Chinese-Japanese Clause Alignment

Xiaojie Wang [1] Fuji Ren [2]
[1]School of Information Engineering, Beijing University of Posts and Telecommunications
Beijing, China, 100876
[2]Department of Information Science & Intelligent Systems, Tokushima University
Tokushima, Japan,
xjwang@bupt.edu.cn ren@is.tokushima-u.ac.jp

Abstract. Bi-text alignment is useful to many Natural Language Processing tasks such as machine translation, bilingual lexicography and word sense disambiguation. This paper presents a Chinese-Japanese alignment at the level of clause. After describing some characteristics in Chinese-Japanese bilingual texts, we first investigate some statistical properties of Chinese-Japanese bilingual corpus, including the correlation test of text lengths between two languages and the distribution test of length ratio data. We then pay more attention to *n*-*m*(*n*>1 or *m*>1) alignment modes which are prone to mismatch. We propose a similarity measure based on Hanzi characters information for these kinds of alignment modes. By using dynamic programming, we combine statistical information and Hanzi character information to find the overall least cost in aligning. Experiments show our algorithm can achieve good alignment accuracy.

1 Introduction

Text alignment is an important task in Natural Language Processing (NLP). It can be used to support many other NLP tasks. For example, it can be utilized to construct statistical translation models (Brown et al. 1991), and to acquire translation examples for example-based machine translation (Kaji et al. 1992). It can be helpful in bilingual lexicography (Tiedemann 2003). It is also used to improve monolingual word sense disambiguation (Diab and Resnik 2002).

The approaches to text alignment can be classified into two types: statistical-based and lexical-based. The statistical-based approaches rely on non-lexical information (such as sentence length, co-occurrence frequency, etc.) to achieve an alignment task. As illustrated in the research of Gale and Church (1991) for the sentence-level alignment, they start from the fact that the length of a source text sentence is highly correlated with the length of its target text translation.

The method proposed in Kay and Roscheisen (1993) is based on the assumption that in order for the sentences in a translation to correspond, the words in them must also correspond. Their method makes use of lexical anchor points to lead an alignment at the sentence level.

It has been shown that different language pairs are in favor of different information in alignment. For example, Wu (1994) found that the sentence-length correlation between English and Chinese is not as good as between English and French. Also, there is less cognate information between Chinese-English pair than that in English-French