Mining Atomic Chinese Abbreviation Pairs: A Probabilistic Model for Single Character Word Recovery

Jing-Shin Chang
Department of Computer Science &
Information Engineering
National Chi-Nan University
Puli, Nantou, Taiwan, ROC.
jshin@csie.ncnu.edu.tw

Wei-Lun Teng
Department of Computer Science &
Information Engineering
National Chi-Nan University
Puli, Nantou, Taiwan, ROC.
S3321512@ncnu.edu.tw

Abstract

An HMM-based Single Character Recovery (SCR) Model is proposed in this paper to extract a large set of "atomic abbreviation pairs" from a large text corpus. By an "atomic abbreviation pair," it refers to an abbreviated word and its root word (i.e., unabbreviated form) in which the abbreviation is a single Chinese character.

This task is interesting since the abbreviation process for Chinese compound words seems to be "compositional"; in other words, one can often decode an abbreviated word, such as "台大" (Taiwan University), character-by-character back to its root form. With a large atomic abbreviation dictionary, one may be able to recover multiple-character abbreviations more easily.

With only a few training iterations, the acquisition accuracy of the proposed SCR model achieves 62% and 50 % precision for training set and test set, respectively, from the ASWSC-2001 corpus.

1 Introduction

Chinese abbreviations are widely used in the modern Chinese texts. They are a special form of *unknown words*, which cannot be exhaustively enumerated in an ordinary dictionary. Many of them originated from important lexical units such as *named entities*. However, the sources for Chinese abbreviations are not solely from the noun class, but from most major categories, including verbs, adjectives adverbs and others. No matter what lexical or syntactic structure a string of characters could be, one can almost always find a way to abbreviate it into a shorter form. Therefore, it may be necessary to handle them beyond a class-based model. Furthermore, abbreviated words are semantically ambiguous. For example, "清大" can be the abbreviation for "清華大學" or "清潔大隊"; on the opposite direction, multiple choices for abbreviating a word are also possible. For instance, "台北大學" may be abbreviated as "台大", "北大" or "台北大". This results in difficulty for correct Chinese processing and applications, including word segmentation, information retrieval, query expansion, lexical translation and much more. An abbreviation model or a large abbreviation lexicon is therefore highly desirable for Chinese language processing.

Since the smallest possible Chinese lexical unit into which other words can be abbreviated is a single character, identifying the set of multi-character words which can be abbreviated into a single character is especially interesting. Actually, the abbreviation of a compound word can often be acquired by the principle of *composition*. In other words, one can decompose a compound word into its constituents and then concatenate their single character equivalents to form its abbreviated form. The reverse process to predict the unabbreviated form from an abbreviation shares the same compositional property.

The Chinese abbreviation problem can be regarded as an *error recovery* problem in which the suspect root words are the "errors" to be recovered from a set of candidates. Such a problem can be mapped to an HMM-based generation model for both abbreviation identification and root word recovery; it can also

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 17-24,
COLING-ACL 2006, Sydney, Australia, 22-23, July 2006. © 2006 Association for Computational Linguistics.

be integrated as part of a unified word segmentation model when the input extends to a complete sentence. As such, we can find the most likely root words, by finding those candidates that maximizes the likelihood of the whole text. An abbreviation lexicon, which consists of the root-abbreviation pairs, can thus be constructed automatically.

In a preliminary study (Chang and Lai, 2004), some probabilistic models had been developed to handle this problem by applying the models to a parallel corpus of compound words and their abbreviations, without knowing the context of the abbreviation pairs. In this work, the same framework is extended and a method is proposed to automatically acquire a large abbreviation lexicon for indivisual characters from web texts or large corpora, instead of building abbreviation models based on aligned abbreviation pairs of short compound words. Unlike the previous task, which trains the abbreviation model parameters from a list of known abbreviation pairs, the current work aims at extracting abbreviation pairs from a corpus of free text, in which the locations of prospective abbreviations and full forms are unknown and the correspondence between them is not known either.

In particular, a Single Character Recovery (SCR) Model is exploited in the current work to extract "atomic abbreviation pairs" from a large text corpus. With only a few training iterations, the acquisition accuracy achieves 62% and 50 % precision for training set and test set from the ASWSC-2001 corpus.

1.1 Chinese Abbreviation Problems

The modern Chinese language is a highly abbreviated one due to the mixed uses of ancient single character words as well as modern multi-character words and compound words. The abbreviated form and root form are used interchangeably everywhere in the current Chinese articles. Some news articles may contain as high as 20% of sentences that have suspect abbreviated words in them (Lai, 2003). Since abbreviations cannot be enumerated in a dictionary, it forms a special class of *unknown words*, many of which originate from *named entities*. Many other open class words are also abbreviatable. This particular class thus introduces complication for Chinese language processing, including the fundamental *word segmentation* process (Chiang *et al.*, 1992; Lin *et al.*, 1993; Chang and Su, 1997) and many word-based applications. For instance, a keyword-based information retrieval system may requires the two forms, such as "中研院" and "中央研究院" ("Academia Sinica"), in order not to miss any relevant documents. The Chinese word segmentation process is also significantly degraded by the existence of unknown words (Chiang *et al.*, 1992), including unknown abbreviations.

There are some heuristics for Chinese abbreviations. Such heuristics, however, can easily break (Sproat, 2002). Unlike English abbreviations, the abbreviation process of the Chinese language is a very special word formation process. Almost all characters in all positions of a word can be omitted when used for forming an abbreviation of a compound word. For instance, it seems that, by common heuristics, "most" Chinese abbreviations could be derived by keeping the first characters of the constituent words of a compound word, such as transforming '台灣大學' into '台大', '清華大學' into '清大' and '以色列(及)巴勒斯坦' into '以巴'. Unfortunately, it is not always the case. For example, we can transform '台灣香港' into '台港', '中國石油' into '中油', and, for very long compounds like '雲林嘉義台南' into '雲嘉南' (Sproat, 2002). Therefore, it is very difficult to predict the possible surface forms of Chinese abbreviations and to guess their base (non-abbreviated) forms heuristically.

P(bit|n)	Score	Examples
P(10|2)	0.87	(德|德國),(美|美國)
P(101|3)	0.44	(宜縣|宜蘭縣), (限級|限制級)
P(1010|4)	0.56	(公投|公民投票), (清大|清華大學)
P(10101|5)	0.66	(環保署|環境保護署), (航警局|航空警察局)
P(101001|6)	0.51	(化工系|化學工程學系), (工工系|工業工程學系)
P(1010001|7)	0.55	(國科會|國家科學委員會), (中科院|中山科學研究院)
P(10101010|8)	0.21	(一中一台|一個中國一個台灣), (一大一小|一個大人一個小孩)

Table 1. High Frequency Abbreviation Patterns [by P(bit|n)] (Chang and Lai, 2004)

The high frequency abbreviation patterns revealed in (Chang and Lai, 2004) further break the heuristics quantitatively. Table 1 lists the distribution of the most frequent abbreviation patterns for word of length 2~8 characters.

The table indicates which characters will be deleted from the root of a particular length (n) with a bit ‘0’; on the other hand, a bit ‘1’ means that the respective character will be retained. This table does support some general heuristics for native Chinese speaker quantitatively. For instance, there are strong supports that the first character in a two-character word will be retained in most cases, and the first and the third characters in a 4-character word will be retained in 56% of the cases. However, the table also shows that around 50% of the cases cannot be uniquely determined by character position simply by consulting the word length of the un-abbreviated form. This does suggest the necessity of either an abbreviation model or a large abbreviation lexicon for resolving this kind of unknown words and named entities.

There are also a large percentage (312/1547) of “*tough*” abbreviation patterns (Chang and Lai, 2004), which are considered “*tough*” in the sense that they violate some simple assumptions, and thus cannot be modeled in a simple way. For instance, some tough words will actually be *recursively* abbreviated into shorter and shorter lexical forms; and others may change the word order (as in abbreviating “第一核能發電廠” as “核一廠” instead of “一核廠”.). As a result, the abbreviation process is much more complicated than a native Chinese speaker might think.

1.2 Atomic Abbreviation Pairs

Since the abbreviated words are created continuously through the abbreviation of new (mostly compound) words, it is nearly impossible to construct a complete abbreviation lexicon. In spite of the difficulty, it is interesting to note that the abbreviation process for Chinese compound words seems to be “compositional”. In other words, one can often decode an abbreviated word, such as “台大” (“Taiwan University”), character-by-character back to its root form “台灣大學” by observing that “台” can be an abbreviation of “台灣” and “大” can be an abbreviation of “大學” and “台灣大學” is a frequently observed character sequence.

Since character is the smallest lexical unit for Chinese, no further abbreviation into smaller units is possible. We therefore use “atomic abbreviation pair” to refer to an abbreviated word and its root word (i.e., unabbreviated form) in which the abbreviation is a single Chinese character.

On the other hand, abbreviations of multi-character compound words may be synthesized from single characters in the “atomic abbreviation pairs”. If we are able to identify all such “atomic abbreviation pairs”, where the abbreviation is a single character, and construct such an atomic abbreviation lexicon, then resolving multiple character abbreviation problems, either by heuristics or by other abbreviation models, might become easier.

Furthermore, many ancient Chinese articles are composed mostly of single-character words. Depending on the percentage of such single-character words in a modern Chinese article, the article will resemble to an ancient Chinese article in proportional to such a percentage. As another application, an effective single character recovery model may therefore be transferred into an auxiliary translation system from ancient Chinese articles into their modern versions. This is, of course, an overly bold claim since lexical translation is not the only factor for such an application. However, it may be consider as a possible direction for lexical translation when constructing an ancient-to-modern article translation system. Also, when a model for recovering atomic translation pair is applied to the “single character regions” of a word segmented corpus, it is likely to recover unknown abbreviated words that are previously word-segmented incorrectly into individual characters.

An HMM-based Single Character Recovery (SCR) Model is therefore proposed in this paper to extract a large set of “atomic abbreviation pairs” from a large text corpus.

1.3 Previous Works

Currently, only a few quantitative approaches (Huang *et al.*, 1994a; Huang *et al.*, 1994b) are available in predicting the presence of an abbreviation. There are essentially no prior arts for automatically extracting atomic abbreviation pairs. Since such formulations regard the word segmentation process and abbreviation identification as two independent processes, they probably cannot optimize the identification process jointly with the *word segmentation* process, and thus may lose some useful contextual information. Some class-based segmentation models (Sun *et al.*, 2002; Gao *et al.*, 2003) well integrate the identification of some regular non-lexicalized units (such as named entities). However, the abbreviation process can be applied to almost all word forms (or classes of words). Therefore, this particular word formation process may have to be handled as a separate layer in the segmentation process.

To resolve the Chinese abbreviation problems and integrate its identification into the word segmentation process, (Chang and Lai, 2004) proposes to regard the abbreviation problem in the word segmentation process as an "error recovery" problem in which the suspect root words are the "errors" to be recovered from a set of candidates according to some generation probability criteria. This idea implies that an HMM-based model for identifying Chinese abbreviations could be effective in either identifying the existence of an abbreviation or the recovery of the root words from an abbreviation.

Since the parameters of an HMM-like model can usually be trained in an unsupervised manner, and the "output probabilities" known to the HMM framework will indicate the likelihood for an abbreviation to be generated from a root candidate, such a formulation can easily be adapted to collect highly probable root-abbreviation pairs. As a side effect of using HMM-based formulation, we expect that a large abbreviation dictionary could be derived automatically from a large corpus or from web documents through the training process of the unified word segmentation model.

In this work, we therefore explore the possibility of using the theories in (Chang and Lai, 2004) as a framework for constructing a large abbreviation lexicon consisting of all Chinese characters and their potential roots. In the following section, the HMM models as outlined in (Chang and Lai, 2004) is reviewed first. We then described how to use this framework to construct an abbreviation lexicon automatically. In particular, a Single Character Recovery (SCR) Model is exploited for extracting possible root (un-abbreviated) forms for all Chinese characters.

2 Chinese Abbreviation Models

2.1 Unified Word Segmentation Model for Abbreviation Recovery

To resolve the abbreviation recovery problem, one can identify some root candidates for suspect abbreviations (probably from a large abbreviation dictionary if available or from an ordinary dictionary with some educated guesses), and then confirm the most probable root by consulting local context. This process is identical to the operation of many error correction models, which generate the candidate corrections according to a *reversed* word formation process, then justify the best candidate.

Such an analogy indicates that we may use an HMM model (Rabiner and Juang, 1993), which is good at finding the best *unseen* state sequence, for root word recovery. There will be a direct map between the two paradigms if we regard the observed input character sequence as our "observation sequence", and regard the unseen word candidates as the underlying "state sequence".

To integrate the abbreviation process into the word segmentation model, firstly we can regard the segmentation model as finding the best underlying words $w_1^m \equiv w_1, \cdots, w_m$ (which include only base/root forms), given the surface string of characters $c_1^n \equiv c_1, \cdots, c_n$ (which may contain abbreviated forms of compound words.) The segmentation process is then equivalent to finding the best (un-abbreviated) word sequence $\vec{w}*$ such that:

$$\begin{aligned}\vec{w}* &= \mathop{\arg\max}_{w_1^m : w_1^m \Rightarrow c_1^n} P\left(w_1^m \mid c_1^n\right) \\ &= \mathop{\arg\max}_{w_1^m : w_1^m \Rightarrow c_1^n} P\left(c_1^n \mid w_1^m\right) \times P\left(w_1^m\right) \\ &= \mathop{\arg\max}_{w_1^m : w_1^m \Rightarrow c_1^n} \prod_{\substack{i=1,m \\ w_i \Rightarrow \vec{c}_i}} P\left(\vec{c}_i \mid w_i\right) \times P\left(w_i \mid w_{i-1}\right)\end{aligned}$$

Equation 1. Unified Word Segmentation Model for Abbreviation Recovery

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 17-24,
COLING-ACL 2006, Sydney, Australia, 22-23, July 2006. © 2006 Association for Computational Linguistics.

where $\vec{c}_i$ refers to the surface form of w_i, which could be in an abbreviated or non-abbreviated root form of w_i. The last equality assumes that the generation of an abbreviation is independent of context, and the language model is a word-based bigram model.

If no abbreviated words appears in real text, such that all surface forms are identical to their "root" forms, we will have $P(\vec{c}_i \mid w_i) = 1$, $\forall i = 1, m$, and **Equation 1** is simply a word bigram model for word segmentation (Chiang *et al.*, 1992). In the presence of abbreviations, however, the generation probability $P(\vec{c}_i \mid w_i)$ can no longer be ignored, since the probability $P(\vec{c}_i \mid w_i)$ is not always 1 or 0.

As an example, if two consecutive $\vec{c}_i$ are '台' and '大' then their roots, w_i, could be '台灣' plus '大學' (Taiwan University) or '台灣' plus '大聯盟' (Taiwan Major League). In this case, the parameters in P(大學|台灣) x P(台|台灣) x P(大|大學) and P(大聯盟|台灣) x P(台|台灣) x P(大|大聯盟) will indicate how likely '台大' is an abbreviation, and which of the above two compounds is the root form.

Notice that, this equation is equivalent to an HMM (Hidden Markov Model) (Rabiner and Juang, 1993) normally used to find the best "state" sequence given the observation symbols. The parameters $P(w_i \mid w_{i-1})$ and $P(\vec{c}_i \mid w_i)$ represent the transition probability and the (word-wise) output probability of an HMM, respectively; and, the formulations for $P(w_1^m)$ and $P(c_1^n \mid w_1^m)$ are the respective "language model" of the Chinese language and the "generation model" for the abbreviated words (i.e., the "abbreviation model" in the current context). The "state" sequence in this case is characterized by the hidden root forms $w_1^m \equiv w_1, \cdots, w_m$; and, the "observation symbols" are characterized by $c_1^n \equiv c_1, \cdots, c_n \equiv \vec{c}_1, \cdots, \vec{c}_m$, where the surface form $\vec{c}_i \equiv c_{b(i)}^{e(i)}$ is a chunk of characters beginning at the b(i)-th character and ending at the e(i)-th character.

The word-wise transition probability $P(w_i \mid w_{i-1})$ in the language model is used to provide contextual constraints among root words so that the underlying word sequence forms a legal sentence with high probability.

Notice that, after applying the word segmentation model **Equation 1** to the word lattice, some of the above candidates may be preferred and others be discarded, by consulting the neighboring words and their transition probabilities. This makes the abbreviation model *jointly* optimized in the word segmentation process, instead of being optimized independent of context.

2.2 Simplified Abbreviation Models

Sometimes, it is not desirable to use a generation probability that is based on the root-abbreviation pairs, since the number of parameters will be huge and estimation error due to data sparseness might be high. Therefore, it is desirable to simplify the abbreviation probability by using some simpler features in the model. For instance, many 4-character compound words are abbreviated as 2-character abbreviations (such as in the case for the <台灣大學, 台大> pair.) It was also known that many such 4-character words are abbreviated by preserving the first and the third characters, which can be represented by a '1010' bit pattern, where the '1' or '0' means to preserve or delete the respective character. Therefore, a reasonable simplification for the abbreviation model is to introduce the *length* and the positional *bit pattern* as additional features, resulting in the following augmented model for the abbreviation probability.

$$
\begin{aligned}
&P(\vec{c} \mid w) = P(c_1^m, bit, m \mid r_1^n, n) \\
&\equiv P(c_1^m \mid r_1^n) \times P(bit \mid n) \times P(m \mid n) \\
&\text{where } \begin{cases} c_1^m : \text{ surface characters.} \\ r_1^n : \text{ root word characters.} \\ m\text{: length of surface characters.} \\ n\text{: length of root word characters.} \\ \text{bit: bit pattern of abbreviation} \end{cases}
\end{aligned}
$$

Equation 2. Abbreviation Probability using Abbreviation Pattern and Length Features.

All these three terms can be combined freely to produce as many as 7 sub-models for the abbreviation model. Note, the first term $\Pr\left(c_1^m \mid r_1^n\right)$ plays the same role as the older notation of $\Pr(\vec{c} \mid w)$. To use the simple length and position features, this term can be unused in the above augmented abbreviation model.

3 The Single Character Recovery (SCR) Model

As mentioned earlier, many multiple character words are frequently abbreviated into a single Chinese character. Compound words consisting of a couple of such multiple character words are then abbreviated by concatenating all the single character abbreviations. This means that those N-to-1 abbreviation patterns may form the basis for the underlying Chinese abbreviation process. The other M-to-N abbreviation patterns might simply be a composition of such basic N-to-1 abbreviations. The N-to-1 abbreviation patterns can thus be regarded as the atomic abbreviation pairs.

Therefore, it is interesting to apply the abbreviation recovery model to acquire all basic N-to-1 abbreviation patterns, in the first place, so that abbreviations of multi-character words can be detected and predicted more easily.

Such a task can be highly simplified if each character in a text corpus is regarded as an abbreviated word whose root form is to be recovered. In other words, the surface form $\vec{c}_i$ in Equation 1 is reduced to a single character. The abbreviation recovery model based on this assumption will be referred to as the SCR Model.

The root candidates for each single character will form a word lattice, and each path of the lattice will represent a non-abbreviated word sequence. The underlying word sequence that is most likely to produce the input character sequence will then be identified as the best word sequence. Once the best word sequence is identified, the model parameters can be re-estimated. And the best word sequence is identified again. Such process is repeated until the best sequence no more changes. In addition, the corresponding <root, abbreviation> pairs will be extracted as atomic abbreviation pairs, where all the abbreviations are one character in size.

While it is overly simplified to use this SCR model for conducting a general abbreviation enhanced word segmentation process (since not all single characters are abbreviated words), the single character assumption might still be useful for extracting roots of real single-character abbreviations. The reason is that one only care to use the contextual constraints around a true single character abbreviation for matching its root form against the local context in order to confirm that the suspect root did conform to its neighbors (with a high language model score). An alternative to use a two-stage recovery strategy, where the first stage applies a baseline word segmentation model to identify most normal words and a second stage to identify and recover incorrectly segmented single characters, is currently under investigation with other modes of operations. For the present, the SCR model is tested first as a baseline.

The HMM-based recovery models enables us to estimate the model parameters using an unsupervised training method that is directly ported from the Baum-Welch re-estimation formula (Rabiner and Juang, 1993) or a generic EM algorithm (Dempster *et al.*, 1977). Upon convergence, we should be able to acquire a large corpus of atomic abbreviation pairs from the text corpus.

If a word-segmented corpus is available, we can also use such re-estimation methods for HMM parameter training in an unsupervised manner, but with initial word transition probabilities estimated in a supervised manner from the seed.

The initial candidate <root, abbreviation> pairs are generated by assuming that all word-segmented words in the training corpus are potential roots for each of its single-character constituents. For example, if we have "台灣" and "大學" as two word-segmented tokens, then the abbreviation pairs <台, 台灣>, <灣, 台灣>, <大, 大學> and <學, 大學> will be generated. Furthermore, each single character by default is its own abbreviation.

To estimate the abbreviation probabilities, each abbreviation pair is associated with a frequency count of the root in the word segmentation corpus. This means that each single-character abbreviation candidate is equally weighted. The equal weighting strategy may not be absolutely true (Chang and Lai, 2004). In fact, the character position and word length features may be helpful as mentioned in **Equation 2**. The initial probabilities are therefore weighted differently according to the position of the character and the length of the root. The weighting factors are directly acquired from a previous work in (Chang and Lai, 2004). Before the initial probabilities are estimated, Good-Turning smoothing (Katz, 1987) is applied to the raw frequency counts of the abbreviation pairs.

4 Experiments

To evaluate the SCR Model, the Academia Sinica Word Segmentation Corpus (dated July, 2001), ASWSC-2001 (CKIP, 2001), is adopted for parameter estimation and performance evaluation. Among the 94 files in this balanced corpus, 83 of them (13,086KB) are randomly selected as the training set and 11 of them (162KB) are used as the test set.

Table 3 shows some examples of atomic abbreviation pairs acquired from the training corpus. The examples here partially justify the possibility to use the SCR model for acquiring atomic abbreviation pairs from a large corpus.

Abbr : Root	Example	Abbr : Root	Example
籃:籃球	籃賽	宣:宣傳	文宣
檢:檢查	安檢	農:農業	農牧
媒:媒體	政媒	艙:座艙	艙壓
宿:宿舍	男舍	攜:攜帶	攜械
臺:臺灣	臺幣	汽:汽車	汽機車
漫:漫畫	動漫	咖:咖啡店	網咖
港:香港	港人	職:職業	現職
網:網路	網咖	設:設計	工設系
股:股票	股市	湖:澎湖	臺澎
韓:韓國	韓流	海:海洋	海生館
祕:祕書	主秘	文:文化	文建會
植:植物	植被	生:學生	新生
儒:儒家	新儒學	新:新加坡	新國
盜:強盜	盜匪	花:花蓮	花東
房:房間	機房	資:資訊	資工

Table 3. Examples of atomic abbreviation pairs.

The iterative training process converges quickly after 3~4 iterations. The numbers of unique abbreviation patterns for the training and test sets are 20,250 and 3,513, respectively. Since the numbers of patterns are large, a rough estimate on the acquisition accuracy rates is conducted by randomly sampling 100 samples of the <root, abbreviation> pairs. The pattern is then manually examined to see if the root is correctly recovered. The precision is estimated as 50% accuracy for the test set, and 62% for the training set on convergence. Although a larger sample may be necessary to get a better estimate, the preliminary result is encouraging. Figures 1~4 demonstrate the curve of convergence for the iterative training process in terms of pattern number and accuracy.

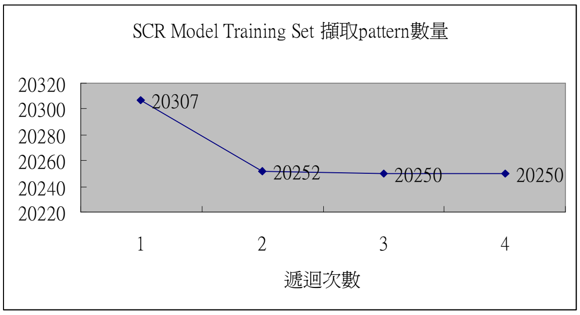

Figure 1. Numbers of abbreviation patterns in each iteration. (Training Set)

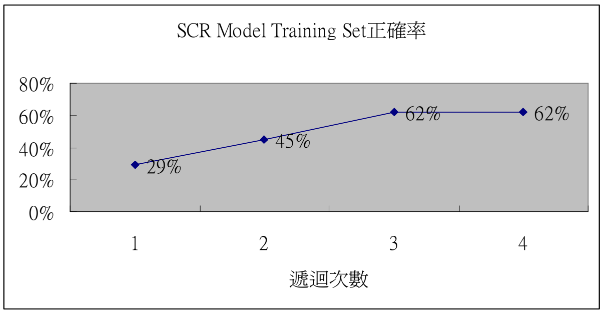

Figure 2. Acquisition accuracy for each iteration. (Training Set)

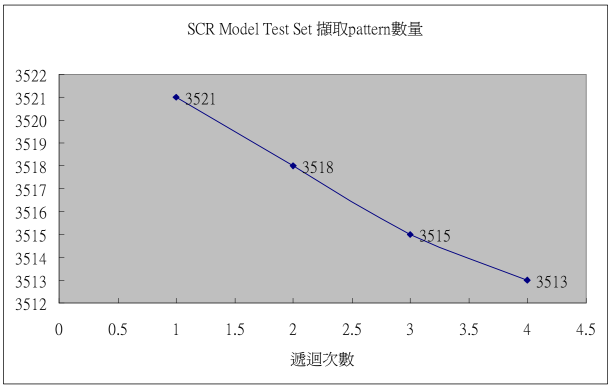

Figure 3. Numbers of abbreviation patterns in each iteration. (Test Set)

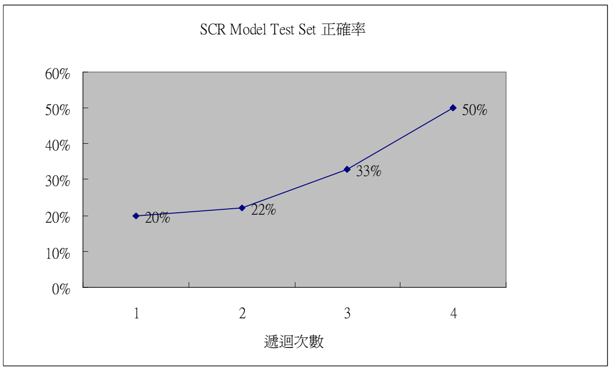

Figure 4. Acquisition accuracy for each iteration. (Test Set)

5 Concluding Remarks

Chinese abbreviations, which form a special kind of unknown words, are widely seen in the modern Chinese texts. This results in difficulty for correct Chinese processing. In this work, we had applied a unified word segmentation model developed in a previous works (Chang and Lai, 2004), which was able to handle the kind of "errors" introduced by the abbreviation process. An iterative training process is developed to automatically acquire an abbreviation dictionary for single-character abbreviations from large corpora. In particular, a Single Character Recovery (SCR) Model is exploited. With only a few training iterations, the acquisition accuracy achieves 62% and 50 % precision for training set and test set from the ASWSC-2001 corpus. For systems that choose to lexicalize such lexicon entities, the automatically constructed abbreviation dictionary will be an invaluable resource to the systems. And, the proposed recovery model looks encouraging.

Acknowledgements

The current work was supported by the National Science Council (NSC), Taiwan, Republic of China (ROC), under the contract NSC 93-2213-E-260-015.

References

Chang, Jing-Shin and Keh-Yih Su, 1997. "An Unsupervised Iterative Method for Chinese New Lexicon Extraction", *International Journal of Computational Linguistics and Chinese Language Processing (CLCLP)*, 2(2): 97-148.

Chang, Jing-Shin and Yu-Tso Lai, 2004. "A Preliminary Study on Probabilistic Models for Chinese Abbreviations." *Proceedings of the Third SIGHAN Workshop on Chinese Language Learning,* pages 9-16, ACL-2004, Barcelona, Spain.

Chiang, Tung-Hui, Jing-Shin Chang, Ming-Yu Lin and Keh-Yih Su, 1992. "Statistical Models for Word Segmentation and Unknown Word Resolution," *Proceedings of ROCLING-V*, pages 123-146, Taipei, Taiwan, ROC.

CKIP 2001, *Academia Sinica Word Segmentation Corpus, ASWSC-2001, (中研院中文分詞語料庫)*, Chinese Knowledge Information Processing Group, Acdemia Sinica, Tiapei, Taiwan, ROC.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society*, 39 (b): 1-38.

Gao, Jianfeng, Mu Li, Chang-Ning Huang, 2003. "Improved Source-Channel Models for Chinese Word Segmentation," *Proc. ACL 2003*, pages 272-279.

Huang, Chu-Ren, Kathleen Ahrens, and Keh-Jiann Chen, 1994a. "A data-driven approach to psychological reality of the mental lexicon: Two studies on Chinese corpus linguistics." In *Language and its Psychobiological Bases*, Taipei.

Huang, Chu-Ren, Wei-Mei Hong, and Keh-Jiann Chen, 1994b. "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, Japan.

Katz, Slava M., 1987. "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. ASSP-35 (3).*

Lai, Yu-Tso, 2003. *A Probabilistic Model for Chinese Abbreviations,* Master Thesis, National Chi-Nan University, ROC.

Lin, Ming-Yu, Tung-Hui Chiang and Keh-Yih Su, 1993. "A Preliminary Study on Unknown Word Problem in Chinese Word Segmentation," *Proceedings of ROCLING VI*, pages 119-142.

Rabiner, L., and B.-H., Juang, 1993. *Fundamentals of Speech Recognition*, Prentice-Hall.

Sun, Jian, Jianfeng Gao, Lei Zhang, Ming Zhou and Chang-Ning Huang, 2002. "Chinese named entity identification using class-based language model," *Proc. of COLING 2002*, Taipei, ROC.

Sproat, Richard, 2002. "Corpus-Based Methods in Chinese Morphology", *Pre-conference Tutorials*, COLING-2002, Taipei, Taiwan, ROC.