Translation Divergence Analysis and Processing for Mandarin-English Parallel Text Exploitation

Shun-Chieh Lin, Jia-Ching Wang, and Jhing-Fa Wang
Department of Electrical Engineering, National Cheng Kung University.
701 No.1, Ta-Hsueh Road, Tainan City Taiwan R.O.C.
Tel: 886-6-2757575 Ext. 62341, Fax: 886-6-2761693
E-mail: wangjf@csie.ncku.edu.tw

Abstract

Previous work shows that the process of parallel text exploitation to extract transfer mappings between language pairs raises the capability of language translation. However, while this process can be fully automated, one thorny problem called "divergence" causes indisposed mapping extraction. Therefore, this paper discuss the issues of parallel text exploitation, in general, with special emphasis on divergence analysis and processing. In the experiments on a Mandarin-English travel conversation corpus of 11,885 sentence pairs, the perplexity with the alignments in IBM translation model is reduced averagely from 13.65 to 5.18 by sieving out inappropriate sentences from the collected corpus.

1. Introduction

Over the past decade, research has focused on the automatic acquisition of translation knowledge from parallel text corpora. Statistical-based systems build alignment models from the corpora without linguistic analysis [1,2]. Another class of systems analyzes sentences in parallel texts to obtain transfer structures or rules [6]. Previous work shows that the process of parallel text exploitation to extract transfer mappings (models or rules) between language pairs can raise the capability of language translation.

However, previous work is still hampered by the difficulties in transfer mapping extraction of achieving accurate lexical alignment and acquiring reusable structural correspondences. Although automatic extraction methods of lexical alignment and structural correspondences are introduced, they are not capable of handling exceptional cases like "divergence" presented in [4]. In general, divergence arises with variant lexical usage of role, position, and morphology between two languages. Therefore, while mapping extraction can be fully automated from parallel texts, divergence causes indisposed mapping extraction. Furthermore, the existence of translation divergences also makes adaptation from source structures into target structures difficult [5,7,8]. For parallel text exploitation, these divergences make the training process of transfer mapping extraction between languages impractical including parsing and word-level alignment, lexical-semantic lexicography, and syntactic structures. Therefore, study of parallel text exploitation needs a careful study of translation divergence.

The framework of this paper is as follows. A brief overview of parallel text exploitation is discussed in Section 2. In Section 3, translation divergence analysis and processing for Mandarin-English parallel texts is presented. Section 4 shows experimental results with the alignments in IBM translation model. Finally, generalized conclusions are presented in Section 5.

2. Overview of Statistical-based Parallel Text Exploitation

The goal of parallel text exploitation is to acquire the knowledge for translation of a text given in some source (“Mandarin”) string of words, *m* into a target (“English”) string of words, *e*. For the presented statistical approach [1] to string translation of $\Pr(e \mid m)$, among all possible target strings, the string will be chosen with the highest probability which is given by Bayes’ decision rule as follows:

$$\hat{e} = \arg\max_{e} \Pr(e)\Pr(m \mid e) \tag{1}$$

Pr(*e*) is the language model of target language and Pr(*m*| *e*) is the translation model. In order to estimate the correspondence between the words of the target sentence and the words of the source sentence, a sort of pair-wise dependence by considering all word pairs for a given sentence pair [*m*, *e*] is assumed, referred to as alignment models. Figure 1 shows an example for the translation parameters of a sentence pair. In general, these parameters are lexicon probability, ex. $p(m_j \mid e_i)$, sentence length probability, ex. $p(l_m \mid l_e)$, and alignment probability, ex. $p(j \mid i, l_m, l_e)$. Therefore, given more parallel texts, more probability parameters could be estimated for translation.

Mandarin: $l_m = 4$

$m = m_1 m_2 \cdots m_j \cdots m_{l_m}$:(一)$_1$ (晚)$_2$ (多少)$_3$ (錢)$_4$?

English: $l_e = 4$

$e = e_1 e_2 \cdots e_i \cdots e_{l_e}$:(*How much*)$_1$ (*for*)$_2$ (*a*)$_3$ (*night*)$_4$?

$p(m_2 \mid e_4)$: lexicon probability

$p(l_m = 4 \mid l_e = 4)$: sentence length probability

$p(j = 2 \mid i = 4, l_m, l_e)$: alignment probability

Fig. 1. An example for the translation parameters of a sentence pair

However, it is difficult to achieve straightforward and correct estimation for these probability parameters. In the above example, the English word “for” is one major factor called “*divergence*” makes the estimation process between sentence pairs impractical. Therefore, in the next section, we present the analysis and processing of the translation divergence for improving the performance on parallel text exploitation.

3. Translation Divergence Analysis and Processing

3.1 Analysis of Divergence Problems

Dorr’s work [3] of divergence analysis is based on English-Spanish and English-German translations. Based on these two language pairs, 5 different categories have been identified. In this section, we discuss more multiform examples among the 5 types of divergences in Mandarin-English parallel texts. For each example, three sentences are given: *e* means an original English sentence in parallel texts, *m* means a Mandarin sentence, and *ẽ* means an amended English sentence which is better for translation parameter training with *m*.

3.1.1 Identification of Thematic Divergence

Thematic divergence often involves a “swap” of the subject and object position and obtains unpredictable word-level alignment. For example,

e: (*Is*)$_1$ (*credit card*)$_2$ (*acceptable*)$_3$ (*to*)$_4$ (*them*)$_5$?
m: (*他們*)$_1$ (*接受*)$_2$ (*信用卡*)$_3$ (*嗎*)$_4$?
ẽ: (*Do*)$_1$ (*they*)$_2$ (*accept*)$_3$ (*credit card*)$_4$?

Here, credit card appears in subject position in *e* and in object position (“*信用卡*”) in *m*; analogously, the object them appears as the subject they (“*他們*”). Therefore, for the thematic divergence, the position alignments of 2↔3 and 5↔1 are obtained in a sentence pair [*m*, *e*]. However, if a sentence pair [*m*, *ẽ*] can be provided, the position alignments of 1↔2, 2↔3, and 3↔4 are better for straightforward parameter estimation of $p(j \mid i, l_m, l_e)$.

3.1.2 Identification of Morphological Divergence

Morphological divergence involves the selection of a target-language word that is a morphological variant of the source-language equivalent and it raises the ambiguity of lexical-semantic lexicography.

e: (*May*)$_1$ (*I*)$_2$ (*have*)$_3$ (*your*)$_4$ (*signature*)$_5$ (*here*)$_6$?
m: (*請*)$_1$ (*你*)$_2$ (*在*)$_3$ (*這*)$_4$ (*簽名*)$_5$ (*好嗎*)$_6$?
ẽ: (*Could*)$_1$ (*you*)$_2$ (*sign*)$_3$ (*here*)$_4$?

In this example, the predicate is nominal (*signature*) in *e* but verbal (“*簽名*”) in *m*. While inputting two sentence pairs [*m*, *e*] and [*m*, *ẽ*], the parameter estimation of $p(m_j \mid e_i)$ should be reformulated with two morphological translation conditions: $p(m_j, m_j \in V \mid e_i, e_i \in N)$ and $p(m_j, m_j \in V \mid e_i, e_i \in V)$. Therefore, with growing of various morphological translations, more conditions would raise more complexity of lexicon transfer parameter estimation and cause more ambiguity of lexical-semantic lexicography.

3.1.3 Identification of Structural Divergence

In structural divergence, a verbal argument has a different syntactic realization in the target language and the appearance of the divergence causes additional syntactic structural mapping constructions.

e: (*About*)$_1$ (*the*)$_2$ (*center*)$_3$.
m: (*大概*)$_1$ (*在*)$_2$ (*中間*)$_3$.
$\tilde{e}$: (*About*)$_1$ (*in*)$_2$ (*the*)$_3$ (*center*)$_4$.

Observe that the place object is realized as a noun phrase (*the center*) in *e* and as a prepositional phrase ("*在 中間*") in *m*. For this example, the divergence causes the alignment of 0←2, which is a null mapping for Mandarin lexicon "*在*". In addition, the divergence also causes alignments of 2↔3 and 3↔3, which result in non-equal mapping number *q*-to-*n* (*q*>1, *n*>1, and *q*≠*n*). For a raised null mapping, the parameter estimation of $p(j \mid i, l_m, l_e)$ and $p(l_m \mid l_e)$ become more complicated by further considering translation of lexicon insertion (*i*=0) and deletion (*j*=0). More raised non-equal mapping number in parallel texts, more parameter estimation of $p(l_m \mid l_e)$ and more length generation condition for translation.

3.1.4 Identification of Conflational Divergence

Conflation is the incorporation of necessary participants (or arguments) of a given action. A conflational divergence arises when there is a difference in incorporation properties between two languages. In addition, there are word compounds in Chinese language by embedding some semantic contiguity. For this divergence, the complexity of training process for transfer mapping extraction is extremely increased.

e: (*Please*)$_1$ (*have*)$_2$ (*him*)$_3$ (*call*)$_4$ (*me*)$_5$.
m: (*請*)$_1$ (*轉告*)$_2$ (*他*)$_3$ (*回*)$_4$ (*個*)$_5$ (*電話*)$_6$ (*給*)$_7$ (*我*)$_8$.
$\tilde{e}$: (*Please*)$_1$ (*tell*)$_2$ (*him*)$_3$ (*to*)$_4$ (*give*)$_5$ (*me*)$_6$ (*a*)$_7$ (*call*)$_8$.

This example illustrates the conflation of a constitution in *e* that must be overly realized in *m*: the effect of the action (*give me a call*) is indicated by the word "*回 個 電話 給 我*" whereas this information is incorporated into the main verb (*call me*) in *e*. Therefore, this divergence causes most complexity on parameter estimation of translation including $p(m_j \mid e_i)$, $p(l_m \mid l_e)$, and $p(j \mid i, l_m, l_e)$.

3.1.5 Identification of Lexical Divergence

For lexical divergence, the event is lexically realized as the main verb in one language but as a different verb in other language. It typically raises the ambiguity of lexical-semantic lexicography and also can be viewed as a side effect of other divergences. Thus, the formulation thereof is considered to be some combination of those given above, such as a conflational divergence forces the occurrence of a lexical divergence.

e: (*Nothing*)$_1$ (*can*)$_2$ (*beat*)$_3$ (*Phantom of the Opera*)$_4$.
m: (*沒有*)$_1$ (*什麼*)$_2$ (*比得上*)$_3$ (*歌劇魅影*)$_4$.
ẽ: (*Nothing*)$_1$ (*can*)$_2$ (*compare*)$_3$ (*with*)$_4$ (*Phantom of the Opera*)$_5$.

Here the main verb "*beat*" in *e* but as a different verb "*比得上*" (*to compare with*) in *m*. Other examples are like "*cash*", "*have*", "*take*", and etc. in English but "*兌換 成 現金*", "*轉告*", "*坐*", and etc. in Mandarin, respectively.

3.2 Processing of Divergence Evaluation

According to the above divergence analysis, the divergent mappings between sentence pairs are composed of non-equal mapping number (*q*-to-*n*, *q*>1, *n*>1, *q*≠*n*), different position mapping (*i*↔*j*, *i*≠ *j*), and null mapping (*i*→0 or 0→*j*). Unlike non-equal mapping number and different position mapping, the null mapping cannot provide target language translation information for lexical item selection and position generation. Therefore, we want to use a simple and straightforward measurement method to evaluate the possible null mappings.

For example to the Mandarin-English parallel text corpus, given a Mandarin sentence $m = m_1 m_2 \cdots m_j \cdots m_{l_m}$ and an English sentence $e = e_1 e_2 \cdots e_i \cdots e_{l_e}$:, direct lexical mappings in the mapping space can be extracted using the relevant bilingual dictionary [13]. The mapping function is defined as follows:

$$\tau\left(m_j, e_i\right) = \delta\left(m_j - \sigma_k\right) = \begin{cases} 1 & \text{if } \exists \sigma_{\mathrm{k}} \in \Theta_{p_i}, \ni m_j = \sigma_k \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where m_j is *j*-th Mandarin segmented term; e_i is the *i*-th English phrase, and Θ_{p_i} is represented as a Mandarin lexicon set of the English phrase e_i in the chosen bilingual dictionary. The mapping function $\tau\left(m_j, e_i\right)$ has the factor σ_k , which represents *k*-th Mandarin lexicon in Θ_{p_i} . Therefore, if the translation of e_i found in the bilingual dictionary is the same to m_j, $\tau(m_j, e_i)$ is assigned to 1; otherwise, $\tau(m_j, e_i)$ is assigned to 0. And we can obtain the direct lexical mapping sequence

$$\Delta_M = \{a_j^i \mid 0 \le i \le I \text{ and } 0 \le j \le J\} \tag{3}$$

where a_j^i is a mapping referred to as the alignment $i \rightarrow j$ if $\tau(m_j, e_i) = 1$ or $i \rightarrow 0$ and $0 \rightarrow j$ if $\tau(m_j, e_i) = 0$.

If the lexical mapping sequence Δ_M contains more than a particular number, named ε_n, of null mappings ($i \rightarrow 0$ and $0 \rightarrow j$), then the degree of divergence between the sentence pairs [*m*, *e*] becomes significant. Hence, the content of *m* or *e* should be updated to improve the accuracy and effectiveness of exploration of mapping order between word sequences and derivation of transfer mappings. In this paper, we choose to sieve out the divergent sentence pairs from the parallel texts.

4. Experimental Results

Table 1 shows the basic characteristics of the collected parallel texts extended by travel conversation [11]. The Mandarin words in the corpora were obtained automatically using a Mandarin morphological analyzer at CKIP [10] and an English morphological analyzer referred to LinkGrammar [12].

Table 1. Basic characteristics of the collected parallel texts

	Mandarin	English
Number of sentences	11,885	11,885
Total number of words	80,699	66915
Number of word entries	6,278	5,118
Average number of words per sentence	6.79	5.63

The percentage of the various types of divergences for the collected parallel texts is shown on Fig. 2. For the collected corpus of travel conversation, almost two out of three parallel sentences (65 percent) occur the conflational divergence and less than one out of five parallel sentences (19 percent) occur the lexical divergence. In order to assess the effect of translation divergence in the parallel texts, the system also utilizes an alignment training tool called GIZA, which is a program in an EGYPT toolkit designed by the Statistical Machine Translation team [9][1]. Based on segmented Chinese, we use the original GIZA for testing in this paper. In relation to the IBM models in GIZA, this study uses models 1-4 and ten iterations of each training models for the collected corpus. The parallel sentences with various types of divergences are sieved out from the collected corpus and perplexity in IBM original GIZA training model with comparison of sieving various types of divergences is shown on Table 2. The perplexity with sieving thematic divergence is similar to that with sieving structural divergence and the perplexity with sieving morphological divergence is similar to that with sieving lexical divergence. For sieving conflational divergence, a noticeable perplexity reduction is obtained among other types of divergence but the cost is that almost two out of three parallel sentences (65 percent) are sieved out from the collected corpus. Table 3 lists the perplexity of the original parallel sentences and that of the evaluated parallel sentences from GIZA. The results demonstrate that more null mappings can result in higher perplexity, i.e. more translation choices for a lexical item, thus increasing the translation ambiguity and lowering the accuracy of lexical mapping extraction. Two amended translation probabilities with evaluation of ε_n <1 are shown in Table 4. The number of translation choices of "*have*" and "*back*" are reduced from 7 to 4 and 7 to 3, respectively. After evaluating the divergence of each sentence pair in parallel texts and retaining those with ε_n <1, i.e. no null mappings in a sentence pair, the perplexity in the alignment training model can be reduced from 13.65 to 5.18 on average.

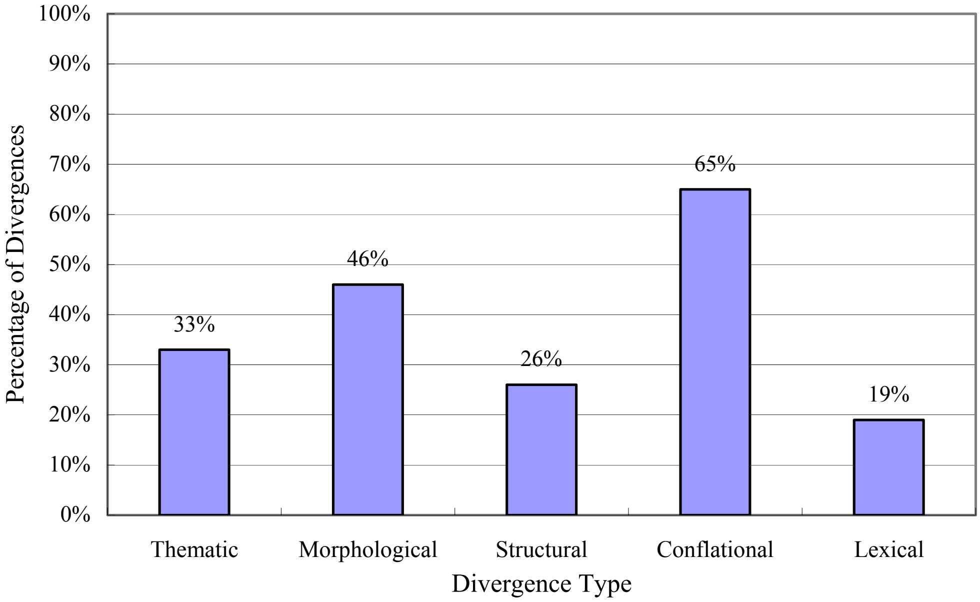

Fig. 2. The percentage of the various types of divergences for the collected parallel texts

[1] This toolkit could be downloaded from http://www.clsp.jhu.edu/ws99/projects/mt/toolkit/

Table 2. Perplexity in IBM original GIZA training model with comparison of sieving various types of divergences.

	Thematic Divergence	Morphological Divergence	Structural Divergence	Conflational Divergence	Lexical Divergence
Model 1	9.91	10.17	9.41	8.46	10.70
Model 2	7.89	10.22	7.84	6.51	9.72
Model 3	8.72	11.83	8.56	7.48	12.35
Model 4	8.65	11.79	8.61	7.44	12.29
Average	8.79	11.00	8.61	7.47	11.26

Table 3. Perplexity in IBM original GIZA training model with comparison of original (ε_n <∞**)** /evaluated parallel sentences.

	ε_n <∞	ε_n <4	ε_n <3	ε_n <2	ε_n <1
No. of (*m*, *e*)	11,885	10,976	9,874	8,618	7,639
Model 1	10.94	10.09	8.26	6.79	5.98
Model 2	12.92	8.52	6.57	5.24	4.43
Model 3	15.39	9.47	7.13	6.21	5.16
Model 4	15.33	9.45	7.11	6.20	5.15
Average	13.65	9.38	7.27	6.11	5.18

Table 4. Examples of two amended English word translation probabilities.

Have			
Translation probability trained with original parallel sentences		Translation probability trained with evaluated parallel sentences	
已經	0.4312746	已經	0.4612446
有	0.346279	有	0.398176
給	0.1231011	給	0.1035049
你	0.0975747	叫	0.0370745
我	0.00146905		
轉告	0.000294352		
在	2.95704e-08		

Back			
Translation probability trained with original parallel sentences		Translation probability trained with evaluated parallel sentences	
回來	0.937283	回來	0.9392834
給	0.0379813	給	0.042981
能	0.01650713	能	0.01871713
錢	0.00786959		
在	2.5874e-06		
轉告	0.000294352		
何時	1.66024e-07		

5. Conclusion

In this work, we discuss one issue of parallel text exploitation, in general, with special emphasis on divergence analysis and processing. Experiments were performed for the languages of Mandarin and

English with the travel conversation corpus of 11,885 sentence pairs. The experimental results show that the analysis and evaluation of divergence for retaining low divergent parallel sentences can reduce the perplexity in IBM translation model averagely from 13.65 to 5.18. For sieving conflational divergence, a noticeable perplexity reduction is obtained among other types of divergence but the cost is that almost two out of three parallel sentences (65 percent) are sieved out from the collected corpus. Future studies will attempt to implement a translation decoder to assess the influence of divergence evaluation on BLEU sore.

References

[1] H. Ney, S. Nießen, F. J. Och, H. Sawaf, C. Tillmann, and S. Vogel, "Algorithms for statistical translation of spoken language," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 1, pp. 24–36, Jan. 2000.

[2] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[3] B. J. Dorr, P. W. Jordan and J. W. Benoit, "A survey of current paradigms in machine translation," in *Advances in Computers*, vol. 49, M. V. Zelkowitz, Ed. Academic Press, 1999.

[4] B. J. Dorr, *Machine translation: A view from the lexicon*. Cambridge, MA: The MIT press, 1993.

[5] B. J. Dorr, "Machine Translation Divergences: A Formal Description and Proposed Solution," *ACL* Vol. 20, No. 4, pp. 597–631, 1994.

[6] A. Menezes and S. D. Richardson, "A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora," *in Proc. Workshop on Data-driven Machine Translation at 39th Annual Meeting of the Association for Computational Linguistics*, 2001, pp. 39–46.

[7] J. F. Wang and S. C. Lin, "Bilingual corpus evaluation and discriminative sentence vector expansion for machine translation," in *Proc. ICAIET*, 2002, pp.117–120.

[8] D. Gupta and N. Chatterjee, "Study of divergence for example based English-Hindi machine translation," in *Proc. STRANS*, 2002, pp. 132-140.

[9] *EGYPT toolkit*, developed by the Statistical Machine Translation team, Center for Language and Speech Processing, Johns-Hopkins University, MD, 1999.

[10] L. L. Chang, "The modality words in modern Mandarin," Chinese Knowledge Information Processing Group, Institute of Information Science Academia Sinica, Taiwan, Tech. Rep. 93-06, 1993.

[11] 徐歡, *旅遊英文 Easy Go*, 廣讀書城出版社, 2001.

[12] D. D. Sleator and D. Temperley, "Parsing English with a Link Grammar," Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-CS-91-196, 1993.

[13] *Dr. Eye 譯典通 6.0*, developed by Inventec Corporation, 2004.