A domain adaption Word Segmenter

For Sighan Bakeoff 2010

Guo jiang
Institute of Intelligent Information Processing, Beijing Information Science & Technology University, Beijing, China, 100192
Guojiang132@gmail.com

Su Wenjie
Institute of Intelligent Information Processing, Beijing Information Science & Technology University, Beijing, China, 100192
dev.sunflower@gmail.com

Yangsen Zhang
Institute of Intelligent Information Processing, Beijing Information Science & Technology University, Beijing, China, 100192
zhangyangsen@163.com

Abstract

We present a Chinese word segmentation system which ran on the closed track of the simplified Chinese Word Segmentation task of CIPS-SIGHAN-CLP 2010 bakeoffs. Our segmenter was built using a HMM. To fulfill the cross-domain segmentation task, we use semi-supervised machine learning method to get the HMM model. Finally we get the mean result of four domains: P=0.719, R=0.72

1 Introduction

The 2010 Sighan Bakeoff included two types of evaluations:
(1) Closed training: In the closed training evaluation, participants can only use data provided by organizers to train their systems specifically, the following data resources and software tools are not permitted to be used in the training:

1) Unspecified corpus;
2) Unspecified dictionary, word list or character list: include the dictionaries of named entity, character lists for specific type of Chinese named entities, idiom dictionaries, semantic lexicons, etc.
3) Human-encoded rule bases;
4) Unspecified software tools, include word segmenters, part-of-speech taggers, or parsers which are trained using unspecified data resources.

The character type information to distinguish the following four character types can be used in training: Chinese characters, English letters, digits and punctuations.

(2) Open training: In the open training evaluation, participants can use any language resource, including the training data provided by organizers

We prefer character-based Tagging than dictionary based word segmentation in closed training, for we can only use the provide train corpus and scale of the corpus is not large enough. If we select dictionary based method we will encounter the out-of-vocabulary problem. But in character-based Tagging method we can yield a better performance than the dictionary based method for such problem.

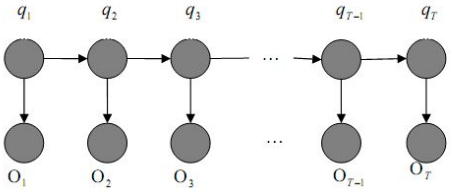
2 Algorithm

Ever before 2002 almost all word segment method is based on dictionary. In SIGHAN 2003 bakeoff, a character-based Tagging method was proposed and since then the character-based Tagging method became more and more popular. HMM (Hidden Markov Model) has been used extensively in speech recognition, pos tagging and get good grades. So we chose HMM as our machine learning method to fulfill our task.
We formally define the elements of an HMM, and explain how the model generates an observation sequence.
An HMM is characterized by the following:

1) N, the number of states in the model. we denote the individual states as $s=\{s_1, s_2, \ldots, s_N\}$,and the state at time t as q_t
2) M, the number of distinct observation symbols per state. we denote the individual symbols as $v=\{v_1, v_2, \ldots, v_M\}$

This work was supported by the national natural science foundation of China (60873013)。 Beijing natural science foundation (KZ200811232019); The Open Project Program of the Key Laboratory of Computational Linguistics (Peking University), Ministry of Education;Funding Project for Academic Human Resources Devel-opment in Institutions of Higher Learning Under the Juris-diction of Beijing Municipality (PHR201007131)

3) The state transition probability distribution A= { a_{ij} } where a_{ij} =P [q $_1$ s_j|q s_i], 1<i,j<N.
4) The observation symbol probability distribution in state j, B={b_j k }, where b_j k P v at t|q s_j
5) The initial state distribution π π_i where π_i P q_1 s_i



Graph1

For convenience, we use the compact notation A, B, π) to indicate the complete parameter set of the model.

There are three basic problems for HMM, for problem 1 we use forward-backward algorithm, for problem 2 we use Viterbi algorithm, for problem 3 we use Baum-Welch algorithm.

To application HMM to our task we define the HMM five factors as blow:

1) We define the whole labels set as Q={B, M, E, S}, B represents word's begin, M represents word's middle, E represents word's end and S represents single word.
2) We define all Unicode characters as O
3) We define A={a_{ij}}, where a_{ij}=P[prior token=s_i|posterior label =s_j]
4) We define B={ b_j k }, where b_j k =P[current character= v |current label =s_j]
5) We define a sentence as a train sample. So π={sentences start with s, s Q}.

Through the design we transform the character-based tagging problem to HMM problem 2. So we can solve this problem with Viterbi algorithm.

3 Experiment

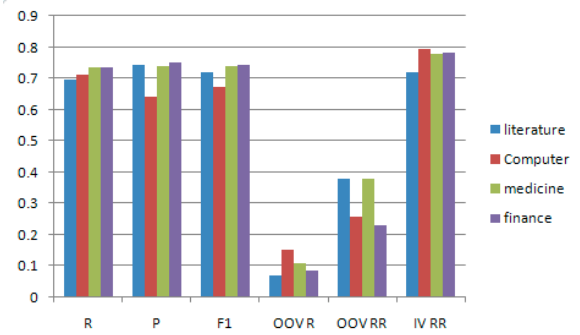
We use HMM to establish the Word Segment prototype system and make use of the Labeled supplied by the Chinese Academy of Sciences to train the HMM and get the model parameters which will be used for the next iterative scaling. After that, we can get a system based on HMM model. Then, with the help of the gotten system, we process the unlabeled corpus. Once it is finished, we should add the processed corpus to the labeled corpus and get a larger corpus with which we can retrain the HMM. All these steps have been done according four test corpuses: literature, computer, medicine, finance. In the table, R indicates the recall rate, P indicates the precision rate, F1 indicates the macro average, OOV R indicates the out-of-vocabulary (OOV) rate, OOV RR indicates the out-of-vocabulary (OOV) self repair rate, IV RR indicates the out-of-vocabulary (OOV) self repair rate. In order to more easily view data, we have presented the Graph2.

From the table and graph, we can see that the finance corpus has a better result, the computer corpus don't show a good result for the R, P, F1. Generally speaking, this result is a reflection for the difference between the dictionary based Tagging method and character-based Tagging method. After recheck our corpus, we can find that there are more technical terms in the computer corpus than finance corpus. The explanation for the result is that if the system encounter a technical terms, the character-based Tagging method will have a bad performance. In such situation, dictionary based Tagging method may have a better performance. For the OOV R and OOV RR, the system has a not bad performance. Table I and Graph2 show the detailed experimental data.

The results of four test corpus as follow:

Type	**R**	**P**	**F1**	**OOV R**	**OOV RR**	**IV RR**
literature	0.695	0.744	0.719	0.069	0.381	0.719
Computer	0.713	0.641	0.675	0.152	0.257	0.795
medicine	0.735	0.74	0.738	0.11	0.378	0.779
finance	0.736	0.752	0.744	0.087	0.23	0.784

Table1



Graph2

4 Conclusion

Our system used a HMM and semi-supervised learning for domain adapting. Our final system achieved a P=0.719, R=0.72. There exist two ways to improve our system performance one is instead our model of CRF, the other is change another way to use the unlabeled data. Because the inherent shortage of HMM we could not get a precise model, and the way we use the unlabeled data can import err to labeled data.

References

Lawrence R. Rabiner. 1989,2. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*. IEEE, VOL.77,No.2,pp:257-286.

Huang Changning, HaoHai. 2007. *Ten Years of Chinese word segmentation*. Vol. 21, No. 3. *JOURNAL OF CHINESE INFORMATION PROCESSING*

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, Christopher Manning. *A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005*.

Blum A, MITCHELL T. *Combining labeled and unlabeled data with co-training* Proceeding of the 11th Annual conference on Computational Learning Theory.

Holmes, W., Russell, M., 1995b. Speech recognition using a linear dynamic segmental HMM. In: Internat. Conf. on Acoust. Speech Signal Process. 1995, Detroit, MI.