Model Transfer for Tagging Low-resource Languages using a Bilingual Dictionary

Meng Fang and **Trevor Cohn**
School of Computing and Information Systems
The University of Melbourne
meng.fang@unimelb.edu.au, t.cohn@unimelb.edu.au

Abstract

Cross-lingual model transfer is a compelling and popular method for predicting annotations in a low-resource language, whereby parallel corpora provide a bridge to a high-resource language and its associated annotated corpora. However, parallel data is not readily available for many languages, limiting the applicability of these approaches. We address these drawbacks in our framework which takes advantage of cross-lingual word embeddings trained solely on a high coverage bilingual dictionary. We propose a novel neural network model for joint training from both sources of data based on cross-lingual word embeddings, and show substantial empirical improvements over baseline techniques. We also propose several active learning heuristics, which result in improvements over competitive benchmark methods.

1 Introduction

Part-of-speech (POS) tagging is an important first step in most natural language processing (NLP) applications. Typically this is modelled using sequence labelling methods to predict the conditional probability of taggings given word sequences, using linear graphical models (Lafferty et al., 2001), or neural network models, such as recurrent neural networks (RNN) (Mikolov et al., 2010; Huang et al., 2015). These supervised learning algorithms rely on large labelled corpora; this is particularly true for state-of-the-art neural network models. Due to the expense of annotating sufficient data, such techniques are not well suited to applications in low-resource languages.

Prior work on low-resource NLP has primarily focused on exploiting parallel corpora to project information between a high- and low-resource language (Yarowsky and Ngai, 2001; Täckström et al., 2013; Guo et al., 2015; Agić et al., 2016; Buys and Botha, 2016). For example, POS tags can be projected via word alignments, and the projected POS is then used to train a model in the low-resource language (Das and Petrov, 2011; Zhang et al., 2016; Fang and Cohn, 2016). These methods overall have limited effectiveness due to errors in the alignment and fundamental differences between the languages. They also assume a large parallel corpus, which may not be available for many low-resource languages.

To address these limitations, we propose a new technique for low resource tagging, with more modest resource requirements: 1) a bilingual dictionary; 2) monolingual corpora in the high and low resource languages; and 3) a small annotated corpus of around $1,000$ tokens in the low-resource language. The first two resources are used as a form of distant supervision through learning cross-lingual word embeddings over the monolingual corpora and bilingual dictionary (Ammar et al., 2016). Additionally, our model jointly incorporates the language-dependent information from the small set of gold annotations. Our approach combines these two sources of supervision using multi-task learning, such that the kinds of errors that occur in cross-lingual transfer can be accounted for, and corrected automatically.

We empirically demonstrate the validity of our observation by using distant supervision to improve POS tagging performance with little supervision. Experimental results show the effectiveness of our approach across several low-resource languages, including both simulated and true low-resource settings. Furthermore, given the clear superiority of training with manual annotations, we compare several active learning heuristics. Active learning using uncertainty sampling with a word-

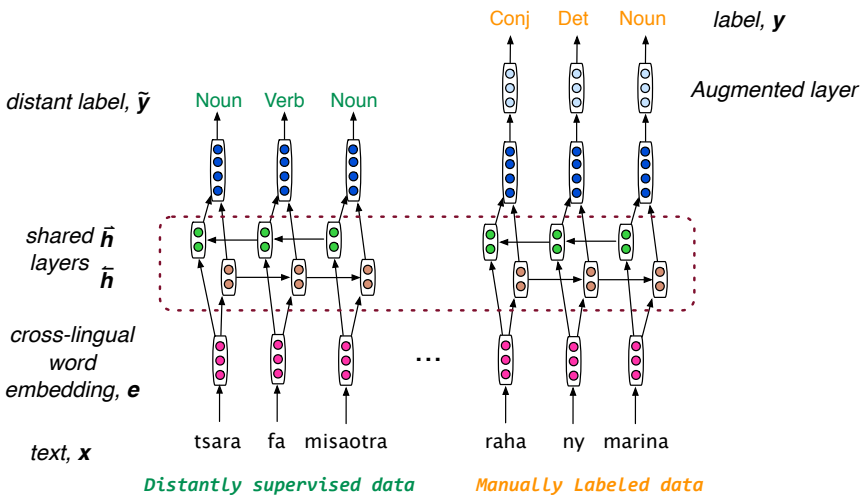

Figure 1: Illustration of the architecture of the joint model, which performs joint inference over both distant supervision (left) and manually labelled data (right).

type bias leads to substantial gains over benchmark methods such as token or sentence level uncertainty sampling.

2 Related work

POS tagging has been studied for many years. Traditionally, probabilistic models are a popular choice, such as Hidden Markov Models (HMM) and Conditional Random Fields (CRF) (Lafferty et al., 2001). Recently, neural network models have been developed for POS tagging and achieved good performance, such as RNN and bidirectional long short-term memory (BiLSTM) and CRF-BiLSTM models (Mikolov et al., 2010; Huang et al., 2015). For example, the CRF-BiLSTM POS tagger obtained the state-of-the-art performance on Penn Treebank WSJ corpus (Huang et al., 2015).

However, in low-resource languages, these models are seldom used because of limited labelled data. Parallel data therefore appears to be the most realistic additional source of information for developing NLP systems in low-resource languages (Yarowsky and Ngai, 2001; Das and Petrov, 2011; Täckström et al., 2013; Fang and Cohn, 2016; Zhang et al., 2016). Yarowsky and Ngai (2001) pioneered the use of parallel data for projecting POS tag information from one language to another language. Das and Petrov (2011) used parallel data and exploited graph-based label propagation to expand the coverage of labelled tokens. Täckström et al. (2013) constructed tag dictionaries by projecting tag information from a high-resource language to a low-resource language via alignments in the parallel text. Fang and Cohn (2016) used parallel data to obtain projected tags as distant labels and proposed a joint BiLSTM model trained on both the distant data and $1,000$ tagged tokens. Zhang et al. (2016) used a few word translations pairs to find a linear transformation between two language embeddings. Then they used unsupervised learning to refine embedding transformations and model parameters. Instead we use minimal supervision to refine 'distant' labels through modelling the tag transformation, based on a small set of annotations.

3 Model

We now describe the modelling framework for POS tagging in a low-resource language, based on very limited linguistic resources. Our approach extends the work of Fang and Cohn (2016), who present a model based on distant supervision in the form of cross-lingual projection and use projected tags generated from parallel corpora as distant annotations. There are three main differences between their work and ours: 1) We do not use parallel corpora, but instead use a bilingual dictionary for knowledge transfer. 2) Our model uses a more expressive multi-layer perceptron when generating the gold standard tags. The multi-layer perceptron can capture both language-specific infor-

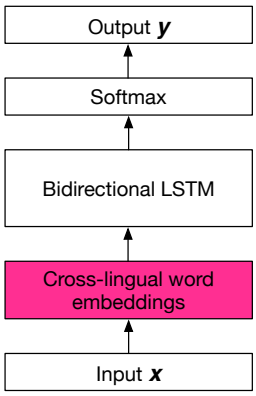

Figure 2: Architecture of the universal POS tagger. Cross-lingual word embeddings are pretrained using monolingual corpora and bilingual dictionaries.

mation and consistent tagging errors arising from this method of supervision. 3) We propose a number of active learning methods to further reduces the annotation requirements. Our method is illustrated in Figure 1, and we now elaborate on the model components.

Distant cross-lingual supervision In order to transfer tag information between the high- and low-resource languages, we start by learning cross-lingual word embeddings, which operate by learning vector valued embeddings such that words and their translations tend to be close together in the vector space. We use the embeddings from Ammar et al. (2016) which trains monolingual `word2vec` distributional representations, which are then projected into a common space, learned from bilingual dictionaries.

We then train a POS tagger on the high-resource language, using the cross-lingual word embeddings as the first, fixed, layer of a bidirectional LSTM tagger. The tagger is a language-universal model based on cross-lingual word embeddings, for processing an arbitrary language, given a monolingual corpus and a bilingual dictionary, as shown in Figure 2. Next we apply this tagger to unannotated text in the low-resource language; this application is made possible through the use of cross-lingual word embeddings. We refer to text tagged this way as *distantly supervised data*, and emphasize that although much better than chance, the outputs are often incorrect and are of limited utility on their own.

As illustrated in Figure 1, the distant components are generated directly as softmax outputs, $y_t \sim \text{Categorial}(\mathbf{o}_t)$, with parameters $\mathbf{o}_t = \text{Softmax}(W\mathbf{h}_t + \mathbf{b})$ as a linear classifier over a sentence encoding, $\mathbf{h}_t$, which is the output of a bidirectional LSTM encoder over the words.

Ground truth supervision The second component of the model is manually labelled text in the low-resource language. To model this data we employ the same model structure as above but augmented with a second perceptron output layer, as illustrated in Figure 1 (right). Formally, $\tilde{y}_t \sim \text{Categorial}(\tilde{\mathbf{o}}_t)$ where $\tilde{\mathbf{o}}_t = \text{MLP}(\mathbf{o}_t)$ is a single hidden layer perceptron with $\tanh$ activation and softmax output transformation. This component allows for a more expressive label mapping than Fang and Cohn (2016)'s linear matrix translation.

Joint multi-task learning To combine the two sources of information, we use a joint objective,

$$\mathcal{J} = -\gamma \sum_{t \in \mathcal{N}} \langle \tilde{y}_t, \log \tilde{o}_t \rangle - \sum_{t \in \mathcal{M}} \langle y_t, \log o_t \rangle \,, \quad (1)$$

where $\mathcal{N}$ and $\mathcal{M}$ index the token positions in the distant and ground truth corpora, respectively, and γ is a constant balancing the two components which we set for uniform weighting, $\gamma = \frac{|\mathcal{M}|}{|\mathcal{N}|}$.

Consider the training effect of the true POS tags: when performing error backpropagation, the cross-entropy error signal must pass through the transformation linking $\tilde{o}$ with o, which can be seen as a language-specific step, after which the generalised error signal can be further backpropagated to the rest of the model.

Active learning Given the scarcity of ground truth labels and the high cost of annotation, a natural question is whether we can optimise which text to be annotated in order achieve the high accuracy for the lowest cost. We now outline a range of active learning approaches based on the following heuristics, which are used to select the instances for annotation from a pool of candidates:

TOKEN Select the token x_t with the highest uncertainty, $H(\mathbf{x}, t) = -\sum_y P(y|\mathbf{x}, t) \log P(y|\mathbf{x}, t)$;

SENT Select the sentence $\mathbf{x}$ with the highest aggregate uncertainty, $H(\mathbf{x}) = \sum_t H(\mathbf{x}, t)$;

FREQTYPE Select the most frequent unannotated word type (Garrette and Baldridge, 2013), in which case all token instances are

annotated with the most frequent label for the type in the training corpus;[1]

SUMTYPE Select a word type, z, for annotation with the highest aggregate uncertainty over token occurrences, $H(z) = \sum_{i \in \mathcal{D}} \sum_{x_{i,t}=z} H(\mathbf{x}_i, t)$, which effectively combines uncertainty sampling with a bias towards high frequency types; and

RANDOM Select word types randomly.

4 Experiments

We evaluate the effectiveness of the proposed model for several different languages, including both simulated low-resource and true low-resource settings. The first evaluation set uses the CoNLL-X datasets of European languages (Buchholz and Marsi, 2006), comprising Danish (da), Dutch (nl), German (de), Greek (el), Italian (it), Portuguese (pt), Spanish (es) and Swedish (sv). We use the standard corpus splits. The first 20 sentences of training set are used for training as the tiny labelled (gold) data and the last 20 sentences are used for development (early stopping). We report accuracy on the held-out test set.

The second evaluation set includes two highly challenging languages, Turkish (tk) and Malagasy (mg), both having high morphological complexity and the latter has truly scant resources. Turkish data was drawn from CoNLL 2003[2] and Malagasy data was collected from Das and Petrov (2011), in both cases using the same training configuration as above.

In all cases English is used as the source 'high resource' language, on which we train a tagger using the Penn Treebank, and we evaluate on each of the remaining languages as an independent target. For cross-lingual word embeddings, we evaluate two techniques from Ammar et al. (2016): CCA-based word embeddings and cluster-based word embeddings. Both types of word embedding techniques are based on bilingual dictionaries. The dictionaries were formed by translating the $20k$ most common words in the English monolingual corpus with Google Translate.[3] The monolingual corpora were constructed from a combination of text from the Leipzig Corpora Collection and Europarl. We trained the language-universal POS tagger based on the cross-lingual word embeddings with the universal POS tagset (Petrov et al., 2011), and then applied to the target language using the embedding lookup table for the corresponding language embeddings. We implement our learning procedure with the DyNet toolkit (Neubig et al., 2017).[4] The BiLSTM layer uses 128 hidden units, and 32 hidden units for the transformation step. We used SGD with momentum to train models, with early stopping based on development performance.

For benchmarks, we compare the proposed model against various state-of-the-art supervised learning methods, namely: a BILSTM tagger, BILSTM-CRF tagger (Huang et al., 2015), and a state-of-the-art semi-supervised POS tagging algorithm, MINITAGGER (Stratos and Collins, 2015), which is also focusing on minimising the amount of labelled data. Note these methods do not use cross-lingual supervision. For a more direct comparison, we include BILSTM-DEBIAS (Fang and Cohn, 2016), applied using our proposed cross-lingual supervision based on dictionaries, instead of parallel corpora; accordingly the key difference is their linear transformation for the distant data, versus our non-linear transformation to the gold data.

Results Table 1 reports the tagging accuracy, showing that our models consistently outperform the baseline techniques. The poor performance of the supervised methods suggests they are overfitting the small training set, however this is much less of a problem for our approach (labelled Joint). Note that distant supervision alone gives reasonable performance (labelled DISTANT) however the joint modelling of the ground truth and distant data yields significant improvements in almost all cases. BILSTM-DEBIAS (Fang and Cohn, 2016) performs worse than our proposed method, indicating that a linear transformation is insufficient for modelling distant supervision. The accuracies are higher overall for the European cf. Turkic languages, presumably because these languages are

[1] We could support more than one class label, by marginalising over the set of valid labels for all tokens in the training objective.

[2] http://www.cnts.ua.ac.be/conll2003/ner/

[3] Although the use of a translation system conveys a dependence on parallel text, high quality word embeddings can be learned directly from bilingual dictionaries such as Panlex (Kamholz et al., 2014).

[4] Code available at https://github.com/mengf1/trpos

	da	nl	de	el	it	pt	es	sv	tk	mg
Random	23.2	30.5	27.1	23.2	25.9	24.3	26.9	21.6	36.9	34.5
BILSTM	61.8	62.1	60.5	70.1	73.6	67.6	63.6	57.2	44.0	63.4
BILSTM-CRF	46.3	47.7	53.2	35.1	41.2	44.1	25.5	54.9	43.1	41.4
MINITAGGER	77.0	72.5	75.9	75.7	67.3	75.1	73.5	77.7	49.8	67.2
DISTANT +CCA	73.5	64.5	57.7	53.1	59.5	67.8	63.5	66.0	57.2	49.7
DISTANT +Cluster	70.4	61.7	65.9	65.5	64.8	66.9	68.4	64.1	51.7	50.2
BILSTM-DEBIAS +CCA	73.2	72.8	72.5	71.2	70.7	72.1	71.1	73.1	49.2	65.9
BILSTM-DEBIAS +Cluster	72.5	70.1	71.2	68.7	69.1	72.5	70.6	73.3	48.7	64.5
JOINT +CCA	81.1	**82.3**	76.1	77.5	75.9	**82.1**	79.7	**78.1**	**72.6**	75.3
JOINT +Cluster	**81.9**	81.5	**78.9**	**80.1**	**81.9**	76.7	**81.2**	78.0	70.4	**75.7**

Table 1: POS tagging accuracy on over the ten target languages, showing first approaches using only the gold data; next methods using only distant cross-lingual supervision, and lastly joint multi-task learning. English is used as the source language and columns correspond to a specific target language.

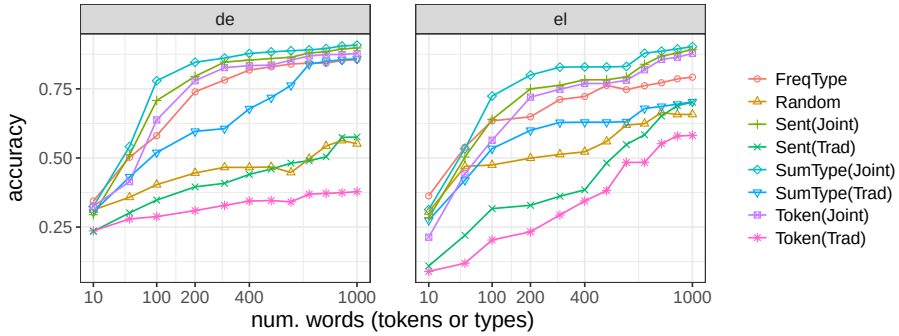

Figure 3: Active learning evaluation on German and Greek, using CCA trained cross-lingual word embeddings. Trad means traditional active learning; Joint means joint multi-task learning.

closer to English, have higher quality dictionaries and in most cases are morphologically simpler. Finally, note the difference between CCA and Cluster methods for learning word embeddings which arise from the differing quality of distant supervision between the languages.

Figure 3 compares various active learning heuristics (see §3) based on different taggers, either a supervised BILSTM (labelled Trad) or our multi-task model which also includes cross-lingual supervision (JOINT).

Traditional uncertainty-based sampling strategies (TOKEN(Trad) and SENT(Trad)) do not work well because models based on limited supervision do not provide accurate uncertainty information,[5] and moreover, annotating at the type rather than token level provides a significantly stronger supervision signal. The difference is apparent from the decent performance of Random sampling over word types. Overall, SUMTYPE(Joint) outperforms the other heuristics consistently, underlining the importance of cross-lingual distant supervision, as well as combining the benefits of uncertainty sampling, type selection and a frequency bias. Comparing the amount of annotation required between the best traditional active learning method SUMTYPE(Trad) and our best method SUMTYPE(Joint), we achieve the same performance with an order of magnitude less annotated data (100 vs. 1,000 labelled words).

[5]Sentence level annotation is likely to be much faster than token or type level annotation, however even if it were an order of magnitude faster it is still not a competitive active learning strategy.

5 Conclusion

In this paper, we proposed a means of tagging a low-resource language without the need for bilingual parallel corpora. We introduced a new cross-lingual distant supervision method based on a bilingual dictionary. Furthermore, deep neural network models can be effective with limited supervision by incorporating distant supervision, in the form of model transfer with cross-lingual word embeddings. We show that traditional uncertainty sampling strategies do not work well on low-resource settings, and introduce new methods based around labelling word types. Overall our approach leads to consistent and substantial improvements over benchmark methods.

Acknowledgments

This work was sponsored by the Defense Advanced Research Projects Agency Information Innovation Office (I2O) under the Low Resource Languages for Emergent Incidents (LORELEI) program issued by DARPA/I2O under Contract No. HR0011-15-C-0114. The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. Trevor Cohn was supported by the Australian Research Council Future Fellowship (project number FT130101105).

References

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.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *Transactions of the Association for Computational Linguistics* 4:431–444.

Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pages 149–164.

Jan Buys and Jan A. Botha. 2016. Cross-lingual morphological tagging for low-resource languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, Berlin, Germany, pages 1954–1964.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*. pages 600–609.

Meng Fang and Trevor Cohn. 2016. Learning when to trust distant supervision: An application to low-resource pos tagging using cross-lingual projection. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*. Berlin, Germany.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Citeseer, pages 138–147.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, pages 1234–1244.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991* .

David Kamholz, Jonathan Pool, and Susan M Colowick. 2014. Panlex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*. pages 3145–3150.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 8th International Conference on Machine Learning (ICML)*. volume 1, pages 282–289.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*. volume 2, page 3.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980* .

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086* .

Karl Stratos and Michael Collins. 2015. Simple semi-supervised pos tagging. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. pages 79–87.

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.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP brackets via robust projection across aligned corpora. In *Proceedings of the 2001 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag–multilingual pos tagging via coarse mapping between embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. pages 1307–1317.