Bits and Pieces: Investigating the Effects of Subwords in Multi-task Parsing Across Languages and Domains

Daniel Dakota, Sandra Kübler
Indiana University
{ddakota,skuebler}@iu.edu

Abstract

Neural parsing is very dependent on the underlying language model. However, very little is known about how choices in the language model affect parsing performance, especially in multi-task learning. We investigate questions on how the choice of subwords affects parsing, how subword sharing is responsible for gains or negative transfer in a multi-task setting where each task is parsing of a specific domain of the same language. More specifically, we investigate these issues across four languages: English, German, Italian, and Turkish. We find a general preference for averaged or last subwords across languages and domains. However, specific POS tags may require different subwords, and the distributional overlap between subwords across domains is perhaps a more influential factor in determining positive or negative transfer than discrepancies in the data sizes.

Keywords: neural parsing, subwords, multi-task learning, domain adaptation

1. Introduction

Multi-task learning (MTL) is one of the major approaches to handle domain adaptation issues. In many cases, both tasks are very similar, the only difference being the domains from which the data were sampled. The issue of negative transfer is an established issue in transfer learning paradigms (Rosenstein et al., 2005), including MTL, but identifying the causes of negative transfer is still an open research question. Strategies such as saliency (Li et al., 2016) and attention (Vaswani et al., 2017) have started to provide a great deal of insight into the internal working of models, but have predominantly benefited classification tasks. For structured prediction tasks, such as dependency parsing, there are many open questions about how information is shared and transferred, particularly in transfer learning approaches. In a multi-task setting (Caruana, 1997), the issue is compounded since tasks are being learned jointly rather than sequentially; consequently, they are able to simultaneously positively and negatively impact one another to different levels (Wu et al., 2020), and the impact on various tasks is still uncertain.

To investigate possible areas of transfer, we focus on the representation of subword information in MTL parsing where each task corresponds to one genre of a given language. Subwords in particular may represent words in non-intuitive ways, sometimes yielding unexpected regularities, particularly for words outside of model's base vocabulary and domain, and can be very sensitive to slight variations. For example, while the adjective *amazing* is not split into subwords, the informal spelling *amazin* is represented as am-##azi-##n by `bert-base-cased`. In this case, selecting the first subword "am" may conflict with selecting the actual word "am", which represents a different part of speech (POS). These representations may interact with annotations or linguistic characteristics (since subwords do not align with morphology), which may subsequently improve or degrade parsing performance in an MTL setting.

We perform a set of experiments in which we pair two treebanks of the same language in an MTL dependency parsing architecture, where the two tasks correspond to two genres of this language. We use English, German, Italian, and Turkish[1]. Results suggest using embeddings from either the average over all subwords or selecting only the last subword tends to yield better performance across languages and domains than selecting only the first which tends to be the default. Additionally, while treebank sizes play a role in performance gains in MTL parsing, the overall overlap of the subwords between treebanks tends to be a better indicator 1) wrt. to the source of possible positive and negative transfer, irrespective of treebank size, and 2) wrt. how an individual treebank may improve or degrade in performance.

2. Related Work

2.1. Embeddings and Language Models

With the predominant use of embeddings and pretrained language models (PLMs), a simple approach for many tasks is to finetune a PLM on the target task, though this often results in subpar performance if there are domain shifts (Ma et al., 2019).

[1]This selection was mainly based on availability of two UD treebanks in the same language.

LREC-COLING 2024, pages 2397–2409
20-25 May, 2024.

This can be partially alleviated with additional pre-training on domain specific data (Ma et al., 2019; Gururangan et al., 2020; Rietzler et al., 2020).

Parsing has traditionally focused on methods for selecting the best source data for a given target domain (Plank and van Noord, 2011; McDonald et al., 2011; Rosa and Žabokrtský, 2015; Falenska and Çetinoğlu, 2017). However, the full potential of PLMs in domain adaptation is still an evolving avenue of research, particularly in parsing where the target and source domains may be distinct from those of the language model (Dakota, 2021), adding complexity. While contextualized embeddings have reduced the gap between domains in constituency parsing, additional strategies are necessary for more distant domains (Joshi et al., 2018), and out-of-domain performance has not shown the expected gains (Fried et al., 2019). Treebank embeddings (Stymne et al., 2018; Wagner et al., 2020), domain embeddings (Li et al., 2019), and fine-tuned contexualized embeddings with adversarial methods (Li et al., 2020) have all shown benefits in dependency parsing. Such observations have resulted in a great deal of diversification of the data used to create language models for various languages (Virtanen et al., 2019; Cui et al., 2020; Martin et al., 2020) and many now are domain specific (Alsentzer et al., 2019; Liu et al., 2020).

2.2. Multi-task Learning

MTL learns tasks jointly, and while problems may overlap with sequential transfer learning, new ones also arise. Much MTL work in domain-adaption has focused on which layers to share; lower layer sharing is often found to be optimal (Søgaard and Goldberg, 2016; Peng and Dredze, 2017). The relationship between data sizes of individual tasks was shown to influence performance across a range of tasks (Luong et al., 2015; Benton et al., 2017; Augenstein and Søgaard, 2017; Schulz et al., 2018). To mitigate negative influence, Dakota et al. (2021) used weighted multi-task learning to weight the contribution of each task on the loss function between imbalanced domains, which improved dependency parsing performance on both the target and source treebanks.

2.3. Subword Representations

Choices in tokenization and the resultant subword representations have been shown to have a noticeable effect on performance in morphology (Church, 2020; Klein and Tsarfaty, 2020; Hofmann et al., 2021) and POS tagging (Blaschke et al., 2023), but for parsing, the situation is less clear (Kitaev et al., 2019). The efficacy of subword representations generated from internal PLM tokenizers has shown varied results in cross-lingual and multilingual settings (Pires et al., 2019). While subword overlap between a target and source language is suggested to be a good indicator for selecting sources (Wu and Dredze, 2019), many subwords may not be relevant or missing all together (Wu and Dredze, 2020). Domain specific vocabulary augmentation, as performed by Sachidananda et al. (2021), yielded competitive performance across various classification tasks compared to additional domain specific model training, at a fraction of the computational resources. For zero-shot POS tagging, Blaschke et al. (2023) found that the absolute difference between the proportion of words split into subwords in the source and target is a strong indicator of performance of a pre-trained language model on non-standard varieties of a language. For English constituency parsing, the choice of first or last subwords did not appear to make a substantial performance difference (Kitaev et al., 2019).

3. Research Questions

Choosing which subwords to use to represent words is an important choice for parsing. If we use the first subword, we may retain information about prefixes or stems; if we use the last subword, we may retain suffix information. In standard transfer learning, subword disparities between the source and target can be more easily handled as there is only one learning objective at a time. In an MTL setting across domains, however, the situation becomes more difficult, since subwords of one domain also simultaneously influence parsing of the second domain. Since there are lexical and syntactic differences across domains, the best subword representation may have to be chosen based on robustness across domains rather than based on coverage within one domain.

We investigate the following questions:

1. How does the selection of different subword tokens impact parsing performance in an MTL parsing?

2. Do certain subword representations show more consistency in performance across languages and domains?

3. What is the interaction between treebank sizes and subword overlap wrt. negative or positive transfer?

4. Methodology

Data We use Universal Dependencies (UD) v2.11 treebanks (Nivre et al., 2020; de Marneffe et al.,

Lang.	Treebank	Train	Dev	Test
EN	Tweebank	1 639	710	1201
	EWT eql	1 600	700	1 200
	EWT full	12 543	2 001	2 077
DE	tweeDe	1 000	150	151
	GSD eql	1 000	150	150
	GSD full	13 814	799	977
IT	Twittirò	1 138	144	142
	ISDT eql	1 100	150	150
	ISDT full	13 121	564	482
TR	BOUN	7 803	600	979
	Penn eql	7 800	622	924
	Penn full	14 850	622	924

Table 1: Number of sentences per treebank across splits.

Hyperparameters	**Value**
Bert Mapping Dimension	100
Number of BERT Layers Used	4
Number of LSTM Layers	3
LSTM Hidden Layer Dimension	400
Optimizer	Adam
Patience	100
Batch Size	20k tokens
Learning Rate	2e-3
Seeds	10, 20, 30

Table 2: Hyperparameter settings for multiparser.

2021). More specifically, we use three Twitter specific treebanks: the German tweeDe treebank (Rehbein et al., 2019), the English Tweebank v2[2] (Liu et al., 2018) and the Italian Twittirò (Cignarella et al., 2019). Each is paired with a UD treebank of a different domain: German-GSD (news, reviews, and wikipedia), Italian-ISDT (news, legal, and wikipedia), and English-EWT (blogs, emails, reviews, social media, and web). In addition, we use the Turkish BOUN treebank (biographical texts, national newspapers, instructional texts, popular culture articles, and essays) and the Turkish Penn treebank (translated newspaper data). We recognize that BOUN is less genre specific in comparison to the Twitter treebanks, but we are limited wrt. availability of language specific treebank resources.

To control for data imbalance, we also use a setting where we reduce the larger treebank of a language to the size of the corresponding smaller one, by selecting the first N sentences (to ensure replicability). For an overview of languages and data sizes, see Table 1.

Parser We use multiparser (Sayyed and Dakota, 2021), which is an extension of the biaffine graph-based neural dependency parser (Dozat and Manning, 2017; Dozat et al., 2017) but includes a multi-task architecture. We allow treebanks to share the MLP layers and modify the base architecture to allow the use of embeddings produced by the various PLMs as input. We use language specific BERT models and their tokenizers[3]. We have further modified multiparser to include options modifying which subword pieces are utilized to generate the token level embeddings: the first subword, the average over all subwords, or the last subword (see Table 2 for hyperparameters). The final embeddings are generated via a scalar mix (Peters et al., 2018; Tenney et al., 2019a,b) of the last four layers. Words consisting of one subword token will have the same representation in all versions.

The parser also provides an option to weight the influence of each task on the loss function. For weighted settings, the Twitter treebank is given a weight of 0.1 and the non-Twitter treebank 0.9 following Dakota et al. (2021). For Turkish, BOUN is given 0.33 and Penn 0.67 to reflect their relative sizes.

Evaluation We perform each experiment with three random initializations and report averages, using the CoNLL2018 scorer (Zeman et al., 2018) for UAS and LAS (punctuation is included in the evaluation), as well as scores for label accuracy (LA) per POS tag. We focus on LA because the choice of subword representations has the most direct influence on the dependency labels, which is reflected in LA. Note that we use gold POS tags only for calculating the LA scores, the parser is not given any POS tags. We generally show POS tags that occur more than 50 times in the smaller corpus of a language. We highlight *major* differences between highest (blue) and lowest results (red) per POS tag when the difference is greater than 2.5 points within either the equal or the full setting.

5. General Trends Across Languages

For each language, we compare 1) the single task setting (STL; i.e., training and testing on the same domain) with MTL (training on both domains and testing on one); 2) two data sizes (equal size of training sets, or full sets); 3) for full sets, we use an unweighted or weighted setting, the latter gives more weight to the majority task.

We first discuss the big picture across languages. In the following sections, we then discuss results for individual languages, followed by a closer look

[2]https://github.com/Oneplus/Tweebank

[3]Models are accessed via Hugging Face (Wolf et al., 2020) and are bert-base-cased (English), dbmdz/bert-base-german-cased (German), dbmdz/bert-base-italian-cased (Italian), and dbmdz/bert-base-turkish-cased (Turkish).

Tweebank		NOUN	VERB	ADJ	PROPN	ADV	INTJ	ADP	SCONJ	AUX	DET	PRON	UAS	LAS
subw.	counts	2669	1985	955	1640	834	218	1189	209	919	826	1716		
first	STL	71.14	66.31	72.88	68.43	81.14	71.71	91.84	79.27	88.54	95.84	90.52	78.70	73.42
	MTL eql.	73.31	69.57	75.99	69.23	83.01	70.34	92.37	84.85	89.66	96.37	91.20	80.50	75.52
avg.	STL	72.06	67.69	73.65	70.39	79.98	68.96	92.54	80.22	88.61	96.29	90.60	79.32	74.19
	MTL eql.	74.45	70.26	77.24	70.06	82.05	68.20	92.63	83.41	90.13	96.45	91.41	81.07	76.25
last	STL	71.39	67.04	74.14	67.44	80.74	68.04	91.87	81.18	87.74	95.80	90.75	78.90	73.50
	MTL eql.	73.76	70.38	76.86	68.74	81.65	66.97	92.54	83.09	90.50	96.85	91.14	80.81	75.84
first	MTL full	74.87	71.00	76.65	70.91	83.33	70.49	93.69	85.33	91.55	96.89	91.92	81.61	76.91
	MTL full+W	75.03	72.26	77.35	71.06	82.57	68.04	93.44	85.81	92.17	97.30	92.44	81.60	77.01
avg.	MTL full	75.95	72.11	77.42	71.95	83.29	71.25	94.14	85.49	92.35	96.81	91.88	82.08	77.59
	MTL full+W	76.47	72.61	77.56	70.83	83.33	69.11	93.52	87.08	93.36	97.18	92.35	82.22	77.71
last	MTL full	75.27	71.55	77.59	69.63	83.09	70.18	93.78	85.01	91.91	96.73	91.69	81.76	76.97
	MTL full+W	76.22	73.33	78.32	70.08	84.25	69.88	93.64	87.08	92.02	96.81	92.02	82.24	77.61

Table 3: Results on English, testing on Tweebank. STL: single task baseline, MTL eql.: multi-task with equal size EWT, MTL full: MTL with full EWT, MTL full+W: MTL with full EWT and weights.

EWT		NOUN	VERB	ADJ	PROPN	ADV	INTJ	ADP	SCONJ	AUX	DET	PRON	UAS	LAS
subw.	counts	2452	1549	916	1524	565	77	1298	299	878	1132	1182		
first	STL eql.	80.57	82.35	85.59	76.16	84.66	42.86	94.30	89.63	94.04	99.15	91.68	84.04	79.27
	MTL eql.	80.82	83.49	86.39	76.60	86.25	71.86	95.69	87.63	94.68	99.47	92.89	84.46	80.38
avg.	STL eql.	81.14	82.83	85.88	77.54	85.31	46.75	94.27	90.08	94.65	99.32	91.77	84.27	79.72
	MTL eql.	81.89	83.99	87.19	77.82	86.43	74.03	95.87	89.63	95.14	99.38	92.78	85.12	81.27
last	STL eql.	80.98	82.70	86.57	76.90	84.31	44.16	94.84	89.19	94.72	99.26	91.74	84.10	79.67
	MTL eql.	81.55	83.86	87.08	77.34	85.55	68.40	95.40	88.52	94.91	99.32	92.81	84.92	80.81
	counts	4136	2639	1782	1985	1147	120	2030	443	1509	1898	2158		
first	STL full	87.23	88.08	89.08	83.26	87.94	83.33	96.27	94.81	96.84	99.17	95.58	89.61	86.64
	MTL full	86.81	87.66	88.76	82.35	87.74	84.44	96.21	93.00	96.80	99.21	95.15	89.21	86.20
	MTL full+W	87.05	88.44	88.93	83.27	88.46	85.83	96.37	93.45	96.97	99.17	95.69	89.62	86.69
avg.	STL eql.	87.61	88.62	90.18	82.54	88.06	84.44	96.68	94.06	96.80	99.10	95.86	89.86	86.98
	MTL full	87.02	87.92	89.41	81.58	87.42	84.17	96.06	92.10	96.75	99.26	95.10	89.21	86.30
	MTL full+W	87.69	88.15	89.64	83.29	88.61	85.00	96.47	93.75	96.84	99.30	95.60	89.77	86.94
last	STL full	87.10	88.33	89.88	82.87	88.75	82.78	96.75	94.51	96.75	99.19	95.55	89.67	86.86
	MTL full	86.88	87.50	88.96	81.48	87.79	83.61	96.32	92.85	96.62	99.17	95.06	89.09	86.17
	MTL full+W	87.43	88.52	90.07	82.94	88.17	85.83	96.95	93.83	96.71	99.14	95.72	89.64	86.91

Table 4: Results on English, testing on EWT. STL: single task baseline, MTL: multi-task setup, eql.: EWT equal in size to Tweebank, full: using the full EWT, full+W: using the full EWT and multi-task weights.

at the distributional overlap in subwords across domains/treebanks of a language.

When looking at the big picture, we see the following trends (based on the results in Tables 3–10): Rather unsurprisingly, the smaller treebanks tend to profit more from the MTL setting than the larger ones. When the larger treebanks are reduced to an equal size, they do profit more from MTL. However, the opposite trend holds for the Turkish treebanks, which indicates that that genre plays a role here. Our hypothesis is that more general genres or a mix of genres is a more useful addition for a very genre specific treebank (such as Twitter or news).

When we look at the larger treebank in the full setting, we see evidence of negative transfer, i.e., a small decrease in LAS and UAS when going from STL to non-weighted MTL, showing that the addition of more data (from a different domain) is not helpful. The decrease is distributed across most POS tags, which suggests that the negative transfer is equally distributed across the larger treebank; there is no single source of subword or annotation sharing from the smaller treebank that drives the performance loss. In the weighted MTL setting, the results are only minimally worse than those in the STL setting, or even minimally surpass the STL setting. The only exception is Turkish, where even the non-weighted MTL task outperforms the STL setting (see Section 9 for a more detailed discussion).

While the choice of these subword representations clearly has an impact on results and can cause negative transfer, it is difficult to determine factors for the preferences we observe. Overall averaged and last subwords perform more robustly across languages. However, the linguistically motivated hypothesis that morphologically richer languages, especially ones with a preference for suffixes, would clearly prefer last tokens is not borne out. Turkish would be expected to have a pronounced preference for last tokens, which is not the case, while German, being morphologically richer than English, should have a stronger preference for last or potentially averaged tokens, which is again not observed.

Work by Church (2020); Klein and Tsarfaty (2020); Hofmann et al. (2021) has all suggested that infusing morphological and linguistic information into subword representations used in PLMs can further enhance their performance on tasks. While it would not guarantee performance gains in MTL, it would potentially better align the morphological

information shared across domains and languages and is an area requiring additional investigation.

6. English Results

The results of the parsing experiments for English when testing on Tweebank are shown in Table 3[4]; the results for testing on EWT in Table 4.

When looking at the experiments on Tweebank in Table 3, we see that the MTL setting improves performance of the parser, especially in the equal setting (where Tweebank and EWT are balanced in size). Here, the main sources for improvement are nouns, verbs, adjectives, and adverbs, but also subordinating conjunctions, and auxiliaries. The only exception is interjections (ITJ), which show the opposite trend (LA in STL: 71.71 vs. LA in MTL eql.: 66.97). Using the full training set when evaluating on Tweebank plus all experiments evaluating on EWT show the same trend, but the differences in performance are less pronounced. For EWT, there are only two POS tags showing a major difference, for interjections in both the equal and full setting, and for subordinating conjunctions (SCONJ) in the full setting. Since Tweebank has more interjections than EWT, the multi-task setting allows the EWT task to improve. Interestingly, using the full EWT treebank (lowest LA: 83.33) is more effective than adding the Tweebank task (highest LA: 74.03) for this POS tag. For subordinating conjunctions, adding the Tweebank task decreases performance on EWT. We assume that this may be the case because conjunctions in Tweebank show more spelling variation (e.g., 'cuz', 'cos') than those in EWT.

Next, we look at the difference between the different subword tokens used in the experiments. Overall, we see that this setting influences the results. Using the first subword token to represent a word does not work as well as using either the average or the last one. When evaluating on Tweebank, using the average tokens in an MTL setting results in the highest LAS and UAS. The same is true for evaluating on EWT, but the differences are considerably smaller. In terms of LA, for most POS tags, average tokens work best, the exceptions are verbs and auxiliaries (AUX; in the equal setting), which prefer the last tokens, and subordinating conjunctions, which prefer the first token (in the equal setting).

For EWT, the only major differences are for interjections in both settings (they prefer averaged tokens in the equal setting and either first or last tokens in the full setting), and subordinating conjunctions in the full setting, where they prefer first tokens. Note that the highest LAS on EWT is reached by the single task full setting (86.98), showing that EWT cannot profit from having Tweebank added as a second task.

[4]For all results tables, we report LAS and UAS, plus LA per POS tag).

7. German Results

The results for the German experiments when evaluating on tweeDe are shown in Table 5, the results on GSD in Table 6.

When we look at the results on tweeDe, we again see that the single task setting often results in low performance. Exceptions are proper nouns (PROPN), where we find the lowest score in the weighted MTL full setting (89.30 LA) when using the last subword token, and the overall best score (98.17) in the same setting but using the first token. In the equal setting, using the first token in the single task gives the best results (97.55 LA).

Overall, the major gains on tweeDe between STL and MTL settings involve nouns, verbs, adjectives (ADJ), adverbs (ADV)), and pronouns. The results in the full setting still show more sizable differences in comparison to English. This may be due to the smaller size of tweeDe as compared to Tweebank (see Table 1).

In contrast to English, the evaluation on GSD also shows considerable differences in the equal setting, for nouns, verbs, adjectives, proper nouns, auxiliaries, and pronouns (PRON). In the full setting, only proper nouns show such a difference (for an explanation see below). Similar to English EWT, the highest LAS on GSD is reached by the single task full setting (83.45), showing that GSD also cannot profit from having tweeDe added.

We also look at the comparison of using different subword tokens. German mostly prefers average or last tokens. In the equal setting and in terms of LAS, both tweeDe and GSD prefer the last tokens (77.65 and 78.40) while in the full setting, average tokens perform better (81.84 and 83.45). The same pattern is found for the LA of different POS tags. One noticeable exception is found in proper nouns in tweeDe, which have a clear preference for the first token. This can be explained by the '@' at the beginning of proper nouns (e.g., @Peter). When the first token is used on such a word, it results in '@', which is then a strong indicator for a noun reading. The average also includes this token, but when we use the last token, '@' is not included, resulting in the loss of this correlation.

8. Italian Results

The results for Italian when evaluating on Twittirò are shown in Table 7, the results on ISDT in Table 8.

The results on Twittirò show a similar trend with regard to the STL and MTL settings with MTL settings outperforming STL. The LAS increases from

	tweeDe	NOUN	VERB	ADJ	PROPN	ADV	ADP	AUX	DET	PRON	UAS	LAS
subw.	counts	213	139	82	109	133	94	76	122	137		
first	STL	69.33	68.59	71.14	97.55	83.71	90.43	88.60	90.16	79.32	82.45	74.45
	MTL eql.	71.67	71.22	73.98	95.41	87.47	90.78	90.79	90.71	81.51	83.10	76.02
avg.	STL	69.48	68.11	69.11	95.11	86.97	91.84	89.47	92.62	78.59	82.03	74.87
	MTL eql.	75.27	71.22	76.83	96.33	86.97	90.78	91.23	92.35	80.54	84.52	77.47
last	STL	70.58	66.91	70.33	90.52	87.47	91.13	92.98	89.62	77.62	82.05	74.10
	MTL eql.	75.74	73.14	73.98	91.74	87.47	91.49	92.11	90.98	81.02	85.06	77.65
first	MTL full	77.31	74.58	79.27	98.17	88.22	92.55	95.18	93.17	80.78	87.50	80.83
	MTL full+W	79.03	78.66	80.89	98.17	87.47	92.55	94.30	92.35	81.75	87.61	81.33
avg.	MTL full	78.09	75.78	80.08	98.17	88.47	93.62	95.18	92.08	81.02	87.24	80.70
	MTL full+W	81.22	77.70	82.52	96.33	88.22	91.49	94.74	93.44	82.48	87.96	81.84
last	MTL full	77.00	73.62	81.30	93.27	86.72	93.62	92.98	90.98	82.24	86.47	79.57
	MTL full+W	9.50	77.70	86.59	89.30	87.72	92.20	93.42	92.08	82.24	87.52	80.93

Table 5: Results on German, testing on tweeDe. STL: single task baseline, MTL eql.: multi-task with equal size GSD, MTL full: MTL with full GSD, MTL full+W: MTL with full GSD and weights.

	GSD	NOUN	VERB	ADJ	PROPN	ADV	ADP	AUX	DET	PRON	UAS	LAS
subw.	counts	361	193	125	29	272	168	124	231	191		
first	STL eql.	77.38	69.26	85.07	62.33	89.22	97.22	90.59	98.70	84.47	82.25	77.08
	MTL eql.	75.66	70.64	85.87	64.37	90.32	97.42	93.28	99.42	86.04	83.20	77.95
avg.	STL eql.	76.45	68.05	85.07	70.11	90.32	97.42	91.94	98.99	82.90	82.50	77.19
	MTL eql.	78.39	69.43	85.87	58.62	89.95	97.02	93.28	99.57	86.21	82.70	77.85
last	STL eql.	76.82	67.01	87.20	67.82	89.22	97.42	93.55	98.70	83.25	81.60	76.53
	MTL eql.	78.58	70.29	88.27	67.82	89.58	97.22	92.47	99.57	87.61	83.29	78.40
	counts	3110	1331	1020	1025	1283	1064	686	2016	915		
first	STL full	84.61	86.00	90.62	77.43	89.95	97.63	93.44	92.66	84.95	87.75	82.95
	MTL full	83.14	84.72	89.61	75.19	89.22	97.51	93.15	92.63	85.61	86.84	81.73
	MTL full+W	84.58	85.88	90.42	77.37	89.82	97.67	93.78	92.54	85.25	87.74	82.92
avg.	STL full	85.38	85.65	90.78	78.18	90.34	97.65	94.27	92.74	85.25	88.08	83.45
	MTL full	83.57	84.87	90.13	76.36	90.39	97.49	93.10	92.64	85.65	87.17	82.25
	MTL full+W	85.11	85.67	90.95	78.08	90.54	97.73	93.93	92.71	85.14	87.89	83.29
last	STL full	85.29	85.65	90.75	78.57	90.23	97.46	93.49	92.76	84.92	87.93	83.30
	MTL full	83.71	84.22	89.84	76.62	89.97	97.59	93.83	92.64	85.97	86.99	82.20
	MTL full+W	85.28	85.62	90.82	78.11	90.31	97.65	93.54	92.79	85.68	87.70	83.20

Table 6: Results on German, testing on GSD. STL: single task baseline, MTL: multi-task setup, eql.: GSD equal in size to tweeDe, full: using the full GSD, full+W: using the full GSD and multi-task weights.

73.90 to 77.34 in the equal setting and to 78.76 for the weighted full setting. We also see similar trends in that nouns, verbs, adjectives, adverbs, and pronouns are the main source of increase. The exception for Italian are coordinating conjunctions, which reach the lowest results in the MTL equal setting (84.31) and the highest results in the STL setting (89.71). While both treebanks include different dependency relations for CCONJ; 'cc' or 'discourse' are the majority of cases in Twittirò, whereas CCONJ does not have a 'discourse' reading in ISDT. This introduces a possible source of negative information. We also see sizable improvements for verbs, adjectives, proper nouns, and pronouns in the full setting, similar to the German treebanks. Again, a possible explanation can be found in the relative sizes of ISDT and the considerably smaller Twittirò.

ISDT also shows major improvements in the equal setting, mostly for nouns, verbs, adjectives, auxiliaries, and pronouns. In the full setting, the only major improvement is for pronouns.

In the comparison of subword tokens, there are differences between the equal and the full setting. Overall, Italian disprefers first tokens, but in the equal setting, averaged subwords work best for Twittirò. For ISDT, there is no difference between average and last tokens. In the full settings, last tokens work somewhat better. However, when we use weighting, the differences are small. We also note that individual POS tags prefer different representations. Nouns, for example, prefer averages for Twittirò in the equal setting and for ISDT in the full setting, but last tokens in the other two settings.

9. Turkish Results

The results for Turkish when evaluating on BOUN are shown in Table 9, the results on the Turkish Penn Treebank in Table 10.

The results show that for the two Turkish treebanks, there are considerably fewer differences between the individual settings. The only major differences in BOUN occur for proper nouns in

	Twittirò	NOUN	VERB	ADJ	PROPN	ADV	ADP	CCONJ	AUX	DET	PRON	UAS	LAS
subw.	counts	428	271	144	203	139	311	68	109	334	145		
first	STL	71.34	70.11	75.69	81.94	84.41	94.96	86.76	91.74	97.50	69.43	80.80	73.90
	MTL eql.	74.69	72.94	80.32	82.43	83.93	96.89	86.27	93.58	97.90	71.72	83.05	76.56
avg.	STL	72.74	72.32	79.17	84.07	84.41	96.25	89.71	91.13	97.70	69.43	82.15	75.35
	MTL eql.	75.39	74.78	83.33	82.59	85.13	97.21	85.78	92.66	98.00	71.26	84.08	77.34
last	STL	72.59	71.22	78.94	82.10	82.73	97.00	85.78	92.05	97.41	68.97	81.50	74.49
	MTL eql.	74.69	74.05	82.41	81.61	85.37	97.96	84.31	93.88	98.20	69.89	83.06	76.22
first	MTL full	75.55	76.14	81.94	83.91	85.37	97.00	84.80	95.72	97.90	74.48	84.16	78.10
	MTL full+W	75.78	77.98	83.10	82.76	86.09	97.32	84.80	96.33	98.40	73.10	85.18	78.63
avg.	MTL full	76.09	76.26	84.49	85.22	86.33	97.00	85.78	95.41	97.90	72.64	84.59	78.42
	MTL full+W	75.93	76.38	85.19	83.42	85.85	97.21	86.27	97.25	98.20	73.33	85.24	78.76
last	MTL full	74.38	75.15	84.03	85.39	84.17	97.32	86.27	95.11	97.80	75.63	84.21	77.73
	MTL full+W	76.40	77.98	82.18	84.07	85.37	98.29	85.78	95.72	98.20	73.79	85.07	78.64

Table 7: Results on Italian, testing on Twittirò. STL: single task baseline, MTL eql.: multi-task with equal size ISDT, MTL full: MTL with full ISDT, MTL full+W: MTL with full ISDT and weights.

	ISDT	NOUN	VERB	ADJ	PROPN	ADV	ADP	CCONJ	AUX	DET	PRON	UAS	LAS
subw.	counts	609	282	193	147	153	481	95	126	491	133		
first	STL eql.	79.97	75.77	79.97	84.58	87.80	98.34	100	92.06	99.46	74.69	85.77	81.10
	MTL eql.	81.88	79.67	82.56	85.26	89.32	97.92	100	95.77	99.32	77.19	87.55	83.50
avg.	STL eql.	80.73	76.12	81.00	83.67	88.67	98.13	100	100	94.97	72.68	86.13	81.86
	MTL eql.	82.92	79.79	82.73	83.67	88.89	98.27	100	95.50	99.32	76.44	87.46	83.51
last	STL eql.	81.94	78.01	83.42	83.90	88.24	98.20	100	92.33	99.12	70.68	86.14	81.75
	MTL eql.	83.14	78.49	85.32	82.77	89.76	97.99	100	93.65	99.39	77.44	87.39	83.51
	counts	2070	863	680	505	401	1643	263	405	1712	411		
first	STL full	90.58	90.73	90.74	89.31	93.35	99.09	100	98.11	99.75	87.59	93.13	90.94
	MTL full	89.65	89.22	89.80	87.82	93.27	99.09	99.87	97.53	99.69	84.18	92.29	89.80
	MTL full+W	90.72	91.35	90.44	89.31	93.68	99.15	100	100	97.78	86.62	93.07	90.96
avg.	STL full	91.32	91.58	91.72	89.50	95.10	99.11	100	98.52	99.65	87.51	93.38	91.39
	MTL full	90.16	90.38	91.27	88.51	94.60	99.07	100	98.19	99.67	85.81	92.60	90.33
	ITL full+W	91.21	92.04	92.21	89.70	94.93	99.11	100	98.77	99.67	86.35	93.53	91.56
last	STL full	90.92	91.23	92.06	89.17	94.76	99.11	100	97.61	99.67	87.75	93.40	91.30
	iMTL full	89.66	89.80	90.74	88.25	94.18	99.01	99.87	97.78	99.67	85.56	92.35	90.02
	MTL full+W	91.06	91.35	91.62	89.87	95.10	99.15	100	100	98.02	87.19	93.31	91.29

Table 8: Results on Italian, testing on ISDT. STL: single task baseline, MTL: multi-task setup, eql.: ISDT equal in size to Twittirò, full: using the full ISDT, full+W: using the full ISDT and multi-task weights.

the equal setting, they increase from 74.94 LA to 79.52. Surprisingly, we see more major differences in the Penn Treebank, even in the full setting, proper nouns and pronouns profit from the MTL setting. This may be due to the mix of genres in BOUN, which provides more varied input for training.

Since the Turkish experiments show smaller differences than the other languages, it is more difficult to identify a preference for a specific subword representation. When looking at the LAS, BOUN prefers last tokens in the equal setting while Penn prefers first tokens. In the full setting, both treebanks prefer averaged tokens.

10. Distributional Overlap

We now have a closer look at the overlap of subword tokens between treebanks (i.e., we calculate the percentage of subwords in the training data of one treebank that also occur in the training data of the other treebank). We assume that differences in the overlap may be one of the reasons why the large treebanks suffer from negative transfer in the full MTL setting without weighting.

Figure 1 presents the overlap of subwords for the training sets for the individual languages. When examining German in the equal setting, we can see that the percentages shared between the treebanks is relatively equal (between 65% and 76%). This may explain why in the equal MTL settings, both treebanks benefit, as both treebanks have room for improvement in terms of unknown words, while neither have a dominant influence on the shared vocabulary, allowing for better mutual sharing. When we look at the overlap for the full GSD treebank, a considerably larger percentage of tweeDe is covered, which, while not unexpected given the size differences, is still surprising because of the difference in genres between the two treebanks. However, this also reveals that in over 90% of cases, tweeDe has an impact on tokens that are also in GSD. This may explain why the MTL setting without weights results in performance degradation for the GSD treebank, as tweeDe is able to have too much influence on the shared tokens given its relative size. English and Italian show similar trends. As noted by Dakota et al. (2021), reducing the weight of the smaller treebanks seems to minimize degrada-

	BOUN	NOUN	VERB	ADJ	PROPN	ADV	ADP	CCONJ	AUX	DET	PRON	UAS	LAS
subw.	counts	3952	2199	681	677	479	264	337	240	546	321		
first	STL	75.38	78.84	78.90	74.94	82.88	82.07	85.36	79.58	94.63	81.10	79.48	71.94
	MTl eql.	75.95	78.78	79.54	77.20	81.91	83.21	85.26	80.14	94.57	80.27	79.71	72.28
avg.	STL	76.21	79.37	79.39	77.30	83.23	83.46	85.56	79.86	94.38	81.31	79.89	72.65
	MTL eql.	76.69	79.31	79.74	78.34	83.09	83.71	85.36	80.28	94.20	80.89	80.00	72.89
last	STL	75.84	79.57	79.20	79.37	82.60	83.59	87.14	80.00	94.44	82.35	79.86	72.62
	MTL eql.	76.60	79.41	80.76	79.52	83.44	83.46	85.76	80.42	94.57	81.83	80.08	72.96
first	MTL full	76.00	78.94	79.69	77.89	82.25	83.59	85.56	80.83	94.44	80.27	80.07	72.68
	MTL full+W	76.27	78.66	79.54	77.35	82.67	83.21	85.46	80.14	94.69	80.27	79.87	72.54
avg.	MTL full	76.54	79.61	80.32	79.57	83.16	83.71	85.06	80.83	94.32	80.27	80.32	73.22
	MTL full+W	76.69	79.72	80.27	79.47	83.37	83.21	85.86	81.11	94.44	80.58	80.36	73.24
last	MTL full	76.64	79.87	80.37	79.42	83.02	83.71	85.66	80.69	94.51	81.83	80.13	73.09
	MTL full+W	76.61	79.90	81.06	79.37	82.46	83.08	85.36	80.56	94.63	80.37	80.29	73.16

Table 9: Results on Turkish, testing on BOUN. STL: single task baseline, MTL eql.: multi-task with equal size Penn, MTL full: MTL with full Penn, MTL full+W: MTL with full Penn and weights.

	Penn	NOUN	VERB	ADJ	PROPN	ADV	ADP	CCONJ	AUX	DET	PRON	UAS	LAS
subw.	counts	3463	819	1073	997	587	206	296	37	407	146		
first	STL eql.	68.89	91.94	68.81	72.38	70.58	69.74	75.23	73.87	90.83	75.57	83.18	69.29
	MTL eql	70.22	92.39	70.02	74.56	72.35	67.31	75.45	73.87	92.14	77.85	84.59	70.84
avg.	STL eql.	69.97	92.31	69.59	74.22	72.17	68.93	74.77	74.77	90.75	75.11	83.56	69.62
	MTL eql.	70.11	92.47	69.77	76.53	71.21	67.31	75.23	75.68	91.56	78.31	84.64	70.74
last	STL eql.	69.16	92.10	69.77	71.95	71.78	68.45	75.00	74.77	90.58	76.03	83.10	69.18
	MTL eql.	70.05	92.27	69.06	74.36	73.25	68.12	74.44	73.87	92.22	78.08	84.25	70.34
first	STL full	71.49	92.47	70.43	74.42	72.86	71.20	75.00	75.68	90.75	74.20	85.30	71.49
	MTL full	71.93	92.59	70.43	76.30	73.31	70.06	75.45	75.68	91.07	76.71	85.80	72.24
	MTL full+W	72.04	92.71	71.08	76.23	73.37	70.55	74.21	75.68	91.40	77.40	85.96	72.40
avg.	STL full	71.68	92.63	70.86	76.20	73.03	70.55	73.87	75.68	90.91	73.06	85.64	71.83
	MTL full	72.04	92.88	70.77	77.47	73.08	71.36	74.32	75.68	90.66	77.17	86.05	72.38
	MTL full+W	72.12	92.84	71.05	77.50	73.59	71.04	74.55	75.68	91.07	76.94	86.02	72.51
last	STL full	71.38	92.51	71.02	76.33	72.69	69.26	75.56	75.68	90.25	75.34	85.38	71.66
	MTL full	71.70	93.04	71.23	76.93	73.14	71.20	73.99	75.68	91.07	77.17	85.76	72.12
	MTL full+W	72.06	92.76	70.52	76.23	73.59	71.36	74.10	75.68	90.99	77.85	85.59	71.96

Table 10: Results on Turkish, testing on Penn. STL: single task baseline, MTL: multi-task setup, eql.: Penn equal in size to BOUN, full: using the full Penn, full+W: using the full Penn and multi-task weights.

tion in performance in the larger treebanks, which can be interpreted as reducing the impact on the proportionally high number of shared tweeDe tokens on optimization.

When we compare the other languages to Turkish, we see in that for Turkish in the full setting, the overlap for both is around 90%. Thus, while BOUN is smaller, it does not control a disproportionate number of tokens that also impact Penn, and we see that Penn ultimately benefits in a non-weighted MTL setup as well. Therefore, while the discrepancy in treebank size certainly plays a role, the distributional subword overlap appears to be a more important factor in influencing directions of positive and negative transfer.

11. Conclusion and Future Work

We conducted a set of experiments to better understand possible sources of transfer in MTL parsing of treebanks of different domains. Experiments showed trends, such as a preference for either using the average over all subwords or the last subword. But there are also influences that can be partially attributed to annotation decisions, which may result in mixed signals if they are in conflict (Sayyed and Dakota, 2021). Importantly, while relative treebank size is still important, results indicate that the degree of overlap of subwords can serve as a better indicator of the degree to which each task in MTL influences the other, and thus which treebank may more strongly drive positive or negative transfer. This aligns with work indicating the important role of subwords in transfer learning paradigms (Wu and Dredze, 2019; Blaschke et al., 2023). For future work, we will examine the interaction of cross-lingual behavior with multiple domains and restrictions on information sharing.

Limitations

The major limitation of this work is the limited availability of treebanks in different genres. Ideally, the comparison should have included more typologically diverse languages. However, we are already stretching the boundaries by including Turkish, for which no UD Twitter Treebank exists.

Another limitation results from differences in annotations. For example, Twittirò contains a high number of words annotated as SYM on the POS

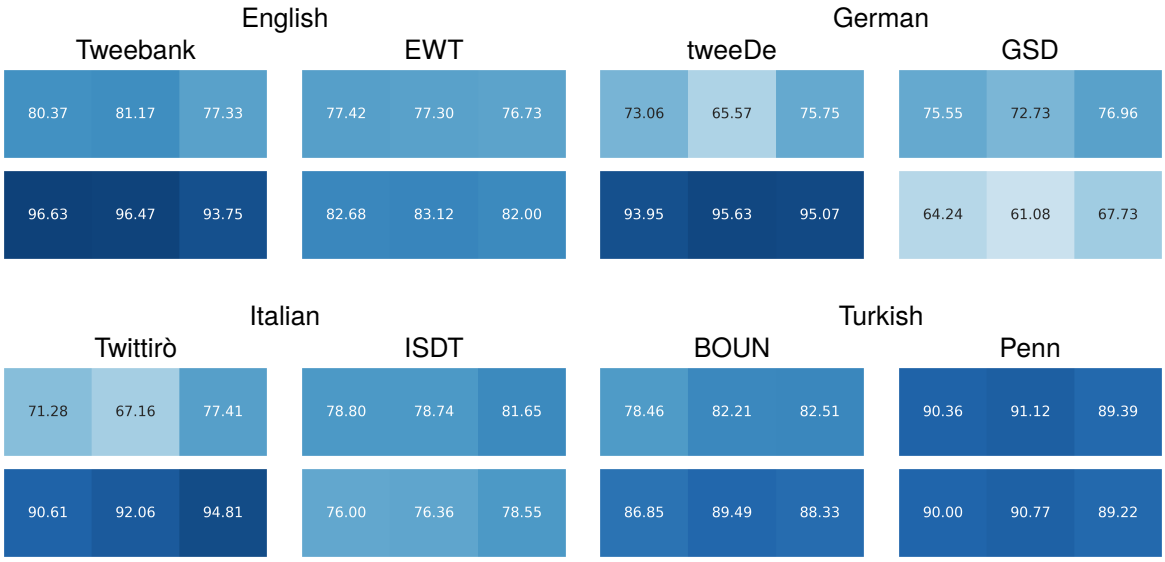

Figure 1: Percentage of token overlap for English (Tweebank train and EWT train), German (tweeDe train and GSD train), and Italian (Twittirò train and ISDT train), and Turkish (BOUN train and Penn train): equal (top) full (bottom) setting; for first subword (left), average subwords (middle), and last subword (right).

level while there are no SYM POS tags in the ISDT subset of equal size. This is due to the decision in Twittirò to annotate Twitter handles and hashtags as SYM rather than as nouns, verbs, or X, as in Tweebank and tweeDe. Such differences in annotation can have a significant influence on the usefulness of MTL.

We also acknowledge that different language models can possess different tokenizers and would further contribute to the complexity of the problem since they may represent the same text very differently, but view this as necessary future work.

Ethics Statement

We are not aware of any ethical concerns.

12. Bibliographical References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Isabelle Augenstein and Anders Søgaard. 2017. Multi-task learning of keyphrase boundary classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 341–346, Vancouver, Canada.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 152–162, Valencia, Spain.

Verena Blaschke, Hinrich Schütze, and Barbara Plank. 2023. Does manipulating tokenization aid cross-lingual transfer? A study on POS tagging for non-standardized languages. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 40–54, Dubrovnik, Croatia.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Kenneth Ward Church. 2020. Emerging trends: Subwords, seriously? *Natural Language Engineering*, 26(3):375–382.

Alessandra Teresa Cignarella, Cristina Bosco, and Paolo Rosso. 2019. Presenting TWITTIRÒ-UD: An Italian Twitter treebank in Universal Dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 190–197, Paris, France.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Daniel Dakota. 2021. Genres, parsers, and BERT: The interaction between parsers and BERT models in cross-genre constituency parsing in English and Swedish. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 59–71, Online. Association for Computational Linguistics.

Daniel Dakota, Zeeshan Ali Sayyed, and Sandra Kübler. 2021. Bidirectional domain adaptation using weighted multi-task learning. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 93–105, Online. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Timothy Dozat and Christopher Manning. 2017. Deep biaffine attention for neural dependency parsing. In *5h International Conference on Learning Representations (ICLR 2017)*, Toulon, France.

Timothy Dozat, Peng Qi, and Christopher D. Manning. 2017. Stanford's graph-based neural dependency parser at the CoNLL 2017 shared task. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30, Vancouver, Canada.

Agnieszka Falenska and Özlem Çetinoğlu. 2017. Lexicalized vs. delexicalized parsing in low-resource scenarios. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 18–24, Pisa, Italy.

Daniel Fried, Nikita Kitaev, and Dan Klein. 2019. Cross-domain generalization of neural constituency parsers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 323–330, Florence, Italy.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3594–3608, Online. Association for Computational Linguistics.

Vidur Joshi, Matthew Peters, and Mark Hopkins. 2018. Extending a parser to distant domains using a few dozen partially annotated examples. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1190–1199, Melbourne, Australia.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Stav Klein and Reut Tsarfaty. 2020. Getting the ##life out of living: How adequate are word-pieces for modelling complex morphology? In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 204–209, Online. Association for Computational Linguistics.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California.

Ying Li, Zhenghua Li, and Min Zhang. 2020. Semi-supervised domain adaptation for dependency parsing via improved contextualized word representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, page 38, Barcelona, Spain (Online).

Zhenghua Li, Xue Peng, Min Zhang, Rui Wang, and Luo Si. 2019. Semi-supervised domain adaptation for dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2386–2395, Florence, Italy.

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A. Smith. 2018. Parsing tweets into Universal Dependencies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 965–975, New Orleans, Louisiana. Association for Computational Linguistics.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4513–4519. International Joint Conferences on Artificial Intelligence Organization. Special Track on AI in FinTech.

Minh-Thang Luong, Quoc Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. In *Proceedings of ICLR*, San Juan, Puerto Rico.

Xiaofei Ma, Peng Xu, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2019. Domain adaptation with BERT-based domain classification and data selection. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 76–83, Hong Kong, China. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, Scotland, UK.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, 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)*, pages 4034–4043, Marseille, France.

Nanyun Peng and Mark Dredze. 2017. Multi-task domain adaptation for sequence tagging. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 91–100, Vancouver, Canada.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Barbara Plank and Gertjan van Noord. 2011. Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1566–1576, Portland, Oregon, USA.

Ines Rehbein, Josef Ruppenhofer, and Bich-Ngoc Do. 2019. tweeDe – a Universal Dependencies treebank for German tweets. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 100–108, Paris, France.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4933–4941, Marseille, France. European Language Resources Association.

Rudolf Rosa and Zdeněk Žabokrtský. 2015. KL-cpos3 - a language similarity measure for delexicalized parser transfer. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 243–249, Beijing, China.

Michael T Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G Dietterich. 2005. To transfer or not to transfer. In *NIPS 2005 Workshop on Transfer Learning*, volume 898, Vancouver, Canada.

Vin Sachidananda, Jason Kessler, and Yi-An Lai. 2021. Efficient domain adaptation of language models via adaptive tokenization. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 155–165, Virtual. Association for Computational Linguistics.

Zeeshan Ali Sayyed and Daniel Dakota. 2021. Annotations matter: Leveraging multi-task learning to parse UD and SUD. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3467–3481, Online. Association for Computational Linguistics.

Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. 2018. Multi-task learning for argumentation mining in low-resource settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 35–41, New Orleans, LA.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 231–235, Berlin, Germany.

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 (ACL)*, pages 619–625, Melbourne, Australia.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, Thomas R. McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations (ICLR) 2019*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS 2017)*, volume 30.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish.

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. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Sen Wu, Hongyang R. Zhang, and Christopher Ré. 2020. Understanding and improving information transfer in multi-task learning. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia.

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

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium.

13. Language Resource References

Cignarella, Alessandra Teresa and Bosco, Cristina and Rosso, Paolo. 2019. *TWITTIRÒ-UD: An Italian Twitter Treebank in Universal Dependencies*. LINDAT/CLARIN.

Liu, Yijia and Zhu, Yi and Che, Wanxiang and Qin, Bing and Schneider, Nathan and Smith, Noah A. 2018. *Parsing Tweets into Universal Dependencies*. https://github.com/Oneplus/Tweebank.

Nivre, Joakim and de Marneffe, Marie-Catherine and Ginter, Filip and Hajič, Jan and Manning, Christopher D. and Pyysalo, Sampo and Schuster, Sebastian and Tyers, Francis and Zeman, Daniel. 2020. *Universal Dependencies v2: An*

Evergrowing Multilingual Treebank Collectio. LINDAT/CLARIN.

Rehbein, Ines and Ruppenhofer, Josef and Do, Bich-Ngoc. 2019. *tweeDe – A Universal Dependencies treebank for German tweets*. University of Heidelberg.