Key Problems in Conversion from Simplified to Traditional Chinese Characters

Xiaodong Shi, Yidong Chen
Department of Cognitive Science
Xiamen University
Xiamen 361005, P. R. China
{mandel,ydchen}@xmu.edu.cn

Xiuping Huang
Xiamen Local Taxation Bureau
Xiamen, Fujian
Xiamen 361012, P. R. China
cpp.huang@gmail.com

Abstract

In this paper we tackle the problem of character conversion from simplified Chinese to traditional Chinese. Of those simplified characters that need conversion, about 9.5% of them have more than 2 counterparts in the traditional scripts. We improve upon the previous log-linear approach first used in (Chen et al 2011) by utilizing more data sets and better translation models. We also show that automatic classification and noise reduction of corpus can achieve better performance. As a proof of the validity of our approach, we scored No. 1 in a recent evaluation of simplified to traditional character conversion systems organized by the Chinese Information Processing Society of China.

1 Introduction

People in mainland China and Singapore typically use simplified Chinese characters and those of Taiwan, Hong Kong, and Macao typically use traditional Chinese characters (henceforth we use the word traditional to mean traditional Chinese scripts). Although simplified Chinese characters come from traditional ones, their correspondences are not one-to-one. So conversion between the two Chinese scripts is necessary if Taiwanese need to read the Chinese newspapers or ordinary mainland Chinese need to read the Taiwanese newspapers and ancient Chinese books written in traditional scripts. Although conversion both ways are needed, the situation is more severe when converting from simplified Chinese characters to traditional ones (henceforth called s2t), as about 9.5% of simplified Chinese characters have more than 2 traditional counterparts (our statistics, papers differ in this respect). Although lots of commercial products (e.g. MS Word) and online web sites (e.g. Google Translate) offer s2t character conversion, their performance is still not satisfactory. In this paper we only focus on character conversion from s2t and do not consider word/term conversion as the former is a more fundamental problem.

Upon careful inspection, the s2t conversion can be classified into 3 types:

1) Conversion from modern simplified Chinese to modern traditional Chinese. The situation arises when a Taiwanese wants to read a mainland Chinese newspaper.

2) Conversion from non-modern simplified Chinese to non-modern traditional Chinese. The situation arises when one wants to recover the traditional Chinese texts from simplified Chinese e-texts because the latter are far easier to input into the computer. Depending on how we define non-modern, the tasks can be further classified into 2 subtypes:

2.1) Conversion from simplified classical Chinese into traditional one. The classical Chinese is formal written Chinese and very different from modern Chinese.

2.2) Conversion from pre-1949 informal simplified Chinese into traditional one. The pre-1949 informal Chinese (up to Song Dynasty) is a written form of oral Chinese more similar to modern Chinese than classical Chinese in syntax.

The classification is useful because the 3 traditional sublanguages have very different characteristics, as the following statistics collected from our corpus suggests (Table 1):

So if we want to do s2t conversion the first step is to do task classification. As far as we know, this step has never been mentioned in previous approaches to s2t conversion. One requirement is that we must have enough data to train separate sub-models and reliably classify test cases. If a sub-model is small we might have

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 287–293.

to borrow from similar sub-models to ameliorate the data sparseness problem.

Table 1 Statistics for 3 types of corpora

	Corpora		
	CNA News	Classical Chinese	Pre-1949 informal Chinese
Corpus size in characters	16,083,000	118,646,765	9,120,340
Top ten characters	的中國台人在日年大會 8.85345%	之不一以也而者人有為 10.74683%	了不一的人道是來他有 13.13563%
Unique characters	5747	19992	7272

In this paper we improve upon the previous log-linear approach in s2t conversion proposed by (Chen et al., 2011) by utilizing more data sets and building better translation models. Because the corpus we collected from the web is prone to noise (errors), we must also reduce the impact of erroneous data. In a recent s2t evaluation organized by Chinese Information Processing Society of China, our approach in this paper achieved better performance than competing systems.

The rest of this paper is organized as follows: Section 2 briefly reviews the related work on s2t conversion. Section 3 describes our corpus collection effort and especially on the evaluation data sets. In Section 4, we present our main contributions in detail: how we improve the overall conversion model by data classification and using a better translation model, also we show how to use the corpus when it contains lots of errors. Our experiments are also described and the results analyzed in this section. Finally in Section 5 we give the conclusion and propose some ideas for future work.

2 Related Work

The following review is not meant to be comprehensive, and our emphasis is on methods proposed.

(Wang and Wang, 2005) use MS Word 2003 to convert novels totaling about 1.5 million characters in Simplified Chinese to traditional and found many mistakes. They listed many examples of incorrect conversion. A few later works on s2t uses the data set in this paper.

Considering that characters surrounding the ambiguous character can help disambiguate it, (Xin et al., 2005) used a larger unit (e.g. words) to help conversion, e.g. while 台 (tai) can convert to 颱臺檯台 depending the context, 台风 (tai-feng "typhoon") can only convert to 颱風 (example from (Hao et al., 2013)). So to implement this approach, we need to do Chinese word segmentation first.

(Wong et al., 2007) proposed a machine learning approach - the maximum Entropy model - to treat the conversion problem as a classification task. Their results are comparable to the conversion system provided by MS Word.

(Li and Wu, 2010) first leveraged "language model" in s2t conversion besides Chinese word segmentation. The language model is a powerful device in statistical natural language processing and their system achieved the state-of-art performance (95.77% precision) for the (Wang et al 2005) data set.

(Chen et al., 2011) first used the popular log-linear model, which has been commonly used in statistical machine translation, in s2t conversion. The feature functions used in the log-linear framework include lexical semantics consistency weight (similar to translation model), language model, sentence length and phrase count. The latter two are used because term conversion is also performed. The results (97.03% precision) are better than that of (Li and Wu, 2010).

(Hao et al., 2013) proposed a new priority-based multi-data resources management model and a new Fused Conversion algorithm from Multi-Data resources which achieved 91.5% precision in the (Wang and Wang, 2005) data set and 99.8% overall precision in another document from the National Palace Museum website. Their model also uses pattern matching in conversion.

Our work follows that of (Chen et al., 2011) in that we use the same log-linear framework but a different feature function sets. The model can be succinctly described by the following formula:

$$\log p(\mathbf{t} \mid \mathbf{s}) = \alpha \sum_i \lambda_i \log h_i(\mathbf{s}, \mathbf{t}) \quad (1)$$

where **s** and **t** represent simplified and traditional Chinese sentences and h_i are features functions defined on them, and λ_i are weights and α is a normalization factor. Different feature functions define different models. In our s2t approach, we only used two features: a language model and a

translation model, both of which are described in Section 4.

3 Resources for s2t Conversion

We began our corpus collection effort when the Chinese Information Processing Society of China (CIPSC) announced earlier this year to evaluate s2t systems, and we found that the test data seemed to include both modern and classical text.

Obviously we need large amounts of text data. For conversion from modern Simplified Chinese to traditional, we need data mainly from Taiwan. For conversion from non-modern Simplified Chinese to traditional, we need more ancient corpus.

(Hao et al., 2013) used the Tagged Chinese Gigaword second edition which contains news-wire texts from Central News Agency (Taiwan), Xinhua News Agency and Lianhe Zaobao. The untagged version is LDC2005T14, and the tagged versions are LDC2007T03 and LDC2009T14. We used LDC2005T14 only in the experiments because there are lots of errors in LDC2007T03 or LDC2009T14. We found these errors because we initially tried to use the word segmented version. We list, in Figure 1, some of the errors for the character 发 (fa) which can be converted to 發髮 and should be 發 (deliver, develop) in the following examples but was written as 髮 (hair).

(Nb) 州(Ncb)	長髮表	(Na)	就職(VA4)	
暫停(VH16)	核髮	(Na)	香港(Nca)	
結束(VH16)	髮	(Nab)	言(VE2)	，
(Nbc) 秘書(Nab)	長髮	(Nab)	表(Na)	專
的(DE)	陳金髮	(Nb)	是(SHI)	合
(Dbb) 必須(Dbab)	研髮	(Na)	菜單(Nab)	
地產(Naeb)	開髮	(Na)	結構(Nad)	

Figure 1 Some errors of data from LDC2007T03

We suspected that the text was first converted from traditional Chinese into simplified and then word segmented using a simplified Chinese segmentor and then finally the segmented text was converted back to traditional. It's ironical that s2t conversion should have played a bad role in the quality of the tagged corpus. Had anyone used the Tagged Gigaword in training a language model for statistical machine translation, the performance would have been compromised. This fact also shows that s2t conversion is far from perfect. So we used the non tagged Gigaword version (LDC2005T14) in our experiments.

Although there are two large data sets for the non-modern Chinese: one is China Basic Classical Texts (中国基本古籍库) from mainland China which contained about 1.7 billion Chinese characters, and the other is Hanji Scripta Sinica database (漢籍全文資料庫) from Academia Sinica of Taiwan which contained about 0.44 billion Chinese characters, they are both not free and hard to get. So we build our own Traditional Chinese Corpus by using the e-texts of the Traditional Chinese portion of the URead e-library[1] (which contained about 0.28 billion Chinese characters from about 4,000 books) and by crawling the web which resulted in another traditional Chinese corpus (0.063 billion Chinese characters). We also have two big dictionaries (漢語大詞典, 漢語大字典) which are not used in the CIPSC evaluation but in some experiments described in this paper.

All the data we collected can be accessed freely[2] (totaling about 0.42 billion Chinese characters).

We noticed that there are many errors in the collected corpus. e.g. 后 (hou) can be converted to 后後 depending on context, e.g. 後來 (after) and 皇后 (queen). We searched the wrong combination 皇後 in our corpus and found more than 500 occurrences in the URead corpus and even 14 occurrences in the relatively high-quality CNA corpus. These errors would inevitably bear their marks in the trained language model if no measures are taken.

The most important resource of s2t conversion is the conversion table. How many characters need to be converted? How many of them have more than one candidate in conversion? Chinese linguists have long collected various lists and argued about their correctness. Table 2 lists the statistics for some of them which were compared in (Chen, 2009), from which we can find that even the number of the ambiguous pairs is not settled by the linguists.

As we have demonstrated above that s2t conversion is not a monolithic task, we think there is no unique conversion table. Depending on different subtasks, we may have to use different tables to get the best performance. Assume we have a large tasked-oriented traditional corpus, we can

[1] http://uread.superfection.com/

[2] http://corpus.superfection.com/corpus_tc.html

derive the conversion table in the following simple steps:

1) Convert the traditional corpus into a simplified one;

2) Discover the differences in characters and collect the s2t conversion table;

3) For those simplified Chinese characters in the conversion table, if they occur in the traditional corpus, add the same character to the conversion candidates.

Table 2 Statistics for some lists of ambiguous s2t character pairs

Author	**# of pairs**
Feng, 1997	117
Zhou, 2009	183
Su, 2004	133
Zhang, 2004	194
Lian, 2004	148
Yang, 2004	121
Li, 2005	247
Feng, 2007	121
Hu, 2007	123
Chinese Language Review Express, 2008	274
Guo and Ye, 2004	1065
Chen, 2009	195

The first step is simple as one-to-many conversions are few and so less error-prone. The last step is necessary to ensure correction conversion for many characters which exist in both simplified and traditional corpus. One drawback is that the conversion tables thus generated does not contain characters outside the traditional corpus used. So to make the list complete we may have to augment it. For example, the CNA portion of the Gigaword corpus second edition contains 9039 unique character types and we know that they were originally coded in BIG5 code (Huang, 2009) and there are 13,053 unique character types in this coding standard. So the remaining 4014 characters are not included in the conversion table, which may have simplified versions.

The s2t conversion table we used in the CIPSC evaluation is collected this way using the Traditional Chinese portion of the URead e-library and augmented by other lists collected by the Chinese linguists and Wikipedia and those offered by the CIPSC. The official Complete Table of Simplified Characters (简化字总表)[3] is also consulted but it is by no means complete and should not be used alone.

Another important data set is the evaluation data. As the review on previous work has shown, the (Wang and Wang, 2005) data set is very popular. However, the reference answers provided in (Li and Wu, 2010) are problematic. Since the set is modern Chinese, we used the CNA corpus (cna2013 we downloaded from the web) and found the following problems (Table 3):

Table 3 Problems for data from (Wang et al., 2005)

Word or character	The reference answer with occurrences in cna2013	Alternate answer with occurrences in cna2013
重　复 (repeat)	重復 0	重複 206
里(in)	裏 110	裡 3704

So by the Taiwanese standard the reference answers contain at least 25 errors, which is more than 9% of test cases. So unless the papers considered the alternate answers as also correct, their reported precision should be placed under further scrutiny. Besides, the test set is very small compared with the huge volume of the train set, we think it's at least inappropriate to use it as a test set.

On the other hand, the test set of CIPSC evaluation contains both ancient Chinese texts and modern ones. As we later learned, they came from examples in 漢語大字典. As is well-known, the dictionary data does not reflect the actual statistical usage of word and characters; we think it's necessary to create our own data sets. In order to test s2t in the two scenarios (modern/ancient), we use a portion of CNA (which is not in the train data) and a portion of Hanji Scripta (which we downloaded from the web) as our test sets. We converted the traditional test data to simplified and then converted them to the traditional and compare the differences and report the whole text precision. We also selected a few characters that are often the sources of errors and report their results.

4 Improvement of the s2t Conversion Model

We describe 3 aspects which lead to better performance in s2t conversion: data classification; a new translation model, and noise reduction. We also show experimental results.

[3] http://www.stlcls.org/s-words/Simplified_word.htm

4.1 Data Classification

As mentioned in Section 1, the s2t conversion systems for modern Chinese and non-modern Chinese would be quite different. Therefore, one may reasonably guess that a preprocessing step that doing the data classification will help improving the performance for the conversion system. In this section, experiments are conducted to test this assumption.

The statistics for the data that we used in the experiments are listed in Table 4.

Table 4 Statistics for the data used in these experiments

Name	Source	Amount
Training Set	Uread e-library	641M
Non-modern Test Set	Hanji Scripta Sinica database portion before Song Dynasty	0.07M
Modern Test Set	Hanji Scripta Sinica database portion after Song Dynasty	0.16M

We first classified the training set into two parts using a classification system based on the text categorization algorithm presented in Cavnar and Trenkle (1994), which was trained using a seed non-modern dataset (10.2MB) and a seed modern dataset (9.8MB) from URead e-library. After the classification, the training set was divided into two parts, i.e., the non-modern part (271M) and the modern part (370M). Then, four experiments were carried out and the results are listed in Table 5.

Table 5 Experimental results for the data classification

Test Set	Training Set	Accuracy (%)
Non-modern	Non-modern	**98.785**
Non-modern	Modern	98.369
Modern	Non-modern	98.153
Modern	Modern	**98.205**

From the experimental results above, we can learn that the data classification did help improve the performance.

4.2 A New Enhanced Translation Model

The s2t conversion can be also tackled from the machine learning approach: it's a classification task similar to Word Sense Disambiguation WSD). Lots of machine learning approaches have been proposed to solve this problem; the most popular ones are Naïve Bayes and Maximum Entropy. We choose the former because it's very simple to implement and yet efficient. Our contribution is two-fold in this aspect:

1) We treat the classification model as a translation model in a log-linear frame-work.

2) We choose features not covered by the language model, e.g. the previous and next characters are not considered as these two can be better handled by a trigram language model. In fact, we use pairwise mutual information[4] to select characters that appear in a certain window near the characters to be converted.

The enhanced translation model is used as a feature of formula (1) in this way:

$$h_{tm}(t,s) = \alpha P_{ml}(t \mid s) W(P_{nb}(t \mid s)) \quad (2)$$

where $P_{nb}(t \mid s)$ is a Naive Bayes classifier on whether the simplified character s is converted to a traditional character t. W is a weighting function which is greater than 1 if $P_{nb}(t \mid s)$ is greater than 0.5, and α is a normalization factor to make the feature a probability. Thus the classification is used to "boost" the probabilities of the *a priori* translation probability $P_{ml}(t \mid s)$, which is calculated by the relative frequency of possible candidates (maximum likelihood estimation).

In the following table we show the results of the various translation models: no translation model, *a priori* translation model, and the enhanced translation model on two data sets: modern and non-modern.

Table 6 Experimental results for various Translation Models (TM)

Models	Test Set	Accuracy (%)
Without TM	Modern	98.205
a priori TM	Modern	98.497
enhanced TM	Modern	**98.611**
Without TM	Non-modern	98.785
a priori TM	Non-modern	**98.950**
enhanced TM	Non-modern	98.935

From the experimental results above, we conclude that the translation model was useful in

[4] $pmi(x,y) = \log \frac{P(x,y)}{P(x)P(y)}$, here x, y are words

data classification and did help improving the performance.

For the non-modern part, the classification boosted TM worked worse. The possible reason is that long distance contexts are not helpful, since the ancient Chinese is less redundant.

Currently in the log-linear model the weights of the language model and the translation model are set manually and we think further tuning can enhance the accuracy further.

4.3 Noise Reduction of the Train Data

When we are experimenting with various data sets, we find all data contained errors more or less. So adding more data does not necessarily lead to higher performance. However, we need not to throw away data if we can utilize the reliable portion. We consider this problem in the context of the n-gram language model.

How can we tell if a bigram (which contains a traditional Chinese character) is correct? We have the following guidelines:

- The bigram is a head word in a reliable dictionary.
- The bigram appears in more than one sources and is very frequent

How can we tell if a bigram is incorrect? We also can examine the following points:

- The bigram is a not a head word in a reliable dictionary.
- The simplified form the bigram appears as a head word.
- The context of the traditional bigram is very different.

Initial experiments have confirmed that by identifying the incorrect bigrams and discount them we can have a better language model. The details on building automatic classifiers to identify the suspicious bigrams and experiments on how to discount these pairs are still in the progress and will be described in another paper.

We have made all our conversion tool publicly available and also offer free web based s2t conversion service at the following url: http://corpus.superfection.com/s2t.html.

5 Conclusions

The paper improves upon the previous log-linear approach to simplified-to-traditional Chinese character conversion by automatic text classification, better translation models and more data sets. The work on data noise reduction is still under way and we are confident it will further improve the performance by squeezing as much information as available from the data. In future we plan also use more linguistic knowledge (e.g. person name identification) and perform word segmentation to get even better results.

Acknowledgement The work described in this paper is supported by the Key Technologies R&D Program of China (Grant No. 2012BAH14F03), the Natural Science Foundation of China (Grant No. 61005052), the Fundamental Research Funds for the Central Universities (Grant No. 2010121068) and the Natural Science Foundation of Fujian Province of China (Grant No. 2011J01369).

References

Cavnar, W. B. and J. M. Trenkle. 1994. N-Gram-Based Text Categorization. *In: Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, NV, pp. 161-175.

Chen, Mingran. 2009. Research on Non One-to-One Simplified-Traditional Character Pairs. *In: Proceedings of 12th Symposium on Shu Tong Wen for Chinese*, Qinhungdao, China.

Chen, Yidong, Shi, Xiaodong, and Zhou, Changle. 2011. A Simplified-Traditional Chinese Character Conversion Model Based on Log-Linear Models. *In: Proceedings in International Conference on Asian Language Processing 2011*, Penang, Malaysia, pp. 3-6.

Hao, Tianyong and Zhu Chunshen. 2013. Toward a Professional Platform for Chinese Character Conversion. *ACM Transactions on Asian Language Information Processing*, 12(1): 1-22.

Huang, C. R. 2009. Tagged Chinese Gigaword Version 2.0. *Linguistic Data Consortium*, Philadelphia.

Li, Min-Hsiang, and Wu Shih-Hung. 2010. Chinese Characters Conversion System based on Lookup Table and Language Model, *In: Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing*, Taiwan, China.

Liu, H. D. and Wu, J. 2008. A multi-layer system of simplified-traditional Chinese character conversion based on word disambiguation. *In Proceedings of the 5th Chinese Digitization Forum (CDF'08)*. 156–167.

Wang, Ning. 2007. Principle of Parallel Word Base Constructing Based on Chinese Conversion, *In:*

*Proceedings of the 4th Chinese Digitization Foru*m, Macao, China.

Wang, N. and Wang, X. M. 2005. The conversion of Chinese characters and its communication in greater China. *In: Proceedings of the 3rd Chinese Digitization Forum (CDF'05)*. 1-20.

Wang, X. M. and Wei, L. M. 2008. Discussion of key problems in simplified-traditional Chinese character conversion. *In: Proceedings of the 5th Chinese Digitization Forum (CDF'08)*. 148–155.

Wong, Fai, Dong, Mingchui, Leong, Kaseng, and Cheong, Haocheong. 2007. Chinese Conversion Based on Statistic Model, *In: Proceedings of the 5th Chinese Digitization Forum*, Anhui, China.