Context-Based Sentence Alignment in Parallel Corpora

Ergun Biçici

Koç University
Rumelifeneri Yolu 34450
Sariyer, Istanbul, Turkey
ebicici@ku.edu.tr

Abstract. This paper presents a language-independent context-based sentence alignment technique given parallel corpora. We can view the problem of aligning sentences as finding translations of sentences chosen from different sources. Unlike current approaches which rely on pre-defined features and models, our algorithm employs features derived from the distributional properties of words and does not use any language dependent knowledge. We make use of the context of sentences and the notion of Zipfian word vectors which effectively models the distributional properties of words in a given sentence. We accept the context to be the frame in which the reasoning about sentence alignment is done. We evaluate the performance of our system based on two different measures: sentence alignment accuracy and sentence alignment coverage. We compare the performance of our system with commonly used sentence alignment systems and show that our system performs 1.2149 to 1.6022 times better in reducing the error rate in alignment accuracy and coverage for moderately sized corpora.

Keywords
sentence alignment, context, Zipfian word vectors, multilingual

1 Introduction

Sentence alignment is the task of mapping the sentences of two given parallel corpora which are known to be translations of each other to find the translations of corresponding sentences. Sentence alignment has two main burdens: solving the problems incurred by a previous erroneous sentence splitting step and aligning parallel sentences which can later be used for machine translation tasks. The mappings need not necessarily be 1-to-1, monotonic, or continuous. Sentence alignment is an important preprocessing step that affects the quality of parallel text.

A simple approach to the problem of sentence alignment would look at the lengths of each sentence taken from parallel corpora and see if they are likely to be translations of each other. In fact, it was shown that paragraph lengths for the English-German parallel corpus from the economic reports of Union Bank of Switzerland (UBS) are highly correlated with a correlation value of 0.991 [1]. A more complex approach would look at the neighboring sentence lengths as well. Our approach is based on this knowledge of context for given sentences from each corpus and the knowledge of distributional features of words, which we name Zipfian word vectors, for alignment purposes. A Zipfian