Evaluation of alignment methods for HTML parallel text[1]

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

Transducens Group, Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant (Spain)

FinTAL – Turku/Åbo – August 23, 2006

[1] Funded through Spanish Govt. grant TIC2003-08681-C02-01.

Contents

1. Introduction
2. Geometric alignment: edit-distance
3. The proposal: using (X)HTML structure
4. Experiments
5. Concluding remarks and future work

Parallel texts on the Internet

- Corpus-based machine translation relies on the availability of large parallel text corpora.
- The Internet contains abundant parallel text that may be harvested for that purpose.

The need for alignment

- In particular, translation learning or computer-aided translation usually require **aligned** corpora (for instance, sentence-aligned corpora).
- Aligners usually require the left and right text to be previously **segmented** in sentences.
    - *Geometrical* aligners use the relationships between the lengths of left and right sentences which are mutual translation (Brown et al. 1991, Gale & Church, 1991, 1993).
    - Alignments may improve if *linguistic* (e.g. lexical) information is added to them.

Using document structure to align

Our focus is on aligners which are

- *linguistically independent* (no linguistic information is required)
- but take advantage of *document structure* (HTML tag structure)

Our main hypothesis: taking document structure into account improves alignment.

Geometric alignment: edit distance /1

- $L = (l_1, l_2, \ldots, l_{|L|})$, left text split in $|L|$ segments.
- $R = (r_1, r_2, \ldots, r_{|R|})$, left text split in $|R|$ segments.
- R may be obtained from L through a sequence $S = (s_1, s_2, \ldots, s_{|S|})$ of *edit operations*:
    - segment insertions,
    - segment deletions, or
    - segment substitutions

Geometric alignment: edit distance /2

- For L and R there is an edit operation sequence

$$A(L,R) = S^* = (s^*_1, s^*_2, \ldots, s^*_{|S^*|}),$$

which is *optimal* in the sense that

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

is the minimum over all possible edit operation sequences.

Sentence splitting heuristics

Sentence splitting is performed at “!”, “?” and “.”.
For “.”s, validate using a threshold (−0.2) and empirical scores:

Characters before	Characters after	Points
-	number	−0.5
-	a blank	+0.5
-	a non-capital letter	−0.2
-	another “.”	−0.5
-	blank + capital letter	+0.5
-	a blank + non-capital letter	−0.2
a capital letter	-	−0.5
a word of 3 characters or less	-	−0.5
a blank	-	+0.2
a “' ” or “"” character	a “' ” or “"” character	−0.5
another “.”	-	+0.4

Geometric alignment of (X)HTML texts /1

Two baseline approaches (for comparison):

- *Remover*: remove tags, split in sentences, and align.
- *Replacer*: replace `hr`, `br`, `p`, `li`, `ol`, `ul`, `tr`, `td`, `th` and `div` by sentence boundaries, split, and align.

These baselines will be used to evaluate a family of tag-driven alignment algorithms.

Geometric alignment of (X)HTML texts /2

We use the following classification of (X)HTML tags:

- *Structural* tags: `blockquote, body, caption, col, colgroup, dd, dir, div, dl, dt, h1—h6, head, hr, html, li, menu, noframes, noscript, ol, optgroup, option, p, q, select, table, tbody, td, tfoot, th, thead, tr, ul.`
- *Format* tags: `abbr, acronym, b, big, center, cite, code, dfn, em, font, i, pre, s, small, span, strike, strong, style, sub, sup, tt, u.`
- *Content* tags: `a, area, fieldset, form, iframe, img, input, isindex, label, legend, map, object, param, textarea, title.`
- *Irrelevant* tags (will be removed): `address, applet, base, basefont, bdo, br, button, del, ins, kbd, link, meta, samp, script, var.`

Geometric alignment of (X)HTML texts /3

How is tag information used to align?

- Forbidden alignments:
  - Tag–text segment (and vice versa)
  - Structural tag–tag of another class
  - Format tag–tag of another class
  - Content tag–different content tag
- Structural tag–structural tag alignment expensive unless they are the same tag.
- Format tag–format tag alignment cheap
- The cost of text alignments should be lower than tag–tag costs.

Geometric alignment of (X)HTML texts /4

Costs of edit operations (empirically adjusted):

	Insert	<strct>	<frmt>	<cntnt>	**Text**
Delete	-	1	0.75	1.25	$0.01\ \lvert l \rvert$
<strct>	1	1.5	1.75	H	H
<frmt>	0.75	1.75	0.4	H	H
<cntnt>	1.25	H	H	H	H
Text	$0.01\ \lvert r \rvert$	H	H	H	Δ

- H is large enough for that operation to be always avoided
- Δ will be different in each variant of the algorithm (next slide)

Geometric alignment of (X)HTML texts /5

Three variants of the tag-driven alignment:

2-in-1: Split texts in tags and text segments and then align. Uses $\Delta = 0.015\ (\mathrm{abs}(|l| - |r|))$ (factor 0.015 empirically obtained)

2-steps-L: Split text by tags; align tags and text segments, split text segments in sentences and align. Uses $\Delta = 0.015\ (\mathrm{abs}(|l| - |r|))$.

2-steps-AD: Same as 2-steps-L but uses $\Delta = 0.01\ D(A(l, r))$ (the alignment distance between the text segments; factor 0.01 empirically determined).

Availability

An implementation of the tag-driven aligners described is available as open-source software (under the GPL license) from `tag-aligner.sourceforge.net`.

Corpora

Three corpora:

- 86.1 MB of parallel HTML downloaded using Bitextor (`http://www.sf.net/projects/bitextor/`) from the bilingual `es–ca` daily `http://www.elperiodico.com` ; [“easy”: 2.14% sentences align to null].
- a small fragment of the Quixote (196 kB: `en`, `es`) from a digital Library, `http://www.cervantesvirtual.com/` [“medium”: 19.08% sentences align to null]
- Help texts of the mIRC program (96 kB: `es`, `pt`, `it`, `ca`, `gl`). [“hard”: 26.42% sentences aligned to null]

Building a reference alignment

Reference alignment obtained by post-editing the automatic alignment:

- Human editor corrects incorrect alignments into correct ones.
- Human editor seldom splits (“refines”) a correct alignment. Just in case, concatenation of reference alignments will be allowed during evaluation (not common).

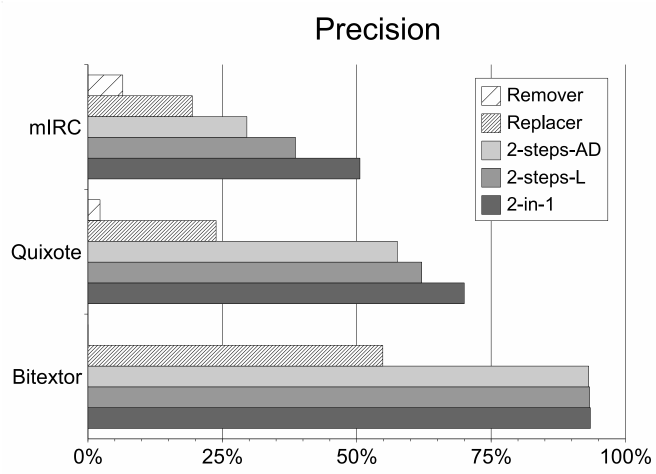
Metrics

$$\text{precision} = \frac{\text{\# correct alignments}}{\text{\# proposed alignments}}$$

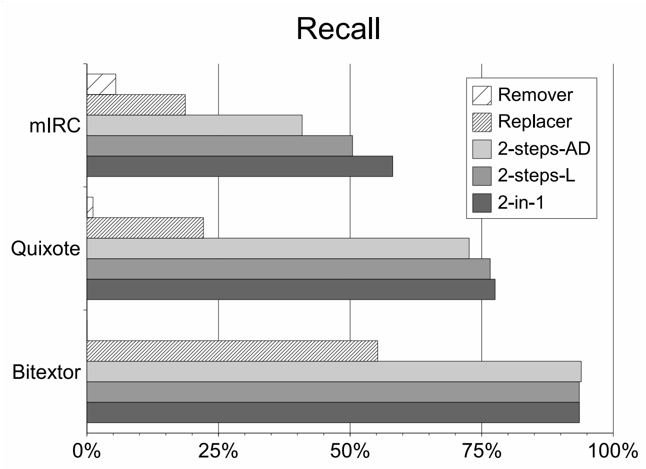
$$\text{recall} = \frac{\text{\# correct alignments}}{\text{\# reference alignments}}$$

$$F = 2 \times \frac{\text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

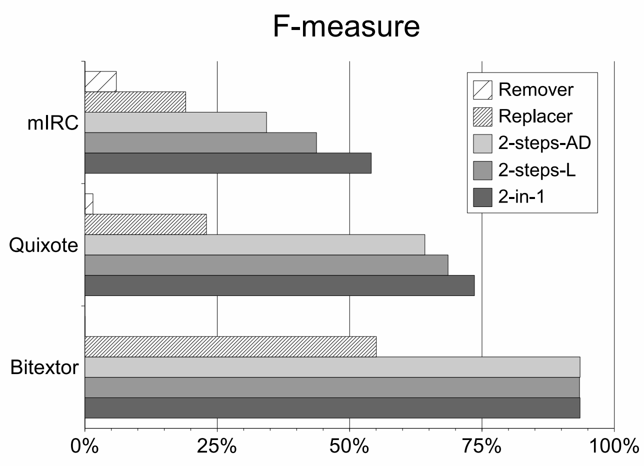
Results: precision



Results: recall



Results: *F*-measure



Concluding remarks

- Aligners using (X)HTML tag information are clearly better than the basic geometric aligners. . .
  . . . at only a twofold increase in processing time.
- The best is the 2-in-1 aligner (the one that splits texts at sentence boundaries *and* tags and then aligns).
- As expected, results are worse for “harder” bitext corpora.

Future work

- Incorporating the same tag-based strategies into existing open-source computer-aided translation tools:
  - the `bitext2tmx` text aligner (`bitext2tmx.sf.net`).
  - the `OmegaT` computer-aided translation tool (`omegat.sf.net`)
- Extending the aligner to other XML-based formats (e.g., DocBook, OpenDocument).
- Task-oriented evaluation of automatically-generated TMX files in real computer-aided translation applications.
- Developers welcome at `tag-aligner.sf.net`!