Chinese Conversion Based on Statistic Model

Fai Wong, Mingchui Dong, Kaseng Leong, Haocheong Cheong
Faculty of Science and Technology, University of Macao
Av. Padre Tomás Pereira, Taipa, Macau, China
{derek, dmc}@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 business organizations and translation companies. Different from the approaches of many conventional code conversion systems, which relies 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 conversion system provided by MS Word.

1. Introduction

In Chinese computing, Traditional and Simplified Chinese 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. As addressed by Wang [1] in the meeting of 4th Chinese Digitization Forum, although there are many conversion systems are implemented and can be found from the market, neither one of them can produce the conversion result with satisfaction. Reviewing the nature of this problem, Simplified Chinese actually is simply a simpler version of Traditional Chinese. The number of simplified characters in use is reduced, that is two different 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. 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 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 ambiguous, and this is the main obstacle of the conversion task for Simplified Chinese to Traditional Chinese translation.

The conventional techniques used to automatically convert 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. 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). This method relies on a sophisticated Chinese word segmenter [3] to identify 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 system 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 by MS Word [6].

However, these approaches suffer from several limitations: 1) it highly relies on human constructed knowledge from lexicon level 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 the 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 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, 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:

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

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

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 $\sum_y p_\lambda(y|x) = 1$ 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 seeking for the optimized set of weighting parameters λ that associated with the features. That 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 $H(p)$:

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

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 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" and "髮" 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 Figure 1. In the sentence, there are three ambiguous characters "发 $_1$", "发 $_2$" and "脏", and their corresponding traditional characters are "發", "髮" and "髒", 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

Figure 1 Format of labeled sentence for simplified characters

Based on tagged corpus, context information and features are collected to encode the useful information for the tagging process. With the model trained with suitable context and features, given a simplified sentence, the model 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 work

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)廠, (4)廠
别 → (1)別, (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:

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

Where *info*(*x*) would be substituted with different expression, and is referring as feature temple 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", to be learned and that *i* is related to 0, then the previous character of it is "头" 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 is also considered to the construction of the conversion model. Actually, each template groups several sets of features. Take the character sequence "你的头发有点脏" as 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". While features obtained based on the second template consist: $C_{-2}C_{-1}$ = "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 to the training the model based on the data in the corpus.

5. Model Evaluation

This section presents the results of the evaluation over the training and test data. This involves the preparation of training data and test data for the model. 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 set of selected ambiguous simplified characters and their correspondences of traditional characters, due to fact that, there are several hundreds of corresponding traditional characters for the ambiguous simplified characters, including both frequently and infrequently use characters. For those of infrequent characters, the data is not enough to cover every single character for testing the model. Thus, we select 30 characters of simplified form and their corresponding 68 characters in traditional form for evaluating the model. For each traditional character, we collect the related sample fragments of sentences from the online corpus of *Chinese Character Frequency* S*tatistics* [10]. Figure 2 shows the sample of collected sentence fragments that form the corpus to be used for constructing the conversion system. The next step is to convert the corpus by adding tag

information that is the corresponding sequence number to each character as described in Section 3.

铺上胶地板的混凝土
撞向天花板男子被爆
贯的树样板和树典型
中共是铁板一块不观
昏迷在地板上紧急送
的餐馆老板说中国餐
懂摊大手板毫不脸红
章多用慢板奏出必有
一块腊笔板随时计价

Figure 2 The fragments of sentences containing the traditional character "板"

For the test data, sentences are collected from several online newspapers of *Jornal Cheng Pou* (正報), *Jornal Cidadão* (市民日報), *Jornal Informação* (訊報), *Jornal San Wa Ou* (新華澳報), *Jornal Tai Chung* (大眾報) and *Macau Daily News* (澳門日報) between 8th April 2008 and 8th August 2008. The data covers all the 68 traditional characters. The relative data set sizes are presented in Table 2. This includes the count of all characters, as well as the interested ambiguous characters for testing.

Table 2 Sizes of used corpora

	Training Data		**Test Data**	
Size (Characters)	919215	92.29%	862586	98.05%
Ambiguous Characters	70839	7.71%	16795	1.95%
Sentences	-		3027	

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 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 test 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 conversion 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 conversion model is comparable to conversion system based on other conversion methodologies.

6. Conclusion

In this paper, a statistic approach based on Maximum Entropy model to construct a Chinese conversion system is proposed. Similar to other Natural Language Processing tasks, the Chinese to Chinese conversion processing is treated as tagging 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 MS Word.

References

[1] 王寧, "基於簡繁漢字轉換的平行詞語庫建設原則", *第四届兩岸四地中文數字化研討會(4th CDF)* 澳門, 24-26 January, 2007.

[2] J. Halpern, and J. Kerman, "The Pitfalls and Complexities of Chinese to Chinese Conversion", *Proceedings of Fourteenth International Unicode Conference*, Cambridge, Massachusetts, 1999.

[3] K.S. Leong, F. Wong, Y.P. Li, and M.C. Dong. "Integration of Named Entity Information for Chinese Word Segmentation Based on Maximum Entropy." In *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues*, 962-969. Shanghai, China: Springer Berlin / Heidelberg, 2008.

[4] 辛春生, 孫玉芳, "漢語簡繁體轉換與語詞切分", *小型微型計算機系統*, Vol. 21, No. 9, 2000, pp. 982-985.

[5] 辛春生, 孫玉芳, "簡繁漢字轉換系統的設計與實現", *軟件學報*, Vol. 11, No. 11, 2000, pp. 1534-1540.

[6] A. Wu, "Chinese Word Segmentation in MSR-NLP", *Proceedings of The second SIGHAN workshop on Chinese language processing*, Sapporo, Japan, 11-12 July, 2003, pp. 172-175.

[7] A.L. Berger, V.J.D. Pietra, and S.A.D. Pietra, "A Maximum Entropy Approach to Natural Language Processing", *Computational Linguistics*, Vol. 22, No. 1, 1996, pp. 39-71.

[8] A. Ratnaparkhi, "A maximum entropy model for part-of-speech tagging", *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, New Brunswick, New Jersey, 1996, pp. 133-142.

[9] A. Suárez, and M. Palomar, "A Maximum Entropy-based Word Sense Disambiguation system", *Proceedings of The 19th International Conference on Computational Linguistics*, Taipei, Taiwan, 24 August - 1 September, 2002, pp. 960–966.

[10] "Hong Kong, Mainland China & Taiwan: Chinese Character Frequency." Chinese University of Hong Kong, http://humanum.arts.cuhk.edu.hk/Lexis/chifreq/.