



Visual Evaluation of Translation Alignment Data

T. Yousef¹  and S. Jänicke² 

¹Leipzig University, Institute for Computer Science, Germany

²University of Southern Denmark, Department of Mathematics and Computer Science, Denmark

Abstract

Translation alignment plays a crucial role in various applications in natural language processing and digital humanities. With the recent advance in neural machine translation and contextualized language models, numerous studies have emerged on this topic, and several models and tools have been proposed. The performance of the proposed models has been always tested on standard benchmark data sets of different language pairs according to quantitative metrics such as Alignment Error Rate (AER) and F1. However, a detailed explanation on what alignment features contribute to these scores is missing. In order to allow analyzing the performance of alignment models, we present a visual analytics framework that aids researchers and developers in visualizing the output of their alignment models. We propose different visualization approaches that support assessing their own model's performance against alignment gold standards or in comparison to the performance of other models.

CCS Concepts

• **Human-centered computing** → **Visual analytics; Visualization design and evaluation methods;** • **Computing methodologies** → **Machine translation;**

1. Introduction

Translation alignment aims to find word-level translation equivalents between parallel texts in multiple languages. It is an essential task in natural language processing and digital humanities. Next to its key role in statistical and neural machine translation [ABP*16], related applications include cross-lingual annotation projection [XQV*21], translation dictionaries generation [APPG14, YPSF22, SZW21], language learning [PFY21].

The development of automatic alignment systems began in 1993 when Brown et al. [BDPDPM93] proposed five statistical models with minimal linguistic content called the *IBM models* to extract translation alignments from a large bilingual corpus. Later, Och and Ney [ON00] implemented *Giza++* using the IBM models and the Hidden-Markov alignment model [VNT96]. Later, Dyer et al. [DCS13] introduced the *fast_align* alignment model, a fast and effective log-linear reparameterization of IBM Model 2. *eflomal* [ÖT16] proposed an efficient and accurate alignment model that outperformed the previous models using Bayesian model with Markov Chain Monte Carlo inference. Recently, a new era of translation alignment studies has started with the advances of neural machine translation systems and the rise of contextualized language models. The recent studies propose exploiting the multilingual contextualized word embeddings [JSDYS20, DN21] or the attention weights between the encoder and decoder of neural machine translation models [CLC*20] to capture semantically related words from two bilingual parallel texts. Some other models combine sta-

tistical and neural models to enhance the accuracy for low- and high-resource languages [SLW21].

Alignment gold standards are fundamental resources to assess the performance of automatic translation alignment systems [ON00, MP03, Mar08, Mac10, HA11]. Alignment gold standards are manually created alignments by two or more annotators using translation alignment tools [YPSF22] following predefined alignment guidelines [LDGBM05]. An Inter-Annotators Agreement must be computed to reflect the accuracy and reliability of the gold standards data sets. Most alignment guidelines follow the *Sure/Possible* annotation schema, which distinguishes between two alignment categories, namely, sure alignment pairs and possible alignment pairs. Several formats are used to encode the alignment gold standards, such as NAACL and TALP formats [MP03]. The alignment of a sentence is represented as a list of Zero-indexed indices pairs connected with "-". The letters *p* and *s* can follow the alignment pair to indicate a possible or sure alignment category.

The performance of alignment models has been measured using several quantitative metrics such as *Precision*, *Recall*, and *F1*. Besides, Och and Ney proposed the Alignment Error Rate (AER) [ON03], a metric to evaluate the performance of automatic alignment models against alignment gold standards. To our best knowledge, no work has inspected and analysed the alignment output to discover alignment failures or frequent alignment errors or to compare the performance of different models. Developers of automatic word alignment models rely totally on AER to confirm that one model produces better alignments than other models.

Our research background in text alignment prepared the ground for this paper. We developed an alignment model to align parallel texts in classical languages automatically [YPWB22]. In this context, we needed to support the following tasks:

- explore the alignment gold standard dataset (T1),
- analyze the output of alignment models in order to conduct a qualitative evaluation and identify alignment errors (T2),
- compare the output of different models at sentence-level (T3),
- compare the performance of different models on corpus- and sentence-level (T4), or compare the performance of same model with different parameters to notice the progress or decline achieved with specific parameters,
- analyze the correlation between sentence length and model performance (T5).

We developed a visual analytics framework and proposed several overview and detail visualizations that aid translation alignment researchers. Our usage scenarios document that the insights gained with our solution can be pivotal for improving automatic alignment models. An online demo is available at <http://www.vis4nlp.com/alignmenteval/>.

2. Related Work

Translation alignment visualization was subject to several research studies. Yousef and Jänicke [YJ20] surveyed the existing tools and the utilized visualization approaches. *YAWAT* [Ger08], *SWIFT Aligner* [GSK14], *CLUE-Aligner* [BRL16], and *UGARIT* [YPSF22] are annotation tools for manual translation alignment used to create alignment gold standards. However, the tools provide different means of visualization of the aligned parallel sentences.

WA-Continuum [SS15] is a tool to visualise word alignments across multiple parallel sentences, it uses the grid view for this purpose. [JSD*21] developed the *Parallel Corpus Explorer (Par-CourE)* which allows browsing word-aligned parallel texts. *Par-CourE* uses the parallel view with connecting lines among the aligned words among two or more parallel sentences. Similarly, *SimAlign* developers [JSDYS20] visualize the alignment output of two parallel sentences. Further, [ACL20] proposed *AlignVis*, an alignment and visualization tool of parallel translations using vertical parallel views.

Our tool is distinguished from the previously mentioned works in that it visualizes the output of automatic alignment models aiming to qualitatively assess their performance against other models or against alignment gold standards to overcome the limitations of quantitative evaluation metrics.

3. Visual Design

Translation alignment is the process of finding word-level equivalents between the source sentence $S = (s_1, s_2, \dots, s_n)$ and its translation $T = (t_1, t_2, \dots, t_m)$ [BDPDPM93]. The alignment process can be considered as a function, its inputs are S and T and its output is a set $A(S, T) = \{(s_i, t_j) : s_i \in S, t_j \in T\}$ where t_j is a word-level translation equivalent of s_i .

Developers of translation alignment models test the performance

of their models against benchmark data sets. The majority of the models are tested on the same collection of gold standard data sets [ON00, MP03, Mar08, Mac10, HA11] using *AER* and *F1* as basic evaluation metrics. To make sense of these measures and to direct the developer to alignment pairs that increase the *AER* of the model, we developed a framework that provides a comprehensive overview of the compared models with different visualization approaches. We offer corpus-level overview visualizations as well as sentence-level detail visualizations, and adapt Schneiderman's Information Seeking Mantra "*Overview first, zoom and filter, then details-on-demand*" [Shn03] to facilitate interactive navigation through the benchmark data set.

3.1. Corpus-level Visualization

The overview consists of a **bar chart** that shows the different models according to their quantitative metrics values, namely, *AER*, *Precision*, *Recall*, *F1*, and the number of translation pairs (T4). From the buttons above users can switch between the different metrics. Every model is assigned a unique color (Figure 1a). Users can select a model (as a reference model) by clicking on the corresponding bar in the bar chart. This will load a **scatter plot** (Figure 1c) with one dot per sentence of the underlying data set according to the selected metric (y-axis) and either sentence Id, source sentence length, or target sentence length (x-axis). The scatter plot supports model developers in detecting outliers and interesting observations such as the relation between the *Recall* and the sentence length (T5). Moreover, a range selector allows filtering multiple sentences with similar features resulting in a paginated list of sentence-level views for more detailed inspection.

Additionally, the framework provides useful information about the overlap among different models, which presented as a **heat map** (Figure 1b) (T4). A color gradient defines different shades of green to indicate this overlap. The more saturated, the more shared alignment pairs are detected across the two models juxtaposed.

3.2. Sentence-level Visualization

The sentence-level views aim to show the alignment among words of the source and target sentences. The framework provides two sentence-level views that allow visualizing the gold standards, single model alignment, and two models alignment with their agreement. The views are accompanied with a bar chart (Figure 1e) showing the sentence-level evaluation metrics of the available models and enabling users to select a model to visualize its output for the corresponding sentence.

3.2.1. Grid View

In this view, the two sentences are represented in the form of a grid (Figure 1d). The rows represent the source sentence's tokens $S = (s_1, s_2, \dots, s_n)$, and the columns represent the translation's tokens $T = (t_1, t_2, \dots, t_m)$.

The gold standard translation alignments are projected on the grid in the form of black-bordered cells with big dots indicating *Sure alignments* and tiny dots to show *Possible alignments* (T1). Colored cells represent alignment pairs of the model(s) under investigation. A cell (i, j) shows an alignment between $s_i \in S$ and



Figure 1: Overview and Detail Views. a) Bar Chart to compare the different quantitative metrics among the alignment models. b) Heat Map to show the relatedness among the models by visualizing their output overlap. c) Scatter Plot to show the relation between the evaluation metrics and the sentence length. d) Sentence-Level Grid View to visualize the alignments of one or two models with the gold standard. e) Parallel View for alignment visualization f) Bar Chart to show the sentence-level evaluation metrics.

$t_j \in T$. In case only one model is observed, cells appear in the model's assigned color, but when two models' outputs are compared, we use green colored cells to indicate an alignment agreement between the two models (T2, T3, T4). A tooltip containing the translation pairs will appear when hovering an alignment cell to investigate the aligned words.

3.2.2. Parallel View

Here, the two sentences are laid out horizontally, the upper row representing the source sentence S and the lower row representing the target sentence T . An alignment between $s_i \in S$ and $t_j \in T$ is represented as a line drawn between them (Figure 1f) (T2, T3, T4). This view uses the same coloring scheme like in the Grid View. For better readability, we visualize the gold standard on demand to reduce the number of crossed lines. Parallel view shares the same coloring schema with the grid view. The *Sure* alignment pairs are presented as dashed lines, whereas the *Possible* ones are presented as dotted lines (T1).

4. Usage Scenarios

S1: Exploring gold standard dataset

This scenario is the simplest one, where no reference model is selected. The grid view shows the gold standard as bordered cells

with big dots for sure alignments and small dots for possible alignments. Likewise, the parallel view visualizes the sure alignments as dashed lines and possible alignments as dotted lines.

S2: Comparing corpus-level models performance

The interface shows the different corpus-level quantitative metrics as a bar chart (Figure 1a), in addition to a heat map that shows the overlap between the models output. Figure 1b shows that the *SoftMax* translation pair set totally contains the *EntMax* set, which explains why *SoftMax* always outperforms *EntMax* regarding *Recall* but underperforms it regarding *Precision*.

S3: Analyzing the output of a single alignment model

Users can select a reference model from the bar chart and filter the data set by selecting a range of points on the scatter plot that meets their query (Figure 1c), resulting in a paginated list of sentence-views with different filtering and sorting options. A detailed grid and parallel views are generated for each sentence, and the gold standard alignments are projected, enabling the user to detect correct and wrong alignments visually (Figures 1d and 1e).

S4: Comparing the output multiple alignment models

We compare the outputs of AWESOME [DN21] and SIMALIGN [JDYS20] aligners, both tools use the embeddings based similarity matrix and apply several extraction methods to capture the alignments; AWESOME employs *Softmax*, *Entmax15* extraction strategy, whereas SIMALIGN uses *ArgMax*, *IterMax*, *MWMF* (Maximum-Weight Maximal Matching). We use mBert [DCLT18] to get the multilingual contextualized embeddings. The models overview shows that *ArgMax* outperforms *Softmax* regarding the *Precision* but *Softmax* outperforms *ArgMax* regarding the *AER*. The heat map shows a great overlap between the outputs of the two models but *Softmax* generates more translation pairs. Moreover, The sentence-level views (Figure 2) show that *Softmax* generates more correct translation pairs.



Figure 2: Scenario 4: Comparing multiple alignment models.

S5: Comparing the output of an alignment model with different parameters

We compare the output of AWESOME with the *Softmax* extraction method based on three different language models, mBert [DCLT18], XLM-R [CKG*19], and afine-tuned mBert model provided by the developers of the tool. Figure 3 shows that the fine-tuned model generates more translation pairs than the other models, but at the same time, it achieves better Recall and Precision. That means the newly generated translation pairs are correct in most cases. Thus, the fine-tuning led to a better language model, which resulted in better alignment results.

Alignment Errors

The following three groups of alignment failures get salient through our framework:

- **Wrong Alignments:** In this case, the translation pairs are incorrect translation equivalents. In Figure 4, the translation pairs *etwas-action* is a wrong alignment.
- **Wrong Positioning:** The translation pairs are semantically related, but they are not supposed to be aligned because they appear in different positions of the sentences. In Figure 4, the pair *Das-the* is a correct translation pair, and it is aligned by the model *Itermax*, but in this context, it is considered as a wrong alignment.



Figure 3: Scenario 5: Comparing the outputs of same model with different embeddings.

- **Missing Alignments:** Two words must be aligned, but the alignment models fail to align them. For instance, the pair *muß-need* in Figure 4.

5. Conclusion and Future Work

Evaluating translation alignment models is a nontrivial task. Despite that, quantitative metrics are widely used, but they are insufficient. Moreover, they do not reflect the actual quality of the model and do not distinguish among different alignment error types. For instance, aligning the German word *Blau* (Blue) to the English word *Red* will have the same penalty as aligning a comma to a full stop. Our framework supports detecting alignment errors and thus helps understanding why some models perform better or worse on specific sentences. For now, in all our experiments, we used one benchmark data set (German-English) and models from recently published papers. We intend to convert it to a generic web-based tool that can parse and visualize any gold standard data set and develop a pipeline to allow developers to upload and share the outputs of their models and get them visualized.

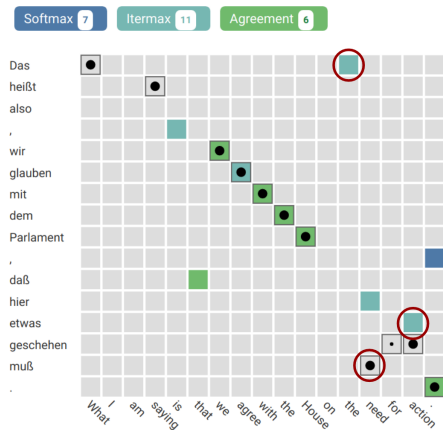


Figure 4: Three types of alignment errors

References

- [ABP*16] ALKHOULI T., BRETSCNER G., PETER J.-T., HETHNAWI M., GUTA A., NEY H.: Alignment-based neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers* (2016), pp. 54–65. 1
- [ACL20] ALHARBI M., CHEESMAN T., LARAMEE R. S.: Alignvis: Semi-automatic alignment and visualization of parallel translations. *2020 24th International Conference Information Visualisation (IV)* (2020), 98–108. 2
- [APPG14] AKER A., PARAMITA M. L., PINNIS M., GAIZAUSKAS R.: Bilingual dictionaries for all eu languages. In *LREC 2014 Proceedings* (2014), European Language Resources Association, pp. 2839–2845. 1
- [BDPDPM93] BROWN P. F., DELLA PIETRA S. A., DELLA PIETRA V. J., MERCER R. L.: The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 19, 2 (1993). 1, 2
- [BRL16] BARREIRO A., RAPOSO F., LUÍS T.: Clue-aligner: An alignment tool to annotate pairs of paraphrastic and translation units. In *10th Language Resources and Evaluation Conference (LREC)* (2016), pp. 7–13. 2
- [CKG*19] CONNEAU A., KHANDLWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTMLOYER L., STOYANOV V.: Unsupervised cross-lingual representation learning at scale. *CoRR abs/1911.02116* (2019). 4
- [CLC*20] CHEN Y., LIU Y., CHEN G., JIANG X., LIU Q.: Accurate word alignment induction from neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Online, Nov. 2020), Association for Computational Linguistics. 1
- [DCLT18] DEVLIN J., CHANG M., LEE K., TOUTANOVA K.: BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805* (2018). 4
- [DCS13] DYER C., CHAHUNEAU V., SMITH N. A.: A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Atlanta, Georgia, June 2013), Association for Computational Linguistics. 1
- [DN21] DOU Z.-Y., NEUBIG G.: Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (Online, Apr. 2021), Association for Computational Linguistics. 1, 4
- [Ger08] GERMAN U.: Yawat: yet another word alignment tool. In *Proceedings of the ACL-08: HLT demo session* (2008), pp. 20–23. 2
- [GSK14] GILMANOV T., SCRIVNER O., KÜBLER S.: Swift aligner, a multifunctional tool for parallel corpora: Visualization, word alignment, and (morpho)-syntactic cross-language transfer. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (2014), pp. 2913–2919. 2
- [HA11] HOLMQVIST M., AHRENBERG L.: A gold standard for English-Swedish word alignment. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)* (Riga, Latvia, May 2011), Northern European Association for Language Technology (NEALT). 1, 2
- [IJS*21] IMANI GOOGHARI A., JALILI SABET M., DUFTER P., CYSOU M., SCHÜTZE H.: ParCourE: A parallel corpus explorer for a massively multilingual corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations* (Online, Aug. 2021), Association for Computational Linguistics. 2
- [JSDYS20] JALILI SABET M., DUFTER P., YVON F., SCHÜTZE H.: SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (Online, Nov. 2020), Association for Computational Linguistics. 1, 2, 4
- [LDGBM05] LAMBERT P., DE GISPERT A., BANCHS R., MARINO J. B.: Guidelines for word alignment evaluation and manual alignment. *Language resources and evaluation* 39, 4 (2005), 267–285. 1
- [Mac10] MACKEN L.: An annotation scheme and gold standard for dutch-english word alignment. In *7th conference on International Language Resources and Evaluation (LREC 2010)* (2010), European Language Resources Association (ELRA), pp. 3369–3374. 1, 2
- [Mar08] MAREČEK D.: *Automatic Alignment of Tectogrammatical Trees from Czech-English Parallel Corpus*. Master's thesis, Charles University, MFF UK, 2008. 1, 2
- [MP03] MIHALCEA R., PEDERSEN T.: An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond* (2003), pp. 1–10. 1, 2
- [ON00] OCH F. J., NEY H.: Improved statistical alignment models. *Association for Computational Linguistics*. 1, 2
- [ON03] OCH F. J., NEY H.: A systematic comparison of various statistical alignment models. *Computational Linguistics* 29, 1 (2003). 1
- [ÖT16] ÖSTLING R., TIEDEMANN J.: Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics* 106 (October 2016), 125–146. 1
- [PFY21] PALLADINO C., FORADI M., YOUSEF T.: Translation alignment for historical language learning: a case study. *Digital Humanities Quarterly* 15, 3 (2021). 1
- [Shn03] SHNEIDERMAN B.: The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*. Elsevier, 2003, pp. 364–371. 2
- [SLW21] STEINGRÍMSSON S., LOFTSSON H., WAY A.: CombAlign: a tool for obtaining high-quality word alignments. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)* (Reykjavik, Iceland (Online), May 31–2 June 2021), Linköping University Electronic Press, Sweden. 1
- [SS15] STEELE D., SPECIA L.: WA-continuum: Visualising word alignments across multiple parallel sentences simultaneously. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations* (Beijing, China, July 2015), Association for Computational Linguistics and The Asian Federation of Natural Language Processing. 2
- [SZW21] SHI H., ZETTMLOYER L., WANG S. I.: Bilingual lexicon induction via unsupervised bitext construction and word alignment. *CoRR abs/2101.00148* (2021). 1
- [VNT96] VOGEL S., NEY H., TILLMANN C.: HMM-based word alignment in statistical translation. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics* (1996). 1
- [XQV*21] XIA P., QIN G., VASHISHTHA S., CHEN Y., CHEN T., MAY C., HARMAN C., RAWLINS K., WHITE A. S., VAN DURME B.: Lome: Large ontology multilingual extraction. *arXiv preprint arXiv:2101.12175* (2021). 1
- [YJ20] YOUSEF T., JANICKE S.: A survey of text alignment visualization. *IEEE transactions on visualization and computer graphics* 27, 2 (2020), 1149–1159. 2
- [YPSF22] YOUSEF T., PALLADINO C., SHAMSIAN F., FORADI M.: Translation alignment with ugarit. *Information* 13, 2 (2022). URL: <https://www.mdpi.com/2078-2489/13/2/65>, doi: 10.3390/info13020065. 1, 2
- [YPWB22] YOUSEF T., PALLADINO C., WRIGHT D. J., BERTI M.: Automatic translation alignment for ancient greek and latin, Apr 2022. URL: osf.io/8epsy, doi:10.31219/osf.io/8epsy. 2