Cross-lingual Annotation Projection Is Effective for Neural Part-of-Speech Tagging

Matthias Huck and Diana Dutka and Alexander Fraser
Center for Information and Language Processing
LMU Munich
Munich, Germany
{mhuck,fraser}@cis.lmu.de diana.dutka@gmail.com

Abstract
We tackle the important task of part-of-speech tagging using a neural model in the zero-resource scenario, where we have no access to gold-standard POS training data. We compare this scenario with the low-resource scenario, where we have access to a small amount of gold-standard POS training data. Our experiments focus on Ukrainian as a representative of under-resourced languages. Russian is highly related to Ukrainian, so we exploit gold-standard Russian POS tags. We consider four techniques to perform Ukrainian POS tagging: zero-shot tagging and cross-lingual annotation projection (for the zero-resource scenario), and compare these with self-training and multilingual learning (for the low-resource scenario). We find that cross-lingual annotation projection works particularly well in the zero-resource scenario.

1 Introduction
Little or no hand-annotated part-of-speech training data exists for the vast majority of languages in the world. This work investigates POS-tagging for under-resourced languages with a state-of-the-art neural network model. We consider how best to deal with the zero-resource scenario (i.e., no availability of any POS-labeled training data for the targeted language). To better understand this scenario, we compare it with the low-resource scenario (i.e., availability of a small POS-labeled training corpus). We thoroughly compare four techniques, including: zero-shot tagging and cross-lingual annotation projection from a linguistically related higher-resource language (for the zero-resource scenario), as well as self-training and multilingual learning (for the low-resource scenario).

A controlled experimental design is established for our study. We aim for immediate comparability of all tested tagging strategies of both scenarios, zero-resource and low-resource. We therefore opt to carry out both the zero-resource and the low-resource experiments on the same language, Ukrainian, and measure tagging accuracy on one common test set. A small amount of manually POS-annotated Ukrainian training data is available, which we use for supervised low-resource training. We simulate the zero-resource scenario by not using any POS-annotated Ukrainian training data. Russian is a higher-resource language which is linguistically closely related to Ukrainian. We use a larger POS-annotated Russian corpus for multilingual learning and zero-shot tagging experiments, and an unlabeled Russian–Ukrainian parallel corpus for the cross-lingual projection annotation experiment. To strengthen the upper-bound result for low-resource tagging, we consider the improvements possible through self-training, for which we use the Ukrainian side of the Russian–Ukrainian parallel corpus in order to maintain comparability. Our experimental design allows us to directly assess whether the tagging quality of any zero-resource strategy is approaching the accuracies of supervised low-resource strategies. We find that zero-shot tagging does not yield satisfactory quality, even if we operate on a higher linguistic abstraction level with word stems, which are often very similar in Ukrainian and Russian. But the empirical results show that annotation projection from a closely-related language is a very effective strategy for training neural POS taggers.

2 Related Work
Annotation projection for POS-tagging was first explored by Yarowsky and Ngai (2001) for cross-lingual transfer from English to French. Our basic approach shares much of Yarowsky and Ngai’s
original idea and reaffirms the efficacy of annotation projection also with a state-of-the-art neural sequence tagging model (Wang et al., 2015) and on the modern universal POS-annotation scheme (Petrov et al., 2012).

Since 2001, in addition to POS-tagging, annotation projection has been successfully applied to other tasks such as named entity recognition (Yarowsky et al., 2001; Enghoff et al., 2018), word sense tagging (Bentivogli et al., 2004), semantic role labeling (Pado and Lapata, 2005, 2009; van der Plas et al., 2011; Aminian et al., 2017), or dependency parsing (Hwa et al., 2005; Tiedemann, 2014; Rasooli and Collins, 2015; Agić et al., 2016; Aufrant et al., 2016). Kim et al. (2011) presented an integration into a full pipeline for information extraction. Open-source software tools for annotation projection are now available online (Akbik and Vollgraf, 2018, 2017).

To avoid unnecessarily noisy data, unlike previous authors, Lacroix et al. (2016) did not apply heuristics to fix certain word alignment links that pose difficulties to annotation projection. They demonstrated that it is simpler and more effective to ignore unaligned words as well as many-to-many alignments. In our work, we likewise settle on a simple technique based on a one-directional word alignment.

Xi and Hwa (2005) have combined projected POS-annotation with a small manually annotated corpus in a low-resource scenario. Newer research on annotation projection for POS-tagging has looked at historical languages (Meyer, 2011; Sukhareva et al., 2017) and sign language (Östling et al., 2015). Notable exceptions are the works of Wisniewski et al. (2014), examining annotation projection for a CRF tagging model (Lavergne et al., 2010) on living spoken languages, and of Agić et al. (2015). Meyer (2011) tags Old Russian via annotation projection from modern Russian translations. Sukhareva et al. (2017) POS-tag the extinct Hittite language through projection from German. Recent related work on neural POS-tagging has mostly focused on robustness through character-level modeling (Heigold et al., 2016, 2018; dos Santos and Zadrozny, 2014; Labeau et al., 2015) or on architectural improvements (Huang et al., 2015; Ma and Hovy, 2016; Yasunaga et al., 2018). Kim et al. (2017) have proposed an interesting neural tagging architecture that allows for multilingual learning with a language-specific component integrated with another cross-lingually shared component. We are however not aware of many prior studies that systematically explore annotation projection for cross-lingual transfer in neural POS-tagging of living spoken languages. Steps in this direction have been taken only lately by Fang and Cohn (2016), Plank and Agić (2018) and Anastasopoulos et al. (2018). We follow up on this line of research with our work.

3 Methods

Research questions. We ask two central research questions in this work, one for each of the considered scenarios:

Low-resource scenario: When the amount of hand-labeled training data is small for the targeted language, how effectively can we further improve the tagger by employing auxiliary resources? Specifically, how helpful is the use of additional unlabeled corpora (self-training) and corpora in a different language (multilingual learning)?

Zero-resource scenario: When there isn’t any hand-labeled training data available for the targeted language, how effectively can we harness knowledge from annotated corpora in a different, but related language? Specifically, is tagging quality close to supervised low-resource conditions attainable with either a plain foreign-language tagging model (zero-shot tagging) or via annotation projection from a foreign language (cross-lingual transfer)?

Neural tagging model. Depending on the context, the part-of-speech of a word may vary. E.g., the English word “green” takes a different POS (adjective, noun, verb) in each of the following three sentences:

- The recipe requires green mangoes.
- She took 63 shots to reach the green.
- How can we green our campus?

The need to resolve such ambiguities is one of the challenges in POS-tagging, and is the reason why the task requires sequence labeling instead of just a simple dictionary lookup. Another challenge is imposed by words that are out-of-vocabulary (OOV) to the tagger—a pressing issue especially under low-resource conditions, where many valid word forms of the language are not observed in training data.
We utilize a Bidirectional Long Short-Term Memory (BLSTM) neural network model (Hochreiter and Schmidhuber, 1997) to build our sequence taggers. BLSTMs are recurrent neural networks (RNNs) that are capable of learning long-term dependencies, taking into account both the previous and the following context. RNNs generally show great results at processing sequential data. They are widely adopted in natural language processing, including the POS-tagging task (Wang et al., 2015). Other statistical sequence labeling methods, such as maximum entropy tagging models (Ratnaparkhi, 1996) or conditional random fields (Lafferty et al., 2001; Lavergne et al., 2010), are nowadays often outperformed by neural network methods (Collobert et al., 2011).

3.1 Self-Training

Given sufficient amount of labeled data, it is possible to build high-performance tools with direct supervision, but since there are languages that do not have enough suitable data to train a model, it is reasonable to employ semi-supervised methods. Those include self-training, which was previously discussed by McClosky et al. (2006), inter alia. Self-training requires labeled and unlabeled data and can be applied to low-resource languages. “Semi-supervised and unsupervised methods are important because good labeled data is expensive, whereas there is no shortage of unlabeled data” (McClosky et al., 2006).

3.2 Multilingual Learning

The multilingual learning method is suitable for under-resourced languages with little annotated data. The training set is enlarged through the texts of a related language. The idea is to shuffle original Ukrainian training sentences with the Russian labeled data to get more annotated texts.

3.3 Zero-shot Tagging

A zero-shot strategy can be pursued in case no annotated text exists for the resource-poor language. The zero-shot approach applies a tagging model trained for a closely related language.

There is quite some vocabulary intersection between Ukrainian and Russian (cf. Section 4.3), and the grammatical structure and word order of sentences are expected to be similar in the two related languages. We will however determine in the experimental section that these similarities are not strong enough to be able to use a model trained for Russian to tag Ukrainian sentences (Section 5.2.4).

3.4 Cross-lingual Transfer

The cross-lingual transfer approach relies on the availability of cross-lingual supervision and is suitable for languages that do not have any annotated data, but for which there is an available parallel corpus with a high-resource language. A POS-tagger for the high-resource language can be applied to automatically annotate the source side (here: Russian) of the parallel corpus. The source annotation is then projected to the target side (here: Ukrainian) (Yarowsky and Ngai, 2001). After that, a tagger for the resource-poor language can be trained on the target side of the parallel corpus with its associated projected automatic source-side annotation. This provides another solution in the case of a complete lack of gold-standard training data, the zero-resource scenario.

4 Corpus-linguistic Analysis

Ukrainian, as an under-resourced language, has a relatively small amount of suitable data that can be freely obtained from the web. There are two main data sources that are used throughout this work: annotated Ukrainian and Russian texts from the Universal Dependencies project and a Russian–Ukrainian parallel corpus of news texts. This section provides a description of the data as well as a quantitative comparison of the Russian and Ukrainian data sets.

4.1 Data

Universal Dependencies. The annotated data used to train taggers is taken from the Universal Dependencies corpora for Russian and
Universal Dependencies (UD) is a project based on open collaboration that is developing cross-linguistically consistent treebank annotation for many languages. The annotation scheme is based on an evolution of Stanford dependencies (de Marneffe et al., 2006; de Marneffe and Manning, 2008; de Marneffe et al., 2014) and Google universal part-of-speech tags (Petrov et al., 2012). The 17 UD core part-of-speech categories are listed in Table 1. Additional lexical and grammatical properties of words are distinguished by extra features that are not part of the tag set.

Russian–Ukrainian parallel corpus. The Russian–Ukrainian parallel corpus was created by ElVisti Information Center. A fragment of 100,000 sentences is freely available for scientific and educational purposes. The corpus consists of web publications of news articles and was created as a resource for building machine translation systems (Lande and Zhygalo, 2008).

4.2 Relatedness of Ukrainian and Russian

Slavic languages descend from a common predecessor, called Proto-Slavonic. Russian and Ukrainian belong to East Slavic, one of three regional subgroups of Slavic languages, which is also the largest group as for the number of speakers (Carlton, 1991).

Alphabet. Both Russian and Ukrainian use the Cyrillic script and have 33 letters each. However, there are differences in their alphabets. Unlike Russian, the letters Ёё, ъ, Ыы, Ээ are not used in Ukrainian, and Ukrainian has extra letters Ії, Єє, Ії, Ії, which are not found in Russian. The apostrophe occurs in words of both languages, but in Russian it is not very common and mainly used in foreign proper nouns.

Vocabulary. Despite the fact that the languages share some of their vocabularies with similar pronunciation and spelling, they often have different semantic shades. Having a common predecessor language, Russian and Ukrainian have retained many identical word stems. Stemming techniques will be explored in this work in order to capitalize on such similarities between the two related languages and improve Ukrainian POS-tagging.

Morphosyntax. Russian and Ukrainian also have similarities in their morphosyntactic features. For example, in both languages, the adjective, participle and possessive pronoun agree with the noun in case, gender and number. The verb has separate forms for different genders in the past but does not have gender variations in other tenses. There are three persons and two numbers.

4.3 Quantitative Comparison

Amount of data. The annotated UD data set for the Russian language is an order of magnitude bigger than the Ukrainian. The Ukrainian training corpus contains 85K annotated tokens in 5K sentences, the Russian corpus 1M tokens in 61K sentences.

Tag statistics. The distribution of tags in the Ukrainian and in the Russian UD training sets is quite similar, as can be seen in Figure 1. The most frequent tags in both corpora are NOUN and PUNCT, which account for nearly 25% and 20% of the tokens, respectively. Together with VERB, ADJ and ADP, they cover over 70% of the texts. The rank-frequency distribution of POS-tags approximately complies with Zipf’s law (Zipf, 1932).

The words in both Russian and Ukrainian are mostly unambiguous. The bigger part of the training data vocabulary is always annotated with the same tag (Table 2). Some words occur with up to five different tags in Ukrainian and up to six in Russian, but those are quite rare cases.
Table 3: Shared vocabulary before and after stemming.

<table>
<thead>
<tr>
<th></th>
<th>Words</th>
<th>Stems</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ukr</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Types</td>
<td>2998</td>
<td>3442</td>
</tr>
<tr>
<td></td>
<td>/ 26376</td>
<td>/ 15821</td>
</tr>
<tr>
<td></td>
<td>11.4 %</td>
<td>21.8 %</td>
</tr>
<tr>
<td>Tokens</td>
<td>35395</td>
<td>47789</td>
</tr>
<tr>
<td></td>
<td>/ 85469</td>
<td>/ 85469</td>
</tr>
<tr>
<td></td>
<td>41.4 %</td>
<td>55.9 %</td>
</tr>
<tr>
<td><strong>Rus</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Types</td>
<td>2998</td>
<td>3442</td>
</tr>
<tr>
<td></td>
<td>/ 130949</td>
<td>/ 48652</td>
</tr>
<tr>
<td></td>
<td>2.3 %</td>
<td>7.1 %</td>
</tr>
<tr>
<td>Tokens</td>
<td>392319</td>
<td>550297</td>
</tr>
<tr>
<td></td>
<td>/ 1087260</td>
<td>/ 1087260</td>
</tr>
<tr>
<td></td>
<td>36.1 %</td>
<td>50.6 %</td>
</tr>
</tbody>
</table>

**Shared vocabulary.** Taking into account that Ukrainian and Russian are related, it makes sense to examine their lexicons for common words. An overview of the shared vocabulary is given in Table 3 (left-hand column). There are only 2998 words appearing in lexicons of both languages. However, when counting their actual occurrences in the text we can see that common words are frequent throughout the texts. In Ukrainian, for example, 41% of the training texts consist of words that can be found in both languages.

These words are also mainly tagged in the same way. About 79% (2358 out of 2998) are tagged in both languages with the same tag (or tags). Another 13% (388 out of 2998) are tagged with the partially same tags. The rest of the words of the shared vocabulary (about 8%) are annotated with completely different tags in each language.

**Stemming.** Since both Russian and Ukrainian are richly inflected languages, but closely related to each other, many differences in their word surface form vocabularies might be caused by inflection diversities. Table 3 (right-hand column) provides statistics of the stemmed Russian and Ukrainian training sets to examine whether the amount of shared vocabulary is higher after the words are reduced to their stem forms. Russian text is stemmed with the Snowball stemmer (Porter, 1980) from the NLTK package. The stemmer for Ukrainian is an implementation found on GitHub posted by Kyrylo Zakharov.

The size of the shared vocabulary rises from 2998 to 3442 types after stemming. Although this increase in vocabulary overlap seems marginal, in terms of the occurrences there is a more significant change. After the stemming, the shared vocabulary tokens in the training sets of both languages amount to over 50%.

5 Experiments

5.1 Experimental Setup

5.1.1 BLSTM Tagger

An open-source re-implementation of Wang et al.’s BLSTM tagging architecture is used for our experiments. We configure a hidden layer size of 100, embedding dimensions of 300, a maximum training sequence length of 100, and a batch size of 32. We optimize with RMSprop, a variation of RProp (Riedmiller and Braun, 1993), at a learning rate of 0.001. A sample sized 20% of the training data is removed and used for validation. To counteract overfitting, we store model checkpoints and do early stopping.

**Word embeddings.** Word embeddings help render more information regarding the word since they carry semantic and syntactic information and capture the meaning of words, the relationship between words, and the context of different words. This is useful for tagging and many other tasks in natural language processing (Plank et al., 2016; Wiegandt et al., 2017).

Pre-trained embeddings used in this work were downloaded from an open repository provided by Facebook Research. These embeddings were trained with fastText on Wikipedia using the skip-gram model with default parameters (Bojanowski et al., 2017).

5.1.2 Frequency Tagger

We additionally built a simple Frequency tagger that annotates each word in isolation with its most frequent tag. The only calculations that are required are tag counts per word in the training data. As soon as the occurrences are counted, the Frequency tagger is ready to annotate sentences. OOVs are tagged with the majority class, which in both languages is NOUN. There are 3771 words in the Ukrainian test set that are new to the Frequency tagger, which means that 25.8% of the text cannot be tagged based on evidence. In Russian,
Table 4: Overview of the conducted tagging experiments on Russian test data (and trained on Russian data): PE - pre-trained embeddings; RE - randomly initialized embeddings; Stem - stemming.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency Tagger</td>
<td>90.7 %</td>
</tr>
<tr>
<td>BLSTM + RE</td>
<td>91.3 %</td>
</tr>
<tr>
<td>BLSTM + PE</td>
<td>94.4 %</td>
</tr>
<tr>
<td>Stem BLSTM + RE</td>
<td>92.3 %</td>
</tr>
</tbody>
</table>

Table 5: Overview of the conducted tagging experiments on Ukrainian test data.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency Tagger</td>
<td>81.6 %</td>
</tr>
<tr>
<td>BLSTM + RE</td>
<td>80.0 %</td>
</tr>
<tr>
<td>BLSTM + PE</td>
<td>85.4 %</td>
</tr>
<tr>
<td>Self-trained BLSTM + PE</td>
<td>86.2 %</td>
</tr>
<tr>
<td>Stem BLSTM + RE</td>
<td>84.1 %</td>
</tr>
<tr>
<td>trained on Ukrainian data</td>
<td></td>
</tr>
<tr>
<td>Zero-shot BLSTM + PE</td>
<td>51.5 %</td>
</tr>
<tr>
<td>Zero-shot Stem BLSTM + RE</td>
<td>56.1 %</td>
</tr>
<tr>
<td>Cross-lingual Transfer</td>
<td>84.4 %</td>
</tr>
<tr>
<td>Multilingual BLSTM + PE</td>
<td>86.4 %</td>
</tr>
<tr>
<td>Multilingual Stem BLSTM + RE</td>
<td>87.3 %</td>
</tr>
</tbody>
</table>

Table 6: Prediction quality per part-of-speech (of the Ukrainian BLSTM + PE tagging model).

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>~99%</td>
<td>NOUN, PUNCT</td>
</tr>
<tr>
<td>91-99%</td>
<td>PRON, CCONJ, SCONJ, AUX, ADP</td>
</tr>
<tr>
<td>71-90%</td>
<td>PART, ADV, NUM, DET</td>
</tr>
<tr>
<td>51-70%</td>
<td>–</td>
</tr>
<tr>
<td>41-50%</td>
<td>VERB, PROPN, SYM</td>
</tr>
<tr>
<td>31-40%</td>
<td>ADJ, INTJ</td>
</tr>
<tr>
<td>11-30%</td>
<td>–</td>
</tr>
<tr>
<td>&lt;10%</td>
<td>X</td>
</tr>
</tbody>
</table>

5.2 Experimental Results

We now present the results for all the investigated techniques. The tagging accuracies for all experiments on the Ukrainian test set are collectively shown in Table 5. Some further empirical observations, e.g. on the taggers’ ability to correctly handle OOV words, will also be discussed below. Supplementary tagging accuracies of Russian POS-tagger measured on a Russian test set are reported in Table 4.

5.2.1 Low-resource Supervision Results

The baseline taggers (Frequency and BLSTM) for Russian and for Ukrainian are trained on annotated UD data for the respective language. For the BLSTM models, there are two flavors: one with randomly initialized embeddings (BLSTM + RE) and one with pre-trained word embeddings (BLSTM + PE). The Russian BLSTM taggers are built with the exact same hyperparameters as the Ukrainian BLSTM taggers but show better results in the evaluation. This is because the Russian model is trained on more data.

On Ukrainian, the BLSTM with randomly initialized embeddings (RE) achieves better results on tag prediction for OOVs than the Frequency tagger (53% vs. 40% correct), but surprisingly does not outperform the Frequency tagger in overall accuracy (BLSTM + RE: 80.0%, Frequency: 81.6%; Table 5). However, the use of pre-trained embeddings in the BLSTM model increases the overall accuracy by about +5% absolute (BLSTM + PE: 85.4%). OOV tag prediction is boosted further to 58% of unknowns correctly labeled. Table 6 shows the prediction quality per individual POS of the Ukrainian BLSTM + PE model. 11 out of 17 tags are predicted with accuracies above 70%. The most inaccurate predictions are made for the X tag which is used for cases of code-switching. Since the tag is used when it is not possible (or meaningful) to analyze the word, it is difficult for a neural network to learn to recognize it without additional features.

5.2.2 Self-Training Results

In self-training, the existing model first labels unlabeled data. We apply our BLSTM + PE model to automatically tag the Ukrainian side of the Russian–Ukrainian parallel corpus. This step provides us with new synthetically annotated data, which is then treated as truth and appended to the original training corpus to re-train the tagger.

The tagger trained with additional synthetically annotated data improves just moderately over the tagger trained on only the hand-labeled UD corpus (86.2% vs. 85.4% overall accuracy; Table 5). Self-training is thus barely effective despite the 20-fold augmentation of training instances through...
the synthetic corpus. Clark et al. (2003) have previously reported similar findings. In the literature, inefficacy of self-training is occasionally attributed to a domain mismatch of the synthetically annotated data. In our case, all corpora are from the same domain (news text), though. The main benefit of self-training that we observe is an increase of correctly tagged OOVs (of around +5% absolute, from 58% to 63%).

5.2.3 Multilingual Learning Results

The multilingual learning approach yields an improvement of one percentage point (86.4% accuracy) compared to the low-resource BLSTM + PE tagger trained on only the Ukrainian data.

We oversampled the Ukrainian corpus to balance out the fraction of data from each language and avoid a bias towards Russian. The Ukrainian data was copied and added to the mixed training set until it reached the size of the Russian data. We also tried undersampling of Russian data and plain concatenation. The differences in tagging accuracy were minor (undersampling: 86.0%, concatenation: 86.2%), but oversampling of Ukrainian worked best.

5.2.4 Zero-shot Tagging Results

In the zero-shot tagging experiment, the BLSTM model trained on the Russian UD corpus (with pre-trained word embeddings) is applied to the Ukrainian test set. The Russian model’s accuracy on the Russian test set had reached 94.4% (Table 4). Yet, when being run on the related Ukrainian language, just over 50% of Ukrainian words are correctly annotated by the Russian tagger (Table 5). This cannot be considered a satisfactory outcome.

5.2.5 Cross-lingual Transfer Results

The idea of the cross-lingual transfer is to project tags from the annotated part of the parallel corpus to its unlabeled translation to produce training data for the under-resourced language. The success of cross-lingual transfer depends not only on the quality of the source language annotation, but also on the reliability of the annotation projection.

We rely on standard statistical word alignment algorithms (Brown et al., 1993) as the basis of POS annotation projection from Russian to Ukrainian. The parallel corpus is aligned with fast_align, an unsupervised word aligner introduced by Dyer et al. (2013). For phrase-based machine translation, the two alignment directions (forward and reverse) are typically combined to a symmetrized alignment. But for annotation projection, it is more convenient to use one-directional alignment with one Ukrainian token never being aligned to multiple tokens on the Russian side. The annotation projection across the alignment then becomes straightforward. No disambiguation heuristics are necessary, which could be a source of additional errors.

The BLSTM tagger supervised with gold-standard Ukrainian annotation (Section 5.2.1) outperforms the cross-lingual transfer tagger by only one percentage point (Table 5), despite the latter not requiring and not using any manually annotated Ukrainian training data. The confusion matrix heatmaps in Figure 2 visually illustrate the situation.
periority of cross-lingual transfer over zero-shot tagging, and how the two compare to the low-resource supervision baseline BLSTM. The result highlights that a competitive neural tagger can be trained even under zero-resource conditions. A parallel corpus with a related language and the existence of a tagger for that related language enable effective cross-lingual transfer. A BLSTM model trained on projected annotation seems to cope very well with the language transfer.

5.2.6 Stemming Results

In Section 4.3 it was demonstrated that the number of common words grew after stemming was applied. We now test whether stemming has a positive impact on tagging quality. Since the pre-trained word embeddings were trained on full word surface forms, the embeddings for these experiments are randomly initialized.

The Stem BLSTM + RE result in Table 4 shows that compared to the previous taggers trained on random embeddings, the accuracy for Russian grows by about one percentage point. There are even bigger improvement for the Ukrainian tagger, which reaches 84.1% accuracy (Table 5).

Stemming benefits the performance of the POS-tagger, since the number of unknown tokens in the test data is reduced. The number of OOVs that are tagged correctly in Ukrainian increases to 61% from the initial 53%. The error rate among the known vocabulary is reduced by 2% absolute compared to the non-stemmed model.

Applying the Russian stem POS-tagger to the Ukrainian stemmed test set results in a nice accuracy improvement (about +4%) over the previous zero-shot attempt on full word forms. The zero-shot tagging quality remains weak, though, even with stemming.

In order to also examine the multilingual learning strategy over stem forms, the last model in this series of experiments is trained on concatenated stemmed Ukrainian and stemmed Russian data. The model achieves about +1% absolute improvement compared to the previous best result for Ukrainian. Tagging accuracy is reaching 87.3%, beating the result with the model trained on full forms of the same concatenation of corpora. We found that the stem system version is actually slightly worse at predicting tags of known Ukrainian words, but OOVs are handled much better (69% vs. 57% correct tags for unseen Ukrainian words).

6 Summary of Findings

The observations that have been made in the course of this work can be briefly summarized as follows: 1) Pre-trained word embeddings are important for better tagging quality since they represent contextual similarities between words. 2) A semi-supervised approach (self-training) showed only moderate gains despite a notable increase of the training corpus with synthetically labeled data. 3) Mixing larger related-language annotated data into the training corpus (multilingual learning) slightly improved the tagging accuracy for the low-resource language. 4) Applying a Russian tagger on Ukrainian (zero-shot) did not show satisfactory results, which could be due to the relatively small amount of shared vocabulary and certain differences in grammar. 5) Given a parallel corpus, a competitive neural POS-tagger can be trained without any initial annotated data (using cross-lingual transfer via annotation projection), which can be viewed as a good solution in the zero-resource scenario. 6) Bridging words by reducing them to their stems has a positive influence since both languages are highly inflected. The number of types is lowered and the tagger can abstract from the sparsity of inflected surface forms.

The best accuracy for Ukrainian (87.3%) was achieved when a multilingual model was trained on both Russian and Ukrainian stemmed training corpora. Potentially, through a combination of stemmed words and pre-trained stem embeddings, further improvements could be attained. For the important zero-resource scenario, cross-lingual projection worked best, and we achieved an accuracy rate of 84.4%. Here there is likely to be room for further improvement by tailoring the word alignment more to the task.

7 Conclusion

We carried out an evaluation on Ukrainian neural POS-tagging for both low-resource and zero-resource scenarios. For low-resource, multilingual learning works best, suggesting that even for languages which do have some gold-standard POS training data, multilingual learning through combining the training data with data from closely related languages is of strong interest. For zero-resource, cross-lingual annotation projection works best, suggesting that where parallel corpora with a related language are available, cross-lingual projection should be strongly considered.
Acknowledgment

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement № 640550).

References


Georg Heigold, Stalin Varanasi, Günter Neumann, and Josef Genabith. 2018. How Robust Are Character-Based Word Embeddings in Tagging and


