Massively Multilingual Word Embeddings

Waleed Ammar **George Mulcaire** **Yulia Tsvetkov**
Guillaume Lample Chris Dyer Noah A. Smith

Abstract

We introduce new methods for estimating and evaluating embeddings of words from dozens of languages in a single shared embedding space. Our estimation methods, multiCluster and multiCCA, use dictionaries and monolingual data; they do not require parallel data. Our new evaluation method, multiQVEC+, is shown to correlate better than previous ones with two downstream tasks (text categorization and parsing). On this evaluation and others, our estimation methods outperform existing ones. We also describe a web portal for evaluation that will facilitate further research in this area, along with open-source releases of all our methods.

1 Introduction

Vector-space representations of words are widely used in statistical models of natural language. In addition to improvements on standard monolingual NLP tasks (Collobert and Weston, 2008), shared representation of words *across* languages offer intriguing possibilities (Klementiev et al., 2012). For example, in machine translation, translating a word never seen in parallel data may be overcome by seeking its vector-space neighbors, provided the embeddings are learned from both plentiful monolingual corpora and more limited parallel data. A second opportunity comes from transfer learning, in which models trained in one language can be deployed in other languages. While previous work has used hand-engineered features that are cross-linguistically stable as the basis model transfer (Zeman and Resnik, 2008; McDonald et al., 2011), automatically learned embeddings offer the promise of better generalization at lower cost (Klementiev et al., 2012; Hermann and Blunsom, 2014; Guo et al., 2016). We therefore conjecture that developing estimation methods for "massively" multilingual word embeddings (i.e., embeddings for words in a large number of languages) will play an important role in the future of multilingual NLP.

This paper makes the following contributions to this area. First, we articulate the desiderata for multilingual embeddings and propose two new estimation methods that fulfill these (§2). These methods are designed to use only monolingual data in each language and pairwise parallel dictionaries (no parallel corpora are required), and they scale to any number of languages (a number of previous models have been limited to only pairs of languages). Second, we propose an automatic evaluation methodology designed to test how well these goals are fulfilled (§3). This includes **multiQVEC+**, an inexpensive to compute evaluation which correlates well with performance on two downstream multilingual tasks, cross-lingual document categorization and cross-lingual parsing. Although intrinsic evaluations will never be perfect,[1] a standard set of evaluation metrics will help drive research. We evaluate our two proposed methods and two existing methods on various sets of languages consisting of 3, 12, and 59 languages (§5). Our proposed methods outperform existing methods on existing intrinsic metrics, the two extrinsic tasks, and our newly proposed multiQVEC+. Finally, in addition to an open-source

[1] Goodhart's eponymous law warns that "When a measure becomes a target, it ceases to be a good measure."

implementation of our methods, we include a link to a public web portal for uploading arbitrary multilingual embeddings and evaluating them automatically using a suite of intrinsic and extrinsic evaluation methods (§4).

2 Estimating Multilingual Embeddings

Let $\mathcal{L}$ be a set of languages, and let $\mathcal{V}^m$ be the set of surface forms (word types) in $m \in \mathcal{L}$. Let $\mathcal{V} = \bigcup_{m \in \mathcal{L}} \mathcal{V}^m$. Our goal is to estimate a partial **embedding** function $E : \mathcal{L} \times \mathcal{V} \nrightarrow \mathbb{R}^d$ (allowing a surface form that appears in two languages to have different vectors in each). We would like to estimate this function such that: (i) semantically similar words in the same language are nearby, (ii) translationally equivalent words in different languages are nearby, and (iii) the domain of the function covers as many words in $\mathcal{V}$ as possible.

We use distributional similarity in a monolingual corpus M^m to model semantic similarity between words in the same language. For cross-lingual similarity, either a parallel corpus $P^{m,n}$ or a bilingual dictionary $D^{m,n} \subset \mathcal{V}^m \times \mathcal{V}^n$ can be used. Our methods focus on the latter, in some cases extracting $D^{m,n}$ from a parallel corpus.[2]

With three notable exceptions (see §2.3, §2.4, §6), previous work on multilingual embeddings only considered the bilingual case, $\mid \mathcal{L} \mid = 2$. In this section, we focus on estimating multilingual embeddings for $\mid \mathcal{L} \mid > 2$ and describe two novel methods (multiCluster and multiCCA), then review the translation-invariance matrix factorization method (Gardner et al., 2015) and a variant of the multiSkip method (Guo et al., 2016).[3]

2.1 Multilingual cluster (multiCluster) embeddings

In this approach, we decompose the problem into two simpler subproblems: $E = E_{\text{embed}} \circ E_{\text{cluster}}$, where $E_{\text{cluster}} : \mathcal{L} \times \mathcal{V} \nrightarrow \mathcal{C}$ deterministically maps words to multilingual clusters $\mathcal{C}$, and $E_{\text{embed}} : \mathcal{C} \rightarrow \mathbb{R}^d$ assigns a vector to each cluster. We use a bilingual dictionary to find clusters of translationally equivalent words, then use distributional similarities of the clusters in monolingual corpora from all languages in $\mathcal{L}$ to estimate an embedding for each cluster. By forcing words from different languages in a cluster to share the same embedding, we create anchor points in the vector space to bridge languages.

More specifically, we define the clusters as the connected components in a graph where nodes are (language, surface form) pairs and edges correspond to translation entries in $D^{m,n}$. We assign arbitrary IDs to the clusters and replace each word token in each monolingual corpus with the corresponding cluster ID, and concatenate all modified corpora. The resulting corpus consists of multilingual cluster ID sequences. We can then apply any monolingual embedding estimator; here, we use the skip-gram model from Mikolov et al. (2013a).

2.2 Multilingual CCA (multiCCA) embeddings

Faruqui and Dyer (2014) proposed a bilingual embedding estimation method based on canonical correlation analysis (CCA) and showed that the resulting embeddings for English words outperform monolingually-trained English embeddings on word similarity tasks. First, they use monolingual corpora to train monolingual embeddings for each language independently (E^m and E^n), capturing semantic similarity within each language separately. Then, using a bilingual dictionary $D^{m,n}$, they use CCA to estimate linear projections from the ranges of the monolingual embeddings E^m and E^n, yielding a bilingual embedding $E^{m,n}$. The linear projections are defined by $T_{m \rightarrow m,n}$ and $T_{n \rightarrow m,n} \in \mathbb{R}^{d \times d}$; they are selected to maximize the correlation between $T_{m \rightarrow m,n} E^m(u)$ and $T_{n \rightarrow m,n} E^n(v)$ where $(u, v) \in D^{m,n}$. The bilingual embedding is then defined as $E_{\text{CCA}}(m, u) = T_{m \rightarrow m,n} E^m(u)$ (and likewise for $E_{\text{CCA}}(n, v)$).

In this work, we use this method as a building block to construct multilingual embeddings for more languages. We let the vector space of the initial (monolingual) English embeddings serve as the multilingual vector space (since English typically offers the largest corpora and wide availability of bilingual dictionaries). We then estimate projections from the monolingual embeddings of the other languages into

[2] To do this, we align the corpus using fast_align (Dyer et al., 2013) in both directions. The estimated parameters of the word translation distributions are used to select pairs: $D^{m,n} = \{(u, v) \mid u \in \mathcal{V}^m, v \in \mathcal{V}^n, p_{m|n}(u \mid v) \times p_{n|m}(v \mid u) > \tau\}$, where the threshold τ trades off dictionary recall and precision. We fixed $\tau = 0.1$ early on based on manual inspection of the resulting dictionaries.

[3] We developed the multiSkip method independently of Guo et al. (2016).

the English space.

We start by estimating, for each $m \in \mathcal{L} \setminus \{\text{en}\}$, the two projection matrices: $T_{m\to m,\text{en}}$ and $T_{\text{en}\to m,\text{en}}$; these are guaranteed to be non-singular. We then define the multilingual embedding as $E_{\text{CCA}}(\text{en}, u) = E^{\text{en}}(u)$ for $u \in \mathcal{V}^{\text{en}}$, and $E_{\text{CCA}}(m, v) = T^{-1}_{\text{en}\to m,\text{en}} T_{m\to m,\text{en}} E^m(v)$ for $v \in \mathcal{V}^m, m \in \mathcal{L} \setminus \{\text{en}\}$.

Though not explored here, this approach generalizes even without a single "hub" language (English) with which every other language shares a bilingual dictionary. If the languages are all connected by bilingual dictionaries, then we can select any spanning tree of the "language graph" induced by the bilingual dictionaries, and any language as the "root." Words in any language can be iteratively projected into the vector spaces along the path to the root using the technique described above. In future work, non-linear transformations might be explored as well.

2.3 MultiSkip embeddings

Luong et al. (2015b) proposed a method for estimating a bilingual embedding which only makes use of parallel data; it extends the skipgram model of Mikolov et al. (2013a). The skipgram model defines a distribution over words u that occur in a context window (of size K) of a word v:

$$p(u \mid v) = \frac{\exp E_{\text{skipgram}}(m, v)^\top E_{\text{context}}(m, u)}{\sum_{u' \in \mathcal{V}^m} \exp E_{\text{skipgram}}(m, v)^\top E_{\text{context}}(m, u')}$$

In practice, this distribution can be estimated using a noise contrastive estimation approximation (Gutmann and Hyvärinen, 2012) while maximizing the log-likelihood:

$$\sum_{i \in \text{pos}(M^m)} \sum_{k \in \{-K,\ldots,-1,1,\ldots,K\}} \log p(u_{i+k} \mid u_i)$$

where $\text{pos}(M^m)$ are the indices of words in the monolingual corpus M^m.

To establish a bilingual embedding, with a parallel corpus $P^{m,n}$ of source language m and target language n, Luong et al. (2015b) estimate conditional models of words in both source and target positions. The source positions are selected as sentential contexts (similar to monolingual skipgram), and the bilingual contexts come from aligned words. The bilingual objective is to maximize:

$$\begin{aligned}
&\sum_{i \in m\text{-pos}(P_{m,n})} \sum_{k \in \{-K,\ldots,-1,1,\ldots,K\}} \begin{array}{l} \log p(u_{i+k} \mid u_i) \\ +\log p(v_{a(i)+k} \mid u_i) \end{array} \\
&+ \sum_{j \in n\text{-pos}(P_{m,n})} \sum_{k \in \{-K,\ldots,-1,1,\ldots,K\}} \begin{array}{l} \log p(v_{j+k} \mid v_j) \\ +\log p(u_{a(j)+k} \mid v_j) \end{array}
\end{aligned} \tag{1}$$

where $m\text{-pos}(P_{m,n})$ and $n\text{-pos}(P_{m,n})$ are the indeces of the source and target tokens in the parallel corpus respectively, $a(i)$ and $a(j)$ are the positions of words that align to i and j in the other language. It is easy to see how this method can be extended for more than two languages by summing up the bilingual objective in Eq. 1 for all available parallel corpora.

2.4 Translation-invariant matrix factorization

Gardner et al. (2015) proposed that multilingual embeddings should be translation invariant. Consider a matrix $X \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ which summarizes the pointwise mutual information statistics between pairs of words in monolingual corpora, and let $UV^\top$ be a low-rank decomposition of X where $U, V \in \mathbb{R}^{|\mathcal{V}|\times d}$. Now, consider another matrix $A \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ which summarizes bilingual alignment frequencies in a parallel corpus. Gardner et al. (2015) solves for a low-rank decomposition $UV^\top$ which both approximates X as well as its transformations $A^\top X$, XA and $A^\top XA$ by defining the following objective:

$$\begin{aligned}
&min_{U,V} \|X - UV^\top\|^2 + \|XA - UV^\top\|^2 \\
&+ \|A^\top X - UV^\top\|^2 + \|A^\top XA - UV^\top\|^2
\end{aligned}$$

The multilingual embeddings are then taken to be the rows of the matrix U.

3 Evaluating Multilingual Embeddings

One of our main contributions is to streamline the evaluation of multilingual embeddings. In addition to assessing goals (i–iii) stated in §2, a good evaluation metric should also (iv) show good correlation with performance in downstream applications and (v) be computationally efficient.

It is easy to evaluate the coverage (iii) by counting the number of words covered by an embedding

function in a closed vocabulary. Intrinsic evaluation metrics are generally designed to be computationally efficient (v) but may or may not meet the goals (i, ii, iv). By design, standard (monolingual) word similarity tasks meet (i) while cross-lingual word similarity tasks and the word translation tasks meet (ii). We propose another evaluation method (multiQVEC+), designed to simultaneously assess goals (i, ii). MultiQVEC+ extends QVEC (Tsvetkov et al., 2015), a recently proposed monolingual evaluation method, addressing fundamental flaws and extending it to multiple languages. To assess the degree to which these evaluation metrics meet (iv), in §5 we perform a correlation analysis looking at which intrinsic metrics are best correlated with downstream task performance—i.e., we evaluate the evaluation metrics.

3.1 Word similarity

Word similarity datasets such as WS-353-SIM (Agirre et al., 2009) and MEN (Bruni et al., 2014) provide human judgments of semantic similarity. By ranking words by cosine similarity and by their empirical similarity judgments, a ranking correlation can be computed that assesses how well the estimated vectors capture human intuitions about semantic relatedness.

Some of previous work on bilingual and multilingual embeddings has focused on monolingual word similarity to evaluate embeddings (e.g., Faruqui and Dyer., 2014). This approach is limited because it cannot measure the degree to which embeddings from different languages are similar (ii). For this paper, we report results on an English word similarity task, the Stanford RW dataset (Luong et al., 2013), as well as a combination of several cross-lingual word similarity datasets (Camacho-Collados et al., 2015).

3.2 Word translation

This task directly assesses the degree to which translationally equivalent words in different languages are nearby in the embedding space. The evaluation data consists of word pairs which are known to be translationally equivalent. The score for one word pair $(l_1, w_1), (l_2, w_2)$ both of which are covered by an embedding E is 1 if $\text{cosine}(E(l_1, w_1), E(l_2, w_2)) \geq \text{cosine}(E(l_1, w_1), E(l_2, w'_2)) \; \forall \, w'_2 \in G^{l_2}$ where G^{l_2} is the set of words of language l_2 in the evaluation dataset, and cosine is the cosine similarity function. Otherwise, the score for this word pair is 0. The overall score is the average score for all word pairs covered by the embedding function. This is a variant of the method used by Mikolov et al. (2013b) to evaluate bilingual embeddings.

3.3 Correlation-based evaluation

We introduce QVEC+—an intrinsic evaluation measure of the quality of monolingual and multilingual word embeddings. Our method is a monolingual improvement and a multilingual extension of QVEC—a recently proposed monolingual evaluation based on alignment of embeddings to a matrix of features extracted from a linguistic resource (Tsvetkov et al., 2015). We review QVEC, and then describe QVEC+.

QVEC. The main idea behind QVEC is to quantify the linguistic content of word embeddings by maximizing the correlation with a manually-annotated linguistic resource. Let the number of common words in the vocabulary of the word embeddings and the linguistic resource be N. To quantify the semantic content of embeddings, a semantic linguistic matrix $\mathbf{S} \in \mathbb{R}^{P \times N}$ is constructed from a semantic database, with a column vector for each word. Each word vector is a distribution of the word over P linguistic properties, based on annotations of the word in the database. Let $\mathbf{X} \in \mathbb{R}^{D \times N}$ be embedding matrix with every row as a dimension vector $\mathbf{x} \in \mathbb{R}^{1 \times N}$. D denotes the dimensionality of word embeddings. Then, $\mathbf{S}$ and $\mathbf{X}$ are aligned to maximize the cumulative correlation between the aligned dimensions of the two matrices. Specifically, let $\mathbf{A} \in \{0,1\}^{D \times P}$ be a matrix of alignments such that $a_{ij} = 1$ iff $\mathbf{x}_i$ is aligned to $\mathbf{s}_j$, otherwise $a_{ij} = 0$. If $r(\mathbf{x}_i, \mathbf{s}_j)$ is the Pearson's correlation between vectors $\mathbf{x}_i$ and $\mathbf{s}_j$, then QVEC is defined as:

$$\text{QVEC} = max_{\mathbf{A}: \sum_j a_{ij} \leq 1} \sum_{i=1}^{X} \sum_{j=1}^{S} r(\mathbf{x}_i, \mathbf{s}_j) \times a_{ij}$$

The constraint $\sum_j a_{ij} \leq 1$, warrants that one distributional dimension is aligned to at most one linguistic dimension.

QVEC has been shown to correlate strongly with downstream semantic tasks. However, it suffers

from two major weaknesses. First, it is not invariant to linear transformations of the embeddings' basis, whereas the bases in word embeddings are generally arbitrary (Szegedy et al., 2014). Second, a sum of correlations produces an unnormalized score: the more dimensions in the embedding matrix the higher the score. This precludes comparison of models of different dimensionality. QVEC+ simultaneously addresses both problems.

QVEC+. To measure correlation between the embedding matrix $\mathbf{X}$ and the linguistic matrix $\mathbf{S}$, instead of cumulative dimension-wise correlation we employ CCA. CCA finds two sets of basis vectors, one for $\mathbf{X}^\top$ and the other for $\mathbf{S}^\top$, such that the correlations between the projections of the matrices onto these basis vectors are maximized. Formally, CCA finds a pair of basis vectors $\mathbf{v}$ and $\mathbf{w}$ such that

$$\text{QVEC+} = \text{CCA}(\mathbf{X}^\top, \mathbf{S}^\top) = max_{\mathbf{v},\mathbf{w}}\, r(\mathbf{X}^\top\mathbf{v}, \mathbf{S}^\top\mathbf{w})$$

Thus, QVEC+ ensures invariance to the matrices bases rotation, and since it is a single correlation, it produces a score in $[-1, 1]$. Both QVEC and QVEC+ rely on a matrix of linguistic properties constructed from a manually crafted linguistic resource. In this paper, instead of only constructing the linguistic matrix based on monolingual annotations, we use supersense tag annotations for English (Miller et al., 1993), Danish (Martınez Alonso et al., 2015) and Italian (Montemagni et al., 2003) to create extensions of QVEC and QVEC+ for the multilingual case; henceforth, multiQVEC and multiQVEC+.

3.4 Extrinsic tasks

In order to evaluate how useful the word embeddings are for a downstream task, we use the embedding vector as a dense feature representation of each word in the input, and deliberately remove any other feature available for this word (e.g., prefixes, suffixes, part-of-speech). For each task, we train one model on the aggregate training data available for several languages, and evaluate on the aggregate evaluation data in the same set of languages. We apply this for multilingual document classification and multilingual dependency parsing.

For document classification, we follow Klementiev et al. (2012) in using the RCV corpus of newswire text, and train a classifier which differentiates between four topics. While most previous work which used this data only in a bilingual setup, we simultaneously train the classifier on documents in seven languages,[4] and evaluate on the development/test section of those languages. For this task, we report the average classification accuracy on the test set.

For dependency parsing, we train the stack-LSTM parser of Dyer et al. (2015) on a subset of the languages in the universal dependencies v1.1[5], and test on the same languages, reporting unlabeled attachment scores. We remove all part-of-speech and morphology features from the data, and prevent the model from optimizing the word embeddings used to represent each word in the corpus, thereby forcing the parser to rely completely on the provided (pretrained) embeddings as the token representation.

4 Evaluation Portal

In order to facilitate future research on multilingual word embeddings, we developed a web portal[6] to enable researchers who develop new estimation methods to evaluate them using a suite of evaluation tasks. The portal serves the following purposes:

- Download the monolingual and bilingual data we used to estimate multilingual embeddings in this paper,
- Download standard development/test data sets for each of the evaluation metrics to help researchers working in this area report trustworthy and replicable results,[7]
- Upload arbitrary multilingual embeddings, scan which languages are covered by the embeddings, allow the user to pick among the compatible evaluation tasks, and receive evaluation scores for the selected tasks, and
- Register a new evaluation data set or a new evaluation metric via the github repository

[4] Danish, German, English, Spanish, French, Italian and Swedish.

[5] http://hdl.handle.net/11234/LRT-1478

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

[7] Except for the original RCV documents, which are restricted by the Reuters license and cannot be republished. All other data is available for download.

metric	language ISO 639-1 codes
document classification	da, de, en, it, fr, sv
dependency parsing	bg, cs, da, de, el, en, es, fi, fr, hu, it, sv
(multi)QVEC+/(multi)QVEC	da, en, it
word similarity	de, en, es, fa, fr, it, pt
word translation	bg, cs, da, de, el, en, es, fi, fr, hu, it, sv, zh, af, ca, iw, cy, ar, ga, zu, et, gl, id, ru, nl, pt, la, tr, ne, lv, lt, tg, ro, is, pl, yi, be, hy, hr, jw, ka, ht, fa, mi, bs, ja, mg, tl, ms, uz, kk, sr, mn, ko, mk, so, uk, sl, sw

Table 1: Evaluation metrics on the corpus and languages for which evaluation data are available.

which mirrors the backend of the web portal.

Table 1 lists the evaluation metrics used on the web portal along with the languages currently available.

5 Experiments

Our experiments are designed to show two primary sets of results: (i) how well the proposed intrinsic evaluation metrics correlate with downstream tasks that use multilingual word vectors (§5.1) and (ii) which estimation methods work best (§5.2).

5.1 Correlations between intrinsic vs. extrinsic evaluation metrics

In this experiment, we consider four intrinsic evaluation metrics (cross-lingual word similarity, word translation, multiQVEC and multiQVEC+) and two extrinsic evaluation metrics (multilingual document classification and multilingual parsing).

Data: All evaluation data sets we used are available for download on the evaluation portal. For the cross-lingual word similarity task, we use the 307 English-Italian word pairs in the multilingual MWS353 dataset (Leviant and Reichart, 2015). For the word translation task, we use a subset of 647 translation pairs from Wiktionary in English, Italian and Danish. For multiQVEC and multiQVEC+, we used the 41 supersense tag annotations (26 for nouns and 15 for verbs) as described in §3. For the downstream tasks, we use the English, Italian and Danish subsets of the RCV corpus and the universal dependencies v1.1.

Setup: Estimating correlations between the proposed intrinsic evaluation metrics and downstream task performance requires a sample of different vector embeddings with their intrinsic and extrinsic task scores. To create this sample, we trained a total of 17 different multilingual embeddings[8] for three languages (English, Italian and Danish).

(→) extrinsic task (↓) intrinsic metric	document classification	dependency parsing
word similarity	0.386	0.007
word translation	0.066	-0.292
multiQVEC	0.635	0.444
multiQVEC+	0.896	0.273

Table 2: Correlations between intrinsic evaluation metrics (rows) and downstream task performance (columns).

Results: Table 2 shows the correlations of the four intrinsic metrics against the performance of the vectors on the two downstream tasks. We establish (i) that intrinsic methods used in the literature (cross-lingual word similarity and word translation) are poorly correlated with downstream tasks, and (ii) that both intrinsic methods we propose for evaluating multilingual word embeddings (i.e., multiQVEC and multiQVEC+) strongly correlate with both multilingual document classification and multilingual dependency parsing.

5.2 Evaluating multilingual estimation methods

We now turn to evaluating multilingual embeddings obtained using the estimation methods in §2.

Languages: We compare the four estimation methods in §2 on three language sets of $\{3, 12, 59\}$ languages.[9] Since the multiSkip and translation-invariance methods require word translation probabilities, we were only able to use them with the $\{3, 12\}$-language sets for which we have parallel corpora.

[8] $17 = 12$ multiCluster embeddings $+1$ multiCCA embeddings $+1$ multiSkip embeddings $+2$ translation-invariance embeddings.

[9] The ISO 639-1 codes of the three language sets we used are: {da, en, it}, {bg, cs, da, de, el, en, es, fi, fr, hu, it, sv}, and { bg, cs, da, de, el, en, es, fi, fr, hu, it, sv, zh, af, ca, iw, cy, ar, ga, zu, et, gl, id, ru, nl, pt, la, tr, ne, lv, lt, tg, ro, is, pl, yi, be, hy, hr, jw, ka, ht, fa, mi, bs, ja, mg, tl, ms, uz, kk, sr, mn, ko, mk, so, uk, sl, sw }.

Data: As mentioned in §2, the multiCluster and multiCCA estimation methods only require monolingual corpora and bilingual dictionaries, while the multiSkip and translation-invariance methods require parallel data. Details and pointers for downloading the data used to estimate and evaluate embeddings in each set of languages can be found on the evaluation portal.

Setup: All embeddings trained for this evaluation have 40 dimensions. We used the development section of the evaluation methods (see §4) for tuning hyperparameters. All skipgram-based models (multiCCA, multiSkip, and multiCluster) were trained using 10 epochs of stochastic gradient descent. We used a context window size of 3 for the translation-invariance method.[10] For the other methods, we used a context window size of 3 (for the 3-language and 59-language embeddings) and 5 (for the 12-language embeddings). We only estimated embeddings for words/clusters which occur 5 times or more in the monolingual corpora.

Results: We report results of this experiment separately for each set of languages in Tables 3, 4, and 5. Looking at the performance of downstream tasks in Tables 3, 4, we establish that both our proposed dictionary-based methods (multiCCA and multiCluster) outperform the multiSkip and translation-invariance methods. This is consistent for both classification and parsing. It is also clear that multiCCA outperforms multiCluster on most tasks. However, the intrinsic metrics do not always agree with the extrinsic results.

Although the results are not always comparable across the three tables,[11] some of them are. For instance, the results of (multi)QVEC and (multi)QVEC+ are comparable across the three tables, since the semantic annotations required for computing the score are only available in Danish, English and Italian. We can see that the performance tends to decline as we go from three to twelve to fifty-nine languages. This is especially true for the multiCluster method because using bilingual dictionaries for more languages result in conflating more and more words in the same cluster, and all words in the same cluster share the same embedding. To avoid this problem, it may be worth exploring better ways of constructing multilingual word clustering from bilingual dictionaries (e.g., spectral clustering).

6 Previous Work

We focused on methods for training multilingual embeddings for many languages, but there is a rich body of literature on bilingual embeddings, including work on machine translation (Zou et al., 2013; Hermann and Blunsom, 2014; Cho et al., 2014; Luong et al., 2015b; Luong et al., 2015a, *inter alia*),[12] cross-lingual dependency parsing (Guo et al., 2015; Guo et al., 2016), and cross-lingual document classification (Klementiev et al., 2012; Gouws et al., 2014; Kociskỳ et al., 2014). Word clusters is a related form of distributional representation; in clustering, cross-lingual distributional representations were proposed as well (Täckström et al., 2012).

7 Conclusion

We introduced two estimation methods for multilingual word embeddings, multiCCA and multiCluster, which only require bilingual dictionaries and monolingual corpora, and used them to train embeddings for 59 languages. We found the embeddings estimated using our dictionary-based methods to outperform those estimated using other methods for two downstream tasks: multilingual dependency parsing and multilingual document classification. We also developed a new intrinsic method to evaluate multilingual embeddings and showed that it strongly correlates with downstream tasks (and runs faster). Finally, in order to help future research in this area, we created a web portal for users to upload their multilingual embeddings and easily evaluate them on nine evaluation metrics, with two modes of operation (development and test) to encourage sound experimentation practices.

[10] Constructing the pointwise mutual information matrix for larger context window sizes was computaionally challenging.

[11] For example, the evaluation set used for word translation task in Tables 3, 4, and 5 uses word pairs in 3, 12, and 59 languages, respectively.

[12] Hermann and Blunsom (2014) showed that the bicvm method can be extended to more than two languages, but the released software library only supports bilingual embeddings. We tried following the first author's recommendation at `https://github.com/karlmoritz/bicvm/issues/4`, but we were not able to reproduce their results.

	Task	multiCluster	multiCCA	multiSkip	invariance
extrinsic metrics	dependency parsing	0.723	**0.745**	0.651	0.612
	document classification	**0.823**	0.802	0.770	0.759
intrinsic metrics	English word similarity	0.309	0.371	0.327	**0.436**
	multilingual word similarity	0.411	0.442	0.407	**0.450**
	word translation	0.422	0.107	**0.470**	0.464
	monolingual QVEC	0.173	**0.174**	0.129	0.166
	multiQVEC	**0.171**	0.162	0.124	0.116
	monolingual QVEC+	0.269	0.282	0.241	**0.302**
	multiQVEC+	**0.233**	0.225	0.197	0.226

Table 3: Results for multilingual embeddings that cover Danish, English and Italian. Each row corresponds to one of the embedding evaluation metrics we use (higher is better). Each column corresponds to one of the embedding estimation methods we consider; i.e., numbers in the same row are comparable.

	Task	multiCluster	multiCCA	multiSkip	invariance
extrinsic metrics	dependency parsing	0.669	**0.708**	0.607	0.670
	document classification	**0.843**	0.842	0.817	0.821
intrinsic metrics	monolingual word similarity	0.290	0.376	0.402	**0.491**
	multilingual word similarity	0.347	0.522	0.457	**0.523**
	word translation	0.324	0.2	0.479	**0.593**
	monolingual QVEC	0.154	**0.183**	0.142	0.163
	multiQVEC	0.151	**0.170**	0.124	0.133
	monolingual QVEC+	0.250	0.280	0.254	**0.304**
	multiQVEC+	0.211	0.226	0.202	**0.253**

Table 4: Results for multilingual embeddings that cover Bulgarian, Czech, Danish, Greek, English, Spanish, German, Finnish, French, Hungarian, Italian and Swedish. Each row corresponds to one of the embedding evaluation metrics we use (higher is better). Each column corresponds to one of the embedding estimation methods we consider; i.e., numbers in the same row are comparable.

	Task	multiCluster	multiCCA
extrinsic metrics	dependency parsing	0.653	**0.700**
	document classification	0.809	**0.840**
intrinsic metrics	monolingual word similarity	0.093	**0.376**
	multilingual word similarity	0.035	**0.521**
	word translation (Wiktionary)	**0.237**	0.204
	word translation (Google)	0.087	**0.264**
	monolingual QVEC	0.115	**0.183**
	multiQVEC	0.135	**0.170**
	monolingual QVEC+	0.189	**0.280**
	multiQVEC+	0.177	**0.226**

Table 5: Results for multilingual embeddings that cover 59 languages. Each row corresponds to one of the embedding evaluation metrics we use (higher is better). Each column corresponds to one of the embedding estimation methods we consider; i.e., numbers in the same row are comparable.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. A framework for the construction of monolingual and cross-lingual word similarity datasets.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proc. of EMNLP*.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proc. of ICML*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. In *Proc. of NAACL*.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A Smith. 2015. Transition-based dependency parsing with stack long short-term memory. In *Proc. of ACL*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. *Proc. of EACL. Association for Computational Linguistics*.

Matt Gardner, Kejun Huang, Evangelos Papalexakis, Xiao Fu, Partha Talukdar, Christos Faloutsos, Nicholas Sidiropoulos, and Tom Mitchell. 2015. Translation invariant word embeddings. In *Proc. of EMNLP*.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2014. Bilbowa: Fast bilingual distributed representations without word alignments. *arXiv preprint arXiv:1410.2455*.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proc. of ACL*.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016. A representation learning framework for multi-source transfer parsing. In *Proc. of AAAI*.

Michael U Gutmann and Aapo Hyvärinen. 2012. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. In *Proc. of JMLR*.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual Models for Compositional Distributional Semantics. In *Proc. of ACL*.

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

Tomáš Kociskỳ, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning bilingual word representations by marginalizing alignments.

Ira Leviant and Roi Reichart. 2015. Judgment language matters: Towards judgment language informed vector space modeling. In *arXiv preprint arXiv:1508.00106*.

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

Minh-Thang Luong, Ilya Sutskever, Quoc V Le, Oriol Vinyals, and Wojciech Zaremba. 2015a. Addressing the rare word problem in neural machine translation. In *Proc. of ACL*.

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

Héctor Martınez Alonso, Anders Johannsen, Sussi Olsen, Sanni Nimb, Nicolai Hartvig Srensen, Anna Braasch, Anders Sgaard, and Bolette Sandford Pedersen. 2015. Supersense tagging for danish. In *Proc. of NODALIDA*, page 21.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 62–72. Association for Computational Linguistics.

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

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. In *arXiv preprint arXiv:1309.4168v1*.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proc. of HLT*, pages 303–308.

Simonetta Montemagni, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Alessandro Lenci, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, et al. 2003. Building the italian syntactic-semantic treebank. In *Treebanks*, pages 189–210. Springer.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *Proc. of ICLR*.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proc. of NAACL*, pages 477–487.

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

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proc. of IJCNLP*, pages 35–42.

Will Y Zou, Richard Socher, Daniel M Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proc. of EMNLP*, pages 1393–1398.