A New Model for Automatic Sentence Segmentation

Fukun Xing
Department of General Courses, PLA Foreign Languages University
E-mail: xingfukun001@gmail.com

Received: June 8, 2011 Accepted: June 20, 2011 doi:10.5539/cis.v4n4p134

Abstract

Context Overlapping Model (COM) is presented in this article for the task of Automatic Sentence Segmentation (ASS). Comparing with HMM, COM expands observation from single word unit to n-gram unit and there is an overlapping part between the neighboring units. Due to the co-occurrence constraint and transition constraint, COM model reduces the search space and improves accuracy of segmentation. In this research we treated ASS as a task of sequence labeling and applied 2-gram COM to it. The experiment results show that the overall correct rate of the open test is as high as 90.11%, which is significantly higher than the baseline model (second order HMM), which is 85.16%.

Keywords: COM model, Automatic Sentence Segmentation, Natural Language Processing

1. Introduction

Automatic sentence segmentation (ASS) is an important step in the Automatic Speech Recognition (ASR). Due to the lack of morphological hints such as capitalization the task is more difficult than the sentence boundary detection (SBD). Moreover there are always many misrecognized words, which make ASS more difficult. Little work has been done in this area but recently it gained more interest from the research community.(Mikheev,2003:216) CYBERPUNC (Beeferman, Berger, and Lafferty, 1998) is a system which aims to segment sentences in the speech transcripts. This system was designed to augment a standard trigram language model of a speech recognizer with information about sentence splitting. CYBERPUNC was evaluated on the WSJ corpus and achieved a precision of 75.6% and recall of 65.6%. Other than this experiment few of English experiments are reported on this task. Tanev and Mitkov (2000) had an evaluation of a sentence segmentation system for Slavonic languages. This system used nine main end-of-sentence rules with a list of abbreviations and achieved 92% in precision and 99% in recall measured on a text of 190 sentences. In Chinese, ASS is more difficult due to the ambiguity of sentence boundary in the sense of linguistics. Until now there is no formal rule for the definition of sentence in Chinese. Since Chinese sentences are always segmented by comma as well as period, question mark or exclaimer, comma is always regarded as the sentence boundary. Moreover, some researches (Stevenson and Gaizauskas, 2000) show that this task is not only difficult to computer but also to human. The performance of human is that the precision is 90% and the recall is 75%. And there are substantial disagreements among human annotators.

In our research we regarded the ASS as a task of annotation. We transfer the task of segmenting sentences to the task of determining the state of each word in the sequence. The states include terminal (denote by T) and non-terminal (denote by C). If there is a word sequence w1w2…wn, we want to get a state sequence s1s2…sn. If the state of wi is T, then wi is the end of a sentence. If the state of wi is C, then there is no stop after wi. With such a transfer, we can apply the general tagging model such as Hidden Markov Model (HMM) to this task. HMM is widely used in the task of part of speech tagging. But in ASS the number of state is much fewer than pos tags and state is much dependent on the neighboring words. And some experiments results show that HMM has a poor performance in ASS for the observation independence assumption.

In order to improve the performance of ASS, we create a new model based on HMM, which is named Context Overlapping Model (COM). COM expands the observation from one single word unit to n-gram unit and between the neighboring units there is an n-1 gram part, which is shared by the neighboring units. For the overlapping part the model uses the neighboring observations to determine the current state. The result comparing with HMM shows that COM outperforms HMM in ASS significantly.

The structure of the following part of this article is: in the second part we will introduce COM model. The third part will address how to estimate parameters and handle sparseness data. The fourth part is about evaluation criteria and the fifth part presents the experiments and results. The final part is some discussions and future work to do.

2. COM Model

COM model is based on HMM. HMM is a form of generative model, that defines a joint probability distribution p(X,Y) where X and Y are random variables respectively ranging over observation sequences and their corresponding state sequences. There is an assumption for HMM that the observation element at any given time may only directly depend on the state at that time.

Concerning with observation independence assumption, COM is different from HMM. COM can be divided into different kinds in terms of the length of observation unit. Here we present the formalism of 2-gram COM, in which the length of observation unit is 2 words. The formalisms of other n-gram COM (n>2) can be gotten according to the formalism of 2-gram model.

In the 2-gram COM there is a basic state set $Q=\{q_1,q_2,...q_s\}$. The observation sequence is S= $w_1 \ldots w_h$. The corresponding state of a 2-gram observation unit $w_{i-1}w_i$ (2≤i≤h) is a state set $e_i=\{q_{i-1}^j q_i^j\}$, in which q_{i-1}^j is one of the basic states of w_{i-1} and q_i^j is one of the basic states of w_i. The state sequence $q_{i-1}^j q_i^j$ is called one state unit of the observation unit $w_{i-1}w_i$. It is notable that e_i is the state set when the word w_{i-1} and w_i co-occur, which is called Co-occurrence Constraint(CC). When w_{i-1} and w_i co-occur the amount of possible states of $w_{i-1}w_i$ will not be more than the amount of the combination of states of w_{i-1} and w_i.

The search for the state sequence with the highest joint probability can be computed as:

$$\widehat{Q}=\arg\max \mathrm{P(Q\mid S)}=\arg\max \mathrm{P(Q)P(S\mid Q)}\approx$$

$$\arg\max_{q_{i-1},q_i}\ (p(q_1)p(q_2\mid q_1)\prod_{i=3}^{h}p(q_{i-1}q_i\mid q_{i-2}q_{i-1})p(o_1\mid q_1)\prod_{i=2}^{h}p(o_{i-1}o_i\mid q_{i-1}q_i))$$

Q denotes the state sequence and S denotes the observation sequence. $\widehat{Q}$ denotes the final state sequence, whose joint probability is the highest.

For the convenience of computation, we insert 2 "*B*", whose state is "B" at the beginning of the sequence and insert 2 "*E* ", whose state is "E" at the end of the sequence. And then the above formula will be:

$$\hat{Q}=\arg\max_{q_{i-1},q_i}\ (\prod_{i=1}^{h+2}p(q_{i-1}q_i\mid q_{i-2}q_{i-1})\prod_{i=1}^{h+2}p(o_{i-1}o_i\mid q_{i-1}q_i))$$

In this model there is an overlapping part between the neighboring observation units $w_{i-2}w_{i-1}$ and $w_{i-1}w_i$. For w_{i-1} is shared by the neighboring units, the corresponding states units of $w_{i-2}w_{i-1}$ and $w_{i-1}w_i$ should also share the same overlapping state. If $q_{i-2}^k q_{i-1}^k$ is one state of $w_{i-2}w_{i-1}$ and $q_{i-1}^j q_i^j$ is one state of $w_{i-1}w_i$, then only if q_{i-1}^k is the same as q_{i-1}^j then it is possible to transmit from state $q_{i-2}^k q_{i-1}^k$ to $q_{i-1}^j q_i^j$, otherwise there is no transition path from $q_{i-2}^k q_{i-1}^k$ to $q_{i-1}^j q_i^j$. The constraint $q_{i-1}^k=q_{i-1}^j$ is called Transition Constraint (TC).

$\widehat{Q}$ is a sequence consisting of h+1 2-gram state units like:

$B\hat{q}_1,\ \hat{q}_1\hat{q}_2,\ \hat{q}_2\hat{q}_3,...,\ \hat{q}_{h-1}\hat{q}_h,\ \hat{q}_h E\ \ (\hat{q}_i\in Q)$

It is obvious that the final state sequence can be gotten from the above sequence.

3. Parameters estimation and Evaluation Criteria

There are 2 main parameters to be estimated in COM:

(1) P_t:State transition probability;

(2) P_e:State emission probability.

We apply the maximum likelihood to estimate these parameters from the tagged corpus. The details of the estimation will not be introduced here.

For the expansion of the observation the sparseness problem in n-gram COM is more serious than that in HMM. COM applies back-off strategy to deal with the sparseness data. The main idea is that if n-gram (n>2) $w_{i-n+1}...w_i$ is not in the n-gram vocabulary, which is gotten from the training corpus, it will be replaced by n-1 gram $w_{i-n+2}...w_i$. And if $w_{i-1}w_i$ is not in the 2-gram vocabulary then the state units of $w_{i-1}w_i$ will be replaced by the combination of states of w_{i-1} and w_i. If w_i is not in the unigram vocabulary it will be handled as same as in HMM.

We use the following criteria to evaluate the performances of COM in ASS.

(1) Overall Correct rate (C):

$$P = \frac{Correct\_Tags}{Total\_Original\_Tags}$$

(2) Precision rate (P) of terminal tags:

$$P = \frac{Correct\_Ter\min al\_Tags}{Total\_Original\_Ter\min al\_Tags}$$

Correct_Terminal_Tags denotes the number of correct sentence stop tags by the tagging model. The Total_Terminal_Original_Tags denotes the total terminal tags in the original text.

(3) Recall rate(R) of terminal tags:

$$R = \frac{Correct\_Ter\min al\_Tags}{Total\_Ter\min al\_Tags\_by\_Model}$$

Total_Terminal_Tags_by_Model denotes the total sentence terminal tags by the model.

(4) F score:

$$F = \frac{2(P*R)}{(P+R)}$$

F score denotes the average performance of the model.

4. Experiments and Discussion

4.1 Corpus and Preprocessing

We apply COM to the Chinese ASS task. For we have no speech transcripts and we just focus on the performance of COM without any usage of the speech information, we take the general Chinese corpus as the training and test corpus. The training and test data are all taken from the People's Daily of year 2000 , which has been segmented and manually assigned PoS tags by the Peking University. The division of corpus is displayed in table 1.

Before training and tagging the corpus is preprocessed. First, all the named entities such as personal names, location names, organization names and all the digits are replaced by some particular symbols. For example, personal names are all replaced by "*PerN*". Second, we transfer the comma, period, question mark, exclaim mark, semi-colon, colon to one tag "T" as the sentence terminal.

The baseline model is the 2nd order HMM, whose results will be compared with that of 2-gram COM.

4.2 Results

The results are shown in table2. The precision and recall rate of COM all outperform HMM significantly.

It is interesting to see that the gap between the C rates of both models is only about 5 per cent, but gap between other evaluators are much bigger than 15 per cent. The possible reason is the unbalance distribution of word states. In the test corpus the number of T is 20134 and N is 134164. The number of state N is much more than state T. So even if we guess all the states to be N the overall correct rate will not lower than 85%. For this reason and the observation independence assumption of HMM, HMM has a poor performance in ASS. HMM assign much more probability to state N both in observation and transition probability regardless the neighboring words. But in COM it takes the neighboring words into the model and the observation and transition probability will both be influenced. In this sense COM outperforms HMM in the guessing of sentence terminal.

4.3 Discussion

COM is not only suitable to the task of ASS. We have applied it to the Chinese word segmentation, part of speech tagging and chunk detection, in which COM also achieves satisfactory results. Comparing with HMM, COM has the advantages of smaller search space and higher tagging precision rate. Comparing with the discriminative models such as CRF and Maximum Entropy, COM has the advantages of less training time and comparable precision rate. All of these prove that COM is a general, efficient and robust model for sequence labeling.

5. Conclusion

In this paper we apply the COM model to the task of ASS and achieve better performance than the HMM model. COM is superior to HMM because it overcomes the limit of observation independence assumption with the expansion of observation unit and construct an overlapping part between neighboring unit. In the further studies, we will explore possible applications of COM in the sequence labeling task in natural language processing.

References

Beeferman, D., A. Berger, and J. Lafferty. (1998). 'CYBERPUNC: a lightweight punctuation annotation system for speech'. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (Seattle), 689-92.

L. Rabiner. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proc. of the IEEE, 77(2).

Mitkov, Ruslan. (2003). *The Oxford Handbooks of Computational Linguistics.* New York: Oxford University Press. 201-218.

Stevenson, A. and R. Gaizauska. (2000). Experimenting on Sentence Boundary Detection. Proceedings of the 6th Applied National Language Processing Conference (Seattle), 84-9.

Tenev, H. and R. Gaizauskas. (2000). LINGUA: a robust architechture for text processing and anaphora resolution in Bulgarian. Proceedings of the International Conference on Machine Translation and Multilingual Applications (Exeter), 20.1-20.8.

Table 1. Division of corpus

Group	Usage of corpus	Months	Amount of tokens
1	Training	Feb.-June.	6142402
2	Open Test	Jan.	1235628

Table 2. Results of the open test

	C	P	R	F
2nd order HMM	85.16%	43.17%	43.40%	43.29%
2-gram COM	90.11%	60.98%	67.27%	63.97%