Towards a Sharable and Reusable Lexical List: The Construction of a Standard Reference Lexicon for Chinese NLP

Chu-Ren Huang, Zhao-ming Gao, Claude C.C. Shen, and Keh-jiann Chen, Academia Sinica

A lexicon is a sharable natural language resource that is information-rich and sensitive to variations. A lexicon will change across time, topic domain, register, speaker, and age. A lexicon is also a knowledge base that both human understanding and computer-processing must refer too. No information-preserving sharing of natural language resources is possible unless the same standard lexicon is referred to.

The construction of a standard reference lexicon for Chinese is even more complicated than for English since there is no conventionalized wordbreaks in Chinese to identify new words in a text. Hence the selection of lexical entries involves both the definition of a word and a principled way to screen candidate entries. We argue that such reference lexicons must be judged by their cross-domain portability, expressive adequacy, and reusability. Thus principles for lexical selection must also be driven by these criteria.

A sharable lexicon must have high cross-domain portability. This cannot be achieved by simply including entries from all possible domains. Such exhaustive approach will not only entail high ambiguity but also generate high a noise level (i.e. high probability of false word-matches). Hence the foundation of a sharable lexicon is a core lexicon that is frequently used in all domains. To acquire this core lexical list, we examined both a balanced corpus (the 5 million word Sinica Corpus) and expert-compiled MRD's. We define our core lexicon as consisting of the entries whose frequency is greater than 10 in Sinica Corpus and which are included in all 5 authoritative MRD's that are compiled for different Chinese speaking communities. Tests show that these entries also occur across domains and genres.

A reference lexicon must be expressively adequate. Based on the authoritative Chinese Thesaurus of Tongyici Cilin, we showed conceptually basic terms are not necessarily frequent. Actually, conceptual primes distribute according to Zipf's Law, just like all terms. To make sure that all concepts can be expressed, conceptual primes must also included in the lexicon regardless of its frequency in a lexicon. Thus our standard lexicon is supplemented with the conceptual primes from Cilin.

The reusability of the lexicon depends not only on the highly portable core lexical list but also on how easy it is to supplement for specific domain usages. Since our lexicon is corpus-based in essence and since it is expressively adequate with the inclusion of all conceptual primes, it can be easily adapted for domain-specific use by including entries statistically extracted from a special domain corpus. As an illustration, we apply our lexicon to a general reading domain by adding entries from the Sinica Corpus whcih are more frequent than the threshold of 5 but do not qualify for the core lexicon.

1. Introduction

Since words are not conventionally marked in Chinese texts, segmentation is a pre-requisite step for Chinese NLP and setting a standard to define and measure segmentation results becomes necessary both for evaluation and for resource-sharing. However, as noted by standard-setters both in Mainland China (Liu et al. 1994) and in Taiwan (Huang et al. 1996), no segmentation standard can be successfully implemented and evaluated until it is accompanied by a wide-coverage reference lexicon. Huang et al. (1996) argued that segmentation standards must include a sharable and adaptable lexicon in order to apply across variations such as domain, genre, and time. On the other hand, for data-sharing, a standard lexicon is essential to ensure that texts from different sources can be uniformly tokenized. Thus, there is a consensus that an empirically compiled reference lexicon is an indispensible part of a Chinese segmentation standard (e.g. Lin and Miao 1997,

Sun and Zhang 1997). An additional benefit of such a lexicon is that it can be shared without much additional cost and thus saves NLP researchers the time and cost in building this essential infrastructure. To serve the above dual purposes, this standard lexicon must be selected in a principled way in order to best test validity and usefulness.

However, even though computational lexicography offers a rich literature on the structure and content of a lexical entry, there is hardly any discussion of a principled way of lexical entry selection (Armstrong-Warwick 1995). We only see discussion in the context of a terminology lexicon (Nagao 1994) or a reference segmentation lexicon (Liu et al. 1992). Both assume that the lexicon is built from scratch. We suggest that there are three criteria to judge the merit of a lexicon: reusability, expressive adequacy, and domain-portablitity. Based on these three criteria, we propose a principled way to construct a standard reference lexicon for Chinese NLP.

2. Word, Segmentation Unit and Lexical Entry

Determining whether a string is a (new) word is trivial in many languages such as English. However, it is not easy in Chinese because of the lack of conventional demarcation and of native speakers' consensus of what is a word. Once a string is identified as a unit, a further decision needs to made as to whether it should be listed in the lexicon (e.g. Wang et al. 1994).

In this study, we stipulate that all entries must be segmentation units defined in Huang et al. (1996, 1997). Notice that even though Huang et al. propose to take the notion of linguistic word as the theoretical foundation of the the definition of segmentation unit, it is obvious that certain non-words, such as (derivational) affixes, must also be treated as segmentation units. Thus they must be listed in the reference lexicon for segmentation.

The motivation of such a stipulation is two fold. First, it ensures uniformity of the segmentation criteria and the reference database

within the segmentation standard. Second, this allows the reference lexicon to list non-words such as derivational affixes, and thus will provide crucial information to account for the productive morpho-lexical processes.

3. Reusability: Corpus Base of Lexical Selection

That a corpus is the best source of lexical entries has been the cornerstone of recent developments of corpus linguistics (e.g. Sinclair 1987). Making a balanced corpus as the basis of a standard reference lexicon also makes it possible to automatically update the lexicon for different domain or for language changes. Either a monitor corpus will be maintained to indicate any change in the language, or a comparable corpora from separate domains can be maintained, and new entries can be extracted by the same automatic procedure to augment and revise the standard set.

Our current lexicon is based on the Sinica Corpus (Chen et al. 1996), a tagged balanced corpus of Taiwan Mandarin Chinese containing 5 million words. 146,876 different words appear in the corpus. We are current testing the entry sets defined by different frequency threshold to determine the optimal one, presumably between 5 (45,443 entries) and 10 (28,564 entries).

4. Expressive Adequacy: Conceptual Primes and Lexical Selection

Selecting lexical entries by frequency threshold based on corpus calculation is a dependable way to ensure relatively high coverage of the lexicon. However, since lexical information is not available in NLP unless it is encoded in the lexicon, high coverage does not necessarily translate into successful application if conceptually crucial items are missing. Thus, we propose that a standard reference lexicon must achieve expressive adequacy. Our hypothesis is that such adequacy can be ensured when entries representing conceptual primes are exhaustively included.

The conceptual primes that we adopt are the 3,922 covering terms of *Tongyici Cilin* (Mei et al. 1983, CILIN hereafter), the most widely used thesaurus in Chinese NLP. We treat them as if they are covering terms in semantic fields, assuming these terms alone will be adequate to express concepts represented by embedded terms in their field. Thus a lexicon containing all these terms will be expressively adequate.

A possible objection to adopting such an heuristic method independent of corpus-based stochastic approaches is that the same goal could be achieved without the heuristic. In other words, is there any evidence to prove that these conceptual primes cannot be satisfactorily extracted from corpora?

Diagram 1 shows the frequency/rank correlation of the CILIN conceptual primes based on their occurrences in Sinica Corpus. If conceptual primes were to be reliably extracted from corpora, they must fall (almost) exclusively in mid to high frequency rank. However, diagram 1 follows Zipf's law. In other words, these conceptual primes are as widely distributed as other lexemes. Any corpus-based frequency threshold will unfortunately exclude the lower frequency conceptual primes.

In fact, only 3,501 of the CILIN covering terms occur in Sinica Corpus, meaning that 421 terms are missing. These missing terms cannot be attributed solely to the lexical difference between Mainland China and Taiwan. Two authoritative dictionaries that consulted corpus extensively also do not enter all the CILIN primes. The 57,624 entry *Xiandaihanyu Cidian* (XHCD hereafter) lacks 241 of them while the 39,025 entry Segmentation Standard Lexical List (Liu et al. 1994, GB hereafter) misses 546 of them. There does seem to be a correlation between the degree of human intervention with the completeness of conceptual primes though. XHCD is compiled by linguists who consulted corpora, while GB is extracted from a corpus and augmented with thought-up lexical items.

In diagram 2, an addition test is conducted on the distribution of these conceptual primes. The diagram shows the number of conceptual primes per every 1,000 words in Sinica Corpus ordered according to frequency rank. As suspected, a high proportion of the most frequent words are conceptual primes (382 of the first 1,000), while the proportion descends dramatically. The diagram shows the slope smoothes at around rank 1,500 and levels well before rank 10,000.

Two important pieces of information can be inferred from diagram 2. First, it offers an intuitive support of the reliability of the CILIN primes. Since conceptual primes are the most economic (and often necessary) way to express ideas, they are more likely to be frequently used. Thus, we expect a valid set conceptual primes to be dominated by high frequency words. The CILIN distribution confirms such prediction.

Second, the steep descend and quick leveling suggests that it will be impractical to discover conceptual primes with pure stochastic approach. Since these conceptually primary terms are sparsely distributed in mid to lower frequency range, it would be quite impossible to achieve any reasonable recall and precision at the same time. In other words, for the moment at least, conceptual primes must be acquired independent of a corpus.

5. Portability: Bootstrapping with Existing Lexicons

It is impossible for a corpus, with finite total words, to cover all possible topics, genre etc. Hence it is most likely that some significant lexemes are not represented in a corpus. In other words, how can a standard lexicon be portable among all domains given the fact that the corpus it based on does not contain texts from all possible domains? This problem could be aggravated if a corpus is relatively small and geographically restricted.

The case is even worse for a Chinese lexicon because of the fact that there exist substantial lexical differences between Mainland and Taiwan Mandarin. Thus it would be futile to construct a

corpus that could represent both dialects. However, it is also well-known that mutual lexical borrowings are easy and frequent because both sides of the Taiwan Strait speak the same language and because of increasingly frequent contacts. Thus any purely Taiwan or Mainland corpus faces the dilemma of under-representing a critical segment of lexemes.

To solve this dilemma, we propose to bootstrap with existing lexicons. We will compare the entries of two Mainland lexicons: GB and XHCD, two Taiwan lexicons: the CKIP electronic lexicon (over 80,000 entries, CKIP 1995), and the Ministry of Education on-line dictionary (over 120,000), as well as the ABC Chinese-English Dictionary by DeFrancis (roughly 60,000 entries). Such a method allows us to tap existing knowledge and human-intensive reources. Our rationale is that any lexemes listed in at least three of the above dictionaries are generally accepted by speakers and must be included in a standard reference lexicon. This rationale is supported by our pilot studies, which shows that the intersecting sets of any two dictionaries are much smaller than the original and seem to genuinely represent core uses. For instance, there are only 28,728 common entries between GB and XHCD, which is 73.61% of the smaller GB lexicon.

6. Summary: the Proposed Methodology Towards a Standard Reference Lexicon

To meet the criteria of reusability, expressive adequacy, and cross-domain portability, we combine a three step algorithm for constructing a standard reference lexicon for Chinese NLP. First, lexical entries are automatically extracted from a balanced tagged corpus if their frequencies are higher than a stochastically determined threshold. The corpus-based generation allows automatic updating and adaptation to specific domains. Second, the automatically generated lexicon is augmented with a small set of conceptual primes to ensure expressive adequacy. Last, it is further augmented with entries obtained from intersection of 5 lexicons from different sources to ensure cross-domain portability.

The expected outcome of this study is a comprehensive reference lexicon that will be part of the segmentation standard for Chinese NLP in Taiwan. The standard lexicon will contain 60,000 to 100,000 entries. There will also be a 30,000 set of core entries that will be highly portable regardless of topic, genre, etc.

7. Verification and Expendability

To verify that our standard reference lexicon does meet the requirements set out by the three criteria, we will do both inside and outside tests. Tests are performed with an automatic segmentation procedure to determine coverage of the lexicon of all words appearing in their language. Inside tests will be performed on texts extracted from Sinica Corpus, which are marked with topic, genre, style, media etc. Our aim will be to ensure that consistently high coverage is achieved across all possible variations. Outside tests will be performed with texts not included in Sinica Corpus, especially texts from Mainland China as well as texts extracted from WWW.

This standard reference lexicon will be periodically updated and maintained to enure wide applicability in Chinese NLP as well as successful implementation of the segmentation standard that it supports. It is proposed that updates will be performed every 3 to 5 years based on new corpus data as well as revisions of the lexical databases originally consulted.

Bibliography

Armstrong-Warwick, S. 1995. Automated Lexical Resources in Europe: A Survey. In D.E. Walker, A. Zampolli, and N. Calzolari Eds. Automating the Lexicon. 397-403. Oxford: Oxford U. Press.

Chen, K.-j., C.-R. Huang, L.-P. Chang, and H.-L. Hsu. 1996. SINICA CORPUS: Design

Methodology for Balanced Corpora. In B.-S. Park and J.-B. Kim Eds. Language, Information, and Computation. Selected Papers from the 11th PACLIC. Seoul: Kynung Hee U.

Chinese Academy of Social Sciences. 1996. Xiandaihanyu Cidian [A Dictionary of Contemporary Chinese (Revised Edition)]. Beijing: Shangwu.

Chinese Knowledge Information Processing Group. 1996. ShouWen JieZi - A Study of Chinese Word Boundaries and Segmentation Standard for Information Processing [In Chinese]. CKIP Technical Report 96-01. Taipei: Academia Sinica.

Huang, C.-R., 1995. The Grammatical Categories of Mandarin Chinese.[in Chinese] CKIP Technical Report 95-03. Taipei: Academia Sinica.

Huang, C.-R., K.-j. Chen. F.-y. Chen, W.-J. Wei, and L. Chang. 1997. The Design Criteria and Content of the Segmentation Standard for Chinese Information Processing [in Chinese]. Yuyan Wenzi Yingyong. 1997.1.92-100.

Huang, C.-R., K.-j. Chen and L. Chang. 1996. Segmentation Standard for Chinese Natural Language Processing. COLING-96. 1045-48.

Lin, X.G., and C.J. Miao. 1997. Guifan+Cibiao yu Jinyen+Tongji. Yuyan Wenzi Yingyong. 1997.1.87-91.

Liu, Y., Q. Tan, and X. Shen. 1994. Segmentation Standard for Modern Chinese Information Processing and Automatic Segmentation Methodology.[in Chinese] Beijing: Qinghua U. Press.

Liu, Y., N. Liang, and Q. Tan. 1991. Lexical Selection Criteria for `A Lexicon of Frequent Modern Mandarin Words for Information Processing'. Proceedings of the Tenth Anniversary of Chinese Information Society of China. 127-141.

Mei, J., Y. Zhu, Y. Gao, and H. Yin. 1983. Tongyici Cilin. Shanghai: Shangwu Press and Shanghai Dictionaries.

Nagao, M. 1994. A Methodology for the Construction of a Terminology Dictionary. In B.T.S. Atkins and A. Zampolli Eds. Computational Approaches to the Lexicon. 397-412. Oxford: Oxford U. Press.

Sinclair, J. M. 1987. Ed. Looking Up--An account of the COBUILD Project in Lexical Computing. London: Collins.

Sproat, R. 1992. Morphology and Computation. Cambridge: MIT Press.

Sun, M.S., and L. Zhang. 1997. Renjibingcun, Zhiliangheyi - tantan zhiding xinxi chuliyong hanyu cibiao de celue. Yuyan Wenzi Yingyong. 1997.1.79-86.

Wang, M.-C., C.-R. Huang, and K.-j. Chen. 1995. The Identification and Classification of Unknown Words in Chinese: A N-gram-Based Approach. In A. Ishikawa and Y. Nitta Eds. The Proceedings of the 1994 Kyoto Conference. A Festschrift for Professor Akira Ikeya. 113-123. Tokyo: The Logico-Linguistics Society of Japan.

Zhang, Y. and X. Qi. 1997. The Statistic[s] and Analysis of Words Included in Several Chinese Dictionaries.[In Chinese] In L.W. Chen and Q. Yuan Eds. Language Engineering. 82-87. Beijing: Qinghua U. Press.

ıgram 1. CILIN entries frequency distribution in Sinica Corpus (Zipf's Law)

$Y = log\ f,\ X = rank$

ıgram 2. CILIN entries distribution by frequency range in Sinica Corpus
(number of CILIN entries per every 1,000 rank interval)