Yawat: Yet Another Word Alignment Tool

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

Abstract

Yawat[1] is a tool for the visualization and manipulation of word- and phrase-level alignments of parallel text. Unlike most other tools for manual word alignment, it relies on dynamic markup to visualize alignment relations, that is, markup is shown and hidden depending on the current mouse position. This reduces the visual complexity of the visualization and allows the annotator to focus on one item at a time. For a bird's-eye view of alignment patterns within a sentence, the tool is also able to display alignments as alignment matrices. In addition, it allows for manual labeling of alignment relations with customizable tag sets. Different text colors are used to indicate which words in a given sentence pair have already been aligned, and which ones still need to be aligned. Tag sets and color schemes can easily be adapted to the needs of specific annotation projects through configuration files. The tool is implemented in JavaScript and designed to run as a web application.

1 Introduction

Sub-sentential alignments of parallel text play an important role in statistical machine translation (SMT). Aligning parallel data on the word- or phrase-level is typically one of the first steps in building SMT systems, as those alignments constitute the basis for the construction of probabilistic translation dictionaries. Consequently, considerable effort has gone into devising and improving automatic word alignment algorithms, and into evaluating their performance (e.g., Och and Ney, 2003; Taskar *et al.*, 2005; Moore *et al.*, 2006; Fraser and Marcu, 2006, among many others). For the sake of simplicity, we will in the following use the term "word alignment" to refer to any form of alignment that identifies words or groups of words as translations of each other.

Any explicit evaluation of word alignment quality requires human intervention at some point, be it in the direct evaluation of candidate word alignments produced by a word alignment system, or in the creation of a gold standard against which candidate word alignments can be compared automatically. This human intervention works best with an interactive, visual interface.

2 Word alignment visualization

Over the years, numerous tools for the visualization and creation of word alignments have been developed (e.g., Melamed, 1998; Smith and Jahr, 2000; Ahrenberg *et al.*, 2002; Rassier and Pedersen, 2003; Daumé; Tiedemann; Hwa and Madnani, 2004; Lambert, 2004; Tiedemann, 2006). Most of them employ one of two visualization techniques. The first is to draw lines between associated words, as shown in Fig. 1. The second is to use an alignment matrix (Fig. 2), where the rows of the matrix correspond to the words of the sentence in one language and the columns to the words of that sentence's translation into the other language. Marks in the matrix's cells indicate whether the words represented by the row and column of the cell are linked or not. A third technique, employed in addition to drawing lines by Melamed (1998) and as the sole mechanism by Tiedemann (2006), is to use colors to indicate which words correspond to each other on the two sides of the parallel corpus.

The three techniques just mentioned work reasonably well for very short sentences, but reach their limits quickly as sentence length increases. Alignment visualization by coloring schemes requires as many different colors as there are words in the (shorter) sentence. Alignment visualization by drawing lines and alignment matrices both require that each of the two sentences in each sentence pair is

[1] Yawat was first presented at the 2007 *Linguistic Annotation Workshop* (Germann, 2007).

Proceedings of the ACL-08: HLT Demo Session (Companion Volume), pages 20–23,
Columbus, June 2008.

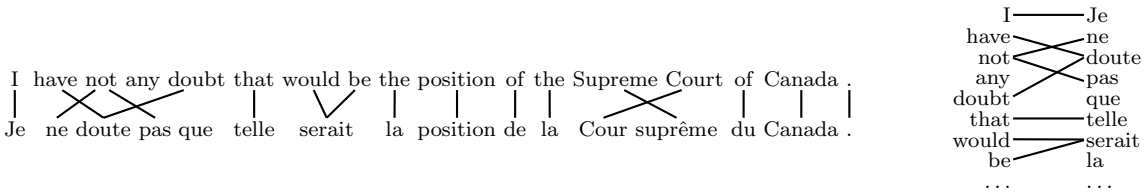

Figure 1: Visualization of word alignments by drawing lines.

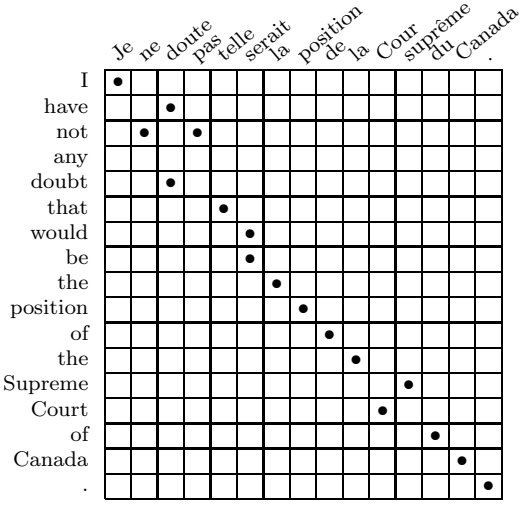

Figure 2: Visualization of word alignments with an alignment matrix.

presented in a single line or column. Pairs of long sentences therefore often cannot be shown entirely on the screen. Aligning pairs of long sentences then requires scrolling back and forth, especially when there are considerable differences in word order between the two languages. Moreover, as sentence length increases, visualization by drawing lines quickly becomes cluttered, and alignment matrices become hard to track. We believe that it is not only because of the intrinsic difficulties of explaining translations by word alignment but also because of such interface issues that aligning words manually has the reputation of being a very tedious task.

3 Yawat

Yawat (Yet Another Word Alignment Tool) was developed to remedy this situation by providing an efficient interface for creating and editing word alignments manually. It is implemented as web application with a thin CGI script on the server side and a browser-based[2] client written in JavaScript. This setup facilitates collaborative efforts with multiple annotators working remotely without the overhead of needing to organize the transfer of alignment data separately. The server-side data structure was deliberately kept small and simple, so that the tool or some of its components can be used as a visualization front-end for existing word alignments.

Yawat's most prominent distinguishing feature is

[2]Unfortunately, differences in the underlying DOM implementations make it laborious to implement truly browser-independent web applications in JavaScript. *Yawat* was developed for FireFox and currently won't work in Internet Explorer.

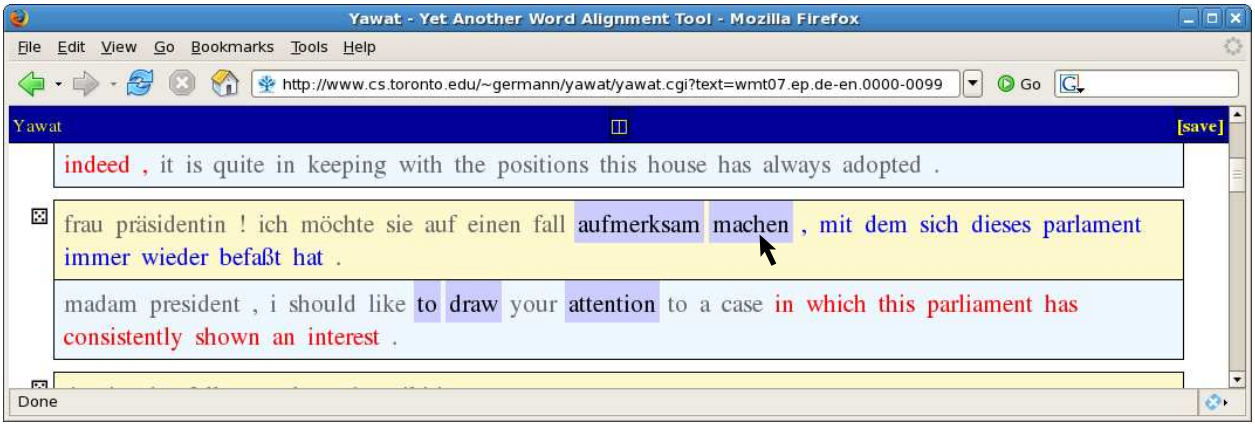

Figure 3: Alignment visualization with *Yawat*. As the mouse is moved over a word, the word and all words linked with it are highlighted. The highlighting is removed when the mouse leaves the word in question. This allows the annotator to focus on one item at a time, without any distracting visual clutter from other word alignments.

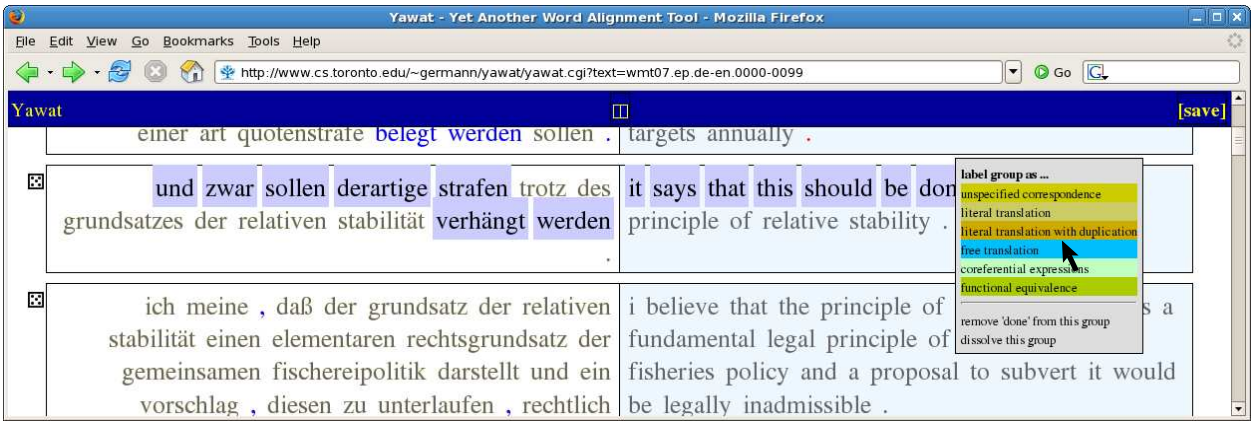

Figure 4: *Yawat* allows alignment relations to be labeled via context menues. Parallel text can be displayed side-by-side as in this screenshot or stacked as in Fig. 3.

the use of dynamic instead of static visualization. Rather than showing alignment links permanently by drawing lines or showing marks in an alignment matrix, associated words are shown only for one word at a time, as determined by the location of the mouse pointer. When the mouse is moved over a word in the text, the word and all the words associated with it are highlighted; when the mouse is moved away, the highlighting is removed. Figure 3 gives a snapshot of the tool in action.

Designed primarily as a tool for creating word alignments, one design objective was to minimize mouse travel required to align words. The interface therefore has no 'link words' button but uses mouse clicks on words directly to establish alignment links. A left-click on a word puts the tool into *edit* mode and opens an 'alignment group' (i.e., a set of words that supposedly constitute the expression of a concept in the two languages). Additional left-clicks on other words add them to or remove them from the current alignment group. A final right-click closes the group and puts the tool back into *view* mode. The typical case of aligning just two individual words thus takes only a single click on each of the two words: a left-click on the first word and a right-click on the second. As words are aligned, their color changes to indicate that they have been dealt with, so that the annotator can easily keep track of which words have been aligned, and which ones still need to be aligned. Notice the difference in color (or shading in a gray-scale printout) in the sentences in Fig. 3, whose first halves have been aligned while their latter halves are still unaligned.

In *view* mode, alignment groups can be labeled with a customizable set of tags via a context menu

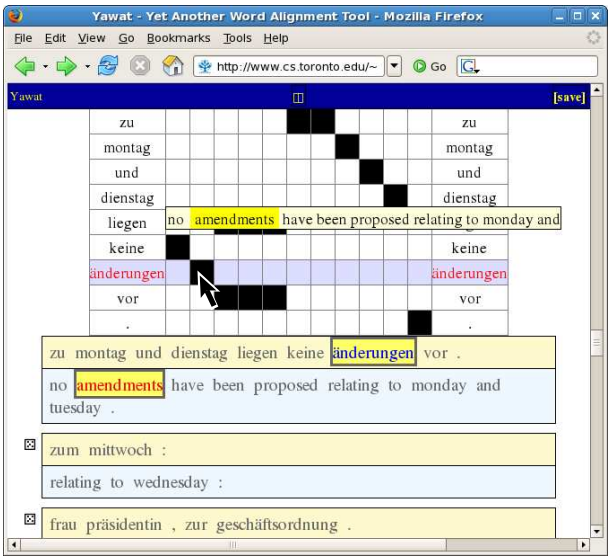

Figure 5: *Yawat* can also show alignments as alignment matrices. The tooltip-like floating bar above the mouse pointer provides column labels.

triggered by a right-click on a word (Fig. 4). For example, one might want to classify translational correspondences as 'literal', 'non-literal / free', or 'coreferential without intensional equivalence'. Different colors are used to indicate different types of alignment; color schemes and tag sets can be configured on the server side.

3.1 Alignment matrix display

One of the drawbacks of the dynamic visualization scheme employed in *Yawat* is that it provides no bird's-eye view of the overall alignment structure, as

it is provided by alignment matrices. We therefore decided to add alignment matrices as an additional visualization option. Alignment matrices are created on demand and can be switched on and off for each sentence pair. Word alignments can be edited in the alignment matrix view by clicking into the respective matrix cells to link or unlink words. Alignments matrices and the normal side-by-side or top-and-bottom display of the sentence pair in question are interlinked, so that an changes in the alignment matrix are immediately visible in the 'normal' display and vice versa (see Fig. 5).

4 Conclusion

We presented *Yawat*, a tool for the creation and visualization of word- and phrase alignments. An on-line demo is currently available at `http://www.cs.toronto.edu/~germann/yawat/yawat.cgi`. A package including the server-side scripts and the client-side code is available upon request.

References

Ahrenberg, Lars, Mikael Andersson, and Magnus Merkel. 2002. "A system for incremental and interactive word linking." *Third International Conference on Linguistic Resources and Evaluation (LREC-2002)*, 485–490. Las Palmas, Spain.

Daumé, Hal. "HandAlign." `http://www.cs.utah.edu/~hal/HandAlign/`.

Fraser, Alexander and Daniel Marcu. 2006. "Semi-supervised training for statistical word alignment." *Joint 44th Annual Meeting of the Association for Computational Linguistics and 21th International Conference on Computational Lignuistics (COLING-ACL '98)*, 769–776. Sydney, Australia.

Germann, Ulrich. 2007. "Two tools for creating and visualizing sub-sentential alignments of parallel text." *Linguistic Annotation Workshop (LAW '07)*, 121–124. Prague, Czech Republic.

Hwa, Rebecca and Nitin Madnani. 2004. "The umiacs word alignment interface." `http://www.umiacs.umd.edu/~nmadnani/alignment/forclip.htm`.

Lambert, Patrik. 2004. "Alignment set toolkit." `http://gps-tsc.upc.es/veu/personal/lambert/software/AlignmentSet.html`.

Melamed, I. Dan. 1998. *Manual Annotation of Translational Equivalence: The Blinker Project.* Technical Report 98-07, Institute for Research in Cognitive Science (IRCS), Philadelphia, PA.

Moore, Robert C., Wen-tau Yih, and Andreas Bode. 2006. "Improved discriminative bilingual word alignment." *Joint 44th Annual Meeting of the Association for Computational Linguistics and 21th International Conference on Computational Lignuistics (COLING-ACL '98)*, 513–520. Sydney, Australia.

Och, Franz Josef and Hermann Ney. 2003. "A systematic comparison of various statistical alignment models." *Computational Linguistics*, 29(1):19–51.

Rassier, Brian and Ted Pedersen. 2003. "Alpaco: Aligner for parallel corpora." `http://www.d.umn.edu/~tpederse/parallel.html`.

Smith, Noah A. and Michael E. Jahr. 2000. "Cairo: An alignment visualization tool." *Second International Conference on Linguistic Resources and Evaluation (LREC-2000).*

Taskar, Ben, Simon Lacoste-Julien, and Dan Klein. 2005. "A discriminative matching approach to word alignment." *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP '05)*, 73–80. Morristown, NJ, USA.

Tiedemann, Jörg. "UPlug: Tools for linguistic corpus processing, word alignment and term extraction from parallel corpora." `http://stp.ling.uu.se/cgi-bin/joerg/Uplug`.

Tiedemann, Jörg. 2006. "ISA & ICA — Two web interfaces for interactive alignment of bitexts." *Fifth International Conference on Linguistic Resources and Evaluation (LREC-2006).* Genoa, Italy.