Automatic Chinese Abbreviation Generation Using Conditional Random Field

Dong Yang, Yi-cheng Pan, and Sadaoki Furui
Department of Computer Science
Tokyo Institute of Technology
Tokyo 152-8552 Japan
{raymond,thomas,furui}@furui.cs.titech.ac.jp

Abstract

This paper presents a new method for automatically generating abbreviations for Chinese organization names. Abbreviations are commonly used in spoken Chinese, especially for organization names. The generation of Chinese abbreviation is much more complex than English abbreviations, most of which are acronyms and truncations. The abbreviation generation process is formulated as a character tagging problem and the conditional random field (CRF) is used as the tagging model. A carefully selected group of features is used in the CRF model. After generating a list of abbreviation candidates using the CRF, a length model is incorporated to re-rank the candidates. Finally the full-name and abbreviation co-occurrence information from a web search engine is utilized to further improve the performance. We achieved top-10 coverage of 88.3% by the proposed method.

1 Introduction

Long named entities are frequently abbreviated in oral Chinese language for efficiency and simplicity. Therefore, abbreviation modeling is an important building component for many systems that accept spoken input, such as directory assistance and voice search systems.

While English abbreviations are usually formed as acronyms, Chinese abbreviations are much more complex, as shown in Figure 1. Most of the Chinese abbreviations are formed by selecting several characters from full-names, which are not necessarily the first character of each word. Usually the original character order in the full-name is preserved in the abbreviation. However, re-ordering of characters as shown in the third example in Figure 1 where characters "三" and "医" are swapped in the abbreviation, also happens.

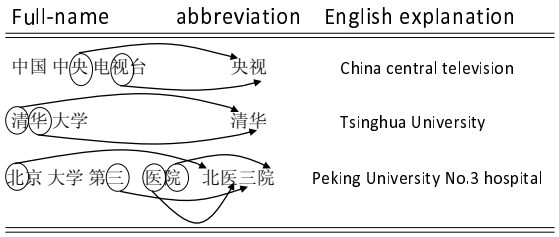

Figure 1: Chinese abbreviation examples

There has been a considerable amount of research on extracting full-name and abbreviation pairs in the same document for obtaining abbreviations (Li and Yarowsky, 2008; Sun et al., 2006; Fu et al., 2006). However, generation of abbreviations given a full-name is still a non-trivial problem. Chang and Lai (Chang and Lai, 2004) have proposed using a hidden Markov model to generate abbreviations from full-names. However, their method assumes that there is no word-to-null mapping, which means that every word in the full-name has to contribute at least one character to the abbreviation. This assumption does not hold for organizations' names which have many word skips in the abbreviation generation.

The CRF was first introduced to natural language processing (NLP) by (Lafferty et al., 2001) and has been widely used in word segmentation, part-of-speech (POS) tagging, and some other NLP tasks. In this paper, we convert the Chinese abbreviation generation process to a CRF tagging problem. The key problem here is how to find a group of discrim-

Proceedings of NAACL HLT 2009: Short Papers, pages 273–276,
Boulder, Colorado, June 2009.

inant and robust features. After using the CRF, we get a list of abbreviation candidates with associate probability scores. We also use the prior conditional probability of the length of the abbreviations given the length of the full-names to complement the CRF probability scores. Such global information is hard to include in the CRF model. In addition, we apply the full-name and abbreviation candidate co-occurrence statistics obtained on the web to increase the correctness of the abbreviation candidates.

2 Chinese Abbreviation Introduction

Chinese abbreviations are generated by three methods (Lee, 2005): reduction, elimination, and generalization.

Both in the reduction and elimination methods, characters are selected from the full-name, and the order of the characters is sometime changed. Note that this paper does not cover the case when the order is changed. The elimination means that one or more words in the full-name are ignored completely, while the reduction requires that at least one character is selected from each word. All the three examples in Figure 1 are produced by the elimination, where at least one word is skipped.

Generalization, which is used to abbreviate a list of similar terms, is usually composed of the number of terms and a shared character across the terms. A example is "三军" (three forces) for "陆军，海军，空军" (land force, sea force, air force). This is the most difficult scenario for the abbreviations and is not considered in this paper.

3 CRF Model for Abbreviation Modeling

3.1 CRF model

A CRF is an undirected graphical model and assigns the following probability to a label sequence $L = l_1 l_2 \dots l_T$, given an input sequence $C = c_1 c_2 \dots c_T$,

$$P(L|C) = \frac{1}{Z(C)} exp(\sum_{t=1}^{T} \sum_{k} \lambda_k f_k(l_t, l_{t-1}, C, t)) \quad (1)$$

Here, f_k is the feature function for the *k*-th feature, λ_k is the parameter which controls the weight of the *k*-th feature in the model, and *Z*(*C*) is the normalization term that makes the summation of the probability of all label sequences to 1. CRF training is usually performed through the typical L-BFGS algorithm (Wallach, 2002) and decoding is performed by Viterbi algorithm (Viterbi, 1967). In this paper, we use an open source toolkit "crf++".

3.2 Abbreviation modeling as a tagging problem

In order to use the CRF method in abbreviation generation, the abbreviation generation problem was converted to a tagging problem. The character is used as a tagging unit and each character in a full-name is tagged by a binary variable with the values of either Y or N: Y stands for a character used in the abbreviation and N means not. An example is given in Figure 2.

Figure 2: Abbreviation in the CRF tagging format

3.3 Feature selection for the CRF

In the CRF method, feature function describes a co-occurrence relation, and it is defined as $f_k(l_t, l_{t-1}, C, t)$ (Eq. 1). f_k is usually a binary function, and takes the value 1 when both observation c_t and transition $l_{t-1} \rightarrow l_t$ are observed. In our abbreviation generation model, we use the following features:

1. Current character The character itself is the most important feature for abbreviation as it will be either retained or discarded. For example, "局" (bureau) and "所" (institue), indicating a government department, are very common characters used in abbreviations. When they appear in full-names, they are likely to be kept in abbreviations.

2. Current word In the full name of "中国农业大学" (China Agricultural university), the word "中国" (China) is usually ignored in the abbreviation, but the word "农业" (agriculture) is usually kept. The length (the number of characters) is also an important feature of the current word.

3. Position of the current character in the current word Previous work (Chang and Lai, 2004) showed that the first character of a word has high possibility to form part of the abbreviation and this is also true for the last character of a three-character word.

4. Combination of feature 2. and 3. above Combination of the features 2 and 3 is expected to improve the performance, since the position infor-

mation affects the abbreviation along with the current word. For example, ending character in "大学" (university) and that in "研究院" (research institute) have very different possibilities to be selected for abbreviations.

Besides the features above, we have examined context information (previous word, previous character, next character, etc.) and other local features like the length of the word, but these features did not improve the performance. The reason may be due to the sparseness of the training data.

4 Improvement by a Length Model and a Web Search Engine

4.1 Length model

There is a strong correlation between the length of organizations' full-names and their abbreviations. We use the length modeling based on discrete probability of $P(M|L)$, in which the variables *M* and *L* are lengths of abbreviations and full-names, respectively. Since it is difficult to incorporate length information into the CRF model explicitly, we use $P(M|L)$ to rescore the output of the CRF.

In order to use the length information, we model the abbreviation process with two steps:

- 1st step: evaluate the length in abbreviation according to the length model $P(M|L)$;
- 2nd step: choose the abbreviation, given the length and full-name.

We assume the following approximation:

$$P(A|F) \simeq P(M|L) \cdot P(A|M,F) \qquad (2)$$

in which variable *A* is the abbreviation and *F* is the full-name; $P(M|L)$ is the length model, and the second probability can be calculated according to the Bayesian rule:

$$P(A|M,F) = \frac{P(A,M|F)}{P(M|F)} = \frac{P(A,M|F)}{\sum_{length(A')=M} P(A',M|F)} \qquad (3)$$

It is obvious that $P(A,M|F) = P(A|F)$ (as *A* contains the information *M* implicitly) and $P(A|F)$ can be obtained from the output of the CRF.

4.2 Web search engine

Co-occurrence of a full-name and an abbreviation candidate can be a clue of the correctness of the abbreviation. We use the "abbreviation candidate"+ "full-name" as queries and input them to the most popular Chinese search engine (www.baidu.com), and then we use the number of hits as the metric to perform re-ranking. The hits is theoretically related to the number of pages which contain both the full-name and abbreviation. The bigger the value of hits, the higher probability that the abbreviation is correct.

We then simply multiply the previous probability score, obtained from Eq. 2, by the number of hits and re-rank the top-30 candidates accordingly.

There are some other ways to use information retrieval methods (Mandala et al., 2000). Our method has an advantage that the access load to the web search engine is relatively small.

5 Experiment

5.1 Data introduction

The corpus we use in this paper comes from two sources: one is the book "modern Chinese abbreviation dictionary" (Yuan and Ruan, 2002) and the other is the wikipedia. Altogether we collected 1945 pairs of organization full-names and their abbreviations.

The data is randomly divided into two parts, a training set with 1298 pairs and a test set with 647 pairs. Table 1 shows the length mapping statistics of the training set. It can be seen that the average length of full-names is about 7.29. We know that for a full-name with length *N*, the number of abbreviation candidates is about $2^N - 2 - N$ (exclude length of 0, 1, and *N*) and we can conclude that the average number of candidates for organization names in this corpus is more than 100.

5.2 Results

The abbreviation method described is part of a project to develop a voice-based search application. For our name abbreviation system we plan to add 10 abbreviation candidates for each organization name into the vocabulary of our voice search application, hence here we consider top-10 coverage.

length of full-name	length of abbreviation					sum
	2	3	4	5	>5	
4	107	1	0	0	0	108
5	89	140	0	0	0	229
6	96	45	46	0	0	187
7	60	189	49	16	0	314
8	48	29	60	3	6	146
9	10	47	35	12	2	106
10	18	11	29	8	6	73
others	21	43	38	17	14	133
average length of the full-name						**7.27**
average length of the abbreviation						**3.01**

Table 1: Length statistics on the training set

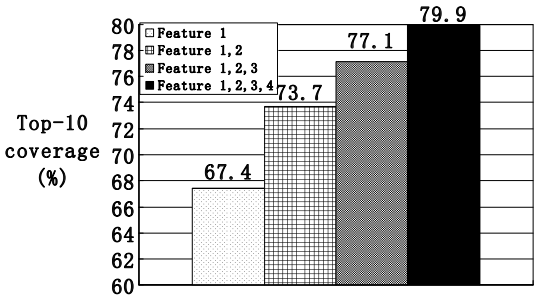

Figure 3: Contribution of features in CRF

Figure 3 shows the result for various combinations of features introduced in Section 3.3.

Figure 4 displays the coverage results obtained using the CRF method and the improvements gained from the inclusion of the length feature and the web search hits. As we can see the CRF gives a coverage 79.9%. Both length model and web search engine show significant improvement over the CRF baseline and the coverage increases to 88.3%.

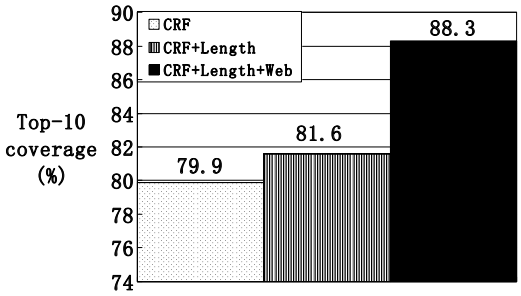

Figure 4: Results of different methods

6 Conclusions and Future work

The CRF works well in generating abbreviations for organization names, while both length model and web search engine further improve the performance.

We are going to perform word clustering or character clustering to alleviate the data sparseness problem. Also we notice that multiple abbreviations for single full-name is very common, such as "中国中央电视台" (China central television) with abbreviations "央视" and "中央台". We plan to collect multiple abbreviations for reference. After that we are going to combine the abbreviation modeling in the voice search system to alleviate the weakness of speech recognition for unknown abbreviation words, which are unlikely to be correctly recognized due to the out of vocabulary problem.

References

Jing-shin Chang and Yu-Tso Lai 2004. *A Preliminary Study on Probabilistic Models for Chinese Abbreviations*. Proceedings of ACL SIGHAN Workshop 2004, pages 9-16.

Guohong Fu, Kang-Kwong Luke, GuoDong Zhou and Ruifeng Xu 2006. *Automatic Expansion of Abbreviations in Chinese News Text*. Lecture Notes in Computer Science, Washington, DC.

John Lafferty, Andrew McCallum, and Fernando Pereira 2001. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.*, In Proceedings of International Conference on Machine Learning 2001, pages 282-289

Hiu Wing Doris Lee 2005. *A Study of Automatic Expansion of Chinese Abbreviations*. MA Thesis, The University of Hong Kong.

Zhifei Li and David Yarowsky. 2008. *Unsupervised Translation Induction for Chinese Abbreviations using Monolingual Corpora*. Proceedings of ACL 2008, pages 425-433.

Rila Mandala, Takenobu Tokunaga, and Hozumi Tanaka 2000. *Query expansion using heterogeneous thesauri.*, In Information Processing and Management Volume 36 , Issue 3 2000,Pages 361 - 378

Xu Sun, Houfeng Wang and Yu Zhang 2006. *Chinese Abbreviation-Definition Identification: A SVM Approach Using Context Information*. Lecture Notes in Computer Science,Volume 4182/2006, pages 530-536.

Andrew J. Viterbi 1967. *Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm*. in IEEE Transactions on Information Theory, Volume IT-13, in April, 1967,pages 260-269,

Hanna Wallach 2002. *Efficient Training of Conditional Random Fields*. M. Thesis, University of Edinburgh, 2002.

Hui Yuan and Xianzhong Ruan 2002. *Modern Chinese abbreviation dictionary*. Yuwen press, Beijing, China.