Distant Supervision from Disparate Sources for Low-Resource Part-of-Speech Tagging

Abstract

We introduce DSDS: a cross-lingual neural part-of-speech tagger that learns from disparate sources of distant supervision, and realistically scales to hundreds of low-resource languages. The model exploits annotation projection, instance selection, tag dictionaries, morphological lexicons, and distributed representations, all in a uniform framework. The approach is simple, yet surprisingly effective, resulting in a new state of the art without access to any gold annotated data.

1 Introduction

Low-resource languages lack manually annotated data to learn even the most basic models such as part-of-speech (POS) taggers. To compensate for the absence of direct supervision, work in crosslingual learning and distant supervision has discovered creative use for a number of alternative data sources to learn feasible models:

However, only one or two compatible sources of distant supervision are typically employed. In reality severely under-resourced languages may require a more pragmatic "take what you can get" viewpoint. Our results suggest that combining supervision sources is the way to go about creating viable low-resource taggers.

We propose a method to strike a balance between model simplicity and the capacity to easily integrate heterogeneous learning signals.

system is a uniform neural model for POS tagging that learns from disparate sources of distant supervision (DSDS). We use it to combine: i) multi-source annotation projection, ii) instance selection, iii) noisy tag dictionaries, and iv) distributed word and sub-word representations. We examine how far we can get by exploiting only the wide-coverage resources that are currently readily available for more than 300 languages, which is the breadth of the parallel corpus we employ.

DSDS yields a new state of the art by jointly leveraging disparate sources of distant supervision in an experiment with 25 languages. We demonstrate: i) substantial gains in carefully selecting high-quality instances in annotation projection, ii) the usefulness of lexicon features for neural tagging, and iii) the importance of word embeddings initialization for faster convergence.

2 Method

DSDS is illustrated in Figure 1. The base model is a bidirectional long short-term memory network (bi-LSTM)

Annotation projection. Ever since the seminal work of projecting sequential labels from source to target languages has been one of the most prevalent approaches to crosslingual learning. Its only requirement is that parallel texts are available between the languages, and that the source side is annotated for POS.

We apply the approach by where labels are projected from multiple sources and then decoded through weighted majority voting with word alignment probabilities and source POS tagger confidences. We

exploit their widecoverage Watchtower corpus (WTC), in contrast to the typically used Europarl data. Europarl covers 21 languages of the EU with 400k-2M sentence pairs, while WTC spans 300+ widely diverse languages with only 10-100k pairs, in effect sacrificing depth for breadth, and introducing a more radical domain shift. However, as our results show little projected data turns out to be the most beneficial, reinforcing breadth for depth.

While selected 20k projected sentences at random to train taggers, we propose a novel alternative: selection by coverage. We rank the target sentences by percentage of words covered by word alignment from 21 sources and select the top k covered instances for training. In specific, we employ the mean coverage ranking of target sentences, whereby each target sentence is coupled with the arithmetic mean of the 21 individual word alignment coverages for each of the 21 source-language sentences. We show that this simple approach to instance selection offers substantial improvements: across all languages, we learn better taggers with significantly fewer training instances.

Dictionaries. Dictionaries are a useful source or distant supervision. There are several ways to exploit such information: i) as type constraints during encoding, ii) to guide unsupervised learning, or iii) as addiional signal at training. We focus on the latter and evaluate two ways to integrate lexical knowledge into neural models, while comparing to the former wo: a) by representing lexicon properties as n-hot vector (e.g., if a word has two properties according to lexicon src, it results in a 2-hot vector, if the word is not present in src, a zero vector), with m the number of lexicon properties; b) by embedding the lexical features, i.e., is a lexicon src embedded into an /-dimensional space. We represent as concatenation of all embedded m properies of length [, and a zero vector otherwise. Tuning on the dev set, we found the second embedding approach to perform best, and simple concatenaion outperformed mean vector representations.

We evaluate two dictionary sources, motivated by ease of accessibility to many languages: WIKTIONARY, a word type dictionary that maps tokens to one of the 12 Universal POS tags; and UNIMORPH, a morphological dictionary that provides inflectional paradigms across 350 languages. For Wiktionary, we use the freely available dictionaries from The size of the dictionaries ranges from a few thousands (e.g., Hindi and Bulgarian) to 2M (Finnish UniMorph). Sizes are provided in Table 1, 1st columns. UniMorph covers between 8-38 morphological properties (for English and Finnish, respectively).

Word embeddings. Embeddings are available for many languages. Pre-initialization of offers consistent and considerable performance improvements in our distant supervision setup (Section 4). We use off-the-shelf Polyglot embeddings, which performed consistently better than FastText.

3 Experiments

Baselines. We compare to the following weaklysupervised POS taggers: AGIC: Multi-source annotation projection with Bible parallel data DAS: The label propagation approach by over Europarl data. GARRETTE: The approach by that works with projections, dictionaries, and unlabeled target text. LI: Wiktionary supervision.

Data. Our set of 25 languages is motivated by accessibility to embeddings and dictionaries. In all experiments we work with the 12 Universal POS tags. For development, we use 21 dev sets of the Universal Dependencies 2.1. We employ UD test sets on additional languages as well as the test sets of to facilitate comparisons. Their test sets are a mixture of CoNLL and HamleDT test data, and are more distant from the training and development data.

Model and parameters. We extend an off-theshelf state-of-the-art bi-LSTM tagger with lexicon information. The code is available at: https:// github.com/bplank/bilstm-aux. The parameter l=40 was set on dev data across all languages. Besides using 10 epochs, word dropout rate (p=.25) and 40-dimensional lexicon embeddings, we use the parameters from For all experiments, we average over 3 randomly seeded runs, and provide mean accuracy. For the learning curve, we average over 5 random samples with 3 runs each.

4 Results

Table 1 shows the tagging accuracy for individual languages, while the means over all languages are given in Figure 2. There are several take-aways.

Data selection. The first take-away is that coverage-based instance selection yields substan-

tially better training data. Most prior work on annotation projection resorts to arbitrary selection; informed selection clearly helps in this noisy data setup, as shown in Figure 2 (a). Training on 5k instances results in a sweet spot; more data (10k) starts to decrease performance, at a cost of runtime. Training on all WTC data (around 120k) is worse for most languages. From now on we consider the 5k model trained with Polyglot as our baseline (Table 1, column "5k"), obtaining a mean accuracy of 83.0 over 21 languages.

Embeddings initialization. Polyglot initialization offers a large boost; on average +3.8% absolute improvement in accuracy for our 5k training scheme, as shown in Figure 2 (b). The big gap in low-resource setups further shows their effectiveness, with up to 10% absolute increase in accuracy when training on only 500 instances.

Lexical information. The main take-away is that lexical information helps neural tagging, and embedding it proves the most helpful. Embedding Wiktionary tags reaches 83.7 accuracy on average, versus 83.4 for n-hot encoding, and 83.2 for type constraints. Only on 4 out of 21 languages are type constraints better. This is the case for only one language for n-hot encoding (French). The best approach is to embed both Wiktionary and Unimorph, boosting performance further to 84.0, and resulting in our final model. It helps the most on morphological rich languages such as Uralic.

On the test sets (Table 4, right) DSDS reaches 87.2 over 8 test languages intersecting and. It reaches 86.2 over the more commonly used 8 languages of, compared to their 83.4. This shows that our novel "soft" inclusion of noisy dictionaries is superior to a hard decoding restriction, and including lexicons in neural taggers helps. We did not assume any gold data to further enrich the lexicons, nor fix possible tagset divergences.

5 Discussion

Analysis. The inclusion of lexicons results in higher coverage and is part of the explanation for the improvement of DSDS; see correlation in Figure 3 (a). What is more interesting is that our model benefits from the lexicon beyond its content: OOV accuracy for words not present in the lexicon overall improves, besides the expected improvement on known OOV, see Figure 3 (b).

More languages. All data sources employed in our experiment are very high-coverage. However, for true low-resource languages, we cannot safely assume the availability of all disparate information sources. Table 2 presents results for four additional languages where some supervision sources are missing. We observe that adding lexicon information always helps, even in cases where only 1k entries are available, and embedding it is usually the most beneficial way. For closely-related languages such as Serbian and Croatian, using resources for one aids tagging the other, and modern resources are a better fit. For example, using the Croatian WTC projections to train a model for Serbian is preferable over in-language Serbian Bible data where the OOV rate is much higher.

How much gold data? We assume not having access to any gold annotated data. It is thus interesting to ask how much gold data is needed to reach our performance. This is a tricky question, as training within the same corpus naturally favors the same corpus data. We test both in-corpus (UD)

and out-of-corpus data (our test sets) and notice an important gap: while in-corpus only 50 sentences are sufficient, outside the corpus one would need over 200 sentences. This experiment was done for a subset of 18 languages with both inand out-ofcorpus test data.

Further comparison. In Table 1 we directly report the accuracies from the original contributions by DAS, LI, GARRETTE, and AGIC over the same test data. We additionally attempted to reach the scores of LI by running their tagger over the Table 1 data setup. The results are depicted in Figure 4 as mean accuracies over EM iterations until convergence. We show: i) LI peaks at 10 iterations for their test languages, and at 35 iterations for all the rest. This is in slight contrast to 50 iterations that recommend, although selecting 50 does not dramatically hurt the scores; ii) Our replication falls Ž23c5 points short of their 84.9 accuracy. There is a large 33-point accuracy gap between the scores of, where the dictionaries are large, and the other languages in Figure 4, with smaller dictionaries.

Compared to DAS, our tagger clearly benefits from pre-trained word embeddings, while theirs relies on label propagation through Europarl, a much cleaner corpus that lacks the coverage of the noisier WTC. Similar applies to as they use 1-5M near-perfect parallel sentences. Even if we use much

Table 1: Results on the development sets and comparison of our best model to prior work. LEX: Size (word types) of dictionaries (W: Wiktionary, U: UniMorph). TC: type-constraints using Wiktionary; (embedded Wiktionary tags), DSDS: our model with ;. Results indicated by use W only. Best result in boldface; in case of equal means, the one with lower std is boldfaced. Averages over language families (with two or more languages in the sample, number of languages in parenthesis).

	LEX (10%)		DEV SETS (UD2.1)					TEST SETS				
LANGUAGE	W	U	5k	TCw	n-hot	Ew	DSDS	DAS	LI	GARRETTE	AGIC	D
Bulgarian (bg)	3	47	88.6	88.6	88.9	89.6	**89.7**			83.1	7.7	8
Croatian (hr)	20		84.9	85.4	84.9	84.8	84.8				67.1	7
Czech (cs)	14	72	86.6	86.6	86.9	87.6	**87.2**				73.3	8
Danish (da)	22	24	89.6	89.0	89.8	90.2	**90.0**	83.2	83.3	78.8	79.0	8
Dutch (nl)	52	26	88.3	88.9	89.0	89.7	**89.8**	79.5	86.3			8
English (en)	358	91	86.5	87.4	86.8	87.3	**87.3**		87.1	80.7	73.6	8
Finnish (fi)	104	2,345	81.5	81.2	81.8	82.4	**82.4**					
French (fr)	17	274	91.0	89.6	**91.7**	91.2	91.4			85.5	76.6	8
German (de)	62	71	85.0	86.4	85.5	86.0	**86.7**	82.8	85.8	87.1	80.2	8
Greek (el)	21		80.6	85.7	80.2	80.5	80.5	79.2	64.4	52.3		8
Hebrew (he)	3	12	76.0	76.1	75.5	74.9	**75.3**					
Hindi (hi)	2	26	64.6	64.6	64.8	65.4	**66.2**				67.6	6
Hungarian (hu)	13	13	75.6	75.6	75.3	75.7	**77.9**			77.9	72.0	7
Italian (it)	478	410	91.9	91.7	93.4	93.5	**93.7**	86.8	83.5	76.9		9
! Norwegian (no)	47	18	90.9	90.9	90.9	91.0	**91.5**			84.3	76.7	8
Persian (fa)	4	26	42.8	43.0	43.7	43.5	**59.6**				59.6	4
Polish (pl)	6	132	84.7	84.6	84.2	84.8	**86.0**				75.1	8
Portuguese	41	211	91.4	91.5	92.3	92.9	**92.2**	87.9	84.5	87.3	83.8	8
Romanian (ro)	7	4	83.9	83.9	84.8	85.3	**86.3**					
Spanish (es)	234	324	90.4	88.6	91.0	91.5	**92.0**	84.2	86.4	88.7	81.4	9
Swedish (sv)	89	67	88.9	88.9	89.6	89.9	**89.9**	80.5	86.1	76.1	75.2	8
AVG(21)			83.0	83.2	83.4	83.7	**84.0**					
AVG(8: DAS)								**83.4**	84.8	80.8	75.5	8
AVG(8: LI/AGIC)									**84.9**	80.8	75.2	8
GERMANIC (6)			88.2	88.6	88.6	89.0	**89.2**					
GERMANIC (4: DAS)								81.5	**85.4**			8
ROMANCE (5)			89.7	89.0	90.6	90.9	**91.1**					
ROMANCE (3: DAS)								86.3	85.8	86.5	80.7	**9**
SLAVIC (4)			86.2	86.3	86.2	86.7	**86.9**					
INDO-IRANIAN (2)			53.7	53.8	54.3	54.4	**62.9**					
URALIC (2)			78.5	78.4	78.6	79.0	**80.1**					

smaller and noisier data sources, DSDS is almost on par: 86.2 vs. 87.3 for the 8 languages from Das and , and we even outperform theirs on four languages: Czech, French, Italian, and Spanish.

6 Related Work

Most successful work on low-resource POS tagging is based on projection, tag dictionaries, annotation of seed training data or even more recently some combination of these, e.g., via multi-task learning. Our paper contributes to this literature by leveraging a range of prior directions in a unified, neural test bed.

Most prior work on neural sequence prediction follows the commonly perceived wisdom that hand-crafted features are unnecessary for deep learning methods. They rely on end-to-end training without resorting to additional linguistic resources. Our study shows that this is not the case. Only few prior studies investigate such sources, e.g., for MT and for POS tagging use lexicons, but only as n-hot features and without examining the cross-lingual aspect.

Table 2: Results for languages with missing data sources: WTC projections, Wiktionary (W), or UniMorph (U). Test sets (TEST), projection sources (PROJ), and embeddings languages (EMB) are indicated. Comparison to TnT trained on PROJ. Results indicated by †use W only.

			TEST SETS					
LANGUAGE	TEST	PROJ	Ew	TnT	TCw	n-hot	Ew	DSDS
Basque (eu)	UD	Bible	57.5	61.8	61.8	61.4	62.7	62.7
Basque (eu)	CoNLL	Bible	57.0	60.3	60.3	60.3	61.3	61.3
Estonian (et)	UD	WTC	79.5	80.6				81.5
Serbian (sr)	UD	WTC (hr)	84.0	84.7	85.5	85.1	85.2	85.2
Serbian (sr)	UD	Bible (sr)	77.1	78.9	79.4	80.5	80.7	80.7
Tamil (ta)	UD	WTC	58.2	61.2				

7 Conclusions

We show that our approach of distant supervision from disparate sources (DSDS) is simple yet surprisingly effective for low-resource POS tagging. Only 5k instances of projected data paired with off-the-shelf embeddings and lexical information integrated into a neural tagger are sufficient to reach a new state of the art, and both data selection and embeddings are essential components to boost neural tagging performance.