Proceedings of the 13th Workshop on Building and Using Comparable Corpora, pages 1–5
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Line-a-line: A Tool for Annotating Word-Alignments

Maria Skeppstedt, Magnus Ahltorp, Gunnar Eriksson, Rickard Domeij
The Language Council of Sweden, The Institute for Language and Folklore
Box 20057, 104 60 Stockholm, Sweden
firstname.lastname@isof.se

Abstract

We here describe *line-a-line*, a web-based tool for manual annotation of word-alignments in sentence-aligned parallel corpora. The graphical user interface, which builds on a design template from the Jigsaw system for investigative analysis, displays the words from each sentence pair that is to be annotated as elements in two vertical lists. An alignment between two words is annotated by drag-and-drop, i.e. by dragging an element from the left-hand list and dropping it on an element in the right-hand list. The tool indicates that two words are aligned by lines that connect them and by highlighting associated words when the mouse is hovered over them. Line-a-line uses the efmaral library for producing pre-annotated alignments, on which the user can base the manual annotation. The tool is mainly planned to be used on moderately under-resourced languages, for which resources in the form of parallel corpora are scarce. The automatic word-alignment functionality therefore also incorporates information derived from non-parallel resources, in the form of pre-trained multilingual word embeddings from the MUSE library.

Keywords: Word-alignments, parallel corpora, annotation tools, multilingual word embeddings

1. Introduction

Word-aligned parallel corpora form useful resources for several tasks, e.g. for bilingual dictionary construction (Bourgonje et al., 2018), for studies of language typology (Dahl and Wälchli, 2016), for translation studies (Merkel et al., 2003), as well as for those types of machine translation systems that use word-aligned corpora as an intermediate step (Alkhouli et al., 2016).

For constructing alignment gold standards, e.g. for evaluating the performance of automatic word-aligners, there is a need for tools by which manual annotations of word-alignments can be performed. There exist many such word-alignment annotation tools, but these tools are typically either (i) several years old (Merkel et al., 2003; Zhang et al., 2008; Hung-Ngo and Winiwarter, 2012), or (ii) not targeting the core task of word-alignment of sentences belonging to two different languages (Wirén et al., 2018).

Annotation tools whose interfaces are not being modernised according to the possibilities offered by more recent graphical user interface libraries might, however, be perceived as not adhering to current graphical user interface conventions. This might in turn decrease the usability of, and trust in, these older tools, also when they offer a functionality that objectively should be adequate for performing the manual alignment annotations.

As an alternative to these older tools, we have used current libraries for web development for constructing an annotation tool to use for the task of word-alignment in sentence-aligned texts. With the aim of increasing the usability of the tool, we have used a design template from the field of visualisation research as an inspiration for the user interface design. To further facilitate the annotation, a selectable pre-annotation in the form of an automatic word-alignment is provided, on which the user can base their manual annotation.

We plan to use word-aligned corpora for performing translation studies, including research on the application of official terminologies in translated texts (Dahlberg, 2017). We will particularly focus on moderately under-resourced languages and under-resourced language pairs, for which small monolingual corpora exist and only very small parallel corpora. The line-a-line tool therefore allows the user to choose between several methods for the pre-annotations of the word-alignments, i.e. the user can select the alignment method that is found most useful for the language pair targeted.

2. Previous tools

The following four tools form examples of tools developed for word-alignment between sentence-aligned parallel texts, or for related tasks.

The I*Link tool (Merkel et al., 2003) for word-alignment annotation proposes alignment candidates, using bilingual resources and built-in heuristics, and the user can then accept, revise or reject these proposals. The tool also saves the user's alignment choices and adapts new alignment suggestions to previous choices made. The sentence pair is displayed in two horisontal rows, and the colour in which the words are written is used for indicating which words that are aligned, i.e. aligned words are displayed with the same, unique colour.

Zhang et al. (2008) developed a word-alignment annotation tool targeted towards Japanese-Chinese parallel corpora. The sentence pairs are provided with pre-alignments through the GIZA++ word-alignment tool. The sentence pair is displayed in two horizontal rows, and alignments are indicated through connecting lines. The user can optionally create chunks of tokens in the individual languages, and align chunks instead of words.

The tool by Hung-Ngo and Winiwarter (2012) also displays the sentence pair in two horizontal rows, and uses connecting lines to show alignments. The sentences are pre-annotated through the use of bilingual dictionaries, and parse trees for the two sentences are also generated and displayed. Annotation is carried out through drag-and-drop of nodes that symbolise the words or other levels in the parse

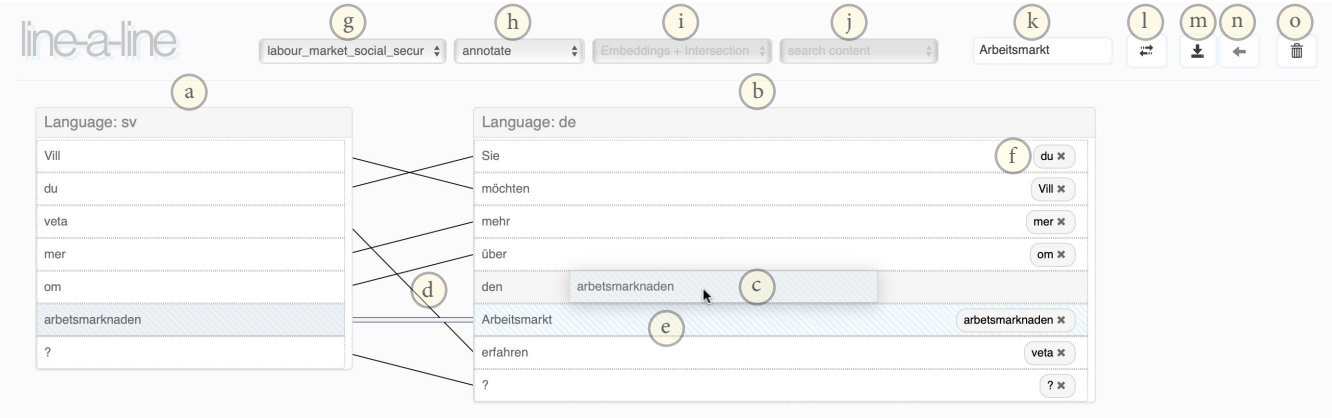

Figure 1: The user interface, showing a Swedish-German sentence pair. The following selections have been made by the user: (i) Annotation mode **(h)**, (ii) to let the tool provide pre-annotated alignments through a union of intersection-symmetrising and the alignments produced during the dictionary creation **(i)**, (iii) to select what to annotate through searching for word in the corpus **(j - k)**. An alignment between two words is annotated through dragging the element from the left-hand list and dropping it on an element on the right-hand list. The figure shows how the user is dragging the element representing *'arbetsmarknaden'* with the aim of dropping it on the *'den'* element. When the user has dropped the element, an alignment between these two words will have been created.

tree. The tool also provides a visualisation of alignments in a matrix format.

The SVALA tool is constructed using current libraries for web development. Its purpose is, however, not to align sentences written in two different languages, but to correct and annotate text written by second-language learners. While the user is correcting the text, the tool maintains an automatic word-alignment between the original text and the corrected version. When necessary, it is also possible to manually correct the automatic word-alignments provided. Also this tool displays the sentence pairs in two horizontal rows, with connecting lines that indicate alignments.

3. The implemented word-alignment annotation tool

The line-a-line tool consists of a web-based front-end written in JavaScript/D3, and a back-end based on Python/Flask, a PyMongo database, as well as on the efmaral library (Östling and Tiedemann, 2016) for producing the automatic word-alignments used for the pre-annotations. Apart from being provided with the information available in the parallel texts, the automatic word-alignment functionality is also provided with data derived from a multilingual embedding space.

During the development of the tool, we used 559 sentence pairs from automatically sentence-aligned Swedish-German parallel texts, which have been collected from translations of Swedish government agency texts (Dahlberg and Domeij, 2017). We also used the Swedish-German multilingual embedding space available from the MUSE library.

3.1. Front-end

The interface for carrying out a manual word alignment is shown in Figure 1. The figure shows the tool applied to the corpus used during the tool development, with the Swedish text to the left and the German text to the right.

The interface contains the following components: **(a)** The sentence belonging to one of the languages. **(b)** The sentence belonging to the other language. **(c)** An alignment between two words is created by drag-and-drop, i.e. by dragging an element from the left-hand list and dropping it on an element in the right-hand list. **(d)** Alignments are shown by lines that connect the associated list elements. **(e)** In addition, when the user hovers the mouse over an element, its associated elements, and the lines indicating the association, 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 for choosing either annotation mode or to browse previously annotated sentences in read-only mode. **(i)** Drop-down list for choosing which word-alignment method to use for the pre-annotation. There are three different word-alignments to choose from (see section 3.3. below), and the user can also choose to annotate the alignments from scratch without any pre-annotations. **(j)** Drop-down list for choosing the criterium by which the next sentence pair to annotate is to be selected. The user can choose the order in which the sentence pairs are to be annotated, i.e. to choose to start with the ones that the pre-alignment system estimates to be easiest or estimates to be most difficult, or to annotate the sentence pairs in the order in which they appear in the corpus. The user can also choose to annotate sentence pairs that contain a specific word, and the word to search for is specified in the text field **(k)**. **(l)** Reverse the order in which sentences belonging to the two languages are displayed. That is, the German text would in this case be displayed to the left and the Swedish text to the right, if the order were reversed. **(m)** Save the alignment annotation. **(n)** Redo, i.e. go back to the previously annotated sentence pair. **(o)** Remove the sentence pair from the annotation task

(e.g. when the sentence pair stems from an incorrect sentence alignment).

To be able to choose a sentence-aligned corpus for manual annotation – in the drop-down list (**g**) above – the Python script provided for loading it into line-a-line's database must have been executed. The sentence-aligned corpus must be provided in the Translation Memory eXchange (TMX) format. The loading script tokenises the sentences using NLTK's TweetTokenizer[1] (Bird, 2002), and saves the tokenised sentence pairs in the PyMongo database.

The user interface builds on a design template from a system constructed within the field of information visualisation research; the Jigsaw system for investigative analysis (Stasko, 2008). The Jigsaw system includes a list view user interface for visualising connections between different types of entities (e.g. people, places, dates and organisations) that are mentioned in a text collection. The interface displays each type of entity in separate lists, and associations between entities in the different lists are indicated by highlighting the entities and the lines that connect them. The same design template has also been used for visualising associations between information entities extracted from large text collections by the use of topic modelling (Skeppstedt et al., 2018).

Lists of words that form sentences in two different languages, and where some of the word-pairs in these two lists are connected, form a data set that is similar to the connected entity data of Jigsaw's list view interface. We therefore found the list view template suitable for the word-alignment task, where alignments are indicated by connecting lines and by highlighting of associated words and of lines that connect these words (shown in Figure 1).

To display the two paired sentences in the form of two vertical lists differs from the approach used in the systems mentioned above, which either display word-alignments through lines between two horizontal sentences, or in a matrix format. By instead arranging the words vertically, as we have done here, the display of the word associations becomes more compact for most writing systems, which has the potential to make it easier to trace the connecting lines. While this vertical view potentially de-emphasises the sentence, it instead emphasises the individual tokens, which might make it easier to focus on the parts of the sentences that are relevant for the immediate alignment connections that are created or inspected by the annotator.

3.2. An automatically created dictionary from multilingual word embeddings

As stated above, we plan to apply the tool on pairs of texts in moderately under-resourced languages, for which parallel resources are scarce. To improve the pre-annotation for these languages, information from monolingual resources should also be included in the automatic word-alignment functionality. To achieve this, the tool uses pre-trained multilingual embeddings from the MUSE library.[2]

By using the MUSE library, multilingual word embeddings can be constructed from independent monolingual resources. A multilingual word embedding is constructed from two separate monolingual word embedding spaces for the two languages in question. That is, each one of the embedding spaces is trained independently on a monolingual corpus. The embeddings for the two monolingual spaces constructed are then automatically aligned, i.e. pairs of corresponding embedding vectors are found in the two spaces. If there is a bilingual dictionary available, the alignment can be carried out in a supervised fashion. A subset of the embeddings can then be aligned with the use of the dictionary, and the alignments of other embeddings can thereafter be adapted to these points. The process can also be carried out in an unsupervised fashion without a dictionary, using a similarity measure called 'cross-domain similarity local scaling' for finding alignments between embeddings (Conneau et al., 2017).

The resulting multilingual word embedding space can then be queried for a word in one of the languages, which results in an output in the form of a list of the nearest neighbours to this word in the other language. We use this functionality to automatically generate a corpus-specific bilingual dictionary, which we give as an extra parallel data input to the word-alignment functionality described below. The method used for incorporating the embeddings is somewhat inspired by the work by Pourdamghani et al. (2018). They, however, use similarity in two monolingual spaces for inferring word-alignments.

The automatic creation of the corpus-specific bilingual dictionary is carried out as follows: For each sentence pair in the parallel corpus, i.e. the pair of two vectors of words, one belonging to language $\mathbf{a}$ and the other to language $\mathbf{b}$, the following is carried out. All possible tuples consisting of one word from the sentence belonging to language $\mathbf{a}$ and one word from the sentence belonging to language $\mathbf{b}$ are constructed. For each such tuple (a_i, b_j), the multilingual word embedding space is used to check whether a_i is included among the top n nearest neighbours to b_j and whether b_i is included among the top n nearest neighbours to a_j. If both conditions are fulfilled, the tuple is added to the automatically constructed bilingual dictionary. The detected tuple is also recorded as a word-alignment for this specific sentence-pair, and this alignment is later used as a component in one of the pre-alignment options provided by the tool.[3]

If no match is found for any of the words in a sentence, we also allow for a search on subwords in the embedding space. Thereby, some morphological variations and compound words that are present in the sentences that are to be aligned, but not included in the embedding space, can be included.[4]

[1] https://www.nltk.org/api/nltk.tokenize.html. The use of tokeniser will later be made configurable, as there are many languages for which the TweetTokenizer it is not suitable. For instance, Japanese and Chinese, which do not use white space to indicate token segmentation.

[2] https://github.com/facebookresearch/MUSE

[3] When developing the system we used $n = 2$, and if no match was found for a word, we allowed for an $n = 20$ for a word pair to be included in the dictionary. The cut-off used should, however, be allowed to be adjusted by the user. Punctuation characters and stop words are excluded from the dictionary construction process.

[4] We here used a minimum allowed length of 4 characters for a

The word-alignments and the automatically created dictionary are then used as components for producing pre-annotated word-alignments.

3.3. Back-end with pre-annotated word-alignments

The corpus loading script also runs an automatic word-alignment, which is used for the pre-annotated alignments on which the user bases their manual annotations. The main method for producing the automatic word-alignments is the efmaral system (Östling and Tiedemann, 2016; Tiedemann et al., 2016)[5]. The efmaral system uses a Bayesian model with Markov Chain Monte Carlo (MCMC) inference for producing the word-alignments. The corpus-specific dictionary, automatically produced through the MUSE library, is used as additional input, i.e. as aligned data, for the efmaral word-alignment.

The efmaral alignment is run twice, first with one of the languages as source language and the other as target language, and thereafter with reversed language order. Three different methods are then available for symmetrising the alignments, (i) a simple intersection of the two alignment predictions, i.e. only keeping the alignments that are predicted by both models, or (ii) symmetrising using the GDFA word-alignment symmetrisation algorithm as implemented in NLTK[6] (Axelrod et al., 2005), (iii) a union of the intersection-symmetrising alignments and the alignments produced during the dictionary creation. The user can thereby choose the type of pre-annotation that is found most useful for the corpus that is being annotated. The user can also choose to carry out the annotation without using a pre-annotation of the alignments, as it is likely that there are circumstances when the pre-annotations would not be found useful. For instance, when the tool is applied on languages with very few existing resources, which would render low quality pre-annotations. None of the methods provided for pre-alignment rely on the existence of heavy resource-demanding language models, e.g., BERT models, as such models would be unobtainable for the low resource language pairs that form the target of the tool.

A difficulty score is computed for the alignments, by measuring the number of word pairs that are included in the intersection set in relation to the total number of words in the two sentences. This difficulty score is used for sorting the sentence pairs that are given to the user for manual alignment. Depending on the choice made by the user in the drop-down list **(i)**, the back-end either delivers the most difficult un-annotated sentence pairs or the easiest ones. The user can also choose not to use this difficulty score for selecting sentences to annotate, but to use the original corpus order of the sentences. There is also a forth option in the drop-down lists, which lets the user search for a specific word in the corpus, and annotate all sentence pairs in which this word is included.

sequence of characters to be considered a subword, but this figure should also be allowed to be adjusted to the language pairs used.

[5]https://github.com/robertostling/efmaral

[6]https://www.nltk.org/_modules/nltk/translate/gdfa.html

4. Concluding remarks

With line-a-line, we have provided a tool that we hope will form a useful resource for annotating word-alignments in sentence-aligned parallel corpora.[7]

Whether the pre-annotations available will have a quality that is high enough to be found helpful when annotating, will depend on the resources available, i.e. on which language pairs that are to be aligned, on the size of the parallel corpus available, and on the quality of the multilingual word embedding space. A key functionality of the line-a-line tool is therefore to provide several methods for pre-annotation, and let the user choose the one that is found most helpful for performing the annotation.

For instance, we perceived the pre-annotations constructed by a union of intersection-symmetrising and dictionary creation alignments to be most useful during the tool development. In contrast, pre-annotations constructed through the GDFA symmetrisation were perceived as not useful, as they contained too many false positives for our small Swedish-German parallel corpus.

Acknowledgements

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

5. Bibliographical References

Alkhouli, T., Bretschner, G., Peter, J.-T., Hethnawi, M., Guta, A., and Ney, H. (2016). Alignment-based neural machine translation. In *ACL 2016 First Conference on Machine Translation*, pages 54–65, Berlin, Germany, August.

Axelrod, A., Mayne, R. B., Callison-burch, C., Osborne, M., and Talbot, D. (2005). Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *In Proc. International Workshop on Spoken Language Translation (IWSLT)*.

Bird, S. (2002). NLTK: The natural language toolkit. In *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bourgonje, P., Hoek, J., Evers-Vermeul, J., Redeker, G., Sanders, T., and Stede, M. (2018). Constructing a lexicon of Dutch discourse connectives. *Computational Linguistics in the Netherlands Journal*, 8:163–175, 12/2018.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Dahl, Ö. and Wälchli, B. (2016). Perfects and iamitives : two gram types in one grammatical space. *Letras de Hoje*, 51(3):325–348.

Dahlberg, S. and Domeij, R. (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*.

[7]The tool will be made freely available at: https://github.com/mariask2/line-a-line.

Dahlberg, S. (2017). Tre svenska myndigheters strategier för termöversättning till spanska och franska. Bachelor's thesis, Stockholm University, Department of Linguistics.

Hung-Ngo, Q. and Winiwarter, W. (2012). A visualizing annotation tool for semi-automatically building a bilingual corpus. In *The Fifth Workshop on Building and Using Comparable Corpora (5th BUCC within the LREC2012)*, pages 67–74.

Merkel, M., Petterstedt, M., and Ahrenberg, L. (2003). Interactive word alignment for corpus linguistics. In *Proceedings of Corpus Linguistics 2003, 28-31st March, 2003, Lancaster UK. UCREL Technical Papers.*, pages 533–542. UCREL (University Centre for Computer Corpus Research on Language). ISBN 1 86220 131 5.

Östling, R. and Tiedemann, J. (2016). Efficient word alignment with Markov Chain Monte Carlo. *The Prague Bulletin of Mathematical Linguistics*, 106:125–146, 10.

Pourdamghani, N., Ghazvininejad, M., and Knight, K. (2018). Using word vectors to improve word alignments for low resource machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 524–528, New Orleans, Louisiana, June. Association for Computational Linguistics.

Skeppstedt, M., Kucher, K., Stede, M., and Kerren, A. (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.

Stasko, J. (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*.

Tiedemann, J., Cap, F., Kanerva, J., Ginter, F., Stymne, S., Östling, R., and Weller-Di Marco, M. (2016). Phrase-based SMT for Finnish with more data, better models and alternative alignment and translation tools. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 391–398, Berlin, Germany, August. Association for Computational Linguistics.

Wirén, M., Matsson, A., Rosén, D., and Volodina, E. (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, Pisa, 8-10 October 2018*, number 159, pages 227–239. Linköping University Electronic Press, Linköpings universitet.

Zhang, Y., Wang, Z., Uchimoto, K., Ma, Q., and Isahara, H. (2008). Word alignment annotation in a Japanese-Chinese parallel corpus. In *LREC 2008*.