Synthetic Word Parsing Improves Chinese Word Segmentation

Fei Cheng **Kevin Duh** **Yuji Matsumoto**
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0192, Japan
{fei-c,kevinduh,matsu}@is.naist.jp

Abstract

We present a novel solution to improve the performance of Chinese word segmentation (CWS) using a synthetic word parser. The parser analyses the internal structure of words, and attempts to convert out-of-vocabulary words (OOVs) into in-vocabulary fine-grained sub-words. We propose a pipeline CWS system that first predicts this fine-grained segmentation, then chunks the output to reconstruct the original word segmentation standard. We achieve competitive results on the PKU and MSR datasets, with substantial improvements in OOV recall.

1 Introduction

Since Chinese has no spaces between words to indicate word boundaries, Chinese word segmentation is a task to determine word boundaries between characters. In recent years, research in Chinese word segmentation has progressed significantly, with state-of-the-art performing at around 96% in precision and recall (Xue, 2003; Zhang and Clark, 2007; Li and Sun, 2009).

However, frequent OOVs are still a crucial issue that causes low accuracy in word segmentation. Li and Zhou (2012) defined those words that are OOVs but consisting of frequent internal parts as pseudo-OOV words and estimated that over 60% of OOVs are pseudo-OOVs in five common Chinese corpora. For instance, PKU corpus does not contain the word 陈列室 (exhibition room), even though the word 陈列 (exhibit) and 室 (room) appear hundreds of times. Goh et al. (2006) also claimed that most OOVs are proper nouns taking the form of Chinese synthetic words.

These previous works suggest that by analysing the internal structure of the synthetic words, we can transform pseudo-OOVs into in-vocabulary words (IVs). By running a synthetic word parser on each of the words in a CWS training set, we can generate a fine-grained segmentation standard that contains more IVs. Since the current conditional random field (CRF) word segmenters (Tseng et al., 2005; Sun and Xu, 2011) perform well on IVs, this transforming process can conceivably improve the handling of pseudo-OOV words, as long as we can recover the original word segmentation standard from the fine-grained sub-word segmentation.

In recent years, some related works about improving OOV problem in CWS have been ongoing. Sun et al. (2012) presented a joint model for Chinese word segmentation and OOVs detection. Their models achieved fast training speed, high accuracies and increase on OOV recall. Sun (2011) proposed a similar sub-word structure which is generated by merging the segmentations provided by different segmenters (a word-based segmenter, a character-based segmenter and a local character classifier). However, her models does not predict the sub-words of all the synthetic words, but those words with different segmented results of the three segmenters. Her work maximizes the agreement of different models to improve CWS performance. Different from her work, we aim to provide an unified way to incorporate morphological information of the synthetic words into the CWS task.

In this paper, we propose a pipeline word segmentation system to address the pseudo-OOV problem. Our word segmentation system first converts the original training data into a fine-grained standard by parsing all words with a synthetic word parser (Section 2.1), then trains a CRF-based sub-word segmenter (Section 2.2). A second CRF chunker is trained to recover the original word segmentation given the fine-grained results of the first CRF. The intuition is that fine-grained sub-word segmentations resolve pseudo-OOVs into IVs, which are easier to predict correctly by the first CRF. Secondly, by training an-

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 262–267,
Beijing, China, July 26-31, 2015.

other CRF that predicts the original word segmentation given the fine-grained segmentation as input, we can recover the fine-grained output into original word segmentation standard (Section 2.3). The flow chart of our word segmentation system is shown in Figure 1.

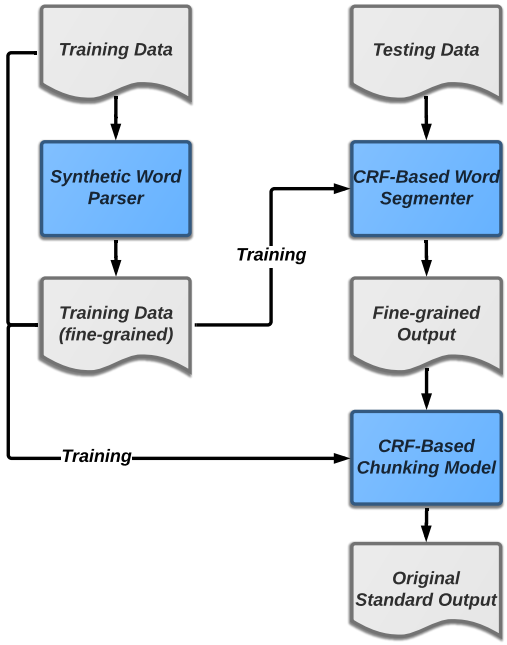


Figure 1: The Flow Chart of the Chinese Word Segmentation System.

2 System Components

2.1 Synthetic Word Parser

Intuitively, Chinese synthetic words contain internal morphological information that is helpful to recognize OOVs. Cheng et al. (2014) proposed a character-based parser to parse the internal tree structure of words. For instance, the tree and flat segmented result of the word 市政府 (municipal government) are shown in Figure 2. In this work, we train a graph-based parser (McDonald, 2006) on the data released by Cheng et al. (2014) and include the dictionary (NAIST Chinese Dictionary[1]) features and Brown clustering features extracted from a large unlabeled corpus (Chinese Gigaword Second Edition[2]) as described in Cheng et al. (2014).

For native Chinese speakers, single character and two character words are usually treated as the smallest units. In this work, we parse all the words in the PKU and MSR training data with character length greater than two. By replacing the words with the flat segmented results, we convert the training data into a fine-grained word segmentation standard as shown in Figure 3.

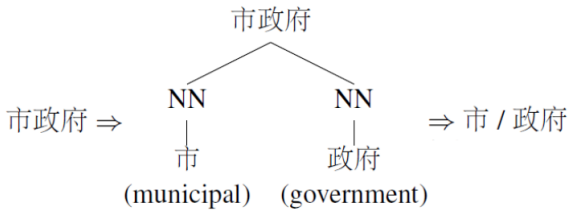


Figure 2: The Tree Structure of a Sample Word and the Flat Segmented Result.

Original CWS tags	市政府 / 办公厅 / 等 / 单位 B I E / B I E / S / B E
Fine-grained CWS tags	市 / 政府 / 办公 / 厅 / 等 / 单位 S / B E / B E / S / S / B E

Figure 3: A Sample Sentence of Labeling Chinese word segmentation tags on the Original and Fine-grained Standard. In this work, we adopt 4-tag set for word segmentation. "B" denotes the beginning character of a word. "I" denotes the middle character of a word. "E" denotes the end character of a word. "S" denotes a single character word.

2.2 CRF-based Word Segmenter

Xue et al. (2003) proposed a method which treated Chinese word segmentation as a character-based sequential labeling problem and exploited several discriminative learning algorithms. Tseng et al. (2005) adopted the CRFs as the learning method and obtained the best results in the second international Chinese word segmentation bakeoff-2005. Moreover, Sun and Xu (2011) attempted to extract information from large unlabeled data to enhance the Chinese word segmentation results.

In this work, we train a traditional CRF-based supervised model on the fine-grained training data, include the dictionary (NAIST Chinese Dictionary) features and access variety features extracted from a large unlabeled corpus (Chinese Gigaword Second Edition) as described in Sun and Xu (2011).

[1] http://cl.naist.jp/index.php?%B8%F8%B3%AB%A5%EA%A5%BD%A1%BC%A5%B9%2FNCD

[2] https://catalog.ldc.upenn.edu/LDC2005T14

2.3 CRF-based Chunking Model

In order to obtain the word segmentation result with original word segmentation standard, we train a CRF-based chunking model on the original and fine-grained training data. We show a sample sentence of labeling chunking tags in Figure 4. Comparing two sentences, we label all common units with the tag "S". The words 市 and 政府 are tagged as "B" and "E", since 市 is the beginning part of the synthetic word 市政府 and 政府 is the ending part. In the chunking process, the frequent prefix 市 is coordinated with neighbouring units to compose the synthetic word 市政府.

For each labeling, we include previous, current and next word as the features for the chunking model.

Original	市政府 / 办公厅 / 等 / 单位
Fine-grained	市 / 政府 / 办公 / 厅 / 等 / 单位
Chunking tags	B / E / B / E / S / S

Figure 4: A Sample Sentence of Labeling Chunking Tags. In this work, we adopt 4-tag set for chunking. "B" denotes the beginning part of a synthetic word. "I" denotes the middle part. "E" denotes the end part. "S" denotes a single word.

3 Experiments

3.1 Settings

Cheng et al. (2014) released a dictionary of 31,849 synthetic words with internal structure annotated. Since transliteration words (e.g. 贝克汉姆 Becham) exist in Chinese, our synthetic word parser should perform well not only on synthetic words but also on transliteration words. We extracted 6,574 transliteration words from the NAIST Chinese Dictionary and automatically assigned flat structures for these words. As a result, we obtained 38,423 words as the training data for our parser.

The second international Chinese word segmentation bakeoff-2005 provided two annotated simplified Chinese corpora: PKU and MSR. We conducted all word segmentation experiments on these two corpora.

We used CRF++[3] (version 0.58) as the implementation of CRFs in our experiments with the default regularization algorithm L2.

[3]The CRF++ package can be found in the following website: http://taku910.github.io/crfpp/

3.2 Word Segmentation Results

Table 1 summarizes the word segmentation results on PKU and MSR corpora. For comparison, we give a baseline result by training a CRF word segmenter on the original PKU and MSR data sets with the same features. Our proposed system is expected to improve the word segmentation performance on pseudo-OOVs. Compared to the baseline, there are significant increases on OOV recall from 0.792 to 0.822 on PKU and 0.682 to 0.717 on MSR. We also evaluated the pseudo-OOV recall and observed 4% increases from the baseline to the proposed system. Our proposed system achieves higher F-score with 0.961 on PKU and 0.971 on MSR. Comparing to other systems, our proposed method obtains the state-of-the-art F-score as the results of Zhang et al. (2013) who extracted dynamic statistical features from both in-domain and out-domain corpus and our OOV recall significantly outperforms theirs with a 9% lead. In MSR, we obtain very close OOV recall and slightly lower F-score than the state-of-the-art system (Sun et al., 2009), which adopted a latent variable CRF model. However, our system significantly outperforms their system in PKU. In both corpora, our proposed system outperforms the best "Bakeoff-2005" results.

We also test the statistical significance of the results by using the criterion (Sproat and Emerson, 2003; Emerson, 2005). The 95% confidence interval is given as $\pm 2\sqrt{p(1-p)/n}$, where n is the number of words in the test data. They treat two systems as significantly different (at the 95% confidence level), if at least one of their precision-based confidences "C_p" or recall-based "C_r" are different. As the results shown in Table 2, the baseline and proposed method are significantly different on precision and recall in both PKU and MSR corpus. In conclusion, our proposed method significantly outperforms the baseline.

3.3 Additional Experiments

We conducted additional experiments to evaluate the performance of the synthetic word parser and CRF-based chunking model.

First, we are interested in how much parsing accuracy is needed for good results. Figure 5 displays the OOV recall results of our word segmentation system when the synthetic word parser is trained with amounts of labeled synthetic words data. As the data size increases, our word segmen-

System	PKU					MSR				
	P	**R**	**F**	$\mathbf{R_{oov}}$	$\mathbf{R_{pseudo}}$	**P**	**R**	**F**	$\mathbf{R_{oov}}$	$\mathbf{R_{pseudo}}$
Baseline	0.957	0.960	0.959	0.792	0.797	0.971	0.968	0.970	0.682	0.689
Proposed method	0.960	**0.962**	**0.961**	**0.822**	0.838	0.972	0.970	0.971	0.717	0.73
Zhang et al. (2013)	**0.965**	0.958	**0.961**	0.731	-	-	-	-	-	-
Sun et al. (2009)	0.956	0.948	0.952	0.778	-	**0.973**	**0.973**	**0.973**	**0.722**	-
Bakeoff-2005	0.953	0.946	0.950	0.636	-	0.962	0.966	0.964	0.717	-

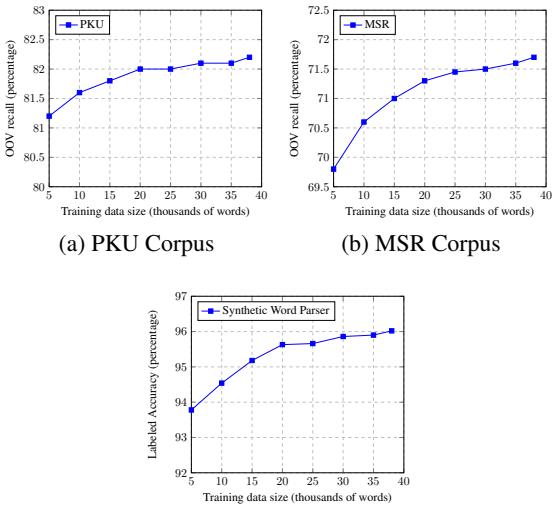
Table 1: Comparison of the Proposed Method to the Baseline and Previous works on PKU and MSR Corpora. Here, "R_{pseudo}" denotes the recall of pseudo-OOV words. "Bakeoff-2005" denotes the best results of the second international Chinese word segmentation bakeoff-2005 on two corpora. Since we use extra resources and our proposed method replies on the synthetic word parser trained on an dictionary with internal structure annotated, the results cannot be directly compared with the state-of-the-art systems.

System	PKU					MSR				
	Words	**P**	$\mathbf{C_p}$	**R**	$\mathbf{C_r}$	**Words**	**P**	$\mathbf{C_p}$	**R**	$\mathbf{C_r}$
Baseline	104372	0.957	±0.00126	0.960	±0.00121	106873	0.971	±0.00103	0.968	±0.00108
Proposed	104372	0.960	±0.00121	0.962	±0.00118	106873	0.972	±0.00101	0.970	±0.00104

Table 2: The Statistical Significance Test of the Word Segmentation Results on PKU and MSR Corpora.

tation system obtains consistent gains on OOV recall on both corpora. On the whole 38K words training data, our system reaches the highest OOV recall. An interesting observation is that the OOV recall on MSR is more sensitive on data size changing. The main reason is the different annotation standard of the two corpus. PKU is a correspondingly fine-grained annotated corpus with shorter average word length than MSR. Our synthetic word parser reaches high parsing accuracy on short length words (three-character and four-character words) even with a small training data size. With the increase of word length, the parser needs more training data. These factors cause that our system reaches high OOV recall on PKU starting from a small training data size and obtains more OOV recall gains on MSR when increasing the training data size.

Our pipeline system adopts a chunking model to recover the original standard from the fine-grained standard. One question is how difficult is this task. Unfortunately, we do not have the gold fine-grained input to evaluate the performance of our chunking model directly; i.e. it is not clear whether a segmentation error is due to mis-predictions in the first or second CRF. Therefore, we use the synthetic word parser to parse all the words in the gold testing data and generate an artificial gold fine-grained input for the chunking model. This data keeps the original word boundaries and can be used to observe the chunking performance. Table 3 shows that the chunking model on the artificial data obtains a 0.822 to 0.847 improvement in OOV recall. We can interpret this to mean that 0.025 improvement is possible if the first CRF was perfect; on the other hand, the gap between 0.847 and 1.0 shows that potentially the second CRF is a harder task. However, the real



(a) PKU Corpus (b) MSR Corpus

(c) Parsing Performance

Figure 5: The OOV Recall Evaluation and the Character Labeled Accuracy (5-fold cross-validation) of the Synthetic Word Parser on Training Data Size.

gap is less for the lose of the parsing step and the existence of non-pseudo OOVs.

System	PKU		MSR	
	F	$\mathbf{R_{oov}}$	**F**	$\mathbf{R_{oov}}$
Proposed	0.961	0.822	0.971	0.717
Artifical gold	0.965	0.847	0.973	0.743

Table 3: The Word Segmentation evaluation of the Chunking Model. "Artificial gold" denotes the word segmentation result when the chunking model runs on the artificial gold input.

3.4 Analysis

As we expected, the proposed method obtains significant improvement on OOV recall. In both corpora, we observed a number of OOVs are segmented correctly. For instance, 管理法 (management law) is an OOV word in PKU corpus. In this word, 管理 (management) appears frequently and 法 (law) is a common suffix in Chinese synthetic words, such as 行政法 (administrative law) or 国际法 (international law). This type of pseudo-OOVs share a major contribution to upgrade the system performance. We also observed that some polysemous words bring ambiguities to the chunking step. The character 会 carries the meanings "will" as an auxiliary verb or "meeting" in a synthetic word 运动会 (sports meeting).

4 Conclusion

In this paper, we presented a series processes to reduce OOV rate and extract morphological information inside Chinese synthetic words on a fine-grained word segmentation standard. As a result, we can improve the Chinese word segmentation performance (especially on pseudo-OOVs) without introducing any new feature types. Our proposed method achieved the state-of-the-art F-score and OOV recall on two common corpus PKU and MSR. However, note that we only exploited the flat segmented results of internal word structure here. As future work, we plan to exploit the full tree structure of synthetic words to improve not only CWS but also additional downstream tasks such as sentence parsing.

References

Fei Cheng, Kevin Duh, and Yuji Matsumoto. 2014. Parsing chinese synthetic words with a character-based dependency model. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Thomas Emerson. 2005. The second international chinese word segmentation bakeoff. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*, volume 133.

Chooi-Ling Goh, Masayuki Asahara, and Yuji Matsumoto. 2006. Machine learning-based methods to chinese unknown word detection and pos tag guessing. *Journal of Chinese Language and Computing*, 16(4):185–206.

Zhongguo Li and Maosong Sun. 2009. Punctuation as implicit annotations for chinese word segmentation. *Computational Linguistics*, 35(4):505–512.

Zhongguo Li and Guodong Zhou. 2012. Unified dependency parsing of chinese morphological and syntactic structures. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1445–1454. Association for Computational Linguistics.

Ryan McDonald. 2006. *Discriminative learning and spanning tree algorithms for dependency parsing*. Ph.D. thesis, PhD Thesis. University of Pennsylvania.

Richard Sproat and Thomas Emerson. 2003. The first international chinese word segmentation bakeoff. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, pages 133–143. Association for Computational Linguistics.

Weiwei Sun and Jia Xu. 2011. Enhancing chinese word segmentation using unlabeled data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 970–979. Association for Computational Linguistics.

Xu Sun, Yaozhong Zhang, Takuya Matsuzaki, Yoshimasa Tsuruoka, and Jun'ichi Tsujii. 2009. A discriminative latent variable chinese segmenter with hybrid word/character information. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 56–64. Association for Computational Linguistics.

Xu Sun, Houfeng Wang, and Wenjie Li. 2012. Fast online training with frequency-adaptive learning rates for chinese word segmentation and new word detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 253–262. Association for Computational Linguistics.

Weiwei Sun. 2011. A stacked sub-word model for joint chinese word segmentation and part-of-speech tagging. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1385–1394. Association for Computational Linguistics.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*, volume 171.

Nianwen Xue. 2003. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing*, 8(1):29–48.

Yue Zhang and Stephen Clark. 2007. Chinese segmentation with a word-based perceptron algorithm. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 840.

Longkai Zhang, Houfeng Wang, Xu Sun, and Mairgup Mansur. 2013. Exploring representations from unlabeled data with co-training for Chinese word segmentation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 311–321, Seattle, Washington, USA, October. Association for Computational Linguistics.