Abstract

There is an increasing amount of evidence that in cases with little or no data in a target language, training on a different language can yield surprisingly good results. However, currently there are no established guidelines for choosing the training (source) language. In an attempt to solve this issue we thoroughly analyze a state-of-the-art multilingual model and try to determine what impacts good transfer between languages. As opposed to the majority of multilingual NLP literature, we don’t only train on English, but on a group of almost 30 languages. We show that looking at particular syntactic features is 2-4 times more helpful in predicting the performance than an aggregated syntactic similarity. We find out that the importance of syntactic features strongly differs depending on the downstream task - no single feature is a good performance predictor for all NLP tasks. As a result, one should not expect that for a target language \( L_1 \) there is a single language \( L_2 \) that is the best choice for any NLP task. Currently, such a decision is based on intuition (either linguistic or researcher experience). This paper aims to provide an empirical framework on how to tackle such a scenario.

1 Introduction

A vast majority of currently available NLP datasets is in English. However, in many applications around the world it is crucial to use the local language. At the same time, creating new datasets requires a lot of resources and time. This is why recently multilingual language models are gaining more and more attention in academia and industry. Using them allows to train well-performing models with little or no data in the target language (usually low-resourced) given a corresponding dataset in the source language (usually high-resourced). This technique is called zero-shot learning.

We explore how differences in linguistic features correlate with or potentially affect performance on various kinds of downstream tasks. To present how such analysis can be helpful, let us consider the following scenario: We want to train a model in a task \( T \) in language \( L_1 \), but we do not have access to a dataset for \( T \) in \( L_1 \). However, we have \( n \) datasets for \( T \) in languages \( L_2, L_3, \ldots, L_{n+1} \). We want to choose one of these languages that has the best ‘transferability’ to \( L_1 \) for this specific task \( T \). Currently, such a decision is based on intuition (either linguistic or researcher experience). This paper aims to provide an empirical framework on how to tackle such a scenario.

Over the past years several models for zero-shot learning were introduced such as LASER (Artetxe and Schwenk, 2019), MultiFit (Eisenschlos et al., 2019) or XLM-R (Conneau et al., 2019). For our analysis, we decided to use the XLM-R model as it achieves the best performance. To facilitate our experiments we take advantage of XTREME (Hu et al., 2020) which is a recently published benchmark for multilingual models covering 40 languages on 8 diverse NLP tasks. However, it allows only training on English. To boost research in multilingual direction, we extend this benchmark to make it suitable for training in languages other than English and make the code publicly available\(^1\). Additionally, we publish results of XLM-R Base model for 3 out of 8 tasks, for all available languages, to encourage further analysis of the nature of interactions between languages.

To conclude, this work aims to examine the connection between linguistic features and the quality of transfer between languages on a variety of downstream tasks.

The following research questions are posed:

- Does the inclusion of fine-grained linguistic

\(^1\)https://github.com/blazejdolicki/multilingual-analysis
features improve prediction of language transfer compared to an aggregated syntactic distance between languages?

- Which linguistic features have the largest impact on ‘transferability’ and performance on 3 different tasks?

- Can we identify some general linguistic features that determine good transfer for many tasks or is the feature importance task-specific?

2 Related work

2.1 Zero-shot learning

Multilingual language models during training create a common vector space for all languages used. This makes them especially relevant for zero-shot learning. The goal is to achieve similar vectors for sentences or words in different languages which are semantically equivalent (i.e. which are translations of each other).

One of the first multilingual models that gained traction in academic circles was m-BERT (Devlin et al., 2018). It provided a shared vocabulary for all languages and was trained on a concatenation of corpora in about 100 different languages. It used a standard language model objective, not designed for cross-lingual purposes. Even though, it achieved surprising results which inspired many scientists to investigate this direction. Artetxe and Schwenk (2019) published a multilingual language model (henceforth LASER) which outperformed m-BERT (Devlin et al., 2018). The model was trained on as many as 93 languages using a machine translation objective with BiLSTMs. Another interesting model (Eisenschlos et al., 2019)(henceforth MultiFit) is monolingual, but thanks to a bootstrapping method (Ruder and Plank, 2018), it works well with zero-shot learning. Namely, it employs LASER to predict so called pseudo-labels of the dataset in the target language and by using them in training it actually surpasses LASER performance.

The most recent model was proposed by Conneau et al. (2019), dubbed XLM-R. It is based on the RoBERTa architecture (Liu et al., 2019). XLM-R has a translation language modeling (TLM) objective (Conneau et al., 2019) in which parallel sentences are fed together as examples and by masking some words in one language the model indirectly uses the translation from the other language. It is the first multilingual model that matches performance of strong monolingual models (Conneau et al., 2019). For our experiments we used XLM-R with a newly published cross-lingual benchmark called XTREME (Hu et al., 2020) which allows to easily evaluate XLM-R on up to 40 languages and 8 diverse NLP tasks.

2.2 What determines good transfer performance?

Lin et al. (2019) created an automatic approach to predict transfer performance in zero-shot learning setups. They examined the impact of a wide variety of factors, ranging from word overlap or type-token ratio in the training set to geographic, genetic or syntactic distance between languages. They evaluated their approach on four tasks: machine translation, entity linking, part of speech tagging and dependency parsing and for each they used a task-specific architecture without incorporating any pretrained embeddings. Two other papers performed various experiments connected to a similar topic on m-BERT (Devlin et al., 2018) using this single model on different tasks. Pires et al. (2019) focused on named entity recognition (NER) and part-of-speech tagging (POS) while Tran and Bisazza (2019) covered dependency parsing. Below we discuss their findings grouped by different factors.

2.2.1 Word overlap

Results of Lin et al. (2019) show that word overlap seems to be one of the best features which suggests that it is related to transfer - high word overlap between the source and the target language leads to good performance. On the other hand, Pires et al. (2019) showed that there is little correlation between word overlap and performance of m-BERT (however this relation was strongly linear for the original version of BERT that was only trained on English). They used NER because they expected “zero-shot performance on NER to be highly dependent on word piece overlap”(Pires et al., 2019). These results are confirmed by Tran and Bisazza (2019) on dependency parsing.

2.2.2 Dataset size

Lin et al. (2019) uses as one of the features the ratio between training size of the source and training languages. It seems to have very little correlation with the performance for all evaluated tasks. In our view, the size of the training set is a much more
important factor than the ratio. Let us consider if training sets of both source and target languages are big and of similar size, then the ratio is large and the performance will probably be high. If we take two small languages, then the ratio is also large, but the performance will probably be low. This clearly suggests that this measure will not be strongly correlated with the performance, without doing any experiments. Hu et al. (2020) calculated Pearson correlation between Wikipedia size (which is not used for pretraining XLM-R) and performance on all their tasks and the outcome is very interesting. For both text classification datasets (XNLI (Conneau et al., 2018) and PAWSX (Yang et al., 2019)) the correlation was very strong - about 0.80, while for POS (Nivre et al., 2018) and NER (Pan et al., 2017) it was as little as 0.30. This indicates that the significance of dataset size varies significantly between tasks.

2.2.3 Syntactic features

There is a lot of evidence that syntax helps in NLP tasks such as POS tagging or dependency parsing (Ammar et al., 2016; Naseem et al., 2012; Zhang and Barzilay, 2015; Östling, 2015). However, it is usually explicitly incorporated as features which is not the case for models analyzed by (Lin et al., 2019; Pires et al., 2019; Tran and Bisazza, 2019).

WALS (World Atlas of Language Structures) (Dryer and Haspelmath, 2013) is a very useful database containing 192 linguistic features for 2662 different languages which was often used in previous literature. Pires et al. (2019) grouped languages with regard to 6 WALS features regarding word order and concluded that the more features the source and target languages have in common, the better the performance. Tran and Bisazza (2019) showed that in their experiments syntactic distance is not correlated with performance. Syntactic distance is computed as follows: for each language in the source-target language pair a one-hot encoded vector with WALS syntactic features is retrieved and then a distance between these two vectors is computed. In experiments performed by Lin et al. (2019), syntactic distance was the 4th (out of 10) most important feature after word overlap, genetic and geographical distance respectively.

When using syntactic distance, all WALS features contribute equally to the score, however in reality particular features have different importance. Previously, most often only the features related to word order were used (O’Horan et al., 2016) which suggests that these features are more important than others and shouldn’t be treated with equal weights. That is why in this work we examine separate features and how using them compares to an aggregated distance in predicting the transfer quality. Moreover, to verify if the significance of different factors varies among downstream task we cover POS tagging, named entity recognition and text classification. Furthermore, as opposed to Lin et al. (2019), we use a single, state-of-the-art language model (XLM-R) for all downstream tasks which decreases the chance that the differences in performance are due to differences in model architecture rather than tasks.

3 Datasets

3.1 Universal Dependencies 2.0 (POS tagging)

We use Universal Dependencies (Nivre et al., 2018) for part of speech tagging on 29 languages. There are 17 tags (categories) universal for all languages.\(^2\)

3.2 Panx (NER)

For named entity recognition, we use Wikiann (Pan et al., 2017), more precisely the balanced sample made by (Rahimi et al., 2019). There are three types of entities: Location, Person and Organization. The entities were automatically labeled from Wikipedia.

3.3 XNLI (NLI)

This task (Conneau et al., 2018) concerns natural language inference (NLI), i.e. for each example there is a pair of sentence and there are three possible labels describing the relationship of this pair: contradiction, entailment or neutral. It is a special case of text classification, yet it is substantially more difficult than standard tasks from this group. For instance, while for sentiment analysis often one strongly polarized word is enough to predict the correct class, here the distinction between two sentences in a pair is much more subtle. The dataset is available in 15 languages. Due to its large size and limited resources we downsample randomly each training set from almost 400k examples to 100k (since the classes are balanced in the original dataset, after random sampling they remain balanced).

\(^2\)To learn more about UD tags and what each abbreviation means visit: https://universaldependencies.org/u/pos/
4 Analysis

We use the XTREME framework to evaluate 3 standard NLP downstream tasks on XLM-R for all available languages (and trying all source-target pairs possible). As we used 28 languages for POS tagging, 37 languages for NER and 15 for XNLI showing results for all pairs wouldn’t be feasible, therefore in the paper we decided to show for each test language its supervised result compared with top 3 source languages in Table 1, Table 2 and Table 3. For abbreviation, we use ISO 693-1 language codes.

4.1 Results on downstream tasks

For POS tagging (Table 1), some outcomes are in line with the expectations (e.g. Portuguese (pt) with the top-3 languages as Spanish (es), Italian (it) and French (fr)) while others are not. All languages apart from Marathi (mr) obtain better results in supervised setting than zero-shot. Its best source language is German (de) which is quite surprising as Marathi is part of the Indic language family represented in the dataset by two other languages - Urdu (ur) and Hindi (hi). French and Italian obtain better results on Bulgarian (a Slavic language) than the other Slavic language in the dataset, Russian. Other interesting pairs include German being the best language for Greek (el), Urdu the best for Basque (eu) - although not that surprising keeping in mind that Basque is a language isolate, Dutch (nl) the best for Hungarian (another language isolate), Basque the best for Korean (ko) and Estonian (et) best for Tamil (ta, a Dravidian language spoken in India and Sri Lanka). Interestingly, some of the results seem related to colonialism - for instance English being the best language for Indonesian (id) and French being the best for Vietnamese (vi).

While for POS tagging many outcomes were clear and we only pointed out the outliers, it is much more difficult to find patterns for NER (Table 2). Hungarian is the third best for Afrikaan and German, this time Italian is the best for Greek, Indonesian for English, French for Farsi (fa) and Italian (instead of a Germanic language) for Dutch.

For XNLI (Table 3) as many as 5 out of 15 languages obtain worse results in supervised setting than zero-shot: Bulgarian, Spanish, Hindi, Russian and Urdu. We can also notice that XNLI is more challenging than POS or NER - while for the
first two tasks accuracy is generally above 0.80, for XNLI, none of the language pairs exceeds that number.

After some initial, qualitative remarks, we follow up with more quantitative analysis. To investigate whether the downstream task is important in choosing the source language we inspect Table 5 which shows that accuracy between NER and POS tasks is much stronger than between those two and XNLI. This means that there is no guarantee that if a source language achieves good performance on a target language for one NLP task, it will also perform well on a different task.

Table 4 presents how performance on downstream tasks correlates with three most important linguistic distance measures based on (Lin et al., 2019): syntactic (syn), geographic (geo) and genetic (gen) distances. As mentioned in previous sections, syntactic distance is the aggregate of a group of linguistic features, geographic distance is the distance between geographic locations where two languages are natively spoken while genetic distance expresses how related the languages are - French and Spanish are closely related as they both belong to the Romance language family while French and Chinese are from different language families, thus their genetic distance is larger. We can clearly see that POS is strongly correlated with all three distances. XNLI correlates the most with geographical distance and for NER all correlations are relatively weak. In the following section we focus on particular linguistic features and examine their importance to explore their effect on transfer between languages.

Table 2: Top 3 languages (train) for each target language (test) together with its supervised result on NER (accuracy)
| test  | #   | train | acc  | test   | #   | train   | acc  | test   | #   | train   | acc  | test   | #   | train   | acc  |
|-------|-----|-------|------|--------|-----|---------|------|--------|-----|---------|------|--------|-----|---------|------|
| ar    | 1   | bg    | 0.713| en     | sup | en      | 0.815| ru     | sup | ru      | 0.739| ur     | sup | ur      | 0.624|
| ar    | 2   | el    | 0.731| en     | 2   | fr      | 0.806| ru     | 2   | bg      | 0.738| ur     | 2   | ru      | 0.677|
| ar    | 3   | ru    | 0.711| en     | 3   | bg      | 0.804| ru     | 3   | es      | 0.736| ur     | 3   | th      | 0.673|
| bg    | sup | bg    | 0.747| es    | sup | es      | 0.763| sw    | sup | sw      | 0.656| vi    | sup | vi      | 0.74  |
| bg    | 1   | th    | 0.759| es    | 1   | ru      | 0.870| sw    | 1   | de      | 0.645| vi    | 1   | bg      | 0.739 |
| bg    | 2   | ru    | 0.751| es    | 2   | de      | 0.777| sw    | 2   | el      | 0.643| vi    | 2   | de      | 0.738 |
| bg    | 3   | vi    | 0.748| es    | 3   | zh      | 0.678| sw    | 3   | th      | 0.636| vi    | 3   | es      | 0.737 |
| de    | sup | de    | 0.763| fr    | sup | fr      | 0.761| th    | sup | th      | 0.746| zh    | sup | zh      | 0.749 |
| de    | 1   | fr    | 0.757| fr    | 1   | el      | 0.760| th    | 1   | zh      | 0.730| zh    | 1   | th      | 0.748 |
| de    | 2   | ar    | 0.755| fr    | 2   | de      | 0.759| th    | 2   | bg      | 0.728| zh    | 2   | ru      | 0.744 |
| de    | 3   | tr    | 0.755| fr    | 3   | ru      | 0.759| th    | 3   | ar      | 0.727| zh    | 3   | de      | 0.739 |
| el    | sup | el    | 0.748| hi    | sup | hi      | 0.701| tr    | sup | tr      | 0.735|      |      |         |      |
| el    | 1   | de    | 0.747| hi    | 1   | ru      | 0.707| tr    | 1   | th      | 0.725|      |      |         |      |
| el    | 2   | es    | 0.745| hi    | 2   | th      | 0.705| tr    | 2   | fr      | 0.724|      |      |         |      |
| el    | 3   | fr    | 0.745| hi    | 3   | tr      | 0.705| tr    | 3   | bg      | 0.722|      |      |         |      |

Table 3: Top 3 languages (train) for each target language (test) together with its supervised result on NLI (accuracy)

| task | syn   | geo   | gen   |
|------|-------|-------|-------|
| NER  | -0.183454 | -0.197949 | -0.198537 |
| POS  | -0.292522 | -0.378323 | -0.319140 |
| NLI  | -0.157961 | -0.274467 | -0.100264 |

Table 4: Performance correlation between tasks and distances

|       | NER     | POS     | NLI     |
|-------|---------|---------|---------|
| NER   | 1.000000| 0.495913| 0.179361|
| POS   | 0.495913| 1.000000| 0.065942|
| NLI   | 0.179361| 0.065942| 1.000000|

Table 5: Performance correlation between tasks

4.2 Feature importance for transfer learning

To measure the importance of particular linguistic features (more specifically WALS features) with regard to performance we use two approaches. The first approach uses a statistical test. ANOVA is often used to measure correlation between a categorical independent variable and a continuous dependent variable in many fields, including NLP (Saha and Das Mandal, 2015; Li et al., 2014; Malmasi and Dras, 2014). However, we cannot make assumptions about distribution of our data, so we use Kruskal-Wallis test which is a non-parametric equivalent of ANOVA and tests if the difference in means between groups is significant. Since we are mostly interested in choosing the right source language for a given target language, we examined the correlation between features of the source language and the accuracy. The second approach measures feature importance after training a simple, interpretable machine learning (ML) model. It is very important to distinguish the models used in the Section 4.1 and in this section. In Section 4.1 we used a multilingual language model, XLM-R, to make predictions on 3 NLP downstream tasks: POS tagging, NER and NLI. In this case, we use an interpretable model such as Random Forest on tabular data. A single data point fed into this model consists of the source language, the target language, linguistic features of both languages and the score for one of the downstream tasks predicted by the first model. Therefore, the second model aims to predict the performance of the first model using linguistic features. Although we focus on feature analysis, as a side effect we can use the second model to predict transfer between languages. In Section 4.2.4 we show that using such a model allows better prediction of transfer performance than only using syntactic distance.

Both approaches have pros and cons. The former method only captures simple correlation patterns. The latter method can capture non-linear correlations, but its reliability depends on the quality of the model. Results for both methods are presented in the following sections.

4.2.1 Feature importance with Kruskal-Wallis test

We can see the important features indicated by Kruskal-Wallis test in Table 6. The most important features for POS tagging align with the results obtained with the ML method when using one-hot encoding (discussed in the next subsection). The first feature for NER is also found to be important by the second method. Similarly, XNLI has one overlapping feature between two methods.
4.2.2 Feature importance with ML models

To evaluate feature importance we create a dataset where each row is corresponding to a train-test language pair together with its score and linguistic features of both languages. We train different interpretable models (Linear Regression, Random Forest and XGBoost) that predict accuracy of a given setting and as a result obtain feature importance values that show which language features determine the score in the most accurate way. We deleted rows for which the training and test language is the same because we are focusing on zero-shot learning. From all available WALS features (Dryer and Haspelmath, 2013) we only discarded phonological features as they are irrelevant for text processing together with features that had missing values for all languages. As aforementioned, such a model not only allows us to interpret linguistic features, but also can be used to predict the performance for given source and target languages. Table 7 renders the results. For POS tagging the obtained test RMSE is 0.0423 which can be roughly interpreted that on average our model was off by about 4% of accuracy. For NER it is quite similar - 0.0468. For XNLI, which has a smaller subset of languages and lower variance between scores we achieved RMSE of 0.007936 (with linear regression). We experimented with two approaches for encoding WALS features - ordinal encoding - where we assign an integer to each category and one-hot encoding. Both achieved similar RMSE, however feature importance varies between those two so below we selected features that were at the top for both approaches. Table 8 indicates the top-5 important features per task and encoding. Features that were found as important by both encodings or by both Kruskal-Wallis test and one of the encodings are presented in Table 9. For each feature we present the significant category that tends to yield lower accuracy and languages that fall in that category. This allows us to suggest rules-of-thumb for predicting the performance, for instance: “If the source language has Object-Verb and Relative Clause-Noun orders, it might lead to low accuracy, therefore one should avoid using languages in this group”.

4.2.3 Feature importance by feature group

WALS features are divided in multiple groups: Morphology, Nominal Categories, Nominal Syntax, Verbal Categories, Word Order, Simple Clauses, Complex Sentences, Lexicon, Sign Languages and Others. We separately trained our best model on each group to determine which one is the most salient. For POS tagging and NER, the winning category is Word Order (with RMSE of 0.0401...
Table 9: Selected important features with categories that yield lower performance.

| Task | Feature | Significant category | Languages |
|------|---------|----------------------|-----------|
| POS  | 96A Relationship between the Order of Object and Verb and the Order of Relative Clause and Noun | source language with Object-Verb and Relative Clause-Noun orders | Basque, Korean, Malayalam, Marathi, Burmese, Tamil, Telugu, Turkish |
|      | 93A Interrogative Phrases in Content Questions | source language with not initial Interrogative Phrase | Basque, Persian, Hindi, Hungarian, Japanese, Georgian, Korean, Malayalam, Marathi, Burmese, Swahili, Tamil, Telugu, Thai, Turkish, Urdu, Mandarin |
| NER  | 89A: Order of Numeral and Noun | target language with Noun-Numeral order | Burmese, Swahili, Thai, Yoruba |
|      | 66A The past tense | source language with no present-past tense distinction | Indonesian, Burmese, Thai, Tagalog, Vietnamese, Yoruba |
| XNLI | 45A: Politeness Distinctions in Pronouns | source language with multiple politeness distinctions or pronouns avoided for politeness | Hindi, Urdu, Thai, Vietnamese |

and 0.0487 respectively) which confirms previous literature. For XNLI, Word order is much less significant (only the 4th best result among groups) and the best group is Verbal Categories (RMSE of 0.00772).

4.2.4 Comparing fine-grained linguistic features to URIEL syntactic distance

Furthermore, we move to answer the question whether using fine-grained linguistic features helps more in predicting the zero-shot performance compared to an aggregated syntactic distance used by (Lin et al., 2019). To verify that we use the same syntactic distance computed with the lang2vec library (Littell et al., 2017). As shown in Table 7, in all experiments syntactic distance itself was worse at predicting the zero-shot performance than the separate features (with RMSE equal to 0.10, 0.12 and 0.04 for POS tagging, NER and XNLI respectively). Therefore we conclude that replacing a syntactic distance with fine-grained features in systems such as LangRank (Lin et al., 2019) would most certainly improve predictions.

5 Conclusion

In this paper we explore the effect of linguistic features (for different languages) on zero-shot learning with the goal to identify which language is best to transfer from.

Importance of particular features appears to heavily depend on the downstream task - there is no single WALS feature that would be the most important for multiple tasks. This insight followed by the computed correlations between tasks shows that one should not expect that for a target language \( L_1 \) there is a single language \( L_2 \) that is the best choice for any NLP task.

Most importantly, we prove that using fine-grained linguistic features allows to predict the performance 2-4 times better than an aggregated syntactic distance and confirm that word order is paramount for POS tagging. Therefore in automatic systems like the one proposed by Lin et al. (2019) particular linguistic features should be included to improve performance.

Although this work sheds some light on the unexplored topic of the effect of linguistic features on transfer learning, there is still much work to be done. One of the issues is that apart from POS tagging, NER and dependency parsing there are hardly any tasks with datasets covering more than 15 languages. Using more languages would make the results more reliable. To encourage the community to perform further analyses we publish the predictions and the code that was used for analysis. Moreover, we acknowledge that there might be other, more complex factors than the ones discussed in this paper that provide better explanation to the performance differences between languages. Therefore, as this topic lays in the intersection of natural language processing and linguistics, cooperation between experts from both fields can lead to interesting findings.

References

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.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics, 4:431–444.
Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. *Unsupervised cross-lingual representation learning at scale.*

Shervin Malmasi and Mark Dras. 2014. *From visualisation to hypothesis construction for second language acquisition.* In Proceedings of TextGraphs-9: the workshop on Graph-based Methods for Natural Language Processing, pages 56–64, Doha, Qatar. Association for Computational Linguistics.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. *Selective sharing for multilingual dependency parsing.* In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 629–637, Jeju Island, Korea. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. *BERT: pre-training of deep bidirectional transformers for language understanding.* CoRR, abs/1810.04805.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyeh, and Mohammed Attia. 2018. Universal dependencies 2.2.

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

Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kardas, Sylvain Gugger, and Jeremy Howard. 2019. *Multitf: Efficient multi-lingual language model fine-tuning.* arXiv preprint arXiv:1909.04761.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. *Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.*

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014. *Assessing the discourse factors that influence the quality of machine translation.* In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 283–288, Baltimore, Maryland. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zipri Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, and Xuezhe Ma. 2019. Choosing transfer languages for cross-lingual learning. arXiv preprint arXiv:1905.12688.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. *Roberta: A robustly optimized bert pretraining approach.*
Ke Tran and Arianna Bisazza. 2019. Zero-shot dependency parsing with pre-trained multilingual sentence representations. *arXiv preprint arXiv:1910.05479*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proceedings of EMNLP 2019*, pages 3685–3690.

Yuan Zhang and Regina Barzilay. 2015. Hierarchical low-rank tensors for multilingual transfer parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1857–1867, Lisbon, Portugal. Association for Computational Linguistics.