

STCP: Simplified-Traditional Chinese Conversion and Proofreading

Jiarui Xu¹ and Xuezhe Ma¹ and Chen-Tse Tsai² and Eduard Hovy¹

¹ Language Technologies Institute, Carnegie Mellon University

² Department of Computer Science, University of Illinois at Urbana-Champaign

{jiarui, xuezhe}@cs.cmu.edu, ctsai12@illinois.edu, hovy@cmu.edu

Abstract

This paper aims to provide an effective tool for conversion between Simplified Chinese and Traditional Chinese. We present STCP, a customizable system comprising statistical conversion model, and proofreading web interface. Experiments show that our system achieves comparable character-level conversion performance with the state-of-art systems. In addition, our proofreading interface can effectively support diagnostics and data annotation. STCP is available at <http://lagos.lti.cs.cmu.edu:8002/>

1 Introduction

There are two standard character sets of the contemporary Chinese written language: Simplified Chinese and Traditional Chinese. Simplified Chinese is officially used in mainland China and Singapore, while Traditional Chinese is used in Taiwan, Hong Kong, and Macau. The conversion has become an essential problem with the increasing communication and collaboration among Chinese-speaking regions.

Although several conversion systems have been made available to the public, the conversion problem, however, remains unsolved. In this paper, we present an open-source system that provides a statistical model for conversion, as well as a web interface for proofreading. Our system achieves comparable performance with state-of-art systems. To the best of our knowledge, it is the first open-source statistical conversion system.

Another contribution of our system is the proofreading web interface. It is important for users to proofread the converted result and to make edits based on the linguistic information.


2 Levels of Conversion

Halpern and Kerman (1999) discussed the pitfalls and complexities of Chinese-to-Chinese conversion and introduced four conversion levels: code level, orthographic level, lexemic level, and contextual level, respectively. In this paper, we compact them into two levels of conversion: character level and word level.

2.1 Character level

There exists a mapping between Simplified Chinese characters and Traditional Chinese characters. Most characters only have a single corresponding character, while some characters may have multiple corresponding characters. In Simplified-to-Traditional conversion, characters with one-to-many mappings constitute about 12% of commonly used Chinese characters (Halpern and Kerman, 1999). Such phenomenon exists in Traditional-to-Simplified as well but to a much lesser extent. Character-level conversion of a given sentence involves both replacing characters that have one-to-one mapping with corresponding characters and disambiguating characters that have one-to-many mappings.

2.2 Word level

A concept may have different string surfaces due to the differences in word usage among various Chinese-speaking areas. For example,  is referred to as "football" in British English but "soccer ball" in American English. Such phenomenon is quite typical in Chinese-speaking areas. For example, Sydney is 悉尼 in mainland China but 雪梨 in Taiwan. Word-level conversion of a sentence involves determining if a word should be replaced with a corresponding word in look-up table. Disambiguation is also necessary if there are multiple corresponding words.

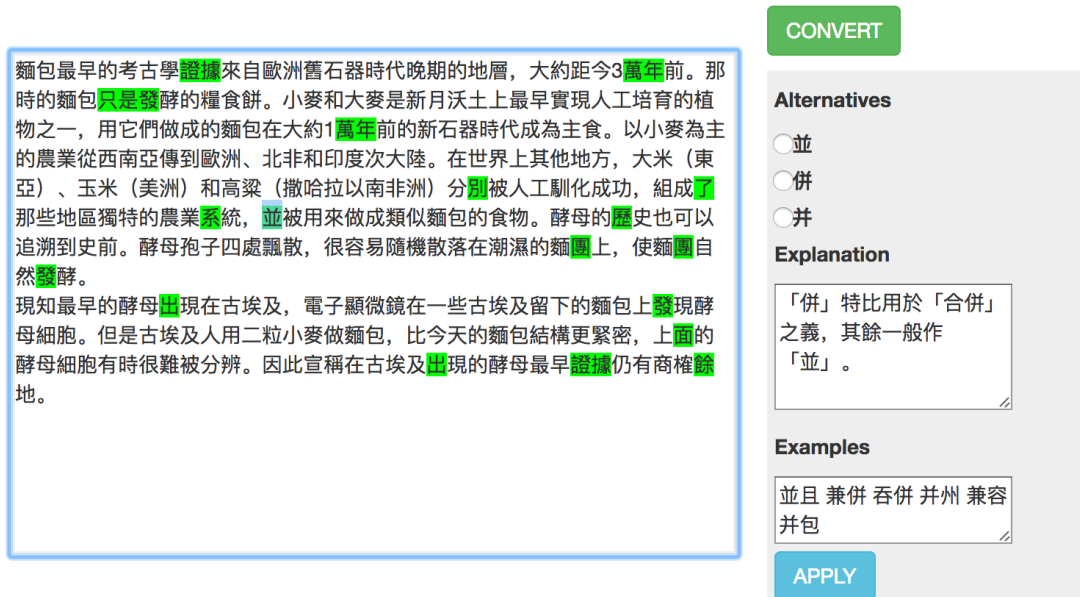


Figure 1: Screenshot of proofreading interface

We use the following sentence in Simplified Chinese to elaborate the conversion process:

我 了解 云端 软件
I know cloud software

Based on look-up tables of the character mappings, we list characters with one-to-one mappings in the above example sentence in Table 1 and those with one-to-many mappings in Table 2. Word mapping is shown in Table 3.

SC	我	解	端	软	件
TC	我	解	端	軟	件

Table 1: one-to-one character mapping

SC	TC	English
了	了	(auxiliary)
	瞭	know
云	云	say
	雲	cloud

Table 2: one-to-many character mapping

SC	软件
TC	軟體

Table 3: Word mapping in example sentence.

We decide to replace ‘软件’ with ‘軟體’ and

finally get the target sentence in Traditional Chinese: 我瞭解雲端軟體

3 System Architecture

3.1 Model

The Simplified-Traditional conversion problem is formulated as a translation problem (Brown et al., 1990):

Given a sentence s from source language (e.g. Simplified Chinese), return a sentence t in target language (e.g. Traditional Chinese) that maximizes the conditional probability:

$$P(t|s) = \frac{P(t)P(s|t)}{P(s)} \propto P(t)P(s|t)$$

Here we let $P(s|t)$ be the same for any candidate sentence t . Therefore, $P(t|s) \propto P(t)$ and the goal is to find:

$$t^* = \operatorname{argmax}_t P(t)$$

We describe how to generate candidate sentences through word and character conversion in section 3.1.1 and 3.1.2. The language model we used is briefly introduced in section 3.1.3.

3.1.1 Word Conversion

We tokenize the source sentence s into word sequence w_1, w_2, \dots, w_n . In our system, we use Jieba¹ Chinese text segmentation. For each word

¹<https://github.com/fxsjy/jieba>

w_i , if there exists a mapping of w_i in mapping table, we convert w_i into word w'_i in target language.

3.1.2 Character Conversion

After word conversion, the characters in words that have not been converted have one-to-one or one-to-many mapping. We generate candidate set T that contains all possible sentences by combining every possible conversions of each character.

3.1.3 Language Model

By default, the system uses a character-level language model with order of 5, estimated by KenLM (Heafield, 2011; Heafield et al., 2013). We choose KenLM because of its advantage in time and storage efficiency. User can substitute it with other trained language model.

3.2 Proofreading Interface

We provide a web-based proofreading interface that allows users to correct the converted text. Automatic conversion between Simplified Chinese and Traditional Chinese can never achieve 100% accuracy and we believe that, in many scenarios, such as government, commercial and legal document conversion, it is important to convert all characters and words as accurately as possible. Characters and words that have alternatives will be highlighted. When user selects these ambiguous fragments, explanation and example will be displayed and user can easily choose an alternative to replace the automatic results. Example proofreading of a paragraph and its highlights are shown in Figure 1.

4 Experimentation

Ministry of Education of the P.R.C. and Chinese Information Processing Society of China held a competition on the Evaluation of Intelligent Conversion System of Simplified Chinese and Traditional Chinese ² (MOE-CIPSC) in 2013. There are two core tasks: Character Conversion and Terminology Conversion. Few high-quality parallel corpus is available (Chang and Kung, 2007) and it is expensive to build one. Most websites that claim to have both Simplified Chinese and Traditional Chinese versions are using automatic systems without proofreading, thus are prone to errors. Our evaluation strategy adopts the task one of MOE evaluation.

²http://www.moe.edu.cn/s78/A19/A19_gggs/s8478/201302/t20130225_181150.html

4.1 Data

We use the Chinese Gigaword Fifth Edition (Parker et al., 2011) produce by the Linguistic Data Consortium (LDC). We select documents of type ‘story’ from Central News Agency (CMA), Taiwan after 2004, which are written in Traditional Chinese. In order to evaluate character conversion, we need to assume that there is no difference in word usage. Since conversion from Traditional Chinese to Simplified Chinese is not problematic on character level (Halpern and Kerman, 1999), we convert the CMA corpus into Simplified Chinese and use it as source language text set. The original CMA corpus becomes the target language text set. We split the entire data set into 80% training and 20% testing data randomly.

4.2 Evaluation

MOE-CIPSC evaluation provides a list of characters that have one-to-many mapping³. Overall accuracy is defined as: ($\#$ correctly converted ambiguous characters) / ($\#$ ambiguous characters). We also use Macro-average accuracy to evaluate performance across different characters.

4.3 Results and Analysis

Accuracies on character conversion are reported in Table 4 and Table 5. Note that XMUCC is a pre-trained system and OpenCC is a rule-based system. STCP outperforms OpenCC in terms of both accuracies and achieved comparable accuracy with XMUCC. Comparisons of these system are in section 5.

	OpenCC	XMUCC	STCP
Overall Accuracy	98.90	99.81	99.64

Table 4: Overall accuracies

	OpenCC	XMUCC	STCP
Macro-avg Acc.	91.75	96.98	95.73

Table 5: Macro-average accuracies

5 Related Work

There are several statistical approaches that have been proposed. Chen et al. (2011) integrates statistical features, including language models and lex-

³<http://bj.bcebos.com/cips-upload/dzb.txt>

ical semantic consistencies, into log-linear models. Li et al. (2010) uses look-up tables retrieved from Wikipedia to perform word substitution and disambiguate characters through language model. We adopt this method to build our conversion model. We use different look-up tables and we use higher order language model while they only use bigram and unigram.

The four most popular and publicly available systems are Google Translate, Microsoft Translator, Open Chinese Convert (OpenCC), and a system co-developed by Xiamen University, Ministry of Education of The People's Republic of China, and Beijing Normal University (XMUCC). OpenCC⁴ is an open-source project that performs conversion based on lookup tables constructed manually. XMUCC⁵ integrates language models and lexical semantic consistencies into log-linear models (Chen et al., 2011). However, XMUCC can be accessed through web interface and but it can only be executed in Windows command line as standalone program.

In end-use applications, especially when high quality conversion is required, human proofreading is required. Compared to (Zhang, 2011, 2014), our conversion is based on language model, instead of simply choosing the most frequent target characters. In addition, our proofreading interface highlights not only ambiguous characters, but also words. Users can also customize the system by importing look-up tables and language model, which can be useful for particular domains, such as science, business, and law.

6 Conclusion and Future Work

We develop an open-source customizable Chinese conversion system that is based on look-up tables and language model with a proofreading interface that assists end-use application. For future work, we will experiment with different language modeling approaches, such as neural language model. We will use the proofreading interface to construct parallel corpus of high quality to evaluate word-level conversion.

Acknowledgments

References

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek,

⁴<https://github.com/BYVoid/OpenCC>

⁵<http://jf.cloudtranslation.cc/>

John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Comput. Linguist.* 16(2):79–85. <http://dl.acm.org/citation.cfm?id=92858.92860>.

Jing-Shin Chang and Chun-Kai Kung. 2007. A chinese-to-chinese statistical machine translation model for mining synonymous simplified-traditional chinese terms. *Proceedings of Machine Translation Summit XI*.

Yidong Chen, Xiaodong Shi, and Changle Zhou. 2011. A simplified-traditional chinese character conversion model based on log-linear models. In *2011 International Conference on Asian Language Processing*, pages 3–6. <https://doi.org/10.1109/IALP.2011.15>.

Jack Halpern and Jouni Kerman. 1999. Pitfalls and complexities of chinese to chinese conversion. In *International Unicode Conference (14th) in Boston*.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pages 187–197.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn. 2013. Scalable modified kneser-ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696.

Min-Hsiang Li, Shih-Hung Wu, Yi-Ching Zeng, Pingche Yang, and Tsun Ku. 2010. Chinese characters conversion system based on lookup table and language model. *Computational Linguistics and Chinese Language Processing* 15(1):19–36.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword fifth edition LDC2011T07. DVD. *Philadelphia: Linguistic Data Consortium*.

Xiaoheng Zhang. 2011. A simplified-traditional chinese conversion tool with a supporting environment for human proofreading. *The 11th Chinese National Conference on Computational Linguistics, CNCCL*.

Xiaoheng Zhang. 2014. *A Comparative Study on Simplified-Traditional Chinese Translation*, Springer International Publishing, Cham, pages 212–222. https://doi.org/10.1007/978-3-319-12277-9_19.