Divergence Analysis and Processing for Mandarin-English Parallel Text Exploitation

Shun-Chieh Lin and Jhing-Fa Wang

Department of Electrical Engineering
National Cheng Kung University
No.1, Dasyue Rd., East District, Tainan City 701, Taiwan, R.O.C.
Tel: 886-6-2757575 ext. 62341 Fax: 886-6-2746867, E-mail : linsj@csie.ncku.edu.tw

ABSTRACT

Previous work shows that the process of parallel text exploitation to extract 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 4.18.

Keywords
Divergence Analysis and Processing, Parallel Text Exploitation

1.0 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 (Brown et al., 1993; Ney et al., 2000). Another class of systems analyzes sentences in parallel texts to obtain transfer structures or rules (Menezes and Richardson, 2001). 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 (Dorr, 1993). 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 (Dorr, 1994; Gupta and Chatterjee, 2002). 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 divergence.

The framework of this paper is as follows. A brief overview of parallel text exploitation is discussed in Section 2. In Section 3, 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 6.

2.0 OVERVIEW OF 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 statistical approach to translation (Brown et al., 1993), among all possible target strings, we will choose the string with the highest probability which is given by Bayes' decision rule as follows:

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

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.

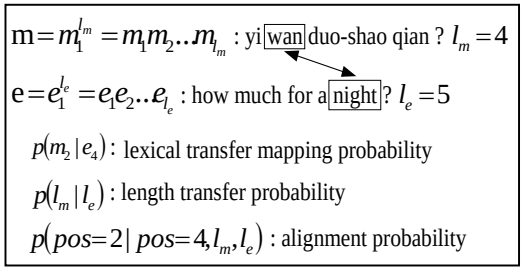

Figure 1: an example for the translation parameters of a sentence pair

However, it is difficult to achieve straightforward 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 divergence.

3.0 DIVERGENCE ANALYSIS AND PROCESSING

3.1 Analysis of Divergence Problems

Dorr's work (Dorr et al., 1999) 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.

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.

E: Is credit card acceptable to them?
C: "ta-men jie-shou xin-yong-ka ma"?
'Do they accept credit card?'

Here, credit card appears in subject position in English and in object position ("xin-yong-ka") in Mandarin; analogously, the object them appears as the subject they ("ta-men").

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 I have your signature here?
C: "qing ni zai zhe-er qian-ming hao ma"?
'Could you sign here?'

In this example, the predicate is nominal (signature) in English but verbal ("qian-ming") in Mandarin.

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 the center.
C: "da-gai zai zhong-jian"
'About in the center.'

Observe that the place object is realized as a noun phrase (the center) in English and as a prepositional phrase ("zai zhong-jian") in Mandarin.

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 have him call me.
C: "qing zhuan-gao ta hui ge dian-hua gei wo"
'Please tell him to give me a call.'

This example illustrates the conflation of a constitution in English that must be overly realized in Chinese: the effect of the action (give me a call) is indicated by the word "hui ge dian-hua gei wo" whereas this information is incorporated into the main verb (call me) in English.

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 can beat 'Phantom of the Opera' "
C: "mei-you she-me bi-de-shang'ge ju mei ying'"
' Nothing can compare with 'Phantom of the Opera'

Here the main verb "beat" in English but as a different verb "bi-de-shang" (to compare with) in Mandarin. Other examples are like "cash", "have", "take", and etc. in English but "dui-huan cheng xian-jin", "zhuan-gao", "zuo", 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 null mappings (1-to-0 or 0-to-1) and non-straightforward mappings. Here, we want to use a simple and straightforward measurement method to analyze the possible null mappings. For example to the Mandarin-English parallel text corpus, given a Mandarin sentence $T_1^J = t_1 t_2 ... t_J$ and an English sentence $P_1^I = p_1 p_2 ... p_I$, direct lexical mappings in the mapping space $\left[T_1^J, P_1^I\right]$ can be extracted using the relevant bilingual dictionary. The mapping function is defined as follows:

$$\tau(t_j, p_i) = \delta(t_j - \sigma_k) = \begin{cases} 1 & \text{if } \exists\, \sigma_k \in \Theta_{p_i}, \ni t_j = \sigma_k \\ \phi_s & \text{if } \forall_i \Theta_{p_i}, \ni t_j \neq \sigma_k, \phi_s : \text{null string} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where t_j is j-th Mandarin segmented term; p_i is the i-th English phrase, and Θ_{p_i} is represented as a Mandarin lexicon set of the English phrase p_i in the chosen bilingual dictionary. The mapping function $\tau(t_j, p_i)$ has the factor σ_k, which represents k-th Mandarin lexicon in Θ_{p_i}. And we can obtain the direct lexical mapping sequence

$$\Delta_M = \left\{ a_j^i \mid 0 \le i \le I \text{ and } 0 \le j \le J \right\} \quad (3)$$

where a_j^i is a mapping referred to as the alignment $i \to j$ and $i \to 0$ ($0 \to j$) if $\tau(t_j, p_i) = \phi_s$.

If the lexical mapping sequence Δ_M contains more than a particular number of null mappings, then the degree of divergence between the sentence pairs $\left[T_1^J, P_1^I\right]$ becomes significant. Hence, the content of T_1^J or P_1^I should be updated to improve the accuracy and effectiveness of exploration of mapping order between word sequences and derivation of transfer mappings.

4.0 EXPERIMENTAL RESULTS

Table 1 shows the basic characteristics of the collected parallel texts extended by travel conversation. The Mandarin words in the corpora were obtained automatically using a Mandarin morphological analyzer at CKIP (Chang, 1993).

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

In order to evaluate the effect of divergence existed in the collected parallel texts, we use the alignments training process in IBM translation model. Therefore, a tool called GIZA, which is a program in EGYPT toolkit developed by the Statistical Machine Translation team. We use 10 iterations of the training models for the collected data. Table 2 shows the perplexity of the

Mandarin text given the English text in the original parallel texts and divergence analyzed parallel texts. The results are revealed that divergence can cause a low occurrence probability of the word sequence in the collected original parallel texts compared with the analyzed parallel texts.

Table 2: Perplexity of IBM translation model.

	Original Parallel Texts	Analyzed Parallel Texts
Model 1	10.94	5.09
Model 2	12.92	3.53
Model 3	15.39	4.06
Model 4	15.33	4.05
Average	13.65	4.18

5.0 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 processing of divergence can reduce the perplexity in IBM translation model averagely from 13.65 to 4.18.

6.0 REFERENCES

P. F. Brown, 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.

Herrmann Ney, Sonja Nießen, Franz Josef Och, Hassan Sawaf, Christoph Tillmann, and Stephan Vogel. (2000). Algorithms for Statistical Translation of Spoken Language. IEEE Transaction on Speech and Audio Processing, Vol. 8, No. 1:24-36.

Bonnie J. Dorr, Pamela W. Jordan and John W. Benoit. (1999). A Survey of Current Paradigms in Machine Translation. In Advances in Computers, vol. 49, Academic Press.

Bonnie Jean Dorr (1993). Machine Translation: A View from the Lexicon, The MIT press.

Bonnie J. Dorr (1994). Machine Translation Divergences: A Formal Description and Proposed Solution, ACL Vol. 20, No. 4, pp. 597-631.

Arul Menezes and Stephen D. Richardson (2001). A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora, In Proceedings of the Workshop on Data-driven Machine Translation at 39th Annual Meeting of the Association for Computational Linguistics (ACL2001), pp: 39-46.

Jhing-Fa Wang and Shun-Chieh Lin. (2002). Bilingual Corpus Evaluation and Discriminative Sentence Vector Expansion for Machine Translation, Int. Conf. on Artificial Intelligence in Engineering and Technology (ICAIET-2002), University Malaysia Sabah, Kota Kinabalu, Malaysia.

Deepa Gupta and Niladri Chatterjee (2002). Study of divergence for Example based English-Hindi Machine Translation», Proceedings of Symposium on Translation Support Systems (STRANS-2002), pp: 132-140.

The EGYPT toolkit developed by the Statistical Machine Translation team, WS'99, CLSP/JHU. http://www.clsp.jhu.edu/ws99/projects/mt

L. L. Chang, “The Modality Words in Modern Mandarin,” CKIP Tech. Rep. No. 93-06, 1993.