Generalized Abbreviation Prediction with Negative Full Forms and Its Application on Improving Chinese Web Search

Xu Sun[†], **Wenjie Li**[‡], **Fanqi Meng**[‡], **Houfeng Wang**[†],
[†]Key Laboratory of Computational Linguistics (Peking University), Ministry of Education, China
[‡]Department of Computing, The Hong Kong Polytechnic University
xusun@pku.edu.cn cswjli@comp.polyu.edu.hk mengfanqi928@163.com wanghf@pku.edu.cn

Abstract

In Chinese abbreviation prediction, prior studies are limited on positive full forms. This lab assumption is problematic in real-world applications, which have a large portion of negative full forms (NFFs). We propose solutions to solve this problem of generalized abbreviation prediction. Experiments show that the proposed unified method outperforms baselines, with the full-match accuracy of 79.4%. Moreover, we apply generalized abbreviation prediction for improving web search quality. Experimental results on web search demonstrate that our method can significantly improve the search results, with the search F-score increasing from 35.9% to 64.9%. To our knowledge, this is the first study on generalized abbreviation prediction and its application on web search.

1 Introduction

Abbreviations increase the ambiguity in a text. Associating abbreviations with their fully expanded forms is important in various natural language processing applications (Pakhomov, 2002; Yu et al., 2006; HaCohen-Kerner et al., 2008). Chinese abbreviations represent fully expanded forms (e.g., the left side of Figure 1) through the use of shortened forms (e.g., the right side of Figure 1). Chinese abbreviations are derived via a generative lexical process. Although native speakers may possess intuitions of the generative process, it cannot be adequately explained by any linguistic theory (Chang and Lai, 2004; Chang and Teng, 2007).

Abbreviation prediction (i.e., predicting abbreviation of a given full form) is important in Chinese natural language processing applications. For example, it is helpful for information retrieval if we can estimate the abbreviation of a query. For

Figure 1: An example of abbreviation prediction.

the data of one month's People's Daily, only 17% of the documents contain the full form in Figure 1, while more than 70% of the articles contain only the abbreviation in Figure 1. It is expected that abbreviation prediction can improve the recall in information retrieval. In addition, (Yang et al., 2009b) in speech studies showed that Chinese abbreviation prediction can improve voice-based search quality.

The study of Chinese abbreviation prediction is still in an early stage. Chinese abbreviation prediction is quite different from English ones, because of its specific characteristics (Sun et al., 2008; Huang et al., 1994; Chang and Teng, 2007; Yang et al., 2009b; Yang et al., 2009a). For example, Chinese abbreviations are not necessarily from the initials of words. They frequently take non-initial characters from the words in the full form. In addition, the Chinese full form does not have word boundaries.

To our knowledge, all of the prior studies of Chinese abbreviation prediction (Sun et al., 2008; Sun et al., 2009; Sun et al., 2013; Huang et al., 1994; Chang and Teng, 2007; Yang et al., 2009b; Yang et al., 2009a) have focused on positive full forms with valid abbreviations. This implicit lab assumption is quite limited in real-world applications, because real-world Chinese full forms contain a large portion of negative full forms (NFFs), which have no abbreviation at all. Abbreviation prediction becomes more difficult by considering NFFs, because of the strong noise. This difficulty is one of the reason of the lab setting on considering only positive full forms. Another reason is

International Joint Conference on Natural Language Processing, pages 641–647,
Nagoya, Japan, 14-18 October 2013.

probably the difficulty of data collection. To our knowledge, there is no existing collection of abbreviation prediction data with NFFs.

We aim at solving this abbreviation prediction problem with generalized assumption (hereinafter *generalized abbreviation prediction*). We manually collected a large dataset for this study, which contains 10,786 entries including NFFs. To deal with the strong noise from NFFs, we propose a variety of solutions. We also apply generalized abbreviation prediction for web search and we show that it can significantly improve web search quality.

2 Proposed Method

2.1 Preprocessing

The preprocessing step includes word segmentation and part-of-speech tagging for the input abbreviation prediction full forms. The word segmentation and part-of-speech tagging is done via the tool ICTCLAS [1].

2.2 Abbreviation Prediction

2.2.1 Simple Heuristic System

The simple heuristic system means always choosing initial characters of words in the segmented full form. This is because the most natural abbreviating heuristic is to produce the first character of each word in the original full form. This is just the simplest baseline.

2.2.2 Unified System

We present a unified system for generalized abbreviation prediction with NFFs. The unified system can conduct the abbreviation prediction with a single step. We cast abbreviation prediction as a sequential labeling task. Following (Sun et al., 2013), each character in the full form is tagged with a label, $y \in \{\texttt{P}, \texttt{S}\}$, where the label *P produces the current character* and the label *S skips the current character*.

As for NFFs, we need a special encoding of labeling $\boldsymbol{E}$ to represent "*no valid abbreviation*". Since there is no prior work, we need to study this "*no valid abbreviation*" issue. Given a full form $\boldsymbol{F}$, its valid abbreviation $\boldsymbol{A}$ should have the character number constraints: $0 < |\boldsymbol{A}| < |\boldsymbol{F}|$. On the other hand, we can assume that a negative full form has an "invalid abbreviation" $\boldsymbol{A}$ with $|\boldsymbol{A}| = 0$ or $|\boldsymbol{A}| = |\boldsymbol{F}|$. Those two kinds of interpretations actually represent two different answers to the question "why some full forms do not have valid abbreviations":

- `Assumption-1 with` $|\boldsymbol{A}| = 0$: It assumes that a negative full form F is "abbreviated" to nothing, i.e., with the abbreviation $A =$*NULL*.
- `Assumption-2 with` $|\boldsymbol{A}| = |\boldsymbol{F}|$: It assumes that a negative full form F is "abbreviated" to itself, i.e., with the abbreviation $A = F$.

We want to find out which assumption leads to better performance.

With those interpretations, invalid abbreviations are treated as special forms of abbreviations, thus positive and negative full forms can be modeled in a unified framework via sequential labeling. For simplicity, we use the well-known conditional random fields (CRFs) (Lafferty et al., 2001) for sequential labeling. Assuming a feature function that maps a pair of observation sequence $\boldsymbol{x}$ (characters of a full form) and label sequence $\boldsymbol{y}$ (label encoding based on abbreviations) to a feature vector $\boldsymbol{f}$, the probability function is defined as follows:

$$P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) = \frac{\exp\left[\boldsymbol{w}^\top \boldsymbol{f}(\boldsymbol{y}, \boldsymbol{x})\right]}{\sum_{\forall \boldsymbol{y}'} \exp\left[\boldsymbol{w}^\top \boldsymbol{f}(\boldsymbol{y}', \boldsymbol{x})\right]}, \quad (1)$$

where $\boldsymbol{w}$ is a parameter vector.

Given a training set consisting of n labeled sequences, $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, for $i = 1 \ldots n$, parameter estimation is performed by maximizing the objective function,

$$\mathcal{L}(\boldsymbol{w}) = \sum_{i=1}^{n} \log P(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{w}) - R(\boldsymbol{w}). \quad (2)$$

The second term is a regularizer, typically an L_2 (Gaussian) norm, $R(\boldsymbol{w}) = \frac{||\boldsymbol{w}||^2}{2\sigma^2}$.

We use features as follows:

- `Character feature` This feature records the input characters x_{i-1}, x_i and x_{i+1}.
- `Character bi-gram` The character bigrams starting at $(i-2) \ldots i$.
- `Numeral` Whether or not the x_i is a numeral.

[1] `http://ictclas.org/`

- `Organization name suffix` Whether or not the x_i is a suffix of traditional Chinese organization names.

- `Location name suffix` Whether or not the x_i is a suffix of traditional Chinese location names.

- `Word segmentation information` After the word segmentation step, whether or not the x_i is the beginning character of a word.

- `Part-of-speech information` The part-of-speech tag information of x_i.

i denotes the current position for extracting features.

The character unigram and bi-gram feature is to capture character-based information in the abbreviating process. For example, some special characters are more likely to be chosen in abbreviating. The named entity suffix features are used because a named entity suffix character is more likely to be chosen in abbreviating. The word segmentation information is also important because the beginning character of a word is more likely to be chosen in abbreviating.

2.3 Label Encoding with Global Information

As a common practice to reduce complexity, only local information based on Markov assumption is used for sequential labeling. Nevertheless, the Chinese abbreviation generation process is highly dependent on global information. An example of global information is the number of characters of the generated abbreviations.

For better performance, we try to model global information to make the system be "aware" of the number of characters being generated. We use a simple, effective, and tractable solution for modeling global information in abbreviation prediction: label encoding with global information (GI) (Sun et al., 2013).

In this approach, the label y_i at position i will be encoded with the global information of its previous labels, $y_1, y_2, \ldots, y_{i-1}$. Note that, while directly increasing the Markov order is untractable, the GI label encoding is tractable. More detailed description of the GI method is in (Sun et al., 2013).

Category	Portion (%)
Noun Phrase	52.01%
Verb Phrase	13.72%
Organization Name	26.84%
Location Name	5.28%
Person Name	0.32%
Others	1.80%

Table 1: Distribution of the full forms in the data.

2.4 Abbreviation Prediction for Web Search

Abbreviation prediction should be helpful for information search, but we find there is almost no prior work on this. It is probably because the traditional abbreviation prediction is not so applicable in real-world data, which includes lots of NFFs. Since we have solved this problem via generalized abbreviation prediction, we hope to apply generalized abbreviation prediction on improving information search.

In particular, we apply generalized abbreviation prediction for "query expansion" in Chinese web search. In this method, the original queries are treated as full forms (with NFFs) for generating abbreviations. Given a query as an input, the generalized abbreviation prediction system outputs an abbreviation candidate or a NULL string. If the output is an empty string, it means the query is an NFF. Finally, the derived abbreviations, together with the original query terms, are used for web search, and their search results (web pages) are simply added together. For the negative full forms, only the original full forms are used for the web search. The simple architecture of the query expansion system is summarized in Figure 2.

In addition, to make clear the role of the predicted abbreviations in web search, we can remove the full form information in web search and check the difference. In this way, the method turns to a "query alternation" method, which uses the predicted abbreviation to *replace* the positive full form for the web search. For a negative full form, the query alternation acts the same like the query expansion.

3 Experiments

Here we describe our collected data for generalized abbreviation prediction. First, we extract long phrases and terms from Chinese natural language processing corpora, including People's Daily cor-

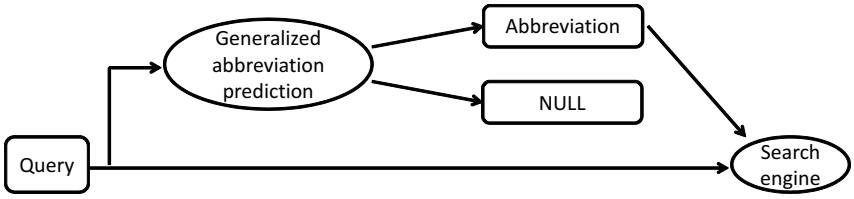

Figure 2: Generalized abbreviation prediction for improving web search.

Positive/Negative Full Forms	Abbreviation
磷酸氢二钠	X
君主专制制	X
珠穆朗玛峰	珠峰
天公不作美	X
中国社会科学院	中国社科院
新时期的总任务	X
自由民主党	自民党
车辆发动机	X
复员退伍军人安置办公室	复退办
车尔尼雪夫斯基	X
持谨慎态度	X
土产日用品杂品公司	土杂公司
一叶蔽目不见泰山	X
打击黑势力扫除恶势力	打黑扫恶

Figure 3: Samples of the collected data with NFFs. The "X" means no valid abbreviation.

pora[2] and SIGHAN word segmentation corpora[3]. Then, we classify the collected phrases and terms into positive and negative full forms. For the negative full forms, no further annotation is required. For the positive full forms, their abbreviations are annotated.

We build a dataset containing 10,786 full forms, including 8,015 positive full forms and 2,771 negative full forms. Samples of the data are shown in Figure 3. The dataset is made up of phrases and terms, including noun phrases, verb phrases, organization names, location names, and so on. The distribution is shown in Table 1. For experiments, we randomly sampled 8,629 samples (80% of the full dataset) for training and 2,157 (20% of the full dataset) for testing.

For experiments on web search, we simply use the 2,157 testing samples as query terms for web search. The evaluation is on the news domain[4] of the well-known web search engine "baidu.com"[5]. The Baidu news search engine has two alternative options: "title search" and "full content search". Since abbreviations are more common in news titles, we adopt the option of title search.

3.1 Experimental Settings

For evaluating abbreviation prediction quality, the systems are evaluated using the following two metrics:

- **All-match accuracy (All-Acc)**: The number of correct outputs (i.e., label strings) generated by the system divided by the total number of full forms in the test set.[6]
- **Character accuracy (Char-Acc)**: The number of correct labels (i.e., a classification on a character) generated by the system divided by the total number of characters in the test set.

For evaluating web search quality based on a given query, the following metrics are used:

- **Precision** P: The number of correct search results returned by the query divided by the total number of search results returned by the query.
- **Recall** R: The number of correct search results returned by the query divided by the total number of existing correct search results based on the query.
- **F-Score** F: $F = 2PR/(P + R)$.

[2] http://icl.pku.edu.cn/icl_res

[3] http://www.sighan.org/bakeoff2005

[4] We choose news domain because abbreviations are mainly from named entities, and named entities are important in news domain.

[5] http://news.baidu.com

[6] There is only one label string for a full form, and a label string corresponds to a unique abbreviation candidate. A label string is deemed as correct if and only if *all* of the labels are correct.

Method	Discriminate Acc (%)	Overall All-Acc	Overall Char-Acc
Heuristic System	73.20	25.77	65.79
Unified-Assum.1 (Perc)	87.48	54.89	87.02
Unified-Assum.1 (MEMM)	86.97	50.16	85.92
Unified-Assum.1 (CRF-ADF)	87.80	56.69	87.20
Unified-Assum.1-GI (Perc)	91.93	75.42	90.23
Unified-Assum.1-GI (MEMM)	88.59	70.32	88.21
Unified-Assum.1-GI (CRF-ADF)	91.05	**79.46**	**91.61**
Unified-Assum.2 (Perc)	86.83	55.86	82.20
Unified-Assum.2 (MEMM)	87.52	56.18	82.27
Unified-Assum.2 (CRF-ADF)	87.11	56.97	82.54
Unified-Assum.2-GI (Perc)	90.35	71.85	88.04
Unified-Assum.2-GI (MEMM)	87.99	63.74	83.77
Unified-Assum.2-GI (CRF-ADF)	90.77	74.78	89.19

Table 2: Results on comparing different methods on generalized abbreviation prediction. *Assum.1* and *Assum.2* represent the two assumptions on NFFs discussed in Section 2.2.2. *GI* means the integration with global information. As we can see, the *Unified-Assum.1-GI (CRF)* system has the best performance.

For evaluating web search quality based on a set of queries, we use the macro-averaging and micro-averaging of the precision, recall, and F-score based on a single query. Hence, we finally have six metrics: macro-precision, macro-recall, macro-F-score, micro-precision, micro-recall, and micro-F-score. We use the novel training method, adaptive online gradient descent based on feature frequency information (ADF) (Sun et al., 2012), for fast and accurate training of the CRF model.

To study the performance of other machine learning models, we also implement on other well-known sequential labeling models, including maximum entropy Markov models (MEMMs) (McCallum et al., 2000) and averaged perceptrons (Perc) (Collins, 2002).

3.2 Results on Abbreviation Prediction

The experimental results are shown in Table 2. In the table, the *overall accuracy* is most important and it means the final accuracy achieved by the systems in generalized abbreviation prediction with NFFs. For the completeness of experimental information, we also show the *discriminate accuracy*. The *discriminate accuracy* checks the accuracy of discriminating positive and negative full forms, without comparing the generated abbreviations with the gold-standard abbreviations.

As we can see from Table 2, first, the best system is the system *Unified-Assum.1-GI (CRF)*. Results demonstrate that incorporating global information can always improve the accuracy for the unified methods. Second, the unified system with *assumption-1* has better accuracy than the one with *assumption-2*. This result suggests that *assumption-1* works better in practice. It is interesting that assumption-1 is more useful. A probable reason is that those negative full forms have no similar patterns with the real abbreviations. For example, the number of characters in NFFs is very different compared with that of real abbreviations. Also, the NFFs contain more formal word units. Real abbreviations contain much less word units. Thus, assumption-2 will have the inconsistency problem between abbreviations generated from NFFs and real abbreviations. As a result, assumption-2 works worse than assumption-1 which gives no abbreviations. Finally, the CRF model outperforms the MEMM and averaged perceptron models. To summarize, the unified system with assumption-1, global information, and CRF model has the best performance.

3.3 Results on Web Search

We use the 2,157 testing samples as query terms for web search. We test the original query terms, the query alternation, and the query expansion methods. Some search results actually do not match the query. For example, given a query abc, some search results do not contain abc, but with the expression "$ab \ldots c$" or "$a \ldots bc$", where "$\ldots$" means other characters. In this case, the search results are incorrect. Since the number of the search results is massive, we need to evaluate the web

Method	Micro Prec	Micro Rec	Micro F1	Macro Prec	Macro Rec	Macro F1
Original query	48.51	18.14	26.41	47.84	28.76	35.92
Query alternation	47.73	54.04	50.69	62.84	61.12	61.97
Query expansion	47.93	72.18	**57.60**	53.70	82.02	**64.90**

Table 3: Results on comparing different methods on web search quality.

Method	Micro Prec	Micro Rec	Micro F1	Macro Prec	Macro Rec	Macro F1
Original query	48.51	18.14	26.41	47.84	28.76	35.92
Query alternation (gold-standard)	72.31	81.86	76.79	83.07	79.11	**81.04**
Query expansion (gold-standard)	66.40	100.00	**79.81**	65.56	100.00	79.19

Table 4: Results on comparing different methods on web search quality.

search quality in an efficient way. The correctness of the search results is evaluated by an automatic postprocessing scoring system, which crawls the search results from the `baidu.com` site. Then, the system runs text matching analysis to check if the search query matches the retrieved web pages.

In traditional web search studies, many queries are phrases (e.g., NP+VP) with ambiguous senses. In this case, improving search precision via contextual information is important. However, for abbreviation processing, most abbreviations are from named entities (the data contains phrases but most of them are NFFs), and the major problem of named entities is variational expressions. In this case, the search recall is more important. To calculate the recall rate in web search, we need to estimate the total number of correct web pages N relating to a query Q and its abbreviation A. We can estimate N via summing up the correct web pages of Q and the correct web pages of the gold-standard abbreviation A.

The precision, recall, F-scores are shown in Table 3. As we can see, the query expansion method based on generalized abbreviation prediction achieves significantly better F-scores on web search quality than using the original queries. We find the major improvement is from the recalls. As expected, the query alternation has lower recall rate than the query expansion method, because the full form information is removed. Nevertheless, the query alternation method is also better than using the original queries. This result emphasizes that the abbreviations are helpful in title-based news search.

Finally, we check the "up-bound" of the performance of generalized abbreviation prediction for web search. The up-bound is achieved by the 100% correct "gold-standard" system, in which gold-standard abbreviations labeled in the data are used. The results are shown in Table 4. As we can see, the up-bound of the micro-F-score and the macro-F-score is 79.81% and 81.04%, respectively. Thus, the web search quality of automatic generalized abbreviation prediction still has a large space to be improved, possibly via a larger training data set in the future.

4 Conclusions and Future Work

This paper is dedicated on generalized abbreviation prediction and its application on improving web search. Experiments demonstrate that the unified system based on global information outperforms the baselines. Experiments also demonstrate that generalized abbreviation prediction can improve web search qualities. As future work, we try to improve the performance via collecting more training data or via semi-supervised learning methods.

Acknowledgments

This work was supported by the Hong Kong Polytechnic University Internal Grant (4-ZZD5), Major National Social Science Fund of China (No.12&ZD227), National High Technology Research and Development Program of China (863 Program) (No.2012AA011101), and National Natural Science Foundation of China (No.91024009).

References

Jing-Shin Chang and Yu-Tso Lai. 2004. A preliminary study on probabilistic models for chinese abbreviations. In *Proceedings of the Third SIGHAN Workshop on Chinese Language Learning*, pages 9–16.

Jing-Shin Chang and Wei-Lun Teng. 2007. Mining atomic chinese abbreviation pairs: A probabilistic model for single character word recovery. *Language Resources and Evaluation*, 40:367–374.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of EMNLP'02*, pages 1–8.

Yaakov HaCohen-Kerner, Ariel Kass, and Ariel Peretz. 2008. Combined one sense disambiguation of abbreviations. In *Proceedings of ACL'08: HLT, Short Papers*, pages 61–64, June.

C.R. Huang, W.M. Hong, , and K.J. Chen. 1994. Suoxie: An information based lexical rule of abbreviation. In *Proceedings of the Second Pacific Asia Conference on Formal and Computational Linguistics II*, pages 49–52.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML'01)*, pages 282–289.

Andrew McCallum, Dayne Freitag, and Fernando Pereira. 2000. Maximum entropy Markov models for information extraction and segmentation. In *Proc. 17th International Conf. on Machine Learning*, pages 591–598. Morgan Kaufmann, San Francisco, CA.

Serguei Pakhomov. 2002. Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. In *Proceedings of ACL'02*, pages 160–167.

Xu Sun, Houfeng Wang, and Bo Wang. 2008. Predicting chinese abbreviations from definitions: An empirical learning approach using support vector regression. *Journal of Computer Science and Technology*, 23(4):602–611.

Xu Sun, Naoaki Okazaki, and Jun'ichi Tsujii. 2009. Robust approach to abbreviating terms: A discriminative latent variable model with global information. In *Proceedings of the ACL'09*, pages 905–913, Suntec, Singapore, August.

Xu Sun, Houfeng Wang, and Wenjie Li. 2012. Fast online training with frequency-adaptive learning rates for chinese word segmentation and new word detection. In *Proceedings of ACL'12*, pages 253–262.

Xu Sun, Naoaki Okazaki, Jun'ichi Tsujii, and Houfeng Wang. 2013. Learning abbreviations from chinese and english terms by modeling non-local information. *ACM Trans. Asian Lang. Inf. Process.*, 12(2):5.

Dong Yang, Yi-Cheng Pan, and Sadaoki Furui. 2009a. Automatic chinese abbreviation generation using conditional random field. In *Proceedings of HLT-NAACL'09 (Short Papers)*, pages 273–276.

Dong Yang, Yi-Cheng Pan, and Sadaoki Furui. 2009b. Vocabulary expansion through automatic abbreviation generation for chinese voice search. In *Proceedings of INTERSPEECH'09*, pages 728–731. ISCA.

Hong Yu, Won Kim, Vasileios Hatzivassiloglou, and John Wilbur. 2006. A large scale, corpus-based approach for automatically disambiguating biomedical abbreviations. *ACM Transactions on Information Systems (TOIS)*, 24(3):380–404.