Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)

Sketches of a graphical user interface for word alignment annotation

Maria Skeppstedt, Magnus Ahltorp, Gunnar Eriksson, Rickard Domeij
The Language Council of Sweden, The Institute for Language and Folklore, Sweden
firstname.lastname@sprakochfolkminnen.se

Abstract

We describe ideas for the graphical user interface of a tool for word alignment annotation. Our prototype interface builds on a design template from the Jigsaw system for investigative analysis, and is implemented as a web application.

1 Introduction

Although many models for machine translation no longer rely on training data in the form of word-aligned corpora, the concept of word alignment – and thereby the need for manually aligned corpora – still exists. For instance, manually annotated corpora are required for evaluating the alignment quality of translation systems that use word alignment as an intermediate step (Alkhouli et al., 2016) or systems that use automatically aligned corpora for dictionary construction (Bourgonje et al., 2018).

There exist several tools for word alignment annotation. Many of these tools are, however, either (i) several years old, and to the best of our knowledge no longer updated (Tiedemann, 2002; Zhang et al., 2008) or, (ii) targeting a more complex annotation task, in which the word alignment task is only a subtask (Hung-Ngo and Winiwarter, 2012; Wirén et al., 2018).

We therefore aim to construct an annotation tool that uses current libraries[1] for web development, and which is solely focused on the task of word alignment in sentence-aligned texts.

2 Design ideas for the user interface

Our prototype interface for word alignment annotation builds on a design template from the Jigsaw system for investigative analysis. This template has previously been used for visualising associations between entities extracted from text collections (Stasko, 2008; Skeppstedt et al., 2018). The template dictates that the entities, which here correspond to words that are to be aligned, are displayed in separate lists, and that associations between elements in the different lists are indicated by lines that connect them and by highlighting.

By arranging the words vertically, the display of the word associations becomes more compact for most writing systems, which we hypothesise will make it easier to trace the connecting lines. While this potentially de-emphasises the sentence, it instead emphasises the individual tokens.

Figure 1 shows the prototype applied to one sentence-pair in a corpus of parallell texts collected from translations made at Swedish government agencies (Dahlberg and Domeij, 2017).

The prototype interface contains the following components: **(a)** The sentence in the first language. **(b)** The sentence in the second language. **(c)** An alignment between two words is created by drag-and-drop, i.e. by dragging an element in the left-hand list and dropping it on an element in the right-hand list. **(d)** Alignment is shown by a line that connects the two list elements. **(e)** In addition, when the user hovers the mouse over an element in one of the lists, the associated elements in the other list are highlighted. **(f)** An alignment is removed by clicking on the corresponding delete button. **(g)** Drop-down list for choosing which corpus to annotate. **(h)** Drop-down list for choosing the criterium[2] by which the next sentence-pair to annotate is to be selected. **(i)** The user can choose either annotation mode or to browse previously annotated sentences in read-only mode. **(j)** Swap the two languages. **(k)** Save annotation. **(l)** Go back to the previously annotated sentence-pair. **(m)** Remove sentence-pairs from the annotation task (e.g. when the sentence-pair stems from an incorrect sentence alignment).

[1] D3 and Flask.

[2] We plan to add the functionality of pre-annotated alignments. The user should then be able to select the order in which the sentence-pairs are to be annotated, e.g. to choose to start with the ones that the pre-alignment system estimates to be easiest or estimates to be most difficult.

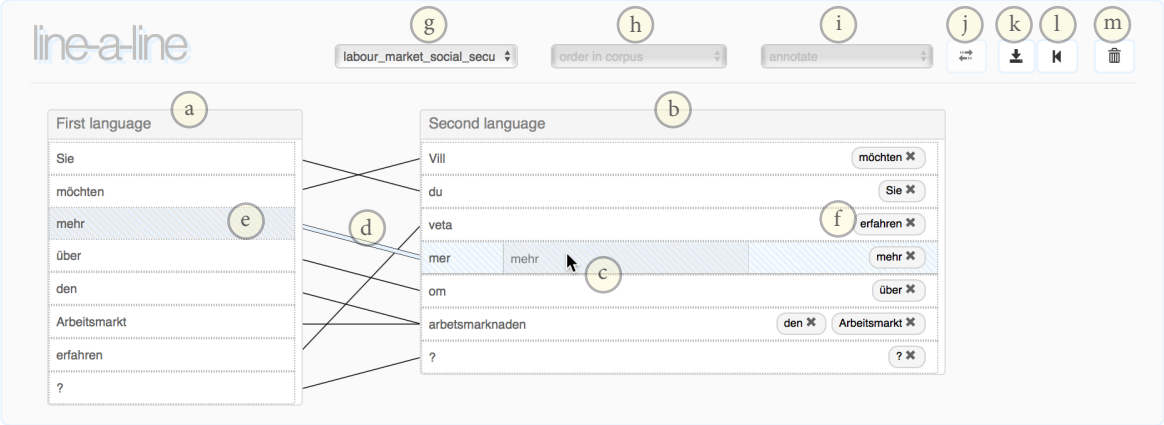

Figure 1: User interface prototype, showing a Swedish-German sentence-pair annotated for word alignment.

3 Next steps

The immediate next steps will consist of refining and evaluating the user interface. We also plan to evaluate the usefulness of the alignment functionality for tasks beyond core natural language processing, e.g. usefulness for translation studies and for research on the application of official terminologies in translated texts (Dahlberg, 2017).

The annotation tool should also be able to provide an optional automatic pre-alignment, i.e. an alignment on which the user can base their manual annotation. To enable pre-alignment for language pairs for which no large parallell corpora exist, the alignment functionality should preferably be based on methods suitable also for small corpora.

Acknowledgements

This work was funded by the Swedish Research Council (project number 2017-00626).

References

[Alkhouli et al.2016] Tamer Alkhouli, Gabriel Bretschner, Jan-Thorsten Peter, Mohammed Hethnawi, Andreas Guta, and Hermann Ney. 2016. Alignment-based neural machine translation. In *ACL 2016 First Conference on Machine Translation*, pages 54–65.

[Bourgonje et al.2018] Peter Bourgonje, Jet Hoek, Jacqueline Evers-Vermeul, Gisela Redeker, Ted Sanders, and Manfred Stede. 2018. Constructing a lexicon of dutch discourse connectives. *Computational Linguistics in the Netherlands Journal*, 8:163–175, 12/2018.

[Dahlberg and Domeij2017] Simon Dahlberg and Rickard Domeij. 2017. Översättning av termer i myndighetstexter: En studie om översättning av myndighetstermer i arbetet med nationell språkinfrastruktur på språkrådet. In *Workshop Termplanering och termbruk i svenskan på Svenskans beskrivning 36*.

[Dahlberg2017] Simon Dahlberg. 2017. Tre svenska myndigheters strategier för termöversättning till spanska och franska. Master's thesis, Stockholm University, Department of Linguistics.

[Hung-Ngo and Winiwarter2012] Quoc Hung-Ngo and Werner Winiwarter. 2012. A visualizing annotation tool for semi-automatically building a bilingual corpus. In *The Fifth Workshop on Building and Using Comparable Corpora*, pages 67–74.

[Skeppstedt et al.2018] Maria Skeppstedt, Kostiantyn Kucher, Manfred Stede, and Andreas Kerren. 2018. Topics2Themes: Computer-Assisted Argument Extraction by Visual Analysis of Important Topics. In *Proceedings of the LREC Workshop on Visualization as Added Value in the Development, Use and Evaluation of Language Resources*, pages 9–16.

[Stasko2008] John Stasko. 2008. Jigsaw: Investigative analysis on text document collections through visualization. In *DESI II: Second International Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings*.

[Tiedemann2002] Jörg Tiedemann. 2002. Uplug a modular corpus tool for parallel corpora. *Language and Computers*, pages 181–197, 07.

[Wirén et al.2018] Mats Wirén, Arild Matsson, Dan Rosén, and Elena Volodina. 2018. Svala: Annotation of second-language learner text based on mostly automatic alignment of parallel corpora. In *Selected papers from the CLARIN Annual Conference 2018*, number 159, pages 227–239. Linköping University Electronic Press, Linköpings universitet.

[Zhang et al.2008] Yujie Zhang, Zhulong Wang, Kiyotaka Uchimoto, Qing Ma, and Hitoshi Isahara. 2008. Word alignment annotation in a Japanese-Chinese parallel corpus. In *LREC 2008*.