Cross-Cultural Similarity Features for Cross-Lingual Transfer Learning of Pragmatically Motivated Tasks

Jimin Sun[1]* Hwijeen Ahn[2]* Chan Young Park[3]*
Yulia Tsvetkov[3] David R. Mortensen[3]
[1]Seoul National University, Republic of Korea
[2]Sogang University, Republic of Korea
[3]Language Technologies Institute, Carnegie Mellon University, USA
jiminsun@dm.snu.ac.kr, hwijeen@sogang.ac.kr,
{chanyoun, ytsvetko, dmortens}@cs.cmu.edu

Abstract

Much work in cross-lingual transfer learning explored how to select better transfer languages for multilingual tasks, primarily focusing on typological and genealogical similarities between languages. We hypothesize that these measures of linguistic proximity are not enough when working with pragmatically-motivated tasks, such as sentiment analysis. As an alternative, we introduce three linguistic features that capture cross-cultural similarities that manifest in linguistic patterns and quantify distinct aspects of language pragmatics: language context-level, figurative language, and the lexification of emotion concepts. Our analyses show that the proposed pragmatic features do capture cross-cultural similarities and align well with existing work in sociolinguistics and linguistic anthropology. We further corroborate the effectiveness of pragmatically-driven transfer in the downstream task of choosing transfer languages for cross-lingual sentiment analysis.

1 Introduction

Hofstede et al. (2005) defined culture as the collective mind which "distinguishes the members of one group of people from another." Cultural idiosyncrasies affect and shape people's beliefs and behaviors. Linguists have particularly focused on the relationship between culture and language, revealing in qualitative case studies how cultural differences are manifested as linguistic variations (Siegel, 1977).

Quantifying cross-cultural similarities from linguistic patterns has largely been unexplored in NLP, with the exception of studies that focused on cross-cultural differences in word usage (Garimella et al., 2016; Lin et al., 2018). In this work, we aim to quantify cross-cultural similarity, focusing on *semantic* and *pragmatic* differences across languages.[1] We devise a new distance measure between languages based on linguistic proxies of culture. We hypothesize that it can be used to select transfer languages and improve cross-lingual transfer learning, specifically in pragmatically-motivated tasks such as sentiment analysis, since expressions of subtle sentiment or emotion—such as subjective well-being (Smith et al., 2016), anger (Oster, 2019), or irony (Karoui et al., 2017)—have been shown to vary significantly by culture.

We focus on three distinct aspects in the intersection of language and culture, and propose features to operationalize them. First, every language and culture rely on different levels of *context in communication*. Western European languages are generally considered low-context languages, whereas Korean and Japanese are considered high-context languages (Hall, 1989). Second, similar cultures construct and construe *figurative language* similarly (Casas and Campoy, 1995; Vulanović, 2014). Finally, *emotion semantics* is similar between languages that are culturally-related (Jackson et al., 2019). For example, in Persian, 'grief' and 'regret' are expressed with the same word whereas 'grief' is co-lexified with 'anxiety' in Dargwa. Therefore, Persian speakers may perceive 'grief' as more similar to 'regret,' while Dargwa speakers may associate the concept with 'anxiety.'

We validate the proposed features qualitatively, and also quantitatively by an extrinsic evaluation method. We first analyze each linguistic feature

*The first three authors contributed equally.

[1] In linguistics, *pragmatics* has both a broad and a narrow sense. Narrowly, the term refers to formal pragmatics. In the broad sense, which we employ in this paper, pragmatics refers to contextual factors in language use. We are particularly concerned with cross-cultural pragmatics and finding quantifiable linguistic measures that correspond to aspects of cultural context. These measures are not the cultural characteristics that would be identified by anthropological linguists themselves but are rather intended to be measurable correlates of these characteristics.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2403–2414
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

to confirm that they capture the intended cultural patterns. We find that the results corroborate the existing work in sociolinguistics and linguistic anthropology. Next, as a practical application of our features, we use them to rank transfer languages for cross-lingual transfer learning. Lin et al. (2019) have shown that selecting the right set of transfer languages with syntactic and semantic language-level features can significantly boost the performance of cross-lingual models. We incorporate our features into Lin et al. (2019)'s ranking model to evaluate the new cultural features' utility in selecting better transfer languages. Experimental results show that incorporating the features improves the performance for cross-lingual sentiment analysis, but not for dependency parsing. These results support our hypothesis that cultural features are more helpful when the cross-lingual task is driven by pragmatic knowledge. [2]

2 Pragmatically-motivated Features

We propose three language-level features that quantify the cultural similarities across languages.

Language Context-level Ratio A language's *context-level* reflects the extent to which the language leaves the identity of entities and predicates to context. For example, an English sentence *Did you eat lunch?* explicitly indicates the pronoun *you*, whereas the equivalent Korean sentence 점심 먹었니? (= *Did eat lunch?*) omits the pronoun. Context-level is considered one of the distinctive attributes of a language's pragmatics in linguistics and communication studies, and if two languages have similar levels of context, their speakers are more likely to be from similar cultures (Nada et al., 2001).

The language context-level ratio (LCR) feature approximates this linguistic quality. We compute the pronoun- and verb-token ratio, $\texttt{ptr}(l_k)$ and $\texttt{vtr}(l_k)$ for each language l_k, using part-of-speech tagging results. We first run language-specific POS-taggers over a large mono-lingual corpus for each language. Next, we compute ptr as the ratio of count of pronouns in the corpus to the count of all tokens. vtr is obtained likewise with verb tokens. Low ptr, vtr values may indicate that a language leaves the identity of entities and predicates, respectively, to context. We then compare these values between the *target language* l_{tg} and *transfer language* l_{tf}, which leads to the following definition of LCR:

$$\texttt{LCR-pron}(l_{tf}, l_{tg}) = \texttt{ptr}(l_{tg})/\texttt{ptr}(l_{tf})$$
$$\texttt{LCR-verb}(l_{tf}, l_{tg}) = \texttt{vtr}(l_{tg})/\texttt{vtr}(l_{tf})$$

Literal Translation Quality Similar cultures tend to share similar figurative expressions, including idiomatic multiword expressions (MWEs) and metaphors (Kövecses, 2003, 2010). For example, *like father like son* in English can be translated word-by-word into a similar idiom *tel père tel fils* in French. However, in Japanese, a similar idiom 蛙の子は蛙 (*Kaeru no ko wa kaeru*) "A frog's child is a frog." cannot be literally translated.

Literal translation quality (LTQ) feature quantifies how well a given language pair's MWEs are preserved in literal (word-by-word) translation, using a bilingual dictionary. A well-curated list of MWEs is not available for the majority of languages. We thus follow an automatic extraction approach of MWEs (Tsvetkov and Wintner, 2010). First, a variant of pointwise mutual information, PMI[3] (Daille, 1994) is used to extract noisy lists of top-scoring n-grams from two large monolingual corpora from different domains, and intersecting the lists filters out domain-specific n-grams and retains the language-specific top-k MWEs. Then, a bilingual dictionary between l_{tf} and l_{tg} and a parallel corpus between the pair are used. [3] For each n-gram in l_{tg}'s MWEs, we search for its literal translations extracted using the dictionary in parallel sentences containing the n-gram. For any word in the n-gram, if there is a translation in the parallel sentence, we consider this as hit, otherwise as miss. And we calculate *hit ratio* as $\frac{hit}{(hit+miss)}$ for each n-gram found in the parallel corpus. Finally, we average the hit ratios of all n-grams and z-normalize over the transfer languages to obtain $\texttt{LTQ}(l_{tf}, l_{tg})$.

Emotion Semantics Distance Emotion semantic distance (ESD) measures how similarly emotions are lexicalized across languages. This is inspired by Jackson et al. (2019) who used colexification patterns (i.e., when different concepts are expressed using the same lexical item) to capture the semantic similarity of languages. However, colexification patterns require human annotation,

[2] Code and data are publicly available at https://github.com/hwijeen/langrank.

[3] While dictionaries and parallel corpora are not available for many languages, they are easier to obtain than the task-specific annotations of MWEs.

and existing annotations may not be comprehensive. We extend Jackson et al. (2019)'s method by using cross-lingual word embeddings.

We define ESD as the average distance of emotion word vectors in transfer and target languages, after aligning word embeddings into the same space. More specifically, we use 24 emotion concepts defined in Jackson et al. (2019) and use bilingual dictionaries to expand each concept into every other language (e.g., *love* and *proud* to *Liebe* and *stolz* in German). We then remove the emotion word pairs from the bilingual dictionaries, and use the remaining pairs to align word embeddings of source into the space of target languages. We hypothesize that if words correspond to the same emotion concept in different languages (e.g., *proud* and *stolz*) have similar meaning, they should be aligned to the same point despite the lack of supervision. However, because each language possesses different emotion semantics, emotions are scattered into different positions. We thus define ESD as the average cosine distance between languages:

$$\mathtt{ESD}(l_{tf}, l_{tg}) = \sum_{e \in E} \cos(\mathbf{v}_{tf,e}, \mathbf{v}_{tg,e})/|E|$$

where E is the set of emotion concepts and $\mathbf{v}_{tf,e}$ is the aligned emotion word vector of language l_{tf}.

3 Feature Analysis

In this section, we evaluate the proposed pragmatically-motivated features intrinsically. Throughout the analyses, we use 16 languages listed in Figure 4 which are later used for extrinsic evaluation (§5).

3.1 Implementation Details

We used multilingual word tokenizers from NLTK and RDR POS Tagger (Nguyen et al., 2014) for most of the languages except for Arabic, Chinese, Japanese, and Korean, where we used PyArabic, Jieba, Kytea, and Mecab, respectively. For monolingual corpora, we used the news-crawl 1M corpora from Leipzig (Goldhahn et al., 2012) for both LCR and LTQ. We used bilingual dictionaries from Choe et al. (2020) and TED talks corpora (Qi et al., 2018) for both parallel corpora and an additional monolingual corpus for LTQ. We focused on bigrams and trigrams and set k, the number of extracted MWEs, to 500. We followed Lample et al. (2018) to generate the supervised cross-lingual word embeddings for ESD.

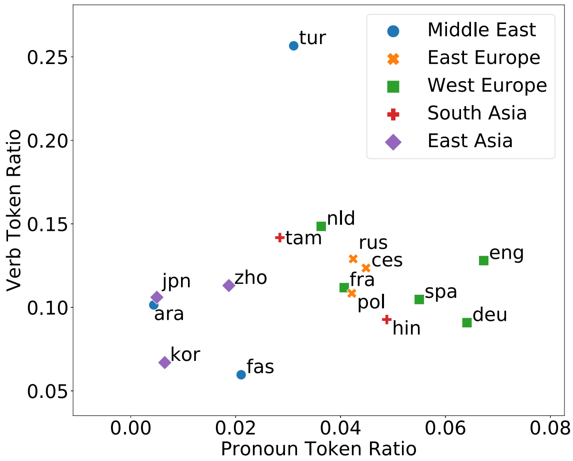


Figure 1: Plot of languages in ptr and vtr plane. Languages are color-coded according to the cultural areas defined in Siegel (1977).

3.2 LCR and Language Context-level

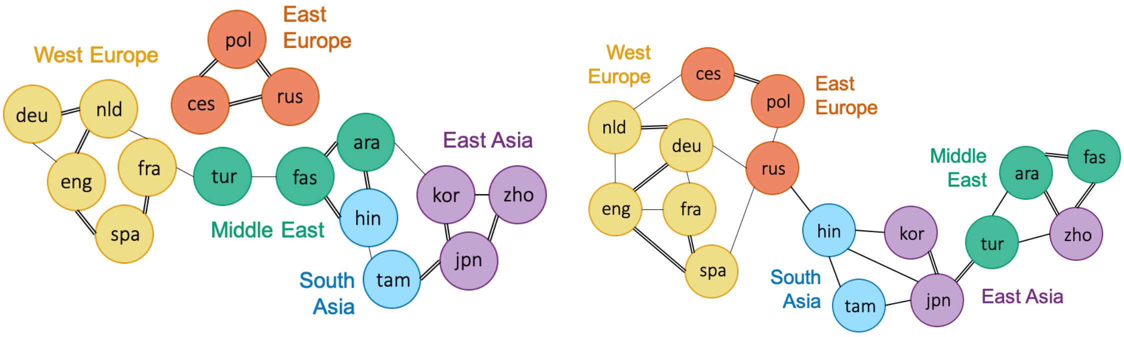
ptr approximates how often discourse entities are indexed with pronouns rather than left conjecturable from context. Similarly, vtr estimates the rate at which predicates appear explicitly as verbs. In order to examine to which extent these features reflect context-levels, we plot languages on a two-dimensional plane where the x-axis indicates ptr and the y-axis indicates vtr in Figure 1.

The plot reveals a clear pattern of context-levels in different languages. Low-context languages such as German and English (Hall, 1989) possess the largest values of ptr. On the other extreme are located Korean and Japanese with low ptr, which are representative of high-context languages. One thing to notice is the isolated location of Turkish with a high vtr. This is morphosyntactically plausible as a lot of information is expressed by the affixation to verbs in Turkish.

3.3 LTQ and MWEs

LTQ uses n-grams with high PMI scores as proxies for figurative language MWE (PMI MWEs). We evaluate the quality of selected MWEs and the resulting LTQ by comparing with human-curated list of figurative language MWE (gold MWEs) that are available in some languages. We collected gold MWEs in multiple languages from Wiktionary[4]. We discarded languages with less than 2,000 phrases on the list, resulting in four languages (English, French, German, Spanish) for

[4] For example, https://en.wiktionary.org/wiki/Category:English_idioms



(a) Network based on Emotion Semantics Distance.

(b) Network based on syntactic distance.

Figure 2: Network of languages color-coded by their cultural areas. An edge is added between the two languages if a language is ranked in the top-2 closest languages of the other language in terms of feature value.

analysis.

First, we check how many PMI MWEs are actually in the gold MWEs. Out of the top-500 PMI bigrams and trigrams, 19.0% of bigrams and 3.8% of trigrams are included in the gold MWE list (averaged over four languages). For example, the trigrams in the PMI MWEs, *keep an eye* and *take into account*, are considered to be in the gold MWEs as *keep an eye peeled* and *take into account* are in the list. The seemingly low percentages are reasonable, regarding that the PMI scores are designed to extract collocations patterns rather than figurative languages themselves.

Secondly, to validate using PMI MWEs as proxies, we compare the LTQ of PMI MWEs with the LTQ using gold MWEs. Specifically, we obtained the LTQ scores of each language pair with target languages limited to the four European languages mentioned above. Then for each target language, we measured Pearson correlation coefficient between the two LTQ scores based on the two MWE lists. The average coefficient was 0.92, which indicates a strong correlation between the two resulting LTQ scores, and thus justifies using PMI MWEs for all other languages.

3.4 ESD and Cultural Grouping

We investigate what is carried by ESD by visualizing and looking at the nearest neighbors of emotion vectors.[5] Jackson et al. (2019) used word colexification patterns to reveal that the same emotion concepts cluster with different emotions according to the language family they belong to. For instance, in Tai-Kadai languages, *hope* appears in the same cluster as *want* and *pity*, while *hope* associates with *good* and *love* in the Nakh-Daghestanian language family. Our results derived from ESD do not rely on colexification patterns, but also support this finding. The nearest neighbors of the Chinese word for *hope* was *want* and *pity*, while they were found as *love* and *joy* for *hope* in Arabic.

In Figure 2, we compare ESD to the syntactic distance between languages by constructing two networks of languages based on each feature. Figure 2a uses ESD as reference while Figure 2b uses the syntactic distance from the URIEL database (Littell et al., 2017). Each node represents a language, color-coded by its cultural area. For each language, we sort the other languages according to the distance value. When a language is in the list of top-k closest languages, we draw an edge between the two. We set $k = 2$.

We see that languages in the same cultural areas tend to form more cohesive clusters in Figure 2a compared to Figure 2b. The portion of edges *within* the cultural areas is 76% for ESD while it is 59% for syntactic distance. These results indicate that ESD effectively extracts linguistic information that aligns well with the commonly shared perception of cultural areas.

3.5 Correlation with Geographical Distance

Regarding the language clusters in Figure 2a, some may suspect that geographic distance can substitute the pragmatically-inspired features. For Chinese, Korean and Japanese are the closest languages by ESD, which can also be explained by their geographical proximity. Do our features add additional pragmatic information, or can they simply be replaced by geographical distance?

To verify this speculation, we evaluate Pearson's correlation coefficient of each pragmatic feature

[5] A visualization demo of emotion vectors can be found at https://bit.ly/emotion_vecs.

value with geographical distance from URIEL. The feature with the strongest correlation was `ESD` (r=0.4). The least correlated was `LCR-verb` (r=0.03), followed by `LCR-pron` (r=0.17) and `LTQ` (r=−0.31)[6]. The results suggest that the pragmatic features contain extra information that cannot be subsumed by geographic distance.

4 Extrinsic Evaluation: Ranking Transfer Languages

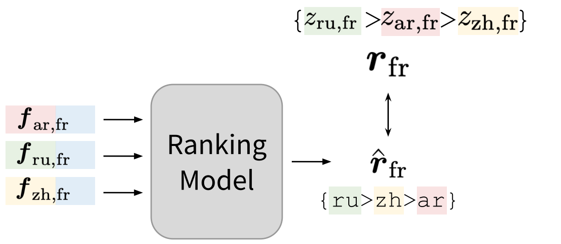
To demonstrate the utility of our features, we apply them to a *transfer language ranking* task for cross-lingual transfer learning. We first present the overall task setting, including the datasets and models used for the two cross-lingual tasks. Next, we describe the transfer language ranking model and its evaluation metrics.

4.1 Task Setting

We define our task as the *language ranking* problem: given the target language l_{tg}, we want to rank a set of n candidate transfer languages $\mathcal{L}_{tf}=\{l_{tf}^{(1)},\ldots,l_{tf}^{(n)}\}$ by their usefulness when transferred to l_{tg}, which we refer to as *transferability* (illustrated in Figure 3). The effectiveness of cross-lingual transfer is often measured by evaluating the joint training or zero-shot transfer performance (Wu and Dredze, 2019; Schuster et al., 2019). In this work, we quantify the effectiveness as the zero-shot transfer performance, following Lin et al. (2019). Our goal is to train a model that ranks available transfer languages in $\mathcal{L}_{tf}$ by their transferability for a target language l_{tg}.

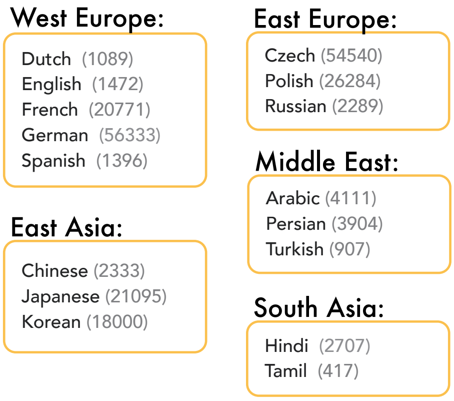
To train the ranking model, we first need to find the ground-truth transferability rankings, which operate as the model's training data. We evaluate the zero-shot performance $z_{tf,tg}$ by training a task-specific cross-lingual model solely with transfer language l_{tf} and testing on l_{tg}. After evaluating $z_{tf,tg}$ for each candidate transfer language in $\mathcal{L}_{tf}$, we obtain the optimal ranking of languages $\boldsymbol{r}_{tg}$ by sorting languages according to the measured $z_{tf,tg}$. Note that $\boldsymbol{r}_{tg}$ also depends on downstream task.

Next, we train the language ranking model. The ranking model predicts the transfer ranking of candidate languages. Each source, target pair (l_{tf}, l_{tg}) is represented as a vector of language features $\boldsymbol{f}_{tf,tg}$, which may include phonological similarity, typological similarity, word-overlap to name a few. The ranking model takes $\boldsymbol{f}_{tf,tg}$ of every l_{tf} as input, and predicts the transferability ranking $\hat{\boldsymbol{r}}_{tg}$. Using $\boldsymbol{r}_{tg}$ from the previous step as training data, the model learns to find optimal transfer languages based on $\boldsymbol{f}_{tf,tg}$. The trained model can either be used to select the optimal set of transfer languages, or to decide which language to additionally annotate during the data creation process.



{$z_{\text{ru,fr}}$ >$z_{\text{ar,fr}}$>$z_{\text{zh,fr}}$}
$\boldsymbol{r}_{\text{fr}}$
$\boldsymbol{f}_{\text{ar,fr}}$
$\boldsymbol{f}_{\text{ru,fr}}$
$\boldsymbol{f}_{\text{zh,fr}}$
Ranking Model
$\hat{\boldsymbol{r}}_{\text{fr}}$
{ru>zh>ar}

Figure 3: Illustration of transfer language ranking problem when the target language is French (fr) and there are three available transfer languages: Arabic (ar), Russian (ru), and Chinese (zh). The output ranking $\hat{\boldsymbol{r}}_{\text{fr}}$ is compared to the ground truth ranking $\boldsymbol{r}_{\text{fr}}$ which is determined by the zero-shot performance $\boldsymbol{z}$ of cross-lingual models.



West Europe:

Dutch (1089)
English (1472)
French (20771)
German (56333)
Spanish (1396)

East Asia:

Chinese (2333)
Japanese (21095)
Korean (18000)

East Europe:

Czech (54540)
Polish (26284)
Russian (2289)

Middle East:

Arabic (4111)
Persian (3904)
Turkish (907)

South Asia:

Hindi (2707)
Tamil (417)

Figure 4: Languages used throughout the experiments are grouped by their cultural areas (Siegel, 1977). The numbers indicate the size of each dataset.

4.2 Task & Dataset

We apply the proposed features to train a ranking model for two distinctive tasks: multilingual sentiment analysis (SA) and multilingual dependency parsing (DEP). The tasks are chosen based on our hypothesis that high-order information such as pragmatics would assist sentiment analysis while it may be less significant for dependency parsing, where lower-order information such as syntax is relatively stressed.

SA As there is no single sentiment analysis dataset covering a wide variety of languages, we

[6] When two languages are more similar, LTQ is higher whereas geographic distance is smaller.

collected various review datasets from different sources.[7] All samples are labeled as either positive or negative. In case of datasets rated with a five-point Likert scale, we mapped 1–2 to negative and 4–5 to positive. We settled on a dataset consist of 16 languages categorized into five distinct cultural groups: West Europe, East Europe, East Asia, South Asia, and Middle East (Figure 4).

DEP To compare the effectiveness of the proposed features on syntax-focused tasks, we chose datasets of the same set of 16 languages from Universal Dependencies v2.2 (Nivre et al., 2018).

4.3 Task-Specific Cross-Lingual Models

SA Multilingual BERT (mBERT) (Devlin et al., 2019), a multilingual extension of BERT pretrained with 104 different languages, has shown strong results in various text classification tasks in cross-lingual settings (Sun et al., 2019; Xu et al., 2019; Li et al., 2019). We use mBERT to conduct zero-shot cross-lingual transfer and to extract optimal transfer language rankings: fine-tune mBERT on transfer language data and test it on target language data. The performance is measured by the macro F1 score on the test set.

DEP We adopt the setting from Ahmad et al. (2018) to perform cross-lingual zero-shot transfer. We train deep biaffine attentional graph-based models (Dozat and Manning, 2016) which achieved state-of-the-art performance in dependency parsing for many languages. The performance is evaluated using labeled attachment scores (LAS).

4.4 Ranking Model & Evaluation

Ranking Model For the language ranking model, we employ gradient boosted decision trees, LightGBM (Ke et al., 2017), which is one of the state-of-the-art models for ranking tasks.[8]

Ranking Evaluation Metric We evaluate the ranking models' performance with two standard metrics for ranking tasks: Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain at position p (NDCG@p) (Järvelin and Kekäläinen, 2002). While MAP assumes a binary concept of relevance, NDCG is a more fine-grained measure that reflects the ranking positions. The *relevant* languages for computing MAP are defined as the top-k languages in terms of zero-shot performance in the downstream task. In our experiments, we set k to 3 for MAP. Similarly, we use NDCG@3.

We train and evaluate the model using leave-one-out cross-validation: where one language is set aside as the test language while other languages are used to train the ranking model. Among the training languages, each language is posited in turn as the *target* language while others are the *transfer* languages.

5 Experiments

5.1 Baselines

LangRank LangRank (Lin et al., 2019) uses 13 features to train the ranking model: The dataset size in transfer language (`tf_size`), target language (`tg_size`), and the ratio between the two (`ratio_size`); Type-token-ratio (`ttr`) which measures lexical diversity and `word_overlap` for lexical similarity between a pair of languages; various distances between a language pair from the URIEL database (geographic `geo`, genetic `gen`, inventory `inv`, syntactic `syn`, phonological `phon` and featural `feat`).

MTVec Malaviya et al. (2017) proposed to learn a language representation while training a neural machine translation (NMT) system in a simliar fashion to Johnson et al. (2017). During training, a language token is prepended to the source sentence and the learned token's embedding becomes the language vector. Bjerva et al. (2019) has shown that such language representations contain various types of linguistic information ranging from word order to typological information. We used the one released by Malaviya et al. (2017) which has the dimension of 512.

5.2 Individual Feature Contribution

We first look into whether the proposed features are helpful in ranking transfer languages for sentiment analysis and dependency parsing (Table 1). We add all three features (Prag) to the two baseline features (LangRank, MTVec) and compare the performance in the two tasks. Results show that our features improve both baselines in SA, implying that the pragmatic information captured by our features is helpful for discerning the subtle differences in sentiment among languages.

In the case of DEP, including our features brings inconsistent results to performance. The features

[7] Details are provided in Appendix A. Note that the difference in domain and label distribution of data can also affect the transferability, and a related discussion is in §5.4

[8] More details on the cross-lingual models, ranking model, and their training can be found in Appendix B.

	SA		DEP	
	MAP	NDCG	MAP	NDCG
LangRank	71.3	86.5	**63.0**	**82.2**
LangRank+Prag	**76.0**	**90.9**	61.7	80.5
- `LCR`	75.0	88.3	60.3	79.6
- `LTQ`	72.4	89.3	63.1*	81.3*
- `ESD`	77.7*	92.1*	58.2	78.5
MTVec	71.1	89.5	43.0	69.7
MTVec+Prag	**74.3**	**90.8**	**49.7**	**74.8**
- `LCR`	72.9	90.1	54.1*	76.3*
- `LTQ`	71.2	89.0	53.0*	78.6*
- `ESD`	73.1	90.7	45.3	73.9

Table 1: Evaluation results of our features (Prag) added to each baseline. The higher scores are **boldfaced**. Rows in gray indicate ablation studies. * is marked when improvements are made compared to LangRank+Prag, MTVec+Prag respectively.

	SA		DEP	
	MAP	NDCG	MAP	NDCG
Pretrain-specific	39.0	55.5	-	-
Data-specific	68.0	85.4	37.2	55.0
Typology	44.9	60.7	**58.0**	**79.8**
Geography	24.9	55.0	32.3	65.1
Orthography	34.2	56.6	35.5	60.5
Pragmatic	**73.0**	**88.0**	46.5	71.8

Table 2: Ranking performance using each feature group as input to the ranking model.

help the performance of MTVec while they deteriorate the performance of LangRank. Although some performance increase was observed when applied to MTVec, the performance of MTVec in DEP remains extremely poor. These conflicting trends suggest that pragmatic information is not crucial to less pragmatically-driven tasks, represented as dependency parsing in our case.

The low performance of MTVec in DEP is noticeable as MTVec is generally believed to contain a significant amount of syntactic information, with much higher dimensionality than LangRank. It also suggests the limitation of using distributional representations as language features; their lack of interpretability makes it difficult to control the kinds of information used in a model.

We additionally conduct ablation studies by removing each feature from the +Prag models to examine each feature's contribution. The SA results show that `LCR` and `LTQ` significantly contribute to overall improvements achieved by adding our features, while `ESD` turns out to be less helpful. Sometimes, removing `ESD` resulted in a better performance. In contrast, the results of DEP show that `ESD` consistently made a significant contribution, and `LCR` and `LTQ` were not useful. The results imply that the emotion semantics information of languages is surprisingly not useful in sentiment analysis, but more so in dependency parsing.

5.3 Group-wise Contribution

The previous experiment suggests that the same pragmatic information can be helpful to different extents depending on the downstream task. We further investigate to what extent each kind of information is useful to each task by conducting group-wise comparisons. To this end, we group the features into five categories: Pretrain-specific, Data-specific, Typology, Geography, Orthography, and Pragmatic. Pretrain-specific features cover factors that may be related to the performance of pretrained language models used in our task-specific cross-lingual models. Specifically, we used the size of the Wikipedia training corpus of each language used in training mBERT.[9] Note that we do not measure this feature group's performance on DEP as no pretrained language model was used in DEP. Data-specific features include `tf_size`, `tg_size`, and `ratio_size`. Typological features include `geo`, `syn`, `feat`, `phon`, and `inv` distances. Geography includes `geo` distance in isolation. Orthographic feature is the `word_overlap` between languages. Finally, the Pragmatic group consists of `ttr` and the three proposed features, `LCR`, `LTQ`, and `ESD`. `ttr` is included in Pragmatic as Richards (1987) have suggested that it encodes a significant amount of cultural information.

Table 2 reports the performance of ranking models trained with the respective feature category. Interestingly, the two tasks showed significantly different results; the Pragmatic group showed the best performance in SA while the Typology group outperformed all other groups in DEP. This again confirms that the features indicating cross-lingual transferability differ depending on the target task. Although the Pretrain-specific features were more predictive than the Geography and Orthography features it was not as helpful as the Pragmatic features.

[9] https://meta.wikimedia.org/wiki/List_of_Wikipedias

5.4 Controlling for Dataset Size

The performance of cross-lingual transfer depends not only on the cultural similarity between transfer and target languages but also on other factors, including dataset size and label distributions. Although our model already accounts for the dataset size to some extent by including `tf_size` as input, we conduct a more rigorous experiment to better understand the importance of cultural similarity in language selection. Specifically, we control the data size by down-sampling all SA data to match both the size and label distribution of the second smallest Turkish dataset.[10] We then trained two ranking models equipped with different sets of features: LANGRANK and LANGRANK+PRAG.

In terms of languages, we focus on a setting where Turkish is the target and Arabic, Japanese and Korean are the transfer languages. This is a particularly interesting set of languages because the source languages are similar/dissimilar to Turkish in different aspects; Korean and Japanese are typologically similar to Turkish, yet in cultural terms, Arabic is more similar to Turkish.

In this controlled setting, the ground-truth ranking reveals that the optimal transfer language among the three is Arabic, followed by Korean and Japanese. It indicates the important role of cultural resemblance in sentiment analysis which encapsulates the rich historical relationship shared between Arabic- and Turkish-speaking communities. LANGRANK+PRAG chose Arabic as the best transfer language, suggesting that the imposed cultural similarity information from the features helped the ranking model learn the cultural tie between the two languages. On the other hand, LANGRANK ranked Japanese the highest over Arabic, possibly because the provided features mainly focus on typological similarity over cultural similarity.

6 Related Work

Quantifying Cross-cultural Similarity A few recent work in psycholinguistics and NLP have aimed to measure cultural differences, mainly from word-level semantics. Lin et al. (2018) suggested a cross-lingual word alignment method that preserves the cultural, social context of words. They derive cross-cultural similarity from the embeddings of a bilingual lexicon in the shared representation space. Thompson et al. (2018) computed similarity by comparing the nearest neighborhood of words in different languages, showing that words in some domains (e.g., time, quantity) exhibit higher cross-lingual alignment than other domains (e.g., politics, food, emotions). Jackson et al. (2019) represented each language as a network of emotion concepts derived from their colexification patterns and measured the similarity between networks.

Auxiliary Language Selection in Cross-lingual tasks There has been active work on leveraging multiple languages to improve cross-lingual systems (Neubig and Hu, 2018; Ammar et al., 2016). Adapting auxiliary language datasets to the target language task can be practiced through either language-selection or data-selection. Previous work on language-selection mostly relied on leveraging syntactic or semantic resemblance between languages (e.g. ngram overlap) to choose the best transfer languages (Zoph et al., 2016; Wang and Neubig, 2019). Our approach extends this line of work by leveraging cross-cultural pragmatics, an aspect that has been unexplored by prior work.

7 Future Directions

Typology of Cross-cultural Pragmatics The features proposed here provide three dimensions in a provisional quantitative cross-linguistic typology of pragmatics in language. Having been validated, both intrinsically and extrinsically, they can be used in studies as a stand-in for cross-cultural similarity. They also open a new avenue of research, raising questions about what other quantitative features of language are correlates of cultural and pragmatic difference.

Model Probing Fine-tuning pretrained models to downstream tasks has become the de facto standard in various NLP tasks, and their success has promoted the development of their multilingual extensions (Devlin et al., 2019; Lample and Conneau, 2019). While the performance gains from these models are undeniable, their learning dynamics remain obscure. This issue has prompted various probing methods designed to test what kind of linguistic information the models retain, including syntactic and semantic knowledge (Conneau et al., 2018; Liu et al., 2019; Ravishankar et al., 2019; Tenney et al., 2019). Similarly, our features can be employed as a touchstone to evaluate a model's knowledge in cross-cultural pragmatics. Investigating how different pretraining tasks affect the

[10] The size of the smallest language (Tamil; 417 samples) was too small to train an effective model.

learning of pragmatic knowledge will also be an interesting direction of research.

8 Conclusion

In this work, we propose three pragmatically-inspired features that capture cross-cultural similarities that arise as linguistic patterns: language context-level ratio, literal translation quality, and emotion semantic distance. Through feature analyses, we examine whether our features can operate as valid proxies of cross-cultural similarity. From a practical standpoint, the experimental results show that our features can help select the best transfer language for cross-lingual transfer in pragmatically-driven tasks, such as sentiment analysis.

Acknowledgements

The authors are grateful to the anonymous reviewers for their invaluable feedback. This material is based upon work supported by the National Science Foundation under Grant No. IIS2007960. We would also like to thank Amazon for providing AWS credits.

References

Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard H. Hovy, Kai-Wei Chang, and Nanyun Peng. 2018. Near or far, wide range zero-shot cross-lingual dependency parsing. *CoRR*, abs/1811.00570.

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

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. What do language representations really represent? *Computational Linguistics*, 45(2):381–389.

Christopher J. Burges, Robert Ragno, and Quoc V. Le. 2007. Learning to rank with nonsmooth cost functions. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 193–200. MIT Press.

Rafael Monroy Casas and JM Hernández Campoy. 1995. A sociolinguistic approach to the study of idioms: Some anthropolinguistic sketches. *Cuadernos de Filología inglesa*, 4.

Yo Joong Choe, Kyubyong Park, and Dongwoo Kim. 2020. word2word: A collection of bilingual lexicons for 3,564 language pairs. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Béatrice Daille. 1994. *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques*. Ph.D. thesis, Ph. D. thesis, Université Paris 7.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Timothy Dozat and Christopher D. Manning. 2016. Deep biaffine attention for neural dependency parsing. *CoRR*, abs/1611.01734.

Aparna Garimella, Rada Mihalcea, and James Pennebaker. 2016. Identifying cross-cultural differences in word usage. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 674–683, Osaka, Japan. The COLING 2016 Organizing Committee.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Edward Twitchell Hall. 1989. *Beyond culture*. Anchor.

Geert H Hofstede, Gert Jan Hofstede, and Michael Minkov. 2005. *Cultures and organizations: Software of the mind*, volume 2. Mcgraw-hill New York.

Joshua Conrad Jackson, Joseph Watts, Teague R. Henry, Johann-Mattis List, Robert Forkel, Peter J. Mucha, Simon J. Greenhill, Russell D. Gray, and Kristen A. Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.

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 V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, 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.

Jihen Karoui, Farah Benamara, Véronique Moriceau, Viviana Patti, Cristina Bosco, and Nathalie Aussenac-Gilles. 2017. Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 262–272, Valencia, Spain. Association for Computational Linguistics.

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.

Zoltán Kövecses. 2003. Language, figurative thought, and cross-cultural comparison. *Metaphor and Symbol*, 18(4):311–320.

Zoltán Kövecses. 2010. *Metaphor: A practical introduction*. Oxford University Press.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 34–41, Hong Kong, China. Association for Computational Linguistics.

Bill Yuchen Lin, Frank F. Xu, Kenny Zhu, and Seung-won Hwang. 2018. Mining cross-cultural differences and similarities in social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 709–719, Melbourne, Australia. Association for Computational Linguistics.

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, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

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*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535, Copenhagen, Denmark. Association for Computational Linguistics.

Korac-Kakabadse Nada, Kouzmin Alexander, Korac-Kakabadse Andrew, and Savery Lawson. 2001. Low-and high-context communication patterns: towards mapping cross-cultural encounters. *Cross Cultural Management: An International Journal*, 8(2):3–24.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

Dat Quoc Nguyen, Dai Quoc Nguyen, Dang Duc Pham, and Son Bao Pham. 2014. RDRPOSTagger: A ripple down rules-based part-of-speech tagger. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 17–20, Gothenburg, Sweden. Association for Computational Linguistics.

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

Ulrike Oster. 2019. Cross-cultural semantic and pragmatlc profiling of emotion words. regulation and expression of anger in Spanish and German. *Current Approaches to Metaphor Analysis in Discourse*, 39:35.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Vinit Ravishankar, Memduh Gökırmak, Lilja Øvrelid, and Erik Velldal. 2019. Multilingual probing of deep pre-trained contextual encoders. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 37–47, Turku, Finland. Linköping University Electronic Press.

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

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

Bernard J. Siegel. 1977. Encyclopedia of anthropology. David E. Hunter and Phillip Whitten, eds. New York. *American Anthropologist*, 79(2):452–454.

Laura Smith, Salvatore Giorgi, Rishi Solanki, Johannes Eichstaedt, H Andrew Schwartz, Muhammad Abdul-Mageed, Anneke Buffone, and Lyle Ungar. 2016. Does 'well-being'translate on twitter? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2042–2047.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics*, pages 194–206, Cham. Springer International Publishing.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

B. Thompson, S. Roberts, and G. Lupyan. 2018. Quantifying semantic similarity across languages. *Proceedings of the 40th Annual Conference of the Cognitive Science Society (CogSci 2018)*.

Yulia Tsvetkov and Shuly Wintner. 2010. Extraction of multi-word expressions from small parallel corpora. In *Coling 2010: Posters*, pages 1256–1264, Beijing, China. Coling 2010 Organizing Committee.

Jelena Vulanović. 2014. Cultural markedness and strategies for translating idiomatic expressions in the epic poem "The Mountain Wreath" into English. *Mediterranean Journal of Social Sciences*, 5(13):210.

Xinyi Wang and Graham Neubig. 2019. Target conditioned sampling: Optimizing data selection for multilingual neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5823–5828, Florence, Italy. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

A Dataset for Sentiment Analysis

Dataset	Languages	Domain	Size	POS/NEG
SemEval-2016 Aspect Based Sentiment Analysis	Chinese	electronics	2333	1.53
	Arabic	hotel	4111	1.54
	English	restaurant	1472	2.14
	Dutch	restaurant	1089	1.43
	Spanish	restaurant	1396	2.82
	Russian	restaurant	2289	3.81
	Turkish	restaurant	907	1.32
SentiPers	Persian	product	3904	1.8
Amazon Customer Reviews	French	product	20771	8.0
	German	product	56333	6.56
	Japanese	product	21095	8.05
CSFD CZ	Czech	movie	54540	1.04
Naver Sentiment Movie Corpus	Korean	movie	18000	1.0
Tamil Movie Review Dataset	Tamil	movie	417	0.48
PolEval 2017	Polish	product	26284	1.38
Aspect based Sentiment Analysis	Hindi	product	2707	3.22

Table 3: Datasets for sentiment analysis.

B Task-Specific Models Details

SA Cross-lingual Model We performed supervised fine-tuning of multilingual BERT (mBERT) (Devlin et al., 2019) for the sentiment analysis task, as the model showed strong results in various text classification tasks in cross-lingual settings (Sun et al., 2019; Xu et al., 2019; Li et al., 2019). mBERT is pretrained with 104 different languages, including the 16 languages we used throughout our experiment. We used a concatenation of mean and max pooled representations from mBERT's penultimate layer, as it outperformed the standard practice of using the last layer's `[CLS]` token. The representation was passed to a fully connected layer for prediction. To extract optimal transfer rankings, we conducted zero-shot transfer with mBERT: fine-tuned mBERT on transfer language data and tested it on target language data.

Ranking Model We used LightGBM (Ke et al., 2017) with LambdaRank (Burges et al., 2007) algorithm. The model consists of 100 decision trees with 16 leaves each, and it was trained with the learning rate of 0.1. We optimized NDCG to train the model (Järvelin and Kekäläinen, 2002).