhù jīn, a charity) a single word when 援助 and 金 are both legitimate words? One answer is that translating the partially-tokenized segments individually can yield "assistance gold" or "aid currency", instead of the unquestionably correct "charity" or "charity fund". Another answer is that a speech synthesizer should never pause between the two segments; otherwise 援助 is taken as a verb and 金 as a surname, changing the meaning to "help Gold". A blind evaluation paradigm is needed that accommodates disagreement between human judges, yet does not bias the judges to accept the computer's output too generously.

We have devised a procedure called *nk-blind* that uses n blind judges' standards. The n judges each hand-segment the test sentences independently, before the algorithm is run. Then, the algorithm's output is compared against the judges'; for each segment produced by the algorithm, the segment is considered to be a correct token if at least k of the n judges agree. Thus, more than one segmentation may be considered correct if we set k such that $k \leq \lfloor \frac{n}{2} \rfloor$. If k is set to 1, it is sufficient for any judge to sanction a segment. If $k = n$, all the judges must agree. Under the nk-blind method a precision rate can be given under any chosen (n, k) setting.

The experiments below were conducted with 100 pairs of sentences from the corpus containing between 2,000 and 2,600 words, sampled randomly with replacement. All results reported in Figure 1 give the precision rates for $n = 8$ judges with all values of k between 1 and n. Note the tendency of higher values of k to reduce precision estimates. The wide variance with different k (between 30% and 90%) underscores the importance of more rigorous evaluation methodology.

EXPERIMENT I

Tokenizing independently derived test data. The unknown word problem is now widely recognized, but we believe its severity is still greatly underestimated. As an "acid test", we tokenized a corpus that was derived completely independently of the dictionary that our tokenizer is based on. We used a statistical dictionary-based tokenizer designed to be representative of current tokenizing approaches, which chooses the segmentation that maximizes the product of the individual words' probabilities. The baseline dictionary used by the tokenizer is the BDC dictionary (BDC 1992), containing 89,346 unique orthographic forms. The text, drawn from the HKUST English-Chinese Parallel Bilingual Corpus (Wu 1994), consists of transcripts from the parliamentary proceedings of the Hong Kong Legislative Council. Thus, the text can be expected to contain many references to subjects outside the domains under consideration by our dictionary's lexicographers in Taiwan. Regional usage differences are also to be expected.

The results (see Figure 1) show accuracy rates far below the 90-99% range which is typically reported. Visual inspection of tokenized output showed that an overwhelming majority of the errors arose from missing dictionary entries. Tokenization performance on realistic unrestricted text is still seriously compromised.

EXPERIMENT II

Tokenization with statistical lexicon augmentation. To alleviate the unknown word problem, we next experimented with augmenting the tokenizer's dictionary using **CXtract**, a statistical tool that finds morpheme sequences likely to be Chinese words (Fung & Wu 1994). In the earlier work we found CXtract to be a good generator of previously unknown lexical entries, so overall token recall was expected to improve. However, it was not clear whether the gain would outweigh errors introduced by the illegitimate lexical entries that CXtract also produces.

The training corpus consisted of approximately 2 million Chinese characters drawn from the Chinese half of our bilingual corpus. The unsupervised training procedure is described in detail in Fung & Wu (1994). The training suggested 6,650 candidate lexical entries. Of these, 2,040 were already present in the dictionary, leaving 4,610 previously unknown new entries.

The same tokenization experiment was then run, using the augmented dictionary instead. The results shown in Figure 1 bear out our hypothesis that augmenting the lexicon with CXtract's statistically generated lexical entries would improve the overall precision, reducing error rates as much as 32.0% for $k = 2$.

EXPERIMENT III

Morphosyntactic filters for lexicon candidates. CXtract produces excellent recall but we wished to improve precision further. Ideally, the false candidates should be rejected by some automatic means, without eliminating valid lexical entries. To this end, we investigated a set of 34 simple filters based on linguistic principles. Space precludes a full listing; selected filters are discussed below.

Our filters can be extremely inexpensive because CXtract's statistical criteria are already tuned for high precision. The filtering process first segments the candidate using the original dictionary, to identify the component words. It then applies morphological and syntactic constraints to eliminate (a) sequences that should remain multiple segments and (b) ill-formed sequences.

Morphological constraints. The morphologically-based filters reject a hypothesized lexical entry if it matches any filtering pattern. The particular characters in these filters are usually classified either as morphological affixes, or as individual words. We reject any sequence with the affix on the wrong end (the special case of the genitive 的 (*de*) is considered below). Because morphemes such as the plural marker 們 (*mén*) or the instance marker 次 (*cì*) are suffixes, we can eliminate candidate sequences that begin with them. Similarly, we can reject sequences that end with the ordinal prefix 第 (*dì*) or the preverbial durative 在 (*zài*).

Filtering characters cannot be used if they are polysemous or homonymous and can participate in legitimate words in other uses. For example, the durative 著 (*zhe*) is not a good filter because the same character (with varying pronunciations) can be used to mean "apply", "trick", or "touch", among others.

Any candidate lexical entry is filtered if it contains the genitive/associative 的 (*de*). This includes, for example, both ill-formed boundary-crossing patterns like 的危險 (*de wéi xiǎn*, danger of), and phrases like 香港的前途 (*xiāng gǎng de qián tú*, Hong Kong's future) which should properly be segmented 香港 的 前途. In addition, because the compounding process does not involve two double-character words as frequently as other patterns, such sequences were rejected.

Closed-class syntactic constraints. The closed-class filters operate on two distinct principles. Sequences ending with strongly prenominal or preverbial words are rejected, as are sequences beginning with postnominals and postverbials. A majority of the filtering patterns match *correct* syntactic units, including prepositional, conjunctive, modal, adverbial, and verb phrases. The rationale for rejecting such sequences is that these closed-class words do not satisfy the criteria for being bound into compounds, and just co-occur with some sequences by chance because of their high frequency.

Results. The same tokenization experiment was run using the filtered augmented dictionary. The filters left 5,506 candidate lexical entries out of the original 6,650, of which 3,467 were previously unknown. Figure 1 shows significantly improved precision in every measurement except for a very slight drop with $k = 8$, with an error rate reduction of 49.4% at $k = 2$. Thus any loss in token recall due to the filters is outweighed by the gain in precision. This may be taken as indirect evidence that the loss in recall is not large.

CONCLUSION

We have introduced a blind evaluation method that accommodates multiple standards and gives some indication of how well algorithms' outputs match human preferences.

We have demonstrated that pure statistically-based lexical acquisition on the same corpus being tokenized can significantly reduce error rates due to unknown words. We also demonstrated empirically the effectiveness of simple morphosyntactic filters in improving the precision of a hybrid statistical/linguistic method for generating new lexical entries. Using linguistic knowledge to construct *filters* rather than generators has the advantage that applicability conditions do not need to be closely checked, since the training corpus presumably already adheres to any applicability conditions.

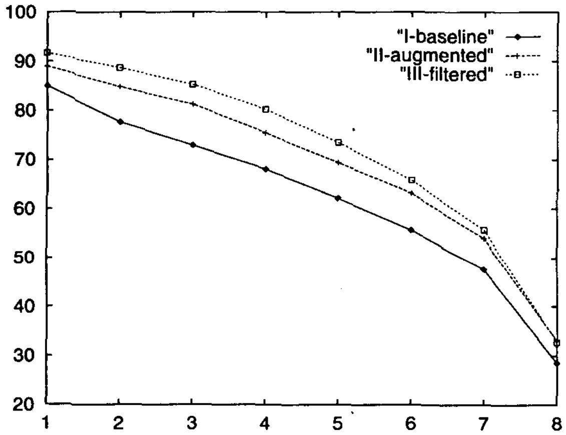

Figure 1. Comparison of nk-Blind Precision Percentages

REFERENCES

BDC. 1992. *The BDC Chinese-English electronic dictionary (version 2.0).* Behavior Design Corporation.

CHANG, CHAO-HUANG & CHENG-DER CHEN. 1993. HMM-based part-of-speech tagging for Chinese corpora. In *Proceedings of the Workshop on Very Large Corpora*, 40–47, Columbus, Ohio.

CHIANG, TUNG-HUI, JING-SHIN CHANG, MING-YU LIN, & KEH-YIH SU. 1992. Statistical models for word segmentation and unknown resolution. In *Proceedings of ROCLING-92*, 121–146.

FUNG, PASCALE & DEKAI WU. 1994. Statistical augmentation of a Chinese machine-readable dictionary. In *Proceedings of the Second Annual Workshop on Very Large Corpora*, 69–85, Kyoto.

LIN, MING-YU, TUNG-HUI CHIANG, & KEH-YIH SU. 1993. A preliminary study on unknown word problem in Chinese word segmentation. In *Proceedings of ROCLING-93*, 119–141.

SPROAT, RICHARD, CHILIN SHIH, WILLIAM GALE, & NANCY CHANG. 1994. A stochastic word segmentation algorithm for a Mandarin text-to-speech system. In *Proceedings of the 32nd Annual Conference of the Association for Computational Linguistics*, 66–72, Las Cruces, New Mexico.

WU, DEKAI. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd Annual Conference of the Association for Computational Linguistics*, 80–87, Las Cruces, New Mexico.

WU, ZIMIN & GWYNETH TSENG. 1993. Chinese text segmentation for text retrieval: Achievements and problems. *Journal of The American Society for Information Science*, 44(9):532–542.