Natural Language Engineering **26** (2): 163–184. Printed in the United Kingdom

Finding next of kin: Cross-lingual embedding spaces for related languages

Serge Sharoff
Centre for Translation Studies
University of Leeds

(*Received 15 October 2018; Revised version received 5 April 2019; Accepted 15 June 2019*)

Abstract

Some languages have very few NLP resources, while many of them are closely related to better resourced languages. This paper explores how the similarity between the languages can be utilised by porting resources from better to lesser resourced languages. The paper introduces a way of building a representation shared across related languages by combining cross-lingual embedding methods with a lexical similarity measure which is based on the Weighted Levenshtein Distance. One of the outcomes of the experiments is a Panslavonic embedding space for nine Balto-Slavonic languages. The paper demonstrates that the resulting embedding space helps in such applications as morphological prediction, Named Entity Recognition and genre classification.

1 Introduction

The total number of living languages in the world is estimated at more than 7,000 (Simons and Fennig, 2017). If we only include the top 100 languages with the largest number of native speakers, they cover about 85% of the world population. Many languages do not have sufficient NLP resources, such as annotated word lists, syntactic parsers or Named Entity Recognition (NER) tools. For example, Balochi, Belarusian and Konkani share the rank of 98–100 in this list with ≈8M speakers each, which is more than the number of speakers of much better resourced languages such as Danish or Finnish, while they have almost no resources. Similarly, Ukrainian with its 30M native speakers occupies the 40^{th} position in this list (the 8^{th} position in Europe), while having very minimal NLP resources.

One of the ways for addressing this issue involves relying on language families, so that the NLP tools for lesser resourced languages can be developed by using better resourced typologically related languages. For example, Belarusian and Ukrainian belong to the Slavonic family, in which Czech and Russian have sufficiently large resources, such as treebanks or annotated translated texts, see Table 1. This paper refers to this method as Language Adaptation, in which the resources are transferred from better resourced languages (donors) to lesser resourced ones (recipients) in a way similar to Domain Adaptation, which is aimed at transferring the models across the domains.

The tradition of developing NLP resources across languages is quite long, see Section 4 for a broader overview. The emphasis of this paper is on the usefulness of typological links

in building and using a shared representation. The specific mechanism of transfer proposed in this paper is based on building cross-lingual embedding spaces, in which words similar in their form and meaning are located close to each other across closely related languages.

The study presented in the paper enriches existing techniques of building cross-lingual embeddings from comparable corpora by introducing the Weighted Levenshtein Distance (WLD), when the weights are also obtained from the same seed dictionaries as used for aligning the spaces, see Section 2.2 below. In addition to an intrinsic evaluation of the parameters of bilingual lexicon induction, cross-lingual embeddings have been evaluated extrinsically through their use in downstream tasks, in particular, via prediction of morphological properties of word forms (Section 3.1), Named Entity Recognition (Section 3.2) and genre classification (Section 3.3).

With respect to data needed for transferring the model, this study assumes a mid-resource setting:

1. a reasonably large (> 40 M words) raw text corpus without annotations is used to build a monolingual word embedding space for each language;
2. a corpus with annotations is available for a donor language, while a much smaller corpus *can* be available for a recipient language, at least for testing;
3. a small seed dictionary of bilingual equivalents is used to establish a cross-lingual embedding space.

This allows a semi-supervised setup: a large raw text corpus helps in achieving good lexical coverage and robustness by accounting for more typical contexts in comparison to smaller annotated corpora. At the same time, an annotated donor corpus provides data for learning a model for the phenomenon of interest, such as morphological properties or features of genres. A seed dictionary (of about 500-2000 words) is used for mapping the embedding spaces between the languages.

In this study, large raw text corpora come from Wikipedia. However, this should not necessarily be the case. A crawl of available Web resources, e.g., the Wacky corpora (Baroni et al., 2009), is equally suitable for the first step. The annotated corpora used in the studies below depend on the task, for example, the morphological annotation experiment uses the respective Universal Dependencies (UD, v.2.0) corpora (Nivre et al., 2016), the Named Entity Recognition experiment is based on a Slovenian NER corpus (Krek et al., 2012), while the text classification experiment uses a Russian collection of genre annotated texts (Sharoff, 2018). When large parallel corpora are not available, the seed dictionaries can be derived from the links between the Wikipedia pages in the donor and recipient languages.

2 Induction of cross-lingual embeddings using cognates

2.1 Cross-lingual embedding spaces

A vector space for words represents each word as a vector of a fixed dimensionality with the aim of grouping semantically similar words closer to each other in this space (Rapp, 1995). Modern methods use neural networks for building such embedding spaces from raw text corpora (Bengio et al., 2003). Out of many methods for building *monolingual* embedding spaces, this study primarily uses FastText (Bojanowski et al., 2016), a recent approach,

Table 1. *Available corpora*

Languages	UD	Wiki	PEMT
Romance			
Catalan	531K	181M	
French	1134K	667M	432K
Italian	502K	433M	329K
Portuguese	570K	222M	321K
Romanian	356K	70M	
Spanish	1004K	530M	265K
Slavonic			
Belarusian	8K	23M	
Bulgarian	124K	60M	
Croatian	197K	40M	
Czech	2222K	120M	183K
Polish	70K	242M	213K
Russian	1247K	460M	266K
Slovak	106K	321M	
Slovenian	170K	351M	
Ukrainian	100K	193M	

Table 2. *Alignments from Wikipedia for titles and words*

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

Word forms aligned for the seed dictionary:

Polish	Russian	English
Budapeszt	Будапешт	Budapest
kapral	капрал	corporal
marionetek	марионеток	marionettes
organizacyjnego	организационного	organisational
patriarchy	патриарха	patriarch
tropikalny	тропический	tropical

Character alignment for word forms:

m a r i o n e t e k
м а р и о н е т о к

which combines the traditional skip-gram model with a model for building the embedding vectors for character n-grams within words. This incorporates some information from the subword level into the word embedding vector.

A commonly used model for building a *cross-lingual* embedding space is based on con-

structing a linear transformation matrix **W** for transforming one of the monolingual spaces to the other one by minimising the following objective:

$$\min_{\mathbf{W}} \sum ||\mathbf{W}e_i - f_i||^2 \quad (1)$$

where $e_i \in \mathbf{E}$ and $f_i \in \mathbf{F}$ are the respective embedding vectors in the two languages for words, which are supposed to be translations of each other according to the seed dictionary. This study uses a method for building **W** via SVD (Artetxe et al., 2016), which ensures that **W** is an orthogonal matrix built using a closed form solution:

$$\mathbf{W} = \mathbf{V} \times \mathbf{U}^T \quad (2)$$

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

2.2 Cross-lingual mapping using cognates

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 (LD) from the seed dictionaries;
3. alignment of monolingual embeddings by linear transformation using orthogonalisation and Weighted LD (WLD);

The seed dictionaries can be provided by word alignment of large parallel corpora. In a low resource setting, the seed dictionaries can be obtained from the titles of interlinked Wikipedia articles in two languages (iWiki links),[1] see examples of aligned titles in Table 2. This helps in modelling scenarios when few parallel texts are available, such as for the Polish-Russian pair. Even though Polish is included in Europarl, and Russian is in the UN corpus, very few reliable resources are available for the Polish-Russian pair itself. 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, which are not representative for the properties of the general lexicon. Table 2 lists a random selection of the word forms aligned for the Polish-Russian pair.

In addition to providing the training lexicon, a seed dictionary can also be used to provide a character-level model for matching the cognates via WLD, see the part of Table 2 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. The character alignment model is particularly important for establishing orthographic similarity

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

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 weighted by the probabilities of their character-level alignments:

$$WLD(s_e, s_f) = \frac{\sum_{(e,f)\in al(s_e,s_f)}(1 - p(f|e))}{\max(len(s_e), len(s_f))} \tag{3}$$

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.

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 2, the process of getting the Levenshtein weights ran in two steps. In the first step, an initial estimate of the character translation probabilities was produced from *all* word pairs 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:

$$score(s_e, s_f) = \alpha \cos(v_e, v_f) + (1 - \alpha)(1 - WLD(s_e, s_f)) \tag{4}$$

where s_e and s_f are words in the two languages, v_e and v_f are their embedding vectors in the cross-lingual embedding space, while α is the relative weight of the cosine similarity. Unlike the cosine similarity measure, the WLD value is greater for more remote strings.

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 (2) helps in improving alignment quality in the general case. However, it does not allow adjusting the transformation matrix by taking into account the orthographic 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 (4) 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 transformation matrix from (2) 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 transformation matrix. Again, while repeated application of these steps is possible, it did not produce better results, so the experiments below present the results obtained after two iterations.

Table 3. *Prec@1 for En-It dictionary induction*

W2V vectors from (Dinu et al., 2014) Full test set	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
	Orth+WLD	**0.531**
FT vectors from (Mikolov et al., 2017) Full test set	FT+TM	0.461
	FT+Orth	0.529
	FT+Orth+WLD	**0.616**
	MUSE (Conneau et al., 2017)	0.683
FT vectors from (Mikolov et al., 2017) Reduced test set with cognates	FT+TM	0.550
	FT+GC	0.575
	FT+Orth	0.614
	LD $\alpha = 0$	0.298
	WLD $\alpha = 0$	0.339
	FT+Orth+WLD $\alpha = 0.5$	0.584
	FT+Orth+LD $\alpha = 0.73$	0.669
	FT+Orth+WLD $\alpha = 0.73$	**0.692**
	MUSE	0.719

2.3 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.

In addition to the standardised embeddings as used in (Dinu et al., 2014; Artetxe et al., 2016), a new set of embeddings produced by FastText has been added to the English-Italian experiments (labelled as FT in Table 3). 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 dictionaries are not always available, so the seed dictionaries for building 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).

Table 4. *Dictionary induction results for Slavonic languages*

	sl-hr	sl-cs	sl-pl	sl-ru	ru-uk	cs-sk
Dictionary induction without WLD						
#Train dic	2510	3328	3047	4356	2617	11400
#Cognates	38247	24918	24215	32935	153644	74542
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

2.4 Experimental results

The results listed in Table 3 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 3) also make a considerable positive impact. Adding the constraint of having orthographic 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* $\rightarrow$ *esaminato* or *Jerusalem* $\rightarrow$ *Gerusalemme*).

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. The same value of $\alpha = 0.73$ has been used throughout the remaining experiments. 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 4). The row labelled #Cognates lists the number of WLD cognates retrieved for the second iteration of alignment. The amount of useable cognates depends on the size of the Wiki corpora used for training, see Table 1, as well as on the typological distance between the languages. Comparison of the Slavonic dictionary induction results 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. The Wikipedia corpus used for Croatian is quite small for reliable training of monolingual embeddings, so incorporating the WLD score contributes to improving the initial deficiencies of its space.

The FastText vectors of 300 dimensions built from the Wikipedias for the selected Balto-Slavonic languages (Belarussian, Czech, Croatian, Lithuanian, Polish, Slovak, Slovene, Ukrainian) have been transformed into a shared Panslavonic embedding space. For convenience of running cross-lingual experiments, English has also been added to the shared

embedding space. In spite of the fact that it is not a closely related language, its alignment to the Slavonic languages benefits from the WLD because of a large number of cognates such as the names of locations, personal names and borrowings. Another shared embedding space was produced for selected Romance languages.

2.5 True cognates and false friends

It is well known that even closely related languages have a number of false friends, for example, *Mist* in German means 'manure' unlike *mist* as used in English. However, a closer look at the list of cognates shows that there is a cline of cases:

1. *consistently* false friends, e.g., *bezcenny* means 'worthless' in Polish and 'invaluable' in Czech;
2. *partial* false friends, e.g., e.g., *žena* can mean either 'wife' or 'woman' in a number of Slavonic languages, e.g., Croatian, while its cognate *жена* in Russian always means 'wife';
3. *actual* cognates with uncommon divergent senses, e.g., similarly to *жена* in Russian, in Polish *żona* means 'wife', while rarely it can also mean 'woman'.

Therefore, the boundary between true cognates and false friends is quite flexible. This can lead to some disagreement between the annotators with respect to what constitutes false friends, see also a discussion in (Fišer and Ljubešić, 2013).

Monolingual word embeddings are built to reflect the similarity of the most common contexts via the distance between the embedding vectors, so the false friends are likely to have fairly distant vectors, as indicated by low cosine similarity values. However, the WLD reflects the similarity of the word forms, thus leading to the possibility of selecting false friends as possible translation equivalents. Therefore, the Panslavonic embedding space has been tested against available lists of Slavonic false friends to determine the amount of non-cognate noise introduced through the use of WLD.

A useful testbed is provided by the False Friends of the Slavist,[2] which covers most of the language pairs for the Panslavonic set, even though its coverage differs across the language pairs. The first two columns in Table 5 list the false friends for the Russian-Czech direction provided in the dictionary. They can be ranked by their similarity scores (Column 'Cos') with the top words corresponding to consistently false friends, as their contexts typically differ. The words at the bottom of the list tend to be actual cognates, which have been included in the gold-standard lists because they also have some divergent uses. While the words at the bottom of the lists can be potential false friends, corpus evidence for the most common senses suggests that their divergent semantic components are uncommon, at least in the Wikipedia corpus used for building the embeddings. Often there is a mismatch between dictionary definitions and the actual corpus use. For example, while the Russian word *запах* 'smell' is neutral in its dictionary sense, the majority of its collocations are negative ('unpleasant', 'pungent', 'foul', similarly to the collocations of the word *odour* in English), thus leading to its embedding vector being closer to *zápach* in Czech, which means 'unpleasant smell'.

[2] `https://en.wikibooks.org/w/index.php?oldid=3417664`

Table 5. *Ordering false friends in cognate lists*

Russian	Czech False	WLD	Cos	W+C	Best Cos		Best Cos+WLD	
заход	záchod	0.473	0.009	0.149	mezipřistání	0.411	*hod*	0.359
рок	rok	0.112	0.037	0.267	punkrockové	0.658	**rock**	0.580
обход	obchod	0.287	0.084	0.254	**obcházení**	0.467	**obcházení**	0.429
штука	štuka	0.204	0.103	0.290	pochopitelná	0.419	taky	0.410
столица	stolice	0.248	0.106	0.280	*město*	0.489	*město*	0.423
заказ	zákaz	0.417	0.131	0.253	**zakázka**	0.608	**zakázka**	0.562
урок	úrok	0.289	0.131	0.288	školník	0.383	školník	0.368
дело	dělo	0.272	0.154	0.309	obvinění	0.361	delikt	0.361
красный	krásný	0.443	0.155	0.264	**červený**	0.599	**červený**	0.503
выход	východ	0.439	0.166	0.273	**výstup**	0.404	přechod	0.384
повесть	pověst	0.345	0.185	0.312	**povídka**	0.698	**povídka**	0.640
живот	život	0.219	0.197	0.354	nohy	0.542	nohy	0.444
родина	rodina	0.123	0.199	0.382	**domovina**	0.447	**domovina**	0.457
худой	chudý	0.623	0.206	0.252	zběhlý	0.345	hodný	0.343
глава	hlava	0.276	0.207	0.347	**starosta**	0.490	**starosta**	0.441
власть	vlast	0.256	0.209	0.353	svrchovanost	0.590	**vláda**	0.518
страна	strana	0.108	0.209	0.394	republika	0.473	ukrajina	0.421
град	hrad	0.270	0.222	0.359	**krupobití**	0.346	**grad**	0.463
ставка	stávka	0.286	0.225	0.357	úroková	0.478	splátka	0.414
жизнь	žízeň	0.682	0.235	0.258	**život**	0.635	**život**	0.564
ел	jel	0.351	0.235	0.346	vypil	0.416	**jedl**	0.428
век	věk	0.394	0.238	0.337	stol	0.454	**století**	0.386
скоро	skoro	0.132	0.245	0.413	brzy	0.595	**brzo**	0.508
князь	kněz	0.489	0.261	0.329	**kníže**	0.703	**kníže**	0.635
враг	vrah	0.304	0.281	0.393	**nepřítel**	0.624	**nepřítel**	0.486
злодей	zloděj	0.380	0.314	0.396	padouch	0.513	**zloduch**	0.474
склеп	sklep	0.157	0.323	0.463	**hrob**	0.583	**hrob**	0.475
петроград	petrohrad	0.201	0.330	0.457	bolševiků	0.390	**petrohrad**	0.457
свет	svět	0.325	0.336	0.428	**světlo**	0.596	**světlo**	0.565
пара	pára	0.252	0.349	0.457	**dvojice**	0.514	**pár**	0.509
мрак	mrak	0.096	0.360	0.507	temnota	0.510	mrak	0.507
час	čas	0.255	0.371	0.472	**hodina**	0.594	**hodina**	0.481
запомнить	zapomenout	0.454	0.390	0.432	**zapamatovat**	0.633	**zapamatovat**	0.566
младенец	mládenec	0.251	0.395	0.491	chlapec	0.500	mládenec	0.491
муж	muž	0.194	0.398	0.508	**manžel**	0.696	**manžel**	0.602
ужасный	úžasný	0.484	0.400	0.432	**příšerný**	0.620	**děsivý**	0.500
тыква	tykev	0.531	0.411	0.426	kdoule	0.463	**tykve**	0.436
словенский	slovenský	0.321	0.415	0.486	chorvatský	0.703	**slovinský**	0.635
стул	stůl	0.277	0.419	0.501	stůl	0.419	stůl	0.501
палец	palec	0.135	0.428	0.546	**prst**	0.552	palec	0.546
постель	postel	0.230	0.490	0.566	**postel**	0.490	**postel**	0.566
запах	zápach	0.461	0.509	0.517	**vůně**	0.521	*zápach*	0.517
овощи	ovoce	0.417	0.518	0.535	**zeleniny**	0.633	ovoce	0.535
угол	úhel	0.617	0.611	0.549	**úhel**	0.611	**úhel**	0.549
слышать	slyšet	0.468	0.625	0.600	**slyšet**	0.625	**slyšet**	0.600

Higher orthographic similarity (lower WLD) increases the final score for all false friends. However, the consistently false friends have very low cosine similarity scores, so that the

Table 6. *Forms of adjectives in Russian and Ukrainian*

Forms of *green*	Russian		Ukrainian	
	Masc	Fem	Masc	Fem
Nominative	зелёный	зелёная	зелений	зелена
Genitive	зелёного	зелён**ой**	зеленого	зеленої
Dative	зелёному	зелён**ой**	зелен**ому**	зеленій
Instrumental	зелёным	зелён**ой**	зеленим	зеленою
Locative	зелёном	зелён**ой**	зелен**ому**	зеленій

weighted sum needs to compete with other vectors, which are closer semantically. The last four columns in Table 5 list the Czech vectors closest to the Russian keywords according to the plain cosine measure, as well as its weighted sum with the WLD ($\alpha = 0.73$). The correct dictionary translations are indicated in bold, while the partially correct translations, such as *запах* vs *zápach* are in italics. In some cases, the WLD helps in correcting the raw cosine measures (seven instances, e.g., *slovenský* vs *slovinský*), while in three cases using the weighted sum deselects the correct choice, but only when the initial similarity score was high (*prst* 'finger' vs *palec* 'thumb'). We can conclude that the WLD score tends to be helpful even in the difficult case of dealing with false friends.

3 Application studies

3.1 Morphology prediction

3.1.1 Prediction of syncretism

The previous section shows that a procedure for aligning the cross-lingual embedding spaces can benefit from taking the similarity between the languages into account. So far the proposed procedure assumed a one-to-one mapping, namely that one form in the donor language corresponds to one form in the recipient language. While the problem with homonymy and polysemy of translation equivalents is important in the general case, this problem is relatively minor in related languages, because their words tend to keep the same distribution of meanings with relatively few exceptions, see the study of false friends above.

However, a far more common problem concerns differences in syncretism, i.e., when one form can serve several grammatical functions. For example, the verbs in French have the same endings for the first- and third-person forms, while these forms are different in Spanish:

Fr:*je/il anticip**e*** vs Es:*yo anticipo/el anticipa*

Therefore, a single form in French needs to be similar to two forms in Spanish.

Syncretism is very common in Slavonic languages as well. Table 6 shows the different case endings for the Russian and Ukrainian adjectives. In Russian, all feminine non-nominative forms of зеленый ('green') have the same ending, while the endings in Ukrainian differ in each case. The reverse is true for the masculine dative and locative forms, which are different in Russian and identical in Ukrainian. So cross-lingual mapping between the forms needs to address the problem of variations in syncretism even across closely related languages.

Table 7. *Proportion of OOV words in the lexicons*

	Cs	Ru	Pl	Sk	Be	Uk
Train	108257	97749	19344	19100	1628	5080
Dev	32461	26567	4778	5425	662	271
OOV #	7891	8034	2327	3385	436	192
OOV %	24.31%	30.24%	48.70%	62.40%	65.86%	70.85%

3.1.2 Experimental setup

A possible way of addressing this problem is by inferring information about morphology from the embeddings. It is known that the embeddings do keep information about the underlying morphology of the word forms, e.g., (Belinkov et al., 2017). Therefore, we can set the task of predicting morphological properties from the embeddings. For example, we train a model to predict the case, gender and number for the two fairly close embedding vectors from the Panslavonic space:

ru зелёному=(-0.047 -0.032 -0.101 0.007 0.021 -0.046 0.0066 0.095…)
→ Case=Dat|Gender=Masc,Neut|Number=Sing
uk зеленому=(-0.044 -0.062 -0.137 -0.035 -0.019 0.058 0.106 0.017…)
→ Case=Dat|Gender=Masc,Neut|Number=Sing
→ Case=Loc|Gender=Masc,Neut|Number=Sing

This expertimental setting helps in two ways. First, it tests the possibility to determine morphological properties within each language even after the cross-lingual transformation in order to assess the difference between the forms or to assign the right translation given a context. Second, it can help in populating the lexicons for POS taggers and parsers. Training corpora, especially for lesser resourced languages, are quite small, see the corpus sizes in Table 1, while the prediction setup using embeddings benefits from more contexts available in large raw text corpora.

Table 7 demonstrates the difference between the training and development parts of the UD corpora with respect to their lexicon. The smaller corpora have a substantial rate of Out-Of-Vocabulary (OOV) words, which makes the tagging task harder, especially given that their tagging models are based on very sparse data.

The experimental setup tested in this study involves predicting properties for nouns, adjectives and verbs from the Panslavonic vectors (300 dimensions) using the UD training sets for training and their development sets for testing morphological predictions. The UD test sets have been reserved for testing the accuracy of POS tagging and parsing. Prediction has been done by a Multi-Layer Perceptron (MLP) with a single hidden layer of 150 neurons using tanh as the activation function and the Adam optimiser. Experiments with other hyperparameter settings did not change the results significantly. Two models have been tested:

R training using the original UD lexicon for each **r**ecipient language;
D training using cross-lingual embedding by transfer from related **d**onor languages: Cs→Pl,Sk, Ru→Be,Sk,Uk

Given the multilabel setup, the evaluation metric is Average Precision for prediction

Table 8. *Morphology prediction results*

Language	POS	#T$_R$	#T$_D$	Train$_R$	PerT$_R$	Train$_D$	PerT$_D$	Test	AP$_R$	AP$_{D,O}$	AP$_{D,W}$
R: Be	adj	23	52	357	16	8,067	155	69	13%	46%	46%
D: Ru	noun	39	77	898	23	14,810	192	196	25%	41%	49%
	verb	26	65	325	13	9,358	144	66	3%	48%	68%
R: Uk	adj	51	52	3,481	68	11,783	227	425	61%	55%	63%
D: Ru	noun	62	77	7,099	115	19,878	258	1,047	41%	49%	53%
	verb	35	65	3,209	92	14,519	223	405	80%	71%	81%
R: Pl	adj	61	245	3,043	50	14,979	61	417	37%	26%	34%
D: Cs	noun	69	140	7,959	115	22,489	161	1,129	49%	40%	40%
	verb	21	65	1,926	92	7,253	112	235	85%	73%	81%
R: Sk	adj	64	245	2,664	42	14,199	58	654	39%	29%	30%
D: Cs	noun	49	140	6,198	126	20,091	144	1,680	43%	48%	53%
	verb	15	65	590	39	6,630	102	133	33%	67%	64%
R: Sk	adj	64	52	2,664	42	11,451	220	654	39%	36%	33%
D: Ru	noun	49	77	6,198	126	21,744	282	1,680	43%	44%	51%
	verb	15	65	590	39	12,659	195	133	33%	33%	75%

(Sorower, 2010). For example, when the model predicts four labels for a word form, three of which are correct, the precision for this prediction is 0.75.

3.1.3 Prediction results

Table 8 presents the results of prediction. In this table and in the discussion below R stands for Recipient, D for Donor. D, O for the Donor part corresponds to predictions using the joint embedding space produced via orthogonal transform as in (Artetxe et al., 2016), D, W corresponds to the join embedding space produced by using WLD-induced cognates.

Column #T$_R$ indicates the number of tags in the training corpus for an individual recipient language, while #T$_D$ indicates the number of tags in the donor language corpus. If the recipient corpus is small, e.g., for Be, it covers only a small portion of possible tags. The number of examples available for training can be significantly increased via the donor language, see Columns Train$_R$ and Train$_D$ (many more examples are available in the donor corpora, so addition of examples was limited to provide at most 400 examples per tag). The donor language also provides more examples per individual tag, see Columns PerT$_R$ and PerT$_D$. The test examples (Test) were selected from the development parts of the respective UD corpora for words not attested in the training corpora.

Table 8 shows that prediction usually improves by taking more data from the donor language. When the initial training set is very small, as it was the case for Belarussian, the improvement is dramatic, e.g., from 3% to 68% for Belarussian verbs. The original Belarussian UD corpus contains merely 13 examples per verbal tag on average, which is not enough for training a classifier. Comparison of the AP$_{D,O}$ vs AP$_{D,W}$ columns (Orthogonalisation vs WLD) shows overall improvement.

Table 9. *F1 strict prediction scores for NER at BSNLP*

EC news:						
cs	hr	pl	ru	sl	uk	
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
39.7	40.4	26.8	30.2	58.4	16.0	Orth
47.6	44.3	44.2	33.6	**59.5**	13.7	Orth+WLD
Trump:						
cs	hr	pl	ru	sl	uk	
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
45.1	51.6	39.0	19.7	**62.7**	21.9	Orth
52.6	**52.4**	**55.2**	21.0	62.6	20.7	Orth+WLD

In the case of Czech, the UD tags make heavy use of features specific to the Czech training corpus, e.g., *Style* (with such values as *Coll*oquial, *Arch*aic, *Rare*, etc) and *NameType* (*Geo, Giv*en name, *Sur*name, etc), which are not used in the available feature sets in other related languages. These two specific morphological attributes have been removed before training. However, the number of Czech tags is still quite high, compare the numbers for $\#T_R$ vs $\#T_D$ in Table 8 for Polish and Slovak ($\#T_D$ is for Czech as the Donor). In the end, many Czech tags do not contribute to predicting the tags for Polish and Slovak in the cross-lingual setting. Another observation is that the gold standard is derived from an annotated corpus, which does not necessarily cover the entire paradigm for each test item. This means that the prediction model often produces correct results without receiving credit for this. For example, *антропологический* ('anthropological') in Russian in the gold standard corpus is annotated as:

ADJ Case=Nom|Gender=Masc|Number=Sing

while the predicted annotation is equally correct:

ADJ Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing

3.2 Named Entity Recognition

3.2.1 Training setup

The cross-lingual embedding space has been also tested through the Named Entity Recognition (NER) task, which is aimed at detecting and labelling 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 easily available features, such as character-level embeddings or the presence of capitalisation. The taggers for individual languages were trained from an

existing NER-annotated corpus from (Krek et al., 2012) in Slovenian using the Panslavonic embedding space.

Small samples from each language have been added to the Slovenian training corpus in order to provide at least some information for the character-level embeddings. The small additional samples were derived from the Wikipedia title names in the respective languages for the articles which categories matched such patterns as 'Births' (for person names), 'Organizations' (for organisations) and 'Countries' or 'Villages' (for locations, since the Wikipedia articles usually lack a more generic category of locations). For example, an entry for a sample of Russian person names looks like:

Игорь	B-Per	Igor
Ларионов	I-Per	Larionov
говорит	O	says

The entry contains the likely first and following elements of a named entity (B-Per and I-Per, respectively), and it ends with a third person verb, which helps in learning typical conditions when a named entity ends. The most common verbs and prepositions were used as the ending elements as selected from the respective UD corpora.

3.2.2 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 news reporting about the European Commission, 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 from JHU (Mayfield et al., 2017) was based on projection of NER labels via word-aligned parallel corpora, see Table 9, as well as a brief explanation of the projection approach in Section 4.

The shared embedding space is surprisingly efficient. The training corpus was for Slovenian, so it provides the upper baseline for language adaptation. Czech, Croatian and Polish are sufficiently similar typologically, so the accuracy on those languages is only 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 transfer from 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 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 relatively small training set. Despite such limitations, the transfer model which only used the Slovenian training corpus was on average more successful than the projection-based model.

3.3 Genre classification

Text classification is one of the commonly used tasks in NLP. A more specific task concerns classification of texts into genres (Santini et al., 2010), since genre annotation provides useful information for understanding kinds of texts a corpus consists of in addition to un-

Table 10. *Genre annotated corpora*

	Russian		Ukrainian	
	#doc	#Words	#doc	#Words
News (A)	100	39583	18	6767
Discussion (B)	218	306063	20	65345
Reviews (C)	46	62072	29	44760
Information (Wikipedia)	236	475128	48	63319
Instructions (E)	62	107652	43	71973
Academic (J)	34	271150	18	14040
Legal (H)	48	277619	6	36024
Fiction (K)	86	196576	50	10001
Personal	205	216822	23	65291
Promotion, ads	46	27334	29	82617
Total	1081	1979999	284	460137

derstanding the structure of its topics via Topic Modeling (Blei et al., 2003; Sharoff, 2013). Unlike topic modeling, which usually uses unsupervised topic discovery via detection of keywords, the relationship between topics and genres is not well defined, since keywords from the same topic are often used in texts of different genres. Instead, genre classification requires a supervised approach to learn the association between *stylistic* features and genre labels.

A supervised approach needs a training set, which might be available for some specific languages and specific genre classification schemes, but not for others. The Language Adaptation framework can be used to solve this problem as well: training is done using the available donor resources within the shared cross-lingual embedding space, while the resulting model is applicable to the recipient language. As an example of such study, a Russian genre-annotated corpus (Sharoff, 2018) has been applied to classify Ukrainian texts into genres.

For evaluating the resulting classifier a small testing corpus is still required. The Ukrainian corpus for this study has been collected from the Web to provide a sample of the major genres represented in the respective Russian corpus, see Table 10. For the ease of interpretation, the category labels given in brackets in Table 10 roughly correspond to the categories of the Brown Corpus (Kučera and Francis, 1967), whenever possible. The Personal category (missing in the Brown corpus) primarily contains personal blog entries and personal messages from social networks.

In comparison to other supervised text classification setups, such as sentiment analysis, genre classification can be biased by the topical words in the training corpus (Petrenz and Webber, 2010). A convenient representation, which can use cross-lingual embeddings and at the same time can have the capacity to generalise a genre across topics represented in the training set, is a mixed feature set (Baroni and Bernardini, 2006), which is produced by replacing the less frequent words with their POS codes, while leaving the most common words in their original form. The POS codes have been taken from the UD set to ensure their transfer across languages. For example, a review text (in English for illustration purposes):

Table 11. *Genre classification results: Precision for Russian and Ukrainian*

	Ru	Ukrainian		
			Ru-Uk transfer	
	CV	CV	Orth	WLD
News (A)	0.928	0.102	0.091	0.286
Discussion (B)	0.594	0.072	0.000	0.109
Reviews (C)	0.744	0.102	0.247	0.253
Information	0.481	0.588	0.321	0.225
Instructions (E)	0.957	0.060	0.474	1.000
Academic (J)	0.932	0.244	0.188	0.067
Legal (H)	0.966	0.500	0.000	1.000
Fiction (K)	0.868	0.667	0.000	1.000
Personal	0.584	0.309	0.321	0.412
Promotion	0.906	0.072	0.400	0.667
Average P	0.796	0.272	0.204	0.502
Hamming loss	0.056	0.134	0.182	0.160

It won the SCBWI Golden Kite Award for best nonfiction book of 1999 and has sold about 50,000 copies.

converts into a mixed representation as

It won the PROPN ADJ NOUN NOUN for best NOUN NOUN of [#] and has sold about [#] NOUN.

This representation makes it easier to compare this review snippet to other reviews without relying too much on the specific keywords and numerical values, while it keeps important lexical features for detecting genres, see (Petrenz and Webber, 2010) for further discussion concerning the importance of non-topical representations for genre classification.

As for the machine learning approach, the genre classification experiment reported here uses a simple Feed Forward network inspired by FastText (Joulin et al., 2017). In this setup, we start with pre-trained word embeddings from Section 2.2 to build a document embedding representation, doc2vec. Then, simple Feed Forward neural networks are used for multi-labelled text classification. This method has been shown to be robust and efficient in a number of sentiment classification tasks for English, while achieving comparable accuracy in comparison to more complicated neural models based on CNN or LSTM (Joulin et al., 2017). The specific implementation in this study is based on Keras.[3]

Table 11 presents the results of classification in terms of average precision, which was the objective for optimising the training pipeline. Given the vast amount of texts on the Web, optimising for precision helps in extraction of useful sample texts in a specific genre, in contrast to retrieving all texts in this genre. In the multilabel context, the overall quality of classification can be described in terms of its Hamming loss, which computes the proportion of *irrelevant* predictions (Sorower, 2010), thus the lower the better.

[3] `https://github.com/keras-team/keras`

The first two columns in Table 11 (marked as CV) show the results of training classifiers on the respective training corpora with 10-fold cross-validation. The bigger Russian corpus quite predictably produces a much better model. The last two columns show the results of training on the Russian corpus with the two versions of the cross-lingual embedding space with and without WLD. In the same pattern as with the NER task, transferring data from the donor language usually helps, and the transfer accompanied with the WLD cognates helps even more. For example, fiction, legal and instructive texts can be detected reliably, so the genre classifier is useful for selecting their samples from the Ukrainian Web. At the same time the resulting Ukrainian models suffer from the mismatch between the original Russian training set and the Ukrainian testing sample. In particular, the academic texts in the Ukrainian testing corpus came primarily from popular science sources, while the Russian model has been trained on a range of research articles.

4 Related studies

The possibility of developing resources across languages has been recognised quite early in the NLP community, e.g., (Wu, 1997). In a rule-based approach, having a shared representation can be interpreted as a system of shared rules with some language-specific constraints when necessary (Bateman et al., 2000).

In the modern Machine Learning paradigm there are several approaches to building multilingual models. One set of approaches uses parallel corpora for projecting automatic annotations in one language to others, e.g., for POS tagging (Das and Petrov, 2011), parsing (Täckström et al., 2013; Tiedemann, 2014) and NER (Mayfield et al., 2017). In the projection approaches, the donor part of a parallel corpus is annotated with an existing tool. The labels are projected into the recipient language via word alignments with possible adjustments of labels in the case of alignments other than one to one. This creates a training corpus for the recipient language. The problem with using parallel corpora in this task is related to their limitations in terms of topics and genres even for better resourced languages, e.g., resources are much scarcer outside of the official documents of Europarl and the United Nations. 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 reliable parallel corpus, which contains this specific language pair.

Another set of approaches uses monolingual comparable corpora, which should help in improving robustness of transfer by accounting for more typical contexts for more language pairs. 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 built via predicting context words (Bengio et al., 2003) has recently become the standard way of representing meanings of words as the distance between their embedding vectors. 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 transformation 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 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). The cross-lingual embedding space has been shown to be useful in topic and sentiment classification tasks, e.g., (Klementiev et al., 2012), but it has not been tested for genres.

Feature spaces with a large number of dimensions (100-500) 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 such vectors more common choices in the lexical retrieval tasks leading to more 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 $|\mathcal{N}_{20}|$ hubness index on the English Wikipedia is 5. Dinu et al. (2014) observe that the hubness phenomenon becomes more pronounced after linear transformation, since the objective for building the transformation matrix **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) 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.

In addition to a model with a seed bilingual dictionary, the initial study by (Mikolov et al., 2013) also introduced constraints on what its authors call "morphological structure" (actually the Levenshtein Distance) for keeping only the cognate words in the output. However, this worked as a filter to reduce the amount of errors rather than to help with improving the dictionary. Further work on bilingual lexicon induction did not include the use of cognates, especially in the context of related languages.

Detection of cognates across related and non-related languages has been also studied recently, e.g. (Frunza and Inkpen, 2009). Some studies relied on using bilingual corpora (Kondrak, 2013), while others used embeddings from comparable corpora. For example, 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). A study aimed at detecting false friends via embeddings (Fišer and Ljubešić, 2013) treated the false friends only among homographs (identically spelled words), not among cognates.

There have been also various studies aimed at providing quantitative analysis of embeddings by training predictors for various classification tasks, e.g., (Belinkov et al., 2017; Köhn, 2015). The specific contribution of this study consists in investigation of transferring such predictors across the related languages using a shared annotation framework, such as UD.

5 Conclusions and further work

The key take-home message from this study is as follows: when cross-lingual embedding spaces for related languages are built by taking into account lexical similarity between the cognates, the resulting models can be more successful in transferring the resources from the donor languages. This study illustrates this claim for a number of language pairs and application domains, such as the dictionary induction task or morphosyntactic prediction.

In particular, the results in the dictionary induction task improve the state of the art considerably, for example, from 0.429 to 0.840 for the Slovenian-Croatian pair when a corpus is too small for reliable training of monolingual embeddings. Incorporating the WLD score contributes to improving the initial deficiencies of a small corpus.

The tools for aligning the monolingual embedding spaces for related languages, the resulting embeddings, as well as the trained NLP models transferred to the recipient languages are available under permissive licenses.[4] In addition to the NER and genre classification tasks, as shown in this paper, the cross-lingual spaces can be used for improving coverage of existing resources, such as POS taggers (Straka et al., 2016) or MT for related languages (Forcada et al., 2011).

The resulting Panslavonic space can be easily expanded to accommodate a new language, e.g., Rusyn or Sorbian, when a reasonable monolingual corpus is available to train the embeddings for this language, and when a reliable seed dictionary exists between this language and one of the other languages in the current Panslavonic space (the Wikipedia iWiki lists for such languages are too short to produce useful seed dictionaries).

There are several extensions possible for this line of research. First, the setup for building cross-lingual embeddings involves a number of hyper-parameters which deserve a separate study. This concerns:

Seed dictionary There can be different sources for choosing the seed dictionary, such as alignments from parallel corpora, existing traditional dictionaries, alignments from comparable Wikipedia titles (as used in this study). In addition to this, there can be variation in their size or contents, which might in turn lead to investigation of their components, such as common names, borrowings or proper names. For example, it is relatively easy to collect large lists of proper names from such sources as Wikipedia titles via the iwiki links. The current study filtered many of them through a frequency list. However, their presence mihgt benefit downstream tasks, such as NER.

WLD contribution The best value of α has been estimated on the development set for one language pair and used in other experiments. However, the optimal balance between the embedding scores and WLD depends on the quality of the seed dictionary and the typological distance between the languages.

Monolingual embedding spaces There are numerous methods and parameters for building embedding spaces, which can impact their usefulness for the cross-lingual embedding task. For each language pair, this study used embeddings from a single source without comparing different settings to the individual tasks.

Second, the cross-lingual spaces in the current study are constructed in iterations by means of a closed-form method for building the transformation matrix. This closed-form method cannot take into account the lexical similarity, which needs to be introduced via a separate dictionary update. A useful experiment would be to use an adversarial training technique (Conneau et al., 2017) in order to transform the monolingual spaces while adding lexical similarity measures such as WLD. As shown in other studies, adversarial training outperforms the orthogonal transform (see the rows marked as MUSE in Tables 3 and 4)

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

and allows incorporation of other cost functions. Another possibility for using WLD in the process of aligning monolingual embedding spaces is by iterative learning of a **non**linear transformation (Di Marzio et al., 2018).

Third, morphological prediction can be improved if done in a multitask fashion, when the tasks concern predictions of the individual features, such as case, gender and number (Augenstein et al., 2018). In the current study, the tags were predicted as a whole. Other kinds of multi-task and multi-domain experiments are also possible. For example, the current study does not make a distinction between the embeddings for different languages, so the shared space is considered to be the same for all languages, even though the semantic and grammatical properties of individual languages are likely to differ. It has been shown that applying autoencoders over the feature spaces in two languages leads to a better feature space for the target model, because this helps in generalising language-specific variations in the monolingual feature spaces (Rios and Sharoff, 2016).

Finally, the current model has been tested with relatively well-defined synchronic languages, such as representatives from the Slavonic family. It is interesting to experiment with languages diachronically by building better models for earlier stages of language development, e.g., for medieval English, from the more abundant models existing for modern languages, see (Piotrowski, 2012). A related experiment would involve building models for dialects. A problem to be tackled in this case concerns the need to build a monolingual embedding space for a recipient language variety from a small amount of available raw texts.

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.

Augenstein, I., Ruder, S., and Søgaard, A. (2018). Multi-task learning of pairwise sequence classification tasks over disparate label spaces. In *Proc NAACL*, pages 1896–1906, New Orleans.

Baroni, M. and Bernardini, S. (2006). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Bateman, J. A., Kruijff, G.-J., Kruijff-Korbayová, I., Skoumalová, H., Sharoff, S., and Teich, E. (2000). Resources for multilingual text generation in three slavic languages. In *Proc Second International Conference on Language Resources and Evaluation (LREC)*.

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., and Glass, J. (2017). What do neural machine translation models learn about morphology? In *Proc ACL*, Vancouver, Canada.

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

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

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

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*.

Das, D. and Petrov, S. (2011). Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proc ACL*, Portland, Oregon.

Di Marzio, M., Panzera, A., and Taylor, C. C. (2018). Nonparametric rotations for sphere-sphere regression. *Journal of the American Statistical Association*.

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.

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

Fišer, D. and Ljubešić, N. (2013). Best friends or just faking it? Corpus-based extraction of Slovene-Croatian translation equivalents and false friends. *Slovenščina 2.0*, 1.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Frunza, O. and Inkpen, D. (2009). Identification and disambiguation of cognates, false friends, and partial cognates using machine learning techniques. *International Journal of Linguistics*, 1(1).

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.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proc EACL*, Valencia.

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

Köhn, A. (2015). What's in an embedding? analyzing word embeddings through multilingual evaluation. In *Proc EMNLP*, pages 2067–2073, Lisbon, Portugal.

Kondrak, G. (2013). Word similarity, cognation and translational equivalence. In *Approaches to measuring linguistic differences*. Walter de Gruyter, Berlin.

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

Kučera, H. and Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press, Providence.

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.

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

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proc LREC 2016*, Portorož, Slovenia.

Petrenz, P. and Webber, B. (2010). Stable classification of text genres. *Computational Linguistics*, 34(4).

Piotrowski, M. (2012). *Natural language processing for historical texts*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, Moscow.

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.

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. (2016). Language adaptation for extending post-editing estimates for closely related languages. *The Prague Bulletin of Mathematical Linguistics*, 106:181–192.

Santini, M., Mehler, A., and Sharoff, S. (2010). Riding the rough waves of genre on the web. In Mehler, A., Sharoff, S., and Santini, M., editors, *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York.

Sharoff, S. (2013). Measuring the distance between comparable corpora between languages. In Sharoff, S., Rapp, R., Zweigenbaum, P., and Fung, P., editors, *BUCC: Building and Using Comparable Corpora*, pages 113–129. Springer.

Sharoff, S. (2018). Functional text dimensions for the annotation of Web corpora. *Corpora*, 13(1):65–95.

Simons, G. F. and Fennig, C. D., editors (2017). *Ethnologue: Languages of the World*. SIL International, 20 edition.

Sorower, M. S. (2010). A literature survey on algorithms for multi-label learning. Technical report, Oregon State University.

Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proc LREC 2016*, Portorož, Slovenia.

Täckström, O., McDonald, R., and Nivre, J. (2013). Target language adaptation of discriminative transfer parsers. In *Proc NAACL HLT*, pages 1061–1071, Atlanta.

Tiedemann, J. (2014). Rediscovering annotation projection for cross-lingual parser induction. In *Proc COLING*, pages Proc 1854–1864, Dublin.

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

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational linguistics*, 23(3):377–403.