

# Evaluation of Alignment Methods for HTML Parallel Text

Enrique Sánchez-Villamil, Susana Santos-Antón,  
Sergio Ortiz-Rojas, and Mikel L. Forcada

Transducens group, Departament de Llenguatges i Sistemes Informàtics  
Universitat d'Alacant, E-03071 Alacant, Spain  
{esvillamil, ssantos, sortiz, mlf}@dlsi.ua.es

**Abstract.** The Internet constitutes a potential huge store of parallel text that may be collected to be exploited by many applications such as multilingual information retrieval, machine translation, etc. These applications usually require at least sentence-aligned bilingual text. This paper presents new aligners designed for improving the performance of classical sentence-level aligners while aligning structured text such as HTML. The new aligners are compared with other well-known geometric aligners.

## 1 Introduction

Many machine translation applications are based on machine learning on parallel corpora. The amount of parallel text required to obtain accurate translations using these applications is quite high (up to hundreds of megabytes) although it seems possible to generate such large corpora using the Internet. The utility of the corpora increases dramatically when they are aligned at sentence or word levels.

A number of sentence-alignment approaches have been developed during the last years. The first effective approach at aligning large corpora was based on modeling the relationship between the lengths of sentences that are mutual translations (Brown et al., 1991; Gale and Church, 1991, 1993). Chen (1993) used a different approach, based on lexical information to improve accuracy, but it was slower than sentence-length-based algorithms. Some years later, Melamed (1996) developed a method based on word correspondences and supported by external linguistic knowledge.

All these aligners are designed to work with text segmented in sentences. In our case, collections of hundreds of megabytes of downloaded webpages, which are not segmented, have to be aligned at sentence-level. These pages are turned into XML<sup>1</sup> using the `tidy` program,<sup>2</sup> which may be used to turn HTML into XHTML.<sup>3</sup>

---

<sup>1</sup> <http://www.w3.org/TR/2004/REC-xml-20040204/>

<sup>2</sup> <http://www.w3.org/People/Raggett/tidy/>

<sup>3</sup> XHTML is a stricter and cleaner XML-version of HTML.

The aligners proposed in this paper are being used to generate a large collection of aligned text corpora. The corpora will be segmented, and segments will be aligned to build translation units. The resulting translation units may be used to train translation applications.

In particular, this paper presents a type of aligners that combine sentence-splitting and alignment generation, and take advantage of the structured nature of web documents to improve the accuracy of sentence-aligned text in the absence of linguistic knowledge. The aligners are compared to classical approaches in the experiments.

## 2 Notation

In this paper, we define the *alignment* as a sequence of edit operations, that is, a sequence of insertions, deletions and substitutions of segments.<sup>4</sup> Let  $L = (l_1, l_2, \dots, l_{|L|})$  and  $R = (r_1, r_2, \dots, r_{|R|})$  be two parallel texts split in segments and  $S = (s_1, s_2, \dots, s_{|S|})$ , a sequence of edit distance operations, where  $s_i$  can be an insertion ( $m_i$ ), a deletion ( $m_d$ ) or a substitution ( $m_s$ ) of a segment. It is straightforward to obtain the aligned segment pairs  $(l_i, r_i)$  using the edit distance sequence. We define  $A$  as the function returning the edit-distance alignment of two texts, so that  $A(L, R) \longrightarrow S$ .

Additionally, we define the alignment distance  $D$  that is considered as a measure of the similarity of the texts that have been aligned. The distance  $D$  is defined as the addition of the differences in length of all aligned segments:

$$D(S) = \sum_{i=1}^{|S|} \text{abs}(|l_i| - |r_i|) \quad (1)$$

where bars  $|\cdot|$  are used to represent the length of a text segment. The  $m_i$  and  $m_d$  operations where either  $l_i$  or  $r_i$  would be the empty string are also taken into account.

## 3 Classical Geometric Aligners

Geometric aligners are based only in geometric properties of the documents, such as sentence lengths, word lengths, paragraph lengths, etc. They are fully independent of the language because they do not use linguistic information.

Classical geometric aligners were designed to align plain text segmented in sentences. However, they can be adapted to marked-up corpora, such as XHTML, in several ways. The simplest approach would be the removal of all tags in both sides, so that a pair of plain texts would be obtained and would then be split; such aligner is called *Remover*. A more elaborate algorithm would require the substitution of some tags by sentence boundaries<sup>5</sup> (and the removal of the rest

<sup>4</sup> This definition induces a monotone alignment.

<sup>5</sup> The tags replaced are `hr`, `br`, `p`, `li`, `ul`, `ol`, `tr`, `td`, `th`, `div`.