A Maximum Entropy (ME) Based Translation Model for Chinese Characters Conversion

Fai Wong, Sam Chao, Cheong Cheong Hao and Ka Seng Leong

Faculty of Science and Technology of University of Macau,
Av. Padre Tomás Pereira S.J., Taipa, Macao
{derekfw, lidiasc}@umac.mo

Abstract. As the growth of exchange activities between four regions of cross strait, the problem to correctly convert between Traditional Chinese (TC) and Simplified Chinese (SC) is getting important and attention from many people, especially in business organizations and translation companies. Different from the approaches of many conventional code conversion systems, which rely on various levels of human constructed knowledge (from character set to semantic level) to facilitate the translation purpose, this paper proposes a Chinese conversion model based on Maximum Entropy (ME), a Machine Learning (ML) technique. This approach uses tagged corpus as the only information source for creating the conversion model. The constructed model is evaluated with selected ambiguous characters to investigate the recall rate as well as the conversion accuracy. The experiment results show that the proposed model is comparable to the state of the art conversion system.

Keywords: Maximum Entropy, Machine Learning, Natural Language Processing, Chinese translation, Traditional Chinese, Simplified Chinese.

1 Introduction

Modern Chinese typically involves two main dialects of writing, Traditional Chinese (TC) and Simplified Chinese (SC). In Chinese computing, these two systems adapt different coding schema for the computer to process the corresponding Chinese information. Traditional Chinese uses Big5 encoding while Simplified Chinese uses GB. For a Simplified Chinese document to be opened and read in a computer with Traditional Chinese operating system, conversion from Simplified Chinese encoding system into Traditional Chinese encoding is necessary for the purpose that the document can be further processed under the Traditional Chinese computer environment, and vice versa. As addressed by Wang [1] in the meeting of the 4th Chinese Digitization Forum, although there are many conversion systems implemented and available in the market, neither one of them can produce the conversion result with satisfaction. Reviewing the nature of this problem, Simplified Chinese is actually a simpler version of Traditional Chinese. It differs in two ways from the Traditional writing system: 1) a reduction of the number of strokes per character and 2) the reduction of the number of characters in use that is two different

© A. Gelbukh (Ed.)
Advances in Computational Linguistics.
Research in Computing Science 41, 2009, pp. 267-276

Received 06/11/08
Accepted 11/12/08
Final version 02/02/09

characters (of TC) are now written with the same character (in SC). The relationship between these two writing systems is not one-to-one mapping. In numerous situations, one simplified character corresponds to two or more traditional forms, e.g. simplified “发” maps to traditional “發(emit)” and “髮(hair)”. Normally only one of these is the correct one depending on the context. In some cases, one simplified character may map to multiple traditional forms, e.g. the simplified character “表” of the fragment “ 有表” may map to traditional “表(form)” and “錶(watch)” and any of which may be correct according to context. There are hundreds of simplified characters which correspond to two or more traditional ones, leading to ambiguity and this is the main obstacle of the conversion task for Simplified Chinese to Traditional Chinese translation.

The conventional techniques used to automatically translate Simplified Chinese to Traditional Chinese can be classified into three different approaches [2]: code conversion, orthographic (dictionary) conversion and lexemic conversion. Code conversion is also known as character based substitution, where the code of one character set is being substituted with a target code of another character set based on mapping table between the GB and Big5 encoding systems. This straightforward conversion methodology produces most unreliable result since the mapping table translates each simplified character to one target traditional character only and ignores the other possible candidates, this frequently results in incorrect conversion. The orthographic approach does the conversion based on larger unit of compound characters instead of single character by looking up the unit from a mapping table (simplified - traditional lexicon). The unit can be a meaningful character or combination of characters, and even idiomatic phrases. This method relies on a sophisticated Chinese word segmenter [3] that identifies the boundaries of words from the stream of text before the conversion of correspondences between simplified and traditional units taken place. The conversion system developed by Xing et al. [4] is based on this paradigm. The third approach is based on lexemic conversion. This kind of conversion systems actually covers the conversion processes of orthographic and code conversions, and in addition, the system also takes the deviations of terminologies and words used for the same concept into consideration during the conversion process, e.g. in Simplified Chinese, the word computer is written as “计算机”, while in Traditional Chinese, it is written as “電腦”. The systems reported by Halpern et al. [2] and Xing et al. [5] are based on this conversion methodology, including the conversion tool provided in Microsoft Word [6].

However, these approaches suffer from several limitations: 1) they highly rely on human constructed knowledge from lexicon to semantic level in order to achieve high conversion accuracy. The creation of these kinds of knowledge is too labor-intensive and time-consuming. 2) Consistency of knowledge formulated in rule is difficult to maintain and sometimes could contradict with each other and thus, affect the overall system performance. In this work, we formulate the Chinese conversion as a sequential tagging problem and use a supervised machine learning (ML) technique, Maximum Entropy (ME), to construct a Chinese conversion system. The ME model is a kind of feature-based model which is flexible to include arbitrary features to help in selecting the correct correspondence for simplified character during the conversion. The major features of this model are the tags and context words from a sentence.

This paper is organized as follows. Section 2 presents the general model of Maximum Entropy. Section 3 discusses the modeling of Chinese conversion problem, and the formulation of features for constructing the ME-based conversion model will be discussed in Section 4. The experiments based on the real text collected from newspapers will be discussed in Section 5 and Section 6, followed by a conclusion to end this paper.

2 Maximum Entropy Modeling

Maximum Entropy was first presented by Jaynes and has been applied successfully in many natural language processing (NLP) tasks[7], such as Part-of-Speech (POS) tagging [8], word sense disambiguation [9], and Chinese word segmentation [3]. ME model is a feature-based probabilistic model which bases on history and is able to flexibly use arbitrary number of context features (unigram, bigram word features and tag features) to the classification process that other generative models like N-gram model, Hidden Markov Model (HMM) cannot. The model is defined over $X \times Y$, where X is the set of possible histories and Y is the set of allowable outcomes or classes for the token or character in our case of Chinese conversion problem. The conditional probability of the model of a history x and a class y is defined as:

$$p_\lambda(y \mid x) = \frac{\prod_i \lambda_i^{f_i(x,y)}}{Z_\lambda(x)} \quad \text{.} \tag{1}$$

$$Z_\lambda(x) = \sum_y \prod_i \lambda_i^{f_i(x,y)} \quad \text{.} \tag{2}$$

where λ is a parameter which acts as a weight for the feature in the particular history. The equation (1) states that the conditional probability of the class given the history is the product of the weightings of all features which are active under the consideration of (x, y) pair, normalized over the sum of the products of the weightings of all the classes given the history x as the equation (2) above. The normalization constant is determined by requiring that $\sum_y P_\lambda(y \mid x) = 1$ for all x.

In ME model, the useful information to predict the outcome y by the equation (1) based on history features is represented by binary feature functions $f()$. Given a set of features and a training corpus, the ME estimation process produces a model which allows us to compute the conditional probability of equation (1). This actually is the process to seek for the optimized set of weighting parameters λ that is associated with the features. In other words, the process is to maximize the likelihood of the training data using p:

$$L(p)=\prod_{i=1}^{n}p_{\lambda}(x_i,y_i)=\prod_{i=1}^{n}\frac{1}{Z_{\lambda}(x)}\prod_{j=1}^{m}\lambda_j^{f_j(x_i,y_i)} \quad . \tag{3}$$

A number of models can be qualified from Equation (3). But according to the ME principle, the target is to generate a model *p* with the maximum conditional entropy *H*(*p*):

$$H(p)=-\sum_{x\in X,y\in Y}p(x,y)\log p(x,y) \text{ where } 0 <= H(p) <= \log |y| \text{ .} \tag{4}$$

3 Chinese Conversion as Tagging Problem

To model the Chinese conversion as a tagging problem, a manually tagged corpus with mapping relationships between simplified character and traditional character is required for training the conversion model based on the Maximum Entropy framework. In this work, we treat each character as a token, and it is assigned with a label sequence number, which represents the corresponding character in Traditional Chinese. For example, the simplified character "发" may map to "發(emit)" and "髮(hair)" in traditional forms. Thus in the labeled format, each ambiguous simplified character is assigned a number representing the mapping character in traditional one, as shown in Fig. 1. In the sentence, there are three ambiguous characters "发$_1$", "发$_2$" and "脏", and their corresponding traditional characters are "發(emit)", "髮(hair)" and "髒(dirty)", and are represented by the sequence number, "/1", "/2" and "/2" for each character, while the other unambiguous characters, including the punctuation marks, is assigned with "/0". The sequence number starts from 1 for each ambiguous character, until *n*, the possible number of candidates in the traditional forms. **Table 1** gives some exampled simplified characters and its correspondences in traditional form together with sequence number.

发/1 ！/0 你/0 的/0 头/0 发/2 有/0 点/0 脏/2 。/0

(Fat! Your hair is dirty.)

Fig. 1. The format of labeled sentence.

Based on tagged corpus, context information and features are collected to encode the useful information for the tagging process. In the model trained with suitable context and features, given a simplified sentence, it is able to predict each character with sequence number as the possible outcome from the tag set.

Table 1. Example of simplified characters with its possible corresponding traditional forms and sequences defined in our model.

板 → (1)板, (2)闆	参 → (1)參, (2)蔘
辟 → (1)辟, (2)闢	尝 → (1)嘗, (2)甞, (3)嚐
表 → (1)表, (2)錶	厂 → (1)厂, (2)庵, (3)厰, (4)廠
别 → (1)别, (2)彆	冲 → (1)沖, (2)衝
并 → (1)并, (2)並, (3)併, (4)竝	虫 → (1)虫, (2)蟲
卜 → (1) 卜, (2)蔔	丑 → (1)丑, (2)醜
布 → (1)布, (2)佈	仇 → (1)仇, (2)讎
才 → (1)才, (2)纔	出 → (1)出, (2)齣
采 → (1)采, (2)埰, (3)寀, (4)採	呆 → (1)呆, (2)獃
彩 → (1)彩, (2)綵	当 → (1)當, (2)噹

4 Feature Description

An important issue in the implementation of Maximum Entropy framework is the form of the function which calculates each feature. These functions are defined in the training phase and depend upon the data in the corpus. The function takes the form of Equation (5) as shown below, which is a binary-valued function:

$$f(x,y)=\begin{cases}1 & \mathit{if}\ y'=y \quad \mathit{and} \quad \mathit{info}(x)=v \\ 0 & \qquad\qquad \mathit{otherwise}\end{cases} \quad . \tag{5}$$

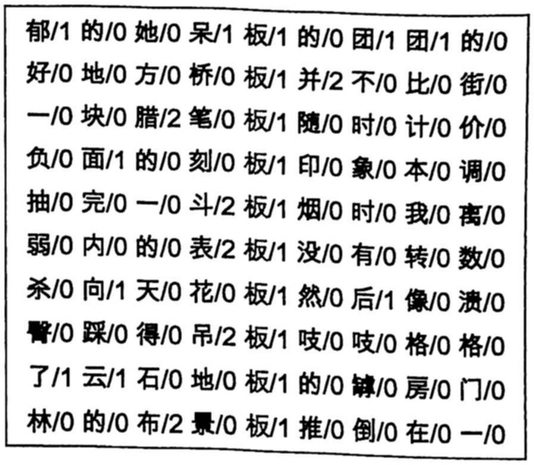
Where *info*(x) would be substituted with different expressions, and is referring as feature template in our work, which focuses on specific interested property that can be found from the context x, and v is a predefined value. For example, if we consider that 0 is the position of the active character, say "发" from the context "你的头发有点脏 (Fat! Your hair is dirty.)", to be learned and that i is related to 0, then the previous character of it is "头(head)" given by expression *PrevChr*(x,-1)="头". The set of features defined for the training of the conversion system mainly focus on characters, and collocations in the local context. In this work, two feature templates are adapted: C_i (i = -2 to 2), and C_iC_{i+1} (i = -2 to 1). Here C_0 represents the current character; C_i/C_{-i} represents the character which is at the i^{th} position to the right/left of C_0. These templates are basically character based features. They capture the contexts of surrounding information regarding the current character, including the form of character itself, which is also considered to the construction for the conversion model. Actually, each template groups several sets of features. Take the character sequence " 你的头发有点脏" as an example, features that will be generated by Equation (5) based on the first template are: C_{-2} = “的”, C_{-1} = “头”, C_0 = “发”, C_1 = “有” and C_2 = “点”. On the other hand, features obtained based on the second template consist of: $C_{-2}C_{-1}$ = “的头”, $C_{-1}C_0$ = “头发”, C_0C_1 = “发有”, C_1C_2 = “有点”. Therefore, for each context, there will be 10 different features in total obtained and used in the training the model based on the data in the corpus.

5 Data Preparation

In order to evaluate the proposed model, we need a corpus for constructing the model, especially with tagged information. This step involves the preparation of training data and test data. Since there is no any corpus intended for the purpose of Chinese conversion from Simplified Chinese to Traditional Chinese, we prepare these data by ourselves for this evaluation purpose. In this work, both the training and test data are created based on ambiguous simplified characters and their correspondences of traditional characters. For each traditional character that corresponds to an ambiguous character in its simplified format, we collect the related sample fragments of sentences from the online corpus of *Chinese Character Frequency Statistics*[10]. The corpus covers the articles from mainland China, Taiwan and Hong Kong, of different time frames from 60’s to 90’s. Basically, we can obtain enough data for majority of the ambiguous characters. For the other characters that are “ancient” or infrequently uses, we try to search from both the Internet and dictionaries. The idea is to collect enough text or examples for all characters. Figure 2 shows the sample of collected sentence fragments for the traditional character “板(plank)” that forms the training corpus to be used for constructing the translation system.

郁的她呆板的团团的
好地方桥板并不比街
一块腊笔板随时计价
负面的刻板印象本调
抽完一斗板烟时我离
弱内的表板没有转数
杀向天花板然后像溃
臂踩得吊板吱吱格格
了云石地板的睡房门
林的布景板推倒在一

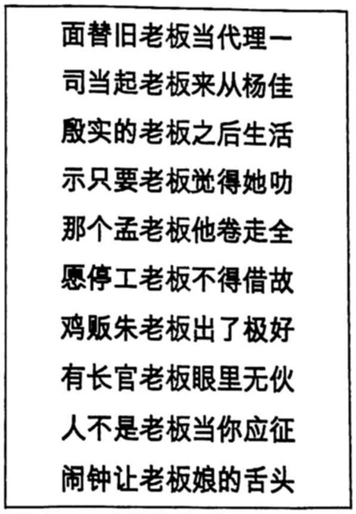
Fig. 2. The fragments of sentences containing the traditional character “板(plank)”.



郁/1 的/0 她/0 呆/1 板/1 的/0 团/1 团/1 的/0
好/0 地/0 方/0 桥/0 板/1 并/2 不/0 比/0 街/0
一/0 块/0 腊/2 笔/0 板/1 随/0 时/0 计/0 价/0
负/0 面/1 的/0 刻/0 板/1 印/0 象/0 本/0 调/0
抽/0 完/0 一/0 斗/2 板/1 烟/0 时/0 我/0 离/0
弱/0 内/0 的/0 表/2 板/1 没/0 有/0 转/0 数/0
杀/0 向/1 天/0 花/0 板/1 然/0 后/1 像/0 溃/0
臂/0 踩/0 得/0 吊/2 板/1 吱/0 吱/0 格/0 格/0
了/1 云/1 石/0 地/0 板/1 的/0 铺/0 房/0 门/0
林/0 的/0 布/2 景/0 板/1 推/0 倒/0 在/0 一/0

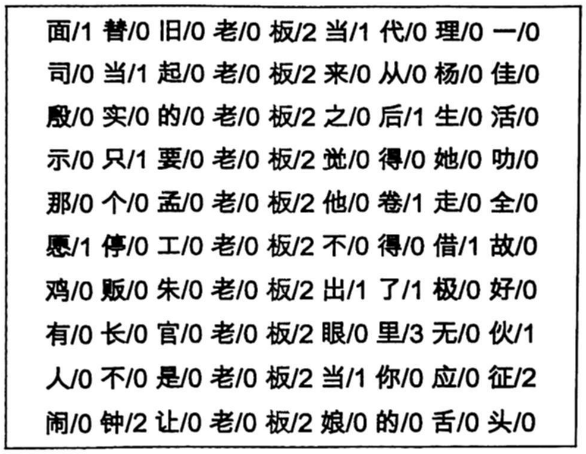
Fig. 3. The processed fragments for the character "板(plank)" after adding related tag information to the characters.

The next step is to convert the corpus by adding related tag information that is the corresponding sequence number to each character as described in Section 3, shown in Fig. 3. From the sample fragments, the unambiguous characters are labeled with "/0", while the ambiguous ones are marked with sequences number representing to its character correspondence in the traditional format. Fig. 4 and Fig. 5 present the set of collected and processed data fragments for another possible translation of simplified character "板" in its traditional form "闆(boss)". In this case, the character is marked with "/2" instead of "/1" as in previous case.



面替旧老板当代理一
司当起老板来从杨佳
殷实的老板之后生活
示只要老板觉得她叻
那个孟老板他卷走全
愿停工老板不得借故
鸡贩朱老板出了极好
有长官老板眼里无伙
人不是老板当你应征
闹钟让老板娘的舌头

Fig. 4. The fragments of sentences for the traditional character "闆(boss)".



面/1 替/0 旧/0 老/0 板/2 当/1 代/0 理/0 一/0
司/0 当/1 起/0 老/0 板/2 来/0 从/0 杨/0 佳/0
殿/0 实/0 的/0 老/0 板/2 之/0 后/1 生/0 活/0
示/0 只/1 要/0 老/0 板/2 觉/0 得/0 她/0 叻/0
那/0 个/0 孟/0 老/0 板/2 他/0 卷/1 走/0 全/0
愿/1 停/0 工/0 老/0 板/2 不/0 得/0 借/1 故/0
鸡/0 贩/0 朱/0 老/0 板/2 出/1 了/1 极/0 好/0
有/0 长/0 官/0 老/0 板/2 眼/0 里/3 无/0 伙/1
人/0 不/0 是/0 老/0 板/2 当/1 你/0 应/0 征/2
闹/0 钟/2 让/0 老/0 板/2 娘/0 的/0 舌/0 头/0

Fig. 5. The fragments for the character "■(boss)" after processed.

Table 2 gives the size of the data set used for training the model. Actually, there are more than 300 interested characters that may cause ambiguity during the translation between traditional and simplified forms. Therefore, the collection of data is based on this set of characters. For each of these characters, a number of sentences are gathered to train up a conversion model for the disambiguation purpose when a simplified character is going to be converted into the traditional form.

Table 2. Size of training corpus

	Characters	Ratio
Size	919215	92.29%
Ambiguous Characters	70839	7.71%

For the test data, sentences are collected from several online Chinese newspapers of *Jornal Cheng Pou* (Cheng Pou Journal), *Jornal Cidadão* (Citizen Journal), *Jornal Informação* (Information Journal), *Jornal San Wa Ou* (San Wa Ou Journal), *Jornal Tai Chung* (Tai Chung Journal) and *Macau Daily News* (Macao Daily News) between 8th April 2008 and 8th August 2008. There are around 3,027 sentences in total. The data covers most of interested ambiguous characters. The relative data set size is presented in Table 3. This includes the count of all characters, as well as the ambiguous characters for testing.

Table 3. Size of test corpus

	Characters	Ratio
Size	862586	98.05%
Ambiguous Characters	16795	1.95%

6 Model Evaluation

Two experiments are carried out to investigate the recall rate and the conversion accuracy of the model. In both cases, only the counts of ambiguous characters are used for calculating the recall and precision, and excluding out the counts of unambiguous characters. Otherwise, the system will always obtain very high conversion accuracy, since the percentage of unambiguous characters is much higher than that of the ambiguous ones, as illustrated in Table 2 and Table 3 for different corpora.

The first experiment evaluates the recall rate. The model is trained and tested by using the training data as presented in Table 2. That is, the same data set is used to evaluate the performance of the model. The conversion accuracy (recall rate) is 99.84%.

In the second experiment, we construct the model based on the training data set and use another data set (test data) to evaluate the model's conversion precision. The accuracy of the conversion results reaches 89.94%. In order to give an idea of our model's performance, we use the tool provided by Microsoft Word to do the conversion for the same set of test data. The accuracy of the conversion result is 87.86%. This illustrates that our proposed model is comparable to systems based on other conversion methodologies.

7 Conclusion

Most of the code translation or conversion tools developed to handle the Chinese conversion problem are simply based on mapping table of code or lexicon. More sophisticated conversion systems adapt the deep analysis approach, where various Chinese analysis systems are used as the preprocessing steps, such as the segmentation of word, labeling of syntax categories (part of speech), even the syntactic and semantic analyzers of sentence. However, robust system is not always available, especially for different analytic systems. Moreover, the management of the ambiguities in these language analyzers has to tackle the combination of overall ambiguities. In this paper, a statistic approach based on Maximum Entropy model is proposed to construct a Chinese translation system for the conversion of characters between traditional and simplified forms. Similar to other Natural Language Processing tasks, the Chinese to Chinese conversion processing is transformed into a labeling problem. Experiments were performed to evaluate the performance of the constructed model in terms of recall rate and the conversion accuracy. The empirical results show that the proposed model is comparable to the conversion system provided by the MS Word.

References

1. Wang, N.: The Principle of Building Parallel Term Corpus for Simplified Chinese Characters Conversion. In: Proceedings of The 4th Chinese Digitization Forum (CDF) Macao, SAR, China (2007)
2. Halpern, J., Kerman, J.: The Pitfalls and Complexities of Chinese to Chinese Conversion. In: Proceedings of Fourteenth International Unicode Conference. Cambridge, Massachusetts (1999)
3. Leong, K.S., Wong, F., Li, Y.P., Dong, M.C.: Integration of Named Entity Information for Chinese Word Segmentation Based on Maximum Entropy. Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues, vol. 5226, pp. 962--969: Springer Berlin / Heidelberg, Shanghai, China, (2008)
4. Xin, C.S., Sun, Y.F.: Simplified-Unsimplified Chinese Conversion and Word Segmentaion. Mini-Micro System, vol. 21, no. 9, pp. 982--985 (2000)
5. Xin, C.S., Sun, Y.F.: Design and Implementation of a Simplified-Unsimplified Chinese Character Conversion System. Journal of Software, vol. 11, no. 11, pp. 1534--1540 (2000)
6. Wu, A.: Chinese Word Segmentation in MSR-NLP. In: Proceedings of The second SIGHAN workshop on Chinese language processing, pp. 172--175. Sapporo, Japan (2003)
7. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A Maximum Entropy Approach to Natural Language Processing. Computational Linguistics, vol. 22, no. 1, pp. 39--71 (1996)
8. Ratnaparkhi, A.: A maximum entropy model for part-of-speech tagging. In: Proceedings of Empirical Methods in Natural Language Processing (EMNLP), pp. 133--142. Association for Computational Linguistics, New Brunswick, New Jersey (1996)
9. Suárez, A., Palomar, M.: A Maximum Entropy-based Word Sense Disambiguation system. In: Proceedings of the 19th International Conference on Computational Linguistics, pp. 960--966. Taipei, Taiwan (2002)
10. "Hong Kong, Mainland China & Taiwan: Chinese Character Frequency." Chinese University of Hong Kong, http://humanum.arts.cuhk.edu.hk/Lexis/chifreq/.