Cross-lingual Transfer Learning

Peifeng Wang
University of Southern California
peifengw@usc.edu

Abstract

One of the benefits brought by transfer learning is to leverage knowledge which is learned from one domain with large resources to better solve problems in the other domains with low resources. This is also happens in the cross-lingual scenario given that the available training data is disproportionately biased towards English while there are plenty of other language being neglected. In this note, we introduce several strategies for cross-lingual transfer learning driven by different tasks. The common assumption which they rely on is that there exists a same or similar property shared by different languages which the strategies take advantage of to conduct transfer. Depending on the nature of tasks, these strategies exploit the unique properties owned by the different tasks as the supervision signals for the low-resource languages. Based on the specific supervision signals, we categorize several cross-lingual transfer learning methods and introduce them as below.

1 Cross-lingual Language Model

Deep pre-trained language models have brought significant improvements in the NLP field. These models is trained on large unlabeled data and later fine-tuned to specific NLP tasks. However, most of the pre-trained language models focus on English corpus and introduce English-centric bias. There are attempts which try to generalize the monolingual language model to the universal language setting. Here, we introduce two of them from the works of Lample and Conneau [2019] and Pires et al. [2019]. The former requires parallel data partially while the latter does not. Still, both of them use word piece vocabulary shared across languages so that part of the knowledge could be transferred via word piece overlap.

In particular, Lample and Conneau [2019] implement word piece by byte-pair encoding (BPE) [Sennrich et al., 2015] which favors low-resource languages. During learning BPE splits, sentences are sampled according to a multinomial distribution with probabilities

$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^{N} p_j^\alpha} \tag{1}$$

where p_i is the frequency of language i. Sampling with this distribution increases the number of tokens associated to low-resource languages and alleviates the bias towards high-resource languages.

1.1 Translation Language Modeling

When parallel data is at hand, Lample and Conneau [2019] introduces a new translation language modeling objective. Firstly, they concatenate parallel sentences as illustrated in Figure 1. Words from different language are differentiated via language embeddings. Position embeddings of the target sentence are also reset to facilitate the alignment. Then then randomly mask words in both the source and target sentences. Thus, to predict a word masked in the English sentence, the language model could depend on either the source or target context. In particular, when the source context is not sufficient, the model could then leverage the target one. In this way, the model is encouraged to align the source and target representations.

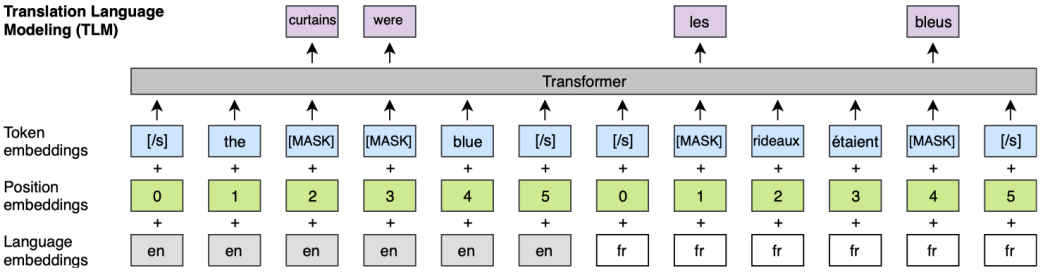

Figure 1: Translation Language Modeling

Fine-tuning \ Eval	EN	DE	NL	ES
EN	**90.70**	69.74	77.36	73.59
DE	73.83	**82.00**	76.25	70.03
NL	65.46	65.68	**89.86**	72.10
ES	65.38	59.40	64.39	**87.18**

Table 1: NER F1 results on the CoNLL data.

Fine-tuning \ Eval	EN	DE	ES	IT
EN	**96.82**	89.40	85.91	91.60
DE	83.99	**93.99**	86.32	88.39
ES	81.64	88.87	**96.71**	93.71
IT	86.79	87.82	91.28	**98.11**

Table 2: POS accuracy on a subset of UD languages.

1.2 Multilingual BERT

The paper Sennrich et al. [2015] is more like a probing work which investigates how well Multilingual BERT (M-BERT) performs on zero-shot cross-lingual model transfer. Instead of leveraging parallel data, M-BERT is trained on the Wikipedia pages of 104 languages with a shared word piece vocabulary. It does not use any marker denoting the input language, and does not have any explicit mechanism to encourage translation-equivalent pairs to have similar representations. When fine-tuning, M-BERT is trained on the task-specic supervised training data from one language and later used for evaluation of the same task in another language.

The results as in Table 1, 2 show that M-BERT is able to perform cross-lingual generalization surprisingly well. To further investigate how the model is able to perform this transfer, the paper provides some probing experiments. Here are some major observations.

Word Piece Overlap. One hypothesis is that zero-shot performance on NER is highly dependent on word piece overlap. To measure the effect of vocabulary memorization, the paper computes the overlap between the sets of word pieces used in entities in the training and evaluation datasets: $overlap = |E_{train} \cap E_{eval}|/|E_{train} \cup E_{eval}|$. Figure 2 plots NER F1 score versus entity overlap for zero-shot transfer between every language pair in a dataset of 16 languages, for both M-BERT and EN-BERT. M-BERT's performance is flat for a wide range of overlaps, and even for language pairs with almost no lexical overlap, showing that M-BERT's pre-training on multiple languages has enabled a representational capacity deeper than simple vocabulary memorization.

Language Similarity. The paper also compares languages on a subset of the WALS features relevant to grammatical ordering. Figure 3 plots POS zero-shot accuracy against the number of common WALS features. As expected, performance improves with similarity, showing that it is easier for M-BERT to map linguistic structures when they are more similar, although it still does a decent job for low similarity languages when compared to EN-BERT.

Multilingual characterization of the feature space. The paper further study the structure of M-BERT's feature space. If it is multilingual, then the transformation mapping between the same sentence in 2 languages should not depend on the sentence itself, just on the language pair. In specific, the paper firstly feeds some sampled sentence to M-BERT. Then they extract the hidden feature activations at each layer for each of the sentences, average the representations for the input tokens to get a vector for each sentence, at each layer l, $v^{(l)}_{LANG}$. For each pair of sentences, e.g. $(v^{(l)}_{EN_i}, v^{(l)}_{DE_i})$, they compute the vector translation and average it over all pairs: $\vec{v}^{(l)}_{EN \to DE} = \frac{1}{M} \sum_i (v^{(l)}_{EN_i} - v^{(l)}_{DE_i})$. Finally, they translate each sentence, v_{EN_i}, by $\vec{v}^{(l)}_{EN \to DE}$, find the closest German sentence vector, and measure the fraction of times the nearest neighbour is the correct pair. The results are ploted in

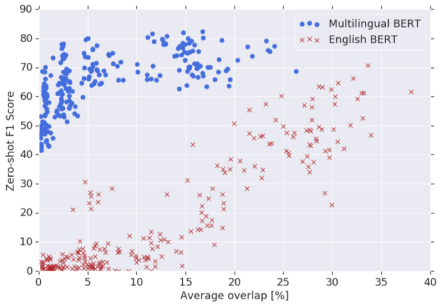

Figure 2: Zero-shot NER F1 score versus entity word piece overlap among 16 languages.

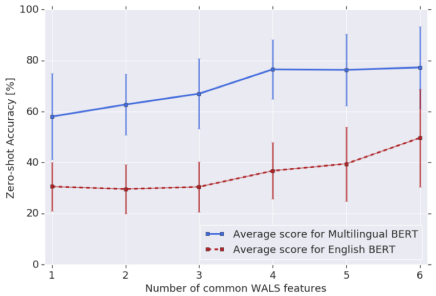

Figure 3: Zero-shot POS accuracy versus number of common WALS features.

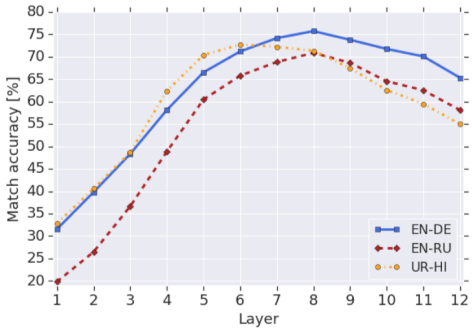

Figure 4: Accuracy of nearest neighbor translation for EN-DE, EN-RU, and HI-UR.

Figure 4. It implies that the hidden representations share a common subspace that represents useful linguistic information, in a language-agnostic way.

2 Adversarial Transfer

Adversarial training has been extensively studied and applied for cross-lingual or more generally, cross domain transfer. It allows the model automatically to induce bilingual and multilingual word representations without using any parallel corpora. It also allows the model to extract language and domain-agnostic features for cross-lingual and cross-domain adaptation. One recent work Huang et al. [2019] takes advantage of these two benefits of adversarial transfer to enhance low-resource name tagging.

2.1 Word-level Transfer

Mikolov et al. [2013] first noticed that the geometric relations that hold between words are similar across languages as illustrated in Figure 5. That is, the continuous word embedding spaces exhibit similar structures across languages, even when considering distant language pairs like English and Vietnamese. They exploit this similarity by learning a linear mapping from a source to a target embedding space as

$$W^* = \arg\min_{W \in \mathcal{R}^d} \|WX - Y\|_F, \tag{2}$$

where X and Y are two embedding matrix for the aligned words in parallel vocabulary.

Conneau et al. [2017] took a step further to use adversarial training to avoid the need for parallel data, where a discriminator is trained to discriminate embeddings sampled from WX and Y while the generator aims at preventing the discriminator from doing so by making WX and Y as similar as possible. The same method is conducted as the first step in [Huang et al., 2019] to ensure that the words from the source and target languages could share a common semantic space.

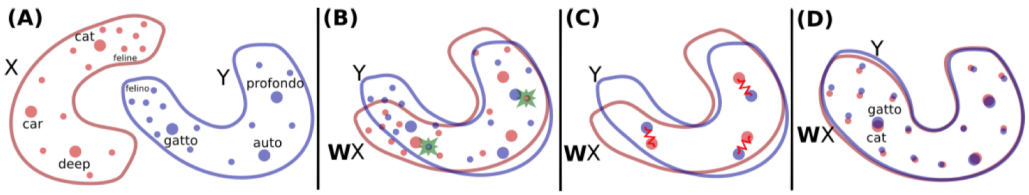

Figure 5: Mapping one distribution of embeddings to another one.

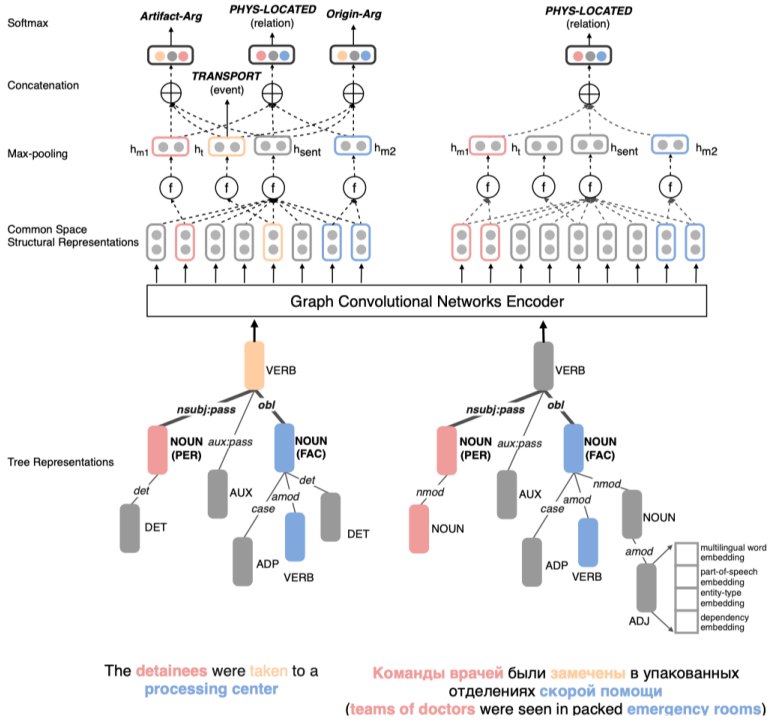

Figure 6: Multilingual common semantic space and cross-lingual structure transfer.

2.2 Sentence-level Transfer

Unifying word-level features is not enough since name tagging also relies on sequential contextual features for entity type classification. Therefore, a similar adversarial training is also conducted on a sentence level. They first feed the sequence of vector representations into a weight sharing BiLSTM encoder E to obtain sequential features and then use a discriminator to predict the language source of each sentence. This encourages the encoder to extract language-agnostic sequential features.

3 Structure Transfer

Unlike sequence representations, tree representations such as constituency trees and dependency trees are typically constructed following a combination of syntactic principles and annotation guidelines designed by linguists. The resulting structures, such as the verb – subject relation and the verb – object relation, are found across languages. This structure consistency is particular suitable for transfer learning in relation/event extraction. This is because relational facts are often expressed in a similar pattern across languages. As shown in Figure 6, even for distinct pairs of entity mentions (colored pink and blue, in both English and Russian), the structures share similar language-universal symbolic features, such as a common labeled dependency path.

Train \ Test	English	Chinese	Arabic
English	68.2	**42.5**	**58.7**
Chinese	**62.6**	69.4	54.0
Arabic	58.6	35.2	67.4

Table 3: Relation Extraction: overall performance (Fscore %).

Train \ Test	English	Chinese	Arabic
English	63.9	**59.0**	**61.8**
Chinese	**51.6**	59.3	60.6
Arabic	43.1	50.1	64.0
English + Chinese	–	–	**63.1**
English + Arabic	–	**60.1**	–
Chinese + Arabic	**51.9**	–	–

Table 4: Event Argument Role Labeling results (F1 %).

In the work of Subburathinam et al. [2019], they exploit language universal features relevant to relation and event argument identification and classification, by way of both symbolic and distributional representations as follows.

Symbolic Representation. To project the multi-lingual data into a common semantic space, one of their steps is to leverage the structure similarity explicitly. They choose dependency trees as the sentence representations because the community has made great efforts at developing language-universal dependency parsing resources across 83 languages. By doing so, the sentences are not regarded as a linear sequence of words which incorporate language-specific information such as word order. Instead, they are represented as the language-universal trees.

Distributional Representation. To further make this tree representation universal across languages, they convert each tree node into a vector which is a concatenation of three language-universal representations at wordlevel: multilingual word embedding, POS embedding, entity-type embedding, and dependency relation embedding. Then a share-weights GCN Encoder is applied on the trees to obtain a contextualized word representations by leveraging neighbors in dependency trees for each node.

Application to Relation Extraction and Event Argument Role Labeling. The learned representations are then fed to the downstream models for relation extraction and event argument role labeling as in Figure 6. The promising results from Table 3, 4 show that the models trained from English are best, followed by Chinese, and then Arabic. The model also benefits from the combination of training data of multiple languages.

References

A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

L. Huang, H. Ji, and J. May. Cross-lingual multi-level adversarial transfer to enhance low-resource name tagging. In *Proceedings of NAACL-HLT*, 2019.

G. Lample and A. Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

T. Mikolov, Q. V. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

T. Pires, E. Schlinger, and D. Garrette. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*, 2019.

R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

A. Subburathinam, D. Lu, H. Ji, J. May, S.-F. Chang, A. Sil, and C. Voss. Cross-lingual structure transfer for relation and event extraction. In *Proceedings of the EMNLP-IJCNLP*, 2019.