Chinese New Word Detection Based on HHMM

ZOU Gang, MENG Yao, YU Hao[*1]

ZHANG Yujie, ISAHARA Hitoshi[*2]

USHIODA Akira[*3]

*1: Fujitsu Research and Development Centre, Beijing, China

*2: NICT

*3: Fujitsu Laboratory

Abstract

In this paper, we propose a method for detecting new words. Firstly, we exploit the knowledge of internal structures of Chinese words to generate new word candidates. Secondly, we use HHMM to rank them and get n-best results. Thirdly, we use bi-gram extracted from annotated corpus to re-rank these n-best results. The Experimental results show that the proposed approach yields 50.6% precision and 71.4% recall when the top results are evaluated, and nearly 90% recall when n-best results are all evaluated (n = 4).

1. Introduction

Since there is no delimiter between words in written Chinese, word segmentation is required in Chinese processing. In word segmentation, new words and ambiguities are two main problems. Sometimes, they occur at the same time, which brings many difficulties for word segmentation and it is hard to be solved. Therefore, new word detection is one of the essential tasks.

New words usually refer to OOV (out of vocabulary) words. General speaking, there are five types of new words from semantic perspective.[1] [2]

(a) Named entities: person name, location name, organization name.

(b) Numeric words: e.g. 15%

(c) Abbreviation: e.g. 亚 冬 会 (Winter Asian Games), 春晚(Spring Festival Party), 非典(SARS)

(d) Derived words: e.g. 常态性(always)

(e) Compound: e.g. 版面费(publishing fee), 拆装(remove and install),

From a surface pattern perspective, new words can be classified into the following types[1]. (1) NW_11 (two-character new word, ‘1+1’), e.g. 拆|装 (2) NW_12 (a single character followed with a bi-character word, ‘1+2’), e.g. 大|世界(Grand World) (3) NW_21 (a bi-character word followed with a single character, ‘2+1’), e.g.版面|费 (4) NW_111 (three-character words, ‘1+1+1’), e.g. 亚|冬|会 (4) NW_22 (two bi-character words, ‘2+2’), e.g. 交通|部门 (5) Others.

As it is reported before,[1] the majority of the new words are NW_11, NW_111, NW_12 and NW_21, so in this paper we focus on detection of NW_11, NW_111, NW_12 and NW_21 of type (c), (d), (e). In the following content, new words refer to these types. There are a few researches on detection of these new words. There are mainly three approaches. (1) Word frequency is used to identify new words.[3][4] (2) Character-based tagging and chunking are used to identify new words.[5] (3) IWP (Independent Word Probability) and SVM are applied to identify new words.[1][6] Because the reported results are obtained in different corpus, it is hard to tell which approach is better. In this paper, we try to apply HHMM(Hierarchical Hidden Markov Model)[7] to this task.

In the rest of the paper, Section 2 presents the framework of our method; Section 3 describes generation of new word candidates; Section 4 describes the application of HHMM to new word detection task; Section 5 introduces re-rank method; Section 6 reports experimental results and error analysis; Section 7 is the conclusion and future works.

2. Framework of New Word Detection

In the dictionary-based word segmentation method, new word will be split into pieces in most cases. That is to say, if single-character sequences appear in the segmentation result, it is possible that there exist new

words, so new word detection can be a post-process on segmentation result. We regard every single-character sequence as candidate sequence containing new words, and then process those sequences. In this way, time and space will be reduced a lot, compared with new word detection without word segmentation. New word detection, therefore, becomes a problem of deciding whether those single-character words should be combined or not and how they should be combined. So input of the process is word segmentation result, output is new word list. The framework of the whole process is shown as Figure 1.

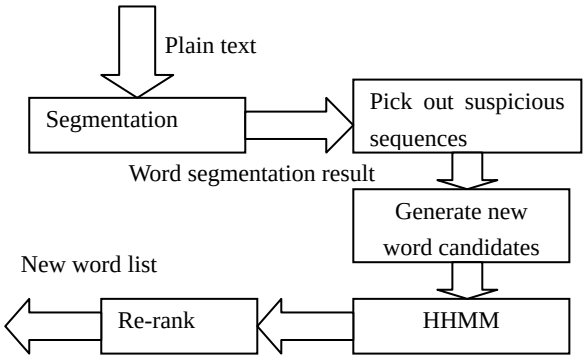

Figure 1. Framework of new word detection

3. Generation of New Word Candidates

There are internal structures in the words, e.g. in the word '中国' (China), '国'(country) often appears in the right of the word, such as '美国'(America), '法国'(France). Therefore, the basic assumption of our approach is every character has its roles in forming a word and appears in the certain position. Thus, we use 'L_', 'R_', 'M_', 'P_' and 'S_' to denote the character's various roles. Here, 'L_', 'M_', 'R_' means left, middle, right position of the word separately, 'P_' and 'S_' mean the positions of single character in '1+2' and '2+1' respectively. Table 1 shows some examples of this expression. Part of speech tag set of Peking University is adopted in Table 1.

Those character roles are used to 'guess' new word during detection process. E.g. in Table 1, '铲' has the role of 'L_v' and '射' has the role of 'R_v'. When '铲' and '射' appear in the sequence as two continuous single-character words, '铲射' (shovel shot) is a possible new word with the POS 'v'. In this way, new word candidates can be generated.

Table 1. Character roles

Type	Word: POS	Character: Role
NW_11	黄海: ns (Yellow Sea)	黄:L_ns 海:R_ns
	周报: n (weekly)	周:L_n 报:R_n
	界定: v （distinguish）	界:L_v 定:R_v
	铲除: v (eradicate)	铲:L_v 除:R_v
	发射: v (launch)	发:L_v 射:R_v
NW_12	大世界: n	大:P_n 世:L_n 界:R_n
	全过程: n (whole procedure)	全:P_n 过:L_n 程:R_n
	半公开: v （semi-overt）	半:P_v 公:L_v 开:R_v
NW_21	黄金周: n (Golden Week)	黄:L_n 金:R_n 周:S_n
	爆破工: n (shotfirer)	爆:L_v 破:R_v 工:S_n
NW_111	氟里昂: n (Chlorofluorocarbon)	氟:L_n 里:M_n 昂:R_n
	亚冬会: n	亚:L_n 冬:M_n 会:R_n

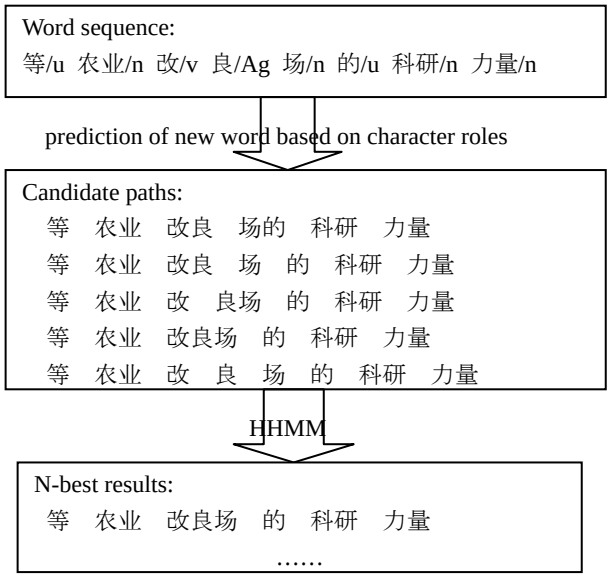

Figure 2. An example of process

4. Application of HHMM

Usually single character has various roles as seen in Table 1, e.g. character '黄'(L_ns, L_n), '周'(L_n, S_n), therefore, lots of candidate paths occur. An example is shown in Figure 2. Consequently, we use HHMM for ranking those paths. HHMM is an

improvement of HMM. It has the ability of describing hierarchical structures. Because we split word states into character roles, it is the same as adding one layer between the layer of final output symbol and the layer of word state. So HHMM is suitable for the new word model.

The state transition of HHMM is described as Figure 3.

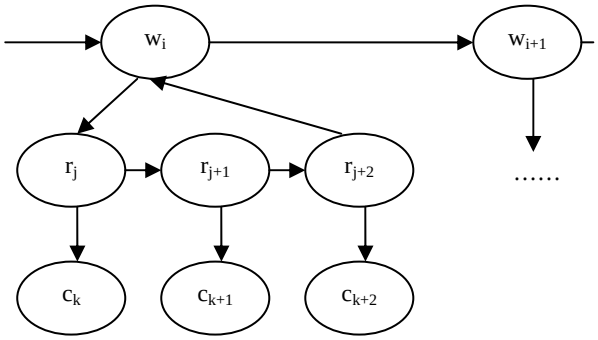

Figure 3 State Transition of HHMM
w is word state; r is character role state; c is final output symbol which is a character.

An observation sequence is denoted by $\bar{O} = o_1 o_2 ... o_T$. An HHMM state is denoted by q_i^d ($d \in \{1,..., D\}$), where i is the state index and d is the hierarchy index. The hierarchy index of the root is 1 and of the production is D. In our approach, D is 2. The state transition probability matrix is denoted by $A^{q^d} = \{a_{ij}^{q^d}\} = \{P(q_j^{d+1} \mid q_i^{d+1})\}$. And the probability vector of making a vertical transition is denoted by $\prod^{q^d} = \{\pi^{q^d}(q_i^{d+1})\} = \{P(q_i^{d+1} \mid q^d)\}$, and the output probability vector is denoted by $B^{q^D} = \{b^{q^D}(k)\} = \{P(\sigma_k \mid q^D)\}$, where σ_k is the final output symbol.[7]

The entire set of parameters is denoted by $\lambda = \{\{A^{q^d}\}_{d\in\{1...D-1\}}, \{\prod^{q^d}\}_{d\in\{1...D-1\}}, \{B^{q^D}\}\}$, which is acquired in the training stage.

In preparing training data, we should label each character of words with roles. For labeling '2+1' and '1+2', a two-character word dictionary is used to identify the boundary of bi-character word automatically.

In the decoding stage, the likelihood of every candidate path is calculated, and then n-best candidates are picked out. An example is shown as Figure 2.

5. Re-rank Process

In an investigation on the best output of HHMM, we find that over 90% of errors are incorrect combinations of single-character words. Therefore, re-rank is applied to n-best results of HHMM in order to remove these incorrectly combinations.

We use bi-gram to remember some existing single-character or bi-character words collocation appearing in the training data. For every new word candidate $w_{nw}=c_0 \ldots c_k$, we calculate its whole probability according to $P(w_{nw}) = \prod_{i=1}^{k} P(c_i \mid c_{i-1})$, and then remove the candidate if the whole probability is above zero. In another word, if the candidate appears in the corpus as word sequence, it means it is not a new word. Experimental result shows about 75% of errors are removed from the results with only a few losses of new words.

6. Experiment and Error Analysis

We use People's Daily corpus (1998) developed by PKU in our experiment.

The data is separated into two parts, training data and evaluation data. To make evaluation data, the first thing is to label new words, and the second thing is to split the labeled new words into characters. In labeling new word, we search the words which don't appear in the training data and take them as new words. In this way, we get evaluation data automatically. The evaluation results are shown as Table 2.

Table 2. Evaluation Results

Data		Details		Precision	Recall
Training	Eval.	#Seq.	#NW		
Jan.	Dec.	85088	8475	49.3%	70.6%
Jan.-Feb.	Dec.	81035	6094	45.5%	69.5%
Jan.-Mar.	Dec.	79299	5010	42.7%	68.6%
Feb.	Dec.	85451	8571	49.7%	69.0%
Mar.	Dec.	86265	9480	49.5%	69.0%
Jan.-Nov.	Dec.	74807	2097	33.3%	63.7%
Jan.-Oct.	Nov.	85049	2477	30.7%	61.8%
Jan.-Sep.	Oct.	78721	2154	27.2%	64.2%
Jan.	Feb.	82300	8869	49.4%	70.7%
Jan.	Feb.-Mar.	163490	15000	50.6%	71.4%

In Table 2, #Seq. is the number of single-character sequence. #NW means the number of new word.

When n-best results are all evaluated, in all the evaluation data, we get recall of about 90% (the average n is 4.), which means the coverage of character roles is high.

There are four kinds of errors. The first kind is inconsistency of the corpus, which results in 5%-10% drop of recall. In almost same contexts, some words are split into '2+1' or '1+1' somewhere in the corpus, while they are taken as words in other places, e.g. '波旁宫'(Bourbon Palace), '奔驰车'(Benz car), '测量船'(survey ship), '厂点'(address of the factory), '电火锅'(electronic chafing dish), '钢片'(steel piece), '穿过'(perforate).

The second kind of error is caused by the definition of words. E.g. in the corpus, '巴西人'(Brazilian), '澳门人'(Macao people) are treated as words, but '中国人'(Chinese), '美国人'(American) are separated into '中国 人', '美国 人'.

The third kind of error is incorrect combination of characters. E.g. '鱼鳞坑'(scale pit) is incorrectly identified as '鱼' and '鳞坑'('挖 鱼鳞坑') ,

The last kind is majority of errors, which is caused by incorrectly combining those single-character words. E.g. '当属'('最 引人注目 的 当 属 国产 ＧＳＭ'), '深有'('的 条件 深 有 体会 。').

7. Conclusion and Future Work

In this paper, we propose a method for detecting new words. The Experimental results show that the proposed approach yields 50.6% precision and 71.4% recall when the top results are evaluated, and nearly 90% recall when n-best results are all evaluated (n = 4). In practice, our approach is applied to the construction of mono-dictionary.

In the future, we will be concentrated on improving the precision of new words.

References

1. Hongqiao Li, etc. The Use of SVM for Chinese New Word Identification, IJCNLP-04, pp.723-732, 2004
2. Keh-Jiann Chen, etc. Unknown Word Detection for Chinese by a Corpus-based Learning Method, TALIP, pp.34-64, 2002
3. 刘挺，吴岩，王开铸. 串频统计和词形匹配相结合的汉语自动分词系统. 中文信息学报，第 12 卷第 1 期：p.17-25, 1998
4. 邹纲，刘洋，刘群等. 面向 Internet 的中文新词语检测. 中文信息学报，第 18 卷第 6 期：p.1-9, 2004
5. GOH Chooi Ling, etc. Chinese Unknown Word Identification Using Character-based Tagging and Chunking, ACL, pp.197-200, 2003
6. Andi Wu, etc. Statistically-Enhanced New Word Identification in a Rule-Based Chinese System. In Proceedings of the Second Chinese Language Processing Workshop, pp. 46-51, 2000
7. S. Fine, etc. The hierarchical hidden markov model: Analysis and applications. Machine Learning, 32(1):41-62, 1998