Sequence Tagging - Using Transfer Learning

Anshuman Mourya
Computer Science and Automation
Indian Institute of Science , Bangalore
anshumanm@iisc.ac.in

Praveen Gupta
Computer Science and Automation
Indian Institute of Science , Bangalore
praveengupta@iisc.ac.in

Abstract

Neural networks obtain state-of-the-art performance on several different sequence tagging task. However, it is unclear if such systems can be used for tasks without large amounts of training data. We explore the problem of transfer learning for neural sequence taggers, where a source task with plentiful annotations (e.g., NER tagging on CONLL2003) is used to improve performance on a target task with fewer available annotations (e.g., NER tagging for Twitter data). We examine the effects of transfer learning for deep hierarchical recurrent networks across domains and languages and compare them to the state-of-the-art.

1 Introduction

Sequence tagging is an important problem in natural language processing, which has wide applications including part-of-speech (POS) tagging, text chunking, and named entity recognition (NER). Given a sequence of words, sequence tagging aims to predict a linguistic tag for each word/group of words such as the NER tag. An important challenge for sequence tagging is how to transfer knowledge from one task to another, which is often referred to as transfer learning (Pan and Yang, 2010). Transfer learning can be used in settings involving low resource languages and low resource domains such as biomedical corpora and Twitter corpora. In these cases, transfer learning can improve performance by taking advantage of more plentiful labels from related tasks.

A number of approaches based on deep neural network have been developed to address the problem of sequence tagging. These neural networks require minimal assumptions about the task at hand and thus demonstrate significant generality one single model can be applied to multiple applications in multiple languages without changing the architecture. Thus we address the question - whether the representation learned from one task can be useful for another task or is there a way we can exploit the generality of neural networks to improve task performance by sharing model parameters and feature representations with another task.

To address these questions, we study the transfer learning setting with the aim to improve performance on target task by joint training with source task. We have used deep hierarchical recurrent networks that shares the hidden feature representations and some of the model parameters between the source and target task. We explore cross-domain and cross-lingual settings in this project. Experiments performed shows that performance on target task can be improved even when the target tasks has few labels and is related to source task. These results are found comparable to state-of-the-art.

2 Related Work

Transfer learning for Natural language Processing tasks are classified into two paradigms : resource-based transfer learning and model-based transfer. Resource-based transfer utilizes additional linguistic annotations as weak supervision for transfer learning, such as cross-lingual dictionaries (Zirikly and Hagiwara", 2015), corpora (Wang and Manning", 2014), and word alignments (David Yarowsky and Wicentowski, 2001). Resource-based methods demonstrate considerable success in cross-lingual transfer, but are quite sensitive to the

scale and quality of the additional resources. It is mostly limited to cross-lingual transfer in previous works, and there is not extensive research on extending resource-based methods to cross-domain and cross-application settings.

Model-based transfer, does not require additional resources. It exploits the similarity and relatedness between the source task and the target task by adaptively modifying the model architectures, training algorithms, or feature representation. (Ando and Zhang, 2005) proposed a transfer learning framework that shares structural parameters across multiple tasks, and improve the performance on various tasks including NER; (Peng and Dredze, 2016) studied transfer learning between named entity recognition and word segmentation in Chinese based on recurrent neural networks. Cross-domain transfer, or domain adaptation, is also a well-studied branch of model-based transfer in NLP. Techniques in cross-domain transfer include the design of robust feature representations (Schnabel and Schtze, 2014), co-training (Minmin Chen and Blitzer, 2011), and hierarchical Bayesian prior (Finkel and Manning, 2009).

We have used model-based transfer for all three settings. (Ronan Collobert and Kuksa, 2011)develop end-to-end neural networks for sequence tagging without hand-engineered features. We have used this approach to some extent

3 Method

In this section we first mention the general approach used and then describe the approach for used in different settings.

3.1 Base Model

Most of the models can be described with the hierarchical framework illustrated in Figure 1. A character-level layer takes a sequence of characters (represented as embeddings) as input, and outputs a representation that encodes the morphological information at the character level. A word-level layer subsequently combines the character-level feature representation and a word embedding, and further incorporates the contextual information to output a new feature representation. After two levels of feature extraction (encoding), the feature representation output by the word-level

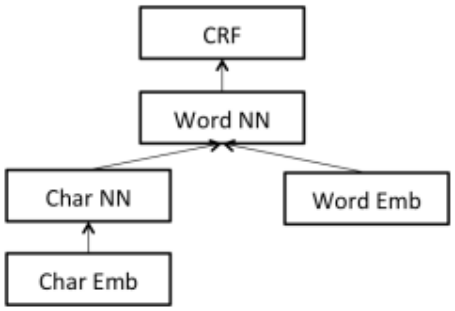

Figure 1: Baseline

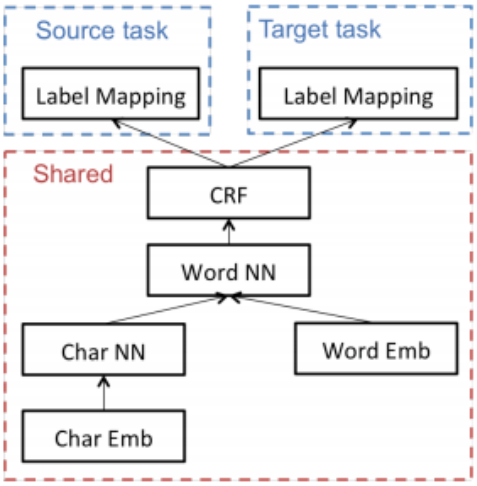

Figure 2: Model-A:Cross-domain transfer when label mapping is possible

layer is fed to a conditional random field (CRF) layer that outputs the label sequence. Both of the word-level layer and the character-level layer are implemented recurrent neural networks (RNNs) (Ronan Collobert and Kuksa, 2011); (Chiu and Nichols, 2015); (Guillaume Lample and Dyer, 2016).

3.2 Transfer Learning Architectures

We have used three architectures shown in figures. All three are extensions of the base model mentioned before.

3.2.1 Cross-domain Transfer

Since different domains are sub-languages that have domain-specific regularities, sequence taggers trained on one domain might not have optimal performance on another domain. The goal of cross-domain transfer is to learn a sequence tagger that transfers knowledge from a source domain to a target domain. We have assumed that few labels are available in the target domain. There are

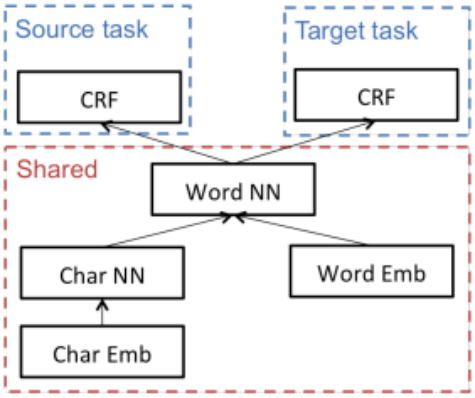

Figure 3: Model-B:Cross-domain transfer when label mapping is not possible

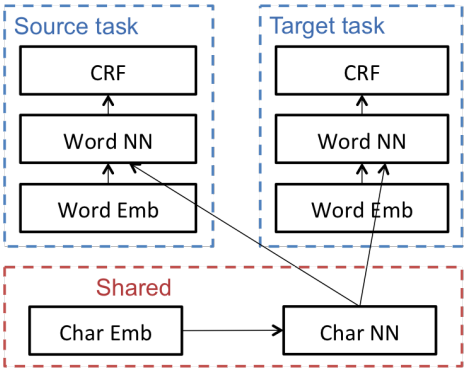

Figure 4: Model-C:Cross-lingual transfer

two cases of cross-domain transfer. The two domains can have label sets that can be mapped to each other, or disparate label sets. For example, POS tags in the Genia biomedical corpus can be mapped to Penn Treebank tags , while some POS tags in Twitter (e.g., URL) cannot be mapped to Penn Treebank tags. If the two domains have mappable label sets, we share all the model parameters and feature representation in the neural networks, including the word and character embedding, the word-level layer, the character-level layer, and the CRF layer. We perform a label mapping step on top of the CRF layer. This becomes the model-A as shown in Figure 2. If the two domains have disparate label sets, we untie the parameter sharing in the CRF layer i.e., each task learns a separate CRF layer. This parameter sharing scheme reduces to model-B as shown in Figure 3.

3.2.2 Cross-lingual Transfer

Though cross-lingual transfer is usually accomplished with additional multi-lingual resources, these methods are sensitive to the size and quality of the additional resources (David Yarowsky and Wicentowski, 2001); (Wang and Manning", 2014). We explored a method that exploits the cross-lingual regularities purely on the model level.

Our approach focuses on transfer learning between languages with similar alphabets, such as English and Spanish, since it is very difficult for transfer learning between languages with disparate alphabets (e.g., English and Chinese or English and Chinese) to work without additional resources (Zirikly and Hagiwara", 2015). Model-level transfer learning is achieved through exploiting the morphologies shared by the two languages. For example, Canada in English and Canad in Spanish refer to the same named entity, and the morphological similarities can be leveraged for NER and also POS tagging with nouns. Thus we share the character embeddings and the character-level layer between different languages for transfer learning, which is illustrated as model-C in Figure 4.

3.3 Training

In previous section, we introduced three neural architectures with different parameter sharing schemes, designed for different transfer learning settings. Now we describe how we train the neural networks jointly for two tasks. Suppose we are transferring from a source task s to a target task t, with the training instances being X_s and X_t . Let W_s and W_t denote the set of model parameters for the source and target tasks respectively. The model parameters are divided into two sets, task specific parameters and shared parameters, i.e.,

$$W_s = W_{s,spec} \cup W_{shared}$$
$$W_t = W_{t,spec} \cup W_{shared}$$

where shared parameters W_shared are jointly optimized by the two tasks, while task specific parameters $W_{s,spec}$ and $W_{t,spec}$ are trained for each task separately. The training procedure is as follows. At each iteration, we sample a task (i.e., either s or t) from s, t based on a binomial distribution (the binomial probability is set as a hyper-parameter). Given the sampled task, we sample a batch of training instances from the given task, and then perform a gradient update according to the loss function of the given task. We update both the shared parameters and the task specific parameters. We repeat the above iterations until stopping. Since the source and target tasks might have

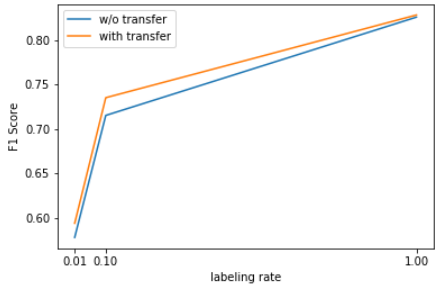

Figure 5: English NER to Spanish NER transfer

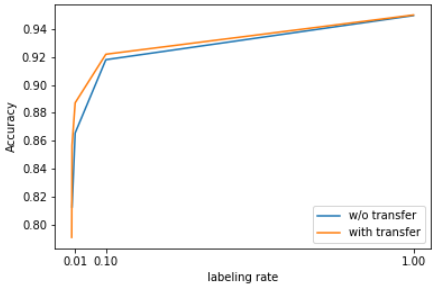

Figure 7: PTB to Genia POS

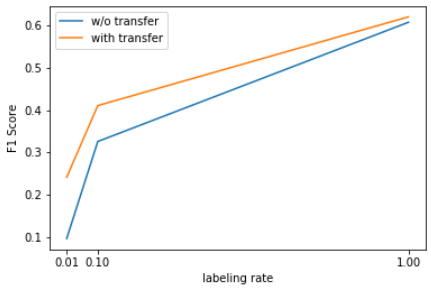

Figure 6: NER to Twitter NER

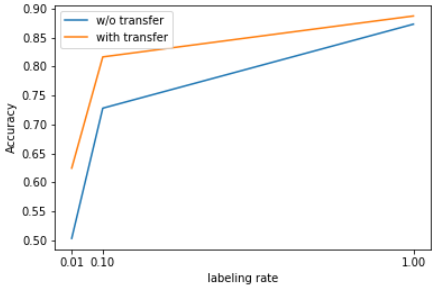

Figure 8: PTB to Twitter POS

different convergence rates, we do early stopping on the target task performance.

3.4 Model Implementation

Both the character-level and word-level neural networks are implemented as RNNs. More specifically, we employ gated recurrent units (GRUs) . Let $(x_1, x_2, , x_T)$ be a sequence of inputs that can be embeddings or hidden states of other layers. Let h_t be the GRU hidden state at time step t. Formally, a GRU unit at time step t can be expressed as

$$r_t = \sigma(W_{rx}x_t + W_{rh}h_{t-1})$$

$$z_t = \sigma(W_{zx}x_t + W_{zh}h_{t-1})$$

$$\bar{h_t} = tanh(W_{hx}x_t + W_{hh}(r_t \odot h_{t-1}))$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \bar{h_t}$$

where Ws are model parameters of each unit, $\bar{h_t}$ is a candidate hidden state that is used to compute h_t , σ is an element-wise sigmoid logistic function defined as $\sigma(x) = 1(1 + \exp x)$, and denotes element-wise multiplication of two vectors. The update gate z_t controls how much the unit updates its hidden state, and the reset gate r_t determines how much information from the previous hidden state needs to be reset. The input to the character-level GRUs is character embeddings, while the input to the word-level GRUs is the concatenation of character-level GRU hidden states and word embeddings. Both GRUs are bi-directional and have two layers. Given an input sequence of words, the word-level GRUs and the character-level GRUs together learn a feature representation h_t for the t^{th} word in the sequence, which forms a sequence $h = (h_1, h_2, , h_T)$. Let $y = (y_1, y_2, , y_T)$ denote the tag sequence. Given the feature representation h and the tag sequence y for each training instance, the CRF layer defines the objective function to maximize based on a max-margin principle (Gimpel and Smith, 2010) as:

$$f(h, y) - \log \sum_{y' \epsilon \gamma(h)} \exp(f(h, y') + cost(y, y'))$$

where f is a function that assigns a score for each pair of h and y, and $\gamma(h)$ denotes the space of tag sequences for h. The cost function $cost(y, y')$ is added based on the max-margin principle that high-cost tags y' should be penalized more heavily

4 Results

Following benchmark datasets are used in our experiments: Penn Treebank (PTB) POS tag-

Table 1: Dataset Statistics

Benchmark	Task	Language	# Training Tokens	# Dev Tokens	# Test Tokens
PTB 2003	POS Tagging	English	91,204	13,432	12,822
CoNLL 2003	NER	English	201,576	51,578	46,666
CoNLL 2002	NER	Spanish	207,484	51,645	52,098
Genia	POS Tagging	English	400,658	50,525	49,761
Twitter	POS Tagging	English	12,196	1,362	1,627
Twitter	NER	English	36,936	4,612	4,921

Table 2: Experiments for various settings

Source	Target	Labeling Rate	Without Transfer	With Transfer
ENG NER	Twitter NER	0.01	0.0963	0.2415
ENG NER	Twitter NER	0.1	0.3261	0.4111
ENG NER	Twitter NER	1.0	0.6084	0.6213
ENG NER	Span NER	0.01	0.5783	0.5942
ENG NER	Span NER	0.1	0.7153	0.7352
ENG NER	Span NER	1.0	0.8257	0.8281
PTB	Genia	0.001	0.8126	0.8565
PTB	Twitter POS	0.1	0.7280	0.8165

ging, CoNLL 2003 English NER, CoNLL 2002 Spanish NER, the Genia biomedical corpus and a Twitter corpus. The Statistics of dataset is given in Table 1. For CONLL 2003 English NER dataset, we appended one-hot gazetteer features to the input of CRF layer. PTB POS , CONLL 2003 English NER and CONLL 2002 Spanish NER have standard dataset splits, but for Gemia and Twitter, we randomly split dataset as 80% training, 10%test and 10%dev set.

We have fixed the following hyperparameters : character embedding dimension at 25, word embedding dimenesion at 50 for English and 64 for Spanish, dimension of hidden states of the character-level GRUs at 80, dimension of hidden states of the word-level GRUs at 300, and the initial learning rate at 0.01. To simulate a low-resource setting we use labeling rate. Given a lableing rate r, we randomly sample ratio r of the sentences from the training set and discard the rest.

We have used labeling rate of 1,0.1,0.01,0.001 and 0.0 for different settings. We use labeling rates only for target tasks, for source tasks it is fixed to 1.0. Table 2 shows the numbers obtained in both with and without transfer on target tasks. Figure 4 ,5,6 and 7 shows how the measuring criteria varies with labeling rate. We have used F1-score for NER tasks and accuracy for POS-tagging. For lower labeling rates we observe substantial improvement in transfer learning. For cross-domain transfer, we obtained substantial improvement on the Genia and Twitter corpora by transferring the knowledge from PTB POS tagging and CoNLL 2003 NER. For example, as shown in Table 2, we can obtain an tagging accuracy of 85% with only 0.001 labels when transferring from PTB to Genia. Our transfer learning approach can improve the performance on Twitter POS tagging and NER for all labeling rates, and the improvements with 0.1 labels are more than 8% for both datasets. Cross-lingual transfer can improve the performance when few labels are available.

Improvements in different models are in following order $model - A > model - B > model - C$. This phenonemenon is because model-A shares the most model parameters while model-C shares the least. Trasnfer learning in Cross-lingual can use only model-C because source task and target tasks are less similar.

5 Discussion

We have used publicly available pretrained word embeddings as initialization. On the English datasets we have used 50-dimensional SENNA embeddings. For Spanish, we use the 64-dimensional Polyglot embeddings. We set the hidden state dimensions to be 300 for the word-level GRU. We use the development set to tune the other hyperparameters of our model. We don't have standard splits for Twitter and Genia dataset. So we compared only the results on CONLL Spanish dataset. The proposed approach could not reach state-of-the art results but is found comparable to it.

6 Future Work

In this project, we have performed experiments only in two settings: Cross-domain transfer and Cross-lingual transfer. These works can be extended to Cross-application transfer setting such as to transfer from POS source task to NER target task or vice versa. Model-B can be modified for this setting as it contains different CRFs for unmappable datasets. Mixture of different settings can be experimented as well. For ex. transferring from Spanish NER to Genia POS. Model-C can be modified for this experimentation.

For Cross-lingual setting, we have used languages that share some alphabets. We would like to explore the possibility of transfer learning between languages such as English and Chinese or English and Hindi. This will involve combining model-based transfer and resource-based transfer.

References

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *JMLR*.

Jason PC Chiu and Eric Nichols. 2015. Named entity recognition with bidirectional lstm-cnns. *arXiv preprint arXiv:1511.08308*.

Grace Ngai David Yarowsky and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. *HLT*.

Jenny Rose Finkel and Christopher D Manning. 2009. Hierarchical bayesian domain adaptation. *HLT, pp. 602 610*.

Kevin Gimpel and Noah A Smith. 2010. Softmax-margin crfs: Training log-linear models with cost functions. *NAACL, pp. 733736*.

Sandeep Subramanian Kazuya Kawakami Guillaume Lample, Miguel Ballesteros and Chris Dyer. 2016. Neural architectures for named entity recognition. *NAACL*.

Kilian Q Weinberger Minmin Chen and John Blitzer. 2011. Co-training for domain adaptation. *NIPS, pp. 2456 2464*.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. knowledge and data engineering. *IEEE Transactions on, 22(10):13451359*.

Nanyun Peng and Mark Dredze. 2016. Improving named entity recognition for chinese social media with word segmentation representation learning. *ACL*.

Leon Bottou Michael Karlen Koray Kavukcuoglu Ronan Collobert, Jason Weston and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *JMLR*.

Tobias Schnabel and Hinrich Schtze. 2014. Flors: Fast and simple domain adaptation for part-of-speech tagging. *TACL, 2:1526*.

"Mengqiu Wang and Christopher D Manning". 2014. Cross-lingual pseudo-projected expectation regularization for weakly supervised learning. *TACL*.

"Ayah Zirikly and Masato Hagiwara". 2015. Cross-lingual transfer of named entity recognizers without parallel corpora. *ACL*.