Segmentation Standard for Chinese Natural Language Processing

Chu-Ren Huang[*], **Keh-jiann Chen**[#], **and Li-Li Chang**[#]
[*]Institute of History and Philology, and [#]Institute of Information Science
Academia Sinica, Nankang, Taipei, Taiwan 115
hschuren@ccvax.sinica.edu.tw, kchen@iis.sinica.edu.tw, lili@iis.sinica.edu.tw

Abstract

This paper proposes a segmentation standard for Chinese natural language processing. The standard is proposed to achieve linguistic felicity, computational feasibility, and data uniformity. Linguistic felicity is maintained by defining a segmentation unit to be equivalent to the theoretical definition of word, and by providing a set of segmentation principles that are equivalent to a functional definition of a word. Computational feasibility is ensured by the fact that the above functional definitions are procedural in nature and can be converted to segmentation algorithms, as well as by the implementable heuristic guidelines which deal with specific linguistic categories. Data uniformity is achieved by stratification of the standard itself and by defining a standard lexicon as part of the segmentation standard.

I. Introduction

One important feature of Chinese texts is that they are character-based, not word-based. Each Chinese character stands for one phonological syllable and in most cases represents a morpheme. The fact that Chinese writing does not mark word boundaries poses the unique question of word segmentation in Chinese computational linguistics (e.g. Sproat and Shih 1990, and Chen and Liu 1992). Since words are the linguistically significant basic elements that are entered in the lexicon and manipulated by grammar rules, no language processing can be done unless words are identified. In theoretical terms, the primacy of the concept of word can be more firmly established if its existence can be empirically supported in a language that does not mark it conventionally in texts (e.g. Bates et al. 1993, Huang et al. 1993). In computational terms, no serious Chinese language processing can be done without segmentation. No efficient sharing of electronic resources or computational tools is possible unless segmentation can be standardized. Evaluation, and thus comparisons and improvements, are also impossible in Chinese computational linguistics without standardized segmentation.

Since the proposed segmentation standard is intended for Chinese natural language processing, it is very important that it reflects linguistic reality as well as computational applicability. Hence we stipulate that the proposed standard must be linguistically felicitous, computationally feasible, and must ensure data uniformity.

I.1.Components of the Segmentation Standard

Our proposed segmentation standard consists of two major components to meet the goals discussed above. The modularization of the components will facilitate revisions and maintenance in the future. The two major components of the segmentation standards are the segmentation criteria and the (standard) lexicon. The tripartite segmentation criteria consist of a definition of the segmentation unit, two segmentation principles, and a set of heuristic guidelines. The segmentation lexicon contains a list of Mandarin Chinese words and other linguistic units that the heuristic guidelines must refer to.

II.Segmentation Standard Part I: Segmentation Criteria

II.1. A Definition of the Segmentation Unit

Given Bloomfield's (1933) definition of words as 'the smallest units of speech that can meaningfully stand by their own,' they are natural units for segmentation in language processing. However, as Chao (1968) observes, sociological words and linguistic words very often do not match up. In English, a sociological word can be defined by the delimitation of blanks in writing. It is nevertheless possible for a linguistic word such as a compound to be composed of more than one sociological words, such as 'the White House.' Since these cases represent only a relatively small portion of English texts, sociological words are taken as the default standard for segmentation units as well as a reasonable approximation to linguistic words in English language processing.

Chinese, on the other hand, defines its sociological words in terms of characters, in spite of the fact that grammatical words may be made up of one or more characters. In fact, one-character words represent slightly less than 10% of all lexical entries, while two-character words take up more than 65%. Similarly, one-character words are estimated to take up only 50% of all texts in Chinese (Chen et al., 1993). Since the notion of the one-word-per-character sociological word is not a good working hypothesis for linguistic words, and since there is no fixed length for words, a crucial issue is whether the notion of linguistic words can be directly used as the standard for segmentation unit.

Computational linguistic works suggest that linguistic words are not the perfect units for natural language processing. For instance, the necessity for lemmatization attests to the fact that some linguistically dependent units may have independent grammatical function and meaning and need to be treated as basic units in language processing (e.g. Sproat 1992). We follow the above findings and define the standard segmentation unit as a close approximation of linguistic words with emphasis on functional rather than phonological or morphological independency.

1) **Segmentation Unit$_{def}$** is the smallest string of character(s) that has both an independent meaning and a fixed grammatical category.

There are two points worth remarking involving the above definition. First, nontechnical terms are deliberately chosen such that even developers in information industries with little or no linguistic background could follow this standard. Second, it follows from this definition that many of the so-called particles, which show various levels of linguistic dependencies but represent invariant grammatical functions, will be treated as segmentation units. They include *le* 'perfective marker', and *de* 'relative clause marker'.

II. 2. Segmentation Principles

We propose two segmentation principles to define the two basic concepts underlining the definition: independent meaning and fixed grammatical category. The principles also provide a functional/procedural algorithm for identifying segmentation units.

2) **Segmentation Principles**

a) A string whose meaning cannot be derived by the sum of its components should be treated as a segmentation unit.

b) A string whose structural composition is not determined by the grammatical requirements of its components, or a string which has a grammatical category other than the one predicted by its structural composition should be treated as a segmentation unit.

Take note that characters are the basic processing units when segmentation is involved. Thus the two principles address the question of which strings of characters can be further combined to form a segmentation unit. Principles 2a) and b) elaborate on the semantic (independent meaning) and syntactic (fixed category) components of the definition of segmentation unit.

Because of the procedural nature of the two principles, they provide the basis for segmentation algorithm. Since a character could be a lexical or sub-lexical element, the basic decision in segmentation is whether the relation between two characters are morpholexical or syntactic. For instance, with a VO sequence such as *lai-dian* come-electricity 'to strike a chord with, to mutually attract', principle 2b) applies to predict that the string is a segmentation unit since *lai* is an intransitive verb and do not take an object.

II.3.Segmentation Guidelines

Even though the above principled definition provides a broad direction for standardized segmentation, it lacks the nuance for guiding actual segmentation. The definition of segmentation units and the segmentation principles are essentially language independent formalizations of information units (i.e. words). Thus they will not vary with linguistic change, and need not be revised for specific applications. However, this universal nature also prevents them from referring to specific details. Hence we propose that a set of Segmentation Guidelines be included in our segmentation standard to reflect heuristic knowledge that is dependent on actual linguistic data. These guidelines can be added, deleted, or altered as necessitated by the linguistic data we are dealing with. Since all essential linguistic knowledge is encoded in the lexicon, it follows that the guidelines will have to refer to a Mandarin lexicon. In contrast, the broad linguistic concepts in the definition and principles do not refer to specific lexical information. Last, we also envision that the guidelines are quantifiable. They are quantifiable because more guidelines a string satisfies, the more likely it is to be a segmentation unit.

3) **Segmentation Guidelines**

a) Bound morphemes should be attached to neighboring words to form a segmentation unit when possible.

b) A string of characters that has a high frequency in the language or high cooccurrence frequency among the components should be treated as a segmentation unit when possible.

c) Strings separated by overt segmentation markers should be segmented.

d) Strings with complex internal structures should be segmented when possible.

III. Segmentation Standard Part II: The Standard Lexicon

We propose that a standard lexicon be included as part of the segmentation standard. This lexicon will list words as well as productive morpho-lexical affixes. It will also contain the list of mandatory segmentation markers, such as the end of sentence marker (.), ($_o$) etc. All derived words can be covered simply by firing all derivational rules governing the list of bound morphemes. However, non-derived words are trickier since they cannot be predicted with generative rules. The only way to verify that they are segmentation units is to consult a lexical list, which is finite and incomplete by nature.

The incompleteness of the lexical list underlines the importance of conforming to the segmentation criteria while compiling the standard lexicon. An entry is entered in the lexicon only when it qualifies as a segmentation unit. The segmentation guidelines 3a)-3c) are the same heuristic guidelines for selecting lexical entries. However, since all language lexicons are constantly changing, an entry in the lexicon is determined by its frequency and usage of the time. The standard lexicon will be updated and maintained regularly to keep up with the evolution of the language. In addition, application of the segmentation standard in any specific domain may require a new special domain lexicon.

IV. Three Levels of Segmentation Standard

A central concern in proposing any standard is whether this standard can be successfully and consistently followed. We took into consideration of the state of art of automatic segmentation in Chinese NLP as well as the technology level of information industries dealing with Chinese natural languages and proposed the following stratification of three levels of instantiations for the Segmentation Standard. It is hoped that this stratification will ensure successful standardization as well as lead to eventual identification of segmentation units with linguistics words.

5) **Levels of Segmentation Standard**

a) **Faithful**[*xin4*]: All segmentation units listed in the standard lexicon should be successfully segmented.

b) **Truthful**[*da2*]: All segmentation units identified at the Faithful level as well as all segmentation units derivable by morphological rules should be successfully segmented.

c) **Graceful**[*ya3*]: All linguistic words are successfully identified as segmentation units.

The goal of the Faithful level is to define a segmentation standard such that uniformity of electronic texts can be achieved even when they are prepared with the lowest possible computational sophistication. In other words, the standard must be as easy to follow as the convention of inserting blanks at wordbreaks in English text processing. At this level, unless it matches a lexical entry, a string will simply be segmented into individual characters. Notice that this is NOT a trivial level since possible ambiguous segments take up as high as 25% of Chinese texts (Chen and Liu 1992). Various automatic segmentation programs reported over 99% precision rate when unknown words are not taken into account (e.g., Chiang et al. 1992). This will be the default segmentation level for the exchange of electronic texts.

The goal of the Truthful level is to define a segmentation standard for most computational linguistic applications. The coverage of the Faithful level is too low for most NLP applications. For instance, unknown words can be left unidentified for data exchange but not for machine translation. Wang et al. (1995) classified unknown words into three types. The first type are the words that are generated by morphological rules. They are productive and cannot be exhaustively listed in the lexicon. The second type are the words whose derivation is either context-dependent or cannot be captured by familiar morphological rules. A good example is the *suoxie* abbreviation where a character from each compound or phrase component is selected to form a new word (Huang et al. 1993), such as deriving *hua2hang2* from *zhong1hua2 hang2kong1* 'China Airlines.' The third type are the unknown words which are not derived by any rules, such as proper names(Chen et al. 1994). Only the first type of unknown words can be comfortably dealt with by current Chinese NLP technology. Thus, at the Truthful level of segmentation, we stipulate that all lexical entries as well as all morphologically derivable unknown words should be properly segmented. This level will offer a wide enough coverage for most NLP applications and yet a reasonably high consistency can still be achieved with current automatic segmentation technology. Since a finite state machine implementing the morphological rules on top of a finite lexicon listing can generate all the segmentation units, the only technical challenge would be to resolve ambiguities among the above units.

Lastly, the Graceful level of segmentation standard will have to deal with the two remaining types of unknown words. It may not be too long before reasonable consistency can be achieved at this level of standard for fully automated language understanding.

V. Concluding Remarks

In this paper, we propose a Segmentation Standard for Chinese language processing composed of two distinct parts: a) the language and lexicon-independent definition and principles, and b) the lexicon-dependent guidelines. The definition and principles offer the conceptual basis of segmentation and are the unifying idea behind resolution of heuristic conflicts. The lexicon-dependent guidelines, as well as the data-dependent lexicon, allows the standard to be easily adaptable to linguistic and sub-language changes.

Bibliography

Bates, E., S. Chen, P. Li, M. Opie, O. Tzeng. 1993. Where is the Boundary between Compounds and Phrases in Chinese? Brain and Language. 45:94-107.

Bloomfield, L. 1933. Language. New York: Holt, Rinehart, and Winston.

Chao, Y. R. 1968. A Grammar of Spoken Chinese. Berkeley: U. of California Press.

Chen, C., S. Tseng, C.-R. Huang and K.-J. Chen. 1993. Some Distributional Properties of Mandarin Chinese-A Study Based on the Academia Sinica Corpus. Proc. of the 1st PACFoCoL. 81-95. Taipei.

Chen, H. and C. Li. 1994. Recognition of Text-based Organization Names in Chinese. [in Chinese] Communications of COLIPS. 4.2.131-142.

Chen, K.-J. and S.-H. Liu. 1992. Word Identification for Mandarin Chinese Sentences. COLING-92. 101-105. Nantes, France.

Chiang, T.-H., J.-S. Chang, M.-Y. Lin, and K.Y. Su. 1992. Statistical Models for Word Segmentation and Unknown Word Resolution. Proc. of ROCLING V. 121-146.

Huang, C.-R., K. Ahrens, and K.-J. Chen. 1993. A Data-driven Approach to Psychological Reality of the Mental Lexicon: Two Studies in Chinese Corpus Linguistics. Proc. of the International Conference on the Biological Basis of Language. 53-68. Chiayi: Center of Cognitive Science, National Chung Cheng U.

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

____ and C. Shih. 1990. A Statistical Method for Finding Word Boundaries in Chinese Text. Computer Processing of Chinese and Oriental Languages. 4.4:336-351.

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. Proc. of the 1994 Kyoto Conference. A Festschrift for Professor Akira Ikeya. 113-123. Tokyo: The Logico-Linguistics Society of Japan.