Effective Document-Level Features for Chinese Patent Word Segmentation

Si Li
Chinese Language Processing Group
Brandeis University
Waltham, MA 02453, USA
lisi@brandeis.edu

Nianwen Xue
Chinese Language Processing Group
Brandeis University
Waltham, MA 02453, USA
xuen@brandeis.edu

Abstract

A patent is a property right for an invention granted by the government to the inventor. Patents often have a high concentration of scientific and technical terms that are rare in everyday language. However, some scientific and technical terms usually appear with high frequency only in one specific patent. In this paper, we propose a pragmatic approach to Chinese word segmentation on patents where we train a sequence labeling model based on a group of novel document-level features. Experiments show that the accuracy of our model reached 96.3% (F_1 score) on the development set and 95.0% on a held-out test set.

1 Introduction

It is well known that Chinese text does not come with natural word delimiters, and the first step for many Chinese language processing tasks is word segmentation, the automatic determination of word boundaries in Chinese text. Tremendous progress was made in this area in the last decade or so due to the availability of large-scale human segmented corpora coupled with better statistical modeling techniques. On the data side, there exist a few large-scale human annotated corpora based on established word segmentation standards, and these include the Chinese TreeBank (Xue et al., 2005), the Sinica Balanced Corpus (Chen et al., 1996), the PKU Peoples' Daily Corpus (Duan et al., 2003), and the LIVAC balanced corpus (T'sou et al., 1997). Another driver for the improvement in Chinese word segmentation accuracy comes from the evolution of statistical modeling techniques. Dictionaries used to play a central role in early heuristics-based word segmentation techniques (Chen and Liu, 1996; Sproat et al., 1996). Modern word segmentation systems have moved away from dictionary-based approaches in favor of character tagging approaches. This allows the word segmentation problem to be modeled as a sequence labeling problem, and lends itself to discriminative sequence modeling techniques (Xue, 2003; Peng et al., 2004). With these better modeling techniques, state-of-the-art systems routinely report accuracy in the high 90%, and a few recent systems report accuracies of over 98% in F_1 score (Sun, 2011; Zeng et al., 2013b).

Chinese word segmentation is not a solved problem however and significant challenges remain. Advanced word segmentation systems perform very well in domains such as newswire where everyday language is used and there is a large amount of human annotated training data. There is often a rapid degradation in performance when systems trained on one domain (let us call it the *source* domain) are used to segment data in a different domain (let us call it the *target* domain). This problem is especially severe when the target domain is distant from the source domain. This is the problem we are facing when we perform word segmentation on Chinese patent data. The word segmentation accuracy on Chinese patents is very poor if the word segmentation model is trained on the Chinese TreeBank data, which consists of data sources from a variety of genres but no patents. To address this issue, we annotated a corpus of 142 patents which contain about 440K words according to the Chinese TreeBank standards. We trained a character-tagging based CRF model for word segmentation, and based on the writing style of patents, we propose a group of document-level features as well as a novel character part-of-speech feature (C_POS). Our results show these new features are effective and we are able to achieve an accuracy of 96.3% (F_1 score) on the development set and 95% (F_1 score) on the test set.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 199–205,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

2 Method

We adopt the character-based sequence labeling approach, first proposed in (Xue, 2003), as our modeling technique for its simplicity and effectiveness. This approach treats each sentence as a sequence of characters and assigns to each character a label that indicates its position in the word. In this paper, we use the *BMES* tag set to indicate the character positions. The tag set has four labels that represent for possible positions a character can occupy within a word: *B* for beginning, *M* for middle, *E* for ending, and *S* for a single character as a word. After each character in a sentence is tagged with a BMES label, a sequence of words can be derived from this labeled character sequence.

We train a Conditional Random Field (CRF) (Lafferty et al., 2001) model for this sequence labeling. When extracting features to train a CRF model from a sequence of n characters $C_1C_2...C_{i-1}C_iC_{i+1}...C_n$, we extract features for each character C_i from a fixed window. We start with a set of core features extracted from the annotated corpus that have been shown to be effective in previous works and propose some new features for patent word segmentation. We describe each group of features in detail below.

2.1 Character features (CF)

When predicting the position of a character within a word, features based on its surrounding characters and their types have shown to be the most effective features for this task (Xue, 2003). There are some variations of these features depending on the window size in terms of the number of characters to examine, and here we adopt the feature templates used in (Ng and Low, 2004).

Character N-gram features The N-gram features are various combinations of the surrounding characters of the candidate character C_i. The 10 features we used are listed below:

- Character unigrams: C_k $(i-3 < k < i+3)$
- Character bigrams: C_kC_{k+1} $(i-3 < k < i+2)$ and $C_{k-1}C_{k+1}$ $(k=i)$

Character type N-gram features We classify the characters in Chinese text into 4 types: Chinese characters or *hanzi*, English letters, numbers and others. T_i is the character type of C_i. The character type has been used in the previous works in various forms (Ng and Low, 2004; Jiang et al., 2009), and the 4 features we use are as follows:

- Character type unigrams: T_k $(k=i)$
- Character type bigrams: T_kT_{k+1} $(i-2 < k < i+1)$ and $T_{k-1}T_{k+1}$ $(k=i)$

Starting with this baseline, we extract some new features to improve Chinese patent word segmentation accuracy.

2.2 POS of single-character words (C_POS)

Chinese words are composed of Chinese hanzi, and an overwhelming majority of these Chinese characters can be single-character words themselves in some context. In fact, most of the multi-character words are compounds that are 2-4 characters in length. The formation of these compound words is not random and abide by word formation rules that are similar to the formation of phrases (Xue, 2000; Packard, 2000). In fact, the Chinese TreeBank word segmentation guidelines (Xia, 2000) specify how words are segmented based on the part-of-speech (POS) of their component characters. We hypothesize that the POS tags of the single-character words would be useful information to help predict how they form the compound words, and these POS tags are more fine-grained information than the character type information described in the previous section, but are more robust and more generalizable than the characters themselves.

Since we do not have POS-tagged patent data, we extract this information from the Chinese TreeBank (CTB) 7.0, a 1.2-million-word out-of-domain dataset. We extract the POS tags for all the single-character words in the CTB. Some of the single-character words will have more than one POS tag. In this case, we select the POS tag with the highest frequency as the C_POS tag for this character. The result of this extraction process is a list of single-character Chinese words, each of which is assigned a single POS tag.

When extracting features for the target character C_i, if C_i is in this list, the POS tag of C_i is used as a feature for this target character.

2.3 Document-level features

A patent is a property right for an invention granted by the government to the inventor, and many of the patents have a high concentration of scientific and technical terms. From a machine learning perspective, these terms are hard to detect and segment because they are often "new words" that are not seen in everyday language. These technical

Algorithm 1 Longest n-gram sequence extraction.

Input:

Sentences $\{s_i\}$ in patent P_i;

Output:

Longest n-gram sequence list for P_i;

1: **For** each sentence s_i in P_i **do**:
   n-gram sequence extraction
   ($2 \leq n \leq length(s_i)$);
2: Count the frequency of each n-gram sequence;
3: Delete the sequence if its frequency<2;
4: Delete sequence i if it is contained in a longer sequence j;
5: All the remaining sequences form a longest n-gram sequence list for P_i;
6: **return** Longest n-gram sequences list.

terminologies also tend to be very sparse, either because they are related to the latest invention that has not made into everyday language, or because our limited patent dataset cannot possibly cover all possible technical topics. However, these technical terms are also topical and they tend to have high relative frequency within a patent document even though they are sparse in the entire patent data set. We attempt to exploit this distribution property with some document-level features which are extracted based on each patent document.

Longest n-gram features (LNG) We propose a longest n-gram (LNG) feature as a document-level feature. Each patent document is treated as an independent unit and the candidate longest n-gram sequence lists for each patent are obtained as described in Algorithm 1.

For a given patent, the LNG feature value for the target character C_i's LNG is set to 'S' if the bigram (C_i, C_{i+1}) are the first two characters of an n-gram sequence in this patent's longest n-gram sequence list. If (C_{i-1}, C_i) are the last two characters of an n-gram sequence in this patent's longest n-gram sequence list, the target character C_i's LNG is set to 'F'. It is set to 'O' otherwise. If C_i can be labeled as both 'S' and 'F' at the same time, label 'T' will be given as the final label. For example, if 'α' is the target character C_i in patent A and the sequence 'α一千扰素' is in patent A's longest n-gram sequence list. If the character next to 'α' is '一', the value of the LNG feature is set to 'S'. If the next character is not '一', the value of the LNG feature is set to 'O'.

Algorithm 2 Pseudo KL divergence.

Input:

Sentences $\{s_i\}$ in patent P_i;

Output:

Pseudo KL divergence values between different characters in P_i;

1: **For** each sentence s_i in P_i **do**:
   trigram sequences extraction;
2: Count the frequency of each trigram;
3: Delete the trigram if its frequency<2;
4: **For** C_i in trigram $C_iC_{i+1}C_{i+2}$ **do** :

$$PKL(C_i, C_{i+1}) = p({C_i}^1) log \frac{p({C_i}^1)}{p({C_{i+1}}^2)} \quad (1)$$

$$PKL(C_i, C_{i+2}) = p({C_i}^1) log \frac{p({C_i}^1)}{p({C_{i+2}}^3)} \quad (2)$$

   The superscripts {1,2,3} indicate the character position in trigram sequences;
5: **return** $PKL(C_i, C_{i+1})$ and $PKL(C_i, C_{i+2})$ for the first character C_i in each trigram.

Pseudo Kullback-Leibler divergence (PKL) The second document-level feature we propose is the Pseudo Kullback-Leibler divergence feature which is calculated following the form of the Kullback-Leibler divergence. The relative position information is very important for Chinese word segmentation as a sequence labeling task. Characters XY may constitute a meaningful word, but characters YX may not be. Therefore, if we want to determine whether character X and character Y can form a word, the relative position of these two characters should be considered. We adopt a pseudo KL divergence with the relative position information as a measure of the association strength between two adjacent characters X and Y. The pseudo KL divergence is an asymmetric measure. The PKL value between character X and character Y is described in Algorithm 2.

The PKL values are real numbers and are sparse. A common solution to sparsity reduction is binning. We rank the PKL values between two adjacent characters in each patent from low to high, and then divide all values into five bins. Each bin is assigned a unique ID and all PKL values in the same bin are replaced by this ID. This ID is then used as the PKL feature value for the target character C_i.

Pointwise Mutual information (PMI) Pointwise Mutual information has been widely used in previous work on Chinese word segmentation (Sun and Xu, 2011; Zhang et al., 2013b) and it is a measure of the mutual dependence of two strings and reflects the tendency of two strings appearing in one word. In previous work, PMI statistics are gathered on the entire data set, and here we gather PMI statistics for each patent in an attempt to capture character strings with high PMI in a particular patent. The procedure for calculating PMI is the same as that for computing pseudo KL divergence, but the functions (1) and (2) are replaced with the following functions:

$$PMI(C_i, C_{i+1}) = log\frac{p(C_i{}^1, C_{i+1}{}^2)}{p(C_i{}^1)p(C_{i+1}{}^2)} \quad (3)$$

$$PMI(C_i, C_{i+2}) = log\frac{p(C_i{}^1, C_{i+2}{}^3)}{p(C_i{}^1)p(C_{i+2}{}^3)} \quad (4)$$

For the target character C_i, we obtain the values for $PMI(C_i, C_{i+1})$ and $PMI(C_i, C_{i+2})$. In each patent document, we rank these values from high to low and divided them into five bins. Then the PMI feature values are represented by the bin IDs.

3 Experiments

3.1 Data preparation

We annotated 142 Chinese patents following the CTB word segmentation guidelines (Xia, 2000). Since the original guidelines are mainly designed to cover non-technical everyday language, many scientific and technical terms found in patents are not covered in the guidelines. We had to extend the CTB word segmentation guidelines to handle these new words. Deciding on how to segment these scientific and technical terms is a big challenge since these patents cover many different technical fields and without proper technical background, even a native speaker has difficulty in segmenting them properly. For difficult scientific and technical terms, we consult BaiduBaike ("Baidu Encyclopedia")[1], which we use as a scientific and technical terminology dictionary during our annotation. There are still many words that do not appear in BaiduBaiKe, and these include chemical names and formulas. These chemical names and formulas (e.g., "１－溴－３－氯丙烷/1-bromo-3-chloropropane") are usually very long, and unlike everyday words, they often have numbers and punctuation marks in them. We decided not to try segmenting the internal structures of such chemical terms and treat them as single words, because without a technical background in chemistry, it is very hard to segment their internal structures consistently.

[1] http://baike.baidu.com/

Table 1: Training, development and test data on Patent data

Data set	# of words	# of patent
Training	345336	113
Devel.	46196	14
Test	48351	15

The annotated patent dataset covers many topics and they include chemistry, mechanics, medicine, etc. If we consider the words in our *annotated dataset* but not in CTB 7.0 data as *new words* (or out-of-vocabulary, OOV), the new words account for 18.3% of the patent corpus by token and 68.1% by type. This shows that there is a large number of words in the patent corpus that are not in the everyday language vocabulary. Table 1 presents the data split used in our experiments.

3.2 Main results

We use CRF++ (Kudo, 2013) to train our sequence labeling model. $Precision$, $recall$, F_1 score and R_{OOV} are used to evaluate our word segmentation methods, where R_{OOV} for our purposes means the recall of new words which do not appear in CTB 7.0 but in patent data.

Table 2 shows the segmentation results on the development and test sets with different feature templates and different training sets. The CTB training set includes the entire CTB 7.0, which has 1.2 million words. The model with the CF feature template is considered to be the baseline system. We conducted 4 groups of experiments based on the different datasets: (1) patent training set + patent development set; (2) patent training set + patent test set; (3) CTB training set + patent development set; (4) CTB training set + patent test set.

The results in Table 2 show that the models trained on the patent data outperform the models trained on the CTB data by a big margin on both the development and test set, even if the CTB training set is much bigger. That proves the importance of having a training set in the same do-

Table 2: Segmentation performance with different feature sets on different datasets.

Train set	Test set	Features	P	R	F_1	R_{OOV}
Patent train	Patent dev.	CF	95.34	95.28	95.32	90.02
		CF+C_POS	95.58	95.40	95.49	90.40
		CF+C_POS+LNG	96.32	96.00	96.15	91.22
		CF+C_POS+PKL	95.62	95.41	95.51	90.40
		CF+C_POS+PMI	95.65	95.40	95.53	89.94
		CF+C_POS+PMI+PKL	95.72	95.53	95.62	90.37
		CF+C_POS+LNG+PMI	96.42	96.09	96.26	91.66
		CF+C_POS+LNG+PMI+PKL	96.48	96.12	96.30	91.69
Patent train	Patent test	CF	93.98	94.49	94.23	85.19
		CF+C_POS+LNG+PKL+PMI	94.89	95.10	95.00	87.89
CTB train	Patent dev.	CF+C_POS+LNG+PKL+PMI	89.04	90.75	89.89	72.80
CTB train	Patent test	CF+C_POS+LNG+PKL+PMI	87.88	89.03	88.45	70.89

main. The results also show that adding the new features we proposed leads to consistent improvement across all experimental conditions, and that the LNG features are the most effective and bring about the largest improvement in accuracy.

4 Related work

Most of the previous work on Chinese word segmentation focused on newswire, and one widely adopted technique is character-based representation combined with sequential learning models (Xue, 2003; Low et al., 2005; Zhao et al., 2006; Sun and Xu, 2011; Zeng et al., 2013b; Zhang et al., 2013b; Wang and Kan, 2013). More recently, word-based models using perceptron learning techniques (Zhang and Clark, 2007) also produce very competitive results. There are also some recent successful attempts to combine character-based and word-based techniques (Sun, 2010; Zeng et al., 2013a).

As Chinese word segmentation has reached a very high accuracy in the newswire domain, the attention of the field has started to shift to other domains where there are few annotated resources and the problem is more challenging, such as work on the word segmentation of literature data (Liu and Zhang, 2012) and informal language genres (Wang and Kan, 2013; Zhang et al., 2013a). Patents are distinctly different from the above genres as they contain scientific and technical terms that require some special training to understand. There has been very little work in this area, and the only work that is devoted to Chinese word segmentation is (Guo et al., 2012), which reports work on Chinese patent word segmentation with a fairly small test set without any annotated training data in the target domain. They reported an accuracy of 86.42% (F_1 score), but the results are incomparable with ours as their evaluation data is not available to us. We differ from their work in that we manually segmented a significant amount of data, and trained a model with document-level features designed to capture the characteristics of patent data.

5 Conclusion

In this paper, we presented an accurate character-based word segmentation model for Chinese patents. Our contributions are two-fold. Our first contribution is that we have annotated a significant amount of Chinese patent data and we plan to release this data once the copyright issues have been cleared. Our second contribution is that we designed document-level features to capture the distributional characteristics of the scientific and technical terms in patents. Experimental results showed that the document-level features we proposed are effective for patent word segmentation.

Acknowledgments

This paper is supported by the Intelligence Advanced Research Projects Activity (IARPA) via a contract NO. D11PC20154. All views expressed in this paper are those of the authors and do not necessarily represent the view of IARPA, DoI/NBC, or the U.S. Government.

References

Keh-Jiann Chen and Shing-Huan Liu. 1996. Word Identification for Mandarin Chinese Sentences. In *Proceedings of COLING'92*, pages 101–107.

Keh-Jiann Chen, Chu-Ren Huang, Li-Ping Chang, and Hui-Li Hsu. 1996. Sinica Corpus: Design Methodology for Balanced Corpora. In *Proceedings of the 11 th Pacific Asia Conference on Language, Information and Computation*, pages 167–176.

Huiming Duan, Xiaojing Bai, Baobao Chang, and Shiwen Yu. 2003. Chinese word segmentation at Peking University. In *Proceedings of the second SIGHAN workshop on Chinese language processing*, pages 152–155.

Zhen Guo, Yujie Zhang, Chen Su, and Jinan Xu. 2012. Exploration of N-gram Features for the Domain Adaptation of Chinese Word Segmentation. In *Proceedings of Natural Language Processing and Chinese Computing Natural Language Processing and Chinese Computing*, pages 121–131.

Wenbin Jiang, Liang Huang, and Qun Liu. 2009. Automatic Adaptation of Annotation Standards: Chinese Word Segmentation and POS Tagging - A Case Study. In *Proceedings of ACL'09*, pages 522–530.

Taku Kudo. 2013. CRF++: Yet Another CRF toolkit.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML'01*, pages 282–289.

Yang Liu and Yue Zhang. 2012. Unsupervised Domain Adaptation for Joint Segmentation and POS-Tagging. In *Proceedings of COLING'12*, pages 745–754.

Jin Kiat Low, Hwee Tou Ng, and Wenyuan Guo. 2005. A Maximum Entropy Approach to Chinese Word Segmentation. In *Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing*, pages 970–979.

Hwee Tou Ng and Jin Kiat Low. 2004. Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based? In *Proceedings of EMNLP'04*, pages 277–284.

Jerome Packard. 2000. *The Morphology of Chinese: a cognitive and linguistic approach*. Cambridge University Press.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese Segmentation and New Word Detection using Conditional Random Fields. In *Proceedings of COLING'04*.

Richard Sproat, Chilin Shih, William Gale, and Nancy Chang. 1996. A Stochastic Finite-State Word-Segmentation Algorithm for Chinese. *Computational Linguistics*, 22(3):377–404.

Weiwei Sun and Jia Xu. 2011. Enhancing Chinese Word Segmentation Using Unlabeled Data. In *Proceedings of EMNLP'11*, pages 970–979.

Weiwei Sun. 2010. Word-based and character-based word segmentation models: Comparison and combination. In *Proceedings of ACL'10*, pages 1211–1219.

Weiwei Sun. 2011. A Stacked Sub-Word Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging. In *Proceedings of ACL'11*, pages 1385–1394.

Benjamin K. T'sou, Hing-Lung Lin, Godfrey Liu, Terence Chan, Jerome Hu, Ching hai Chew, and John K.P. Tse. 1997. A Synchronous Chinese Language Corpus from Different Speech Communities: Construction and Application. *International Journal of Computational Linguistics and Chinese Language Processing*, 2(1):91–104.

Aobo Wang and Min-Yen Kan. 2013. Mining Informal Language from Chinese Microtext: Joint Word Recognition and Segmentation. In *Proceedings of ACL'13*, pages 731–741.

Fei Xia. 2000. The segmentation guidelines for the Penn Chinese Treebank (3.0).

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005. The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, 11(2):207–238.

Nianwen Xue. 2000. *Defining and identifying words in Chinese.* Ph.D. thesis, University of Delaware.

Nianwen Xue. 2003. Chinese Word Segmentation as Character Tagging. *International Journal of Computational Linguistics and Chinese Language Processing*, 8(1):29–48.

Xiaodong Zeng, Derek F. Wong, Lidia S. Chao, and Isabel Trancoso. 2013a. Co-regularizing character-based and word-based models for semi-supervised Chinese word segmentation. In *Proceedings of ACL'13*, pages 171–176.

Xiaodong Zeng, Derek F. Wong, Lidia S. Chao, and Isabel Trancoso. 2013b. Graph-based Semi-Supervised Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging. In *Proceedings of ACL'13*, pages 770–779.

Yue Zhang and Stephen Clark. 2007. Chinese Segmentation Using a Word-based Perceptron Algorithm. In *Proceedings of ACL'07*, pages 840–847.

Longkai Zhang, Li Li, Zhengyan He, Houfeng Wang, and Ni Sun. 2013a. Improving Chinese Word Segmentation on Micro-blog Using Rich Punctuations. In *Proceedings of ACL'13*, pages 177–182.

Longkai Zhang, Houfeng Wang, Xu Sun, and Mairgup Mansur. 2013b. Exploring Representations from Unlabeled Data with Co-training for Chinese Word Segmentation. In *Proceedings of EMNLP'13*, pages 311–321.

Hai Zhao, Chang-Ning Huang, and Mu Li. 2006. An improved Chinese word segmentation system with conditional random field. In *Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing*, pages 162–165.