Chinese word segmentation model using bootstrapping

Baobao Chang and Mairgup Mansur
Institute of Computational Linguistics, Peking University
Key Laboratory of Computational Linguistics(Peking University),
Ministry Education, China
chbb@pku.edu.cn, mairgup@yahoo.com.cn

Abstract

We participate in the CIPS-SIGHAN-2010 bake-off task of Chinese word segmentation. Unlike the previous bakeoff series, the purpose of the bakeoff 2010 is to test the cross-domain performance of Chinese segmentation model. This paper summarizes our approach and our bakeoff results. We mainly propose to use χ^2 statistics to increase the OOV recall and use bootstrapping strategy to increase the overall F score. As the results shows, the approach proposed in the paper does help, both of the OOV recall and the overall F score are improved.

1 Introduction

After more than twenty years of intensive researches, considerable progress has been made in improving the performance of Chinese word segmentation. The bakeoff series hosted by the ACL SIGHAN shows that high F scores can be achieved in the closed test tracks, in which only specified training materials can be used in learning segmentation models.

Instead of using lexicon-driven approaches, state-of-art Chinese word segmenter now use character tagging model as Xue(2003) firstly proposed. In character tagging model, no predefined Chinese lexicons are required; a tagging model is learned using manually segmented training texts. The model is then used to assign each character a tag indicating the position of this character within word. Xue's approach has been become the most popular approach to Chinese word segmentation for its high performance and unified way to deal with OOV issues. Most of the segmentation works since then follow this approach. Major improvements in this line of research including: 1) More sophisticated learning models were introduced instead of the maximum entropy model that Xue used, like conditional random fields (CRFs) model which fit the sequence tagging tasks much better than maximum entropy model (Tseng et al.,2005). 2) More tags were introduced, as Zhao et al. (2006) shows 6 tags are superior to 4 tags in achieving high performance. 3) New feature templates were added, such as templates used in representing numbers, dates, letters etc. (Low et al., 2005)

Usually, the performance of segmentation model is evaluated on a test set from the same domain as the training set. Such evaluation does not reveal its ability to deal with domain variation. It is believed that, when test set is from other domains than the domain where training set is from, the learned model normally underperforms substantially.

The CIPS-SIGHAN-2010 bake-off task of Chinese word segmentation is set to focus on the cross-domain performance of Chinese word segmentation model.

We participate in the closed test track for simplified Chinese. Different with the previous bakeoffs, CIPS-SIGHAN-2010 bake-off provides both label corpus and unlabeled corpora. The labeled corpus is composed of texts from newspaper and has about 1.1 million words in total. The two unlabeled corpora cover two domains: literature and computer science, and each domain have about 100K characters in size. The test corpora cover four domains, two of which are literature and computer science, and the other two domains are unknown before releasing.

We build the Chinese word segmenter following the character tagging model. Instead of using CRF model, we use the hidden Markov support vector machines (Altun et al., 2003), which is also a sequence labeling model like CRF. We just show it can also be used to model Chinese segmentation tasks as an alternative other than CRF. To increase the ability of the model to recall OOV words, we propose to use χ^2 statistics and bootstrapping strategy to the overall performance of the model to out-of-domain texts.

2 The hidden Markov support vector machines

The hidden Markov support vector machine (SVM-HMM) is actually a special case of the structural support vector machines proposed by Tsochantaridis et al.(2005) which is a powerful model to structure predication problem. It differs from support vector machine in its ability to model complex structured problems and shares the max-margin training principles with support vector machines. The hidden Markov support vector machine model is inspired by the hidden Markov model and is an instance of structural support vector machine dedicated to solve sequence labeling learning, a problem that CRF model is assumed to solve. In the SVM-HMM model, the sequence labeling problems is modeled by learning a discriminant function F: $X \times Y \rightarrow R$ over the input sequence and the label sequence pairs, thus prediction of label sequence can be derived by maximizing F over all possible label sequences for a specific given input sequence $\mathbf{x}$.

$$f(\mathbf{x};\mathbf{w}) = \arg\max_{\mathbf{y}\in Y} F(\mathbf{x},\mathbf{y};\mathbf{w})$$

In the structural SVMs, F is assumed to be linear in some combined feature representation of the input sequence and the label sequence $\psi(\mathbf{x},\mathbf{y})$, i.e.

$$F(\mathbf{x},\mathbf{y};\mathbf{w}) = \langle \mathbf{w}, \boldsymbol{\psi}(\mathbf{x},\mathbf{y}) \rangle$$

where $\mathbf{w}$ denotes a parameter vector. For the SVM-HMMs, the discriminant function is defined as follows.

$$F(x,y;\mathbf{w}) = \sum_{t=1..T} \sum_{y\in\Sigma} \langle \overline{\mathbf{w}}_y, \boldsymbol{\Phi}(\mathbf{x}^t) \rangle \delta(y^t, y) + \eta \sum_{t=1..T-1} \sum_{y\in\Sigma} \sum_{y'\in\Sigma} \hat{\mathbf{w}}_{y,y'} \delta(y^t, y)\delta(y^{t+1}, y')$$

Here $\mathbf{w} = (\overline{\mathbf{w}}, \hat{\mathbf{w}})$, $\boldsymbol{\Phi}(\mathbf{x}^t)$ is the vector of features of the input sequence.

Like SVMs, parameter vector $\mathbf{w}$ is learned with maximum margin principle using training data. To control the complexity of the training problem, cutting plane method is proposed to solve the resulted constrained optimization problem. Thus only small subset of constraints from the full-sized optimization is checked to ensure a sufficiently accurate solution. Roughly speaking, SVM-HMM differs with CRF in its principle of training, both of them could be used to deal with sequence labeling problem like Chinese word segmentation.

3 The tag set and the basic feature templates

As most of other works on segmentation, we use a 4-tag tagset, that is S for character being a single-character-word by itself, B for character beginning a multi-character-word, E for character ending a multi-character-word and M for a character occurring in the middle of a multi-character-word.

We use the following feature template, like most of segmentation works widely used:

(a) C_n (n = -2, -1, 0, 1, 2)
(b) C_nC_{n+1} (n = -2, -1, 0, 1)
(c) $C_{-1}C_{+1}$

Here C refers to character; n refers to the position index relative to the current character. By setting the above feature templates, we actually set a 5-character window to extract features, the current character, 2 characters to its left and 2 characters to its right.

In addition, we also use the following feature templates to extract features representing character type. The closed test track of CIPS-SIGHAN-2010 bake-off allows participants to use four character types, which are Chinese Character, English Letter, digits and punctuations:

(d) T_n (n = -2, -1, 0, 1, 2)
(e) T_nT_{n+1} (n = -2, -1, 0, 1)
(f) $T_{-1}T_{+1}$

Here T refers to character type, its value can be digit, letter, punctuation or Chinese character.

4 The χ^2 statistic features

One of reasons of the performance degradation lies in the model's ability to cope with OOV words while working with the out-of-domain texts. Aiming at preventing the OOV recall from dropping sharply, we propose to use χ^2 statistics as features to the segmentation model.

χ^2 test is one of hypothesis test methods, which can be used to test if two events co-occur just by chance or not. A lower χ^2 score normally means the two co-occurred events are independent; otherwise they are dependent on each other. Hence, χ^2 statistics could also be used to deal with the OOV issue in segmentation models. The idea is very straightforward. If two adjacent characters in the test set have a higher χ^2 score, it is highly likely they form a word or are part of a word even they are not seen in the training set.

We only compute χ^2 score for character bigrams in the training texts and test texts. The χ^2 score of a bigram C_1C_2 can be computed by the following way.

$$\chi^2(C_1,C_2)=\frac{n\times(a\times d-b\times c)^2}{(a+b)\times(a+c)\times(b+d)\times(c+d)}$$

Here,

a refers to all counts of bigram C_1C_2 in the text;

b refers to all counts of bigrams that C_1 occurs but C_2 does not;

c refers to all counts of bigrams that C_1 does not occur but C_2 occurs;

d refers to all counts of bigrams that both C_1 and C_2 do not occur.

n refers to total counts of all bigrams in the text, apparently, $n=a+b+c+d$.

We do the χ^2 statistics computation to the training texts and the test texts respectively. To make the χ^2 statistics from the training texts and test texts comparable, we normalize the χ^2 score by the following formula.

$$\chi^2_{norm}(C_1,C_2)=\left\lfloor\frac{\chi^2(C_1,C_2)-\chi^2_{\min}}{\chi^2_{\max}-\chi^2_{\min}}\times 10\right\rfloor$$

Then we incorporate the normalized χ^2 statistics into the SVM-HMM model by adding two more feature templates as follows:

(g) X_nX_{n+1} (n = -2, -1, 0, 1)
(h) $X_{-1}X_{+1}$

The value of the feature X_nX_{n+1} is the normalized χ^2 score of the bigram C_nC_{n+1}. Note we also compute the normalized χ^2 score to bigram $C_{-1}C_{+1}$.

Because the normalized χ^2 score is one of 11 possible values 0, 1, 2, …, 10, templates (g)-(h) generate 55 features in total.

All features generated from the templates (a)-(f) together with the 55 χ^2 features form the whole feature set. The training texts and test texts are then converted into their feature representations. The feature representation of the training texts is then used to learn the model and the feature representation of the test texts is used for segmentation. By this way, we expect that an OOV word in the test texts might be found by the segmentation model if the bigrams extracted from this word take higher χ^2 scores.

5 the bootstrapping strategy

The addition of the χ^2 features can be also harmful. Even though it could increase the OOV recall, it also leads to drops in IV recall as we found. To keep the IV recall from falling down, we propose to use bootstrapping strategy. Specifically, we choose to use both models with $\chi2$ features and without $\chi2$ features. We train two models firstly, one is $\chi2$-based and another not. Then we do the segmentation to the test text with the two models simultaneously. Two segmentation results can be obtained. One result is produced by the $\chi2$-based model and has a high OOV recall. The other result is produced by the non-$\chi2$-based model and has higher IV recall. Then we do intersection operation to the two results. It is not difficult to understand that the intersection of the two results has both high OOV recall and high IV recall. We then put the intersection results into the training texts to form a new training set. By this new training set, we train again to get two new models, one $\chi2$-based and another not. Then the two new models are used to segment the test texts. Then we do again intersection to the two results and the common parts are again put into the training texts. We

Table-2. The bakeoff results

test set	R	P	F	Riv	Roov
A	0.925	0.931	0.928	0.944	0.667
B	0.941	0.916	0.928	0.967	0.796
C	0.928	0.918	0.923	0.953	0.730
D	0.948	0.928	0.937	0.965	0.761

repeat this process until a plausible result is obtained.

The whole process can be informally described as the following algorithm:

1. let training set T to be the original training set;
2. for I = 0 to K
   1) train a χ^2-based model and a non-χ^2-base model separately using training set T;
   2) use both models to segment test texts;
   3) do intersection to the two segmentation results
   4) put the intersection results into the training set and get the enlarged training set T
3. train the non-χ^2-based model using training set T, and take the output of this model as the final output;
4. end.

6 The evaluation results

The labeled training texts released by the bakeoff are mainly composed of texts from newspaper. A peculiarity of the training data is that all Arabic numbers, Latin letters and punctuations in the data are double-byte codes. As in Chinese texts, there are actually two versions of codes for Arabic numbers, Latin letters and punctuations: one is single-byte codes defined by the western character encoding standard; another is double-byte codes defined by the Chinese character encoding standards. Chinese normally use both versions without distinguishing them strictly.

The four final test sets released by the bakeoff cover four domains, the statistics of the test sets are shown in table-1. (the size is measured in characters)

Table-1. Test sets statistics

test set	domain	size	OOV rate
A	Literature	51K	0.069
B	Computer	64K	0.152
C	Medicines	52K	0.110
D	Finance	56K	0.087

We train all models using SVM-HMMs[1], we set ε to 0.25. This is a parameter to control the accuracy of the solution of the optimization problem. We set C to half of the number of the sentences in the training data. The C parameter is set to trade off the margin size and training error. We also set a cutoff frequency to feature extraction. Only features are seen more than three times in training data are actually used in the models. We set K = 3 and run the algorithm shown in section 5. This gives our final bakeoff results shown in Table-2.

To illustrate whether the χ^2 statistics and bootstrapping strategy help or not, we also show two intermediate results using the online scoring system provided by the bakeoff[2].Table-3 shows the results of the initial non-χ^2-based model using feature template (a)-(f), table-4 shows results of the initial χ^2-based model using feature template (a)-(h).

As we see from the table-1, table-3 and table-4, the approach present in this paper does improve both the overall performance and the OOV recalls in all four domains.

Table-3 Results of initial non-χ^2-based model

test set	R	P	F	Roov
A	0.921	0.924	0.923	0.632
B	0.930	0.904	0.917	0.758
C	0.919	0.906	0.913	0.687
D	0.946	0.924	0.935	0.750

[1] http://www.cs.cornell.edu/People/tj/svm_light/svm_hmm.html

[2] http://nlp.ict.ac.cn/demo/CIPS-SIGHAN2010/#

Table-4 Results of initial χ^2-based model

test set	R	P	F	Roov
A	0.898	0.921	0.910	0.673
B	0.925	0.914	0.920	0.801
C	0.916	0.922	0.919	0.764
D	0.931	0.937	0.934	0.821

We also do a rapid manual check to the final results; one of the main sources of errors lies in the approach failing to recall numbers encoded by one-byte codes digits. For the labeled training corpus provided by the bakeoff almost do not use one-byte codes for digits, and the type feature seems do not help too much. Actually, such numbers can be recalled by simple heuristics using regular expressions. We do a simple number recognition to the test set of domain D. this will increase the F score from 0.937 to 0.957.

7 Conclusions

This paper introduces the approach we used in the CIPS-SIGHAN-2010 bake-off task of Chinese word segmentation. We propose to use χ^2 statistics to increase OOV recall and use bootstrapping strategy to increase the overall performance. As our final results shows, the approach works in increasing both of the OOV recall and overall F-score.

We also show in this paper that hidden Markov support vector machine can be used to model the Chinese word segmentation problem, by which high f-score results can be obtained like CRF model.

Acknowledgements

This work was supported by National Natural Science Foundation of China under Grant No. 60975054 and National Social Science Foundation of China under Grant No. 06BYY048.

We want to thank Professor Duan Huiming and Mr. Han Dongxu for their generous help at the data preprocessing works.

References

Liang, Nanyuan, 1987. ‘‘written Cinese text segmentation system--cdws”. Journal of Chinese Information Processing, Vol.2, NO.2,pp44–52.(in Chinese)

Gao, Jianfeng et al., 2005, Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach, Computational Linguistics,Vol.31, No.4, pp531-574.

Huang, Changning et al. 2007, Chinese word segmentation: a decade review. Journal of Chinese Information Processing, Vol.21, NO.3,pp8–19.(in Chinese)

Tseng, Huihsin et al., 2005, A conditional random field word segmenter for SIGHAN 2005, Proceedings of the fourth SIGHAN workshop on Chinese language processing. Jeju Island, Korea. pp168-171

Xue, Nianwen, 2003, Chinese Word Segmentation as Character Tagging, Computational Linguistics and Chinese Language Processing. Vol.8, No.1, pp29-48.

Zhao, Hai et al., 2006, Effective tag set selection in Chinese word segmentation via conditional random field modeling, Proceedings of the 20th Pacific Asia Conference on language, Information and Computation (PACLIC-20), Wuhan, China, pp87-94

Tsochantaridis,Ioannis et al., 2005, Large Margin Methods for Structured and Interdependent Output Variables, Journal of Machine Learning Research (JMLR), No.6, pp1453-1484.

Altun,Yasemin et al.,2003, Hidden Markov Support Vector Machines. Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

Low, Jin Kiat et al.,2005, A Maximum Entropy Approach to Chinese Word Segmentation. Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, Jeju Island, Korea,. pp161-164