and D_{bi} is the set of pairs of words and the corresponding dependency-based cross-lingual contexts. $L_{w,c}$ is defined in Equation (3). Bilingual word embeddings can be learned by optimizing the dependency-based joint learning objective above, and the algorithm is called DepBiWE.

Cross-lingual phrase-level regularization

We can further augment the DepBiWE model and enhance the quality of cross-lingual embeddings by making full use of word alignment information in the parallel corpus with a cross-lingual regularization in terms of phrase-level semantic similarities.

In the dependency parse-tree of the parallel sentence pairs, we define the dependency-phrase p as a word pair $(w, w_{dr^{-1}})$, where $w_{dr^{-1}}$ is the head of w and dr^{-1} denotes their inverse dependency relation. The representation of a dependency-phrase $\mathbf{p}$ can be represented as the sum of two word vectors, i.e., $\mathbf{p} = \mathbf{w} + \mathbf{w}_{dr^{-1}}$. By incorporating the phrase-level semantic information, we encourage the representations of similar dependency-phrases to be close, as we can derive the aligned dependency-phrases from the aligned words in the parallel sentence pairs. For example, (*review*, *word*) and (*wiederholen*, *Wörter*) in Figure 1 are aligned as dependency-phrases. The more dependency-phrase pairs are identified in the parallel corpus, the closer the embeddings for the two dependency-phrases will be pushed together. By minimizing the distance between aligned dependency-phrases, the auxiliary cross-lingual regularization term can be written as:

$$L_R = \gamma_R \sum_{(p_i^{l_1}, p_j^{l_2}) \in D_p} ||\mathbf{p}_i^{l_1} - \mathbf{p}_j^{l_2}||^2, \qquad (5)$$

where D_p is a set of aligned dependency-phrase pairs extracted from the parallel corpus. The regularization term is combined with the joint objective in Equation (4) to learn bilingual word embeddings (DepBiWE+R), where γ_R is a tradeoff parameter to control the contribution of the phrase-level regularization term.

3.3 Integration of Semantic Spaces

Dependency parse-trees can be regarded as the supervised information from corpus which is valuable yet expensive to obtain, and only applies to small-scale data. This prohibits the dependency-based bilingual word embedding model from being applied to large-scale corpus. On the other hand, the quality of the parsers affects the performance of dependency-based embedding methods. Fortunately, the BoW-based embeddings learned from large-scale monolingual corpus can be incorporated as unsupervised information without parsers, which can be combined with the supervised dependency-based embeddings via joint learning and make the bilingual word embedding model more robust to parsing error.

l_1-l_2	#S	#l_1-W	#l_2-W	#l_1-V	#l_2-V
en-de	1.9M	55M	52M	40k	50k
en-fr	2.0M	50M	51M	40k	50k
en-es	1.9M	49M	51M	40k	50k

Table 1: The size of the parallel corpus of three language pairs after preprocessing the data. #S denotes the number of sentence pairs, and #l_i-W represents the number of tokens of the parallel corpus in language l_i, while #l_i-V is the vocabulary size.

Specifically, the dependency-based bilingual embedding matrix $\mathbf{W}_s$ learned with supervised dependency parse-tree information and the BoW-based monolingual embedding matrix $\mathbf{W}_u$ learned from large-scale unsupervised data represent two different semantic vector spaces respectively. To integrate the two different semantic spaces for the better word representations, we propose a joint learning scheme to encourage the model to learn similar representations in both $\mathbf{W}_s$ and $\mathbf{W}_u$. Two corresponding context matrices $\mathbf{C}_s$ and $\mathbf{C}_u$ are learned simultaneously by optimizing the joint objective,

$$L_C = (L_{w_u,c_u} + L_{w_s,c_s}) + \gamma_C(L_{w_u,c_s} + L_{w_s,c_u}) \qquad (6)$$

where w_u and w_s denote two different representations of the same target word w, while c_u and c_s correspond to the BoW context and the DEP context of the target word respectively. L_{w_u,c_u} and L_{w_s,c_s} are the loss functions corresponding to BoW-based and dependency-based bilingual embedding learning respectively, while L_{w_u,c_s} and L_{w_s,c_u} are the loss functions integrating the supervised dependency-based embeddings and the BoW-based embeddings learned from large-scale monolingual corpus, which encourage the model to learn similar representations in both $\mathbf{W}_s$ and $\mathbf{W}_u$. γ_C is a tradeoff parameter of the integrated model DepBoW.

4 Experiments

4.1 Data and Setup

We train our dependency-based bilingual models for the English-German (en-de), English-French (en-fr) and English-Spanish (en-es) language pairs on the Europarl v7 parallel corpus[1] [Koehn, 2005]. To preprocess the dataset, we lowercase and tokenize all words and select the top words according to their term frequencies in the training corpus. The words with low frequencies for all languages are mapped to <unk>. The statistics of the parallel corpus for all language pairs are summarized in Table 1.

In our experiments, the Europarl corpus is used for both monolingual training and bilingual training. Parameters for bilingual embedding learning are set as suggested in BiSkip [Luong *et al.*, 2015] and fixed for all experiments. The subsampling rate, negative sampling size are set to $1e$-4 and 30 respectively; the default learning rate of Stochastic Gradient Decent (SGD) is set to 0.025 and gradually decreases to $2.5e$-6 when training is finished. The dimensionality of all embedding vectors d is set to 200, and experiments are run for 10 epochs. We set the monolingual weight α and bilingual weight β in Equation (4) to 1.0 and 4.0 respectively, with the regularization weight γ_R =0.1. Word alignments are obtained with FastAlign [Dyer *et al.*, 2013], and a python library spaCy[2] is employed to produce the dependency parse-trees for all languages in the parallel corpus for the dependency-based models.

We compare our proposed bilingual word embedding models based on syntactic dependencies with baselines including SGNS [Mikolov *et al.*, 2013c] and DepWE [Levy

[1] http://www.statmt.org/europarl/

[2] https://spacy.io/docs/usage/dependency-parse