<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Chinese unknown word identification as known word tagging</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Fu, GH; Luke, KK</td>
</tr>
<tr>
<td><strong>Citation</strong></td>
<td>Proceedings Of 2004 International Conference On Machine Learning And Cybernetics, 2004, v. 4, p. 2612-2617</td>
</tr>
<tr>
<td><strong>Issued Date</strong></td>
<td>2004</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://hdl.handle.net/10722/47018">http://hdl.handle.net/10722/47018</a></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>Creative Commons: Attribution 3.0 Hong Kong License</td>
</tr>
</tbody>
</table>
CHINESE UNKNOWN WORD IDENTIFICATION AS KNOWN WORD TAGGING

GUO-HONG FU, KANG-KWONG LUKE

Department of Linguistics, The University of Hong Kong, Pokfulam Road, Hong Kong
E-MAIL: ghfu@hkucc.hku.hk, kkluke@hkuusa.hku.hk

Abstract:

This paper presents a tagging approach to Chinese unknown word identification based on lexicalized hidden Markov models (LHMMs). In this work, Chinese unknown word identification is represented as a tagging task on a sequence of known words by introducing word-formation patterns and part-of-speech. Based on the lexicalized HMMs, a statistical tagger is further developed to assign each known word an appropriate tag that indicates its pattern in forming a word and the part-of-speech of the formed word. The experimental results on the Peking University corpus indicate that the use of lexicalization technique and the introduction of part-of-speech are helpful to unknown word identification. The experiment on the SIGHAN-PK open test data also shows that our system can achieve state-of-art performance.

Keywords:

Unknown word identification; lexicalized HMMs; known word tagging; Chinese word segmentation

1. Introduction

Unknown word identification (UWI) is an important and difficult problem in Chinese word segmentation. On the one hand, most current systems for Chinese word segmentation are based on a predefined machine-readable dictionary. However, no dictionary can be complete. In general, some 8-10% of words in real text are out of the dictionaries in use. Therefore, a practical system for Chinese word segmentation must be capable of detecting these out-of-vocabulary or unknown words. On the other hand, Chinese UWI is by no means a trivial task in that Chinese unknown words are constructed dynamically and freely. In theory, any combination of Chinese characters or lexicon words may be a potential unknown word. However, there lack of enough explicit marks in plain Chinese texts, such as capitalization in English that can be used directly to identify unknown words. Consequently, the exploration of more potential features is usually an effective way to improve the systems for Chinese UWI.

In the past years, a variety of techniques have been proposed to address the problem of Chinese UWI. Each technique has its own deficiencies while offering its advantages. Wu and Jiang took word segmentation and UWI as an integral part of full sentence analysis [1]. This method is proved to be powerful in segmentation disambiguation and UWI. However, the coverage of the parser may restrict its applications in practical NLP systems. Zhang et al presented a novel method to Chinese UWI based on role tagging [2]. They defined a set of unknown word roles about varied internal components and contexts. As a result, their system can detect different types of unknown words in real text. However, an additional role-tagged corpus is needed to learn role knowledge, which is not always available in practice. Xue recently reported a supervised machine-learning approach to Chinese word segmentation [3]. In his work, Chinese word segmentation is re-formulated as a problem of tagging Chinese character with position-of-character (POC) tags. This approach does not need a dictionary at all, so it is effective in principle for UWI. However, this method is purely based on character tagging, which may lose the important word-level features for correct disambiguation and UWI. More recently, Fu and Luke proposed a modified class-based LM approach to Chinese UWI [4]. In their work, Chinese UWI is viewed as a classification problem, and a number of different features, including contextual class feature, word juncture model and word formation patterns, are combined in a class-based LM framework to identify different unknown words. However, it is still an open problem to normalize different probabilistic distributions of different dimensions in an optimal way.

In this paper, we propose a lexicalized hidden Markov model (LHMM) approach to Chinese UWI. In this work, Chinese UWI is represented as a tagging task on a sequence of known words by introducing word-formation patterns. To do this, a tagger is thus developed based on the lexicalized HMMs to assign each known word in input an appropriate tag that indicates its patterns in forming a word and the part-of-speech of this formed word. In comparison with standard HMMs, the lexicalized HMMs can handle
richer contextual information, both contextual words and
tags for correct tagging of known words. In addition,
part-of-speech tags are also introduced and incorporated
with the word-formation pattern tags. In this way, most
Chinese unknown words can be resolved effectively.

The rest of this paper is organized as follows: Section
2 discusses how Chinese UWI can be reformulated
as known word tagging. Section 3 presents the lexicalized
HMMs for unknown word tagging. In section 4, the tagging
algorithm is given in brief. Finally, the experimental results
and some conclusions on this work will be given
respectively in section 5 and section 6.

2. Chinese UWI as known word tagging

In this section, Chinese UWI is represented as known
word tagging by introducing word-formation pattern and
part-of-speech tags.

2.1. Representing segmented words with pattern-tags

In practice, known words and unknown words in a
sentence can be represented by means of word-formation
pattern tags. As discussed in [4], a lexicon word $w$ has
four possible word-formation patterns to present itself after
UWI: (1) $w$ is an independent segmented known word by
itself; (2) $w$ is at the beginning of an unknown word. (3)
$w$ is at the middle of an unknown word. (4) $w$ is at the
end of an unknown word. For convenience, these patterns
are denoted respectively by four tags, i.e. ISW, BOW, MOW
and EOW.

Obviously, an unknown word will be resolved once
the relevant word-formation patterns of its components are
determined. At this point, Chinese UWI is equivalent to a
process of assigning of word-formation pattern tags on a
sequence of known words. More formally, a lexicon word
may be tagged with four possible tags shown above in
terms of its patterns during UWI: It will be tagged as ISW if
it is recognized as an independent known word during UWI;
On the contrary, it will be tagged as BOW, MOW or EOW
respectively if it present itself at the beginning, middle or
end of an unknown word after UWI.

For example, the segmented sentence “中国/国家/主
席/胡/锦涛/朝/北朝鲜/领导人/正” (Chinese President Hu Jintao held talks with North Korean
leader Kim Jong-Il) can be represented using the pattern
tags as follows:

$<$ISW$>$中国/$<$ISW$>$国家/$<$ISW$>$主席
$<$ISW$>$胡/$<$ISW$>$锦/$<$ISW$>$涛
$<$EOW$>$正

Differing from Xue's formulation [3], our formulation
is based on known word tagging, which has two main
advantages: Firstly, the word-based formulation is more
general in that any unknown word must be made up of a
number of known words, including single-character or
multi-character known words. The second advantage of the
formulation based on known word tagging is that it allows
the use of more important word-level information such as
contextual words and tags for ambiguity resolution and
UWI.

2.2. Incorporating POS-tag with pattern-tag for UWI

It has been proved that part-of-speech is another
important information for correct UWI [1][4],
part-of-speech tags are accordingly introduced in this work.
For convenience, we merge part-of-speech tags and the
pattern tags by using following format: T1-T2. Where T1
denotes a part-of-speech tag and T2 denotes a
word-formation pattern tag. Note that the Peking University
part-of-speech tag-set is used in our system, which contains
48 different tags. With this combined tag-set, the previous
e Xue can be further represented as follows:

$<$ISW$>$中国/$<$ISW$>$国家/$<$ISW$>$主席/$<$ISW$>$胡/$<$ISW$>$锦/$<$ISW$>$涛/$<$ISW$>$正

3. Lexicalized HMMs for Chinese UWI

The lexicalized HMM approach has been widely used
in POS tagging [5], shallow parsing [6] and Chinese
prosodic phrase prediction [7]. In this section, we continue
to apply it to perform the known tagging for Chinese UWI.

3.1. Lexicalized HMMs

From the statistical point of view, the task of known
word tagging for Chinese UWI can be defined as the
process of finding an appropriate tag sequence $\widehat{T} = t_1 t_2 \ldots t_n$
that maximizes the conditional probability $P(T|W)$, given
a sequence of known words $W = w_1 w_2 \ldots w_n$, namely,

$$\widehat{T} = \arg \max_{T} P(T|W) = \arg \max_{T} \frac{P(W|T)P(T)}{P(W)} \quad (1)$$
For a specific sequence of known words \( w \), the probability \( p(w) \) is fixed. Therefore, it can be dropped from the above equation. Thus, we have a general statistical model for known word tagging as follows:

\[
\hat{T} = \arg \max_T P(W \mid T) P(T)
\]

\[
= \arg \max_T P(w_{1:n}, t_{1:n})
\]

\[
= \arg \max_T \prod_{i=1}^{n} P(w_i \mid w_{1:i-1}, t_i) P(t_i \mid w_{1:i-1}, t_{i-1})
\]

(2)

However, this general model is not computable in practice because it has too many parameters. To address this problem, two types of approximations are employed here to make it applicable.

The first approximation is based on the independent hypothesis in standard HMMs: The appearance of current word \( w_i \) depends only on current tag \( t_i \) during known word tagging, and the assignment of current tag \( t_i \) depends only on its previous tag \( t_{i-1} \). Thus,

\[
\hat{T} = \arg \max_T \prod_{i=1}^{n} P(w_i \mid t_i) P(t_i \mid t_{i-1})
\]

(3)

Equation (3) actually presents a first-order HMMs for known word tagging. Where, \( P(w_i \mid t_i) \) is the so-called lexical probability; and \( P(t_i \mid t_{i-1}) \) denotes the contextual tag probability.

The second type of approximation follows the notion of the lexicalized HMMs. In this approximation, the appearance of current word \( w_i \) is assumed to depend not only on current tag \( t_i \) but also its previous word \( w_{i-1} \), and the assignment of current tag \( t_i \) is supposed to depend both its previous word \( w_{i-1} \) and its previous tag \( t_{i-1} \). Thus, we have the lexicalized HMMs for UWI as follows:

\[
\hat{T} = \arg \max_T \prod_{i=1}^{n} P(w_i \mid w_{i-1}, t_i) P(t_i \mid w_{i-1}, t_{i-1})
\]

(4)

In comparison with the standard HMMs, the lexicalized HMMs can provide richer contextual information for the assigning of tags to known words, including both contextual words and contextual tags, which will result in improvement of accuracy in UWI.

### 3.2. Parameter estimation and data smoothing

Parameter estimation and data smoothing are two main steps in our system. In our implementation, we employ the linear interpolation smoothing technique to avoid this problem of data sparseness. As shown in equation (7), higher-order parameters in HMMs are smoothed with the relevant lower-order probabilities.

\[
P'(w_i \mid t_i) = \lambda P(w_i \mid t_i) + \frac{1-\lambda}{\text{Count}(t_i)}
\]

\[
P'(t_i \mid t_{i-1}) = \mu P(t_i \mid t_{i-1}) + (1-\mu) P(t_i)
\]

(7)

4. The tagging algorithm

Based on the models in equation (3) or (4), the tagging algorithm aims to score all possible candidate sequences of tags and find the best one that has the maximum score. In our system, this task is done by the classical Viterbi algorithm, which consists of two main steps: (1) The generation of candidate tags: The first step generates all possible candidate tags for each known word in the input by looking up the system dictionary or the library of lexical probabilities. All these candidate tags are stored in a lattice structure. (2) The decoding of best tags: In this step, the Viterbi algorithm scores all candidate tags with HMMs or LHMMs, and then searches the best path through the lattice built in the first step that maximizes the score. This path contains the best sequence of tags for the input sequence of known word sequence.

With this tagging algorithm, we develop a complete Chinese word segmenter using the two-stage strategy [4].
This system works in three main phrases, namely known word segmentation, tagging, and the conversion of known word tagged result to a sequence of segmented words. In order to yield correct segmentations for some complicated cases such as a mixture of ambiguities and unknown words in real texts, a pure known-word based n-gram is applied here to perform known word segmentation.

Similar to the work in [3], inconsistent tagging may occur in our system. In practice, there are two types of inconsistent tagging in this work, namely the pattern inconsistency and the POS inconsistency. Pattern inconsistency arises when two adjacent known words are assigned inconsistent pattern tags such as “ISW:MOW” or “ISW:EOW”. The part-of-speech inconsistency means that two adjacent known words are tagged with different part-of-speech while at the same time, they are assigned the pattern tags indicating they should occur in one unknown word. For example, the tag pair “a-BOW:n-EOW” is inconsistent in part-of-speech tagging. Since it has been proved that the inconsistent tagging hardly exerts any influence on the final results [3], we leave the inconsistent tagging as it is in our implementation. In fact, few inconsistent tags can occur in the final result because they usually have lower probabilities, and will be mostly blocked by the decoder.

5. Experiments

In evaluating our approach, we conduct two experiments respectively on the Peking University corpus (January 1998 of the People’s Daily) [8] and the PK-open test corpus for the First International Word Segmentation Bakeoff [9]. This section reports the relevant results of these experiments.

5.1. Experimental data and evaluation measures

In our experiments, we use the same corpora as used in [4], which come from two resources: The first one is from the Peking University corpus, which contains one month (January 1998) of news texts from the People’s Daily, and has been manually segmented and tagged with part-of-speech by Peking University [8]. As shown in Table 1, this corpus is separated into two parts: The larger part (viz. the Corpus A) is used to train our system, and the smaller part (viz. the Corpus B) is used for the closed-test. Furthermore, Corpus A is automatically labeled with word-formation pattern tags by using the forward maximum matching technique. The second source (viz. the Corpus C) is from SIGHAN bakeoff data, which is first used for the PK-open test at the First International Chinese Word Segmentation Bakeoff sponsored by SIGHAN [9], and is used here for the open comparison test.

<table>
<thead>
<tr>
<th>Corpora</th>
<th># words</th>
<th>#OOV words</th>
<th>OOV rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus A</td>
<td>998,085</td>
<td>68,638</td>
<td>6.88</td>
</tr>
<tr>
<td>Corpus B</td>
<td>112,373</td>
<td>7,444</td>
<td>6.62</td>
</tr>
<tr>
<td>Corpus C</td>
<td>17,605</td>
<td>1,619</td>
<td>9.20</td>
</tr>
</tbody>
</table>

In addition to the above corpora, we also use a lexicon in our system, which is mainly built from the Peking University dictionary. In order to process the non-standard Chinese words in real texts, a number of non-Hanzi characters are also added in it. Consequently, the final dictionary contains about 65,270 different word-forms in all. Furthermore, all possible part-of-speech candidates of a word-form are also defined in it. Based on this lexicon, the relevant out-of-vocabulary rates (OOV rate for short) of the three corpora in Table 1 are 6.88%, 6.62% and 9.20% respectively.

In evaluating the effectiveness of our system, three measures are computed in our experiments, including recall (R), precision (P) and F-score (F). Here, recall (R) is defined as the number of correctly segmented words divided by the total number of words in the manually annotated corpus, and precision (P) is defined as the number of correctly segmented words divided by the total numbers of words segmented automatically by the system. As for F-score (denoted by F), it is the weighted harmonic mean of precision and recall that is formulated as follows:

\[ F = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 \times R + P} \]  

(9)

Here, we employ the balanced F-score (viz. \( \beta^2 = 1 \)) to evaluate the overall performance of our system in word segmentation and UWI in that it is still not clear whether recall or precision is more important in evaluating a word segmentation system.

5.2. Experimental results and discussions

As mentioned above, the lexicalization technique and part-of-speech tags are introduced into the proposed approaches. The first experiment is therefore conducted to test how the introduction of the lexicalization technique or part-of-speech tags improves the performance of our system in word segmentation and UWI. The results are presented in Table 2. Each row in this table contains three lines of numbers, which denote the accuracy of the relevant approach respectively in word segmentation, known word segmentation and UWI.
Table 2. Experimental results on PKU corpus

<table>
<thead>
<tr>
<th>Methods</th>
<th>Recall(%)</th>
<th>Precision(%)</th>
<th>F-score(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMMs</td>
<td>94.65</td>
<td>93.12</td>
<td>93.88</td>
</tr>
<tr>
<td>without</td>
<td>97.57</td>
<td>94.13</td>
<td>95.82</td>
</tr>
<tr>
<td>POS</td>
<td>54.34</td>
<td>73.63</td>
<td>62.53</td>
</tr>
<tr>
<td>HMMs</td>
<td>96.07</td>
<td>95.24</td>
<td>95.66</td>
</tr>
<tr>
<td>With</td>
<td>97.83</td>
<td>96.06</td>
<td>96.94</td>
</tr>
<tr>
<td>POS</td>
<td>71.83</td>
<td>82.11</td>
<td>76.63</td>
</tr>
<tr>
<td>LHMMs</td>
<td>97.06</td>
<td>96.72</td>
<td>96.89</td>
</tr>
<tr>
<td>without</td>
<td>98.01</td>
<td>97.42</td>
<td>97.72</td>
</tr>
<tr>
<td>POS</td>
<td>84.05</td>
<td>86.54</td>
<td>85.28</td>
</tr>
<tr>
<td>LHMMs</td>
<td>97.32</td>
<td>96.91</td>
<td>97.12</td>
</tr>
<tr>
<td>With</td>
<td>98.13</td>
<td>97.46</td>
<td>97.79</td>
</tr>
<tr>
<td>POS</td>
<td>86.20</td>
<td>89.03</td>
<td>87.59</td>
</tr>
</tbody>
</table>

The data in Table 2 reveals two main findings. Firstly, the lexicalized HMMs perform better than the non-lexicalized HMMs. As can be seen in Table 2, the lexicalized HMMs improve the F-measure in UWI by 10.96 percent for the tag-set with part-of-speech and 23.75 percent for the tag-set without part-of-speech. Furthermore, the improvements of accuracy in UWI will contribute further 1.46 or 3.01 percent to the relevant overall F-score in word segmentation. Secondly, the introduction of part-of-speech tags is helpful to improve unknown word identification. It is shown in Table 2 that the introduction of part-of-speech leads to improvement of F-score in UWI by about 13.1 percent for HMMs and 2.31 percent for the lexicalized HMMs. As for the overall F-score in word segmentation, the improved number is 1.78 percent for HMMs and 0.23 percent for LHMMs.

In addition to the above experiment, we also conduct an open evaluation using the test-corpus for the track of PK-open in 2003 SIGHAN Bakeoff and compare our system with other public systems in the track. We have two reasons for the selection of this corpus: The first reason is that both the training data of this work and the SIGHAN-PK open test data are from Peking University. The second one is that some other resources such as the part-of-speech lexicon are used in our work, which does not satisfy the requirements for the closed tests at SIGHAN bakeoff. The results are presented in Table 3 and Table 4.

Actually, Table 3 presents the results of the open test of the proposed methods on SIGHAN bakeoff data, which is parallel to Table 2. Comparing the data in the two tables, we find that they show the similar trends for different methods under discussion. We also notice that our system yields worse results in the open test. Our further error analysis shows that the drop of performance is caused by three main factors, namely the inconsistent segmentations between the training data and the open test data, the problem of data sparseness and the complicated cases that are beyond current methods.

Table 3. Experimental results on the corpus for SIGHAN-PK open test

<table>
<thead>
<tr>
<th>Methods</th>
<th>Recall(%)</th>
<th>Precision(%)</th>
<th>F-score(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMMs</td>
<td>93.31</td>
<td>91.09</td>
<td>92.19</td>
</tr>
<tr>
<td>without</td>
<td>97.20</td>
<td>92.04</td>
<td>94.55</td>
</tr>
<tr>
<td>POS</td>
<td>55.96</td>
<td>77.64</td>
<td>65.04</td>
</tr>
<tr>
<td>HMMs</td>
<td>93.73</td>
<td>91.90</td>
<td>92.80</td>
</tr>
<tr>
<td>With</td>
<td>97.40</td>
<td>93.14</td>
<td>95.23</td>
</tr>
<tr>
<td>POS</td>
<td>58.43</td>
<td>75.68</td>
<td>65.95</td>
</tr>
<tr>
<td>LHMMs</td>
<td>96.07</td>
<td>93.86</td>
<td>94.40</td>
</tr>
<tr>
<td>without</td>
<td>96.92</td>
<td>94.91</td>
<td>95.91</td>
</tr>
<tr>
<td>POS</td>
<td>76.10</td>
<td>82.63</td>
<td>79.23</td>
</tr>
<tr>
<td>LHMMs</td>
<td>96.19</td>
<td>94.09</td>
<td>96.44</td>
</tr>
<tr>
<td>With</td>
<td>97.02</td>
<td>95.02</td>
<td>96.01</td>
</tr>
<tr>
<td>POS</td>
<td>77.58</td>
<td>84.24</td>
<td>80.77</td>
</tr>
</tbody>
</table>

Table 4 presents the result of the comparison of our system with other open systems for the SIGHAN-PK open test. As shown in Table 5, our system ranks the third in terms of the overall F-score in word-segmentation, which indicates in a sense that the proposed approach can yield results that are comparable to other state-of-the-arts approaches. Here, the system S10 is developed based on the full sentence parsing technique and its results usually depend on a complicated fine tuning [10]; The system S01 is based on role-tagging technique, which need an additional role-tagged corpus for training [2]. In comparison with the two systems, our system is purely based on known word tagging and the lexicalized HMMs, and can be built efficiently on a manually segmented and part-of-speech corpus. This kind of corpus is now available for Chinese such as the PKU corpus and the CKIP corpus. At this point, our approach is more applicable.

Table 4. The comparison with the systems for the SIGHAN-PK open test

<table>
<thead>
<tr>
<th>Systems</th>
<th>ROOV(%)</th>
<th>Rv(%)</th>
<th>R(%)</th>
<th>P (%)</th>
<th>F (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S10</td>
<td>79.9</td>
<td>97.5</td>
<td>96.3</td>
<td>95.6</td>
<td>95.9</td>
</tr>
<tr>
<td>S01</td>
<td>74.3</td>
<td>98.0</td>
<td>96.3</td>
<td>94.3</td>
<td>95.3</td>
</tr>
<tr>
<td>S08</td>
<td>67.5</td>
<td>95.9</td>
<td>93.9</td>
<td>93.8</td>
<td>93.8</td>
</tr>
<tr>
<td>S04</td>
<td>71.2</td>
<td>94.9</td>
<td>93.3</td>
<td>94.2</td>
<td>93.7</td>
</tr>
<tr>
<td>S03</td>
<td>64.7</td>
<td>96.2</td>
<td>94.0</td>
<td>91.1</td>
<td>92.5</td>
</tr>
<tr>
<td>S11</td>
<td>50.3</td>
<td>93.4</td>
<td>90.5</td>
<td>86.9</td>
<td>88.6</td>
</tr>
<tr>
<td>Our system</td>
<td>77.58</td>
<td>97.02</td>
<td>95.19</td>
<td>94.09</td>
<td>94.64</td>
</tr>
</tbody>
</table>
6. Conclusions

In this paper, we have presented a lexicalized hidden Markov model approach to Chinese UWJ. In this work, Chinese UWJ is represented as a tagging task on a sequence of known words by introducing word-formation patterns. To do this work, a LHMM tagger is further developed on a manually part-of-speech tagged corpus to assign each known word in input an appropriate tag that indicates its patterns in forming a word and the part-of-speech of this formed word. In comparison with standard HMMs, the lexicalized HMMs can handle richer contextual information, both contextual words and tags for correct tagging of known words. In addition, part-of-speech tags are also introduced and further incorporated with the word-formation pattern tags. In this way, most Chinese unknown words can be resolved effectively. The experimental results on Peking University corpus indicate that the use of lexicalization technique and the introduction of POS are helpful to unknown word identification. The experiment on SIGHAN-PK open test data also shows that our system can achieve state-of-art performance. In practice, the proposed approach also provides a framework for part-of-speech tagging, in particular for unknown word tagging. In future work, we hope to apply it in Chinese part-of-speech tagging and other NLP applications such as named entity recognition.

Acknowledgements

We would like to thank the Institute of Computational Linguistics, Peking University for providing part-of-speech tagset, lexicon and corpus for our experiments.

References


