Chinese Word Segmentation in ICT-NLP

ShuangLong Li
Institute of Computing Technology,Beijing
University of Science and Technology Beijing,Beijing
slli@ict.ac.cn

Abstract

Chinese word segmentation is always much accounted of in ICT-NLP. In this bakeoff, two different systems in ICT-NLP participated. The one is SYSTEM_#1 evaluated in three tracks -- PK-close, MSR-close and MSR-open, and SYSTEM_#2 PK-open. Through this bakeoff , the development of Chinese segmentation is learned and the problems are found in our systems.

1 System Description

Two different systems in ICT-NLP participated the second bakeoff.

1.1 SYSTEM_#1

The SYSTEM_#1 is implemented mainly based on the log-linear model CRFs(Conditional Random Fields). CRFs are arbitrary undirected graphical models trained to maximize the conditional probability of the desired outputs given the corresponding inputs. We cast the segmentation as one of sequence tagging.

The conditional probability for the tag sequence $T = t_1 t_2 ... t_n$ given a input Chinese sentence $C = c_1 c_2 ... c_n$ is defined by a linear-chain CRF with parameters $\lambda = \{\lambda_1 \lambda_2 ... \lambda_m\}$ to be

$$P_\Lambda(T \mid C) = \frac{1}{Z_c} \exp\left(\sum_{i=1}^{n} \sum_{m} \lambda_m f_m(t_{i-1}, t_i, C, i) \right)$$

Where Z_c is the per-input normalization that makes the probability of all state sequences sum to one. $f_m(t_{i-1}, t_i, C, i)$ is a feature function which is can be any real number.

The most probable tag sequence for an input C,

$$T^* = \arg\max_T P_\Lambda(T \mid C)$$

is determined using the Viterbi algorithm, An *N*-best list of tagging sequences is obtained using modi-fied Viterbi algorithm.

Six tags according to the different positions of one character in a word are used in this model, such as #START(beginning of one sentence), B(beginning of one word), M(middle of one word), E(end of one word), and #END(end of one sentence).

The feature templates used in this model are listed in Table 1.

Description	Feature
current state	t_i, c_i
current & previous states	t_{i-1}, t_i, c_i
current & two previous states	$t_{i-2}, t_{i-1}, t_i, c_i$
state transitions	t_{i-1}, t_i
second previous character	t_i, c_{i-2}
previous character	t_i, c_{i-1}
next character	t_i, c_{i+1}
second next character	t_i, c_{i+2}
previous two characters	t_i, c_{i-2}, c_{i-1}
next two characters	t_i, c_{i+1}, c_{i+2}
previous current & next character	$t_i, c_{i-1}, c_i, t_{i+1}$
current and previous character	t_i, c_{i-1}, c_i
current and next character	t_i, c_i, c_{i+1}
character types of current & two previous & two next characters	$t_i, T_{i-2}, T_{i-1}, T_i, T_{i+1}, T_{i+2}$

Table 1. Feature templates used in SYSTEM_#1

- Any Arabia number like “０，１，２…” is replaced by mark #N;
- Some punctuation like “。，！，？…” is replaced by mark #C;

- Any Chinese letter like “a, A ,b, B …” is replaced by mark #L;
- All the characters are classified to be 7 types, such as : number, letter, time-suffix, etc.

In this model, the parameters are estimated by one Perceptron Algorithm. The parameters can also be trained through the maximum entropy learning algorithms like GIS or IIS, but performed poorly in our tests.

1.2 SYSTEM_#2

The SYSTEM_#2 is mainly a HHMM-based Chinese segmentation system which has participated the 1st bakeoff in 2003. And the improvement has been made by some post processes. The system structure is shown in Figure 1

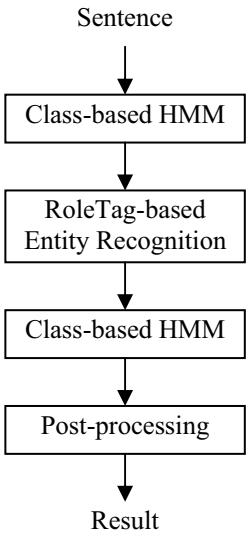

Figure 1 . SYSTEM_#2 structure

Details of the post-processing are listed below.

- TBL(Transformation Based Learning) is applied for adaptation to different segmentation standards.
- The repeating strings extracted in the test data using Accessor Variety correct the inconsistent segmentation.
- A filtering approach to find the unknown words in the single character strings is used.

2 Track

Here ,we introduce the operation of the different tracks. Table 2 gives the results of the tracks.

Tracks	SYSTEM_#1	SYSTEM_#2
PK-close	P: 0.942 R: 0.938 F: 0.940	--
PK-open	--	P: 0.961 R: 0.944 F: 0.952
MSR-close	P: 0.942 R: 0.948 F: 0.945	--
MSR-open	P: 0.933 R: 0.916 F: 0.924	--

Table 2 . The results of the tracks

2.1 Closed Tracks

Only the SYSTEM_#1 participated the closed tracks(PK and MSR) because SYSTEM_#2 must utilize the POS information in the entity recognition process.

All the features information used in the SYSTEM_#1 were trained automatic for one time.

We split the training randomly to two parts—80% used for training and 20% used for estimating the parameters.

2.2 Open Tracks

We participated two open tracks--SYSTEM_#1 MSR and SYSTEM_#2 PK.

For SYSTEM_#1, an external dictionary which contains 140,332 entries with POS is used. However, because of the standard conflict, the open result is weaker than the closed result.

For SYSTEM_#2, the system is trained on six-months tagged news corpus of People Daily 1998 and an external dictionary which contains 22,858 entries with POS.

3 Conclusion

The result in this bakeoff is not so satisfied as there are also many problems in our systems. Through this bakeoff, we learn more about the the development of Chinese segmentation. So the future research is needed to improve our work.