Bilingual Learning of Multi-sense Embeddings with Discrete Autoencoders

Simon Šuster
University of Groningen
Netherlands
s.suster@rug.nl

Ivan Titov
University of Amsterdam
Netherlands
titov@uva.nl

Gertjan van Noord
University of Groningen
Netherlands
g.j.m.van.noord@rug.nl

Abstract

We present an approach to learning multi-sense word embeddings relying both on monolingual and bilingual information. Our model consists of an encoder, which uses monolingual and bilingual context (i.e. a parallel sentence) to choose a sense for a given word, and a decoder which predicts context words based on the chosen sense. The two components are estimated jointly. We observe that the word representations induced from bilingual data outperform the monolingual counterparts across a range of evaluation tasks, even though crosslingual information is not available at test time.

1 Introduction

Approaches to learning word embeddings (i.e. real-valued vectors) relying on word context have received much attention in recent years, and the induced representations have been shown to capture syntactic and semantic properties of words. They have been evaluated intrinsically (Mikolov et al., 2013a; Baroni et al., 2014; Levy and Goldberg, 2014) and have also been used in concrete NLP applications to deal with word sparsity and improve generalization (Turian et al., 2010; Collobert et al., 2011; Bansal et al., 2014; Passos et al., 2014). While most work to date has focused on developing embedding models which represent a word with a single vector, some researchers have attempted to capture *polysemy* explicitly and have encoded properties of each word with multiple vectors (Huang et al., 2012; Tian et al., 2014; Neelakantan et al., 2014; Chen et al., 2014; Li and Jurafsky, 2015).

In parallel to this work on multi-sense word embeddings, another line of research has investigated integrating *multilingual* data, with largely two distinct goals in mind. The first goal has been to obtain representations for several languages in the same semantic space, which then enables the transfer of a model (e.g., a syntactic parser) trained on annotated training data in one language to another language lacking this annotation (Klementiev et al., 2012; Hermann and Blunsom, 2014; Gouws et al., 2014; Chandar A P et al., 2014). Secondly, information from another language can also be leveraged to yield better first-language embeddings (Guo et al., 2014). Our paper falls in the latter, much less explored category. We adhere to the view of multilingual learning as a means of language grounding (Faruqui and Dyer, 2014b; Zou et al., 2013; Titov and Klementiev, 2012; Snyder and Barzilay, 2010; Naseem et al., 2009). Intuitively, polysemy in one language can be at least partially resolved by looking at the translation of the word and its context in another language (Kaji, 2003; Ng et al., 2003; Diab and Resnik, 2002; Ide, 2000; Dagan and Itai, 1994; Brown et al., 1991). Better sense assignment can then lead to better sense-specific word embeddings.

We propose a model that uses second-language embeddings as a supervisory signal in learning multi-sense representations in the first language. This supervision is easy to obtain for many language pairs as numerous parallel corpora exist nowadays. Our model, which can be seen as an autoencoder with a discrete hidden layer encoding word senses, leverages bilingual data in its encoding part, while the decoder predicts the surrounding words relying on the

Proceedings of NAACL-HLT 2016, pages 1346–1356,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

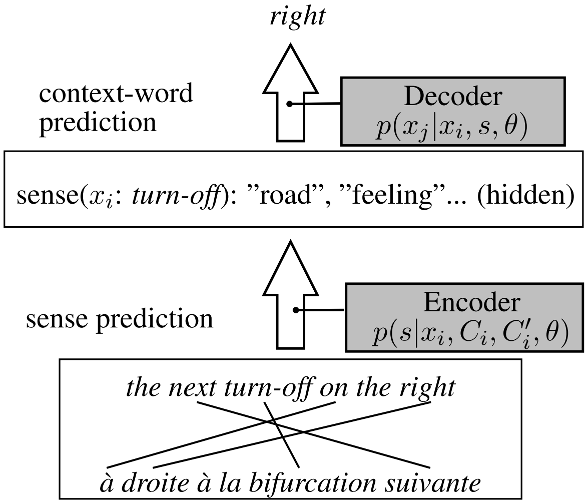

right
context-word prediction
Decoder
$p(x_j|x_i, s, \theta)$
sense(x_i: turn-off): ”road”, ”feeling”... (hidden)
sense prediction
Encoder
$p(s|x_i, C_i, C'_i, \theta)$
the next turn-off on the right
à droite à la bifurcation suivante

Figure 1: Model schema: the sense encoder with bilingual signal and the context-word predictor are learned jointly.

predicted senses. We strive to remain flexible as to the form of parallel data used in training and support both the use of word- and sentence-level alignments.

Our findings are:

- The second-language signal effectively improves the quality of multi-sense embeddings as seen on a variety of intrinsic tasks for English, with the results superior to that of the baseline Skip-Gram model, even though the crosslingual information is not available at test time.
- This finding is robust across several settings, such as varying dimensionality, vocabulary size and amount of data.
- In the extrinsic POS-tagging task, the second-language signal also offers improvements over monolingually-trained multi-sense embeddings, however, the standard Skip-Gram embeddings turn out to be the most robust in this task.

We make the implementation of all the models as well as the evaluation scripts available at `http://github.com/rug-compling/bimu`.

2 Word Embeddings with Discrete Autoencoders

Our method borrows its general structure from neural autoencoders (Rumelhart et al., 1986; Bengio et al., 2013). Autoencoders are trained to reproduce their input by first mapping their input to a (lower dimensional) hidden layer and then predicting an approximation of the input relying on this hidden layer. In our case, the hidden layer is not a real-valued vector, but is a categorical variable encoding the sense of a word. Discrete-state autoencoders have been successful in several natural language processing applications, including POS tagging and word alignment (Ammar et al., 2014), semantic role induction (Titov and Khoddam, 2015) and relation discovery (Marcheggiani and Titov, 2016).

More formally, our model consists of two components: an *encoding* part which assigns a sense to a pivot word, and a *reconstruction* (decoding) part recovering context words based on the pivot word and its sense. As predictions are probabilistic (‘soft’), the reconstruction step involves summation over all potential word senses. The goal is to find embedding parameters which minimize the error in recovering context words based on the pivot word and the sense assignment. Parameters of both encoding and reconstruction are jointly optimized. Intuitively, a good sense assignment should make the reconstruction step as easy as possible. The encoder uses not only words in the first-language sentence to choose the sense but also, at training time, is conditioning its decisions on the words in the second-language sentence. We hypothesize that the injection of crosslingual information will guide learning towards inducing more informative sense-specific word representations. Consequently, using this information at training time would benefit the model even though crosslingual information is not available to the encoder at test time.

We specify the encoding part as a log-linear model:

$$p(s|x_i, C_i, C'_i, \theta) \propto \exp\big(\varphi_{i,s}^{\top}(\frac{1-\lambda}{|C_i|}\sum_{j\in C_i}\gamma_j + \frac{\lambda}{|C'_i|}\sum_{k\in C'_i}\gamma'_k)\big). \tag{1}$$

To choose the sense $s \in \mathcal{S}$ for a word x_i, we use the bag of context words C_i from the first language l, as well as the bag of context words C'_i from the second language l'.[1] The context C_i is defined as a

[1] We have also considered a formulation which included a sense-specific bias $b_{x_i,s} \in \mathbb{R}$ to capture relative frequency of latent senses but it did not seem to affect performance.

multiset $C_i = \{x_{i-n}, \dots, x_{i-1}, x_{i+1}, \dots, x_{i+n}\}$, including words around the pivot word in the window of size n to each side. We set n to 5 in all our experiments. The crosslingual context C'_i is discussed in § 3, where we either rely on word alignments or use the entire second-language sentence as the context. We distinguish between sense-specific embeddings, denoted by $\varphi \in \mathbb{R}^d$, and generic sense-agnostic ones, denoted $\{\gamma, \gamma'\} \in \mathbb{R}^d$ for first and second language, respectively. The number of sense-specific embeddings is the same for all words. We use θ to denote all these embedding parameters. They are learned jointly, with the exception of the pre-trained second-language embeddings.

The hyperparameter $\lambda \in \mathbb{R}, 0 \leq \lambda \leq 1$ weights the contribution of each language. Setting $\lambda = 0$ would drop the second-language component and use only the first language. Our formulation allows the addition of new languages easily, provided that the second-language embeddings live in the same semantic space.

The reconstruction part predicts a context word x_j given the pivot x_i and the current estimate of its s:

$$p(x_j|x_i, s, \theta) = \frac{\exp(\varphi_{i,s}^\top \gamma_j)}{\sum_{k \in |\mathcal{V}|} \exp(\varphi_{i,s}^\top \gamma_k)}, \qquad (2)$$

where $|\mathcal{V}|$ is the vocabulary size. This is effectively a Skip-Gram model (Mikolov et al., 2013a) extended to rely on senses.

2.1 Learning and regularization

As sense assignments are not observed during training, the learning objective includes marginalization over word senses and thus can be written as:

$$\sum_i \sum_{j \in C_{x_i}} \log \sum_{s \in \mathcal{S}} p(x_j|x_i, s, \theta) p(s|x_i, C_i, C'_i, \theta),$$

in which index i goes over all pivot words in the first language, j over all context words to predict at each i, and s marginalizes over all possible senses of the word x_i. In practice, we avoid the costly computation of the normalization factor in the softmax computation of Eq. (2) and use negative sampling (Mikolov et al., 2013b) instead of $\log p(x_j|x_i, s, \theta)$:

$$\log \sigma(\varphi_{i,s}^\top \gamma_j) + \sum_{x \in N} \log \sigma(-\varphi_{i,s}^\top \gamma_x), \qquad (3)$$

where σ is the sigmoid non-linearity function and γ_x is a word embedding from the sample of negative (noisy) words N. Optimizing the autoencoding objective is broadly similar to the learning algorithm defined for multi-sense embedding induction in some of the previous work (Neelakantan et al., 2014; Li and Jurafsky, 2015). Note though that this previous work has considered only monolingual context.

We use a minibatch training regime and seek to optimize the objective function $L(\mathcal{B}, \theta)$ for each minibatch $\mathcal{B}$. We found that optimizing this objective directly often resulted in inducing very flat posterior distributions. We therefore use a form of posterior regularization (Ganchev et al., 2010) where we can encode our prior expectations that the posteriors should be sharp. The regularized objective for a minibatch is defined as

$$L(\mathcal{B}, \theta) + \lambda_H \sum_{i \in \mathcal{B}} H(q_i), \qquad (4)$$

where H is the entropy function and q_i are the posterior distributions from the encoder $(p(s|x_i, C_i, C'_i, \theta))$. This modified objective can also be motivated from a variational approximation perspective, see Marcheggiani and Titov (2016) for details. By varying the parameter $\lambda_H \in \mathbb{R}$, it is easy to control the amount of entropy regularization. For $\lambda_H > 0$, the objective is optimized with flatter posteriors, while $\lambda_H < 0$ infers more peaky posteriors. When $\lambda_H \to -\infty$, the probability mass needs to be concentrated on a single sense, resulting in an algorithm similar to hard EM. In practice, we found that using hard-update training[2], which is closely related to the $\lambda_H \to -\infty$ setting, led to best performance.

2.2 Obtaining word representations

At test time, we construct the word representations by averaging all sense embeddings for a word x_i and weighting them with the sense expectations (Li and Jurafsky, 2015)[3]:

$$\omega_i = \sum_{s \in \mathcal{S}} p(s|x_i, C_i) \varphi_{i,s}. \qquad (5)$$

[2] I.e. updating only that embedding φ_{i,s^*} for which $s^* = \arg\max_s p(s|x_i, C_i, C'_i, \theta)$.

[3] Although our training objective has sparsity-inducing properties, the posteriors at test time are not entirely peaked, which makes weighting beneficial.

Unlike in training, the sense prediction step here does not use the crosslingual context C_i' since it is not available in the evaluation tasks. In this work, instead of marginalizing out the unobservable crosslingual context, we simply ignore it in computation.

Sometimes, even the first-language context is missing, as is the situation in many word similarity tasks. In that case, we just use the uniform average, $1/|\mathcal{S}| \sum_{s \in \mathcal{S}} \varphi_{i,s}$.

3 Word affiliation from alignments

In defining the crosslingual signal we draw on a heuristic inspired by Devlin et al. (2014). The second-language context words are taken to be the multiset of words around and including the pivot affiliated to x_i:

$$C_i' = \{x'_{a_i - m}, ..., x'_{a_i}, ..., x'_{a_i + m}\}, \tag{6}$$

where x'_{a_i} is the word affiliated to x_i and the parameter m regulates the context window size. By choosing $m = 0$, only the affiliated word is used as l' context, and by choosing $m = \infty$, the l' context is the entire sentence (≈uniform alignment). To obtain the index a_i, we use the following:

1) If x_i aligns to exactly one second-language word, a_i is the index of the word it aligns to.
2) If x_i aligns to multiple words, a_i is the index of the aligned word in the middle (and rounding down when necessary).
3) If x_i is unaligned, C_i' is empty, therefore no l' context is used.

We use the cdec aligner (Dyer et al., 2010) to word-align the parallel corpora.

4 Parameters and Set-up

4.1 Learning parameters

We use the AdaGrad optimizer (Duchi et al., 2011) with initial learning rate set to 0.1. We set the mini-batch size to 1000, the number of negative samples to 1, the sampling factor to 0.001 and the window size parameter m to 5. All the embeddings are 50-dimensional (unless specified otherwise) and initialized by sampling from the uniform distribution between $[-0.05, 0.05]$. We include in the vocabulary all words occurring in the corpus at least 20 times. We set the number of senses per word to 3 (see further discussion in § 6.4 and § 7). All other parameters with their default values can be examined in the source code available online.

4.2 Bilingual data

In a large body of work on multilingual word representations, Europarl (Koehn, 2005) is the preferred source of parallel data. However, the domain of Europarl is rather constrained, whereas we would like to obtain word representations of more general language, also to carry out an effective evaluation on semantic similarity datasets where domains are usually broader. We therefore use the following parallel corpora: News Commentary (Bojar et al., 2013) (NC), Yandex-1M[4] (RU-EN), CzEng 1.0 (Bojar et al., 2012) (CZ-EN) from which we exclude the EU legislation texts, and GigaFrEn (Callison-Burch et al., 2009) (FR-EN). The sizes of the corpora are reported in Table 1. The word representations trained on the NC corpora are evaluated only intrinsically due to the small sizes.

Corpus	Language	Words	Sent.
NC	Fr, Ru, Cz, De, Es	3-4 M	.1-.2 M
RU-EN	Ru	24 M	1 M
CZ-EN	Cz	126 M	10 M
FR-EN	Fr	670 M	23 M

Table 1: Parallel corpora used in this paper. The word sizes reported are based on the English part of the corpus. Each language pair in NC has a different English part, hence the varying number of sentences per target language.

5 Evaluation Tasks

We evaluate the quality of our word representations on a number of tasks, both intrinsic and extrinsic.

5.1 Word similarity

We are interested here in how well the semantic similarity ratings obtained from embedding comparisons correlate to human ratings. For this purpose, we use a variety of similarity benchmarks for English and report the Spearman ρ correlation scores between the human ratings and the cosine ratings obtained from our word representations. The **SCWS** benchmark (Huang et al., 2012) is probably the most suitable

[4] https://translate.yandex.ru/corpus

similarity dataset for evaluating multi-sense embeddings, since it allows us to perform the sense prediction step based on the sentential context provided for each word in the pair.

The other benchmarks we use provide the ratings for the word pairs without context. WS-353 contains 353 human-rated word pairs (Finkelstein et al., 2001), while Agirre et al. (2009) separate this benchmark for similarity (WS-SIM) and relatedness (WS-REL). The RG-65 (Rubenstein and Goodenough, 1965) and the MC-30 (Miller and Charles, 1991) benchmarks contain nouns only. The MTurk-287 (Radinsky et al., 2011) and MTurk-771 (Halawi et al., 2012) include word pairs whose similarity was crowdsourced from AMT. Similarly, MEN (Bruni et al., 2012) is an AMT-annotated dataset of 3000 word pairs. The YP-130 (Yang and Powers, 2006) and Verb-143 (Baker et al., 2014) measure verb similarity. Rare-Word (Luong et al., 2013) contains 2034 rare-word pairs. Finally, SimLex-999 (Hill et al., 2014b) is intended to measure pure similarity as opposed to relatedness. For these benchmarks, we prepare the word representations by taking a uniform average of all sense embeddings per word. The evaluation is carried out using the tool described in Faruqui and Dyer (2014a). Due to space constraints, we report the results by averaging over all benchmarks (**Similarity**), and include the individual results in the online repository.

5.2 Supersense similarity

We also evaluate on a task measuring the similarity between the embeddings—in our case uniformly averaged in the case of multi-sense embeddings—and a matrix of supersense features extracted from the English SemCor, using the **Qvec** tool (Tsvetkov et al., 2015). We choose this method because it has been shown to output scores that correlate well with extrinsic tasks, e.g. text classification and sentiment analysis. We believe that this, in combination with word similarity tasks from the previous section, can give a reliable picture of the generic quality of word embeddings studied in this work.

5.3 POS tagging

As our downstream evaluation task, we use the learned word representations to initialize the embedding layer of a neural network tagging model. We use the same convolutional architecture as Li and Jurafsky (2015): an input layer taking a concatenation of neighboring embeddings as input, three hidden layers with a rectified linear unit activation function and a softmax output layer. We train for 10 epochs using one sentence as a batch. Other hyperparameters can be examined in the source code. The multi-sense word embeddings are inferred from the sentential context (weighted average), as for the evaluation on the SCWS dataset. We use the standard splits of the Wall Street Journal portion of the Penn Treebank: 0–18 for training, 19–21 for development and 22–24 for testing.

6 Results

We compare three embeddings models, Skip-Gram (SG), Multi-sense (MU) and Bilingual Multi-sense (BIMU), using our own implementation for each of them. The first two can be seen as simpler variants of the BIMU model: in SG we omit the encoder entirely, and in MU we omit the second-language (l') part of the encoder in Eq. (1). We train the SG and the MU models on the English part of the parallel corpora. Those parameters common to all methods are kept fixed during experiments. The values λ and m for controlling the second-language signal in BIMU are set on the POS-tagging development set (cf. § 6.3).

The results on the **SCWS** benchmark (Table 2) show consistent improvements of the BIMU model over SG and MU across all parallel corpora, except on the small CZ-EN (NC) corpus. We have also measured the 95% confidence intervals of the difference between the correlation coefficients of BIMU and SG, following the method described in Zou (2007). According to these values, BIMU significantly outperforms SG on RU-EN, and on French, Russian and Spanish NC corpora.[5]

Next, ignoring any language-specific factors, we would expect to observe a trend according to which the larger the corpus, the higher the correlation score. However, this is not what we find. Among the largest corpora, i.e. RU-EN, CZ-EN and FR-EN, the models trained on RU-EN perform surprisingly well, practically on par with the 23-times larger FR-EN corpus. Similarly, the quality of the embeddings trained on CZ-EN is generally lower than when trained on the

[5] I.e. counting those results in which the CI of the difference does not include 0.

Task	Corpus	SG	MU	BIMU	BIMU-SG
SCWS	RU-EN	54.8	57.3	**59.5**	$4.7_{0.9}^{9.8}$
	CZ-EN	51.2	54.0	**55.3**	$4.1_{-0.6}^{8.8}$
	FR-EN	58.8	60.4	**60.5**	$1.7_{-2.6}^{5.9}$
	FR-EN (NC)	47.2	52.4	**54.3**	$7.1_{2.2}^{12.0}$
	RU-EN (NC)	47.3	**54.0**	**54.0**	$6.7_{0.6}^{12.8}$
	CZ-EN (NC)	47.7	**52.1**	51.9	$4.2_{-2.0}^{10.3}$
	DE-EN (NC)	48.5	52.9	**54.0**	$5.5_{-0.6}^{11.6}$
	ES-EN (NC)	47.2	53.2	**54.5**	$7.3_{1.1}^{13.3}$
Similarity	RU-EN	37.8	41.2	**46.3**	
	CZ-EN	39.5	36.9	**41.9**	
	FR-EN	**46.3**	42.0	43.5	
	FR-EN (NC)	17.9	26.0	**27.6**	
	RU-EN (NC)	19.3	27.3	**28.4**	
	CZ-EN (NC)	15.8	**26.6**	25.4	
	DE-EN (NC)	20.7	28.4	**30.8**	
	ES-EN (NC)	19.9	27.2	**31.2**	
Qvec	RU-EN	55.8	56.0	**56.5**	
	CZ-EN	**56.6**	56.5	55.9	
	FR-EN	57.5	57.1	**57.6**	
POS	RU-EN	**93.5**	93.2	93.3	
	CZ-EN	**94.0**	93.7	**94.0**	
	FR-EN	**94.1**	93.8	94.0	

Table 2: Results, per-row best in bold. SG and MU are trained on the English part of the parallel corpora. In BIMU-SG, we report the difference between BIMU and SG, together with the 95% CI of that difference. The **Similarity** scores are averaged over 12 benchmarks described in § 5.1. For POS tagging, we report the accuracy.

Model (300-dim.)	SCWS
SG	65.0
MU	66.7
BIMU	69.0
Chen et al. (2014)	68.4
Neelakantan et al. (2014)	69.3
Li and Jurafsky (2015)	69.7

Table 3: Comparison to other works (reprinted), for the vocabulary of top-6000 words. Our models are trained on RU-EN, a much smaller corpus than those used in previous work.

10 times smaller RU-EN corpus. One explanation for this might be different text composition of the corpora, with RU-EN matching the domain of the evaluation task better than the larger two corpora. Also, FR-EN is known to be noisy, containing web-crawled sentences that are not parallel or not natural language (Denkowski et al., 2012). Furthermore, language-dependent effects might be playing a role: for example, there are signs of Czech being the least helpful language among those studied. But while there is evidence for that in all intrinsic tasks, the situation in POS tagging does not confirm this speculation.

We relate our models to previously reported SCWS scores from the literature using 300-dimensional models in Table 3. Even though we train on a much smaller corpus than the previous works,[6] the BIMU model achieves a very competitive correlation score.

The results on **similarity** benchmarks and **qvec** largely confirm those on SCWS, despite the lack of sentential context which would allow to weight the contribution of different senses more accurately for the multi-sense models. Why, then, does simply averaging the MU and BIMU embeddings lead to better results than when using the SG embeddings? We hypothesize that the single-sense model tends to over-represent the dominant sense with its generic, one-vector-per-word representation, whereas the uniformly averaged embeddings yielded by the multi-sense models better encode the range of potential senses. Similar observations have been made in the context of selectional preference modeling of polysemous verbs (Greenberg et al., 2015).

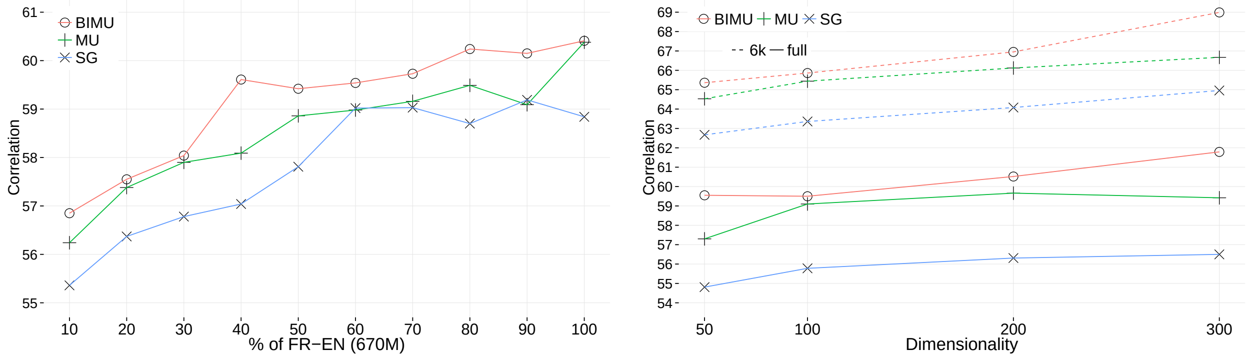
In **POS** tagging, the relationship between MU and BIMU models is similar as discussed above. Overall, however, neither of the multi-sense models outperforms the SG embeddings. The neural network tagger may be able to implicitly perform disambiguation on top of single-sense SG embeddings, similarly to what has been argued in Li and Jurafsky (2015). The tagging accuracies obtained with MU on CZ-EN and FR-EN are similar to the one obtained by Li and Jurafsky with their multi-sense model (93.8), while the accuracy of SG is more competitive in our case (around 94.0 compared to 92.5), although they use a larger corpus for training the word representations.

In all tasks, the addition of the bilingual component during training increases the accuracy of the encoder for most corpora, even though the bilingual information is not available during evaluation.

6.1 The amount of (parallel) data

Fig. 2a displays how the semantic similarity as measured on SCWS evolves as a function of increasingly

[6] For example, Li and Jurafsky (2015) use the concatenation of Gigaword and Wikipedia with more than 5B words.



(a) (b)

Figure 2: (a) Effect of amount of data used in learning on the SCWS correlation scores. (b) Effect of embedding dimensionality on the models trained on RU-EN and evaluated on SCWS with either full vocabulary or the top-6000 words.

larger sub-samples from FR-EN, our largest parallel corpus. The BiMu embeddings show relatively stable improvements over Mu and especially over Sg embeddings. The same performance as that of Sg at 100% is achieved by Mu and BiMu sooner, using only around 40/50% of the corpus.

6.2 The dimensionality and frequent words

It is argued in Li and Jurafsky (2015) that often just increasing the dimensionality of the Sg model suffices to obtain better results than that of their multi-sense model. We look at the effect of dimensionality on semantic similarity in fig. 2b, and see that simply increasing the dimensionality of the Sg model (to any of 100, 200 or 300 dimensions) is not sufficient to outperform the Mu or BiMu models. When constraining the vocabulary to 6,000 most frequent words, the representations obtain higher quality. We can see that the models, especially Sg, benefit slightly more from the increased dimensionality when looking at these most frequent words. This is according to expectations—frequent words need more representational capacity due to their complex semantic and syntactic behavior (Atkins and Rundell, 2008).

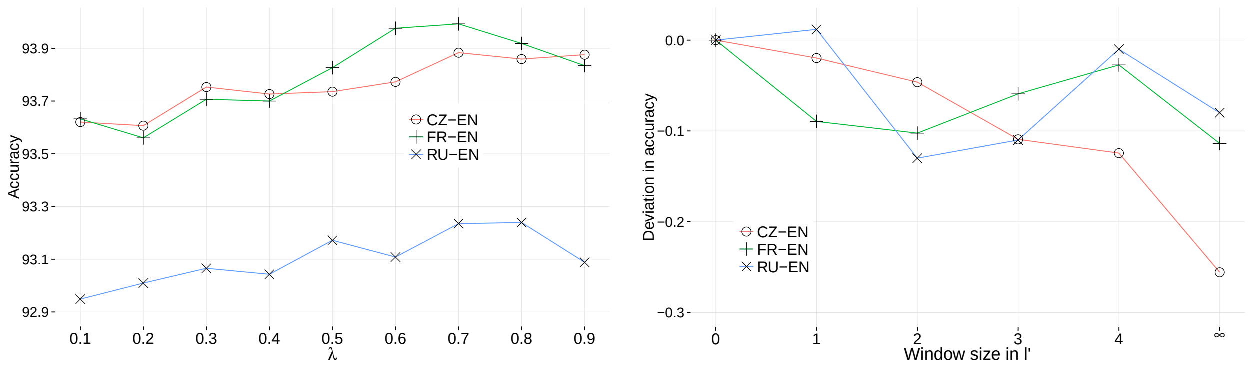
6.3 The role of bilingual signal

The degree of contribution of the second language l' during learning is affected by two parameters, λ for the trade-off between the importance of first and second language in the sense prediction part (encoder) and the value of m for the size of the window around the second-language word affiliated to the pivot. Fig. 3a suggests that the context from the second language is useful in sense prediction, and that it should be weighted relatively heavily (around 0.7 and 0.8, depending on the language).

Regarding the role of the context-window size in sense disambiguation, the WSD literature has reported both smaller (more local) and larger (more topical) monolingual contexts to be useful, see e.g. Ide and Véronis (1998) for an overview. In fig. 3b we find that considering a very narrow context in the second language—the affiliated word only or a $m = 1$ window around it—performs the best, and that there is little gain in using a broader window. This is understandable since the l' representation participating in the sense selection is simply an average over all generic embeddings in the window, which means that the averaged representation probably becomes noisy for large m, i.e. more irrelevant words are included in the window. However, the negative effect on the accuracy is still relatively small, up to around -0.1 for the models using French and Russian as the second languages, and -0.25 for Czech when setting $m = \infty$. The infinite window size setting, corresponding to the sentence-only alignment, performs well also on SCWS, improving on the monolingual multi-sense baseline on all corpora (Table 4).

Model	RU-EN	CZ-EN	FR-EN
Mu	63.29	59.12	64.19
BiMu, $m = \infty$	**65.61**	**62.07**	**64.36**

Table 4: Comparison of SCWS correlation scores of BiMu trained with infinite l' window to the Mu baseline (vocabulary of top-6000 words).



(a) (b)

Figure 3: Controlling the bilingual signal. (a) Effect of varying the parameter λ for controlling the importance of second-language context (0.1-least important, 0.9-most important). (b) Effect of second-language window size m on the accuracy. In both (a) and (b) the reported accuracies are measured on the POS-tagging development set.

6.4 The number of senses

In our work, the number of senses k is a model parameter, which we keep fixed to 3 throughout the empirical study. We comment here briefly on other choices of $k \in \{2, 4, 5\}$. We have found $k = 2$ to be a good choice on the RU-EN and FR-EN corpora (but not on CZ-EN), with an around 0.2-point improvement over $k = 3$ on SCWS and in POS tagging. With the larger values of k, the performance tends to degrade. For example, on RU-EN, the $k = 5$ score on SCWS is about 0.6 point below our default setting.

7 Additional Related Work

Multi-sense models. One line of research has dealt with sense induction as a separate, clustering problem that is followed by an embedding learning component (Huang et al., 2012; Reisinger and Mooney, 2010). In another, the sense assignment and the embeddings are trained jointly (Neelakantan et al., 2014; Tian et al., 2014; Li and Jurafsky, 2015; Bartunov et al., 2015). Neelakantan et al. (2014) propose an extension of Skip-Gram (Mikolov et al., 2013a) by introducing sense-specific parameters together with the k-means-inspired ‘centroid’ vectors that keep track of the contexts in which word senses have occurred. They explore two model variants, one in which the number of senses is the same for all words, and another in which a threshold value determines the number of senses for each word. The results comparing the two variants are inconclusive, with the advantage of the dynamic variant being virtually nonexistent. In our work, we use the static approach. Whenever there is evidence for less senses than the number of available sense vectors, this is unlikely to be a serious issue as the learning would concentrate on some of the senses, and these would then be the preferred predictions also at test time. Li and Jurafsky (2015) build upon the work of Neelakantan et al. with a more principled method for introducing new senses using the Chinese Restaurant Processes (CRP). Our experiments confirm the findings of Neelakantan et al. that multi-sense embeddings improve Skip-gram embeddings on intrinsic tasks, as well as those of Li and Jurafsky, who find that multi-sense embeddings offer little benefit to the neural network learner on extrinsic tasks. Our discrete-autoencoding method when viewed without the bilingual part in the encoder has a lot in common with their methods.

Multilingual models. The research on using multilingual information in the learning of *multi-sense* embedding models is scarce. Guo et al. (2014) perform a sense induction step based on clustering translations prior to learning word embeddings. Once the translations are clustered, they are mapped to a source corpus using WSD heuristics, after which a recurrent neural network is trained to obtain sense-specific representations. Unlike in our work, the sense induction and embedding learning components are entirely separated, without a possibility for one to influence another. In a similar vein, Bansal et al. (2012) use bilingual corpora to perform soft word clustering, extending the previous work on the monolingual case of

Lin and Wu (2009). *Single-sense* representations in the multilingual context have been studied more extensively (Lu et al., 2015; Faruqui and Dyer, 2014b; Hill et al., 2014a; Zhang et al., 2014; Faruqui and Dyer, 2013; Zou et al., 2013), with a goal of bringing the representations in the same semantic space. A related line of work concerns the crosslingual setting, where one tries to leverage training data in one language to build models for typically lower-resource languages (Hermann and Blunsom, 2014; Gouws et al., 2014; Chandar A P et al., 2014; Soyer et al., 2014; Klementiev et al., 2012; Täckström et al., 2012).

The recent works of Kawakami and Dyer (2015) and Nalisnick and Ravi (2015) are also of interest. The latter work on the infinite Skip-Gram model in which the embedding dimensionality is stochastic is relevant since it demonstrates that their embeddings exploit different dimensions to encode different word meanings. Just like us, Kawakami and Dyer (2015) use bilingual supervision, but in a more complex LSTM network that is trained to predict word translations. Although they do not represent different word senses separately, their method produces representations that depend on the context. In our work, the second-language signal is introduced only in the sense prediction component and is flexible—it can be defined in various ways and can be obtained from sentence-only alignments as a special case.

8 Conclusion

We have presented a method for learning multi-sense embeddings that performs sense estimation and context prediction jointly. Both mono- and bilingual information is used in the sense prediction during training. We have explored the model performance on a variety of tasks, showing that the bilingual signal improves the sense predictor, even though the crosslingual information is not available at test time. In this way, we are able to obtain word representations that are of better quality than the monolingually-trained multi-sense representations, and that outperform the Skip-Gram embeddings on intrinsic tasks. We have analyzed the model performance under several conditions, namely varying dimensionality, vocabulary size, amount of data, and size of the second-language context. For the latter parameter, we find that bilingual information is useful even when using the entire sentence as context, suggesting that sentence-only alignment might be sufficient in certain situations.

Acknowledgments

We would like to thank Jiwei Li for providing his tagger implementation, and Robert Grimm, Diego Marcheggiani and the anonymous reviewers for useful comments. The computational work was carried out on Peregrine HPC cluster of the University of Groningen. The second author was supported by NWO Vidi grant 016.153.327.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *NAACL-HLT*.

Waleed Ammar, Chris Dyer, and Noah A. Smith. 2014. Conditional random field autoencoders for unsupervised structured prediction. In *NIPS*.

Sue B. T. Atkins and Michael Rundell. 2008. *The Oxford guide to practical lexicography*. Oxford University Press.

Simon Baker, Roi Reichart, and Anna Korhonen. 2014. An unsupervised model for instance level subcategorization acquisition. In *EMNLP*.

Mohit Bansal, John Denero, and Dekang Lin. 2012. Unsupervised translation sense clustering. In *NAACL-HLT*.

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *ACL*.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL*.

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. 2015. Breaking sticks and ambiguities with adaptive skip-gram. *arXiv preprint arXiv:1502.07257*.

Yoshua Bengio, Aaron Courville, and Pierre Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2012. The Joy of Parallelism with CzEng 1.0. In *LREC*.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia

Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *WMT*.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. 1991. Word-sense disambiguation using statistical methods. In *ACL*.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *ACL*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *WMT*.

Sarath Chandar A P, Stanislas Lauly, Hugo Larochelle, Mitesh M. Khapra, Balaraman Ravindran, Vikas C. Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *NIPS*.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *EMNLP*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Ido Dagan and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.

Michael Denkowski, Greg Hanneman, and Alon Lavie. 2012. The CMU-Avenue French-English Translation System. In *WMT*.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *ACL*.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *ACL*.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *ACL*.

Manaal Faruqui and Chris Dyer. 2013. An information theoretic approach to bilingual word clustering. In *ACL*.

Manaal Faruqui and Chris Dyer. 2014a. Community evaluation and exchange of word vectors at wordvectors.org. In *ACL System Demonstrations*.

Manaal Faruqui and Chris Dyer. 2014b. Improving vector space word representations using multilingual correlation. In *EACL*.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *WWW*.

Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 11:2001–2049.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2014. BilBOWA: Fast Bilingual Distributed Representations without Word Alignments. *arXiv preprint arXiv:1410.2455*.

Clayton Greenberg, Asad Sayeed, and Vera Demberg. 2015. Improving unsupervised vector-space thematic fit evaluation via role-filler prototype clustering. In *NAACL*.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning sense-specific word embeddings by exploiting bilingual resources. In *COLING*.

Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. 2012. Large-scale learning of word relatedness with constraints. In *KDD*.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *ACL*.

Felix Hill, Kyunghyun Cho, Sébastien Jean, Coline Devin, and Yoshua Bengio. 2014a. Embedding word similarity with neural machine translation. *arXiv preprint arXiv:1412.6448*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014b. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *arXiv preprint arXiv:1408.3456*.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *ACL*.

Nancy Ide and Jean Véronis. 1998. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational linguistics*, 24(1):2–40.

Nancy Ide. 2000. Cross-lingual sense determination: Can it work? *Computers and the Humanities*, 34(1-2):223–234.

Hiroyuki Kaji. 2003. Word sense acquisition from bilingual comparable corpora. In *NAACL-HLT*.

Kazuya Kawakami and Chris Dyer. 2015. Learning to represent words in context with multilingual supervision. *arXiv preprint arXiv:1511.04623*.

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

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

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. *CoNLL*.

Jiwei Li and Dan Jurafsky. 2015. Do multi-sense embeddings improve natural language understanding? In *EMNLP*.

Dekang Lin and Xiaoyun Wu. 2009. Phrase clustering for discriminative learning. In *ACL-IJCNLP of AFNLP*.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *NAACL*.

Minh-Thang Luong, Richard Socher, and Christopher D Manning. 2013. Better word representations with recursive neural networks for morphology. In *CoNLL*.

Diego Marcheggiani and Ivan Titov. 2016. Discrete-state variational autoencoders for joint discovery and factorization of relations. *Transactions of the Association for Computational Linguistics*, 4.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *ICLR Workshop Papers*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *NIPS*.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

Eric Nalisnick and Sachin Ravi. 2015. Infinite dimensional word embeddings. *arXiv preprint arXiv:1511.05392*.

Tahira Naseem, Benjamin Snyder, Jacob Eisenstein, and Regina Barzilay. 2009. Multilingual part-of-speech tagging: Two unsupervised approaches. *Journal of Artificial Intelligence Research*, 36:1–45.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient nonparametric estimation of multiple embeddings per word in vector space. In *EMNLP*.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *ACL*.

Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. In *CoNLL*.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *WWW*.

Joseph Reisinger and J. Raymond Mooney. 2010. Multi-prototype vector-space models of word meaning. In *NAACL-HLT*.

Herbert Rubenstein and John B Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning internal representations by error propagation. In David E. Rumelhart, James L. McClelland, and PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*. MIT Press.

Benjamin Snyder and Regina Barzilay. 2010. Climbing the Tower of Babel: Unsupervised Multilingual Learning. In *ICML*.

Hubert Soyer, Pontus Stenetorp, and Akiko Aizawa. 2014. Leveraging monolingual data for crosslingual compositional word representations. *CoRR*, abs/1412.6334.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *NAACL-HLT*.

Fei Tian, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu. 2014. A probabilistic model for learning multi-prototype word embeddings. In *COLING*.

Ivan Titov and Ehsan Khoddam. 2015. Unsupervised induction of semantic roles within a reconstruction-error minimization framework. In *NAACL*.

Ivan Titov and Alexandre Klementiev. 2012. Crosslingual induction of semantic roles. In *ACL*.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *EMNLP*.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *ACL*.

Dongqiang Yang and David M. W. Powers. 2006. Verb similarity on the taxonomy of wordnet. In *GWC*.

Jiajun Zhang, Shujie Liu, Mu Li, Ming Zhou, and Chengqing Zong. 2014. Bilingually-constrained phrase embeddings for machine translation. In *ACL*.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*.

Guang Yong Zou. 2007. Toward using confidence intervals to compare correlations. *Psychological methods*, 12(4).