Statistical Properties of Overlapping Ambiguities in Chinese Word Segmentation and a Strategy for Their Disambiguation*

Wei Qiao[1], Maosong Sun[1], Wolfgang Menzel[2]

[1] State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Sci. & Tech., Tsinghua University, Beijing 100084, China.
qiaow04@mails.tsinghua.edu.cn, sms@mail.tsinghua.edu.cn
[2] Department of Informatik, Hamburg University, Hamburg, Germany.
menzel@informatik.uni-hamburg.de

Abstract. Overlapping ambiguity is a major ambiguity type in Chinese word segmentation. In this paper, the statistical properties of overlapping ambiguities are intensively studied based on the observations from a very large balanced general-purpose Chinese corpus. The relevant statistics are given from different perspectives. The stability of high frequent maximal overlapping ambiguities is tested based on statistical observations from both general-purpose corpus and domain-specific corpora. A disambiguation strategy for overlapping ambiguities, with a predefined solution for each of the 5,507 pseudo overlapping ambiguities, is proposed consequently, suggesting that over 42% of overlapping ambiguities in Chinese running text could be solved without making any error. Several state-of-the-art word segmenters are used to make comparisons on solving these overlapping ambiguities. Preliminary experiments show that about 2% of the 5,507 pseudo ambiguities which are mistakenly segmented by these segmenters can be properly treated by the proposed strategy.

Key words: overlapping ambiguity, statistical property, disambiguation strategy, domain-specific corpora

1 Introduction

Word segmentation is the initial stage of many Chinese language processing tasks and has drawn a large body of research. Overlapping ambiguity (OA) is one of the basic types of segmentation ambiguities. A string in Chinese text is called an overlapping ambiguity string (OAS) if it satisfies following definition: suppose S is a string of Chinese characters, D is a Chinese wordlist, and S is not in D. S is an OAS if there exists a sequence of words in D denoted $w_1, w_2, \ldots, w_m (m > 2)$that exactly cover S, and adjacent words $w_i, w_i + 1$ $(1 \leqslant i < m)$ intersect but do not cover each other.

* The research is supported by the National Natural Science Foundation of China under grant number 60573187 and the CINACS project.

Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): TSD 2008, LNAI 5246, pp. 1–10, 2008.

It is reported that OA constitutes 90% of segmentation ambiguities in Chinese running text [6].

Previous work on solving overlapping ambiguities can be roughly classified into two categories: rule-based and statistical approaches.

Maximum Matching (MM) can be viewed as the simplest rule-based OA disambiguation strategy. [2] indicates that MM can only achieve an accuracy of 73.1% for OA strings. A set of manually generated rules are used in [15] and reported an accuracy of 81.0%. A lexicon-based method is presented in [12], achieving an accuracy of 95.0%. A general scheme in statistical approach is to use character or word N-gram models or POS N-gram models to find the best segmentation in the candidate segmentation space of an input sentence. For example, [10] presents a character bigram method and reports an accuracy of 90.3% for OA strings. Another general scheme is to define segmentation disambiguation as a binary classification problem and use a classifier to solve it. For example, [4] uses Support Vector Machine with mutual information between Chinese character pairs as features, achieving an accuracy of 92.0% for OA strings.

It is worth noting that [11] finds that the 4,619 most frequent OA strings, which fall into an ambiguity category named "pseudo segmentation ambiguity" (see Section 2 for detail), can be disambiguated in a context-free way, and these strings can cover 53.35% of OA tokens in a news corpus. A so-called "memory-based model" is proposed to solve these OAs. The work of [5] continues in this line: totally 41,000 most frequent pseudo OA strings are identified and a solution called "lexicalized rules" is proposed. Experimental results show that it can benefit word segmentation significantly.

The research here will follow and extend that of [11] and [5]. Three basic issues remain unsolved in their work:

Basic issue 1: The corpora used in either [11] or [5] only include news data. Obviously, the findings in these works should be further validated by adopting more "appropriate" data, otherwise, they still seems to be too restricted.

Basic issue 2: Even if the above "conclusion" can really work based on the observation from more "appropriate" data, we further need to determine the stable core of pseudo OA strings, to test its coverage and check its stability in Chinese running text, both general-purpose and domain-specific.

Basic issue 3: Once the core of pseudo OA strings is determined, we need to see if there exists a disambiguation strategy that can solve them effectively.

The remainder of this paper is organized as follows: Section 2 presents the related terms. Section 3 gives distributions of OA strings based on general-purpose corpus. Section 4 observes the stability of high frequent OA strings on both general-purpose corpus and domain-specific corpora. A resulting disambiguation strategy is proposed in Section 5. Section 6 is the conclusion.

2 Related Terms

In this section, we introduce some related terms (concepts) which are defined in [9] for describing various aspects of overlapping ambiguities.

We first give five basic terms:

Length of an OAS is the number of characters it contains; Each word in an OAS is called ***Span of an OAS***; The number of spans it contains is the ***Order of an OAS***; The common part of two adjacent spans is ***Intersection of spans***; The totality of the spans of an OAS constitutes its structure and is called ***Structure of an OAS***.

Three important concepts are further introduced as follows:

Maximal overlapping ambiguity string (MOAS): Let S_1 be an OAS and occurs as a substring of a sentence S. If an OAS containing S_1 never exists in S then S_1 is called a MOAS in the context of S.

Take the sentence "他为推广普通话费尽心血"(He tried his best to popularize the Mandarin) as an example, both the string of "普通话费" and "普通话费尽心血" in this sentence are OASs whereas "普通话费" is not maximal because it is included in "普通话费尽心血".

Real segmentation ambiguity: A segmentation ambiguity is said to be real if at least its two distinct possible segmentations can be realized in running Chinese texts depending on its contexts.

For example, "其次要" is said "real", because two segmentations "其|次要(the subordination)" and "其次|要(secondly should)" can be realistic:

a.先解决其主要问题，再解决其|次要问题。(First of all we should solve the main problem, and then consider the subordination one.)

b.首先要关注整体框架，其次|要注意细节。(Firstly we should focus on the whole framework, secondly should notice the details.)

Pseudo segmentation ambiguity: A segmentation ambiguity is said to be pseudo if only one of its distinct segmentations can be realized in running text.

For example, "部长篇小说" is said "pseudo" because only the latter of its two possible segmentations, "部长(minister)| 篇(measure word)小说(novel)" and "部(measure word)|长篇小说(long novel)", can be realized in text.

The advantage of distinguishing MOAS from OAS is that the latter is comparatively isolated from its context and thus readily available for independent study. Clearly, defining an ambiguity as a maximal overlapping ambiguity provides an adequate and quite stable processing unit for further investigation of its properties, for example, whether it is real or pseudo.

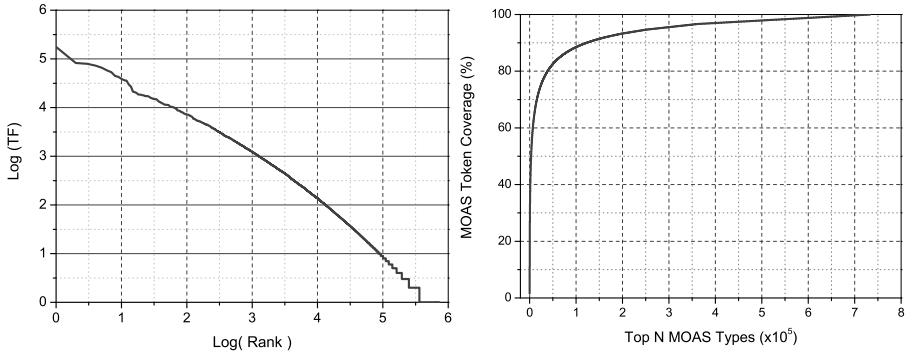
3 Statistical Properties of MOAS

Targeting at the basic issue 1 and 2 mentioned in Section 1, first of all, we design and construct a huge balanced Chinese corpus, CBC. CBC is very rich in content as it contains the collection of Chinese literature since 1920's, and it is well balanced, covering rich categories such as novel, essay, news, entertainment and texts from the web. The total size of CBC is 929,963,468 characters. The Chinese wordlist we used in this paper is developed by Peking University [14], with 74,191 entries (word types), denoted CWL here. Based on CWL, we extract all of the MOASs from CBC. A total of 733,066 distinct MOAS types are obtained at last, forming a complete MOAS type set, denoted CS-MOAS. These MOAS types have 11,103,551 occurrences in CBC, covering 39,432,267 Chinese characters, which constitutes 4.24% of CBC.

We then systematically observe the statistical properties of overlapping ambiguities through their distributions in CBC.

3.1 MOAS and Zipf's Law

Figure 1(a) shows the relationship between the rank of a MOAS type and its token frequency over CS-MOAS, in a log-log scale. We can see that this relationship roughly obeys Zipf's Law, i.e., rank $\times\ TF \approx C$, where the constant C is roughly 1.11×10^6.



(a) MOAS and Zipf's Law.

(b) Coverage of top N frequent MOAS types.

Fig. 1. MOAS and Zipf's Law.

The coverage of the top N frequent MOAS types to MOAS tokens over CBC (i.e., the total occurrences of CS-MOAS in CBC) is shown in Figure 1(b), where the MOAS types are sorted by their token frequencies. It indicates that, as expected, the coverage of top 3,500, 7,000 and 40,000 frequent MOAS types is 50.78%, 60.43% and 80.39% respectively. These top frequent MOASs are thus possibly regarded as a "core" of MOAS in running text.

3.2 The Detailed MOAS Distributions

We give the detailed distributions of MOAS from following four perspectives:

Perspective 1: The distribution of MOAS length over CBC (Table 1).

As shown in Table 1, the type and token coverage of MOASs with length below 6 are as high as 98.26% and 99.72%. Obviously, these MOAS types are of higher significance in OA disambiguation.

Perspective 2: The distribution of MOAS order over CBC (Table 2).

Table 2 shows that the type and token coverage of MOASs with order 2 and 3 are 89.14% and 98.17% respectively. And that of MOASs with orders below 5 are as high as 99.63% and 99.95% respectively. They show an even more concentrated distribution compared to that of MOAS length.

Perspective 3: The distribution of intersection length over CBC (Table 3).

Table 1. Distribution of MOAS length over CBC.

Length	#. of MOAS types	Coverage to MOAS types	Coverage to MOAS tokens
3	211,270	28.82%	55.22%
4	348,065	47.48%	36.86%
5	108,786	14.84%	5.83%
6	52,233	7.13%	1.81%
7∼12	12712	1.74%	0.28%
Total	733,066	100.00%	100.00%

Table 2. Distribution of MOAS order over CBC.

Order	#. of MOAS types	Coverage to MOAS types	Coverage to MOAS tokens
2	269,717	36.79%	64.49%
3	383,794	52.35%	33.68%
4	53,748	7.33%	1.27%
5	23,171	3.16%	0.51%
6∼10	2,636	0.35%	0.05%
Total	733,066	100.00%	100.00%

Table 3. Distribution of MOAS intersection length over CBC.

Length	#. of types	Coverage to tokens	#. of tokens	Coverage to tokens
1	1,300,736	99.66%	15,224,314	99.39%
2	4,408	0.34%	92,434	0.60%
3	30	0.00%	1,060	0.01%
Total	1,305,174	100%	15,317,808	100%

The distribution of intersection length is much more concentrated, with 99.66% of MOAS has intersection length of 1.

Perspective 4: The distribution of MOAS structure over CBC (Table 4).

The MOAS structure is in fact a combination of the order and length of spans of an MOAS, for example, the structure of the OAS "其次要" is (0-2,1-3)(note that the notation i-j means there exists a span which starts from location i and ends at location j of the given OAS). We totally find 97 different structure types for CS-MOAS over CBC. The top 3 major structure types are listed in Table 4.

Table 4. Distribution of MOAS structure over CBC.

Structure	#. of MOAS types	Coverage to MOAS types	Coverage to MOAS tokens
(0-2,1-3,2-4)	306,627	41.83%	29.63%
(0-2,1-3)	211,270	28.82%	55.22%
(0-2,1-3,2-4,3-5)	40,482	5.52%	0.94%
others	174,687	23.83%	14.21%
Total	733,066	100%	100%

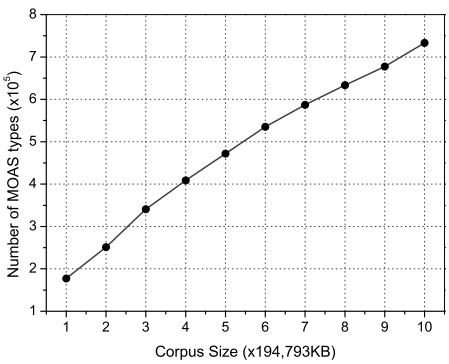
4 Stability of the Top N Frequent MOAS Types

The distributions given in Section 3 demonstrate that MOASs exhibit very strong centralization tendencies. It suggests that there may exist a "core" of MOASs with relatively small size meanwhile with quite powerful coverage capability for running texts. Regarding the top 3,500, 7,000 and 40,000 frequent MOASs as the candidates of "core", a question then comes: are these MOASs stable in Chinese running text? We observe the stability of these MOASs from following two perspectives.

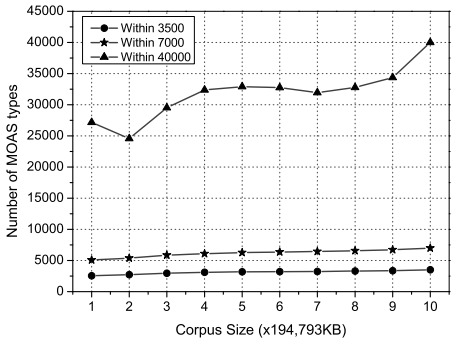
4.1 Perspective 1: Stability vs. Corpus Size

Firstly, we want to study the influence of corpus size on the stability of a potential MOAS core. We randomly divide CBC into ten equal parts, each of which has a size of 194,793KB. The experiments here start from any part of CBC, with one more part added in next round.

Figure 2(a) shows that the number of MOAS types increases almost linearly with the corpus size. On the contrary, the situation is totally different in Figure 2(b), where the vertical axis stands for the number of common part between the top N frequent MOASs in the current corpus and the top N frequent MOASs in CBC: the curves for the top 3,500 and 7,000 are almost flat, meanwhile the curve for the top 40,000 is with obvious fluctuation.



(a) #. of MOAS types vs. corpus size.



(b) #. of MOAS types falling into the top 3,500, 7,000, and 40,000 vs. corpus size.

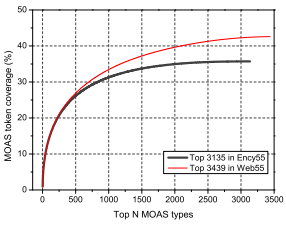
Fig. 2. Stability vs. corpus size.

4.2 Perspective 2: Stability vs. Domain-Specific Corpora

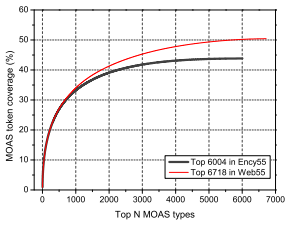
Secondly, we try to test the stability of a potential MOAS core on domain-specific corpora. In order to do this we design and conduct the other two domain-specific corpora: Ency55 and Web55. Ency55 is the electronic version of Chinese Encyclopedia, with 90.02 million characters while Web55 is collected from the web, with 54.97 million characters. They are all organized into 55 technical domains such as Geography, Communication, Mechanics, Chemistry etc. (Both corpora are fully independent to CBC).

There are totally 168,478 and 119,663 MOAS types in Ency55 and Web55 respectively, in terms of CWL. The MOAS types in Ency55 include a total of 3,783,164 characters, covering 4.2% of the corpus, and that in Web55 includes 2,028,053 characters, covering 3.7% of the corpus.

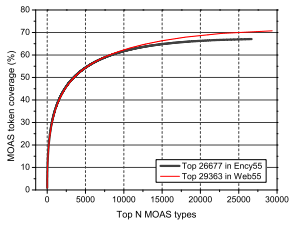
The coverage test of the top 3,500, 7,000 and 40,000 frequent MOASs in CBC to Ency55 and Web55 is carried out, as shown in Figure 3(a), 3(b) and 3(c). As can be seen, the token coverage can still reach 35.72%, 43.84% and 67.08%, which has a 13% $\sim$ 16% drop from that of CBC (50.78%, 60.43% and 80.39%) respectively. This is another evidence for that the potential MOAS cores are quite stable in Chinese.



(a) N=3,500



(b) N=7,000



(c) N=40,000

Fig. 3. Coverage of top N frequent MOASs in CBC over Ency55 and Web55.

Table 5. Token coverage of PM and RM over MOASs in CBC.

	PM	RM	Top 7,000
#. of MOAS Types	5,507	1,493	7,000
Token Coverage on MOASs	52.73%	7.70%	60.43%

Table 6. Token coverage of high frequent PMs in CBC over Ency55 and Web55.

	Ency55 MOAS	Web55 MOAS
#. of Common PMs	4,342	5,079
Token Coverage on MOASs	42.21%	47.89%

Due to the fact that the top 7,000 MOASs can cover 60.43% MOAS tokens in CBC, 43.84% MOAS tokens in Ency55 and 50.43% in Web55, we thus choose the top 7,000 as the core of MOASs which will serve as a basis for Section 5.

5 Disambiguation Strategy Inspired by Statistical Properties of MOAS

The top 7,000 MOASs can be further divided into two categories: pseudo MOAS (PM) and real MOAS (RM). Sometimes however it is not always easy to decide if an MOAS is pseudo or real in the strict sense. Here we make a relaxation on "pseudo": an RM which has very strong tendency to have only one segmentation way in corpus will be treated as an PM.

Consider the RM "出国门": in almost all the cases it is realized as "出| 国门(go abroad)", while in very rare cases, there still exists a very small possibility for it to be realized as "出国| 门(the way to go abroad)", as in the sentence "他想出国门都没有(For him there is no way to go abroad)". "出国门" will thus be thought of as an PM according to the above relaxation.

In terms of the relaxation definition, 5,507 out of the top 7,000 frequent MOAS are manually judged to be PM. The token coverage of PMs and RMs on all MOASs in CBC is listed in Table 5.

Table 6 shows the token coverage of the high frequent PM in CBC to all the MOASs over Ency55 and Web55.

It is promising to see that the PMs in the top 7,000 frequent MOASs can still cover 42.21% and 47.89% MOASs in domain-specific corpora. This indicates that this set of PMs is quite stable. Thus if these PMs are solved, it can be expected about 42% of overlapping ambiguities in Chinese text can be perfectly solved.

One thing that deserves to be mentioned is, it is possible for an PM in general-purpose corpus to change into an RM in domain-specific corpora. Based on our observation on 5,507 PMs, only a few of them fall into this case.

Concerning the basic issue 3 in Section 1, since the resolution of pseudo segmentation ambiguities is independent of their contexts, a basic strategy can be formulated: for

a highly frequent pseudo MOAS, its disambiguation can be performed by simply looking up a table in which its solution is pre-stored. In essence, this is an individual-based strategy, with the following merits: quite satisfactory token coverage to MOASs, full correctness for segmentation of pseudo MOASs, and low cost in time and space complexity.

Experiments have been performed to compare the performance of existing word segmenters with that of our strategy, focusing on PM resolution. Two state-of-the-art Chinese word segmenters, ICTCLAS1.0[1] and MSRSeg1.0[2], which separately scores the top one in the SIGHAN-bakeoff 2003 [8] and 2006 [1], have been chosen for comparison. For each of the 5,507 PM types, we randomly select a sample sentence which contains the PM from Web55. The results show that proposed strategy can handle all the samples perfectly, whereas, about 2.6% of them are mistakenly segmented by ICTCLAS1.0 and 2.3% of them by MSRSeg1.0.

Here we give out some typical segmentation errors produced by ICTCLAS1.0 and MSRSeg1.0 (the underlined):

公安局 长 是 主管 这一 事故 的. (From ICTCLAS1.0)

(The police chief(公安 局长) is the person in charge of this accident).

核电站 的 特殊性 质. (From MSRSeg1.0)

(The special properties(特殊 性质) of nuclear power station).

As Conditional Random Fields (CRF) [3] is the state-of-the-art machine learning model on solving sequence labeling problem [7], the performance on PMs is also tested by using CRF. The toolkit CRF++0.50[3] is used to build our CRF-based word segmenter. The window size is set five and four tag-set is used to distinguish the position of character. The training set provided by MSRA in SIGHAN-bakeoff 2005 is used to train the CRF model. The basic feature template adopted from [13] is used. The experimental result shows that totally 2.1% of 5,507 PM types are mistakenly segmented by CRF-based word segmenter.

A typical segmentation error (the underlined) is given here:

这一 现状 先 天地 决定 了 他们 的 使命

(This situation congenitally(先天 地) makes them to take the mission).

The improvement of 2% on PMs seems trivial, but it is a net gain. The strategy could be more effective when facing running text.

6 Conclusion

In this paper the statistical properties of overlapping ambiguities are intensively studied based on observations from a very large balanced general-purpose Chinese corpus. The stability of high frequent MOAS is tested based on statistical observations on both general-purpose corpus and domain-specific corpora. A disambiguation strategy is proposed consequently. Experiments show that over 42% of overlapping ambiguities in running text can be solved without making any error. About 2% mistakes produced by

[1] ICTCLAS 1.0.: http://www.nlp.org.cn

[2] MSRSeg.v1.: http://research.microsoft.com/-S-MSRSeg

[3] http://crfpp.sourceforge.net/

state-of-the-art Chinese word segmenters on MOASs can be solved by this strategy. We are now confident to claim that the basic issues addressed in Section 1 have been settled quite satisfactorily.

References

1. Emerson T.. 2005. *The second international Chinese word segmentation bakeoff.* In: Proceedings of the 4th SIGHAN Workshop, pages 123–133.
2. Huang C. N.. 1997. *Segmentation Problems in Chinese Processing.* Applied Linguistics 1:72–78. (In Chinese).
3. Lafferty J., A. McCallum, and F. Pereira. 2001. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* In Proceedings of 18th International Conference of ICML, pages 282–289.
4. Li R., S.H. Liu, S.W. Ye, and Z.Z. Shi. 2001. *A method for resolving overlapping ambiguities in Chinese word segmentation based on SVM and k-NN.* Journal of Chinese Information Processing, 15(6): 13–18. (In Chinese).
5. Li M., J.F. Gao, C.N. Huang, and J.F. Li. 2003. *Unsupervised training for overlapping ambiguity resolution in Chinese word segmentation.* In: Proceedings of SIGHAN 2003, pages 1–7.
6. Liang N.Y.. 1987. *A Chinese automatic segmentation system for written texts - CDWS.* Journal of Chinese Information Processing, 1(2):44–52. (In Chinese).
7. Peng F.C., F.F. Feng, and A. McCallum. 2004. *Chinese segmentation and new word detection using conditional random fields.* In: Proceedings of COLING 2004, pages 562–568, Geneva, Switzerland.
8. Sproat R. and T. Emerson. 2003. *The first international Chinese word segmentation bakeoff.* In: Proceedings of the 2nd SIGHAN Workshop, pages 133–143.
9. Sun M.S. and Z.P. Zuo. 1998. *Overlapping ambiguities in Chinese text.* Quantitative and Computational Studies on the Chinese Language, pages 323–338.
10. Sun M.S., C.N. Huang, and B.K.Y. T'sou. 1997. *Using character bigram for ambiguity resolution In Chinese word segmentation.* Computer Research and Development, 34(5): 332–339. (In Chinese).
11. Sun M.S., Z.P. Zuo and B.K.Y. T'sou. 1999. *The role of high frequent maximal crossing ambiguities in Chinese word segmentation.* Journal of Chinese Information Processing, 13(1): 27–37. (In Chinese).
12. Swen B. and S.W. Yu. 1999. *A graded approach for the efficient resolution of Chinese word segmentation ambiguities.* In: Proceedings of 5th Natural Language Processing Pacific Rim Symposium, pages 19–24.
13. Xue N.W.. 2003. *Chinese word segmentation as character tagging.* In: International Journal of Computational Linguistics, 8(1): 29–48.
14. Yu S.W. and X.F. Zhu. 2003. *Grammatical Information Dictionary for Contemporary Chinese, 2nd edition.* Tsinghua University Press. (In Chinese).
15. Zheng J.H. and K.Y. Liu. 1997. *Research on ambiguous word segmentation technique for Chinese text.* Language Engineering, pages 201–206, Tsinghua University Press, Beijing. (In Chinese).