EVALIGN: Visual Evaluation of Translation Alignment Models

Tariq Yousef † **Gerhard Heyer** † **Stefan Jänicke**‡

†Leipzig University ‡University of Southern Denmark

<tariq.yousef@uni-leipzig.de>

Abstract

This paper presents EVALIGN, a visual analytics framework for quantitative and qualitative evaluation of automatic translation alignment models. EVALIGN offers various visualization views enabling developers to visualize their models' predictions and compare the performance of their models with other baseline and state-of-the-art models. Through different search and filter functions, researchers and practitioners can also inspect the frequent alignment errors and their positions. EVALIGN hosts nine gold standard datasets and the predictions of multiple alignment models. The tool is extendable, and adding additional datasets and models is straightforward. EVALIGN can be deployed and used locally and is available on GitHub[1].

1 Introduction

Translation Alignment is the process of finding and linking translation equivalents between a text and its translations. It can be performed on different granularity levels. Word-level Translation Alignment plays a key role in several NLP and Digital Humanities tasks such as statistical machine translation (Brown et al., 1993; Koehn et al., 2003), cross-lingual transfer (Hinrichs et al., 2022; Jacqmin et al., 2021), classical language learning (Palladino et al., 2021; Palladino, 2020), dynamic dictionaries induction (Shi et al., 2021), Word Sense Disambiguation (Luan et al., 2020)and analyzing neural machine translation systems (Alkhouli et al., 2016).

The work on automatic translation alignment started 30 years ago when Brown et al. (1993) introduced the first statistical alignment models called IBM models. Later, several tools and models such as Giza++ (Och and Ney, 2003) and fast_align (Dyer et al., 2013) were developed based on Brown's models with different improvements and optimization additions. With the recent advances in neural machine translation systems and the emergence of pre-trained multilingual transformer models (Devlin et al., 2018; Conneau et al., 2019), it has been possible to develop neural alignment models that significantly outperform the statistical models without needing extensive training datasets.

Performance evaluation of alignment models is essential, and many ground truth datasets have been developed for this purpose. Precision, Recall, F1, and Alignment Error Rate (AER) are used as indicators of the alignment quality. Although they are widely used, quantitative metrics have their limitations (Ayan and Dorr, 2006; Vilar et al., 2006; Lambert et al., 2005). Thus, additional qualitative evaluation is required for a better understanding of the models behaviors.

For this purpose, we introduce EVALIGN, a tool for quantitative and qualitative evaluation of automatic alignment models that allows developers to estimate the quality of alignment models and get insights into their performance. With multiple visualization approaches and tailored views, the proposed framework helps researchers and developers working on automatic translation alignment models inspect their predictions with different gold standard data sets and compare their performance to other baseline and state-of-the-art models quantitatively and qualitatively. Further, it supports non-experts who want to employ alignment models in their research or business to explore different alignment models and their performance on texts in different languages to choose the suitable model for their purpose. EVALIGN is available online[2] and the online demo hosts nine benchmark datasets and five alignment methods combined with four different embeddings models (20 models in total); EVALIGN can be deployed locally, and users can add new datasets and import new models.

[1] https://github.com/TariqYousef/EVALign

[2] http://evalign.info/

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics System Demonstrations, pages 277–297
May 2-4, 2023 ©2023 Association for Computational Linguistics

2 Related Works

Employing visualization for exploring benchmark data sets, analyzing models' behaviour, and conducting qualitative evaluation is common practice in NLP. The Language Interpretability Tool *LIT* (Tenney et al., 2020) offers several interactive visualization techniques for a broad range of NLP tasks. *DeepCompare* (Murugesan et al., 2019) supports visual and interactive performance comparison of deep learning models. *SummVis* (Vig et al., 2021) and *Summary Explorer* (Syed et al., 2021) support qualitative evaluation for the summarization task. Paper with Code[3] platform allows to track state-of-the-art performance on benchmark datasets for different NLP tasks. *Vis-Eval* (Steele and Specia, 2018), *ASIYA* (Gonzàlez et al., 2012) and *MT-ComparEval* (Klejch et al., 2015) allow for systematic comparison and evaluation of various machine translation models.

Visualizing word-level alignments was the aim of many tools such as *Ugarit* (Yousef et al., 2022b) and *WA-Continuum* (Steele and Specia, 2015), which visualizes word alignment of automatically aligned sentences to facilitate their evaluation. ImaniGooghari et al. (2021) introduced the *Parallel Corpus Explorer* which supports exploring a word-aligned parallel corpus.

To our best knowledge, EVALIGN is the first system that allows researchers and practitioners to qualitatively evaluate the performance of alignment models on multiple gold standard datasets.

3 Automatic Alignment Models

Automatic translation alignment models can be categorized into three main categories: **Statistical models** such as Giza++, fast_align (Dyer et al., 2013), and eflomal (Östling and Tiedemann, 2016). They have been widely used and achieved state-of-the-art performance until recently and are currently used as a baseline. However, the performance of the statistical models is governed by the availability of training corpora in the form of parallel sentences. **Neural Models** utilize neural machine translation models or multilingual transformer models to capture word-level translation alignment. Different workflows are available. For instance, extracting alignment using embeddings-based semantic similarity by employing pre-trained and fine-tuned multilingual contextualized embeddings such as SIMALIGN (Jalili Sabet et al., 2020), AWEASOME (Dou and Neubig, 2021), XLM-ALIGN (Chi et al., 2021), and MirrorAlign(Wu et al., 2022). **Hybrid Models** combine statistical and neural models aiming for better performance, for instance, by using the output of statistical models as supervision to train neural models (Alkhouli et al., 2018).

[3] https://paperswithcode.com/

4 Evaluation

4.1 Alignment Gold Standards

Gold standards are the main components for evaluating the performance of NLP models. Developing alignment gold standards involves multiple domain experts (at least 2) to avoid any bias in the manual annotation process. Annotators must follow predefined guidelines to reduce disagreements and ensure consistency and quality of the manual alignments. Moreover, Inter-Annotator Agreement (IAA) can be computed to validate the reliability and quality of the alignment guidelines and gold standard. The gold standard dataset is a list of manually aligned sentences, each sentence has a list of translation pairs, and each translation pair is assigned one of two categories, SURE (S) or POSSIBLE (P).

Table 1 shows that most literature papers evaluated their models mainly on three alignment datasets, German-English, English-French (Och and Ney, 2003), and Romanian-English (Mihalcea and Pedersen, 2003). However, several datasets are available in various languages (Table 2), but they have not yet been used for performance evaluation. For this reason, EVALIGN hosts some of the "unused" datasets, and we generated the alignments of different alignment models for these datasets and made them available for users for further experiments.

4.2 Evaluation Metrics

In addition to the classical quantitative evaluation metrics Precision, Recall, and F1; researchers utilize the Alignment Error Rate (AER) (Och and Ney, 2003). These metrics are based on the overlap between the model's predictions A with the SURE (S), and POSSIBLE (P) alignment sets of the gold standards (Equation 1). Lambert et al. (2005) studied the influence of the amount of SURE and POSSIBLE alignments in the gold standard on AER and concluded that AER would be smaller when the S/P ratio is low, and vice versa. Figure 4 confirms this conclusion.

$$AER = 1 - \frac{|A \cap P| + |A \cap S|}{|A| + |S|} \quad (1)$$

4.3 Limitations

Evaluating the performance of translation alignment models is a complex task, even for humans. In many cases, it is challenging to tell if an alignment between two tokens/phrases is entirely correct because that relies on several factors, mainly the text genre, context, translation quality, and human annotator's knowledge.

AER is highly affected by the gold standard dataset, i.e., the selection of sure and possible translation pairs and their proportions of the whole dataset. And the gold standard alignments are subject to the alignment guidelines and annotators' agreement, which is also influenced by the characteristics of the selected corpus, the annotators' knowledge, and the target application. That means it might be possible to have correct alignments predicted by the models, but the gold standards do not consider them. Thus, AER will treat them as incorrect alignments. We encountered such cases repeatedly while inspecting the existing gold standard datasets.

AER is intolerant; it considers all tokens equally important and there is no distinction between function words and context words. In the example illustrated in Figures 5A and 5C, AER penalizes a missing alignment of the full-stop the same as missing alignment of $Madrid$.

Further, AER fails to capture phrase misalignments. In Figure 5B, the German word *auch* must be aligned to the English phrase $aswell$, producing two sure alignments $auch - as$ and $auch - well$. Nevertheless, If an alignment model aligns only a part of the phrase, $auch - well$, this will be considered a correct alignment, while it is not because there is no constraint saying that the model must produce the two sure alignments together in order to count them as correct alignments. Also, AER does not consider null-alignments, i.e. tokens with no translation equivalents in the parallel sentence. Thus, quantitative evaluation gives an overview of models' performance, but it is limited and must be accompanied by qualitative evaluation.

All these reasons motivated us to develop EVALIGN. The framework allows users to explore quantitative evaluation metrics and also provides the ability to conduct an extensive qualitative evaluation using different interactive visualization views and filtering options. Additionally, we proposed two metrics to overcome the limitations above. The ALIGNMENT COVERAGE represents the portion of the aligned tokens out of all tokens in the dataset. It can computed for the gold standard dataset and for models' predictions. We compute *Coverage* as follows:

$$Coverage = 1 - \frac{|S_n| + |T_n|}{|S| + |T|} \quad (2)$$

Where S and T are the sets of all tokens in the source and target sentences, respectively, S_n and T_n are the sets of null-alignments in the source sentences and target sentences.

The PHRASE ALIGNMENT ACCURACY (PAC) measures the model's ability to align phrases. Phrase alignment appears when a token in one sentence is aligned to multiple tokens in the corresponding translation (one-to-many or many-to-one), or when multiple tokens in one sentence are aligned to multiple tokens in the corresponding sentence (many-to-many). Our definition of phrase does not constrain that the tokens must be consecutive. However, the phrase is correctly aligned if all its tokens are aligned with each other. For instance, the English phrase $public\ health\ policy$ and the German equivalent $Gesundheitspolitik$ are aligned correct if, and only if the model predicts $public - Gesundheitspolitik$, $health - Gesundheitspolitik$ and $policy - Gesundheitspolitik$ pairs. Because all tokens contribute to the meaning of the phrases, and missing any token changes the meaning or make it incomplete. We compute *PAC* as stated in Equation 3:

$$PAC = \frac{|P_m \cap P_{gs}|}{|P_{gs}|} \quad (3)$$

Where P_{gs} is the aligned phrases set of the gold standard, and P_m is the set of predicted aligned phrases by the model. Figure 14 compares the performance of the best five alignment models on the German-English dataset, the models use the fine-tuned mBERT embeddings.

5 Implementation Details

We surveyed the automatic alignment papers published after 2019 (Table 1). Most researchers evaluate their models performance on at least three benchmark datasets, mainly German-English, French-English, and Romanian-English. We used these three datasets in addition to six other datasets (English-French, English-Spanish, English-Portuguese, Spanish-French, Portuguese-Spanish, and Portuguese-French) that have not been used before for evaluation.

Regarding the alignment models, we selected embeddings-based $Softmax$, $Entmax$ (Dou and Neubig, 2021), $Argmax$, $Itermax$, and $Match$ (Jalili Sabet et al., 2020) with different contextualized embeddings, namely, mBERT, XLM-R, fine-tuned mBERT[4], and XLM-Align[5] (Chi et al., 2021). In addition to Giza++, fast_align, and EfLoMAl for the datasets DE-EN, EN-FR, and RO-EN. Our selection was subject to the implementation availability and reproducibility. We used the default implementations provided by authors in their GitHub repositories with the default parameters. The backend API is implemented using Django and Postgres database, while the visualization views are created with React JS and D3.js.

5.1 User Interface

Figure 1 illustrates EVALIGN usage workflow; users start navigating through the tool by selecting a dataset from the landing page, which lists all hosted benchmark datasets or selecting a model from the models page, which lists all hosted models or selecting a model or dataset from the aggregated overview. EVALIGN offers five main views:

Single Dataset vs Multiple Models (V1). This view provides a performance overview of the alignment models hosted on EVALIGN over the selected dataset using a *bar chart*. The overview allows users to select among different quantitative evaluation metrics, namely, Precision, Recall, F1, AER, Coverage, PAC, and the number of translation pairs. The view also visualizes all sentences of the selected dataset using a grid view allowing users to inspect the possible and sure alignments and assess their correctness and coverage (Figure 7A) . From this view, users can select a single model to inspect its performance on a specific dataset.

Single Model vs Multiple Datasets (V2). This view provides a summarized performance overview of the selected alignment model over different benchmark datasets using a *bar chart* that allows switching among different evaluation metrics. Selecting a dataset will forward the user to *Single Models vs Single Dataset* view.

Single Model vs Single Dataset (V3). This view offers various corpus-level and sentence level visualization views providing the user with all needed functions to inspect the the dataset sentences and explore the alignments predicted by the selected model. This view aggregate wrong alignments, missing alignments and correct alignments to facilitate the analysis of the model performance. Further it shows the relation of the different evaluation metrics with sentence lengths. The predicted alignments are visualized with *Grid*, *Side-by-side*, and *Table* views. Moreover, it offers various sorting, filtering and searching options to support qualitative evaluation (Figures 7B and 7C).

Two Models vs Single Dataset (V4). In this view, users can compare two models at sentence-level using the *Grid* and *Table* view which show the agreement and disagreement between the two models (figure 7D and 16).

All Models vs all Datasets (V5). In this aggregated view, all hosted datasets and models are presented in a table. Users can switch between different quantitative metrics with different sorting options (figure 13).

5.2 Visual Design

EVALIGN offers a variety of corpus and sentence level visualization views in addition to several searching and filtering functions. When designing EVALIGN, we consulted the text alignment visualization survey (Yousef and Jänicke, 2020) and adapted Schneiderman's Information Seeking Mantra (Shneiderman, 2003) "Overview first, zoom and filter, then details-on-demand" to facilitate interactive navigation through the benchmark datasets and alignment models.

5.2.1 Corpus-level Views

Corpus-level views provide comprehensive overviews of the compared models by visualizing aggregated statistics and evaluation metrics at the dataset level. A *bar chart* on the dataset page will be shown in the upper left corner, allowing users to compare available alignment models. A button bar is located above the bar chart, allowing users to switch between the evaluation metric. Each model is assigned a unique color (Figure 6A). A *bar chart* on the dataset page is placed in the upper

[4]https://github.com/neulab/awesome-align
[5]https://huggingface.co/microsoft/xlm-align-base

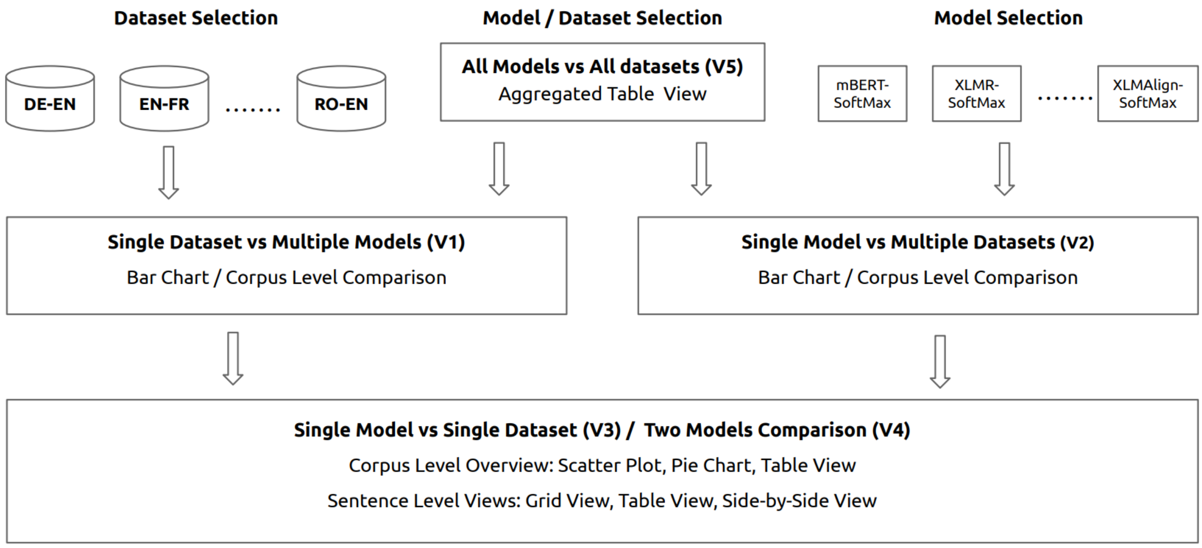

Figure 1: Overview of EVALIGN

left corner, allowing users to compare available alignment models. A button bar is located above the bar chart, allowing users to switch between seven evaluation metrics: AER, Precision, Recall, F1, Number of Translation Pairs, Coverage, and PAC. Each model is assigned a unique color (Figure 6A) and hovering a bar will show a tooltip with the corresponding information.

Selecting a model with a mouse click on the corresponding bar will load a *pie chart* that shows a comparison between the model predictions and the gold standard. We distinguish among three sets (Figure 3): i) *correct alignments*, where the model predictions match the gold standard, shown in green. ii) *wrong alignments*, where the model failed to align the translation pairs correctly, shown in red. iii) *missing alignments*, the pairs the model was supposed to align, shown in orange. Clicking on any of the three sets will load the corresponding translation pairs in the neighbouring table, which aggregates the translation pairs and shows them with their frequency. Moreover, the translation pairs are clickable, and the corresponding gold standard sentences with sentence-level views will be displayed when clicked.

Further, users can switch between the *pie chart* and the *scatter plot*, which displays the relation between the sentence length (x-axis) and the selected evaluation metric (y-axis) of the selected alignment model; each sentence is presented as one dot (Figure 6B). The *scatter plot* helps users detect outliers and interesting observations, such as the relation between the AER and the sentence length. Moreover, a range selector allows filtering of the dataset by selecting multiple sentences to be visualized at the sentence level for more detailed inspection. Further, the evaluation metrics will be calculated for the selected sentences and displayed under the scatter plot. This allows users to eliminate subsets (for example, short or long sentences) and see how these subsets affect the quantitative evaluation metrics. The selected sentences will be displayed as paginated list of sentence-level views. The tool provides sorting options according to the selected metric.

5.2.2 Sentence-level Views

The sentence-level views aim to show the alignment among words of the source and target sentences. The framework provides two sentence-level views, namely, *grid view* and *side-by-side view*. The views are accompanied with a *bar chart* showing the sentence-level evaluation metrics of the hosted models and enabling users to select a model to visualize its output for the corresponding sentence. The *grid view* places the two sentences as a grid. The source sentence tokens are placed vertically, and the target tokens are placed horizontally. The gold standard Sure and Possible alignments will be displayed in the corresponding cells as big and small dots, respectively. The *grid view* is suitable for visualizing the alignments of a single model by coloring the corresponding cells with the model's unique color. It is also appropriate to visualize the alignments of two models and their agreement (Figure 2).

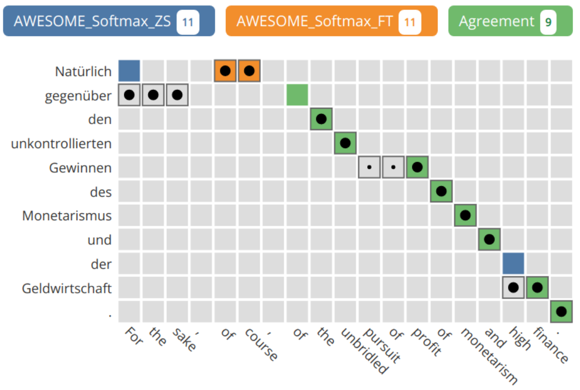

Figure 2: A Grid View to visualize the alignment at sentence level.

The *side-by-side view* places the two sentences alongside each other; it utilizes the mouse hover to highlight the hovered token and the aligned tokens in the parallel sentence. The current implementation of this view allows visualizing the alignment of a single model, and users can switch between models via a neighboring *bar chart*.

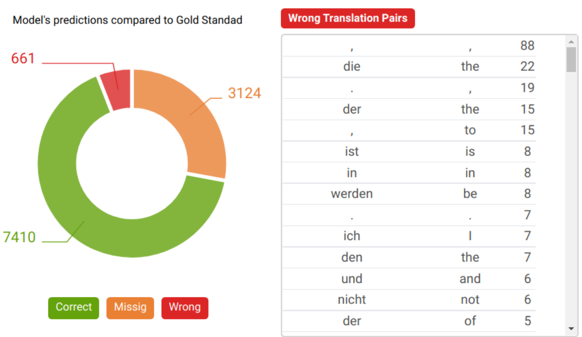

Figure 3: A Pie Chart shows the overlap between the model's predictions and the gold standard in three categories. The neighboring table shows the translation pairs with their frequency of a selected category.

6 Usage Scenarios

The framework offers a variety of usage scenarios that can be summarized as follows:

Gold standard quality control. Visualizing the gold standard datasets using the *Grid View* allowed us to inspect their accuracy and assess their quality. The analysis of the English-French (EN-FR) dataset showed that the dataset contains several single or two token sentences, for which the alignment will always be correct (figure 9C-D). Moreover, some sentences occur more than once in the dataset, and that would affect the evaluation process since they increase recall and precision and consequently reduce AER (figure 9B,C,E). The inspection showed that there are several sentences with plenty of *Possible* alignments and few or no *Sure* alignments (figure 9A,F).

Comparing datasets' characteristics. Users can see all hosted datasets on the datasets page with different statistics on the number of sentences, tokens, sure and possible alignments and their percentages. For instance, the English-French (EN-FR) dataset has significantly more possible alignments than sure alignments (figure 15). This explains why all alignment models have the lowest AER on this dataset compared to all other datasets. The same applies to the Romanian-English (RO-EN) dataset since it only has sure alignments, which explains why the AER is always higher than the other datasets.

Comparing model performance with different configurations. As an example, we compare the performance of $Softmax$ with two different embeddings models,namely, mBERT and a fine-tuned mBERT, to estimate the improvement achieved with the fine-tuning process. In addition to comparing all quantitative metrics, the framework allows filtering sentences where model A outperforms model B. Figure 2 shows that the fine-tuning enhanced the overall alignment accuracy and allowed to predict two more correct *Sure* alignments and eliminate two incorrect ones.

Comparing quantitative metrics. The framework provides different options to compare the models performance using different quantitative metrics at corpus and sentence levels using bar chart and table views. The aggregated results in the table view (V5) reveals that the fine-tuned mBert achieved the best results in all datasets regarding AER. While $Itermax$ achieved the best Recall on all datasets, $Argmax$ with fine-tuned mBert embeddings achieved best precision on 7 datasets and second best precision on 2 datasets. Further, $Itermax$ with the fine-tuned mBERT embeddings achieved the best Phrase Alignment Accuracy on all datasets. The $Match$ algorithm generates more translation pairs than all other algorithms, and $Entmax$ with XLM-RoBERTa embeddings generates always less translation pairs that all other algorithms.

Analyzing alignment errors. From the pie chart provided for the *Single Dataset – Single Model* view (figure 3), we can click on the red arc that represents the wrong alignment pairs to list all incorrect pairs produced by the model. Our analysis of $Itermax$ with the fine-tuned mBert model on the German-English (DE-EN) dataset revealed the following:

- The most frequent wrong pairs involve a punctuation mark in one or both languages. However, such issues can be avoided by adding constraints that prevent aligning punctuation to a word (Figure 18).

- Long sentences with repeated tokens are more likely to produce incorrect alignments despite that the pairs are correct translations, but their positions in the two sentences do not correspond (Figure 19).

- The majority of wrong pairs are function words, such as articles, pronouns, prepositions, and conjunctions, and most of them are semantically correct translations such as ($nicht$ - not) (Figure 12C).

- The German-English (DE-EN) dataset contains incorrect alignments. For instance, in sentence *10*, the model generated the correct pair *präzise - precise*. However, it is classified as wrong because the gold standard aligns *präzise* with $very$ and $sind$ with $precise$, which is incorrect. Moreover, some sentences are not entirely aligned, and many tokens are left. For example, in sentence *40* (Figure 8), there are many *correct* translation pairs predicted by the model such as $Soziale - social$ and $Sicherheit - security$, but they are not included in the gold standard. However, these errors are not model-specific but apply to different alignment models and datasets (Figure 11).

7 Conclusion

Evaluating translation alignment models is a non-trivial task. Qualitative evaluation is needed because quantitative evaluation metrics do not reflect the real quality of the alignment models due to many factors. For this purpose, we presented the framework EVALIGN that supports quantitative and qualitative evaluation of automatic alignment systems. EVALIGN hosts several evaluation datasets and various alignment models. It offers different visualization views and filtering functions to help users to investigate alignment datasets and models and conduct various quality analyses. Moreover, we presented different usage scenarios that showcase the use and effectiveness of the tool.

Our analyses revealed that gold standard datasets, especially the German-English (DE-EN) and French-English (FR-EN), which have been used almost in all related works on automatic alignment, contain plenty of errors and need to be revised and corrected by linguists and domain experts. In future work, we aim to incorporate morphological features such as POS, lemma and named entities to assess model performances and classify alignment errors.

Finally, we will keep the tool updated by adding new datasets and/or models, and we encourage researchers to send us the output of their new models to publish them on EVALIGN. A short video demonstrating the tool is available on youtube https://youtu.be/hfii6x0bktw

Limitations

Some literature papers do not share the source code or their models' output. Therefore, we could not host their models on EVALIGN. Also, not all datasets mentioned in the literature are accessible.

Regarding the visualization views, the current tool implementation allows for comparing two alignment models simultaneously at the sentence level. Also, the sentence-level side-by-side visualize only one model's alignments. The view does not allow comparing two or more models. The grid view is not suitable for long sentences.

Ethics Statement

The datasets hosted on EVALIGN are downloaded from their authors' websites. The datasets are well-known and have been used for evaluation in most literature papers. Model predictions are generated using the code published on developers' Github repositories. We have not retrained or fine-tuned any language models and used the publicly available language models on Huggingface. The tool offers visualization views to facilitate the performance evaluation to get a better understanding of models' behaviours.

References

Tamer Alkhouli, Gabriel Bretschner, and Hermann Ney. 2018. On the alignment problem in multi-head attention-based neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 177–185, Brussels, Belgium. Association for Computational Linguistics.

Tamer Alkhouli, Gabriel Bretschner, Jan-Thorsten Peter, Mohammed Hethnawi, Andreas Guta, and Hermann

Ney. 2016. Alignment-based neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 54–65.

Necip Fazil Ayan and Bonnie Dorr. 2006. Going beyond aer: An extensive analysis of word alignments and their impact on mt. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 9–16.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Chi Chen, Maosong Sun, and Yang Liu. 2021. Mask-align: Self-supervised neural word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4781–4791, Online. Association for Computational Linguistics.

Yun Chen, Yang Liu, Guanhua Chen, Xin Jiang, and Qun Liu. 2020. Accurate word alignment induction from neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 566–576, Online. Association for Computational Linguistics.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, He-Yan Huang, and Furu Wei. 2021. Improving pretrained cross-lingual language models via self-labeled word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3418–3430.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Shuoyang Ding, Hainan Xu, and Philipp Koehn. 2019. Saliency-driven word alignment interpretation for neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 1–12, Florence, Italy. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. 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*, pages 2112–2128, Online. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. 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*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4453–4462, Hong Kong, China. Association for Computational Linguistics.

Meritxell Gonzàlez, Jesús Giménez, and Lluís Màrquez. 2012. A graphical interface for MT evaluation and error analysis. In *Proceedings of the ACL 2012 System Demonstrations*, pages 139–144, Jeju Island, Korea. Association for Computational Linguistics.

Joao Graca, Joana Paulo Pardal, Luísa Coheur, and Diamantino Caseiro. 2008. Building a golden collection of parallel multi-language word alignment. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Nway Nway Han and Aye Thida. 2019. Annotated guidelines and building reference corpus for myanmar-english word alignment. *arXiv preprint arXiv:1909.11288*.

Nicolás Hinrichs, Maryam Foradi, Tariq Yousef, Elisa Hartmann, Susanne Triesch, Jan Kaßel, and Johannes Pein. 2022. Embodied metarepresentations. *Frontiers in neurorobotics*, 16.

Anh Khoa Ngo Ho and François Yvon. 2020. Generative latent neural models for automatic word alignment. *arXiv preprint arXiv:2009.13117*.

Maria Holmqvist and Lars Ahrenberg. 2011. A gold standard for English-Swedish word alignment. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 106–113, Riga, Latvia. Northern European Association for Language Technology (NEALT).

Ayyoob ImaniGooghari, Masoud Jalili Sabet, Philipp Dufter, Michael Cysou, and Hinrich Schütze. 2021. 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*, pages 63–72, Online. Association for Computational Linguistics.

Léo Jacqmin, Gabriel Marzinotto, Justyna Gromada, Ewelina Szczekocka, Robert Kołodyński, and Géraldine Damnati. 2021. SpanAlign: Efficient sequence tagging annotation projection into translated data applied to cross-lingual opinion mining. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 238–248, Online. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Ondrej Klejch, Eleftherios Avramidis, Aljoscha Burchardt, and Martin Popel. 2015. Mt-compareval: Graphical evaluation interface for machine translation development. *Prague Bull. Math. Linguistics*, 104:63–74.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Ivana Kruijff-Korbayová, Klára Chvátalová, and Oana Postolache. 2006. Annotation guidelines for Czech-English word alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Patrik Lambert, Adria DE GISPERT, Rafael BANCHS, and José B MARINO. 2005. Guidelines for word alignment evaluation and manual alignment. *Language resources and evaluation*, 39(4):267–285.

Yixing Luan, Bradley Hauer, Lili Mou, and Grzegorz Kondrak. 2020. Improving word sense disambiguation with translations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4055–4065.

Lieve Macken. 2010. An annotation scheme and gold standard for Dutch-English word alignment. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Kelly Marchisio, Conghao Xiong, and Philipp Koehn. 2021. Embedding-enhanced giza++: Improving alignment in low-and high-resource scenarios using embedding space geometry. *arXiv preprint arXiv:2104.08721*.

Rada Mihalcea and Ted Pedersen. 2003. 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*, pages 1–10.

Sugeerth Murugesan, Sana Malik, Fan Du, Eunyee Koh, and Tuan Manh Lai. 2019. Deepcompare: Visual and interactive comparison of deep learning model performance. *IEEE Computer Graphics and Applications*, 39(5):47–59.

Masaaki Nagata, Katsuki Chousa, and Masaaki Nishino. 2020. A supervised word alignment method based on cross-language span prediction using multilingual BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 555–565, Online. Association for Computational Linguistics.

Anh Khoa Ngo Ho and François Yvon. 2021. Optimizing word alignments with better subword tokenization. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 256–269, Virtual. Association for Machine Translation in the Americas.

Franz Josef Och and Hermann Ney. 2000. A comparison of aligment models for statistical machine translation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-ACL 2000)*.

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

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Chiara Palladino. 2020. Reading texts in digital environments: Applications of translation alignment for classical language learning. *J. Interact. Technol. Pedagog*, 18:724–731.

Chiara Palladino, Maryam Foradi, and Tariq Yousef. 2021. Translation alignment for historical language learning: a case study. *Digital Humanities Quarterly*, 15(3).

Haoyue Shi, Luke Zettlemoyer, and Sida I. Wang. 2021. Bilingual lexicon induction via unsupervised bitext construction and word alignment. *CoRR*, abs/2101.00148.

Ben Shneiderman. 2003. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*, pages 364–371. Elsevier.

David Steele and Lucia Specia. 2015. WA-continuum: Visualising word alignments across multiple parallel sentences simultaneously. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 121–126, Beijing, China. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

David Steele and Lucia Specia. 2018. Vis-eval metric viewer: A visualisation tool for inspecting and evaluating metric scores of machine translation output. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 71–75, New Orleans, Louisiana. Association for Computational Linguistics.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2021. CombAlign: a tool for obtaining high-quality word alignments. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 64–73, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Elias Stengel-Eskin, Tzu-ray Su, Matt Post, and Benjamin Van Durme. 2019. A discriminative neural model for cross-lingual word alignment. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 910–920, Hong Kong, China. Association for Computational Linguistics.

Shahbaz Syed, Tariq Yousef, Khalid Al Khatib, Stefan Jänicke, and Martin Potthast. 2021. Summary explorer: Visualizing the state of the art in text summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 185–194, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. 2020. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 107–118, Online. Association for Computational Linguistics.

Jesse Vig, Wojciech Kryscinski, Karan Goel, and Nazneen Rajani. 2021. SummVis: Interactive visual analysis of models, data, and evaluation for text summarization. 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*, pages 150–158, Online. Association for Computational Linguistics.

David Vilar, Maja Popovic, and Hermann Ney. 2006. AER: do we need to "improve" our alignments? In *Proceedings of the Third International Workshop on Spoken Language Translation: Papers*, Kyoto, Japan.

Di Wu, Liang Ding, Shuo Yang, and Mingyang Li. 2022. MirrorAlign: A super lightweight unsupervised word alignment model via cross-lingual contrastive learning. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 83–91, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Minhan Xu and Yu Hong. 2020. Improving word alignment with contextualized embedding and bilingual dictionary. In *CCF Conference on Big Data*, pages 180–194. Springer.

Tariq Yousef and Stefan Jänicke. 2020. A survey of text alignment visualization. *IEEE transactions on visualization and computer graphics*, 27(2):1149–1159.

Tariq Yousef, Chiara Palladino, Farnoosh Shamsian, Anise d'Orange Ferreira, and Michel Ferreira dos Reis. 2022a. An automatic model and gold standard for translation alignment of ancient greek. In *Proceedings of the Language Resources and Evaluation Conference*, pages 5894–5905, Marseille, France. European Language Resources Association.

Tariq Yousef, Chiara Palladino, Farnoosh Shamsian, and Maryam Foradi. 2022b. Translation alignment with ugarit. *Information*, 13(2).

Tariq Yousef, Chiara Palladino, David J. Wright, and Monica Berti. 2022c. Automatic translation alignment for ancient greek and latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 101–107, Marseille, France. European Language Resources Association.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2020. End-to-end neural word alignment outperforms GIZA++. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1605–1617, Online. Association for Computational Linguistics.

Vilém Zouhar and Daria Pylypenko. 2021. Leveraging neural machine translation for word alignment. *arXiv preprint arXiv:2103.17250*.

A Appendix

Paper	EN-CZ	EN-DE	EN-FR	EN-HI	EN-RO	EN-JA	EN-ZH	EN-IC	EN-VI	EN-FA	EN-AR
(Stengel-Eskin et al., 2019)							x				x
(Garg et al., 2019)		x	x		x						
(Ding et al., 2019)		x	x		x						
(Jalili Sabet et al., 2020)	x	x	x	x	x					x	
(Zenkel et al., 2020)		x	x		x						
(Chen et al., 2020)		x	x		x		x				
(Ho and Yvon, 2020)			x		x						
(Xu and Hong, 2020)			x	x	x						
(Nagata et al., 2020)		x	x		x	x	x				
(Dou and Neubig, 2021)		x	x		x	x	x				
(Zouhar and Pylypenko, 2021)	x	x									
(Steingrímsson et al., 2021)	x	x	x					x			
(Marchisio et al., 2021)		x	x		x						
(Ngo Ho and Yvon, 2021)	x	x	x		x	x			x		
(Chen et al., 2021)		x	x		x		x				
(Chi et al., 2021)		x	x	x	x						
(Wu et al., 2022)		x	x		x						

Table 1: An overview of gold standard datasets that have been used for performance evaluation in the literature papers.

Source	Language Pair	# Sentences	IAA	Text Type
(Och and Ney, 2000)	English-German	508		Verbmobil
(Mihalcea and Pedersen, 2003)	Romanian-English	248		
	English-French	447		Hansard
(Lambert et al., 2005)	English-Spanish	500		Europarl
(Kruijff-Korbayová et al., 2006)	Czech-English	2400	93%	Penn Treebank corpus (WSJ)
(Graca et al., 2008)	English-Portuguese	100	89.5 %	Europarl
	English-Spanish	100	86.7 %	Europarl
	English-French	100	90.8 %	Europarl
	Portuguese-Spanish	100	93.2 %	Europarl
	Portuguese-French	100	93.5 %	Europarl
	Spanish-French	100	96.5 %	Europarl
(Macken, 2010)	Dutch-English	1500	84-94 %	Journalistic texts, Newsletters, and Medical Reports
(Holmqvist and Ahrenberg, 2011)	English–Swedish	1164	91.3%	Europarl
(Steingrímsson et al., 2021)	Icelandic-English	604		ParIce Corpus[6] Project[7]
(Yousef et al., 2022a)	Ancient Greek-English	275	86.17%	Perseus Digital Library
	Ancient Greek-Portuguese	183	83.31%	Perseus Digital Library
(Yousef et al., 2022c)	Ancient Greek-Latin	100	90.50%	DFHG Project[8]
(Han and Thida, 2019)	Myanmar-English	500	91.56%	Myanmar- English ALT parallel corpus

Table 2: An overview of the existing alignment gold standard datasets.

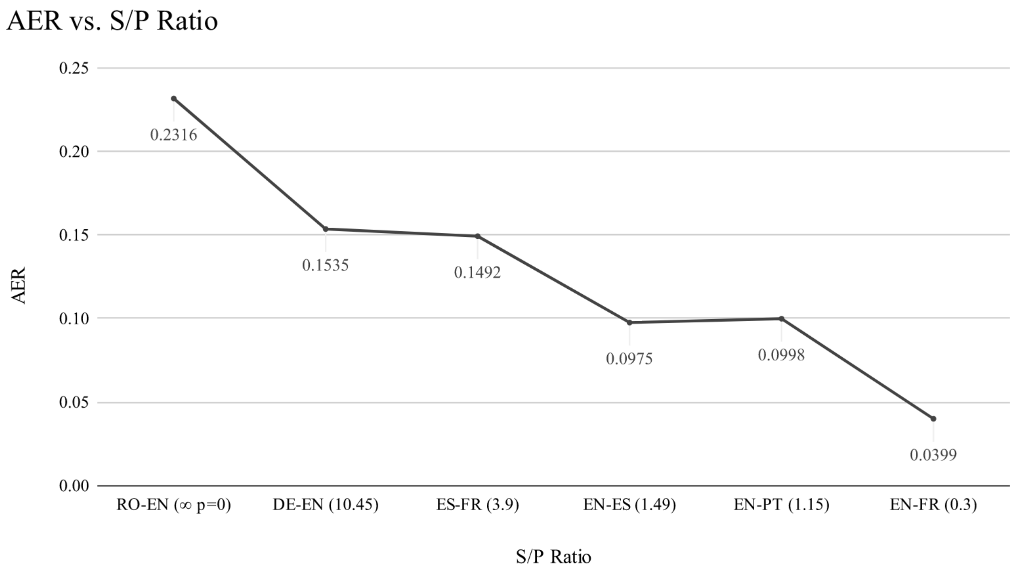

Figure 4: The correlation between AER and S/P Ratio. The alignment model used for this illustration uses Argmax method with fine-tuned mBERT Embeddings.

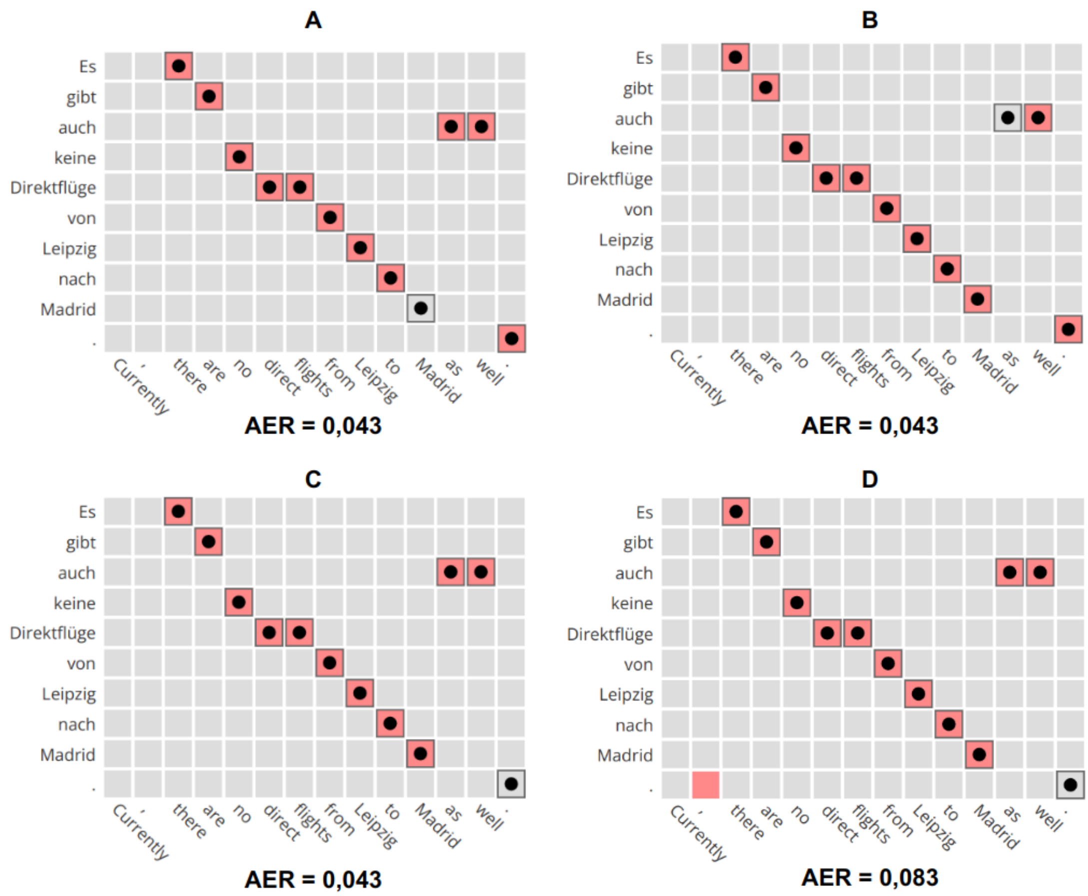

Figure 5: AER Limitations, the bold circles means gold standard sure alignments and colored cells represent model's output. A) The model failed to align *Madrid*. B) The model failed to align *auch* to *as well*. C) The model failed to align ".". D) The model aligned "." incorrectly.

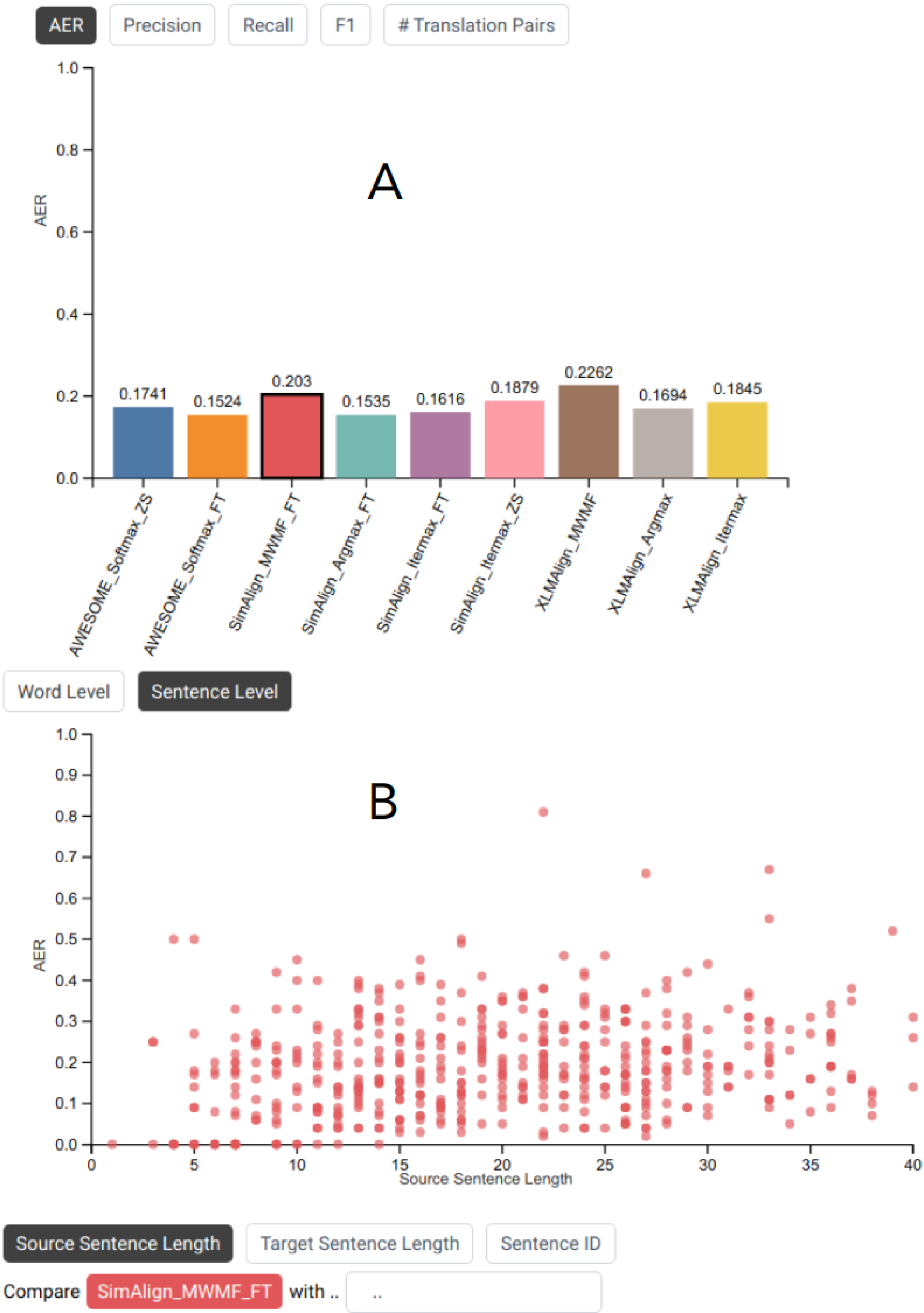

Figure 6: Single Dataset-Single Model view, A) Bar Char to compare the performance of different alignment models according to a selected metric. B) Scatter Plot shows the relation between sentences length and the selected metric.

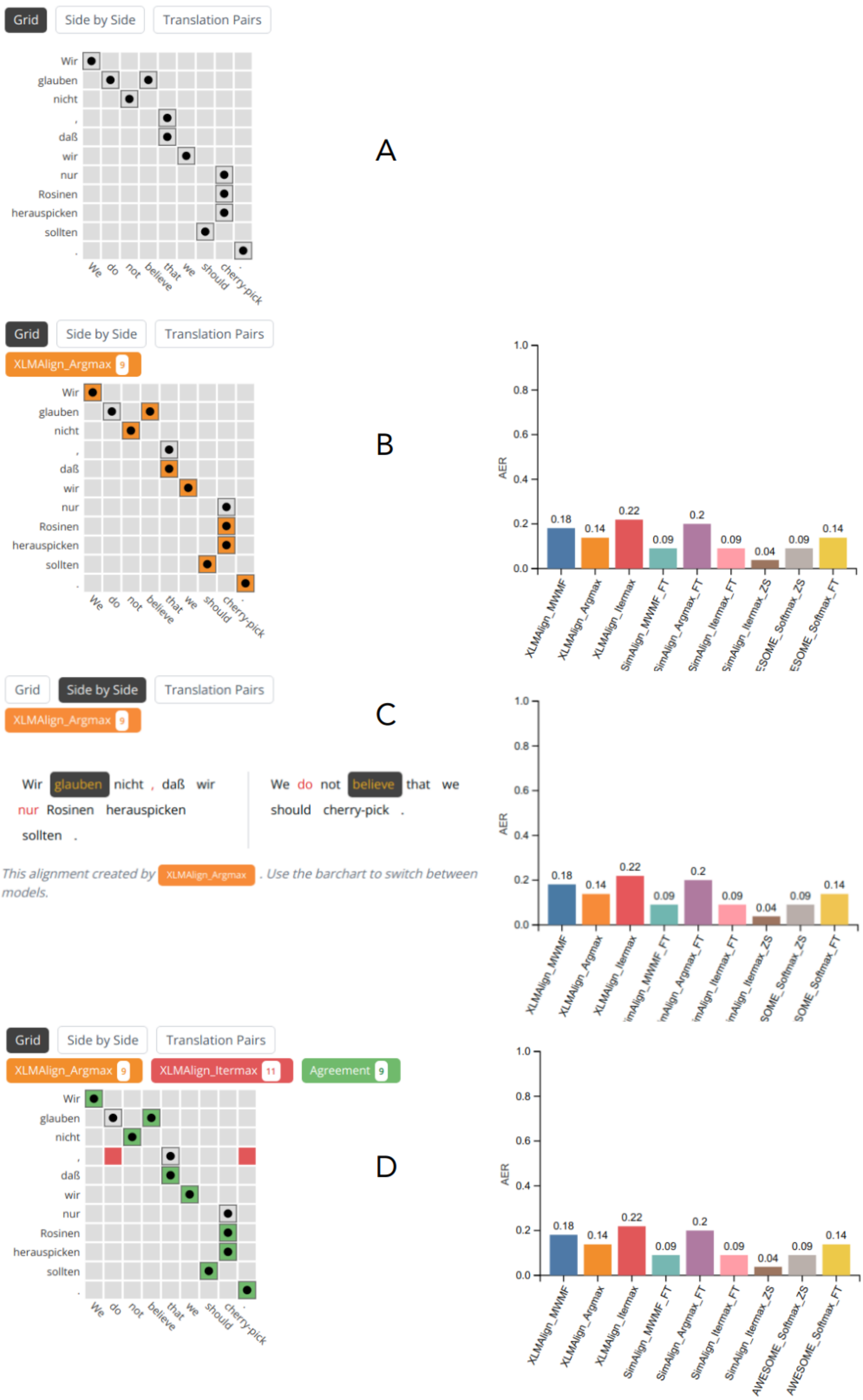

Figure 7: Sentence-level views, A) The default view when no model is selected, showing the sure (big dots) and possible (small dots) alignments. B) Visualizing the alignment of one model. C) the side-by-side view. D) the gird view visualizing the alignments of two models and their agreement.

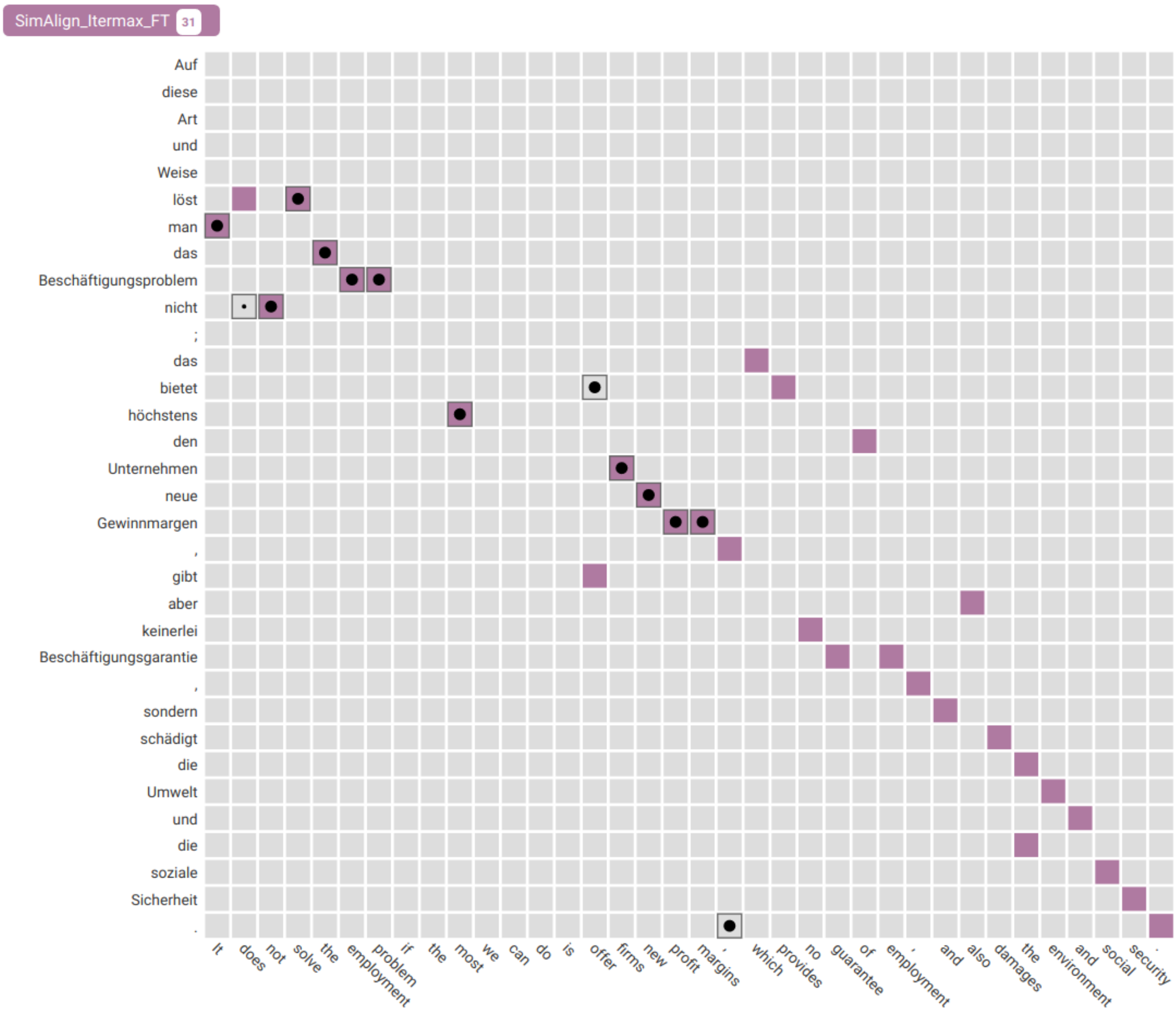

Figure 8: Sentence 40 in the DE-EN dataset, an example of incorrect/incomplete annotation of the gold standard sentence. The model predicts correct translation pairs but they are counted incorrect since they are not included in the gold standard.

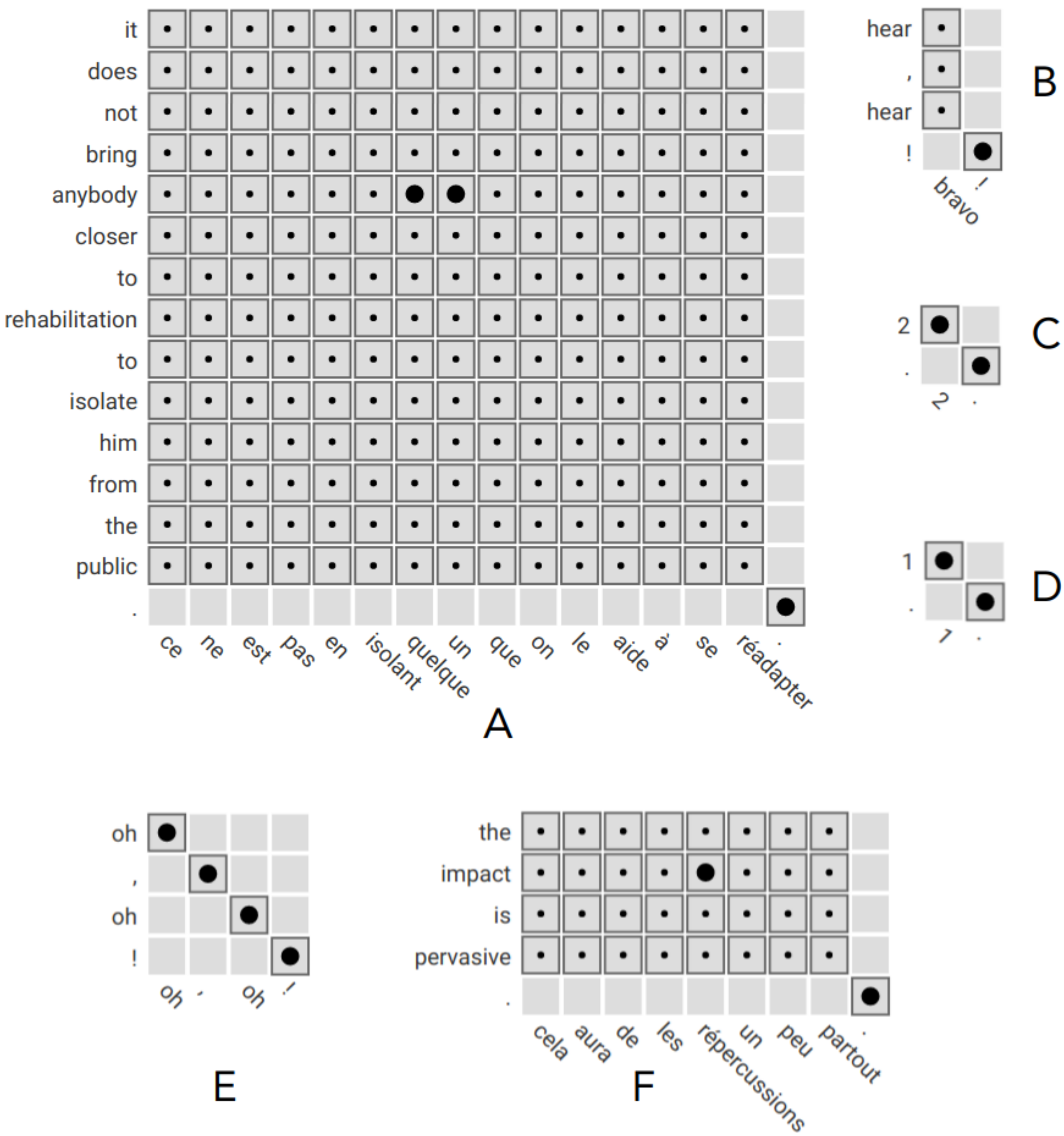

Figure 9: Examples from the EN-FR dataset. A) Sentence *0027* with too many possible links. B) This sentence is repeated 4 times in the datasets in sentences *0007*, *0008*, *0045*, and *0046*. C) This sentence is repeated twice in sentences *0001* and *0002*. D) Sentence *0011*, another example of short sentences with a number and a full stop. E) Short sentence with non-informative tokens repeated 3 times in sentences *0003*, *0004*, and *0005*. F) Sentence *0223*, another example of sentences with too many possible links.

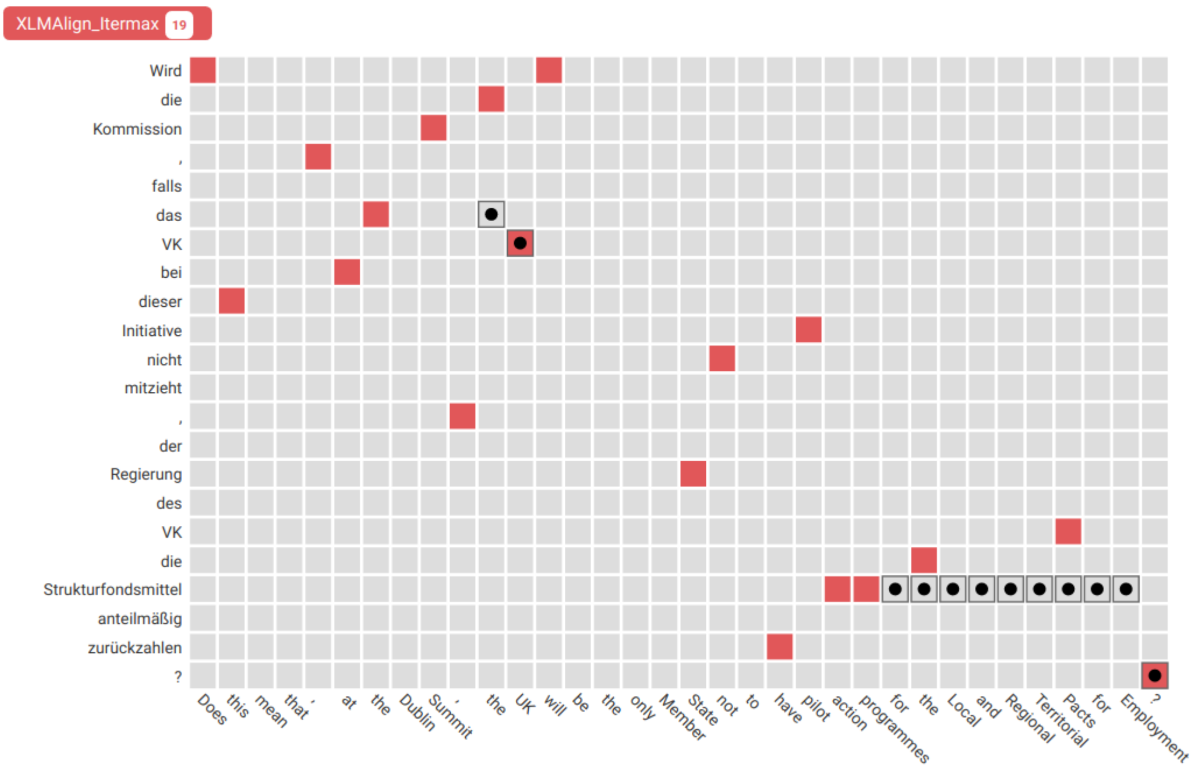

Figure 10: Sentence 202 in the DE-EN dataset, an example of incorrect/incomplete annotation of the gold standard sentence.

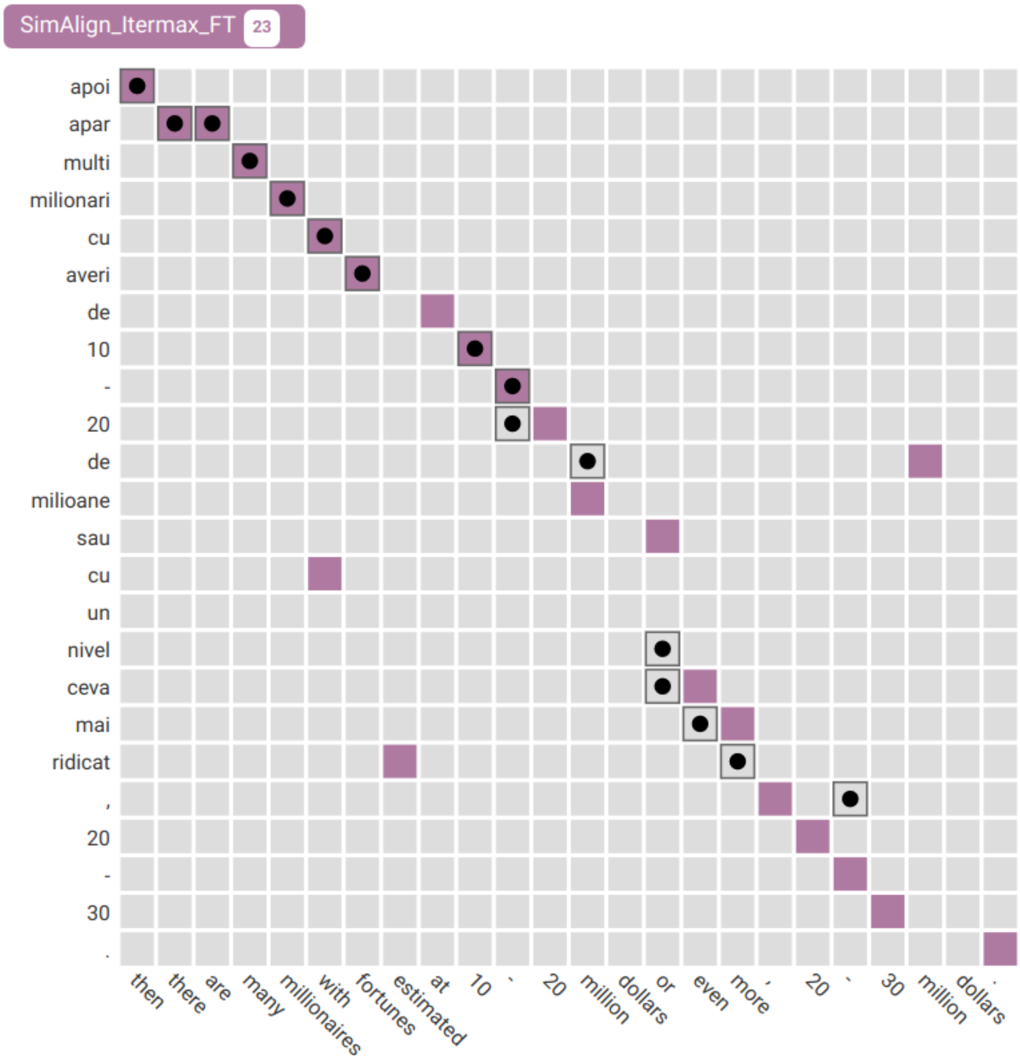

Figure 11: Sentence *101* in the RO-EN dataset; An example of incorrect/incomplete annotation of the gold standard sentence. The Romanian word *mollioane* is translated to *million* but the gold standard aligns the word *de* to *million* instead. Moreover, the sentence is not entirely aligned.

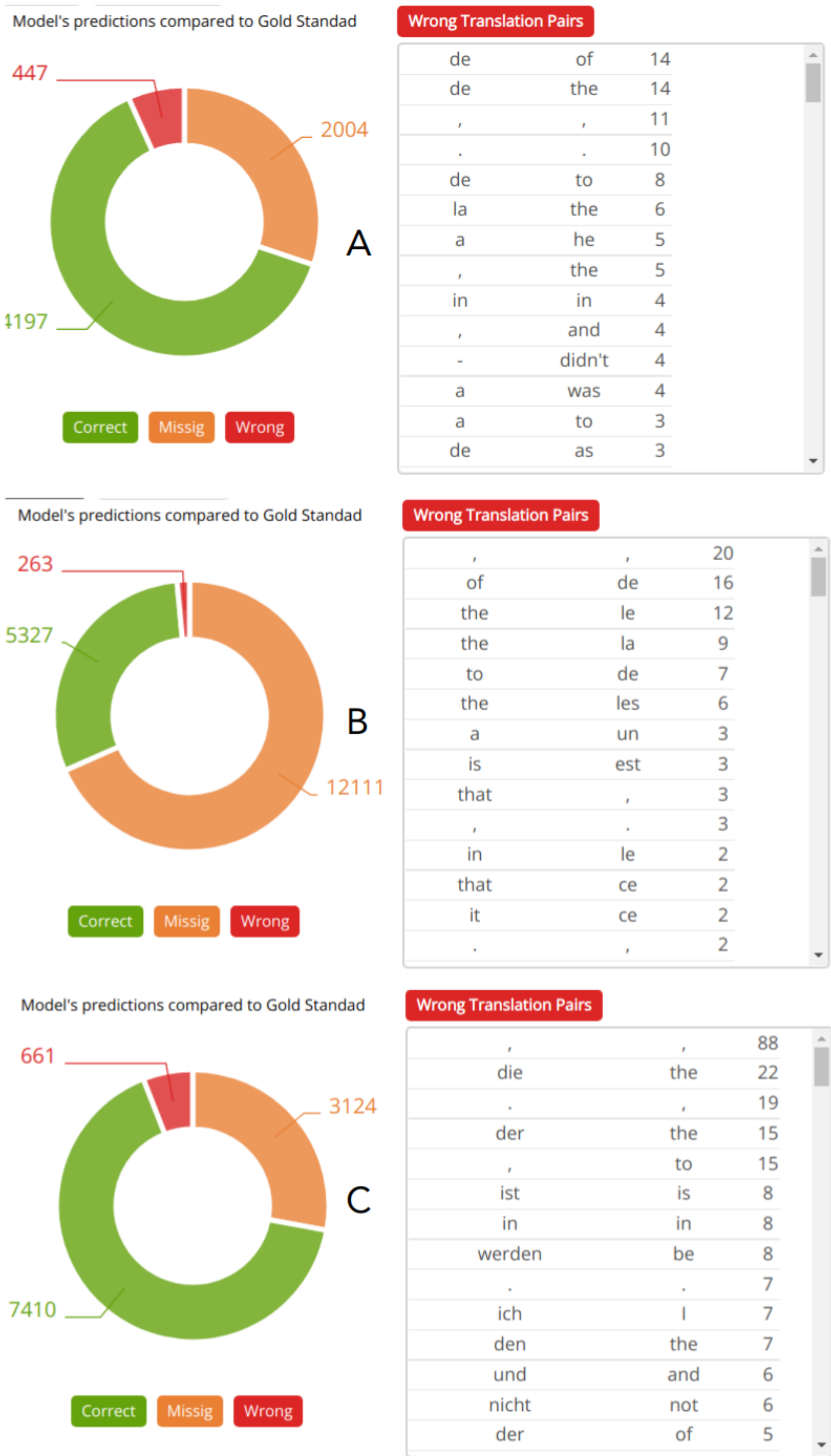

Figure 12: Frequent Alignment Errors, A) The alignment produced by *XLMAlign_Argmax* on RO-EN dataset. B) The alignment produced by *XLMAlign_Argmax* on EN-FR dataset. C) The alignment produced by XLMAlign_Argmax on DE-EN dataset.

AER | Precision | Recall | F1 | # Translation Pairs

Models/Datasets	EN-FR	EN-FR-100 ↑	EN-PT-100	ES-FR-100	PT-ES-100	PT-FR-100	DE-EN	EN-ES-100	RO-EN
Softmax_FT_mBERT	0.0407	0.133	0.1027	0.1403	0.0589	0.1564	0.1524	0.0958	0.226
Entmax_FT_mBERT	0.0392	0.1354	0.1013	0.1446	0.0611	0.1593	0.1539	0.0954	0.2293
Argmax_FT_mBERT	0.0399	0.1359	0.0998	0.1492	0.0659	0.16	0.1535	0.0975	0.2316
Itermax_FT_mBERT	0.067	0.1362	0.1207	0.1328	0.0739	0.1528	0.1616	0.109	0.221
Softmax_mBERT	0.0514	0.1487	0.1179	0.1478	0.0777	0.1595	0.1828	0.1093	0.2457
Entmax_mBERT	0.0511	0.149	0.1162	0.1533	0.0799	0.1686	0.1889	0.1107	0.2547
Argmax_mBERT	0.0519	0.1579	0.1245	0.1613	0.0797	0.1701	0.1971	0.1219	0.2645
Itermax_mBERT	0.0757	0.1612	0.1397	0.1559	0.0918	0.1747	0.1943	0.1292	0.2413
Argmax_XLMAlign	0.0569	0.165	0.1336	0.1721	0.0873	0.1937	0.1694	0.1165	0.2535
Softmax_XLMAlign	0.0792	0.1697	0.146	0.1637	0.0937	0.1944	0.1846	0.1272	0.2527
Entmax_XLMAlign	0.0736	0.1714	0.1407	0.172	0.0949	0.1974	0.1879	0.1221	0.2664
Itermax_XLMAlign	0.0903	0.1753	0.1631	0.1693	0.1109	0.2113	0.1845	0.1375	0.2473
Match_FT_mBERT	0.097	0.1774	0.1502	0.1831	0.1265	0.2187	0.203	0.1372	0.2493
Match_mBERT	0.1132	0.1837	0.1827	0.1887	0.1304	0.2233	0.2264	0.1499	0.2635
Argmax_XLMR	0.0655	0.1924	0.1324	0.1768	0.0743	0.2022	0.1923	0.1222	0.2591
Itermax_XLMR	0.0902	0.1951	0.1564	0.1784	0.1018	0.2204	0.1986	0.1418	0.2537
Softmax_XLMR	0.0918	0.1953	0.157	0.1771	0.0891	0.2162	0.2306	0.1406	0.2864
Entmax_XLMR	0.0901	0.2115	0.1693	0.1923	0.0925	0.2248	0.245	0.1463	0.302
Match_XLMAlign	0.1317	0.2122	0.1907	0.1976	0.1378	0.2523	0.2262	0.1711	0.2753
Match_XLMR	0.1357	0.2192	0.1864	0.2016	0.1329	0.2541	0.2369	0.1704	0.2781

Figure 13: Aggregated table view allows to compare the quantitative metrics of all models on all datasets.

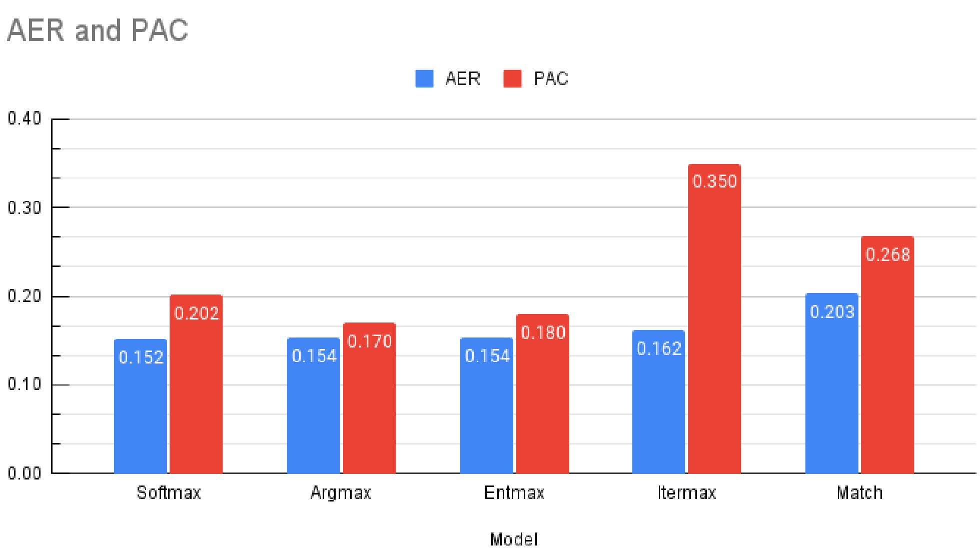

Figure 14: Comparison among five alignment models on the German-English dataset regarding AER and PAC.

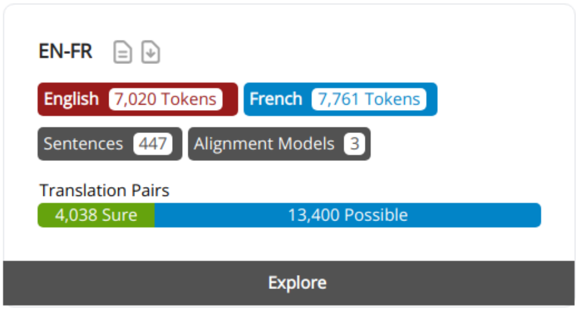

Figure 15: A dataset card contains all related information such as languages, number of tokens, sentences, Sure, and Possible pairs

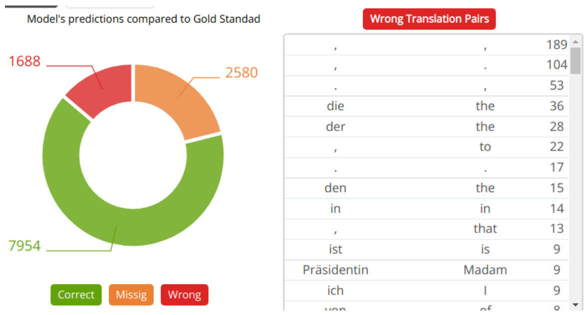

Figure 18: Frequent alignment errors produced $Itermax$ model and fine-tuned mBert model on the DE-EN dataset.

Figure 16: Table View shows the agreement of two models predictions at sentence level. Sentence 1 form DE-EN dataset

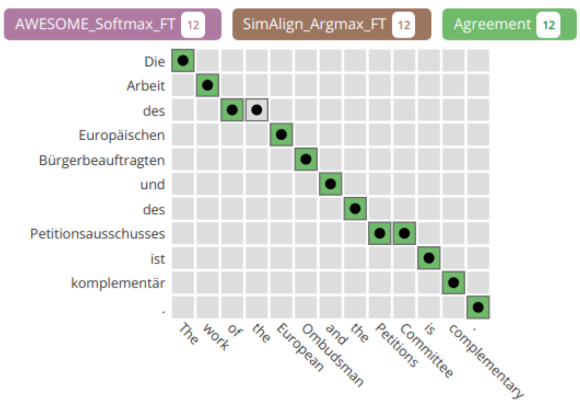

Figure 17: Sentence 46 from DE-EN, comparing two models at sentence level.

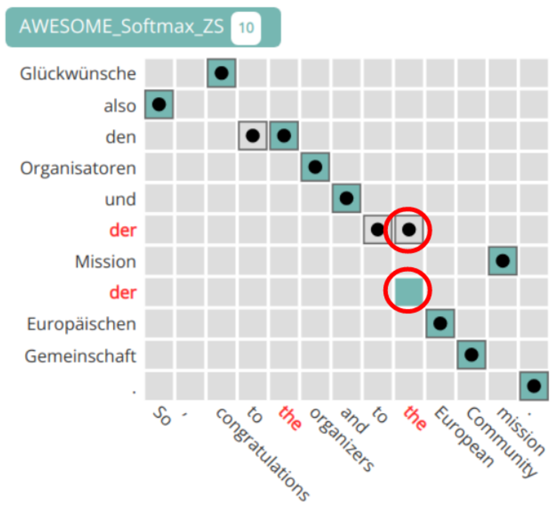

Figure 19: Sentence 127 from DE-EN, incorrect alignment of repeated tokens