BosonNLP: An Ensemble Approach for Word Segmentation and POS Tagging

Kerui Min(✉), Chenggang Ma, Tianmei Zhao, and Haiyan Li

BosonData, Inc., Shanghai, China
{kerui.min,chenggang.ma,tianmei.zhao,haiyan.li}@bosondata.com.cn

Abstract. Chinese word segmentation and POS tagging are arguably the most fundamental tasks in Chinese natural language processing. In this paper, we show an ensemble approach for segmentation and POS tagging, combining both discriminative and generative methods to get the advantage of both worlds. Our approach achieved the F1-score of 96.65% and 91.55% for segmentation and tagging respectively in the contest of NLPCC 2015 Shared Task 1, obtained the 1st place for both tasks.

1 Introduction

Chinese word segmentation and POS tagging are arguably the most fundamental tasks in Chinese natural language processing [4]. The problem has been studied for a long time in formal language domains such as newspapers and radio reports. Recently, informal text domain such as micro-blog started to attract researchers' attention, due to its practical and industrial value. As a result, the Shared Task 1 in NLPCC 2015 tries to provide a platform for evaluation [1].

In addition to other widely used dataset such as *the Chinese TreeBank (CTB)*, the difficulty of the task comes from the following:

1. Relatively small corpus for training, which only contains 10,000 sentences and 215,027 words.
2. High rate of OOV (out-of-vocabulary) words: it is estimated that the OOV rate for the test data is 7.25%.
3. Sentences contain informal words such as "萌萌哒", *e.g.*, 【/PU 美声/NN 版/NN 《/PU 小/JJ 苹果/NN 》/PU 传递/VV 萌萌哒/JJ 关爱/NN 】/PU..
4. The POS tag-set defines two types of adjectives: JJ and VA, distinguishing a normal adjective from predicative adjective, which requires longer dependency syntactic information for inference.

In this paper, we show an ensemble approach for segmentation and POS tagging, combining both discriminative and generative methods to get the advantage of both worlds. Evaluated on the Sina Weibo data, our approach achieved the F1-score of 96.65% and 91.55% for segmentation and tagging respectively, in the contest of NLPCC 2015 Shared Task 1, obtained the 1st place for both tasks.

J. Li et al. (Eds.): NLPCC 2015, LNAI 9362, pp. 520–526, 2015.
DOI: 10.1007/978-3-319-25207-0_48

Type	Template
Unigram	$C[-2], C[-1], C[0], C[1], C[2]$
Bigram	$C[-2..-1], C[-1..0], C[0..1], C[1..2]$
Trigram	$C[-2..0]$
Skipgram	$C[-1]C[1]$
CharType	$T[-2..2]$
Punctuation	True if $C[0]$ is a punctuation
LastChar	True if $C[0]$ is the last char of $\mathbf{x}$
Repeat-1	True if $C[0] = C[-1]$
Repeat-2	True if $C[0] = C[-2]$

Fig. 1. Feature template for character-level features.

2 Backbone Algorithm

As usual, we treat both word segmentation and POS tagging as sequence labeling tasks. We use conditional random fields (CRF) as the backbone algorithm for ensemble learning, combining results from different models.

In this section, we describe our approach for the backbone algorithm in three steps: pre-processing, statistical modeling, and post-processing. In the next section, we show how to modify this algorithm to build an ensemble model, where the latter one will be used for open track evaluation.

2.1 Pre-processing

The training data was given in the format that each character was associated with one of the four states $\{\mathbb{B}, \mathbb{M}, \mathbb{E}, \mathbb{S}\}$ for word segmentation, representing the beginning, inside, ending, isolation of a word respectively. In literature, this is usually referred as the 4-tag labeling [3]. In addition to the 4-tag labeling, we also generated the following formats:

1. 3-tag: $\{\mathbb{M}, \mathbb{E}, \mathbb{S}\}$, where the state "$\mathbb{M}$" represents both the beginning and inside states.
2. 5-tag: $\{\mathbb{B}, \mathbb{C}, \mathbb{M}, \mathbb{E}, \mathbb{S}\}$: where "$\mathbb{C}$" indicates the second character in a word.
3. 6-tag: $\{\mathbb{B}, \mathbb{C}, \mathbb{D}, \mathbb{M}, \mathbb{E}, \mathbb{S}\}$: where "$\mathbb{D}$" indicates the thrid character in a word.

Larger tag-sets allow more detailed feature representation, at the cost of potentially higher variance. This is the well-known *Bias-Variance Tradeoff* [5]. In our experiments, we found that 5-tag consistently provides the best results for the given corpus, often 0.1% to 0.2% higher than other representations. Therefore, we use the 5-tag representation for both segmentation and POS tagging. As the shared task requires to submit evaluation results using the 4-tag representation, we simply convert the 5-tag to 4-tag before submission.

2.2 Statistical Modeling

As we have mentioned earlier, we use the second-order linear-chain conditional random fields (CRF) as the backbone algorithm, as exact inference can be done in

Type	Template
PREFIX	P-k=TRUE if $C[0..k-1]$ is k-prefix $(2 \leq k \leq 10)$
SUFFIX	S-k=TRUE if $C[-k+1..0]$ is k-suffix $(2 \leq k \leq 10)$
SINGLETON	TRUE if $C[0] \in \mathfrak{D}$
PREFIXTAG	P-k-t=TRUE if $C[0..k-1]$ is k-prefix and $t \in \mathfrak{D}[C[0..k-1]]$ $(2 \leq k \leq 10)$
SUFFIXTAG	S-k-t=TRUE if $C[1-k..0]$ is k-suffix and $t \in \mathfrak{D}[C[1-k..0]]$ $(2 \leq k \leq 10)$
MORPHOLOGY1	M-t-$C[k]$=TRUE $t \in \mathfrak{D}[C[0..k-1]]$ and $C[0..k] \notin \mathfrak{D}$
MORPHOLOGY2	M-$C[0]$-t=TRUE $t \in \mathfrak{D}[C[1..k]]$ and $C[0..k] \notin \mathfrak{D}$

Fig. 2. Feature template for dictionary features.

polynomial time. Specifically, the conditional probability of the hidden sequence $\mathbf{y}$, given the observation sequence $\mathbf{x}$, is defined as

$$p(\mathbf{y}|\mathbf{x}) = \frac{\exp\left(\sum_i \sum_k \lambda_k f_k\left(\mathbf{y}_{i-1}, \mathbf{y}_i, \mathbf{x}\right)\right)}{Z_{\mathbf{x}}}, \tag{1}$$

where $Z_{\mathbf{x}}$ is the normalization factor.

Notice that from the above equation, the observation $\mathbf{x}$ influences $\mathbf{y}$ through feature functions $\{f_k(\cdot)\}_k$. The problem boils down to good feature extraction $\{f_k(\mathbf{y}_{i-1}, \mathbf{y}_i, \mathbf{x})\}_{k=1}^m$, which has been well-studied. Similar to previous work, we use $C[i]$ to indicate the character of the i-th position, centered at the current position, and $T[i]$ to indicate the type of $C[i]$. Below, five types of character are defined.

1. $T[i] = P$: if $C[i]$ is either an English or Chinese punctuation.
2. $T[i] = N$: if $C[i]$ is either an Arabic digit or Chinese number, *e.g.*, "0123 $\cdots$ 零一二**…**".
3. $T[i] = A$: if $C[i]$ is in alphabet set, $[a-z][A-Z]$, and full-width characters like "A B C D $\cdots$" are included.
4. $T[i] = D$: if $C[i]$ is in the following set "年月日时分秒".
5. $T[i] = O$: for any other characters.

Given the above definition, we are able to define effective feature template as in Fig 1. To deal with OOV words with repeat-character patterns, we define the REPEAT-1, REPEAT-2 to distinguish word of "AABB" type from "ABAB" type, where the former one is more likely an adjective and the later one a verb [2].

Notice that the above features only require character-level information, which is good for its simplicity. However, in order to get relatively accurate parameter estimation for weights $\{\lambda_k\}_{k=1}^m$, a minimal occurrence number of ~ 10 is often required for *maximum-likelihood estimation*. In other words, if a word, say $C[-1..0]$, occurred only a few times in training, it is unlikely to obtain accurate and unbiased estimation for it. Therefore, if we know $C[-1..0]$ can be a word, we would like to incorporate the information that $C[-1..0]$ occurred as a word in training directly, in addition to the BIGRAM features.

Let $\mathfrak{D}$ be the dictionary extracted from training corpus. We say a substring $C[i..i+k-1]$ is a k-prefix if $C[i..i+k-1] \in \mathfrak{D}$; a substring $C[i-k+1..i]$ is

a k-suffix if $C[i-k+1..i] \in \mathfrak{D}$. Furthermore, we define $\mathfrak{D}[s]$ to be the set of all possible POS tags of the given word s in our dictionary. If $s \notin \mathfrak{D}$, we have $\mathfrak{D}[s] = \emptyset$.

Here we introduce another family of feature to alleviate the difficulty of OOV words. Although Chinese is not a *morphologically rich language* (MRL) [6], we do observe several common patterns that might be useful for segmentation and tagging. For example, there are a number of words has the structure of "NN+ 们", such as "童鞋们/NN 动物们/NN 亲们/NN 鬣狗们/NN". During evaluation, if we ever encounter an OOV word "设计师们", we can infer that the it is likely to be a NN word if NN NN $\in \mathfrak{D}$[设计师]. Inspired by it, we designed the morphological features, as listed in Fig 2.

In the experiment, we will see that the combination of both character-level features and dictionary features provides effective information for our model to produce accurate prediction.

2.3 Post-processing

From the formulation of Eq (1), the character-based CRF model is a probabilistic model. We impose the following rules as post-processing to further improve the accuracy of the prediction.

1. the prediction of the last character of $\mathbf{y}$ must end with $\{\mathbb{S}, \mathbb{E}\}$.
2. never mark the end of a number if the next character is also a number.
3. never mark the end of a English word if the next character is in alphabet.

The above post-processing rules make the final prediction more coherent. We also tried to use *regular expression* to recognize URLs, which gave limited improvement, since URLs are very rare in the corpus.

3 Ensemble Model for Open Track

The backbone algorithm described in Section 2 is able to produce relatively accurate prediction, as will be found in the experiment section. However, due to the small training size, the variance of the model parameters is high. In the open track of the shared task, it is allowed to use other information and tools to improve the prediction. In addition to the backbone algorithm, we use the output from the following models:

- *Model A*: a HMM-based model trained on People's daily corpus[1].
- *Model B*: a CRF-based model trained on *the Chinese TreeBank (CTB) 7.0* [7]. The features used in this model are the same as in Section 2.
- *Model C*: a third-order discriminative sequence model trained on People's daily corpus, plus 100,000 sentences of more recent news collected and annotated by BosonData, Inc. This model will be released at www.bosonnlp.com for industrial developers and researchers to use for free.

[1] Newspaper text from People's Daily 1998.

Dataset	Sents	Words	Chars	Word Types	Char Types	OOV Rate
Training	10,000	215,027	347,984	28,208	39,71	-
Test	5,000	106,327	171,652	18,696	3,538	7.25%
Total	15,000	322,410	520,555	35,277	4,243	-

Fig. 3. Statistical information of dataset.

For a given input $\mathbf{x}$, denote the output of the above models as $\mathbf{y}_A, \mathbf{y}_B$ and $\mathbf{y}_C$. For each feature described in Section 2, we concatenate the original feature with the output from the above models during feature extraction, *i.e.* using $\{f_k(\mathbf{y}_{i-1}, \mathbf{y}_i, \mathbf{x}, \mathbf{y}_A, \mathbf{y}_B, \mathbf{y}_C)\}_k$. For example, for the original TRIGRAM feature, $C[-2..0]$, we enhance it by the following features:

$$C[-2..0] \circ \mathbf{y}_A[i], C[-2..0] \circ \mathbf{y}_B[i], C[-2..0] \circ \mathbf{y}_C[i],$$

where i is the current position for feature extraction. Finally, we train the same model with these extra information to get better prediction. We call it the ensemble model. In the next section, we will show that although different corpus has its own segmentation and POS tagging standard, the ensemble model could benefit from them.

4 Experiments

As we have mentioned earlier, the dataset used for this shared task is relatively informal, collected from *Sina Weibo*. The training and test data consist of microblogs from various topics, such as finance, sports, entertainment, and so on. Basic statistics of the dataset can be found in Fig 3.

Next, we show the performance of our models, BosonNLP, with other top competitors for this shared task. Notice that these evaluation scores were released from the official program committee of NLPCC 2015.

4.1 Closed Track Evaluation

During the closed track, competitors are not allowed to use external datasets or tools. We therefore apply the backbone algorithm described in Section 2 on the given training corpus. Notice that the dictionary $\mathfrak{D}$ can be obtained from the training coropus. The evaluation result can be found in Fig 4.

From the result, one can see that the proposed algorithm is very effective. Our result is on a par with the best solution (with 0.09% difference) for segmentation, and obtained the best result for POS tagging with 0.74% higher in terms of the F-1 score.

4.2 Open Track Evaluation

As shown in Fig 5, the evaluation result for the open track clearly demonstrate the advantage of our ensemble approach, compared with other teams.

Task	Team	Precision	Recall	F-1
POS	BosonNLP **(1st)**	88.91	88.95	88.93
POS	XUPT **(2nd)**	88.54	87.83	88.19
POS	WHU **(3rd)**	88.28	87.67	87.97
SEG	NJU **(1st)**	95.14	95.09	95.12
SEG	BosonNLP **(2nd)**	95.03	95.03	95.03
SEG	BUPT **(3rd)**	94.78	94.42	94.60

Fig. 4. Closed track result for NLPCC 2015 Shared Task 1.

Task	Team	Precision	Recall	F-1
POS	BosonNLP **(1st)**	91.42	91.68	91.55
POS	SZU **(2nd)**	88.93	89.05	88.99
POS	BJTU **(3rd)**	79.85	83.51	81.64
SEG	BosonNLP **(1st)**	96.56	96.75	96.65
SEG	NJU **(2nd)**	96.03	96.15	96.09
SEG	SZU **(3rd)**	95.52	95.64	95.58

Fig. 5. Open track result for NLPCC 2015 Shared Task 1.

We obtained the *1st* place for both segmentation and POS tagging, with significantly higher F-1 score (2.56% higher than the second place team). For the POS tagging task, we obtained the F-1 score of 91.55%, which is the only team with the score above 90%.

As an example, we compare the output from our proposed algorithm for the closed track, our ensemble algorithm, and *Model C* output, in order to get some insights.

- **Input:** 据气象部门预告，哈尔滨有瞬时风速6-7级左右的大风，并可能伴有短时强降水，雷电或冰雹等强对流天气。
- *Backbone algorithm* output: 据/P 气象/NN 部门/NN 预告/VV ，/PU 哈尔滨/LOC **有瞬/VV 时风速/NN** 6-7/CD 级/NN 左右/LC 的/DSP 大风/NN ，/PU 并/AD 可能/MV 伴有/VV 短时/JJ 强/JJ 降水/NN ，/PU **雷电/PER** 或/CC 冰雹/NN 等/ETC 强对流/NN 天气/NN 。/PU
- *Ensemble model* output: 据/P 气象/NN 部门/NN 预告/VV ，/PU 哈尔滨/LOC 有/VV **瞬时/NN 风速/NN** 6-7/CD 级/NN 左右/LC 的/DSP 大风/NN ，/PU 并/AD 可能/MV 伴有/VV 短时/JJ 强/JJ 降水/NN ，/PU **雷电/NN** 或/CC 冰雹/NN 等/ETC 强对流/NN 天气/NN 。/PU
- *Model C* output: 据/p 气象/n 部门/n 预告/v ，/wd 哈尔滨/ns 有/vyou **瞬时/t 风速/n** 6/m -/wp 7/m 级/q 左右/m 的/ude 大风/n ，/wd 并/c 可能/v 伴有/v 短时/b 强/a 降水/n ，/wd **雷电/n** 或/c 冰雹/n 等/udeng 强/a 对流/n 天气/n 。/wj

We highlighted the difference between models. Observe that the backbone algorithm made a segmentation mistake at "有瞬/VV 时风速/NN", which was a typical OOV word mistake. The result was corrected by the ensemble model to be "瞬时/NN 风速/NN". Notice that although the POS tagging standard, including

the word segmentation standard for *Model C* is different ("瞬时/t 风速/n"), the output is helpful to correct the backbone algorithm output. The ensemble model proposed in Section 3 provides a probabilistic model to *fuse* the information.

5 Conclusion

In this paper, we described our approach for the word segmentation and POS tagging approach for the NLPCC 2015 Shared Task 1. Although some of the features and techniques were designed for this particular task, they are also useful for general word segmentation and POS tagging tasks with varying standards. We plan to release our solution at www.bosonnlp.com, to allow other researchers and developers to exploit.

References

1. Qiu, X., Qian, P., Yin, L., Huang, X.: Overview of the NLPCC 2015 Shared Task: Chinese Word Segmentation and POS Tagging for Micro-blog Texts (2015). http://arxiv.org/abs/1505.07599
2. Huang, C.-T.J., Li, Y.-H.A., Li, Y.: The Syntax of Chinese (Cambridge Syntax Guides). Cambridge University Press (2009)
3. Zhao, H., Huang, C.-N., Li, M., Lu, B.-L.: Effective tag set selection in chinese word segmentation via conditional random field modeling. In: The 20th Pacific Asia Conference on Language, Information and Computation (PACLIC-2006) (2006)
4. Wong, K.-F., Li, W., Xu, R., Zhang, Z.-S.: Introduction to Chinese Natural Language Processing (Synthesis Lectures on Human Language Technologies). Morgan & Claypool Publishers (2009)
5. Bias-variance decomposition. In: Sammut, C., Webb, G.I. (eds.) Encyclopedia of Machine Learning. Springer (2011)
6. Sarikaya, R., Kirchhoff, K., Schultz, T., Hakkani-Tur, D.: Introduction to the special issue on processing morphologically rich languages. IEEE Transactions on Audio Speech, and Language Processing (2009)
7. Xue, N., Xia, F., Chiou, F.-D., Palmer, M.: The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. Journal of Natural Language Engineering (2005)