Choosing Transfer Languages for Cross-Lingual Learning

Yu-Hsiang Lin*, **Chian-Yu Chen***, **Jean Lee***, **Zirui Li***, **Yuyan Zhang***,
Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma,
Antonios Anastasopoulos, Patrick Littell†, Graham Neubig
Language Technologies Institute, Carnegie Mellon University
†National Research Council, Canada

Abstract

Cross-lingual transfer, where a high-resource *transfer* language is used to improve the accuracy of a low-resource *task* language, is now an invaluable tool for improving performance of natural language processing (NLP) on low-resource languages. However, given a particular task language, it is not clear *which* language to transfer from, and the standard strategy is to select languages based on *ad hoc* criteria, usually the intuition of the experimenter. Since a large number of features contribute to the success of cross-lingual transfer (including phylogenetic similarity, typological properties, lexical overlap, or size of available data), even the most enlightened experimenter rarely considers all these factors for the particular task at hand. In this paper, we consider this task of automatically selecting optimal transfer languages as a ranking problem, and build models that consider the aforementioned features to perform this prediction. In experiments on representative NLP tasks, we demonstrate that our model predicts good transfer languages much better than *ad hoc* baselines considering single features in isolation, and glean insights on what features are most informative for each different NLP tasks, which may inform future *ad hoc* selection even without use of our method.[1]

1 Introduction

A common challenge in applying natural language processing (NLP) techniques to low-resource languages is the lack of training data in the languages in question. It has been demonstrated that through cross-lingual transfer, it is possible to leverage one or more similar high-resource languages to improve the performance on the low-resource languages in several NLP tasks, including machine translation (Zoph et al., 2016; Johnson et al., 2017; Nguyen and Chiang, 2017; Neubig and Hu, 2018), parsing (Täckström et al., 2012; Ammar et al., 2016; Ahmad et al., 2019; Ponti et al., 2018), part-of-speech or morphological tagging (Täckström et al., 2013; Cotterell and Heigold, 2017; Malaviya et al., 2018; Plank and Agić, 2018), named entity recognition (Zhang et al., 2016; Mayhew et al., 2017; Xie et al., 2018), and entity linking (Tsai and Roth, 2016; Rijhwani et al., 2019). There are many methods for performing this transfer, including joint training (Ammar et al., 2016; Tsai and Roth, 2016; Cotterell and Heigold, 2017; Johnson et al., 2017; Malaviya et al., 2018), annotation projection (Täckström et al., 2012; Täckström et al., 2013; Zhang et al., 2016; Ponti et al., 2018; Plank and Agić, 2018), fine-tuning (Zoph et al., 2016; Neubig and Hu, 2018), data augmentation (Mayhew et al., 2017), or zero-shot transfer (Ahmad et al., 2019; Xie et al., 2018; Neubig and Hu,

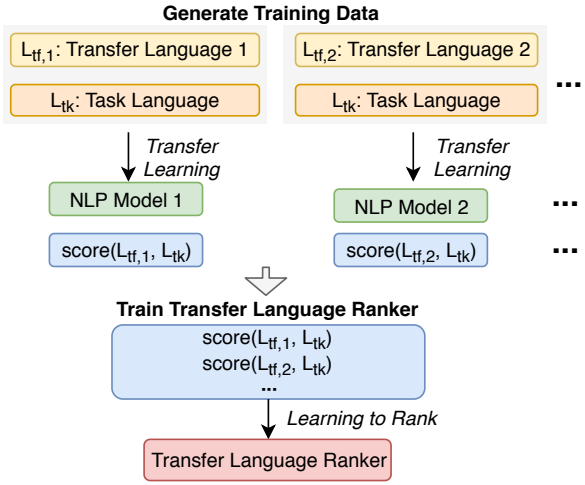

Figure 1: Workflow of learning to select the transfer languages for an NLP task: (1) train a set of NLP models with all available transfer languages and collect evaluation scores, (2) train a ranking model to predict the top transfer languages.

*Equal contribution

[1]Code, data, and pre-trained models are available at https://github.com/neulab/langrank

2018; Rijhwani et al., 2019). The common thread is that data in a high-resource *transfer language* is used to improve performance on a low-resource *task language*.

However, determining the best transfer language for any particular task language remains an open question – the choice of transfer language has traditionally been done in a heuristic manner, often based on the intuition of the experimenter. A common method of choosing transfer languages involves selecting one that belongs to the same language family or has a small phylogenetic distance in the language family tree to the task language (Dong et al., 2015; Johnson et al., 2017; Cotterell and Heigold, 2017). However, it is not always true that all languages in a single language family share the same linguistic properties (Ahmad et al., 2019). Therefore, another strategy is to select transfer languages based on the typological properties that are relevant to the specific NLP task, such as word ordering for parsing tasks (Ammar et al., 2016; Ahmad et al., 2019). With several heuristics available for selecting a transfer language, it is unclear *a priori* if any single attribute of a language will be the most reliable criterion in determining whether cross-lingual learning is likely to work for a specific NLP task. Other factors, such as lexical overlap between the training datasets or size of available data in the transfer language, could also play a role in selecting an appropriate transfer language. Having an empirical principle regarding how to choose the most promising languages or corpora to transfer from has the potential to greatly reduce the time and effort required to find, obtain, and prepare corpora for a particular language pair.

In this paper, we propose a framework, which we call LANGRANK, to empirically answer the question posed above: *given a particular task low-resource language and NLP task, how can we determine which languages we should be performing transfer from?* We consider this language prediction task as a ranking problem, where each potential transfer language is represented by a set of attributes including typological information and corpus statistics, such as word overlap and dataset size. Given a task language and a set of candidate transfer languages, the model is trained to rank the transfer languages according to the performance achieved when they are used in training a model to process the task low-resource language. These models are trained by performing a computation- and resource-intensive exhaustive search through the space of potential transfer languages, but at test time they can rapidly predict optimal transfer languages, based only on a few dataset and linguistic features, which are easily obtained.

In experiments, we examine cross-lingual transfer in four NLP tasks: machine translation (MT), entity linking (EL), part-of-speech (POS) tagging and dependency parsing (DEP). We train gradient boosted decision trees (GBDT; Ke et al. (2017)) to select the best transfer languages based on the aforementioned features. We compare our ranking models with several reasonable baselines inspired by the heuristic approaches used in previous work, and show that our ranking models significantly improve the quality of the selection of the top languages for cross lingual transfer. In addition, through an ablation study and examining the learned decisions trees, we glean insights about which features were found to be useful when choosing transfer languages for each task. This may inform future attempts for heuristic selection of transfer languages, even in the absence of direct use of LANGRANK.

2 Problem Formulation

We define the *task* language t as the language of interest for a particular NLP task, and the *transfer* language a as the additional language that is used to aid in training models. Formally, during the training stage of transfer learning, we perform a model training step:

$$M_{t,a} = \text{train}(\langle x_t^{(trn)}, y_t^{(trn)} \rangle, \langle x_a^{(trn)}, y_a^{(trn)} \rangle),$$

where $x^{(trn)}$ and $y^{(trn)}$ indicate input and output training data for each training language, and $M_{t,a}$ indicates the resulting model trained on languages t and a. The actual model and training procedure will vary from task to task, and we give several disparate examples in our experiments in §5.1. The model can then be evaluated by using it to predict outputs over the test set, and evaluating the results:

$$\hat{y}_{t,a}^{(tst)} = \text{predict}(x_t^{(tst)}; M_{t,a})$$
$$c_{t,a} = \text{evaluate}(y_t^{(tst)}, \hat{y}_{t,a}^{(tst)}),$$

where $c_{t,a}$ is the resulting test-set score achieved by using a as an transfer language.

Assuming we want to get the highest possible performance on task language t, one way to do so

is to exhaustively enumerate over every single potential transfer language a, train models, and evaluate the test set. In this case, the optimal transfer language for task language t can be defined as:

$$a_t^* = \text{argmax}_a c_{t,a}.$$

However, as noted in the introduction, this brute-force method for finding optimal transfer languages is not practical: if resources for many languages are available *a priori*, it is computationally expensive to train all of the models, and in many cases these resources are not-available *a priori* and need to be gathered from various sources before even starting experimentation.

Thus, we turn to formulating our goal as a ranking task: given an NLP task, a low-resource task language t, and a list of J available high-resource transfer languages $a_1, a_2, \ldots, a_J$, attempt to predict their ranking according to their expected scores $c_{t,a_1}, c_{t,a_2}, \ldots, c_{t,a_J}$ without actually calculating the scores themselves. To learn this ranker, we need to first create training data for the ranker, which we create by doing an exhaustive sweep over a set of *training* task languages $t_1, t_2, \ldots, t_I$, which results in sets of scores $\{c_{t_1,a_1}, \ldots, c_{t_1,a_J},\}, \ldots, \{c_{t_I,a_1}, \ldots, c_{t_I,a_J}\}$. These scores that can be used to train a ranking system, using standard methods for learning to rank (see, e.g., Liu et al. (2009)). Specifically, these methods work by extracting features from the pair of languages $\langle t_i, a_j \rangle$:

$$\phi_{t_i,a_j} = \text{feat\_extract}(t_i, a_j)$$

and then using these features to predict a relative score for each pair of task and transfer languages

$$r_{t_i,a_j} = \text{rank\_score}(\phi_{t_i,a_j}; \theta)$$

where θ are the parameters of the ranking model. These parameters θ are learned in a way such that the order of the ranking scores $r_{t_i,a_1}, \ldots, r_{t_i,a_J}$ match as closely as possible with those of the gold-standard evaluation scores $c_{t_i,a_1}, \ldots, c_{t_i,a_J}$.

Now that we have described the overall formulation of the problem, there are two main questions left: how do we define our features ϕ_{t_i,a_j}, and how do we learn the parameters θ of the ranking model?

3 Ranking Features

We represent each language pair/corpus by a set of features, split into two classes: dataset-dependent and dataset-independent.

3.1 Data-dependent Features

Dataset-dependent features are statistical features of the particular corpus used, such as dataset size and the word overlap between two corpora. Importantly, these features require the dataset to already be available for processing and thus are less conducive to use in situations where resources have not yet been acquired. Specifically, we examine the following categories:

Dataset Size: We denote the number of training examples in the transfer and task languages by s_{tf} and s_{tk}, respectively. For MT, POS and DEP, this is the number of sentences in a corpus, and for EL the dataset size is the number of named entities in a bilingual entity gazetteer. In our experiments, we also consider the ratio of the dataset size, s_{tf}/s_{tk}, as a feature, since we are interested in how much bigger the transfer-language corpus is than the task-language corpus.

Type-Token Ratio (TTR): The TTR of the transfer- and task-language corpora, t_{tf} and t_{tk}, respectively, is the ratio between the number of types (the number of unique words) and the number of tokens (Richards, 1987). It is a measure for lexical diversity, as a higher TTR represents higher lexical variation. We also consider the distance between the TTRs of the transfer- and task-language corpora, which may very roughly indicate their morphological similarity:

$$d_{ttr} = \left(1 - \frac{t_{tf}}{t_{tk}}\right)^2.$$

Transfer and task languages that have similar lexical diversity are expected to have d_{ttr} close to 0.

The data for the entity linking task consists only of named entities, so the TTR is typically close to 1 for all languages. Therefore, we do not include TTR related features for the EL task.

Word Overlap and Subword Overlap: We measure the similarity between the vocabularies of task- and transfer-language corpora by word overlap o_w, and subword overlap o_{sw}:

$$o_w = \frac{|T_{tf} \cap T_{tk}|}{|T_{tf}| + |T_{tk}|}, \quad o_{sw} = \frac{|S_{tf} \cap S_{tk}|}{|S_{tf}| + |S_{tk}|},$$

where T_{tf} and T_{tk} are the sets of types in the transfer- and task-language corpora, and S_{tf} and S_{tk} are their sets of subwords. The subwords are obtained by an unsupervised word segmentation algorithm (Sennrich et al., 2016; Kudo,

2018). Note that for EL, we do not consider subword overlap, and the word overlap is simply the count of the named entities that have exactly the same representations in both transfer and task languages. We also omit subword overlap in the POS and DEP tasks, as some low-resource languages do not have enough data for properly extracting subwords in the corpora used for training the POS and DEP models in our experiments.

3.2 Dataset-independent Features

Dataset-independent features are measures of the similarity between a pair of languages based on phylogenetic or typological properties established by linguistic study. Specifically, we leverage six different linguistic distances queried from the URIEL Typological Database (Littell et al., 2017):

Geographic distance (d_{geo}): The orthodromic distance between the languages on the surface of the earth, divided by the antipodal distance, based primarily on language location descriptions in Glottolog (Hammarström et al., 2018).

Genetic distance (d_{gen}): The genealogical distance of the languages, derived from the hypothesized tree of language descent in Glottolog.

Inventory distance (d_{inv}): The cosine distance between the phonological feature vectors derived from the PHOIBLE database (Moran et al., 2014), a collection of seven phonological databases.

Syntactic distance (d_{syn}): The cosine distance between the feature vectors derived from the syntactic structures of the languages (Collins and Kayne, 2011), derived mostly from the WALS database (Dryer and Haspelmath, 2013).

Phonological distance (d_{pho}): The cosine distance between the phonological feature vectors derived from the WALS and Ethnologue databases (Lewis, 2009).

Featural distance (d_{fea}): The cosine distance between feature vectors combining all 5 features mentioned above.

4 Ranking Model

Having defined our features, the next question is what type of ranking model to use and how to learn its parameters θ. As defined in §2, the problem is a standard learning-to-rank problem, so there are a myriad of possibilities for models and learning algorithms (Liu et al., 2009), and any of them would be equally applicable to our task.

We opt to use the GBDT (Ke et al., 2017) model with LambdaRank as our training method (Burges, 2010). This method works by learning an ensemble of decision-tree-based learners using gradient boosting, and specifically in our setting here has two major advantages. First, its empirical performance – it is currently one of the state-of-the-art methods for ranking, especially in settings that have few features and limited data. Second, but perhaps more interesting, is its interpretability. Decision-tree based algorithms are relatively interpretable, as it is easy to visualize the learned tree structure. One of our research goals is to understand what linguistic or statistical features of a dataset play important roles in transfer learning, so the interpretable nature of the tree-based model can provide valuable insights, which we elaborate further in §6.2.

5 Experimental Settings

5.1 Testbed Tasks

We investigate the performance of LANGRANK on four common NLP tasks: machine translation, entity linking, POS tagging, and dependency parsing. We briefly outline the settings for all four NLP tasks, which are designed based on previous work on transferring between languages in these settings (Neubig and Hu, 2018; Rijhwani et al., 2019; Kim et al., 2017; Ahmad et al., 2019).

Machine Translation We train a standard attention-based sequence-to-sequence model (Bahdanau et al., 2015), using the XNMT toolkit (Neubig et al., 2018). We perform training on the multilingual TED talk corpus of Qi et al. (2018), using 54 task and 54 transfer languages, always translating into English, which results in 2,862 task/transfer pairs and 54 single-source training settings. Transfer is performed by joint training over the concatenated task and transfer corpora, and subwords are learned over the concatenation of both corpora (Sennrich et al., 2016).

Entity Linking The cross-lingual EL task involves linking a named entity mention in the task language to an English knowledge base. We train two character-level LSTM encoders, which are trained to maximize the cosine similarity between parallel (i.e., linked) entities (Rijhwani

et al., 2019). We use the same dataset as Rijhwani et al. (2019), which contains language-linked Wikipedia article titles from 9 low-resource task languages and 53 potential transfer languages, resulting in 477 task/transfer pairs. We perform training in a zero-shot setting, where we train on corpora only in the transfer language, and test entity linking accuracy on the task language without joint training or fine-tuning.

POS Tagging We train a bi-directional LSTM-CNNs-CRF model (Ma and Hovy, 2016) on word sequences without using pre-trained word embeddings. The implementation is based on the NCRF++ toolkit (Yang and Zhang, 2018). We perform training on the Universal Dependencies v2.2 dataset (Nivre et al., 2018), using 26 languages that have the least training data as task languages, and 60 transfer languages,[2] resulting in 1,545 pairs of transfer-task languages. Transfer is performed by joint training over the concatenated task and transfer corpora if the task language has training data, and training only with transfer corpora otherwise. The performance is measured by POS tagging accuracy on the task language.

Dependency Parsing For the dependency parsing task, we utilize a deep biaffine attentional graph-based model (Dozat and Manning, 2016). We select 30 languages from Universal Dependencies v2.2 (Nivre et al., 2018), resulting in 870 pairs of transfer-task languages. The selection basically follows the settings of Ahmad et al. (2019), but we exclude Japanese (ja) since we observe unstable results on it. For this task, transfer is performed in the zero-shot setting where no task language annotations are available in training. We rely on the multi-lingual embeddings which are mapped into the same space with the offline method of Smith et al. (2017) and directly adopt the model trained with the transfer language to task languages. The performance is measured by LAS (Labeled Attachment Accuracy) excluding punctuation.

5.2 Evaluation Protocol

We evaluate all our models on all NLP tasks with leave-one-out cross validation. For each cross-validation fold, we leave one language $\ell^{(tst)}$ out from the N languages we have as the test set, and train our ranking model $\theta_{\ell^{(tst)}}$ using all remaining languages, $\{\ell_1^{(trn)}, \ldots, \ell_{N-1}^{(trn)}\}$, as the training set. During training, each $\ell_i^{(trn)}$ is treated as the task language in turn, and the other $N-2$ languages in the training set as transfer languages. We then test the learned model $\theta_{\ell^{(tst)}}$ by taking $\ell^{(tst)}$ as the task language, and $\{\ell_1^{(trn)}, \ldots, \ell_{N-1}^{(trn)}\}$ as the set of transfer languages, and predict the ranking scores $\{r_{\ell^{(tst)},\ell_1^{(trn)}}, \ldots, r_{\ell^{(tst)},\ell_{N-1}^{(trn)}}\}$. We repeat this process with each language in all N languages as the test language $\ell^{(tst)}$, and collect N learned models.

We use Normalized Discounted Cumulative Gain (NDCG) (Järvelin and Kekäläinen, 2002) to evaluate the performance of the ranking model. The NDCG at position p is defined as:

$$\text{NDCG}\ @p = \frac{\text{DCG}\ @p}{\text{IDCG}\ @p},$$

where the Discounted Cumulative Gain (DCG) at position p is

$$\text{DCG}\ @p = \sum_{i=1}^{p} \frac{2^{\gamma_i} - 1}{\log_2(i+1)}.$$

Here γ_i is the relevance of the language ranked at position i by the model being evaluated. We keep only the top-$\gamma_{\max}$ transfer languages as our learning signal: the true best transfer language has $\gamma = \gamma_{\max}$, and the second-best one has $\gamma = \gamma_{\max} - 1$, and so on until $\gamma = 1$, with the remaining languages below the top-$\gamma_{\max}$ ones all sharing $\gamma = 0$. The Ideal Discounted Cumulative Gain (IDCG) uses the same formula as DCG, except it is calculated over the gold-standard ranking. When the predicted ranking matches the "true" ranking, then NDCG is equal to 1.

5.3 Method Parameters and Baselines

We use GBDT to train our LANGRANK models. For each LANGRANK model, we train an ensemble of 100 decision trees, each with 16 leaves. We use the LightGBM implementation (Ke et al., 2017) of the LambdaRank algorithm in our training. In our experiments, we set $\gamma_{\max} = 10$, and evaluate the models by NDCG@3. The threshold of 3 was somewhat arbitrary, but based on our intuition that we would like to test whether LANGRANK can successfully recommend the best transfer language within a few tries, instead of testing its ability to accurately rank *all* available transfer languages. The results in Table 1

[2] For each language, we choose the treebank that has the least number of training instances, which results in 60 languages with training data and 11 without training data.

	Method	MT	EL	POS	DEP
dataset	word overlap o_w	28.6	30.7	13.4	52.3
	subword overlap o_{sw}	29.2	–	–	–
	size ratio s_{tf}/s_{tk}	3.7	0.3	9.5	24.8
	type-token ratio d_{ttr}	2.5	–	7.4	6.4
ling. distance	genetic d_{gen}	24.2	50.9	14.8	32.0
	syntactic d_{syn}	14.8	46.4	4.1	22.9
	featural d_{fea}	10.1	47.5	5.7	13.9
	phonological d_{pho}	3.0	4.0	9.8	43.4
	inventory d_{inv}	8.5	41.3	2.4	23.5
	geographic d_{geo}	15.1	49.5	15.7	46.4
LANGRANK (all)		51.1	**63.0**	**28.9**	**65.0**
LANGRANK (dataset)		**53.7**	17.0	26.5	**65.0**
LANGRANK (URIEL)		32.6	58.1	16.6	59.6

Table 1: Our LANGRANK model leads to higher average NDCG@3 over the baselines on all four tasks: machine translation (MT), entity linking (EL), part-of-speech tagging (POS) and dependency parsing (DEP).

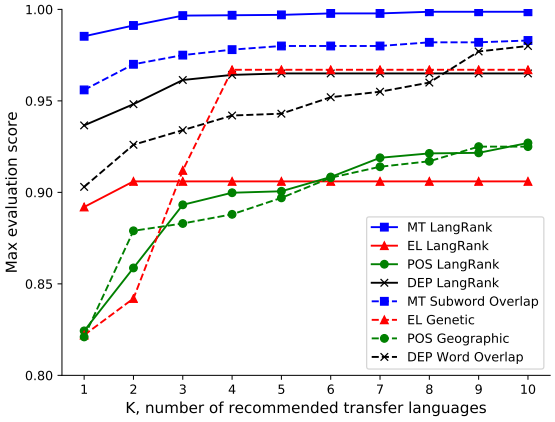

Figure 2: The best evaluation score (BLEU for MT, accuracy for EL and POS, and LAS for DEP) attainable by trying out the top K transfer languages recommended by the LANGRANK models and the single-feature baselines.

report the average NDCG@3 across all cross-validation folds. For LANGRANK (all) we include all available features in our models, while for LANGRANK (dataset) and LANGRANK (ling) we include only the subsets of dataset-dependent and dataset-independent features, respectively.

We consider the following baseline methods:

- Using a single dataset-dependent feature: While dataset-dependent features have not typically been used as criteria for selecting transfer languages, they are a common feature in data selection methods for cross-domain transfer (Moore and Lewis, 2010). In view of this, we include selecting the transfer languages by sorting against each single one of o_w, o_{sw}, and s_{tf}/s_{tk} in descending order, and sorting against d_{ttr} in ascending order, as baseline methods.

- Using a single linguistic distance feature: More common heuristic criteria of selection the transfer languages are choosing ones that have small phylogenetic distance to the task language (Dong et al., 2015; Cotterell and Heigold, 2017). We therefore include selecting the transfer languages by sorting against each single one of d_{gen}, d_{syn}, d_{fea}, d_{pho}, d_{inv}, and d_{geo} in ascending order as our baseline methods.

6 Results and Analysis

6.1 Main Results

The performance of predicting transfer languages for the four NLP tasks using single-feature baselines and LANGRANK is shown in Table 1. First, using LANGRANK with either all features or a subset of the features leads to substantially higher NDCG than using single-feature heuristics. Although some single-feature baselines manage to achieve high NDCG for some tasks, the predictions of LANGRANK consistently surpass the baselines on all tasks. In fact, for the MT and POS tagging tasks, the ranking quality of the best LANGRANK model is almost double that of the best single-feature baseline.

Furthermore, using dataset-dependent features on top of the linguistic distance ones enhances the quality of the LANGRANK predictions. The best results for EL and POS tagging are obtained using all features, while for MT the best model is the one using dataset-only features. The best performance on DEP parsing is achieved with both settings. LANGRANK with only dataset features outperforms the linguistics-only LANGRANK on the MT and POS tagging tasks. It is, however, severely lacking in the EL task, likely because EL datasets lack most dataset features as discussed in the previous section; the EL data only consists of pairs of corresponding entities and not complete sentences as in the case of the other tasks' datasets.

In addition, it is important to note that LANGRANK with only linguistic database information still outperforms all heuristic baselines on all tasks. This means that our model is potentially useful even before any resources for the language and task of interest have been collected, and could inform the data creation process.

Finally, from a potential user's point of view,

Task Lang	LANG RANK	Best Dataset	Best URIEL	True Best
		o_w	d_{fea}	
MT	tur (1)	tur (1)	ara (32)	tur (1)
aze	fas (3)	hrv (5)	fas (3)	kor (2)
	hun (4)	ron (31)	sqi (22)	fas (3)
		o_w	d_{geo}	
MT	hun (1)	vie (3)	mya (30)	hun (1)
ben	tur (2)	ita (20)	hin (27)	tur (2)
	fas (4)	por (18)	mar (41)	vie (3)
		o_w	d_{inv}	
EL	amh (6)	amh (6)	pan (2)	hin (1)
tel	orm (40)	swa (32)	hin (1)	pan (2)
	msa (7)	jav (9)	ben (5)	mar (3)

Table 2: Examples of predicted top-3 transfer languages (and true ranks). The languages are denoted by the ISO 639-2 Language Codes. The first two task languages (aze, ben) are on the MT task, and the last one (tel) is on the EL task.

a practical question is: *If we train models on the top K transfer languages suggested by the ranking model and pick the best one, how good is the best model expected to be?* If a user could obtain a good transfer model by trying out only a small number of transfer languages as suggested by our ranking model, the overhead of searching for a good transfer language is immensely reduced.

Figure 2 compares the BLEU score (for MT), accuracy (for EL and POS) and LAS (for DEP) of the best transfer model attainable by using one of the top K transfer languages recommended by LANGRANK (all) and by the best single feature baseline. We plot the ratio of the best score to that of the ground-truth best transfer model c_{t,a_t^*}, averaged over all task languages. On the MT task, the best transfer models obtained by the suggestions of our LANGRANK (all) model constantly outperforms the models obtained from the best baseline. On the POS tagging task, the best transfer models obtained by our ranking model are generally comparable to those using baseline suggestions.

We note that in the EL task, after looking beyond the top 3 LANGRANK predictions, the best baseline models on average seem to give more relevant transfer language suggestions than our LANGRANK models. However, this is a case where averaging is possibly misleading. In fact, the LANGRANK model manages to select the correct top-1 language for 7 of the 9 task languages. The other two languages (Telugu and Uyghur) do not have any typologically similar languages in the small training set, and hence the learned model fails to generalize to these languages.

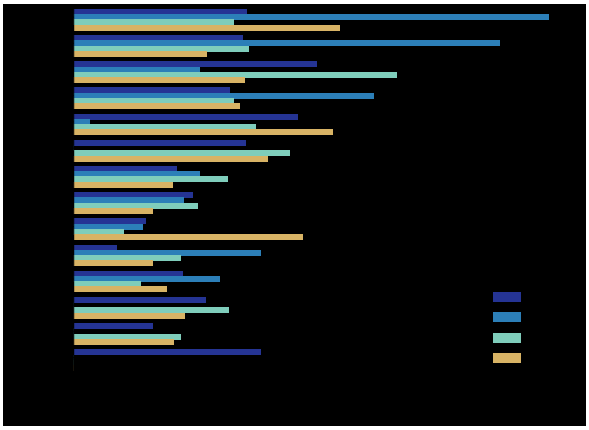

Figure 3: Normalized feature importance for the MT, EL, POS and DEP tasks.

In Table 2 we include a few representative examples of the top-3 transfer languages selected by LANGRANK and the baselines.[3] In the first case (aze) LANGRANK outperforms the already strong baselines by being able to consider both dataset and linguistic features, instead of considering them in isolation. In the second case (ben) where no baselines provide useful recommendations, LANGRANK still displays good performance; interestingly Turkish and Hungarian proved good transfer languages for a large number of task languages (perhaps to large data size and difficulty as tasks), and LANGRANK was able to learn to fall back to these when it found no good typological or dataset-driven matches otherwise – behavior that would have be inconceivable without empirical discovery of transfer languages. The final failure case (tel), as noted above, can be attributed to overfitting the small EL dataset, and may be remedied by either creating larger data or training LANGRANK jointly over multiple tasks.

6.2 Towards Better Educated Guesses for Choosing Transfer Languages

Our transfer language rankers are trained on a few languages for the particular tasks. It is possible that our models will not generalize well on a different set of languages or on other NLP tasks. However, generating training data for ranking with exhaustive transfer experiments on a new task or set of languages will not always be feasible. It could, therefore, be valuable to analyze the learned models and extract "rules of thumb" that can be

[3] Detailed results are in the supplementary material.

used as educated guesses in choosing transfer languages. They might still be ad-hoc, but they may prove superior to the intuition-based heuristic approaches used in previous work. To elucidate how LANGRANK determines the best transfer languages for each task, Figure 3 shows the feature importance for each of the NLP tasks. The feature importance is defined as the number of times a feature is chosen to be the splitting feature in a node of the decision trees.

For the MT task, we find that dataset statistics features are more influential than the linguistic features, especially the dataset size ratio and the word overlap. This indicates that a good transfer language for machine translation depends more on the dataset size of the transfer language corpus and its word and subword overlap with the task language corpus. This is confirmed by results of the LANGRANK (dataset) model in Table 1, which achieves the best performance by only using the subset of dataset statistics features. At the same time, we note that the dataset size ratio and TTR distance, although of high importance among all features, when used alone result in very poor performance. This phenomenon may be understood by looking at an example of a small decision tree in Figure 4: a genetic distance of less than 0.4 would produce a high ranking regardless of dataset size. The dataset feature in this tree provides a smaller gain than two typological features, although it still informs the decision.

For POS tagging, the two most important features are dataset size and the TTR distance. On the other hand, the lack of rich dataset-dependent features for the EL task leads to the geographic and syntactic distance being most influential. There are several relatively important features for the DEP parsing task, with geographic and genetic distance standing out, as well as word overlap. These are features that also yield good scores on their own (see Table 1) but LANGRANK is able to combine them and achieve even better results.

7 Related Work

Cross-lingual transfer has been extensively used in several NLP tasks. In Section 1, we provided a (non-exhaustive) list of examples that employ cross-lingual transfer across several tasks. Other work has performed large-scale studies on the importance of appropriately selecting a transfer language, such as Paul et al. (2009), which performed

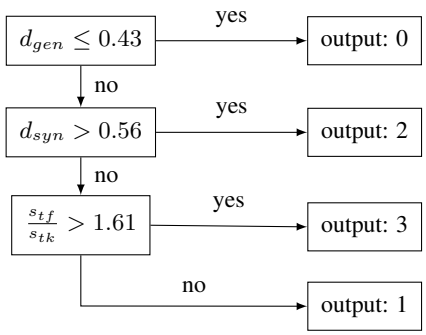

Figure 4: An example of the decision tree learned in the machine translation task for Galician as task language.

an extensive search for a "pivot language" in statistical MT, but without attempting to actually learn or predict which pivot language is best.

Typologically-informed models are another vein of research that is relevant to our work. The relationship between linguistic typology and statistical modeling has been studied by Gerz et al. (2018) and Cotterell et al. (2018), with a focus on language modeling. Tsvetkov et al. (2016b) used typological information in the target language as additional input to their model for phonetic representation learning. Ammar et al. (2016) and Ahmad et al. (2019) used similar ideas for dependency parsing, incorporating linguistically-informed vectors into their models. O'Horan et al. (2016) survey typological resources available and their utility in NLP tasks.

Although not for cross-lingual transfer, there has been prior work on data selection for training models. Tsvetkov et al. (2016a) and Ruder and Plank (2017) use Bayesian optimization for data selection. van der Wees et al. (2017) study the effect of data selection of neural machine translation, as well as propose a dynamic method to select relevant training data that improves translation performance. Plank and van Noord (2011) design a method to automatically select domain-relevant training data for parsing in English and Dutch.

8 Conclusion

We formulate the task of selecting the optimal transfer languages for an NLP task as a ranking problem. For machine translation, entity linking, part-of-speech tagging, and dependency parsing, we train ranking models to predict the most promising transfer languages to use given a task language. We show that by taking multiple dataset statistics and language attributes into

consideration, the learned ranking models recommend much better transfer languages than the ones suggested by considering only single language or dataset features. Through analyzing the learned ranking models, we also gain some insights on the types of features that are most influential in selecting transfer languages for each of the NLP tasks, which may inform future *ad hoc* selection even without using our method.

Acknowledgments

This project was supported in part by NSF Award No. 1761548 "Discovering and Demonstrating Linguistic Features for Language Documentation," and the Defense Advanced Research Projects Agency Information Innovation Office (I2O) Low Resource Languages for Emergent Incidents (LORELEI) program under Contract No. HR0011-15-C0114. The views and conclusions contained in this doc- ument are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Proceedings of NAACL*.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *ICLR 2015 (arXiv:1409.0473)*.

Chris J.C. Burges. 2010. From RankNet to LambdaRank to LambdaMART: An overview. Technical report, Microsoft Research.

Chris Collins and Richard Kayne. 2011. Syntactic structures of the world's languages.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual character-level neural morphological tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759, Copenhagen, Denmark. Association for Computational Linguistics.

Ryan Cotterell, Sebastian J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541. Association for Computational Linguistics.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327. Association for Computational Linguistics.

Harald Hammarström, Robert Forkel, and Martin Haspelmath. 2018. Glottolog 3.3. Max Planck Institute for the Science of Human History.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernand a Vigas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3146–3154. Curran Associates, Inc.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for POS tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on*

Empirical Methods in Natural Language Processing, pages 2832–2838, Copenhagen, Denmark. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75. Association for Computational Linguistics.

Paul M Lewis. 2009. *Ethnologue: Languages of the World Sixteenth edition*. Dallas, Texas: SIL International.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 8–14.

Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1064–1074.

Chaitanya Malaviya, Matthew R. Gormley, and Graham Neubig. 2018. Neural factor graph models for cross-lingual morphological tagging. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2653–2663, Melbourne, Australia. Association for Computational Linguistics.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Steven Moran, Daniel McCloy, and Richard Wright, editors. 2014. *PHOIBLE Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium.

Graham Neubig, Matthias Sperber, Xinyi Wang, Matthieu Felix, Austin Matthews, Sarguna Padmanabhan, Ye Qi, Devendra Singh Sachan, Philip Arthur, Pierre Godard, John Hewitt, Rachid Riad, and Liming Wang. 2018. XNMT: The extensible neural machine translation toolkit. In *Conference of the Association for Machine Translation in the Americas (AMTA) Open Source Software Showcase*, Boston.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proc. IJCNLP*, volume 2, pages 296–301.

Joakim Nivre, Mitchell Abrams, Željko Agić, and et al. 2018. Universal dependencies 2.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Helen O'Horan, Yevgeni Berzak, Ivan Vulic, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1297–1308, Osaka, Japan. The COLING 2016 Organizing Committee.

Michael Paul, Hirofumi Yamamoto, Eiichiro Sumita, and Satoshi Nakamura. 2009. On the importance of pivot language selection for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 221–224. Association for Computational Linguistics.

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 and Gertjan van Noord. 2011. Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1566–1576, Portland, Oregon, USA. Association for Computational Linguistics.

Edoardo Maria Ponti, Roi Reichart, Anna Korhonen, and Ivan Vuli. 2018. Isomorphic transfer of syntactic structures in cross-lingual nlp. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1531–1542, Melbourne, Australia. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful 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 2 (Short Papers), pages 529–535. Association for Computational Linguistics.

Brian Richards. 1987. Type/token ratios: What do they really tell us? *Journal of child language*, 14(2):201–209.

Shruti Rijhwani, Jiateng Xie, Graham Neubig, and Jaime Carbonell. 2019. Zero-shot neural transfer for cross-lingual entity linking. In *Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, Honolulu, Hawaii.

Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with Bayesian Optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 372–382, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*.

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.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 477–487, Montréal, Canada. Association for Computational Linguistics.

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 589–598, San Diego, California. Association for Computational Linguistics.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Brian MacWhinney, and Chris Dyer. 2016a. Learning the Curriculum with Bayesian Optimization for Task-Specific Word Representation Learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 130–139, Berlin, Germany. Association for Computational Linguistics.

Yulia Tsvetkov, Sunayana Sitaram, Manaal Faruqui, Guillaume Lample, Patrick Littell, David Mortensen, Alan W Black, Lori Levin, and Chris Dyer. 2016b. Polyglot neural language models: A case study in cross-lingual phonetic representation learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1357–1366, San Diego, California. Association for Computational Linguistics.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic Data Selection for Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1410, Copenhagen, Denmark. Association for Computational Linguistics.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379. Association for Computational Linguistics.

Jie Yang and Yue Zhang. 2018. Ncrf++: An open-source neural sequence labeling toolkit. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Dongxu Zhang, Boliang Zhang, Xiaoman Pan, Xiaocheng Feng, Heng Ji, and Weiran XU. 2016. Bitext name tagging for cross-lingual entity annotation projection. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 461–470, Osaka, Japan. The COLING 2016 Organizing Committee.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.