CHINESE SEGMENTATION DISAMBIGUATION

Wanying Jin
Computing Research Laboratory
New Mexico State University
wanying@crl.nmsu.edu

Abstract

A technique of reasoning under uncertainty is studied in an attempt to solve disambiguation problems of Chinese segmentation. A knowledge-based inexact reasoning theory incorporating knowledge in morphology, syntax, semantics and pragmatics is presented.

1 Introduction

Processing Chinese texts is specifically difficult in its computation because normally sentences in Chinese texts are represented as strings of Chinese characters without spaces to indicate word boundaries. This causes a problem for Chinese machine translation, statistical analysis of Chinese corpora, Chinese information retrieval, etc.; as usually these projects are based on the assumption that all lexicon distinctions have been recognized in advance.

Several approaches aimed to transfer a Chinese character string into a word string have been studied in recent decades. Two competing approaches commonly used for Chinese text segmentation are the statistical approach (Chang, et al, 1991; Sproat and Shih, 1991; Chiang, et al, 1992) and the heuristic approach (Chen and Liu, 1992; He, et al, 1991; Jin and Nie, 1993; Jin, 1992; Liang and Zhen, 1991; Wang, et al, 1991). Although a high degree of precision has been reported for both methods, each has its limitations particularly in identifying unknown words and disambiguating multiple segmentations. Recently, a hybrid approach incorporating heuristics with statistics has been studied in an attempt to solve unknown word recognition problems (Chen and Liu, 1992; Nie and Jin, 1994). However, ambiguous segmentation is still a difficult problem.

In this paper a method of reasoning under uncertainty intending to disambiguate Chinese segmentation is presented. A model of evidential strength in inexact reasoning has been studied by (Buchanan and Shortliffe, 1984). In the process of Chinese segmentation knowledge in morphology, syntax, semantics and pragmatics is used as evidence to support the disambiguation hypotheses. The similarity of uncertain knowledge and inexact reasoning between medical diagnosis and natural language interpretation makes it possible to apply MYCIN technique to Chinese text segmentation.

2 Difficulties in Chinese segmentation

As claimed in (Liu, 1987),the main causes of segmentation ambiguity are vagueness in word definition and the phenomenon of word chains. The vagueness of the word definitions causes segmentation ambiguities, as in the string 现代化工厂. It can stands either for 现代化 工厂 (*modern factory*) or for 现代 化工厂 (*modern chemical factory*). A word chain is a sequence of Chinese characters from which several words can be produced with or without overlap. Two types of word chains have been recognized in Chinese literature, i.e. multisense combinations and intersection combinations (Huang and Liu, 1988). The string

冰箱 is an example of multi-sense combination; 冰 (ice), 箱(box) and 冰箱(refrigerator) are all words. The character string 球拍卖 is an example of intersection combination; 球拍(paddle) is a word and 拍卖(sell-at-sale-price) is also a word, whereas 拍 is the intersection character. The example of the string 乒乓球拍卖完了 illustrates the typical segmentation ambiguity caused by word chains. The segmentation of this string can be either

乒乓球 拍卖 完了。

(*The ping-pong-balls were sold outat sale price.*) or

乒乓 球拍 卖 完了。

(*The paddles for table tennis were sold out.*)

Some ambiguities can be solved by word structure knowledge. Others can be disambiguated by syntactic and/or semantic knowledge. The most difficult disambiguation is that requiring contextual or pragmatic knowledge to arrive at an appropriate interpretation as in the string 学生会写文章 which can be segmented into:

学生 会 写 文章

(*students will write a paper.*) or

学生会 写 文章

(*student-association writes a paper.*)

Both are syntactically and semantically correct. In this case, contextual information would allow the reader to trace the information claimed in the previous statements to solve ambiguity problems.

3 Reasoning theory for Chinese segmentation disambiguation

A model of evidential strength in inexact reasoning studied by (Buchanan and Shortliffe, 1984) has been successfully implemented in the MYCIN system. The theory is that, if a hypothesis can be derived from various types of mutually exclusive evidence, then the strength of truth of the hypothesis can be increased to reach a plausible conclusion.

Two concepts MB[h,e] and MD[h,e] have been introduced as the measures of belief and disbelief. MB[h,e] means the measure of increased belief in the hypothesis h, based on the evidence e. MD[h,e] means the measure of increased disbelief in the hypothesis h, based on the evidence e. To facilitate comparison of the evidential strength of competing hypotheses, certainty factor CF is introduced to combine degrees of belief and disbelief as follows:

$$CF[h,e] = MB[h,e] - MD[h,e]$$

In the case that a hypothesis is derived from a number of mutually exclusive observations, the combining functions are defined as:

if $MD[h, e_1 \& e_2] = 1$
then $MB[h, e_1 \& e_2] = 0$
otherwise
$MB[h, e_1 \& e_2]$
$= MB[h, e_1] + MB[h, e_2] * (1 - MB[h, e_1])$

if $MB[h, e_1 \& e_2] = 1$
then $MD[h, e_1 \& e_2] = 0$
otherwise
$MD[h, e_1 \& e_2]$
$= MD[h, e_1] + MD[h, e_2] * (1 - MD[h, e_1])$

In the case that two hypotheses are established with positive evidence from syntactic and semantic knowledge with the same degree, no discrimination of the strength of truth hypotheses can be drawn. If world knowledge provides positive evidence for the first hypothesis and negative evidence to the second; then the strength of the first hypothesis is stronger than that of the second. Therefore, the first hypothesis would be the most likely correct segmentation.

A weighted certainty factor is proposed here to represent the importance of various linguistic aspects. The weight is a vector of four elements representing the importance of morphology, syntax, semantics and pragmatics, respectively, which total 1, i.e.

$$CF_i[h,e] = W_i * CF[h,e]$$

where W_i is the weight of the certainty factor CF_i in hypothesis h supported by the evidence e with respect to one of the linguistic

aspects. Suppose the weight vector (0.1, 0.2, 0.3, 0.4) is assigned for morphology, syntax, semantics and pragmatics, respectively, then the following example illustrates the function of the weighted certainty factor $CF_i[h, e]$.

For the string 我们公司的第三把手没有多大权力 (the third leader in our company does not have much power) the word chain 把手 produces two segmentations:

我们 公司 的 第三 把手 没有 多大 权力
(the third leader in our company does not have much power) or
我们 公司 的 第三 把 手 没有 多大 权力
(the third piece-of hand in our company does not have much power)

To estimate the strength of truth of the first hypothesis, suppose:

- the word structure rule gives the evidential strength (0.5) for the hypothesis because the word chain 把手 can be either 把 手 (piece-of hand) or 把手 (leader). Therefore,

  $CF_1[h, e_1] = W_1 * CF[h, e_1] = 0.05$ and
  $CF[h, e_1] = MB[h, e_1] - MD[h, e_1]$
  $= 0.05$

- the syntactic rule gives the evidential strength (1) because it definitely is a grammatical sentence. Therefore,

  $CF_2[h, e_2] = W_2 * CF[h, e_2] = 0.2$ and
  $CF[h, e_1\&e_2]$
  $= MB[h, e_1\&e_2] - MD[h, e_1\&e_2]$
  $= 0.24$

- the semantic rule gives the evidential strength 1) since 把手(the leader) can have power. Therefore,

  $CF_3[h, e_3] = W_3 * CF[h, e_3] = 0.3$ and
  $CF[h, e_1\&e_2\&e_3]$
  $= MB[h, e_1\&e_2\&e_3] - MD[h, e_1\&e_2\&e_3]$
  $= 0.46$

- the world knowledge rule gives the evidential strength (0.8) because it is quite true that the leader has less power than that of the first or second leader. Therefore,

  $CF_4[h, e_4]$
  $= W_4 * CF[h, e_4] = 0.32$ and
  $CF[h, e_1\&e_2\&e_3\&e_4]$
  $= MB[h, e_1\&e_2\&e_3\&e_4]$
  $- MD[h, e_1\&e_2\&e_3\&e_4]$
  $= 0.63$

The certainty factor CF of the hypothesis 我们 公司 的 第三 把手 没有 多大 权力 is 0.63. Therefore, this segmentation is likely to be a coherent string.

To estimate the evidential strength of the second hypothesis, suppose:

- the word structure rule gives the evidential strength (0.5) for this hypothesis since 把手 can be either 把 手(piece-of hand) or 把手 (leader). Therefore,

  $CF_1[h, e_1]$
  $= W_1 * CF[h, e_1] = 0.05$ and
  $CF[h, e_1]$
  $= MB[h, e_1] - MD[h, e_1] = 0.05$

- the syntactic rule gives the evidential strength (1) because it is a grammatical sentence. Therefore,

  $CF_2[h, e_2]$
  $= W_2 * CF[h, e_2] = 0.2$ and
  $CF[h, e_1\&e_2]$
  $= MB[h, e_1\&e_2] - MD[h, e_1\&e_2] = 0.24$

- the semantic rule gives the negative evidential strength (-1) because the phrase *the hand of a company* violates the semantic constraint. Therefore,

  $CF_3[h, e_3]$
  $= W_3 * CF[h, e_3] = -0.3$ and
  $CF[h, e_1\&e_2\&e_3]$
  $= MB[h, e_1\&e_2\&e_3] - MD[h, e_1\&e_2\&e_3]$
  $= -0.06$

- the world knowledge rule gives a negative evidential strength (-1) because a company does not have a hand as one of its components.

  $CF_4[h, e_4] = -0.4$ and
  $CF[h, e_1\&e_2\&e_3\&e_4]$
  $= MB[h, e_1\&e_2\&e_3\&e_4]$
  $- MD[h, e_1\&e_2\&e_3\&e_4]$
  $= -0.34$

The certainty factor CF of the hypothesis 我们 公司 的 第三 把 手 没有 多大 权力 is - 0.34.

Therefore, this segmentation is unlikely to be a coherent string.

4 Discussion

The assignment for the weight vector is empirical. It is based on the following analysis in which '1's reresent the truth of each evidence/hypothesis and '0's represent the false. Since the segmentation algorithm always produces a segmented string, it is assumed that the evidence from morphology is true in varying degrees depending on the complexity of the word chain. The justification of a hypothesis is based on the evidence presented by the pragmatic, semantic and syntactic aspects shown in the following table.

	pragmtc	semtc	syntc	hypths
(1)	0	0	0	0
(2)	0	0	1	0
(3)	0	1	0	0
(4)	0	1	1	0
(5)	1	0	0	1
(6)	1	0	1	1
(7)	1	1	0	1
(8)	1	1	1	1

- Case(1) indicates that if no evidence can prove the truth of the hypothesis, then the hypothesis is false.
- Case(2) indicates that if the evidence supports an incoherent grammatical sentence inconsistent with the context/circumstance, then the hypothesis is false as in the case of 香蕉吃猴子(a banana ate a monkey).
- Case(3) indicates that if the evidence supports a meaningful but ungrammatical string inconsistent with the context/circumstance, then the hypothesis is false, i.e. 他坏蛋 (he wretch) against the real fact that he is a nice guy.
- Case(4) indicates that even if the evidence supports a grammatical meaningful sentence but is inconsistent with the context/circumstance, then the hypothesis is false, i.e., 总统 下台 平 民 愤 (the president's forced resignation makes people angry) violates the circumstance that people hate the president.
- Case(5) indicates the case of an idiomatic expression where the string is literally ungrammatical and incoherent, but as a whole it can be interpreted figuratively to make perfect sense. Therefore, we assume that the hypothesis is true as in the case of 车水马龙, literally means "car-water-horse-dragon", but figuratively, it means "very crowded".
- Case(6) indicates the case of a metaphor or metonymy which superficially it is an incoherent grammatical string, but by reasoning with the support of world knowledge it can be interpreted as a meaningful string. Then, it is assumed that the hypothesis is true, i.e., 我喝西北风 (I drink North-West wind) means "I have nothing to eat".
- Case(7) indicates that the evidence supports a meaningful but ungrammatical string consistent with the context/circumstance. then the hypothesis is true as in 他坏蛋 (he wretch) is consistent with the real fact that he is a bad guy.
- Case(8) indicates that if all evidence gives positive support to the hypothesis, then the hypothesis is true.

From the analysis, it seems to be that pragmatic knowledge provides the strongest evidence for the hypothesis. Therefore, the highest weight is assigned to the pragmatic aspect of the certainty factor. In the absence of pragmatic information a default assumption, that semantic evidence is more important than syntactic evidence, is made. This can be observed in daily life people communicate through many ungrammatical expressions without having a problem of transferring the message such as a brief email message:[1] *DRAFT-comments-hard copy best-asap to yw pls.* It means "To

[1] A brief e-mail message from Dr. Yorick Wilks to the researchers in Computing Research Laboratory at New Mexico State University.

write the comment for the DRAFT on the hard copy would be the best. Please return it to Yorick Wilks as soon as possible."

The certainty factor *CF* is used under the premise that all of the evidence is rendered by mutually exclusive observations. Since language is an expression integrating syntactic, semantic and pragmatic information, is the syntactic, semantic and pragmatic evidence mutually exclusive? This is not so clear. All knowledge is culturally dependent, i.e. one particular instance may be acceptable in one culture but not in another. In this research a default assumption is made that the observations from various language aspects are independent. The question is left open for further discussion.

5 References

Buchanan, B. and E. Shortliffe. (1984). Uncertainty and Evidential Support. In B. G. Buchanan and E. H. Shortliffe Ed., *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, Addison-Wesley Publishing Company., pp. 209-232.

Chang, J. S., et al. (1991). Chinese word segmentation through constraint satisfaction and statistical optimization, *Proc. of the 4th R.O.C. Computational Linguistics Conference*, pp. 147-165.

Chen, K. J. and S. H. Liu. (1992). Word Identification for Mandarin Chinese Sentences. *Proc. of the 5th International Conference on Computational Linguistics, Vol. 1*, pp. 101-107.

Chiang, T. H., et al. (1992). Statistical models for segmentation and unknown word resolution. *Proc. of the 5th R.O.C. Computational Linguistics Conference*, pp. 123-146.

He, K. K.,et al. (1991). The Design Principle for a Written Chinese Automatic Segmentation Expert System. *Journal of Chinese Information Processing, vol.5, No.2*, pp. 1-14.

Huang, X. X. and D. Y., Liu. (1988). The Phenomenon of Word Chain and the Automatic Segmentation in Written Chinese. *Journal of the Development of Knowledge Engineering*, pp. 287-291.

Jin, W. and J. Y. Nie. (1993). Segmentation du Chinois - une Etape Cruciale vers la Traduction Automatique du Chino is. In P.Bouillon and A. Clas Ed., *La Traductique*. Les presses de l'Universite de Montreal, pp. 349-363.

Jin, W. (1992). A Case Study: Chinese Segmentation and its Disambiguation. *MCCS-92-227*, Computing Research Laboratory, New Mexico State University.

Liang, N. Y. and Y. B., Zhen. (1991). A Chinese Word Segmentation Model and a Chinese Word Segmentation System PC-CWSS. *Proc. of COLIPS, Vol. 1, No. 1*, pp.51-55.

Liu, Y. Q. (1987). Difficulties in Chinese Language Processing and Method to their Solution. *Proc. of 1987 International Conference on Chinese Information Processing, Vol. 2*, pp. 125-126.

Nie, J. Y. and W. Jin. (1994). A Hybrid Approach to Unknown Word Detection and Segmentation of Chinese, Appear in *Proc. of International Conference on Chinese Computing'94 (ICCC94)*.

Sproat, R. and C., Shih. (1991). A statistical method for finding word boundaries in Chinese text,*Computer Processing of Chinese and Oriental Languages, Vol. 4, No. 4*, pp. 336-351.

Wang, L. J., et al. (1991). A Parsing Method for Identifying Words in Mandarin Chinese Sentences. *Proc. of the 12th International Joint Conference on Artificial Intelligence, Vol. 2*, pp. 1018-1023.