

Sentence Alignment of Hungarian-English Parallel Corpora Using a Hybrid Algorithm

Krisztina Tóth, Richárd Farkas, and András Kocsor*

Abstract

We present an efficient hybrid method for aligning sentences with their translations in a parallel bilingual corpus. The new algorithm is composed of a length-based and anchor matching method that uses Named Entity recognition. This algorithm combines the speed of length-based models with the accuracy of anchor finding methods. The accuracy of finding cognates for Hungarian-English language pair is extremely low, hence we thought of using a novel approach that includes Named Entity recognition. Due to the well selected anchors it was found to outperform the best two sentence alignment algorithms so far published for the Hungarian-English language pair.

Keywords: sentence segmentation, sentence alignment, length-based alignment, hybrid method, Named Entity recognition, anchor, cognates, dynamic programming

1 Introduction

In the last few years parallel corpora have become evermore important in natural language processing. There are many applications which could benefit from parallel texts like (i) automatic translation programs (as machine learning algorithms) that are used as training databases, (ii) translation support tools that can be obtained from them (translation memories, bilingual dictionaries) and (iii) Cross Language Information Retrieval methods. These applications require a high-quality correspondence of text segments like sentences. Sentence alignment establishes relations between sentences of a bilingual parallel corpus. This relation may not have just a one-to-one correspondence between sentences; there could be a many-to-zero (in the case of insertion or deletion), many-to-one (if there is a contraction or an expansion) or even many-to-many alignments.

Various methods have been proposed to solve the sentence alignment task. These are all derived from two main classes: length-based and lexical methods,

*Research Group on Artificial Intelligence of the Hungarian Academy of Sciences and University of Szeged H-6720 Szeged, Aradi vértanúk tere 1., Hungary, E-mail: {tothkr, rfarkas, kocsor}@inf.u-szeged.hu