Simple task-specific bilingual word embeddings*

Stephan Gouws
Stellenbosch University
stephan@ml.sun.ac.za

Anders Søgaard
University of Copenhagen
soegaard@hum.ku.dk

Abstract

We introduce a simple wrapper method that uses off-the-shelf word embedding algorithms to learn task-specific bilingual word embeddings. We use a small dictionary of easily-obtainable task-specific word equivalence classes to produce mixed context-target pairs that we use to train off-the-shelf embedding models. Our model has the advantage that it (a) is independent of the choice of embedding algorithm, (b) does not require parallel data, and (c) can be adapted to specific tasks by re-defining the equivalence classes. We show how our method outperforms off-the-shelf bilingual embeddings on the task of unsupervised cross-language part-of-speech (POS) tagging, as well as on the task of semi-supervised cross-language super sense (SuS) tagging.

1 Introduction

Using multi-layered neural networks to learn word embeddings has become standard in NLP (Turian et al., 2010; Guo et al., 2014). While there is still some controversy whether such methods are superior to older methods (Levy and Goldberg, 2014; Baroni et al., 2014), there is little doubt that continuous word representations can potentially solve some of the data sparsity problems inherent in NLP.

Most research on word embeddings has focused on learning representations for the words in a single language, making syntactically or semantically similar words appear close in the embedding space. Embeddings have been applied to many tasks, from named entity recognition (Turian et al., 2010) to dependency parsing (Bansal et al., 2014). It has furthermore been shown that weakly supervised embedding algorithms can also lead to huge improvements for tasks like sentiment analysis (Tang et al., 2014). In this work, we also use weak or distant supervision, relying on small dictionary seeds.

This paper, however, considers the problem of learning *bilingual* word embeddings, i.e., word embeddings such that similar words in two different languages end up close in the embedding space. Such bilingual word embeddings can potentially be used for better cross-language transfer of NLP models, as we show in this paper. Previous work on bilingual word embeddings have defined similar words as translation equivalents and evaluated embeddings in the context of document classification tasks (Klementiev et al., 2012; Kocisky et al., 2014). In this paper, we present a simple wrapper method to existing monolingual word embedding algorithms that can learn task-specific bilingual embeddings, e.g., for POS tagging, named entity recognition, or sentiment analysis. Our algorithm is simpler and performs better on the tasks where we could compare performance to existing algorithms. Also, we note that our approach, unlike existing algorithms (Klementiev et al., 2012), is as fast as learning monolingual embeddings.

Our contributions In this paper we introduce a new approach for learning bilingual word embeddings and revisit the task of unsupervised cross-language POS tagging (Das and Petrov, 2011). Our bilingual embedding model, which we call *Bilingual Adaptive Reshuffling with Individual Stochastic Alternatives* (BARISTA), takes two (non-parallel) corpora and a small dictionary as input. The dictio-

*The authors contributed equally to this work. The second author is funded by the ERC Starting Grant LOWLANDS No. 313695.

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1386–1390,
Denver, Colorado, May 31 – June 5, 2015.

nary is essentially a list of words in the two languages that are equivalent with respect to some task, e.g., English *car* and French *maison* ('house') are both nouns, and hence "equivalent" in POS tagging; English *clerk* and *chauffeur* are both persons, and hence "equivalent" in SuS tagging; *house* and *maison* are equivalent in machine translation. BARISTA has the advantage that it (a) is independent of the choice of embedding algorithm, (b) does not require parallel data, and (c) can be adapted to specific tasks by using appropriate dictionaries. We use the bilingual embeddings directly to train a target language POS tagger on source language training data. Instead of lexical features, we use the bilingual embeddings. We show our bilingual embedding method outperforms using off-the-shelf bilingual embeddings on this task, and that our system is competitive to state-of-the-art approaches for cross-language POS tagging. Finally, we show that the same embeddings also lead to significantly better performance in semi-supervised cross-language SuS tagging. The code will be made publicly available at `https://github.com/gouwsmeister/barista`.

2 Our approach

Standard monolingual neural language models are unsupervised models that train on raw text, learning word features that enable the model to predict the next word (the target) from a sequence of words (the context). In the process, the model learns to cluster words into soft equivalence classes (words that have similar distributions).

Several authors have proposed bilingual clustering and embedding algorithms based on parallel data (Täckström et al., 2012; Klementiev et al., 2012; Zou et al., 2013; Kocisky et al., 2014; Hermann and Blunsom, 2014b). These authors have all evaluated their embeddings on document classification and machine translation, and not yet structured prediction tasks like POS/SuS tagging or syntactic parsing. A notable exception is Hermann and Blunsom (2014a), who do not rely on parallel data and do not use word alignments, but they still use comparable data and sentence alignments, and they only evaluate their embeddings in document classification.

The assumption that large amounts of parallel data exists for a language pair of interest is sometimes too strong (Hermann and Blunsom, 2014a). On the other hand, we often have access to small samples of near-equivalences from knowledge bases of various forms. For example, for POS tagging, we often have access to small-to-sizeable crowd-sourced tag dictionaries (e.g. Wiktionary). For SuS tagging, which is the other example considered in this paper, we sometimes have access to WordNets or similar resources. If we have such resources for both source and target language, we can extract word-equivalences from them and use these to learn bilingual embeddings using our proposed method.

In this paper, we experiment with using *both* equivalences based on word alignments (e.g., *house* $\sim$ *maison*), and equivalences based on knowledge bases (e.g., *car* $\sim$ *maison*). Crucially, our approach to learning bilingual embeddings only assumes a small seed of equivalences, no parallel data. It then uses these to produce a set of mixed context-target pairs.

Our input is a source corpus C_s and a target corpus C_t, as well as a set of bilingual equivalences $R_\sim$. We begin by shuffling the concatenation of C_s and C_t. We then pass over this mixed corpus, and for each word w, if $\{w' \mid w, w' \in R_\sim\}$ is non-empty and of cardinality k, i.e., w is in the seed list of equivalences, we replace w with w' with probability $1/2k$. In other words, we flip a coin whether to replace w, and then randomly choose one of its equivalences as our replacement. For example, using translation equivalence classes, one could generate any of the following mixed texts from the English sentence *build the house*: *construire the house*, *build la maison*, *build the maison*, etc., or any other combination of English and French words with the English word order. With POS equivalence classes, any of the words in *build the house* can be replaced with words with overlapping syntactic categories, e.g., *build the voiture*.

3 Experiments

In our experiments we balance the source and target corpora, by subsampling from the bigger corpus. The vocabularies for all models are kept unrestricted, and result in around $1M$ words per language pair. We train with a window of 4 words on either side of the target word, using linear discounting of the initial learning rate of 0.1. These parameters were set on the Spanish POS data (see §3.2). We

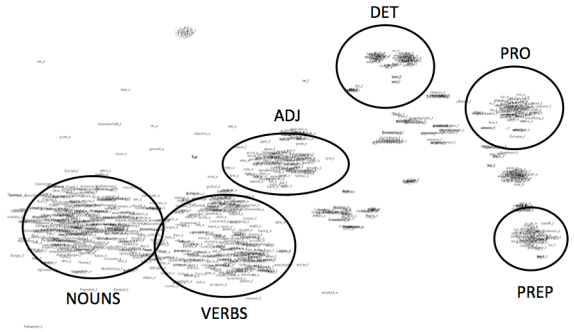

Figure 1: t-SNE visualization of BARISTA embeddings trained with POS classes.

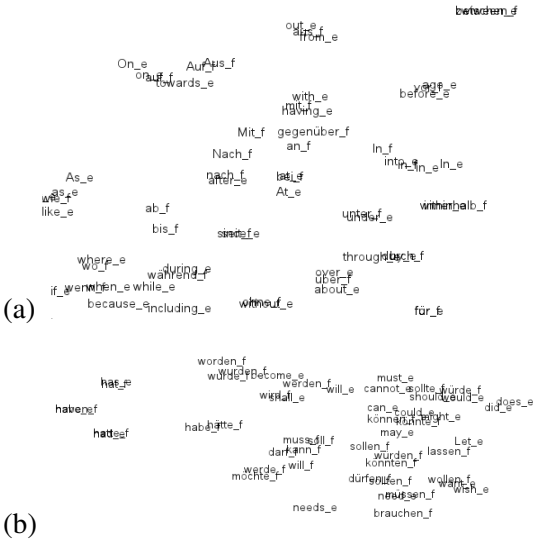

Figure 2: t-SNE visualizations of BARISTA embedding subspaces for (a) prepositions and (b) modals.

use the CBOW model in `word2vec` for training the word embeddings.[1]

Both our POS tagging evaluation datasets, as well as the Wiktionaries,[2] are mapped to Google's universal tagset (Petrov et al., 2011). We use the derived dictionaries for extracting bilingual POS equivalence classes. For SuS tagging, we only consider English-Danish and extract equivalence classes from Princeton WordNet and DanNet. For translation equivalents, we made use of Google Translate.

3.1 Qualitative Evaluation

In our main experiments (§3.2–3.3), we use data from Wikipedia, but we first present a qualitative evaluation of English-German bilingual embeddings learned from the smaller Europarl corpus.[3] Note that while this is parallel data, we do not exploit its parallel nature.

POS classes The embeddings learned from English-German Europarl using POS classes from Wiktionary were visualized using the t-SNE technique (Van der Maaten and Hinton, 2008), and are shown in Fig 1. The model learns to cluster words very distinctively by their POS tag. Monolingual word embedding models with short context windows typically cluster by POS, but here we see the same effect for bilingual embeddings, i.e. words from **both** languages with the same POS tag cluster together. Individual words, on the other hand, do not appear to retain the fine-grained relationships we normally observe in word embeddings (where similar words cluster closer together). The question thus is whether such embeddings are more or less useful for cross-language POS tagging than bilingual embeddings based on translation equivalencies.

Translation classes Next, we induced English-German bilingual embeddings on Europarl using translations obtained from Google Translate. We derived a dictionary of the top $20k$ most frequent words in English, translated into German. The embeddings are shown in Fig 2. The visualizations show that the models are able to extract very fine-grained bilingual relationships, and some clusters still correspond to POS.

3.2 Cross-language part-of-speech tagging

Next we evaluated the embeddings in the context of unsupervised cross-language POS tagging (Das and Petrov, 2011). The goal is to train a tagger – in our case, we use SEARN (Daume et al., 2009), following (Johannsen et al., 2014)– on labeled English data decorated with bilingual embeddings, and then evaluate the model on another target language. We use data from Danish, German, Spanish, Italian, Dutch, Portuguese, and Swedish. Training and test data, which are the same used in Das and Petrov (2011), were converted to use the same 12 universal parts of speech proposed by Petrov et al. (2011).

All results in this section were obtained by training bilingual embedding models on the publicly-

[1] `https://code.google.com/p/word2vec/`

[2] `https://code.google.com/p/wikily-supervised-pos-tagger/`

[3] http://www.statmt.org/europarl/

Language	TC-Perc	Random	Klmtv	POS-50	POS-300	Tr-50	Tr-300	DP	B-K
Spanish	80.6	81.8	79.8	82.4	81	81.6	82.6	**84.2**	80.2
German	80.4	82.7	82.8	81.8	84.1	82.6	**84.8**	82.8	81.3
Danish	63	68.9	-	68.9	72.4	71.8	78.4	**83.2**	69.1
Swedish	71.6	73.7	-	75	76	75.4	77.5	**80.5**	70.1
Italian	80.1	81.3	-	82.1	80.9	82.1	80.7	**86.8**	68.1
Dutch	74.5	77.2	-	78.3	77.4	78.7	**80.3**	79.5	65.1
Portuguese	76.9	78.1	-	77.3	76.1	80.6	80.5	**87.9**	78.4
Avg	75.3	77.7	-	78	78.3	79	80.7	**83.6**	73.2

Table 1: Cross-language POS tagging. TC-Perc: type-constrained structured perceptron.

available, pre-tokenized versions of Wikipedia,[4] and then using the trained embeddings as features in a publicly available implementation of the structured perceptron.[5] We use ortographic features, as well as the embedding vector of the target word. In addition, we use type constraints from Wiktionary to prune the search lattice during decoding (Täckström et al., 2013). We scaled the embedding features in the same way as Turian et al. (2010) with scaling parameter 0.01 (set on Spanish POS data).

Our baseline method is a type-constrained structured perceptron with only ortographic features, which are expected to transfer across languages. We also experiment with using *random* embeddings, as well as the embeddings provided by Klementiev et al. (2012) (Klmtv).[6] Our results are displayed in Table 1. POS-X refer to X-dimensional BARISTA embeddings trained with POS equivalence classes, and Tr-X is X-dimensional BARISTA embeddings trained using translation equivalence classes. We note that random embeddings improve over our baseline, suggesting that the random features act as regularizers. The embeddings provided by Klementiev et al. (2012) seem to lead to worse performance than random embeddings, presumably because they capture mostly semantic (topic) similarity. We also compare our results to those reported by Berg-Kirkpatrick et al. (2010) (B-K) and Das and Petrov (2011) (DP), but note that their approaches require in-sample unlabeled data, and in the latter case, also parallel bilingual data.

While training with POS classes improves over the random baseline, training with translation equivalence classes gives even better performance. For both approaches, using more embedding features improves the performance (500 dimensions did not improve significantly over 300). Our model is generally better than Berg-Kirkpatrick et al. (2010) – but worse than Das and Petrov (2011).

3.3 Cross-language super sense tagging

Finally, we also tried using the BARISTA-embeddings for English-Danish (with parameters still set on Spanish POS) on *another* task, namely super sense tagging (Ciaramita and Altun, 2006). We train a system on a mixture of 1000 randomly sampled sentences from English SemCor[7]and 320 labeled Danish sentences (see below) and compare using bilingual embeddings trained with equivalence classes from English and Danish wordnets (WN-300), to embeddings trained using translation equivalence classes (Tr-300). We use 300 dimensional embeddings. We use a POS-sensitive most frequent sense baseline (MFS), as well as structured perceptron model trained only with ortographic and POS features, as well as MFS features (Johannsen et al., 2014). Our metric is a weighted average over F_1-scores for the (41) semantic classes. Note that using the knowledge base is superior to using translation equivalences, but both embeddings are superior to both our baselines. The Danish training (newswire only) and test data (six different domains) – is made publicly available.[8]

[4] https://sites.google.com/site/rmyeid/projects/polyglot

[5] https://github.com/coastalcph/rungsted

[6] http://people.mmci.uni-saarland.de/~aklement/data/distrib/

[7] http://web.eecs.umich.edu/~mihalcea/downloads.html#semcor

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

	Baselines		BARISTA	
	MFS	TC-Searn	Tr-300	WN-300
blogs	49.1	46.7	50.5	**61.9**
forum	44.5	41.2	45.0	**53.9**
magazine	46.5	45.2	50.4	**51.5**
newswire	48.4	45.4	52.7	**60.9**
reviews	48.4	44.9	50.4	**55.8**
speech	51.1	48.4	51.5	**58.4**
Avg	48.1	45.4	50.5	**58.3**

Table 2: Cross-language SuS tagging.

4 Conclusions

We presented a simple approach, BARISTA, to learning bilingual embeddings. BARISTA has the advantages that it (a) is independent of the choice of embedding algorithm, (b) does not require parallel data, and (c) can be adapted to specific tasks by using appropriate dictionaries. Our embeddings proved useful for cross-language POS/SuS tagging.

References

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

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

Taylor Berg-Kirkpatrick, Alexandre Cote, John DeNero, and Dan Klein. 2010. Painless unsupervised learning with features. In *NAACL*.

Massimiliano Ciaramita and Yasemin Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proc. of EMNLP*, pages 594–602, Sydney, Australia, July.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *ACL*.

Hal Daume, John Langford, and Daniel Marcu. 2009. Search-based structured prediction. *Machine Learning*, 75:297–325.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Revisiting embedding features for simple semi-supervised learning. In *EMNLP*.

Karl Hermann and Phil Blunsom. 2014a. Multilingual distributed representations without word alignment. In *ICLR*.

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

Anders Johannsen, Dirk Hovy, Hector Martinez Alonso, and Anders Søgaard. 2014. More or less supervised super-sense tagging of twitter. In **SEM*.

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

Tomas Kocisky, Moritz Hermann, and Phil Blunsom. 2014. Learning bilingual word representations by marginalizing alignments. In *ACL*.

Omer Levy and Yoav Goldberg. 2014. Neural word embeddings as implicit matrix factorization. In *NIPS*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. CoRR abs/1104.2086.

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

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *TACL*, 1:1–12.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for Twitter sentiment classification. In *ACL*.

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

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85.

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