Cross-Lingual Treebank Combination for Speech Dependency Parsing

Sara Stymne and Joakim Nivre
Department of Linguistics and Philology
Uppsala University
{sara.stymne,joakim.nivre}@lingfil.uu.se

Abstract

Parsing speech data is challenging, especially for the large number of languages lacking syntactically annotated speech data. In this paper we show that we can often improve parsing in such cases by combining in-language written treebanks with speech treebanks from other languages. We formulate this in a parsing framework with treebank embeddings, and propose the use of a parameter free method, which can be applied to a language without development data. Finally we apply this method to Swedish, performing on error analysis for a challenging set of sentences, and show that the treatment of fillers is improved, whereas reparandum relations are still challenging.

1 Introduction

Recent advances in dependency parsing have enabled high-quality parsing for a relatively high number of languages. However, satisfactory results are mainly limited to text types for which there are treebanks for a specific language. Even for high-resource languages, treebanks are typically only available for a small number of domains and genres. In this work we propose a data combination method utilizing data for the desired text type across language boundaries. We note that treebanks for a specific text type often exist for some language(s), and we show that we can take advantage of such data for parsing this text type in other languages. Our main focus is on the case where we want to parse data for a language that has some resources, but none for the text type in question.[1]

Although the method applies to arbitrary text types, in this paper we focus on one low-resource text type, which stands out from canonical written texts: (transcribed) speech, for which annotated treebanks exist only for a small number of languages. Speech tends to be more informal than written texts, and contains features such as fillers, restarts, and reparandums. We use the Universal Dependencies (UD) (Nivre et al., 2020) treebanks for our experiments. We apply our methods to three languages where speech treebanks are available, and also perform a small manual evaluation for Swedish, since Swedish does not have a UD treebank for speech data, with the exception of a small sample of Europarl data, which is highly edited speech.[2] In this work, we combine adaption to specific text types and cross-lingual parsing. While there is plenty of work both on cross-lingual parsing (Ammar et al., 2016; Ahmad et al., 2019; Kondratyuk and Straka, 2019) and domain adaptation for parsing (Nivre et al., 2007; Kim et al., 2016; Sato et al., 2017; Xiuming et al., 2019), there is to the best of our knowledge no attempts to combine these approaches in a uniform framework for dependency parsing.

We adapt the parsing framework of Smith et al. (2018a), which incorporates treebank embeddings to represent treebanks, similarly to language embeddings (Ammar et al., 2016; de Lhoneux et al., 2017a). Each parsing model is trained on a concatenation of different treebanks, and the representation of each input token includes an embedding representing the treebank from which the token comes from. Depending on the mix of treebanks, the treebank embedding can encode aspects such as differences between languages, domains, and annotation style. Parsing with treebank embeddings has previously been applied monolingually (Stymne et al., 2018; Wagner et al., 2020) and cross-lingually for related languages, but without taking domain into

[1] This abstract overlaps with Stymne (2020), who also presents results with the same method on Twitter data.

[2] There is also treebanked speech in Talbanken (Einarsson, 1976), but these sections of Talbanken have not been converted to UD.

Language	Speech Treebank	Train	Dev	Test	Additional data
French	Spoken	15.0K	10.2K	10.2K	GSD (364K), *Partut* (24.9K), Sequoia (51.9K)
Norwegian	NynorskLIA	35.2K	10.2K	10.0K	*Nynorsk* (245K), Bokmaal (244K)
Slovenian	SSJ	18.6K	906	10.0K	*SST* (113K), Croatian_SET (153K), Serbian_SET (74.3K)
Swedish	Our custom	–	–	129	Talbanken (66.7K), LinES (48.3K)

Table 1: Treebanks and number of tokens in train, dev, and test data sets for the target treebanks. Additional data lists treebanks used for each target treebank, which is in-language unless otherwise noted, and the number of tokens in the training set for each treebank. Treebanks in italics are used in the contrastive data sets.

account (Smith et al., 2018a; Lim et al., 2018),[3] In this paper, we show that joint training with treebank embeddings can be applied both across languages and domains, in effect addressing the task of cross-lingual domain adaptation. It is a simple and efficient method, which does not require expensive pre-processing, pre-training, or translation.

When the input sentence is not from a treebank used during training there is a need to determine the treebank embedding. One option is to use a *proxy* treebank (Stymne et al., 2018), i.e. to choose the embedding from one of the treebanks used for training, which can be decided based on development data. Wagner et al. (2020) showed that it is often advantageous to instead interpolate the embeddings of the treebanks used for training. They show in a monolingual setting how interpolation weights can be learnt based on sentence similarity. However, their equal weight baseline performs just as well in the majority of cases, and avoids the need of learning interpolation weights, which would also be less straight-forward in the cross-lingual setting. We thus adopt equal-weight interpolation.

In this paper, we explore this method on transcribed speech data. Our parameter free model allows us to directly train a model for Swedish, for which we have no annotated speech data at all, and can potentially be applied to other languages and text types as well.

2 Experimental Setup

Data We use data from the Universal Dependencies (UD) project (Nivre et al., 2016, 2020), version 2.4 (Nivre et al., 2019). Our main focus is on languages with a single-domain dependency treebank with speech data, including both training and test data and treebank data for other domains. French, Norwegian, and Slovenian fulfills our criteria. In addition to the in-domain treebanks we use additional treebanks from the same language, and two additional related languages for Slovenian, which has quite small treebanks. We also train models for Swedish, which we use for a small error analysis. We selected and annotated 13 challenging sentences taken from Eklund (2004). The sentences are sampled from travel dialogues and all contain disfluencies and repairs of different kinds.

Table 1 lists the data used for each language. Note that in all cases, the additional data is much larger than the speech data, which is typically quite small. For Slovenian SST, no development data was available, so we split off 5% of the training data. In all other cases we use the original splits. While UD treebanks have standard annotation guidelines, there are several inconsistencies between the treebanks used, especially for the rather unusual features of speech data.

To be able to compare the effect of adding in-domain data, we also create a contrastive dataset for each language with the same number of tokens as in the corresponding speech treebank. For each language, we sample data from the treebank marked with italics in Table 1.

Parser We use the transition-based uuparser[4] (de Lhoneux et al., 2017b). It uses a BiLSTM as a feature extractor followed by a multi-layer perceptron predicting transitions, in the style of Kiperwasser and Goldberg (2016). Each word, w_i, is represented by the concatenation of a word embedding, $e_w(w_i)$, a character-level embedding, obtained by running a BiLSTM over the characters ch_j ($1 \leq j \leq m$) of w_i, where m is the word length in characters, and a treebank embedding, $e_{tb}(t^*)$:

$$e_i = [e_w(w_i); \text{BiLSTM}(ch_{1:m}); e_{tb}(t^*)] \quad (1)$$

The treebank embedding represents a treebank, t^*, which is chosen among the set of k treebanks used when training the model. During training, t^* is chosen as the treebank to which the current word/sentence belongs. When applying the model,

[3] With the exception of a footnote in Smith et al. (2018a), where this type of data combination is mentioned for spoken French and Naija. However, no details or experimental results are provided.

[4] https://github.com/UppsalaNLP/uuparser

Same language		Other language		French		Norwegian		Slovenian	
Speech	Written	Speech	Written	UAS	LAS	UAS	LAS	UAS	LAS
–	–	–	X	18.2	2.8	24.2	8.8	25.9	7.3
–	–	X	–	21.8	2.2	19.1	3.6	20.7	4.7
–	X	–	–	74.8	63.4	60.1	52.8	60.2	46.9
–	X	–	X	75.3	64.3	62.2	54.4	60.1	47.6
–	X	X	–	75.9	64.5	59.1	52.0	63.2	52.7

Table 2: Test set scores for spoken data with different combinations of training data, using the best proxy treebank. Note that Same language includes related Slavic languages for Slovenian.

the treebank of the sentence can be used only if the test sentence comes from a treebank that was used during training. In other cases some other method has to be used. In this work we explore the following methods:

- Proxy treebank: when dev data is available, we can try all possible proxy treebanks i.e. all treebanks used during training the model, and choose the treebank, t^*, which performs best on dev data.
- Interpolation: We interpolate the embeddings from all treebanks used during training by averaging them with equal weights: $(t^* = \sum_{t=1}^{k} \frac{1}{k} e_{tb}(t))$

Note that we only apply these techniques at test time. The interpolation method only requires a single test run. Proxy treebank requires k dev test runs, followed by a single test run. Interpolation has the advantage of being parameter free, while proxy treebanks require dev data.

We use the default hyperparameters of uuparser, as specified in Smith et al. (2018a). Note that no POS-tags are used, since POS-tagging in these difficult domains would lead to the same issues as for parsing. In addition, character embeddings compensate for the lack of POS-tags to a large extent across several typologically different languages (Smith et al., 2018b), and in order for universal POS-tags, the most feasible choice cross-lingually, to be useful for parsing, the tagging quality has to be prohibitively high (Gómez-Rodríguez, 2020). The parser is trained end-to-end on treebank data, without any pre-training. All embeddings are initialized randomly at training time. Each model is trained for 30 epochs, and the best epoch is chosen based on average development scores.

3 Results

We use the full set of treebanks from Table 1 in our experiments.[5] For other language written data, we use the contrastive datasets sampled from the same languages as the other language speech data. We use labelled and unlabelled attachment score (LAS/UAS) for evaluation.

We first present results using different sources of training data, speech or written, from the same or another language, choosing the best proxy treebank based on development UAS scores. Our main interest is the bottom part of Table 2, where we investigate the effect of adding speech data from other languages to in-language written data. For Slovenian and French, adding out-of-language speech data to the in-language written data leads to gains compared to both variants trained only on written data. It is thus not the extra data alone, which helps here, but only data from a matching text type. For Norwegian, however, it is for some reason better to use additional written data, and speech data does not help. We leave an investigation of this to future work. As a point of comparison, if we train a model only on the relatively small in-language speech treebank results go up with 4.9–17.4 LAS points compared to our proposed model, showing the importance of using appropriate data when available.

The top of Table 2 shows results when no in-language data is available. As expected these scores are considerably lower than when adding in-language written data, being so poor that these parsers are hardly useful, confirming previous research, e.g. Meechan-Maddon and Nivre (2019) and Vania et al. (2019). In this case speech data actually gives worse LAS results for all languages.

Next, we focus on our main scenario of interest, where we have in-language written data and out-of-language speech data. We investigate how best to apply the model at test time for this case, where the treebank, i.e. the combination of language and text type, has not been seen at training time. We compare using a proxy treebank, matching either language or domain, or interpolation. Table 3 summarizes the results.

When choosing a single proxy it is mostly better to use the same language than the same do-

[5] Using a subset of these treebanks mostly gave lower scores but showed the same trends.

	Proxy Language			Proxy Speech			Interpolation	
	UAS	LAS	Proxy	UAS	LAS	Proxy	UAS	LAS
French	75.9	64.5	fr_partut	76.0	61.7	no_nynorsklia	75.2	63.7
Norwegian	59.1	52.0	no_bokmaal	60.2	51.0	sl_sst	61.2	53.5
Slovenian	63.2	49.6	sr_set	61.3	48.1	no_nynorsklia	63.8	52.2

Table 3: Test scores for models trained on in-language written data and speech data from other languages, using different methods for applying it to the target treebank.

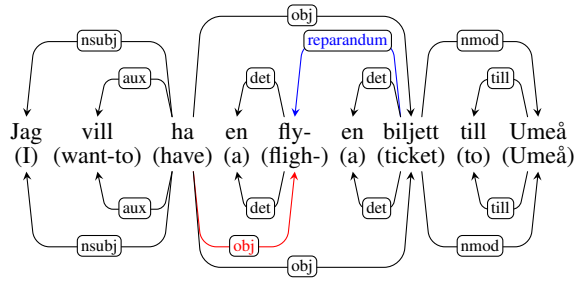

Figure 1: Example sentence from Swedish test set with gold standard tree (top) and parse produced by speech-enhanced parser (bottom).

main. The interpolation method works well on both metrics, giving the best results for Norwegian and Slovenian, and falling only slightly behind for French. Table 3 also shows the best proxy used, either from a matching text type or language. For language proxies we note some surprises, Norwegian Bokmaal is a better fit than the matching language variety Nynorsk, and the Serbian corpus is better than the Slovenian. The differences between proxies are typically small, though.

For Swedish, we only apply the interpolation method, since we have no development data to use for selecting a proxy treebank. We evaluate three different models, one trained using only Swedish written data, and two models where we added the spoken treebanks or the contrastive data from the other three languages. The Swedish test set, consisting of only 13 sentences and 129 tokens, is too small to make quantitative evaluation really meaningful, so we concentrate instead on a qualitative error analysis.

The overall impression is that the parsers, regardless of training data, cope well with regular syntactic structures but struggle with disfluencies, in particular repairs. A typical example is shown in Figure 1, where the interrupted noun phrase *en fly-* (a fligh-) is repaired by the noun phrase *en biljett till Umeå* (a ticket to Umeå). According to the UD guidelines, this structure should be analyzed by attaching (the head of) the aborted phrase to the phrase that replaces it with the *reparandum* relation. All three parsers instead parse the first noun phrase as a direct object of the verb *boka* (book) and then adds the repair as a second object.[6] It is not surprising that a parser trained on only written language data will produce this type of error, since the *reparandum* relation does not occur at all in written texts. However, adding speech data from other languages does not seem to solve this problem, probably because the *reparandum* relation is too rare for the parser to generalize across languages. By contrast, adding speech data helps with respect to fillers like *eh* and *um*, which should be annotated using the *discourse* relation in UD. While the baseline parser trained on only Swedish written data rarely uses this relation, the parser trained on speech data from other languages gets the label correct in almost all cases but sometimes attaches the filler to the wrong head.[7] Finally, it is worth noting that adding non-speech data from other languages has no positive effect and leads to lower accuracy for several sentences.

4 Conclusion

We have shown how we can improve parsing for transcribed speech by combining speech data from other languages with in-language written data. We show that it is possible to do so using a single parsing model with treebank embeddings, and that this mostly leads to improvements. We also propose the use of a parameter free method for applying treebank embeddings to new data at test time, which gives competitive results, and which allows as to apply the model directly to a language without annotated speech data, like Swedish.

In future work we want to apply our methods also to other text types and to explore how the data selection strategies work with other parsing frameworks. We also want to extend the work on weighted interpolation by Wagner et al. (2020) to the cross-lingual case, to be able to combine it with the proposed methods.

[6] This is in fact a violation of the UD guidelines, which does not allow a verb to have two direct objects.

[7] Admittedly, the notion of syntactic head is not always clear-cut for fillers of this kind.

References

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 2440–2452, Minneapolis, Minnesota, US.

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.

Jan Einarsson. 1976. Talbankens talspråkskonkordans. Lund University, Department of Scandinavian Languages.

Robert Eklund. 2004. *Disfluency in Swedish human–human and human–machine travel booking dialogues*. Ph.D. thesis, Linköping University.

Mark Anderson Carlos Gómez-Rodríguez. 2020. On the frailty of universal POS tags for neural UD parsers. In *Accepted to CoNLL 2020*.

Young-Bum Kim, Karl Stratos, and Ruhi Sarikaya. 2016. Frustratingly easy neural domain adaptation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 387–396, Osaka, Japan.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 2779–2795, Hong Kong, China.

Miryam de Lhoneux, Yan Shao, Ali Basirat, Eliyahu Kiperwasser, Sara Stymne, Yoav Goldberg, and Joakim Nivre. 2017a. From raw text to universal dependencies - look, no tags! In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 207–217, Vancouver, Canada.

Miryam de Lhoneux, Sara Stymne, and Joakim Nivre. 2017b. Arc-hybrid non-projective dependency parsing with a static-dynamic oracle. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 99–104, Pisa, Italy.

KyungTae Lim, Cheoneum Park, Changki Lee, and Thierry Poibeau. 2018. SEx BiST: A multi-source trainable parser with deep contextualized lexical representations. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 143–152, Brussels, Belgium.

Ailsa Meechan-Maddon and Joakim Nivre. 2019. How to parse low-resource languages: Cross-lingual parsing, target language annotation, or both? In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 112–120, Paris, France.

Joakim Nivre, Mitchell Abrams, Željko Agić, et al. 2019. Universal dependencies 2.4. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 915–932, Prague, Czech Republic.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 4034–4043, Marseille, France.

Motoki Sato, Hitoshi Manabe, Hiroshi Noji, and Yuji Matsumoto. 2017. Adversarial training for cross-domain universal dependency parsing. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 71–79, Vancouver, Canada.

Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. 2018a. 82 treebanks, 34 models: Universal dependency parsing with multi-treebank models. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 113–123, Brussels, Belgium.

Aaron Smith, Miryam de Lhoneux, Sara Stymne, and Joakim Nivre. 2018b. An investigation of the interactions between pre-trained word embeddings, character models and POS tags in dependency parsing.

In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2711–2720, Brussels, Belgium.

Sara Stymne. 2020. Cross-lingual domain adaptation for dependency parsing. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 62–69, Düsseldorf, Germany. Association for Computational Linguistics.

Sara Stymne, Miryam de Lhoneux, Aaron Smith, and Joakim Nivre. 2018. Parser training with heterogeneous treebanks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 619–625, Melbourne, Australia.

Clara Vania, Yova Kementchedjhieva, Anders Sogaard, and Adam Lopez. 2019. A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. 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 1105–1116, Hong Kong, China.

Joachim Wagner, James Barry, and Jennifer Foster. 2020. Treebank embedding vectors for out-of-domain dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8812–8818, Online.

Qiao Xiuming, Zhang Yue, and Zhao Tiejun. 2019. Learning domain invariant word representations for parsing domain adaptation. In *Natural Language Processing and Chinese Computing (NLPCC 2019)*, pages 801–813, Dunhuang, China.