基于标注语料库的组合歧义检测与消解

孙承杰 [1]，黄昌宁 [2]，关毅 [1]

（1. 哈尔滨工业大学，哈尔滨　150001；2. 微软亚洲研究院，北京　100080）

摘　要： 本文调查了不同的汉语分词标注语料库中的组合歧义的分布情况。基于调查结果，提出了一种利用一个大规模高质量的标注语料库中的知识来进行组合歧义的检测与消解的新方法。实验表明，我们的方法可以比基于实例的方法覆盖更多的组合歧义情况，在 SIGHAN bakeoff2 微软亚洲研究院的测试集上，组合歧义消解的 F-measure 为 70.9%。

关键词： 组合歧义；自动分词；带标语料库

Combinative Ambiguity String Detection and Resolution Based on Annotated Corpus

Chengjie Sun[1], Chang-Ning Huang[2], Yi Guan[1]

(1. Harbin Institute of Technology, Harbin 150001; 2.Microsoft Research Asia, Beijing 10008)

Abstract: This paper investigates the distribution of combinative ambiguity string (CAS) in different annotated Chinese word segmentation corpora. Based on these investigations, proposed a new approach to detect and solve combinative ambiguity utilizing the knowledge from a large and high quality annotated Chinese segmentation corpus. Experiments show that the new approach can cover more CAS than other case by case method and the CAS disambiguation F-measure in MSRA test data for SIGHAN Bakeoff2 is 70.9%.

Keywords: combinative ambiguity string; automatic word segmentation; annotated corpus

1 Introduction

Automatic word segmentation is a prerequisite for natural language processing in Chinese. Resolving segmentation ambiguities is one of the fundamental tasks for Chinese word segmentation and has received considerable attention in the research community. Word segmentation ambiguities can be roughly classified into two classes: overlapping ambiguity (OA) and combinative ambiguity (CA) [1]. The researches on CA detection and resolution are still far to satisfaction [2]. We focus on combinative ambiguity in the paper.

CA is firstly defined in [3]. A character string AB is called a combinative ambiguity string (CAS), if either A, B, or AB are lexicon words. Here we use the definition in [4] because this definition is much formal than the one given in [3]. Given a word $w \in W$, W is a Chinese lexicon, if w is a concatenation of multiple words $w_1 \Lambda\ w_n (n \ge 2)$, $w_i \in W (i = 1 \Lambda\ n)$, and in addition, both the word w and the sequence $w_1 \Lambda\ w_n$ can be realized in some sentences,

基金资助：国家自然科学基金（60504021），中国教育部微软基金（01307620）.

作者简介：孙承杰（1980-），男，山东烟台，博士生，E-mail：cjsun@insun.hit.edu.cn.

then we term w a "combinative ambiguity" (we also refer to w an "CAS type"), meanwhile, term the sequence w and $w_1 \Lambda w_n$ the combined form and the separated form of the CAS type w respectively.

Given the definition, people want to know what the percentage of CAS is in real text. One answer was given in [3]. Combinative ambiguities constitute about 10 percent of segmentation ambiguities in Chinese running text. However, investigation in [3] was base on a corpus only including 48,092 Chinese characters at that time. Today we have had annotated corpus including millions of Chinese words, the distribution of CAS can be acquired through the investigation on a larger annotated corpus. In this paper we investigate CAS distribution on 5 different corpora in SIGHAN Bakeoff test[1]. We report the CAS types, instances and tokens in each corpus and give some conclusions drawn from the investigation results.

Previous methods of resolving combinative ambiguity can be grouped into rule-based approach [5] and statistical approaches [4, 6]. This research proposes a method to detect and resolve the combinative ambiguity based on a CAS list learned from a large scale and high quality annotation Chinese segmentation corpus in this paper. Experiments show that the new approach can resolve more CAS with high precision than previous methods in real text.

The remainder of this paper is structured as follows. In section 2, CASs distributions in different corpus are investigated. In section 3, a new approach is proposed to detect and resolve the CASs in running text. Experiment results are shown in section 4. Section 5 is a brief conclusion.

2 CASs distribution in different corpus

In order to know the distribution of CASs in real text, we firstly have our investigations on MSRA training data for Bakeoff2 (hereafter MSRA-train). The reasons lie in three folders. First we have the lexicon for MSRA-train and the lexicon is necessary according to our CA definition in section 1. Second, MSRA-train is a high quality annotated corpus. The error rate of MSRA-train is lower than 0.1% through manual checking on random samples [7]. The last reason is that MSRA-train is a large scale corpus with 2,370,000 word tokens. The investigation results on MSRA-train are shown in Table 1.

Tab. 1 CAS types and tokens distribution in MSRA-train

# of CAS types	Lexicon size	ratio
1,015	93,729	1.08% (1,015/93,729)
	# of Lexicon word type in corpus	ratio
	42,360	2.40% (1,015/42,360)
# of CAS tokens	# of tokens in whole corpus	ratio
85,362	2,370,000	3.60% (85,362/2,370,000)
# of separated form tokens in all CAS tokens	# of tokens in whole corpus	ratio
13,538	2,370,000	0.57%(13.538/2.370,000)

From Table 1, we can see the ratio of number of CAS types to Lexicon size is 1.08%. The figure is very close to 0.98%, i.e. 859 types, reported by Fuji and Nie [8] though an investigation within a machine readable dictionary (87,599 entries). The number of separated form tokens in all CAS tokens only occupied 0.57% in the corpus. This means that if we keep all CAS in combined form, the loss to segmentation recall will not more than 0.57%.

In order to show the CAS distributions in corpora with different sizes, we investigate another 5 corpora: PK-test (Peking University test corpus for Bakeoff-2003), PK-train (Peking University training corpus for Bakeoff-2003), AS-test (Academia Sinica test corpus for Bakeoff-2003), AS-train (Academia Sinica training corpus for Bakeoff-2003) and MSRA-test (MSRA test corpus for Bakeoff-2005). Because we don't have the real lexicon of PK-train, PK-test and AS-train, so we use the lexicon gotten from the corresponding corpus. For MSRA-train and MSRA-test, we obey the

[1] http://www.sighan.org/bakeoff2003/ and http://www.sighan.org/bakeoff2005/

same rule in order to compare with other corpora. Thus the actual CAS figures should be smaller than what in table 2 because we introduce lots of out of vocabulary (OOV) words.

Figure 1 shows that the number of CAS instances and tokens are increase with the increase of corpus scale. The strange thing is that the numbers of CAS types in MSRA-train and PK-train are almost same although the size of MSRA-train is twice as large as PK-train's. We think the reason may lie in that both of the two corpora come from People Daily. In Table 2, the values in the last row are too large compare to other lines. The 11614 CAS types in AS-train consist of 6753 two-character CAS types and 4861 long CAS types (a CAS includes more than two characters). We find out that most of the long CAS types are not real CAS. They are caused by the annotation. For example, "豁然开朗" is segmented as "豁然开朗" and "豁然/开朗" in AS-train. According to our definition, we find "豁然开朗" as a CAS type because of the wrong annotations. We find 74 CAS types turn out to be wrongly segmented strings in randomly select 87 long CAS types; the ratio is about 85%.

Besides the total distribution of CAS in corpora, we also investigate the distribution of CAS instances in a corpus. The results are shown in Table 3. From Table 3, the distribution of CAS tokens in a corpus is skew heavily and fit with Zipf's Law. In another word, about 60-70% of total CAS instances occur only once in a given corpus. Thus the CAS problem is not only hard to be solved with a statistical approach, but also set a precision limitation to word segmentation.

Tab. 2 CASs distribution in different corpora

Corpus name	# of total tokens	# of CAS types	# of CAS instances	# of CAS tokens	separated form CAS tokens ratio
AS-test	12,000	21	100	137	0.62%
PK-test	17,000	21	87	120	0.38%
MSRA-test	106,000	70	798	1029	0.43%
PK-train	1,100,100	1498	35762	42736	1.26%
MSRA-train	2,300,000	1492	84055	93582	0.80%
AS-train	5,800,000	11614	633562	778254	4.90%

$$separated\ form\ CAS\ tokens\ ratio = \frac{\#\ of\ separated\ form\ CAS\ tokens}{\#\ of\ total\ tokens} \qquad (1)$$

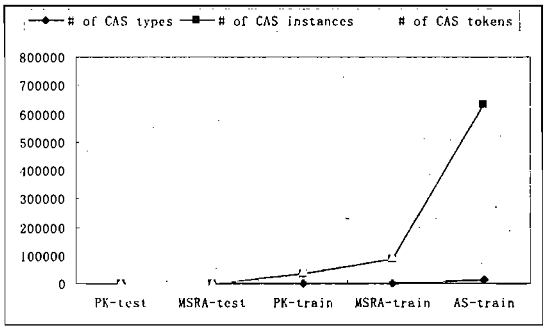

Fig. 1 CASs distribution in different corpus

Tab. 3 distribution of CAS instances in MSRA-train and PK-train

	MSRA-train	PK-train
# of CAS types	1,015	1,498
# of CAS instances occur only once	602	1,071
Ratio1	59.3%	71.5%
# of CAS instances occur less than 3	773	1,269
Ratio2	76.2%	84.7%

3 CASs detection and resolution

In [4], 90 CASs are resolved [0]with high precision. We investigate these CASs in MSRA-train and PK -train. The 21 CAS types listed by [4] but not found in MSRA-train are:*条约 一时 自我 总会 年会 年产 一定 一生 最近 名将 大小 名人 中等 是以 存在 行人 难以 要求 列车 比分 前提*. The 22 CAS types listed by [4] but not found in PK-train are: *学会 一时 已经 自我 下来 年会 到底 东西 部长 一定 前来更新 都会 中学 是以 处在 行人 会同 一度 列车 前提 中长期*.

Tab. 4 the coverage of [4]'s CAS list in MSRA-train

	# of types	# of tokens
All CASs in MSRA-train	1015	85,362
CASs within Luo's list	69	21,168
Ratio	6.80%	24.8%

Tab. 5 the coverage of [4]'s CAS list in PK-train

	# of types	# of tokens
All CASs in PK-train	1,498	42,736
CASs within Luo's list	68	10137
ratio	4.54%	23.7%

Table 4 and 5 show that 90 CAS list only covers a little fraction (lower than 7% types and nearly 25% tokens) of CASs in real corpora (such as MSRA-train and PK-train). Although the CASs in Luo's [4] list have been studied accordingly, but the majority of CASs in real corpora is still need to be studied with specially. Also, the CASs out of Luo's list often occurs sparsely, it is hard to use statistics method to solve them. Because the number of types of these CASs is large, there are also challenges to rule based methods.

We propose an approach based on above investigation. In our approach, we use the CAS list detected from a high quality annotated corpus as the knowledge to detect and resolve CA in new data. The strategy for resolving is majority first, i.e., if the occurrence times of combined form of current CAS is more than the separated form in the training corpus, we choose combined form as the default segmentation form of current CAS. For example, CAS "及其" has 242 combined form and 3 separated form in MSRA-train, we will segment "及其" as one word when we encounter it in new data. Our list includes 1015 CAS types found in Table 1.

4 Experiments and results analysis

In order to show the validity of our approach, we compare our approach with baseline method in four corpora: AS-test, PK-test, PK-train and MSRA-test. The Baseline method is just keeping all CASs in the list in its combined form. The results are shown in Table 6.

Tab. 6 the precision of baseline method and our approach

	AS-test	PK-test	PK-train	MSRA-test
Baseline	0.474	0.481	0.694	0.565
Ours	0.651	0.481	0.795	0.772

Tab. 7 the type coverage ratios of our approach

corpus	AS-test	PK-test	PK-train	MSRA-test
TCR	47.6%	50%	17.8%	67.1%

Tab. 8 the comparison of Luo's method and our approach

		AS-test	PK-test	PK-train	MSRA-test
Luo's method	P	0.966	0.966	0.966	0.966
	R	0.241	0.235	0.238	0.304
	F	0.386	0.378	0.381	0.462
Our Approach	P	0.650	0.532	0.795	0.772
	R	0.394	0.287	0.367	0.656
	F	0.491	0.300	0.502	0.709

From Table 6, we can see our approach has a huge promotion in PK-train and MSRA-test compare with baseline method. In PK-test, the precision of our approach is just comparable with the baseline. This is because the total number of separated form tokens of CASs in PK-test is very small. So it is hard to see the effect of our approach in PK-test. The same situation also occurs in Table 8. The type coverage ratios (TCR) of our approach in different corpus are listed in Table 7. From Table 7, we can see that our approach can cover more CAS type than previous method mentioned in section 3.

We also compare our approach with Luo's method [4]. The results are shown in Table 8. Although our approach's precision is lower than Luo's (we used the highest precision in [4], i.e. 0.966, as its precision in our experiments), the recall is much higher than Luo's. So our approach's F-measure is much higher than Luo's in PK-train and MSRA-test. The P, R and F stand for precision, recall and F-measure respectively. They are calculated with the following formulas:

$$P = \frac{\#\ of\ correctly\ segmentated\ CAS\ tokens}{\#\ of\ total\ segmentated\ CAS\ tokens} \quad R = \frac{\#\ of\ correctly\ segmentated\ CAS\ tokens}{\#\ of\ total\ \ CAS\ tokens} \quad F = \frac{2PR}{P+R}$$

As mention in Table 1, the percentage of all CAS tokens with separated forms is only 0.57% of all CAS tokes. That ratio is the loss to word segmentation system recall if we keep all CAS in combined form. Our approach's loss ratios in each corpus are listed in Table 9, which are lower than separated form CAS tokens ratio. This also shows that our approach is effective in improving the performance of word segmentation system.

Tab. 9 our approach's loss ratio

	AS-test	PK-test	PK-train	MSRA-test
Error CAS- tokens	62	64	12284	316
separated form CAS tokens ratio	0.62%	0.38%	1.26%	0.43%
Our approach's loss ratio	0.52%	0.37%	1.12%	0.30%

5 Conclusion

A new approach for CAS detection and resolution based on large high quality annotated corpus is proposed in this paper. The main contributions of this paper are: 1) Report the distribution features of CAS in different scale corpus. Unlike the common notion that CAS is only a kind of rare ambiguities in Chinese word segmentation, the CAS tokens reach 3.6% of total tokens in the corpus. 2) Table 4 shows that the coverage of 90 CAS list [4] is 6.80% in types and 24.8% in tokens respectively within MSRA-train corpus. It means that only one fourth of the CAS tokens or 7% of CAS types have been studied by researchers so far. 3) Propose a new approach to detect and resolve CAS in real text. Our approach just needs a large scale and high quality annotated corpus. Experiments show the effective of our approach. In MSRA-test, the F-measure is 70.9%.

Reference

[1] Mu Li, Jianfeng Gao, Changning Huang and Jianfeng Li. 2003. Unsupervised training for overlapping ambiguity resolution in Chinese word segmentation[A]. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing [C], Japan, 2003. 1-7.

[2] 孙茂松, 邹嘉彦.汉语自动分词研究评述[J]. 当代语言学, 2001, (1): 22-32.

[3] 梁南元. 书面汉语自动分词系统—CDWS[J]. 中文信息学报, 1987, (2): 44-52.

[4] Xiao Luo, Maosong Sun and Benjamin K Tsou. Covering ambiguity resolution in Chinese word segmentation based on contextual information[A]. In Proceedings of COLING2002[C]. Taipei, 2002. 598~604.

[5] Andi Wu and Zixin Jiang. Word Segmentation in Sentence Analysis[A]. In Proceedings of the 1998 International Conference on Chinese Information Process[C]. Beijing, China, 1998. 169-180.

[6] 曲维光, 董宇, 陈钟等. 基于语境计算模型的词义消歧[A]. 孙茂松, 陈群秀. 自然语言理解与大规模内容计算[C]. 北京：清华大学出版社, 2005. 134-139.

[7] Chengjie Sun, Chang-ning Huang, Xiaolong Wang et al. Detecting segmentation errors in Chinese annotated corpus[A]. Proceedings of Fourth Workshop on Chinese Language Processing[C]. Jeju island, Korea, 2005. 1-8.

[8] Fuji Ren, Jianyun Nie. The concept of sensitive word in Chinese--Survey in a machine readable dictionary[J]. Journal of Natural Language Processing (Published in Japan), 1999, 6(1): 59-78.