A Survey of Cross-lingual Word Embedding Models

Sebastian Ruder SEBASTIAN@RUDER.IO
Insight Research Centre, National University of Ireland, Galway, Ireland
Aylien Ltd., Dublin, Ireland

Ivan Vulić IV250@CAM.AC.UK
Language Technology Lab, University of Cambridge, UK

Anders Søgaard SOEGAARD@DI.KU.DK
University of Copenhagen, Copenhagen, Denmark

Abstract

Cross-lingual representations of words enable us to reason about word meaning in multilingual contexts and are a key facilitator of cross-lingual transfer when developing natural language processing models for low-resource languages. In this survey, we provide a comprehensive typology of cross-lingual word embedding models. We compare their data requirements and objective functions. The recurring theme of the survey is that many of the models presented in the literature optimize for the same objectives, and that seemingly different models are often equivalent, *modulo* optimization strategies, hyper-parameters, and such. We also discuss the different ways cross-lingual word embeddings are evaluated, as well as future challenges and research horizons.

1. Introduction

In recent years, (monolingual) vector representations of words, so-called *word embeddings* (Mikolov, Chen, Corrado, & Dean, 2013a; Pennington, Socher, & Manning, 2014) have proven extremely useful across a wide range of natural language processing (NLP) applications. In parallel, the public awareness of the digital language divide[1], as well as the availability of multilingual benchmarks (Hovy, Marcus, Palmer, Ramshaw, & Weischedel, 2006; Sylak-Glassman, Kirov, Yarowsky, & Que, 2015; Nivre, de Marneffe, Ginter, Goldberg, Hajic, Manning, McDonald, Petrov, Pyysalo, Silveira, et al., 2016a), has made cross-lingual transfer a popular NLP research topic. The need to transfer lexical knowledge across languages has given rise to *cross-lingual word embedding models*, i.e., cross-lingual representations of words in a joint embedding space, as illustrated in Figure 1.

Cross-lingual word embeddings are appealing for two reasons: First, they enable us *to compare the meaning of words across languages*, which is key to bilingual lexicon induction, machine translation, or cross-lingual information retrieval, for example. Second, cross-lingual word embeddings *enable model transfer between languages*, e.g., between resource-rich and low-resource languages, by providing a common representation space. This duality is also reflected in how cross-lingual word embeddings are evaluated, as discussed in Section 10.

Many models for learning cross-lingual embeddings have been proposed in recent years. In this survey, we will give a comprehensive overview of existing cross-lingual word embedding models. One of the main goals of this survey is to show the similarities and differences between

[1] E.g., `http://labs.theguardian.com/digital-language-divide/`

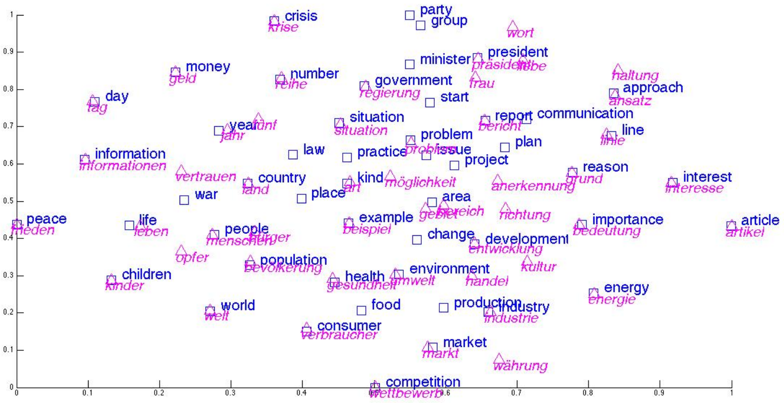


Figure 1: A shared embedding space between two languages (Luong et al., 2015)

these approaches. To facilitate this, we first introduce a common notation and terminology in Section 2. Over the course of the survey, we then show that existing cross-lingual word embedding models can be seen as optimizing very similar objectives, where the main source of variation is due to the data used, the monolingual and regularization objectives employed, and how these are optimized. As many cross-lingual word embedding models are inspired by monolingual models, we introduce the most commonly used monolingual embedding models in Section 3. We then motivate and introduce one of the main contributions of this survey, a typology of cross-lingual embedding models in Section 4. The typology is based on the main differentiating aspect of cross-lingual embedding models: the nature of the data they require, in particular the type of alignment across languages (alignment of words, sentences, or documents), and whether data is assumed to be parallel or just comparable (about the same topic). The typology allows us to outline similarities and differences more concisely, but also starkly contrasts focal points of research with fruitful directions that have so far gone mostly unexplored.

Since the idea of cross-lingual representations of words pre-dates word embeddings, we provide a brief history of cross-lingual word representations in Section 5. Subsequent sections are dedicated to each type of alignment. We discuss cross-lingual word embedding algorithms that rely on word-level alignments in Section 6. Such methods can be further divided into mapping-based approaches, approaches based on pseudo-bilingual corpora, and joint methods. We show that these approaches, *modulo* optimization strategies and hyper-parameters, are nevertheless often equivalent. We then discuss approaches that rely on sentence-level alignments in Section 7, and models that require document-level alignments in Section 8. In Section 9, we describe how many bilingual approaches that deal with a pair of languages can be extended to the multilingual setting. We subsequently provide an extensive discussion of the tasks, benchmarks, and challenges of the evaluation of cross-lingual embedding models in Section 10 and outline applications in Section 11. We present general challenges and future research directions in learning cross-lingual word representations in Section 12. Finally, we provide our conclusions in Section 13.

This survey makes the following contributions:

1. It proposes a general typology that characterizes the differentiating features of cross-lingual word embedding models and provides a compact overview of these models.

2. It standardizes terminology and notation and shows that many cross-lingual word embedding models can be cast as optimizing nearly the same objective functions.

3. It provides an informal proof that connects the three types of word-level alignment models and shows that these models are optimizing roughly the same objective.

4. It critically examines the standard ways of evaluating cross-lingual embedding models.

5. It describes multilingual extensions for the most common types of cross-lingual embedding models.

6. It outlines outstanding challenges for learning cross-lingual word embeddings and provides suggestions for fruitful and unexplored research directions.

Disclaimer Neural Machine Translation (NMT) is another area that has received increasing interest. NMT approaches *implicitly* learn a shared cross-lingual embedding space by optimizing for the Machine Translation (MT) objective, whereas we will focus on models that *explicitly* learn cross-lingual word representations throughout this survey. These methods generally do so at a much lower cost than MT and, in terms of speed and efficiency, can be considered to be to MT what word embedding models (Mikolov et al., 2013a; Pennington et al., 2014) are to language modeling.

2. Notation and Terminology

For clarity, we show all notation used throughout this survey in Table 1. Let $\mathbf{X}^l \in \mathbb{R}^{|V^l| \times d}$ be a word embedding matrix that is learned for the l-th of L languages where V^l is the corresponding vocabulary and d is the dimensionality of the word embeddings. We will furthermore refer to $\mathbf{X}^l_{i,:}$, that is, the word embedding of the i-th word in language l with the shorthand $\mathbf{x}^l_i$ or $\mathbf{x}_i$ if the language is unambiguous. We will refer to the word corresponding to the i-th word embedding $\mathbf{x}_i$ as w_i. Some monolingual word embedding models use a separate embedding for words that occur in the context of other words. We will use $\tilde{x}_i$ as the embedding of the i-th context word and detail its meaning in the next section. Most approaches only deal with two languages, a source language s and a target language t.

Some approaches learn a matrix $\mathbf{W}^{s \to t}$ that can be used to transform the word embedding matrix $\mathbf{X}^s$ of the source language s to that of the target language t. We will designate such a matrix by $\mathbf{W}^{s \to t} \in \mathbb{R}^{d \times d}$ and $\mathbf{W}$ if the language pairing is unambiguous. These approaches often use n source words and their translations as seed words. In addition, we will use τ as a function that maps from source words w^s_i to their translation w^t_i: $\tau : V^s \to V^t$. Approaches that learn a transformation matrix are usually referred to as *offline* or *mapping* methods. As one of the goals of this survey is to standardize nomenclature, we will use the term *mapping* in the following to designate such approaches.

Some approaches require a monolingual word-word co-occurrence matrix $\mathbf{C}^s$ in language s. In such a matrix, every row corresponds to a word w^s_i and every column corresponds to a context word w^s_j. $\mathbf{C}^s_{ij}$ then captures the number of times word w_i occurs with context word

Symbol	Meaning
$\mathbf{X}$	word embedding matrix
V	vocabulary
d	word embedding dimensionality
$\mathbf{X}^l_{i,:}$ / $\mathbf{x}^l_i$ / $\mathbf{x}_i$	word embedding of the i-th word in language l
$\tilde{x}_i$	word embedding of the i-th context word
w_i	word pertaining to embedding $\mathbf{x}_i$
s	source language
t	target language
$\mathbf{W}^{s\to t}$ / $\mathbf{W}$	learned transformation matrix between space of s and t
n	number of words used as seed words for learning $\mathbf{W}$
τ	function mapping from source words to their translations
$\mathbf{C}^s$	monolingual co-occurrence matrix in language s
C	size of context window around a center word
$\mathcal{C}$	corpus of words / aligned sentences used for training
$\mathbf{A}^{s\to t}$	cross-lingual co-occurrence matrix / alignment matrix
$sent^s_i$	i-th sentence in language s
$\mathbf{y}^s_i$	representation of i-th sentence in language s
doc^s_i	i-th document in language s
$\mathbf{z}^s_i$	representation of i-th document in language s
$\underline{\mathbf{X}^s}$	$\mathbf{X}^s$ is kept fixed during optimization
$\underbrace{\mathcal{L}^1}_{1} + \underbrace{\mathcal{L}^2}_{2}$	$\mathcal{L}^1$ is optimized before $\mathcal{L}^2$

Table 1: Notation used throughout this survey.

w_j usually within a window of size C to the left and right of word w_i. In a cross-lingual context, we obtain a matrix of alignment counts $\mathbf{A}^{s \rightarrow t} \in \mathbb{R}^{|V^t| \times |V^s|}$, where each element $\mathbf{A}^{s \rightarrow t}_{ij}$ captures the number of times the $i-$th word in language t was aligned with the j-th word in language s, with each row normalized to 1.

Finally, as some approaches rely on pairs of aligned sentences, we designate $sent^s_1, \ldots, sent^s_n$ as sentences in source language s with representations $\mathbf{y}^s_1, \ldots, \mathbf{y}^s_n$, and analogously refer to their aligned sentences in the target language t as $sent^t_1, \ldots, sent^t_n$ with representations $\mathbf{y}^t_1, \ldots, \mathbf{y}^t_n$. We adopt an analogous notation for representations obtained by approaches based on alignments of documents in s and t: $doc^s_1, \ldots, doc^s_n$ and $doc^t_1, \ldots, doc^t_n$ with document representations $\mathbf{z}^s_1, \ldots, \mathbf{z}^s_n$ and $\mathbf{z}^t_1, \ldots, \mathbf{z}^t_n$ respectively.

Different notations make similar approaches appear different. Using the same notation across our survey facilitates recognizing similarities between the various cross-lingual word embedding models. Specifically, we intend to demonstrate that cross-lingual word embedding models are trained by minimizing roughly the same objective functions, and that differences in objective are unlikely to explain the observed performance differences (Levy, Søgaard, & Goldberg, 2017).

The class of objective functions minimized by most cross-lingual word embedding methods (if not all), can be formulated as follows:

$$J = \mathcal{L}^1 + \ldots + \mathcal{L}^L + \Omega \tag{1}$$

where $\mathcal{L}^l$ is the monolingual loss of the l-th language and Ω is a regularization term. A similar loss was also defined by Upadhyay, Faruqui, Dyer, and Roth (2016). As recent work (Levy & Goldberg, 2014; Levy, Goldberg, & Dagan, 2015) shows that many monolingual objectives are very similar, one of the main contributions of this survey is to condense the difference between approaches into a regularization term and to detail the assumptions that underlie different regularization terms.

Importantly, how this objective function is optimized is a key characteristic and differentiating factor between different approaches. The joint optimization of multiple non-convex losses is difficult. Most approaches thus take a step-wise approach and optimize one loss at a time while keeping certain variables fixed. In most cases, we will thus use a longer formulation such as the one below, which makes clear in what order the losses are optimized and which variables they depend on:

$$J = \underbrace{\mathcal{L}(\mathbf{X}^s) + \mathcal{L}(\mathbf{X}^t)}_{1} + \underbrace{\Omega(\underline{\mathbf{X}^s}, \underline{\mathbf{X}^t}, \mathbf{W})}_{2} \tag{2}$$

The underbraces indicate that the two monolingual loss terms on the left, which depend on $\mathbf{X}^s$ and $\mathbf{X}^t$ respectively, are optimized first. Subsequently, Ω is optimized, which depends on $\underline{\mathbf{X}^s}, \underline{\mathbf{X}^t}, \mathbf{W}$. Note that underlined variables are kept fixed during optimization of the loss.

The monolingual objectives are optimized by training one of several monolingual embedding models on a monolingual corpus. These models are outlined in the next section.

3. Monolingual Embedding Models

The majority of cross-lingual embedding models take inspiration from and extend monolingual word embedding models to bilingual settings, or explicitly leverage monolingually trained models. As an important preliminary, we thus briefly introduce monolingual embedding models that have been used in the cross-lingual embeddings literature.

Latent Semantic Analysis (LSA) Latent Semantic Analysis (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) has been one of the most widely used methods for learning dense word representations. Given a sparse word-word co-occurrence matrix $\mathbf{C}$ obtained from a corpus, we replace every entry in $\mathbf{C}$ with its pointwise mutual information (PMI) (Church & Hanks, 1990) score, thus yielding a PMI matrix $\mathbf{P}$.[2] We factorize $\mathbf{P}$ using singular value decomposition (SVD), which decomposes $\mathbf{P}$ into the product of three matrices:

$$\mathbf{P} = \mathbf{U}\boldsymbol{\Psi}\mathbf{V}^{\top} \tag{3}$$

where $\mathbf{U}$ and $\mathbf{V}$ are in column orthonormal form and $\boldsymbol{\Psi}$ is a diagonal matrix of singular values. We subsequently obtain the word embedding matrix $\mathbf{X}$ by reducing the word representations to dimensionality k the following way:

$$\mathbf{X} = \mathbf{U}_k\boldsymbol{\Psi}_k \tag{4}$$

where $\boldsymbol{\Psi}_k$ is the diagonal matrix containing the top k singular values and $\mathbf{U}_k$ is obtained by selecting the corresponding columns from $\mathbf{U}$.

Max-margin hinge loss (MMHL) Collobert and Weston (2008) learn word embeddings by training a model on a corpus $\mathcal{C}$ to output a higher score for a correct word sequence than for an incorrect one. For this purpose, they use a max-margin hinge loss:

$$\mathcal{L}_{\text{MMHL}} = \sum_{i=C}^{|\mathcal{C}|-C} \sum_{w' \in V} \max(0, 1 - f([w_{i-C}, \ldots, w_i, \ldots, w_{i+C}]) + f([w_{i-C}, \ldots, w', \ldots, w_{i+C}])) \tag{5}$$

The outer sum iterates over all words in the corpus $\mathcal{C}$, while the inner sum iterates over all words in the vocabulary. Each word sequence consists of a center word w_i and a window of C words to its left and right. Note that we usually use the index i to indicate the i-th word in the vocabulary V, while we use it here to designate the i-th word in the corpus $\mathcal{C}$. $f(\cdot)$ is a neural network that outputs a score given a word sequence and is trained to output a higher score for a word sequence occurring in the corpus (the left term) than a word sequence where the center word is replaced by an arbitrary word w' from the vocabulary (the right term).

Skip-gram with negative sampling (SNGS) Skip-gram with negative sampling (Mikolov et al., 2013a) is arguably the most popular method to learn monolingual word embeddings due to its training efficiency and robustness (Levy et al., 2015). SGNS approximates a language model but focuses on learning efficient word representations rather than accurately modeling word probabilities. It induces representations that are good at predicting surrounding context words given a target word w_t. The objective is shown in Figure 2. To this end, it minimizes

[2] Positive PMI is used by Levy et al. (2017).

the negative log-likelihood of the training data under the following *skip-gram* objective:

$$\mathcal{L}_{\text{SGNS}} = -\frac{1}{|\mathcal{C}|} \sum_{t=1}^{|\mathcal{C}|} \sum_{-C \leq j \leq C, j \neq 0} \log P(w_{t+j} \mid w_t) \tag{6}$$

$P(w_{t+j} \mid w_t)$ is computed using the softmax function:

$$P(w_{t+j} \mid w_t) = \frac{\exp(\tilde{\mathbf{x}}_{t+j}{}^\top \mathbf{x}_t)}{\sum_{i=1}^{|V|} \exp(\tilde{\mathbf{x}}_i{}^\top \mathbf{x}_t)} \tag{7}$$

where $\mathbf{x}_i$ and $\tilde{\mathbf{x}}_i$ are the word and context word embeddings of word w_i respectively. This formulation makes the meaning of the context word embeddings clearer. The skip-gram architecture can be seen as a neural network without a hidden layer. The word embedding $\mathbf{x}_i$ of the input word w_i is then the same as the hidden state of the model. This word embedding $\mathbf{x}_i$ is then fed into a softmax layer, where each word has a *separate* representation $\tilde{\mathbf{x}}_i$, which represents how it behaves in the context of the input word. Generally, $\mathbf{x}_i$ is used as the final word representation, although combining both $\mathbf{x}_i$ and $\tilde{\mathbf{x}}_i$ can be beneficial (Levy et al., 2015). In language modeling, recent approaches in fact constrain $\mathbf{x}_i$ and $\tilde{\mathbf{x}}_i$ to be the same (Inan, Khosravi, & Socher, 2016).

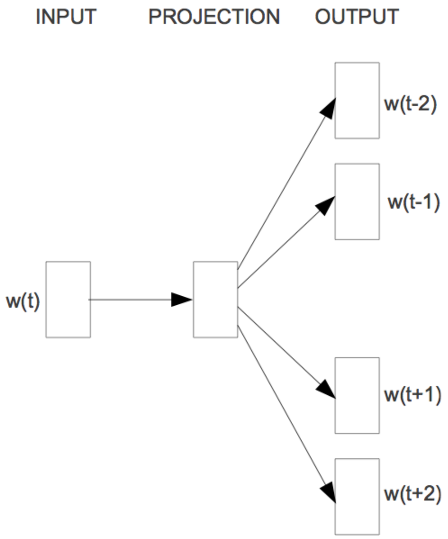


Figure 2: The SGNS monolingual embedding model (Mikolov et al., 2013a)

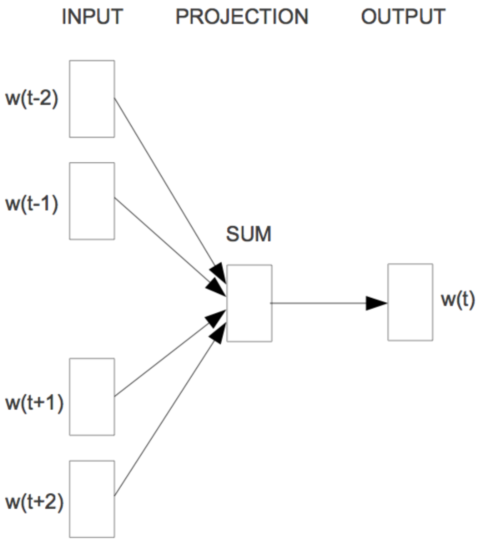


Figure 3: The CBOW monolingual embedding model (Mikolov et al., 2013a)

As the partition function in the denominator of the softmax is expensive to compute, SGNS uses Negative Sampling, which approximates the softmax to make it computationally more efficient. Negative sampling is a simplification of Noise Contrastive Estimation (Gutmann & Hyvärinen, 2012), which was applied to language modeling by Mnih and Teh (2012). Similar to noise contrastive estimation, negative sampling trains the model to distinguish a target word w_t from negative samples drawn from a noise distribution P_n. In this regard, it is similar to MMHL as defined above, which ranks true sentences above noisy sentences. Negative sampling is defined as follows:

$$P(w_{t+j} \mid w_t) = \log \sigma(\tilde{\mathbf{x}}_{t+j}{}^\top \mathbf{x}_t) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n} \log \sigma(-\tilde{\mathbf{x}}_i{}^\top \mathbf{x}_t) \tag{8}$$

where σ is the sigmoid function $\sigma(x) = 1/(1 + e^{-x})$ and k is the number of negative samples. The distribution P_n is empirically set to the unigram distribution raised to the $3/4^{th}$ power. Levy and Goldberg (2014) observe that negative sampling does not in fact minimize the negative log-likelihood of the training data as in Equation 6, but rather implicitly factorizes a shifted PMI matrix, very similar to LSA.

Continuous bag-of-words (CBOW) Continuous bag-of-words can be seen as the inverse of the skip-gram architecture: The model receives as input a window of C context words and seeks to predict the target word w_t by minimizing the CBOW objective:

$$\mathcal{L}_{\text{CBOW}} = -\frac{1}{|\mathcal{C}|} \sum_{t=1}^{|\mathcal{C}|} \log P(w_t \mid w_{t-C}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+C}) \tag{9}$$

$$P(w_t \mid w_{t-C}, \ldots, w_{t+C}) = \frac{\exp(\tilde{\mathbf{x}}_t{}^\top \mathbf{x}_s)}{\sum_{i=1}^{|V|} \exp(\tilde{\mathbf{x}}_i{}^\top \mathbf{x}_s)} \tag{10}$$

where $\mathbf{x}_s$ is the sum of the word embeddings of the words $w_{t-C}, \ldots, w_{t+C}$, i.e. $\mathbf{x}_s = \sum_{-C \leq j \leq C, j \neq 0} \mathbf{x}_{t+j}$. This is depicted in Figure 3. The CBOW architecture is typically also trained with negative sampling for the same reason as the skip-gram model.

Global vectors (GloVe) Global vectors (Pennington et al., 2014) allows us to learn word representations via matrix factorization. GloVe minimizes the difference between the dot product of the embeddings of a word w_i and its context word c_t and the logarithm of their number of co-occurrences within a certain window size[3]:

$$\mathcal{L}_{\text{GloVe}} = \sum_{i,j=1}^{|V|} f(\mathbf{C}_{ij})(\mathbf{x}_i{}^\top \tilde{\mathbf{x}}_j + b_i + \tilde{b}_j - \log \mathbf{C}_{ij})^2 \tag{11}$$

where b_i and $\tilde{b}_j$ are the biases corresponding to word w_i and its context word w_j, $\mathbf{C}_{ij}$ captures the number of times word w_i occurs with context word w_j, and $f(\cdot)$ is a weighting function that assigns relatively lower weight to rare and frequent co-occurrences.

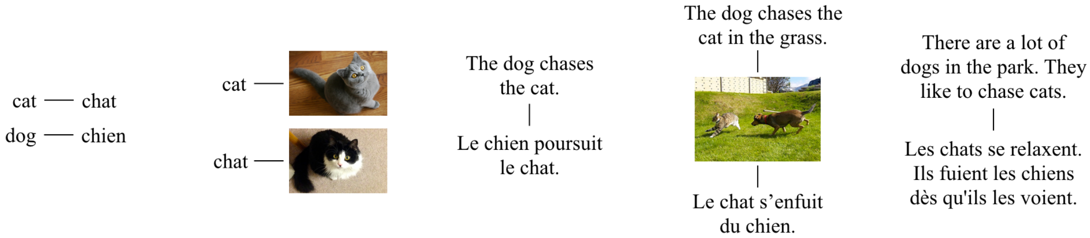
4. Cross-Lingual Word Embedding Models: Typology

Recent work on cross-lingual embedding models suggests that the actual choice of bilingual supervision signal – that is, the data a method requires to learn to align a cross-lingual representation space – is more important for the final model performance than the actual underlying architecture (Levy et al., 2017). Similar conclusions can be drawn from empirical

[3] GloVe favors slightly larger window sizes (up to 10 words to the right and to the left of the target word) compared to SGNS (Levy et al., 2015).

	Parallel	Comparable
Word	Dictionaries	Images
Sentence	Translations	Captions
Document	-	Wikipedia

Table 2: Nature and alignment level of bilingual data sources required by cross-lingual embedding models.


cat — chat
dog — chien
cat —
chat —
The dog chases the cat.
Le chien poursuit le chat.
The dog chases the cat in the grass.
Le chat s'enfuit du chien.
There are a lot of dogs in the park. They like to chase cats.
Les chats se relaxent. Ils fuient les chiens dès qu'ils les voient.

(a) Word, par. (b) Word, comp. (c) Sentence, par. (d) Sentence, comp. (e) Doc., comp.

Figure 4: Examples for the nature and type of alignment of data sources. Par.: parallel. Comp.: comparable. Doc.: document. From left to right, word-level parallel alignment in the form of a bilingual lexicon (4a), word-level comparable alignment using images obtained with Google search queries (4b), sentence-level parallel alignment with translations (4c), sentence-level comparable alignment using translations of several image captions (4d), and document-level comparable alignment using similar documents (4e).

work in comparing different cross-lingual embedding models (Upadhyay et al., 2016). In other words, large differences between models typically stem from their data requirements, while other fine-grained differences are artifacts of the chosen architecture, hyper-parameters, and additional tricks and fine-tuning employed. This directly mirrors the argument raised by Levy et al. (2015) regarding monolingual embedding models: They observe that the architecture is less important as long as the models are trained under identical conditions on the same type (and amount) of data.

We therefore base our typology on the data requirements of the cross-lingual word embedding methods, as this accounts for much of the variation in performance. In particular, methods differ with regard to the data they employ along the following two dimensions:

1. **Type of alignment**: Methods use different types of bilingual supervision signals (at the level of words, sentences, or documents), which introduce stronger or weaker supervision.

2. **Comparability**: Methods require either *parallel* data sources, that is, exact translations in different languages or *comparable* data that is only similar in some way.

In particular, there are three different types of alignments that are possible, which are required by different methods. We discuss the typical data sources for both parallel and comparable data based on the following alignment signals:

1. **Word alignment**: Most approaches use parallel word-aligned data in the form of bilingual or cross-lingual dictionary with pairs of translations between words in different languages (Mikolov, Le, & Sutskever, 2013b; Faruqui & Dyer, 2014b). Comparable word-aligned data, even though more plentiful, has been leveraged less often and typically involves other modalities such as images (Bergsma & Van Durme, 2011; Kiela, Vulić, & Clark, 2015).

2. **Sentence alignment**: Sentence alignment requires a parallel corpus, as commonly used in MT. Methods typically use the Europarl corpus (Koehn, 2005), which consists of sentence-aligned text from the proceedings of the European parliament, and is perhaps the most common source of training data for MT models (Hermann & Blunsom, 2013; Lauly, Boulanger, & Larochelle, 2013). Other methods use available word-level alignment information (Zou, Socher, Cer, & Manning, 2013; Shi, Liu, Liu, & Sun, 2015). There has been some work on extracting parallel data from comparable corpora (Munteanu & Marcu, 2006), but no-one has so far trained cross-lingual word embeddings on such data. Comparable data with sentence alignment may again leverage another modality, such as captions of the same image or similar images in different languages, which are not translations of each other (Calixto, Liu, & Campbell, 2017; Gella, Sennrich, Keller, & Lapata, 2017).

3. **Document alignment**: Parallel document-aligned data requires documents in different languages that are translations of each other. This is rare, as parallel documents typically means sentences can be aligned (Hermann & Blunsom, 2014). Comparable document-aligned data is more common and can occur in the form of documents that are topic-aligned (e.g. Wikipedia) or class-aligned (e.g. sentiment analysis and multi-class classification datasets) (Vulić & Moens, 2013b; Mogadala & Rettinger, 2016).

We summarize the different types of data required by cross-lingual embedding models along these two dimensions in Table 2 and provide examples for each in Figure 4. Over the course of this survey we will show that models that use a particular type of data are mostly variations of the same or similar architectures. We present our complete typology of cross-lingual embedding models in Table 3, aiming to provide an exhaustive overview by classifying each model (we are aware of) into one of the corresponding model types. We also provide a more detailed overview of the monolingual objectives and regularization terms used by every approach towards the end of this survey in Table 5.

5. A Brief History of Cross-Lingual Word Representations

We provide a brief overview of the historical lineage of cross-lingual word embedding models. In brief, while cross-lingual word embeddings is a novel phenomenon, many of the high-level ideas that motivate current research in this area, can be found in work that pre-dates the popular introduction of word embeddings. This includes work on learning cross-lingual word representations from seed lexica, parallel data, or document-aligned data, as well as ideas on learning from limited bilingual supervision.

Language-independent representations have been proposed for decades, many of which rely on abstract linguistic labels instead of lexical features (Aone & McKee, 1993; Schultz &

	Parallel	**Comparable**
Word	Mikolov et al. (2013)	Bergsma and Van Durme (2011)
	Dinu et al. (2015)	Vulić et al. (2016)
	Lazaridou et al. (2015)	Kiela et al. (2015)
	Xing et al. (2015)	Vulić et al. (2016)
	Zhang et al. (2016)	Gouws and Søgaard (2015)
	Artexte et al. (2016)	Duong et al. (2015)
	Smith et al. (2016)	
	Vulić and Korhonen (2016)	
	Artexte et al. (2017)	
	Hauer et al. (2017)	
	Mrkšić et al. (2017)	
	Faruqui and Dyer (2014)	
	Lu et al. (2015)	
	Ammar et al. (2016)	
	Xiao and Guo (2014)	
	Gouws and Søgaard (2015)	
	Duong et al. (2016)	
	Adams et al. (2017)	
	Klementiev et al. (2012)	
	Kočiský et al. (2014)	
Sentence	Zou et al. (2013)	Calixto et al. (2017)
	Shi et al. (2015)	Gella et al. (2017)
	Gardner et al. (2015)	
	Vyas and Carpuat (2016)	
	Guo et al. (2015)	
	Hermann and Blunsom (2013)	
	Hermann and Blunsom (2014)	
	Soyer et al. (2015)	
	Lauly et al. (2013)	
	Chandar et al. (2014)	
	Gouws et al. (2015)	
	Luong et al. (2015)	
	Coulmance et al. (2015)	
	Pham et al. (2015)	
	Levy et al. (2017)	
	Rajendran et al. (2016)	
Document		Vulić and Moens (2016)
		Vulić and Moens (2013)
		Vulić and Moens (2014)
		Søgaard et al. (2015)
		Mogadala and Rettinger (2016)

Table 3: Cross-lingual embedding models ordered by data requirements.

Waibel, 2001). This is also the strategy used in early work on so-called *delexicalized* cross-lingual and domain transfer (Zeman & Resnik, 2008; Søgaard, 2011; McDonald, Petrov, & Hall, 2011; Cohen, Das, & Smith, 2011; Täckström, McDonald, & Uszkoreit, 2012; Henderson, Thomson, & Young, 2014), as well as in work on inducing cross-lingual word clusters (Täckström et al., 2012; Faruqui & Dyer, 2013), and cross-lingual word embeddings relying on syntactic/POS contexts (Duong, Cohn, Bird, & Cook, 2015; Dehouck & Denis, 2017).[4] The ability to represent lexical items from two different languages in a shared cross-lingual space supplements seminal work in cross-lingual transfer by providing fine-grained word-level links between languages; see work in cross-lingual dependency parsing (Ammar, Mulcaire, Ballesteros, Dyer, & Smith, 2016a; Zeman et al., 2017) and natural language understanding systems (Mrkšić, Vulić, Ó Séaghdha, Leviant, Reichart, Gašić, Korhonen, & Young, 2017b).

Similar to our typology of cross-lingual word embedding models outlined in Table 3 based on bilingual data requirements from Table 2, one can also arrange older cross-lingual representation architectures into similar categories. A traditional approach to cross-lingual vector space induction was based on high-dimensional context-counting vectors where each dimension encodes the (weighted) co-occurrences with a specific context word in each of the languages. The vectors are then *mapped* into a single cross-lingual space using a seed bilingual dictionary containing paired context words from both sides (Rapp, 1999; Gaussier, Renders, Matveeva, Goutte, & Déjean, 2004; Laroche & Langlais, 2010; Tamura, Watanabe, & Sumita, 2012, inter alia). This approach is an important predecessor to the cross-lingual word embedding models described in Section 6. Similarly, the bootstrapping technique developed for traditional context-counting approaches (Peirsman & Padó, 2010; Vulić & Moens, 2013b) is an important predecessor to recent iterative self-learning techniques used to limit the bilingual dictionary seed supervision needed in mapping-based approaches (Hauer, Nicolai, & Kondrak, 2017; Artetxe, Labaka, & Agirre, 2017). The idea of CCA-based word embedding learning (see later in Section 6) (Faruqui & Dyer, 2014b; Lu, Wang, Bansal, Gimpel, & Livescu, 2015) was also introduced a decade earlier (Haghighi, Liang, Berg-Kirkpatrick, & Klein, 2008); their word additionally discussed the idea of combining orthographic subword features with distributional signatures for cross-lingual representation learning: This idea re-entered the literature recently (Heyman, Vulić, & Moens, 2017), only now with much better performance.

Cross-lingual word embeddings can also be directly linked to the work on word alignment for statistical machine translation (Brown, Pietra, Pietra, & Mercer, 1993; Och & Ney, 2003). Levy et al. (2017) stress that word translation probabilities extracted from sentence-aligned parallel data by IBM alignment models can also act as the cross-lingual semantic similarity function in lieu of the cosine similarity between word embeddings. Such word translation tables are then used to induce bilingual lexicons. For instance, aligning each word in a given source language sentence with the most similar target language word from the target language sentence is exactly the same greedy decoding algorithm that is implemented in IBM Model 1. Bilingual dictionaries and cross-lingual word clusters derived from word alignment

[4] Along the same line, the recent initiative on providing cross-linguistically consistent sets of such labels (e.g., Universal Dependencies (Nivre et al., 2016b)) facilitates cross-lingual transfer and offers further support to the induction of word embeddings across languages (Vulić, 2017; Vulić, Schwartz, Rappoport, Reichart, & Korhonen, 2017).

links can be used to boost cross-lingual transfer for applications such as syntactic parsing (Täckström et al., 2012; Durrett, Pauls, & Klein, 2012), POS tagging (Agić, Hovy, & Søgaard, 2015), or semantic role labeling (Kozhevnikov & Titov, 2013) by relying on shared lexical information stored in the bilingual lexicon entries. Exactly the same functionality can be achieved by cross-lingual word embeddings. However, cross-lingual word embeddings have another advantage in the era of neural networks: the continuous representations can be plugged into current end-to-end neural architectures directly as sets of lexical features.

A large body of work on multilingual probabilistic topic modeling (Vulić, De Smet, Tang, & Moens, 2015; Boyd-Graber, Hu, & Mimno, 2017) also extracts shared cross-lingual word spaces, now by means of conditional latent topic probability distributions: two words with similar distributions over the induced latent variables/topics are considered semantically similar. The learning process is again steered by the data requirements. The early days witnessed the use of pseudo-bilingual corpora constructed by merging aligned document pairs, and then applying a monolingual representation model such as LSA (Landauer & Dumais, 1997) or LDA (Blei, Ng, & Jordan, 2003) on top of the merged data (Littman, Dumais, & Landauer, 1998; De Smet, Tang, & Moens, 2011). This approach is very similar to the pseudo-cross-lingual approaches discussed in Section 6 and Section 8. More recent topic models learn on the basis of parallel word-level information, enforcing word pairs from seed bilingual lexicons (again!) to obtain similar topic distributions (Boyd-Graber & Blei, 2009; Zhang, Mei, & Zhai, 2010; Boyd-Graber & Resnik, 2010; Jagarlamudi & Daumé III, 2010). In consequence, this also influences topic distributions of related words not occurring in the dictionary. Another group of models utilizes alignments at the document level (Mimno, Wallach, Naradowsky, Smith, & McCallum, 2009; Platt, Toutanova, & Yih, 2010; Vulić, De Smet, & Moens, 2011; Fukumasu, Eguchi, & Xing, 2012; Heyman, Vulić, & Moens, 2016) to induce shared topical spaces. The very same level of supervision (i.e., document alignments) is used by several cross-lingual word embedding models, surveyed in Section 8. Another embedding model based on the document-aligned Wikipedia structure (Søgaard, Agić, Alonso, Plank, Bohnet, & Johannsen, 2015) bears resemblance with the cross-lingual Explicit Semantic Analysis model (Gabrilovich & Markovitch, 2006; Hassan & Mihalcea, 2009; Sorg & Cimiano, 2012).

All these “historical” architectures measure the strength of cross-lingual word similarities through metrics defined in the cross-lingual space: e.g., Kullback-Leibler or Jensen-Shannon divergence (in a topic space), or vector inner products (in sparse context-counting vector space),and are therefore applicable to NLP tasks that rely cross-lingual similarity scores. The pre-embedding architectures and more recent cross-lingual word embedding methods have been applied to an overlapping set of evaluation tasks and applications, ranging from bilingual lexicon induction to cross-lingual knowledge transfer, including cross-lingual parser transfer (Täckström et al., 2012; Ammar et al., 2016a), cross-lingual document classification (Gabrilovich & Markovitch, 2006; De Smet et al., 2011; Klementiev, Titov, & Bhattarai, 2012; Hermann & Blunsom, 2014), cross-lingual relation extraction (Faruqui & Kumar, 2015), etc. In summary, while sharing the goals and assumptions of older cross-lingual architectures, cross-lingual word embedding methods have capitalized on the recent methodological and algorithmic advances in the field of representation learning, owing their wide use to their simplicity, efficiency and handling of large corpora, as well as their relatively robust performance across domains.

6. Word-Level Alignment Models

In the following, we will now discuss different types of the current generation of cross-lingual embedding models, starting with models based on word-level alignment. Among these, models based on parallel data are more common.

6.1 Word-level Alignment Methods with Parallel Data

We distinguish three methods that use parallel word-aligned data:

a) **Mapping-based approaches** that first train monolingual word representations independently on large monolingual corpora and then seek to learn a transformation matrix that maps representations in one language to the representations of the other language. They learn this transformation from word alignments or bilingual dictionaries (we do not see a need to distinguish between the two).

b) **Pseudo-multi-lingual corpora-based approaches** that use monolingual word embedding methods on automatically constructed (or corrupted) corpora that contain words from both the source and the target language.

c) **Joint methods** that take parallel text as input and minimize the source and target language monolingual losses jointly with the cross-lingual regularization term.

We will show that *modulo* optimization strategies, these approaches are equivalent.

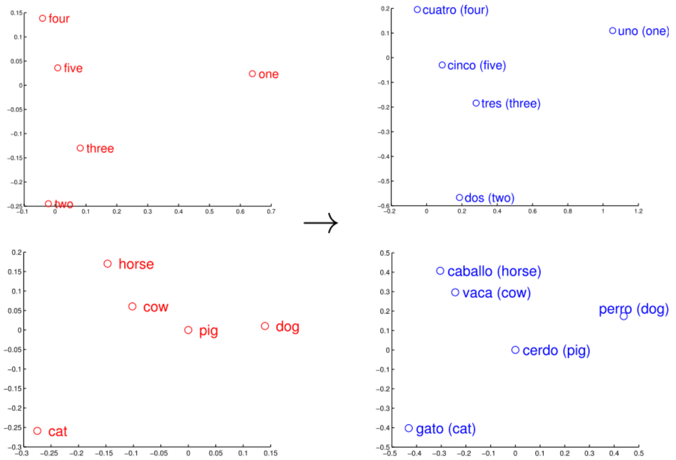


Figure 5: Similar geometric relations between numbers and animals in English and Spanish (Mikolov et al., 2013b)

6.1.1 Mapping-based approaches

Minimizing mean squared error Mikolov et al. (2013) notice that the geometric relations that hold between words are similar across languages: for instance, numbers and animals in English show a similar geometric constellation as their Spanish counterparts, as illustrated in Figure 5. This suggests that it is possible to transform the vector space of a source language s to the vector space of the target language t by learning a linear projection with a transformation matrix $\mathbf{W}^{s \rightarrow t}$. We use $\mathbf{W}$ in the following if the direction is unambiguous.

They use the $n = 5000$ most frequent words from the source language $w_1^s, \ldots, w_n^s$ and their translations $w_1^t, \ldots, w_n^t$ as seed words. They then learn $\mathbf{W}$ using stochastic gradient descent by minimising the squared Euclidean distance (mean squared error, MSE) between the previously learned monolingual representations of the source seed word $\mathbf{x}_i^s$ that is transformed using $\mathbf{W}$ and its translation $\mathbf{x}_i^t$ in the bilingual dictionary:

$$\Omega_{\text{MSE}} = \sum_{i=1}^{n} \|\mathbf{W}\mathbf{x}_i^s - \mathbf{x}_i^t\|^2 \tag{12}$$

This can also be written in matrix form as minimizing the squared Frobenius norm of the residual matrix:

$$\Omega_{\text{MSE}} = \|\mathbf{W}\mathbf{X}^s - \mathbf{X}^t\|_F^2 \tag{13}$$

where $\mathbf{X}^s$ and $\mathbf{X}^t$ are the embedding matrices of the seed words in the source and target language respectively. $\mathbf{W}$ can be computed more efficiently by taking the Moore-Penrose pseudoinverse $\mathbf{X}^+ = (\mathbf{X}^{s\top}\mathbf{X}^s)^{-1}\mathbf{X}^{S\top}$ as $\mathbf{W} = \mathbf{X}^+\mathbf{X}^t$ (Artetxe, Labaka, & Agirre, 2016).

In our notation, the MSE mapping approach can be seen as optimizing the following objective:

$$J = \underbrace{\mathcal{L}_{\text{SGNS}}(\mathbf{X}^s) + \mathcal{L}_{\text{SGNS}}(\mathbf{X}^t)}_{1} + \underbrace{\Omega_{\text{MSE}}(\underline{\mathbf{X}}^s, \underline{\mathbf{X}}^t, \mathbf{W})}_{2} \tag{14}$$

First, each of the monolingual losses is optimized independently. Second, the regularization term Ω_{MSE} is optimized while keeping the induced monolingual embeddings fixed. Several extensions to the basic mapping model as framed by Mikolov et al. (2013) have been proposed. We discuss the developments in the following. Some extensions focus on improving or altering the objective function, while another set of models puts focus on the seed lexicon used for learning the mapping.

Max-margin with intruders Dinu, Lazaridou, and Baroni (2015) discover that using MSE as the sub-objecive for learning a projection matrix leads to the issue of *hubness*: some words tend to appear as nearest neighbours of many other words (i.e., they are hubs). As translations are typically generated by choosing the nearest neighbour of a source embedding, hubs reduce the quality of the embedding space. They propose a globally corrected neighbour retrieval method to overcome this issue. Lazaridou, Dinu, and Baroni (2015) show that optimizing a max-margin based ranking loss instead of MSE reduces hubness and consequently improves performance. This max-margin based ranking loss is essentially the same as the MMHL (Collobert & Weston, 2008) used for learning monolingual embeddings. Instead of assigning higher scores to correct sentence windows, we now try to assign a higher cosine similarity to word pairs that are translations of each other ($\mathbf{x}_i^s, \mathbf{x}_i^t$; first term below) than random word pairs ($\mathbf{x}_i^s, \mathbf{x}_j^t$; second term):

$$\Omega_{\text{MMHL}} = \sum_{i=1}^{n} \sum_{j \neq i}^{k} \max\{0, \gamma - \cos(\mathbf{W}\mathbf{x}_i^s, \mathbf{x}_i^t) + \cos(\mathbf{W}\mathbf{x}_i^s, \mathbf{x}_j^t)\} \tag{15}$$

The choice of the k negative examples, which we compare against the translations is crucial. Dinu et al. propose to select negative examples that are more informative by being near the current projected vector $\mathbf{W}\mathbf{x}_i^s$ but far from the actual translation vector $\mathbf{x}_i^t$. Unlike random intruders, such intelligently chosen intruders help the model identify training instances where the model considerably fails to approximate the target function. In the formulation adopted in this article, their objective becomes:

$$J = \underbrace{\mathcal{L}_{\text{CBOW}}(\mathbf{X}^s) + \mathcal{L}_{\text{CBOW}}(\mathbf{X}^t)}_{1} + \underbrace{\Omega_{\text{MMHL-I}}(\underline{\mathbf{X}^s}, \underline{\mathbf{X}^t}, \mathbf{W})}_{2} \tag{16}$$

where $\Omega_{\text{MMHL-I}}$ designates Ω_{MMHL} with intruders as negative examples.

Smith, Turban, Hamblin, and Hammerla (2017) propose a similar solution to the hubness issue in the framework of mapping-based approaches: they invert the softmax used for finding the translation of a word at test time and normalize the probability over source words rather than target words.

Normalization and orthogonality constraint Xing, Liu, Wang, and Lin (2015) argue that there is a mismatch between the comparison function used for training word embeddings with SGNS, that is, the dot product and the function used for evaluation, which is cosine similarity. They suggest to normalize word embeddings to be unit length to address this discrepancy. In order to preserve unit length after mapping, they propose, in addition, to constrain $\mathbf{W}$ to be orthogonal: $\mathbf{W}^\top\mathbf{W} = \mathbf{I}$. The exact solution under this constraint is $\mathbf{W} = \mathbf{V}\mathbf{U}^\top$ and can be efficiently computed in linear time with respect to the vocabulary size using SVD where $\mathbf{X}^{t\top}\mathbf{X}^s = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. An orthogonality constraint is also used by Zhang, Gaddy, Barzilay, and Jaakkola (2016) for learning cross-lingual embeddings for POS tagging.

Artetxe et al. (2016) further demonstrate the similarity between different approaches by showing that the mapping model variant of Xing et al. (2015) optimizes the same loss as Mikolov et al. (2013) with an orthogonality constraint and unit vectors. In addition, they empirically show that orthogonality is more important for performance than length normalization. They also propose dimension-wise mean centering in order to capture the intuition that two randomly selected words are generally expected not to be similar and their cosine similarity should thus be zero. Smith et al. (2017) also learn a linear transformation with an orthogonality constraint and use identical character strings as their seed lexicon.

Using highly reliable seed lexicon entries The previous mapping approaches used a bilingual dictionary as an inherent component of their model, but did not pay much attention to the quality of the dictionary entries, using either automatic translations of frequent words or word alignments of all words. Vulić and Korhonen (2016) emphasize the role of the seed lexicon that is used for learning the projection matrix. They propose a hybrid model that initially learns a first shared bilingual embedding space based on an existing cross-lingual embedding model. They then use this initial vector space to obtain translations for a list of frequent source words by projecting them into the space and using the nearest neighbor in the target language as translation. They then use these translation pairs as seed words for learning a mapping. In addition, they propose a symmetry constraint: it enforces that words are included in the seed lexicon if and only if their projections are nearest neighbors of each other in the first embedding space.

Bootstrapping from few seed entries Recently, there have been initiatives towards enabling embedding induction using only a small number of seed translation pairs. If effective, such approaches would boost the induction process for truly low-resource language pairs, where only very limited amounts of bilingual data are at hand. The core idea behind these bootstrapping approaches is to start from a few seed words initially, which they then iteratively expand. Artetxe et al. (2017) propose a mapping model that relies only on a small set of shared words (e.g., identically spelled words or only shared numbers) to seed the procedure. The model has multiple bootstrapping rounds where it gradually adds more and more bilingual translation pairs to the original seed lexicon and refines it.

Smith et al. (2017) and Hauer et al. (2017) propose a method that creates seed lexicons by identifying cognates in the vocabularies of related languages. In contrast to Mikolov et al. (2013), they learn not only a transformation matrix $\mathbf{W}^{s \to t}$ that transforms embeddings of language s to embeddings of language t, but also a matrix $\mathbf{W}^{t \to s}$ that transforms embeddings in the opposite direction. Starting from a small set of automatically extracted seed translation pairs, they iteratively expand the size of the lexicon.

As discussed in Section 5, the bootstrapping idea is conceptually similar to the work of Peirsman and Padó (2010, 2011) and Vulić and Moens (2013), with the difference that earlier approaches were developed within the traditional cross-lingual distributional framework (mapping vectors into the count-based space using a seed lexicon).

Cross-lingual embeddings via retro-fitting Learning a mapping between monolingual embedding spaces can also be framed as retro-fitting (Faruqui, Dodge, Jauhar, Dyer, Hovy, & Smith, 2015), which is used to inject knowledge from semantic lexicons into pre-trained word embeddings. Retro-fitting tries to create a new word embedding matrix $\hat{\mathbf{X}}$ whose embeddings $\hat{\mathbf{x}}_i$ are both close to the corresponding learned monolingual word embeddings $\mathbf{x}_i$ as well as close to neighbors $\mathbf{x}_j$ in a knowledge graph:

$$\Omega_{\text{retro}} = \sum_{i=1}^{|V|} [\alpha_i \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|^2 + \sum_{(i,j) \in E} \beta_{ij} \|\hat{\mathbf{x}}_i - \mathbf{x}_j\|^2] \tag{17}$$

where E is the set of edges in the knowledge graph and α and β control the strength of the contribution of each term.

Mrkšić et al. (2017) similarly derive cross-lingual synonymy and antonymy constraints from BabelNet. They then use these constraints to bring the monolingual vector spaces of two different languages together into a shared embedding space. Such retro-fitting approaches employ MMHL with a careful selection of intruders, similar to the work of Lazaridou et al. (2015). In contrast to previous work, retro-fitting approaches use constraints on each word rather than a translation matrix $\mathbf{W}$ to arrive at a new cross-lingual vector space. While these constraints can capture relations that are more complex than a linear transformation, they are limited to words that are contained in the semantic lexicons.

CCA-based mapping Haghighi et al. (2008) are the first to use canonical correlation analysis (CCA) for learning translation lexicons for the words of different languages. Faruqui and Dyer (2014) later apply CCA to project words from two languages into a shared embedding space. Whereas linear projection only learns one transformation matrix $\mathbf{W}^{s \to t}$ to project the embedding space of a source language into the space of a target language, CCA learns a transformation matrix for the source and target language $\mathbf{W}^{s\rightarrow}$ and $\mathbf{W}^{t\rightarrow}$ respectively to project them into a new joint space that is different from both the space of s and of t. We can write the correlation between a source language embedding vector $\mathbf{x}_i^s$ and its corresponding target language embedding vector $\mathbf{x}_i^t$ as:

$$\rho(\mathbf{x}_i^s, \mathbf{x}_i^t) = \frac{\text{cov}(\mathbf{x}_i^s, \mathbf{x}_i^t)}{\sqrt{\text{cov}(\mathbf{x}_i^{s2})\text{cov}(\mathbf{x}_i^{t2})}} \tag{18}$$

where $\text{cov}(\mathbf{x}_i^s, \mathbf{x}_i^t)$ is the covariance of $\mathbf{x}_i^s$ and $\mathbf{x}_i^t$. CCA then aims to minimize the following:

$$\Omega_{\text{CCA}} = -\sum_{i=1}^{n} \rho(\mathbf{W}^{s\rightarrow}\mathbf{x}_i^s, \mathbf{W}^{t\rightarrow}\mathbf{x}_i^t) \tag{19}$$

We can write their objective in our notation as the following:

$$J = \underbrace{\mathcal{L}_{\text{LSA}}(\mathbf{X}^s) + \mathcal{L}_{\text{LSA}}(\mathbf{X}^t)}_{1} + \underbrace{\Omega_{\text{CCA}}(\underline{\mathbf{X}^s}, \underline{\mathbf{X}^t}, \mathbf{W}^{s\rightarrow}, \mathbf{W}^{t\rightarrow})}_{2} \tag{20}$$

As CCA sorts the projection matrices $\mathbf{W}^{s\rightarrow}$ and $\mathbf{W}^{t\rightarrow}$ in descending order, Faruqui and Dyer find that using the 80% projection vectors with the highest correlation generally yields the highest performance and that CCA helps to separate synonyms and antonyms in the source language.

Lu et al. (2015) extend Bilingual CCA to Deep Bilingual CCA by introducing non-linearity in the mapping process: they train two *deep* neural networks to maximize the correlation between the projections of both monolingual embedding spaces. Finally, Artetxe et al. (2016) show that their objective, built on top of the original or "standard" Mikolov-style mapping idea, and which uses dimension-wise mean centering is directly related to bilingual CCA. The only fundamental difference is that the CCA-based model does not guarantee monolingual invariance, while this property is enforced in the model of Artetxe et al.

6.1.2 Word-level approaches based on pseudo-bilingual corpora

Rather than learning a mapping between the source and the target language, some approaches use the word-level alignment of a seed bilingual dictionary to construct a pseudo-bilingual corpus by randomly replacing words in a source language corpus with their translations. Xiao and Guo (2014) propose the first such method. They first construct a seed bilingual dictionary by translating all words that appear in the source language corpus into the target language using Wiktionary, filtering polysemous words as well as translations that do not appear in the target language corpus. From this seed dictionary, they create a joint vocabulary, in which each translation pair occupies the same vector representation. They train this model using MMHL (Collobert & Weston, 2008) by feeding it context windows of both the source and target language corpus.

Other approaches explicitly create a pseudo-bilingual corpus: Gouws and Søgaard (2015) concatenate the source and target language corpus and replace each word that is part of a translation pair with its translation equivalent with a probability of $\frac{1}{2k_t}$, where k_t is the total number of possible translation equivalents for a word, and train CBOW on this corpus. Ammar, Mulcaire, Tsvetkov, Lample, Dyer, and Smith (2016) extend this approach to multiple languages: Using bilingual dictionaries, they determine clusters of synonymous words in different languages. They then concatenate the monolingual corpora of different languages and replace tokens in the same cluster with the cluster ID. Finally, they train SGNS on the concatenated corpus.

Duong, Kanayama, Ma, Bird, and Cohn (2016) propose a similar approach. Instead of randomly replacing every word in the corpus with its translation, they replace each center word with a translation on-the-fly during CBOW training. In addition, they handle polysemy explicitly by proposing an EM-inspired method that chooses as replacement the translation w_i^t whose representation is most similar to the sum of the source word representation $\mathbf{x}_i^s$ and the sum of the context embeddings $\mathbf{x}_s^s$ as in Equation 10:

$$w_i^t = \text{argmax}_{w' \in \tau(w_i)} \cos(\mathbf{x}_i + \mathbf{x}_s^s, \mathbf{x}') \tag{21}$$

They jointly learn to predict both the words and their appropriate translations using PanLex as the seed bilingual dictionary. PanLex covers around 1,300 language with about 12 million expressions. Consequently, translations are high coverage but often noisy. Adams, Makarucha, Neubig, Bird, and Cohn (2017) use the same approach for pre-training cross-lingual word embeddings for low-resource language modeling.

In what follows, we now show that these pseudo-bilingual models are in fact optimizing for the same objective as the mapping models discussed earlier (Mikolov et al., 2013b).

Proof for the occasional equivalence of mapping and pseudo-bilingual approaches Recall that in the negative sampling objective of SGNS in Equation 8, the probability of observing a word w with a context word c with representations $\mathbf{x}$ and $\tilde{\mathbf{x}}$ respectively is given as $P(c \mid w) = \sigma(\tilde{\mathbf{x}}^\top \mathbf{x})$, where σ denotes the sigmoid function. We now sample a set of k negative examples, that is, contexts c_i with which w does not occur, as well as actual contexts C consisting of (w_j, c_j) pairs, and try to maximize the above for actual contexts and minimize it for negative samples. Recall that Mikolov et al. (2013) obtain cross-lingual embeddings by running SGNS over two monolingual corpora of two different languages at the same time with the constraint that words known to be translation equivalents have the same representation. We will refer to this as Constrained Bilingual SGNS. Further, recall that τ is a function from words w into their translation equivalents w' (if any are available in the dictionary seed or word alignments) with the representation $\mathbf{x}'$. With some abuse of notation, this can be written as the following *joint* objective for the source language (idem for the target language):

$$\sum_{(w_j, c_j) \in C} \log \sigma({\tilde{\mathbf{x}}_j}^\top \mathbf{x}_j) + \sum_{i=1}^{k} \log \sigma(-{\tilde{\mathbf{x}}_i}^\top \mathbf{x}_j) + \Omega_\infty \sum_{w' \in \tau(w_j)} |\mathbf{x}_j - \mathbf{x}'_j| \tag{22}$$

Alternatively, we can sample sentences from the corpora in the two languages. When we encounter a word w for which we have a translation, that is, $\tau(w) \neq \emptyset$ we flip a coin and if heads, we replace w with a randomly selected member of $\tau(w)$.

In the case, where τ is bijective as in the work of Xiao and Guo (2014), it is easy to see that the two approaches are equivalent. In the limit, sampling mixed $\langle w, c \rangle$-pairs, w and $\tau(w)$ will converge to the same representations. If we loosen the requirement that τ is bijective, establishing equivalence becomes more complicated. It suffices for now to show that, regardless of the nature of τ, methods based on mixed corpora can be equivalent to methods based on regularization, and can as such also be presented and implemented as joint, regularized optimization problems.

We provide an example and a simple proof sketch here. In Constrained Bilingual SGNS, we can conflate translation equivalents; in fact, it is a common way of implementing this method. So assume the following word-context pairs: $\langle a, b\rangle, \langle a, c\rangle, \langle a, d\rangle$. The vocabulary of our source language is $\{a, b, d\}$, and the vocabulary of our target language is $\{a, c, d\}$. Let a_s denote the source language word in the word pair a; etc. To obtain a mixed corpus, such that running SGNS directly on it, will induce the same representations, in the limit, simply enumerate all combinations: $\langle a_s, b\rangle, \langle a_t, b\rangle, \langle a_s, c\rangle, \langle a_t, c\rangle, \langle a_s, d_s\rangle$, $\langle a_s, d_t\rangle, \langle a_t, d_s\rangle, \langle a_t, d_t\rangle$. Note that this is exactly the mixed corpus you would obtain in the limit with the approach by Xiao and Guo (2014). Since this reduction generalizes to all examples where τ is bijective, this suffices to show the two approaches are equivalent.

6.1.3 Joint models

While the previous approaches either optimize a set of monolingual losses and then the cross-lingual regularization term (mapping-based approaches) or optimize a monolingual loss and implicitly—via data manipulation—a cross-lingual regularization term, joint models optimize monolingual and cross-lingual objectives at the same time jointly. In what follows, we discuss a few illustrative example models which sparked this sub-line of research.

Bilingual language model Klementiev et al. (2012) cast learning cross-lingual representations as a multi-task learning problem. They jointly optimize a source language and target language model together with a cross-lingual regularization term that encourages words that are often aligned with each other in a parallel corpus to be similar. The monolingual objective is the classic LM objective of minimizing the negative log likelihood of the current word w_i given its C previous context words:

$$\mathcal{L} = -\log P(w_i \mid w_{i-C+1:i-1}) \tag{23}$$

For the cross-lingual regularization term, they first obtain an alignment matrix $\mathbf{A}^{s\rightarrow t}$ that indicates how often each source language word was aligned with each target language word from parallel data such as the Europarl corpus (Koehn, 2009). The cross-lingual regularization term then encourages the representations of source and target language words that are often aligned in $\mathbf{A}^{s\rightarrow t}$ to be similar:

$$\Omega_s = -\sum_{i=1}^{|V|^s} \frac{1}{2} \mathbf{x}_i^{s\top}(\mathbf{A}^{s\rightarrow t} \otimes \mathbf{I})\mathbf{x}_i^s \tag{24}$$

where $\mathbf{I}$ is the identity matrix and $\otimes$ is the Kronecker product. The final regularization term will be the sum of Ω_s and the analogous term for the other direction (Ω_t). Note that Equation (24) is a weighted (by word alignment scores) average of inner products, and hence, for unit length normalized embeddings, equivalent to approaches that maximize the sum of the cosine similarities of aligned word pairs. Using $\mathbf{A}^{s\rightarrow t} \otimes \mathbf{I}$ to encode interaction is borrowed from linear multi-task learning models (Cavallanti, Cesa-Bianchi, & Gentile, 2010). There, an interaction matrix $\mathbf{A}$ encodes the relatedness between tasks. The complete objective is the following:

$$J = \mathcal{L}(\mathbf{X}_s) + \mathcal{L}(\mathbf{X}_t) + \Omega(\underline{\mathbf{A}^{s\to t}}, \mathbf{X}_s) + \Omega(\underline{\mathbf{A}^{t\to s}}, \mathbf{X}_t) \tag{25}$$

Joint learning of word embeddings and word alignments Kočiský, Hermann, and Blunsom (2014) simultaneously learn word embeddings and word-level alignments using a distributed version of FastAlign (Dyer, Chahuneau, & Smith, 2013a) together with a language model.[5] Similar to other bilingual approaches, they use the word in the source language sentence of an aligned sentence pair to predict the word in the target language sentence.

They replace the standard multinomial translation probability of FastAlign with an energy function that tries to bring the representation of a target word w_i^t close to the sum of the context words around the word w_i^s in the source sentence:

$$E(w_i^s, w_i^t,) = -(\sum_{j=-C}^{C} \mathbf{x}_{i+j}^{s}{}^\top \mathbf{T})\mathbf{x}_i^t - \mathbf{b}^\top \mathbf{x}_i^t - b_{w_i^t} \tag{26}$$

where $\mathbf{x}_{i+j}^s$ and $\mathbf{x}_i^t$ are the representations of source word w_{i+j}^s and target word w_i^t respectively, $\mathbf{T} \in \mathbb{R}^{d\times d}$ is a projection matrix, and $\mathbf{b} \in \mathbb{R}^d$ and $b_{w_i^t} \in \mathbb{R}$ are representation and word biases respectively. The authors speed up training by using a class factorization strategy similar to the hierarchical softmax and predict frequency-based class representations instead of word representations. For training, they use EM but fix the alignment counts of the E-step learned by FastAlign that was initially trained for 5 epochs. They then optimize the word embeddings in the M-step only. Note that this model is conceptually very similar to bilingual models that discard word-level alignment information and learn solely on the basis of sentence-aligned information, which we discuss in Section 7.1.

Proof for the occasional equivalence of mapping-based and joint approaches We provide another informal proof, which demonstrates that the joint models optimize similar objectives as the mapping-based approaches. We call it a proof of *occasional* equivalence, since we only focus on the modeling objectives, disregarding optimization strategies and hyper-parameters, and we only discuss a particular instance of the mapping-based and joint approaches.

First recall the Constrained Bilingual SGNS from above. This model is a simple application of SGNS with the constraint that word pairs that are translation equivalents in our dictionary seed use the same representation. We now loosen the constraint that translation equivalents must have the same representation and posit instead that the distance between (the vector representation of) a word w and its translation $\tau(w)$ must be smaller than ϵ. This introduces a sphere around the null model. Fitting to the monolingual objectives now becomes a matter of finding the optimum within this bounding sphere.

Intuitively, we can think of mapping approaches as projecting our embeddings back into this bounding sphere, after fitting the monolingual objectives. Note also that this approach introduces an additional inductive bias, from the mapping algorithm itself. While joint approaches are likely to find the optimum within the bounding sphere, it is not clear that there is a projection (within the class of possible projections) from the fit to the monolingual

[5] FastAlign is a fast and effective variant of IBM Model 2.

objectives and into the optimum within the bounding sphere. It is not hard to come up with examples where such an inductive bias would hurt performance, but it remains an empirical question whether mapping-based approaches are therefore inferior on average.

In some cases, however, it is possible to show an equivalence between mapping approaches and joint approaches. Consider the mapping approach in Faruqui et al. (2015) (*retro-fitting*) and Constrained Bilingual SGNS (Xiao & Guo, 2014).

Retro-fitting requires two monolingual embeddings. Let us assume these embeddings were induced using SGNS with a set of hyper-parameters $\mathcal{H}$. Retro-fitting minimizes the weighted sum of the Euclidean distances between the seed words and their translation equivalents and their neighbors in the monolingual embeddings, with a parameter α that controls the strength of the regularizer. As this parameter goes to infinity, translation equivalents will be forced to have the same representation. As is the case in Constrained Bilingual SGNS, all word pairs in the seed dictionary will be associated with the same vector representation.

Since retro-fitting only affects words in the seed dictionary, the representation of the words not seen in the seed dictionary is determined entirely by the monolingual objectives. Again, this is the same as in Constrained Bilingual SGNS. In other words, if we fix $\mathcal{H}$ for retro-fitting and Constrained Bilingual SGNS, and set the regularization strength $\alpha = \Omega_\infty$ in retro-fitting, retro-fitting is equivalent to Constrained Bilingual SGNS.

6.2 Word-Level Alignment Methods with Comparable Data

All previous methods required word-level *parallel* data. We categorize methods that employ word-level alignment with *comparable* data into two types:

a) **Language grounding models** anchor language in images and use image features to obtain information with regard to the similarity of words in different languages.

b) **Comparable feature models** that rely on the comparability of some other features. The main feature that has been explored in this context is part-of-speech (POS) tag equivalence.

Grounding language in images Most methods employing word-aligned comparable data ground words from different languages in image data. The idea in all of these approaches is to use the image space as the shared cross-lingual signals. For instance, bicycles always look like bicycles even if they are referred to as “fiets”, “Fahrrad”, “bicikl”, “bicicletta”, or “velo”. A set of images for each word is typically retrieved using Google Image Search. Bergsma and Van Durme (2011) calculate a similarity score for a pair of words based on the visual similarity of their associated image sets. They propose two strategies to calculate the cosine similarity between the color and SIFT features of two image sets: They either take the average of the maximum similarity scores (AvgMax) or the maximum of the maximum similarity scores (MaxMax). Kiela et al. (2015) propose to do the same but use CNN-based image features. Vulić, Kiela, Clark, and Moens (2016) in addition propose to combine image and word representations either by interpolating and concatenating them or by interpolating the linguistic and visual similarity scores.

A similar idea of grounding language for learning multimodal multilingual representations can be applied for sensory signals beyond vision, e.g. auditive or olfactory signals (Kiela &

Clark, 2015). This line of work, however, is currently under-explored. Moreover, it seems that signals from other modalities are typically more useful as an additional source of information to complement the uni-modal signals from text, rather than using other modalities as the single source of information.

POS tag equivalence Other approaches rely on comparability between certain features of a word, such as its part-of-speech tag. Gouws and Søgaard (2015) create a pseudo-cross-lingual corpus by replacing words based on part-of-speech equivalence, that is, words with the same part-of-speech in different languages are replaced with one another. Instead of using the POS tags of the source and target words as a bridge between two languages, we can also use the POS tags of their contexts. This makes strong assumptions about the word orders in the two languages, and their similarity, but Duong et al. (2015) use this to obtain cross-lingual word embeddings for several language pairs. They use POS tags as context features and run SGNS on the concatenation of two monolingual corpora. Note that under the (too simplistic) assumptions that all instances of a part-of-speech have the same distribution, and each word belongs to a single part-of-speech class, this approach is equivalent to the pseudo-cross-lingual corpus approach described before.

7. Sentence-Level Alignment Methods

Thanks to research in MT, large amounts of sentence-aligned parallel data are available for European languages, which has led to much work focusing on learning cross-lingual representations from sentence-aligned parallel data. For low-resource languages or new domains, sentence-aligned parallel data is more expensive to obtain than word-aligned data as it requires fine-grained supervision. Only recently have methods started leveraging sentence-aligned comparable data.

7.1 Sentence-Level Methods with Parallel data

Methods leveraging sentence-aligned data are generally extensions of successful monolingual models. We have detected four main types:

a) **Word-alignment based matrix factorization approaches** apply matrix factorization techniques to the bilingual setting and typically require additional word alignment information.

b) **Compositional sentence models** use word representations to construct sentence representations of aligned sentences, which are trained to be close to each other.

c) **Bilingual autoencoder models** reconstruct source and target sentences using an autoencoder.

d) **Bilingual skip-gram models** use the skip-gram objective to predict words in both source and target sentences.

Word-alignment based matrix factorization Several methods directly leverage the information contained in an alignment matrix $\mathbf{A}^{s \rightarrow t}$ between source language s and target language t respectively. $\mathbf{A}^{s \rightarrow t}$ is generally automatically derived from sentence-aligned parallel text using an unsupervised word alignment model such as FastAlign (Dyer, Chahuneau, & Smith, 2013b). $\mathbf{A}_{ij}^{s\rightarrow t}$ captures the number of times the i-th word in language t was aligned with the j-th word in language s, with each row normalized to 1. The intuition is that if a word in the source language is only aligned with one word in the target language, then those words should have the same representation. If the target word is aligned with more than one source word, then its representation should be a combination of the representations of its aligned words. Zou et al. (2013) represent the embeddings $\mathbf{X}^s$ in the target language as the product of the source embeddings $\mathbf{X}^s$ with the corresponding alignment matrix $\mathbf{A}^{s\rightarrow t}$. They then minimize the squared difference between these two terms in both directions:

$$\Omega_{s\rightarrow t} = ||\mathbf{X}^t - \mathbf{A}^{s\rightarrow t}\mathbf{X}^s||^2 \tag{27}$$

Note that $\Omega_{s\rightarrow t}$ can be seen as a variant of Ω_{MSE}, which incorporates soft weights from alignments. In contrast to mapping-based approaches, the alignment matrix, which transforms source to target embeddings, is fixed in this case, while the corresponding source embeddings $\mathbf{X}^s$ are learned:

$$J = \underbrace{\mathcal{L}_{\text{MMHL}}(\mathbf{X}^t)}_{1} + \underbrace{\Omega_{s\rightarrow t}(\underline{\mathbf{X}^t}, \underline{\mathbf{A}^{s\rightarrow t}}, \mathbf{X}^s)}_{2} \tag{28}$$

Shi et al. (2015) also take into account monolingual data by placing cross-lingual constraints on the monolingual representations and propose two alignment-based cross-lingual regularization objectives. The first one treats the alignment matrix $\mathbf{A}^{s\rightarrow t}$ as a cross-lingual co-occurrence matrix and factorizes it using the GloVe objective. The second one is similar to the objective by Zou et al. (2013) and minimizes the squared distance of the representations of words in two languages weighted by their alignment probabilities.

Gardner, Huang, Paplexakis, Fu, Talukdar, Faloutsos, Sidiropoulos, Mitchell, and Sidiropoulos (2015) extend LSA as translation-invariant LSA to learn cross-lingual word embeddings. They factorize a multilingual co-occurrence matrix with the restriction that it should be invariant to translation, i.e., it should stay the same if multiplied with the respective word or context dictionary.

Vyas and Carpuat (2016) propose another method based on matrix factorization that enables learning sparse cross-lingual embeddings. As the sparse cross-lingual embeddings are different from the monolingual embeddings $\mathbf{X}$, we diverge slightly from our notation and designate them as $\mathbf{S}$. They propose two constraints: The first constraint induces monolingual sparse representations from pre-trained monolingual embedding matrices $\mathbf{X}^s$ and $\mathbf{X}^t$ by factorizing each embedding matrix $\mathbf{X}$ into two matrices $\mathbf{S}$ and $\mathbf{E}$ with an additional ℓ_1 constraint on $\mathbf{S}$ for sparsity:

$$\mathcal{L} = \sum_{i=1}^{|V|} \|\mathbf{S}_i\mathbf{E}^\top - \mathbf{X}_i\|_2^2 + \lambda\|\mathbf{S}_i\|_1 \tag{29}$$

To learn bilingual embeddings, they add another constraint based on the alignment matrix $\mathbf{A}^{s\rightarrow t}$ that minimizes the ℓ_2 reconstruction error between words that were strongly aligned to each other in a parallel corpus:

$$\Omega = \sum_{i=1}^{|V^s|}\sum_{j=1}^{|V^t|} \frac{1}{2}\lambda_x \mathbf{A}_{ij}^{s\rightarrow t}\|\mathbf{S}_i^s - \mathbf{S}_j^t\|_2^2 \tag{30}$$

The complete optimization then consists of first pre-training monolingual embeddings $\mathbf{X}^s$ and $\mathbf{X}^t$ with GloVe and in a second step factorizing the monolingual embeddings with the cross-lingual constraint to induce cross-lingual sparse representations $\mathbf{S}^s$ and $\mathbf{S}^t$:

$$J = \underbrace{\mathcal{L}_{\text{GloVe}}(\mathbf{X}^s) + \mathcal{L}_{\text{GloVe}}(\mathbf{X}^t)}_{1} + \underbrace{\mathcal{L}(\underline{\mathbf{X}^s}, \mathbf{S}^s, \mathbf{E}^s) + \mathcal{L}(\underline{\mathbf{X}^t}, \mathbf{S}^t, \mathbf{E}^t) + \Omega(\underline{\mathbf{A}^{s \to t}}, \mathbf{S}^s, \mathbf{S}^t)}_{2} \tag{31}$$

Guo, Che, Yarowsky, Wang, and Liu (2015) similarly create target a target language word embedding $\mathbf{x}_i^t$ of a source word w_i^s by taking the average of the embeddings of its translations $\tau(w_i^s)$ weighted by their alignment probability with the source word:

$$\mathbf{x}_i^t = \sum_{w_j^t \in \tau(w_i^s)} \frac{\mathbf{A}_{i,j}}{\mathbf{A}_{i,\cdot}} \cdot \mathbf{x}_j^t \tag{32}$$

They propagate alignments to out-of-vocabulary (OOV) words using edit distance as an approximation for morphological similarity and set the target word embedding $\mathbf{x}_k^t$ of an OOV source word w_k^s as the average of the projected vectors of source words that are similar to it based on the edit distance measure:

$$\mathbf{x}_k^t = \frac{1}{|E_k|} \sum_{w^s \in E_k} \mathbf{x}^t \tag{33}$$

where $\mathbf{x}^t$ is the target language word embedding of a source word w^s as created above, $E_k = \{w^s \mid \mathit{EditDist}(w_k^s, w^s) \leq \chi\}$, and χ is set empirically to 1.

Compositional sentence model Hermann and Blunsom (2013) train a model to bring the sentence representations of aligned sentences sent^s and sent^t in source and target language s and t respectively close to each other. The representation $\mathbf{y}^s$ of sentence sent^s in language s is the sum of the embeddings of its words:

$$\mathbf{y}^s = \sum_{i=1}^{|\mathit{sent}^s|} \mathbf{x}_i^s \tag{34}$$

They seek to minimize the distance between aligned sentences sent^s and sent^t:

$$E_{\mathit{dist}}(\mathit{sent}^s, \mathit{sent}^t) = \|\mathbf{y}^s - \mathbf{y}^t\|^2 \tag{35}$$

They optimize this distance using MMHL by learning to bring aligned sentences closer together than randomly sampled negative examples:

$$\mathcal{L} = \sum_{(\mathit{sent}^s, \mathit{sent}^t) \in \mathcal{C}} \sum_{i=1}^{k} \max(0, 1 + E_{\mathit{dist}}(\mathit{sent}^s, \mathit{sent}^t) - E_{\mathit{dist}}(\mathit{sent}^s, s_i^t)) \tag{36}$$

where k is the number of negative examples. In addition, they use an ℓ_2 regularization term for each language $\Omega = \frac{\lambda}{2}\|\mathbf{X}\|^2$ so that the final loss they optimize is the following:

$$J = \mathcal{L}(\mathbf{X}^s, \mathbf{X}^t) + \Omega(\mathbf{X}^s) + \Omega(\mathbf{X}^t) \tag{37}$$

Note that compared to most previous approaches, there is no dedicated monolingual objective and all loss terms are optimized jointly. Note that in this case, the ℓ_2 norm is applied to representations $\mathbf{X}$, which are computed as the difference of sentence representations.

This regularization term *approximates* minimizing the mean squared error between the pair-wise interacting source and target words in a way similar to Gouws, Bengio, and Corrado (2015). To see this, note that we minimize the squared error between source and target representations, i.e. Ω_{MSE}—this time only not with regard to word embeddings but with regard to sentence representations. As we saw, these sentence representations are just the sum of their constituent word embeddings. In the limit of infinite data, we therefore implicitly optimize Ω_{MSE} over word representations.

Hermann and Blunsom (2014) extend this approach to documents, by applying the composition and objective function recursively to compose sentences into documents. They additionally propose a non-linear composition function based on bigram pairs, which outperforms simple addition on large training datasets, but underperforms it on smaller data:

$$\mathbf{y} = \sum_{i=1}^{n} \tanh(\mathbf{x}_{i-1} + \mathbf{x}_i) \tag{38}$$

Soyer, Stenetorp, and Aizawa (2015) augment this model with a monolingual objective that operates on the phrase level. The objective uses MMHL and is based on the assumption that phrases are typically more similar to their sub-phrases than to randomly sampled phrases:

$$\mathcal{L} = [\max(0, m + \|\mathbf{a}_o - \mathbf{a}_i\|^2 - \|\mathbf{a}_o - \mathbf{b}_n\|^2) + \|\mathbf{a}_o - \mathbf{a}_i\|^2]\frac{n_i}{n_o} \tag{39}$$

where m is a margin, $\mathbf{a}_o$ is a phrase of length n_o sampled from a sentence, $\mathbf{a}_i$ is a sub-phrase of $\mathbf{a}_o$ of length n_i, and $\mathbf{b}_n$ is a phrase sampled from a random sentence. The additional loss terms are meant to reduce the influence of the margin as a hyperparameter and to compensate for the differences in phrase and sub-phrase length.

Bilingual autoencoder Instead of minimizing the distance between two sentence representations in different languages, Lauly et al. (2013) aim to reconstruct the target sentence from the original source sentence. Analogously to Hermann and Blunsom (2013), they also encode a sentence as the sum of its word embeddings. They then train an auto-encoder with language-specific encoder and decoder layers and hierarchical softmax to reconstruct from each sentence the sentence itself and its translation. In this case, the encoder parameters are the word embedding matrices $\mathbf{X}^s$ and $\mathbf{X}^t$, while the decoder parameters are transformation matrices that project the encoded representation to the output language space. The loss they optimize can be written as follows:

$$J = \mathcal{L}_{\text{AUTO}}^{s\rightarrow s} + \mathcal{L}_{\text{AUTO}}^{t\rightarrow t} + \mathcal{L}_{\text{AUTO}}^{s\rightarrow t} + \mathcal{L}_{\text{AUTO}}^{t\rightarrow s} \tag{40}$$

where $\mathcal{L}_{\text{AUTO}}^{s\rightarrow t}$ denotes the loss for reconstructing from a sentence in language s to a sentence in language t. Aligned sentences are sampled from parallel text and all losses are optimized jointly.

Chandar, Lauly, Larochelle, Khapra, Ravindran, Raykar, and Saha (2014) use a binary BOW instead of the hierarchical softmax. To address the increase in complexity due to the

Model	Alignment model	Monolingual losses	Cross-lingual regularizer
BilBOWA (Gouws et al., 2015)	Uniform	$\mathcal{L}^s_{\text{SGNS}} + \mathcal{L}^t_{\text{SGNS}}$	Ω_{BILBOWA}
Trans-gram (Coulmance, Marty, Wenzek, & Benhalloum, 2015)	Uniform	$\mathcal{L}^s_{\text{SGNS}} + \mathcal{L}^t_{\text{SGNS}}$	$\Omega^{s\rightarrow t}_{\text{SGNS}} + \Omega^{t\rightarrow s}_{\text{SGNS}}$
BiSkip (Luong et al., 2015)	Monotonic	$\mathcal{L}^s_{\text{SGNS}} + \mathcal{L}^t_{\text{SGNS}}$	$\Omega^{s\rightarrow t}_{\text{SGNS}} + \Omega^{t\rightarrow s}_{\text{SGNS}}$

Table 4: A comparison of similarities and differences of the three bilingual skip-gram variants.

higher dimensionality of the BOW, they propose to merge the bags-of-words in a mini-batch into a single BOW and to perform updates based on this merged bag-of-words. They also add a term to the objective function that encourages correlation between the source and target sentence representations by summing the scalar correlations between all dimensions of the two vectors.

Bilingual skip-gram Several authors propose extensions of the monolingual skip-gram with negative sampling (SGNS) model to learn cross-lingual embeddings. We show their similarities and differences in Table 4. All of these models jointly optimize monolingual SGNS losses for each language together with one more cross-lingual regularization terms:

$$J = \mathcal{L}^s_{\text{SGNS}} + \mathcal{L}^t_{\text{SGNS}} + \Omega \tag{41}$$

Another commonality is that these models do not require word alignments of aligned sentences. Instead, they make different assumptions about the alignment of the data. Bilingual Bag-of-Words without Word Alignments (BilBOWA) (Gouws et al., 2015) assumes each word in a source sentence is aligned with *every* word in the target sentence. If we knew the word alignments, we would try to bring the embeddings of aligned words in source and target sentences close together. Instead, under a uniform alignment model which perfectly matches the intuition behind the simplest (lexical) word alignment IBM Model 1 (Brown et al., 1993), we try to bring the *average* alignment close together. In other words, we use the means of the word embeddings in a sentence as the sentence representations $\mathbf{y}$ and seek to minimize the distance between aligned sentence representations:

$$\mathbf{y}^s = \frac{1}{|sent^s|} \sum_{i=1}^{|sent^s|} \mathbf{x}^s_i \tag{42}$$

$$\Omega_{\text{BILBOWA}} = \sum_{(sent^s, sent^t) \in \mathcal{C}} \|\mathbf{y}^s - \mathbf{y}^t\|^2 \tag{43}$$

Note that this regularization term is very similar to the objective used in the compositional sentence model (Hermann & Blunsom, 2013) (Equations 34 and 35); the only difference is that we use the mean rather than the sum of word embeddings as sentence representations.

Trans-gram (Coulmance et al., 2015) (2015) also assumes uniform alignment but uses the SGNS objective as cross-lingual regularization term. Recall that skip-gram with negative sampling seeks to train a model to distinguish context words from negative samples drawn from a noise distribution based on a center word. In the cross-lingual case, we aim to predict words in the aligned target language sentence based on words in the source sentence. Under uniform alignment, we aim to predict *all* words in the target sentence based on each word in the source sentence:

$$\Omega^{s \to t}_{\text{SGNS}} = - \sum_{(sent^s, sent^t) \in \mathcal{C}} \frac{1}{|sent^s|} \sum_{t=1}^{|sent^s|} \sum_{j=1}^{|sent^t|} \log P(w_{t+j} \mid w_t) \tag{44}$$

where $P(w_{t+j} \mid w_t)$ is computed via negative sampling as in Equation 8.

BiSkip (Luong et al., 2015) uses the same cross-lingual regularization terms as Trans-gram, but only aims to predict monotonically aligned target language words: Each source word at position i in the source sentence $sent^s$ is aligned to the target word at position $i \cdot \frac{|sent^s|}{|sent^t|}$ in the target sentence $sent^t$. In practice, all these bilingual skip-gram models are trained by sampling a pair of aligned sentences from a parallel corpus and minimizing for the source and target language sentence the respective loss terms.

In a similar vein, Pham, Luong, and Manning (2015) propose an extension of paragraph vectors (Le & Mikolov, 2014) to the multilingual setting by forcing aligned sentences of different languages to share the same vector representation.

Other sentence-level approaches Levy et al. (2017) use another sentence-level bilingual signal: IDs of the aligned sentence pairs in a parallel corpus. Their model provides a strong baseline for cross-lingual embeddings that is inspired by the Dice aligner commonly used for producing word alignments for MT. Observing that sentence IDs are already a powerful bilingual signal, they propose to apply SGNS to the word-sentence ID matrix. They show that this method can be seen as a generalization of the Dice Coefficient.

Rajendran, Khapra, Chandar, and Ravindran (2016) propose a method that exploits the idea of using pivot languages, also tackled in previous work, e.g., Shezaf and Rappoport (2010). The model requires parallel data between each language and a pivot language and is able to learn a shared embedding space for two languages without any direct alignment signals as the alignment is implicitly learned via their alignment with the pivot language. The model optimizes a correlation term with neural network encoders and decoders that is similar to the objective of the CCA-based approaches (Faruqui & Dyer, 2014b; Lu et al., 2015). We discuss the importance of pivoting for learning multilingual word embeddings later in Section 9.

7.2 Sentence Alignment with Comparable Data

Grounding language in images Similarly to approaches based on word-level aligned comparable data, methods that learn cross-lingual representations using sentence alignment with comparable data do so by associating sentences with images (Elliott & Kádár, 2017). The associated image captions/annotations can be direct translations of each other, but are not expected to be in general. The images are then used as pivots to induce a shared multimodal embedding space. These approaches typically use Multi30k (Elliott, Frank, Sima'an, & Specia, 2016), a multilingual extension of the Flickr30k dataset (Young, Lai, Hodosh, & Hockenmaier, 2014), which contains 30k images and 5 English sentence descriptions and their German translations for each image. Calixto et al. (2017) represent images using features from a pre-trained CNN and model sentences using a GRU. They then use MMHL to assign a higher score to image-description pairs compared to images with a random description. Gella et al. (2017) augment this objective with another MMHL term that also brings the representations of translated descriptions closer together, thus effectively combining learning signals from both visual and textual modality.

8. Document-Level Alignment Models

Models that require parallel document alignment presuppose that sentence-level parallel alignment is also present. Such models thus reduce to parallel sentence-level alignment methods, which have been discussed in the previous section. Comparable document-level alignment, on the other hand, is more appealing as it is often cheaper to obtain. Existing approaches generally use Wikipedia documents, which they either automatically align, or they employ already theme-aligned Wikipedia documents discussing similar topics.

8.1 Document Alignment with Comparable Data

We divide models using document alignment with comparable data into three types, some of which employ similar general ideas to previously discussed word and sentence-level parallel alignment models:

a) **Approaches based on pseudo-bilingual document-aligned corpora** automatically construct a pseudo-bilingual corpus containing words from the source and target language by mixing words from document-aligned documents.

b) **Concept-based methods** leverage information about the distribution of latent topics or concepts in document-aligned data to represent words.

c) **Extensions of sentence-aligned models** extend methods using sentence-aligned parallel data to also work without parallel data.

Pseudo-bilingual document-aligned corpora The approach of Vulić and Moens (2016) is similar to the pseudo-bilingual corpora approaches discussed in Section 6. In contrast to previous methods, they propose a *Merge and Shuffle* strategy to merge two aligned documents of different languages into a pseudo-bilingual document. This is done by concatenating the documents and then randomly shuffling them by permuting words. The intuition is that as most methods rely on learning word embeddings based on their context, shuffling the documents will lead to robust bilingual contexts for each word. As the shuffling step is completely random, it might lead to sub-optimal configurations.

For this reason, they propose another strategy for merging the two aligned documents, called *Length-Ratio Shuffle*. It assumes that the structures of the two documents are similar: words are inserted into the pseudo-bilingual document by alternating between the source and the target document relying on the order in which they appear in their monolingual document and based on the monolingual documents' length ratio. The whole process can be seen in Figure 6.

Concept-based models Some methods for learning cross-lingual word embeddings leverage the insight that words in different languages are similar if they are used to talk about or evoke the same multilingual concepts or topics. Vulić and Moens (2013) base their method on the cognitive theory of semantic word responses. Their method centers on the intuition

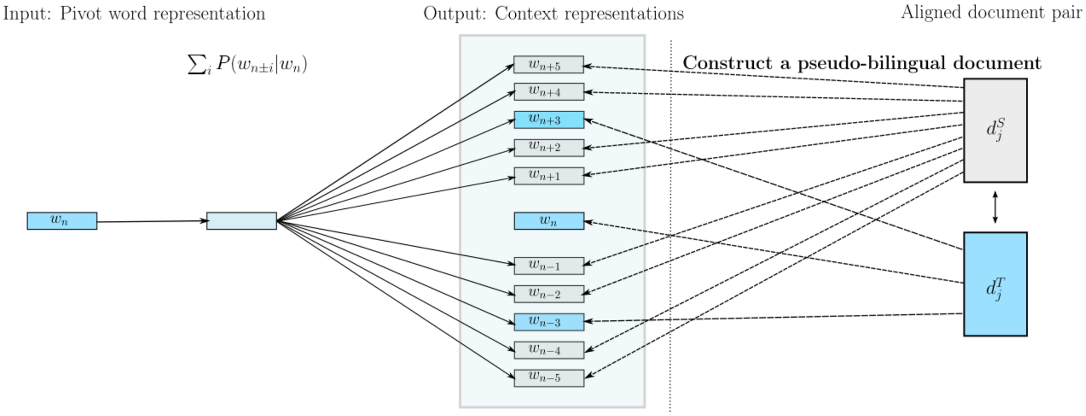


Figure 6: The Length-Ratio Shuffle strategy (Vulić & Moens, 2016)

that words in source and target language are similar if they are likely to generate similar words as their top semantic word responses. They utilise a probabilistic multilingual topic model again trained on aligned Wikipedia documents to learn and quantify semantic word responses. The embedding $\mathbf{x}_i^s \in \mathbb{R}^{|V^s|+|V^t|}$ of source word w_i is the following vector:

$$\mathbf{x}_i^s = [P(w_1^s|w_i), \ldots, P(w_{|V^s|}^s|w_i), P(w_1^t|w_i) \ldots, P(w_{|V^t|}^t|w_i)] \tag{45}$$

where $[\cdot, \cdot]$ represents concatenation and $P(w_j|w_i)$ is the probability of w_j given w_i under the induced bilingual topic model. The sparse representations may be turned into dense vectors by factorizing the constructed word-response matrix.

Søgaard et al. (2015) propose an approach that relies on the structure of Wikipedia itself. Their method is based on the intuition that similar words are used to describe the same concepts across different languages. Instead of representing every Wikipedia concept with the terms that are used to describe it, they use an inverted index and represent a word by the concepts it is used to describe. As a post-processing step, dimensionality reduction on the produced word representations in the word-concept matrix is performed. A very similar model by (Vulić et al., 2011) uses a bilingual topic model to perform the dimensionality reduction step and learns a shared cross-lingual topical space.

Extensions of sentence-alignment models Mogadala and Rettinger (2016) extend the approach of Pham et al. (2015) to also work without parallel data and adjust the regularization term Ω based on the nature of the training corpus. Similar to previous work (Hermann & Blunsom, 2013; Gouws et al., 2015), they use the mean of the word embeddings of a sentence as the sentence representation $\mathbf{y}$ and constrain these to be close together. In addition, they propose to constrain the sentence paragraph vectors $\mathbf{p}^s$ and $\mathbf{p}^t$ of aligned sentences $sent^s$ and $sent^t$ to be close to each other. These vectors are learned via paragraph vectors (Le & Mikolov, 2014) for each sentence and stored in embedding matrices $\mathbf{P}^s$ and $\mathbf{P}^t$. The complete regularizer then uses elastic net regularization to combine both terms:

$$\Omega = \sum_{(sent^s, sent^t) \in \mathcal{C}} \alpha||\mathbf{p}^s - \mathbf{p}^t||^2 + (1-\alpha)\|\mathbf{y}^s - \mathbf{y}^t\|^2 \tag{46}$$

The monolingual paragraph vector objectives $\mathcal{L}_{\text{SGNS-P}}$ are then optimized jointly with the cross-lingual regularization term:

$$J = \mathcal{L}^{s}_{\text{SGNS-P}}(\mathbf{P}^{s}, \mathbf{X}^{s}) + \mathcal{L}^{t}_{\text{SGNS-P}}(\mathbf{P}^{t}, \mathbf{X}^{t}) + \Omega(\mathbf{P}^{s}, \mathbf{P}^{t}, \mathbf{X}^{s}, \mathbf{X}^{t}) \tag{47}$$

To leverage data that is not sentence-aligned, but where an alignment is still present on the document level, they propose a two-step approach: They use Procrustes analysis (Schönemann, 1966), a method for statistical shape analysis, to find the most similar document in language t for each document in language s. This is done by first learning monolingual representations of the documents in each language using paragraph vectors on each corpus. Subsequently, Procrustes analysis aims to learn a transformation between the two vector spaces by translating, rotating, and scaling the embeddings in the first space until they most closely align to the document representations in the second space. In the second step, they then simply use the previously described method to learn cross-lingual word representations from the alignment documents, this time treating the entire documents as paragraphs.

As a final overview, we list all approaches with their monolingual objectives and regularization terms in Table 5. The table is meant to reveal the *high-level* objectives and losses each model is optimizing. It also indicates for each method whether all objectives are jointly optimized; if they are, both monolingual losses and regularization terms are optimized jointly; otherwise the monolingual losses are optimized first and the monolingual variables are frozen, while the cross-lingual regularization constraint is optimized. The table obscures smaller differences and implementation details, which can be found in the corresponding sections of this survey or by consulting the original papers. We use Ω_{∞} to represent an infinitely stronger regularizer that enforces equality between representations. Regularizers with a * imply that the regularization is achieved in the limit, e.g. in the pseudo-bilingual case, where examples are randomly sampled with some equivalence, we obtain the same representation in the limit, without strictly enforcing it to be the same representation.

As we have demonstrated, most approaches can be seen as optimizing a combination of monolingual losses with a regularization term. As we can see, some approaches do not employ a regularization term; notably, a small number of approaches, i.e., those that ground language in images, do not optimize a loss but rather use pre-trained image features and a set of similarity heuristics to retrieve translations.

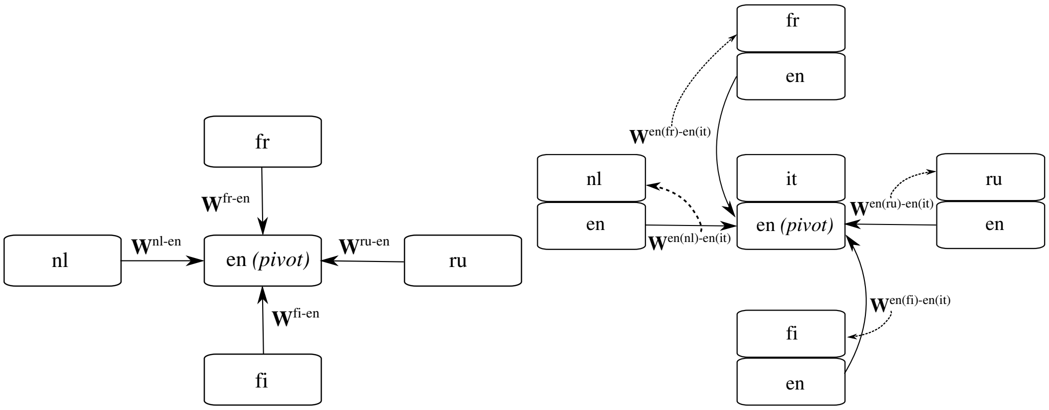
9. From Bilingual to Multilingual Training

So far, for the sake of simplicity and brevity of presentation, we have put focus on models which induce cross-lingual word embeddings in a shared space comprising only two languages. This standard bilingual setup is also in the focus of almost all research in the field of cross-lingual embedding learning. However, notable exceptions such as the recent work of Levy et al. (2017) and Duong, Kanayama, Ma, Bird, and Cohn (2017) demonstrate that there are clear benefits to extending the learning process from bilingual to *multilingual* settings, with improved results reported on standard tasks such as word similarity prediction, bilingual dictionary induction, document classification and dependency parsing.

The usefulness of multilingual training for NLP is already discussed by, e.g., Naseem, Snyder, Eisenstein, and Barzilay (2009) and Snyder and Barzilay (2010). They corroborate a hypothesis that "variations in ambiguity" may be used as a form of naturally occurring

Approach	Monolingual	Regularizer	Joint?	Description
Klementiev et al. (2012)	$\mathcal{L}_{\text{MLE}}$	Ω_{MSE}	✓	Joint
Mikolov et al. (2013)	$\mathcal{L}_{\text{SGNS}}$	Ω_{MSE}		Projection-based
Zou et al. (2013)	$\mathcal{L}_{\text{MMHL}}$	Ω_{MSE}		Matrix factorization
Hermann and Blunsom (2013)	$\mathcal{L}_{\text{MMHL}}$	Ω^*_{MSE}	✓	Sentence-level, joint
Hermann and Blunsom (2014)	$\mathcal{L}_{\text{MMHL}}$	Ω^*_{MSE}	✓	Sentence-level + bigram composition
Soyer et al. (2015)	$\mathcal{L}_{\text{MMHL}}$	Ω^*_{MSE}	✓	Phrase-level
Shi et al. (2015)	$\mathcal{L}_{\text{MMHL}}$	Ω_{MSE}		Matrix factorization
Dinu et al. (2015)	$\mathcal{L}_{\text{SGNS}}$	Ω_{MSE}		Better neighbour retrieval
Gouws et al. (2015)	$\mathcal{L}_{\text{SGNS}}$	Ω_{MSE}	✓	Sentence-level
Vyas and Carpuat (2016)	$\mathcal{L}_{\text{GloVe}}$	Ω_{MSE}		Sparse matrix factorization
Hauer et al. (2017)	$\mathcal{L}_{\text{SGNS}}$	Ω_{MSE}		Cognates
Mogadala and Rettinger (2016)	$\mathcal{L}_{\text{SGNS-P}}$	Ω_{MSE}	✓	Elastic net, Procrustes analysis
Xing et al. (2015)	$\mathcal{L}_{\text{SGNS}}$	Ω_{MSE} s.t. $\mathbf{W}^\top\mathbf{W} = \mathbf{I}$		Normalization, orthogonality
Zhang et al. (2016)	$\mathcal{L}_{\text{SGNS}}$	Ω_{MSE} s.t. $\mathbf{W}^\top\mathbf{W} = \mathbf{I}$		Orthogonality constraint
Artexte et al. (2016)	$\mathcal{L}_{\text{SGNS}}$	Ω_{MSE} s.t. $\mathbf{W}^\top\mathbf{W} = \mathbf{I}$		Normalization, orthogonality, mean centering
Smith et al. (2017)	$\mathcal{L}_{\text{SGNS}}$	Ω_{MSE} s.t. $\mathbf{W}^\top\mathbf{W} = \mathbf{I}$		Orthogonality, inverted softmax identical character strings
Artexte et al. (2017)	$\mathcal{L}_{\text{SGNS}}$	Ω_{MSE} s.t. $\mathbf{W}^\top\mathbf{W} = \mathbf{I}$		Normalization, orthogonality, mean centering, bootstrapping
Lazaridou et al. (2015)	$\mathcal{L}_{\text{CBOW}}$	Ω_{MMHL}		Max-margin with intruders
Mrkšić et al. (2017)	$\mathcal{L}_{\text{SGNS}}$	Ω_{MMHL}		Semantic specialization
Calixto et al. (2017)	$\mathcal{L}_{\text{RNN}}$	Ω_{MMHL}	✓	Image-caption pairs
Gella et al. (2017)	$\mathcal{L}_{\text{RNN}}$	Ω_{MMHL}	✓	Image-caption pairs
Faruqui and Dyer (2014)	$\mathcal{L}_{\text{LSA}}$	Ω_{CCA}		-
Lu et al. (2015)	$\mathcal{L}_{\text{LSA}}$	Ω_{CCA}		Neural CCA
Rajendran et al. (2016)	$\mathcal{L}_{\text{AUTO}}$	Ω_{CCA}		Pivots
Ammar et al. (2016)	$\mathcal{L}_{\text{LSA}}$	Ω_{CCA}		Multilingual CCA
Søgaard et al. (2015)	-	Ω_{SVD}	✓	Inverted indexing
Levy et al. (2017)	$\mathcal{L}_{\text{PMI}}$	Ω_{SVD}	✓	
Levy et al. (2017)	-	Ω_{SGNS}	✓	Inverted indexing
Lauly et al. (2013)	$\mathcal{L}_{\text{AUTO}}$	Ω_{AUTO}	✓	Autoencoder
Chandar et al. (2014)	$\mathcal{L}_{\text{AUTO}}$	Ω_{AUTO}	✓	Autoencoder
Vulić and Moens (2013a)	$\mathcal{L}_{\text{LDA}}$	Ω^*_∞	✓	Document-level
Vulić and Moens (2014)	$\mathcal{L}_{\text{LDA}}$	Ω^*_∞	✓	Document-level
Xiao and Guo (2014)	$\mathcal{L}_{\text{MMHL}}$	Ω_∞	✓	Pseudo-multilingual
Gouws and Søgaard (2015)	$\mathcal{L}_{\text{CBOW}}$	Ω^*_∞	✓	Pseudo-multilingual
Luong et al. (2015)	$\mathcal{L}_{\text{SGNS}}$	Ω^*_∞		Monotonic alignment
Gardner et al. (2015)	$\mathcal{L}_{\text{LSA}}$	Ω^*_∞		Matrix factorization
Pham et al. (2015)	$\mathcal{L}_{\text{SGNS-P}}$	Ω_∞	✓	Paragraph vectors
Guo et al. (2015)	$\mathcal{L}_{\text{CBOW}}$	Ω_∞		Weighted by word alignments
Coulmance et al. (2015)	$\mathcal{L}_{\text{SGNS}}$	Ω^*_∞	✓	Sentence-level
Ammar et al. (2016)	$\mathcal{L}_{\text{SGNS}}$	Ω_∞	✓	Pseudo-multilingual
Vulić and Korhonen (2016)	$\mathcal{L}_{\text{SGNS}}$	Ω_∞		Highly reliable seed entries
Duong et al. (2016)	$\mathcal{L}_{\text{CBOW}}$	Ω_∞	✓	Pseudo-multilingual, polysemy
Vulić and Moens (2016)	$\mathcal{L}_{\text{SGNS}}$	Ω_∞	✓	Pseudo-multilingual documents
Adams et al. (2017)	$\mathcal{L}_{\text{CBOW}}$	Ω_∞	✓	Pseudo-multilingual, polysemy
Bergsma and Van Durme (2011)	-	-	✓	SIFT image features, similarity
Kiela et al. (2015)	-	-	✓	CNN image features, similarity
Vulić et al. (2016)	-	-	✓	CNN features, similarity, interpolation
Gouws and Søgaard (2015)	$\mathcal{L}_{\text{CBOW}}$	POS-level Ω^*_∞	✓	Pseudo-multilingual
Duong et al. (2015)	$\mathcal{L}_{\text{CBOW}}$	POS-level Ω^*_∞	✓	Pseudo-multilingual

Table 5: Overview of approaches with monolingual objectives and regularization terms, with an indication whether the order of optimization matters and short descriptions. Ω_∞ represents an infinitely strong regularizer that enforces equality between representations. * implies that the regularization is achieved in the limit.


fr
$\mathbf{W}^{\text{fr-en}}$
nl
$\mathbf{W}^{\text{nl-en}}$
en (pivot)
$\mathbf{W}^{\text{ru-en}}$
ru
$\mathbf{W}^{\text{fi-en}}$
fi
fr
en
$\mathbf{W}^{\text{en(fr)-en(it)}}$
nl
en
it
en (pivot)
ru
en
$\mathbf{W}^{\text{en(nl)-en(it)}}$
$\mathbf{W}^{\text{en(ru)-en(it)}}$
$\mathbf{W}^{\text{en(fi)-en(it)}}$
fi
en

(a) Starting spaces: monolingual

(b) Starting spaces: bilingual

Figure 7: Learning shared multilingual embedding spaces via linear mapping. (a) Starting from monolingual spaces in L languages, one linearly maps $L-1$ into one chosen pivot monolingual space (typically English); (b) Starting from bilingual spaces sharing a language (typically English), one learns mappings from all other English subspaces into one chosen pivot English subspace and then applies the mapping to all other subspaces.

supervision. In simple words, what one language leaves implicit, another defines explicitly and the target language is thus useful for resolving disambiguation in the source language (Faruqui & Dyer, 2014b). While this hypothesis is already true for bilingual settings, using additional languages introduces additional supervision signals which in turn leads to better word embedding estimates (Mrkšić et al., 2017b).

In most of the literature focused on bilingual settings, English is typically on one side, owing its wide use to the wealth of both monolingual resources available for English as well as bilingual resources, where English is paired with many other languages. However, one would ideally want to also exploit cross-lingual links between other language pairs, reaching beyond English. For instance, typologically/structurally more similar languages such as Finnish and Estonian are excellent candidates for transfer learning. Yet, only few readily available parallel resources exist between Finnish and Estonian that could facilitate a direct induction of a shared bilingual embedding space in these two languages.

A multilingual embedding model which maps Finnish and Estonian to the same embedding space shared with English (i.e., English is used as a resource-rich *pivot language*) would also enable exploring and utilizing links between Finnish and Estonian lexical items in the space (Duong et al., 2017). Further, multilingual shared embedding spaces enable multi-source learning and multi-source transfers: this results in a more general model and is less prone to data sparseness (McDonald et al., 2011; Agić, Johannsen, Plank, Alonso, Schluter, & Søgaard, 2016; Guo, Che, Yarowsky, Wang, & Liu, 2016; Zoph & Knight, 2016; Firat, Cho, & Bengio, 2016).

The purpose of this section is not to demonstrate the multilingual extension of every single bilingual model discussed in previous sections, as these extensions are not always straightforward and include additional modeling work. However, we will briefly present and discuss several best practices and multilingual embedding models already available in the literature, again following the typology of models established in Table 3.

9.1 Multilingual Word Embeddings from Word-Level Information

Mapping-based approaches Mapping L different monolingual spaces into a single multilingual shared space is achieved by: (1) selecting one space as the *pivot space*, and then (2) mapping the remaining $L-1$ spaces into the same pivot space. This approach, illustrated by Figure 7a, requires L monolingual spaces and $L-1$ seed translation dictionaries to achieve the mapping. Labeling the pivot language as l^p, we can formulate the induction of a multilingual embedding space as follows:

$$\mathcal{L}^1 + \mathcal{L}^2 + \ldots + \mathcal{L}^{L-1} + \mathcal{L}^p + \Omega^{l^1 \to l^p} + \Omega^{l^2 \to l^p} + \ldots + \Omega^{l^{L-1} \to l^p} \tag{48}$$

This means that through *pivoting* one is able to induce a shared bilingual space for a language pair without having any directly usable bilingual resources for the pair. Exactly this multilingual mapping procedure (based on minimizing mean squared errors) has been constructed by Smith et al. (2017): English is naturally selected as the pivot, and 89 other languages are then mapped into the pivot English space. Seed translation pairs are obtained through Google Translate API by translating the 5,000 most frequent words in each language to English. The recent work of, e.g., Artetxe et al. (2017) holds promise that seed lexicons of similar sizes may also be bootstrapped for resource-lean languages from very small seed lexicons (see again Section 6). Smith et al. use original fastText vectors available in 90 languages (Bojanowski, Grave, Joulin, & Mikolov, 2017)[6] and effectively construct a multilingual embedding space spanning 90 languages (i.e., 4005 language pairs using 89 seed translation dictionaries) in their software and experiments.[7] The distances in all monolingual spaces remain preserved by constraining the transformation to be orthogonal.

Along the same line, Ammar et al. (2016) introduce a multilingual extension of the CCA-based mapping approach. They perform a multilingual extension of bilingual CCA projection again using the English embedding space as the pivot for multiple *English-*l^t bilingual CCA projections with the remaining $L-1$ languages.

As demonstrated by Smith et al. (2017), the multilingual space now enables reasoning for language pairs not represented in the seed lexicon data. They verify this hypothesis by examining the bilingual lexicon induction task for all $\binom{L}{2}$ language pairs: e.g., BLI Precision-at-one (P@1) scores[8] for Spanish-Catalan without any seed Spanish-Catalan lexicon are 0.82, while the average P@1 score for Spanish-English and Catalan-English bilingual spaces is 0.70. Other striking findings include P@1 scores for Russian-Ukrainian (0.84 vs. 0.59), Czech-Slovak (0.82 vs. 0.59), Serbian-Croatian (0.78 vs. 0.56), or Danish-Norwegian (0.73 vs. 0.67).

A similar approach to constructing a multilingual embedding space is discussed by Duong et al. (2017). However, their mapping approach is tailored for another scenario frequently

[6] The latest release of fastText vectors contains vectors for 204 languages. The vectors are available here: `https://github.com/facebookresearch/fastText`

[7] `https://github.com/Babylonpartners/fastText_multilingual`

[8] P@1 is a standard evaluation measure for bilingual lexicon induction that refers to the proportion of source test words for which the best translation is ranked as the most similar word in the target language.

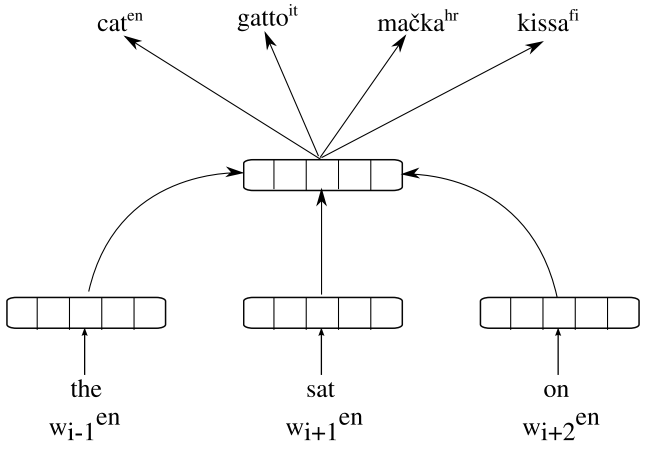


Figure 8: Illustration of the joint multilingual model of Duong et al. (2017) based on the modified CBOW objective; instead of predicting only the English word given the English context, the model also tries to predict its translations in all the remaining languages (i.e., in languages for which the translations exist in any of the input bilingual lexicons).

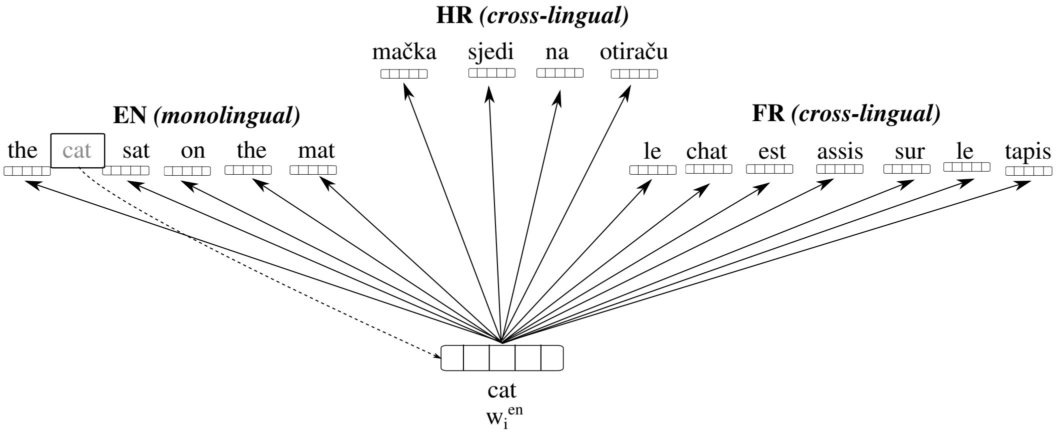


Figure 9: A multilingual extension of the sentence-level TransGram model of Coulmance et al. (2015). Since the model bypasses the word alignment step in its SGNS-style objective, for each center word (e.g., the EN word *cat* in this example) the model predicts *all* words in each sentence (from all other languages) which is aligned to the current sentence (e.g., the EN sentence *the cat sat on the mat*).

encountered in practice: one has to align bilingual embedding spaces where English is one of the two languages in each bilingual space. In other words, our starting embedding spaces are now not monolingual as in the previous mapping approach, but bilingual. The overview of the approach is given in Figure 7b. This approach first selects a pivot bilingual space (e.g., this is the EN-IT space in Figure 7b), and then learns a linear mapping/transformation from the English subspace of all other bilingual spaces into the pivot English subspace. The learned linear mapping is then applied to other subspaces (i.e., “foreign” subspaces such as FI, FR, NL, or RU in Figure 7b) to transform them into the shared multilingual space.

Pseudo-bilingual and joint approaches The two other sub-groups of word-level models also assume the existence of monolingual data plus multiple bilingual dictionaries covering translations of the same term in multiple languages. The main idea behind *pseudo-multilingual* approaches is to "corrupt" monolingual data available for each of the L languages so that words from all languages are present as context words for every center word in all monolingual corpora. A standard monolingual word embedding model (e.g., CBOW or SGNS) is then used to induce a shared multilingual space. First, for each word in each vocabulary one collects all translations of that word in all other languages. The sets of translations may be incomplete as they are dependent on their presence in input dictionaries. Following that, we use all monolingual corpora in all L languages and proceed as the original model of Gouws and Søgaard (2015): (i) for each word w for which there is a set of translations of size k_{mt}, we flip a coin and decide whether to retain the original word w or to substitute it with one of its k_{mt} translations; (ii) in case we have to perform a substitution, we choose one translation as a substitute with a probability of $\frac{1}{2k_{mt}}$. In the limit, this method yields "hybrid" pseudo-multilingual sentences with each word surrounded by words from different languages. Despite its obvious simplicity, the only work that generalizes pseudo-bilingual approaches to the multilingual setting that we are aware of is by Ammar et al. (2016) who replace all tokens in monolingual corpora with their corresponding translation cluster ID, thereby restricting them to have the same representation. Note again that we do not need lexicons for all language pairs in case one resource-rich language (e.g., English) is selected as the pivot language.

Joint multilingual models rely on exactly the same input data (i.e., monolingual data plus multiple bilingual dictionaries) and the core idea is again to exploit multilingual word contexts. An extension of the *joint* modeling paradigm to multilingual settings, illustrated in Figure 8, is discussed by Duong et al (2017). The core model is an extension of their bilingual model (Duong et al., 2016) based on the CBOW-style objective: in the multilingual scenario with L languages, for each language l^i the training procedure consists of predicting the center word in language l^i given the monolingual context in l^i plus predicting all translations of the center word in all other languages, subject to their presence in the input bilingual dictionaries. Note that effectively this MultiCBOW model may be seen as a combination of multiple monolingual and cross-lingual CBOW-style sub-objectives as follows:

$$J = \mathcal{L}^1_{\text{CBOW}} + \mathcal{L}^2_{\text{CBOW}} + \cdots + \mathcal{L}^L_{\text{CBOW}} + \mathcal{L}^{1\to 2}_{\text{CBOW}} + \mathcal{L}^{2\to 1}_{\text{CBOW}} + \cdots \mathcal{L}^{(L-1)\to L}_{\text{CBOW}} + \mathcal{L}^{L\to (L-1)}_{\text{CBOW}} \tag{49}$$

where the cross-lingual part of the objective again serves as the cross-lingual regularizer Ω. By replacing the CBOW-style objective with the SGNS objective, the model described by Equation (49) effectively gets transformed to the straightforward multilingual extension of the bilingual BiSkip model (Luong et al., 2015). Exactly this model, called MultiSkip, is described in the work of Ammar et al. (2016b). Instead of summing contexts around the center word as in CBOW, the model now tries to predict surrounding context words of the center word in its own language, plus its translations and all surrounding context words of its translations in all other languages. Translations are again obtained from input dictionaries or extracted from word alignments as in the original BiSkip and MultiSkip models. The pivot language idea is also applicable with the MultiSkip and MultiCBOW models.

9.2 Multilingual Word Embeddings from Sentence-Level and Document-Level Information

Extending bilingual embedding models which learn on the basis of aligned sentences and documents closely follows the principles already established for word-level models in Section 9.1. For instance, the multilingual extension of the TransGram model from Coulmance et al. (2015) may be seen as MultiSkip without word alignment information (see again Table 4). In other words, instead of predicting only words in the neighborhood of the word aligned to the center word, TransGram predicts all words in the sentences aligned to the sentence which contains the current center word (i.e., the model assumes uniform word alignment). This idea is illustrated by Figure 9. English is again used as the pivot language to reduce bilingual data requirements.

The same idea of pivoting, that is, learning multiple bilingual spaces linked through the shared pivot English space is directly applicable to other prominent bilingual word embedding models such as (Chandar et al., 2014), (Hermann & Blunsom, 2014), (Soyer et al., 2015), (Zou et al., 2013), (Gouws et al., 2015).

The document-level model of Vulić et al. (2016) may be extended to the multilingual setting using the same idea as in previously discussed word-level pseudo-multilingual approaches. Søgaard et al. (2015) and Levy et al. (2017) exploit the structure of a multi-comparable Wikipedia dataset and a multi-parallel Bible dataset respectively to directly build sparse cross-lingual representations using the same set of shared indices (i.e., the former uses the indices of aligned Wikipedia articles while the latter relies on the indices of aligned sentences in the multi-parallel corpus). Dense word embeddings are then obtained by factorizing the multilingual word-concept matrix containing all words from all vocabularies.

10. Evaluation

Given the wide array of cross-lingual models and their potential applications, there is an equally wide array of possible evaluation settings. Cross-lingual embeddings can be evaluated on a set of *intrinsic* and *extrinsic* tasks that directly measure the quality of the embeddings. In addition, as we detail in Section 11, cross-lingual embeddings can be employed to facilitate cross-lingual transfer in any of a wide array of downstream applications.

In the following, we discuss the most common intrinsic and extrinsic evaluation tasks that have been used to test cross-lingual embeddings and outline associated challenges. In addition, we also present benchmarks that can be used for evaluating cross-lingual embeddings and the most important findings of two recent empirical benchmark studies.

10.1 Intrinsic tasks

The first two widely used tasks are *intrinsic* evaluation tasks: They evaluate cross-lingual embeddings in a controlled *in vitro* setting that is geared towards revealing certain characteristics of the representations. The major downside with these tasks is that good performance on them does not generalize necessarily to good performance on downstream tasks (Tsvetkov, Sitaram, Faruqui, Lample, Littell, Mortensen, Black, Levin, & Dyer, 2016; Schnabel, Labutov, Mimno, & Joachims, 2015).

Word similarity This task evaluates how well the notion of word similarity according to humans is emulated in the vector space. Multi-lingual word similarity datasets are multilingual extensions of datasets that have been used for evaluating English word representations. Many of these originate from psychology research and consist of word pairs – ranging from synonyms (e.g., car - automobile) to unrelated terms (e.g., noon - string) – that have been annotated with a relatedness score by human subjects. The most commonly used ones of these human judgement datasets are: a) the RG dataset (Rubenstein & Goodenough, 1965); b) the MC dataset (Miller & Charles, 1991); c) the WordSim-353 dataset (Finkelstein, Gabrilovich, Matias, Rivlin, Solan, Wolfman, & Ruppin, 2002), a superset of MC; and d) the SimLex-999 dataset (Hill, Reichart, & Korhonen, 2015). Extending them to the multilingual setting then mainly involves translating the word pairs into different languages: WordSim-353 has been translated to Spanish, Romanian, and Arabic (Hassan & Mihalcea, 2009) and to German, Italian, and Russian (Leviant & Reichart, 2015); RG was translated to German (Gurevych, 2005), French, (Joubarne & Inkpen, 2011), Spanish and Farsi (Camacho-Collados, Pilehvar, & Navigli, 2015); and SimLex-999 was translated to German, Italian and Russian (Leviant & Reichart, 2015) and to Hebrew and Croatian (Mrkšić et al., 2017b). Other prominent datasets for word embedding evaluation such as MEN (Bruni, Tran, & Baroni, 2014), RareWords (Luong, Socher, & Manning, 2013), and SimVerb-3500 (Gerz, Vulić, Hill, Reichart, & Korhonen, 2016) have only been used in monolingual contexts.

The SemEval 2017 task on cross-lingual and multilingual word similarity (Camacho-Collados, Pilehvar, Collier, & Navigli, 2017) has introduced cross-lingual word similarity datasets between five languages: English, German, Italian, Spanish, and Farsi, yielding 10 new datasets in total. Each cross-lingual dataset is of reasonable size, containing between 888 and 978 word pairs.

Cross-lingual embeddings are evaluated on these datasets by first computing the cosine similarity of the representations of the cross-lingual word pairs. The Spearman's rank correlation coefficient (Myers, Well, & Lorch, 2010) is then computed between the cosine similarity score and the human judgement scores for every word pair. Cross-lingual word similarity datasets are affected by the same problems as their monolingual variants (Faruqui, Tsvetkov, Rastogi, & Dyer, 2016): the annotated notion of word similarity is subjective and is often confused with relatedness; the datasets evaluate semantic rather than task-specific similarity, which is arguably more useful; they do not have standardized splits; they correlate only weakly with performance on downstream tasks; past models do not use statistical significance; and they do not account for polysemy, which is even more important in the cross-lingual setting.

multiQVEC/multiQVEC+ multiQVEC+ is a multilingual extension of QVEC (Tsvetkov, Faruqui, Ling, Lample, & Dyer, 2015), a method that seeks to quantify the linguistic content of word embeddings by maximizing the correlation with a manually-annotated linguistic resource. A semantic linguistic matrix is first constructed from a semantic database. The word embedding matrix is then aligned with the linguistic matrix and the correlation is measured using cumulative dimension-wise correlation. Ammar et al. (2016) propose QVEC+, which computes correlation using CCA and extend QVEC to the multilingual setting (multiQVEC) by using supersense tag annotations in multiple languages to construct the linguistic matrix. While QVEC has been shown to correlate well with certain semantic downstream tasks, as an intrinsic evaluation task it can only approximately capture the performance as it relates to downstream tasks.

10.2 Extrinsic tasks

The two following tasks are not intrinsic in the sense that they measure performance on tasks that are potentially of real-world importance. The tasks, however, are constrained to evaluating cross-lingual word embeddings as they rely on nearest neighbor search in the cross-lingual embedding space to identify the most similar target word given a source word.

Word alignment prediction For word alignment prediction, each word in a given source language sentence is aligned with the most similar target language word from the target language sentence. If a source language word is out of vocabulary, it is not aligned with anything, whereas target language out-of-vocabulary words are given a default minimum similarity score, and never aligned to any candidate source language word in practice (Levy et al., 2017). The inverse of the alignment error rate (1-AER) (Koehn, 2009) is typically used to measure performance, where higher scores mean better alignments. Levy et al. (2017) use alignment data from Hansards[9] and from four other sources (Graça, Pardal, Coheur, & Caseiro, 2008; Lambert, De Gispert, Banchs, & Mariño, 2005; Mihalcea & Pedersen, 2003; Holmqvist & Ahrenberg, 2011).

Bilingual dictionary induction After the shared cross-lingual embedding space is induced, given a list of N source language words $x_{u,1}, \ldots, x_{u,N}$, the task is to find a target language word t for each *query word* x_u relying on the representations in the space. t_i is the target language word closest to the source language word $x_{u,i}$ in the induced cross-lingual space, also known as the *cross-lingual nearest neighbor*. The set of learned N $(x_{u,i}, t_i)$ pairs is then run against a gold standard dictionary.

Bilingual dictionary induction is appealing as an evaluation task, as high-quality, freely available, wide-coverage manually constructed dictionaries are still rare, especially for non-European languages. The task also provides initial intrinsic evidence on the quality of the shared space. Upadhyay et al. (2016) obtain evaluation sets for the task across 26 languages from the Open Multilingual WordNet (Bond & Foster, 2013), while Levy et al. (2017) obtain bilingual dictionaries from Wiktionary for Arabic, Finnish, Hebrew, Hungarian, and Turkish. More recently Wijaya, Callahan, Hewitt, Gao, Ling, Apidianaki, and Callison-Burch (2017) build evaluation data for 28 language pairs (where English is always the target language) by semi-automatically translating all Wikipedia words with frequency above 100. Most previous work (Vulić & Moens, 2013a; Gouws et al., 2015; Mikolov et al., 2013b) filters source and target words based on part-of-speech, though this simplifies the task and introduces bias in the evaluation. Each cross-lingual embedding model is then evaluated on its ability to select the closest target language word to a given source language word as the translation of choice and measured based on precision-at-one (P@1).

[9] `https://www.isi.edu/natural-language/download/hansard/`

11. Applications

Cross-lingual transfer Both word alignment prediction and bilingual dictionary induction rely on (constrained) nearest neighbor search in the cross-lingual word embedding graph based on computed similarity scores. However, cross-lingual word embeddings can also be used directly as features in NLP models. Such models are then defined for several languages, and can be used to facilitate *cross-lingual transfer*. In other words, the main idea is to train a model on data from one language and then to apply it to another relying on shared cross-lingual features. Extrinsic evaluation on such downstream tasks is often preferred, as it directly allows to evaluate the usefulness of the cross-lingual embedding model for the respective task. We briefly describe the cross-lingual tasks that people have used to evaluate cross-lingual embeddings:

- *Document classification* is the task of classifying documents with respect to topic, sentiment, relevance, etc. The task is commonly used following the setup of Klementiev et al. (2012): it uses the RCV2 Reuters multilingual corpus[10]. A document classifier is trained to predict topics on the document representations derived from word embeddings in the source language and then tested on the documents of the target language. Such representations typically do not take word order into account, and the standard embedding-based representation is to represent documents by the TF-IDF weighted average over the embeddings of the individual words, with an averaged perceptron model (or some other standard off-the-shelf classification model) acting as the document classifier. Word embeddings can also be used to seed more sophisticated classifiers based on convolutional or recurrent neural networks (Liu, Qiu, & Huang, 2016; Mandelbaum & Shalev, 2016; Zhang, Lee, & Radev, 2016a). Although it is clear that cross-lingual word embeddings are instrumental to cross-lingual document classification, the task might be considered suboptimal for a full-fledged extrinsic evaluation of embeddings. It only evaluates topical associations and provides a signal for sets of co-occurring words, not for the individual words.

- *Dependency parsing* is the task that constructs the grammatical structure of a sentence, establishing typed relationships between "head" words and words which modify those heads. In a cross-lingual setting Täckström et al. (2012) proposed a parser transfer model that employed cross-lingual similarity measures based on cross-lingual Brown clusters. When relying on cross-lingual word embeddings, similar to cross-lingual document classification, a dependency parsing model is trained using the embeddings for a source language and is then evaluated on a target language. In the setup of Guo et al. (2015), a transition-based dependency parser with a non-linear activation function is trained on Universal Dependencies data (McDonald, Nivre, Quirmbach-Brundage, Goldberg, Das, Ganchev, Hall, Petrov, Zhang, Täckström, et al., 2013), with the source-side embeddings as lexical features[11].

- *POS tagging*, the task of assigning parts-of-speech to words, is usually evaluated using the Universal Dependencies treebanks (Nivre et al., 2016a) as these are annotated with

[10] http://trec.nist.gov/data/reuters/reuters.html

[11] https://github.com/jiangfeng1124/acl15-clnndep

the same universal tag set. Zhang et al. (2016) furthermore map proper nouns to nouns and symbol makers (e.g. "-", "/") and interjections to an X tag as it is hard and unnecessary to disambiguate them in a low-resource setting. Fang and Cohn (2017) use data from the CoNLL-X datasets of European languages (Buchholz & Marsi, 2006), from CoNLL 2003[12] and from Das and Petrov (2011), the latter of which is also used by Gouws and Søgaard (2015).

- *Named entity recognition (NER)* is the task of tagging entities with their appropriate type in a text. Zou et al. (2013) perform NER experiments for English and Chinese on OntoNotes (Hovy et al., 2006), while Murthy, Khapra, and Bhattacharyya (2016) use English data from CoNLL 2003 (Tjong Kim Sang & De Meulder, 2003) and Spanish and Dutch data from CoNLL 2002 (Tjong Kim Sang, 2002).

- *Super-sense tagging* is the task that involves annotating each significant entity in a text (e.g., nouns, verbs, adjectives and adverbs) within a general semantic taxonomy defined by the WordNet lexicographer classes (called super-senses). The cross-lingual variant of the task is used by Gouws and Søgaard (2015) for evaluating their embeddings. They use the English data from SemCor[13] and publicly available Danish data[14].

- *Semantic parsing* is the task of automatically identifying semantically salient targets in the text. Frame-semantic parsing, in particular, disambiguates the targets by assigning a sense (frame) to them, identifies their arguments, and labels the arguments with appropriate roles. Johannsen, Alonso, and Søgaard (2015) create a frame-semantic parsing corpus that covers five topics, two domains (Wikipedia and Twitter), and nine languages and use it to evaluate cross-lingual word embeddings.

- *Discourse parsing* is the task of segmenting text into elementary discourse units (mostly clauses), which are then recursively connected via discourse relations to form complex discourse units. The segmentation is usually done according to Rhetorical Structure Theory (RST) (Mann & Thompson, 1988). Braud, Lacroix, and Søgaard (2017) and Braud, Coavoux, and Søgaard (2017) perform experiments using a diverse range of RST discourse treebanks for English, Portuguese, Spanish, German, Dutch, and Basque.

- *Dialog state tracking (DST)* is the component in task-oriented dialogue statistical systems that keeps track of the belief state, that is, the system's internal distribution over the possible states of the dialogue. A recent state-of-the-art DST model of Mrkšić, Ó Séaghdha, Wen, Thomson, and Young (2017) is based exclusively on word embeddings fed into the model as its input. This property of the model enables a straightforward adaptation to cross-lingual settings by simply replacing input monolingual word embeddings with cross-lingual embeddings. Still an under-explored task, we believe that DST serves as a useful proxy task which shows the capability of induced word embeddings to support more complex language understanding tasks. Mrkšić et al. (2017) use DST for evaluating cross-lingual embeddings on the Multilingual WOZ 2.0 dataset (Wen, Vandyke, Mrkšić, Gašić, Rojas-Barahona, Su, Ultes, & Young, 2017)

[12] http://www.cnts.ua.ac.be/conll2003/ner/

[13] http://web.eecs.umich.edu/Ëœmihalcea/downloads.html#semcor

[14] https://github.com/coastalcph/noda2015_sst

available in English, German, and Italian. Their results suggest that cross-lingual word embeddings boost the construction of dialog state trackers in German and Italian even without any German and Italian training data, as the model is able to also exploit English training data through the embedding space. Further, a multilingual DST model which uses training data from all three languages combined with a multilingual embedding space improves tracking performance in all three languages.

- *Entity linking or wikification* is another task tackled using cross-lingual word embeddings (Tsai & Roth, 2016). The purpose of the task is to ground mentions written in non-English documents to entries in the English Wikipedia, facilitating the exploration and understanding of foreign texts without full-fledged translation systems (Ji, Nothman, Hachey, & Florian, 2015). Such wikifiers, i.e., entity linkers are a valuable component of several NLP and IR tasks across different domains (Mihalcea & Csomai, 2007; Cheng & Roth, 2013).

- *Sentiment analysis* is the task of determining the sentiment polarity (e.g. positive and negative) of a text. Mogadala and Rettinger (2016) evaluate their embeddings on the multilingual Amazon product review dataset of Prettenhofer and Stein (2010).

- *Machine translation* is used to translate entire texts in other languages. This is in contrast to bilingual dictionary induction, which focuses on the translation of individual words. Zou et al. (2013) used phrase-based machine translation to evaluate their embeddings. Cross-lingual embeddings are incorporated in the phrase-based MT system by adding them as a feature to bilingual phrase-pairs. For each phrase, its word embeddings are averaged to obtain a feature vector.

Information retrieval Word embeddings in general and cross-lingual word embeddings in specific have naturally found application beyond core NLP applications. They also offer support to Information Retrieval tasks (IR) (Zamani & Croft, 2016; Mitra & Craswell, 2017, inter alia) serving as useful features which can link semantics of the query to semantics of the target document collection, even when query terms are not explicitly mentioned in the relevant documents (e.g., the query can talk about *cars* while a relevant document may contain a near-synonym *automobile*). A shared cross-lingual embedding space provides means to more general cross-lingual and multilingual IR models without any substantial change in the algorithmic design of the retrieval process (Vulić & Moens, 2015). Semantic similarity between query and document representations, obtained through the composition process as in the document classification task, is computed in the shared space, irrespective of their actual languages: the similarity score may be used as a measure of document relevance to the information need formulated in the issued query.

Multi-modal and cognitive approaches to evaluation Evaluation of monolingual word embeddings is a controversial topic. Monolingual word embeddings are useful downstream (Turian, Ratinov, & Bengio, 2010), but in order to argue that one set of embeddings is better than another, we would like a robust evaluation metric. Metrics have been proposed based on co-occurrences (perplexity or word error rate), based on ability to discriminate between contexts (e.g., topic classification), and based on lexical semantics (predicting links in lexical knowledge bases). Søgaard (2016) argues that such metrics are not valid, because co-occurrences, contexts, and lexical knowledge bases are also used to induce word embeddings, and that downstream evaluation is the best way to evaluate word embeddings. The only task-independent evaluation of embeddings that is reasonable, he claims, is to evaluate word embeddings by how well they predict behavioral observations, e.g. gaze or fMRI data.

For cross-lingual word embeddings, it is easier to come up with valid metrics, e.g., P@k (?, inline) in word alignment and bilingual dictionary induction. Note that these metrics only evaluate cross-lingual neighbors, not whether monolingual distances between words reflect synonymy relations. In other words, a random pairing of translation equivalents in vector space would score perfect precision in bilingual dictionary induction tasks. In addition, if we intend to evaluate the ability of cross-lingual word embeddings to allow for generalizations *within* languages, we inherit the problem of finding valid metrics from monolingual word representation learning.

11.1 Benchmarks

Benchmarks In light of the plethora of both intrinsic and extrinsic evaluation tasks and datasets, a rigorous evaluation of cross-lingual embeddings across many benchmark datasets can often be cumbersome and practically infeasible. To the best of our knowledge, there are two resources available, which facilitate comparison of cross-lingual embedding models: Faruqui and Dyer propose `wordvectors.org`[15], a website for evaluating word representations, which allows the upload and evaluation of learned word embeddings. The website, however, focuses mainly on evaluating monolingual word representations and only evaluates them on word similarity datasets.

The second resource is by Ammar et al. (2016) who make a website[16] available where monolingual and cross-lingual word representations can be uploaded and automatically evaluated on some of the tasks we discussed. In particular, their evaluation suite includes word similarity, multiQVEC, bilingual dictionary induction, document classification, and dependency parsing. As a good practice in general, we recommend to evaluate cross-lingual word embeddings on an intrinsic task that is cheap to compute and on at least one downstream NLP task besides document classification.

Benchmark studies To conclude this section, we summarize the findings of two recent benchmark studies of cross-lingual embeddings: Upadhyay et al. (2016) evaluate cross-lingual embedding models that require different forms of supervision on various tasks. They find that on word similarity datasets, models with cheaper supervision (sentence-aligned and document-aligned data) are almost as good as models with more expensive supervision in the form of word alignments. For cross-lingual classification and bilingual dictionary induction, more informative supervision is more beneficial: word-alignment and sentence-level models score better. Finally, for dependency parsing, models with word-level alignment are able to capture syntax more accurately and thus perform better overall. The findings by Upadhyay et al. strengthen our hypothesis that the choice of the data is more important than the algorithm learning from the same data source.

Levy et al. (2017) evaluate cross-lingual word embedding models on bilingual dictionary induction and word alignment. In a similar vein as our typology that is based on the

[15] `http://wordvectors.org/`

[16] `http://128.2.220.95/multilingual`

type and level of alignment, they argue that whether or not an algorithm uses a particular feature set is more important than the choice of the algorithm. In their experiments, they achieve the best results using sentence IDs as features to represent words, which outperforms using word-level source and target co-occurrence information. These findings lend further evidence and credibility to our typology that is based on the data requirements of cross-lingual embedding models. Models that learn from word-level and sentence-level information typically outperform other approaches, especially for finer-grained tasks such as bilingual dictionary induction. These studies furthermore raise awareness that we should not only focus on developing better cross-lingual embedding models, but also work on unlocking new data sources and new ways to leverage comparable data, particularly for languages and domains with only limited amounts of parallel training data.

12. General Challenges and Future Directions

Subword-level information In morphologically rich languages, words can have complex internal structures, and some word forms can be rare. For such languages, it makes sense to compose representations from representations of lemmas and morphemes. Neural network models increasingly leverage subword-level information (Sennrich, Haddow, & Birch, 2015; Lample, Ballesteros, Subramanian, Kawakami, & Dyer, 2016) and character-based input has been found useful for sharing knowledge in multilingual scenarios (Gillick, Brunk, Vinyals, & Subramanya, 2016; Ling, Trancoso, Dyer, & Black, 2016). Subword-level information has also been used for learning word representations (Ling, Luis, Marujo, Astudillo, Amir, Dyer, Black, & Trancoso, 2015; Bhatia, Guthrie, & Eisenstein, 2016) but has so far not been incorporated in learning cross-lingual word representations.

Multi-word expressions Just like words can be too coarse units for representation learning in morphologically rich languages, words also combine in non-compositional ways to form multi-word expressions such as *ad hoc* or *kick the bucket*, the meaning of which cannot be derived from standard representations of their constituents. Dealing with multi-word expressions remains a challenge for monolingual applications and has only received scarce attention in the cross-lingual setting.

Function words Models for learning cross-linguistic representations share weaknesses with other vector space models of language: While they are very good at modeling the conceptual aspect of meaning evaluated in word similarity tasks, they fail to properly model the functional aspect of meaning, e.g. to distinguish whether one remarks “Give me *a* pencil" or “Give me *that* pencil". Modeling the functional aspect of language is of particular importance in scenarios such as dialogue, where the pragmatics of language must be taken into account.

Polysemy While conflating multiple senses of a word is already problematic for learning monolingual word representations, this issue is amplified in a cross-lingual embedding space: If polysemy leads to m bad word embeddings in the source language, and n bad word embeddings in the target language, we can derive $\mathcal{O}(n \times m)$ false nearest neighbors from our cross-lingual embeddings. While recent work on learning cross-lingual multi-sense embeddings (Li & Jurafsky, 2015) is extremely interesting, it is still an open question whether modern NLP models can infer from context, what they need in order to resolve lexical ambiguities.

Embeddings for specialized domains There are many domains, for which cross-lingual applications would be particularly useful, such as bioinformatics or social media. However, parallel data is scarce in many such domains as well as for low-resource languages. Creating robust cross-lingual word representations with as few parallel examples as possible is thus an important research avenue. An important related direction is to leverage comparable corpora, which are often more plentiful and incorporate other signals, such as from multi-modal contexts.

For many domains or tasks, we also might want to have not only word embeddings, but be able to compose those representations into accurate sentence and document representations. Besides existing methods that sum word embeddings, not much work has been doing on learning better higher-level cross-lingual representations.

Feasibility Learning a general shared vector space for words that reliably captures inter-language and intra-language relations may seem slightly optimistic. Languages are very different, and it is not clear if there is even a definition of *words* that make words commensurable across languages. Note that while this is related to whether it is possible to translate between the world's languages in the first place, the possibility of translation (at document level) does not necessarily entail that it is possible to device embeddings such that translation equivalents in two languages end up as nearest neighbors.

There is also the question of what is the computational complexity of finding an embedding that obeys all our inter-lingual and intra-lingual constraints, say, for example, translation equivalents and synonymy. Currently, many approaches to cross-lingual word embeddings, as shown in this survey, minimize a loss that penalizes models for violating such constraints, but there is no guarantee that the final model satisfies all constraints.

Checking whether all such constraints are satisfied in a given model is trivially done in time linear in the number of constraints, but finding out whether such a model exists is much harder. While the problem's decidability follows from the decidability of two-variable first order logic with equivalence/symmetry closure, determining whether such a graph exists is in fact NP-hard (Eades & Whitesides, 1995).

Non-linear mapping Mapping-based approaches assume that a linear transformation can project the embedding space of one language into the space of a target language. While Mikolov et al. and Conneau, Lample, Ranzato, Denoyer, and Jégou both find that a linear transformation outperforms non-linear transformation learned via a feedforward neural network, assuming a linear transformation between two languages is overly simplistic and ignores important language-specific differences. Nakashole and Flauger (2018) lend further credibility to this intuition by learning neighbourhood-specific linear transformations and showing that these vary across the monolingual word embedding space. However, to the best of our knowledge, there has not been any model yet that leveraged this intuition to construct a more effective mapping model.

Robust unsupervised approaches Recently, word-level mapping-based approaches have become the preferred choice for learning cross-lingual embeddings due to their ease of use and reliance on inexpensive forms of supervision. At the same time, methods for learning with less supervision have been developed: These range from approaches using small seed lexicons (Zhang, Liu, Luan, Liu, & Sun, 2016; Artetxe et al., 2017) to completely unsupervised approaches that seek to match source and target distributions based on adversarial learning (Zhang, Liu, Luan, & Sun, 2017a, 2017b; Conneau et al., 2018) and offer support to neural machine translation and cross-lingual information retrieval from monolingual data only (Lample, Denoyer, & Ranzato, 2018; Artetxe, Labaka, Agirre, & Cho, 2018; Litschko, Glavaš, Ponzetto, & Vulić, 2018). Such unsupervised methods, however, rely on the assumption that monolingual word embedding spaces are approximately isomorphic, which has been shown not to hold in general and for distant language pairs in particular (Søgaard, Ruder, & Vulić, 2018), for which such methods are desired in the first place. In simple terms, although thought-provoking and attractive in theory, such unsupervised methods thus fail when languages are distant. In such cases, using a distantly supervised seed lexicon of identical strings in both languages is preferable (Søgaard et al., 2018).

13. Conclusion

This survey has focused on providing an overview of cross-lingual word embedding models. It has introduced standardized notation and a typology that demonstrated the similarity of many of these models. It provided proofs that connect different word-level embedding models and has described ways to evaluate cross-lingual word embeddings as well as how to extend them to the multilingual setting. It finally outlined challenges and future directions.

Acknowledgements

We thank the anonymous reviewers for their valuable and comprehensive feedback. Sebastian is supported by Irish Research Council Grant Number EBPPG/2014/30 and Science Foundation Ireland Grant Number SFI/12/RC/2289. Ivan's work is supported by the ERC Consolidator Grant LEXICAL: Lexical Acquisition Across Languages (no 648909).

References

Adams, O., Makarucha, A., Neubig, G., Bird, S., & Cohn, T. (2017). Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of EACL*, pp. 937–947.

Agić, Z., Hovy, D., & Søgaard, A. (2015). If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. In *Proceedings of ACL*, pp. 268–272.

Agić, Z., Johannsen, A., Plank, B., Alonso, H. M., Schluter, N., & Søgaard, A. (2016). Multilingual projection for parsing truly low-resource languages. *Transactions of the ACL*, *4*, 301–312.

Ammar, W., Mulcaire, G., Ballesteros, M., Dyer, C., & Smith, N. A. (2016a). Many languages, one parser. *Transactions of the ACL*, *4*, 431–444.

Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C., & Smith, N. A. (2016b). Massively Multilingual Word Embeddings. *CoRR, abs/1602.01925*.

Aone, G., & McKee, D. (1993). A language-independent anaphora resolution system for understanding multilingual texts. In *Proceedings of ACL*, pp. 156–163.

Artetxe, M., Labaka, G., & Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of EMNLP*, pp. 2289–2294.

Artetxe, M., Labaka, G., & Agirre, E. (2017). Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of ACL*, pp. 451–462.

Artetxe, M., Labaka, G., Agirre, E., & Cho, K. (2018). Unsupervised neural machine translation. In *Proceedings of ICLR*.

Bergsma, S., & Van Durme, B. (2011). Learning bilingual lexicons using the visual similarity of labeled Web images. In *Proceedings of IJCAI*, pp. 1764–1769.

Bhatia, P., Guthrie, R., & Eisenstein, J. (2016). Morphological Priors for Probabilistic Neural Word Embeddings. *EMNLP*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the ACL*, *5*, 135–146.

Bond, F., & Foster, R. (2013). Linking and extending an Open Multilingual WordNet. In *Proceedings ACL*, pp. 1352–1362.

Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). *Applications of Topic Models*, Vol. 11 of *Foundations and Trends in Information Retrieval*.

Boyd-Graber, J., & Resnik, P. (2010). Holistic sentiment analysis across languages: Multilingual supervised Latent Dirichlet Allocation. In *Proceedings of EMNLP*, pp. 45–55.

Boyd-Graber, J. L., & Blei, D. M. (2009). Multilingual topic models for unaligned text. In *Proceedings of UAI*, pp. 75–82.

Braud, C., Coavoux, M., & Søgaard, A. (2017a). Cross-lingual RST discourse parsing. In *Proceedings EACL*, pp. 292–304.

Braud, C., Lacroix, O., & Søgaard, A. (2017b). Cross-lingual and cross-domain discourse segmentation of entire documents. In *Proceedings of ACL*, pp. 237–243.

Brown, P. F., Pietra, S. D., Pietra, V. J. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, *19*(2), 263–311.

Bruni, E., Tran, N.-K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, *49*(2014), 1–47.

Buchholz, S., & Marsi, E. (2006). CoNLL-X Shared task on multilingual dependency parsing. In *Proceedings of CoNLL*, pp. 149–164.

Calixto, I., Liu, Q., & Campbell, N. (2017). Multilingual Multi-modal Embeddings for Natural Language Processing. *CoRR*, *abs/1702.01101*.

Camacho-Collados, J., Pilehvar, M. T., Collier, N., & Navigli, R. (2017). SemEval-2017 Task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of SEMEVAL*, pp. 15–26.

Camacho-Collados, J., Pilehvar, M. T., & Navigli, R. (2015). A framework for the construction of monolingual and cross-lingual word similarity datasets. In *Proceedings of ACL*, pp. 1–7.

Cavallanti, G., Cesa-Bianchi, N., & Gentile, C. (2010). Linear algorithms for online multitask classification. *Journal of Machine Learning Research*, *11*, 2901–2934.

Chandar, S., Lauly, S., Larochelle, H., Khapra, M. M., Ravindran, B., Raykar, V., & Saha, A. (2014). An autoencoder approach to learning bilingual word representations. In *Proceedings of NIPS*, pp. 1853–1861.

Cheng, X., & Roth, D. (2013). Relational inference for wikification. In *Proceedings of EMNLP*, pp. 1787–1796.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Cohen, S., Das, D., & Smith, N. (2011). Unsupervised structure prediction with non-parallel multilingual guidance. In *EMNLP*.

Collobert, R., & Weston, J. (2008). A unified architecture for Natural Language Processing. In *Proceedings of ICML*, pp. 160–167.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2018). Word Translation Without Parallel Data. In *Proceedings of ICLR 2018*.

Coulmance, J., Marty, J.-M., Wenzek, G., & Benhalloum, A. (2015). Trans-gram, fast cross-lingual word-embeddings. In *Proceedings of EMNLP*, pp. 1109–1113.

Das, D., & Petrov, S. (2011). Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of ACL*, pp. 600–609.

De Smet, W., Tang, J., & Moens, M. (2011). Knowledge transfer across multilingual corpora via latent topics. In *Proceedings of PAKDD*, pp. 549–560.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Dehouck, M., & Denis, P. (2017). Delexicalized word embeddings for cross-lingual dependency parsing. In *Proceedings of EACL*, pp. 241–250.

Dinu, G., Lazaridou, A., & Baroni, M. (2015). Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of ICLR (Workshop Track)*.

Duong, L., Cohn, T., Bird, S., & Cook, P. (2015). Cross-lingual transfer for unsupervised dependency parsing without parallel data. In *Proceedings of CoNLL*, pp. 113–122.

Duong, L., Kanayama, H., Ma, T., Bird, S., & Cohn, T. (2016). Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of EMNLP*, pp. 1285–1295.

Duong, L., Kanayama, H., Ma, T., Bird, S., & Cohn, T. (2017). Multilingual training of crosslingual word embeddings. In *Proceedings of EACL*, pp. 894–904.

Durrett, G., Pauls, A., & Klein, D. (2012). Syntactic transfer using a bilingual lexicon. In *Proceedings of EMNLP*, pp. 1–11.

Dyer, C., Chahuneau, V., & Smith, N. (2013a). A simple, fast, and effective parameterization of IBM Model 2. In *Proceedings of NAACL-HLT*, pp. 644–649.

Dyer, C., Chahuneau, V., & Smith, N. A. (2013b). A simple, fast, and effective reparameterization of ibm model 2.. Association for Computational Linguistics.

Eades, P., & Whitesides, S. (1995). Nearest neighbour graph realizability is np-hard. In *Latin American Symposium on Theoretical Informatics*, pp. 245–256. Springer.

Elliott, D., Frank, S., Sima'an, K., & Specia, L. (2016). Multi30k: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pp. 70–74.

Elliott, D., & Kádár, Á. (2017). Imagination improves multimodal translation. *CoRR*, *abs/1705.04350*.

Fang, M., & Cohn, T. (2017). Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proceedings of ACL*, pp. 587–593.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., & Smith, N. A. (2015). Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL-HLT*, pp. 1606–1615.

Faruqui, M., & Dyer, C. (2013). An information theoretic approach to bilingual word clustering. In *Proceedings of ACL*, pp. 777–783.

Faruqui, M., & Dyer, C. (2014a). Community evaluation and exchange of word vectors at wordvectors.org. In *Proceedings of ACL: System Demonstrations*, pp. 19–24.

Faruqui, M., & Dyer, C. (2014b). Improving vector space word representations using multilingual correlation. In *Proceedings of EACL*, pp. 462–471.

Faruqui, M., & Kumar, S. (2015). Multilingual open relation extraction using cross-lingual projection. In *Proceedings of NAACL-HLT*, pp. 1351–1356.

Faruqui, M., Tsvetkov, Y., Rastogi, P., & Dyer, C. (2016). Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of REPEVAL*, pp. 30–35.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, *20*(1), 116–131.

Firat, O., Cho, K., & Bengio, Y. (2016). Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of NAACL-HLT*, pp. 866–875.

Fukumasu, K., Eguchi, K., & Xing, E. P. (2012). Symmetric correspondence topic models for multilingual text analysis. In *Proceedings of NIPS*, pp. 1295–1303.

Gabrilovich, E., & Markovitch, S. (2006). Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of AAAI*, pp. 1301–1306.

Gardner, M., Huang, K., Paplexakis, E., Fu, X., Talukdar, P., Faloutsos, C., Sidiropoulos, N., Mitchell, T., & Sidiropoulos, N. (2015). Translation invariant word embeddings. In *Proceedings of EMNLP*, pp. 1084–1088.

Gaussier, É., Renders, J.-M., Matveeva, I., Goutte, C., & Déjean, H. (2004). A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of ACL*, pp. 526–533.

Gella, S., Sennrich, R., Keller, F., & Lapata, M. (2017). Image pivoting for learning multilingual multimodal representations. In *Proceedings of EMNLP*, pp. 2829–2835.

Gerz, D., Vulić, I., Hill, F., Reichart, R., & Korhonen, A. (2016). SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of EMNLP*, pp. 2173–2182.

Gillick, D., Brunk, C., Vinyals, O., & Subramanya, A. (2016). Multilingual Language Processing From Bytes. *NAACL*, 1296–1306.

Gouws, S., Bengio, Y., & Corrado, G. (2015). BilBOWA: Fast bilingual distributed representations without word alignments. In *Proceedings of ICML*, pp. 748–756.

Gouws, S., & Søgaard, A. (2015). Simple task-specific bilingual word embeddings. In *Proceedings of NAACL-HLT*, pp. 1302–1306.

Graça, J., Pardal, J. P., Coheur, L., & Caseiro, D. (2008). Building a golden collection of parallel multi-language word alignment. In *Proceedings of LREC*, pp. 986–993.

Guo, J., Che, W., Yarowsky, D., Wang, H., & Liu, T. (2015). Cross-lingual dependency parsing based on distributed representations. In *Proceedings of ACL*, pp. 1234–1244.

Guo, J., Che, W., Yarowsky, D., Wang, H., & Liu, T. (2016). A representation learning framework for multi-source transfer parsing. In *Proceedings of AAAI*, pp. 2734–2740.

Gurevych, I. (2005). Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of IJCNLP*, pp. 767–778.

Gutmann, M. U., & Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, *13*(1), 307–361.

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., & Klein, D. (2008). Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL*, pp. 771–779.

Hassan, S., & Mihalcea, R. (2009). Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of EMNLP*, pp. 1192–1201.

Hauer, B., Nicolai, G., & Kondrak, G. (2017). Bootstrapping unsupervised bilingual lexicon induction. In *Proceedings of EACL*, pp. 619–624.

Henderson, M., Thomson, B., & Young, S. J. (2014). Robust dialog state tracking using delexicalised recurrent neural networks and unsupervised adaptation. In *Proceedings of IEEE SLT*, pp. 360–365.

Hermann, K. M., & Blunsom, P. (2013). Multilingual distributed representations without word alignment. In *Proceedings of ICLR (Conference Track)*.

Hermann, K. M., & Blunsom, P. (2014). Multilingual models for compositional distributed semantics. In *Proceedings of ACL*, pp. 58–68.

Heyman, G., Vulić, I., & Moens, M. (2016). C-BiLDA: Extracting cross-lingual topics from non-parallel texts by distinguishing shared from unshared content. *Data Mining and Knowledge Discovery*, *30*(5), 1299–1323.

Heyman, G., Vulić, I., & Moens, M.-F. (2017). Bilingual lexicon induction by learning to combine word-level and character-level representations. In *Proceedings of EACL*, pp. 1085–1095.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695.

Holmqvist, M., & Ahrenberg, L. (2011). A gold standard for English–Swedish word alignment. In *Proceedings of NODALIDA*, pp. 106–13.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2006). OntoNotes: The 90% solution. In *Proceedings of NAACL-HLT*, pp. 57–60.

Inan, H., Khosravi, K., & Socher, R. (2016). Tying Word Vectors and Word Classifiers: A Loss Framework for Language Modeling. *arXiv preprint arXiv:1611.01462*.

Jagarlamudi, J., & Daumé III, H. (2010). Extracting multilingual topics from unaligned comparable corpora. In *Proceedings of ECIR*, pp. 444–456.

Ji, H., Nothman, J., Hachey, B., & Florian, R. (2015). Overview of the TAC-KBP2015 entity discovery and linking tasks. In *Proceedings of Text Analysis Conference*.

Johannsen, A., Alonso, H. M., & Søgaard, A. (2015). Any-language frame-semantic parsing. In *Proceedings of EMNLP*, pp. 2062–2066.

Joubarne, C., & Inkpen, D. (2011). Comparison of semantic similarity for different languages using the Google N-gram corpus and second-order co-occurrence measures. In *Proceedings of the Canadian Conference on Artificial Intelligence*, pp. 216–221.

Kiela, D., & Clark, S. (2015). Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of EMNLP*, pp. 2461–2470.

Kiela, D., Vulić, I., & Clark, S. (2015). Visual bilingual lexicon induction with transferred ConvNet features. In *Proceedings of EMNLP*, pp. 148–158.

Klementiev, A., Titov, I., & Bhattarai, B. (2012). Inducing crosslingual distributed representations of words. In *Proceedings of COLING*, pp. 1459–1474.

Kočiský, T., Hermann, K. M., & Blunsom, P. (2014). Learning bilingual word representations by marginalizing alignments. In *Proceedings of ACL*, pp. 224–229.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*, pp. 79–86.

Koehn, P. (2009). *Statistical Machine Translation.* Cambridge University Press.

Kozhevnikov, M., & Titov, I. (2013). Cross-lingual transfer of semantic role labeling models. In *Proceedings of ACL*, pp. 1190–1200.

Lambert, P., De Gispert, A., Banchs, R., & Mariño, J. B. (2005). Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation*, *39*(4), 267–285.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Lample, G., Denoyer, L., & Ranzato, M. (2018). Unsupervised machine translation using monolingual corpora only. In *Proceedings of ICLR*.

Landauer, T. K., & Dumais, S. T. (1997). Solutions to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Laroche, A., & Langlais, P. (2010). Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of COLING*, pp. 617–625.

Lauly, S., Boulanger, A., & Larochelle, H. (2013). Learning multilingual word representations using a bag-of-words autoencoder. In *Proceedings of the NIPS Workshop on Deep Learning*, pp. 1–8.

Lazaridou, A., Dinu, G., & Baroni, M. (2015). Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of ACL*, pp. 270–280.

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of ICML*, pp. 1188–1196.

Leviant, I., & Reichart, R. (2015). Separated by an un-common language: Towards judgment language informed vector space modeling. *CoRR, abs/1508.00106*.

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Proceedings of NIPS*, pp. 2177–2185.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the ACL*, *3*, 211–225.

Levy, O., Søgaard, A., & Goldberg, Y. (2017). A strong baseline for learning cross-lingual word embeddings from sentence alignments. In *Proceedings of EACL*, pp. 765–774.

Li, J., & Jurafsky, D. (2015). Do multi-sense embeddings improve natural language understanding?. *arXiv preprint arXiv:1506.01070*.

Ling, W., Luis, T., Marujo, L., Astudillo, R. F., Amir, S., Dyer, C., Black, A. W., & Trancoso, I. (2015). Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation. In *Proceedings of EMNLP 2015*, pp. 1520–1530.

Ling, W., Trancoso, I., Dyer, C., & Black, A. (2016). Character-based Neural Machine Translation. In *ICLR*, pp. 1–11.

Litschko, R., Glavaš, G., Ponzetto, S. P., & Vulić, I. (2018). Unsupervised cross-lingual information retrieval using monolingual data only. In *Proceedings of SIGIR*.

Littman, M., Dumais, S. T., & Landauer, T. K. (1998). Automatic cross-language information retrieval using Latent Semantic Indexing. In *Chapter 5 of Cross-Language Information Retrieval*, pp. 51–62. Kluwer Academic Publishers.

Liu, P., Qiu, X., & Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. In *Proceedings of IJCAI*, pp. 2873–2879.

Lu, A., Wang, W., Bansal, M., Gimpel, K., & Livescu, K. (2015). Deep multilingual correlation for improved word embeddings. In *Proceedings of NAACL-HLT*, pp. 250–256.

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Bilingual word representations with monolingual quality in mind. In *Proceedings of the Workshop on Vector Modeling for NLP*, pp. 151–159.

Luong, T., Socher, R., & Manning, C. D. (2013). Better word representations with recursive neural networks for morphology. In *Proceedings of CoNLL*, pp. 104–113.

Mandelbaum, A., & Shalev, A. (2016). Word embeddings and their use in sentence classification tasks. *CoRR*, *abs/1610.08229*.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, *8*(3), 243–281.

McDonald, R., Petrov, S., & Hall, K. (2011). Multi-source transfer of delexicalized dependency parsers. In *Proceedings of EMNLP*, pp. 62–72.

McDonald, R. T., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K. B., Petrov, S., Zhang, H., Täckström, O., et al. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of ACL*, pp. 92–97.

Mihalcea, R., & Csomai, A. (2007). Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of CIKM*, pp. 233–242.

Mihalcea, R., & Pedersen, T. (2003). An evaluation exercise for word alignment. In *Proceedings of the Workshop on Building and Using Parallel Texts*, pp. 1–10.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of NIPS*, pp. 3111–3119.

Mikolov, T., Le, Q. V., & Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. *CoRR*, *abs/1309.4168*.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*(1), 1–28.

Mimno, D., Wallach, H., Naradowsky, J., Smith, D. A., & McCallum, A. (2009). Polylingual topic models. In *Proceedings of EMNLP*, pp. 880–889.

Mitra, B., & Craswell, N. (2017). Neural models for Information Retrieval. *CoRR*, *abs/1705.01509*.

Mnih, A., & Teh, Y. W. (2012). A fast and simple algorithm for training neural probabilistic language models. *Proceedings of ICML*, 1751–1758.

Mogadala, A., & Rettinger, A. (2016). Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of NAACL-HLT*, pp. 692–702.

Mrkšić, N., Ó Séaghdha, D., Wen, T.-H., Thomson, B., & Young, S. (2017a). Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of ACL*, pp. 1777–1788.

Mrkšić, N., Vulić, I., Ó Séaghdha, D., Leviant, I., Reichart, R., Gašić, M., Korhonen, A., & Young, S. (2017b). Semantic specialisation of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the ACL*, *5*, 309–324.

Munteanu, D. S., & Marcu, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of ACL*, pp. 81–88.

Murthy, R., Khapra, M., & Bhattacharyya, P. (2016). Sharing network parameters for crosslingual named entity recognition. *CoRR*, *abs/1607.00198*.

Myers, J. L., Well, A., & Lorch, R. F. (2010). *Research Design and Statistical Analysis*. Routledge.

Nakashole, N., & Flauger, R. (2018). Characterizing Departures from Linearity in Word Translation. In *Proceedings of ACL 2018*.

Naseem, T., Snyder, B., Eisenstein, J., & Barzilay, R. (2009). Multilingual part-of-speech tagging: Two unsupervised approaches. *Journal of Artificial Intelligence Research*, *36*, 341–385.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R. T., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016a). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of LREC*.

Nivre et al., J. (2016b). Universal Dependencies 1.4.. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19–51.

Peirsman, Y., & Padó, S. (2010). Cross-lingual induction of selectional preferences with bilingual vector spaces. In *Proceedings of NAACL-HLT*, pp. 921–929.

Peirsman, Y., & Padó, S. (2011). Semantic relations in bilingual lexicons. *ACM Transactions on Speech and Language Processing (TSLP)*, *8*(2), 3.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pp. 1532–1543.

Pham, H., Luong, M.-T., & Manning, C. D. (2015). Learning distributed representations for multilingual text sequences. In *Proceedings of the Workshop on Vector Modeling for NLP*, pp. 88–94.

Platt, J. C., Toutanova, K., & Yih, W.-T. (2010). Translingual document representations from discriminative projections. In *Proceedings of EMNLP*, pp. 251–261.

Prettenhofer, P., & Stein, B. (2010). Cross-language text classification using structural correspondence learning. In *Proceedings of ACL*, pp. 1118–1127.

Rajendran, J., Khapra, M. M., Chandar, S., & Ravindran, B. (2016). Bridge correlational neural networks for multilingual multimodal representation learning. In *Proceedings of NAACL-HLT*, pp. 171–181.

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of ACL*, pp. 519–526.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*(10), 627–633.

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of EMNLP*, pp. 298–307.

Schönemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika*, *31*(1), 1–10.

Schultz, T., & Waibel, A. (2001). Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication*, *35*(1), 31–51.

Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Shezaf, D., & Rappoport, A. (2010). Bilingual lexicon generation using non-aligned signatures. In *Proceedings of ACL*, pp. 98–107.

Shi, T., Liu, Z., Liu, Y., & Sun, M. (2015). Learning cross-lingual word embeddings via matrix co-factorization. In *Proceedings of ACL*, pp. 567–572.

Smith, S. L., Turban, D. H. P., Hamblin, S., & Hammerla, N. Y. (2017). Bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of ICLR (Conference Track)*.

Snyder, B., & Barzilay, R. (2010). Climbing the tower of Babel: Unsupervised multilingual learning. In *Proceedings of ICML*, pp. 29–36.

Søgaard, A. (2011). Data point selection for cross-language adaptation of dependency parsers. In *ACL*.

Søgaard, A. (2016). Evaluating word embeddings with fmri and eye-tracking. *ACL 2016*, 116.

Søgaard, A., Agić, Z., Alonso, H. M., Plank, B., Bohnet, B., & Johannsen, A. (2015). Inverted indexing for cross-lingual NLP. In *Proceedings of ACL*, pp. 1713–1722.

Søgaard, A., Ruder, S., & Vulić, I. (2018). On the Limitations of Unsupervised Bilingual Dictionary Induction. In *Proceedings of ACL 2018*.

Sorg, P., & Cimiano, P. (2012). Exploiting Wikipedia for cross-lingual and multilingual information retrieval. *Data and Knowledge Engineering*, *74*, 26–45.

Soyer, H., Stenetorp, P., & Aizawa, A. (2015). Leveraging monolingual data for crosslingual compositional word representations. In *Proceedings of ICLR (Conference Track)*.

Sylak-Glassman, J., Kirov, C., Yarowsky, D., & Que, R. (2015). A language-independent feature schema for inflectional morphology. In *Proceedings of ACL*, pp. 674–680.

Täckström, O., McDonald, R., & Uszkoreit, J. (2012). Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceeding of NAACL-HLT*, pp. 477–487.

Tamura, A., Watanabe, T., & Sumita, E. (2012). Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of EMNLP*, pp. 24–36.

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition, proceedings of the 6th conference on natural language learning. In *Proceedings of CoNLL*, pp. 1–4.

Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL*, pp. 142–147.

Tsai, C.-T., & Roth, D. (2016). Cross-lingual wikification using multilingual embeddings. In *Proceedings of NAACL-HLT*, pp. 589–598.

Tsvetkov, Y., Faruqui, M., Ling, W., Lample, G., & Dyer, C. (2015). Evaluation of word vector representations by subspace alignment. In *Proceedings of EMNLP*, pp. 2049–2054.

Tsvetkov, Y., Sitaram, S., Faruqui, M., Lample, G., Littell, P., Mortensen, D., Black, A. W., Levin, L., & Dyer, C. (2016). Polyglot neural language models: A case study in cross-lingual phonetic representation learning. In *Proceedings of NAACL-HLT*, pp. 1357–1366.

Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 384–394. Association for Computational Linguistics.

Upadhyay, S., Faruqui, M., Dyer, C., & Roth, D. (2016). Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of ACL*, pp. 1661–1670.

Vulić, I. (2017). Cross-lingual syntactically informed distributed word representations. In *Proceedings of EACL*, pp. 408–414.

Vulić, I., De Smet, W., & Moens, M.-F. (2011). Identifying word translations from comparable corpora using latent topic models. In *Proceedings of ACL*, pp. 479–484.

Vulić, I., De Smet, W., Tang, J., & Moens, M. (2015). Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing and Management*, *51*(1), 111–147.

Vulić, I., Kiela, D., Clark, S., & Moens, M.-F. (2016). Multi-modal representations for improved bilingual lexicon learning. In *Proceedings of ACL*, pp. 188–194.

Vulić, I., & Korhonen, A. (2016). On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of ACL*, pp. 247–257.

Vulić, I., & Moens, M.-F. (2013a). Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *Proceedings of NAACL-HLT*, pp. 106–116.

Vulić, I., & Moens, M.-F. (2013b). A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proceedings of EMNLP*, pp. 1613–1624.

Vulić, I., & Moens, M. (2015). Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of SIGIR*, pp. 363–372.

Vulić, I., & Moens, M.-F. (2016). Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, *55*, 953–994.

Vulić, I., Schwartz, R., Rappoport, A., Reichart, R., & Korhonen, A. (2017). Automatic selection of context configurations for improved class-specific word representations. In *Proceedings of CoNLL*, pp. 112–122.

Vyas, Y., & Carpuat, M. (2016). Sparse bilingual word representations for cross-lingual lexical entailment. In *Proceedings of NAACL-HLT*, pp. 1187–1197.

Wen, T.-H., Vandyke, D., Mrkšić, N., Gašić, M., Rojas-Barahona, L. M., Su, P.-H., Ultes, S., & Young, S. (2017). A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of EACL*, pp. 438–449.

Wijaya, D. T., Callahan, B., Hewitt, J., Gao, J., Ling, X., Apidianaki, M., & Callison-Burch, C. (2017). Learning translations via matrix completion. In *Proceedings of EMNLP*, pp. 1452–1463.

Xiao, M., & Guo, Y. (2014). Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of CoNLL*, pp. 119–129.

Xing, C., Liu, C., Wang, D., & Lin, Y. (2015). Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of NAACL-HLT*, pp. 1005–1010.

Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the ACL*, *2*, 67–78.

Zamani, H., & Croft, W. B. (2016). Embedding-based query language models. In *Proceedings of ICTIR*, pp. 147–156.

Zeman, D., & Resnik, P. (2008). Cross-language parser adaptation between related languages. In *Proceedings of IJCNLP*, pp. 35–42.

Zeman et al., D. (2017). CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 1–19.

Zhang, D., Mei, Q., & Zhai, C. (2010). Cross-lingual latent topic extraction. In *Proceedings of ACL*, pp. 1128–1137.

Zhang, M., Liu, Y., Luan, H., Liu, Y., & Sun, M. (2016). Inducing Bilingual Lexica From Non-Parallel Data With Earth Mover's Distance Regularization. In *Proceedings of COLING 2016*, pp. 3188–3198.

Zhang, M., Liu, Y., Luan, H., & Sun, M. (2017a). Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of ACL*, pp. 1959–1970.

Zhang, M., Liu, Y., Luan, H., & Sun, M. (2017b). Earth Mover ' s Distance Minimization for Unsupervised Bilingual Lexicon Induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Zhang, R., Lee, H., & Radev, D. R. (2016a). Dependency sensitive convolutional neural networks for modeling sentences and documents. In *Proceedings of NAACL-HLT*, pp. 1512–1521.

Zhang, Y., Gaddy, D., Barzilay, R., & Jaakkola, T. (2016b). Ten Pairs to Tag – Multilingual POS tagging via coarse mapping between embeddings. In *Proceedings of NAACL-HLT*, pp. 1307–1317.

Zoph, B., & Knight, K. (2016). Multi-source neural translation. In *Proceedings of NAACL-HLT*, pp. 30–34.

Zou, W. Y., Socher, R., Cer, D., & Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. In *Proceedings of EMNLP*, pp. 1393–1398.