Lang Resources & Evaluation (2006) 40:367–374
DOI 10.1007/s10579-007-9026-8

ORIGINAL PAPER

Mining atomic Chinese abbreviations with a probabilistic single character recovery model

Jing-Shin Chang · Wei-Lun Teng

Published online: 18 July 2007

Abstract An HMM-based single character recovery (SCR) model is proposed in this paper to extract a large set of atomic abbreviations and their full forms from a text corpus. By an ''atomic abbreviation,'' it refers to an abbreviated word consisting of a single Chinese character. This task is important since Chinese abbreviations cannot be enumerated exhaustively but the abbreviation process for compound words seems to be *compositional*. One can often decode an abbreviated word character by character to its full form. With a large atomic abbreviation dictionary, one may be able to handle multiple character abbreviation problems more easily based on the compositional property of abbreviations.

Keywords Abbreviation · Atomic abbreviation · Single character recovery model

1 Motivation

Chinese abbreviations are widely used in the modern Chinese texts. They are a special form of out-of-vocabulary words, which cannot be exhaustively enumerated in a dictionary. A Chinese abbreviation is normally generated by deleting some characters in its unabbreviated full form (hereafter, its ''root'' for short), while retaining representative characters that preserve meaning. Many abbreviations are named entities. However, the sources for Chinese abbreviations are not solely from the *noun* class, but also from most major categories, including *verbs, adjectives, adverbs* and others. In fact, 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

J.-S. Chang (✉) · W.-L. Teng
Department of Computer Science & Information Engineering,
National Chi-Nan University, Puli, Nantou, Taiwan, ROC
e-mail: jshin@csie.ncnu.edu.tw

W.-L. Teng
e-mail: s3321512@ncnu.edu.tw

form. Therefore, it may be necessary to handle them in a separated layer above any classes. Furthermore, abbreviated words are often semantically ambiguous. For example, 清大 tsing-da can represent 清華大學 tsing-hua-da-xue 'Tsing-Hua University' or 清潔大隊 tsing-jie-da-duei 'cleaning team'; on the opposite direction, multiple choices for abbreviating a word are also possible. For instance, 台北大學 tai-bei-da-xue 'Taipei University' may be abbreviated as 台大 tai-da, 北大 bei-da or 台北大 tai-bei-da. This kind of two-way ambiguity makes it difficult either to generate the abbreviated form from a full form or to recover the full form of an abbreviation.

An abbreviation serves many linguistic functions. First of all, it is a synonym of its full form. Secondly, it is also a translation equivalent of its full form in cross-lingual environments. Therefore, they can be used interchangeably in *mono-* or *multi-lingual* applications. As such, it results in difficulty for Chinese language processing and applications, including word segmentation (Chiang et al. 1992, 1996; Lin et al. 1993), information retrieval, query expansion, lexical translation and more. For instance, a keyword-based information retrieval system may require the two forms, such as 中研院 zhong-yian-yuan and 中央研究院 zhong-yang-yian-jiou-yuan 'Academia Sinica', in order not to miss any relevant documents. The Chinese word segmentation process is also significantly degraded by the existence of out-of-vocabulary words (Chiang et al. 1992, 1996), including unregistered abbreviations. An abbreviation model or a large abbreviation lexicon is therefore highly desirable for Chinese abbreviation processing. However, abbreviations cannot be enumerated exhaustively. This implies that we may have to find all the sub-word atomic abbreviation patterns in order to completely solve the problems.

Identifying the set of full forms for all individual Chinese characters is especially interesting since the smallest possible abbreviation target is a single character. Such a single character abbreviation will be referred to as an "atomic abbreviation." The abbreviation and its full form will be referred to as an atomic abbreviation pattern, accordingly.

Actually, 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 tai-da, character by character to its full form 台灣大學 tai-wan-da-xue 'Taiwan University' by observing that 台 tai can be an abbreviation of 台灣 tai-wan 'Taiwan' and 大 da can be an abbreviation of 大學 da-xue 'University' and 台灣大學 tai-wan-da-xue is a frequently observed character sequence in real text. On the other hand, multiple character abbreviations of compound words can often be *synthesized* from single character abbreviations. In other words, one can decompose a compound word into its constituents and then concatenate their single character equivalents to form its abbreviated form. If we are able to identify all atomic abbreviation patterns for all Chinese characters, then multiple character abbreviation problems might be resolved more easily. Therefore, a model for mining the full forms of the finite Chinese character set could be significant.

Two kinds of abbreviation problems can be identified; one is to generate abbreviations from full forms, the other is to recover full forms from abbreviations. Currently, only a few quantitative approaches are available for the *generation* of abbreviations. For instance, Huang et al. (1998) proposed a (binary point-wise) mutual information model for resolving ambiguity with good results in generating

16 abbreviated county names. There are essentially no prior arts for *recovering* abbreviations to their full forms until Lai (2003). In particular, automatically extracting full forms for atomic abbreviations in the full Chinese character set, as addressed in this paper, is not seen. There are various types of abbreviations. The current paper is interested in morphological shortening. Syntactical omission, such as abbreviating 清華大學 tsing-hua-da-xue 'Tsing-Hua University' as 清華 tsing-hua by omitting the organizational title will not be addressed, since it requires word sense disambiguation which is beyond a simple morphological framework. For more interesting types of abbreviations and *tough* abbreviation patterns, Lai (2003) as well as Chang and Lai (2004) have more quantitative analyses.

The Chinese abbreviation recovery problem can be regarded as an *error recovery* problem (Chang and Lai 2004) in which the abbreviations are the "errors" to be recovered to their unseen full forms. Such a problem can be mapped to an HMM-based model for both abbreviation identification and full form recovery by integrating the abbreviation process into a unified word segmentation model. In the unsupervised training process for the model parameters, the most likely full forms can then be automatically extracted by finding candidates that maximize the likelihood of the training sentences. An abbreviation lexicon, which consists of the most probable root-abbreviation pairs, can thus be constructed automatically.

In the following section, the unified word segmentation model with abbreviation recovery capability (Chang and Lai 2004) is reviewed. We then describe how to adapt this general framework to a simplified single character recovery (SCR) model to construct an atomic abbreviation lexicon for all Chinese characters.

2 Unified word segmentation model for abbreviation recovery

To resolve the abbreviation recovery problem, one can identify some candidate full forms for each suspect abbreviation, and then enumerate all possible sequences of such candidates. The most probable root sequence is then confirmed by consulting local context. Such a recovery process can be easily mapped to an HMM model (Rabiner and Juang 1993), which is good at finding the best *unseen* state sequence; the input characters can simply be regarded as the "observation sequence", and the underlying word candidates as the unseen "state sequence". The abbreviation recovery process can thus be integrated into the word segmentation model by regarding the segmentation process as finding the best underlying words $w_1^m \equiv w_1, \ldots, w_m$, given the input characters $c_1^n \equiv c_1, \ldots, c_n \quad \equiv \vec{c}_1, \ldots, \vec{c}_m$. The segmentation process is then equivalent to finding the best *unabbreviated* word sequence $\vec{w}^*$such that:

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

where $\vec{c}_i$ refers to the surface form of w_i, which could be in the abbreviated or unabbreviated 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. The word-wise transition probability $P(w_i \mid w_{i-1})$ in the language model is used to impose contextual constraints over neighboring roots so that the underlying word sequence forms a highly probable sentence. In the absence of abbreviations, such that all surface forms are exactly the full forms, we will have $P(\vec{c}_i|w_i) = 1$. Equation (1) will then simply reduce to a word bigram model for word segmentation (Chiang et al. 1992, 1996). In the presence of abbreviations, however, the generation probability $P(\vec{c}_i|w_i)$ will indicate the strength of the abbreviation pattern.

As an example, if $\vec{c}_i$ and $\vec{c}_{i+1}$ are 台 tai and 大 da, respectively, then their roots, w_i and w_{i+1}, could be 台灣 tai-wan 'Taiwan' plus 大學 da-xue 'University' or 台灣 tai-wan plus 大聯盟 da-lien-meng 'Major League'. In this case, the probability scores P(台|台灣) × P(大|大學) × P(大學|台灣) and P(台|台灣) × P(大|大聯盟) × P(大聯盟|台灣) will indicate how likely 台大 tai-da is an abbreviation, and which of the above two compounds is the more probable full form.

By applying the unified and abbreviation-enhanced word segmentation model to the underlying word lattice, some of the root candidates may be preferred and others be discarded. If the best w_i is not the same as $\vec{c}_i$ then an abbreviation pattern will be identified.

It is desirable to estimate the abbreviation probability using some simple yet useful features, in addition to the lexemes (i.e., the surface character sequences) of the roots and abbreviations. Some heuristics about Chinese abbreviations might suggest such features. For instance, most 4-character words are abbreviated as 2-character abbreviations. Abbreviating into words of other lengths is less probable. It is also known that many 4-character words are abbreviated by preserving the first and the third characters. This can be represented by a '1010' bit pattern, where the '1' or '0' means to preserve or delete the respective character. Therefore, a reasonable abbreviation model is to introduce the *length* and the *positional bit pattern* as additional features, resulting in the following abbreviation probability.

$$\begin{aligned} P(\vec{c}|w) &= P(c_1^m, bit, m|r_1^n, n) \\ &\equiv P(c_1^m|r_1^n) \times P(bit|n) \times P(m|n) \end{aligned} \tag{2}$$

where c_1^m are the characters in the abbreviation of length m, r_1^n are the characters in the full form of length n, and *bit* is the above-mentioned bit pattern associated with the abbreviation process.

3 The SCR (single character recovery) model

The unified abbreviation recovery model allows us to acquire any M-to-N abbreviation patterns if we have enough training data for the language and abbreviation models. For the specific task of mining atomic N-to-1 abbreviation

patterns, it can be greatly simplified if each character in the training corpus is assumed to be a probable abbreviation whose full form is to be recovered. In other words, the surface form $\vec{c}_i$ in Eq. (1) is reduced to a single character abbreviation, and thus the associated abbreviation pattern is an atomic one. The abbreviation recovery model based on this assumption will be referred to as the SCR (single character recovery) model.

To acquire the atomic abbreviation patterns, the following iterative training process can be applied. The root candidates for each single character are enumerated to form a word lattice as the first step. Each path of the lattice will represent an unabbreviated word sequence. The underlying word sequence that is most likely to produce the input character sequence, according to Eq. (1), will then be identified as the best word sequence. Once the best word sequence is identified, the model parameters are re-estimated. And the best word sequence is identified again. Such an iterative process is repeated until the best sequence does not change any more. Upon convergence, the corresponding <root, abbreviation> pairs will be extracted.

It is highly simplified to use this SCR model for conducting a general abbreviation enhanced word segmentation process since not all single characters are really abbreviations. However, the single character assumption might be useful for extracting roots of real single character abbreviations with high demand on recall rate. The reason is that unknown abbreviations will be segmented into single characters with most segmentation algorithms; furthermore, a real root will be extracted only when it has high transition probability against neighboring words in addition to high output probability to produce the input character. Failing to satisfy such contextual constraints, spurious roots will be suppressed automatically. The over-generative assumption may be harmful for the precision rate, but will cover most interesting atomic abbreviations, which might be more important for the current mining task.

The above unsupervised training process can be greatly simplified if a word-segmented corpus is available. This is exactly our situation. Under such circumstances, the *abbreviation probabilities* can be trained iteratively in an unsupervised manner, with the word *transition probabilities* estimated in a supervised manner from the segmented corpus.

Furthermore, given the segmented corpus, the initial candidate <root, abbreviation> pairs can be generated by assuming that all word-segmented tokens are potential roots for each of its single character constituents. For example, if 台灣 tai-wan is a word-segmented token, then the abbreviation pairs <台灣 tai-wan, 台 tai> and <台灣 tai-wan, 灣 wan> can be generated. To handle the case where an input character is not really an abbreviation, each single character is assumed to be its own abbreviation by default.

In addition, to estimate the initial abbreviation probabilities, each abbreviation pair is associated with the frequency count of the root in the word segmentation corpus. This means that each single character abbreviation candidate of a root word is equally weighted initially. The equal weighting strategy, however, may not be appropriate (Chang and Lai 2004). In fact, the character position and word length features, as mentioned in Eq. (2), may be helpful. 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 Chang and Lai (2004). Finally, before the initial probabilities are re-estimated, Good-Turning smoothing (Katz 1987) is applied to the raw frequency counts of the abbreviation patterns in order to smooth unseen patterns.

4 Experiments

To evaluate the SCR model, the Academia Sinica Word Segmentation Corpus, ASWSC-2001 (CKIP 2001), is adopted for parameter estimation and performance evaluation. Among 94 files of this balanced corpus, 83 of them (13,086 KB) are randomly selected as the training set and 11 of them (162 KB) as the test set. Several models using different features for estimating the abbreviation probabilities are investigated. Table 1 shows the main results of the various models (M1 ~ M3). A '1' for each model indicates that the feature in the first row is used in Eq. (2). In short, M1 uses the lexemes as the only feature. M2 adds the positional bit pattern feature for full forms of known lengths (n). M3 further considers the most likely length, m, of the abbreviation, given the length of the full form.

The performance is successively improved with more and more features. Overall, using all the lexeme, positional and length features achieves the best results. The iterative training process, outlined in the previous section, 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, which represent a large set of abbreviation patterns that had rarely noticed in the literature. Table A1 in the Appendix shows some examples of atomic abbreviations acquired from the training corpus. A more complete list can be found in (Teng 2006). Note that the acquired abbreviations are not limited to named entities as previous literatures might expect; a wide variety of word classes have actually been acquired. The examples here partially justify the possibility and usefulness to use the SCR model for acquiring atomic abbreviations and their full forms from a large corpus.

Since the numbers of patterns are large, a rough estimate on the acquisition accuracy rates is conducted by 100 random samples of the <root, abbreviation> pairs. The patterns are then examined subjectively by our team members to see if the full forms are correctly recovered. The best precision rate is estimated to be 50% for the test set, and 62% for the training set. It is hard to estimate the recall for the large corpus. Fortunately, the SCR model uses an over-generative assumption to enumerate potential roots for all characters; the recall is thus expected to be high. Therefore, the recall rate is not particularly interesting. As far as the cost for

Table 1 Accuracy of SCR Model using various features

Models	$P(c_1^m \mid r_1^n)$	P(bit|n)	P(m|n)	Training	Test
M1	1	0	0	30%	25%
M2	1	1	0	48%	38%
M3	1	1	1	62%	50%

compiling an abbreviation lexicon is concerned, the preliminary result is encouraging since an atomic abbreviation pattern can be acquired about every two entries.

Although the mining performance is not directly related to the performance of a word-segmentation system, which can be referred to (Chang and Lai 2004), it is worth mentioning that a large percentage of the segmentation error comes from the generation of spurious root candidates, resulting in the notorious *searching* errors. With the enhancement of the atomic abbreviation lexicon, the reduction of searching error can be well expected.

There are several sources of errors with the current model. Firstly, the word-bigram language model takes more responsibility when the lexemes are the only model feature and each character is weighted equally. Unfortunately, the word-bigram model is sensitive to data sparseness problem. As a result, the abbreviation probabilities might not be well estimated. M2 uses the positional feature to tell which character position is more likely to be retained. This extra feature thus improved M1 significantly by weighting different positions differently. However, this extra feature does not solve all problems. Some 3-character full forms will be incorrectly preferred than 2-character words since P(100|3) is about 3 times larger than P(01|2). This can be partially compensated by introducing the length feature, since P($m = 1|n = 3$) is about 3 times smaller than P($m = 1|n = 2$). M3 thus has the highest performance above all.

In addition to the above modeling and estimation errors, two major sources of searching errors are significant. Firstly, each single character in the corpus is assumed to be a possible abbreviation in the current SCR model. This assumption may result in the extraction of non-atomic abbreviations. On the other hand, each word-segmented token is assumed to be a candidate full form of each of its constituents. This may introduce extra candidates which actually do not have any abbreviated form (like 尼采 ni-tsai 'Nietzsche') or the abbreviated form cannot be derived directly from its surface string (such as 上海 'Shanghai' whose abbreviated form is 滬 hu). A rough estimation shows that personal names have the biggest share (22%) among those unabbreviatable words. Such searching errors can be partially resolved by heuristic filtering when generating the root/abbreviation candidates. A "generation by composition" filter (Teng 2006), for instance, greatly reduces the number of candidate patterns by 10-folds while achieving comparable performance. The training set performance is 67% and the test set performance is 47%.

5 Concluding remarks

In this work, we adapt Chang and Lai's (2004) unified word segmentation model for mining full forms of atomic abbreviations in a large Chinese character set. An iterative training process, based on an SCR model, is developed to acquire an abbreviation dictionary from large corpora. The acquisition accuracy of the proposed SCR model achieves 62% and 50% precision for the training set and the test set, respectively. For systems that need to handle unlimited multiple character abbreviations, the atomic abbreviation dictionary could be invaluable.

Acknowledgements This work is partially supported by the National Science Council (NSC), Taiwan, Republic of China (ROC), under contract NSC 93-2213-E-260-015.

Appendix

Table A1 Examples of atomic abbreviation patterns

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

References

Chang, J.-S., & Lai, Y.-T. (2004). A preliminary study on probabilistic models for Chinese abbreviations. *Proceedings of the Third SIGHAN workshop on Chinese language learning* (pp. 9–16). ACL-2004, Barcelona, Spain.

Chiang, T.-H., Chang, J.-S., Lin, M.-Y., & Su, K.-Y. (1992). Statistical models for word segmentation and unknown word resolution. *Proceedings of ROCLING-V* (pp. 123–146). Taipei, Taiwan, ROC.

Chiang, T.-H., Chang, J.-S., Lin, M.-Y., & Su, K.-Y. (1996). Statistical word segmentation. In C.-R. Huang, K.-J. Chen, & B. K. T'sou (Eds.), *Journal of Chinese Linguistics*, Monograph Series No. 9, Readings in Chinese Natural Language Processing (pp. 147–173). University of California, Berkeley.

CKIP (2001). Academia Sinica word segmentation corpus, ASWSC-2001, (*中研院中文分詞語料庫*). Chinese Knowledge Information Processing Group, Academia Sinica, Taipei, Taiwan, ROC. From http://www.aclclp.org.tw/.

Huang, C.-R., Ahrens, K., & Chen, K.-J. (1998). A data-driven approach to the mental lexicon: Two studies on Chinese corpus linguistics. *Bulletin of the Institute of History and Philology, 69*(1), 151–179.

Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Trans. ASSP-35*(3).

Lai, Y.-T. (2003). *A probabilistic model for Chinese abbreviations*. Master Thesis, CS&IE, National Chi-Nan University, Taiwan, ROC.

Lin, M.-Y., Chiang, T.-H. & Su, K.-Y. (1993). A preliminary study on unknown word problem in Chinese word segmentation. *Proceedings of ROCLING VI* (pp. 119–142).

Rabiner, L., & Juang, B.-H. (1993). *Fundamentals of speech recognition*, Prentice-Hall.

Teng, W.-L. (2006). *Automatic models for mining atomic Chinese abbreviations*. Master Thesis, CS&IE, National Chi-Nan University, Taiwan, ROC.