Two Tools for Creating and Visualizing Sub-sentential Alignments of Parallel Text

Ulrich Germann
University of Toronto
germann@cs.toronto.edu

Abstract

We present two web-based, interactive tools for creating and visualizing sub-sentential alignments of parallel text. *Yawat* is a tool to support distributed, manual word- and phrase-alignment of parallel text through an intuitive, web-based interface. *Kwipc* is an interface for displaying words or bilingual word pairs in parallel, word-aligned context.

A key element of the tools presented here is the interactive visualization: alignment information is shown only for one pair of aligned words or phrases at a time. This allows users to explore the alignment space interactively without being overwhelmed by the amount of information available.

1 Introduction

Sub-sentential alignments of parallel text play an important role in statistical machine translation (SMT). They establish which parts of a sentence correspond to which parts of the sentence's translation, and thus form the basis of a compositional approach to translation that models the translation of a sentence as a sequence of individual translation decisions for basic units of meaning. The simplest assumption is that typographic words, i.e., strings of letters delimited by punctuation and white space, constitute the basic units of translation. In reality, of course, things are more complicated. One word in one language may have to be translated into several in the other or not at all, or several words may form a conceptual unit that cannot be translated word for word. Because of its central role in building machine translation systems and because of the complexity of the task, sub-sentential alignment of parallel corpora continues to be an active area of research (e.g., Moore *et al.*, 2006; Fraser and Marcu, 2006), and this implies a continuing demand for manually created or human-verified gold standard alignments for development and evaluation purposes.

We present here two tools that are designed to facilitate the process and allow human inspection of automatically aligned parallel corpora for the study of translation. The first is a web-based interface for manual sub-sentential alignment of parallel sentences. The second is an extension of the traditional keywords-in-context tools to the bilingual case. A distinctive feature of both tools is that they are based on an interactive process. Rather than showing all alignment information at once, they hide most information most of the time and visualize alignment information only selectively and only on demand.

2 Visualization schemes for sub-sentential text alignment information

In this section, we briefly review existing visualization schemes for word-level alignments.

2.1 Drawing lines

Word alignment visualization by drawing lines is shown in Figure 1. This visualization technique has several limitations.

- The parallel text cannot be wrapped easily. Each sentence has to be represented as a straight line or column of text. If the word

Proceedings of the Linguistic Annotation Workshop, pages 121–124,
Prague, June 2007.

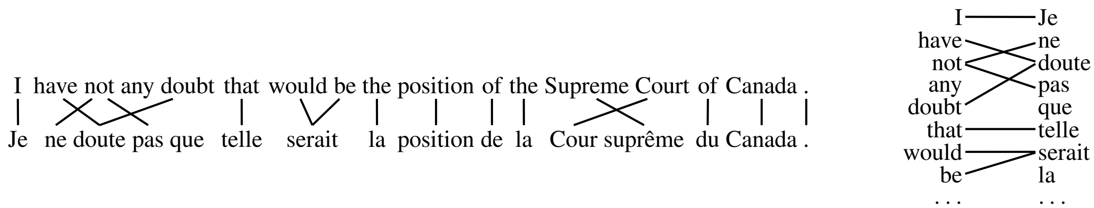

Figure 1: Visualization of word alignments by drawing lines.

alignment is known, it may be possible to pre-segment the parallel text into smaller blocks of text such that all word alignment links are contained within these blocks of text. For manual word alignment from scratch, this is impossible, for lack of prior word alignment information. In consequence, the sentence pair often will not fit on the computer screen entirely, so that users have to scroll back and forth to view and create alignment links.

- Especially when the two aligned sentences show differences in word order, many of the lines representing word alignments will cross one another, leading to a cluttered and hard-to-follow display.

- There is no good way to represent the alignment on the phrase level, especially when the phrases contain gaps. If the phrases involved are contiguous, we can use brackets or boxes to group words into phrases, but this does not work for phrases that contain gaps. Another way to visualize phrase alignments is to link each word in each of the two phrases with each word in the respective other phrase. This acerbates the aforementioned problem of visual clutter.

2.2 Alignment matrices

Alignment matrices such as the one shown in Figure 2 map the words of one sentence onto the rows and the words of the other sentence onto the columns of a two-dimensional table. Each cell (r, c) in the table represents a potential alignment between the word in the r-th position of the first sentence and the word in the c-th position in the second sentence. If the two words are in fact aligned, the respective cell contains a dot, otherwise it is empty. This technique allows the visualization of phrase-level alignments even of discontinuous phrases (by filling the cells representing the cross-product of the two sets of words involved). Fitting the matrix for pairs of long sentences onto the screen is still a problem, however.

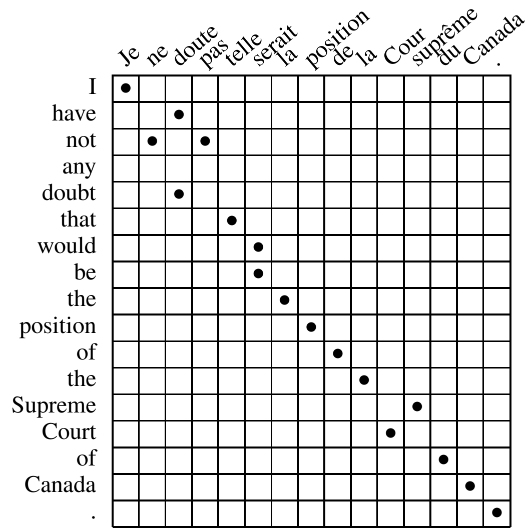

Figure 2: Visualization of word alignments with an alignment matrix.

2.3 Coloring

A third way of visualizing word alignments is the use of colors. This technique has two draw-backs. First, it may be difficult to find enough colors that are easily distinguished to mark up all alignments in pairs of long sentences, and second, actually tracking alignments is tedious and requires a lot of concentration.

2.4 Interactive visualization

Our solution to the visualization problem is to take an interactive approach. We use the coloring approach, but use only one or two colors to mark up

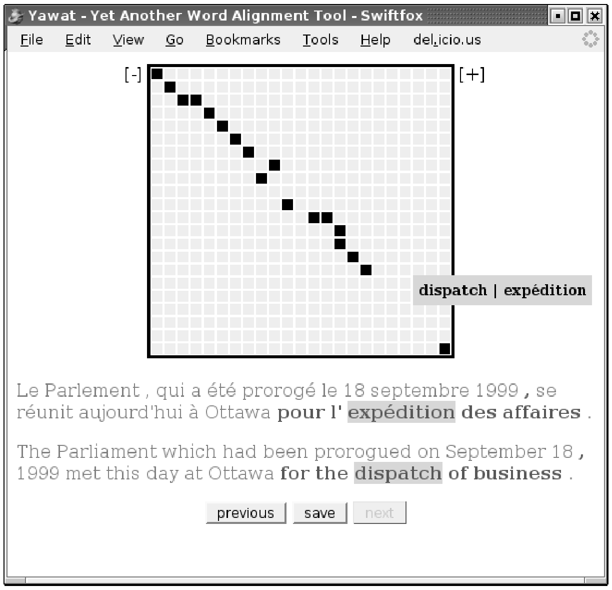

Figure 3: Manual word alignment with *Yawat*. The image shows the state of the screen with the mouse hovering over the alignment matrix cell corresponding to *dispatch* ↔ *expédition*. A click onto the cell links the two words.

alignment pairs, and we mark up alignment pairs only one at a time. By positioning the mouse pointer over a word of interest, the user indicates which alignment he or she would like to see. All other alignments are hidden.

3 The tools

3.1 Yawat

Yawat (***Y**et **A**nother **W**ord **A**lignment **T**ool*) is a tool for manual alignment of parallel sentences. It consists of a cgi-script responsible for retrieving and storing sentence pairs and their alignments from a database on the server side and marking them up in HTML, and client-side functionality that handles the interactive aspects of word-alignment and display and reports changes back to the server-side script.

The user interface combines alignment matrix visualization with interactive colorization. Figure 3 shows the typical *Yawat* interface. The alignment matrix on top gives a birds-eye view of the alignment relations in the sentence. If the mouse is positioned over one of the cells, a tool-tip window pops up showing the row and column labels of the respective cell. If the cell is 'active' (i.e., represents part of an alignment relation), the corresponding alignment pair is highlighted in the text section below. Rows and columns of the alignment matrix are deliberately not labeled so that the alignment matrix can be kept small. Its size is adjustable via the [–] and [+] buttons to its left and right.

The text section below the matrix shows the actual sentence pair. Moving the mouse over an aligned word highlights the respective alignment pair in the text as well as the corresponding cells in the matrix.

The tool was designed to minimize the number of mouse clicks and mouse travel necessary to align words. Clicking on an empty cell in the matrix aligns the respective words. The effect of clicking on an active cell depends on whether the cell represents an exclusive link between two single words, or is part of a larger alignment group. In the former case, the link is simply removed, in the latter, the respective alignment group is opened for editing. Once an alignment group is open for editing, a left-click with the mouse adds or removes words. Selecting a word that is currently part of another alignment group automatically removes it from that group. An alignment group is closed by a right-click on one of its members. A right click on a non-member adds it to the group and then closes the group for editing. This allows us to perform single word alignments with two simple mouse clicks: left-click on the first word and right click on the second, without the need to move the mouse on a visual 'link words' button in the interface.

Unaligned text in the sentence pair is represented in red, aligned text in gray. This allows the annotator to immediately spot unaligned sections without having to refer to the alignment matrix or to scan the text with the mouse to find unaligned words.

We have not performed a formal user study, but we have found the tool very efficient in our own experience.

3.2 Kwipc

Kwipc (***K**ey **W**ords **I**n **P**arallel **C**ontext*) uses the same interactive visualization technique to display word alignments for multiple sentence pairs. It currently uses a very simple search interface that allows the user to specify regular expressions for one or both of the sentences in the sentence pair. The server-side cgi-script searches the corpus lin-

Table 1: Word alignment visualization and editing tools

name	visualization	editing
Cairo[a]	lines	no
Alpaco[b]	lines	yes
Lingua-AlignmentSet[c]	matrix	no
UMIACS WA Interface[d]	lines	yes
HandAlign[e]	lines	yes
Ilink[f]	static colors	yes
UPlug[g]	matrix	yes
ICA[h]	matrix	yes
ReWrite Decoder	interactive, colors	no
Yawat	matrix, interactive, colors	yes
Kwipc	interactive, colors	no

[a] `http://www.clsp.jhu.edu/ws99/projects/mt/toolkit/`

[b] `http://www.d.umn.edu/~tpederse/parallel.html`

[c] `http://gps-tsc.upc.es/veu/personal/lambert/\newlinesoftware/AlignmentSet.html`

[d] `http://www.umiacs.umd.edu/~nmadnani/\newlinealignment/forclip.htm`

[e] `http://www.cs.utah.edu/~hal/HandAlign/`

[f] `http://www.ida.liu.se/~nlplab/ILink/`

[g] `http://stp.ling.uu.se/cgi-bin/joerg/Uplug`

[h] Tiedemann (2006)

early and returns a list of marked-up sentence pairs that contain matching expressions (which are highlighted in red) and provides the same interactive alignment visualization as *Yawat*. For lack of space, we cannot provide a screen shot here.

4 Related work

There are numerous tools available for the visualization and creation of word alignments, most of which are listed on Rada Mihalcea's web site on word alignment at `http://www.cs.unt.edu/~rada/wa/`. A comparison of these tools is shown in Table 1. Most tools use line drawing or alignment matrices for visualization. Only *Ilink* (Ahrenberg *et al.*, 2002) relies on colors to visualize alignments, but it implements a static colorization scheme. The interactive visualization scheme was first used in the HTML output of the *ISI ReWrite Decoder*[1], but the formatting used there relies on an obsolete Document Object Model and is not functional any more. The use of different colors to distinguish aligned and unaligned sections of text can also be found in *HandAlign*.

5 Conclusion

We have presented two web-based tools that use an interactive visualization method to display word- and phrase-alignment information for parallel sentence pairs, thus reducing visual clutter in the display and providing users with focussed access to the alignment information they are actually interested in. The editing tool *Yawat* was designed to minimize unnecessary scrolling, mouse clicks and mouse travel to provide the annotator with an efficient tool to perform manual word- and phrase-alignment of parallel sentences. Delivery of the application through the web browser allows collaborative alignment efforts with a central repository of alignments and without the need to install the software locally.

6 Availability

The tools are available at `http://www.cs.toronto.edu/compling/Software`.

References

Ahrenberg, Lars, Mikael Andersson, and Magnus Merkel. 2002. "A system for incremental and interactive word linking." *Proc. LREC 2002*, 485–490. Las Palmas, Spain.

Fraser, Alexander and Daniel Marcu. 2006. "Semi-supervised training for statistical word alignment." *Proc. COLING-ACL 2006*, 769–776. Sydney, Australia.

Moore, Robert C., Wen-tau Yih, and Andreas Bode. 2006. "Improved discriminative bilingual word alignment." *Proc. COLING-ACL 2006*, 513–520. Sydney, Australia.

Tiedemann, Jörg. 2006. "ISA & ICA — Two web interfaces for interactive alignment of bitexts." *Proc. LREC 2006*. Genoa, Italy.

[1] `http://www.isi.edu/publications/licensed-sw/rewrite-decoder/index.html`