An Double Hidden HMM and an CRF for Segmentation Tasks with Pinyin's Finals

Huixing Jiang **Zhe Dong**
Center for Intelligence Science and Technology
Beijing University of Posts and Telecommunications
Beijing, China
jhx0129@163.com jimmybupt@gmail.com

Abstract

We have participated in the open tracks and closed tracks on four corpora of Chinese word segmentation tasks in CIPS-SIGHAN-2010 Bake-offs. In our experiments, we used the Chinese inner phonology information in all tracks. For open tracks, we proposed a double hidden layers' HMM (DHHMM) in which Chinese inner phonology information was used as one hidden layer and the BIO tags as another hidden layer. N-best results were firstly generated by using DHHMM, then the best one was selected by using a new lexical statistic measure. For close tracks, we used CRF model in which the Chinese inner phonology information was used as features.

1 Introduction

Chinese language has many characteristics not possessed by other languages. One obvious is that the written Chinese text does not have explicit word boundaries like western languages. So word segmentation became very significative for Chinese information processing, and is usually considered as the first step of any further processing. Identifying words has been a basic task for many researchers who have devoted themselves on Chinese text processing.

The biggest characteristic of Chinese language is its trinity of sound, form and meaning (Pan, 2002). Hanyu Pinyin is the form of sound for Chinese text and the Chinese phonology information is explicit expressed by Pinyin which is the inner features of Chinese Characters. And it naturally contributes to the identification of Out-Of-Vacabulary words (OOV).

In our work, Chinese phonology information is used as basic features of Chinese characters in all models. For open tracks, we propose a new double hidden layers HMM in which a new phonology information is built in as a hidden layer, a new lexical association is proposed to deal with the OOV questions and domains' adaptation questions. And for closed tracks, CRF model has been used , combined with Chinese inner phonology information. We used the CRF++ package Version 0.43 by Taku Kudo[1].

In the rest sections of this paper, we firstly introduce the Chinese phonology in Section 2. Then in the Section 3, the models used in our tasks are presented. And the experiments and results are described in Section 4. Finally, we give the conclusions and make prospect on future work.

2 Chinese Phonology

Hanyu Pinyin is the form of sound for Chinese text and the Chinese phonology information is explicit expressed by Pinyin. It is currently the most commonly used romanization system for Standard Mandarin. Hanyu means the Chinese language, and Pinyin means "phonetics", or more literally, "spelling sound" or "spelled sound" (wikipedia, 2010). The system has been employed to teach Mandarin as home language or as second language by China, Malaysia, Singapore et.al. Pinyin has been the most Chinese character's input method for computers and other devices.

[1] http://crfpp.sourceforge.net/

The romanization system was developed by a government committee in the People's Republic of China, and approved by the Chinese government on February 11, 1958. The International Organization for Standardization adopted pinyin as the international standard in 1982, and since then it has been adopted by many other organizations(wikipedia, 2010). In this system, pinyin is composed by initials(pinyin: shengmu), finals(pinyin: yunmu) and tones(pinyin: shengdiao) instead of consonants and vowels used in European language. For example, the Pinyin of "中" is "zhong1" composed by "zh", "ong" and "1". In which "zh" is initial, "ong" is final and "1" is the tone.

Every language has its rhythm and rhyme, so Chinese is no exception. The rhythm system are the driving force from the unconscious habit of language(Edward, 1921). And the Pinyin's finals contribute the Chinese rhythm system, Which is the basic assumption our research based on.

3 Algorithms

Generally the task of segmentation can be viewed as a sequence labeling problem. We first define a tag set as TS = {B, I, E, S}, shown in Table 1.

Table 1: The tag set used in this paper.

Label	Explanation
B	beginning character of a word
I	inner character of a word
E	end character of a word
S	a single character as a word

For the piece "是英国前王妃戴安娜" of the example described in the experiments section, firstly, the TS tags are labeled to it. And its result is "是/S 英/B 国/E 前/S 王/B 妃/E 戴/B 安/I 娜/E". Then the tags are combined sequentially to get the finally result "是_英国_前_王妃_戴安娜".

In this section, A novel HMM solution is presented firstly for open tracks. Then the CRF solution for closed tracks is introduced.

3.1 Double hidden layers' HMM

For a given piece of Chinese sentence, $X = x_1x_2 \dots x_T$, where $x_i, i = 1, \dots, T$ is a Chinese character. Suppose that we can give each Chinese character x_i a Pinyin's final y_i. And suppose the label sequence of X is $S = s_1s_2 \dots s_T$, where $s_i \in TS$ is the tag of x_i. Then what we want to find is an optimal tag sequence S^* which is defined in (1).

$$\begin{aligned} S^* &= \arg\max_S P(S, Y|X) \\ &= \arg\max_S P(X|S, Y)P(S, Y) \end{aligned} \quad (1)$$

The model is described in Fig. 1. For a given piece of Chinese character strings, One hidden layer is label sequence S. Another hidden layer is Pinyin's finals sequence Y. The observation layer is the given piece of Chinese characters X.

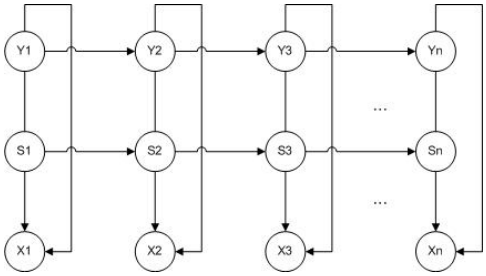

Figure 1: Double Hidden Markov Model

For transition probability, second-order Markov model is used to estimate probability of the double hidden sequences as described in (2).

$$P(S, Y) = \prod_t p(s_t, y_t|s_{t-1}, y_{t-1}) \quad (2)$$

For emission probability, we keep the first-order Markov assumption as shown in (5).

$$P(X|S, Y) = \prod_t p(x_t|s_t, y_t) \quad (3)$$

3.1.1 Nbest results

Based on the work of (Jiang, 2010), a word lattice is also built firstly, then in the second step, the backward A^* algorithm is used to find the top N results instead of using the backward viterbi algorithm to find the top one. The backward A^* search algorithm is described as follow (Wang, 2002; Och, 2001).

3.1.2 Reranking with a new lexical statistic measure

Given two random Chinese characters X and Y and assume that they appears in an aligned region of the corpus. The distribution of the two random Chinese characters could be depicted by a 2 by 2 contingency table shown in Fig. 2(Chang, 2002).

	Y	$\neg Y$
X	a	b
$\neg X$	c	d

Figure 2: A 2 by 2 contingency table

In Fig. 2, a is the counts of X and Y co-occur; b is the counts of the cases that X occurs but Y does not; c is the counts of the cases that X does not occur but Y does; d is the counts of the cases that both X and Y do not occur. The Log-likelihood rate is calculated by (4).

$$\begin{aligned} LLR(x,y) = 2(&a \cdot \log \frac{a \cdot N}{(a+b)\cdot(a+c)} \\ &+ b \cdot \log \frac{b \cdot N}{(a+b)\cdot(b+d)} \\ &+ c \cdot \log \frac{c \cdot N}{(c+d)\cdot(a+c)} \\ &+ d \cdot \log \frac{d \cdot N}{(c+d)\cdot(b+d)}) \end{aligned} \quad (4)$$

For the N-best result described in sec. 3.1.1, they can be re-ranked by (5).

$$S^* = \arg\min_S (score_h(S) + \frac{\lambda}{K}\sum_{k=1}^{K} LLR_(x_k, y_k)) \quad (5)$$

where $score_h$ is the negative log value of $P(S, Y|X)$. K is the number of breaks in X and x_k is the left Chinese character of the k break and y_k is the right Chinese character of the k break. λ is the regulatory factor(in our experiments $\lambda = 0.45$).

Bigger value of $LLR(x_k, y_k)$ means stronger ability in combining of the two characters x_k and y_k, then they should not be segmented.

3.2 CRF model for closed tracks

Conditional random field, as statistical sequence labeling model, has been used widely in segmentation(Lafferty, 2001; Zhao, 2006). In the closed tracks of the paper, we also use it.

3.2.1 Feature templates

We adopted two main kinds of features: n-gram features and Pinyin's finals features. The n-gram feature set is quite orthodox, they are, namely, C-2, C-1, C0, C1, C2, C-2C-1, C-1C0, C0C1, C1C2. The Pinyin's finals feature set is the same as n-gram feature set. They are described in Table. 2.

Table 2: Feature templates

Templates	Category
C-2, C-1, C0, C1, C2	N-gram: Unigram
C-2C-1, C-1C0, C0C1, C1C2	N-gram: Bigram
P-2, P-1, P0, P1, P2	Phonetic: Unigram
P-2P-1, P-1P0, P0P1, P1P2	Phonetic: Bigram

4 Experiments and Results

4.1 Dataset

We build a basic words dictionary for DHHMM and a Pinyin's finals dictionary for both DHHMM and CRF from The Grammatical Knowledge-base of Contemporary Chinese(Yu, 2001). For the finals dictionary, we give each Chinese character a final extracted from its Pinyin. When it comes to a polyphone, we just combine its all finals simply to one. For example, "中{ong}", "差{a&ai&i}".

The training corpus (5,769 KB) we used is the Labeled Corpus provided by the organizer. We firstly add the Pinyin's finals to each Chinese character of it, then we train the parameters of DHHMM and CRF model on it.

And the test corpus contains four domains: Literature (A), Computer (B), Medicine (C) and Finance(D).

The LLR function's parameters{a, b, c, d} are counted from the current test corpus A, B, C, or D. It's means that for segmenting A, the LLR parameters are counted from A, so the same for segmenting B, C and D.

4.2 Preprocessing

The date, time, numbers and symbols information are easily identified by rules. We propose four regular expressions' processes, in which the regular expressions' processes are handled one after another in order of date, time, numbers and symbols. By now, a rough segmentation can be done. For a character stream, the date, time, numbers and symbols are firstly identified, then the whole stream can be divided by these units to some pieces of character strings which will be segment by the models described in sec. 3. For example, a character stream "2009年的8月31日，是英国前王妃戴安娜12周年忌日。" will be divided to "2009年_的_8月_31日_，_是英国前王妃戴安娜_12_周年忌日_。". Then the pieces "的", "是英国前王妃戴安娜", "周年忌日" will be segmented sequentially by the models described in Section 3.

4.3 Results on DHHMM

We evaluate our system by Precision Rate(6), Recall Rate(7), F1 measure(8) and OOV(Out-Of-Vocabulary) Recall rate(9).

$$P = \frac{C(correct\ words\ in\ segmented\ result)}{C(words\ in\ segmented\ result)} \tag{6}$$

$$R = \frac{C(correct\ words\ in\ segmented\ result)}{C(words\ in\ standard\ result)} \tag{7}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{8}$$

$$OR = \frac{C(correct\ OOV\ in\ segmented\ result)}{C(OOV\ in\ standard\ result)} \tag{9}$$

In (6-9), $C(\cdots)$ is the count of $(\cdots)$.

Table 3 are the results of the DHHMM on open tracks.

In Table 3, $OOVRR$ is the recall rate of OOV, $IVRR$ is the recall rate of IV(In Vocabulary).

4.4 Postprocessing for CRF and Results on It

Since the CRF segmenter will not always return a valid tag sequence that can be translated into segmentation result, some corrections should be made if such error occurs. We devised a dynamic programming routine to tackle this problem: first we compute the valid tag sequence that closest to the output of CRF segmenter (by term closest, we mean least hamming distance), if there is a tie, we choose the one has the least 'S' tags, if the tie still exists, we choose the one that comes lexicographically earlier ($B < I < E < S$, described in Table. 1). Table 4 are the results of the CRF on closed tracks.

Table 3: Results of open tracks using DHHMM: Literature (A), Computer (B), Medicine (C) and Finance(D)

	A	B	C	D
R	0.893	0.918	0.917	0.928
P	0.918	0.896	0.907	0.934
F1	0.905	0.907	0.912	0.931
OOV RR	0.803	0.771	0.704	0.808
IV RR	0.899	0.945	0.943	0.939

Table 4: Results of closed tracks using CRF: Literature (A), Computer (B), Medicine (C) and Finance(D)

	A	B	C	D
R	0.945	0.946	0.94	0.956
P	0.946	0.914	0.928	0.952
F1	0.946	0.93	0.934	0.954
OOV RR	0.816	0.808	0.761	0.849
IV RR	0.954	0.971	0.962	0.966

From the results of Table 3 and Table 4, we can observe that the CRF model outperforms the DHHMM by average 2.72% in F1 measure. In the other hand, from Table 5, we can see that the computation cost in DHHMM is less than half of the time cost and lower one-fifth memory cost than CRF model.

Table 5: The computation cost in DHHMM and CRF

	Time cost(ms)	Memory cost(MB)
DHHMM	34398	16.3
CRF	43415	35

5 Conclusions and Future works

This paper has presented a double hidden lawyers HMM for Chinese word segmentation task in SIGHAN bakeoff 2010. It firstly created N top results and then select the best one from it by a new lexical association.

Chinese phonology (specially by Pinyin's final in text) is very useful inner information of Chinese language, which is the first time used in our models. We have used it in both DHHMM and CRF model.

In future work, there are lots of improvements can be done. Firstly, which polyphone's finals should be used in a given context is a visible question. And the strategy to train the parameter λ described in 3.1.2 can also be improved.

Acknowledgments

This research has been partially supported by the National Science Foundation of China (NO. NSFC90920006). We also thank Caixia Yuan for leading our discuss, Li Sun, Peng Zhang, Yaojing Chen, Zhixu Lin, Gan Lin, Guannan Fang for their useful helps in this work.

References

Wenguo Pan. 2002. *zibenwei yu hanyu yanjiu*:120–141. East China Normal University Press.

Sapir Edward 1921. *Language: An introduction to the study of speech*:230. New York: Harcourt, Brace and company.

wikipedia. 2010. *Pinyin*. http://en.wikipedia.org/wiki/Pinyin#cite_note-6.

Baobao Chang, Pernilla Danielsson, and Wolfgang Teubert. 2002. *Extraction of translation unit from chinese-english parallel corpora*, Proceedings of the first SIGHAN workshop on Chinese language processing:1–5.

Huixing Jiang, Xiaojie Wang, Jilei Tian. 2010. *Second-order HMM for Event Extraction from Short Message*, 15th International Conference on Applications of Natural Language to Information Systems, Cardiff, Wales, UK.

Franz Josef Och, Nicola Ueffing, Hermann Ney. 2001. *An Efficient A^* Search Algorithm for Statistical Machine Translation*, Proceedings of the ACL Workshop on Data-Driven methods in Machine Translation 14(Toulouse, France): 1-8.

Ye-Yi Wang, Alex Waibel. 2002. *Decoding Algorithm in Statistical Machine Translation*, Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics: 366-372.

Yu Shiwen, Zhu Xuefeng, Wang Hui. 2001. *New Progress of the Grammatical Knowledge-base of Contemporary Chinese*, ZHONGWEN XINXI XUEBAO, 2001 Vol. 01.

John Lafferty, A.Mccallum, F.Pereira. 2001. *Conditional Random Field: Probabilitic Models for Segmenting and Labeling Sequence Data.*, Proceedings of the Eighteenth International Conference on Machine Learning: 282–289.

Hai Zhao, Changning Huang, Mu Li. 2006. *An Improved Chinese Word Segmentation System with Conditional Random Field*, Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing (SIGHAN-5)(Sydney, Australia):162-165.