Unbalanced Optimal Transport for Unbalanced Word Alignment

Yuki Arase
Osaka University, Japan
arase@ist.osaka-u.ac.jp

Han Bao
Kyoto University, Japan
bao@i.kyoto-u.ac.jp

Sho Yokoi
Tohoku University, Japan
RIKEN, Japan
yokoi@tohoku.ac.jp

Abstract

Monolingual word alignment is crucial to model semantic interactions between sentences. In particular, null alignment, a phenomenon in which words have no corresponding counterparts, is pervasive and critical in handling semantically divergent sentences. Identification of null alignment is useful on its own to reason about the semantic similarity of sentences by indicating there exists information inequality. To achieve *unbalanced* word alignment that values both alignment and null alignment, this study shows that the family of optimal transport (OT), i.e., balanced, partial, and unbalanced OT, are natural and powerful approaches even without tailor-made techniques. Our extensive experiments covering unsupervised and supervised settings indicate that our generic OT-based alignment methods are competitive against the state-of-the-arts specially designed for word alignment, remarkably on challenging datasets with high null alignment frequencies.

1 Introduction

Monolingual word alignment, which identifies semantically corresponding words in a sentence pair, has been actively studied as a crucial technique for modelling semantic relationships between sentences, such as for paraphrase identification, textual entailment recognition, and question answering (MacCartney et al., 2008; Das and Smith, 2009; Wang and Manning, 2010; Heilman and Smith, 2010; Yao et al., 2013a; Feldman and El-Yaniv, 2019). Its ability to declare redundant information in sentences is also useful for summarisation and sentence fusion (Thadani and McKeown, 2013; Brook Weiss et al., 2021). In addition, the alignment information is valuable for interpretability of model predictions (Agirre et al., 2015; Li and Srikumar, 2016) and for realising interactive document exploration as well (Shapira et al., 2017; Hirsch et al., 2021).

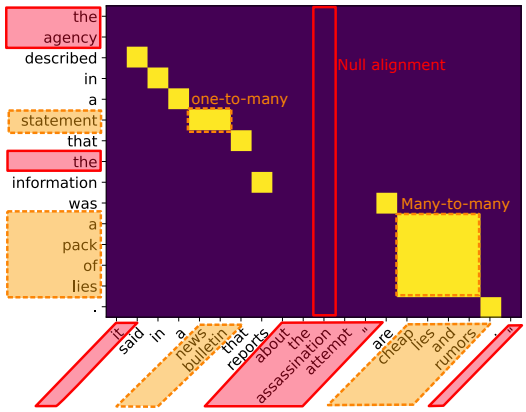


Figure 1: Unbalanced monolingual word alignment matrix; frequent null alignment (enclosed in red boxes) and mapping beyond one-to-one (enclosed in orange dashed boxes) are primary challenges.

Figure 1 illustrates the challenges of monolingual word alignment. The first challenge is the *null alignment*, where words may not have corresponding counterparts, which causes alignment asymmetricity (MacCartney et al., 2008). Null alignment is prevalent in semantically divergent sentences; indeed, the null alignment ratio reaches 63.8% in entailment sentence pairs used in our experiments (shown in the third row of Table 2). Its identification explicitly declares semantic gaps between two sentences and helps reason about their semantic (dis)similarity (Li and Srikumar, 2016). The second challenge is that alignment *beyond one-to-one mapping* needs to be addressed. These challenges constitute an *unbalanced* word alignment problem, where both word-by-word and null alignment should be fully identified.

This study reveals that a family of optimal transport (OT) are suitable tools for unbalanced word alignment. Among the OT problems, balanced OT (**BOT**) (Monge, 1781; Kantorovich, 1942)[1] should be the most prominent in natural language pro-

[1]We explicitly call it *balanced* OT (BOT) to distinguish it from partial and unbalanced OT in this paper.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3966–3986

cessing (NLP) (Wan, 2007; Kusner et al., 2015), which can handle many-to-many alignment. In contrast to BOT that is unable to deal with null alignment, partial OT (**POT**) (Caffarelli and McCann, 2010; Figalli, 2010) and unbalanced OT (**UOT**) (Frogner et al., 2015; Chizat et al., 2016) can handle the asymmetricity as desired in unbalanced word alignment, which has also attracted applications where null alignment is unignorable (Swanson et al., 2020; Zhang et al., 2020a; Yu et al., 2022; Chen et al., 2020b; Wang et al., 2020).

This is the first study that connects these two paradigms of unbalanced word alignment and the family of OT problems that can naturally address null alignment as well as many-to-many alignment. We empirically (1) demonstrate that the OT-based methods are natural and sufficiently powerful approaches to unbalanced word alignment without tailor-made techniques and (2) deliver a comprehensive picture that unveils the characteristics of BOT, POT, and UOT on unbalanced word alignment with different null alignment ratios. We conduct extensive experiments using representative datasets for understanding OT-based alignment: effects of OT formalisation, regularisation, and a heuristic to sparsify alignments. Our primary findings can be summarised as follows. First, in unsupervised alignment, the best OT problem depends on null alignment ratios. Second, simple thresholding on regularised BOT can produce unbalanced alignment. Third, in supervised alignment, simple and generic OT-based alignment shows competitive performance to the state-of-the-art models specially designed for word alignment.

The adoption of well-understood methods with a solid theory like OT is highly valuable in application development and scientific reproducibility. Furthermore, OT-based alignment performs superiorly or competitively to existing methods, despite being independent of tailor-made techniques. Our OT-based alignment methods are publicly available as a tool called *OTAlign*.[2]

2 Related Work

2.1 Word Alignment Problem

Word alignment techniques have been actively studied in the crosslingual setting, primarily for machine translation. Crosslingual and monolingual word alignment tackle relevant problems; however, they have notable differences (MacCartney et al., 2008). Crosslingual alignment can assume that a large-scale parallel corpus exists, which is scarce in the monolingual case. Furthermore, asymmetricity can be more intense in monolingual alignment to handle semantically divergent sentences. A common approach to crosslingual word alignment is unsupervised learning using parallel corpora (Och and Ney, 2003; Dyer et al., 2013; Garg et al., 2019; Zenkel et al., 2020; Dou and Neubig, 2021). Among them, Dou and Neubig (2021) applied BOT for alignment while focussing on fine-tuning multilingual pre-trained language models using a parallel corpus. Among crosslingual word alignment methods, SimAlign (Jalili Sabet et al., 2020) is directly applicable to monolingual alignment because it only uses a multilingual pre-trained language model and simple heuristics without parallel corpora. In addition, the supervised word alignment method proposed by Nagata et al. (2020), which modelled alignment as a question-answering task, is also applicable.

For monolingual word alignment, supervised learning has been commonly used (MacCartney et al., 2008; Thadani and McKeown, 2011; Thadani et al., 2012a). A representative method modelled word alignment using the conditional random field regarding source words as observations and target words as hidden states (Yao et al., 2013a,b). Lan et al. (2021) enhanced this approach by adopting neural models, which constitute the state-of-the-art methods together with Nagata et al. (2020). In monolingual alignment, phrase alignment is also a research focus (Ouyang and McKeown, 2019; Arase and Tsujii, 2020; Culkin et al., 2021), which depends on high-quality parsers and chunkers. In contrast, we use words as an alignment unit and do not assume the availability of such parsers.

2.2 Optimal Transport in NLP

BOT, POT, and UOT have been adopted in various NLP tasks where alignment exists implicitly or explicitly. Most previous studies modelled tasks of their interest, with alignment being a hidden state; further, the alignment quality itself was out of their focus. A typical application is similarity estimation among sentences and documents using BOT (Kusner et al., 2015; Huang et al., 2016; Zhao et al., 2019; Yokoi et al., 2020; Chen et al., 2020b; Alqahtani et al., 2021; Lee et al., 2022; Mysore et al., 2022; Wang et al., 2022) and recently UOT (Chen et al., 2020b; Wang et al., 2020). Such similarity

[2] https://github.com/yukiar/OTAlign

estimation mechanisms can be integrated into language generation models as a penalty (Chen et al., 2019; Zhang et al., 2020b; Li et al., 2020).

Unlike the tasks above, alignment information obtained with BOT, POT, and UOT is the primary concern in extractive summarization (Tang et al., 2022), text matching (Swanson et al., 2020; Zhang et al., 2020a; Yu et al., 2022), and bilingual lexicon induction (Zhang et al., 2017; Grave et al., 2019; Zhao et al., 2020). However, these tasks and unbalanced monolingual word alignment have different objectives. These tasks concern the precision where *alignment exists*. In contrast, unbalanced word alignment aims at exhaustive identification of both alignment and null alignment. Furthermore, there has not been any systematic study that compared the qualities of different OT-based alignment methods, which hinders the understanding of a suitable OT formulation for problems with different null alignment frequencies.

3 Problem Definition

Suppose we have a source and target sentence pair $s = \{s_1, s_2, \ldots, s_n\}$ and $t = \{t_1, t_2, \ldots, t_m\}$ with their word embeddings $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_n\}$ and $\{\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_m\}$, respectively, where $\boldsymbol{s}_i, \boldsymbol{t}_j \in \mathbb{R}^d$. The goal of monolingual word alignment is to identify an alignment $\boldsymbol{P} \in \mathbb{R}_+^{n \times m}$ between semantically corresponding words, where $P_{i,j}$ indicates a likelihood or binary indicator of aligning s_i and t_j.

Evaluation Metrics for Unbalanced Alignment As evaluation metrics, we adopt the macro averages of precision and recall of alignment pairs, and their F1 score (MacCartney et al., 2008). In light of the importance of null alignment, we explicitly incorporate it into the evaluation metrics:

$$\text{precision} = \frac{|\widehat{\mathbb{Y}}_a \cap \mathbb{Y}_a| + |\widehat{\mathbb{Y}}_\emptyset \cap \mathbb{Y}_\emptyset|}{|\widehat{\mathbb{Y}}_a| + |\widehat{\mathbb{Y}}_\emptyset|},$$

$$\text{recall} = \frac{|\widehat{\mathbb{Y}}_a \cap \mathbb{Y}_a| + |\widehat{\mathbb{Y}}_\emptyset \cap \mathbb{Y}_\emptyset|}{|\mathbb{Y}_a| + |\mathbb{Y}_\emptyset|},$$

where $\widehat{\mathbb{Y}}_a$ and $\mathbb{Y}_a$ are sets of word-by-word alignment pairs in prediction and ground-truth, respectively, e.g., $(s_i, t_j) \in \mathbb{Y}_a$ if s_i aligns to t_j. The sets of $\widehat{\mathbb{Y}}_\emptyset$ and $\mathbb{Y}_\emptyset$ correspond to those of null-alignment that regards a word aligning to a null word $w_\emptyset$, e.g., $(s_k, w_\emptyset) \in \mathbb{Y}_\emptyset$ if s_k is null alignment. Hence, $|\mathbb{Y}_\emptyset|$ is equal to the number of null alignment words in the ground-truth. Compared to previous metrics that only consider $\mathbb{Y}_a$ and $\widehat{\mathbb{Y}}_a$ (MacCartney et al., 2008), ours consider null alignment equally as word-by-word alignment.

4 Background: Optimal Transport

OT seeks the most efficient way to move mass from one measure to another. Remarkably, the OT problem induces the OT *mapping* that indicates correspondences between the samples. While OT is frequently used as a distance metric between two measures, OT mapping is often the primary concern in alignment problems.

Formally, the inputs to the optimal transport problem are a cost function and a pair of measures. On the premise of its application to monolingual word alignment, the following explanation assumes that the source and target sentences s and t and their word embeddings are at hand (see §3). A *cost* means a dissimilarity between $\boldsymbol{s}_i$ and $\boldsymbol{t}_j$ (source and target words) computed by a distance metric $c\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$, such as Euclidean and cosine distances. The cost matrix $\boldsymbol{C} \in \mathbb{R}_+^{n \times m}$ summarises the costs of any word pairs, that is, $C_{i,j} = c(\boldsymbol{s}_i, \boldsymbol{t}_j)$. A *measure* means a weight each word has. The concept of measure corresponds to *fertility* introduced in IBM Model 3 (Brown et al., 1993), which defines how many target (source) words a source (target) word can align. In summary, the mass of words in s and t is represented as arbitrary measures $\boldsymbol{a} \in \mathbb{R}_+^n$ and $\boldsymbol{b} \in \mathbb{R}_+^m$, respectively.[3] Finally, the OT problem identifies an alignment matrix $\boldsymbol{P}$ with which the sum of alignment costs is minimised under the cost matrix $\boldsymbol{C}$:

$$L_C(\boldsymbol{a}, \boldsymbol{b}) \coloneqq \min_{\boldsymbol{P} \in \boldsymbol{U}(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{C}, \boldsymbol{P} \rangle ,$$

where $\langle \boldsymbol{C}, \boldsymbol{P} \rangle \coloneqq \sum_{i,j} C_{i,j} P_{i,j}$. The alignment matrix $\boldsymbol{P}$ belongs to $\boldsymbol{U}(\boldsymbol{a}, \boldsymbol{b}) \subseteq \mathbb{R}_+^{n \times m}$ that is a set of valid alignment matrices, introduced in the following sections. With this formulation, we can seek the most plausible word alignment matrix $\boldsymbol{P}$.

4.1 Balanced Optimal Transport

BOT (Kantorovich, 1942) assumes that $\boldsymbol{a}$ and $\boldsymbol{b}$ are probability distributions, i.e., $\boldsymbol{a} \in \Sigma_n$ and $\boldsymbol{b} \in \Sigma_m$, respectively, where Σ_n is the probability simplex: $\Sigma_n \coloneqq \{\boldsymbol{p} \in \mathbb{R}_+^n : \sum_i p_i = 1\}$. In this case, alignment matrices must live in the following space:

$$\boldsymbol{U}^b(\boldsymbol{a}, \boldsymbol{b}) \coloneqq \{\boldsymbol{P} \in \mathbb{R}_+^{n \times m} : \boldsymbol{P}\mathbb{1}_m = \boldsymbol{a}, \boldsymbol{P}^\top \mathbb{1}_n = \boldsymbol{b}\},$$

[3] Subsequently, we use the terms *mass* and *fertility* interchangeably to indicate each element of measures.

where $\mathbb{1}_n$ is the all-ones vector of size n. Under this constraint set, the BOT problem is a linear programming (LP) problem and can be solved by standard LP solvers (Pele and Werman, 2009). However, the non-differentiable nature makes it challenging to integrate into neural models.

Regularised BOT The entropy-regularised optimal transport (Cuturi, 2013), initially aimed at improving the computational speed of BOT, is a differentiable alternative and thus can be directly integrated into neural models. The regularisation makes the objective as follows:

$$L_C^{\varepsilon}(\boldsymbol{a}, \boldsymbol{b}) \coloneqq \min_{\boldsymbol{P} \in \boldsymbol{U}^b(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{C}, \boldsymbol{P} \rangle + \varepsilon H(\boldsymbol{P}),$$

where the function H is the negative entropy of alignment matrix $\boldsymbol{P}$, and ε controls the strength of the regularisation; with sufficiently small ε, L_C^{ε} well approximates the exact BOT. The optimisation problem can be efficiently solved using the Sinkhorn algorithm (Sinkhorn, 1964).

4.2 Relaxation of BOT Constraint

Despite its success, the hard constraint $\boldsymbol{U}^b$ of the BOT that aligns all words often makes it suboptimal for some alignment problems where null alignment is more or less pervasive, such as unbalanced word alignment. POT and UOT introduced subsequently relax this constraint to allow certain words to be left unaligned.

Partial Optimal Transport POT (Caffarelli and McCann, 2010; Figalli, 2010) relaxes BOT by aligning only a fraction m of the fertility, with the following constraint set in place of $\boldsymbol{U}^b$:

$$\boldsymbol{U}^p(\boldsymbol{a}, \boldsymbol{b}) \coloneqq \{\boldsymbol{P} \in \mathbb{R}_+^{n \times m} : \boldsymbol{P}\mathbb{1}_m \leq \boldsymbol{a}, \boldsymbol{P}^\top \mathbb{1}_n \leq \boldsymbol{b}, \\ \mathbb{1}_m^\top \boldsymbol{P}^\top \mathbb{1}_n = m\}.$$

The fraction m is bound as $m \leq \min(\|\boldsymbol{a}\|_1, \|\boldsymbol{b}\|_1)$ where $\|\cdot\|_1$ represents the L^1 norm. While POT can be solved by standard LP solvers, it can also be regularised as in the BOT and solved with the Sinkhorn algorithm (Benamou et al., 2015).

Unbalanced Optimal Transport UOT relaxes BOT by introducing soft constraints that penalise marginal deviation.

$$L_C^{\varepsilon}(\boldsymbol{a}, \boldsymbol{b}) \coloneqq \min_{\boldsymbol{P} \in \mathbb{R}_+^{n \times m}} \langle \boldsymbol{C}, \boldsymbol{P} \rangle + \varepsilon H(\boldsymbol{P}) \\ + \tau_a D(\boldsymbol{P}\mathbb{1}_m, \boldsymbol{a}) + \tau_b D(\boldsymbol{P}^\top \mathbb{1}_n, \boldsymbol{b}),$$

where D is a divergence and τ_a and τ_b control how much mass deviations are penalised. Notice that the unbalanced formulation seeks alignment matrices in the entire $\mathbb{R}_+^{n \times m}$. In this study, we adopt the Kullback–Leibler divergence as D because of its simplicity in computation (Frogner et al., 2015; Chizat et al., 2016).

5 Proposal: Unbalanced Word Alignment as Optimal Transport Problem

In this study, we aim at formulating monolingual word alignment with high null frequencies in a natural way. For this purpose, we leverage OT with different constraints and reveal their features on this problem. We adopt the basic and generic cost matrices and measures instead of engineering them to avoid obfuscating the difference between BOT, POT, and UOT.

5.1 Cost Function and Measures

We obtain contextualised word embeddings from a pre-trained language model as adopted in the previous word alignment methods (Jalili Sabet et al., 2020; Dou and Neubig, 2021). Specifically, we concatenate source and target sentences and input to the language model to obtain the ℓ-th layer hidden outputs. We then compute a word embedding by mean pooling the hidden outputs of its subwords.

As a cost function, we use the cosine and Euclidean distances[4] to obtain the cost matrix $\boldsymbol{C}$, which are the most commonly used semantic dissimilarity metrics in NLP. In addition, we employ the distortion introduced in IBM Model 2 (Brown et al., 1993) when computing a cost to discourage aligning words appearing in distant positions. Jalili Sabet et al. (2020) modelled the distortion between s_i and t_j as $\kappa\,(i/n - j/m)^2$, where κ scales the value to range in $[0, \kappa]$. The value becomes larger if the relative positions of i/n and j/m are largely different. Each entry of the cost matrix is then modulated by the corresponding distortion value. Note that the cost matrix is scaled in its computation process by using the min-max normalisation to make sure all entries lie in $[0, 1]$.[5]

[4] Although the cosine distance is not a proper metric due to its in-satisfaction of the triangle inequality, we adopt it for its prevalence and empirical efficacy in NLP.

[5] The scaling is not mathematically necessary; however, it is crucial for numerical stability in solving OT problems with the Sinkhorn algorithm; without the scaling, parameter tuning of ε becomes challenging as the relative scale of the alignment cost $\langle \boldsymbol{C}, \boldsymbol{P} \rangle$ may deviate a lot.

Fertility determines the likelihood of words having alignment links. We use two standard measures adopted in previous studies that used OT in NLP tasks as discussed in §2.2: the uniform distribution (Kusner et al., 2015) and L^2-norms of word embeddings (Yokoi et al., 2020). On POT and UOT, we directly use these measures without scaling, for which $P_{i,j}$ may have an arbitrary positive value. We scale $\boldsymbol{P}$ by the min-max normalisation so that we can handle alignment matrices of BOT, POT, and UOT by a unified manner.

5.2 Heuristics for Sparse Alignment

While the Sinkhorn algorithm is a powerful tool for solving OT problems, one drawback is that a resultant alignment matrix becomes dense, i.e., each element has a non-zero weight. It is not straightforward to interpret a dense solution as an alignment matrix and thus dense matrices are better avoided. As an empirical remedy, simple heuristics have been commonly used to make alignment matrices sparse: assuming top-k elements based on their mass (Lee et al., 2022; Yu et al., 2022) or elements whose mass are larger than a threshold (Swanson et al., 2020; Dou and Neubig, 2021) are aligned.

We take the latter approach to avoid introducing an arbitrary assumption on fertility, i.e., the number of alignment links that a word can have. Specifically, we derive the final alignment matrix $\hat{\boldsymbol{P}}$ using a threshold λ:

$$\hat{P}_{i,j} = \begin{cases} P_{i,j} & P_{i,j} > \lambda, \\ 0 & \text{otherwise.} \end{cases}$$

Our experiments reveal that this simple 'patch' to obtain a sparse alignment can produce *unbalanced* alignment, rather than just sparse, as we see in §7.2.

5.3 Application to Unsupervised Alignment

We obtain contextualised word embeddings from a pre-trained masked language model without fine-tuning in the unsupervised setting. Such word embeddings are known to show relatively high cosine similarity between any random words (Ethayarajh, 2019), which blurs the actual similarity of semantically corresponding words. Chen et al. (2020a) alleviated this phenomenon with a simple technique of centring the word embedding distribution. We apply the corpus mean centring that subtracts the mean word vector of the entire corpus from each word embedding.

		Train	Val	Test
MSR-RTE		600	200	800
Edinburgh++		514	200	306
MultiMWA	MTRef	2,398	800	800
	Wiki	2,514	533	1,052
	Newsela	–	–	500
	ArXiv	–	–	200

Table 1: Statistics of evaluation datasets

5.4 Application to Supervised Alignment

In the supervised alignment, we adopt linear metric learning (Huang et al., 2016) that learns a matrix $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ defining a generalised distance between two embeddings: $c(\boldsymbol{s}_i, \boldsymbol{t}_j) = c(\boldsymbol{W}\boldsymbol{s}_i, \boldsymbol{W}\boldsymbol{t}_j)$. We train the entire model to learn parameters of $\boldsymbol{W}$ and the pre-trained language model by minimising the binary cross-entropy loss:

$$L(P_{i,j}, Y_{i,j}) = -Y_{i,j} \log P_{i,j} - (1 - Y_{i,j}) \log(1 - P_{i,j}),$$

where $\boldsymbol{P}$ and $\boldsymbol{Y}$ are the predicted and ground-truth alignment matrices, respectively. Specifically, $Y_{i,j} \in \{0, 1\}$ indicates the ground-truth alignment between s_i and t_j; 1 means that alignment exists while 0 means no alignment.

6 Experiment Settings

We empirically investigate the characteristics of BOT, POT, and UOT on word alignment in both unsupervised and supervised settings. We refer to our OT-based alignment methods as OTAlign, hereafter. To alleviate performance variations due to neural model initialisation, all the experiments were conducted for 5 times with different random seeds, and the means of the scores are reported.

Dataset As datasets that provide human-annotated word alignment, we used Microsoft Research Recognizing Textual Entailment (**MSR-RTE**) (Brockett, 2007), **Edinburgh++** (Thadani et al., 2012b), and Multi-Genre Monolingual Word Alignment (**MultiMWA**) (Lan et al., 2021) as shown in Table 1. MultiMWA consists of four subsets according to the sources of sentence pairs: **MTRef**, **Newsela**, **ArXiv**, and **Wiki**. Among them, Newsela and ArXiv are intended for a transfer-learning experiment, on which models trained using MTRef should be tested (Lan et al., 2021). MSR-RTE and Edinburgh++ do not have an official split for validation. Hence, we

subsampled a validation set from the training split, which was excluded from there. As the tradition of word alignment, there are two types of alignment links: sure and possible. The former indicates that an alignment was agreed upon by multiple annotators and thus preserves high confidence. Experiments were conducted in both 'Sure and Possible' and 'Sure Only' settings. For more details, see Appendix A.1.

Pre-trained Language Model OTAlign as well as the recent powerful methods (§2.1) use contextualised word embeddings obtained from pre-trained (masked) language models. As the basic and standard model, we used BERT-base-uncased (Devlin et al., 2019) for all the methods compared to directly observe the capabilities of different alignment mechanisms and to exclude performance differences owing to the underlying pre-trained models.[6] For unsupervised word alignment, we used the 10th layer in BERT that performs strongly in unsupervised textual similarity estimation (Zhang et al., 2020c). For supervised alignment, we used the last hidden layer following the convention.

OTAlign Our OT-based alignment methods require a cost function and fertility. As discussed in §5.1, we experiment with **cosine** and **Euclidean** distances as cost functions and **uniform** distribution and L^2-**norm** of word embeddings as the fertility. Due to the space limitation, the main paper only discusses the results of a cost function and fertility that performed consistently strongly on the validation sets. Appendix B draws the complete picture of different distance metrics and fertility and analyses their effects. We fixed the regularisation penalty ε to 0.1 throughout all experiments. The threshold λ to sparsify alignment matrices was searched in $[0, 1]$ with 0.01 interval to maximise the validation F1. More details of the implementation and experiment environment are described in Appendix A.2.

7 Unsupervised Word Alignment

We reveal features of OT-based methods on monolingual word alignment under unsupervised learning to segregate the effects of supervision on the pre-trained language model. Note that we use the validation sets for hyper-parameter tuning, which may not be strictly 'unsupervised.' We still call this scenario 'unsupervised' for convenience, believing that the preparation of small-scale annotation datasets should be feasible in most practical cases.

7.1 Settings

As a conventional word alignment method, we compared OTAlign to the **fast-align** (Dyer et al., 2013) that implemented IBM Model 2. For fast-align, the training, validation, and test sets were concatenated and used to gather enough statistics for alignment. As the state-of-the-art unsupervised word alignment method, we compared OTAlign to **SimAlign** (Jalili Sabet et al., 2020). While SimAlign was initially proposed for crosslingual word alignment, it is directly applicable to the monolingual setting. SimAlign computes a similarity matrix in the same manner as ours,[7] and aligns words using simple heuristics on the matrix. Specifically, we used the 'IterMax' that performed best across many language pairs. IterMax conducts the 'ArgMax' alignment iteratively (Jalili Sabet et al. (2020) set the iteration as two times), which aligns a word pair if their similarity is the highest among each other. SimAlign has two hyper-parameters: one for the distortion (κ) and another for forcing null alignment if a word is not particularly similar to any target words. These values were tuned using the validation sets.

For OTAlign, POT has a hyper-parameter m to represent a fraction of fertility to be aligned, which we parameterise as $m = \widetilde{m} \min(\|\boldsymbol{a}\|_1, \|\boldsymbol{b}\|_1)$. UOT has marginal deviation penalties of τ_a and τ_b.[8] All of these hyper-parameters were searched in a range of $[0, 1]$ with 0.02 interval to maximise the validation F1. For the distortion strength κ, we applied the same values tuned on SimAlign.

7.2 Results: Primary Observations

Table 2 shows the F1 scores measured on the test sets.[9] We have the following observations: **(i) The best OT problem depends on null alignment ratios.** On datasets with higher null alignment ratios, i.e., Edinburgh++ and MTRef, regularised BOT, regularised POT, and UOT largely outper-

[6] We indeed investigated the performance when adopting SpanBERT (Joshi et al., 2020) as a pre-trained model. In summary, the results indicated that the underlying pre-trained model does not affect the overall trends of alignment methods, while their performance was consistently improved by a few percent owing to the better phrase representation.

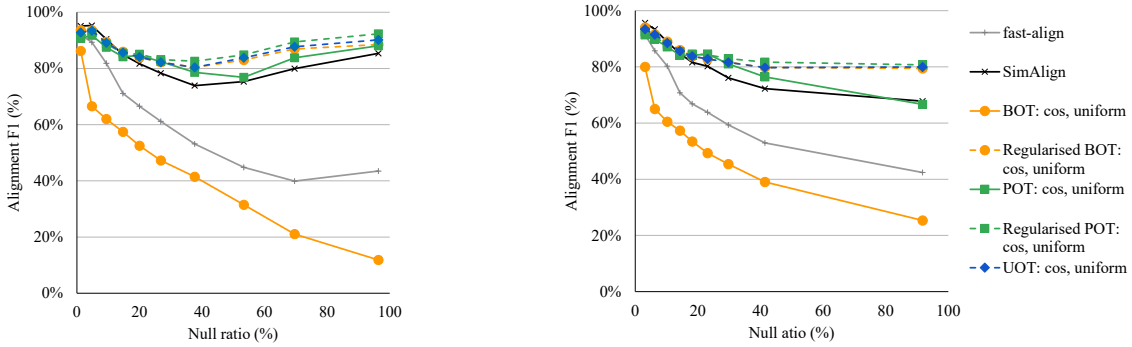
[7] They use cosine *similarity* instead of distance.

[8] We set $\tau_a = \tau_b$ following the common trait to reduce the number of hyper-parameters (Chapel et al., 2021).

[9] The complete sets of results including precision and recall are in Appendix B.

Dataset (sparse ↔ dense)				MSR-RTE		Newsela		EDB++		MTRef		Arxiv		Wiki
Alignment links				S	S + P	S	S + P	S	S + P	S	S + P	S	S + P	S
Null alignment rate (%)				63.8	59.0	33.3	23.5	27.4	19.0	18.7	11.2	12.8	12.2	8.3
fast-align (Dyer et al., 2013)				42.3	41.6	58.4	56.5	59.6	60.8	58.1	58.0	80.5	80.5	87.2
SimAlign (Jalili Sabet et al., 2020)				85.4	81.5	76.7	77.3	74.7	78.9	74.8	75.8	91.7	91.9	94.8
Type	Reg.	cost	mass											
BOT	–	cosine	uniform	20.6	22.5	41.4	46.9	49.0	55.0	50.4	55.5	65.6	66.2	66.5
	Sk	cosine	uniform	88.8	83.0	83.7	79.4	84.4	82.8	77.3	77.2	90.4	90.9	93.9
POT	–	cosine	uniform	89.0	84.0	77.1	76.2	78.4	78.7	75.6	76.2	84.3	89.9	94.5
	Sk	cosine	uniform	92.2	86.4	84.6	79.8	83.8	82.3	77.0	76.6	91.5	90.3	93.9
UOT	Sk	cosine	uniform	90.2	84.5	83.1	79.1	84.7	82.5	77.2	77.1	90.0	89.6	93.8

Table 2: Unsupervised word alignment F1 scores (%) measured on the test sets, where the underlined scores are the best score and those within 1% absolute differences. 'S+P' and 'S' are 'Sure and Possible' and 'Sure Only' settings, respectively. 'Reg.' indicates the regulariser: 'Sk' means the Sinkhorn.



(a) On all datasets

(b) On datasets excluding MSR-RTE

Figure 2: Unsupervised alignment F1 (%) per null ratio

formed SimAlign. In particular, regularised POT performed the best on datasets with significantly high null alignment ratios, i.e., Newsela and MSR-RTE. On the other hand, when null alignment frequency is low, i.e., ArXiv and Wiki, regularised BOT, regularised POT, and UOT performed similarly. On these datasets, SimAlign also performed strongly thanks to the BERT representations.[10]

(ii) Thresholding on the alignment matrix makes it unbalanced. As expected, the unregularised BOT showed the worst performance because its constraint forces to align all words and prohibits null alignment. In contrast, we observe that the performance of unregularised BOT and POT was significantly boosted by regularisation and thresholding on resultant alignment matrices (see the performance differences between with and without regulariser, denoted as '–' and 'Sk', respectively, in Table 2).

For further investigation, we binned the test samples across datasets (under the 'Sure Only' setting) according to their null alignment ratios and observed the performance of the methods.[11] Figure 2 (a) shows the trend of the F1 score on all the datasets according to the null alignment ratios. Although the F1 score of BOT naturally degrades as null alignment rates increase, the F1 score of regularised BOT stays closer to that of UOT owing to thresholding on the regularised solutions. Similarly, the regularised POT outperforms the unregularised POT. Therefore, we argue that thresholding is vital to obtain not only a sparse but also *unbalanced* alignment matrix.

Interestingly, most methods show a v-shape curve in Figure 2 (a), i.e., the F1 score decreases first and then increases again. We identified this trend is due to the characteristics of MSR-RTE. Most sentence pairs with a high ($> 45\%$) null alignment ratio come from MSR-RTE that consists of 'hypothesis' and 'text' of entailment pairs

[10] Table 4 in Appendix B shows that the simplest alignment by thresholding on a cosine similarity matrix using BERT embeddings outperforms SimAlign depending on datasets.

[11] Each bin had roughly the same number of samples.

Dataset (sparse ↔ dense)			MSR-RTE		Newsela		EDB++		MTRef		Arxiv		Wiki
Alignment links			S	S + P	S	S + P	S	S + P	S	S + P	S	S + P	S
Null alignment rate (%)			63.8	59.0	33.3	23.5	27.4	19.0	18.7	11.2	12.8	12.2	8.3
(Lan et al., 2021)			95.1	89.2	86.7	85.3	88.3	87.8	83.4	86.1	95.2	95.0	96.6
(Nagata et al., 2020)			95.0	89.2	79.4	82.4	86.9	87.2	82.9	88.0	89.1	89.5	96.5
Type	cost	mass											
BOT	cosine	norm	94.6	88.4	86.5	84.4	85.7	85.4	82.9	87.3	91.7	93.0	96.5
POT	cosine	norm	94.6	88.4	84.0	81.4	85.5	83.7	82.0	85.2	93.0	92.2	95.5
UOT	cosine	norm	94.8	89.0	86.8	84.7	86.7	86.6	82.9	87.4	92.5	92.8	96.7

Table 3: Supervised word alignment F1 scores (%) measured on the test sets, where the underlined scores are the best score and those within 1% absolute differences.

that tend to have largely different lengths. In these entailment pairs, most words in a (shorter) sentence would align with words in the (longer) pair, while the rest of the words in the pair have to be left unaligned. This characteristic is unique in MSR-RTE, which we conjecture affected the performance. In Figure 2 (b), we removed MSR-RTE and drew trends of the alignment F1 score. The alignment F1 approximately becomes inversely proportional to the null alignment rate as expected.

8 Supervised Word Alignment

In this section, we evaluate the performance of OTAlign in supervised word alignment by comparing it to the state-of-the-art methods.

8.1 Settings

We compared our OT-based alignment methods to the state-of-the-art supervised methods proposed by Lan et al. (2021) and Nagata et al. (2020). The method of Lan et al. (2021) uses a semi-Markov conditional random field combined with neural models to simultaneously conduct word and chunk alignment. Its high computational costs limit a chunk to be at most 3-gram. The method proposed by Nagata et al. (2020) models word alignment as span prediction in the same formulation as the SQuAD (Rajpurkar et al., 2016) style question answering. These previous methods explicitly incorporate chunks in the alignment process. In contrast, our OT-based alignment is purely based on word-level distances represented as a cost matrix.

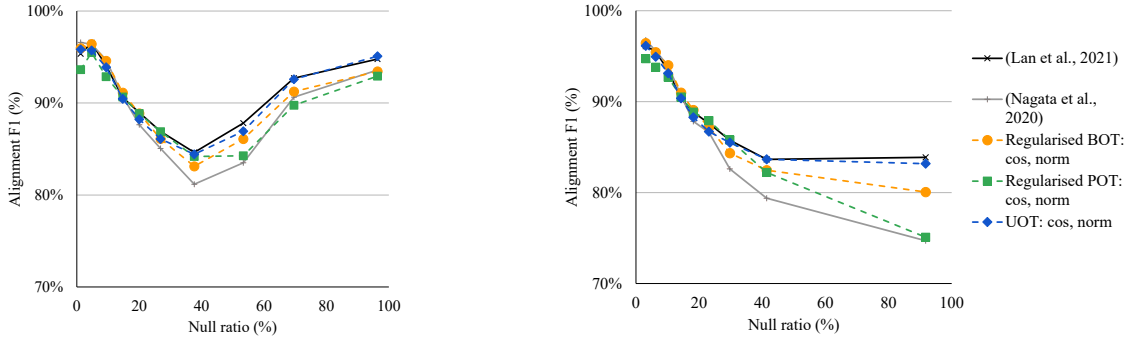
We trained alignment methods of the regularised BOT, regularised POT, and UOT in an end-to-end fashion using the Adafactor (Shazeer and Stern, 2018) optimiser. The batch size was set to 64 for datasets except for Wiki, for which 32 was used due to longer sentence lengths. The training was stopped early, with 3 patience epochs and a minimum delta of 1.0×10^{-4} based on the validation F1 score. Before evaluation, we searched the learning rate from $[5.0 \times 10^{-5}, 2.5 \times 10^{-4}]$ with a 2.0×10^{-5} interval using the validation sets. Other hyper-parameters were set to the same as the unsupervised settings.

8.2 Results

Table 3 shows the alignment F1 (%) measured on the test sets.[12] The UOT-based alignment exhibits competitive performance against these state-of-the-art methods on the datasets with higher null alignment ratios. Notably, despite the simple and generic alignment mechanism based on UOT, it performed on par with Lan et al. (2021) who specially designed the model for monolingual word alignment. The UOT-based alignment also shows consistently higher alignment F1 scores compared to the regularised BOT and POT alignment. These results confirm that UOT well-captures the unbalanced word alignment problem. In addition, OT-based alignment showed better transferability as well as Lan et al. (2021) than Nagata et al. (2020) as demonstrated on results of Newsela and Arxiv. We conjecture this is because our cost matrix has less inductive bias due to its simplicity. In contrast, Nagata et al. (2020) directly learn to predict alignment spans.

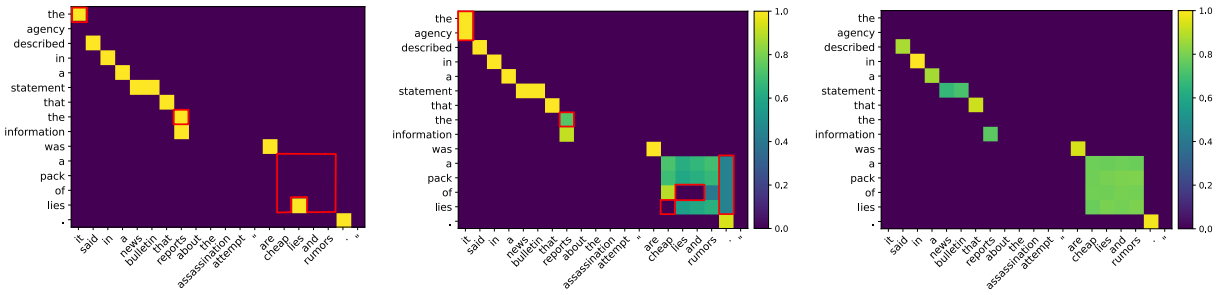
Figure 3 shows the trend of the F1 score according to the null alignment ratios assembled in the same way as Figure 2. Figure 3 (a) shows the trend on all datasets, demonstrating the robustness of OT-based methods against null alignment. OT-based alignment outperformed Nagata et al. (2020) at sen-

[12] The complete sets of results are available in Appendix B.



(a) On all datasets (b) On datasets excluding MSR-RTE

Figure 3: Supervised alignment F1 (%) per null ratio



(a) Alignment by Lan et al. (2021) (b) Alignment by Nagata et al. (2020) (c) Alignment by supervised UOT

Figure 4: Visualization of alignment matrix: red bounding boxes indicate alignment errors. UOT successfully identified many-to-many and null alignment.

tences whose null alignment ratios are higher than 15%. The performances of the regularised BOT and UOT reverse around 20% null alignment ratio; the regularised BOT outperformed UOT on a lower side (0.7% higher F1 on average except for the lowest edge) and UOT outperformed BOT on a higher side (1.0% higher F1 on average). Same with the unsupervised word alignment, all methods show the v-shape trend. Figure 3 (b) shows the trend of the alignment F1 on the datasets except MSR-RTE. The performance of all methods becomes inversely proportional to the null alignment ratio as observed in the unsupervised case.

While UOT and previous methods are competitive on the datasets with high null alignment frequencies, we found that UOT has an advantage in longer phrase alignment. Figure 4 (a), (b), and (c) visualise alignment matrices by Lan et al. (2021), Nagata et al. (2020), and UOT, respectively, of the same sentence pair in Figure 1. The chunk size constraint in Lan et al. (2021), 3-gram at maximum, hindered the many-to-many alignment. Nagata et al. (2020) also failed to complete this alignment because their method conducts one-to-many alignment bidirectionally and merges results. In contrast, UOT can align such a long phrase pair if the cost matrix is sufficiently reliable as demonstrated here.

9 Summary and Future Work

This study revealed the features of BOT, POT, and UOT on unbalanced word alignment and suggested that they perform sufficiently powerfully without tailor-made designs. In future, our OT-based methods can be enhanced for better phrase alignment by employing sophisticated phrase embedding models as discussed in Limitations section. We will also apply OT-based alignment to problems related yet with different constraints and objectives, e.g., crosslingual word alignment and text matching.

Limitations

In this study, we used standard and basic word embeddings to highlight the characteristics of the different OT problems on unbalanced word alignment. This limits the capability of phrasal alignment. Similar to Figure 3 (a), we binned all the test samples across datasets (under the 'Sure Only' setting) according to their phrase alignment ratios and evaluated the performance of the supervised

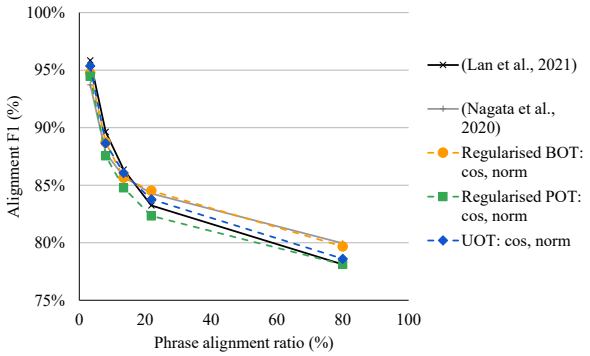


Figure 5: Supervised alignment F1 (%) per phrase alignment ratios

OT-based alignment methods.[13] Specifically, we regarded one-to-many, many-to-one, and many-to-many alignment as phrase alignment. Figure 5 shows the trend of the F1 score according to the phrase alignment ratios. Obviously, the F1 score degrades as more phrase alignment exists in a sentence pair.

One of the straightforward ways to improve the phrase alignment is exploring pre-trained language models enhanced for span representations (Joshi et al., 2020) and sophisticated methods for phrase representation composition (Yu and Ettinger, 2020). In addition, phrase alignment can be addressed from the OT perspective, too, by conducting structure-aware (Alvarez-Melis et al., 2018) and order-aware (Liu et al., 2018) optimal transport. These directions constitute our future work.

Acknowledgements

We sincerely appreciate the anonymous reviewers for their insightful comments and suggestions to improve the paper. This work was partially supported by JSPS KAKENHI Grant Number 22H03654.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pages 252–263.

Sawsan Alqahtani, Garima Lalwani, Yi Zhang, Salvatore Romeo, and Saab Mansour. 2021. Using optimal transport as alignment objective for fine-tuning multilingual contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 3904–3919.

David Alvarez-Melis, Tommi Jaakkola, and Stefanie Jegelka. 2018. Structured optimal transport. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84 of *Proceedings of Machine Learning Research*, pages 1771–1780.

Yuki Arase and Jun'ichi Tsujii. 2020. Compositional phrase alignment and beyond. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1611–1623.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. 2015. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138.

Chris Brockett. 2007. Aligning the RTE 2006 corpus. Technical Report MSR-TR-2007-77, Microsoft Research.

Daniela Brook Weiss, Paul Roit, Ayal Klein, Ori Ernst, and Ido Dagan. 2021. QA-align: Representing cross-text content overlap by aligning question-answer propositions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9879–9894.

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

Luis Caffarelli and Robert J. McCann. 2010. Free boundaries in optimal transport and monge-ampère obstacle problems. *Annals of Mathematics*, 171(2):673–730.

Laetitia Chapel, Rémi Flamary, Haoran Wu, Cédric Févotte, and Gilles Gasso. 2021. Unbalanced optimal transport through non-negative penalized linear regression. In *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*, pages 23270–23282.

Liqun Chen, Yizhe Zhang, Ruiyi Zhang, Chenyang Tao, Zhe Gan, Haichao Zhang, Bai Li, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019. Improving sequence-to-sequence learning via optimal transport. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[13] In the evaluation datasets, phrase alignment is infrequent, which skewed the bin sizes: the first bin (less than 3.3% phrase alignment ratio) size was 2.2k and the rests were $\approx$ 350.

Xi Chen, Nan Ding, Tomer Levinboim, and Radu Soricut. 2020a. Improving text generation evaluation with batch centering and tempered word mover distance. In *Proceedings of the Workshop on Evaluation and Comparison of NLP Systems*, pages 51–59.

Yimeng Chen, Yanyan Lan, Ruinbin Xiong, Liang Pang, Zhiming Ma, and Xueqi Cheng. 2020b. Evaluating natural language generation via unbalanced optimal transport. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3730–3736.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. 2016. Scaling algorithms for unbalanced transport problems. *arXiv*, 1607.05816.

Ryan Culkin, J. Edward Hu, Elias Stengel-Eskin, Guanghui Qin, and Benjamin Van Durme. 2021. Iterative paraphrastic augmentation with discriminative span alignment. *Transactions of the Association of Computational Linguistics (TACL)*, 9:494–509.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*, volume 26.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190.

Dipanjan Das and Noah A. Smith. 2009. Paraphrase identification as probabilistic quasi-synchronous recognition. In *Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 468–476.

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 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2112–2128.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 644–648.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65.

Yair Feldman and Ran El-Yaniv. 2019. Multi-hop paragraph retrieval for open-domain question answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2296–2309.

Alessio Figalli. 2010. The optimal partial transport problem. *Archive for Rational Mechanics and Analysis*, 195(2):533–560.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. 2021. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. 2015. Learning with a Wasserstein loss. In *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*, volume 28.

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

Edouard Grave, Armand Joulin, and Quentin Berthet. 2019. Unsupervised alignment of embeddings with Wasserstein procrustes. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1880–1890.

Michael Heilman and Noah A. Smith. 2010. Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1011–1019.

Eran Hirsch, Alon Eirew, Ori Shapira, Avi Caciularu, Arie Cattan, Ori Ernst, Ramakanth Pasunuru, Hadar Ronen, Mohit Bansal, and Ido Dagan. 2021. iFacetSum: Coreference-based interactive faceted summarization for multi-document exploration. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 283–297.

Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha, and Kilian Q Weinberger. 2016. Supervised word mover's distance. In *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*, volume 29.

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*, pages 1627–1643.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7943–7960.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association of Computational Linguistics (TACL)*, 8:64–77.

L. V. Kantorovich. 1942. On the translocation of masses. *Doklady Akademii Nauk*, 37(2):227–229.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 957–966.

Wuwei Lan, Chao Jiang, and Wei Xu. 2021. Neural semi-Markov CRF for monolingual word alignment. In *Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 6815–6828.

Seonghyeon Lee, Dongha Lee, Seongbo Jang, and Hwanjo Yu. 2022. Toward interpretable semantic textual similarity via optimal transport-based contrastive sentence learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5969–5979.

Jianqiao Li, Chunyuan Li, Guoyin Wang, Hao Fu, Yuhchen Lin, Liqun Chen, Yizhe Zhang, Chenyang Tao, Ruiyi Zhang, Wenlin Wang, Dinghan Shen, Qian Yang, and Lawrence Carin. 2020. Improving text generation with student-forcing optimal transport. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9144–9156.

Tao Li and Vivek Srikumar. 2016. Exploiting sentence similarities for better alignments. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2193–2203.

Bang Liu, Ting Zhang, Fred X. Han, Di Niu, Kunfeng Lai, and Yu Xu. 2018. Matching natural language sentences with hierarchical sentence factorization. In *Proceedings of the World Wide Web Conference (WWW)*, page 1237–1246.

Bill MacCartney, Michel Galley, and Christopher D. Manning. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 802–811.

Gaspard Monge. 1781. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Sheshera Mysore, Arman Cohan, and Tom Hope. 2022. Multi-vector models with textual guidance for fine-grained scientific document similarity. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4453–4470.

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

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

Jessica Ouyang and Kathy McKeown. 2019. Neural network alignment for sentential paraphrases. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4724–4735.

Ofir Pele and Michael Werman. 2009. Fast and robust Earth Mover's Distances. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 460–467.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392.

Ori Shapira, Hadar Ronen, Meni Adler, Yael Amsterdamer, Judit Bar-Ilan, and Ido Dagan. 2017. Interactive abstractive summarization for event news tweets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 109–114.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 4596–4604.

Richard Sinkhorn. 1964. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Annals of Mathematical Statistics*, 35:876–879.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014. Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *Transactions of the Association of Computational Linguistics (TACL)*, 2:219–230.

Kyle Swanson, Lili Yu, and Tao Lei. 2020. Rationalizing text matching: Learning sparse alignments via optimal transport. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5609–5626.

Peggy Tang, Kun Hu, Rui Yan, Lei Zhang, Junbin Gao, and Zhiyong Wang. 2022. OTExtSum: Extractive Text Summarisation with Optimal Transport. In *Findings of the Association for Computational Linguistics: NAACL (Findings-NAACL)*, pages 1128–1141.

Kapil Thadani, Scott Martin, and Michael White. 2012a. A joint phrasal and dependency model for paraphrase alignment. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 1229–1238.

Kapil Thadani, Scott Martin, and Michael White. 2012b. A joint phrasal and dependency model for paraphrase alignment. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 1229–1238.

Kapil Thadani and Kathleen McKeown. 2011. Optimal and syntactically-informed decoding for monolingual phrase-based alignment. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 254–259.

Kapil Thadani and Kathleen McKeown. 2013. Supervised sentence fusion with single-stage inference. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, pages 1410–1418.

Xiaojun Wan. 2007. A novel document similarity measure based on earth mover's distance. *Information Sciences*, 177(18):3718–3730.

Mengqiu Wang and Christopher Manning. 2010. Probabilistic tree-edit models with structured latent variables for textual entailment and question answering. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 1164–1172.

Zihao Wang, Jiaheng Dou, and Yong Zhang. 2022. Unsupervised sentence textual similarity with compositional phrase semantics. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 4976–4995.

Zihao Wang, Datong Zhou, Ming Yang, Yong Zhang, Chenglong Rao, and Hao Wu. 2020. Robust document distance with Wasserstein-Fisher-Rao metric. In *Proceedings of the Asian Conference on Machine Learning*, volume 129, pages 721–736.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Xuchen Yao. 2014. *Feature-driven Question Answering With Natural Language Alignment*. Ph.D. thesis, Johns Hopkins University.

Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013a. A lightweight and high performance monolingual word aligner. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 702–707.

Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013b. Semi-Markov phrase-based monolingual alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 590–600.

Sho Yokoi, Ryo Takahashi, Reina Akama, Jun Suzuki, and Kentaro Inui. 2020. Word rotator's distance. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2944–2960.

Lang Yu and Allyson Ettinger. 2020. Assessing phrasal representation and composition in transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4907.

Weijie Yu, Liang Pang, Jun Xu, Bing Su, Zhenhua Dong, and Ji-Rong Wen. 2022. Optimal partial transport based sentence selection for long-form document matching. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 2363–2373.

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

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1934–1945.

Ruiyi Zhang, Changyou Chen, Xinyuan Zhang, Ke Bai, and Lawrence Carin. 2020a. Semantic matching for sequence-to-sequence learning. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 212–222.

Shuying Zhang, Tianyu Zhao, and Tatsuya Kawahara. 2020b. Topic-relevant response generation using optimal transport for an open-domain dialog system. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 4067–4077.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020c. BERTScore: Evaluating text generation with bert. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578.

Xu Zhao, Zihao Wang, Yong Zhang, and Hao Wu. 2020. A relaxed matching procedure for unsupervised BLI. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3036–3041.

A Details of Evaluation Settings

A.1 Evaluation Datasets

MSR-RTE and Edinburgh++ have been commonly used for evaluating monolingual word alignment models (Yao et al., 2013b; Sultan et al., 2014). MultiMWA is the newest dataset that provides a larger number of examples with improved inter-annotator agreements.

MSR-RTE annotated word alignment on sentence pairs for the 2006 PASCAL RTE challenge (Dagan et al., 2006; Bar-Haim et al., 2006). Due to the nature of RTE pairs, the lengths of the source and target sentences tend to diverge, which results in a higher null-alignment ratio. In contrast, Edinburgh++ annotates word alignment on paraphrase pairs collected from machine translation (MT) references, novels, and news domains. MultiMWA annotated (or corrected the original) word alignment on (1) paraphrases extracted from multiple human references for MT evaluation (Yao, 2014) (MTRef), (2) simple and complex sentence pairs sampled from Newsela-Auto (Jiang et al., 2020) constructed for text simplification (Newsela), (3) aligned sentences extracted from arXiv[14] revision history (arXiv), and (4) aligned Wikipedia[15] sentences based on the edit histories (Wiki).

[14] https://arxiv.org/

[15] https://en.wikipedia.org/

A.2 Implementation and Environment

All the BOT-, POT-, and UOT-based word alignment methods were implemented using PyTorch,[16] PyTorch Lightning,[17] Hugging Face Transformers (Wolf et al., 2020), and Python Optimal Transport (Flamary et al., 2021) libraries.

For the previous studies compared, we used the public implementations released by the authors with the necessary modifications. All the experiments were conducted on an NVIDIA Tesla V100 when GPU computation is needed.

B Additional Experimental Results

Tables 4 to 6 show the F1, precision, and recall scores of unsupervised word alignment with all the cost functions (cosine and Euclidean distances) and measures (the uniform distribution and L^2-norms). Tables 7 to 9 indicate those of supervised word alignment experiments.

B.1 Effects of Cost Function and Measures

Unsupervised Alignment The 7th and 8th rows from the top of Table 4 show the F1 scores of naive baselines that determine word alignment on the cost matrix with simple thresholding. That is, these baselines align words whose costs are smaller than the threshold. Their results indicate that cosine distance consistently outperforms Euclidean distance when word embeddings are naively used to estimate semantic similarity.

The unregularised BOT and POT also prefer cosine distance as the cost function but their performances are more sensitive to measures, i.e., the uniform distribution outperformed the L^2-norms. The effect of the measures is most pronounced on the regularised BOT, likely because the uniform fertility makes the threshold selection simple. In contrast, UOT and regularised POT are less sensitive to the measures, likely because of their internal mechanism to allow null alignment.

Supervised Alignment As shown in Table 7, the UOT significantly improved over the unsupervised setting, outperforming the POT. In the supervised setting, UOT prefers the L^2-norms to model fertility than the uniform distribution. We conjecture that the BERT with linear metric learning successfully learns to adapt norms of word embeddings to well represent the fertility of words.

[16] https://pytorch.org/

[17] https://www.pytorchlightning.ai/

Dataset (sparse ↔ dense)				MSR-RTE		Newsela		EDB++		MTRef		Arxiv		Wiki
Alignment links				S	S + P	S	S + P	S	S + P	S	S + P	S	S + P	S
Null alignment rate (%)				63.8	59.0	33.3	23.5	27.4	19.0	18.7	11.2	12.8	12.2	8.3
fast-align (Dyer et al., 2013)				42.3	41.6	58.4	56.5	59.6	60.8	58.1	58.0	80.5	80.5	87.2
SimAlign (Jalili Sabet et al., 2020)				85.4	81.5	76.7	77.3	74.7	78.9	74.8	75.8	91.7	91.9	94.8
Type	Reg.	cost	mass											
–	–	cosine	–	88.9	83.2	81.3	74.0	81.5	76.0	71.7	67.0	83.3	82.2	87.4
–	–	L2	–	87.3	81.4	79.3	71.3	78.9	73.1	69.2	64.7	75.9	71.9	86.1
BOT	–	cosine	norm	20.4	22.3	40.2	45.6	45.0	50.6	50.3	55.2	56.6	57.1	60.3
	–	cosine	uniform	20.6	22.5	41.4	46.9	49.0	55.0	50.4	55.5	65.6	66.2	66.5
	–	L2	norm	20.4	22.3	40.2	45.6	45.0	50.6	50.2	55.2	56.7	57.0	60.3
	–	L2	uniform	20.6	22.5	41.4	46.8	49.1	54.9	50.4	55.4	65.7	66.2	66.4
	Sk	cosine	norm	86.0	80.6	82.3	79.3	81.5	82.0	76.1	77.1	86.6	90.1	91.1
	Sk	cosine	uniform	88.8	83.0	83.7	79.4	84.4	82.8	77.3	77.2	90.4	90.9	93.9
	Sk	L2	norm	87.9	82.5	83.4	78.9	82.3	81.7	76.7	76.4	87.2	89.9	91.6
	Sk	L2	uniform	89.8	83.8	84.8	79.4	84.9	82.1	77.0	76.6	90.7	90.1	94.0
POT	–	cosine	norm	86.6	81.5	74.1	71.7	75.8	74.3	73.0	72.1	77.2	82.9	89.5
	–	cosine	uniform	89.0	84.0	77.1	76.2	78.4	78.7	75.6	76.2	84.3	89.9	94.5
	–	L2	norm	86.0	80.9	73.5	71.1	75.3	74.4	72.3	71.4	78.5	83.6	89.8
	–	L2	uniform	88.7	83.6	75.8	75.5	78.6	78.3	75.4	75.6	87.2	89.9	94.5
	Sk	cosine	norm	91.3	86.2	83.5	81.2	83.2	82.2	77.0	76.7	89.3	91.6	92.6
	Sk	cosine	uniform	92.2	86.4	84.6	79.8	83.8	82.3	77.0	76.6	91.5	90.3	93.9
	Sk	L2	norm	92.0	86.1	85.2	79.5	83.2	81.2	76.4	75.8	89.3	91.0	92.9
	Sk	L2	uniform	91.3	85.6	84.6	78.9	85.1	81.7	76.1	75.4	91.1	89.6	93.7
UOT	Sk	cosine	norm	90.6	84.7	84.3	79.9	83.7	82.5	77.0	76.8	88.3	90.7	92.6
	Sk	cosine	uniform	90.2	84.5	83.1	79.1	84.7	82.5	77.2	77.1	90.0	89.6	93.8
	Sk	L2	norm	90.4	84.3	84.3	79.3	84.3	82.1	77.1	76.5	89.7	90.3	93.2
	Sk	L2	uniform	90.0	84.2	83.9	79.2	84.5	81.6	77.0	76.4	90.0	88.6	93.5

Table 4: Unsupervised word alignment F1 scores (%) measured on the test sets, where the underlined scores are the best score and those within 1% absolute differences. 'S+P' and 'S' are 'Sure and Possible' and 'Sure Only' settings, respectively. 'Reg.' indicates the regulariser: 'Sk' means the Sinkhorn.

Dataset (sparse ↔ dense)				MSR-RTE		Newsela		EDB++		MTRef		Arxiv		Wiki
Alignment links				S	S + P	S	S + P	S	S + P	S	S + P	S	S + P	S
Null alignment rate (%)				63.8	59.0	33.3	23.5	27.4	19.0	18.7	11.2	12.8	12.2	8.3
fast-align (Dyer et al., 2013)				43.3	43.0	59.8	57.0	60.8	60.9	57.9	57.2	79.7	79.6	86.5
SimAlign (Jalili Sabet et al., 2020)				85.7	82.6	80.0	79.6	77.6	81.4	77.3	77.6	92.3	92.5	95.4
Type	Reg.	cost	mass											
–	–	cosine	–	87.7	82.9	78.4	70.1	79.2	71.7	68.5	62.9	76.6	74.9	82.4
–	–	L2	–	85.4	80.5	75.5	65.9	75.4	67.6	65.0	59.3	65.2	59.6	80.4
BOT	–	cosine	norm	18.4	20.5	33.2	37.5	36.9	41.0	41.4	45.2	44.1	44.4	46.9
	–	cosine	uniform	18.7	20.8	34.9	39.2	42.6	47.1	41.6	45.5	56.2	56.7	54.8
	–	L2	norm	18.4	20.5	33.2	37.4	36.9	41.0	41.4	45.2	44.1	44.4	46.8
	–	L2	uniform	18.7	20.8	34.8	39.2	42.7	47.0	41.6	45.5	56.2	56.6	54.7
	Sk	cosine	norm	85.2	81.2	82.9	80.1	83.5	83.4	76.7	77.8	84.8	89.4	90.2
	Sk	cosine	uniform	88.2	83.3	84.6	80.2	86.0	84.2	78.1	77.9	89.3	90.2	93.6
	Sk	L2	norm	87.2	82.7	83.8	79.3	83.7	82.4	77.1	76.6	85.4	89.0	90.6
	Sk	L2	uniform	89.1	84.0	85.2	79.6	85.9	82.9	77.1	77.0	89.6	88.8	93.5
POT	–	cosine	norm	85.5	82.0	73.4	70.2	75.1	72.7	72.0	70.8	72.8	79.7	86.8
	–	cosine	uniform	88.5	85.3	79.8	79.9	80.3	81.9	76.8	78.6	82.0	89.8	95.5
	–	L2	norm	84.4	80.8	72.6	69.3	74.5	72.8	71.1	69.7	74.2	80.2	87.1
	–	L2	uniform	88.2	84.9	79.3	79.0	80.1	81.4	77.5	77.9	86.1	89.8	95.5
	Sk	cosine	norm	90.5	87.1	84.4	81.9	84.8	83.4	77.3	76.8	87.4	90.5	91.3
	Sk	cosine	uniform	91.3	86.9	85.2	79.9	86.0	83.3	76.9	76.4	90.2	88.0	93.1
	Sk	L2	norm	91.3	86.7	85.3	79.4	84.8	81.8	76.0	75.2	87.4	89.2	91.5
	Sk	L2	uniform	90.2	85.7	84.6	78.2	86.2	82.0	75.6	74.6	89.2	86.6	92.7
UOT	Sk	cosine	norm	90.0	85.3	84.8	80.7	85.4	83.7	77.3	77.4	86.6	90.0	91.4
	Sk	cosine	uniform	89.8	84.9	83.6	79.5	85.8	83.4	77.7	77.5	88.5	88.0	93.2
	Sk	L2	norm	89.6	85.0	85.0	79.6	85.2	82.8	77.5	76.7	88.2	89.1	92.2
	Sk	L2	uniform	89.2	84.4	83.8	78.9	85.5	82.0	76.8	76.3	88.1	86.1	92.6

Table 5: Unsupervised word alignment precisions (%) measured on the test sets, where the underlined scores are the best and those within 1% absolute differences. 'S+P' and 'S' are 'Sure and Possible' and 'Sure Only' settings, respectively. 'Reg.' indicates the OT regulariser: 'Sk' means the Sinkhorn.

Dataset (sparse ↔ dense)				MSR-RTE		Newsela		EDB++		MTRef		Arxiv		Wiki
Alignment links				S	S + P	S	S + P	S	S + P	S	S + P	S	S + P	S
Null alignment rate (%)				63.8	59.0	33.3	23.5	27.4	19.0	18.7	11.2	12.8	12.2	8.3
fast-align (Dyer et al., 2013)				41.4	40.2	57.1	56.0	58.3	60.7	58.3	58.7	81.2	81.3	87.9
SimAlign (Jalili Sabet et al., 2020)				85.2	80.5	73.6	75.0	72.0	76.5	72.4	74.1	91.1	91.4	94.2
Type	Reg.	cost	mass											
–	–	cosine	–	90.2	83.5	84.4	78.5	83.8	80.9	75.3	71.7	91.4	91.1	93.0
–	–	L2	–	89.2	82.4	83.5	77.8	82.8	79.6	73.9	71.1	90.7	90.5	92.7
BOT	–	cosine	norm	22.9	24.5	50.9	58.3	57.7	66.2	63.9	70.9	79.0	79.8	84.5
	–	cosine	uniform	22.9	24.5	51.0	58.4	57.7	66.1	63.9	71.0	78.9	79.7	84.5
	–	L2	norm	22.9	24.5	50.9	58.3	57.7	66.2	63.9	70.9	79.1	79.8	84.5
	–	L2	uniform	22.9	24.5	50.9	58.2	57.8	66.1	63.8	71.0	79.0	79.6	84.5
	Sk	cosine	norm	86.8	80.1	81.8	78.6	79.5	80.7	75.4	76.4	88.5	90.8	92.1
	Sk	cosine	uniform	89.4	82.6	82.9	78.6	82.8	81.5	76.6	76.4	91.5	91.6	94.2
	Sk	L2	norm	88.5	82.3	83.1	78.4	80.9	80.9	76.3	76.2	89.2	90.9	92.5
	Sk	L2	uniform	90.4	83.6	84.3	79.3	83.8	81.2	76.9	76.3	91.8	91.4	94.4
POT	–	cosine	norm	87.6	81.0	74.7	73.2	76.4	76.1	73.9	73.4	82.2	86.4	92.4
	–	cosine	uniform	89.5	82.7	74.5	72.8	76.6	75.8	74.5	74.0	86.7	89.9	93.6
	–	L2	norm	87.7	81.1	74.4	72.9	76.1	76.0	73.5	73.1	83.3	87.2	92.7
	–	L2	uniform	89.3	82.4	72.5	72.3	77.1	75.5	73.4	73.4	88.4	89.9	93.6
	Sk	cosine	norm	92.1	85.4	82.6	80.5	81.6	81.0	76.7	76.7	91.4	92.8	93.8
	Sk	cosine	uniform	93.1	85.9	84.0	79.7	81.7	81.4	77.0	76.7	92.9	92.7	94.7
	Sk	L2	norm	92.8	85.6	85.1	79.5	81.6	80.6	76.9	76.4	91.3	92.8	94.4
	Sk	L2	uniform	92.6	85.5	84.6	79.6	84.1	81.4	76.7	76.3	93.1	92.8	94.6
UOT	Sk	cosine	norm	91.2	84.2	83.8	79.2	82.0	81.5	76.6	76.3	90.1	91.4	93.7
	Sk	cosine	uniform	90.7	84.1	82.5	78.7	83.6	81.7	76.7	76.6	91.6	91.1	94.4
	Sk	L2	norm	91.1	83.7	83.7	79.0	83.3	81.4	76.8	76.3	91.2	91.4	94.2
	Sk	L2	uniform	90.9	84.0	84.0	79.5	83.6	81.2	77.1	76.4	92.1	91.2	94.4

Table 6: Unsupervised word alignment recalls (%) measured on the test sets, where the underlined scores are the best and those within 1% absolute differences. 'S+P' and 'S' are 'Sure and Possible' and 'Sure Only' settings, respectively. 'Reg.' indicates the OT regulariser: 'Sk' means the Sinkhorn.

Dataset (sparse ↔ dense)			MSR-RTE		Newsela		EDB++		MTRef		Arxiv		Wiki
Alignment links			S	S + P	S	S + P	S	S + P	S	S + P	S	S + P	S
Null alignment rate (%)			63.8	59.0	33.3	23.5	27.4	19.0	18.7	11.2	12.8	12.2	8.3
(Lan et al., 2021)			95.1	89.2	86.7	85.3	88.3	87.8	83.4	86.1	95.2	95.0	96.6
(Nagata et al., 2020)			95.0	89.2	79.4	82.4	86.9	87.2	82.9	88.0	89.1	89.5	96.5
Type	cost	mass											
BOT	cosine	norm	94.6	88.4	86.5	84.4	85.7	85.4	82.9	87.3	91.7	93.0	96.5
	cosine	uniform	94.5	88.1	85.9	82.0	85.9	85.1	80.3	84.8	94.2	92.6	96.4
	L2	norm	94.7	88.7	86.4	84.4	85.9	85.8	82.5	87.0	92.2	92.6	96.6
	L2	uniform	94.4	88.0	86.1	83.3	85.9	85.3	81.2	85.1	94.3	93.3	96.6
POT	cosine	norm	94.6	88.4	84.0	81.4	85.5	83.7	82.0	85.2	93.0	92.2	95.5
	cosine	uniform	94.2	88.3	84.3	82.7	85.9	85.1	82.4	86.7	93.5	92.8	96.4
	L2	norm	94.3	88.4	82.8	80.1	85.2	84.5	81.5	85.7	90.6	90.6	95.5
	L2	uniform	94.1	88.3	84.9	83.8	85.8	85.6	82.5	86.9	94.1	93.4	96.6
UOT	cosine	norm	94.8	89.0	86.8	84.7	86.7	86.6	82.9	87.4	92.5	92.8	96.7
	cosine	uniform	95.0	88.9	85.1	83.5	86.5	86.0	81.8	86.8	93.9	92.9	96.7
	L2	norm	94.8	89.1	87.3	85.1	86.6	86.6	82.4	87.3	93.4	93.6	96.8
	L2	uniform	94.6	89.0	86.6	84.5	86.3	86.6	82.0	86.8	94.4	93.6	96.8

Table 7: Supervised word alignment F1 scores (%) measured on the test sets, where the underlined scores are the best score and those within 1% absolute differences.

Dataset (sparse ↔ dense)			MSR-RTE		Newsela		EDB++		MTRef		Arxiv		Wiki
Alignment links			S	S + P	S	S + P	S	S + P	S	S + P	S	S + P	S
Null alignment rate (%)			63.8	59.0	33.3	23.5	27.4	19.0	18.7	11.2	12.8	12.2	8.3
(Lan et al., 2021)			95.0	89.9	88.1	85.7	89.7	88.4	84.1	86.0	95.4	95.1	96.8
(Nagata et al., 2020)			95.0	90.2	80.3	82.2	88.5	88.1	83.3	87.4	87.8	88.0	96.7
Type	cost	mass											
BOT	cosine	norm	94.2	89.4	88.4	85.8	87.5	87.6	84.4	87.9	91.0	93.0	96.8
	cosine	uniform	94.3	89.2	88.1	84.3	87.6	87.5	82.0	86.2	94.5	93.0	96.7
	L2	norm	94.5	89.7	88.6	86.1	87.6	87.9	84.3	88.0	91.8	92.6	96.9
	L2	uniform	94.3	89.0	88.4	85.4	87.8	87.7	83.1	86.6	94.6	93.7	96.9
POT	cosine	norm	94.2	89.1	85.4	81.1	86.2	84.0	82.2	84.1	91.8	91.1	95.5
	cosine	uniform	93.7	89.0	86.5	84.3	87.4	86.8	83.7	87.1	92.9	92.3	96.7
	L2	norm	93.8	89.1	84.0	80.0	86.1	85.0	82.0	85.0	88.9	88.9	95.5
	L2	uniform	93.6	89.0	87.1	85.2	87.2	87.2	83.8	87.1	93.7	92.9	96.9
UOT	cosine	norm	94.7	89.9	88.8	86.1	88.4	88.5	84.4	87.9	92.0	92.7	97.1
	cosine	uniform	94.8	89.8	87.7	85.6	88.5	88.2	83.7	88.1	94.3	93.3	97.1
	L2	norm	94.5	90.2	89.2	86.7	88.5	88.5	84.0	88.1	93.1	93.8	97.1
	L2	uniform	94.6	90.0	88.8	86.4	88.3	88.6	83.7	88.2	94.6	93.9	97.1

Table 8: Supervised word alignment precision (%) measured on the test sets, where the underlined scores are the best and those within 1% absolute differences.

Dataset (sparse ↔ dense)			MSR-RTE		Newsela		EDB++		MTRef		Arxiv		Wiki
Alignment links			S	S + P	S	S + P	S	S + P	S	S + P	S	S + P	S
Null alignment rate (%)			63.8	59.0	33.3	23.5	27.4	19.0	18.7	11.2	12.8	12.2	8.3
(Lan et al., 2021)			95.2	88.5	85.4	84.9	87.0	87.1	82.6	86.2	95.0	94.9	96.4
(Nagata et al., 2020)			95.0	88.3	78.5	82.6	85.4	86.3	82.5	88.6	90.5	91.0	96.3
Type	cost	mass											
BOT	cosine	norm	95.0	87.5	84.7	83.0	84.0	83.2	81.4	86.7	92.3	93.0	96.2
	cosine	uniform	94.7	87.0	83.8	79.9	84.2	82.9	78.6	83.5	93.9	92.1	96.1
	L2	norm	95.0	87.6	84.4	82.7	84.3	83.7	80.7	86.0	92.7	92.7	96.3
	L2	uniform	94.4	86.9	83.9	81.4	84.1	83.1	79.4	83.6	93.9	92.9	96.3
POT	cosine	norm	94.9	87.6	82.7	81.6	84.9	83.5	81.7	86.3	94.2	93.4	95.5
	cosine	uniform	94.7	87.6	82.1	81.2	84.5	83.4	81.2	86.4	94.1	93.3	96.1
	L2	norm	94.7	87.8	81.6	80.2	84.3	84.0	81.0	86.3	92.3	92.5	95.4
	L2	uniform	94.6	87.5	82.8	82.5	84.5	84.0	81.1	86.6	94.4	94.0	96.2
UOT	cosine	norm	95.0	88.1	84.9	83.4	85.0	84.7	81.5	86.9	93.0	93.0	96.4
	cosine	uniform	95.2	88.1	82.7	81.5	84.7	84.0	79.9	85.5	93.5	92.5	96.3
	L2	norm	95.1	88.1	85.5	83.5	84.7	84.7	80.9	86.4	93.6	93.5	96.4
	L2	uniform	94.7	88.0	84.5	82.6	84.5	84.7	80.4	85.5	94.2	93.3	96.4

Table 9: Supervised word alignment recall (%) measured on the test sets, where the underlined scores are the best and those within 1% absolute differences.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
The "Limitations" section comes after Section 9 (we did not number it).

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Sections 6-8

☑ B1. Did you cite the creators of artifacts you used?
Section 6

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 6

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 6 and Appendix A.1

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 6

C ☑ Did you run computational experiments?

Sections 7 and 8

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Sections 5.4 and 6, Appendix A.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Sections 7.1 and 8.1

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Sections 6-8

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.