WASSERSTEIN CROSS-LINGUAL ALIGNMENT FOR NAMED ENTITY RECOGNITION

Rui Wang   *Ricardo Henao*

Duke University

ABSTRACT

Supervised training of Named Entity Recognition (NER) models generally require large amounts of annotations, which are hardly available for less widely used (low resource) languages, *e.g.*, Armenian and Dutch. Therefore, it will be desirable if we could leverage knowledge extracted from a high resource language (source), *e.g.*, English, so that NER models for the low resource languages (target) could be trained more efficiently with less cost associated with annotations. In this paper, we study cross-lingual alignment for NER, an approach for transferring knowledge from high- to low-resource languages, via the alignment of token embeddings between different languages. Specifically, we propose to align by minimizing the Wasserstein distance between the contextualized token embeddings from source and target languages. Experimental results show that our method yields improved performance over existing works for cross-lingual alignment in NER tasks.

Index Terms— Named Entity Recognition, Cross-Lingual Alignment, Wasserstein Distance, Optimal Transport

1. INTRODUCTION

Named Entity Recognition (NER) constitutes the task of recognizing named entities, *e.g.*, *PERSON* and *LOCATION*, from text data, via predicting a label for each token from the input text sequences. Recent advancements of NER can mostly be ascribed to the advent of deep learning models [1, 2]. However, training of such models generally assumes large amounts of NER annotations, *i.e.*, the availability of labels for each word/token of the input sequences in a large corpus. Such annotations are usually expensive and labor-consuming, especially for low resource languages, *e.g.*, Armenian, where expert annotators are less readily available compared with high-resource languages, *e.g.*, English. In order to effectively train NER models for low resource languages with limited annotations, a natural question is whether we can transfer knowledge extracted from a high resource language (source), where annotations are more adequate and abundant, to those low-resource ones for which annotations are scarce (target).

In this paper, we consider such a challenge under the extreme scenario where no NER annotations are available for the target languages, *i.e.*, zero-shot cross-lingual knowledge transfer. To address this problem, one approach is to leverage multi-lingual pretrained models, *e.g.*, mBERT [1], for which data from multiple languages are used to pretrain a model with a common embedding space. In this way, the parameters of these pretrained models encode language-independent patterns that are shared across languages, thus facilitating knowledge transfer across different languages. However, languages involved in the pretraining of such models generally leverage data with no explicit cross-lingual supervision. Consequently, embeddings of corresponding token/words from different languages may not be properly aligned [3, 4].

So motivated, [5, 6] proposed to further conduct cross-lingual alignment during fine-tuning, explicitly aligning the contextualized embeddings of different languages with a *paralleled* corpus, *i.e.*, a set of translated pairs of text sequences for the source and target language. Since there is not a direct word/token-level correspondence between the translated pair of text sequences, these works use *fastAlign* [7], a statistical tool that learns to map a word in the source sequence into its corresponding translation in the target sequence, given the paralleled corpus. However, the mapping from *fastAlign* is entirely based on word-level features and remains fixed throughout training. Therefore, the resulting model may be biased to shallow cross-lingual correspondences during alignment [8]. Another naive way of inducing word/token-level correspondences is to utilize contextualized embeddings from the deep layers of a pretrained model, by mapping a token from the target sequence to its nearest neighbor of tokens from the source sequence in embedding space. In our experiments, we show that the mapping obtained from such an approach is usually noisy and affected by conflicting mappings, *e.g.*, a source token whose embedding, due to abstract semantics, can be mapped with multiple target tokens with distinct meanings, leaving some source token not mapped with any target token.

In this work, we propose to align the token embeddings from the source and target languages by minimizing their Wasserstein distance in representation (embedding) space. Specifically, we first learn a coupling matrix (mapping) between source and target tokens from the translated pair of text sequences, via solving the optimal transport plan between the contextualized source and target tokens embeddings. Then, we minimize the cosine distance between the source token and its corespondent target tokens induced from the coupling matrix. We denote our approach as Wasserstein Cross-lingual Alignment (WAC). Different from *fastAlign*, our coupling matrix is learned with contextualized tokens embeddings, thus less likely to suffer from biases caused by shallow cross-lingual correspondences. Further, our method addresses the mapping conflict issue described above in a principled way, via the optimization constrains imposed while solving the optimal transport. Experimental results show that our approach can produce improved results over previous baselines of zero-shot cross-lingual knowledge transfer.

2. RELATED WORK

Below we briefly describe previous works of zero-shot cross-lingual transfer for NER organized in four categories.

Model-transfer-based: [9, 10] train the mBERT [1] with labeled data of the source language, *i.e.*, the source model, and directly evaluate on target languages, demonstrating the ability of cross-lingual transfer for multi-lingual pretrained models.

Distillation-based: [11] proposes a teacher-student framework that first trains mBERT on the source language, then distills it into the student model using unlabeled data from the target language. This assumes that the student can self-correct the *noisy labels* generated by the teacher on the target data.

Projection-based: [12] generate pseudo-labels of the target language by projecting the source *gold-standard annotations* into the target using a word/entity dictionary, then train a target model on the pseudo-labeled target data. [13] translates the source annotated sentence into target, and finds the target correspondence of the source labeled entities with an additional entity list. Unitrans [14] fine-tunes the source model with the target pseudo-labels and enhances it with model ensembling. Our method can be complementary to these methods when a word/entity dictionary is available.

Alignment-based: Different from *model-transfer based* methods in which the model is trained without being aware of the target data, *alignment-based* methods align the token embeddings of the source and target languages. [5, 6] apply a statistical tool, *fastAlign* [7], to align the word embeddings of MBERT, with a paralleled corpus of the source and target languages. Similarly, our approach also uses an paralleled corpus for alignment. The key difference is that we propose to align by minimizing the Wasserstein distance between the source and target embeddings, with the coupling learnt from solving the optimal transport problem.

3. METHODOLOGY

3.1. Wasserstein Distance

The Wasserstein distance is induced from the optimal transport (OT) theory [15]. It has been previously used for the matching of probability distributions or node embeddings in graphs [16]. Here, we employ it for matching source and target token embeddings in cross-lingual knowledge transfer.

Let $x^s = [x_i^s]_{i=1}^{n^s}$ and $x^t = [x_j^t]_{j=1}^{n^t}$ be a translated pair of text sequences for the source and target languages, respectively, with corresponding source and target token embeddings, $[e_i^s]_{i=1}^{n^s}$ and $[e_j^t]_{j=1}^{n^t}$. Further, n^s and n^t are the length of the source and target text sequences, respectively. Ideally, for cross-lingual knowledge transfer, e_i^s should be close to e_j^t if x_i^s and x_j^t are semantically similar. We introduce two discrete probability distributions, ν^s and ν^t, formulated as $\nu^s = \sum_{i=1}^{n^s} u_i^s \delta_{e_i^s}$ and $\nu^t = \sum_{j=1}^{n^t} u_j^t \delta_{e_j^t}$, respectively, where $u^s = [u_i^s]_{i=1}^{n^s}$ and $u^t = [u_j^t]_{j=1}^{n^t}$ are positive weight vectors, *s.t.*, $\sum_{i=1}^{n^s} u_i^s = \sum_{j=1}^{n^t} u_j^t = 1$, and δ is the delta function. With $c(\cdot, \cdot)$ being a cost metric, the degree of misalignment between $[e_i^s]_{i=1}^{n^s}$ and $[e_j^t]_{j=1}^{n^t}$ can be quantified as the Wasserstein distance between distributions ν^s and ν^t, via

$$\mathcal{D}_w([e_i^s]_{i=1}^{n^s}, [e_j^t]_{j=1}^{n^t}) = \min_{T \in \Pi(u^s, u^t)} \sum_{i=1}^{n^s} \sum_{j=1}^{n^t} T_{i,j} c(e_i^s, e_j^t) \,, \quad (1)$$

where $\Pi(u^s, u^t)$ is the set of all possible couplings given weights u^s and u^t, *i.e.*, $\Pi(u^s, u^t) = \{T \in \mathbb{R}_{\geq 0}^{n^s \times n^t} | T1_{n^t} = u^s, T^\intercal 1_{n^s} = u^t\}$ and 1_{n^r} denotes the n^r-dimensional all-ones vector, for $r = s, t$. The optimal transport matrix T^* that satisfies (1) is defined as the optimal transport plan between ν^s and ν^t given the cost metric $c(\cdot, \cdot)$, controlling, via weighting, the matching between each possible pair of source and target tokens. The value of $T_{i,j}^*$, *i.e.*, the probability mass assigned for (x_i^s, x_j^t), denotes the degree of mapping between source and target tokens (x_i^s, x_j^t). Note that a smaller $c(e_i^s, e_j^t)$ is likely to result in a larger $T_{i,j}^*$, indicating that e_i^s and e_j^t should be matched to a high degree during alignment.

In our setting, we set both the weight vectors u^s and u^t to be uniform. In this way, the sum of each column and row of T^* should not be zero, indicating each soruce/target token should be mapped with some target/source tokens in the learnt T^*. This avoid the problem of conflicting mapping, in which case the probability mass is concentrated on some rows of T^*, while other rows are left zero.

3.2. Cross-lingual Alignment

Let $\{X^l, Y^l\}$ be a label NER dataset for the source language, *e.g.*, English, with $(x^l = [x_k^l]_{k=1}^{n^l}, y^l = [y_k^l]_{k=1}^{n^l}) \sim \{X^l, Y^l\}$ be the input and label sequences for the NER task, respectively. An NER model is expected to predict the label sequence y^l from input x^l, with $y_k^l \in \{0, \cdots, C-1\}$ being the label of token x_k^l, where C is the number of token labels. In cross-lingual transfer, given the source dataset $\{X^l, Y^l\}$, we want to obtain an NER model that works on a target language with lower NER resources, *e.g.*, Dutch. We consider the zero-shot case where no NER labels on the target language are available. Instead, we leverage a paralleled corpus $\{X^s, X^t\}$, where $(x^s, x^t) \sim (X^s, X^t)$ is a pair of translated source and target sentences from machine translation. We use the superscript s

and t for source and target, respectively. In our experiments, we use the paralleled corpus generated from Google Cloud Translation [17]. Such corpora are becoming more widely available for cross-lingual transfer problems with the recent advances of machine translation [5, 13].

Let M be a pretrained mBERT model with a linear layer on the top. The outputs from M are the token logits, denoted as m, *i.e.*, $M(x^r) = [m_k^r]_{k=1}^{n^r}$, for $r = s, t, l$. We can fine-tune M for NER with the source labeled dataset $\{X^l, Y^l\}$, using the cross-entropy loss

$$\mathcal{L}^{ce}(X^l, Y^l) = \\ -\mathbb{E}_{[M(X^l),y)\sim(M(X^l),Y^l)} \sum_{k=1}^{n^l} \log[\mathrm{softmax}(m_k^l)]_{y_k}\,, \tag{2}$$

where $[\cdot]_k$ denotes the k-th element of a vector. As discussed in Section 1, M is not pretrained with explicit cross-lingual supervision. Therefore, the resulting model fine-tuned with only the source data from (2) may produce inferior performance on the target language. Here, to maximally transfer knowledge between the source and target, we introduce Wasserstein Cross-lingual Alignment (WCA).

Given $(x^s, x^t) \sim (X^s, X^t)$, let $M^b(x^s) = [e_i^{s,b}]_{i=1}^{n^s}$ and $M^b(x^t) = [e_j^{t_b}]_{j=1}^{n^t}$ be the contextualized source and target token embeddings, respectively, from the b-th layer of M. Our loss for WCA is defined as,

$$\mathcal{L}^w(X^s, X^t) = \mathbb{E}_{(x^s,x^t)\sim(X^s,X^t)} \sum_{i=1}^{n^s} \sum_{j=1}^{n^t} T_{i,j}^{\mathbb{B},*} \sum_{b\in\mathbb{B}} c(e_i^{s,b}, e_j^{t,b})\,, \tag{3}$$

where we define $c(\cdot,\cdot)$ as the cosine distance and $\mathbb{B}$ is the set of layers for cross-lingual alignment. $T^{\mathbb{B},*}$ is a coupling (mapping) between the source and target shared by layers in $\mathbb{B}$, learnt based on OT.

We first consider a soft version of coupling, $T_{soft}^{\mathbb{B},*}$, by averaging the optimal transport plans for each layer in $\mathbb{B}$,

$$T_{soft}^{\mathbb{B},*} = \frac{1}{|B|} \sum_{b\in\mathbb{B}} T_{soft}^{b,*}, \tag{4}$$

$$T_{soft}^{b,*} = \underset{T\in\Pi(u^s,u^t)}{\arg\min} \sum_{i=1}^{n^s} \sum_{j=1}^{n^t} T_{i,j}\, c(e_i^{s,b}, e_j^{t,b})\,. \tag{5}$$

(5) can be solved with the Sinkhorn algorithm [18]. $T^{\mathbb{B},*}$ is defined as the hard version of $T_{soft}^{\mathbb{B},*}$,

$$T_{i,j}^{\mathbb{B},*} = \begin{cases} 1/n^t, & i = \arg\max_{i'} T_{i',j,soft}^{\mathbb{B},*} \\ 0, & \text{otherwise} \end{cases}. \tag{6}$$

We find $T^{\mathbb{B},*}$ has slightly better performance than $T_{soft}^{\mathbb{B},*}$, probably because (6) can block some noise from OT by modifying each column of $T_{soft}^{\mathbb{B},*}$ to be binary (one-hot).

Table 1: F1 scores (%) for cross-lingual knowledge transfer.

	En-De	**En-Es**	**En-Nl**	**En-Hy**
[9]	69.56	74.96	77.56	62.38
[19]	71.90	74.30	77.60	-
[20]	70.54	75.77	79.03	-
[6]	71.23	75.93	79.90	66.56
[10]	73.16	76.75	80.44	-
[11]	73.22	76.94	80.89	-
Nearest Token	73.57	76.54	80.65	71.87
fastAlign (Ours)	73.24	76.33	80.03	71.45
WCA	**74.03**	**77.01**	**81.10**	**72.24**

Compared with (1), we can find that aligning with (3) is approximately equivalent to minimizing the sum of Wasserstein distance for each layer in $\mathbb{B}$. The overall objective for cross-lingual knowledge transfer is defined as

$$\mathcal{L} = \mathcal{L}^{ce}(X^l, X^l) + \beta\mathcal{L}^w(X^s, X^t)\,, \tag{7}$$

where we simultaneously train on the labeled source dataset $\{X^l, Y^l\}$, while aligning the source and target embeddings with $\{X^s, X^t\}$. β is a balancing parameter.

4. EXPERIMENTS

4.1. Datasets and Training Details

Following the experiment design of [6], we experiment with three datasets with five langauges: CoNLL2003 (English/En and German/De), CoNLL2002 (Spanish/Es and Dutch/Nl), PioNER[1] (Armenian/Hy). We use English as the source language and the English data from CoNLL2003 as the source labeled dataset X^l, Y^l, than conduct cross-lingual transfer to the rest four languages, *i.e.*, En→De, En→Es, En→Nl and En→Hy. For the unlabeled paralleled corpus $\{X^s, X^t\}$, we ignore the NER labels in training dataset of the target languages to obtain the unlabeled target data X^t. Then, X^s is generated by translating X^t into English using Google Translation [17]. Our NER model is a pretrained MBERT with a linear layer on the top, whose outputs are the token logits.

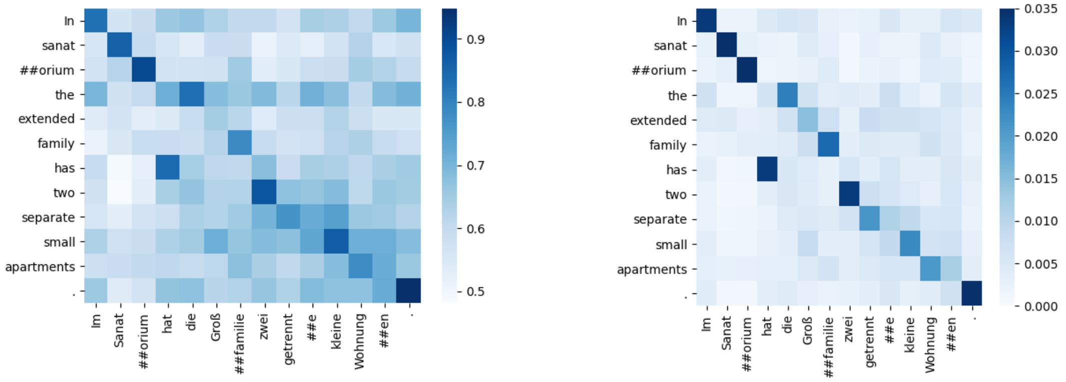
We use a training batch size of 32. Our model is trained with the learning rate of 5e-5 for three epochs. We use the AdaW [21] optimization and empirically set $\beta = 1e - 2$. Following [6], $\mathbb{B}$ contains the last four layers of the mBERT.

4.2. Baselines

We compare with the recent works that use unlabeled target/paralleled data for cross-lingual knowledge transfer. We also implement two additional baselines:

Nearest Token: Let (x^s, x^t) be translation pair for source and target. $(\{e_i^{s,b}\}_{i=1}^{n^s}, \{e_j^{t,b}\}_{j=1}^{n^t})$ are the contextualized token

[1] https://github.com/ispras-texterra/pioner



(a)

(b)

Fig. 1: Illustration of two baseline alignments for En-De. (a) Averaged cosine similarity between token embeddings from pretrained MBERT. Cosine similarity is 1 minus the cosine distance. (b) Averaged coupling learning from OT, *i.e.*, $T^{\mathbb{B},*}_{soft}$.

embeddings from layer b. Instead of learning the coupling via solving OT, we define $T^{\mathbb{B},near}$, a baseline coupling that maps each target token in x^t to its nearest source token,

$$T^{\mathbb{B},near}_{i,j} = \begin{cases} 1/n^t, & i = \arg\min_{i'} \frac{1}{|\mathbb{B}|}\sum_{b\in\mathbb{B}} c(e^{s,b}_i, e^{t,b}_j) \\ 0, & \text{otherwise} \end{cases}. \quad (8)$$

We replace $T^{\mathbb{B},*}$ with $T^{\mathbb{B},near}$. The rest of the procedures for alignment is the same as in (3). Compared with the coupling learnt from OT, the coupling from (8) is not subject to the OT optimization constrains as in (1).

fastAlign (Ours): [5, 6] have used *fastAlign* [7] in cross-lingual transfer problems. We have included their NER results for comparison. Further, we experiment with *fastAlign (Ours)* as our re-implementation of cross-lingual word alignment with *fastAlign*. Following [5, 6], we represent each word by averaging its corresponding token embeddings, generating $(\{w^{s,b}_i\}^{n^s_w}_{i=1}, \{w^{t,b}_j\}^{n^t_w}_{j=1})$ for layer b, where n^s_w and n^t_w are the number of words in the source and target sequences, respectively. Then, we define $T^{\mathbb{B},fa} \in \mathbb{R}^{n^s_w \times n^t_w}$ as,

$$T^{\mathbb{B},fa}_{i,j} = \begin{cases} 1/n^t_w, & (i,j) \in \epsilon^{fa} \\ 0, & \text{otherwise} \end{cases}. \quad (9)$$

ϵ^{fa} is the aligned source and target word pairs from *fastAlign*. For the alignment in (3), we replace $(\{e^{s,b}_i\}^{n^s}_{i=1}, \{e^{t,b}_j\}^{n^t}_{j=1})$ with $(\{w^{s,b}_i\}^{n^s_w}_{i=1}, \{w^{t,b}_j\}^{n^t_w}_{j=1})$, and $T^{\mathbb{B},*}$ with $T^{\mathbb{B},fa}$.

4.3. Result Analysis

Table 1 shows the results of cross-lingual knowledge transfer. Our method is better than the recent works that transfer with unlabeled target/paralleled data. Moreover, it has a large F1 margin over *Nearest Token* and *fastAlign (Ours)*. As an justification of these results, in Figure 1(a), we present the En-De cosine similarity, *i.e.*, one minus the cosine distance, between the source and target token embeddings from the pretrained MBERT, averaged over layers in $\mathbb{B}$. We find that the outputs from pretrained mBERT already induce a reasonable alignment for the paralleled corpus, *e.g.*, the German token *##familie* is mapped to the English word *family*, and *zwei* is mapped to *two*. However, some of the tokens are misaligned. For instance, the German token *Groß* means *extended* in English, but it is more similar to *small* in the English sequence, *i.e.*, the English *small* is the nearest source token of both *Groß* and *kleine* (small in German). Such a conflict in alignment may affect the context aggregation for named entity predictions. In Figure 1(b), we show the averaged coupling from OT of layers in $\mathbb{B}$, *i.e.*, $T^{\mathbb{B},*}_{soft}$. We show the soft version to give a more comprehensive view of the coupling from OT. Additionally, it is a fair comparison with 1(a), which is also a soft version. We do not show the coupling layer by layer due to space limit. It can be observed that the mapping from OT is less noisy than that in Figure 1(a). Specifically, the probability mass assigned to the pair *extended-Groß* is larger than that of *small-Groß*, which is contrary to what is shown in Figure 1(a). Such result is owed to the OT constraint in (1), *i.e.*, $T \in \Pi(u^s, u^t)$, which restricts the sum of each row and column of T to be constant. This reduces the noise from conflicting alignments, *i.e.*, assigning large probability mass from multiple target tokens with different meanings (Groß and *kleine*) into a single source token (*small*).

5. CONCLUSION

In this paper, we propose Wasserstein Cross-lingual Alignment (WAC), aligning contextualized token embeddings between the source and target langauges by minimizing their Wasserstein distance, where the coupling between the source and target tokens is learnt by solving the optimal transport problem. Experiments on zero-shot cross-lingual transfer show that our approach can provide improved results over previous baselines of cross-lingual alignment.

6. REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[2] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[3] Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng, "On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing," *arXiv preprint arXiv:1811.00570*, 2018.

[4] Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R Gormley, and Graham Neubig, "Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces," *arXiv preprint arXiv:1908.06625*, 2019.

[5] Steven Cao, Nikita Kitaev, and Dan Klein, "Multilingual alignment of contextual word representations," *arXiv preprint arXiv:2002.03518*, 2020.

[6] Saurabh Kulshreshtha, José Luis Redondo-García, and Ching-Yun Chang, "Cross-lingual alignment methods for multilingual bert: A comparative study," *arXiv preprint arXiv:2009.14304*, 2020.

[7] Chris Dyer, Victor Chahuneau, and Noah A Smith, "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*, 2013, pp. 644–648.

[8] Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Saloni Potdar, and Mo Yu, "Multilingual bert post-pretraining alignment," *arXiv preprint arXiv:2010.12547*, 2020.

[9] Shijie Wu and Mark Dredze, "Beto, bentz, becas: The surprising cross-lingual effectiveness of bert," *arXiv preprint arXiv:1904.09077*, 2019.

[10] Qianhui Wu, Zijia Lin, Guoxin Wang, Hui Chen, Börje F Karlsson, Biqing Huang, and Chin-Yew Lin, "Enhanced meta-learning for cross-lingual named entity recognition with minimal resources," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 9274–9281.

[11] Qianhui Wu, Zijia Lin, Börje F Karlsson, Jian-Guang Lou, and Biqing Huang, "Single-/multi-source cross-lingual ner via teacher-student learning on unlabeled data in target language," *arXiv preprint arXiv:2004.12440*, 2020.

[12] Stephen Mayhew, Chen-Tse Tsai, and Dan Roth, "Cheap translation for cross-lingual named entity recognition," in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017, pp. 2536–2545.

[13] Alankar Jain, Bhargavi Paranjape, and Zachary C Lipton, "Entity projection via machine translation for cross-lingual ner," *arXiv preprint arXiv:1909.05356*, 2019.

[14] Qianhui Wu, Zijia Lin, Börje F Karlsson, Biqing Huang, and Jian-Guang Lou, "Unitrans: Unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data," *arXiv preprint arXiv:2007.07683*, 2020.

[15] Cédric Villani, "Optimal transport: old and new," *Bull. Amer. Math. Soc*, vol. 47, pp. 723–727, 2010.

[16] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu, "Graph optimal transport for cross-domain alignment," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1542–1553.

[17] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[18] Gabriel Peyré, Marco Cuturi, et al., "Computational optimal transport: With applications to data science," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

[19] Phillip Keung, Yichao Lu, and Vikas Bhardwaj, "Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and ner," *arXiv preprint arXiv:1909.00153*, 2019.

[20] Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime Carbonell, "Cross-lingual alignment vs joint training: A comparative study and a simple unified framework," *arXiv preprint arXiv:1910.04708*, 2019.

[21] Ilya Loshchilov and Frank Hutter, "Fixing weight decay regularization in adam," 2018.