Minimally-Constrained Multilingual Embeddings via Artificial Code-Switching

Michael Wick
Oracle Labs
michael.wick@oracle.com

Pallika Kanani
Oracle Labs
pallika.kanani

Adam Pocock
Oracle Labs
adam.pocock

Abstract

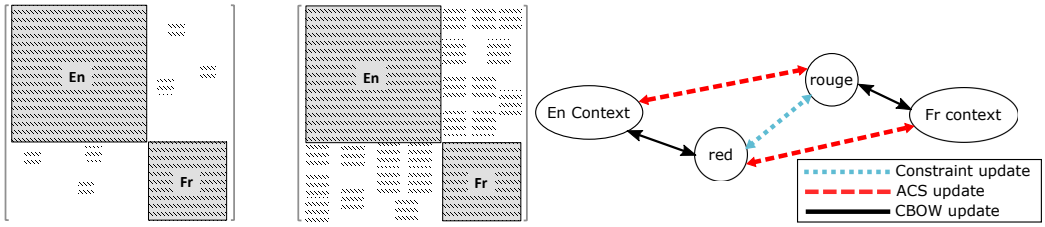
We present a method that consumes a large corpus of multilingual text and produces a single, unified word embedding in which the word vectors generalize across languages. Our method is agnostic about the languages with which the documents in the corpus are expressed, and does not rely on parallel corpora to constrain the spaces. Instead we utilize a small set of human provided word translations to artificially induce code switching; thus, allowing words in multiple languages to appear in contexts together and share distributional information. We evaluate the embeddings on a new multilingual word analogy dataset. We also find that our embeddings allow an NLP model trained in one language to generalize to another, achieving up to 80% of the accuracy of an in-language model.

1 Introduction

An important practical problem in natural language processing (NLP) is to make NLP tools (e.g., named entity recognition, parsers, sentiment analysis) available in every language. Many of the resources available in a language such as English are not available in languages with fewer speakers. One solution is to collect training data in every language for every task for every domain, but such data collection is expensive and time consuming. A second, more feasible solution, is to use large collections of unlabeled multilingual data to find a common representation (e.g., an embedding) in which structure is shared across languages. Under such representations, we can train an NLP model in a language with many resources and generalize that model to work on lower resource languages.

Word embeddings are a promising technique for learning multilingual representations because they map word-types to dense, low dimensional (e.g., 300) vectors [4, 2, 1, 15], and are advantageous for NLP because they help cope with the sparsity problems associated with text. Most approaches exploit the distributional hypothesis of language [9, 6] which stipulates that words are defined by how they are used in a large corpus of text. Embeddings trained simply to predict their context words capture a surprising amount of both syntactic and semantic meaning [15]; in essence, by merely factorizing a matrix of word co-occurrence statistics [13]. However, these techniques are likely to fail on multilingual data because words from different languages rarely appear in the same context together. That is, the co-occurrence matrix exhibits block-diagonal structure (Figure 1a).

Of course, there are some notable exceptions. *Named entities* (e.g., "iPad"), *lexical borrowing* [20], and the act of *code switching* in which a multilingual speaker switches between languages during a dialogue [14], all cause words in different languages to appear in the same contexts. Indeed, these phenomena may allow for a jointly trained embedding model to learn structure that generalizes across languages. Thus, we could try to learn word vectors with an approach such as CBOW on a large enough corpus and see if these phenomena occur with sufficient frequency to learn good multilingual embeddings. Alternatively, we would expect to do even better if we transform the model or data to more directly capture the desired multilingual structure.



(a) Biling. co-oc matrix (b) ACS co-oc matrix. (c) Different updates on word-context pairs.

Figure 1

We propose artificial code switching (ACS), a method that employs dictionaries to swap words in one language with its translation in another. ACS effectively fills in more cells of the co-occurrence matrix (Figure 1b) making the matrix less block diagonal, and thus ripe for learning multilingual representations. In much the same way that machine vision practitioners apply affine transformations to their training to data to learn invariance to rotation and scale, we too apply a special type of transformation (ACS) to our input data to learn invariance to language. Our method improves the quality of multilingual embeddings over a system that relies upon natural cross lingual co-occurrences alone (as measured on a new multilingual word analogy dataset). We can use the embeddings to train a sentiment classifier on a source language that generalizes to multiple target languages (with accuracy as high as 80% of the in-language sentiment classifiers).

2 Related Work

Most work on learning multilingual embeddings require language identification and parallel corpora. The aligned translated documents make it possible to constrain the monolingual outputs of recurrent neural networks (RNNs) [19], multilayer perceptrons [7], auto-encoders [17], multi-task learners [12], and other models [11, 5] to agree on a common bilingual representation. Unfortunately, finding a sufficiently large parallel corpus of paired documents in the source and target languages is difficult, especially for lower resource languages. Contemporaneously, Barista [8] employs an algorithm similar to artificial code switching, but for bilingual data. They demonstrate that the algorithm works remarkably well for the bilingual setting, allowing a model (e.g., a part of speech tagger) trained in one language to generalize to another.

In contrast to the aforementioned approaches which are predominantly bilingual, we combine as many corpora from as many languages as possible into the same embedding space. Combining many languages together is not only more convenient (than managing many bilingual pairs), but has the potential to learn better quality embeddings. For example, on a document similarity task, recent work demonstrates that combining similarity scores across four languages results in a more accurate similarity measure than those based on a single language [10].

3 Multilingual Embeddings

Word embedding models such as CBOW capture a tremendous amount of monolingual structure by updating word vectors to predict their context (see black arrows in Figure 1c). We would like extend these models to generalize such structure across multiple languages. We are particularly interested in the social media domain in which language identification is difficult and parallel corpora are scarce.

3.1 Problem Setting

More formally, suppose we have M languages L_m with corresponding vocabularies V_m, then $V = \cup_{m=1}^{M} V_m$ is the vocabulary of all the languages. We have a large corpus of multilingual text $\mathcal{D}$ with documents $D_i \in \mathcal{D}$ comprised of word sequences $w_1, \cdots, w_{n_i}$ where each $w_j \in V$. We also have a small set of human-provided concept dictionaries $\mathcal{C}$ that link word-types across languages. A concept dictionary $\mathcal{C}$ is a set of concepts in which each concept $C_i \in \mathcal{C}$ is a set of words that all have similar meaning (e.g., a concept set containing "red", "rouge" and "rojo"). Note that we do not necessarily know the language for any given word or document.

	Embedding Data		Sentiment Data			Target-Lang. Baseline	
Language	#Docs (x10^6)	#Concepts	#Train	#Test	%Tw	TwA	TotA
English (en)	4.87	8821	24960	6393	36.0%	63.8%	68.4%
French (fr)	1.62	7408	14284	3081	36.0	69.9	74.9
German (de)	1.82	8258	12247	2596	25.5	70.2	74.9
Spanish (es)	1.18	6501	16506	59529	84.7	64.4	64.2
Bokmål (no)	0.41	5336	1225	1000	0.00	-	72.9

Table 1: Datasets. Sentiment accuracy on just twitter documents (TwA) and all documents (TotA).

Our task is to learn an embedding model $\mathcal{M} : V \rightarrow \mathbb{R}^k$ that maps each word type to a k-dimensional vector such that the vectors capture syntactic and semantic relations between words in a way that generalizes across the languages. We investigate a solution space that is modular in the sense that the multilingual approaches are compatible with many underlying (monolingual) embedding methods. In this way, it is easy to implement the techniques on top of existing embedding implementations such as LSA, RBMs, CBOW, SkipGram, GloVe, or LDA. One approach is to add constraints to the embedding objective that encourage words in the same concept sets to have similar vectors. Then, the resulting updates move these words together (blue arrows in Figure 1c). However, we find that balancing the dictionary-based constraints with the data-based constraints is difficult, and instead employ a somewhat unconventional approach that transforms the input data instead of the model.

3.2 Artificial Code-Switching

Code-switching is the process in which a speaker of multiple languages switches between those languages in discourse. Consider the utterance "*pizza khaneka mood nahi*" which translates to "not in the mood to eat pizza." The Hindi verb "khaneka" which means "to eat" and the English noun "pizza" are able to share distributional information via the code-switched utterance. As we can see, code-switching allows the distributional meaning of a word in one language to borrow from context in another language, thus providing a rich base from which to learn window-based embeddings.

Unfortunately, in written text, code-switching is an infrequent event. Thus, we use our dictionaries to artificially induce extra code-switching in the input data. This process, which we term artificial code-switching (ACS), fills in unobserved cells in the word to context-word co-occurrence matrix (see Figure 1b). This is analogous to having extra recommendations in a recommender system (i.e., recommendations that a word in one language could be substituted for one in another). An interesting question is how to fill the cells of this matrix in a way that most naturally causes the learning of shared structure in the multilingual space. In order to respect the distributional hypothesis, we want the co-occurrence statistics between words of different languages to resemble the co-occurrence statistics of words within a language. A simple way of accomplishing this is to fill in the matrix by randomly replacing a word in one language with its translation in another.

Specifically, we generate a new code-switched corpus $\mathcal{D}'$ by transforming each word $w_i \in \mathcal{D}$ with the following process. First, we draw a variable $s \sim Bernoulli(\alpha)$. If $s = true$ then we code-switch and sample a new word w_i'. To generate w_i' we sample a concept c from $C(w_i)$ then we sample a word w_i' from c. If $s = false$ then we do not code switch. We can then learn multilingual embeddings by running an existing system, such as CBOW, on $\mathcal{D}'$. Looking at Figure 1c again, we can see that the code-switching update moves the English word "red" closer to the French context for the word "rouge" and vice versa. This does not directly affect the relationship between "red" and "rouge" but over repeated updates it causes them to have similar vectors. Of course, more complex models of code-switching are possible, but our goal is not to model this phenomenon; rather, we are more interested in exploiting it to improve our multilingual embedding space.

4 Experiments

We are interested in studying the following questions. *Is it possible to learn a joint multilingual embedding from just a small subset of translated vocabulary words? Does learning a joint multilingual embedding affect the in-language quality of the word vectors? Do the multilingual embeddings allow an NLP model such as sentiment to generalize from one language to another?*

Before we present our results we briefly describe our data and systems. Recall that our approach employs a set of human provided concept dictionaries $\mathcal{C}$ that translate words with similar meaning

Method	Word-pair cos.	En Analogy	Fr Analogy	Mixed En+Fr Analogy
no const	0.286	77.5%	47.8%	39.3%
with const	0.422	52.8	40.4	43.6
ACS	0.439	66.9	53.3	52.6
Monolingual	-0.015	81.1	45.2	NA

Table 2: Comparison of multilingual embeddings.

	(Twitter) % of target-lang. baseline						(All) % of target-lang. baseline					
Embedding	mean	fr	es	de	no	en	mean	fr	es	de	no	en
None	60.6	53.1	59.3	68.5	–	61.6	63.4	57.3	60.4	63.3	87.9	48.2
No Const.	76.0	79.0	**77.4**	**80.9**	–	66.5	70.3	67.8	69.1	62.7	**94.4**	57.6
W/ Const.	76.9	78.4	71.4	76.2	–	**81.6**	**73.7**	**71.4**	67.8	**68.8**	89.0	**71.7**
ACS	**79.6**	**80.5**	76.6	79.5	–	**81.6**	73.6	65.1	**78.4**	66.7	88.8	69.2

Table 3: Cross-lingual sentiment analysis: trained on English; tested on target languages. English results are using model trained on French. Mean is computed across languages for each system.

from one language to another. Such dictionaries are readily available, and for the purpose of these experiments we use OmegaWiki,[1] a community based effort to provide definitions and translations for every language in the world. Only a tiny fraction of our full vocabulary (millions of words) is represented in the concept sets. For training embedding models we use Wikipedia (WP) documents from five different languages (see Table 1). Our multilingual corpus comprises all documents from all five languages. We train three different models on this corpus, all with CBOW [15] as the underlying embedding method: artificial code switching (**ACS**), vanilla CBOW with no constraints (**no const**) and CBOW augmented with hard constraints derived from the dictionary (**with const**). We also train monolingual embeddings with CBOW on each language independently (**monolingual**).

First, we study the quality of the embedding systems via two tasks. In the first we evaluate the accuracy of each model in answering multilingual word analogies of the form *homme:roi::woman:queen* (last column of Table 2). Interestingly, CBOW with no constraints learns a surprising amount of structure, but the two approaches that incorporate dictionaries, and in particular, ACS, have much higher accuracy. Second, we study how well the multilingual embeddings generalize from the small set of seed translation concepts to the rest of the vocabulary. For this, we split OmegaWiki into two sets and use the first set for training the embeddings and the second set for evaluation. We report the average cosine similarity of the evaluation set in the second column of Table 2. Again, CBOW alone learns a surprising amount of structure (0.286 cosine similarity), but ACS again improves the quality further (0.439). We also evaluate the models on English and French monolingual word analogies to see if learning a shared multilingual space affects quality of individual languages. Interestingly, for English, the accuracy of the multilingual embedding is lower than the English-only embedding, but in French, the opposite is true. It appears that lower resource language embeddings benefit from the combined multilingual space, though, at the expense of higher resource languages.

Second, we study the utility of our multilingual embeddings on cross-lingual sentiment analysis in which a model trained on a source language must generalize to a target language. In order to establish in-language baselines for the target languages, we train sentiment classifiers on each of the target language's training data, using unigrams, bigrams, and bias, but no embeddings as features (termed *target language baselines*; we present their accuracy in the last two columns of Table 1). The ultimate goal is to achieve accuracy similar to these systems, but without in-language training data. Our cross-lingual sentiment models use only the embeddings as features: for each document we first normalize each word's vectors, average them together, and use each dimension of the average as a feature. In Table 3, we report the percentage of the respective target-language baseline system's accuracy achieved by each cross-lingual system. The "mean" columns are averaged across all five languages. Note that while ACS does not always perform best, it performs well on average, especially for Twitter subset of the data in which the short documents have high information content.

Discussion We have presented a model for learning multilingual embeddings that does not require parallel corpora or language id. We demonstrate that our approach, based on code-switching, learns high quality multilingual embeddings, but sometimes at the expense of monolingual embedding quality. Further, the embedding allows sentiment analysis models to generalize across languages.

[1] `http://www.omegawiki.org`

References

[1] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

[2] Peter F. Brown, Peter V. deSouza, and Robert L. Mercer. Class-based n-gram models of natural language. In *ACL*, 1992.

[3] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, November 2011.

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

[5] M. Faruqui and C. Syer. Retrofitting word vectors to semantic lexicons. In *EACL*, 2014.

[6] J.R. Firth. Synopsis of linguistic theory. *Studies in Linguistic Analysis*, 1957.

[7] J. Gao, X. He, W. Yih, and L. Deng. Learning continuous phrase representations for translation modeling. In *Proceedings of Association for Computational Linguistics*, 2014.

[8] Stephan Gouws and Anders Søgaard. Simple task-specific bilingual word embeddings. In *NAACL*, pages 1386–1390, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[9] Z.S. Harris. Distributional structure. *Word*, 1954.

[10] S. Hassan, C. Banea, and R. Mihalcea. Measuring semantic relatedness using multilingual representations. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, SemEval '12, pages 20–29. Association for Computational Linguistics, 2012.

[11] Karl Moritz Hermann and Phil Blunsom. Multilingual distributed representations without word alignment. *arXiv preprint arXiv:1312.6173*, 2013.

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

[13] O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185, 2014.

[14] J. Lipski. Code-switching and the problem of bilingual competence. *Aspects of bilingualism*, 250:264, 1978.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *ICLR 2013, Workshop track*, 2013.

[16] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Empiricial Methods in Natural Language Processing (EMNLP 2014)*, volume 12, 2014.

[17] A.P. Sarath Chandar, S. Lauly, H. Larochelle, M. Khapra, B. Ravindran, V.C. Raykar, and A. Saha. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems*, pages 1853–1861, 2014.

[18] Rocher Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, 2013.

[19] I. Sutskever, O. Vinyals, and Q. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014.

[20] Y. Tsvetkov, W. Ammar, and C. Dyer. Constraint-based models of lexical borrowing. In *NAACL*, 2015.

[21] J. Turian, L. Ratinov, and Y. Bengio. Word representations: A simple and general method for semi-supervised learning. In *ACL*, 2010.