SilverAlign: MT-Based Silver Data Algorithm for Evaluating Word Alignment

Abdullatif Köksal[1,2], Silvia Severini[3*], Hinrich Schütze[1,2]

[1]Center for Information and Language Processing (CIS), LMU Munich, Germany
[2]Munich Center for Machine Learning (MCML), Germany
[3]Leonardo Labs, Italy

`akoksal@cis.lmu.de`

Abstract

Word alignments are essential for a variety of NLP tasks. Therefore, choosing the best approaches for their creation is crucial. However, the scarce availability of gold evaluation data makes the choice difficult. We propose SilverAlign, a new method to automatically create silver data for the evaluation of word aligners by exploiting machine translation and minimal pairs. We show that performance on our silver data correlates well with gold benchmarks for 9 language pairs, making our approach a valid resource for evaluation of different languages and domains when gold data is not available. This addresses the important scenario of missing gold data alignments for low-resource languages.

Keywords: word alignment, synthetic data, machine translation

1. Introduction

Word alignments (WA) are crucial for statistical machine translation where they constitute the basis for creating probabilistic translation dictionaries. They are relevant to different tasks such as neural machine translation (NMT) (Alkhouli et al., 2018), typological analysis (Östling, 2015), annotation projection (Huck et al., 2019), bilingual lexicon extraction (Smadja et al., 1996; Ribeiro et al., 2001), and for creating multilingual embeddings (Dufter et al., 2018). Different approaches have been investigated using statistics like IBM models (Brown, 1993) and Giza++ (Och and Ney, 2003). More recently, Östling and Tiedemann (2016) introduced Eflomal, a high-quality word aligner widely used nowadays for its ability to align many languages effectively. Other methods create alignments from attention matrices of NMT models (Zenkel et al., 2019), solve multitask problems (Garg et al., 2019), or leverage multilingual word embeddings (Sabet et al., 2020).

Given the variety of approaches available for aligning words, the choice of the best alignment methods for a certain parallel corpus has gained attention. Such decision requires evaluation data for the pair of languages and the specific domain addressed. Collecting gold data or high-quality word alignment benchmarks requires the work of various annotators as for the Blinker project of Melamed (1998) and WA shared tasks (Mihalcea and Pedersen, 2003; Martin et al., 2005) which can be a time-consuming or an impractical job for lesser spoken languages. Melamed (1998) reports that annotating word alignments for 100 sentences in English-French would take 9 to 22 hours. Addi-

English	Blissymbols
The *dog* plays alone	
The *orchestra* plays alone	
The *cat* plays alone	
The *child* plays alone	

Figure 1: An example of our technique with minimal pairs for a source sentence in English and the target language, Blissymbols.[1] For a word (dog) in the source sentence, we create minimal pairs (orchestra, cat, child), and then we can align the word dog to the correct symbol in Blissymbols with the help of translations.

tionally, the annotation process often leads to conflicts among annotators (Macken, 2010). Hence, gold data is scarce or completely unavailable for many languages, especially low-resource ones and, when dealing with domain-specific data such as medical or legal text, such availability is even less. Therefore analyzing existing word alignment models with a varying number of language pairs in different domains is a challenging task.

We propose SilverAlign, a novel algorithm to cre-

---

*Work was done while at CIS.

[1]Blissymbols is a constructed language established in 1949 to help people with communication difficulties. The `blissonline.se` dictionary is used for the examples.

ate silver evaluation data for guiding the choice of appropriate word alignment methods. Our approach is based on a machine translation model and exploits minimal sentence pairs to create parallel corpora with alignment links. Figure 1 illustrates our core idea with minimal pairs in English and Blissymbols. Our approach is to create alternative sentences in minimal pairs, rely on machine translation models to track changed words for each alternative, and finally align words in the source sentence.

In summary, our contributions are:

1. We find that our silver benchmarks rank methods with high consistency compared to rankings based on gold data. This means that we can identify the best setup based on silver data if there is no gold data available, which is frequently the case in low-resource scenarios for word alignment.

2. We conduct an extensive analysis of our silver resource with respect to gold data for 9 language pairs from different language families and resource availability. We perform various experiments for word alignment models on sub-word tokenization, tokenizer vocabulary size, varying performance of Part-of-Speech tags, and word frequencies.

3. SilverAlign can automatically create large evaluation benchmarks so it supports a more accurate evaluation and a more in-depth analysis than small gold sets (i.e., English-Hindi has only 90 sentences). Also, SilverAlign is robust to domain changes since it shows a high correlation between gold and both in- and out-of-domain silver benchmarks.

4. It has been shown that machine translation performance (including NMT performance) can be improved by choosing a tokenization that optimizes compatibility between source and target languages (Deguchi et al., 2020). We show that SilverAlign can be used to find such a compatible tokenization for each language pair by looking at performance on silver word alignment data.

5. We make our silver data and code available as a resource for future work that takes advantage of our silver evaluation datasets.[2] Our code can be used to create silver benchmarks for multiple languages, and our silver benchmark can be used out-of-the box.

The rest of the paper is organized as follows. Section 2 describes related work. The details of

SilverAlign method are explained in Section 3. Section 4 describes the experimental setup, evaluation metrics and datasets. We compare the results of our silver benchmarks to gold data in Section 5. Finally, we draw conclusions and discuss future work in Section 6.

2. Related Work

2.1. Word alignment analysis

The analysis of word alignment performance with respect to different factors has been analyzed by many works. Ho and Yvon (2019) compare discrete and neural word aligners performance with respect to unaligned words, rare words, Part-of-Speech (PoS) tags, and distortion errors. Asgari et al. (2020) study word alignment results when using subword-level tokenization and show improved performance with respect to word level. Sabet et al. (2020) analyze the performance of word aligners regarding different PoS for English/German and show that Eflomal has low performance when aligning links with high distortion. They also analyze the alignments based on word frequency and show that the performance decreases for rare words when aligning at the word level versus the subword level.

Ho and Yvon (2021) analyze the interaction between alignment methods and subword tokenization (Unigram and Byte Pair Encoding (BPE)). They observe that tokenizing into smaller units helps to align rare and unknown words. They also investigate the effect of different vocabulary sizes and conclude that word-based segmentation is less optimal. We also conduct an experiment in this direction in Section 5.4.

2.2. Silver data creation in NLP

Collecting gold data for evaluating or training systems can be impractical due to its cost and the need for human annotators. To solve these issues, silver data - data generated automatically - has been widely exploited for different tasks and domains. For the Named Entity Recognition (NER) task, Rebholz-Schuhmann et al. (2010) introduce CALBC, a silver standard corpus generated by the harmonization of multiple annotations, Wu et al. (2021) create training data for their NER model through word-to-word machine translation and annotation projection, and Severini et al. (2022) create named entities pairs from co-occurrence statistics and transliteration models. For the medical domain, there exist multiple silver sets due to the difficulty of finding qualified annotators. Examples are the silver corpus of Rashed et al. (2020) for training and evaluating COVID-19-related NLP tools, and DisTEMIST from Miranda-Escalada et al.
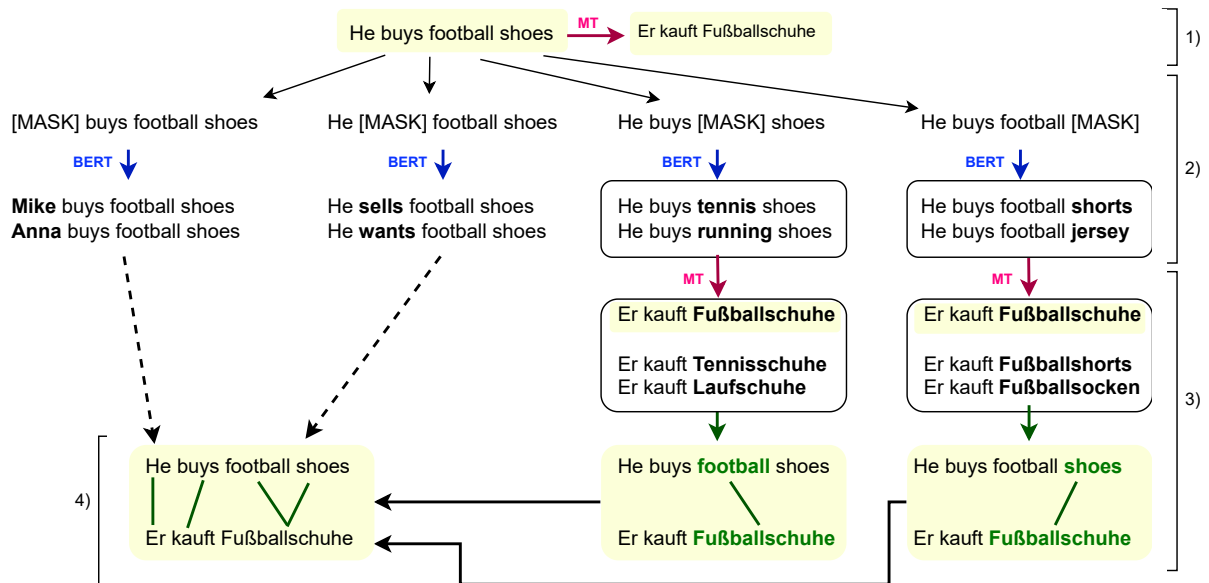
Figure 2: One-to-one and many-to-one examples of English-German alignments according to SilverAlign. The first step translates a given source sentence (s_i) to a sentence (t_i) in the target language via machine translation. The second step finds alternatives for each word in s_i via a foundation model trained with the masked language modeling (MLM) objective. The third step translates all alternative sentences of a given word and tracks which words are changed in t_i. If alternative translations change the same word in t_i, we label the corresponding alignments and merge the affected words in the fourth step. (Details for the first two columns omitted for clarity).

(2022), a multilingual dataset for 6 languages created through annotation transfer and MT for automatic detection and normalization of disease mentions from clinical case documents. Paulheim (2013) introduced DBpedia-NYD for evaluating the semantic relatedness of resources in DBpedia and exploiting web search engines. Baig et al. (2021) propose a silver-standard dependency treebank of Urdu tweets using self-training and co-training to automatically parse big amounts of data. Wang et al. (2022) synthesize labeled data using lexicons to adapt pretrained multilingual models to low-resource languages.

3. Method

The pipeline of our silver data creation algorithm is illustrated in Figure 2. Given a source language S and a target language T, we now describe the steps to create our word alignment silver data for S-T:

1. We first collect monolingual data from the source language, D_S. Given a sentence $s_i = w_1^s, w_2^s, ..., w_N^s \in D_S$ of length N, we use a machine translation system to generate the target sentence $t_i = w_1^t, w_2^t, ..., w_M^t$, and therefore target data D_T.

2. Then, we create minimal pairs for s_i by finding alternative words for each w_j^s in the sentence ($j \in [1, N]$). We use a pretrained Masked Language Model (i.e., English BERT$_{\text{Large}}$) to find alternative words which fit into the context well. For each s_i, we create five alternatives per word by masking one word at a time. Examples of minimal pairs for the sentence "I love pizza" are "You love pizza", "I hate pizza", and "I love apples".

3. In the third step, we use a machine translation system to translate all the alternative sentences to the target language.[3] Based on the changed words in the alternative sentences, we align the words in s_i to the words in t_i. For example, given a sentence of length 4 as in Figure 2 and a masked word in position 1, if the alternatives present the same w_2^t (*kauft*) and w_3^t (*Fußballschuhe*), but words different from w_1^t (*Mike, Anna*), then w_1^t (*Er*) is aligned to the masked word, w_1^s (*He*).

The method works under the following three assumptions. First, each alternative of a word

---

[3]As recent machine translation models (Costa-jussà et al., 2022) support the pairwise translation of more than 200 languages, our method is applicable to several hundred languages and thousands of language pairs.

Language	Gold data	Size	Parallel data	Size
ENG-CES	Mareček (2008)	2501	EuroParl (Koehn, 2005)	648K
ENG-DEU	EuroParl-based[a]	508	EuroParl (Koehn, 2005)	1907K
ENG-FAS	Tavakoli and Faili (2014)	400	TEP (Pilevar et al., 2011)	595K
ENG-FRA	WPT2003, Och and Ney (2000)	447	Hansards (Germann, 2001)	1123K
ENG-HIN	WPT2005[b]	90	Emille (McEnery et al., 2000)	3K
ENG-RON	WPT2005[b]	199	Constitution, Newspaper [b]	39K
ENG-TUR	PBC-based (Our)	100	PBC (Mayer and Cysouw, 2014)	30K
FIN-ELL	Yli-Jyrä et al. (2020)	7,909	PBC (Mayer and Cysouw, 2014)	8K
FIN-HEB	Yli-Jyrä et al. (2020)	22,291	PBC (Mayer and Cysouw, 2014)	22K

[a] www-i6.informatik.rwth-aachen.de/goldAlignment/
[b] http://web.eecs.umich.edu/~mihalcea/wpt05/

Table 1: Overview of gold and parallel datasets. "Size" refers to the number of sentences. Language pairs are represented with their respective ISO 639-3 codes.

Language	Gold		Silver$_{Small}$		Silver$_{Large}$	
	Size	$|A|$	Size	$|A|$	Size	$|A|$
ENG-CES	2,501	67K	1,507	3,852	26K	57K
ENG-DEU	508	11K	227	480	31K	77K
ENG-FAS	400	12K	137	242	16K	27K
ENG-FRA	447	17K	216	359	32K	74K
ENG-HIN	90	1,409	46	87	26K	58K
ENG-RON	199	5,034	69	161	28K	64K
ENG-TUR	100	2,670	50	80	27K	60K
FIN-ELL	7909	161K	1,668	2,230	-	-
FIN-HEB	22,291	405K	4,522	6,396	-	-

Table 2: Overview of gold and silver dataset sizes. Size refers to the number of sentences and $|A|$ is the total number of alignments.

is valid as long as the length of the respective translated sentence is equal to the target sentence's length (i.e. M). Second, at least 4 out of 5 alternative words should be valid. Third, all the valid alternatives should change the same word and only one word in the target language, consistently. Even though these constraints mean that our "yield" is low (many words are not aligned because they do not satisfy the constraints), they are necessary to get confident word alignments. As collecting monolingual data is easy, we overcome this size problem by using a larger monolingual corpus (see 5.4).

4. Finally, we merge the alignment links for each $w_j \in s_i$ to create a set of silver alignments for the sentence pair s_i-t_i.

4. Experimental Setup

Table 1 shows the gold and parallel data sources and statistics. The chosen set of languages rep-resents diverse language families, scripts, and re-source availability. In order to include a challenging benchmark containing a dissimilar language pair, we propose a new gold dataset for English-Turkish (a language with poor morphology and a highly agglutinative language with complex morphology). We first collect 100 random English sentences from the Parallel Bible Corpus of Mayer and Cysouw (2014) and translate them to Turkish with Google Translate. Then, gold word alignments are anno-tated by a native speaker annotator who is also in charge of fixing any translation issues.

Yli-Jyrä et al. (2020) present a gold dataset for Finnish and Ancient Greek. However, we were not able to find a machine translation model from Finnish to Ancient Greek so we compare the Finnish - Modern Greek silver data to the Finnish - Ancient Greek gold data.

In all our experiments, we use Google Translate as our machine translation model. For Hebrew, the model produces words without vowels since this language is standardly written without them. However, this is not the case for (Yli-Jyrä et al., 2020)'s gold dataset, so we pre-process the latter by removing short vowels. We use Eflomal (Östling and Tiedemann, 2016) with the grow-diag-and-final (GDFA) symmetrization method as a word aligner to compare different configurations such as tok-enizers.

To create alternative sentences, we use En-glish BERT$_{Large}$ and Finnish BERT (Virtanen et al., 2019).

4.1. Evaluation

We compare word alignment for different setups with gold data and silver data to show the correla-tions between the latter. For word alignment, we
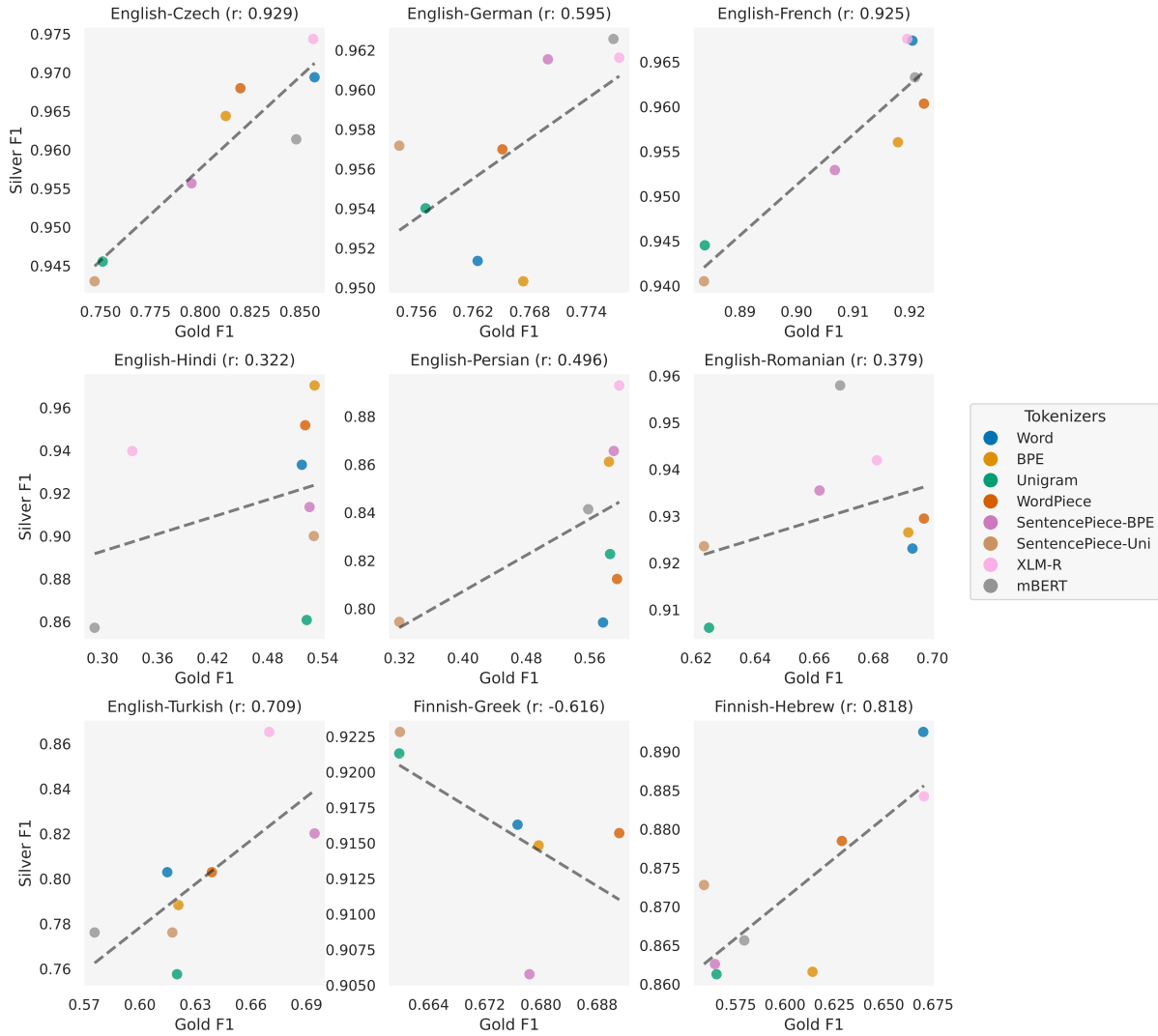
Figure 3: Comparison of tokenizer performance on Gold and Silver$_{\text{Small}}$. Pearson's r correlation coefficient of F_1 scores is reported in the title of each language pair subfigure. The subfigures demonstrate that there is a strong correlation between Silver and Gold when different tokenizers are ranked. Thus, we can identify the best performing tokenizer based on Silver data if Gold data are not available.

report F_1 scores.[4] For a given set of word alignment prediction edges A, a set of sure alignments S, and a set of possible gold (or silver) alignments P, we calculate the F_1 as:

$$F_1 = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}},$$

$$\text{where} \quad \text{precision} = \frac{|A \cap P|}{|A|}, \quad \text{recall} = \frac{|A \cap S|}{|S|}$$

Silver alignments are not complete, so we perform a partial evaluation: we only evaluate with a

subset of predicted edges A which include alignments (s_i, t_j) where s_i is aligned to at least one word in the silver data. Thus, we expect higher F_1 scores for the silver datasets. Our goal in using silver datasets is not an accurate assessment of absolute F_1, but rather an assessment of relative performance, e.g., which tokenization method is better.

For the verification of the silver data quality, we evaluate different sets of settings (e.g. tokenizers) and compare the correlation of scores for gold and silver data for each language pair. We report Pearson correlation coefficient (Pearson's r) (Freedman et al., 2007).

---

[4]Alignment Error Rate (AER) and F_1 scores were calculated for all experiments. Since they led to similar conclusion, we only report F_1 for clarity.

5. Results

We conduct multiple experiments and evaluations to compare the performance of our silver sets and to analyze different tokenizers.

5.1. Silver Dataset

We use the gold and parallel data in Table 1 for our experiments. We apply SilverAlign for creating two silver sets for each language: $Silver_{Small}$ and $Silver_{Large}$. $Silver_{Small}$ uses the source sentences in the gold data (English or Finnish). As our method selects safe alignments with more than three alternatives, it may produce a low number of total alignments and may not generate any alignments for some sentences as shown in Table 2.

However, in the next sections, we will show that $Silver_{Small}$ demonstrates a strong correlation with the gold data even if the relative number of alignments is small.

We introduce $Silver_{Large}$ to better illustrate our findings, especially when the gold sets contain less than 200 sentences. It is created by applying SilverAlign to 50,000 random English sentences from the C4 real news corpus (Raffel et al., 2020). In section 5.5, we show that when the number of monolingual sentences is bigger, our method can find diverse alignment links in terms of frequency and PoS tags. We also generally observe a strong correlation between $Silver_{Large}$ and Gold even though they are sampled from different domains highlighting the wide applicability of SilverAlign. In all the experiments that include $Silver_{Small}$ and $Silver_{Large}$, we use the same parallel data for word alignment training, shown in Table 1.

5.2. Subword Tokenization

In our first experiment, we analyze the effects of subword tokenization. For a given language pair, we train a tokenizer with a shared vocabulary size 50,000 for source and target languages. We compare Byte-Pair Encoding (BPE) (Gage, 1994), Unigram (Kudo, 2018), WordPiece (Schuster and Nakajima, 2012), SentencePiece (Kudo and Richardson, 2018) with Unigram, and SentencePiece with BPE tokenizers. We also include word-level tokenization and the tokenizer of mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020). We use HuggingFace's tokenizer library for the implementation.[5] Please note that the mBERT tokenizer contains an additional splitting mechanism (i.e., hyphen (-) and apostrophes (')) on top of whitespace splitting and the XLM-R tokenizer includes tokens with only whitespace.

---

[5] https://github.com/huggingface/tokenizers
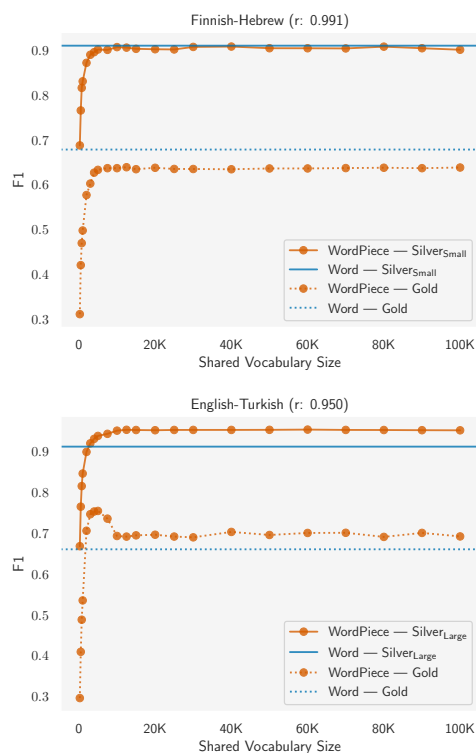
Figure 4: Comparison between word and Word-Piece tokenizers with varying vocabulary sizes. The figure shows that there is a strong correlation (Pearson's r, reported in the title of each subfigure) between F_1 scores of Silver and Gold when word and WordPiece tokenizers with varying number of vocabulary sizes are compared. Therefore, Silver can find the best vocabulary size for a given language pair with a high correlation with Gold.

We consider this when mapping tokens back to words because token ids do not match to index of a character sequence in a tokenized text, split from whitespace. Also, we do not include mBERT and XLM-R tokenizers for Finnish-Greek. The reason is that they support modern Greek but not Ancient Greek, and its use would cause a significant mismatch because of many unknown tokens (41% vs 0% [UNK] ratio) in mBERT and because of aggressive word splitting (3.12 vs. 1.64 token to word ratio) in XLM-R for Ancient Greek compared to modern Greek.

Figure 3 shows tokenizer performance evaluated on Gold and $Silver_{Small}$. Tokenizer rankings are strongly correlated for gold and silver, with a 0.86 average r score (except English-Hindi, probably due to the very low amount of Gold data with 90 sentences). We also observe better correlation in language pairs with high-quality gold data. Gold in English-Czech, English-German, and English-French include possible alignments to take ambiguity of the word alignment task into account (Matusov et al., 2004). Gold sets in Finnish-X
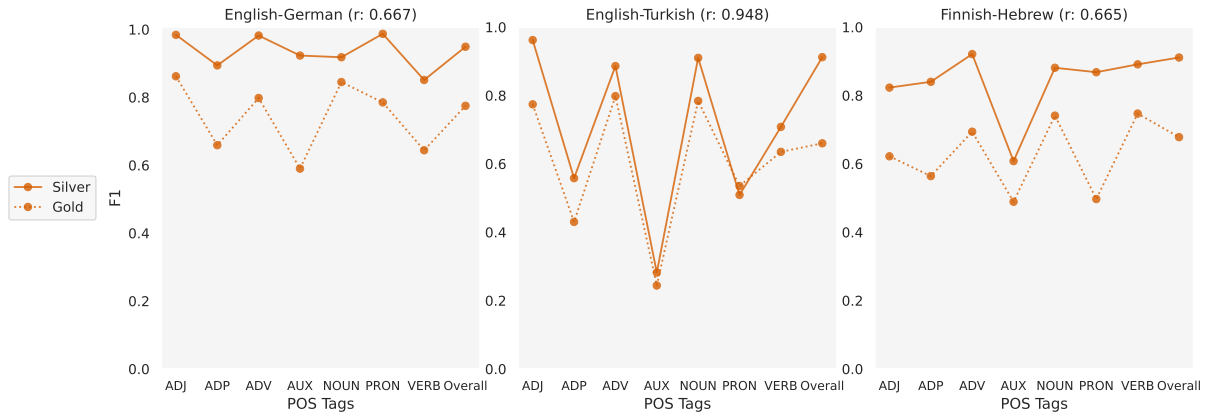
Figure 5: Analysis of word alignment performance with word-level tokenization for different part-of-speech tags. The title for each subfigure gives correlation (Pearson's r) of Silver and Gold F_1 across the different part-of-speech tags. This shows that Silver is able to capture the relative performance of a word alignment method in PoS tags, similarly to Gold.
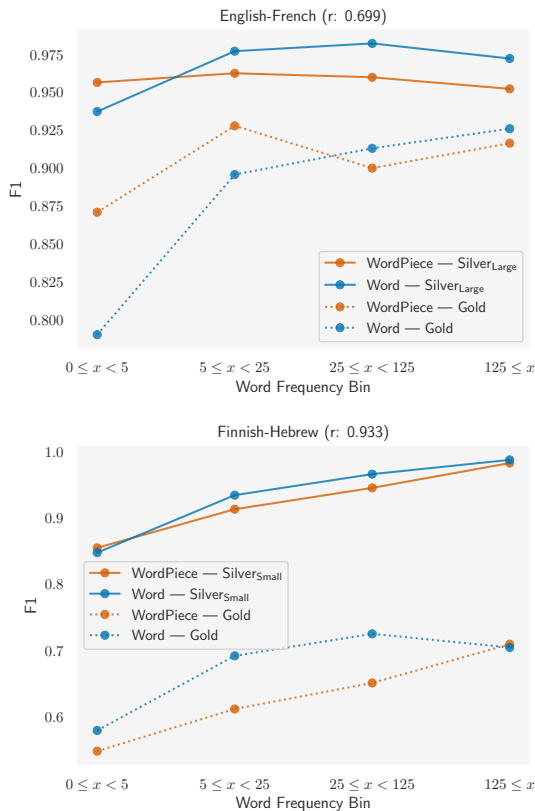

Figure 6: Comparison of the word and WordPiece tokenization with respect to F_1 scores of different word frequency bins on Silver and Gold. This figure demonstrates that similar performance's patterns on different frequencies and tokenizers can be observed between Silver and Gold.

tackle this problem by introducing large-scale gold datasets. We observe that Silver_Small has 0.96 average r score with Gold in these five language

pairs. This shows that SilverAlign automatically creates a benchmark that strongly correlates with high-quality gold datasets.

5.3. Word Frequency

Unigram and SentencePiece-Unigram have the worst performance among all tokenizers when compared on all language pairs. The gold data of English-Hindi are an exception; however, the silver data of English-Hindi align well with these pattern across all languages.

We find that word tokenization is already strong for many languages such as English-Czech, English-German, and English-French. However, we see that subword level tokenization outperforms word tokenization for English-Turkish and Finnish-Greek. In all experiments, XLM-R and mBERT tokenizers achieve comparable results to custom subword tokenization and word tokenization.

Comparisons of Gold with Silver_Large are in the Appendix, Figure 7. We find similar performance as Silver_Small when we compare gold with Silver_Large even though the larger sets come from C4 real news corpus, a domain different from the domain of Gold (i.e., there is a domain adaptation aspect to this evaluation).

In the rest of the results section, we use Silver_Large for English-X and Silver_Small for Finnish-X pairs to perform a more granular analysis.[6]

5.4. Tokenizer Vocabulary Size

As the tokenizer's vocabulary size is an important aspect in word alignment (Ho and Yvon, 2021),

---

[6]Note that there is no Silver_Large for Finnish-X since the small sets already contain more than 7K sentences. That is, for Finnish-X the "small" sets are in reality large.

we experiment with different settings with shared vocabulary size. We use a shared WordPiece tokenizer and illustrate changes in F_1 score at varying vocabulary sizes. Results are depicted in Figure 4. For both silver and gold, there's a strong correlation between word and WordPiece tokenization and vocabulary size change in WordPiece with an average 0.96 r score in all language pairs. We observe that WordPiece performs better than word tokenization for English-Turkish and vice versa for Finnish-Hebrew. F_1 scores converge after around 10K vocabulary size for both language and dataset. There is a slight bump in Gold for English-Turkish around 4K vocabulary size. Since we do not observe a similar pattern in any language pairs in the gold data, (see full plots in Figure 7) we conjecture that the bump is due to the small amount of Gold data in Turkish.

5.5.   Part-of-Speech Tagging Performance

Previous works (Sabet et al., 2020; Ho and Yvon, 2019) show that word aligners' performance can significantly vary with respect to different parts of speech. Therefore, we investigate this aspect of our algorithm. We use the Stanza (Qi et al., 2020) toolkit to tag the source sentences in English and Finnish. We compare Gold and Silver performance for different PoS tags (Silver_Large for English-X and Silver_Small for Finnish-X).

Figure 5 shows similar patterns and high correlation between gold and silver. For example, auxiliaries and adpositions perform significantly worse than other PoS tags in English-Turkish. This is because Turkish is an agglutinative language in which adpositions are usually case marked in noun forms, and auxiliaries are represented in suffixes in verb complexes. Also, word ordering on the English side is not monotonically aligned with the morpheme order of the Turkish counterpart of English adpositions and auxiliaries (El-Kahlout and Oflazer, 2009) which makes accurately aligning words with those tags more difficult.

This experiment also illustrates that a silver dataset created with SilverAlign contains a diverse set of word alignments to infer additional information about linguistic properties. We present PoS distribution with respect to gold and silver for all language pairs in the Appendix, Figure 9.

We compare word and WordPiece tokenizations for different word frequency bins in Figure 6. The frequency of a word is defined as the minimum frequency of source and target words for a predicted alignment. Similar to Sabet et al. (2020), we observe that subword tokenizations, like Word-Piece, perform better than word tokenization on low-frequency words for English-French while we do not observe such major performance difference for Finnish-Hebrew in Figure 6. Even though word

and WordPiece tokenizers perform comparably for these languages, we observe that the impact on low- and high-frequency words might be quite different. Therefore, tokenizers can be selected according to sub-objectives by using Silver data, obviating the need for creating an expensive gold data benchmark.

6.   Conclusion

Since creating a human-annotated word alignment dataset is a challenging task, we propose the SilverAlign method to create a silver benchmark using a Masked Language Model (MLM) and a machine translation model. SilverAlign makes use of MLMs to create minimal pairs with alternatives that fit well into context and find partial alignments based on the changes in the translation of the alternatives via machine translation.

We show that our method can create a high-quality silver benchmark for 9 language pairs including pairs of two non-English languages. We show that the silver benchmark on two different domains (Silver_Small and Silver_Large) can help to compare different configurations and investigate errors with a high correlation to the gold data. We perform experiments on sub-word level tokenization, tokenizer vocabulary size, and performance change with respect to PoS tags and word frequency.

For future work, SilverAlign can be extended to create a specific subset of a general domain dataset to analyze the effects of potential issues in word alignment such as rare words. We believe that SilverAlign can ease up the process of finding issues in existing word alignment models for various language pairs, and it can help to improve both word alignment tools and tasks that use word alignment implicitly or explicitly such as machine translation.

Finally, we believe that our silver data creation algorithm can be helpful for both low- and high-resource language pairs to investigate word alignment without a time-consuming human annotation process. If combined with recent machine translation models (e.g. NLLB (Costa-jussà et al., 2022)), SilverAlign can, in principle, support more than 200 languages. Therefore, we make our silver data and code available as a resource for future work that takes advantage of our silver evaluation datasets.[7]

Limitations

SilverAlign is limited to language pairs with existing machine translation systems and MLMs for the source language. Even though there are recent

---
[7] https://github.com/akoksal/SilverAlign

works and commercial tools that support hundreds of languages for machine translation and MLM, the quality of these systems should be taken into account. For low-quality MT and MLM systems, SilverAlign might require larger monolingual corpora in the source language to create a silver dataset with a good amount of total alignment.

For evaluation purposes, we only evaluate SilverAlign on language pairs for which gold alignments are available. However, our method is applicable to any language pairs for which MT systems and MLMs are available for. Therefore, this includes languages that are more "low-resource" with respect to the one addressed in this paper.

Acknowledgements

7. Bibliographical 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.

Ehsaneddin Asgari, Masoud Jalili Sabet, Philipp Dufter, Christopher Ringlstetter, and Hinrich Schütze. 2020. Subword sampling for low resource word alignment. *arXiv preprint arXiv:2012.11657*.

Amber Baig, Mutee U Rahman, Sirajuddin Qureshi, Saima Tunio, Sehrish Abrejo, and Shadia S Baloch. 2021. Towards silver standard dependency treebank of urdu tweets. *International Journal*, 10(3).

Peter Brown. 1993. Stephen a. and d ella pietra, vincent j., and mercer, robert l. *The mathematics of statistical Machine Translation: Parameter estimation In: C omputational Linguistics*, 19(2):263–311.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Hiroyuki Deguchi, Masao Utiyama, Akihiro Tamura, Takashi Ninomiya, and Eiichiro Sumita. 2020. Bilingual subword segmentation for neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4287–4297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and*

Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Philipp Dufter, Mengjie Zhao, Martin Schmitt, Alexander Fraser, and Hinrich Schütze. 2018. Embedding learning through multilingual concept induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1520–1530, Melbourne, Australia. Association for Computational Linguistics.

Ilknur Durgar El-Kahlout and Kemal Oflazer. 2009. Exploiting morphology and local word reordering in english-to-turkish phrase-based statistical machine translation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1313–1322.

David Freedman, Robert Pisani, and Roger Purves. 2007. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

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.

Anh Khoa Ngo Ho and François Yvon. 2019. Neural baselines for word alignments. In *International Workshop on Spoken Language Translation*.

Anh Khoa Ngo Ho and François Yvon. 2021. Optimizing word alignments with better subword tokenization. In *Proceedings of the 18th Biennial Machine Translation Summit (Volume 1: Research Track)*, pages 256–269.

Matthias Huck, Diana Dutka, and Alexander Fraser. 2019. Cross-lingual annotation projection is effective for neural part-of-speech tagging. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects,* pages 223–233, Ann Arbor, Michigan. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Lieve Macken. 2010. An annotation scheme and gold standard for dutch-english word alignment. In *7th conference on International Language Resources and Evaluation (LREC 2010)*, pages 3369–3374. European Language Resources Association (ELRA).

Joel Martin, Rada Mihalcea, and Ted Pedersen. 2005. Word alignment for languages with scarce resources. In *Proceedings of the ACL workshop on building and using parallel texts*, pages 65–74.

Evgeny Matusov, Richard Zens, and Hermann Ney. 2004. Symmetric word alignments for statistical machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 219–225.

I Dan Melamed. 1998. Manual annotation of translational equivalence: The blinker project. *arXiv preprint cmp-lg/9805005*.

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.

Antonio Miranda-Escalada, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2022. Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*.

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

Robert Östling. 2015. Word order typology through multilingual word alignment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language*

Processing (Volume 2: Short Papers), pages 205–211, Beijing, China. Association for Computational Linguistics.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics*.

Heiko Paulheim. 2013. Dbpedianyd-a silver standard benchmark dataset for semantic relatedness in dbpedia. In *NLP-DBPEDIA@ ISWC*. Citeseer.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Salma Kazemi Rashed, Johan Frid, and Sonja Aits. 2020. English dictionaries, gold and silver standard corpora for biomedical natural language processing related to sars-cov-2 and covid-19. *arXiv preprint arXiv:2003.09865*.

Dietrich Rebholz-Schuhmann, Antonio José Jimeno Yepes, Erik M. van Mulligen, Ning Kang, Jan Kors, David Milward, Peter Corbett, Ekaterina Buyko, Katrin Tomanek, Elena Beisswanger, and Udo Hahn. 2010. The CALBC silver standard corpus for biomedical named entities — a study in harmonizing the contributions from four independent named entity taggers. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

António Ribeiro, Gabriel Lopes, and João Mexia. 2001. Extracting translation equivalents from portuguese-chinese parallel texts. *Journal of Studies in Lexicography*, 11(1):118–194.

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.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Silvia Severini, Ayyoob Imani, Philipp Dufter, and Hinrich Schütze. 2022. Towards a broad coverage named entity resource: A data-efficient approach for many diverse languages. In *Proceedings of the Thirtheenth International Conference on Language Resources and Evaluation*.

Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for finnish. *CoRR*, abs/1912.07076.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. Expanding pretrained models to thousands more languages via lexicon-based adaptation. *arXiv preprint arXiv:2203.09435*.

Qianhui Wu, Zijia Lin, Börje F Karlsson, Biqing Huang, and Jian-Guang Lou. 2021. Unitrans: unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3926–3932.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2019. Adding interpretable attention to neural translation models improves word alignment. *arXiv preprint arXiv:1901.11359*.

8.   Language Resource References

Ulrich Germann. 2001. Aligned hansards of the 36th parliament of canada. *Natural Language Group of the USC Information Sciences Institute*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. *Machine Translation Summit, 2005*, pages 79–86.

David Mareček. 2008. Automatic alignment of tectogrammatical trees from czech-english parallel corpus. In *Master's thesis, Charles University, MFF UK*.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

Anthony McEnery, Paul Baker, Robert Gaizauskas, and Hamish Cunningham. 2000. Emille: Building a corpus of south asian languages. In *Proceedings of the International Conference on Machine Translation and Multilingual Applications in the new Millennium: MT 2000*.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th annual meeting of the association for computational linguistics*, pages 440–447.

Mohammad Taher Pilevar, Heshaam Faili, and Abdol Hamid Pilevar. 2011. Tep: Tehran english-persian parallel corpus. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 68–79. Springer.

Leila Tavakoli and Heshaam and Faili. 2014. Phrase alignments in parallel corpus using bootstrapping approach. *International Journal of Information and Communication Technology Research*, 6(3).

Anssi Yli-Jyrä, Josi Purhonen, Matti Liljeqvist, Arto Antturi, Pekka Nieminen, Kari M. Räntilä, and Valtter Luoto. 2020. HELFI: a Hebrew-Greek-Finnish parallel Bible corpus with cross-lingual morpheme alignment. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4229–4236, Marseille, France. European Language Resources Association.
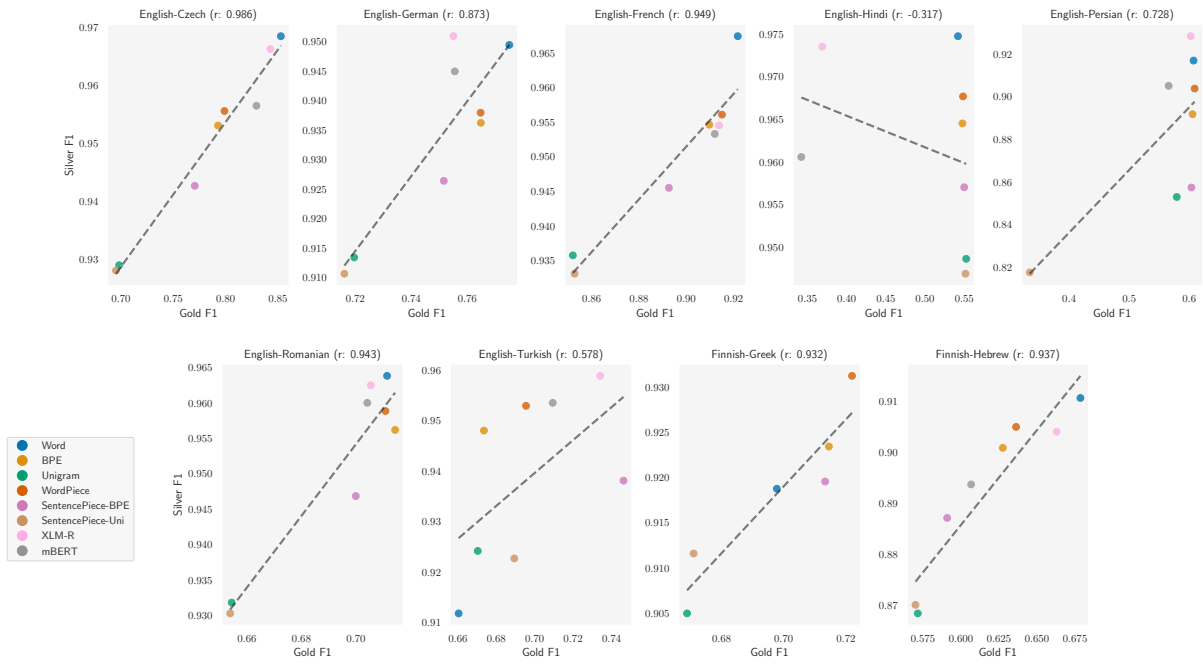
9. Appendix

Figure 7: Comparison of tokenizer performance on Gold and Silver$_{Large}$ (Silver$_{Small}$ for Finnish-X). The Pearson's r of F_1 scores are reported in the title of each language pair subfigure. The figures demonstrate that there's a strong correlation between Silver and Gold when different tokenizers are ranked. Thus, we can identify the best performing methods based on Silver data if Gold data are not available.
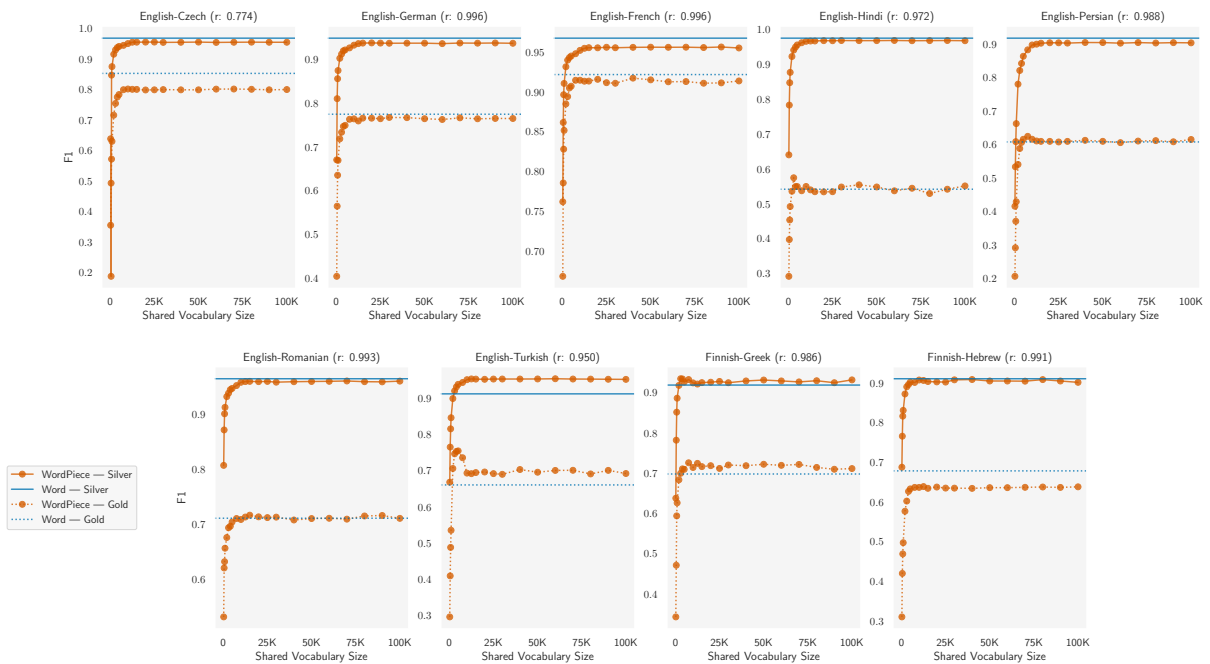

Figure 8: Comparison between word and WordPiece tokenization with varying vocabulary sizes. The figure shows that there's a strong correlation (Pearson's r) between F_1 scores of Silver and Gold when word and WordPiece tokenizer with varying number of vocabulary size are compared, as reported in the title. Therefore, Silver can find the best vocabulary size for a given language pair with a high correlation with Gold.
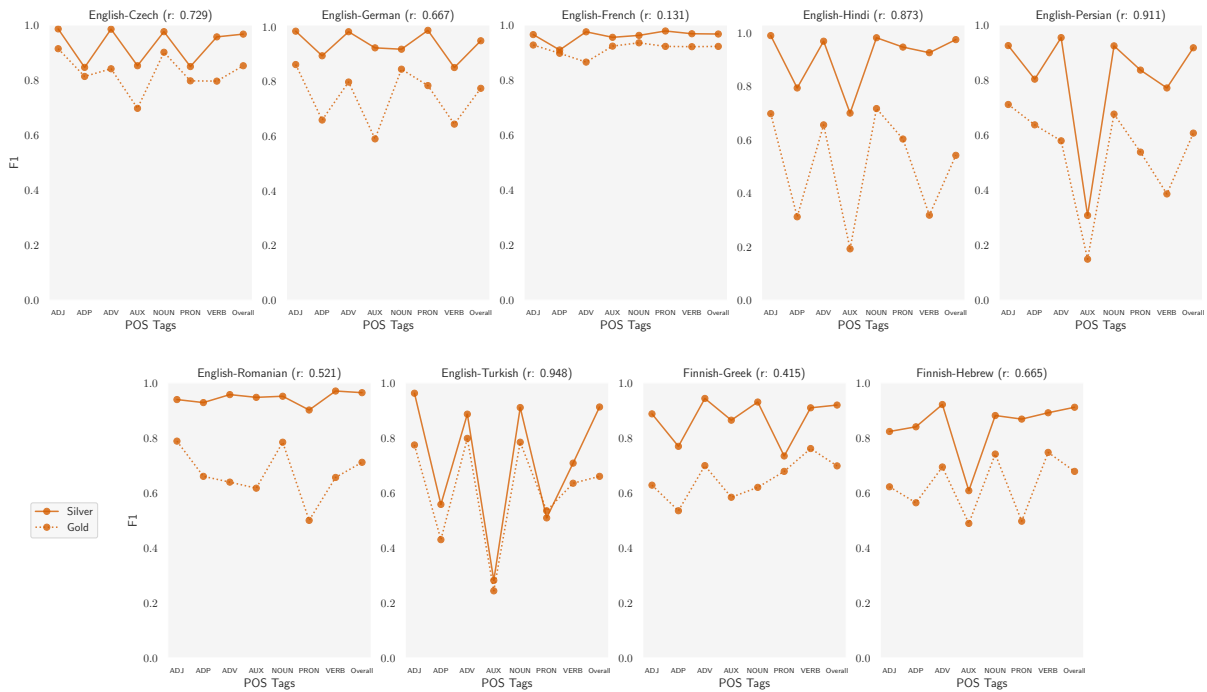
Figure 9: Analyzing word alignment performance with word-level tokenization for different part-of-speech tags. The title for each subfigure gives correlation (Pearson's r) of Silver and Gold F_1 across the different part-of-speech tags. This shows that Silver is able to capture the relative performance of a word alignment method in PoS tags, similarly to Gold.
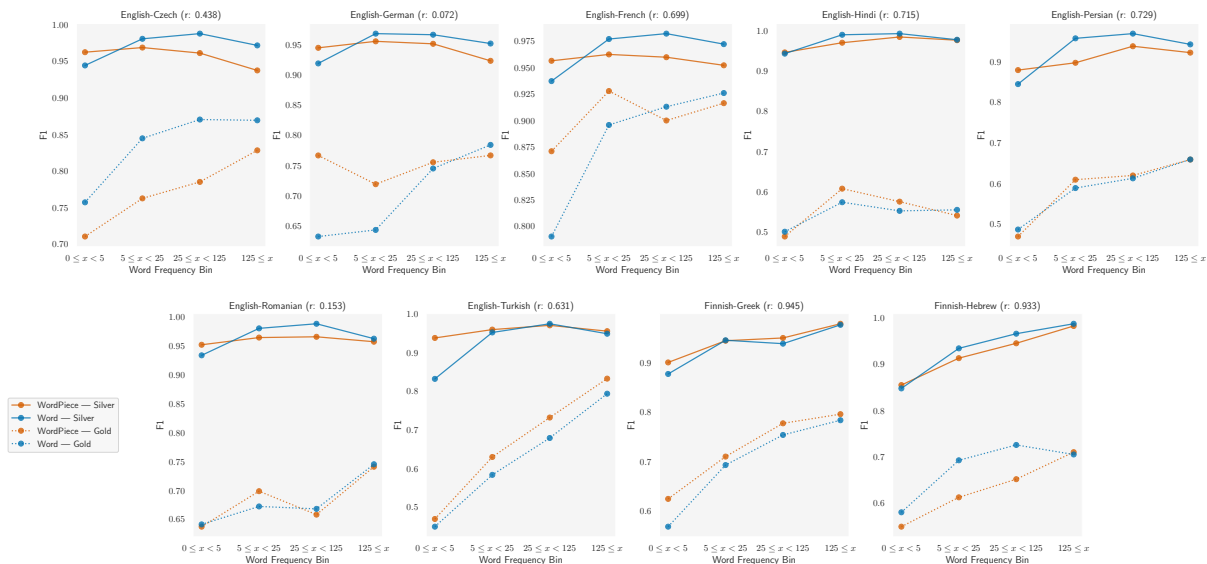

Figure 10: Comparison of the word and WordPiece tokenization with respect to F_1 scores of different word frequency bins on Silver and Gold. This figure demonstrates that similar performance's patterns on different frequencies and tokenizers can be observed between Silver and Gold.