CLUSE: Cross-Lingual Unsupervised Sense Embeddings

Ta-Chung Chi and Yun-Nung Chen
National Taiwan University, Taipei, Taiwan
r06922028@ntu.edu.tw y.v.chen@ieee.org

Abstract

This paper proposes a modularized sense induction and representation learning model that jointly learns bilingual sense embeddings that align well in the vector space, where the cross-lingual signal in the English-Chinese parallel corpus is exploited to capture the collocation and distributed characteristics in the language pair. The model is evaluated on the Stanford Contextual Word Similarity (SCWS) dataset to ensure the quality of monolingual sense embeddings. In addition, we introduce Bilingual Contextual Word Similarity (BCWS), a large and high-quality dataset for evaluating cross-lingual sense embeddings, which is the first attempt of measuring whether the learned embeddings are indeed aligned well in the vector space. The proposed approach shows the superior quality of sense embeddings evaluated in both monolingual and bilingual spaces.[1]

1 Introduction

Word embeddings have recently become the basic component in most NLP tasks for its ability to capture semantic and distributed relationships learned in an unsupervised manner. The higher similarity between word vectors can indicate similar meanings of words. Therefore, embeddings that encode semantics have been shown to serve as the good initialization and benefit several NLP tasks. However, word embeddings do not allow a word to have different meanings in different contexts, which is a phenomenon known as polysemy. For example, "*apple*" may have different meanings in *fruit* and *technology* contexts. Several attempts have been proposed to tackle this problem by inferring multi-sense word representations (Reisinger and Mooney, 2010; Neelakantan et al., 2014; Li and Jurafsky, 2015; Lee and Chen, 2017).

These approaches relied on the "one-sense per collocation" heuristic (Yarowsky, 1993), which assumes that presence of nearby words correlates with the sense of the word of interest. However, this heuristic provides only a weak signal for discriminating sense identities, and it requires a large amount of training data to achieve competitive performance.

Considering that different senses of a word may be translated into different words in a foreign language, Guo et al. (2014) and Šuster et al. (2016) proposed to learn multi-sense embeddings using this additional signal. For example, "*bank*" in English can be translated into *banc* or *banque* in French, depending on whether the sense is financial or geographical. Such information allows the model to identify which sense a word belongs to. However, the drawback of these models is that the trained foreign language embeddings are not aligned well with the original embeddings in the vector space.

This paper addresses these limitations by proposing a bilingual modularized sense induction and representation learning system. Our learning framework is the first pure sense representation learning approach that allows us to utilize two different languages to disambiguate words in English. To fully use the linguistic signals provided by bilingual language pairs, it is necessary to ensure that the embeddings of each foreign language are related to each other (i.e., they align well in the vector space). We solve this by proposing an algorithm that jointly learns sense representations between languages. The contributions of this paper are four-fold:

- We propose the first system that maintains purely sense-level cross-lingual representation learning with linear-time sense decoding.
- We are among the first to propose a single ob-

[1]The code and dataset are available at http://github.com/MiuLab/CLUSE.

jective for modularized bilingual sense embedding learning.

- We are the first to introduce a high-quality dataset for directly evaluating bilingual sense embeddings.
- Our experimental results show the state-of-the-art performance for both monolingual and bilingual contextual word similarities.

2 Related Work

There are a lot of prior works focusing on representation learning, while this work mainly focuses on bridging the work about sense embeddings and cross-lingual embeddings and introducing a newly collected bilingual data for better evaluation.

Sense Embeddings Reisinger and Mooney (2010) first proposed multi-prototype embeddings to address the lexical ambiguity when using a single embedding to represent multiple meanings of a word. Huang et al. (2012); Neelakantan et al. (2014); Li and Jurafsky (2015); Bartunov et al. (2016) utilized neural networks as well as the Bayesian non-parametric method to learn sense embeddings. Lee and Chen (2017) first utilized a reinforcement learning approach and proposed a modularized framework that separates learning of senses from that of words. However, none of them leverages the bilingual signal, which may be helpful for disambiguating senses.

Cross-Lingual Word Embeddings Klementiev et al. (2012) first pointed out the importance of learning cross-lingual word embeddings in the same space and proposed the cross-lingual document classification (CLDC) dataset for extrinsic evaluation. Gouws et al. (2015) trained directly on monolingual data and extracted a bilingual signal from a smaller set of parallel data. Kočiskỳ et al. (2014) used a probabilistic model that simultaneously learns alignments and distributed representations for bilingual data by marginalizing over word alignments. Hermann and Blunsom (2014) learned word embeddings by minimizing the distances between compositional representations between parallel sentence pairs. Šuster et al. (2016) reconstructed the bag-of-words representation of semantic equivalent sentence pairs to learn word embeddings. Shi et al. (2015) proposed a training algorithm in the form of matrix decomposition, and induced cross-lingual constraints for simultaneously factorizing monolingual matrices. Luong et al. (2015) extended the skip-gram model to bilingual corpora where contexts of bilingual word pairs were jointly predicted. Wei and Deng (2017) proposed a variational autoencoding approach that explicitly models the underlying semantics of the parallel sentence pairs and guided the generation of the sentence pairs. Although the above approaches aimed to learn cross-lingual embeddings jointly, they fused different meanings of a word in one embedding, leading to lexical ambiguity in the vector space model.

Cross-Lingual Sense Embeddings Guo et al. (2014) adopted the heuristics where different meanings of a polysemous word usually can be represented by different words in another language and clustered bilingual word embeddings to induce senses. Šuster et al. (2016) proposed an encoder, which uses parallel corpora to choose a sense for a given word, and a decoder that predicts context words based on the chosen sense. Bansal et al. (2012) proposed an unsupervised method for clustering the translations of a word, such that the translations in each cluster share a common semantic sense. Upadhyay et al. (2017) leveraged cross-lingual signals in more than two languages. However, they either used pretrained embeddings or learned only for the English side, which is undesirable since cross-lingual embeddings shall be jointly learned such that they aligned well in the embedding space.

Evaluation Datasets Several datasets can be used to justify the performance of learned sense embeddings. Huang et al. (2012) presented SCWS, the first and only dataset that contains word pairs and their sentential contexts for measuring the quality of sense embeddings. However, it is a monolingual dataset constructed in English, so it cannot evaluate cross-lingual semantic word similarity. On the other hand, while Camacho-Collados et al. (2017) proposed a cross-lingual semantic similarity dataset, it ignored the contextual words but kept only word pairs, making it impossible to judge sense-level similarity. In this paper, we present an English-Chinese contextual word similarity dataset in order to benchmark the experiments about bilingual sense embeddings.

3 CLUSE: Cross-Lingual Unsupervised Sense Embeddings

Our proposed model borrows the idea about modularization from Lee and Chen (2017), which treats the sense induction and representation mod-

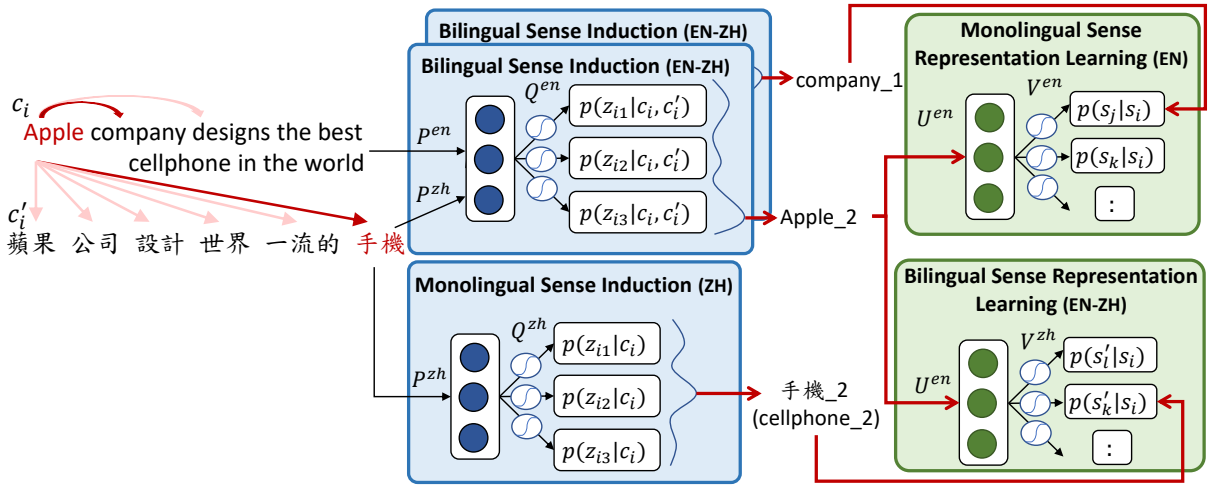

Figure 1: Sense induction modules decide the senses of words, and two sense representation learning modules optimize the sense collocated likelihood for learning sense embeddings within a language and between two languages. Two languages are treated equally and optimized iteratively.

ules separately to avoid mixing word-level and sense-level embeddings together.

Our model consists of four different modules illustrated in Figure 1, where sense induction modules decide the senses of words, and two sense representation learning modules optimize the sense collocated likelihood for learning sense embeddings within a language and between two languages in a joint manner. All modules are detailed below.

3.1 Notations

We denote our parallel corpus without word alignment C, where C^{en} is for the English part and C^{zh} is for the Chinese part. Our English vocabulary is W^{en} and Chinese vocabulary is W^{zh}. Moreover, C_t^{en} and C_t^{zh} are the t-th sentence-level parallel sentences in English and Chinese respectively. In the following sections, we treat English as the major language and Chinese as an additional bilingual signal, while their roles can be mutually exchanged. Specifically, English and Chinese iteratively become the major language during the training procedure.

3.2 Bilingual Sense Induction Module

The bilingual sense induction module takes a parallel sentence pair as input and determines which sense identity a target word belongs to given the bilingual contextual information. Formally, for the t-th English sentence C_t^{en}, we aim to decode the most probable sense $z_{ik} \in Z_i$ for the i-th word $w_i \in W^{en}$ in C_t^{en}, where Z_i is the set of sense candidates for w_i and $1 \leq k \leq |Z_i|$. We assume that the meaning of w_i can be determined by its surrounding words, or the so-called local context, $c_i = \{w_{i-m}, \cdots, w_{i+m}\}$, where m is the size of context window.

Aside from monolingual information, it is desirable to exploit the parallel sentences as additional bilingual contexts to enable cross-lingual embedding learning. Note that word alignment is not required in this work, so we consider the whole parallel bilingual sentence during training. Considering training efficiency, we sample M words in the parallel bilingual sentence with their original relative order or pad it to M for those shorter than M. Formally, given the t-th parallel bilingual sentence C_t^{zh}, the bilingual context of w_i is therefore $c_i' = \{w_0', \cdots, w_{M-1}'\}$ and $w' \in W^{zh}$.

To ensure efficiency, continuous bag-of-words (CBOW) model is applied, where it takes word-level input tokens and outputs sense-level identities. Specifically, given an English *word* embedding matrix P^{en}, the local context can be modeled as the average of word embeddings from its context, $\frac{1}{|c_i|}\sum_{w_j \in c_i} P_j^{en}$. Similarly, we can model the bilingual contextual information given Chinese word embedding matrix P^{zh} using the CBOW formulation and obtain $\frac{1}{M}\sum_{w_j' \in c_i'} P_j^{zh}$. We linearly combine the contextual information from different languages as:

$$\bar{C} = \alpha \cdot \frac{1}{|c_i|} \sum_{w_j \in c_i} P_j^{en} + (1-\alpha) \cdot \frac{1}{M} \sum_{w_j' \in c_i'} P_j^{zh}. \quad (1)$$

The likelihood of selecting each sense identity z_{ik} for w_i can be formulated in the form of *Bernoulli* distribution with a sigmoid function $\sigma(\cdot)$:

$$p(z_{ik} \mid c_i, c_i') = \sigma((Q_{ik}^{en})^T \bar{C}), \quad (2)$$

where Q^{en} is a 3-dimensional tensor with each dimension denotes W^{en}, z_{ik} for a specific word i in W^{en}, and the corresponding latent variable, respectively. Therefore, Q_{ik}^{en} will retrieve the latent variable of k-th sense of i-th English word. Finally, we can induce the sense identity, z_{ik}^*, given the contexts of a word w_i from different languages, c_i and c_i'.

$$z_{ik}^* = \arg\max_{z_{ik}} p(z_{ik} \mid c_i, c_i') \quad (3)$$

In order to allow the module to explore other potential sense identities, we apply an ϵ-greedy algorithm (Mnih et al., 2013) for exploration in the training procedure.

3.3 Monolingual Sense Induction Module

This module is the degraded version of bilingual sense induction module when $\alpha = 1$, which occurs where *no* parallel bilingual signal exists. In other words, every bilingual sense induction module will experience the degradation during the training process presented in Algorithm 1. The only difference is that it cannot access the bilingual information. The purpose of this module is to maintain the stability of sense induction and to decode the sampled bilingual sense identity which will later be used in the bilingual sense representation learning module. As shown in Figure 1, given the monolingual context of a word, this module selects its sense identity using (2) and (3) with $\alpha = 1$.

3.4 Monolingual Sense Representation Learning Module

Given the decoded sense identities from the sense induction module, the skip-gram architecture (Mikolov et al., 2013) is applied considering that it only requires two decoded sense identities for stochastic training. We first create an input English sense representation matrix U^{en} and an English collocation estimation matrix V^{en} as the learning targets. Given a target word w_i and its collocated word w_j in the t-th English sentence C_t^{en}, we map them to their sense identities as $z_{ik}^* = s_i$ and $z_{jl}^* = s_j$ by the sense induction module and maximize the sense collocation likelihood. The skip-gram objective can be formulated as $p(s_j \mid s_i)$:

$$p(s_j \mid s_i) = \frac{\exp((U_{s_i}^{en})^T V_{s_j}^{en})}{\sum_{s_k} \exp((U_{s_i}^{en})^T V_{s_k}^{en})}, \quad (4)$$

where s_k iterates over all possible English sense identities in the denominator. This formulation shares the same architecture as skip-gram but extends to rely on senses. Note that the Chinese sense representation learning module is built similarly.

3.5 Bilingual Sense Representation Learning Module

To ensure sense embeddings of two different languages align well, we hypothesize that the target sense identity s_i not only predicts the sense identity s_j of w_j in C_t^{en} but also one sampled sense identity s_l' of w_l' from the parallel sentence C_t^{zh}, where s_l' is decoded by the Chinese monolingual sense induction module. Specifically, the *bilingual* skip-gram objective can be formulated using the English sense embedding matrix U^{en} and the bilingual collocation estimation matrix V^{zh} as:

$$p(s_l' \mid s_i) = \frac{\exp((U_{s_i}^{en})^T V_{s_l'}^{zh})}{\sum_{s_k'} \exp((U_{s_i}^{en})^T V_{s_k'}^{zh})}, \quad (5)$$

where s_k' iterates over all possible Chinese sense identities in the denominator.

3.6 Joint Learning

In this learning framework, the gradient cannot be back-propagated from the representation module to the induction module due to the usage of $\arg\max$ operator. It is therefore desirable to connect these two modules in a way such that they can improve each other by their own estimations. In one direction, forwarding the prediction of the sense induction module to the sense representation learning module is trivial, while in another direction, we treat the estimated collocation likelihood as the reward for the induction module.

First note that calculating the partition function in the denominator of (4) and (5) is intractable since it involves a computationally expensive summation over all sense identities. In practice, we adopt the negative sampling strategy technique (Mikolov et al., 2013) and rewrite (4) and (5) as:

$$\log p(s_j \mid s_i) = \log \sigma((U^{en}_{s_i})^T V^{en}_{s_j}) + \sum_{k=1}^{N} \mathbb{E}_{s_k \sim p_{\text{neg}}(s)}[\sigma(-(U^{en}_{s_i})^T V^{en}_{s_k})], \tag{6}$$

$$\log p(s'_l \mid s_i) = \log \sigma((U^{en}_{s_i})^T V^{zh}_{s'_l}) + \sum_{k=1}^{N} \mathbb{E}_{s'_k \sim p_{\text{neg}}(s')}[\sigma(-(U^{en}_{s_i})^T V^{zh}_{s'_k})], \tag{7}$$

where $p_{\text{neg}}(s)$ and $p_{\text{neg}}(s')$ is the distribution over all English senses and all Chinese senses for negative samples respectively, and N is the number of negative sample. The rewritten objective for optimizing two sense representation learning modules is the same as maximizing (6) and (7). Moreover, we can utilize the probability of correctly classifying the skip-gram sense pair as the reward signal. The intuition is that a correctly decoded sense identity is more likely to predict its neighboring sense identity compared to incorrectly decoded ones.

This learning framework can now be viewed as a reinforcement learning agent solving one-step Markov Decision Process (Sutton and Barto, 1998; Lee and Chen, 2017). For bilingual modules, the state, action, and reward correspond to bilingual context $\bar{C}$, sense z_{ik}, and $\sigma((U^{en}_{s_i})^T V^{zh}_{s'_l})$ respectively. As for the monolingual modules, the state, action, and reward correspond to monolingual context c_t, sense z_{ik}, and $\sigma((U^{en}_{s_i})^T V^{en}_{s_j})$. Finally, we can optimize both bilingual and monolingual sense induction modules (P and Q from (2) by minimizing the cross entropy loss between decoded sense probability and reward. We also include an entropy regularization term as suggested in (Šuster et al., 2016) to let the sense induction module converge faster and make more confident predictions. Formally,

$$\min H(\sigma((U^{en}_{s_i})^T V^{zh}_{s'_l}), p(z_{ik} \mid c_i, c'_i)) + \lambda E(p(z_{ik} \mid c_i, c'_i)) \tag{8}$$

$$\min H(\sigma((U^{en}_{s_i})^T V^{en}_{s_j}), p(z_{ik} \mid c_i)) + \lambda E(p(z_{ik} \mid c_i)) \tag{9}$$

E is the entropy of selection probability weighted by λ. Note that the major language is switched iteratively among two languages. Algorithm 1 presents the full learning procedure.

Algorithm 1 Bilingual Sense Embedding Learning Algorithm

Input: $C^{en}, C^{zh}, W^{en}, W^{zh}$
Output: $P^{en}, P^{zh}, Q^{en}, Q^{zh}, U^{en}, U^{zh}, V^{en}, V^{zh}$

loop until converge
  MAIN(*en*, *zh*, 0.4) ▷ 0.4 is just an example weight
  MAIN(*zh*, *en*, 0.4)
end loop
function MAIN(maj, bi, α)
  $t, i, j, k, l \leftarrow$ GETTRAINDATA(maj)
  $s_i, pred_i \leftarrow$ INDUCESENSE(maj, bi, t, i, α)
  $s_j, \_ \leftarrow$ INDUCESENSE(maj, bi, t, j, α)
  $s'_l, pred'_l \leftarrow$ INDUCESENSE(bi, bi, t, k, 1.0)
  $s'_k, \_ \leftarrow$ INDUCESENSE(bi, bi, t, l, 1.0)
  $r \leftarrow$ TRAINSRL(maj, maj, s_i, s_j)
  $r' \leftarrow$TRAINSRL(maj, bi, s_i, s'_l)
  $r'' \leftarrow$TRAINSRL(bi, bi, s'_l, s'_k)
  TRAINSI(maj, bi, $r, pred_i$)
  TRAINSI(maj, bi, $r', pred_i$)
  TRAINSI(bi, bi, $r'', pred'_l$)
end function
function INDUCESENSE(maj, bi, t, i, α)
  calculate α-weighted $\bar{C}$ by (1)
  select z^*_{ik} by (2) and (3)
  return $z^*_{ik}, p(z^*_{ik} \mid \bar{C})$
end function
function TRAINSRL(maj, bi, s_i, s_j)
  if maj==bi **then**
    optimize U^{maj}, V^{maj} by (6) given s_i, s_j
  else
    optimize U^{maj}, V^{bi} by (7) given s_i, s_j
  end if
  return collocation prob of (s_i, s_j)
end function
function TRAINSI(maj, bi, r, pred)
  if maj==bi **then**
    optimize P^{maj}, Q^{maj} by (9) given r, pred
  else
    optimize P^{maj}, Q^{bi} by (8) given r, pred
  end if
end function

4 New Dataset—Bilingual Contextual Word Similarity (BCWS)

We propose a new dataset to measure the bilingual contextual word similarity. English and Chinese are chosen as our language pair for three reasons:

1. They are the top widely used languages in the world.
2. English and Chinese belong to completely different language families, making it interesting to explore syntactic and semantic difference among them.
3. Chinese is a language that requires segmentation, this dataset can also help researchers experiment on different segmentation levels and investigate how segmentation affects the

English Sentence	Chinese Sentence	Score
Judges must give both sides an equal opportunity to <**state**> their cases.	我非常喜歡這個故事，它<**告訴**>我們一些重要的啓示。 (I like this story a lot, which <**tells**> us some important inspiration.)	7.00
It was of negligible <**importance**> prior to 1990, with antiquated weapons and few members.	黃斑部病變的預防及早期治療是相當<**重要**>的。 (The prevention and early treatment of macular lesions is very <**important**>.)	6.94
Due to the San Andreas Fault bisecting the hill, one side has <**cold**> water, the other has hot.	水果攤老闆似乎很意外眞有人買這<**冷**>貨，露出「你眞內行」的眼神與我聊了幾句。 (The owner of the fruit stall seemed surprised that someone bought this <**unpopular**> product, talking me few words about "you are such a pro".)	3.70

Table 1: Sentence pair examples and average annotated scores in BCWS.

sense similarity.

This dataset also provides a *direct* measure to determine whether the two language embeddings align well in the vector space. Note that we focus on word-level, and this is different from (Klementiev et al., 2012), which also measured the cross-lingual embedding similarity but rely on the ambiguous document-level classification.

Our dataset contains 2091 question pairs, where each pair consists of exactly one English and one Chinese sentence; note that they are **not** parallel but with their own sentential contexts shown in Table 1. Eleven raters[2] were recruited to annotate this dataset. Each rater gives a score ranging from 1.0 (different) to 10.0 (same) for each question to indicate the semantic similarity of bilingual word pairs based on sentential clues. The annotated dataset shows very high intra-rater consistency; we leave one rater out and calculate Spearman correlation between the rater and the average of the rest, and the average number is about 0.83, indicating the human-level performance (the average number in SCWS is 0.52).

We describe the construction of BCWS below.

Chinese Multi-Sense Word Extraction We utilize the Chinese Wikipedia dump to extract the most frequent 10000 Chinese words that are *nouns*, *adjective*, and *verb* based on Chinese Wordnet (Huang et al., 2010). In order to test the sense-level representations, we discard single-sense words to ensure that the selected words are polysemous. Also, the words with more than 20 senses are deleted, since those senses are too fine-grained and even hard for human to disambiguate. We denote the list of Chinese words l_c.

English Candidate Word Extraction We have to find an English counterpart for each Chinese word in l_c. We utilize *BabelNet* (Navigli and Ponzetto, 2010), a free and open-sourced knowledge resource, to serve as our bilingual dictionary. To be more concrete, we first query the selected Chinese word using the free API call provided by Babelnet to retrieve all *WordNet* senses[3]. For example, the Chinese word "制服" has two major meanings:

- *a type of clothing worn by members of an organization*
- *force to submit or subdue*.

Hence, we can obtain two candidate English words "*uniform*" and "*subjugate*". Each word in l_c retrieves its associated English candidate words and obtain the dictionary D.

Enriching Semantic Relationship Note that D is merely a simple translation mapping between Chinese and English words. It is desirable that we have a more complicated and interesting relationship between bilingual word pairs. Hence, we traverse D and for each English word we find its *hyponyms*, *hypernyms*, *holonyms* and *attributes*, and add the additional words into D. In our example, we may obtain {制服:[uniform, subjugate, livery, clothing, repress, dominate, enslave, dragoon...]}. We sample 2 English words if the number of English candidate words is more than 5, 3 English words if more than 10, and 1 English word oth-

[2]They are all Chinese native speaker whose scores are at least 29 in the TOEFL reading section or 157 in the GRE verbal section.

[3]*BabelNet* contains sense definitions from various resources such as Wordnet, Wikitionary, Wikidata, etc

erwise to form the final bilingual pair. For example, a bilingual word pair (制服, enslave) can be formed accordingly. After this step, we obtain 2091 bilingual word pairs P.

Adding Contextual Information Given the bilingual word pairs P, appropriate contexts should be found in order to form the full sentences for human judgment. For each Chinese word, we randomly sample one example sentence in Chinese WordNet that matches the PoS tag we selected in section 4. For each English word, we traverse the whole English Wikipedia dump to find the sentences that contain the target English word. We then sample one sentence where the target word is tagged as the matched PoS tag[4].

5 Experiments

5.1 Experimental Setup

Two sets of parallel data are used in the experiments, one for English-Chinese (EN-ZH) and another for English-German (EN-DE). UM-corpus (Tian et al.) is used for EN-ZH training, while Europarl corpus (Koehn, 2005) is used for EN-DE training. UM-corpus contains 15,764,200 parallel sentences with 381,921,583 English words and 572,277,658 unsegmented Chinese words. Europarl contains 1,920,209 parallel sentences with 44,548,491 German words and 47,818,827 English words. We evaluate our proposed model on the benchmark monolingual dataset, SCWS, and on the bilingual dataset, our proposed BCWS, where the evaluation metrics are actually introduced in section 5.4.

5.2 Hyperparameter Settings

In our experiments, we use a mini-batch size of 512, context window size for major language is set to $m = 5$ and we sample $M = 20$ words for bilingual context. For the exploration of sense induction module, we set $\epsilon = 0.05$. The λ of entropy regularization is set to 1.[5] For negative sampling in (6) and (7), we pick $N = 25$. The fixed learning rate is set to 0.025. The embedding dimension is 300 and the sense number per word is set to 3 for both Chinese, German, and English ($|Z_i| = 3$). This setting is for a fair comparison with prior works.

[4]We use the NLTK PoS tagger to obtain the tags.

[5]We tried different values of $\lambda = 0.001, 0.5$, and the model converges approximately 12, 5 times slower compared to $\lambda = 1$.

5.3 Baseline

The baselines for comparison can be categorized into three:

- *Monolingual sense embeddings*: Lee and Chen (2017) is the current state-of-the-art model of monolingual sense embedding evaluated on SCWS. We re-train the sense embeddings using the same data but only in English for fair comparison.
- *Cross-lingual word embeddings*: Luong et al. (2015) treated words from different languages the same and trained cross-lingual embeddings in the same space. Conneau et al. (2017) utilized adversarial training to map pretrained word embeddings into another language space.
- *Cross-lingual sense embeddings*: Upadhyay et al. (2017) utilized more than two languages to learn multilingual embeddings. We report the number shown in the paper for comparison.

5.4 Evaluation Metric

Reisinger and Mooney (2010) introduced two contextual similarity estimations, AvgSimC and MaxSimC. AvgSimC is a *soft* measurement that addresses the contextual information with a probability estimation:

$$\text{AvgSimC}(w_i, \bar{C}_t, w_j, \bar{C}_{t'}) = \sum_{k=1}^{|Z_i|}\sum_{l=1}^{|Z_j|} \pi(z_{ik}|\bar{C}_t)\pi(z_{jl}|\bar{C}_{t'})d(z_{ik}, z_{jl}),$$

AvgSimC weights the similarity measurement of each sense pair z_{ik} and z_{jl} by their probability estimations. On the other hand, MaxSimC is a *hard* measurement that only considers the most probable senses:

$$\text{MaxSimC}(w_i, \bar{C}_t, w_j, \bar{C}_{t'}) = d(z_{ik}, z_{jl}),$$
$$z_{ik} = \arg\max_{z_{ik'}} \pi(z_{ik'}|\bar{C}_t),$$
$$z_{jl} = \arg\max_{z_{jl'}} \pi(z_{jl'}|\bar{C}_{t'}).$$

$d(z_{ik}, z_{jl})$ refers to the cosine similarity between $U^{maj}_{z_{ik}}$ and $U^{bi}_{z_{jl}}$ in the bilingual case (BCWS) and $U^{maj}_{z_{ik}}$ and $U^{maj}_{z_{jl}}$ in the monolingual case (SCWS).

5.5 Bilingual Embedding Evaluation

Cross-lingual sense embeddings are the main contribution of this paper. Table 2 shows that all results from the proposed model are significantly

Model	α	EN-ZH		EN-DE
		Bilingual/BCWS	Mono(EN)/SCWS	Mono(EN)/SCWS
1) Monolingual Sense Embeddings				
Lee and Chen (2017)			**66.8** / 65.5	63.8 / 63.4
2) Cross-Lingual Word Embeddings				
Luong et al. (2015)		50.4	61.1	62.1
Conneau et al. (2017)		54.7	65.5	64.0
3) Cross-Lingual Sense Embeddings				
Upadhyay et al. (2017)		-	45.0*	-
Proposed	0.1	58.3 / 58.3	65.8 / 65.8	63.1 / 63.3
	0.3	**58.8 / 58.8**	65.9 / 66.0	63.5 / 63.9
	0.5	58.5 / 58.5	66.7 / **67.0**	63.7 / 64.3
	0.7	58.3 / 58.4	66.3 / 66.6	63.7 / 64.1
	0.9	58.3 / 58.3	66.1 / 66.2	**63.9 / 64.6**

Table 2: Contextual similarity results evaluated on the SCWS/BCWS dataset, where the reported numbers indicate Spearman's rank correlation $\rho \times 100$ on AvgSimC / MaxSimC.* indicates that Upadhyay et al. (2017) trained the sense embeddings using a different parallel dataset.

better than the baselines that learn cross-lingual word embeddings. It indicates that the sense-level information is critical for precise vector representations. In addition, all results for AvgSimC and MaxSimC are the same in the proposed model, showing that the learned selection distribution is reliable for sense decoding.

5.6 Monolingual Embedding Evaluation

Because our model considers multiple languages and learns the embeddings jointly, the multilingual objective makes learning more difficult due to more noises. In order to ensure the quality of the monolingual sense embeddings, we also evaluate our learned English sense embeddings on the benchmark SCWS data. Comparing the results between training on EN-ZH and training on EN-DE, all results using EN-ZH are better than ones using EN-DE. The probable reason is that the language difference between English and Chinese is larger than English and German; parallel Chinese sentences therefore provide informative cues for learning better sense embeddings. Furthermore, our proposed model achieves comparable or superior performance than the current state-of-the-art monolingual sense embeddings proposed by Lee and Chen (2017) when trained on our monolingual data.

5.7 Sensitivity of Bilingual Contexts

To investigate how much the bilingual sense induction module relies on another language, the results with different α are shown in the table.

Model	EN2DE	DE2EN
1) Sentence-Level Training		
Hermann and Blunsom (2014)	83.7	71.4
AP et al. (2014)	**91.8**	72.8
Wei and Deng (2017)	91.0	**80.4**
2) Word-Level Training		
Klementiev et al. (2012)	77.7	71.1
Gouws et al. (2015)	86.5	75.0
Kočiskỳ et al. (2014)	83.1	75.4
Shi et al. (2015)	**91.3**	**77.2**
Luong et al. (2015)	86.4	75.6
Conneau et al. (2017)	78.7	67.1
Proposed	81.8	76.0

Table 3: Accuracy on cross-lingual document classification (%).

To justify the usefulness of utilizing bilingual signal, we compare our model with Lee and Chen (2017), which used monolingual signal in a similar modular framework. Our method outperforms theirs in terms of MaxSimC on both EN-ZH and EN-DE. However, the performance is roughly the same on AvgSimC. The reason may be that the bilingual signal is indicative but noisy, which largely affects AvgSimC due to its weighted sum operation. MaxSimC only picks the most probable senses, which makes it robust to noises.

In addition, our performance improves as α increases for EN-DE, and the best is obtained when α is large. This is interesting if we compare

Target	kNN Senses (EN)	kNN Senses (ZH)
apple_0	fruit, cake, sweet	蘋果, 春天, 蛋糕, iphone, 雞蛋, 巧克力, 葡萄 (apple, spring, cake, iphone, egg, chocolate, purples)
apple_1	iphone, cake, google, stores	蘋果, iphone, 微軟, 競爭對手, 春天, 谷歌 (apple, iphone, microsoft, competitor, spring, google)
uniform_0	dressed, worn, tape, wearing, cloth	均勻,光滑,衣服,鞋子,穿著,服裝 (even, smooth, clothes, shoes, wearing, clothing)
uniform_1	particle, computed, varying, gradient	態,粉末,縱向,等離子體,剪切,剛度 (phase, powder, longitudinal, plasma, cut, stiffness)

Table 4: Words with similar senses obtained by kNN.

$\alpha = 0.9$ to *MUSE*, we can see that AvgSimC is similar but ours outperforms *MUSE* on MaxSimC, indicating this little bilingual signal does help disambiguate senses more confidently. In contrast, the best performance is obtained on EN-ZH when two languages have equal contribution, because English is very different from Chinese, such that it can benefit more from Chinese than from German.

5.8 Extrinsic Evaluation

We further evaluate our bilingual sense embeddings using a downstream task, cross-lingual document classification (CLDC), with a standard setup (Klementiev et al., 2012). To be more concrete, a set of labeled documents in language *A* is available to train a classifier, and we are interested in classifying documents in another language *B* at test time, which tests semantic transfer of information across different languages. We use the averaged sense embeddings as word embeddings for a fair comparison.

The result is shown in Table 3. We can see that our proposed model achieves comparable performance or even superior performance to most prior work on the DE2EN direction; however, the same conclusion does not hold for the EN2DE direction. The reason may be that we test the model that works best on BCWS and hence not able to tune hyperparameters on the development set of CLDC. In addition, we use the average of sense vectors as input word embeddings, which may induce some noises into the resulting vectors. In sum, the comparable performance of the downstream task shows the practical usage and the potential extension of the proposed model.

5.9 Qualitative Analysis

Some examples of our learned sense embeddings are shown in Table 4. It is obvious to see that the first sense of *Apple* is related to *fruit and things to eat*, while the second one means the *tech company Apple Inc*. Most English and Chinese nearest neighbors match the meanings of the induced senses, but there are still some noises that are underlined. For example, *cake* should be the neighbor of the first sense rather than the second one. The same observation applies to *iphone* and *spring*. In our second example for *uniform*, the first sense is related to *outfit and clothes*, while the second is related to *engineering terms*. However, *even* appears in the *outfit and clothes* sense, which is incorrect. The reason may be that the size of the parallel corpus is not large enough for the model to accurately distinguish all senses via unsupervised learning. Hence, utilizing external resources such as bilingual dictionaries or designing a new model that can use existing large monolingual corpora like Wikipedia are our future work.

6 Conclusion

This paper is the first purely sense-level cross-lingual representation learning model with efficient sense induction, where several monolingual and bilingual modules are jointly optimized. The proposed model achieves superior performance on both bilingual and monolingual evluation datasets. A newly collected dataset for evaluating bilingual contextual word similarity is presented, which provides potential research directions for future work.

Acknowledgement

We would like to thank reviewers for their insightful comments on the paper. This work was finan-

cially supported from the Young Scholar Fellowship Program by Ministry of Science and Technology (MOST) in Taiwan, under Grant 107-2636-E-002-004.

References

Sarath Chandar AP, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems*, pages 1853–1861.

Mohit Bansal, John DeNero, and Dekang Lin. 2012. Unsupervised translation sense clustering. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 773–782. Association for Computational Linguistics.

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. 2016. Breaking sticks and ambiguities with adaptive skip-gram. In *Artificial Intelligence and Statistics*, pages 130–138.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *International Conference on Machine Learning*, pages 748–756.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning sense-specific word embeddings by exploiting bilingual resources. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 497–507.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. *arXiv preprint arXiv:1404.4641*.

Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese wordnet: Design, implementation, and application of an infrastructure for cross-lingual knowledge processing. *Journal of Chinese Information Processing*, 24(2):14–23.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving Word Representations via Global Context and Multiple Word Prototypes. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

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

Tomáš Kočiskỳ, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning bilingual word representations by marginalizing alignments. *arXiv preprint arXiv:1405.0947*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Guang-He Lee and Yun-Nung Chen. 2017. MUSE: Modularizing unsupervised sense embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 327–337.

Jiwei Li and Dan Jurafsky. 2015. Do multi-sense embeddings improve natural language understanding? *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1722–1732.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in neural information processing systems*, pages 3111–3119.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *NIPS Deep Learning Workshop*.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Joseph Reisinger and Raymond J Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117. Association for Computational Linguistics.

Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. 2015. Learning cross-lingual word embeddings via matrix co-factorization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 567–572.

Simon Šuster, Ivan Titov, and Gertjan van Noord. 2016. Bilingual learning of multi-sense embeddings with discrete autoencoders. In *Proceedings of NAACL-HLT*, pages 1346–1356.

Richard S Sutton and Andrew G Barto. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Liang Tian, Derek F Wong, Lidia S Chao, Paulo Quaresma, and Francisco Oliveira. Um-corpus: A large english-chinese parallel corpus for statistical machine translation.

Shyam Upadhyay, Kai-Wei Chang, Matt Taddy, Adam Kalai, and James Zou. 2017. Beyond bilingual: Multi-sense word embeddings using multilingual context. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 101–110.

Liangchen Wei and Zhi-Hong Deng. 2017. A variational autoencoding approach for inducing cross-lingual word embeddings. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4165–4171. AAAI Press.

David Yarowsky. 1993. One sense per collocation. In *Proceedings of the workshop on Human Language Technology*, pages 266–271. Association for Computational Linguistics.