Towards a Consistent Segmentation Level across Multiple Chinese Word Segmentation Corpora

Fei Cheng[†], Kevin Duh[††] and Yuji Matsumoto[†]

One of the crucial problems facing current Chinese natural language processing (NLP) is the ambiguity of word boundaries, which raises many further issues, such as different word segmentation standards and the prevalence of out-of-vocabulary (OOV) words. We assume that such issues can be better handled if a consistent segmentation level is created among multiple corpora. In this paper, we propose a simple strategy to transform two different Chinese word segmentation (CWS) corpora into a new consistent segmentation level, which enables easy extension of the training data size. The extended data is verified to be highly consistent by 10-fold cross-validation. In addition, we use a synthetic word parser to analyze the internal structure information of the words in the extended training data to convert the data into a more fine-grained standard. Then we use two-stage Conditional Random Fields (CRFs) to perform fine-grained segmentation and chunk the segments back to the original Peking University (PKU) or Microsoft Research (MSR) standard. Due to the extension of the training data and reduction of the OOV rate in the new fine-grained level, the proposed system achieves state-of-the-art segmentation recall and F-score on the PKU and MSR corpora.

Key Words: *Chinese Word Segmentation, Synthetic Words, Internal Structure*

1 INTRODUCTION

In Chinese, a sentence is written as continuous characters without distinct word boundaries. Therefore, Chinese word segmentation (CWS) is commonly treated as the first step in the natural language processing (NLP) pipeline, before part-of-speech (POS) tagging, parsing, and other processes. Unfortunately, there is no clear and intuitive notion of 'word' in Chinese. A highly controversial part is that Chinese synthetic words have quite complex structures and can be represented by several segmentation levels. For example, the Second International Chinese Word Segmentation Bakeoff 2005 (Bakeoff-2005) provided two annotated simplified Chinese corpora, i.e., the Peking University (PKU) and Microsoft Research (MSR) corpora. In the MSR corpus, the synthetic word 中国国家广播电台 (China Radio International) is considered as a single word. While in the PKU corpus, it appears as four words 中国 (China) / 国家 (International) / 广播

[†] Nara Institute of Science and Technology
[††] Johns Hopkins University

(Broadcast) / 电台 (Station).

In recent years, Chinese word segmentation has progressed significantly and has achieved state-of-the-art performance of approximately 96–97 F-score on the Bakeoff-2005 data. However, some issues remain and we summarize two in the following.

1. **Variety of word segmentation standards**

   Due to the difficulty in defining 'word' in Chinese, resources are annotated in compliance with different 'rules' specified by the providers. In Bakeoff-2005, Peking University supplied a very specific 19-page segmentation guideline for the PKU corpus. For instance, one clause indicates that most community, institution and organization names are synthetic words that should be segmented. Therefore, the word 北京国安队 is segmented into 北京 (Beijing) / 国安 (Guoan) / 队 (club). On the other hand, according to the MSR corpus guideline, a named entity such as 北京国安队 is treated as a single word. Ambiguous word segmentation standards not only make it hard to share current annotated resources, but also have an adverse impact on downstream Chinese NLP tasks.

2. **Low recall of out-of-vocabulary (OOV) words**

   Frequent out-of-vocabulary (OOV) words are a crucial issue that causes low word segmentation accuracy. Li and Zhou (2012) defined those words that are OOVs but consisting of frequent internal parts, i.e., in-vocabulary (IV) words as pseudo-OOVs and estimated that more than 60% of OOVs are pseudo-OOVs in five common Chinese corpora. For instance, the PKU training data 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, Asahara, and Matsumoto (2006) also claimed that most OOVs are proper nouns that take the form of Chinese synthetic words. These studies suggested that analyzing the internal structure of Chinese synthetic words has the potential to improve the OOV problem.

Both issues can be resolved by finding a suitable segmentation level that is consistent across multiple corpora and where a part of OOVs are naturally segmented into IV words. Here, segmentation level is defined as any middle segmentation standard between the most fine-grained (character) and most coarse-grained standard (original corpora). A consistent level is expected to be not only more consistent across different corpora, but also more consistent inside each individual corpus. For instance, as shown in Figure 1, two words 石油天然气集团公司 (Oil and Gas Corporation) and 天然气 (natural gas) inconsistently exist in Standard 1, because 石油天然气集团公司 is a named entity that is treated as a single word. In the consistent level, 石油天然气集团公司 takes a more natural standard 石油 / 天然气 / 集团公司 close to 天然气.

Even in the consistent level, many synthetic words, such as 天然气 and 集团公司, still exist. A

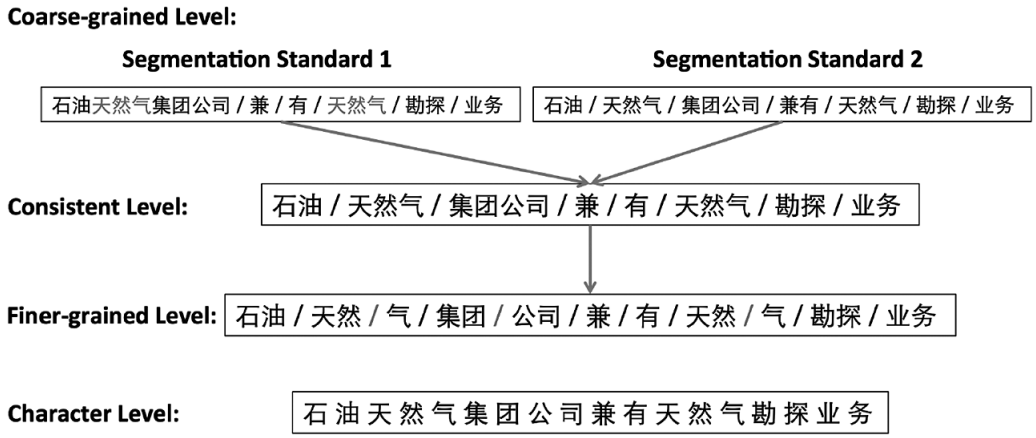


Fig. 1 An example of a sentence in several different segmentation level.

synthetic word parser is used to analyze the internal structure of 天然气 (a possible OOV word) and generate the flat sub-word segmentation 天然 / 气 (natural / gas). Our goal is to create a strategy to find a consistent level automatically and extend training data using heterogeneous data. Then, the internal word structure information helps convert the extended data to a finer-grained level (to reduce OOVs).

In this paper, we propose a pipeline word segmentation system that adapts two different corpora, i.e., PKU and MSR into one consistent segmentation level and improves segmentation performance on each individual corpus. In Section 3, we describe two components of the proposed method. In Section 4.1, we explain how the proposed method maps two different word segmentation corpora to a consistent segmentation level. We further explain how we achieve finer-grained segmentation (to reduce OOVs) using synthetic word parsers (Section 4.2), and how we transform the segments back to the PKU and MSR standards (Section 4.3). In Section 5.3, we compare the final segmentation results obtained by the proposed system to a baseline and state-of-the-art systems.

2 RELATED WORK

In recent years, several studies have investigated multiple segmentation levels. Sun (2011) proposed a sub-word structure that is generated by merging the segmentations provided by different segmenters, i.e., a word-based segmenter, a character-based segmenter and a local character classifier. However, their model does not truely investigate the sub-word structures of all syn-

thetic words, but only those cases with disagreements among the three segmenters. Cheng, Duh, and Matsumoto (2015) presented a method to transform CWS corpora to a fine-grained level by parsing the words with a synthetic word parser. Although these previous studies demonstrated positive results on each individual corpus, it is not guaranteed that different corpora can reach consistent in their segmentation levels.

Another research line is to boost word segmentation involves incorporating heterogeneous data. Jiang, Huang, and Liu (2009) presented a simple strategy to train a source segmenter to segment the target corpus. Then the proposed segmenter is trained on the target corpus including 'source-style' predictions as guide-features. Their method is similar to the ideas in domain adaptation (Daumé III and Marcu 2006; Daume III 2007). Chao, Li, Chen, and Zhang (2015) proposed a coupled Conditional Random Fields model to exploit multiple heterogeneous data to improve segmentation performance on Weibo data. Although our proposed pipeline method contains a similar "source-style" prediction step, the following strategy to find a consistent segmentation level differs from simply treating predictions as guide-features in their work. Our work aims to do a more detailed investigation on the conversions between different segmentation levels. In addition, our idea of a consistent segmentation level is friendly to introduce synthetic words parsing to boost the segmentation performance further.

Recently, studies that explored the use of the internal structures of words to improve Chinese processing have shown promising results. Li and Zhou (2012) claimed the importance of word structures. They proposed a new parsing paradigm, in which the internal structures of words are identified. Zhang, Zhang, Che, and Liu (2013) manually annotated the internal structures of 37,382 words, which covers the entire Chinese TreeBank 5 (CTB5). Then, they constructed a shift-reduce parser with customized actions to jointly perform word segmentation, part-of-speech tagging, and parsing. Their system significantly outperformed current pipeline methods. However, these studies relied on prior knowledge of internal structure information, which is manually annotated. In this work, we employ an automatic parsing mechanism to analyze the internal structures of words to improve word segmentation performance.

3 COMPONENTS

3.1 CRF-based Word Segmenter

Character-based labeling is a dominant approach for Chinese word segmentation. Xue (2003) first proposed a method that treated Chinese word segmentation as a character-based sequential labeling problem and exploited several discriminative learning algorithms. Tseng, Chang,

Andrew, Jurafsky, and Manning (2005) adopted CRFs as the learning method and obtained the best results in Bakeoff-2005. Sun and Xu (2011) attempted to extract statistical information from large unlabeled data to enhance CWS performance. Recently, Liu, Duh, Matsumoto, and Iwakura (2014) introduced character vector representations as a new cluster-based feature for a CRF segmenter. These feature types successfully improved current CWS performance. In this work, we adopt a CRF-based model with state-of-the-art features as a basic segmenter for our pipeline processes. Note that the basic segmenter can also provide the baseline results for comparison.

To demonstrate the features easily, we denote a current character c_i with a context $[...c_{i-2}c_{i-1}c_i c_{i+1}c_{i+2}...]$. $c_{[s:e]}$ denotes a character sequence that starts from c_s and ends at c_e.

- **Character context:**
  character uni-gram: $c_s(i-3<s<i+3)$
  character bi-gram: $c_s c_{s+1}(i-3<s<i+2)$
  whether c_s and c_{s+1} are identical, for $(i-2<s<i+2)$
  whether c_s and c_{s+2} are identical, for $(i-4<s<i+2)$
- **Dictionary:**
  The identity of $c_{[s:i]}(i-5<s<i)$, if it matches a word in a dictionary.
  The identity of $c_{[i:e]}(i<e<i+5)$, if it matches a word in a dictionary.
- **Access Variety:** Feng, Chen, Deng, and Zheng (2004) first introduced accessor variety (AV) to identify meaningful Chinese words. The number of distinct occurrences (i.e., AV value) of character types before or after a target character is an important statistical indicator to evaluate how likely word boundaries surround it. The AV value of a character sequence s is defined as follows (Zhao and Kit 2008).

$$AV(s) = min\{L_{av}(s), R_{av}(s)\} \tag{1}$$

$$f(s) = t, \quad \text{if} \quad 2^t \leq AV(s) < 2^{t+1} \tag{2}$$

  where t is an integer to logarithmize the score. L_{av} and R_{av} are the left and right AV values of a character sequence s, respectively, i.e., the number of its distinct predecessor and successor characters. f(s) is calculated based on AV(s) and used as features. The following features are adopted:
  The $f(s)$ score of a character sequence $c_{[s:i]}(i-5<s<i)$
  The $f(s)$ score of a character sequence $c_{[i:e]}(i<e<i+5)$
- **Character Sequence Vector:** (Liu et al. 2014) introduced vector representations of

character sequences as features. They split a sequence into uni-gram $[C_{i-2}, C_{i-1}, C_i, C_{i+1}, C_{i+2}]$, bi-gram $[C_{i-2}C_{i-1}, C_{i-1}C_i, C_iC_{i+1}, C_{i+1}C_{i+2}]$ and tri-gram $[C_{i-2}C_{i-1}C_i, C_{i-1}C_iC_{i+1}, C_iC_{i+1}C_{i+2}, C_{i+1}C_{i+2}C_{i+3}]$. We separately train dense vectors on these sequences by using word2vec. Then, cluster number features are derived by applying a K-mean method on these vector representations.

uni-gram cluster: $c_s(i-3 < s < i+3)$

bi-gram cluster: $c_sc_{s+1}(i-3 < s < i+2)$

tri-gram cluster: $c_sc_{s+1}c_{s+2}(i-4 < s < i+2)$

3.2 Synthetic Word Parser

It is generally considered that in Chinese, there is no clear notion of 'word'. However, for native Chinese speakers, a word is a lexical entry that represents a complete meaning. Words can be classified as single-morpheme (i.e., the meaning cannot be decomposed, e.g., "Olympic" in Figure 2) or synthetic (i.e., the meaning is composed from multiple individual components, e.g., "Olympic games" in Figure 2). Synthetic words are compositional in the sense that, even if we have not seen the entire word before, we may be able to infer the meaning from its parts. Synthetic word parsing is the process of inferring this internal structure. Building a synthetic word parse tree on character sequences is analogous to building a sentential parse tree on word sequences.

Intuitively, internal structure information is potentially helpful to transform the current word segmentation standard to arbitrary levels. Cheng, Duh, and Matsumoto (2014) proposed a character-based dependency representation to analyze the internal tree structure of words. They investigated the performance of both transition-based (Yamada and Matsumoto 2003; Nivre 2003) and graph-based (McDonald 2006) parsers. Several feature types extracted from extra resources are used to enhance parsing results. A representation of the example word 奥林匹克

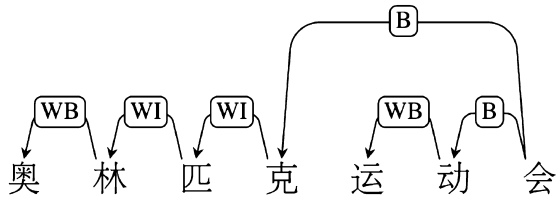


Fig. 2 Character-based dependency representation of an example word. *WB* denotes the beginning character of a single-morpheme word. *WI* denotes the other parts of a single-morpheme word. *B* denotes the branching relation between two words.

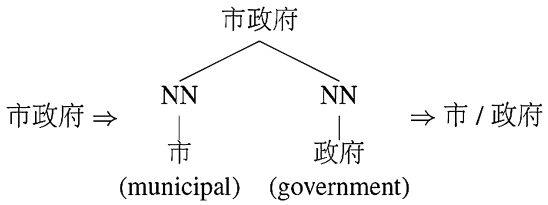


Fig. 3 Internal tree structure of an example word and the flat sub-word segmentation output.

运动会 (Olympic games) is shown in Figure 2. 奥林匹克 with arcs (*WB WI WI*) represents the transliterated word 'Olympic'. 运动会 (sports competition) is composed of two words 运动 (sports) and 会 (competition). 奥林匹克 and 运动会 take a branching relation between them.

In their subsequent work (Cheng et al. 2015), they used the synthetic word parser to analyze the internal structures of words in a corpus and transform the corpus into a fine-grained segmentation level. For instance, an input word 市政府 (municipal government) is parsed into a tree structure and a flat sub-word segmentation 市/政府 (municipal/government) is the output, as shown in Figure 3. Their experiments demonstrated an overall improvement, particularly for OOV recall.

Cheng et al. (2014) released a dictionary of 38,432 synthetic words for which the internal tree structures are manually annotated (Figure 2). In this work, we train a second-order graph-based (McDonald 2006; Carreras 2007) synthetic word parser on this data with the same settings as described in (Cheng et al. 2014). The parser is used to parse the words in a corpus and the internal trees help construct finer-grained data.

For native Chinese speakers, single character and 2-character words are typically treated as basic units. In this paper, the synthetic word parser is only used to parse words with length equal to or greater than 3 characters.

4 METHODS

4.1 Consistent Segmentation Level for PKU and MSR Corpora

In the proposed model, the first step is to find a consistent segmentation level for multiple CWS corpora. In Figure 4, the raw sentences in the MSR train are segmented by a segmenter trained on PKU. We refer to the output as PKU-level MSR train. Our strategy finds a new segmentation level by including word boundaries that appear in either the original MSR annotation or the PKU-level MSR train (Figure 5). The same process is performed on the PKU side. The new

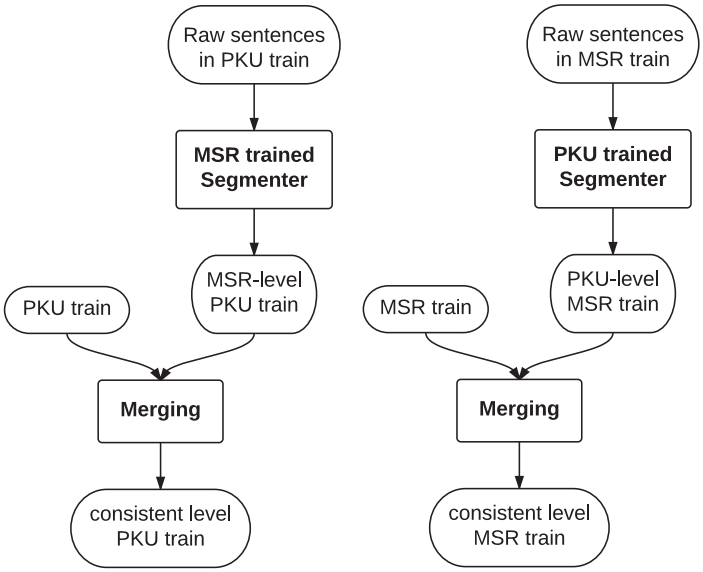


Fig. 4 Workflow of the proposed method to find a consistent segmentation level of multiple CWS corpora. 'PKU train' denotes the original annotated PKU training data and 'MSR train' denotes the original annotated MSR training data.

Raw sentence | 成 都 铁 路 局 开 行 行 包 专 列
Original MSR | 成 都 铁 路 局 / 开 / 行 / 行 / 包 / 专 列
PKU-level MSR | 成 都 / 铁 路 局 / 开 行 / 行 包 / 专 列
Consistent Level | 成 都 / 铁 路 局 / 开 / 行 / 行 / 包 / 专 列

Fig. 5 Example of the strategy to find a new consistent segmentation level.

segmentation level maximizes the number of word boundaries based on the annotation standards of the original corpus and the other corpus, which is expected to be fine-grained compared to the two original standards. We hypothesize that the new segmentation level MSR and PKU training data are consistent and can be easily combined into a larger training dataset.

4.2 Finer-grained Conversion using Synthetic Word Parser

As mentioned in Section 3.2, (Cheng et al. 2015) demonstrated that a fine-grained segmentation level improves word segmentation performance due to the morphological information and low OOV rate. Intuitively, the consistent segmentation level data can be further converted into a finer-grained level using a synthetic word parser. In this work, we train a graph-based parsing model on 38,432 synthetic words with annotated internal structures to perform the conversion.

In this work, the words longer than two characters in the new consistent level data provided by the previous stage are parsed by the synthetic word parser. With the flat sub-word segmentation of each word, the consistent level data are converted to a finer-grained level (Figure 6).

4.3 Finer-grained Word Segmentation and Chunking to Original Segmentation Level

After obtaining finer-grained PKU and MSR data, we combine the data into a larger training dataset. Then, the first-stage CRF-based segmenter predicts the fine-grained output of the test data. The second-stage CRF-based chunker is used to recover the fine-grained output to the original segmentation level. The workflow of this step is shown in Figure 7.

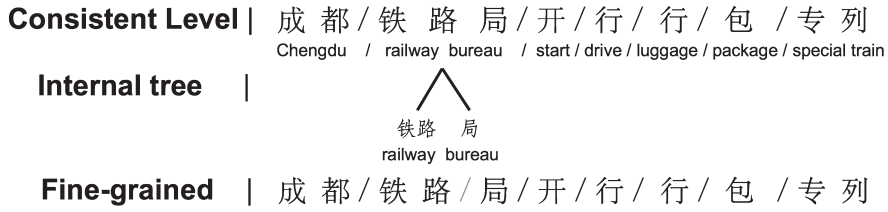


Fig. 6 Example of a sentence with the consistent segmentation level converted to a finer-grained level.

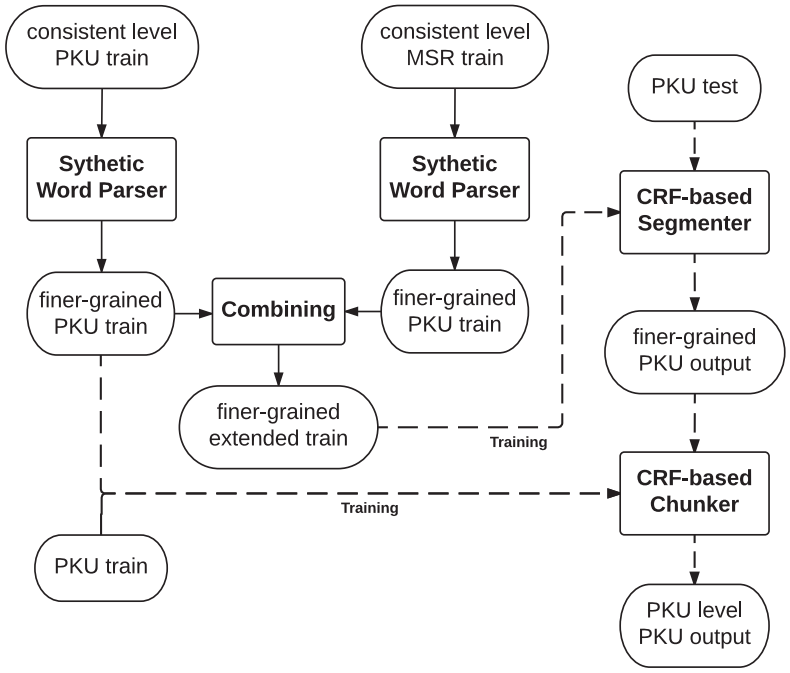


Fig. 7 Workflow of the two-stage word segmenter.

Fine-grained Level | 歌剧 / 院 / 合唱 / 团 / 共 / 200 / 人 / 。
Chunking Tags | B / E / B / E / S / S / S / S
Original Level | 歌剧院 / 合唱团 / 共 / 200 / 人 / 。

Fig. 8 Chunking tags of an example sentence.

The training data of the chunker is constructed on the fine-grained and original standard training data. An example of the chunking tags of a sentence is shown in Figure 8. In this sentence, two words 歌剧 (opera) and 院 (house) are chunked to the synthetic word 歌剧院 (opera house). 合唱 (chorus) and 团 (group) are chunked to the word 合唱团 (chorus group). Note that the same features (Section 3.1) of the first-stage segmenter are used in the second-stage chunker and the basic units of the features for the chunker are words rather than characters.

5 EXPERIMENTS

5.1 Settings

CRFSuite[1] is a speed-oriented implementation of CRFs for labeling sequential data. In this work, CRFSuite with passive-aggressive algorithm is adopted as the basic segmenter in our pipeline processes. We incorporate the dictionary (NAIST Chinese Dictionary[2]) and access variable feature (Chinese Gigaword Second Edition) in the same manner described in (Sun and Xu 2011). Based on (Liu et al. 2014), we use word2vec[3] to train the vector representations of the unigram, bigram and trigram character sequences of Chinese Gigaword Second Edition. The cluster-based results with K=100 are treated as the features for the segmenter. The segmenter also provides the baseline results for comparison.

Cheng et al. (2015) built a 38,423 words dictionary with annotated internal structures. In this work, our synthetic word parser is trained on this data and includes the dictionary (NAIST Chinese Dictionary), access variable and Brown clustering features extracted from a large unlabeled corpus (Chinese Gigaword Second Edition[4]) as described in (Cheng et al. 2014).

In this paper, OOVs are defined as words not seen in the training set; thus, even if a word is in the NAIST dictionary, it could be OOV with respect to the training set. In PKU, there are 2,404 OOVs and among them 1,097 are seen in the dictionary. In MSR, there are 1,960 OOVs and 259 are seen in the dictionary.

[1] http://www.chokkan.org/software/crfsuite/
[2] http://cl.naist.jp/index.php?%B8%F8%B3%AB%A5%EA%A5%BD%A1%BC%A5%B9%2FNCD
[3] https://code.google.com/p/word2vec
[4] https://catalog.ldc.upenn.edu/LDC2005T14

5.2 Consistency of New Segmentation Level

An important hypothesis of this work is that the process described in Section 4.1 on two different corpora may reach a new consistent segmentation level. To prove this, we first combine the consistent level PKU and MSR training data into a single extended dataset. Then we randomly shuffle the order of the sentences and divide the data into 10 equal pieces. We use the basic CRF-based segmenter trained on 90% data do segmentation on the other 10% (10-fold cross-validation). The results are shown in Table 1.

Table 2 shows the character length distribution of the words in PKU, MSR, and the extended data. Since the synthetic word parser is performed on the words with three or more characters, we just ignore the 1-char and 2-char words (generally considered as the smallest units in Chinese) in this table. MSR is obviously coarse standard data compared to PKU with higher rates for nearly all character lengths. Particularly, words with seven or more characters account for approximately 11.6% of all word types in MSR, while the rates of the other two corpora are only 1.6% and 2%. The extended data have the most fine-grained level among the three corpora.

5.3 Main Results

Table 3 summarizes the main segmentation results obtained by the proposed methods. Here, we specify two different settings **Proposed-1** and **Proposed-2** for our model. In the first setting, the system simply uses the consistent level extended training data (Section 4.1) to train the first-stage segmenter (the synthetic word parser process is omitted). **Proposed-1** helps us estimate the benefit of larger extended training data (including heterogeneous data). In the second setting, the extended training data are converted to a finer-grained standard by synthetic word parsing.

Table 1 10-fold cross-validation results on new extended data.

	Precision	Recall	F-score	R_{oov}
10-fold cross-validation	98.25	98.03	98.14	70.21

Table 2 Character length distribution of words in PKU and MSR corpora.

Corpus	3-char		4-char		5-char		6-char		longer	
	Count	Rate	Count	Rate	Count	Rate	Count	Rate	Count	Rate
PKU	11,320	20.5%	6,812	12.3%	1,746	3.2%	611	1.1%	887	1.6%
MSR	17,081	19.4%	12,545	14.2%	6,879	7.8%	5,103	5.8%	10,195	11.6%
Extended	13,706	17.90%	9,053	11.8%	3,689	4.8%	1,465	1.9%	1,536	2%

'Count' denotes the number of the word types with specific character length; 'Rate' denotes the number of word types with specific character length against the total word types in the corpus.

Table 3 Comparison between state-of-the-art Chinese word segmentation results and our system on PKU and MSR corpora.

System	PKU				MSR			
	Precision	Recall	F-score	R_{oov}	Precision	Recall	F-score	R_{oov}
Baseline	96.34	95.69	96.01	81.87	97.13	97.35	97.24	72.5
Sub-word Information								
+ bpe (10 K)	96.21	95.37	95.79	81.58	97.08	97.36	97.22	72.61
+ bpe (20 K)	96.21	95.55	95.88	81.33	97.12	*97.45	97.28	72.43
+ bpe (30 K)	96.28	95.76	96.02	81.17	97.12	*97.47	97.29	72.32
+ bpe (40 K)	96.30	95.81	96.05	81.13	97.13	*97.51	97.32	72.47
+ bpe (50 K)	96.30	95.81	96.05	81.13	97.10	*97.48	97.29	72.12
+ leftmost	96.43	*95.87	*96.15	*83.43	*97.24	*97.45	*97.34	*73.89
+ synthetic words	*96.45	*95.92	*96.19	***83.57**	***97.28**	*97.46	***97.36**	***74.03**
Heterogeneous Data								
Jiang et al. (2009)	***96.57**	*95.96	*96.26	*82.38	97.16	97.33	97.25	*73.5
Proposed-1	96.26	*96.18	*96.22	*82.2	97.03	*97.46	97.23	*73.25
Heterogeneous + Sub-word								
Proposed-2	96.36	***96.27**	***96.31**	*83.07	97.21	***97.52**	***97.36**	*73.38

Proposed-1 can be directly compared to (Jiang et al. 2009) because they use the same baseline segmentater and heterogeneous data. Proposed-2 finally combines heterogenous data and synthetic word parsing. * denotes significance at $p < 0.05$, compared to the baseline.

The improvement of **Proposed-2** compared to **Proposed-1** indicates the benefit of including the internal structure information of words.

Since analyzing the internal structure information of words is a general component in our pipeline framework, the synthetic word parser can be flexibly alternated by other sub-word analysis algorithms. In the 'Sub-word Information' parts, we first investigate the benefits provided by three different sub-word segmentation analyzers. In Table 3, 'synthetic words' denotes the segmentation system (Cheng et al. 2015) based on the baseline segmenter with all the words in the training data transformed into sub-word segmentation by a synthetic word parser, 'bpe' denotes the system with sub-word segmentation predicted by byte-pair encoding[5] (Sennrich, Haddow, and Birch 2016) with different vocabulary settings, and 'leftmost' denotes the system with sub-word segmentation predicted heuristically by a leftmost dictionary match (using the NAIST Chinese Dictionary). 'synthetic words', and 'leftmost' demonstrates improvements on both PKU and MSR, and our synthetic word parser achieves the most overall gains, particularly in OOV Recall.

The results of 'bpe' are not stable. 'bpe' does not show drop on MSR when the vocabulary size

[5] https://github.com/rsennrich/subword-nmt

is 10 K; compared to the very low performance on PKU. As vocabulary size increases from 10 K to 50 K, the recall of 'bpe' increases. However, OOV recall is dropping continuously. When the vocabulary is 40 K, 'bpe' obtains the highest F-score, which is slightly better than the baseline. Although 'leftmost' obtains F-score and OOV recall that are close to 'synthetic words', we find some issues such as 1) 总 / 司令 (chief/commander) is incorrectly split into 总司 (Japanese given name) / 令 (command). 2) the inconsistency of 太仓 / 县 (Taicang/county) and 宣汉县 (Xuanhan county) caused by the absence of 宣汉 in the dictionary. Moreover, the synthetic words parser is designed to predict real tree structure rather than such flat sub-word segmentation.

Although Jiang et al. (2009) obtains the highest F-score improvement +0.8 on Chinese TreeBank 5.0 (CTB5) (guided by People's Daily), the results of re-implementation obtained on PKU and MSR are surprisingly lower. This can be attributed to two points: 1) the difference in the segmentation standards between PKU and MSR is large (as shown in Table 2), which brings a big barrier to benefit each other. 2) the large difference in the sizes of the two corpora (19 K and 87 K sentences). It is difficult for a large corpus to obtain improvements considering the additional loss in the pipeline processing.

Both Jiang et al. (2009) and Chao et al. (2015) intended to improve word segmentation performance on a small data with large heterogeneous data. As Chao et al. (2015) showed in the paper, their coupled CRF obtains an F-score improvement of +0.5 on Chinese TreeBank 7.0 (CTB7, 50 K) with the help of PD (280 K) while the guide-feature method obtains an improvement of +0.34. In our experiments, we also investigate another case, i.e., a large corpus (MSR, 87 K) guided by a small corpus (PKU, 19 K). Jiang's method obtains an improvement of +0.21 on PKU (guided by MSR) and +0.02 on MSR (guided by PKU), while **Proposed-1** shows slightly lower F-scores. The results of the Jiang's method and **Proposed-1** on MSR suggest that the large corpus can hardly obtain an obvious improvement guided by a small corpus, considering the additional loss from different segmentation standards and chunking. However, our consistent level data are friendly to incorporate synthetic word parsing to further boost performance.

Parsing synthetic words (**Proposed-2**) contributes an additional +0.1 F-score on both PKU and MSR, based on **Proposed-1** (with only heterogeneous data). **Proposed-2** finally obtains +0.3 on PKU and +0.12 on MSR, compared to the baseline, which outperforms (Jiang et al. 2009) and **Proposed-1**.

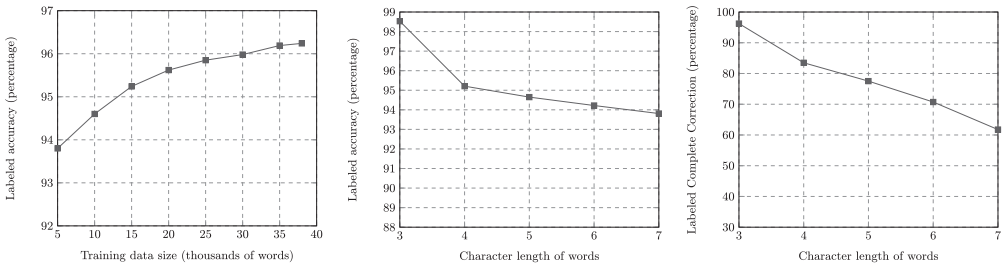
5.4 Analysis

Although we decompose words into the consistent level standard (similar to the PKU standard shown in Table 2), many synthetic words still exist in this data. Synthetic word parsing helps

analyze the internal information of these words, which provides major improvement compared to **Proposed-1** and (Jiang et al. 2009). We manually detect some difference between the results of **Proposed-1** and **Proposed-2**. OOV words such as 紫团 / 山 (Zituan/mountain) and 二 / 进制 (binary/numeral system) are correctly identified with the help of the internal information of words because other mountain names and 十 / 进制 (decimal/numeral system) exist in the training data. The internal information of synthetic words also helps identify IV words such as 数学 / 家 (mathematic/ian), 有钱 / 人 (rich/people), because they get more consistency with other words like 教育 / 家 (educat/or). We also observed that some polysemous characters result in ambiguous errors. For instance, 非 can be a prefix 'non-' in 非 / 军事 (non-/military) or an auxiliary verb 'must'.

5.5 Performance of Synthetic Word Parser

The synthetic word parser is an important component that determines the quality of the conversion from the consistent segmentation level to a finer-grained standard. To evaluate the performance of the parser, we conduct 10-fold cross-validation on all 38 K synthetic words, as shown in Figure 9. As the size of the training data increases, the parser obtains gains on character-level labeled accuracy (LA). It finally achieves 96.24% LA (Figure 9a) trained on the entire data. Figures 9b and 9c show the character-level LA and labeled complete correction (LCC) performance against character length of words. As the character length of words increases, the performance of the model consistently drops. For the words with 7-character or more, the LCC of the parser has dropped to 61.76%. However, considering the low rate of the words with character length equal to 6 or longer in the extended data (Table 2), our parser provides reasonable performance when generating finer-grained data.



(a) Labeled Accuracy on training data size

(b) Labeled Accuracy on character length of words

(c) Labeled Complete Correction on character length of words

Fig. 9 Character-level performance of the synthetic word parser (10-fold cross-validation). In (b) and (c), 'Character length of words' equal to 7 means equal to 7 or greater.

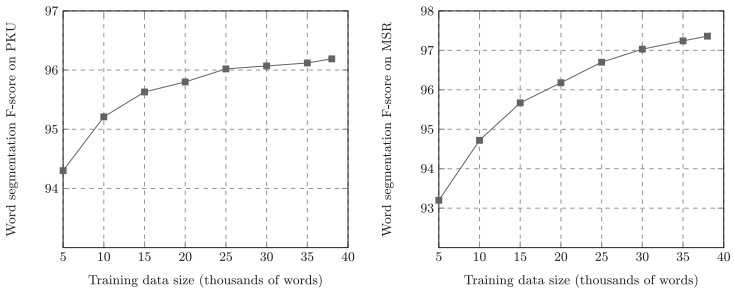


Fig. 10 Word segmentation F-score against the training data size of the synthetic word parser

Figure 10 shows the word segmentation F-score of 'Baseline + synthetic word parsing' when the synthetic word parser is trained with different amounts of training data. As the amount of training data increases, the system continuously improves on both corpora. As a relatively fine-grained corpus compared to MSR, PKU quickly achieves high performance, which suggests that the parser gets high parsing accuracy on short words (3-char, 4-char words) with a small amount of the training data. The system starts from a very low F-score on MSR with 5 K data. More training data are required to reach reasonable parsing accuracy when processing longer words in MSR. With all 38 K data, the system obtains the highest F-scores on both corpora.

6 CONCLUSIONS

In this paper, we have proposed a method to find a new consistent segmentation level across two different corpora. This consistent level makes it possible for multiple corpora to be extended easily. Using a synthetic word parser, we converted the consistent level extended data to a finer-grained level, in which our first-stage segmenter is expected to provide more accurate prediction of both IV and OOV words. Although the second-stage chunking brings an additional loss, the proposed system achieves state-of-the-art recall and F-score results on both PKU and MSR. We also perform additional investigations of the benefits of three different sub-word analysis algorithms: synthetic word parsing, byte-pair encoding and leftmost dictionary match. A further idea for improvement is that part-of-speech information may offer important clues for the second-stage chunking prediction.

Acknowledgement

We thank the anonymous reviewers for the insightful comments.

Reference

Carreras, X. (2007). "Experiments with a Higher-Order Projective Dependency Parser." In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pp. 957–961, Prague, Czech Republic. Association for Computational Linguistics.

Chao, J., Li, Z., Chen, W., and Zhang, M. (2015). "Exploiting Heterogeneous Annotations for Weibo Word Segmentation and POS Tagging." In *National CCF Conference on Natural Language Processing and Chinese Computing*, pp. 495–506. Springer.

Cheng, F., Duh, K., and Matsumoto, Y. (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. European Language Resources Association (ELRA).

Cheng, F., Duh, K., and Matsumoto, Y. (2015). "Synthetic Word Parsing Improves Chinese Word Segmentation." In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 262–267, Beijing, China. Association for Computational Linguistics.

Daume III, H. (2007). "Frustratingly Easy Domain Adaptation." In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 256–263, Prague, Czech Republic. Association for Computational Linguistics.

Daumé III, H. and Marcu, D. (2006). "Domain Adaptation for Statistical Classifiers." *Journal of Artificial Intelligence Research*, **26** (1), pp. 101–126.

Feng, H., Chen, K., Deng, X., and Zheng, W. (2004). "Accessor Variety Criteria for Chinese Word Extraction." *Computational Linguistics*, **30** (1), pp. 75–93.

Goh, C.-L., Asahara, M., and Matsumoto, Y. (2006). "Machine Learning-based Methods to Chinese Unknown Word Detection and POS Tag Guessing." *Journal of Chinese Language and Computing*, **16** (4), pp. 185–206.

Jiang, W., Huang, L., and Liu, Q. (2009). "Automatic Adaptation of Annotation Standards: Chinese Word Segmentation and POS Tagging: A Case Study." In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Confer-*

ence on Natural Language Processing of the AFNLP: Volume 1, pp. 522–530. Association for Computational Linguistics.

Li, Z. and Zhou, G. (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*, pp. 1445–1454. Association for Computational Linguistics.

Liu, X., Duh, K., Matsumoto, Y., and Iwakura, T. (2014). "Learning Character Representations for Chinese Word Segmentation." In *NIPS 2014 Workshop on Modern Machine Learning and Natural Language Processing*.

McDonald, R. (2006). *Discriminative Learning and Spanning Tree Algorithms for Dependency Parsing*. Ph.D. thesis, University of Pennsylvania.

Nivre, J. (2003). "An Efficient Algorithm for Projective Dependency Parsing." In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT)*, pp. 149–160.

Sennrich, R., Haddow, B., and Birch, A. (2016). "Neural Machine Translation of Rare Words with Subword Units." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Sun, W. (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*, pp. 1385–1394, Portland, Oregon, USA. Association for Computational Linguistics.

Sun, W. and Xu, J. (2011). "Enhancing Chinese Word Segmentation Using Unlabeled Data." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 970–979. Association for Computational Linguistics.

Tseng, H., Chang, P., Andrew, G., Jurafsky, D., and Manning, C. (2005). "A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005." In *Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing*, pp. 168–171.

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

Yamada, H. and Matsumoto, Y. (2003). "Statistical Dependency Analysis with Support Vector Machines." In *Proceedings of IWPT*, Vol. 3, pp. 195–206.

Zhang, M., Zhang, Y., Che, W., and Liu, T. (2013). "Chinese Parsing Exploiting Characters." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 125–134, Sofia, Bulgaria. Association for Computational

Linguistics.

Zhao, H. and Kit, C. (2008). "Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition." In *6th SIGHAN Workshop on Chinese Language Processing*, p. 106.

Fei Cheng: received his B.S. degree from Shanghai Donghua University in 2005 and M.S. degree from the Nara Institute of Science and Technology in 2013. He is currently a postdoctoral researcher at the National Institute of Informatics. His research interests include natural language processing and deep learning.

Kevin Duh: received the B.S. degree from Rice University in 2003, and Ph.D. from the University of Washington in 2009, both in electrical engineering. He was working as an assistant professor at the Nara Institute of Science and Technology (NAIST), Graduate School of Information Science, Ikoma, Japan. Before joining NAIST, from 2009 to 2012, he worked at the NTT Communication Science Laboratories. He is currently an assistant research professor at the Johns Hopkins University. His research interests include intersection of natural language processing and machine learning, in particular, in areas relating to machine translation and deep learning.

Yuji Matsumoto: received his M.S. and Ph.D. degrees in information science from Kyoto University in 1979 and in 1989. He joined the Machine Inference Section of Electrotechnical Laboratory in 1979. He has then experienced an academic visitor at the Imperial College of Science and Technology, a deputy chief of the First Laboratory at ICOT, and an associate professor at Kyoto University. He is currently a professor at the Graduate School of Information Science, Nara Institute of Science and Technology. His main research interests are natural language understanding and machine learning.

(Received May 11, 2017)
(Revised July 21, 2017)
(Accepted September 16, 2017)