CRF-based Hybrid Model for Word Segmentation, NER and even POS Tagging

Zhiting Xu, Xian Qian, Yuejie Zhang, Yaqian Zhou
Department of Computer Science & Engineering,
Shanghai Key Laboratory of Intelligent Information Processing,
Fudan University, Shanghai 200433, P. R. China
{zhiting, qianxian, yjzhang, zhouyaqian}@fudan.edu.cn

Abstract

This paper presents systems submitted to the close track of Fourth SIGHAN Bakeoff. We built up three systems based on Conditional Random Field for Chinese Word Segmentation, Named Entity Recognition and Part-Of-Speech Tagging respectively. Our systems employed basic features as well as a large number of linguistic features. For segmentation task, we adjusted the BIO tags according to confidence of each character. Our final system achieve a F-score of 94.18 at CTB, 92.86 at NCC, 94.59 at SXU on Segmentation, 85.26 at MSRA on Named Entity Recognition, and 90.65 at PKU on Part-Of-Speech Tagging.

1 Introduction

Fourth SIGHAN Bakeoff includes three tasks, that is, Word Segmentation, Named Entity Recognition (NER) and Part-Of-Speech (POS) Tagging. In the POS Tagging task, the testing corpora are presegmented. Word Segmentation, NER and POS Tagging could be viewed as classification problems. In a Segmentation task, each character should be classified into three classes, B, I, O, indicating whether this character is the Beginning of a word, In a word or Out of a word. For NER, each character is assigned a tag indicating what kind of Named Entity (NE) this character is (Beginning of a Person Name (PN), In a PN, Beginning of a Location Name (LN), In a LN, Beginning of an Organization Name (ON), In an ON or not-a-NE). In POS tagging task defined by Fourth SIGHAN Bakeoff, we only need to give a POS tag for each given word in a context.

We attended the close track of CTB, NCC, SXU on Segmentation, MSRA on NER and PKU on POS Tagging. In the close track, we cannot use any external resource, and thus we extracted several word lists from training corpora to form multiple features beside basic features. Then we trained CRF models based on these feature sets. In CRF models, a margin of each character can be gotten, and the margin could be considered as the confidence of that character. For the Segmentation task, we performed the Maximum Probability Segmentation first, through which each character is assigned a BIO tag (B represents the Beginning of a word, I represents In a word and O represents Out of a word). If the confidence of a character is lower than the threshold, the tag of that character will be adjusted to the tag assigned by the Maximum Probability Segmentation (R. Zhang et al., 2006).

2 Conditional Random Fields

Conditional Random Fields (CRFs) are a class of undirected graphical models with exponent distribution (Lafferty et al., 2001). A common used special case of CRFs is linear chain, which has a distribution of:

$$P_\Lambda(\vec{y} \mid \vec{x}) = \frac{1}{Z_{\vec{x}}} \exp\left(\sum_{t=1}^{T} \sum_{k} \lambda_k f_k(y_{t-1}, y_t, \vec{x}, t)\right) \quad (1)$$

where $f_k(y_{t-1}, y_t \vec{x}, t)$ is a function which is usually an indicator function; λ_k is the learned weight of feature f_k; and $Z_{\vec{x}}$ is the normalization factor. The feature function actually consists of two kinds of features, that is, the feature of single state and the feature of transferring between states. Features will be discussed in section 3.

Several methods (e.g. GIS, IIS, L-BFGS) could be used to estimate λ_k, and L-BFGS has been showed to converge faster than GIS and IIS. To build up our system, we used Pocket CRF[1].

3 Feature Representation

We used three feature sets for three tasks respectively, and will describe them respectively.

3.1 Word Segmentation

We mainly adopted features from (H. T. Ng et al., 2004, Y. Shi et al., 2007), as following:

a) C_n(n=-2, -1, 0, 1, 2)
b) C_nC_{n+1}(n=-2,-1,0,1)
c) $C_{-1}C_1$
d) $C_nC_{n+1}C_{n+2}$ (n=-1, 0, 1)
e) *Pu*(C_0)
f) $T(C_{-2})T(C_{-1})T(C_0)T(C_1)T(C_2)$
g) $L_{Begin}(C_0)$, $L_{end}(C_0)$
h) *Single*(C_0)

where C_0 represents the current character and C_n represents the n^{st} character from the current character. *Pu*(C_0) indicates whether current word is a punctuation. this feature template helps to indicate the end of a sentence. T(C) represents the type of character C. There are four types we used: (1) Chinese Number ("一/one", "二/two", "十/ten"); (2) Chinese Dates ("日/day", "月/month", "年/year"); (3) English letters; and (4) other characters. The (f) feature template is used to recognize the Chinese dates for the construction of Chinese dates may cause the sparseness problem. $L_{Begin}(C_0)$ represents the maximum length of the word beginning with the character C_0, and $L_{end}(C_0)$ presents the maximum length of the word ending with the character C_0. The (g) feature template is used to decide the boundary of a word. *Single*(C_0) shows whether current character can form a word solely.

3.2 Named Entity Recognition

Most features described in (Y. Wu et al., 2005) are used in our systems. Specifically, the following is the feature templates we used:

a) *Surname*(C_0): Whether current character is in a Surname List, which includes all first characters of PNs in the training corpora.
b) *PersonName*($C_0C_1C_2$, C_0C_1): Whether $C_0C_1C_2$, C_0C_1 is in the Person Name List, which contains all PNs in the training corpora.
c) *PersonTitle*($C_{-2}C_{-1}$): Whether $C_{-2}C_{-1}$ is in the Person Title List, which is extracted from the previous two characters of each PN in the training corpora.
d) *LocationName*(C_0C_1,$C_0C_1C_2$,$C_0C_1C_2C_3$): Whether C_0C_1,$C_0C_1C_2$,$C_0C_1C_2C_3$ is in the Location Name List, which includes all LNs in the training corpora.
e) *LocationSuffix*(C_0): Whether current character is in the Location Suffix List, which is constructed using the last character of each LN in the training corpora.
f) *OrgSuffix*(C_0): Whether current character is in the Organization Suffix List, which contains the last-two-character of each ON in the training corpora.

3.3 Part-Of-Speech Tagging

We employed part of feature templates described in (H. T. Ng et al., 2004, Y. Shi et al., 2007). Since we are in the close track, we cannot use morphological features from external resources such as HowNet, and we used features that are available just from the training corpora.

a) W_n, (n=-2,-1,0,1,2)
b) W_nW_{n+1}, (n=-2,-1,0,1)
c) $W_{-1}W_1$
d) $W_{n-1}W_nW_{n+1}$ (n=-1, 1)
e) $C_n(W_0)$ (n=0,1,2,3)
f) *Length*(W_0)

where C_n represents the n^{th} character of the current word, and *Length*(W_0) indicates the length of the current word.

4 Reliability Evaluation

In the task of Word Segmentation, the label of each character is adjusted according to their reliability. For each sentence, we perform Maximum Probability Segmentation first, through which we can get a BIO tagging for each character in the sentence.

After that, the features are extracted according to the feature templates, and the weight of each feature has already been estimated in the step of training. Then marginal probability for each character can be computed as follows:

[1] http://sourceforge.net/project/showfiles.php?group_id=201943

$$p(y \mid \vec{x}) = \frac{1}{Z(x)} \exp(\lambda_i f_i(\vec{x}, y)) \quad (2)$$

The value of $p(y \mid \vec{x})$ becomes the original reliability value of BIO label y for the current character under the current contexts. If the probability of y with the largest probability is lower than 0.75, which is decided according to the experiment results, the tag given by Maximum Probability Segmentation will be used instead of tag given by CRF. The motivation of this method is to use the Maximum Probability method to enhance the F-measure of In-Vocabulary (IV) Words. According to the results reported in (R. Zhang et al., 2006), CRF performs relatively better on Out-of-Vocabulary (OOV) words while Maximum Probability performs well on IV words, so a model combining the advantages of these two methods is appealing. One simplest way to combine them is the method we described. Besides, there are some complex methods, such as estimation using Support Vector Machine (SVM) for CRF, CRF combining boosting and combining Margin Infused Relaxed Algorithm (MIRA) with CRF, that might perform better. However, we did not have enough time to implement these methods, and we will compare them detailedly in the future work.

5 Experiments

5.1 Results on Fourth SIGHAN Bakeoff

We participated in the close track on Word Segmentation on CTB, NCC and SXU corpora, NER on MSRA corpora and POS Tagging on PKU corpora.

For Word Segmentation and NER, our memory was enough to use all features. However, for POS tagging, we did not have enough memory to use all features, and we set a frequency cutoff of 10; that is, we could only estimate variables for those features that occurred more than ten times.

Our results of Segmentation are listed in the Tabel 1, the results of NER are listed in the Tabel 2, and the results of POS Tagging are listed in the Tabel 3.

	R	P	F	R_{oov}	R_{iv}
CTB	0.9459	0.9418	0.9439	0.6589	0.9628
NCC	0.9396	0.9286	0.9341	0.5007	0.9614
SXU	0.9554	0.9459	0.9507	0.6206	0.9735

Tabel 1. Results of Word Segmentation

MSRA	P	R	F
PER	0.8084	0.8557	0.8314
LOC	0.9138	0.8576	0.8848
ORG	0.8666	0.773	0.8171
Overall	0.873	0.8331	0.8526

Tabel 2. Results of NER

	Total-A	IV-R	OOV-R	MT-R
PKU	0.9065	0.9259	0.5836	0.8903

Tabel 3. Results of POS Tagging

5.2 Errors Analysis

Observing our results of Word Segmentation and POS Tagging, we found that the recall of OOV is relatively low, this may be improved through introducing features aiming to enhance the performance of OOV.

On NER task, we noticed that precision of PN recognition is relative low, and we found that our system may classify some ONs as PNs, such as "吉尼斯(Guinness)/ORG" and "世界记录(World Record)/)". Besides, the bound of PN is sometimes confusing and may cause problems. For example, "胡绳/PER 曾/ 有/ 题词" may be segmented as "胡绳曾/PER 有/ 题词". Further, some words beginning with Chinese surname, such as "丁丑盛夏", may be classified as PN.

For List may not be the real suffix. For example, "玉峰山麓" should be a LN, but it is very likely that "玉峰山" is recognized as a LN for its suffix "山". Another problem involves the characters in the Location Name list may not a LN all the time. In the context "华裔/ 作家/", for example, "华" means Chinese rather than China.

For ONs, the correlative dictionary also exists. Consider sequence "人大代表", which should be a single word, "人大" is in the Organization Name List and thus it is recognized as an ON in our system. Another involves the subsequence of a word. For example, the sequence "湖北钟祥市工业局长", which should be a person title, but "湖北钟祥市工业局" is an ON. Besides, our recall of ON is low for the length of an ON could be very long.

6 Conclusions and Future Works

We built up our systems based on the CRF model and employed multiple linguistics features based on the knowledge extracted from training corpora.

We found that these features could greatly improve the performance of all tasks. Besides, we adjusted the tag of segmentation result according to the reliability of each character, which also helped to enhance the performance of segmentation.

As many other NLP applications, feature plays a very important role in sequential labeling tasks. In our POS tagging task, we could only use features with high frequency, but some low-frequency features may also play a vital role in the task; good non-redundant features could greatly improve classification performance while save memory requirement of classifiers. In our further research, we will focus on feature selection on CRFs.

Acknowledgement

This research was sponsored by National Natural Science Foundation of China (No. 60773124, No. 60503070).

References

O. Bender, F. J. Och, and H. Ney. 2003. Maximum Entropy Models for Named Entity Recognition. *Proceeding of CoNLL-2003*.

A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics, 22(1)*.

H. L. Chieu, H. T. Ng. 2002. Named Entity Recognition: A Maximum Entropy Approach Using Global Information. *International Conference on Computational Linguistics (COLING).*

J. N. Darroch and D. Ratcliff. 1972. Generalized Iterative Scaling for Log-Linear Models. *The Annals of Mathematical Statistics, 43(5)*.

J. Lafferty, A McCallum, and F. Pereira..2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conf. on Machine Learning (ICML).*

R. Li, J. Wang, X. Chen, X. Tao, and Y. Hu. 2004. Using Maximum Entropy Model for Chinese Text Categorization. *Computer Research and Development, 41(4)*.

H. T. Ng and J. K. Low. 2004. Chinese Part-Of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Base or Character-Based? *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP).*

A. Ratnaparkhi. 1997. A Simple Introduction to Maximum Entropy Models for Natural Language Processing. *Institute for Research in Cognitive Science Report, 97(8)*.

F. Sha and F.Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of HLT-NAACL.*

Y. Shi and M. Wang. 2007. A Dual-Layer CRFs Based Joint Decoding Method for Cascaded Segmentation and Labeling Tasks. In *International Joint Conferences on Artificial Intelligence (IJCAI)*.

C. A. Sutton, K. Rohanimanesh, A. McCallum. 2004. Dynamic conditional random fields: factorized probabilistic models for labeling and segmenting sequence data. In *International Conference on Machine Learning (ICML).*

M. Volk, and S. Clematide. 2001. Learn - Filter - Apply -- Forget Mixed Approaches to Named Entity Recognition. *Proceeding of the 6th International Workshop on Applications of Natural Language for Information Systems*.

Y. Wu, J. Zhao, B. Xu and H. Yu. 2005. Chinese Named Entity Recognition Based on Multiple Features. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP).*

H. Zhang, Q. Liu, H. Zhang, and X. Cheng. 2002. Automatic Recognition of Chinese Unknown Words Based on Roles Tagging. *Proceeding of the 19th International Conference on Computational Linguistics.*

R. Zhang, G. Kikui and E. Sumita. 2006. Subword-based tagging by conditional random fields for Chineseword segmentation. Companion volume to theproceedings of the North American chapter of the Association for Computational Linguistics (NAACL).

Y. Zhou, Y. Guo, X. Huang, and L. Wu. 2003. Chinese and English BaseNP Recognition Based on a Maximum Entropy Model. *Journal of Computer Research and Development, 40(3)*.