Semi-Supervised Bilingual Lexicon Induction with Two-way Interaction

Xu Zhao[1], **Zihao Wang**[2], **Hao Wu**[3], **Zhang Yong**[1] †
[1]BNRist, Department of Computer Science and Technology, RIIT,
Institute of Internet Industry, Tsinghua University, Beijing, China
[2]Department of CSE, HKUST, Hong Kong SAR, China
[3]Department of Mathematical Sciences, Tsinghua University, Beijing, China
zhaoxu18@mails.tsinghua.edu.cn
zwangggc@cse.ust.hk
{hwu,zhangyong05}@tsinghua.edu.cn

Abstract

Semi-supervision is a promising paradigm for Bilingual Lexicon Induction (BLI) with limited annotations. However, previous semi-supervised methods do not fully utilize the knowledge hidden in annotated and non-annotated data, which hinders further improvement of their performance. In this paper, we propose a new semi-supervised BLI framework to encourage the interaction between the supervised signal and unsupervised alignment. We design two message-passing mechanisms to transfer knowledge between annotated and non-annotated data, named prior optimal transport and bi-directional lexicon update respectively. Then, we perform semi-supervised learning based on a cyclic or a parallel parameter feeding routine to update our models. Our framework is a general framework that can incorporate any supervised and unsupervised BLI methods based on optimal transport. Experimental results on MUSE and VecMap datasets show significant improvement of our models. Ablation study also proves that the two-way interaction between the supervised signal and unsupervised alignment accounts for the gain of the overall performance. Results on distant language pairs further illustrate the advantage and robustness of our proposed method.

1 Introduction

Bilingual Lexicon Induction (BLI) is of huge interest to the research frontier. BLI methods learn cross-lingual word embeddings from separately trained monolingual embeddings. BLI is believed to be a promising way to transfer semantic information between different languages, and spawns lots of NLP applications like machine translation (Lample et al., 2018b; Artetxe et al., 2018b), Part Of Speech (POS) tagging (Gaddy et al., 2016), parsing (Xiao and Guo, 2014), and document classification (Klementiev et al., 2012).

The key step of BLI is to learn a transformation between monolingual word embedding spaces (Ruder et al., 2019), which could be further used for translation retrieval or cross-lingual analogy tasks. However, it is hard to obtain the high quality transformation with low supervision signals, i.e. with limited annotated lexicon. Thus, some semi-supervised BLI methods (Artetxe et al., 2017; Patra et al., 2019) are proposed to make use of annotated and non-annotated data. Artetxe et al. (2017) bootstrapped the supervised lexicon to enhance the supervision but ignored the knowledge in non-annotated data. Meanwhile, Patra et al. (2019) combined the unsupervised BLI loss that captured the structural similarity in word embeddings (Lample et al., 2018a) with the supervised loss (Joulin et al., 2018). However, this loss combination still performed poorly since the bad supervised optimization under limited annotations, see the Experiment part for details. As a result, existing semi-supervised BLI methods suffer from low effectiveness (Artetxe et al., 2017) or low robustness (Patra et al., 2019).

In this work, we focus on designing a new semi-supervised BLI method to make full use of both annotated and non-annotated data. We propose a novel framework with two different strategies, which exceeds the previous separate (Artetxe et al., 2017; Patra et al., 2019) semi-supervised methods by emphasizing the two-way interaction between the supervised signal and unsupervised alignment. In this framework, supervised training tries to align the parallel lexicon and unsupervised training can exploit the structure similarity between monolingual embedding spaces. The foundation of two-way interaction is in two carefully designed *message passing mechanisms*, see Section 3.1 and 3.2. Two-way interaction enables semi-supervised BLI

† Yong Zhang is the corresponding author.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 2973–2984,
November 16–20, 2020.

to guide the exploitation of structural similarity by *unsupervised procedure* (Grave et al., 2019) and extend insufficient lexicon for *supervised procedure* (Joulin et al., 2018) simultaneously, see Figure 1. In this paper, we only consider the unsupervised BLI methods based on Optimal Transport (OT) (Grave et al., 2019; Alaux et al., 2019; Huang et al., 2019; Alvarez-Melis et al., 2018), which have achieved impressive results on BLI task.

More specifically, the contributions of this paper are listed below.

- We propose the two-way interaction between the supervised signal and unsupervised alignment. It consists of two message passing mechanisms, Prior Optimal Transport (POT) and Bidirectional Lexicon Update (BLU). POT enables the OT-based unsupervised BLI approach to be guided by any prior BLI transformation, i.e. transfers what is learned by supervised BLI method to the unsupervised BLI method. BLU employs the alignment results in bi-directional retrieval to enlarge the annotated data, and thus enhances the supervised training by unsupervised BLI transformation.
- We propose two strategies of semi-supervised BLI framework based on POT and BLU, named by Cyclic Semi-Supervision (CSS) and Parallel Semi-Supervision (PSS). They are recognized by the cyclic and parallel parameter feeding routines, respectively, see Figure 1. Notably, CSS and PSS are universal to admit any supervised BLI methods and OT-based unsupervised BLI methods.
- Extensive experiments on two popular datasets show that CSS and PSS exceed all previous supervised, unsupervised, and semi-supervised approaches and are suitable to different scenarios. Ablation study of CSS and PSS demonstrates that the two-way interaction (POT and BLU) is the key to improve the performance. Results on distant language pairs show the advantage and robustness of our method.

2 Background

In this section, we describe the basic formulation of related supervised and unsupervised BLI methods. We define two embedding matrices $X, Y \in \mathbb{R}^{n\times d}$, where n is the number of words and d is the dimension of the word embedding.

The key to supervised BLI is the parallel lexicon between two languages, say word x_i in X is translated to word y_i in Y. Mikolov et al. (2013) suggested regarding supervised BLI as a regression problem aligning word embeddings by a linear transformation $Q^\star$.

$$Q^\star = \arg\min_{Q\in\mathbb{R}^{d\times d}} \sum_i \|X_i Q - Y_i\|_2^2 \tag{1}$$

Artetxe et al. (2016) introduced the orthogonal constraint on Q. Therefore, Problem (1) has a closed-form solution $Q^\star = UV^\top$, where U, V are defined by the SVD decomposition $Y^\top X = USV^\top$. Joulin et al. (2018) proposed to replace the 2-norm in Problem (1) by Relaxed Cross-domain Similarity Local Scaling (RCSLS) loss to mitigate the hubness problem, which is formulated in Equation (2).

$$\begin{aligned}\texttt{RCSLS}(x_iQ, y_i) = &-2Q^\top x_i^\top y_i \\ &+\frac{1}{k}\sum_{y\in\mathcal{N}_Y(x_iQ)} Q^\top x_i^\top y \\ &+\frac{1}{k}\sum_{xQ\in\mathcal{N}_X(y_j)} Q^\top x^\top y_j\end{aligned} \tag{2}$$

where $\mathcal{N}_X(y)$ represents the set which consists of the k nearest neighbors of y in the point cloud X, so as $\mathcal{N}_Y(x_iQ)$.

For unsupervised BLI, embeddings in X and Y are totally out of order. As a result, unsupervised BLI methods need to model an unknown permutation matrix $P \in \mathcal{P}_n = \{0,1\}^{n\times n}$

$$\min_{Q\in\mathcal{O}_d, P\in\mathcal{P}_n} \|XQ - PY\|_F^2 \tag{3}$$

where $\mathcal{O}_d$ is the set of orthogonal matrices. Problem (3) could be solved by iteratively minimizing Q and P. More specifically, Grave et al. (2019) considered random samples $\bar{X}, \bar{Y} \in \mathbb{R}^{m\times d}$ from X, Y in a stochastic optimization scheme. Minimizing P directly is hard. The key to unsupervised methods is how to solve P approximately, see Section 6. OT based methods solve P by optimal transport (Zhang et al., 2017; Grave et al., 2019; Alaux et al., 2019; Huang et al., 2019). Grave et al. (2019) and Zhang et al. (2017) proposed to solve the Wasserstein problem between the two distributions supported on $\bar{X}Q$ and $\bar{Y}$, respectively.

$$W_2^2(\bar{X}Q, \bar{Y}) = \min_{P\in\Pi} \sum_{i,j} P_{ij} D_{ij} \tag{4}$$

where D_{ij} is the cost between x_iQ and y_j, such as 2-norm, RCSLS loss, or other costs. $P \in \Pi =$

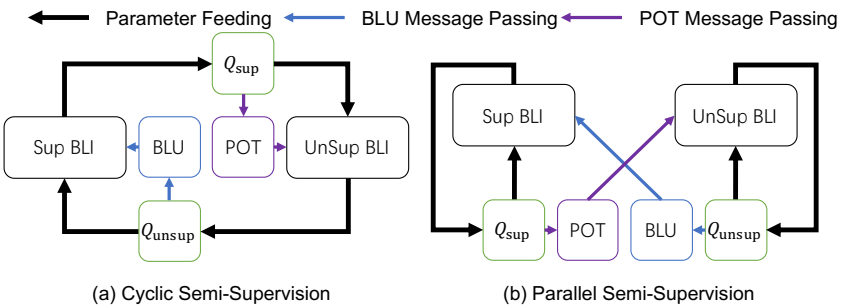

Figure 1: Illustration for Cyclic Semi-Supervision and Parallel Semi-Supervision

$\{P \in \mathbb{R}_+^{m\times m} | \sum_i P_{ij} = \sum_j P_{ij} = 1\}$ is the transport plan (Peyré et al., 2019). OT related metrics can be solved by the entropy regularized Sinkhorn algorithm (Cuturi, 2013).

$$W_2^2(\bar{X}Q, \bar{Y}) = \min_{P\in\Pi} \sum_{i,j} P_{ij}D_{ij} + \epsilon H(P) \quad (5)$$

To summarize, the foundation of supervised BLI is the annotated parallel lexicon for training, and the critical step of OT-based unsupervised BLI is the solution of transport plan P.

3 Message Passing in BLI

In this section, we present two message passing mechanisms for semi-supervised framework, including POT and BLU. POT is proposed to enhance the unsupervised BLI by the knowledge passed from the supervised BLI. Meanwhile BLU enhances the supervised BLI by the additional lexicon based on the unsupervised retrieval results. Therefore, POT and BLU form the two-way interaction between the supervised signal and unsupervised alignment.

3.1 Prior Optimal Transport

We present POT to strengthen the stochastic optimization in unsupervised BLI with prior information from supervised BLI. More specifically, POT is designed to guide the original OT solution of P, see Problem (4). POT can replace the original OT problems in unsupervised BLI models such as (Grave et al., 2019). In this way, we enable the transformation Q_{sup} trained in supervised BLI to enhance the unsupervised BLI.

Given Q_{sup} learned from any supervised BLI and random word embedding samples $\{x_i\}$ and $\{y_j\}$, we compute the cost matrix C between transformed source embeddings and target embeddings. In this work, we choose RCSLS as the specific formulation of C_{ij}

$$C_{ij} = RCSLS(x_i Q_{sup}, y_j). \quad (6)$$

Based on this cost function, we propose to use the Boltzmann Distribution, i.e. softmax function with temperature to construct a prior transport plan Γ:

$$\Gamma_{ij} = \frac{e^{-C_{ij}/T}}{\sum_{1\le k\le n} e^{-C_{ik}/T}}. \quad (7)$$

Γ_{ij} represents the probability that the i-th word in X is a translation of the j-th word in Y. Temperature T controls the significance of translation in Γ. Γ, induced from Q_{sup}, assigns each pair of words with a smaller cost in C a higher probability of forming a lexicon.

Instead of considering Problem (4), we consider the POT regularized by the Kullback-Leibler (KL) divergence between Γ and P.

$$\texttt{POT}(\bar{X}Q, \bar{Y}) = \min_{P\in\Pi} \langle D, P\rangle + \varepsilon KL(P\|\Gamma), \quad (8)$$

where $\langle D, P\rangle = \sum_{i,j} P_{ij}D_{ij}$ is the matrix inner product. We note that KL regularization in POT problem is totally different from the aforementioned entropic regularization (5). For entropy regularized OT, the regularization coefficient ϵ is expected to be as small as possible to approximate the original OT solution. However, for POT discussed in (8), the regularization coefficient ε controls the interpolation of OT transport plan that minimizes Problem (4) and prior transport plan Γ. Therefore, ε does not need to be as small as possible. Instead, it is a proper number to coordinate the effect from prior supervised transformation Q_{sup}.

The key to solving Problem (8) is to decompose the KL divergence into entropic term and linear term. Therefore, Problem (8) is reduced to

$$\min_{P\in\Pi} \langle D - \varepsilon \log \Gamma, P\rangle + \varepsilon H(P). \quad (9)$$

By treating $D - \varepsilon \log \Gamma$ as the Γ-prior cost matrix $D(\varepsilon, \Gamma)$, Problem (8) could also be solved by Sinkhorn algorithm. Again, since POT does not require ε to be closed to zero, the solution of POT Problem (8) will not suffer from numerical instability problems (Peyré et al., 2019).

3.2 Bi-directional Lexicon Update

As stated in Section 2, the key to supervised BLI is its parallel lexicon for training. Therefore, to enhance the supervised BLI, we propose BLU to extend the parallel lexicon by the structural similarity of word embeddings exploited in unsupervised BLI. To distinguish from unsupervised notations in Section 3.1, let $S, T \in \mathbb{R}^{n \times d}$ be the parallel word embedding matrices for source and target languages respectively. The i-th row s_i of S and t_i of T form a translation pair in the annotated lexicon. BLU selects the additional lexicon $S', T' \in \mathbb{R}^{l \times d}$ with high credit scores to extend S and T. Let $S^* = S \oplus S'$ and $T^* = T \oplus T'$ be the extended lexicon, where $\oplus$ denotes the concatenation operation along columns between two matrices.

Given the forward and backward transformations $\overrightarrow{Q}_{unsup}$ and $\overleftarrow{Q}_{unsup}$ between source language and target language from unsupervised BLI. BLU defines the S' and T' by Q_{unsup} and $Q_{unsup}^{\top}$ respectively in four steps:

(1) Compute the forward and backward distance matrices. Forward distance matrix $\overrightarrow{D}$ is defined between transformed source embeddings and target embeddings, while backward distance matrix $\overleftarrow{D}$ is defined between source embeddings and backward transformed target embeddings.

(2) Generate forward and backward translation pairs. Let $\overrightarrow{B} = \{(i,j) | j = \arg\min_k \overrightarrow{D}_{ik}\}$ and $\overleftarrow{B} = \{(i,j) | i = \arg\min_k \overleftarrow{D}_{kj}\}$ be the translation pair sets. Then take the intersection $B = \overrightarrow{B} \cap \overleftarrow{B}$ as the candidate additional lexicon.

(3) Compute the credit score CS for each translation pair. Firstly, we define the forward and backward credit scores for a pair $(i,j) \in B$. Let $\overrightarrow{C}(i)$ be the set of target word indices k, $\overrightarrow{C}(i) = \{k |$ $k \neq j$ and k is among top $K+1$ elements of $-\overrightarrow{D}_{i*}\}$, so as $\overleftarrow{C}(j)$. The forward credit score is defined by $\overrightarrow{CS}_{ij} = \sum_{k \in \overrightarrow{C}(i)} \overrightarrow{D_{ik}}/K - \overrightarrow{D}_{ij}$, and $\overleftarrow{CS}_{ij}$ is similarly defined. Then we define credit score CS_{ij} for $(i,j) \in B$ by $CS_{ij} = \overrightarrow{CS}_{ij} + \overleftarrow{CS}_{ij}$.

(4) Select additional lexicon by credit score. The additional lexicon is selected in descending order of the CS for each translation pair (i,j).

Based on the steps mentioned above, we append the *annotated lexicon* with the *additional lexicon* that contains high credit translation pairs.

This message passing mechanism is related to the bootstrap routine in (Artetxe et al., 2018a). However, we select the credible translation pairs from the intersection, rather than union, of the forward and backward set of translation pairs. In this way, we guarantee the high quality of the additional lexicon.

4 Semi-Supervision with Two-way Interaction

In the previous section, we have presented two message passing mechanisms to enhance supervised BLI and OT-based unsupervised BLI by prior transformation Q_{unsup} and Q_{sup}, respectively. Moreover, recent state-of-the-art (SOTA) supervised(Sup) and unsupervised($UnSup$) approaches are all based on stochastic optimization rather than the closed-form solution. This means that all SOTA Sup and $UnSup$ approaches can be considered as a module that operates on the feed-in parameter Q. Therefore, we propose two different strategies for semi-supervision that emphasize the two-way interaction between the supervised signal and unsupervised alignment based on the message passing mechanisms, see Figure 1. All SOTA Sup and OT-based $UnSup$ methods can be plugged into the proposed framework seamlessly.

4.1 Cyclic Semi-Supervision

The first proposed semi-supervised BLI strategy is CSS, see Figure 1 (a). CSS feeds the parameter Q into Sup and $UnSup$ iteratively in a cyclic parameter feeding routine. Cyclic parameter feeding is a "hard" way to share the parameters and is no more than Patra et al. (2019) itself. Besides parameter feeding, we propose to use the message passing mechanisms BLU and POT to strengthen the Sup and $UnSup$. However, there is no convergence guarantee for this optimization scheme. As a result, it may suffer from limited performance when the BLI task is hard, as will be detailed in Section 5.

4.2 Parallel Semi-Supervision

The second strategy is PSS, see Figure 1 (b), where Sup and $UnSup$ are performed in parallel. The information between Sup and $UnSup$ is only passed by the proposed message passing mechanisms. In

this point of view, Artetxe et al. (2017) only had the Sup part with lexicon update and ignored the UnSup part. Compared to CSS, PSS indirectly shares the information in a "soft" way and may be suitable for some hard BLI tasks. We use the metric formulated in Equation 4 to evaluate Q_{sup} and Q_{unsup} on the word embedding spaces and choose the better one as the final output of PSS.

5 Experiment

In this section, we conduct extensive experiments to evaluate the performance of CSS and PSS. We open the source code on Github*.

5.1 Setup

Baselines We take several methods proposed in recent five years as baselines, including supervised (Artetxe et al., 2016; Joulin et al., 2018; Jawanpuria et al., 2019), unsupervised (Artetxe et al., 2016; Lample et al., 2018a; Mohiuddin and Joty, 2019; Grave et al., 2019; Alaux et al., 2019) and semi-supervised (Artetxe et al., 2017; Patra et al., 2019; Mohiuddin et al., 2020) approaches. Brief introductions could be found in Section 6. The scores of baselines are retrieved from their papers or by running the publicly available codes if necessary. For Mohiuddin et al. (2020), we don't find the released source code.

Datasets We evaluate CSS and PSS against baselines on two popularly used datasets: the MUSE† dataset (Lample et al., 2018a) and the VecMap‡ dataset (Dinu and Baroni, 2015). The MUSE dataset consists of FASTTEXT word embeddings (Bojanowski et al., 2017) trained on Wikipedia corpora and more than 100 bilingual dictionaries of different languages. The FASTTEXT embeddings used in MUSE are trained on very large and highly semantically similar language corpora (Wikipedia), which means the results on MUSE are biased (Artetxe et al., 2018a) and easier to obtain. On the contrary, the VecMap dataset is less biased and harder using CBOW embeddings trained on the WacKy scrawled corpora and bilingual dictionaries obtained from the Europarl word alignments (Dinu and Baroni, 2015). We use the default training and test splits for both datasets.

Evaluation Setting Similar to Mohiuddin et al. (2020), we compare CSS and PSS against baselines on three annotated lexicons with different sizes, including one-to-one and one-to-many mappings: "100 unique" and "5K unique" contain one-to-one mappings of 100 and 5000 source-target pairs respectively, while "5K all" contains one-to-many mappings of all 5000 source and target words, that is, for each source word there may be multiple target words. Moreover, we present the experiment results of five totally unsupervised baselines and three supervised ones. All the accuracies reported in this section are the average of four repetitions. For detailed experimental data, such as the standard deviation, please refer to the tables in appendix.

Hyperparameter Setting We train our models using Stochastic Gradient Descent with a batch size of 400 and a learning rate 1.0 for Sup, a batch size of 8K and a learning rate 500 for UnSup. The temperature T in Equation (7) is 0.1 and the coefficient ε in Equation (8) is 1. The additional lexicon size is set 10000. Each epoch contains 2K supervised iterations and 50 unsupervised iterations. Each case runs 5 epochs. The aforementioned parameters work sufficiently good and we didn't search the best hyperparameters in this work. All the experiments are conducted by 32-core CPU and one NVIDIA Tesla V100 core. Our framework finished in 30 minutes, while the running time for Mohiuddin and Joty (2019) was 3 hours.

5.2 Results on MUSE Dataset

In Table 1, we show the word translation results for five language pairs from the MUSE dataset, including 10 BLI tasks considering bidirectional translation.

With "100 unique" annotated lexicon, CSS outperforms all other semi-supervised methods on every task. The accuracy score of Patra et al. (2019) is less than 3% on all tasks because the limited annotated lexicon is insufficient for effective learning, while Artetxe et al. (2017) avoided this problem by lexicon bootstrap. Both CSS and PSS keep strong performance with insufficient annotated lexicon by the proposed message passing mechanisms, and achieve 2.8% and 0.9% improvement over Artetxe et al. (2017), respectively. Compared to the iterative CSS that feeds parameters by UnSup directly into Sup, the parallel PSS has fewer connections between Sup and UnSup. Thus, CSS performance is better than PSS under low supervision.

We notice that semi-supervised approaches with "100 unique" annotated lexicon are even worse than

*https://github.com/BestActionNow/SemiSupBLI
†https://github.com/facebookresearch/MUSE
‡https://github.com/artetxem/vecmap

Method	EN-ES		EN-FR		EN-DE		EN-RU		EN-IT		Avg.
	→	←	→	←	→	←	→	←	→	←	
Unsupervised Baselines											
Artetxe et al. (2018a)	82.2	84.5	82.5	83.6	75.2	74.2	48.5	65.1	78.9	79.5	75.4
Lample et al. (2018a)	81.7	83.3	82.3	82.1	74.0	72.2	44.0	59.1	78.3	78.1	73.5
Mohiuddin and Joty (2019)	82.6	84.4	83.5	82.4	75.5	73.9	41.2	61.7	78.8	78.5	74.2
Grave et al. (2019)	82.8	84.1	82.6	82.9	75.4	73.3	43.7	59.1	66.6	62.5	71.3
Alaux et al. (2019)	82.4	85.1	82.7	83.4	75.5	74.4	45.8	64.9	79.4	79.4	75.3
Supervised Baselines with "5K all" annotated lexicon											
Artetxe et al. (2016)	81.9	83.4	82.1	82.4	74.2	72.7	51.7	63.7	77.4	77.9	74.7
Joulin et al. (2018)	84.1	86.3	83.3	84.1	**79.1**	76.3	**57.9**	67.2	79.0	81.4	77.9
Jawanpuria et al. (2019)	81.4	85.5	82.1	84.1	74.7	76.7	51.3	67.6	77.8	80.9	76.2
Semi-Supervised Baselines with "100 unique" annotated lexicon											
Artetxe et al. (2017)	79.9	83.2	82.8	83.0	72.9	72.5	38.9	62.2	78.5	77.7	73.2
Patra et al. (2019)	<3	<3	<3	<3	<3	<3	<3	<3	<3	<3	<3
CSS - RCSLS	83.9	85.1	83.7	83.5	77.5	74.6	48.8	63.0	79.9	80.5	76.0
PSS - RCSLS	82.0	83.1	82.1	81.9	74.4	72.2	46.5	61.5	78.7	78.9	74.1
Semi-Supervised Baselines with "5K unique" annotated lexicon											
Artetxe et al. (2017)	82.7	83.3	82.9	83.3	75.9	72.4	47.6	62.3	78.7	77.7	74.7
Patra et al. (2019)	82.2	84.6	82.6	83.9	75.6	73.7	52.2	65.2	77.8	78.6	75.6
Mohiuddin et al. (2020)	80.9	80.8	-	-	74.9	72.3	52.2	64.8	77.1	76.5	72.4
CSS - RCSLS	**84.5**	86.4	84.5	84.9	78.8	77.4	57.0	66.5	**81.4**	82.6	78.4
PSS - RCSLS	83.5	85.9	84.2	84.5	77.1	76.8	56.5	67.1	80.0	82.1	77.8
Semi-Supervised Baselines with "5K all" annotated lexicon											
Artetxe et al. (2017)	82.3	83.5	82.9	82.7	76.3	72.5	48.7	62.3	77.9	78.3	74.7
Patra et al. (2019)	84.3	86.2	83.9	84.7	**79.1**	76.6	57.1	67.7	79.3	82.4	78.1
Mohiuddin et al. (2020)	80.5	82.2	-	-	73.9	72.7	53.5	67.1	76.7	78.3	73.1
CSS - RCSLS	**84.5**	**86.9**	**85.3**	85.3	78.9	**78.7**	57.3	**67.9**	81.2	82.7	**78.9**
PSS - RCSLS	83.7	86.5	84.4	**85.5**	77.6	78.6	56.8	67.4	80.4	**82.8**	78.4

Table 1: Word translation accuracy(@1) of CSS and PSS on the MUSE dataset with RCSLS as their supervised loss. ('EN': English, 'ES': Spanish, 'FR': French, 'DE': German, 'RU': Russian, 'IT': Italian. Underline: the highest accuracy among the group. In bold: the best among all methods).

the unsupervised methods. This indicates that 100 annotation lexicon is too weak for supervised approach to learn meaningful transformation. It does not mean our approach has marginal contribution. On the contrary, these empirical results reveal that bad supervised BLI won't hurt the overall performance of our semi-supervised framework and this is what previous work cannot achieve.

As the annotated lexicon size increases, the dominance of CSS and PSS is still observed. Moreover, the gap between CSS and PSS disappears as the size of annotated lexicon gets larger. With "5K unique" annotated lexicon, CSS and PSS outperform other semi-supervised methods on all tasks. With "5K all" annotated lexicon, CSS and PSS outperform other semi-supervised baselines on 9 of 10 tasks. On average, CSS exceeds Artetxe et al. (2017), Patra et al. (2019) and Mohiuddin et al. (2020) by $4.2\%, 0.8\%, 5.8\%$, respectively.

Taking all methods into consideration, including supervised, semi-supervised and unsupervised, CSS and PSS achieve the highest accuracy on 8 of 10 tasks and the best results on average.

5.3 Results on VecMap Dataset

In Table 2, we show the word translation accuracy for three language pairs, including 6 translation tasks on the harder VecMap dataset (Dinu and Baroni, 2015).

Notably, a couple of unsupervised approaches (Lample et al., 2018a; Mohiuddin and Joty, 2019; Grave et al., 2019; Alaux et al., 2019) are evaluated to have a zero accuracy on some of the language pairs. On the one hand, their valotile results demonstrate the toughness of the VecMap dataset where the structural similarity for unsupervised BLI is very low. On the other hand, unstable performance may be explained by the high dependence of those methods on the initialization. Though the performance of those methods are highest in some cases,

Method	EN-ES		EN-IT		EN-DE		Avg.
	→	←	→	←	→	←	
Unsupervised Baselines							
Artetxe et al. (2018a)	36.9	31.6	47.9	42.3	48.3	44.1	41.9
Lample et al. (2018a)	34.7	0.0	44.9	38.7	0.0	0.0	19.7
Mohiuddin and Joty (2019)	37.4	31.9	47.6	42.5	0.0	0.0	26.6
Grave et al. (2019)	0.0	0.7	40.3	34.8	0.0	37.1	18.8
Alaux et al. (2019)	0.0	**58.3**	**70.0**	**69.5**	0.0	0.0	33.0
Supervised Baselines with "5K all" annotated lexicon							
Artetxe et al. (2016)	19.5	13.7	39.3	20.7	25.4	22.3	23.5
Joulin et al. (2018)	35.5	31.2	44.6	37.6	46.6	41.7	39.5
Jawanpuria et al. (2019)	37.5	33.1	47.6	40.1	48.8	**45.1**	42.0
Semi-Supervised Baselines with "100 unique" annotated lexicon							
Artetxe et al. (2017)	33.1	24.9	43.3	39.2	46.9	42.0	38.2
Patra et al. (2019)	<3	<3	<3	<3	<3	<3	<3
CSS - RCSLS	36.8	31.4	45.4	40.9	48.0	42.2	40.8
PSS - RCSLS	34.6	29.6	45.3	40.5	48.0	42.6	40.1
Semi-Supervised Baselines with "5K unique" annotated lexicon							
Artetxe et al. (2017)	33.3	27.6	43.9	38.4	46.0	41.1	38.4
Patra et al. (2019)	34.3	31.6	41.1	39.3	47.5	43.6	39.6
Mohiuddin et al. (2020)	33.4	27.3	44.1	38.9	42.5	39.4	37.6
CSS - RCSLS	38.1	32.2	46.4	41.2	47.9	43.2	41.5
PSS - RCSLS	**38.9**†	32.9	**47.8**†	41.1	**49.3**†	43.7	**42.3**†
Semi-Supervised Baselines with "5K all" annotated lexicon							
Artetxe et al. (2017)	32.7	28.1	43.8	38.0	47.4	40.8	38.5
Patra et al. (2019)	34.5	32.1	46.2	39.5	48.1	44.1	40.8
Mohiuddin et al. (2020)	33.7	27.9	43.7	38.9	43.6	39.2	37.8
CSS - RCSLS	38.9	32.5	46.6	41.3	48.4	42.5	41.7
PSS - RCSLS	**39.6**	**33.7**†	**47.8**†	**42.1**†	**50.8**	**44.8**†	**43.1**

Table 2: Word translation accuracy(@1) of CSS and PSS on the VecMap dataset with RCSLS as their supervised loss. ('EN': English, 'ES': Spanish, 'DE': German, 'IT': Italian. Underline: the highest accuracy among the group. In bold: the best among all methods. In bold and marked by †: the second-highest among all methods).

e.g. Alaux et al. (2019), due to the unstable nature. We also mark the second-highest score by bold font and † if necessary.

At all supervision levels, CSS and PSS outperform all other semi-supervised approaches. Taking all unsupervised, semi-supervised and supervised methods into account, CSS and PSS achieve SOTA accuracy on average. Notably, PSS gets the highest or the second-highest (except the unstable unsupervised baseline (Alaux et al., 2019)) scores for 5 of 6 language pairs.

The results for "100 unique" annotated lexicon support our finding on the MUSE dataset that CSS learns better at low supervision level. Interestingly, with "5K unique" and "5K all" annotated lexicons, PSS outperforms CSS on almost every task, which is different from the MUSE dataset. Given that the structural similarity of embeddings between different languages in VecMap is very low, UnSup procedure is very unstable. In this case, CSS has lower performance due to the unstable Q_{unsup} is directly fed into Sup, while the parallel strategy of PSS does not suffer from this problem.

5.4 Ablation Study

In the ablation study, we disassemble CSS and PSS into the basic components to analyze the contribution of each component. Specifically, we consider the proposed two message passing mechanisms POT and BLU. For CSS, we also include the effect of Sup or UnSup module in the cyclic parameter feeding. However, if Sup or UnSup in PSS is removed, the framework falls back to the unsupervised or supervised BLI, whose results are already in Table 1 and 2.

We conduct ablation experiments with two annotated lexicons with different sizes, "5K all" and "1K unique" to compare the behavior of CSS and PSS under different annotation level. The experimental setting is the same as the main experiments. The ablation results are presented in Table 3 on four language pairs (2 from MUSE dataset and 2 from VecMap dataset).

Effectiveness of POT and BLU:

Regardless of the annotated lexicon size, removing POT, BLU and both of them from CSS brings 2.4%, 0.9% and 13.0% decline of accuracy respectively on average. Notably, the cyclic parameter feeding does not bring further benefits. Only when combined with at least one message passing mechanism, POT or BLU, the accuracy is improved significantly. For PSS, removal of POT or BLU brings 1.6% and 1.0% decline on the average score respectively.

Moreover, we consider different annotated lexicon sizes. On average, removal of POT, BLU or both from CSS brings 1.2%, 0.7% and 4.2% decline respectively with "5K all" annotated lexicon size, 3.4%, 0.9% and 21.6% decline with "1K unique" annotated lexicon size. The message passing mechanisms contribute drastically with a smaller annotated lexicon size since Sup receives significantly larger additional lexicons from UnSup to strengthen its performance. As for PSS, removal of POT and BLU brings 0.9% and 1.5% decline respectively with "5K all" annotated lexicon size, 2.3% and 0.7% decline with "1K unique" annotated lexicon size. No significant effect of annotation level for PSS is observed in ablation study. For both CSS and PSS, the contribution of POT is slightly larger than that of BLU and the combination of them could bring impressive improvement in general.

Analysis of Sup and UnSup in CSS:

In this step, we remove Sup or UnSup from CSS

Annotated Lexicon Size	5K all									1K unique									
Dataset	MUSE				VecMap				avg	MUSE				VecMap				avg	avg of all
Languages	EN-ES		EN-FR		EN-DE		EN-IT			EN-ES		EN-FR		EN-DE		EN-IT			
	→	←	→	←	→	←	→	←		→	←	→	←	→	←	→	←		
Results of the Ablation to CSS										***Results of the Ablation to CSS***									
CSS - RCSLS	84.5	86.9	85.5	85.3	48.4	42.5	46.6	41.3	65.1	83.8	85.0	83.9	83.7	47.8	42.8	45.3	41.2	64.2	64.7
⊖ POT	83.9	86.6	84.0	84.8	46.6	41.7	43.7	39.5	63.9	81.6	84.8	82.2	83.4	41.8	40.2	36.5	35.7	60.8	62.3
⊖ BLU	83.2	86.6	84.4	84.7	47.4	42.9	45.4	40.4	64.4	82.5	83.7	82.5	82.2	47.7	42.6	45.3	39.7	63.3	63.8
⊖ POT & ⊖ BLU	82.5	84.9	83.0	83.7	41.7	36.6	39.7	34.8	60.9	61.0	62.4	57.4	59.5	28.1	22.4	26.2	23.7	42.6	51.7
⊖ UnSup & ⊖ POT	84.3	86.5	84.8	85.1	45.8	40.1	42.8	39.0	63.6	81.7	83.3	80.5	81.4	40.3	36.1	38.2	36.4	59.7	61.6
⊖ Sup & ⊖ BLU	82.3	83.2	82.5	82.7	47.7	43.2	45.3	40.4	63.4	82.5	83.8	82.2	82.9	47.8	42.8	45.2	39.7	63.4	63.4
Results of the Ablation to PSS										***Results of the Ablation to PSS***									
PSS - RCSLS	83.7	86.5	84.4	85.5	50.8	44.8	47.8	42.1	65.7	82.9	83.8	82.4	83.0	48.4	43.0	46.6	40.1	63.8	64.7
⊖ POT	83.5	85.4	84.4	85.3	49.1	43.1	46.5	40.8	64.8	81.1	82.6	82.4	81.9	45.1	40.7	41.7	36.4	61.5	63.1
⊖ BLU	82.8	85.4	83.0	84.3	48.5	43.7	46.0	39.9	64.2	81.9	83.9	82.2	82.5	48.0	42.6	44.8	39.0	63.1	63.7

Table 3: Ablation Study with "5K all" and "1K unique" annotated lexicon. (⊖: remove specific component from the CSS or PSS. &: remove both components.)

and monitor the performance change. Note that if we remove UnSup from CSS, POT also needs to be removed as we do not need any prior transport plan for UnSup anymore. Removing Sup also means the removal of BLU for a similar reason. After removing UnSup and POT, CSS feeds Q_{sup} exactly to BLU for additional lexicon and then to Sup again, just like Artetxe et al. (2017, 2018a). After removing Sup and BLU, UnSup takes the transformation learned by itself in previous steps to generate the prior transport plan. The average accuracy drops by 1.5% and 4.5% with "5K all" and "1K unique" annotated lexicon respectively after removing UnSup, by 1.7% and 0.8% after removing Sup.

Given the comparison above, Sup contributes less than UnSup with "1K unique" annotated lexicon. Whereas Sup and UnSup contribute comparably with "5K all" annotated lexicon. In other words, at low annotation level, i.e. "1K unique", where Sup BLI does not work well, the participation of UnSup extends the valuable additional lexicon.

5.5 Results on distant language pairs

In this section, We report the tranlation accuracy of our method on five distant language pairs with 5000 lexicon. We choose three methods as baselines: Patra et al. (2019) proposed semi-supervised SOTA method. Jawanpuria et al. (2019) is the supervised SOTA method. Zhou et al. (2019) designed an unsupervised matching procedure with density matching technologies, which achieved significant improvement on distant language pairs. As we need to compare supervised, unsupervised and semi-supervised method simultaneously, we conduct evaluation only on the "5K unique" supervision level.

As shown in Table 4, our method also retains a distinct advantage on these distant language pairs. In the cases between "EN" and "JA", Patra et al. (2019) and Jawanpuria et al. (2019) are completely inefficient. While our method obtains stable results on these cases, which proves the robustness of CSS and PSS. Moreover, our method outperforms Zhou et al. (2019) on most cases. In short, CSS and PSS could obtain stable and better results on various language pairs.

6 Related Work

This paper is mainly related to the following three lines of work.

Supervised methods. Mikolov et al. (2013) pointed out that it was a feasible way to BLI by learning a linear transformation based on the Euclidean distance. Artetxe et al. (2016) applied normalization to word embeddings and imposed an orthogonal constraint on the linear transformation which led to a closed-form solution. Joulin et al. (2018) replaced Euclidean distance with the RCSLS distance to relieve the hubness phenomenon and achieved SOTA results for many languages. Jawanpuria et al. (2019) optimized a Mahalanobis metric along with the transformation to refine the similarity between word embeddings.

Unsupervised methods. Artetxe et al. (2018a) proposed an unsupervised method to generate an initial lexicon by exploiting the similarity in cross-lingual space and applied a robust self-learning to improve it iteratively. Lample et al. (2018a) did the first work for unsupervised BLI which learned a linear transformation by adversarial training and improved it by a refinement procedure. Mohiuddin and Joty (2019) revisited adversarial autoencoder for unsupervised word translation and proposed two novel extensions to it. Moreover, OT-based

Method	EN-ZH		EN-TA		EN-JA		EN-MS		EN-FI		Avg.
	→	←	→	←	→	←	→	←	→	←	
(Semi-)Supervised Baselines with "5K unique" annotated lexicon											
Patra et al. (2019)	42.5	42.8	15.3	22.0	3.3	34.8	50	49.3	48.6	60.9	37.0
Jawanpuria et al. (2019)	43.7	40.1	16.1	22.0	0.0	0.0	51.0	49.8	47.4	65.1	33.5
Zhou et al. (2019)	42.5	40.0	**17.9**	**27.9**	**52.0**	35.6	37.6	40.1	50.7	60.8	40.5
CSS - RCSLS	**48.6**	**46.0**	17.7	24.5	50.3	**41.9**	**54.6**	55.5	**56.0**	65.6	**46.1**
PSS - RCSLS	47.8	45.4	17.3	23.3	50.6	41.3	54.6	**55.6**	55.6	**66.4**	45.8

Table 4: Word translation accuracy(@1) of CSS and PSS on the distant language pairs with RCSLS as their supervised loss. ('EN': English, 'TA': Tamil, 'JA': Japanese, 'MS': Malay, 'FI': Finnish. In bold: the best among all methods).

unsupervised BLI is the central part in this paper. Alvarez-Melis and Jaakkola (2018) exploited the structure similarity of embedding space by minimizing the Gromov-Wasserstein metric between source and target word embedding distributions. Grave et al. (2019) viewed unsupervised BLI task as the minimization of Wasserstein distance between the source and target distributions of word embeddings. They optimized this problem by using Sinkhorn and Procrustes alternatively. Alaux et al. (2019) furthered the work of Grave et al. (2019) by using the RCSLS as the distance metric, which addresses hubness phenomenon better than Euclidean distance. Zhao et al. (2020) proposed an relaxed matching procedure derived from unbalanced OT algorithms and solved the polysemy problem to a certain extent. Xu et al. (2018) used a neural network implementation to calculate the Sinkhorn distance, a well-defined OT-based distributional similarity measure, and optimized the objective through back-propagation.

Semi-supervised methods. Artetxe et al. (2017) proposed a simple self-learning approach that can be combined with any dictionary-based mapping technique and started with almost no lexicon. Patra et al. (2019) proposed a semi-supervised approach that relaxes the isometric assumption and optimizes a supervised loss and an unsupervised loss together.

Notably, comparing with the self-learning method like (Artetxe et al., 2018a) or (Vulic et al., 2019), our framework with two message passing mechanisms is quite different from theirs. Although the lexicon updating procedures in their papers are similar with the BLU that we proposed, there are two main differences: (1) Their approaches use the lexicon from current step to extract the lexicon for next step. Meanwhile, BLU uses unsupervised output to extract lexicon for the supervised part. Our models will degenerate to their situation after removing the unsupervised part and POT. This situation has been discussed in the ablation study in Section 5.4. (2) BLU extracts lexicon according to bidirectional matching information while they only consider one direction. This trick improves the lexicon quality.

Moreover, alignment of word embeddings in latent spaces by Auto-Encoders or other projections is another trend of BLI research. Latent space alignment includes unsupervised variants (Dou et al., 2018; Bai et al., 2019; Mohiuddin and Joty, 2019) and semi-supervised variants (Mohiuddin et al., 2020). We emphasize that the latent space alignment is orthogonal to our proposed framework. Our entire framework can be transferred to any given latent space.

7 Conclusions

In this paper, we introduce the two-way interaction between the supervised signal and unsupervised alignment by proposed POT and BLU message passing mechanisms. POT guides the OT-based unsupervised BLI by prior BLI transformation. BLU employs a bidirectional retrieval to enlarge the annotated data and stabilize the training of supervised BLI approaches. Ablation study shows that the two-way interaction by POT and BLU is the key to significant improvement.

Based on the message passing mechanisms, we design two strategies of semi-supervised BLI to integrate supervised and unsupervised approaches, CSS and PSS, which are constructed on cyclic and parallel strategies respectively. The results show that CSS and PSS achieve SOTA results over two popular datasets. As CSS and PSS are compatible with any supervised BLI and OT-based unsupervised BLI approaches, they can also be applied to the latent space optimization.

References

Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. 2019. Unsupervised hyperalignment for multilingual word embeddings. *ICLR*.

David Alvarez-Melis and Tommi S. Jaakkola. 2018. Gromov-wasserstein alignment of word embedding spaces. In *EMNLP*, pages 1881–1890.

David Alvarez-Melis, Tommi S. Jaakkola, and Stefanie Jegelka. 2018. Structured optimal transport. In *AISTATS*, pages 1771–1780.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *EMNLP*, pages 2289–2294.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *ACL*, pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *ACL*, pages 789–798.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *ICLR*.

Xuefeng Bai, Hailong Cao, Kehai Chen, and Tiejun Zhao. 2019. A bilingual adversarial autoencoder for unsupervised bilingual lexicon induction. *TASLP*, 27(10):1639–1648.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *TACL*, 5:135–146.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, pages 2292–2300.

Georgiana Dinu and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *ICLR*.

Zi-Yi Dou, Zhi-Hao Zhou, and Shujian Huang. 2018. Unsupervised bilingual lexicon induction via latent variable models. In *EMNLP*, pages 621–626.

David M Gaddy, Yuan Zhang, Regina Barzilay, and Tommi S Jaakkola. 2016. Ten pairs to tag-multilingual pos tagging via coarse mapping between embeddings. In *NAACL*.

Edouard Grave, Armand Joulin, and Quentin Berthet. 2019. Unsupervised alignment of embeddings with wasserstein procrustes. In *AISTATS*, pages 1880–1890.

Jiaji Huang, Qiang Qiu, and Kenneth Church. 2019. Hubless nearest neighbor search for bilingual lexicon induction. In *ACL*, pages 4072–4080.

Pratik Jawanpuria, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. 2019. Learning multilingual word embeddings in latent metric space: A geometric approach. *TACL*, 7:107–120.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *EMNLP*, pages 2979–2984.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *COLING*, pages 1459–1474.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018a. Word translation without parallel data. In *ICLR*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *EMNLP*, pages 5039–5049.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Tasnim Mohiuddin, M. Saiful Bari, and Shafiq R. Joty. 2020. Lnmap: Departures from isomorphic assumption in bilingual lexicon induction through non-linear mapping in latent space. *CoRR*, abs/2004.13889.

Tasnim Mohiuddin and Shafiq R. Joty. 2019. Revisiting adversarial autoencoder for unsupervised word translation with cycle consistency and improved training. In *NAACL*, pages 3857–3867.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *ACL*, pages 184–193.

Gabriel Peyré, Marco Cuturi, et al. 2019. Computational optimal transport. *FTML*, 11(5-6):355–607.

Sebastian Ruder, Ivan Vulic, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *JAIR*, 65:569–631.

Ivan Vulic, Goran Glavas, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *EMNLP*, pages 4406–4417.

Min Xiao and Yuhong Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. In *CoNLL*, pages 119–129.

Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. 2018. Unsupervised cross-lingual transfer of word embedding spaces. In *EMNLP*, pages 2465–2474.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *EMNLP*, pages 1934–1945.

Xu Zhao, Zihao Wang, Yong Zhang, and Hao Wu. 2020. A relaxed matching procedure for unsupervised BLI. In *ACL*, pages 3036–3041.

Chunting Zhou, Xuezhe Ma, Di Wang, and Graham Neubig. 2019. Density matching for bilingual word embedding. In *NAACL*, pages 1588–1598.

Appendix

Dataset		MUSE										VecMap						Distant Language Pairs									
Languages		EN-ES		EN-FR		EN-DE		EN-RU		EN-IT		EN-ES		EN-IT		EN-DE		EN-ZH		EN-TA		EN-JA		EN-MS		EN-FI	
		→	←	→	←	→	←	→	←	→	←	→	←	→	←	→	←										
Semi-Supervised Baselines with "100 unique" word dictionary																											
CSS - RCSLS	**best**	84.1	85.2	83.9	83.5	77.7	74.7	49.3	63.1	80.1	80.7	36.9	32.8	45.9	41.1	48.5	42.3										
	avg	83.9	85.1	83.7	83.5	77.5	74.6	48.8	63.0	79.9	80.5	36.8	31.4	45.4	40.9	48.0	42.2										
	st	0.193	0.158	0.240	0.065	0.285	0.122	0.525	0.138	0.175	0.140	0.205	0.930	0.341	0.180	0.375	0.087										
PSS - RCSLS	**best**	82.2	83.3	82.6	82.1	74.87	72.6	47.2	61.6	78.7	79	35.0	29.7	45.7	40.7	48.3	42.7										
	avg	82.0	83.1	82.1	81.9	74.4	72.2	46.5	61.5	78.7	78.9	34.6	29.6	45.3	40.5	48.0	42.6										
	st	0.138	0.167	0.366	0.249	0.358	0.345	0.627	0.136	0.078	0.115	0.553	0.156	0.406	0.178	0.381	0.181										
Semi-Supervised Baselines with "5K unique" word dictionary																											
CSS - RCSLS	**best**	84.7	86.5	84.6	85.1	79.0	77.7	57.5	66.8	81.6	82.7	38.4	32.6	46.5	41.4	48.3	43.3	48.7	46.1	18.1	24.9	50.5	42.7	55.1	56.3	56.5	65.7
	avg	84.5	86.4	84.5	84.9	78.8	77.4	57.0	66.5	81.4	82.6	38.1	32.2	46.4	41.2	47.9	43.2	48.6	46.0	17.7	24.5	50.3	41.9	54.6	55.5	56.0	65.6
	st	0.167	0.050	0.115	0.201	0.158	0.236	0.817	0.339	0.168	0.136	0.318	0.358	0.177	0.313	0.369	0.113	0.196	0.103	0.335	0.524	0.149	0.729	0.409	0.717	0.578	0.191
PSS - RCSLS	**best**	83.7	86.0	84.5	84.7	77.2	77.0	56.7	67.6	80.1	82.1	39.1	33.2	48.2	41.3	49.6	43.8	47.9	45.4	17.3	23.4	50.8	41.6	54.9	55.8	55.9	66.6
	avg	83.5	85.9	84.2	84.5	77.05	76.8	56.5	67.1	80.0	82.1	38.9	32.9	47.8	41.1	49.3	43.7	47.8	45.4	17.3	23.3	50.6	41.3	54.6	55.6	55.6	66.4
	st	0.173	0.115	0.236	0.171	0.129	0.183	0.283	0.395	0.082	0.100	0.168	0.303	0.364	0.308	0.214	0.100	0.075	0.075	0.115	0.107	0.171	0.283	0.196	0.231	0.273	0.236
Semi-Supervised Baselines with "5K all" word dictionary																											
CSS - RCSLS	**best**	84.6	87.1	85.5	85.3	78.9	78.8	57.8	68.2	81.4	82.8	39.0	32.7	47.0	41.4	49.1	42.9										
	avg	84.5	86.9	85.3	85.3	78.9	78.7	57.3	67.9	81.2	82.7	38.9	32.5	46.6	41.3	48.4	42.5										
	st	0.126	0.481	0.238	0.099	0.087	0.115	0.521	0.285	0.198	0.053	0.175	0.295	0.425	0.178	0.500	0.272										
PSS - RCSLS	**best**	83.9	87.0	84.6	85.5	77.7	78.8	57.2	67.5	80.7	82.9	39.7	34.0	48.3	42.5	51.2	45.1										
	avg	83.7	86.5	84.4	85.5	77.6	78.6	56.8	67.4	80.4	82.8	39.6	33.7	47.8	42.1	50.8	44.8										
	st	0.163	0.359	0.150	0.100	0.115	0.206	0.271	0.191	0.25	0.115	0.196	0.368	0.968	0.436	0.513	0.345										

Table 1: Detailed Experimental Results of CSS and PSS on MUSE Dataset, VecMap Dataset and distant language pairs. We repeat the experiment on each language pair four times and report **best**, **avg**, **st** of the four results(**best**: the highest @1 accuracy. **avg**: the average accuracy which is reported in main body of this paper. **st**: the standard deviation.)

Annotated Leixon Size		5K all								1K unique							
dataset		MUSE				VecMap				MUSE				VecMap			
Languages		EN-ES		EN-FR		EN-DE		EN-IT		EN-ES		EN-FR		EN-DE		EN-IT	
		→	←	→	←	→	←	→	←	→	←	→	←	→	←	→	←
Detailed Results of the Ablation to CSS										***Detailed Results of the Ablation to CSS***							
CSS - RCSLS	**best**	84.6	87.1	85.5	85.3	49.1	42.9	47.0	41.4	83.8	85.2	84.0	83.9	48.2	42.9	45.7	41.3
	avg	84.5	86.9	85.5	85.3	48.4	42.5	46.6	41.3	83.8	85.0	83.9	83.7	47.8	42.8	45.3	41.2
	st	0.126	0.481	0.238	0.099	0.500	0.272	0.425	0.178	0.063	0.144	0.175	0.136	0.408	0.071	0.430	0.096
⊖ POT	**best**	84.0	86.7	84.1	85.0	46.9	42.0	44.0	39.6	83.5	85.9	83.0	83.9	42.5	40.9	37.1	37.5
	avg	83.9	86.6	84.0	84.8	46.6	41.7	43.7	39.5	81.6	84.8	82.2	83.4	41.8	40.2	36.5	35.7
	st	0.156	0.099	0.116	0.158	0.330	0.372	0.282	0.093	2.274	1.004	1.372	0.737	1.175	0.700	0.835	1.806
⊖ BLU	**best**	83.3	86.8	84.7	84.9	47.7	43.1	45.6	40.6	82.7	83.9	82.9	82.3	47.9	43.0	46.0	40.1
	avg	83.2	86.6	84.4	84.7	47.4	42.9	45.4	40.4	82.5	83.7	82.5	82.2	47.7	42.6	45.3	39.7
	st	0.083	0.201	0.182	0.180	0.247	0.256	0.219	0.175	0.259	0.197	0.293	0.115	0.268	0.314	0.552	0.343
⊖ POT & ⊖ BLU	**best**	82.9	85.5	83.3	83.9	42.1	37.0	40.2	34.9	62.1	63.2	57.9	60.5	28.8	22.6	26.8	24.1
	avg	82.5	84.9	83.0	83.7	41.7	36.6	39.7	34.8	61.0	62.4	57.4	59.5	28.1	22.4	26.2	23.7
	st	0.242	0.396	0.259	0.205	0.298	0.234	0.539	0.075	0.756	0.628	0.412	0.774	0.476	0.255	0.410	0.260
⊖ UnSup & ⊖ POT	**best**	84.4	86.8	85.5	85.5	45.9	40.6	43.0	39.5	83.4	84.4	83.3	82.7	42.1	36.5	38.6	37.5
	avg	84.3	86.5	84.8	85.1	45.8	40.1	42.8	39.0	81.7	83.3	80.5	81.4	40.3	36.1	38.2	36.4
	st	0.191	0.223	0.500	0.233	0.144	0.365	0.360	0.490	1.523	1.346	3.528	1.506	1.972	0.465	0.368	0.998
⊖ Sup & ⊖ BLU	**best**	82.4	83.3	82.6	82.9	48.1	43.6	45.5	40.9	82.6	83.9	82.3	83.0	47.9	43.0	45.4	40.4
	avg	82.3	83.2	82.5	82.7	47.7	43.2	45.3	40.4	82.5	83.8	82.2	82.9	47.8	42.8	45.2	39.7
	st	0.085	0.115	0.157	0.347	0.337	0.530	0.110	0.388	0.115	0.115	0.135	0.083	0.040	0.249	0.148	0.518
Detailed Results of the Ablation to PSS										***Detailed Results of the Ablation to PSS***							
PSS - RCSLS	**best**	83.9	87.0	84.6	85.5	51.2	45.1	48.3	42.5	83.1	84.5	82.4	83.2	48.8	43.3	47.1	40.5
	avg	83.7	86.5	84.4	85.5	50.8	44.8	47.8	42.1	82.9	83.8	82.4	83.0	48.4	43.0	46.6	40.1
	st	0.163	0.359	0.150	0.100	0.513	0.345	0.968	0.436	0.139	0.476	0.063	0.183	0.409	0.314	0.451	0.334
⊖ POT	**best**	83.7	86.2	84.5	85.5	49.3	43.8	46.8	40.9	81.3	83.5	82.5	82.4	45.7	41.8	42.5	37.1
	avg	83.5	85.4	84.4	85.3	49.1	43.1	46.5	40.8	81.1	82.6	82.4	81.9	45.1	40.7	41.7	36.4
	st	0.162	1.568	0.087	0.300	0.205	0.731	0.446	0.138	0.348	0.839	0.135	0.707	0.816	1.326	1.222	0.830
⊖ BLU	**best**	82.9	85.7	83.1	84.4	48.9	44.0	46.9	40.2	82.1	84.3	82.5	82.9	48.5	42.8	45.3	40.0
	avg	82.8	85.4	83.0	84.3	48.5	43.7	46.0	39.9	81.9	83.9	82.2	82.5	48.0	42.6	44.8	39.0
	st	0.083	0.168	0.083	0.139	0.391	0.297	0.768	0.183	0.258	0.252	0.162	0.301	0.353	0.199	0.531	0.746

Table 2: Detailed Experimental Results Ablation Study. We repeat the experiment on each language pair four times and report **best**, **avg**, **st** of the four results(**best**: the highest @1 accuracy. **avg**: the average accuracy which is reported in main body of this paper. **st**: the standard deviation.)