Designing Special Post-processing Rules for SVM-based Chinese Word Segmentation

Muhua Zhu, Yilin Wang, Zhenxing Wang, Huizhen Wang, Jingbo Zhu
Natural Language Processing Lab
Northeastern University
No.3-11, Wenhua Road, Shenyang, Liaoning, China, 110004
{zhumh, wangyl, wangzx, wanghz}@ics.neu.edu.cn
zhujingbo@mail.neu.edu.cn

Abstract

We participated in the Third International Chinese Word Segmentation Bakeoff. Specifically, we evaluated our Chinese word segmenter NEUCipSeg in the close track, on all four corpora, namely Academis Sinica (AS), City University of Hong Kong (CITYU), Microsoft Research (MSRA), and University of Pennsylvania/University of Colorado (UPENN). Based on Support Vector Machines (SVMs), a basic segmenter is designed regarding Chinese word segmentation as a problem of character-based tagging. Moreover, we proposed post-processing rules specially taking into account the properties of results brought out by the basic segmenter. Our system achieved good ranks in all four corpora.

1 SVM-based Chinese Word Segmenter

We built out segmentation system following (Xue and Shen, 2003), regarding Chinese word segmentation as a problem of character-based tagging. Instead of Maximum Entropy, we utilized Support Vector Machines as an alternate. SVMs are a state-of-the-art learning algorithm, owing their success mainly to the ability in control of generalization error upper-bound, and the smooth integration with kernel methods. See details in (Vapnik, 1995). We adopted svm-light[1] as the specific implementation of the model.

1.1 Problem Formalization

By formalizing Chinese word segmentation into the problem of character-based tagging, we assigned each character to one and only one of the four classes: word-prefix, word-suffix, word-stem and single-character. For example, given a two-word sequence “东南亚人”, the Chinese words for ”Southeast Asia(东南亚) people(人) ”, the character “东” is assigned to the category word-prefix, indicating the beginning of a word; “南” is assigned to the category word-stem, indicating the middle position of a word; “亚” belongs to the category word-suffix, meaning the ending of a Chinese word; and last, “人” is assigned to the category single-character, indicating that the single character itself is a word.

1.2 Feature Templates

We utilized four of the five basic feature templates suggested in (Low et al. , 2005), described as follows:

- $C_n(n = -2, -1, 0, 1, 2)$
- $C_nCn + 1(n = -2, -1, 0, 1)$
- $P_u(C_0)$
- $T(C_{-2})T(C_{-1})T(C_0)T(C_1)T(C_2)$

where C refers to a Chinese character. The first two templates specify a context window with the size of five characters, where C_0 stands for the current character: the former describes individual characters and the latter presents bigrams within the context window. The third template checks if current character is a punctuation or not, and the last one encodes characters' type, including four types: numbers, dates, English letters and the type representing other characters. See detail description and the example in (Low et al. , 2005). We dropped template $C_{-1}C_1$, since,

[1] http://svmlight.joachims.org/

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 217–220,
Sydney, July 2006.

in experiments, it seemed not to perform well when incorporated by SVMs. Slightly different from (Low et al. , 2005), character set representing dates are expanded to include “日”, “月”, “年”, “时”, “分”, “秒”, the Chinese characters for “day”, “month”, “year”, “hour”,“minute”,“second”, respectively.

2 Post-processing Rules

Segmentation results of SVM-based segmenter have their particular properties. In respect to the properties of segmentation results produced by the SVM-based segmenter, we extracted solely from training data comprehensive and effective post-processing rules, which are grouped into two categories: The rules, termed `IV rules`, make efforts to fix segmentation errors of character sequences, which appear both in training and testing data; Rules seek to recall some OOV(Out Of Vocabulary) words, termed `OOV rules`. In practice, we sampled out a subset from training dataset as a development set for the analysis of segmentation results produced by SVM-based segmenter. Note that, in the following, we defined `Vocabulary` to be the collection of words appearing in training dataset and `Segmentation Unit` to be any isolated character sequence assumed to be a valid word by a segmenter. `A segmentation unit` can be a correctly segmented word or an incorrectly segmented character sequence.

2.1 IV Rules

The following rules are named `IV rules`, pursuing the consistence between segmentation results and training data. The intuition underlying the rules is that since training data give somewhat specific descriptions for most of the words in it, a character sequence in testing data should be segmented in accordance with training data as much as possible.

Ahead of post-processing, all words in the training data are grouped into two distinct sets: the `uniquity set`, which consists of words with unique segmentation in training data and the `ambiguity set`, which includes words having more than one distinct segmentations in training data. For example, the character sequence “新世纪” has two kinds of segmentations, as “新 世纪” (new century) and “新世纪” (as a component of some Named-Entity, such as the name of a restaurant).

- For each word in the uniquity set, check whether it is wrongly segmented into more than one segmentation units by the SVM-based segmenter. If true, the continuous segmentation units corresponding to the word are grouped into the united one. The intuition underlying this post-processing rule is that SVM-based segmenter prefers two-character words or single-character words when confronting the case that the segmenter has low self-confidence in some character-sequence segmentation. For example, “复制品” (duplicate) was segmented as “复制 品” and “统一” (unify) was split into “统 一”. This phenomenon is caused by the imbalanced data distribution. Specifically, characters belonging to category `word-stem` are much less than other three categories.

- For each segmentation unit in the result produced by SVM-based segmenter, check whether the unit can be segmented into more than one IV words and, meanwhile, the words exist in a successive form for at least once in training data . If true, replace the segmentation unit with corresponding continuously existing words. The intuition underlying this rule is that SVM-based segmenter tends to combine a word with some suffix, such as “者”、“人”, two Chinese characters representing “person”. For example, “报名 者” (Person in registration) tends to be grouped as a single unit.

- For any sequence in the ambiguity set, such as “新世纪”, check if the correct segmentation can be determined by the context surrounding the sequence. Without losing the generality, in the following explanation, we assume each sequence in the ambiguity set has two distinct segmentations. we collected from training data the word preceding a sequence where each existence of the sequence has one of its segmentations, into a collection, named `preceding word set`, and, correspondingly, the following word into another set, which is termed `following word set`. Analogically, we can produce `preceding word`

`set` and `following word set` for another case of segmentation. When an ambiguous sequence appears in testing data, the surrounding context (in fact, just one preceding word and a following word) is extracted. If the context has overlapping with either of the pre-extracted contexts of the same sequence which are from training data, the segmentation corresponding to one of the contexts is retained.

- More over, we took a look into the annotation errors existing in training data. We assume there unavoidably exist some annotation mistakes. For example, in UPENN, the sequence "中美" (abbreviation for China and America) exists, for eighty-seven times, as a whole word and only one time, exists as "中 美". We regarded the segmentation "中 美" as an annotation error. Generally, when the ratio of two kinds of segmentations is greater than a pre-determined threshold (the value is set seven in our system), the sequence is removed from the ambiguity set and added as a word of unique segmentation into the uniquity set.

2.2 OOV Rules

The following rules are termed `OOV rules`, since they are utilized to recall some of the wrongly segmented OOV words. A OOV word is frequently segmented into two continuous OOV segmentation units. For example, the OOV word "梵蒂冈" (Vatican) was frequently segmented as "梵蒂 冈", where both "梵蒂" and "冈" are OOV character sequences. Continuous OOVs present a strong clue of potential segmentation errors. A rule is designed to merge some of continuous OOVs into a correct segmentation unit. The designed rule is applicable to all four corpora. Moreover, since distinction between different segmentation standards frequently leads to very different segmentation of a same OOV words in different corpora, we designed rules particularly for MSRA and UPENN respectively, to recall more OOVs.

- For two continuous OOVs, check whether at least one of them is a single-character word. If true, group the continuous OOVs into a segmentation unit. The reason for the constraint of at least one of continuous OOVs being single-character word is that not all continuous OOVs should be combined, for example, "德商 拜耳", both "德商" (Germany merchant) and "拜耳" (the company name) are OOVs, but this sequence is a valid segmentation unit. On the other hand, we assume appropriately that most of the cases for character being single-character word have been covered by training data. That is, once a single character is a OOV segmentation unit, there exists a segmentation error with high possibility.

- MSRA has very different segmentation standard from other three corpora, mainly because it requires to group several continuous words together into a Name Entity. For example, the word "中国外交部" (the Ministry of Foreign Affairs of China) appearing in MSRA is generally annotated into two words in other corpora, as "中国" (China) and "外交部" (the Ministry of Foreign Affairs). In our system, we first gathered all the words from the training data whose length are greater than six Chinese characters, filtering out dates and numbers, which was covered by `Finite State Automation` as a pre-processing stage. For each words collected, regard the first two and three characters as `NE prefix`, which indicates the beginning of a Name Entity. The collection of prefixes is termed $S_{p(refix)}$. Analogously, the collection $S_{s(uffix)}$ of suffixes is brought up in the same way. Obviously not all the prefixes (suffixes) are good indicators for Name Entities. Partly inheriting from (Brill, 1995), we applied error-driven learning to filter prefixes in S_p and suffixes in S_s. Specifically, if a prefix and a suffix are both matched in a sequence, all the characters between them, together with the prefix and the suffix, are merged into a single segmentation unit. The resulted unit is compared with corresponding sequence in training data. If they were not exactly matched, the prefix and suffix were removed from collections respectively. Finally resulted S_p and S_s are utilized to recognize Name Entities in the initial segmentation results.

- UPENN has different segmentation standard from other three corpora in that, for some

Corpus	R	P	F	R_{OOV}	R_{IV}
AS	0.949	0.940	0.944	0.694	0.960
MSRA	0.955	0.956	0.956	0.650	0.966
UPENN	0.940	0.914	0.927	0.634	0.969
CITYU	0.965	0.971	0.968	0.719	0.981

Table 1: Our official SIGHAN bakeoff results

`Locations`, such as “北京市” (Beijing) and `Organizations`, such as “外交部” (the Ministry of Foreign Affairs), the last Chinese character presents a clue that the character with high possibility is a suffix of some words. In fact, SVM-based segmenter sometimes mistakenly split an OOV word into a segmentation unit followed by a suffix. Thus, when some suffixes exist as a single-character segmentation unit, it should be grouped with the preceding segmentation unit. Undoubtedly not all suffixes are appropriate to this rule. To gather a clean collection of suffixes, we first clustered together the words with the same suffix, filtering according to the number of instances in each cluster. Second, the same as above, error-driven method is utilized to retain effective suffixes.

3 Evaluation Results

We evaluated the Chinese word segmentation system in the close track, on all four corpora, namely Academis Sinica (AS), City University of Hong Kong (CITYU), Microsoft Research (MSRA), and University of Pennsylvania/University of Colorado (UPENN). The results are depicted in Table 1, where columns R,P and F refer to `Recall`, `Precision`, `F measure` respectively, and R_{OOV}, R_{IV} for the recall of out-of-vocabulary words and in-vocabulary words.

In addition to final results reported in Bakeoff, we also conducted a series of experiments to evaluate the contributions of `IV rules` and `OOV rules`. The experimental results are showed in Table 2, where V1, V2, V3 represent versions of our segmenters, which compose differently of components. In detail, V1 represents the basic SVM-based segmenter; V2 represents the segmenter which applied IV rules following SVM-based segmentation; V3 represents the segmenter composing of all the components, that is, including SVM-based segmenter, IV rules and OOV rules. Since the OOV ratio is much lower than IV correspondence, the improvement made by OOV rules is not so dramatic as IV rules.

Corpus	V1	v2	v3
AS	0.932	0.94	0.944
MSRA	0.939	0.954	0.956
UPENN	0.914	0.923	0.927
CITYU	0.955	0.966	0.968

Table 2: Word segmentation accuracy(F Measure) resulted from post-processing rules

4 Conclusions and future work

We added post-processing rules to SVM-based segmenter. By doing so, we our segmentation system achieved comparable results in the close track, on all four corpora. But on the other hand, post-processing rules have the problems of confliction, which limits the number of rules. We expect to transform rules into features of SVM-based segmenter, thus incorporating information carried by rules in a more elaborate manner.

Acknowledgements

This research was supported in part by the National Natural Science Foundation of China(No. 60473140) and by Program for New Century Excellent Talents in University(No. NCET-05-0287).

References

Nianwen Xue and Libin Shen. 2003. Chinese Word segmentation as LMR tagging. *In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*,pages 176-179.

Vladimir N. Vapnik. 1995. The Nature of Statistical Learning Theory. Berlin: Springer-Verlag.

Jin Kiat Low, Hwee Tou Ng and Wenyuan Guo. 2005. A Maximum Entropy Approach to Chinese Word Segmentation. *In Proceeding of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 161-164.

Eric.Brill. 1995. Transformation-based error-driven learning and natural language processing:A case study in part-of-speech tagging. *Computational Linguistics, 21(4)*:543-565.