Bilingual Segmentation for Alignment and Translation

Chung-Chi Huang, Wei-Teh Chen, and Jason S. Chang

Information Systems and Applications, NTHU, HsingChu, Taiwan 300 R.O.C
{u901571, weitehchen, jason.jschang}@gmail.com

Abstract. We propose a method that bilingually segments sentences in languages with no clear delimiter for word boundaries. In our model, we first convert the search for the segmentation into a sequential tagging problem, allowing for a polynomial-time dynamic-programming solution, and incorporate a control to balance monolingual and bilingual information at hand. Our bilingual segmentation algorithm, the integration of a monolingual language model and a statistical translation model, is devised to tokenize sentences more suitably for bilingual applications such as word alignment and machine translation. Empirical results show that bilingually-motivated segmenters outperform pure monolingual one in both the word-aligning (12% reduction in error rate) and the translating (5% improvement in BLEU) tasks, suggesting monolingual segmentation is useful in some aspects but, in a sense, not built for bilingual researches.

1. Introduction

A statistical translation model (STM) is a model that, relied on lexical information or syntactic structures of languages involved, decodes the process of human translation and that, in turn, detects most appropriate word correspondences in parallel sentences. Ever since the pioneer work of (Brown et al., 1993), the field of STMs has drawn myriads of attention. Some researchers exploited Hidden Markov models to approach relatively monotonic word-aligning problems in similarly-structured language pairs