Lexical Resources for Low-Resource PoS Tagging in Neural Times

Barbara Plank
Department of Computer Science
ITU, IT University of Copenhagen
Denmark
bplank@itu.dk

Sigrid Klerke
Department of Computer Science
ITU, IT University of Copenhagen
Denmark
sikl@itu.dk

Abstract

More and more evidence is appearing that integrating symbolic lexical knowledge into neural models aids learning. This contrasts the widely-held belief that neural networks largely learn their own feature representations. For example, recent work has shown benefits of integrating lexicons to aid cross-lingual part-of-speech (PoS). However, little is known on how complementary such additional information is, and to what extent improvements depend on the coverage and quality of these external resources. This paper seeks to fill this gap by providing a thorough analysis on the contributions of lexical resources for cross-lingual PoS tagging in neural times.

1 Introduction

In natural language processing, the deep learning revolution has shifted the focus from conventional hand-crafted symbolic representations to dense inputs, which are adequate representations learned automatically from corpora. However, particularly when working with low-resource languages, small amounts of symbolic lexical resources such as user-generated lexicons are often available even when gold-standard corpora are not. Recent work has shown benefits of combining conventional lexical information into neural cross-lingual part-of-speech (PoS) tagging (Plank and Agić, 2018). However, little is known on how complementary such additional information is, and to what extent improvements depend on the coverage and quality of these external resources.

The contribution of this paper is in the analysis of the contributions of models' components (tagger transfer through annotation projection vs. the contribution of encoding lexical and morphosyntactic resources). We seek to understand under which conditions a low-resource neural tagger benefits from external lexical knowledge. In particular:

a) we evaluate the neural tagger across a total of 20+ languages, proposing a novel baseline which uses retrofitting;
b) we investigate the reliance on dictionary size and properties;
c) we analyze model-internal representations via a probing task to investigate to what extent model-internal representations capture morphosyntactic information.

Our experiments confirm the synergetic effect between a neural tagger and symbolic linguistic knowledge. Moreover, our analysis shows that the composition of the dictionary plays a more important role than its coverage.

2 Methodology

Our base tagger is a bidirectional long short-term memory network (bi-LSTM) (Graves and Schmidhuber, 2005; Hochreiter and Schmidhuber, 1997; Plank et al., 2016) with a rich word encoding model which consists of a character-based bi-LSTM representation $c\vec{w}$ paired with pre-trained word embeddings $\vec{w}$. Sub-word and especially character-level modeling is currently pervasive in top-performing neural sequence taggers, owing to its capacity to effectively capture morphological features that are useful in labeling out-of-vocabulary (OOV) items. Sub-word information is often coupled with standard word embeddings to mitigate OOV issues. Specifically, i) word embeddings are typically built from massive unlabeled datasets and thus OOVs are less likely to be encountered at test time, while ii) character embeddings offer further linguistically plausible fallback for the remaining OOVs through modeling intra-word relations. Through these approaches, multilingual PoS tagging has seen tangible gains from neural methods in the recent years.

2.1 Lexical resources

We use linguistic resources that are user-generated and available for many languages. The first is WIKTIONARY, a word type dictionary that maps words to one of the 12 Universal PoS tags (Li et al., 2012; Petrov et al., 2012). The second resource is UNIMORPH, a morphological dictionary that provides inflectional paradigms for 350 languages (Kirov et al., 2016). For Wiktionary, we use the freely available dictionaries from Li et al. (2012). UniMorph covers between 8-38 morphological properties (for English and Finnish, respectively).[1] The sizes of the dictionaries vary considerably, from a few thousand entries (e.g., for Hindi and Bulgarian) to 2M entries (Finnish UniMorph). We study the impact of smaller dictionary sizes in Section 4.1.

The tagger we analyze in this paper is an extension of the base tagger, called *distant supervision from disparate sources* (DSDS) tagger (Plank and Agić, 2018). It is trained on projected data and further differs from the base tagger by the integration of lexicon information. In particular, given a lexicon src, DSDS uses $\vec{e}_{src}$ to embed the lexicon into an l-dimensional space, where $\vec{e}_{src}$ is the concatenation of all embedded m properties of length l (empirically set, see Section 2.2), and a zero vector for words not in the lexicon. A property here is a possible PoS tag (for Wiktionary) or a morphological feature (for Unimorph). To integrate the type-level supervision, the lexicon embeddings vector is created and concatenated to the word and character-level representations for every token: $\vec{w} \circ \vec{cw} \circ \vec{e}$.

We compare DSDS to alternative ways of using lexical information. The first approach uses lexical information directly during decoding (Täckström et al., 2013). The second approach is more implicit and uses the lexicon to induce better word embeddings for tagger initialization. In particular, we use the dictionary for retrofitting off-the-shelf embeddings (Faruqui et al., 2015) to initialize the tagger with those. The latter is a novel approach which, to the best of our knowledge, has not yet been evaluated in the neural tagging literature. The idea is to bring the off-the-shelf embeddings closer to the PoS tagging task by retrofitting the embeddings with syntactic clusters derived from the lexicon.

We take a deeper look at the quality of the lexicons by comparing tag sets to the gold treebank data, inspired by Li et al. (2012). In particular, let T be the dictionary derived from the gold treebank (development data), and W be the user-generated dictionary, i.e., the respective Wiktionary (as we are looking at PoS tags). For each word type, we compare the tag sets in T and W and distinguish six cases:

1. NONE: The word type is in the training data but not in the lexicon (out-of-lexicon).
2. EQUAL: $W = T$
3. DISJOINT: $W \cap T = \emptyset$
4. OVERLAP: $W \cap T \neq \emptyset$
5. SUBSET: $W \subset T$
6. SUPERSET: $W \supset T$

In an ideal setup, the dictionaries contain no disjoint tag sets, and larger amounts of equal tag sets or superset of the treebank data. This is particularly desirable for approaches that take lexical information as type-level supervision.

2.2 Experimental setup

In this section we describe the baselines, the data and the tagger hyperparameters.

Data We use the 12 Universal PoS tags (Petrov et al., 2012). The set of languages is motivated by accessibility to embeddings and dictionaries. We here focus on 21 dev sets of the Universal Dependencies 2.1 (Nivre and et al., 2017), test set results are reported by Plank and Agić (2018) showing that DSDS provides a viable alternative.

Annotation projection To build the taggers for new languages, we resort to annotation projection following Plank and Agić (2018). In particular, they employ the approach by Agić et al. (2016), where labels are projected from multiple sources to multiple targets and then decoded through weighted majority voting with word alignment probabilities and source PoS tagger confidences. The wide-coverage Watchtower corpus (WTC) by Agić et al. (2016) is used, where 5k instances are selected via data selection by alignment coverage following Plank and Agić (2018).

Baselines We compare to the following alternatives: type-constraint Wiktionary supervision (Li et al., 2012) and retrofitting initialization.

[1] More details: `http://unimorph.org/`

	DEV SETS (UD2.1)			
LANGUAGE	5k	TC_W	RETRO	DSDS
Bulgarian (bg)	89.8	89.9	87.1	**91.0**
Croatian (hr)	84.7	85.2	83.0	**85.9**
Czech (cs)	**87.5**	**87.5**	84.9	87.4
Danish (da)	89.8	89.3	88.2	**90.1**
Dutch (nl)	88.6	89.2	86.6	**89.6**
English (en)	86.4	**87.6**	82.5	87.3
Finnish (fi)	81.7	81.4	79.2	**83.1**
French (fr)	**91.5**	90.0	89.8	91.3
German (de)	85.8	87.1	84.7	**87.5**
Greek (el)	80.9	**86.1**	79.3	79.2
Hebrew (he)	75.8	75.9	71.7	**76.8**
Hindi (hi)	63.8	63.9	63.0	**66.2**
Hungarian (hu)	**77.5**	**77.5**	75.5	76.2
Italian (it)	92.2	91.8	90.0	**93.7**
Norwegian (no)	91.0	91.1	88.8	**91.4**
Persian (fa)	43.6	43.8	**44.1**	43.6
Polish (pl)	84.9	84.9	83.3	**85.4**
Portuguese	92.4	92.2	88.6	**93.1**
Romanian (ro)	84.2	84.2	80.2	**86.0**
Spanish (es)	90.7	88.9	88.9	**91.7**
Swedish (sv)	89.4	89.2	87.0	**89.8**
AVG(21)	83.4	83.6	81.3	**84.1**
GERMANIC (6)	88.5	88.9	86.3	**89.3**
ROMANCE (5)	90.8	90.1	88.4	**91.4**
SLAVIC (4)	86.7	86.8	84.6	**87.4**
INDO-IRANIAN (2)	53.7	53.8	53.5	**54.9**
URALIC (2)	79.6	79.4	79.2	**79.6**

Table 1: Replication of results on the dev sets. 5k: model trained on only projected data; TC_W: type constraints; Retro: retrofitted initialization.

Hyperparameters We use the same setup as Plank and Agić (2018), i.e., 10 epochs, word dropout rate (p=.25) and l=40-dimensional lexicon embeddings for DSDS, except for downscaling the hidden dimensionality of the character representations from 100 to 32 dimensions. This ensures that our probing tasks always get the same input dimensionality: 64 (2x32) dimensions for $c\vec{w}$, which is the same dimension as the off-the-shelf word embeddings. Language-specific hyperparameters could lead to optimized models for each language. However, we use identical settings for each language which worked well and is less expensive, following Bohnet et al. (2018). For all experiments, we average over 3 randomly seeded runs, and provide mean accuracy.

We use the off-the-shelf Polyglot word embeddings (Al-Rfou et al., 2013). Word embedding initialization provides a consistent and considerable boost in this cross-lingual setup, up to 10% absolute improvements across 21 languages when only 500 projected training instances are available (Plank and Agić, 2018). Note that we empirically find it to be best to *not* update the word embeddings in this noisy training setup, as that results in better performance, see Section 4.4.

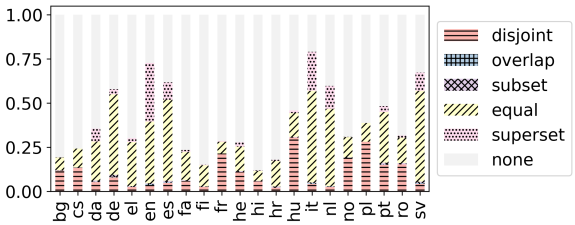


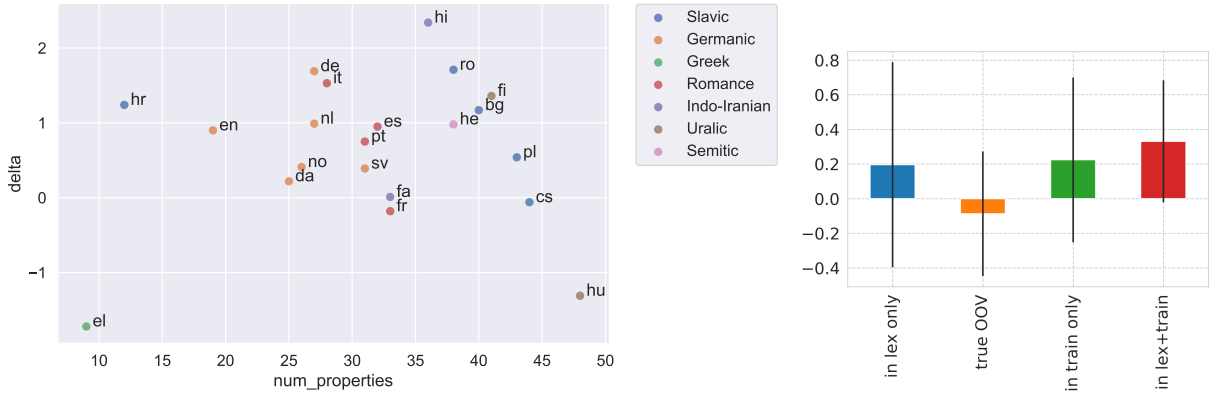
Figure 1: Analysis of Wiktionary vs gold (dev set) tag sets. 'None': percentage of word types not covered in the lexicon. 'Disjoint': the gold data and Wiktionary do not agree on the tag sets. See Section 2.1 for details on other categories.

3 Results

Table 1 presents our replication results, i.e., tagging accuracy for the 21 individual languages, with means over all languages and language families (for which at least two languages are available). There are several take-aways.

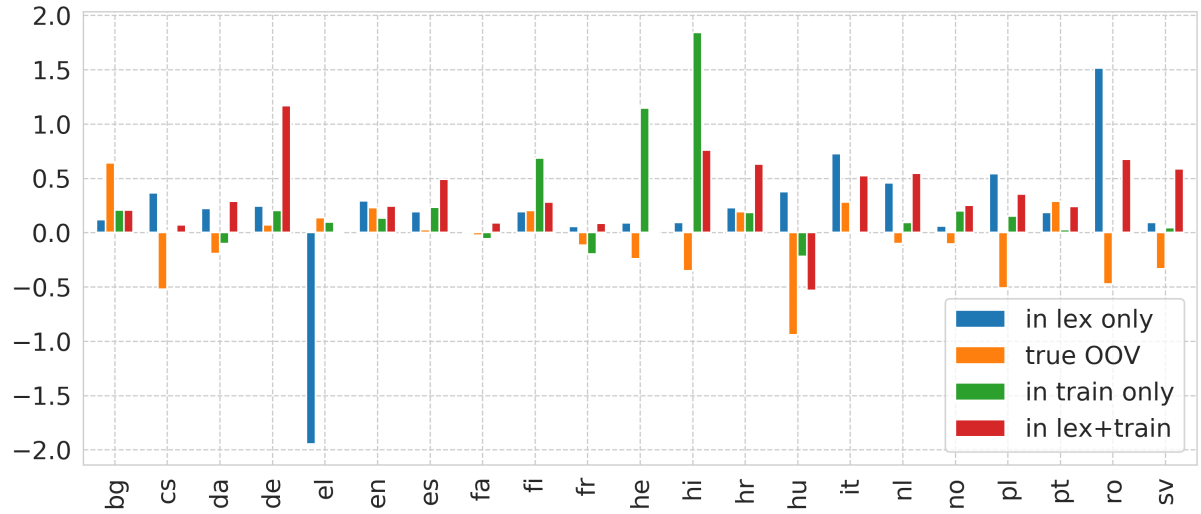
Inclusion of lexical information Combining the best of two worlds results in the overall best tagging accuracy, confirming Plank and Agić (2018): Embedding lexical information into a neural tagger improves tagging accuracy from 83.4 to 84.1 (means over 21 languages). On 15 out of 21 languages, DSDS is the best performing model. On two languages, type constraints work the best (English and Greek). Retrofitting performs best only on one language (Persian); this is the language with the overall lowest performance. On three languages, Czech, French and Hungarian, the baseline remains the best model, none of the lexicon-enriching approaches works. We proceed to inspect these results in more detail.

Analysis Overall, type-constraints improve the baseline but only slightly (83.4 vs 83.6). Intuitively, this more direct use of lexical information requires the resource to be high coverage and a close fit to the evaluation data, to not introduce too many pruning errors during decoding due to contradictory tag sets. To analyze this, we look at the tag set agreement in Figure 1. For languages for which the level of *disjoint* tag set information is low, such as Greek, English, Croatian, Finnish and Dutch, type constraints are expected to help. This is in fact the case, but there are exceptions,



(a) Absolute improvement (delta) vs number of dictionary properties (ρ=0.08).

(b) Absolute improvement per OOV category (21 languages).



(c) Per language analysis: absolute improvements of DsDs over the baseline for words in the lexicon, in the training data, in both or in neither (true OOVs).

Figure 2: Analysis of OOVs and dictionary properties.

such as Finnish. Coverage of the lexicon is also important, and for this morphologically rich language, the coverage is amongst the lowest (c.f. large amount of the 'none' category in Figure 1).

The more implicit use of lexical information in DsDs helps on languages with relatively high dictionary coverage and low tag set disagreement, such as Danish, Dutch and Italian. Compared to type constraints, embedding the lexicon also helps on languages with low dictionary coverage, such as Bulgarian, Hindi, Croatian and Finnish, which is very encouraging and in sharp contrast to type constraints. The only outlier remains Greek.

Figure 2 (a) plots the absolute improvement in tagging accuracy over the baseline versus the number of properties in the dictionaries. Slavic and Germanic languages cluster nicely, with some outliers (Croatian). However, there is only a weak positive correlation (ρ=0.08). More properties do not necessarily improve performance, and lead to sparsity. The inclusion of the lexicons results in higher coverage, which might be part of the explanation for the improvement of DsDs. The question remains whether the tagger learns to rely only on this additional signal, or it generalizes beyond it. Therefore, we first turn to inspecting out-of-vocabulary (OOV) items. OOV items are the key challenge in part-of-speech tagging, i.e., to correctly tag tokens unseen in the training data.

In Figure 2 (b) and (c), we analyze accuracy improvements on different groups of tokens: The *in lex+train* tokens that were seen both in the lexicon and the training data, the *in train only* tokens seen in the training data but not present in the lexicon, the *in lex only* tokens that were present in the lexicon but not seen in the training data and the *true OOV* tokens that were neither seen in training nor present in the lexicon. Figure 2 (b) shows means

over the 21 languages, Figure 2 (c) provides details per language. The first take-away is that in many cases the tagger does learn to use information beyond the coverage of the lexicon. The embedded knowledge helps the tagger to improve on tokens which are in train only (and are thus not in the lexicon, green bars). For true OOVs (orange bars), this is the case for some languages as well Figure 2 (c), i.e., improvements on true OOVs can be observed for Bulgarian, German, Greek, English, Finish, Croatian, Italian and Portuguese. Over all 21 languages there is a slight drop on true OOVs: -0.08, but this is a mean over all languages, for which results vary, making it important to look beyond the aggregate level. Over all languages except for Hungarian, the tagger, unsurprisingly, improves over tokens which are both in the lexicon and in the training data (see further discussion in Section 4).

4 Discussion

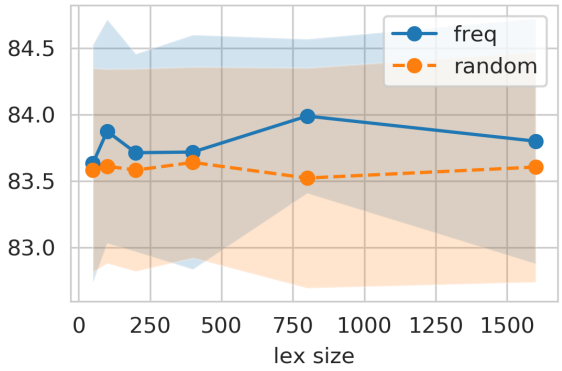
Here we dig deeper into the effect of including lexical information by a) examining learning curves with increasing dictionary sizes, b) relating tag set properties to performance, and finally c) having a closer look at model internal representations, by comparing them to the representations of the base model that does not include lexical information. We hypothesize that when learning from dictionary-level supervision, information is propagated through the representation layers so as to generalize beyond simply relying on the respective external resources.

4.1 Learning curves

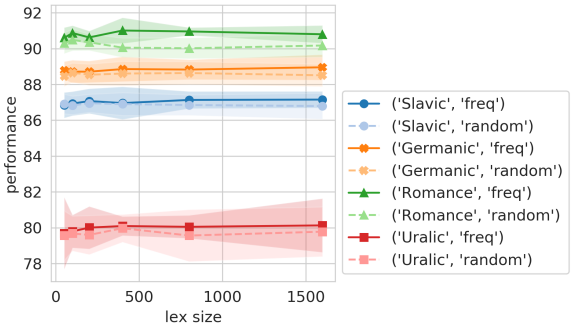
The lexicons we use so far are of different sizes (shown in Table 1 of Plank and Agić (2018)), spanning from 1,000 entries to considerable dictionaries of several hundred thousands entries. In a low-resource setup, large dictionaries might not be available. It is thus interesting to examine how tagging accuracy is affected by dictionary size. We examine two cases: randomly sampling dictionary entries and sampling by word frequency, over increasing dictionary sizes: 50, 100, 200, 400, 800, 1600 word types. The latter is motivated by the fact that an informed dictionary creation (under limited resources) might be more beneficial. We estimate word frequency by using the UD training data sets (which are otherwise not used).

Figure 3 (a) provides means over the 21 languages (with confidence intervals of ± 1 standard deviation based on three runs). We note that sampling by frequency is overall more beneficial than random sampling. The biggest effect of sampling by frequency is observed for the Romance language family, see Figure 3 (b). It is noteworthy that more dictionary data is not always necessarily beneficial. Sometimes a small but high-frequency dictionary approximates the entire dictionary well. This is for instance the case for Danish, where sampling by frequency approximates the entire dictionary well ('all' achieves 90.1, while using 100 most frequent entries is close: 89.93). Frequency sampling also helps clearly for Italian, but here having the entire dictionary results in the overall highest performance.

For some languages, the inclusion of lexical information does not help, not even at smaller dictionary sizes. This is the case for Hungarian, French and Czech. For Hungarian using the entire dictionary drops performance below the baseline. For Czech, this is less pronounced, as the performance stays around baseline. Relating these negative ef-



(a) Average effect over 21 languages of high-freq and random dictionaries



(b) Effect for subset of language families of high-freq and random dictionaries

Figure 3: Learning curves over increased dictionary sizes.

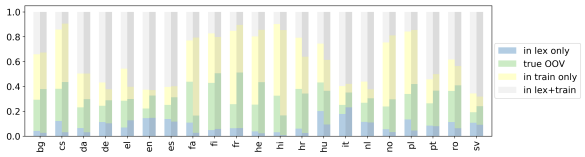


Figure 4: Proportion of tokens unseen in the training data, in the lexicon or in both (true OOV's). Lighter bars are proportion of correctly labeled portion, dark bars are proportion of errors.

fects to the results from the tag set agreement analysis (Figure 1), we note that Hungarian is the language with the largest *disjoint* tag set. Albeit the coverage for Hungarian is good (around .5), including too much contradictory tag information has a clear deteriorating effect. Consequently, neither sampling strategy works. Czech, which has less coverage, sees a negative effect as well: half of the dictionary entries have disjoint tag sets. Italian is the language with the highest dictionary coverage *and* the highest proportion of equal tag sets, thereby providing a large positive benefit.

We conclude that when dictionaries are not available, creating them by targeting high-frequency items is a pragmatic and valuable strategy. A small dictionary, which does not contain too contradictory tag sets, can be beneficial.

4.2 Analysis of correct/incorrect predictions

In the following we analyze correctly and incorrectly labeled tokens. Because we are analyzing differences between languages as well as between errors and successes we abstract away from the underlying sample size variation by comparing proportions.

The analysis inspects the differences in proportions on four subsections of the development set, as introduced above: the *in lex+train* tokens, the *in train only* tokens, the *in lex only* tokens and the *true OOVs*. The proportion of these four data subsets in the correctly and the incorrectly labeled tokens are shown side by side in Figure 4 in lighter and darker shades, respectively. If the OOV-status of a word was unrelated to performance, the lighter and darker bars would be of identical size. This is not the case and we can observe that the true OOVs make up a significantly larger share of the errors than of successes (two-tailed paired Student's t-test: $p = 0.007$). Similarly, seen across all languages the shift in the size of the proportion of true OOVs is made up by more correct labeling of a larger proportion of *in train only* (two-tailed paired Student's t-test: $p = 0.014$) and *in lex only* (two-tailed paired Student's t-test: $p = 0.020$), whereas the proportion of *in lex+train* does not significantly differ between the correctly and incorrectly labeled parts (two-tailed paired Student's t-test: $p = 0.200$).[2]

4.3 Probing word encodings

Probing tasks, or diagnostic classifiers, are separate classifiers which use representations extracted from any facet of a trained neural model as input for solving a separate task. Following the intuition of Adi et al. (2017), if the target can be predicted, then the information must be encoded in the representation. However, the contrary does not necessarily hold: if the model fails it does not necessarily follow that the information is not encoded, as opposed to not being encoded in a useful way for a probing task classifier.

As the internal representations stored in neural models are not immediately interpretable, probing tasks serve as a way of querying neural representations for interpretable information. The probing task objective and training data is designed to model the query of interest. The representation layer we query in this work is the word-level output from the character embedding sub-model. This part of the word-level representation starts out uninformative and thus without prior prediction power on the classifier objectives.

The pre-trained word embeddings stay fixed in our model (see Section 4.4). However, the character-based word encodings get updated: This holds true both for the BASE system and the DSDS tagger. As a target for assessing the flow of information in the neural tagger, we thus focus on the character-based word encodings.

The word-level is relevant as it is the granularity at which the tagger is evaluated. The word embeddings may already have encoded PoS-relevant information and the lexicon embeddings explicitly encodes PoS-type-level information. By contrast, the character-based word encodings are initialized to be uninformative and any encoding of PoS-related information is necessarily a result of the neural training feedback signal.

For these reasons we query the character-based word representations of the tagger in order to com-

[2]Significance based on an α-level of 0.05

pare variation between the base tagger and the DsDs lexicon-enriched architecture.

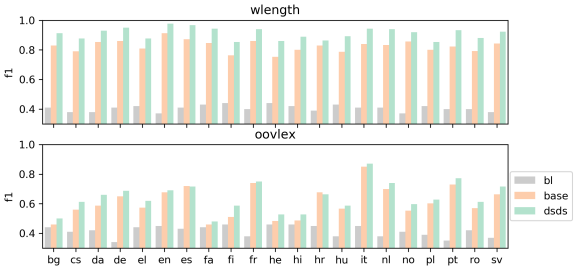


Figure 5: Macro F1 scores for stand-alone classifiers on the probing tasks of predicting which words are long and which are in the lexicon, respectively. The baseline (bl) is a simple majority baseline. The base- and DsDs-informed classifiers were trained on character-based word representations from the neural taggers with and without access to lexical information, respectively.

We employ two binary probing tasks: predicting which words are long, i.e., contain more than 7 characters[3], and predicting which words are in the lexicon. The word length task is included as a task which can be learned independently of whether lexicon information is available to the neural model. Storing length-related information might help the model distinguish suffix patterns of relevance to PoS-tagging.

Following Shi et al. (2016) and Gulordava et al (2018), we use a logistic regression classifier setup and a constant input dimensionality of 64 across tasks (Conneau et al., 2018). The classifiers are trained using 10-fold cross-validation for each of three trained runs of each neural model and averaged. We include a majority baseline and report macro F1-scores, as we are dealing with imbalanced classes. The training vocabulary of both probing tasks is restricted to the neural tagger training vocabulary, that is, all word types in the projected training data, as these are the representations which have been subject to updates during training of the neural model. Using the projected data has the advantage that the vocabulary is similar across languages as the data comes from the same domain (Watchtower).

[3]Considering words of 7 characters or more to be long is based on the threshold that was experimentally tuned in the design of the readability metric LIX (Björnsson, 1983). This threshold aligns well with the visual perceptual span within which proficient readers from grade four and up can be expected to automatically decode a word in a single fixation (Sperlich et al., 2015)

The results on the word length probing task shown on the top half of Figure 5 confirm that information relevant to distinguishing word length is being encoded in the neural representation, as expected. It is intriguing that the lexicon-informed DsDs representation encodes this information even at higher degree.

On the task of classifying which words are in the lexicon, all neural representations beat the majority baseline, but we also see that this task is harder, given the higher variance across languages. With Spanish (es) and Croatian (hr) as the only exceptions, the DsDs-based representations are generally encoding more of the information relevant to distinguishing which words are in the lexicon, confirming our intuitions that the internal representations were altered. Note, however, that even the base-tagger is able to solve this task above chance level. This is potentially an artifact of how lexicons grow where it would be likely for several inflections of the same word to be added collectively to the lexicon at once, and since the character representations can be expected to produce more similar representations of words derived from the same lemma the classifier will be able to generalize and perform above chance level without the base-model representations having ever been exposed to the lexical resource.

4.4 Updating in light of noisy data?

When training a tagger with noisy training data and pre-trained embeddings, the question arises whether it is more beneficial to freeze the word embeddings or update them. We hypothesize that freezing embeddings is more beneficial in noisy training cases, as it helps to stabilize the signal from the pre-trained word embeddings while avoiding updates from the noisy training data. To test this hypothesis, we train the base tagger on high-quality gold training data (effectively, the UD training data sets), with and without freezing the word embeddings layer. We find that updating the word embedding layer is in fact beneficial in the high-quality training data regime: on average +0.4% absolute improvement is obtained (mean over 21 languages). This is in sharp contrast to the noisy training data regime, in which the baseline accuracy drops by as much as 1.2% accuracy. Therefore, we train the tagger with pre-trained embeddings on projected WTC data and freeze the word embeddings lookup layer during training.

5 Related work

In recent years, natural language processing has witnessed a move towards deep learning approaches, in which automatic representation learning has become the de facto standard methodology (Collobert et al., 2011; Manning, 2015).

One of the first works that combines neural representations with semantic symbolic lexicons is the work on *retrofitting* (Faruqui et al., 2015). The main idea is to use the relations defined in semantic lexicons to refine word embedding representations, such that words linked in the lexical resource are encouraged to be closer to each other in the distributional space.

The majority of recent work on neural sequence prediction follows the commonly perceived wisdom that hand-crafted features are obsolete for deep learning methods. They rely on end-to-end training without resorting to additional linguistic resources. Our study contributes to the increasing literature to show the utility of linguistic resources for deep learning models by providing a deep analysis of a recently proposed model (Plank and Agić, 2018). Most prior work in this direction can be found on machine translation (Sennrich and Haddow, 2016; Chen et al., 2017; Li et al., 2017; Passban et al., 2018), work on named entity recognition (Wu et al., 2018) and PoS tagging (Sagot and Martínez Alonso, 2017) who use lexicons, but as n-hot features and without examining the cross-lingual aspect.

Somewhat complementary to evaluating the utility of linguistic resources empirically is the increasing body of work that uses linguistic insights to try to understand what properties neural-based representations capture (Kádár et al., 2017; Adi et al., 2017; Belinkov et al., 2017; Conneau et al., 2018; Hupkes et al., 2018). Shi et al. (2016) and Adi et al. (2017) introduced the idea of probing tasks (or 'diagnostic classifiers'), see Belinkov and Glass for a recent survey (Belinkov and Glass, 2019). Adi et al. (2017) evaluate several kinds of sentence encoders and propose a range of probing tasks around isolated aspects of sentence structure at the surface level (sentence length, word content and word order). This work has been greatly expanded by including both syntactic and semantic probing tasks, careful sampling of probing task training data, and extending the framework to make it encoder agnostic (Conneau et al., 2018). A general observation here is that task-specific knowledge is needed in order to design relevant diagnostic tasks, which is not always straightforward. For example, Gulordava (2018) investigate whether RNNs trained using a language model objective capture hierarchical syntactic information. They create nonsensical construction so that the RNN cannot rely on lexical or semantic clues, showing that RNNs still capture syntactic properties in sentence embeddings across the four tested languages while obfuscating lexical information. There is also more theoretical work on investigating the capabilities of recurrent neural networks, e.g., Weiss et al. (2018) show that specific types of RNNs (LSTMs) are able to use counting mechanisms to recognize specific formal languages.

Finally, linguistic resources can also serve as proxy for evaluation. As recently shown (Agić et al., 2017), type-level information from dictionaries approximates PoS tagging accuracy in the absence of gold data for cross-lingual tagger evaluation. Their use of high-frequency word types inspired parts of our analysis.

6 Conclusions

We analyze DsDs, a recently-proposed low-resource tagger that symbiotically leverages neural representations and symbolic linguistic knowledge by integrating them in a soft manner. We replicated the results of Plank and Agić (2018), showing that the more implicit use of embedding user-generated dictionaries turns out to be more beneficial than approaches that rely more explicitly on symbolic knowledge, such a type constraints or retrofitting. By analyzing the reliance of DsDs on the linguistic knowledge, we found that the composition of the lexicon is more important than its size. Moreover, the tagger benefits from small dictionaries, as long as they do not contain tag set information contradictory to the evaluation data. Our quantitative analysis also sheds light on the internal representations, showing that they get more sensitive to the task. Finally, we found that freezing pre-trained word embeddings complement the learning signal well in this noisy data regime.

Acknowledgements

We kindly acknowledge the support of NVIDIA Corporation for the donation of the GPUs and Amazon for an Amazon Research Award.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *ICLR*.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Alonso Martínez, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Željko Agić, Barbara Plank, and Anders Søgaard. 2017. Cross-lingual tagger evaluation without test data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. *arXiv preprint arXiv:1307.1662*.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872. Association for Computational Linguistics.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7.

C. H. Björnsson. 1983. Readability of newspapers in 11 languages. *Reading Research Quarterly*, 18(4):480–497.

Bernd Bohnet, Ryan McDonald, Goncalo Simoes, Daniel Andor, Emily Pitler, and Joshua Maynez. 2018. Morphosyntactic tagging with a meta-bilstm model over context sensitive token encodings. *arXiv preprint arXiv:1805.08237*.

Huadong Chen, Shujian Huang, David Chiang, and Jiajun Chen. 2017. Improved neural machine translation with a syntax-aware encoder and decoder. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1936–1945. Association for Computational Linguistics.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single .. vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136. Association for Computational Linguistics.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615. Association for Computational Linguistics.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Ákos Kádár, Grzegorz Chrupała, and Afra Alishahi. 2017. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4):761–780.

Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large scale parsing and normalization of wiktionary morphological paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

Junhui Li, Deyi Xiong, Zhaopeng Tu, Muhua Zhu, Min Zhang, and Guodong Zhou. 2017. Modeling source syntax for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–697, Vancouver, Canada. Association for Computational Linguistics.

Shen Li, João Graça, and Ben Taskar. 2012. Wiki-ly supervised part-of-speech tagging. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational*

Natural Language Learning, pages 1389–1398, Jeju Island, Korea. Association for Computational Linguistics.

Christopher D Manning. 2015. Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707.

Joakim Nivre and et al. 2017. Universal dependencies 2.1.

Peyman Passban, Qun Liu, and Andy Way. 2018. Improving character-based decoding using target-side morphological information for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 58–68. Association for Computational Linguistics.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Barbara Plank and Željko Agić. 2018. Distant supervision from disparate sources for low-resource part-of-speech tagging. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 614–620. Association for Computational Linguistics.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418. Association for Computational Linguistics.

Benoît Sagot and Héctor Martínez Alonso. 2017. Improving neural tagging with lexical information. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 25–31, Pisa, Italy. Association for Computational Linguistics.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural mt learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534. Association for Computational Linguistics.

Anja Sperlich, Daniel J. Schad, and Jochen Laubrock. 2015. When preview information starts to matter: Development of the perceptual span in german beginning readers. *Journal of Cognitive Psychology*, 27(5):511–530.

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

Gail Weiss, Yoav Goldberg, and Eran Yahav. 2018. On the practical computational power of finite precision rnns for language recognition. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 740–745. Association for Computational Linguistics.

Minghao Wu, Fei Liu, and Trevor Cohn. 2018. Evaluating the utility of hand-crafted features in sequence labelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2850–2856. Association for Computational Linguistics.