Rethinking Chinese Word Segmentation: Tokenization, Character Classification, or Wordbreak Identification

Chu-Ren Huang
Institute of Linguistics
Academia Sinica,Taiwan
churen@gate.sinica.edu.tw

Petr Šimon
Institute of Linguistics
Academia Sinica,Taiwan
sim@klubko.net

Shu-Kai Hsieh
DoFLAL
NIU, Taiwan
shukai@gmail.com

Laurent Prévot
CLLE-ERSS, CNRS
Université de Toulouse, France
prevot@univ-tlse2.fr

Abstract

This paper addresses two remaining challenges in Chinese word segmentation. The challenge in HLT is to find a robust segmentation method that requires no prior lexical knowledge and no extensive training to adapt to new types of data. The challenge in modelling human cognition and acquisition it to segment words efficiently without using knowledge of wordhood. We propose a radical method of word segmentation to meet both challenges. The most critical concept that we introduce is that Chinese word segmentation is the classification of a string of character-boundaries (CB's) into either word-boundaries (WB's) and non-word-boundaries. In Chinese, CB's are delimited and distributed in between two characters. Hence we can use the distributional properties of CB among the background character strings to predict which CB's are WB's.

1 Introduction: modeling and theoretical challenges

The fact that word segmentation remains a main research topic in the field of Chinese language processing indicates that there maybe unresolved theoretical and processing issues. In terms of processing, the fact is that none of exiting algorithms is robust enough to reliably segment unfamiliar types of texts before fine-tuning with massive training data. It is true that performance of participating teams have steadily improved since the first SigHAN Chinese segmentation bakeoff (Sproat and Emerson, 2004). Bakeoff 3 in 2006 produced best f-scores at 95% and higher. However, these can only be achieved after training with the pre-segmented training dataset. This is still very far away from real-world application where any varieties of Chinese texts must be successfully segmented without prior training for HLT applications.

In terms of modeling, all exiting algorithms suffer from the same dilemma. Word segmentation is supposed to identify word boundaries in a running text, and words defined by these boundaries are then compared with the mental/electronic lexicon for POS tagging and meaning assignments. All existing segmentation algorithms, however, presuppose and/or utilize a large lexical databases (e.g. (Chen and Liu, 1992) and many subsequent works), or uses the position of characters in a word as the basis for segmentation (Xue, 2003).

In terms of processing model, this is a contradiction since segmentation should be the pre-requisite of dictionary lookup and should not presuppose lexical information. In terms of cognitive modeling, such as for acquisition, the model must be able to account for how words can be successfully segmented and learned by a child/speaker without formal training or a priori knowledge of that word. All current models assume comprehensive lexical knowledge.

2 Previous work

Tokenization model. The classical model, described in (Chen and Liu, 1992) and still adopted in many recent works, considers text segmentation as a

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 69–72,
Prague, June 2007.

tokenization. Segmentation is typically divided into two stages: dictionary lookup and out of vocabulary (OOV) word identification. This approach requires comparing and matching tens of thousands of dictionary entries in addition to guessing thousands of OOV words. That is, this is a 10^4x10^4 scale mapping problem with unavoidable data sparseness.

More precisely the task consist in finding all sequences of characters $C_i, \ldots, C_n$ such that $[C_i, \ldots C_n]$ either matches an entry in the lexicon or is guessed to be so by an unknown word resolution algorithm. One typical kind of the complexity this model faces is the overlapping ambiguity where e.g. a string $[Ci-1, Ci, Ci+1]$ contains multiple substrings, such as $[Ci-1, Ci,]$ and $[Ci, Ci+1]$, which are entries in the dictionary. The degree of such ambiguities is estimated to fall between 5% to 20% (Chiang et al., 1996; Meng and Ip, 1999).

2.1 Character classification model

A popular recent innovation addresses the scale and sparseness problem by modeling segmentation as character classification (Xue, 2003; Gao et al., 2004). This approach observes that by classifying characters as word-initial, word-final, penultimate, etc., word segmentation can be reduced to a simple classification problem which involves about 6,000 characters and around 10 positional classes. Hence the complexity is reduced and the data sparseness problem resolved. It is not surprising then that the character classification approach consistently yields better results than the tokenization approach. This approach, however, still leaves two fundamental questions unanswered. In terms of modeling, using character classification to predict segmentation not only increases the complexity but also necessarily creates a lower ceiling of performance In terms of language use, actual distribution of characters is affected by various factors involving linguistic variation, such as topic, genre, region, etc. Hence the robustness of the character classification approach is restricted.

The character classification model typically classifies all characters present in a string into at least three classes: word Initial, Middle or Final positions, with possible additional classification for word-middle characters. Word boundaries are inferred based on the character classes of ‘Initial’ or ‘Final’.

This method typically yields better result than the tokenization model. For instance, Huang and Zhao (2006) claims to have a f-score of around 97% for various SIGHAN bakeoff tasks.

3 A radical model

We propose a radical model that returns to the core issue of word segmentation in Chinese. Crucially, we no longer pre-suppose any lexical knowledge. Any unsegmented text is viewed as a string of character-breaks (CB’s) which are evenly distributed and delimited by characters. The characters are not considered as components of words, instead, they are contextual background providing information about the likelihood of whether each CB is also a wordbreak (WB). In other words, we model Chinese word segmentation as wordbreak (WB) identification which takes all CB’s as candidates and returns a subset which also serves as wordbreaks. More crucially, this model can be trained efficiently with a small corpus marked with wordbreaks and does not require any lexical database.

3.1 General idea

Any Chinese text is envisioned as sequence of characters and character-boundaries $CB_0C1CB_1C_2 \ldots CB_{i-1}C_iCB_i \ldots CB_{n-1}C_nCB_n$ The segmentation task is reduced to finding all CBs which are also wordbreaks WB.

3.2 Modeling character-based information

Since CBs are all the same and do not carry any information, we have to rely on their distribution among different characters to obtain useful information for modeling. In a segmented corpus, each WB can be differentiated from a non-WB CB by the character string before and after it. We can assume a reduced model where either one character immediately before and after a CB is considered or two characters (bigram). These options correspond to consider (i) only word-initial and word-final positions (hereafter the 2-CB-model or 2CBM) or (ii) to add second and penultimate positions (hereafter the 4-CB-model or 4CBM). All these positions are well-attested as morphologically significant.

3.3 The nature of segmentation

It is important to note that in this approaches, although characters are recognized, unlike (Xue, 2003) and Huang et al. (2006), charactes simply are in the background. That is, they are the necessary delimiter, which allows us to look at the string of CB's and obtaining distributional information of them.

4 Implementation and experiments

In this section we slightly change our notation to allow for more precise explanation. As noted before, Chinese text can be formalized as a sequence of characters and intervals as illustrated in we call this representation an *interval form.*

$c_1 I_1 c_2 I_2 \ldots c_{n-1} I_{n-1} c_n$.

In such a representation, each interval I_k is either classified as a plain character boundary (CB) or as a word boundary (WB).

We represent the neighborhood of the character c_i as $(c_{i-2}, I_{i-2}, c_{i-1}, I_{i-1}, c_i, I_i, c_{i+1}, I_{i+1})$, which we can be simplified as $(I_{-2}, I_{-1}, c_i, I_{+1}, I_{+2})$ by removing all the neighboring characters and retaining only the intervals.

4.1 Data collection models

This section makes use of the notation introduced above for presenting several models accounting for character-interval class co-occurrence.

Word based model. In this model, statistical data about word boundary frequencies for each character is retrieved word-wise. For example, in the case of a monosyllabic word only two word boundaries are considered: one before and one after the character that constitutes the monosyllabic word in question.

The method consists in mapping all the Chinese characters available in the training corpus to a vector of word boundary frequencies. These frequencies are normalized by the total frequency of the character in a corpus and thus represent probability of a word boundary occurring at a specified position with regard to the character.

Let us consider for example, a tri-syllabic word $W = c_1 c_2 c_3$, that can be rewritten as the following interval form as $W^I = I^B_{-1} c_1 I^N_1 c_2 I^N_2 c_3 I^B_3$.

In this interval form, each interval I_k is marked as word boundary ^{B} or ^{N} for intervals within words. When we consider a particular character c_1 in W, there is a word boundary at index -1 and 3. We store this information in a mapping $c_1 = \{-1 : 1, 3 : 1\}$. For each occurrence of this character in the corpus, we modify the character vector accordingly, each WB corresponding to an increment of the relevant position in the vector. Every character in every word of the corpus in processed in a similar way.

Obviously, each character yields only information about positions of word boundaries of a word this particular character belongs to. This means that the index I_{-1} and I_3 are not necessarily incremented everytime (e.g. for monosyllabic and bi-syllabic words)

Sliding window model. This model does not operate on words, but within a window of a give size $(span)$ sliding through the corpus. We have experimented this method with a window of size 4. Let us consider a string, $s =$ "$c_1 c_2 c_3 c_4$" which is not necessarily a word and is rewritten into an interval form as $s^I =$ "$c_1 I_1 c_2 I_2 c_3 I_3 c_4 I_4$". We store the co-occurrence character/word boundaries information in a fixed size $(span)$ vector.

For example, we collect the information for character c_3 and thus arrive at a vector $c_3 = (I_1, I_2, I_3, I_4)$, where 1 is incremented at the respective position $if I_k = WB$, zero otherwise.

This model provides slightly different information that the previous one. For example, if a sequence of four characters is segmented as $c_1 I^N_1 c_2 I^B_2 c_3 I^B_3 c_4 I^B_4$ (a sequence of one bi-syllabic and two monosyllabic words), for c_3 we would also get probability of I_4, i.e. an interval with index +2 . In other words, this model enables to learn WB probability across words.

4.2 Training corpus

In the next step, we convert our training corpus into a corpus of interval vectors of specified dimension. Let's assume we are using dimension $span = 4$. Each value in such a vector represents the probability of this interval to be a word boundary. This probability is assigned by character for each position with regard to the interval. For example, we have segmented corpus $C = c_1 I_1 c_2 I_2 \ldots c_{n-1} I_{n-1} c_n$, where each I_k is labeled as B for word boundary or N for non-boundary.

In the second step, we move our 4-sized window through the corpus and for each interval we query a character at the corresponding position from the interval to retrieve the word boundary occurrence probability. This procedure provides us with a vector of 4 probability values for each interval. Since we are creating this training corpus from an already segmented text, a class (B or N) is assigned to each interval.

The testing corpus (unsegmented) is encoded in a similar way, but does not contain the class labels B and N.

Finally, we automatically assign probability of 0.5 for unseen events.

4.3 Predicting word boundary with a classifier

The Sinica corpus contains 6820 types of characters (including Chinese characters, numbers, punctuation, Latin alphabet, etc.). When the Sinica corpus is converted into our interval vector corpus, it provides 14.4 million labeled interval vectors. In this first study we have implement a baseline model, without any pre-processing of punctuation, numbers, names.

A decision tree classifier (Ruggieri, 2004) has been adopted to overcome the non-linearity issue. The classifier was trained on the whole Sinica corpus, i.e. on 14.4 million interval vectors. Due to space limit, actual bakeoff experiment result will be reported in our poster presentation.

Our best results is based on the sliding window model, which provides better results. It has to be emphasized that the test corpora were not processed in any way, i.e. our method is sufficiently robust to account for a large number of ambiguities like numerals, foreign words.

5 Conclusion

In this paper, we presented a radical and robust model of Chinese segmentation which is supported by initial experiment results. The model does not pre-suppose any lexical information and it treats character strings as context which provides information on the possible classification of character-breaks as word-breaks. We are confident that once a standard model of pre-segmentation, using textual encoding information to identify WB's which involves non-Chinese characters, will enable us to achieve even better results. In addition, we are looking at other alternative formalisms and tools to implement this model to achieve the optimal results. Other possible extensions including experiments to simulate acquisition of wordhood knowledge to provide support of cognitive modeling, similar to the simulation work on categorization in Chinese by (Redington et al., 1995). Last, but not the least, we will explore the possibility of implementing a sharable tool for robust segmentation for all Chinese texts without training.

References

Academia Sinica Balanced Corpus of Modern Chinese. http://www.sinica.edu.tw/SinicaCorpus/

Chen K.J and Liu S.H. 1992. *Word Identification for Mandarin Chinese sentences*. Proceedings of the 14th conference on Computational Linguistics, p.101-107, France.

Chiang,T.-H., J.-S. Chang, M.-Y. Lin and K.-Y. Su. 1996. *Statistical Word Segmentation*. In C.-R. Huang, K.-J. Chen and B.K. T'sou (eds.): Journal of Chinese Linguistics, Monograph Series, Number 9, Readings in Chinese Natural Language Processing, pp. 147-173.

Gao, J. and A. Wu and Mu Li and C.-N.Huang and H. Li and X. Xia and H. Qin. 2004. *Adaptive Chinese Word Segmentation*. In Proceedings of ACL-2004.

Meng, H. and C. W. Ip. 1999. *An Analytical Study of Transformational Tagging for Chinese Text*. In. Proceedings of ROCLING XII. 101-122. Taipei

Ruggieri S. 2004. *YaDT: Yet another Decision Tree builder*. Proceedings of the 16th International Conference on Tools with Artificial Intelligence (ICTAI 2004): 260-265. IEEE Press, November 2004.

Richard Sproat and Thomas Emerson. 2003. *The First International Chinese Word Segmentation Bakeoff*. Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, Sapporo, Japan, July 2003.

Xue, N. 2003. *Chinese Word Segmentation as Character Tagging*. Computational Linguistics and Chinese Language Processing. 8(1): 29-48

Redington, M. and N. Chater and C. Huang and L. Chang and K. Chen. 1995. *The Universality of Simple Distributional Methods: Identifying Syntactic Categories in Mandarin Chinese*. Presented at the Proceedings of the International Conference on Cognitive Science and Natural Language Processing. Dublin City University.