Language adaptation experiments via cross-lingual embeddings for related languages

Serge Sharoff
Centre for Translation Studies,
University of Leeds, LS2 9JT, UK
s.sharoff@leeds.ac.uk

Abstract

Language Adaptation (similarly to Domain Adaptation) is a general approach to extend existing resources from a better resourced language (donor) to a lesser resourced one (recipient) by exploiting the lexical and grammatical similarity between them when the two languages are related. The current study improves the state of the art in cross-lingual word embeddings by considering the impact of orthographic similarity between cognates. In particular, the use of the Weighted Levenshtein Distance combined with orthogonalisation of the translation matrix and generalised correction for hubness can considerably improve the state of the art in induction of bilingual lexicons. In addition to intrinsic evaluation in the bilingual lexicon induction task, the paper reports extrinsic evaluation of the cross-lingual embeddings via their application to the Named-Entity Recognition task across Slavonic languages. The tools and the aligned word embedding spaces for the Romance and Slavonic language families have been released.

Keywords: Word embeddings, Related languages, Cognate words, Comparable corpora

1. Introduction

Parallel corpora play an important role in many multilingual NLP applications, such as Machine Translation, Cross-Lingual Text Classification or Information Retrieval. However, the topics and genres of parallel corpora are limited even for better resourced languages, e.g., resources are scarcer outside of the official documents of Europarl and the United Nations (Koehn, 2005; Eisele and Chen, 2010). Also, even if each individual language has reasonably good parallel resources, such as Polish and Russian aligned with English, it is difficult to find a large parallel corpus, which contains this specific, e.g., Polish-Russian, language pair.

Monolingual corpora can be substantially bigger and more varied in comparison to parallel ones. Comparable corpora of different levels of comparability (Sharoff et al., 2013) can be used for induction of bilingual lexicons from small seed dictionaries. The present paper follows an influential study (Mikolov et al., 2013), which presented a method for building *multilingual* embedding spaces. In addition to a model with a seed bilingual dictionary, it also introduced constraints on what its authors call “morphological structure” (actually the Levenshtein Distance) for keeping only the cognate words in the output.

However, further work on bilingual lexicon induction did not include the use of cognates, especially in the context of related languages. The importance of utilising links between related languages can be illustrated by the use of Machine Translation via a pivot language. A simple dictionary transfer from Ukrainian into Russian followed by MT for the better resourced Russian-English pair easily beats MT translating from Ukrainian directly into English using far smaller resources (Babych et al., 2007). Overall, many lesser resourced languages can benefit from Language Adaptation by applying the models developed for the better resourced ones.

The present study advances the state of the art by combining existing techniques of building cross-lingual embedding from comparable corpora with the Weighted Levenshtein distance, when the weights are themselves obtained from the seed dictionaries, see Section 3. In addition to intrinsic evaluation of the parameters of bilingual lexicon induction, the quality of cross-lingual embeddings can be measured extrinsically through accuracy of their use in downstream tasks, in particular, Named Entity Recognition, see Section 4. Both intrinsic and extrinsic evaluations show considerable improvements from the use of Language Adaptation.

2. Related studies

Starting from earlier work on Neural Language Models, a common way of representing word meanings is via word embeddings built from predictions of word neighbours using neural networks (Bengio et al., 2003; Mikolov, 2012). Recently, the Facebook group developed FastText, an updated method for producing monolingual embeddings by using information from character ngrams (Mikolov et al., 2017), i.e., a word embedding vector is:

$$v(w) = w2v(w) + \frac{1}{|\mathcal{N}|} \sum_{n \in \mathcal{N}} x_n \qquad (1)$$

where $w2v$ is the standard word embedding of w (using the skip-gram model), while $\mathcal{N}$ is the set of ngrams derived from this word, x_n are their respective embeddings.

Studies in extraction of bilingual lexicons from comparable corpora can be traced back to at least (Fung, 1995; Rapp, 1995), who described words via a vector of their collocates, translated some words using a seed dictionary and compared the vectors across the languages. Word embeddings offer a better way of building word vectors in comparison to the vectors of collocate counts (Baroni et al., 2014). Word embeddings across languages have been studied since (Klementiev et al., 2012). A seminal study, which transformed the field, was (Mikolov et al., 2013), which used a translation matrix (TM) trained on a seed bilingual dictionary to convert monolingual word embeddings into a shared space. That study was followed by other studies aimed at

Table 1: Alignments from Wikipedia for titles and words

Polish	Russian	English
Z życia marionetek	Из жизни марионеток	From the Life of the Marionettes
Wskaźnik jakości życia	Индекс качества жизни	Quality-of-life index

Character alignment for words:	m	a	r	i	o	n	e	t	e	k	ż	y	c	i	a
	м	а	р	и	о	н	е	т	о	к	ж	и	з	н	и

improving the process of TM production, e.g., via Canonical Correspondence Analysis (Faruqui and Dyer, 2014), Global Correction (Dinu et al., 2014) or TM orthogonalisation (Artetxe et al., 2016).

A traditionally accepted model for this task is based on constructing a linear transformation matrix $\mathbf{W}$ by minimising the following objective:

$$\min_{\mathbf{W}} \sum ||\mathbf{W}e_i - f_i||^2 \tag{2}$$

where e_i and f_i are the respective embedding vectors in the two languages, which are supposed to be translations of each other according to the training set. The differences between the approaches are primarily in the method for building $\mathbf{W}$, e.g., by stochastic gradient descent (Mikolov et al., 2013), CCA (Faruqui and Dyer, 2014), multivariate regression (Dinu et al., 2014) or matrices from the SVD transform (Artetxe et al., 2016). The latter model ensures that $\mathbf{W}$ is an orthogonal matrix built using a closed form solution:

$$\mathbf{W} = \mathbf{V} \times \mathbf{U}^T \tag{3}$$

when $\mathbf{V}$ and $\mathbf{U}$ are the matrices from the SVD factorisation of $\mathbf{F} \times \mathbf{E}^T$, see (Artetxe et al., 2016) for justification and discussion.

The feature spaces with large number of dimensions also demonstrate a phenomenon of *hubness* (Radovanović et al., 2010), i.e., some vectors happen to be in close proximity to many other vectors. This makes them more common choices in the lexical retrieval tasks leading to a larger number of errors. Formally, a word w is mapped to a set of words $\mathcal{N}_k(w)$ for which this word is within their k nearest neighbours. Words with the largest $|\mathcal{N}_k(w)|$ are (typically unwanted) hubs. Often such words have restricted context of their use, e.g., *troops* (183), *retreated* (176), *cavalry* (156) are such hubs in the FastText English space induced from Wikipedia (the numbers in brackets refer to their $|\mathcal{N}_{20}|$ hubness index, i.e., there are 183 words for which the word *troops* is in the list of their 20 closest neighbours), while the median hubness index on the English Wikipedia is 5. Dinu et al. (2014) observe that hubness becomes more pronounced after linear transformation, since the objective for building the transformation matrix $\mathbf{W}$ leads to lower variance of the transformed vectors, which in turn means that the vectors (on average) are closer to each other (Dinu et al., 2014). They suggest a way of mitigating hubness by using Global Correction (GC), i.e., by downgrading the similarity ranks for the items proportionally to their hubness index.

The initial TM study (Mikolov et al., 2013) did suggest the use of the Levenshtein Distance (LD), as a filtering step. Similarly, filtering of cross-lingual embedding spaces via LD for the purposes of Statistical Machine Translation between related languages has been explored in (Rios and Sharoff, 2015). A manually developed set of rules for a Finite State Transducer (FST) was used for identification of cognates and borrowings in (Tsvetkov and Dyer, 2016). However, post hoc filtering improves precision at the expense of reduced coverage. The method suggested below operates at the ranking stage, while it also uses the Weighted Levenshtein Distance (WLD), a simpler alternative to FSTs.

3. Dictionary induction using cognates

3.1. Cross-lingual mapping

The method for cross-lingual mapping across related languages in this study consists of three steps:

1. automated collection of seed bilingual dictionaries;
2. determining weights for the Levenshtein distance from the seed dictionaries;
3. alignment of monolingual embeddings by linear transformation using global correction and weighted LD;

In a low resource setting, the seed dictionaries for related languages can be obtained from the titles of interlinked Wikipedia articles in two languages (iWiki links),[1] see examples of aligned titles in Table 1. This helps in modelling scenarios when few parallel texts are available, e.g., for the Polish-Russian pair (Polish is included in Europarl, Russian is in the UN corpus, but very few reliable resources are available for the Polish-Russian pair). The titles have been word-aligned using FastAlign (Dyer et al., 2013). The resulting word-level dictionaries have been filtered against the respective frequency lists, since the Wikipedia titles are dominated by relatively infrequent proper names.

In addition to providing the training lexicon, a seed dictionary can also be used to provide a character-level model for matching the cognates, see the part of Table 1 for examples of character alignment. The pairs of words from the training dictionary have been aligned on the character level (again using FastAlign in this study) to produce the probabilities of regular correspondences between the characters in the two languages. This character alignment model is particularly important when the two languages use different character sets, such as the case for Polish and Russian. For example, the characters with the highest probability for translating the Russian characters ф and л into Polish are respectively *f* and *ł*.

In the end the standard edit operations for computing the traditional normalised Levenshtein Distance can be

[1] github.com/clab/wikipedia-parallel-titles

weighted by the probabilities of their character-level alignments:

$$WLD(s_e, s_f) = 1 - \frac{\sum_{(e,f)\in al(s_e,s_f)}(1 - p(f|e))}{\max(len(s_e), len(s_f))} \quad (4)$$

where s_e and s_f are words in the two languages, al is a set of their alignments, $p(f|e)$ is the probability from the character alignment model. The distance is normalised by the length of the longest word. For convenience in comparing it with the cosine similarity, the value is flipped to represent greater similarity with larger values.

Given that even correctly aligned words from the Wikipedia titles for related languages are not necessarily cognates e.g., *wskaźnik* vs индекс ('index') from Table 1, the process of getting the Levenshtein weights ran in two steps. In the first step, an initial estimate of the probabilities for characters was produced from *all* words in the seed dictionary. This was used for assessing the rough WLD between them. The most likely cognates according to this rough WLD were used as the input for the second iteration of character-level alignments. The WLD threshold for choosing the most likely cognates was determined for each language pair individually. Repeated application of these steps did not result in any improvements in detecting cognates.

The value of either LD or WLD can be used as a factor for scoring the translation suggestions:

$$score(s_e, s_f) = \alpha \cos(v_e, v_f) + (1-\alpha)WLD(s_e, s_f) \quad (5)$$

where v_e and v_f are vectors for respectively s_e and s_f in the cross-lingual embedding space, while α is the relative weight of the cosine similarity.

While the combined score is useful for producing bilingual dictionaries, it does not affect the bilingual embedding space by itself. A closed form solution for orthogonalisation as used in (3) helps in improving alignment quality in the general case, but it does not allow weight adjustment by taking into account the similarity between the cognates. An easy way for incorporating this information into the cross-lingual embedding space is by aligning the entire lexicons from the cross-lingual space using the WLD score from (5) and selecting the most similar words in this list. This far longer lexicon can be used instead of the seed dictionary for producing a new weight matrix from (3) for re-alignment of the already aligned cross-lingual space from the previous step. The rationale for this iteration is that we want to minimise the distance between the known cognates while preserving the orthogonality of the weight matrix. Again, while repeated application of these steps is possible, it did not produce better results, so the tables below present the results obtained after two iterations.

3.2. Experimental setup

This paper reports two sets of experiments. One experiment involved a replicable setting for the English-Italian language pair with the standardised embeddings and training / test dictionaries initially developed for (Dinu et al., 2014) and used in (Artetxe et al., 2016). Even though English and Italian are not closely related languages (English is a Germanic language, Italian is from the Romance family), a large number of English words are borrowings from Romance languages, primarily from French and Latin, so the WLD approach could work for the En-It pair as well. The test dictionary from (Dinu et al., 2014) includes both cognate word pairs, such as *academy / accademia*, and non-cognate pairs, such as *absolve / esimere* or *abysmally / malo*, which are also often questionable translation equivalents. Therefore, a cognate-only version of the En-It test set was produced by retaining only the words with the WLD value above 0.5, reducing the En-It test dictionary from 1869 down to 818 entries.

Table 2: Prec@1 for En-It

W2V vectors from (Dinu, et al. 2014)	
TM (Mikolov et al., 2013)	0.349
CCA (Faruqui and Dyer, 2014)	0.378
Orth (Artetxe et al., 2016)	0.393
GC (Dinu et al., 2014)	0.377
GC+LD	0.501
GC+WLD	**0.531**
FT vectors from (Mikolov et al., 2017)	
FT+TM	0.461
FT+Orth	0.529
FT+GC	0.477
FT+GC+Orth+WLD	**0.616**
MUSE (Conneau et al., 2017)	0.683
FT vectors for cognates only	
FT+TM	0.550
FT+Orth	0.614
FT+GC	0.575
LD $\alpha = 0$	0.298
WLD $\alpha = 0$	0.339
FT+GC+Orth+WLD $\alpha = 0.5$	0.584
FT+GC+Orth+LD $\alpha = 0.73$	0.669
FT+GC+Orth+WLD $\alpha = 0.73$	**0.692**
MUSE	0.719

A new set of embeddings produced by FastText has been used in the English-Italian experiments (labelled as FT in Table 2) in addition to the standardised embeddings as used in (Dinu et al., 2014; Artetxe et al., 2016). The FT embeddings have been the basis for the experiments with the Slavonic languages.

The experiments with the Slavonic languages also emphasise the low-resource setting, when large parallel corpora for the seed dictionary are not always available, so the seed dictionaries for the Transformation Matrices and the WLD weights came from the iWiki links (the Italian seed dictionary used in (Dinu et al., 2014) and (Artetxe et al., 2016) was derived from aligning Europarl).

3.3. Experimental results

The results listed in Table 2 confirm that orthogonalisation (Artetxe et al., 2016) and global correction (Dinu et al., 2014) improve the accuracy of translation detection in comparison to the baseline of (Mikolov et al., 2013). Embedding vectors produced by incorporating subword information (marked by FT in Table 2) also make a considerable positive impact. Adding the constraint of having or-

Table 3: Dictionary induction results for Slavonic languages

Dictionary induction without WLD

	sl-hr	sl-cs	sl-pl	sl-ru	ru-uk	cs-sk
Prec@1:	0.429	0.611	0.584	0.566	0.929	0.814
Prec@10:	0.688	0.868	0.842	0.818	0.976	0.971
MUSE, Prec@1:			0.724		0.942	

Dictionary induction with WLD

	sl-hr	sl-cs	sl-pl	sl-ru	ru-uk	cs-sk
Prec@1:	0.840	0.763	0.751	0.662	0.945	0.910
Prec@10:	0.963	0.973	0.977	0.883	0.994	0.996

thographic cognates (LD or WLD) improves the accuracy of dictionary induction further, often by a substantial margin. Even for the English-Italian pair, where the languages operate over the same alphabet, WLD outperforms LD because it assigns a very low cost to more common substitutions, e.g., $x \rightarrow s$ or $j \rightarrow g$ (*examined* → *esaminato* or *Jerusalem* → *Gerusalemme*).

Cleaning the existing English-Italian test dictionary for cognates brings further improvement in P@1 to 0.692, so that the resulting dictionaries become acceptable for downstream tasks. The best value of α, the relative weight to balance the contribution between the cosine similarity and the Weighted Levenshtein Distance, was estimated at 0.73 using a development set which was randomly extracted from the training dictionary. Relying exclusively on the orthographic similarity ($\alpha = 0$) leads to relatively poor results.

Given that the FT+Orth+WLD combination results in consistently better performance, the results of dictionary induction across Slavonic languages are shown only for this setup (Table 3). Comparison of the Slavonic dictionaries to the English-Italian pair shows even more significant improvements through the use of WLD, occasionally from 0.429 to 0.840 for the Slovenian-Croatian pair.

FastText vectors of 300 dimensions built from Wikipedias for selected Balto-Slavonic languages (Belorussian, Czech, Croatian, Lithuanian, Polish, Slovak, Slovene, Ukrainian) have been transformed into a shared Panslavonic embedding space with extraction of the full set of possible cognate forms. For convenience of running cross-lingual experiments, English has also been added to the shared embedding space, even though it is not a related language.

If a reasonable monolingual corpus is available to train the embeddings for another Slavonic language, e.g., Rusyn or Sorbian, as well as a reliable dictionary between this language and one of the languages in the current Panslavonic space (the Wikipedia iWiki lists for such languages are too short to produce useful seed dictionaries), a new language can be easily added to this space.

4. Named Entity Recognition

4.1. Training setup

The cross-lingual shared space has been tested through the Named Entity Recognition (NER) task, which consists in detection and labelling of all occurrences of person names, organisations or locations. This is a convenient downstream task for which there are existing methods and test sets. Recently, various neural network approaches produced very convincing results for NER (Collobert et al., 2011). A particular implementation used in the extrinsic evaluation experiment reported below is based on a sequence tagging method, which combines bidirectional LSTM with CRF for making the final prediction (Lample et al., 2016). Each word is represented by its embedding vector from the shared embedding space, in addition to other universal features, such as character-level embeddings or the presence of capitalisation. The tagger was trained on an existing NER-annotated corpus from (Krek et al., 2012) (in Slovenian) with addition of small samples in Croatian, Czech, Polish, Russian and Ukrainian in order to provide at least some information for the character-level embeddings. The samples were derived from the titles of Wikipedia articles in the respective languages for categories matching such patterns as ‘Births’ (for person names), ‘Organisations’ and ‘Countries’ or ‘Villages’ (for the lack of a more generic category of locations in Wikipedia).

4.2. BSNLP NER shared task

The NER shared task at BSNLP’17 contained two separate test sets with no training sets for individual languages. One test set was based on the European Commission reports, another one on news wires concerning Donald Trump. The baseline system (Piskorski et al., 2017) was based on large gazetteers developed by the JRC, while the only other submission covering all Slavonic languages (Mayfield et al., 2017) was based on projection of labels via word-aligned parallel corpora, see Table 4.

The shared embedding space is surprisingly efficient. The Slovenian space was used for training, so it provides the upper baseline for adaptation. Czech, Croatian and Polish are sufficiently similar typologically, so the accuracy on those languages is slightly below what has been achieved for Slovenian. Russian and Ukrainian are East Slavonic languages, further away typologically from the rest, which is probably the main reason for the markedly lower accuracy of adaptation of the Slovenian training set. Across all languages, the NER tagger has a problem with detecting relatively long NERs, which are common in the EC test set, such as *The European Convention for the Protection of Human Rights and Fundamental Freedoms*, while the accuracy is higher on the general newswire texts. Overall, the results are considerably lower than what has been achieved for English, which can be explained by much richer morphology of the Slavonic languages, as well as by a smaller training set.

Table 4: NER recognition results

	cs	hr	pl	ru	sl	uk	
EC news:	47.2	**46.2**	**44.8**	46.5	47.8	10.8	JHU
	41.2	30.0	34.6	**53.7**	37.5	**20.8**	JRC
	47.7	44.3	44.2	33.6	**59.5**	13.7	Sharoff
	cs	hr	pl	ru	sl	uk	
Trump:	46.1	50.4	41.0	41.8	46.2	33.2	JHU
	42.2	37.4	48.0	**55.6**	44.2	**50.8**	JRC
	52.6	**52.4**	**55.2**	21.0	**62.6**	20.7	Sharoff

5. Conclusions

The experiments reported in the paper showed that adding the WLD constraints on aligning word forms is a very efficient way for building cross-lingual embedding spaces and for extracting bilingual dictionaries for related languages. For lesser resourced languages and language pairs, the procedure can rely on readily available Wikipedia corpora and the respective iWiki links. Incorporating the resulting multilingual embedding spaces into downstream tasks, such as NER, is also efficient. The results are competitive with the more commonly used projection methods, which are based on parallel corpora (primarily Europarl, which limits the amount of language pairs). The lexicon induction scripts, the shared cross-lingual embeddings as well as the resulting Panslavonic NER taggers are available under permissive licenses.[2]

When the camera-ready copy of the present paper was ready for the final submission, I learned about MUSE (Conneau et al., 2017), a recently developed approach to producing cross-lingual embeddings. It relies on unsupervised alignment between the probability distributions in the two monolingual spaces using adversarial training: the task is to create a translation matrix which can confuse the discriminating function to distinguish two translations in the shared space. The experiments I was able to run on my data before submitting this paper are marked as MUSE in Tables 2 and 3. The MUSE method does not use information about cognates and it offers comparable or better performance in comparison to WLD. Given that MUSE is based on iterative updates instead of a closed form solution, one possible extension concerns integration of the WLD scoring function into MUSE to improve the accuracy across related languages even further.

Another important extension required for the model concerns reliable mapping across the full paradigm of related lexical items in the two languages. A single form in one language can correspond to a number of forms in another language, e.g., *adequate* in English maps to four cognate forms in Italian: *adeguato, adeguata, adeguate* and *adeguati*, corresponding to the choices of singular vs plural and feminine vs masculine, because the English adjectives do not inflect for number and gender. The potential for such one-to-many matches is smaller for closely related languages, since they usually have the same set of morphological categories. However, differences in the suppletivism of forms are common even across related languages, for example, the feminine adjectival forms ending with ой in Russian (e.g., новой, 'new') are used for any non-nominative case, while unique cognate forms are used in Ukrainian for each grammatical case, e.g., genitive: нової, dative: новій, instrumental: новою, etc. A related problem concerns representation of similarities on the level above words. For example, the meaning of the identical forms *postale* in both French and Italian is the same ('post.adj'), they share a number of collocates with the same meaning, e.g., *adresse postale* vs *indirizzo postale*, so they are likely to be well-aligned in the shared embedding space (either with or without WLD constraints). However, the French form is feminine, while the Italian one is masculine, so the correct embedding space should have mapped *postale* in Italian with *postal* in French.

Therefore, the cross-lingual embedding space needs to be built in a way which takes into account the similarity across the full set of forms with respect to their grammatical functions. Using lemmas only can work fine for remote languages, but this loses information about the correspondence of forms in related languages. One of the ways of achieving this is to follow the line of research by (Avraham and Goldberg, 2017). They suggest morphological decomposition of embeddings, which is similar to lexical decomposition of a word into character ngrams in the FastText model. An alternative approach could involve using classifiers to predict the morphological annotations from the embedding vectors in the monolingual space (Belinkov et al., 2017) in order to align embedding vectors by paying attention to the similarity of their morphological annotations.

References

Artetxe, M., Labaka, G., and Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proc EMNLP*, Austin, Texas, November.

Avraham, O. and Goldberg, Y. (2017). The interplay of semantics and morphology in word embeddings. In *Proc EACL*, pages 422–426, Valencia, Spain, April.

Babych, B., Hartley, A., and Sharoff, S. (2007). Translating from under-resourced languages: comparing direct transfer against pivot translation. In *Proc MT Summit XI*, pages 412–418, Copenhagen.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proc ACL*, Baltimore, Maryland, June.

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., and Glass, J. (2017). What do neural machine translation mod-

[2] `https://github.com/ssharoff/cognates`

els learn about morphology? In *Proc ACL*, Vancouver, Canada, July.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Dinu, G., Lazaridou, A., and Baroni, M. (2014). Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM Model 2. In *Proc NAACL*, Atlanta, Georgia, June.

Eisele, A. and Chen, Y. (2010). MultiUN: A multilingual corpus from United Nations documents. In *Proc Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Faruqui, M. and Dyer, C. (2014). Improving vector space word representations using multilingual correlation. In *Proc EACL*, pages 462–471, Gothenburg, Sweden, April.

Fung, P. (1995). Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In *Proc. Third Annual Workshop on Very Large Corpora*, pages 173–183, Boston, Massachusetts.

Klementiev, A., Titov, I., and Bhattarai, B. (2012). Inducing crosslingual distributed representations of words. In *Proc COLING*, Mumbai, India, December.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proc. MT Summit X*.

Krek, S., Erjavec, T., Dobrovoljc, K., Holz, N., Ledinek, N., and Može, S., (2012). *Učni korpus ssj500k kot podatkovna zbirka*.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *Proc NAACL*, pages 260–270, San Diego, California.

Mayfield, J., McNamee, P., and Costello, C. (2017). Language-independent named entity analysis using parallel projection and rule-based disambiguation. In *Proc BSNLP*, pages 92–96, Valencia, Spain, April.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2017). Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405*.

Mikolov, T. (2012). *Statistical language models based on neural networks*. Ph.D. thesis, Brno University of Technology.

Piskorski, J., Pivovarova, L., Šnajder, J., Steinberger, J., and Yangarber, R. (2017). The first cross-lingual challenge on recognition, normalization, and matching of named entities in slavic languages. In *Proc BSNLP*, pages 76–85, Valencia, Spain, April.

Radovanović, M., Nanopoulos, A., and Ivanović, M. (2010). Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531.

Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Proc. of the 33rd ACL*, pages 320–322, Cambridge, MA.

Rios, M. and Sharoff, S. (2015). Obtaining SMT dictionaries for related languages. In *Proc the Eighth Workshop on Building and Using Comparable Corpora*, pages 68–73, Beijing, China, July.

Sharoff, S., Rapp, R., and Zweigenbaum, P. (2013). Overviewing important aspects of the last twenty years of research in comparable corpora. In Serge Sharoff, et al., editors, *BUCC: Building and Using Comparable Corpora*, pages 1–17. Springer.

Tsvetkov, Y. and Dyer, C. (2016). Cross-lingual bridges with models of lexical borrowing. *JAIR*, 55:63–93.