Cross-Lingual Dependency Parsing with Universal Dependencies and Predicted PoS Labels

Jörg Tiedemann
Uppsala University
Department of Linguistics and Philology
firstname.lastname@lingfil.uu.se

Abstract

This paper presents cross-lingual models for dependency parsing using the first release of the universal dependencies data set. We systematically compare annotation projection with monolingual baseline models and study the effect of predicted PoS labels in evaluation. Our results reveal the strong impact of tagging accuracy especially with models trained on noisy projected data sets. This paper quantifies the differences that can be observed when replacing gold standard labels and our results should influence application developers that rely on cross-lingual models that are not tested in realistic scenarios.

1 Introduction

Cross-lingual parsing has received considerable attention in recent years. The demand for robust NLP tools in many languages makes it necessary to port existing tools and resources to new languages in order to support low-resource languages without starting their development from scratch. Dependency parsing is one of the popular tasks in the NLP community (Kübler et al., 2009) that also found its way into commercial products and applications. Statistical parsing relies on annotated data sets, so-called treebanks. Several freely available data sets exist but still they only cover a small fraction of the linguistic variety in the world (Buchholz and Marsi, 2006; Nivre et al., 2007). Transferring linguistic information across languages is one approach to add support for new languages. There are basically two types of transfer that have been proposed in the literature: data transfer approaches and model transfer approaches. The former emphasizes the projection of data sets to new languages and it usually relies on parallel data sets and word alignment (Hwa et al., 2005; Tiedemann, 2014). Recently, machine translation was also introduced as yet another alternative to data transfer (Tiedemann et al., 2014). In model transfer, one tries to port existing parsers to new languages by (i) relying on universal features (McDonald et al., 2013; McDonald et al., 2011a; Naseem et al., 2012) and (ii) by adapting model parameters to the target language (Täckström et al., 2013). Universal features may refer to coarse part-of-speech sets that represent common word classes (Petrov et al., 2012) and may also include language-set-specific features such as cross-lingual word clusters (Täckström et al., 2012) or bilingual word embeddings (Xiao and Guo, 2014). Target language adaptation can be done using external linguistic resources such as prior knowledge about language families or lexical databases or any other existing tool for the target language.

This paper is focused on data transfer methods and especially annotation projection techniques that have been proposed in the related literature. There is an on-going effort on harmonized dependency annotations that makes it possible to transfer syntactic information across languages and to compare projected annotation and cross-lingual models even including labeled structures. The contributions of this paper include the presentation of monolingual and cross-lingual baseline models for the recently published universal dependencies data sets (UD; release 1.0)[1] and a detailed discussion of the impact of PoS labels. We systematically compare results on standard test sets with gold labels with corresponding experiments that rely on predicted labels, which reflects the typical real-world scenario.

Let us first look at baseline models before starting our discussion of cross-lingual approaches. In all our experiments, we apply the Mate tools (Bohnet, 2010; Bohnet and Kuhn, 2012) for train-

[1] http://universaldependencies.github.io/docs/

Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), pages 340–349,
Uppsala, Sweden, August 24–26 2015.

ing dependency parsers and we use standard settings throughout the paper.

2 Baseline Models

Universal Dependencies is a project that develops cross-linguistically consistent treebank annotation for many languages. The goal is to facilitate cross-lingual learning, multilingual parser development and typological research from a syntactic perspective. The annotation scheme is derived from the universal Stanford dependencies (De Marneffe et al., 2006), the Google universal part-of-speech (PoS) tags (Petrov et al., 2012) and the Interset interlingua for morphological tagsets (Zeman and Resnik, 2008). The aim of the project is to provide a universal inventory of categories and consistent annotation guidelines for similar syntactic constructions across languages. In contrast to previous attempts to create universal dependency treebanks, the project explicitly allows language-specific extensions when necessary. Current efforts involve the conversion of existing treebanks to the UD annotation scheme. The first release includes ten languages: Czech, German, English, Spanish, Finnish, French, Irish, Italian, Swedish and Hungarian. We will use ISO 639-1 language codes throughout the paper (cs, de, en, es, fi, fr, ga, it, sv and hu).

UD comes with separate data sets for training, development and testing. In our experiments, we use the provided training data subsets for inducing parser models and test their quality on the separate test sets included in UD. The data sizes vary quite a lot and the amount of language-specific information is different from language to language (see Table 1. Some languages include detailed morphological information (such as Czech, Finnish or Hungarian) whereas other languages only use coarse PoS labels besides the raw text. Some treebanks include lemmas and enhanced PoS tag sets that include some morpho-syntactic features. We will list models trained on those features under the common label "morphology" below.

The data format is a revised CoNLL-X format which is called CoNLL-U. Several extensions have been added to allow language-specific representations and special constructions. For example, dependency relations may include language-specific subtypes (separated by ":" from the main type) and multiword tokens can be represented by both, the surface form (that might be a contraction of multiple words) and a tokenized version. For multiword units, special indexing schemes are proposed that take care of the different versions.[2] For our purposes, we remove all language-specific extensions of dependency relations and special forms and rely entirely on the tokenized version of each treebank with the standard setup that is conform to the CoNLL-X format (even in the monolingual experiments). In version 1.0, language-specific relation types and CoNLL-U-specific constructions are very rare and, therefore, our simplification does not alter the data a lot.

language	size	lemma	morph.	LAS	UAS	LACC
CS	60k	X	X	85.74	90.04	91.99
DE	14k			79.39	84.38	90.28
EN	13k		(X)	85.70	87.76	93.29
ES	14k			84.05	86.77	92.90
FI	12k	X	X	84.51	86.51	93.53
FR	15k			81.03	84.39	91.02
GA	0.7k	X		72.73	78.75	84.74
HU	1k	X	X	83.19	85.28	92.73
IT	9k	X	X	89.58	91.86	95.92
SV	4k		X	82.66	85.66	91.06

Table 1: Baseline models for all languages included in release 1.0 of the universal dependencies data set. Results on the given test sets in labeled accuracy (LAS), unlabeled accuracy (UAS) and label accuracy (LACC).

After our small modifications, we are able to run standard tools for statistical parser induction and we use the Mate tools as mentioned earlier to obtain state-of-the-art models in our experiments. Table 1 summarizes the results of our baseline models in terms of labeled and unlabeled attachment scores as well as label accuracy. All models are trained with the complete information available in the given treebanks, i.e. including morphological information and lemmatized tokens if given in the data set. For morphologically rich languages such as Finnish or Hungarian these features are very important to obtain high parsing accuracies as we will see later on. In the following, we look at the impact of various labels and compare also the difference between gold annotation and predicted features in monolingual parsing performance.

3 Gold versus Predicted Labels

Parsing accuracy is often measured on test sets that include manually verified annotation of essential features such as PoS labels and morphological

[2] See http://universaldependencies.github.io/docs/format.html for more details.

LAS/ACCURACY	CS	DE	EN	ES	FI	FR	GA	HU	IT	SV
gold PoS & morphology	85.74	—	85.70	—	84.51	—	72.73	83.19	89.58	82.66
gold coarse PoS	80.75	79.39	84.81	84.05	74.62	81.03	71.39	73.39	88.25	81.02
delexicalized & gold PoS	70.36	71.29	76.04	75.47	59.54	74.19	66.97	66.57	79.07	66.95
coarse PoS tagger (accuracy)	98.28	93.19	94.89	95.13	95.69	95.99	91.97	94.69	97.63	96.79
morph. tagger (accuracy)	93.47	—	94.80	—	94.53	—	91.92	91.06	97.50	95.26
predicted PoS & morphology	82.67	—	81.36	—	80.59	—	66.74	75.78	87.16	78.76
predicted coarse PoS	79.41	74.39	80.33	80.16	70.25	78.73	65.93	68.04	85.08	76.42
delexicalized & predicted PoS	62.44	61.82	67.40	69.03	49.79	68.60	55.33	58.90	72.92	61.99

Table 2: The impact of morphology and PoS labels: Comparing gold labels with predicted labels.

properties. However, this setup is not very realistic because perfect annotation is typically not available in real-world settings in which raw text needs to be processed. In this section, we look at the impact of label accuracy and compare gold feature annotation with predicted one. Table 2 summarizes the results in terms of labeled attachment scores.

The top three rows in Table 2 refer to models tested with gold annotation. The first one corresponds to the baseline models presented in the previous section. If we leave out morphological information, we achieve the performance shown in the second row. German, Spanish and French treebanks include only the coarse universal PoS tags. English includes a slightly more fine-grained PoS set besides the universal tag set leading to a modest improvement when this feature is used. Czech, Finnish, Hungarian and Italian contain lemmas and morphological information. Irish include lemmas as well but no explicit morphology and Swedish has morphological tags but no lemmas. The impact of these extra features is as expected and mostly pronounced in Finnish and Hungarian with a drop of roughly 10 points in LAS when leaving them out. Czech also drops with about 5 points without morphology whereas Italian and Swedish do not seem to suffer much from the loss of information. The third row shows the results of delexicalized parsers. In those models, we only use the coarse universal PoS labels to train parsing models that can be applied to any of the other languages as one simple possibility of cross-lingual model transfer. As we can see, this drastic reduction leads to significant drops in attachment scores for all languages but especially for the ones that are rich in morphology and more flexible in word order.

In order to contrast these results with predicted features, we also trained taggers that provide automatic labels for PoS and morphology. We apply Marmot (Müller and Schütze, 2015), an efficient implementation for training sequence labelers that include rich morphological tag sets. The tagger performance is shown in the middle of the table.

The three rows at the bottom of Table 2 list the results of our parsing experiments. The first of them refers to the baseline model when applied to test sets with predicted coarse PoS labels and morphology (if it exists in the original treebank we train on). We can see that we loose 2-4 points in LAS with Irish and Hungarian being a bit stronger effected (showing 5-7 points drop in LAS). Irish and Hungarian treebanks are, however, very small and we cannot expect high tagging accuracies for those languages especially with the rich morphological tag set in Hungarian. In general, the performance is quite a good achievement especially considering the languages that require rich morphological information such as Finnish and Czech and this is due to the high quality of the taggers we apply. As expected, we can observe significant drops again when taking out morphology. The effect is similar to the results with gold labels when looking at absolute LAS differences.

The final row represents the LAS for delexicalized models when tested against data sets with predicted PoS labels. Here, we can see significant drops compared to the gold standard results that are much more severe than we have seen with the lexicalized counterparts. This is not surprising, of course, as these models entirely rely on these PoS tags. However, the accuracy of the taggers is quite high and it is important to stress this effect when talking about cross-lingual parsing approaches. In the next section, we will investigate this result in more detail with respect to cross-lingual models.

4 Cross-Lingual Delexicalized Models

The previous section presented delexicalized models when tested on the same language they are trained on. The primary goal of these models is,

	← target (test) language →									
LAS	CS	DE	EN	ES	FI	FR	GA	HU	IT	SV
CS		48.90	43.78	43.82	42.18	40.70	30.28	32.18	43.93	40.09
DE	47.27		47.80	53.63	33.45	51.60	37.63	39.41	53.63	46.14
EN	44.27	54.27		60.94	38.52	60.53	39.31	34.06	61.88	50.76
ES	48.40	52.59	50.10		32.80	65.40	43.84	34.46	69.54	46.79
FI	43.75	38.31	40.36	30.14		28.54	20.15	37.39	27.49	37.97
FR	43.63	53.04	52.55	66.42	31.44		41.82	34.53	69.62	44.98
GA	23.23	32.10	28.52	45.61	16.19	43.69		18.24	50.21	27.41
HU	31.83	38.42	29.77	31.17	36.68	30.94	17.59		30.42	25.86
IT	47.38	49.68	47.65	64.96	33.03	64.87	43.42	34.39		45.65
SV	41.20	50.48	47.16	51.93	36.46	51.07	37.76	40.48	55.65	

Table 3: Delexicalized models tested with gold PoS labels across languages.

Δ LAS	CS	DE	EN	ES	FI	FR	GA	HU	IT	SV
CS		-9.30	-7.73	-10.27	-7.17	-8.53	-8.85	-4.36	-10.59	-4.05
DE	-6.69		-6.22	-7.28	-6.62	-5.18	-7.77	-8.22	-5.26	-5.09
EN	-3.94	-5.93		-8.42	-5.37	-6.27	-6.99	-2.87	-7.96	-4.87
ES	-3.99	-7.05	-5.46		-4.58	-5.59	-7.28	-4.63	-4.86	-2.31
FI	-2.47	-7.72	-3.94	-3.80		-1.70	-5.39	-5.68	-1.59	-2.28
FR	-4.24	-7.62	-5.24	-7.68	-4.95		-9.50	-4.73	-7.61	-3.51
GA	-2.15	-2.38	-1.42	-6.91	-2.25	-3.57		-3.12	-7.13	-3.01
HU	-2.81	-5.29	-3.14	-2.50	-5.63	-1.64	-2.41		-2.05	-1.62
IT	-8.81	-7.15	-6.19	-6.98	-5.33	-5.84	-8.61	-8.08		-3.98
SV	-2.64	-10.18	-6.13	-14.78	-3.12	-13.11	-10.83	-6.68	-14.09	

Table 4: LAS differences of delexicalized models tested with **predicted** PoS labels across languages compared to gold PoS labels (shown in Table 3).

however, to be applied to other languages with the same universal features they are trained on. Figure 1 illustrates the general idea behind delexicalized parsing across languages and Table 3 lists the LAS's of applying our models across languages with the UD data set.

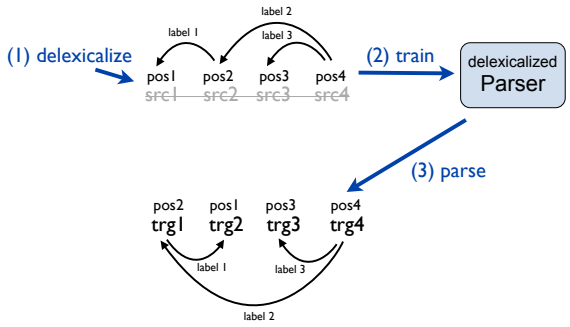

Figure 1: Delexicalized models applied across languages.

The results show that delexicalized models are quite robust across languages, at least for closely related languages like Spanish and Italian, but also for some languages from different language subfamilies such as English and French. The situation is, of course, much worse for distant languages and small training data sets such as Irish models applied to Finnish or Hungarian. Those models are essentially useless. Nevertheless, we can see the positive effect of universal annotation and harmonized annotation guidelines.

However, as argued earlier, we need to evaluate the performance of such models in real-world scenarios which require automatic annotation of PoS labels. Therefore, we used the same tagger models from the previous section to annotate the test sets in each language and parsed those data sets with our delexicalized models across languages. The LAS difference to the gold standard evaluation are listed in Table 4.

With these experiments, we can basically confirm the findings on monolingual parsing, namely that the performance drops significantly with predicted PoS labels. However, there is quite a variation among the language pairs. Models that have been quite bad to start with are in general less effected by the noise of the tagger. LAS reductions up to 14 points are certainly very serious and most models go down to way below 50% LAS. Note that we still rely on PoS taggers that are actually trained on manually verified data sets with over 90% accuracy which we cannot necessarily assume to find for low resource languages.

In the next section, we will look at annotation projection as another alternative for cross-lingual

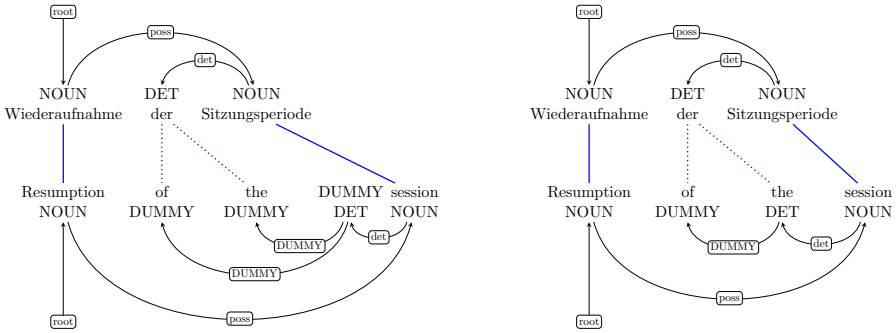

Figure 2: Reduced number of dummy labels in annotation projection as suggested by Tiedemann (2014) (bottom) compared to DCA of Hwa et al. (2005) (top).

parsing using the same setup.

5 Annotation Projection

In annotation projection, we rely on sentence aligned parallel corpora, so-called bitexts. The common setup is that source language data is parsed with a monolingually trained parser and the automatic annotation is then transfered to the target language by mapping labels through word alignment to corresponding target language sentences. The process is illustrated in Figure 3.

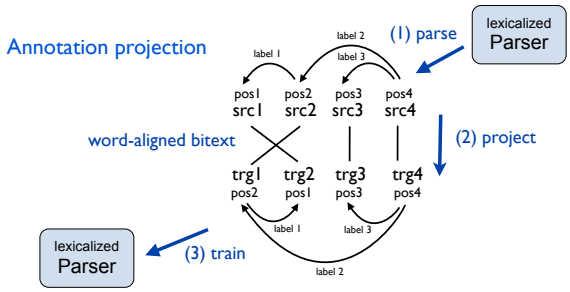

Figure 3: An illustration of annotation projection for cross-lingual dependency parsing.

There are several issues that need to be considered in this approach. First of all, we rely on noisy annotation of the source language which is usually done on out-of-domain data depending on the availability of parallel corpora. Secondly, we require accurate word alignments which are, however, often rather noisy when created automatically especially for non-literal human translations. Finally, we need to define heuristics to treat ambiguous alignments that cannot support one-to-one annotation projection. In our setup, we follow the suggested strategies of Tiedemann (2014), which are based on the projection heuristics proposed by Hwa et al. (2005). The data set that we use is a subset of the parallel Europarl corpus (version 7) which is a widely accepted data set primarily used in statistical machine translation (Koehn, 2005). We use a sample of 40,000 sentences for each language pair and annotate the data with our monolingual source language parsers presented in section 2. For the alignment, we use the symmetrized word alignments that are provided from OPUS (Tiedemann, 2012) that are created with standard statistical alignment tools such as Giza++ (Och and Ney, 2003) and Moses (Koehn et al., 2007). Our projection heuristics follow the direct correspondence assumption (DCA) algorithm of Hwa et al. (2005) but also apply the extensions proposed by Tiedemann (2014) that reduce the number of empty nodes and dummy labels. Figure 2 illustrates the effect of these extensions.

Applying the annotation projection strategy, we obtain the parsing results shown in Table 5. For each language pair, we use the same procedure and the same amount of data taken from Europarl (40,000 sentences).[3]

From the results, we can see that we beat the delexicalized models by a large margin. Some of the language pairs achieve LAS of above 70 which is quite a remarkable result. However, good results are in general only possible for closely related languages such as Spanish, Italian and French whereas more distant languages struggle more (see, for example Czech and Hungarian). For the latter, there is also a strong influence of the rich morphology which is not well supported by the projected information (we only project universal PoS tags and cross-lingually harmonized dependency relations). The results in Table 5 reflect the scores on gold

[3]Unfortunately, we have to leave out Irish as there is no data available in the same collection. The original treebank is, however, so small that the results are not very reliable for this language anyway.

LAS	CS	DE	EN	ES	FI	FR	HU	IT	SV
CS		50.20	47.96	49.17	49.58	46.48	39.34	49.24	46.38
DE	55.08		55.96	63.49	46.90	65.22	48.70	65.40	52.94
EN	57.70	63.31		65.07	48.86	67.48	49.14	68.69	54.01
ES	59.95	60.17	54.02		48.57	66.18	50.09	70.40	50.05
FI	54.67	47.06	45.69	42.37		40.56	41.72	43.06	44.03
FR	58.65	63.75	58.14	69.33	48.61		50.39	70.22	52.56
HU	46.58	48.79	41.07	48.97	40.08	48.23		51.64	38.87
IT	56.80	56.92	52.03	65.76	46.39	64.88	46.42		51.16
SV	51.71	56.37	50.46	59.06	44.51	60.39	46.86	65.15	

Table 5: Cross-lingual parsing with projected annotation (dependency relations and coarse PoS tags). Evaluation with gold PoS labels.

	CS	DE	EN	ES	FI	FR	HU	IT	SV
CS		-4.55	-2.04	-2.34	-2.48	-2.18	-3.71	-1.83	-1.87
DE	-0.71		-2.05	-2.18	-2.51	-1.74	-2.53	-2.15	-2.09
EN	-0.65	-4.43		-2.45	-2.59	-0.92	-2.57	-2.12	-2.18
ES	-1.02	-4.07	-2.08		-2.22	-1.18	-2.79	-1.75	-2.40
FI	-0.54	-3.83	-1.61	-1.41		-1.73	-3.41	-1.72	-1.85
FR	-0.84	-4.01	-2.21	-3.15	-2.70		-3.05	-1.95	-2.14
HU	-0.49	-2.63	-1.15	-1.61	-1.90	-1.67		-1.60	-1.39
IT	-0.77	-3.89	-1.96	-2.45	-3.40	-1.62	-3.78		-1.88
SV	-0.65	-3.53	-1.92	-1.63	-1.98	-2.12	-3.74	-1.43	

	CS	DE	EN	ES	FI	FR	HU	IT	SV
CS		-20.13	-14.71	-12.38	-12.73	-14.07	-14.50	-17.94	-6.37
DE	-15.20		-13.23	-10.34	-13.89	-9.25	-15.53	-9.97	-2.28
EN	-14.60	-14.53		-8.38	-10.85	-8.08	-11.75	-6.50	-0.63
ES	-16.81	-13.12	-10.53		-9.29	-7.22	-17.39	-6.78	-0.96
FI	-24.09	-23.01	-16.81	-18.87		-16.05	-16.55	-20.57	-8.55
FR	-16.13	-12.29	-11.12	-7.52	-11.55		-17.51	-5.79	-1.14
HU	-19.68	-22.76	-15.85	-22.15	-12.61	-20.71		-23.70	-12.15
IT	-17.03	-13.20	-10.18	-9.78	-11.99	-8.37	-15.48		-3.93
SV	-12.08	-10.17	-3.56	-7.00	-6.71	-6.71	-20.73	-10.13	

Table 6: Cross-lingual parsing with **predicted** PoS labels with PoS tagger models trained on verified target language treebanks (left table) and models trained on **projected** treebanks (right table). Differences in LAS compared to the results with gold PoS labels from Table 5.

standard data and the same question as before applies here: What is the drop in performance when replacing gold PoS labels with predicted ones? The answer is in Table 6 (left part). Using automatic annotation leads to substantial drops for most language pairs as expected. However, we can see that the lexicalized models trained through annotation projection are much more robust than the delexicalized transfer models presented earlier. With the drop of up to 3 LAS we are still rather close to the performance on gold annotation.

	CS	DE	EN	ES	FI	FR	HU	IT	SV
CS		70.49	67.59	71.64	79.23	71.47	67.87	72.85	80.96
DE	79.29		74.77	81.36	74.68	83.22	75.06	84.65	80.24
EN	79.22	82.24		83.04	75.08	83.49	76.81	86.97	81.52
ES	79.47	80.03	75.58		75.33	87.86	76.04	90.41	81.58
FI	72.13	62.76	63.03	57.17		58.57	64.76	57.29	69.82
FR	80.99	82.10	76.92	88.26	76.36		76.00	92.41	82.87
HU	70.08	66.48	63.64	66.24	69.45	68.04		67.83	69.43
IT	79.80	80.77	75.14	86.50	75.27	87.37	74.82		80.80
SV	81.25	77.84	74.85	83.39	77.07	83.34	67.97	83.80	

Table 7: Coarse PoS tagger accuracy on test sets from the universal dependencies data set with models trained on projected bitexts.

The experimental results in Table 6 rely on the availability of taggers trained on verified target language annotations. Low resource language may not even have resources for this purpose and, therefore, it is interesting to know if we can even learn PoS taggers from the projected data sets as well. In the following setup, we trained models on the projected data for each language pair to test this scenario. Note that we had to remove all dummy labels and tokens that may appear in the projected data. This procedure certainly corrupts the training data even further and the PoS tagging quality is effected by this noise (see Table 7). Applying cross-lingual parsers trained on the same projected data results in the scores shown in the right part of Table 6. Here, we can see that the models are seriously effected by the low quality provided by the projected PoS taggers. The LAS drops dramatically making any of these models completely useless. This result is, unfortunately, not very encouraging and shows the limitations of direct projection techniques and the importance of proper linguistic knowledge in the target language. Note that we did not spend any time on optimizing projection techniques of PoS annotation but we expect similar drops even with slightly improved cross-lingual methods.

6 Treebank Translation

The possibility of translating treebanks as another strategy for cross-lingual parsing has been proposed by Tiedemann et al. (2014). They apply

LAS	CS	DE	EN	ES	FI	FR	HU	IT	SV
CS		50.37	45.84	49.81	47.36	44.72	36.66	49.53	46.24
DE	55.06		55.89	64.88	42.29	63.95	46.68	66.17	51.76
EN	52.47	61.98		67.20	44.51	67.50	41.58	69.28	56.16
ES	60.40	57.69	54.62		42.60	68.67	30.35	72.39	51.51
FI	49.56	42.98	46.50	36.11		35.39	39.19	37.22	41.45
FR	57.35	61.33	58.12	71.15	42.60		40.33	72.84	51.58
HU	39.89	42.72	38.51	43.16	39.93	39.91		41.74	34.26
IT	58.20	55.60	53.26	68.74	41.95	68.19	39.74		50.62
SV	47.89	55.07	52.86	59.80	42.23	60.64	41.98	66.19	

Table 8: Cross-lingual parsing with translated treebanks; evaluated with gold PoS labels.

	CS	DE	EN	ES	FI	FR	HU	IT	SV
CS		-4.14	-1.72	-1.74	-2.45	-0.90	-3.38	-1.72	-2.42
DE	-0.73		-1.88	-2.54	-1.82	-1.46	-2.53	-2.22	-2.21
EN	-0.48	-4.41		-2.72	-2.85	-0.95	-1.84	-2.00	-2.77
ES	-1.03	-3.51	-2.25		-2.60	-1.22	-1.87	-2.36	-2.31
FI	-0.51	-4.37	-1.99	-1.66		-0.99	-2.68	-1.74	-1.84
FR	-0.98	-3.87	-2.25	-3.45	-2.25		-1.69	-2.11	-1.88
HU	-0.46	-2.73	-1.56	-2.09	-2.39	-0.58		-1.47	-1.57
IT	-0.90	-3.76	-2.55	-2.64	-2.58	-1.81	-2.20		-2.19
SV	-0.50	-3.51	-2.13	-2.39	-2.27	-1.68	-2.42	-1.88	

	CS	DE	EN	ES	FI	FR	HU	IT	SV
CS		-17.74	-11.71	-9.79	-8.65	-10.65	-10.68	-13.23	-4.38
DE	-10.57		-11.25	-11.58	-10.28	-8.55	-11.96	-10.26	-1.46
EN	-13.68	-14.60		-11.02	-8.15	-9.75	-13.54	-10.03	-0.63
ES	-14.91	-11.15	-9.76		-7.86	-6.03	-8.88	-5.62	-2.37
FI	-14.57	-15.92	-14.78	-9.25		-10.88	-12.33	-10.28	-2.15
FR	-14.23	-10.50	-8.72	-7.38	-6.79		-14.27	-4.60	-2.35
HU	-15.29	-15.67	-14.99	-17.35	-13.51	-16.14		-16.19	-9.48
IT	-14.21	-12.07	-8.73	-6.92	-8.24	-5.47	-14.24		-2.04
SV	-7.62	-9.75	-4.44	-8.54	-6.86	-8.80	-19.30	-10.01	

Table 9: Cross-lingual parsing with translated treebanks and **predicted** PoS labels with PoS tagger models trained on verified target language treebanks (left table) and models trained on **projected** treebanks (right table). Differences in LAS compared to the results with gold PoS labels from Table 8.

phrase-based statistical machine translation to the universal dependency treebank (McDonald et al., 2013) and obtain encouraging results. We use a similar setup but apply it to the UD data set testing the approach on a wider range of languages. We follow the general ideas of Tiedemann (2014) and the projection heuristics described there. Our translation models apply a standard setup of a phrase-based SMT framework using the default training pipeline implemented in Moses as well as the Moses decoder with standard settings for translating the raw data sets. We consequently use Europarl data only for all models including language models and translation models. For tuning, we apply 10,000 sentences from a disjoint corpus of movie subtitles taken from OPUS (Tiedemann, 2012). We deliberately use these out-of-domain data sets to tune model parameters in order to avoid domain overfitting. A mixed-domain set would certainly have been even better for this purpose but we have to leave a closer investigation of this effect on treebank translation quality to future work. Similar to the projection approach, we have to drop Irish as there is no training data in Europarl for creating our SMT models.

Translating treebanks can be seen as creating synthetic parallel corpora and the same projection heuristics can be used again to transfer annotation to the target language. The advantage of the approach is that the source language annotation is given and manually verified and that the word alignment is an integral part of statistical machine translation. The general concept of treebank translation is illustrated in Figure 4.

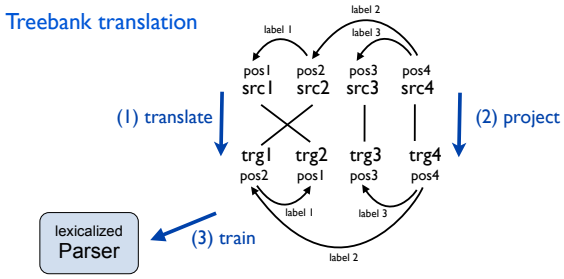

Figure 4: Translating treebanks to project syntactic information.

Applying this approach to the UD data results in the outcome summarized in Table 8. With these experiments, we can confirm the basic findings of related work, i.e. that treebank translation is a valuable alternative to annotation projection on existing parallel data with comparable results and some advantages in certain cases. In general, we can see that more distant languages are worse again mostly due to the lower quality of the basic translation model for those languages.

Similar to the previous approaches, we now test our models with predicted PoS labels. The left part in Table 9 lists the LAS differences when replacing gold annotation with automatic tags. Similar to the annotation projection approach, we can observe drops of around 2 LAS with up to over 4 LAS in some cases. This shows again, that the lexicalized models are much more robust than delexicalized ones and should be preferred when applied in real-world applications.

	CS	DE	EN	ES	FI	FR	HU	IT	SV
CS		72.17	68.80	73.81	80.28	73.72	72.02	77.36	83.27
DE	82.97		77.80	82.65	73.28	84.05	77.23	86.20	81.54
EN	78.84	83.69		83.88	77.21	84.60	74.15	87.04	84.66
ES	82.17	82.56	78.36		76.47	90.66	71.95	92.31	83.00
FI	78.25	67.09	66.70	60.67		61.05	70.80	60.06	72.11
FR	82.02	82.76	78.46	89.23	77.76		75.27	93.52	83.00
HU	71.74	67.62	63.44	65.98	69.35	66.20		68.20	67.97
IT	83.06	81.57	78.50	89.81	76.49	91.80	75.65		83.13
SV	84.62	78.53	75.98	83.97	76.80	83.66	68.74	84.20	

Table 10: Coarse PoS tagger accuracy on test sets from the universal dependencies data set with models trained on translated treebanks.

Finally, we also look at tagger models trained on projected treebanks as well (see Table 10). The parsing results on data sets that have been annotated with those taggers are shown on the right-hand side in Table 9. Not surprisingly, we observe significant drops again in LAS and, similar to annotation projection, all models are seriously damaged by the noisy annotation. Nevertheless, the difference is relatively smaller in most cases when compared to the annotation projection approach. This points to the advantage of treebank translation that makes annotation projection more straightforward due to the tendency of producing rather literal translations that are more straightforward to align than human translations. Surprising is especially the performance of the cross-lingual models from German, English and Italian to Swedish which perform better with projected PoS taggers than with monolingually trained ones. This is certainly unexpected and deserves some additional analyses. Overall, the results are still very mixed and further studies are necessary to investigate the projection quality depending on the cross-lingual parsing approach in more detail.

7 Discussion

Our results illustrate the strong impact of PoS label accuracy on dependency parsing. Our projection techniques are indeed very simple and naive. The performance of the taggers drops significantly when training models on small and noisy data sets such as the projected and translated treebanks. There are techniques that improve cross-lingual PoS tagging using a combination of projection and unsupervised learning (Das and Petrov, 2011). These techniques certainly lead to better parsing performance as shown by McDonald et al. (2011b). Another alternative would be to use the recently proposed models for joint word alignment and annotation projection (Östling, 2015). A thorough comparison with those techniques is beyond the scope of this paper but would also not contribute to the point we would like to make here. Furthermore, looking at the actual scores that we achieve with our directly projected models (see Tables 7 and 10), we can see that the PoS models seem to perform reasonably well with many of them close or above 80% accuracy, which is on par with the advanced models presented by Das and Petrov (2011).

In any case, the main conclusion from our experiments is that reliable PoS tagging is essential for the purpose of dependency parsing especially across languages. To further stress this outcome, we can look at the correlation between PoS tagging accuracy and labeled attachment scores. Figure 5 plots the scores we obtain with our naive direct projection techniques. The graph clearly shows a very strong correlation between both evaluation metrics on our data sets.

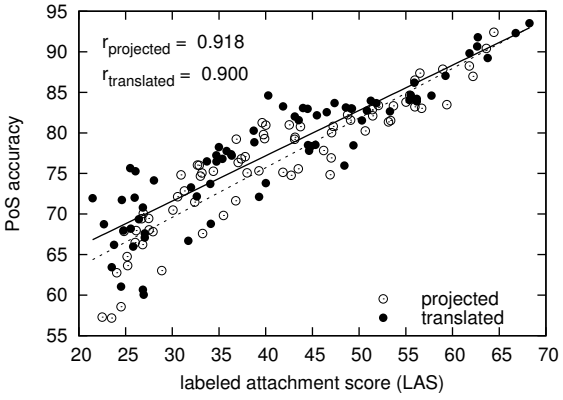

Figure 5: Correlation between PoS tagger accuracy and cross-lingual parsing performance.

Another interesting question is whether the absolute drops we observe in labeled attachment scores are also directly related to the PoS tagging performance. For this, we plot the difference between LAS on test sets with gold PoS labels and test sets with predicted labels in comparison to the PoS tag-

ger performance used for the latter (Figure 6). As we can see, even in this case we can measure a significant (negative) correlation which is, however, not as strong as the overall correlation between PoS tagging and LAS.

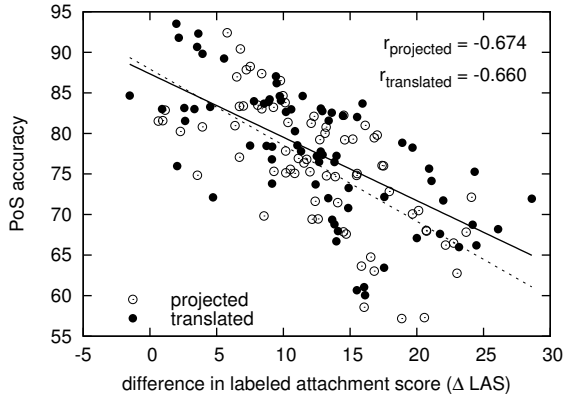

Figure 6: Correlation between PoS tagger accuracy and the **drop** in cross-lingual parsing performance.

Looking at these outcomes, it seems wise to invest some effort in improving PoS tagging performance before blindly trusting any cross-lingual approach to statistical dependency parsing. Hybrid approaches that rely on lexical information, unsupervised learning and annotation projection might be a good strategy for this purpose. Another useful framework could be active learning in which reliable annotation can be created for the induction of robust parser models. We will leave these ideas to future work.

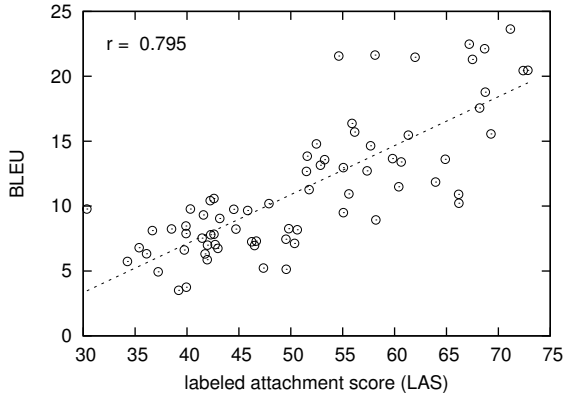

Figure 7: Correlation between translation performance (measured in BLEU) and cross-lingual parsing performance.

Finally, we can also have a look at the correlation between translation performance and cross-lingual parsing. Figure 7 plots the BLEU scores that we obtain on an out-of-domain test set (from the same subtitle corpus we used for tuning) for the phrase-based models that we have trained on Europarl data compared to the labeled attachment scores we achieve with the corresponding models trained on translated treebanks. The figure illustrates a strong correlation between the two metrics even though the results need to be taken with a grain of salt due to the domain mismatch between treebank data and SMT test data, and due to instabilities of BLEU as a general measure of translation performance. Interesting to see is that we obtain competitive results with the translation approach when compared to annotation projection even though the translation performance is really poor in terms of BLEU. Note, however, that the BLEU scores are in general very low due to the significant domain mismatch between training data and test data in the SMT setup.

8 Conclusions

This paper presents a systematic comparison of cross-lingual parsing based on delexicalization, annotation projection and treebank translation on data with harmonized annotation from the universal dependencies project. The main contributions of the paper are the presentations of cross-lingual parsing baselines for this new data set and a detailed discussion about the impact of predicted PoS labels and morphological information. With our empirical results, we demonstrate the importance of reliable features, which becomes apparent when testing models trained on noisy naively projected data. Our results also reveal the serious shortcomings of delexicalization in connection with cross-lingual parsing. Future work includes further investigations of improved annotation projection of morphosyntactic information and the use of multiple languages and prior knowledge about linguistic properties to improve the overall results of cross-lingual dependency parsing. The use of abstract cross-lingual word representations and other target language adaptations for improved model transfer are other ideas that we would like to explore. We would also like to emphasize truly under-resourced languages in further experiments that would require new data sets and manual evaluation. In connection with this we also need to focus on improved models for distant languages that exhibit significant differences in their syntax. Our experiments presented in this paper reveal already that the ex-

isting approaches to cross-lingual parsing have severe shortcomings for languages from different language families. However, we are optimistic that new techniques with stronger target language adaptation and improved transfer mechanisms will be able to support even those cases. In order to show this, we will look at downstream applications that can demonstrate the utility of cross-lingual parsing in other areas of NLP and end-user systems.

References

Bernd Bohnet and Jonas Kuhn. 2012. The Best of Both Worlds – A Graph-based Completion Model for Transition-based Parsers. In *Proceedings of EACL*, pages 77–87.

Bernd Bohnet. 2010. Top Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of COLING*, pages 89–97.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of CoNLL*, pages 149–164.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of ACL*, pages 600–609.

Marie-Catherine De Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of LREC*, pages 449–454.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping Parsers via Syntactic Projection across Parallel Texts. *Natural Language Engineering*, 11(3):311–325.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher J. Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL*, pages 177–180.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit*, pages 79–86.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. *Dependency Parsing*. Morgan & Claypool Publishers.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011a. Multi-Source Transfer of Delexicalized Dependency Parsers. In *Proceedings of EMNLP*, pages 62–72.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011b. Multi-source transfer of delexicalized dependency parsers. In *Proceedings EMNLP*, pages 62–72.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of ACL*, pages 92–97.

Thomas Müller and Hinrich Schütze. 2015. Robust morphological tagging with word representations. In *Proceedings of NAACL*.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective Sharing for Multilingual Dependency Parsing. In *Proceedings of ACL*, pages 629–637.

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 CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–52.

Robert Östling. 2015. *Bayesian Models for Multilingual Word Alignment*. Ph.D. thesis, Stockholm University, Department of Linguistics.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A Universal Part-of-Speech Tagset. In *Proceedings of LREC*, pages 2089–2096.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual Word Clusters for Direct Transfer of Linguistic Structure. In *Proceedings of NAACL*, pages 477–487.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target Language Adaptation of Discriminative Transfer Parsers. In *Proceedings of NAACL*, pages 1061–1071.

Jörg Tiedemann, Željko Agić, and Joakim Nivre. 2014. Treebank Translation for Cross-Lingual Parser Induction. In *Proceedings of CoNLL*, pages 130–140.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of LREC*, pages 2214–2218.

Jörg Tiedemann. 2014. Rediscovering Annotation Projection for Cross-Lingual Parser Induction. In *Proceedings of COLING*, pages 1854–1864.

Min Xiao and Yuhong Guo. 2014. Distributed Word Representation Learning for Cross-Lingual Dependency Parsing. In *Proceedings of CoNLL*, pages 119–129.

Daniel Zeman and Philip Resnik. 2008. Cross-Language Parser Adaptation between Related Languages. In *Proceedings of IJCNLP*, pages 35–42.