# Graph Algorithms for Multiparallel Word Alignment

**Ayyoob Imani**[*1], **Masoud Jalili Sabet**[*1], **Lütfi Kerem Şenel**[1],
**Philipp Dufter**[1], **François Yvon**[2], **Hinrich Schütze**[1]

[1]Center for Information and Language Processing (CIS), LMU Munich, Germany
[2]Université Paris-Saclay, CNRS, LISN, France
`{ayyoob, masoud, lksenel, philipp}@cis.lmu.de,`
`francois.yvon@limsi.fr`

## Abstract

With the advent of end-to-end deep learning approaches in machine translation, interest in word alignments initially decreased; however, they have again become a focus of research more recently. Alignments are useful for typological research, transferring formatting like markup to translated texts and can be used in the decoding of machine translation systems. At the same time, massively multilingual processing is becoming an important NLP scenario and pretrained language and machine translation models that are truly multilingual are proposed. However, most alignment algorithms rely on bitexts only and do not leverage the fact that many parallel corpora are multiparallel. In this work, we exploit multiparallelity of corpora by representing an initial set of bilingual alignments as a graph and then predicting additional edges in the graph. We present two graph algorithms for edge prediction: one inspired by recommender systems and one based on network link prediction. Our experimental results show absolute improvements of $F_1$ of up to 28% over the baseline bilingual word aligner in different datasets.

## 1 Introduction

Word alignment is a challenging NLP task that plays an essential role in statistical machine translation and is useful for neural machine translation (Alkhouli and Ney, 2017; Alkhouli et al., 2016; Koehn et al., 2003). Other applications of word alignments include bilingual lexicon induction, annotation projection, and typological analysis (Shi et al., 2021; Rasooli et al., 2018; Müller, 2017; Lewis and Xia, 2008). With the advent of deep learning, interest in word alignment initially decreased. However, recently a new wave of publications has again drawn attention to the task (Jalili Sabet et al., 2020; Dou and Neubig, 2021; Marchisio et al., 2021; Wu and Dredze, 2020).
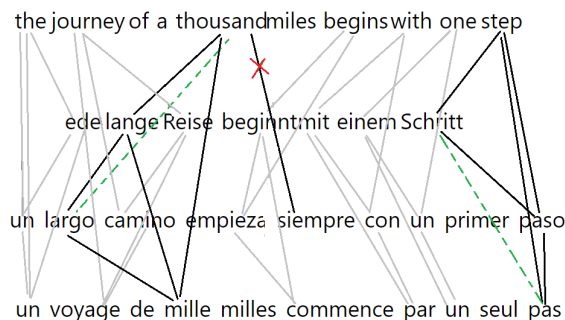


Figure 1: Bilingual alignments of a verse in English, German, Spanish, and French. Two of the alignment edges not found by the bilingual method are German "Schritt" to French "pas" and Spanish "largo" to English "thousand miles". By looking at the structure of the entire graph, one can infer the correctness of these two edges.

In this paper we propose *MPWA* (MultiParallel Word Alignment), a framework that employs graph algorithms to exploit the information latent in a multiparallel corpus to achieve better word alignments than aligning pairs of languages in isolation. Starting from translations of a sentence in multiple languages in a multiparallel corpus, MPWA generates bilingual word alignments for all language pairs using any available bilingual word aligner. MPWA then improves the quality of word alignments for a target language pair by inspecting how they are aligned to other languages. *The central idea is to exploit the graph structure of an initial multiparallel word alignment to improve the alignment for a target language pair.* To this end, MPWA casts the multiparallel word alignment task as a link (or edge) prediction problem. We explore standard algorithms for this purpose: Adamic-Adar and matrix factorization. While these two graph-based algorithms are quite different and are used in different applications, we will show that MPWA effectively leverages them for high-performing word alignment.

---

* Equal contribution - random order.

8457

Link prediction methods are used to predict whether there should be a link between two nodes in a graph. They have various applications like movie recommendations, knowledge graph completion, and metabolic network reconstruction (Zhang and Chen, 2018). We use the Adamic-Adar index (Adamic and Adar, 2003); it is a second-order link prediction algorithm, i.e., it exploits the information of neighbors that are up to two hops aways from the starting target nodes (Zhou et al., 2009). We use a second-order algorithm since a set of aligned words in multiple languages (representing a concept) tends to establish a clique (Dufter et al., 2018). This means that exploring the influence of nodes at a distance of two in the graph provides informative signals while at the same time keeping runtime complexity low.

Matrix factorization is a collaborative filtering algorithm that is most prominently used in recommender systems where it provides users with product recommendations based on their interactions with other products. This method is especially useful if the matrix is sparse (Koren et al., 2009). This is true for our application: Given two translations of a sentence with lengths $M$ and $N$, among all possible alignment links ($M \times N$), only a few ($O(M + N)$) are correct. This is partly due to fertility: words in the source language generally have only a few possible matches in the target language (Zhao and Gildea, 2010).

A multiparallel corpus provides parallel sentences in more than two languages. This type of corpus facilitates the study of multiple languages together, which is especially important for research on low resource languages. As far as we know, out of all available multiparallel corpora, the Parallel Bible Corpus (Mayer and Cysouw, 2014) (PBC) provides the highest language coverage, supporting 1334 different languages, many of which belong to categories 0 and 1 (Joshi et al., 2020) – that is, they are languages for which no language technologies are available and that are severely underresourced.

MPWA has especially strong word alignment improvements for distant language pairs for which existing bilingual word aligners perform poorly. Much work that addresses low resource languages relies on the availabiliy of monolingual corpora. Complementarily, MPWA assumes the existence of a very small (a few 10,000s of sentences in our case) parallel corpus and then takes advantage of information from the other languages in the parallel corpus. This is an alternative approach that is especially important for low resource languages for which monolingual data often are not available.

The PBC corpus does not contain a word alignment gold standard. To conduct the comparative evaluation of our new method, we port three existing word alignment gold standards of Bible translations to PBC, for the language pairs English-French, Finnish-Hebrew and Finnish-Greek. We also create artificial multiparallel datasets for four widely used word alignment datasets using machine translation. We evaluate our method with all seven datasets. Results demonstrate substantial improvements in all scenarios.

Our main contributions are:

1. We propose two graph-based algorithms for link prediction (i.e., the prediction of word alignment edges in the alignment graph), one based on second-order link prediction and one based on recommender systems for improving word alignment in a multiparallel corpus and show that they perform better than established baselines.

2. We port and publish three word alignment gold standards for the Parallel Bible Corpus.

3. We show that our method is also applicable, using machine translation, to scenarios where multiparallel data is not available.

4. We publish our code[1] and data.

## 2 Related Work

**Bilingual Word Aligners** take different approaches. Some are based on statistical analysis, like IBM models (Brown et al., 1993), Giza++ (Och and Ney, 2003a), fast-align (Dyer et al., 2013) and Eflomal (Östling and Tiedemann, 2016). Another more recent group, including SimAlign (Jalili Sabet et al., 2020) and Awesome-align (Dou and Neubig, 2021), utilizes neural language models. The last group is based on neural machine translation (Garg et al., 2019; Zenkel et al., 2020). While neural models outperform statistical models, for cases where only a small parallel dataset is available, statistical models are still superior. In this paper we use PBC, a corpus with 1334 languages, of which only about two hundred are supported by multilingual language models like Bert and XLM-R (Devlin et al., 2019; Conneau et al., 2020). MPWA can

---

[1] https://github.com/cisnlp/graph-align

leverage multiparallelism on top of any bilingual word aligner; in this paper, we use Eflomal and SimAlign.

**Multiparallel corpus alignment.** Most work on word alignment has focused on bilingual corpora. To the best of our knowledge, only one method specifically designed for multiparallel corpora was previously proposed: (Östling, 2014).[2] However this method is outperformed by a "biparallel" method by the same author, Eflomal (Östling and Tiedemann, 2016). We compare with Eflomal in our experiments.

Cohn and Lapata (2007) make use of multiparallel corpora to obtain more reliable translations from small datasets. Kumar et al. (2007) show that multiparallel corpora can be of benefit to reach better performance in phrase-based statistical machine translation (SMT). Filali and Bilmes (2005) present a multilingual SMT-based word alignment model, extending IBM models, based on HMM models and a two step alignment procedure. Since the goal of this research is to tackle word alignment directly without considering machine translation, these works are not considered here.

In another line of research, Lardilleux and Lepage (2008a) introduce a corpus splitting method to come up with a perfect alignment of multiwords. Lardilleux and Lepage (2008b), and Lardilleux and Lepage (2009) suggest to rely only on low frequency terms for a similar purpose: sub-sentential alignment. These methods solve a somewhat different problem than what is addressed by us. Other usages of multiparallel corpora are language comparison (Mayer and Cysouw, 2012), typology studies (Östling, 2015; Asgari and Schütze, 2017; Imani-Googhari et al., 2021) and SMT (Nakov and Ng, 2012; Bertoldi et al., 2008; Dyer et al., 2013)

**Matrix factorization and link prediction.** Matrix factorization is a technique that factors, in the most typical case, a matrix into two lower-ranked matrices in which the latent factors of the original matrix are represented. Matrix factorization approaches have been widely used in document clustering (Xu et al., 2003; Shahnaz et al., 2006), topic modeling (Kuang et al., 2015; Choo et al., 2013) information retrieval (Zamani et al., 2016; Deerwester et al., 1990) and NLP tasks like word sense disambiguation (Schütze, 1998). In 2009, Netflix's recommender system competition revealed that this technique effectively works for collaborative filtering (Koren et al., 2009). Since then it has been a state of the art method in recommender systems.

Link prediction algorithms are widely used in different areas of science since many social, biological, and information systems can be described as networks with nodes and connecting links (Zhou et al., 2009). Link prediction algorithms compute the likelihood of links based on different heuristics. One can categorize available methods based on the maximum number of hops they consider in their computations for each node (Zhang and Chen, 2018). First order algorithms, such as common neighbors (CN), only consider one hop neighborhoods, e.g., (Barabási and Albert, 1999). Second order methods consider two hops, e.g., (Zhou et al., 2009). Finally, higher order methods take the whole network into account for making predictions (Brin and Page, 1998; Jeh and Widom, 2002; Rothe and Schütze, 2014). In this paper, we use a two-hop method since it offers a good tradeoff between effectiveness and efficiency.

## 3 Methods

### 3.1 The MPWA framework

While a bilingual aligner considers each language pair separately, MPWA utilizes the synergy between all language pairs to improve word alignment performance. In Figure 1, Eflomal alignments of a sentence from PBC in four different languages are depicted. Although Eflomal has failed to find the link between German "Schritt" and French "pas", we can easily find this relation by observing that the four nodes "step", "Schritt", "paso", and "pas" are fully connected, except for the edge from "Schritt" to "pas". In this case, the inference amounts to a completion of a clique. However, most cases are not that simple. In the figure, English "thousand miles" is mistakenly aligned to Spanish "siempre" although its alignments to German "lange" and French "mille" are correct. We would like to infer that "thousand miles" should be aligned to "largo", but in this case creating a fully connected subgraph, i.e., a clique (which would include "siempre"), would add many incorrect edges. Given the complexity and error-proneness of initial bilingual alignments, inferring an alignment between two languages from a multiparallel alignment in general is a complex problem.

Starting from a multiparallel corpus, we first generate bilingual alignments for all language pairs.

---

[2] https://github.com/robertostling/eflomal

MPWA then employs a prediction algorithm to find and add new alignment links. In this paper, we focus on two prediction algorithms: non-negative matrix factorization and Adamic-Adar link prediction.

## 3.2 Non-negative matrix factorization

Non-negative matrix factorization (NMF) has been used in many different applications. After discovery of its effectiveness for collaborative recommendation (Koren et al., 2009), it was widely accepted as a standard method for recommender systems.

In a standard recommender system with $m$ users and $n$ items, ratings (a number from 1 to 5) from each user for the items they have seen so far are known. The aim is to predict the ratings the user would give to unseen items and, based on these predictions, recommend new items to the user. As described by (Luo et al., 2014), let $W = [w_{u,i}] \in \mathbb{R}^{m \times n}$ be the matrix of ratings. For NMF to work it is essential that the matrix be sparse, thus if a user's rating for an item is unknown, the corresponding cell is zeroed. The matrix $W$ is then decomposed into two low-rank non-negative matrices, $T = [t_{u,k}] \in \mathbb{R}^{m \times r}$ and $V = [v_{k,i}] \in \mathbb{R}^{r \times n}$ such that $TV \approx W$ and $r \ll \min(m, n)$. $r$ is a hyperparameter. By multiplication of these two matrices we end up with a reduced matrix $W' = TV$ in which each zeroed cell $w_{u,i}$ from matrix $W$ is replaced with a value $w'_{u,i}$ that represents a prediction for the rating that user $u$ would give to item $i$. NMF solves the following optimization program:

$$\underset{T,V}{\operatorname{argmin}} \left( \|W - TV\|^2 \right)$$
$$\text{subject to} \quad T, V \geqslant 0$$

This optimization problem can be solved by gradient descent using the following updates:

$$t_{u,k} \leftarrow t_{u,k} + \eta_{u,k}((WV^T)_{u,k} - (TVV^T)_{u,k})$$
$$v_{k,i} \leftarrow v_{k,i} + \eta_{k,i}((T^TW)_{k,i} - (T^TTV)_{k,i})$$

In this equation, $\eta$ is the learning rate. To guarantee non-negativity, it is defined as:

$$\eta_{u,k} = \frac{t_{u,k}}{(TVV^T)_{u,k}}, \quad \eta_{k,i} = \frac{v_{k,i}}{(T^TTV)_{k,i}}$$

Note that the objective function only takes account of non-zero cells. Luo et al. (2014) propose an approach that takes advantage of the sparseness of the matrix for faster computation. In addition,

|       | I | can | see | ich | kann | es | sehen | je | vois |
|-------|---|-----|-----|-----|------|----|-------|----|------|
| I     | 5 | 1   | 5   |     | 1    |    | 5     | 1  |      |
| can   |   | 5   | 1   |     | 5    | 1  |       |    |      |
| see   | 1 | 1   | 5   |     | 1    |    | 5     | 1  | 5    |
| ich   | 5 | 1   |     | 5   |      |    | 1     | 5  | 1    |
| kann  | 1 | 5   |     | 1   | 5    |    |       |    |      |
| es    |   |     |     |     |      | 5  | 1     |    |      |
| sehen | 1 |     | 5   |     | 1    |    | 5     | 1  |      |
| je    | 5 | 1   |     | 5   | 1    |    | 5     |    | 1    |
| vois  | 1 |     | 5   | 1   |      |    | 5     | 1  | 5    |

Figure 2: An example of how the original matrix is filled for a sentence in three languages. Zero entries are left blank for readability.

Tikhonov regularization is integrated to improve precision, recall, and convergence rate.

We use the implementaion of NMF provided by the Surprise[3] library, with default hyperparameters ($r = 15$, n_epochs $= 50$).

### 3.2.1 NMF in MPWA framework

We create a separate matrix $W$ for each sentence in the multiparallel corpus. Tokens in the sentence play the role of both users and items, i.e., we consider each token both as a row and as a column. Figure 2 shows an example of a sentence in a toy English-German-French multiparallel corpus. If two tokens are aligned using the bilingual aligner, we fill the corresponding cell with the highest rating (5). To give a few negative examples to the algorithm, if a token $x$ from language $L_1$ is aligned to token $y$ in language $L_2$, we pick another random token $z$ from $L_2$ and fill the corresponding cell of $x$ to $z$ with the lowest rating (1). We zero out all other cells. Next we apply the matrix factorization algorithm to this matrix and then compute the reduced matrix $W'$ from the factors. Now we grab the predicted alignment scores between source and target languages from $W'$. To extract new alignment edges we apply the Argmax algorithm (Jalili Sabet et al., 2020). Argmax creates an alignment edge between word $w_i$ from language $L_1$ and word $w_j$ from language $L_2$ if among all words from $L_2$, $w_i$ has the highest alignment score with $w_j$, and among all words from $L_1$, $w_j$ has the highest alignment score with $w_i$.

## 3.3 Link Prediction

A multiparallel sentence annotated with bilingual word alignments can be considered to be a graph with words from all translations as nodes and the

---

[3] http://surpriselib.com/

8460

word alignments as edges. Link prediction algorithms such as Common Neighbors (CN) and Adamic-Adar (AdAd) estimate the likelihood of having an edge between two nodes $x$ and $y$ in the graph based on the similarity of their neighborhoods. The CN index weights all common neighbors equally. In contrast, AdAd gives higher weight to neighbors with low degrees because sharing a neighbor that in turn has few neighbors is more significant. Therefore, we use the AdAd index. It is defined as:

$$\text{AdAd}_{x,y} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|} \qquad (1)$$

where $\Gamma(x)$ is the neighborhood of $x$.

If we use a word aligner that produces a score for each alignment edge, we can use Weighted Adamic-Adar (Lü and Zhou, 2010):

$$\text{WAdAd}_{x,y} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(z,y)}{\log(1 + S(z))} \quad (2)$$

where $w(x,z)$ is the similarity score of $x$ and $z$ generated by the aligner and $S(x) = \sum_{z \in \Gamma(x)} w(x,z)$. For embedding-based aligners we use embedding similarity as the score $w(x,z)$. If an aligner does not provide scores, we set all weights to 1.0.

Given a scored word alignment, we create a multilingual word alignment matrix $W$ for each sentence as shown in Figure 2. Each cell contains 0 or 1 for Adamic-Adar or the alignment score for Weighted Adamic-Adar. We again apply Argmax to extract new alignment edges and then add them to the original alignment.

## 4 Experimental setup

### 4.1 PBC

The PBC corpus (Mayer and Cysouw, 2014) contains 1758 editions of the Bible in 1334 languages. The editions are aligned at the verse level and tokenized. A verse can contain more than one sentence, but we treat it as one unit in the parallel corpus since a true sentence level alignment is not available. There are some errors in tokenization (e.g., for Tibetan, Khmer and Chinese), but the overall quality of the corpus is good. For the majority of languages one edition is provided, while a few languages (in particular, English, French and German) contain up to dozens of editions. The verse coverage also differs from language to language. Some languages have translations of both

New Testament and Hebrew Bible while others contain only one. Table 2 gives corpus statistics.

### 4.2 Word alignment datasets

PBC does not provide gold word alignments. To evaluate MPWA, we port two word alignment gold datasets of the Bible to PBC: Blinker (Melamed, 1998) and the recently published HELFI (Yli-Jyrä et al., 2020). We further experiment with bilingual datasets, using Machine Translation (MT) to create multiparallel corpora. Table 1 gives dataset statistics.

**The HELFI dataset** consists of the Greek New Testament, the Hebrew Bible and translations of both into Finnish. In addition, morpheme alignments are provided for Finnish-Greek and Finnish-Hebrew. We reformatted this dataset to the format used by PBC. In more detail, we added three new editions for the three languages to PBC. We identified the PBC verse identifier for each verse of HELFI to ensure proper verse alignment of these three new editions. The Finnish-Hebrew dataset has 22,291 verses and the Finnish-Greek dataset 7,909. We split these datasets 80/10/10 into train, validation and test.

**The Blinker Bible dataset** provides word level alignments of 250 Bible verses between English and French. The French side of this dataset matches with the edition Louis Segond 1910 in PBC. However, the tokenizations (Blinker vs PBC) are different. We therefore create a mapping of the tokens using character n-gram matching. For English, we created and added a new edition to PBC.

**MT datasets**. To more broadly evaluate MPWA, we also create multiparallel datasets for four non-Bible word alignment gold standards; these are listed in Table 1 as "Non-Bible" corpora. For these gold standards, we translate from English to all languages available in Google Translate, using their API.[4] For the added languages, we create alignments for the gold standard sentences using SimAlign.

### 4.3 Initial word alignments

We compare with two state of the art models, one statistical, one neural. Eflomal (Östling and Tiedemann, 2016) is a Bayesian statistical word aligner using Markov Chain Monte Carlo inference. SimAlign (Jalili Sabet et al., 2020) obtains word align-

---

[4] https://cloud.google.com/translate/docs/basic/translating-text

| Language Pair | | Name | # Sentences (train/valid./test) |
|---|---|---|---|
| **Bible** FIN-HEB | ‖ | HELFI (Yli-Jyrä et al., 2020) | 22291 (17832/2229/2230) |
| FIN-GRC | ‖ | HELFI (Yli-Jyrä et al., 2020) | 7909 (6327/791/791) |
| ENG-FRA | ‖ | BLINKER (Melamed, 1998) | 250 |
| **Non-** ENG-DEU | ‖ | EuroParl-based[a] | 508 |
| **Bible** ENG-FAS | ‖ | (Tavakoli and Faili, 2014) | 400 |
| ENG-HIN | ‖ | WPT2005[b] | 90 |
| ENG-RON | ‖ | WPT2005[b] | 203 |

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

Table 1: Overview of datasets. We use ISO 639-3 language codes. # Sentences: the number of available verses (i.e., sentences). FIN-HEB and FIN-GRC datasets split into train, validation and test.

| # editions | 1758 |
|---|---|
| # languages | 1334 |
| # verses | 20,470,892 |
| # verses / # editions | 11,520 |
| # tokens / # verses | 28.6 |

Table 2: PBC corpus statistics

ments from multilingual pretrained language models with no need for parallel data. For the symmetrization of Eflomal, we use grow-diag-final-and (GDFA) and intersection, and for SimAlign we use Argmax and Itermax. Intersection and Argmax generate accurate alignments while GDFA and Itermax are less accurate but have better coverage (Jalili Sabet et al., 2020).

We evaluate on a *target language pair* parallel sentence as follows: First, we create the matrix (Figure 2) for this sentence for all languages in the multiparallel corpus. Then we run link prediction on the matrix – this accumulates evidence from a set of languages in the multiparallel corpus. Finally, we take the predictions for the target language pair and add them to the original (bilingual) alignment.

NMF works best if it starts with high-accuracy (i.e., non-noisy) bilingual alignments – errors can result in incorrectly predicted alignment edges. We therefore use SimAlign Argmax and Eflomal Intersection, two word alignment methods with high precision, to create the initial alignments that are then fed into NMF. We then add the predictions to any desired original alignments; e.g., NMF (GDFA) uses Eflomal Intersection as the initial alignments and adds the predictions to Eflomal GDFA. See the Appendix for more details.

SimAlign offers high quality word alignments for well-represented languages from pretrained language models; however, our experiments show that its performance is far behind Eflomal for less well resourced languages like Biblical Hebrew and Koine Greek. Also, Eflomal is a better match for

MPWA because it can provide word alignments for all languages available in a multiparallel corpus. In contrast, SimAlign is limited to languages supported by pretrained multilingual embeddings.

To feed Eflomal with enough training data for a target language pair, we use all available data from different translations of the language pair. For example if one language has two translations and the other one has three translations, Eflomal's training data will contain six aligned translation pairs for these two languages.

We use the standard evaluation measures for word alignment: precision, recall, $F_1$ and Alignment Error Rate (AER) (Och and Ney, 2003b; Östling and Tiedemann, 2016; Jalili Sabet et al., 2020).

## 5 Results

### 5.1 Multiparallel corpus results

We perform the first set of experiments on the Blinker Bible and the HELFI gold standards in the PBC. The baseline results are calculated on the original language pairs. MPWA can be applied to both Eflomal and SimAlign alignments. Since the default version of SimAlign can only generate alignments for the 84 languages that multilingual BERT supports,[5] for a better comparison, we use the same set of languages in the alignment graph for both SimAlign and Eflomal.

Table 3 shows the results for our methods applied on SimAlign and Eflomal baselines.[6] AdAd, NMF and WAdAd substantially improve the performance for all language pairs. SimAlign generates high-quality alignments for the English-French dataset, but cannot properly align underresourced languages like Biblical Hebrew and Koine Greek.

---

[5] https://github.com/google-research/bert/blob/master/multilingual.md
[6] We only consider SimAlign IterMax, not SimAlign ArgMax, because IterMax performed better throughout.

| | | FIN-HEB | | | | FIN-GRC | | | | ENG-FRA | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Method | Prec. | Rec. | $F_1$ | AER | Prec. | Rec. | $F_1$ | AER | Prec. | Rec. | $F_1$ | AER |
| Baseline | Eflomal (intersection) | **0.818** | 0.269 | 0.405 | 0.595 | **0.897** | 0.506 | 0.647 | 0.353 | **0.971** | 0.521 | 0.678 | 0.261 |
| | Eflomal (GDFA) | 0.508 | 0.448 | 0.476 | 0.524 | 0.733 | 0.671 | 0.701 | 0.300 | 0.856 | 0.710 | 0.776 | 0.221 |
| | SimAlign | 0.190 | 0.113 | 0.142 | 0.858 | 0.366 | 0.265 | 0.307 | 0.693 | 0.886 | 0.692 | 0.777 | 0.221 |
| Init SimAlign | AdAd | 0.199 | 0.127 | 0.155 | 0.845 | 0.402 | 0.289 | 0.336 | 0.664 | 0.878 | 0.731 | 0.798 | 0.200 |
| | WAdAd | 0.186 | 0.165 | 0.175 | 0.825 | 0.353 | 0.350 | 0.351 | 0.649 | 0.856 | 0.752 | 0.801 | 0.197 |
| | NMF | 0.122 | 0.100 | 0.110 | 0.890 | 0.396 | 0.337 | 0.364 | 0.636 | 0.835 | 0.700 | 0.762 | 0.236 |
| Init Eflomal | WAdAd (intersection) | 0.781 | 0.612 | **0.686** | **0.314** | 0.849 | 0.696 | **0.765** | **0.235** | 0.938 | 0.689 | 0.794 | 0.203 |
| | NMF (intersection) | 0.78 | 0.576 | 0.663 | 0.337 | 0.864 | 0.669 | 0.754 | 0.248 | 0.948 | 0.624 | 0.753 | 0.245 |
| | WAdAd (GDFA) | 0.546 | **0.693** | 0.611 | 0.389 | 0.707 | **0.783** | 0.743 | 0.257 | 0.831 | **0.796** | **0.813** | **0.186** |
| | NMF (GDFA) | 0.548 | 0.646 | 0.593 | 0.407 | 0.72 | 0.759 | 0.739 | 0.261 | 0.844 | 0.767 | 0.804 | 0.195 |

Table 3: Comparison of results from different methods on PBC. The best results in each column are in bold. The three methods exploiting multiparallelism (AdAd, WAdAd, NMF) generally outperform the baselines on $F_1$ and AER.

In such cases, MPWA uses the accumulated information from all other language pairs in the graph to improve the performance. When starting with the SimAlign alignment ("Init SimAlign"), both methods improve the result for both FIN-HEB and FIN-GRC.

Eflomal generates better alignments for FIN-HEB and FIN-GRC. This means that Eflomal also generates better alignments between FIN, HEB and GRC on the one hand and the other languages in the graph on the other hand and therefore can provide a better signal for MPWA. The improvements of our models applied on Eflomal are higher than the ones applied on SimAlign for these language pairs.

When changing the initial alignments from Eflomal (intersection) to Eflomal (GDFA), we see different behaviors: GDFA improves the results for Blinker while it does not help for HELFI. We believe this is caused by the different ways the two datasets were annotated. In Blinker, many phrases are "exhaustively" aligned: if a phrase DE in English is aligned with FG in French then all four alignment edges (D-F, D-G, E-F, E-G) are given as gold edges.[7]

So Blinker contains a lot of many-to-many links. In contrast, most alignments are one-to-one in HELFI. This partially explains why intersection as initial alignment works much better for HELFI than GDFA and vice versa for Blinker.

In summary, compared to the baselines, we see very large improvements through exploiting multiparallelism for one type of alignment methodology (HELFI, $F_1$ improved by up to 20% for FIN-

HEB) and improvements of up to 3.5% for the other (ENG-FRA).

## 5.2 MT dataset results

We perform the second set of experiments on gold standard alignments for language pairs that are not part of a multiparallel corpus such as PBC. To this end, we create artificial multiparallel corpora by translating the English side to all languages available in the Google Translate API. The main goal is to give broader evidence for the effectiveness of our method, beyond the specialized domain of the Bible.

Eflomal's alignments generally have good quality. However, they get worse when less parallel data is available (Jalili Sabet et al., 2020). Since the size of the multiparallel corpus created by machine translation is rather small, we use SimAlign for generating initial alignments. SimAlign has been shown to have good performance even for very small parallel corpora; in fact, it does not need any parallel data at all.

Table 4 shows the results of the experiments. Both NMF and WAdAd, improve the performance of the baseline by using the alignment graph. Improvements range from 0.8% (ENG-DEU) to 3.3% (ENG-HIN). This again demonstrates the utility of exploiting multiparallelism for word alignment. It is worth mentioning that in this case the translations are noisy since they were automatically generated. But even with these noisy translations (instead of a "true" multiparallel corpus), our models effectively leverage multiparallelism.

---

[7]For example, the alignment of the phrases "trembled violently" and "fut saisi d'und grande, d'une violente émotion" consists of $2 \cdot 8$ gold edges.

| | Method | ENG-PES | | | | ENG-HIN | | | | ENG-RON | | | | ENG-DEU | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Prec. | Rec. | $F_1$ | AER | Prec. | Rec. | $F_1$ | AER | Prec. | Rec. | $F_1$ | AER | Prec. | Rec. | $F_1$ | AER |
| Baseline | SimAlign | 0.756 | 0.645 | 0.696 | 0.304 | 0.709 | 0.493 | 0.582 | 0.418 | 0.807 | 0.663 | 0.728 | 0.272 | 0.829 | 0.795 | 0.812 | 0.188 |
| Init SimAlign | AdAd | 0.751 | 0.700 | **0.725** | **0.276** | 0.693 | 0.544 | 0.610 | 0.390 | 0.799 | 0.696 | **0.744** | **0.256** | 0.818 | 0.823 | **0.820** | **0.179** |
| | WAdAd | 0.705 | **0.740** | 0.722 | 0.278 | 0.643 | **0.574** | 0.607 | 0.394 | 0.725 | **0.717** | 0.721 | 0.279 | 0.749 | **0.844** | 0.794 | 0.207 |
| | NMF | 0.734 | 0.698 | 0.716 | 0.284 | 0.684 | 0.559 | **0.615** | **0.385** | 0.780 | 0.696 | 0.736 | 0.265 | 0.804 | 0.827 | 0.815 | 0.185 |

Table 4: Results with gold standards translated into other languages using machine translation. The best results in each column are in bold. The three methods exploiting multiparallelism (AdAd, WAdAd, NMF) outperform the baselines on $F_1$ and AER.
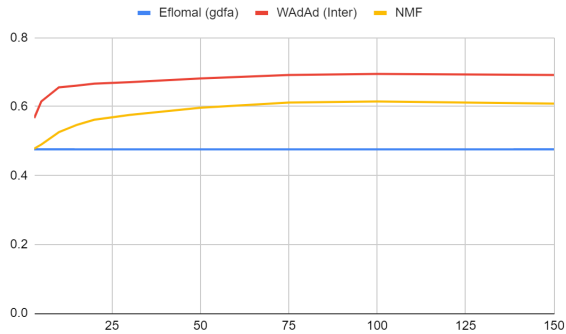


Figure 3: $F_1$ of MPWA for the target language pair FIN-HEB as a function of the number of additional languages. There is a clear rise initially. The curve flattens around 75.



Figure 4: Word alignment $F_1$ on ENG-FRA as a function of the size of the training set, ranging from 30K to 6.4M training sentence pairs

## 5.3 Analysis

### 5.3.1 Effect of number of languages

The effect of adding more languages to the alignment graph is depicted in Figure 3. This plot shows $F_1$ for FIN-HEB. As seen in the figure, the slope is pretty steep up to 25 languages, but even adding just three languages can still improve the results. For 75 languages we have almost reached the peak and after 100, adding more languages is not improving the results. This means that MPWA can also be helpful for corpora with a smaller number of languages – a massively parallel corpus with thousands of languages is not required.

### 5.3.2 Size of the training set

To assess the effect of dataset size on the performance of MPWA, we perform a set of experiments on ENG-FRA and NMF. To this end, we take the training data for ENG-FRA and train models on subsets of it. The training data consists of 6.4M sentence pairs – this number is so high because we use the crossproduct of all editions in English and French (§4.3).

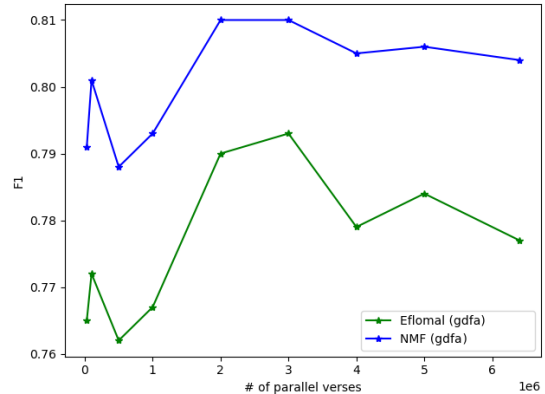The results are shown in Figure 4. Eflomal performance increases with training set size initially

and is then less predictable. NMF consistently improves the scores.

### 5.3.3 Effect of task difficulty

Table 3 shows large improvements for all datasets, and especially for FIN-HEB and FIN-GRC. To get more insight into the reasons for this improvement, we stratify FIN-HEB verses by dividing the interval $[0, 1]$ of initial $F_1$ performance of Eflomal into five equal-sized subintervals: $[0, 0.2], \ldots, (0.8, 1]$.

Figure 5 indicates that MPWA is most effective for difficult verses, but brings little improvement for easy verses. We attribute this to two reasons:

1. An easy to align verse in a language pair cannot use help from other languages since it already has good alignment links (although the language pair would still be of benefit in improving alignments for the sentence in other languages). So there is no way for MPWA to get better results in this scenario.

2. MPWA only tries to get better results by adding new alignments, and as an easy verse already has many alignment links, adding new links almost inevitably results in a drop in pre-
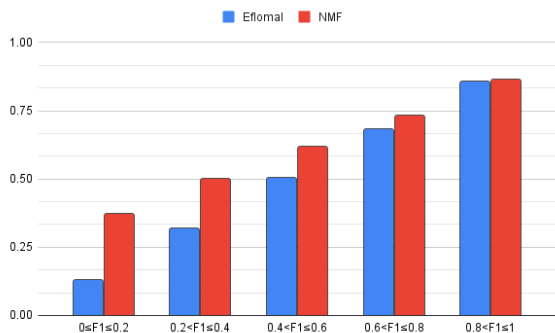
Figure 5: How helpful is MPWA for different difficulty levels? For this analysis, FIN-HEB verses were stratified according to Eflomal $F_1$ (x-axis). We see that MPWA brings the largest improvements for difficult sentences.

| ENG-FRA | | FIN-HEB | | FIN-GRC | |
|---|---|---|---|---|---|
| Lang. | $\Delta$ | Lang. | $\Delta$ | Lang. | $\Delta$ |
| SPA | +2.0% | TGL | +17.7% | LAT | +7.5% |
| ITA | +1.9% | FRY,ELL | +17.3% | ELL | +6.6% |
| DEU | +1.8% | SWE | +17.3% | ENG | +6.1% |
| NLD | +1.4% | NLD | +16.8% | FRY | +5.8% |
| AFR | +1.3% | YOR | +14.2% | BEL | +5.7% |

Table 5: The five most helpful languages and WAdAd's absolute improvements in $F_1$ over the initial bilingual aligner SimAlign. For example, MPWA improves the bilingual FIN-GRC alignment by 7.5% if applied to the trilingual corpus FIN-GRC-LAT, i.e., Latin can be viewed as the best bridge between Finnish and Greek.

cision. It may also be possible to inspect and prune existing Eflomal links using MPWA to get better results in this scenario.

### 5.3.4 Most helpful languages

For each dataset, the five most helpful languages with their corresponding improvements are listed in Table 5. We hypothesize that these languages serve to bridge the typological gap between the two target languages. Table 5 suggests one should be able to achieve excellent results – even for a corpus with a small number of languages – if we utilize an intelligent selection of languages.

### 5.3.5 Multiple translations in two languages

There are some datasets that contain few languages, but many translations of a text in one language. PBC is one example of such a dataset, many literary works another (e.g., many novels have many translations in English). To see whether MPWA can also help in this scenario, we picked all available 49 English and French editions from PBC and used them as additional translations for the ENG-FRA dataset. The outcome of this experiment is

| | Prec. | Rec. | $F_1$ | AER |
|---|---|---|---|---|
| Eflomal (intersection) | 0.971 | 0.521 | 0.678 | 0.319 |
| Eflomal (GDFA) | 0.856 | 0.710 | 0.776 | 0.221 |
| NMF (target languages) | 0.830 | 0.749 | 0.787 | 0.213 |
| NMF (other languages) | 0.837 | 0.753 | 0.793 | 0.205 |

Table 6: $F_1$ for ENG-FRA. MPWA can exploit a multiparallel corpus with languages different from the target languages ("other languages") better than one that contains additional translations in the target languages ("target languages").

compared with the outcome of the same setup, but with translations from languages other than French and English in Table 6. From this table we can conclude that translations from the target language pair can also assist, but not as much as translations from other languages.

## 6 Conclusion and Future Work

We presented MPWA, a framework for leveraging multiparallel corpora for word alignment. We used two prediction methods, one based on recommender systems and one based on link prediction algorithms. By adding new alignment edges to the output of bilingual aligners, both methods show large improvements over the bilingual baselines, with absolute improvements of $F_1$ of up to 20%. We have also ported Blinker and HELFI word alignment gold standards to the Parallel Bible Corpus in the hope that this will help foster more work on exploiting multiparallel corproa.

**Future work.** In this paper, we have mainly focused on *adding* new alignment edges to baseline word alignments based on properties of the multiparallel alignment graph. This increases recall, but can harm precision. In future work, we plan to expand on the possibility of *deleting* edges based on evidence from the multiparallel alignment graph (cf. 5.3.3), thereby potentially improving both precision and recall.

## Acknowledgments

# References

Lada A Adamic and Eytan Adar. 2003. Friends and neighbors on the web. *Social networks*, 25(3):211–230.

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, Berlin, Germany. Association for Computational Linguistics.

Tamer Alkhouli and Hermann Ney. 2017. Biasing attention-based recurrent neural networks using external alignment information. In *Proceedings of the Second Conference on Machine Translation*, pages 108–117, Copenhagen, Denmark. Association for Computational Linguistics.

Ehsaneddin Asgari and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124, Copenhagen, Denmark. Association for Computational Linguistics.

Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science*, 286(5439):509–512.

Nicola Bertoldi, Madalina Barbaiani, Marcello Federico, and Roldano Cattoni. 2008. Phrase-based statistical machine translation with pivot languages. In *International Workshop on Spoken Language Translation (IWSLT) 2008*.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117.

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).

Jaegul Choo, Changhyun Lee, Chandan K Reddy, and Haesun Park. 2013. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE transactions on visualization and computer graphics*, 19(12):1992–2001.

Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 728–735, Prague, Czech Republic. Association for Computational Linguistics.

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.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

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.

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.

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.

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.

Karim Filali and Jeff Bilmes. 2005. Leveraging multiple languages to improve statistical MT word alignments. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, pages 92–97. IEEE.

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.

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 Conference on Natural Language Processing: System Demonstrations*, pages 63–72, 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.

Glen Jeh and Jennifer Widom. 2002. Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

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.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Da Kuang, Jaegul Choo, and Haesun Park. 2015. Non-negative matrix factorization for interactive topic modeling and document clustering. In *Partitional Clustering Algorithms*, pages 215–243. Springer.

Shankar Kumar, Franz J. Och, and Wolfgang Macherey. 2007. Improving word alignment with bridge languages. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 42–50, Prague, Czech Republic. Association for Computational Linguistics.

Adrien Lardilleux and Yves Lepage. 2008a. A truly multilingual, high coverage, accurate, yet simple, sub-sentential alignment method. In *The 8th conference of the Association for Machine Translation in the Americas (AMTA 2008)*, pages 125–132, Waikiki, Honolulu, United States.

Adrien Lardilleux and Yves Lepage. 2008b. Multilingual alignments by monolingual string differences. In *Coling 2008: Companion volume: Posters*, pages 55–58, Manchester, UK. Coling 2008 Organizing Committee.

Adrien Lardilleux and Yves Lepage. 2009. Sampling-based multilingual alignment. In *Proceedings of the International Conference RANLP-2009*, pages 214–218, Borovets, Bulgaria. Association for Computational Linguistics.

William D. Lewis and Fei Xia. 2008. Automatically identifying computationally relevant typological features. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Linyuan Lü and Tao Zhou. 2010. Link prediction in weighted networks: The role of weak ties. *EPL (Europhysics Letters)*, 89(1):18001.

Xin Luo, Mengchu Zhou, Yunni Xia, and Qingsheng Zhu. 2014. An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems. *IEEE Transactions on Industrial Informatics*, 10(2):1273–1284.

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*.

Thomas Mayer and Michael Cysouw. 2012. Language comparison through sparse multilingual word alignment. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 54–62, Avignon, France. Association for Computational Linguistics.

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).

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

Mathias Müller. 2017. Treatment of markup in statistical machine translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 36–46, Copenhagen, Denmark. Association for Computational Linguistics.

Preslav Nakov and Hwee Tou Ng. 2012. Improving statistical machine translation for a resource-poor language using related resource-rich languages. *Journal of Artificial Intelligence Research*, 44:179–222.

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

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

Robert Östling. 2014. Bayesian word alignment for massively parallel texts. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 123–127. The Association for Computer Linguistics.

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*, 106(1).

Mohammad Sadegh Rasooli, Noura Farra, Axinia Radeva, Tao Yu, and Kathleen McKeown. 2018. Cross-lingual sentiment transfer with limited resources. *Machine Translation*, 32(1):143–165.

Sascha Rothe and Hinrich Schütze. 2014. CoSimRank: A flexible & efficient graph-theoretic similarity measure. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1392–1402, Baltimore, Maryland. Association for Computational Linguistics.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Farial Shahnaz, Michael W Berry, V Paul Pauca, and Robert J Plemmons. 2006. Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2):373–386.

Haoyue Shi, Luke Zettlemoyer, and Sida I. Wang. 2021. Bilingual lexicon induction via unsupervised bitext construction and 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 813–826, Online. Association for Computational Linguistics.

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

Shijie Wu and Mark Dredze. 2020. Do explicit alignments robustly improve multilingual encoders? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4471–4482, Online. Association for Computational Linguistics.

Wei Xu, Xin Liu, and Yihong Gong. 2003. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273.

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.

Hamed Zamani, Javid Dadashkarimi, Azadeh Shakery, and W Bruce Croft. 2016. Pseudo-relevance feedback based on matrix factorization. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 1483–1492.

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.

Muhan Zhang and Yixin Chen. 2018. Link prediction based on graph neural networks. *Advances in Neural Information Processing Systems*, 31:5165–5175.

Shaojun Zhao and Daniel Gildea. 2010. A fast fertility hidden Markov model for word alignment using MCMC. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 596–605, Cambridge, MA. Association for Computational Linguistics.

Tao Zhou, Linyuan Lü, and Yi-Cheng Zhang. 2009. Predicting missing links via local information. *The European Physical Journal B*, 71(4):623–630.

## A  Pipeline Details

There are several elements of the MPWA pipeline that can be configured by the user, e.g., depending on whether precision or recall are more important for an application. Here we show in Figures 6 and 7 the two pipeline configurations we used for the results in the paper.
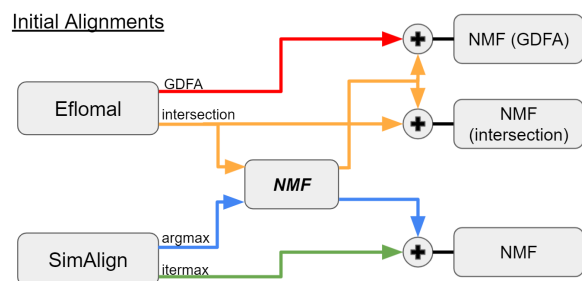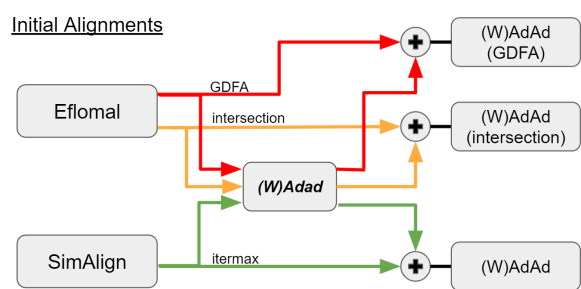


Figure 6: The pipeline for NMF alignments

Figure 7: The pipeline for AdAd and WAdAd alignments