Selecting Languages for Cross-Lingual Dependency Parsing

Yucen Li
Carnegie Mellon University
yucenl@andrew.cmu.edu

Abstract

Dependency parsing is the crucial foundation for a variety of tasks within the field of NLP such as question answering, machine translation, summarization, etc. The continuing developments in cross-lingual dependency parsing, where models train on input data from a combination of languages, have allowed for many advances within the field. However, most of the current research in cross-lingual dependency parsing is focused on creating the best model given a predefined set of languages. In contrast, this research aims to select the optimal set of languages for models to train on given a specific target language. Due to the consideration of linguistic properties and typological features, models trained on the set of languages determined by the research are often more accurate than the models trained on the original predetermined set.

1 Introduction

The dependency parse of a sentence, detailing the grammatical relationships between words in the sentence, is used for many tasks in NLP relating to semantic understanding of text such as text summarization and machine translation. The dependency parse of a sentence is based on the syntactic dependencies between words, such as the subject and object of a verb. Each sentence contains a root, and each word depends on exactly one other word in the sentence to form a connected parse tree.

Typically, parsers are trained on many sentences with labeled dependency parses, called a treebank, and then used to parse new sentences into the correct dependency tree structure and labels. Cross-lingual dependency parsing is when a parser is trained on treebanks from different languages. This parser can then be used to either parse new sentences from a new language or a language that the parser was originally trained on. This can often be beneficial for low-resource languages which may have very small or nonexistent treebanks, as additional languages can be added to supplement the original treebank for the training dataset. Additionally, if two languages have similar typological properties, cross-lingual dependency parsing may lead to a stronger parser.

Currently, a lot of research has been done regarding the development of new models given a predefined set of languages. A lot of work has been put into developing specific cross-lingual parsers between two similar languages, as many features such as pretrained embeddings are very effective given a predefined set of languages. However, there has been very little research done to investigate which combination of languages are actually the best to train on, as typically similar languages seem to be arbitrarily chosen for the models.

This paper chooses to investigate the treebanks and focus on linguistic properties between languages to select the best language to supplement the treebank of a low-resource language. We selected a diverse group of languages in terms of typological features, geographical proximity, language family, and other features. We used data ablation to simulate English as a low-resource language, and then supplemented the low-resource treebank with a larger treebank from our set of languages. Through our experiments, we were able to determine important properties that should be shared between the languages.

2 Related Work

Many approaches have been taken in the field of NLP for the enhancement for cross-lingual dependency parsing.

2.1 Universal Dependency

While languages can be vastly different, basic syntactic similarities do exist between most languages: nouns typically depend on verbs, adjectives depend on nouns, etc. However, linguistics often construct treebanks by focusing on either open class or closed class items, and the differing treebanks often are not created with the same emphasis on certain relationships. This can make cross-lingual dependency difficult, as the treebanks may have different labels and structures. McDonald et al. (2013) presented a major project to make treebanks more uniform by setting a homogenous set of syntactic labels, dependencies, and rules. These treebanks were labeled as ”universal treebanks,” and their consistencies in formation greatly facilitate cross-lingual training for dependency parsing. There are currently 102 different treebanks in 60 languages which follow the UD guidelines.

2.2 Selective Sharing

Naseem et al. (2012) developed a model for cross-lingual dependency parsing by separating the selection of dependencies with their ordering. Because many languages share similar properties in syntactic structure, the selection component is improved through supervised learning on a large set of languages. This enables the selection to have a lot of training data and be highly accurate. Because the ordering of dependencies is vastly different across different language based on typological properties, the ordering of is only based on languages which are similar and have the same features as that of the target language. This separation of parsing leads to good cross-lingual results. Taking advantage of the similarities between languages while separating the differences is crucial to successful cross-lingual parsing.

2.3 Distributed Representations

Guo et al. (2015) approaches cross-lingual dependency parsing by using distributed feature representations and their corresponding compositions to transfer data from a high-resource language such as English to a low-resource language. These distributed representations of different language components such as morphemes and sentences are discretized in a neural network, allowing the model to fully use the representations. Additionally, these representations of the language are required to be precomputed for each language, so a predetermined set of languages need to be selected.

2.4 MaLOPa

Ammar et al. (2016) developed a cross-lingual dependency parser named MaLOPa by using one parser to training on a collection of languages. The object of the model is to train one parser with a union of treebanks, and then test on a collection of sentences of one language. The model is based on a stack LSTM, which updates three main data structures at each time-step:

- a buffer to read the token sequence,
- a stack containing partial parse treebanks, and
- a list of previous actions

. The arc-standard transition system is used as input for the model. At each step, the parser computes the vector representations for each of the data structures, and then learns the associated weights of the neural net. MaLOPa is trained to maximize the log-likelihood of the correct actions for the parser. When updating parameters, MaLOPa uses mini-batches with one sentence from each treebank until all the sentences in the smallest treebank are used. This prevents one language from dominating the parameters of the neural net.

MaLOPa is trained on the concatenation of the UD treebanks for seven different languages: German, English, Spanish, French, Italian, Portuguese, and Swedish, and the development set contains 300 sentences of the development treebank for each language. The language ID of each of these sentences is also included by appending the language prefix to the beginning of each word in the treebank. For example, “dog” in English is changed to “en:dog” for all words in the English treebank. This allows the parser to identify the language of the sentence and better adjust the parameters of the neural net.

3 Set of Languages

For our experiments, we chose languages from the set of treebanks available in UD 2.0. To chose a di-

Language	Word Order	Gen/Noun	Adj/Noun	Agg.	Region	Family
Chinese	SVO	Gen-N	Adj-N	No	East Asia	Sino-Tibetan
English	SVO	Both	Adj-N	No	Europe	Germanic
Finnish	SVO	Gen-N	Adj-N	Yes	Europe	Uralic
Hebrew	SVO	N-Gen	N-Adj	No	Middle East	Semitic
Hindi	SOV	Gen-N	Adj-N	No	South Asia	Indo-Aryan
Indonesian	SVO	N-Gen	N-Adj	No	SE Asia	Austronesian
Irish	VSO	N-Gen	N-Adj	No	Europe	Gaelic
Japanese	SOV	Gen-N	Adj-N	Yes	East Asia	Japonic
Turkish	SOV	Gen-N	Adj-N	Yes	Eurasia	Turkic

Table 1: Features of Languages

Language	Word Length	Sent Length	Depth	Distance	Children (Internal)	Children
Chinese	1.59	26.7	4.3	3.82	2.47	0.96
English	3.54	16.3	3.40	3.32	2.84	0.94
Finnish	5.97	14.8	3.22	2.86	2.55	0.92
Hebrew	3.08	28.4	5.12	3.43	2.33	0.96
Hindi	3.79	22.3	4.03	3.27	2.43	0.95
Indonesian	5.17	23.8	4.57	3.06	2.37	0.95
Irish	4.16	26.4	4.52	3.61	2.36	0.96
Japanese	1.75	24.5	4.38	2.69	2.74	0.96
Turkish	5.25	12.3	2.96	2.45	2.16	0.90

Table 2: Features of UD Treebank

verse collection of languages, we studied different WALS properties (Dryer and Haspelmath, 2013) which are often used in current literature for identifying typological features, such as WALS feature 82A: Order of Subject and Verb, WALS feature 83A: Order of Object and Verb, WALS feature 85A: Order of Adposition and Noun Phrase, WALS feature 86A: Order of Genetive and Noun, and 87A: Order of Adjective and Noun. We also chose languages that were of different language families, derivational morphology type (such as agglutinating vs isolating), as well as geographical region. Furthermore, we looked for languages which had at least 5000 sentences in the treebank and are not low-resource languages. This enables consistency between our languages, and shows the benefits of using a high-resource language to supplement a small treebank. The languages and their associated properties are shown in Table 1.

After selecting our set of languages, we then analyzed the treebanks of each language to see how each one is structured. For each treebank, we analyzed the average word length, average sentence length, average depth of the tree, average distance between word and its dependent in the dependency parse, the average number of children (branching factor) of internal nodes, as well as the average number of children of all nodes. The results can be seen in Table 2. From this analysis, we can see that many of the languages have differently structured treebanks: Finnish as an agglutinating language has long words and short sentences when compared to isolating languages such as Chinese. Additionally, because Finnish has shorter sentences, its trees are also not as deep as that of the other trees in other languages. Furthermore, while Turkish and Japanese have very similar grammars, we can see that the treebanks are still structured differently in terms of sentence length and depth. This result may be due to different treebank practices, as well as different linguistic phenomena between the languages.

4 Experiment

For our experiments, we used a multilingual parser to train on treebanks with different combinations of languages.

4.1 Experimental Design

Rather than use a low-resource language with minimal data, we instead chose to use data ablation on English to simulate a low-resource language. This allowed us to have a better benchmark for the performance of the different language, as we are able to train the test set of English sentences with 5000 sentences. English also aided us with error analysis of parsed sentences, as we could easily identify the correct parse of the sentence and compare to it to the generated parse.

For each language in our test test, we randomly chose 5000 sentences from the full set of sentences in the UD treebank. We then concatenated the treebank of these 5000 sentences with 0, 50, 100, or 250 sentences from the English treebank. After training MaLOPa on this combined treebank, we tested on 300 English sentences to get the labeled attachment score of the output sentences to measure the helpfulness of the supplemental language.

4.2 MaLOPa Setup

We used the multilingual parser MaLOPa for our experiments. We used a language embedding by prefixing each word with the associated language. We did not include the lexical embeddings of the different languages, and we also did not include the Brown embedding clusters or the learned embeddings of the POS tags. We stochastically dropout the embeddings for each token with a 50% probability, and use fine-grained as well as coarse POS tags for each language.

5 Results

As a baseline, we first tested MaLOPa trained on differing numbers of English sentences from the English UD treebank. The results of this are shown in Table 3. We see that the LAS score starts very low when the parser is trained on minimal number of sentences, but quickly increases as more sentences are added. The curve is logarithmic: as the LAS score improves at a slower rate as the number of sentences increase. When trained on 5000 English sentences, the parser was able to obtain an LAS score of 0.823.

The LAS scores of the 5000 sentences of each language in our selected set concatenated with 0, 50, 100, or 250 English sentences can be viewed in Table 4. From the results, we can see that the languages have very different LAS scores when there are no English sentences included with the training data. This result is to be expected, as many of the languages have very different grammatical structures compared to English, and the learned shift reduce actions may not transfer well across languages.

English	LAS
10	0.383
20	0.447
30	0.490
40	0.544
50	0.568
100	0.641
150	0.663
200	0.726
250	0.731
300	0.734
1000	0.765
5000	0.823

Table 3: Baseline LAS scores of MaLOPa trained on differing numbers of English sentences with no other language

When the parser is trained on an additional 50 English sentences, the scores significantly improve compared to the LAS score when trained on zero English sentences. Some of the languages such as Finnish and Indonesian were able to obtain LAS scores that were higher than that of the parser trained on only 50 English sentences. However, perhaps due to the larger number of training data from another languages, some of the languages did not score as high as that of the baseline 50 sentences. This suggests that typological similarities and discrepancies may effect the accuracy of cross-lingual dependency parsing.

As the number of English sentences in the training set increase, the accuracy of the parser for each language also increases when tested on English. Additionally, The LAS scores of the languages seem to converge as the number of English sentences increase. While the languages had very different results for zero sentences, the results seem to be a lot closer when 250 English sentences were included in the training dataset. This may be due to the construction of cross-lingual parser used: because the language embedding was included with the sentence, it is possible that MaLOPa placed a higher emphasis on the English sentences included in the training while setting parameters.

When compared to the baseline model, it appears that adding the 5000 sentences of a different language may cause adverse results compared to

Language	0	50	100	250
Chinese	0.350	0.563	0.635	0.709
Finnish	0.290	0.597	0.648	0.705
Hebrew	0.281	0.559	0.623	0.698
Hindi	0.292	0.541	0.647	0.708
Indonesian	0.192	0.572	0.645	0.728
Irish	0.300	0.561	0.655	0.713
Japanese	0.130	0.471	0.559	0.684
Turkish	0.298	0.562	0.634	0.694

Table 4: LAS of 5000 sentences in language combined with labeled number of English sentences

the model trained on only English. Therefore, it seems that it may not be necessary to supplement treebanks of size 100 or larger with sentences from a different model, as inconsistencies between multiple languages may lead to worse LAS results as the training data is more variable in terms of tree structure.

6 Looking at Japanese

For many of the results, we noticed that many of the languages converged to similar scores; however, the LAS scores for the training data using the Japanese treebank consistently appeared lower than the other scores. After looking into the Japanese treebank, we noticed that there were some inconsistencies between the treebank and the UD guidelines: many of the words were marked as auxilary verbs when they should have been marked as case markers, and a particular case marker 'koto', which is very common in Japanese sentences, was treated as the dependent of the verb rather than its proper placement as head of the phrase.

Because this did not strictly adhere to the UD conventions, we decided to experiment with the importance of UD conventions by removing all instances of 'koto' from the trees: each sentence which originally contained 'koto' in the dependency parse was modified by simply eliminating 'koto' from the sentence. Because 'koto' was always a leaf node, it could be completely removed from the tree with no consequences since it did not have any dependents. Although this removal did form ungrammatical Japanese sentences, the resulting trees in the treebank were more indicative of UD schema. The statistics of this new treebank can be seen in Table 5, compared to that of the original Japanese. As can be expected, the average sentence length, depth, and distance between dependent and head were reduced, as the new sentences contained fewer words due to the elimination.

Property	Original	Modified
Word Length	1.75	1.75
Sent Length	24.5	22.4
Depth	4.38	4.36
Distance	2.69	2.67
Children (Internal)	2.74	2.73
Children	0.96	0.95

Table 5: Features of Japanese Treebanks

Treebank	0	50	100	250
Original	0.130	0.471	0.559	0.684
Modified	0.130	0.499	0.590	0.688

Table 6: LAS of 5000 sentences in different Japanese treebanks combined with labeled number of English sentences

We then tested this modified Japanese treebank with 0, 50, 100, and 250 English sentences. The new LAS results for this treebank are shown in Table 6. We can see that this new treebank significantly improved the LAS score of the test set of English sentences for all instances when English sentences were included in the training set. Because the training dataset better adhered to UD conventions, it appears that it greatly improved the performance of the model. This was only one modification to the Japanese treebank: the treebank still contains a few instances of mislabeled cases which have not been corrected. Therefore, it is possible that the resulting LAS scores would be even higher if the treebank entirely conformed to the UD rules.

7 Conclusions

From our experiments, we conclude that the most important factor when supplementing a low-resource language with additional sentences is conformity of the treebanks to UD conventions. When training a parser on treebanks from different languages, it is very important to keep the labeling consistent, as discrepancies in how the treebanks are set up and what properties are emphasized in the language may lead to poor results.

Furthermore, we were not able to find a strong correlation between parser performance and typological features of the language. Most of the LAS scores seem to converge, and were not significantly different from one another when the treebank was concatenated with 100 or more sentences from the English treebank. Therefore, it seems that the features of the selected language may not be as important as the consistencies between labeling and treebank ideology.

8 Future Work

Cross-lingual dependency parsing is often done on training data consisting of multiple languages, so it may be worthwhile to investigate results of training on a combination of different languages rather than only using 5000 sentences of a single language. For example, it may be possible that training on two or three languages from a language family be helpful to the parsing of another language from the same language family. We can also further investigate different ratios of high-resource and low-resource languages to find the optimal balance which leads to the best results.

Additionally, we can look into different features of the treebanks, such as genre. The treebanks are not from a consistent genre, and it is very possible that a parser trained on reviews or tweets would not transfer well when parsing a news sentence. Therefore, genre of treebanks should also be taken into consideration when selecting languages and treebanks for cross-lingual dependency parsing.

Further work could also be done to test on languages other than English. It may be possible that English has a specialized structure, and it is not evident that our findings can be generalized to that of other languages. Additionally, we could also include different high-resource languages as part of the set of languages to test on to ensure that the languages chosen are indicative of their language family and other typological properties. The current selection of languages of UD treebanks are diverse, but not all-inclusive, and it may be difficult to languages with specific properties from specific regions.

Acknowledgements

I want to thank Lori Levin and David Mortensen for advising me and providing support through this project. Additionally, I'd like to thank Waleed Ammar for helping me set up MaLOPa and making sure the parameters and configurations are properly set up.

References

W. Ammar, G. Mulcaire, M. Ballesteros, C. Dyer, and N. Smith. 2016. One parser, many languages. *CoRR* abs/1602.01595. http://arxiv.org/abs/1602.01595.

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

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, , and Ting Li. 2015. Cross-lingual dependency parsing based on distributed representations. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics* .

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuz man Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Tackstrom, Claudia Bedini, Nuria Bertomeu Castello, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* Volume 2: Short Papers:92–97.

T. Naseem, R. Barzilay, and A. Globerson. 2012. Selective sharing for multilingual dependency parsing. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* pages 692–637.