Text Alignment Methods, Hybrid Paragraph and Sentence Alignment Technique

Gábor Pohl

Péter Pázmány Catholic University, Budapest
Department of Information Technology
pohl@morphologic.hu

Keywords: text alignment, paragraph alignment, sentence alignment, statistically filtered anchors, hybrid alignment technique

Abstract

This paper contains an introduction to text alignment methods and presents a new state of the art hybrid text alignment technique.
In the introductory part, alignment is defined in a generic way as a one-to-one correspondence of alignment units, where an alignment unit is a set of text units (paragraphs, sentences, etc.) that are not contained in other alignment units. After the explanation of the definition two different alignment strategies—the strategy of character length based complete alignment and the strategy of anchor based partial alignment—are described.

After the introduction to alignment techniques a new hybrid method is presented. In order to achieve high precision alignment of text pairs (with omitted and inserted text units) the char-length based method is combined with the use of statistically filtered anchors. In order to find the same anchors in both texts, anchor candidates are also considered in their lemmatized forms. The methods described by Ribeiro et al. have been used to filter anchor candidates. Integrating the two alignment strategies in one alignment framework, we present a new heuristic formula that defines the *revenue of anchors* that can be used in the dynamic programming algorithm used by the length based method.