An Alignment based technique for Text Translation between Traditional Chinese and Simplified Chinese

Sue J. Ker
Department of Computer and Information Science
Soochow University
Taipei, Taiwan, R.O.C.
ksj@cis.scu.edu.tw

Chun-Hsien Lin
Department of Computer and Information Science
Soochow University
Taipei, Taiwan, R.O.C.
jslin@seed.net.tw

Abstract

Aligned parallel corpora have proved very useful in many natural language processing tasks, including statistical machine translation and word sense disambiguation. In this paper, we describe an alignment technique for extracting transfer mapping from the parallel corpus. During building our system and data collection, we observe that there are three types of translation approaches can be used. We especially focuses on Traditional Chinese and Simplified Chinese text lexical translation and a method for extracting transfer mappings for machine translation.

1 Introduction

Aligned parallel corpora have proved very useful in many tasks, including statistical machine translation (Brown et al., 1993; Wu and Ng, 1995, Chen et al., 1997; Moore, 2001) and word sense disambiguation (Chang et al., 1996; Chen and Chang, 1998).

Traditional and Simplified Chinese are two Chinese writing systems that used by Chinese-speaking communities. Since their typefaces are different, foreigners always view these two languages as a family of cognate languages. Aside from differences in typeface, their encoding schemas are also different. For the conversion of text, special utilities or tools are required for mapping the correspondence between the two schemas. At present, the methods used to undertake this mapping are far from perfect. In general, a table-conversion method is used to translate between Traditional and Simplified Chinese text. There are several problems with this method. First, correspondences between Big5 (Traditional Chinese) and GB2312 (Simplified Chinese) code schemas are not one-to-one. Thus, this method can cause mismatches in character translation. Second, the unit of processing should be words instead of morphemes since the meaning of a morpheme can be very ambiguous. Third, conventional language usage is also quite different between Traditional and Simplified Chinese. To tackle these difficulties, it thus seems wise to acquire a set of word mappings between Traditional Chinese and Simplified Chinese from parallel corpora automatically.

This paper discusses the issues mentioned above, and especially focuses on Traditional and Simplified Chinese text lexical translation and a method for building a synonym thesaurus meaning-translation. The character set standards used in this paper are Big5 code for Traditional Chinese and GB2312 code for Simplified Chinese; nevertheless, these issues are code-independent. Simply categorized, there are three methods for Traditional Chinese and Simplified Chinese conversion and translation, each of which satisfies different purposes. These include code schema conversion, word translation, and semantic translation, all of which are described in the present paper.

2 Motivation

As the Internet and World Wide Web become increasingly popular, text documents in electronic form are becoming ever more abundant. Therefore, there is a rapidly developing demand for the translation of text between Traditional Chinese and Simplified Chinese. In addition, there are vast numbers of Traditional Chinese and Simplified Chinese texts in electronic form that require an automatic conversion system for translation. Although several utilities and automatic conversion systems have been designed (Chang, 1998), they are far from perfect and many issues remain unsolved. Indeed, manual correction is typically required when these systems are used. Since the volume of documents is growing rapidly, however, manual correction is not a feasible long-term solution. There is thus an urgent need for a reliable automatic conversion system to deal with Traditional Chinese and Simplified Chinese texts.

3 Observation

In the process of building our system and collecting data, we observe that there are three types of translation approaches can be used. These include code schema conversion, word translation, and semantic translation.

3.1 Code Schema Conversion is not enough

While there are many encoding schema for Chinese character set, we have focused on Big5 and GB2312 because they are most commonly used. Table 1 illustrate how “中” and “文”, which have the same typeface and meaning in both Traditional and Simplified Chinese, are encoded differently in Big5 and GB2312. Many Traditional and Simplified Chinese characters can be conversed simply apply to this approach which building a kind of code conversion table. It is the easiest approach for Traditional Chinese and Simplified Chinese translation. We find out there is many off-the-shelf utilities can help us. For example, Emurasoft (www.emurasoft.com) provides software named EmEditor, which can converse various Chinese code schema. The Microsoft® Word 2000 Traditional Chinese edition, which built-in a Traditional Chinese and Simplified Chinese conversion function. These software can do this kind of work well.

Table 1 The sample of Character Coding between Traditional and Simplified Chinese.

Chinese Character	Big5 Character Set Code	GB2312 Character Set Code
中	0xA4A4	0xD6D0
文	0xA4E5	0xCEC4

Unfortunately, using the code conversion table approach cannot satisfy all translation requirements. Table 2 is a sample of code conversion between Big5 and GB. We can observe that there is a one-to-many relationship between GB and Big5, especially when converting from GB to Big5. Table 3 shows that after word combination, translating from Simplified Chinese into Traditional Chinese only using the code conversion table approach will lead to lexeme mismatches. In Table 3, the translation of “老么(lao-yao)” to “老么(lao-yao)” is correct. The translation of “什么(she-me)”, however, show be “什麼(she-me)”, and not “什么(she-yao)”. Also, “技朮(hi-shu)” and “几乎(ji-hu)” are not translated correctly. Therefore, some type of lexical correction approach should be applied to correct these anomalies.

Table 2 The sample of Code conversion table.

Traditional Chinese	Simplified Chinese
么(yao)	么(yao)
麼(me)	么(me)
朮(shu)	朮(shu)
術(shu)	朮(shu)
几(ji)	几(ji)
幾(ji)	几(ji)
裡(li)	里(li)
里(li)	里(li)

Table 3 Sample of translating from Simplified Chinese into Traditional Chinese by code conversion only.

Simplified Chinese	After code conversion	Correct Traditional Chinese
老么(last child)	老么(Lao-Iao)	老么(Lao-Iao)
什么(what)	什么(She-Iao)	什麼(She-Mo)
技朮(technique)	技朮(Ji-Shu)	技術(Ji-Shu)
几乎(almost)	几乎(Ji-Hu)	幾乎(Ji-Hu)

3.2 Word Translation rather than Morpheme Translation

Differences in word usage are another issue affecting translation between Traditional and Simplified Chinese. For instance, Table 4 shows a brief example of a synonymous lexeme thesaurus. In our experience, building a specified domain synonym thesaurus for word usage translation is necessary.

Table 4 Differences in Traditional Chinese and Simplified Chinese lexeme usage.

Traditional Chinese	Simplified Chinese	GB to Big5
軟體(software)	软件	軟件
影印本 (copy)	复印件	複印件
線上(on-line)	在线	在線
印表機(printer)	打印机	印表機
社區(community)	居住小区	居住小區

In many cases, only using a synonym thesaurus for handling word usage issues in translation is not reliable. Example 1 and Example 2 below show that mismatches in word usage can occur during translation. For instance, in Example 1, “微軟” is the name for Microsoft® in Chinese, but this term is followed by another character that leads to incorrect word division and a resulting mismatch. The sentence marked with an asterisk in Example 1 illustrates this issue. The following is a detailed description of the process by which the mismatch occurs. In the first sentence in Example 1, “微軟” and “體驗” are two Traditional Chinese words. If they are divided into the four Chinese morphemes, “微”, ” 軟”, “體”, and “驗”,they can be converted into their Simplified Chinese forms, “微”, “软”, “体”, and “验”. Using either code schema conversion or word translation, “微软” and “体验” can be converted correctly. But in some cases, “微軟” and “體驗” will be incorrectly divided into “微”, “軟體”, and “驗”. After conversion using word translation, they will then become “微”, “软件”, and “验”. When recombined, the result, is “微软件验” which, as shown in Example 1, produces an incorrect translation. Example 2 and Example 3 also illustrate the same phenomenon. Hence, translating sentences with a synonym thesaurus can work in some cases, but fail in other cases. In general, word translation does produce better results than morpheme translation. On the other hand, building a general synonym thesaurus is arduous work. A specified domain synonym thesaurus for specified domain translation is preferred

Example 1

微軟體驗到顧客的需求是多樣化的
*微软件验到顾客的需求是多样化的

Example 2

第六合作社區分成三大工作小組
*第六合作居住小区分成三大工作小组

Example 3

約旦外交大臣哈提卜
*约旦外交通大学臣哈提卜

3.3 Translation in Accordance with Document Topic

The examples in the previous section imply that using code schema conversion accompanied by word translation is sometimes insufficient. In such cases, properly selecting a synonym thesaurus in accordance with the document's topic can be of great assistance in the translation process. For instance, the GB-to-Big5 translation in Example 4, which is a sentence fragment, describes a social issue. If we use an IT synonym thesaurus to translate this phrase, the incorrect translation, marked with an asterisk, results. The word "循环(xun-huan)" is mapped into "循環(xun-huan)", when using only code schema conversion. When making use of an IT synonym thesaurus, however, "循环(xun-huan)" will be translated into "回圈(hui-quan)", which is incorrect semantically, but correct in terms of synonym thesaurus translation. Example 5 shows the same phenomenon. "活动(huo-dong, meaning 'action')" should be converted to "活動(huo-dong)", rather than "啓動(qi-dong, which means 'enable')". Therefore, the synonym thesaurus used must be selected in accordance with the document's topic..

Example 4

以便制止暴力循环
*以便制止暴力回圈

Example 5

严禁任何未经官方许可的游行示威活动
*嚴禁任何未經官方許可的遊行示威啓動

Semantic translation between Traditional and Simplified Chinese is the most difficult undertaking of all. After using the translation approaches mentioned above, a readable sentence should be created. In experience, there are still a few sentences and phrases that need to be adjusted in terms of syntax and grammar. Table 5 shows some cases in which different sentences in Traditional and Simplified Chinese describe the same thing. In row 1 of Table 5, making "表明(show)" synonymous with "顯示(show)" is feasible. But the synonym thesaurus approach clearly does not work for the sentences in rows 2 through

4. In row 2, it is necessary to add “對(to)” and “造成(let, make)” to the translated sentence and exchange the position of “極大損害” and “世界經濟” to produce a fluent Traditional Chinese sentence. In row 3, “大腕” and “有錢” both mean “rich”, “好一阵兴奋” and “相當興奮” both mean “very exciting”, but they are examples of large differences in literal meaning and grammar between Traditional and Simplified Chinese. The extreme example shown in row 5 illustrates two sentences there are totally different in literal meaning, but the same in figurative meaning. We leave this phenomenon for future work.

Table 5 Literal differences between Traditional Chinese and Simplified Chinese.

Simplified Chinese	Traditional Chinese
迹象表明	跡象顯示
同时会极大损害世界经济	同時會對世界經濟造成極大損害
就连见惯明星大腕的记者们23日见到崔永元也是好一阵兴奋	就連見慣有錢明星的記者們23日見到崔永元也是相當興奮
可就是无法成功将ERP移植入企业	但就是無法將ERP成功移植至企業
人上一百形形色色	一樣米養百樣人

4 The Alignment-based Algorithm

4.1 Estimation of Lexical Translation Probability

In this section, we propose a word-to-word similarity measure between Traditional Chinese and Simplified Chinese words. Let us consider a Traditional Chinese word *t* and a Simplified Chinese word *s*. Let SC_t denote the GB code of word *t*. A similarity measure based on the Dice coefficient (Dice, 1945) can be given as follows:

$$Sim(s,t) = \frac{2 \times |s \cap SC_t|}{|s| + |t|} \qquad \text{(Eq. 1)}$$

where t = the morpheme strings of Traditional Chinese word,
s = the morpheme strings of Simplified Chinese word,
SC_t= the morpheme strings of Simplified Chinese representation of *t*,
$|t|$= The lenth of string *t*.

To illustrate how *Sim* is calculated, consider the word pairs in Table 5. The first column shows some Traditional Chinese words. The second column displays the corresponding Simplified Chinese words for those in column 1. We convert word *t* from BIG5 to GB code, name it SC_t, and present it in the third Column. After applying Equation 1 to calculate the similarity of word *s* and *t*, we show the obtained value in the last column of Table 6.

Table 6 Some word pairs and their Sim value.

Traditional Chinese word *t*	Simplified Chinese *s*	SC_t	*Sim*(*s*,*t*)
軟體(software)	软件	软体	0.50
影印本 (copy)	复印件	影印本	0.33
印表機(printer)	打印机	印表机	0.67
社區(community)	居住小区	社区	0.33

Armed with *Sim*, we then define and estimate the LTP t(*s*, *t*) by the following cases:

Case 1. $Sim(s, t) \geq h_1$,
Case 2. $h_1 > Sim(s, t) \geq h_2$,
Case 3. $h_2 > Sim(s, t) \geq h_3$,
Case 4. $Sim(s, t) < h_3$.

The connections satisfying each condition are given the same probability value determined by maximal likelihood estimation (MLE). For instance, if there are k connections in a sample of n candidates (s, t)· such that $Sim(s, t) \geq h_1$ then all these candidates are given the same MLE value for LTP, i.e. $t(s, t) = t_1 = k/n$. Equation (Eq. 2) sums up the above discussion:

$$t(s,t) = \begin{cases} t_1 & \text{if } Sim(s,t) \geq h_1, \\ t_2 & \text{if } h_1 > Sim(s,t) \geq h_2, \\ t_3 & \text{if } h_2 > Sim(s,t) \geq h_3, \\ t_4 & \text{if } Sim(s,t) < h_3. \end{cases} \qquad \text{(Eq. 2)}$$

By using a small sample of a few hundred sentences, the LTP values t_i for $1 \leq i \leq 4$ can be estimated. Table 7 summarizes the probabilistic values based on likelihood ratio estimated using 200 sentences from corpus.

Table 7 Maximum likelihood estimation (MLE) of LTP estimated based on likelihood ratio, $Sim(s, t)$.

Likelihood Ratio	MLE of LTP(s, t)	
$Sim(s, t) \geq 0.80$	t_1	0.84
$0.80 > Sim(s, t) \geq 0.50$	t_2	0.38
$0.50 > Sim(s, t) \geq 0.25$	t_3	0.23
$Sim(s, t) < 0.25$	t_4	0.02

4.2 Estimation of Distortion probability

We observe that by considering the translational position relative to the immediate left and right neighbors, one obtains a probabilistic distribution function with a smaller deviation, thereby making a tighter estimation possible for $d(i, j)$. To this end, we define dislocation, *dis* for the connection (s_i, t_j) of the i th and j th words in S and T, to denote $|(j - j')-(i - i')|$ where i' is the position of a word s' sharing the minimum syntactic structure with s, and s' translates into t', the j' th word in T. Short of syntactic analysis, $dis(i, j)$ can be calculated with respect to a nearby connection in *CONN*.

For establishing the initial connections, two dummies are replace to the left of the first and to the right of the last word of the Tradition Chinese sentence. Similar two dummies are added to the target sentence. These left dummies in the both sentences are aligned each other. Similarly, the right dummies align with each other. This establishes anchor points for calculating the relative distortion score.

Such treatment closely approximates the dislocation value. In light of this, dislocation can be defined as follows:

$$\text{dis}(i,j) = \min(|d_L|, |d_R|) \qquad \text{(Eq. 3)}$$

where i = the sequence number[1] of s in S,

j = the sequence number of t in T,

$d_L = (j - j_L) - (i - i_L)$,

$d_R = (j - j_R) - (i - i_R)$,

$(i_L, j_L) = \underset{(i',j') \in CONN_{<i}}{\text{argmax}} \; i'$,

$(i_R, j_R) = \underset{(i',j') \in CONN_{>i}}{\text{argmin}} \; i'$,

$CONN_{<i}$ = $\{(k, l) \mid k$ th and l th word in (S, T) form a connection in $CONN$, $k < i\}$,

$CONN_{>i}$ = $\{(k, l) \mid k$ th and l th word in (S, T) form a connection in $CONN$, $k > i\}$,

$CONN$ = the initial connections established according to the two added null anchors.

The distortion function defined by cases can now be given according to *dislocation* values.

$$\text{d}(i,j) = \begin{cases} d_1 & \text{if } \text{dis}(i,j) = 0, \\ d_2 & \text{if } \text{dis}(i,j) = 1, \\ d_3 & \text{if } \text{dis}(i,j) = 2, \\ d_4 & \text{if } \text{dis}(i,j) \geq 3. \end{cases} \qquad \text{(Eq. 4)}$$

The connection candidates with small *dislocation* values tend to be alignment connections. Again, all candidates (s_i, t_j) satisfying a certain case in (Eq. 4) are given the same MLE value. For instance, if there are k true connections in a sample of n candidates (i, j) with 0 *dislocation*, then all these candidates are given the same MLE value for DP, i.e. $\text{d}(i, j) = d_1 = k/n$ for all i and j such that $\text{dis}(i, j) = 0$. By using a small sample of a few hundred sentences, the DP values d_i for $1 \leq i \leq 4$ can be estimated. Table 8 summarizes the probabilistic values based on likelihood ratio estimated using 200 sentences from corpus.

Table 8 Maximum likelihood estimation (MLE) of DP estimated based on likelihood ratio, *dis(i, j)*.

Likelihood Ratio	MLE of DP(*i*,*j*)	
dis(i,j) = 0	d_1	0.94
dis(i,j) = 1	d_2	0.30
dis(i,j) = 2	d_3	0.12
dis(i,j) ≥ 3	d_4	0.01

4.3 The Word Alignment Algorithm

The above descriptions are summarized in the following algorithm:

Step 1: Get a pair of Tradition Chinese and Simplified Chinese sentences from corpus.
Step 2: Perform the word segmentation for both sentences.
Step 3: Two dummies are placed to both end of two sentences, and let them aligned each other.
Step 4: Follow the procedure in Section 4.1 to calculate a lexicon translation probability for each connection candidate according to the Eq. 1 and Eq. 2.

[1] The sequence number of a word is assigned according to the segmentation which satisfies the long-word-first heuristic.

```
        Repeat
Step 5:     Follow the procedure in Section 4.2 to calculate a relative distortion probability for each
            connection candidate according to the Eq. 3 and Eq. 4.
Step 6:     The highest scored candidate is selected and added to CONN.
Step 7:     The connection candidates that are inconsistent with the selected connection are
            removed from the candidate list.
Step 8: Until all words in the source sentence are aligned or no candidate is greater than a preset
        threshold $\theta$.
```

In this paper, we propose a translation model for extracting synonyms from our small parallel corpus. In section 6, we use this small parallel corpus to build-up synonyms thesaurus for experiment.

5 Experimental Results

To assess the proposed method's effectiveness, we have implemented the algorithms described in Section 4 and conducted a series of experiments. A general description of the materials used in the experiments follows. Finally, the success rates are quantitatively evaluated.

5.1 The Experimental Setup

We collect over 5,000 article-pairs (Traditional-Simplified Chinese pair) from news website (http://news.pchome.com.tw/) to form a small parallel corpus. The news topic we focus on Information Technique news. The experimental results obtained from the proposed algorithms are presented in this Section. The training data were used primarily to determine MLE estimates for the cases of LTP and DP. The test·sentences were randomly chosen from unseen data from the same domain.

5.2 Performance Evaluation

Our experiment was designed to demonstrate the effectiveness of the presented algorithm. We used human evaluators. According to the experimental results, over 90% of the source words in test sets are connected to a target and over 90% are correct connections. Table 9 shows some high quality of aligned word pairs.

Table 9 Some transfer mappings was produced by our alignment.

Aligned Chinese Terms		Aligned Chinese Terms	
Traditional	Simplified	Traditional	Simplified
電腦	计算器	密碼	口令
行動電話	手机	著作權	版权
目前	当前	資訊	信息
國語	普通话	軟體	软件
整合	集成	印表機	打印机
產生	生成	取代	替换
個人電腦	个人计算器	螢幕	屏幕
水準	水平	大專	高校
網路	网络	呆帳	坏帐
記憶體	内存	新力	索尼
晶片	芯片	模組	模块
上班族	工薪阶级	班機	航班
資訊	信息	高水準	高水平
浴室	卫生间	元件	组件
黨團	党派	免洗	卫生

廁所	卫生间	拋售	甩卖
互動	交互	晶片組	芯片组
品質	质量	影像	图像
支援	支持	列印	打印
大學聯考	高考	顯示器	监视器
區域網路	局域网	紐西蘭	新西兰
電腦週邊	计算器周边	環保	绿色
電腦網路	计算器网络	程式	程序
電腦廠	计算器厂	超商	超市
保險套	避孕套	網際網路	网际网络
伺服器	服务器	輪胎	轮带
用戶	客户		

6 Concluding Remarks

This paper has presented some examples and explains issues between text Traditional Chinese and Simplified Chinese translation approaches. An accurately specified synonyms thesaurus can raise lexical translation accuracy. By difference of Traditional Chinese and Simplified Chinese phrase usage, a good semantic translation is needed. Using semantic translation, fluently sentence can be generated. We leave this part as future work.

Acknowledgments

This work was partially supported by ROC NSC grants, 90-2213-E-031-005. We are also thankful to anonymous reviewers for many useful suggestions.

References

Brown, P.F., S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics*, 19(2): 263-311.

Brown, P.F., V.J. Della Pietra, P.V. deSouza, J.C. Lai, and R.L. Mercer. 1992. A Statistical Approach to Machine Translation, *Computational Linguistics*, 16(2):79-85.

Chang, J.S., J.N. Chen, H.H. Sheng and S.J. Ker, 1996. Combining Machine Readable Lexical Resources and Bilingual Corpora for Board Word Sense Disambiguation, In Proceedings of the second Conference of the Association for Computational Linguistics, 9-16.

Chao-Huang Chang, 1998. Noisy Channel Models for Corrupted Chinese Text Restoration and GB-to-Big5 Conversion, *Computational Linguistics and Chinese Processing*, 3(2):79-92.

Chen J.N. and J.S. Chang, 1998. TopSense : A Topical Sense Clustering Method Based on Information Retrieval Techniques on Machine Readable Resources, *Computational Linguistics*, 24(1):61-95.

Chen, M.H., J.S. Chang, S.J. Ker and J-N Chen, 1997, TopAlign: Word Alignment for Bilingual Corpora based on Topical Clusters of Dictionary Entries and Translations, In *Proceedings of the Eighth International Conference on Theoretical and Methodological Issues in Machine Translation* (*TMI-97*), 127-134.

Dice, L.R. 1945. Measures of the Amount of Ecologic Association between Species, *Journal of Ecology*, 26: 297-302.

Emurasoft, Inc., http://www.emurasoft.com/index.htm.

Ker, S.J. and J.S. Chang, 1997. A Class-base Approach to Word Alignment, *Computational Linguistics*, 23(2): 313-343.

Moore, R.C., Towards a SIMPLE AND accurate Statistical Approach to Learning Translation Relationships among Words, In *proceedings of the Workshop on Data-Driven Machine Translation*, ACL 20001.

Wu D. and C. Ng, 1995. Using Brackets to Improve Search for Statistical Machine Translation. In *Proceedings of the 10th Pacific Asia Conference on Language, Information and Computation*, 195-204.