Challenges to Open-Domain Constituency Parsing

Sen Yang[♠], **Leyang Cui**[♡♠], **Ruoxi Ning**[♡], **Di Wu**[△], **Yue Zhang**[♠◇*]
[♠]School of Engineering, Westlake University
[♡]Zhejiang University [△]Sichuan University
[◇]Institute of Advanced Technology, Westlake Institute for Advanced Study
senyang.stu@gmail.com {cuileyang, zhangyue}@westlake.edu.cn
lune_rgb@163.com wuuuudle@gmail.com

Abstract

Neural constituency parsers have reached practical performance on news-domain benchmarks. However, their generalization ability to other domains remains weak. Existing findings on cross-domain constituency parsing are only made on a limited number of domains. Tracking this, we manually annotate a high-quality constituency treebank containing five domains. We analyze challenges to open-domain constituency parsing using a set of linguistic features on various strong constituency parsers. Primarily, we find that 1) BERT significantly increases parsers' cross-domain performance by reducing their sensitivity on the domain-variant features. 2) Compared with single metrics such as unigram distribution and OOV rate, challenges to open-domain constituency parsing arise from combinations of factors, including cross-domain lexical and constituent structure variations.

1 Introduction

Constituency parsing is a fundamental task in NLP that has received constant research attention (Cross and Huang, 2016; Liu and Zhang, 2017; Stern et al., 2017; Kitaev and Klein, 2018). As shown in Figure 1, given a sentence, the task is to identify hierarchical phrase structures that reflect its syntax, such as prepositional phrases (PP; e.g., "*in late 1991*"), noun phrases (NP; e.g., "*late 1991*") and verb phrases (VP; e.g., "*scheduled for delivery in late 1991*"). Constituent structures have been shown useful for downstream tasks including machine translation (Wang et al., 2018), natural language inference (Chen et al., 2017), text summarization (Xu and Durrett, 2019). In addition, they can be transformed into dependency tree structures (Zhang and Clark, 2008), which have been shown to be useful for a wide range of NLP tasks.

The dominant approach to constituency parsing employs a neural model with pre-trained token representation (Kitaev et al., 2019), training the network parameters over manually labeled constituent structures from the Penn Treebank (PTB) (Marcus et al., 1993). As labeled constituent trees can be costly to obtain, most work makes use of the PTB data for training, which is financial news. The current state-of-the-art F-scores reach over 95% on the training domain (i.e., newswire) and are around 88% for biomedical and web test data (Tateisi et al., 2005; Silveira et al., 2014). Compared with parser performance decades ago, accuracies around 90% nowadays is much more useful for downstream applications. Fried et al. (2019) showed that pre-training is a key factor that brings consistent cross-domain performance improvements by using BERT (Devlin et al., 2019).

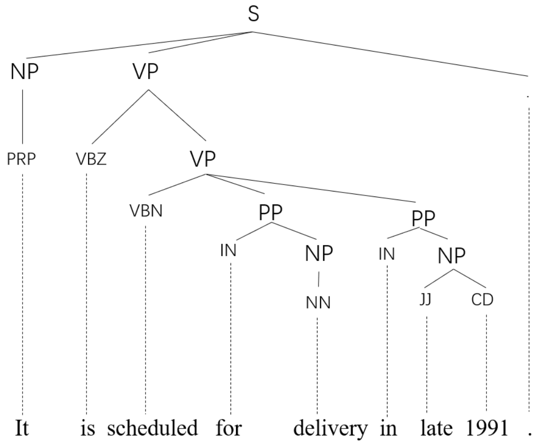


Figure 1: An example of constituency parse tree.

Ideally, a constituency parser should give robust performance *in the open domain*, so that both domain-specific applications (Zhang et al., 2021) and open-domain NLP tasks (Hu et al., 2019) can benefit from syntactic structures. The above observations, however, are made on a rather limited (i.e., 3) number of domains. In addition, there has been relatively little study on the key factors to the performance gap between financial news test and

[*] Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2022, pages 112 - 127
May 22-27, 2022 ©2022 Association for Computational Linguistics

test on other domains, when the model is trained on financial news. It remains an interesting research question to understand the performance of constituency parsing with regard to a wider range of domains and text genres in order to understand the boundaries and existing techniques and identify the main challenges for robust open-domain constituent parsing. Such knowledge can be informative for guiding the design of robust open-domain parsers.

To this end, we evaluate three strong constituency parsers on these domains, as well as the existing news, biomedical and web domains. The parsers include the non-neural BLLIP parser (Charniak and Johnson, 2005), the in-order transition-based parser (Liu and Zhang, 2017) and the Berkeley neural chart-based parser (Kitaev and Klein, 2018). For the test domains, we include the vast majority of existing cross-domain test data in the literature, which cover the biomedical, web text, literature fiction and telephone conversations. In addition, given much research interest in NLP for dialogue (Budzianowski et al., 2018), law (Chalkidis et al., 2019) and review (Oved and Levy, 2021) domains, we manually label constituent structures for five typical domains (i.e., dialogue, forum, law, literature, review), resulting in a test set of 1,000 sentences for each domain. Empirically, we aim to answer the following research questions.

First, *what are the parser performances in the open domain, and which domains are the most challenging for constituent parsing?* We find that the parser performance varies from 83% to 93% under different domains, and the most challenging text genres are review, dialogue and literature. The low results on these domains mean that open-domain constituency parsing is still a challenge.

Second, *what are the relative strengths of different parser models, and does BERT give similar improvements for all domains?* We find that the parsers that give stronger results on PTB do not necessarily give stronger results on various other domains, which reflects limitations of evaluating parser performances only on PTB data. Besides, we show that BERT benefits parsers on cross-domain performance by reducing their sensitivity on domain-variant features.

Third, *what are the main challenges for cross-domain parsing?* By analyzing a set of linguistic features, we find that compared with single metrics such as unigram distribution and OOV rate, challenges to cross-domain constituency parsing arise from combinations of factors, including cross-domain lexical and constituent structure variations.

To our knowledge, we are the first to construct constituency parsing test data for the forum and law domains and the first to analyze the factors that make open-domain parsing challenging by extensive empirical evaluation. We release our dataset and results at `https://github.com/RingoS/multi-domain-parsing-analysis`.

2 Related Work

2.1 Cross-domain Treebanks

Penn Treebank (Marcus et al., 1993) was the very first large-scale dataset that enables researchers to implement statistical constituency parsers that achieve high accuracy on phrase structure prediction (Charniak, 1997; Klein and Manning, 2003). Encouraged by the success of PTB, treebanks on other domains have been developed. Brown corpus (Marcus et al., 1993) was created to assess the cross-domain generalization ability of parsers trained on the newswire data of PTB. Switchboard contains transcripts from telephone conversations. BNC (Foster and van Genabith, 2008) consists of 1,000 hand-corrected British National Corpus parse trees. English Web Treebank (EWT) (Silveira et al., 2014) contains phrase structure annotations from five genres of web media: weblogs, newsgroups, emails, reviews, and Yahoo! answers. Genia (Tateisi et al., 2005) is based on biomedical literatures and was created to support the development of NLP for the domain of molecular biology. Our MCTB is constructed to cover a variety of domains for test interest. Some MCTB test domains turn out to be more challenging, as shown in Tables 1 and 3.

2.2 Cross-domain Syntactic Parsing

There has been work considering cross-domain constituent parsing with parser combinations. McClosky et al. (2010) investigated multiple source parser adaptation, which trains several parsers on many different domains. A linear regression model is adopted to predict the combination of these parsers. Their work is different from ours in that: 1) they make use of both PTB and cross-domain training data; In contrast, we consider PTB training to study domain difference in more isolation; 2) Our goal is to systemically compare parser per-

Dataset	# Instance	Avg # token per sent	Avg # cons per sent	Avg # token per cons	# Total tokens	# Total cons	Max # token of sent	Min # token of sent	Avg # token of NP	Avg # token of VP	Avg # token of PP
PTB-train (News)	39,832	23.85	18.62	7.44	950,028	741,833	141	1	4.13	10.56	5.96
PTB-dev (News)	1,700	23.60	18.02	7.30	40,117	30,633	118	1	4.11	10.38	5.69
PTB-test (News)	2,416	23.46	18.33	7.43	56,684	44,276	67	1	4.11	10.54	5.81
Genia-test (Biomedical)	1,360	26.21	21.77	7.46	35,639	29,602	164	2	4.36	11.94	6.38
Brown-test (Mixed)	2,425	18.95	15.83	6.37	45,950	38,380	128	1	3.20	8.62	4.92
Brown-all (Mixed)	24,243	18.94	16.84	7.07	459,148	408,198	172	1	3.20	8.49	5.05
Brown-cf (Lore)	3,164	23.42	20.22	7.82	74,114	63,984	122	1	3.70	9.95	5.53
Brown-cg (Biography)	3,279	25.55	22.18	8.12	83,769	72,728	142	1	3.94	10.38	5.80
Brown-ck (GeneralFic)	3,881	17.24	15.50	6.74	66,890	60,166	112	1	2.95	7.89	4.74
Brown-cl (MysteryFic)	3,714	15.71	14.40	6.33	58,362	53,489	172	1	2.63	7.47	4.37
Brown-cm (ScienceFic)	881	16.59	14.68	6.67	14,613	12,934	144	1	3.06	7.67	4.54
Brown-cn (AdventureFic)	4,415	16.00	14.41	6.30	70,654	63,607	144	1	2.69	7.29	4.43
Brown-cp (RomanceStory)	3,942	17.45	15.79	6.67	68,771	62,242	124	1	2.75	7.75	4.51
Brown-cr (Humor)	967	22.72	19.70	7.90	21,975	19,048	130	1	3.56	9.81	5.75
EWT-all-test (WebText)	8,309	15.24	13.25	6.09	126,593	110,086	135	1	3.05	8.30	4.87
EWT-answers-test	1,709	16.70	15.12	5.64	28,542	25,846	135	1	2.63	7.25	4.14
EWT-email-test	2,450	11.70	10.12	5.91	28,676	24,784	91	1	2.89	8.43	4.80
EWT-newsgroup-test	1,195	17.28	14.49	6.77	20,651	17,318	104	1	3.54	9.64	5.38
EWT-reviews-test	1,906	14.74	12.98	5.57	28,086	24,733	85	1	2.71	7.39	4.36
EWT-weblog-test	1,014	20.07	16.91	7.06	20,356	17,146	95	1	3.73	10.07	5.72
BNC (British English)	1,000	28.31	23.55	7.83	28,311	23,547	130	2	3.94	11.04	6.09
Switchboard (Spoken)	110,503	9.41	9.33	5.31	1,040,013	1,031,528	114	1	2.25	6.88	4.16
Dialogue	1,000	13.51	12.49	5.19	13,509	12,490	89	2	2.65	6.56	4.17
Forum	1,000	22.01	20.39	6.14	22,012	20,386	95	2	2.71	7.56	4.75
Law	1,000	25.59	20.24	7.50	25,585	20,241	66	5	4.10	10.52	5.66
Literature	1,000	23.24	18.59	6.71	23,238	18,585	184	2	3.21	8.20	4.93
Review	1,000	13.30	11.68	5.21	13,297	11,677	106	2	2.96	6.23	4.62

Table 1: Dataset statistics. "# Instance" — the number of sentences in the corresponding dataset. "Avg" — to average. "# token" and "# cons" — the numbers of tokens and constituents, respectively. "Sent" — sentence. "Fic" in Brown dataset means fiction.

formance for understanding the challenges, and thus we consider more parsers and domains, but no innovative models. Joshi et al. (2018) empirically found that contextualized word representations improves domain adaptation when the target domain is syntactically similar to the source domain. They also proposed to make use of a dozen partial annotations to improve cross-domain performance on syntactically-distant domains. Fried et al. (2019) conducted a systematic analysis on cross-domain parsing. They found that: 1) neural models and non-neural models generalize similarly to new domains; 2) large-scale pretraining improves domain adaptation; 3) structured models (e.g., in-order parser) generalizes better to new domains. Our analysis differs from previous work on the follows: 1) we empirically analysis what factors make cross-domain constituency parsing challenging; 2) we conduct experiments on more domains and datasets, which provide more comprehensive understanding for the open-domain setting.

Cross-domain parsing has also been investigated on other grammar formalisms, in particular dependency syntax. Blodgett et al. (2018) broadened English dependency parsing to handle social media English, especially social media African-American English (AAE). They released a dataset which contains 500 tweets along with their dependency annotations. Li et al. (2019) investigated a semi-supervised approach for domain adaptation in dependency parsing. They combined data from source and target domains using a domain embedding approach. Rotman and Reichart (2019) proposed Deep Contextualized Self-training (DCST), which utilizes representation models trained on sequence labeling tasks that are derived from the parser's output when applied to unlabeled data, and integrates these models with the base parser through a gating mechanism.

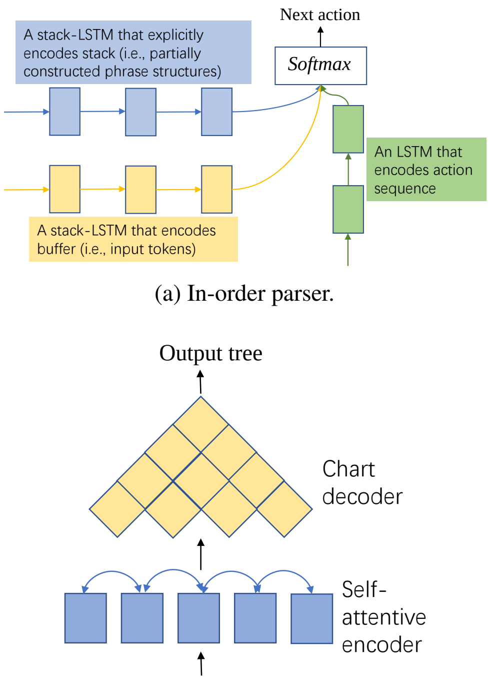
3 Methods and Settings

3.1 Models

We experiment with a strong non-neural parser and recent SOTA neural parsers. The neural parsers are additionally augmented with pretrained BERT (Devlin et al., 2019).

BLLIP Parser. The BLLIP parser (Charniak and Johnson, 2005) is a statistical parser that includes a generative parser (first-stage) and a maximum entropy based re-ranker (second-stage). It first calculates the n-best (typically n = 50) parses, and then re-ranks all produced parses with weighted-averaged scores that are produced by a set of manually-designed features.

In-order Parser. The in-order parser (Liu and Zhang, 2017) is a transition-based parser that traverses the parse tree in an in-order sequence. As


Next action
Softmax
A stack-LSTM that explicitly encodes stack (i.e., partially constructed phrase structures)
An LSTM that encodes action sequence
A stack-LSTM that encodes buffer (i.e., input tokens)
Output tree
Chart decoder
Self-attentive encoder
Input

(a) In-order parser.

(b) Berkeley neural parser.

Figure 2: Structures of the two adopted neural parsers.

shown in Figure 2a, it adopts a stack-LSTM to encode partially constructed tree structures, a stack LSTM to encode input buffer and an LSTM to encode action sequence. In this way, it explicitly models the output phrase structures.

Berkeley Neural Parser. As shown in Figure 2b, Berkeley Neural Parser (Kitaev and Klein, 2018) is a chart-based parser that adopts a self-attentive encoder and a chart-based decoder. Different from in-order parser, it predicts the span labels solely based on local span representations and does not explicitly model the output tree structure.

3.2 Experimental Settings

For BLLIP[1], we adopt their released parser "WSJ-PTB3". For in-order[2], we use their released code, model checkpoints and word embeddings. The embeddings are pretrained on the AFP portion of English Gigaword. The in-order parser requires part-of-speech (POS) tags, for which we adopt a transformer-based tagger trained on the PTB training set. As for the BERT-augmented in-order parser, we adopt the open-sourced code and model checkpoints from Fried et al. (2019)[3]. We train the Berkeley neural parser without and with BERT, respectively, using their released code[4]. The non-BERT Berkeley parser uses randomly initialized embeddings, which differs from the in-order parser. All parsers are trained on standard PTB training set and validated on PTB development set (Marcus et al., 1993).

We evaluate the parsers on 25 test sets, including PTB, Brown (Marcus et al., 1993), Genia (Tateisi et al., 2005), EWT (Silveira et al., 2014), BNC (Foster and van Genabith, 2008), Switchboard and our newly annotated test set. Some of these datasets have multiple subdomains (i.e., Brown and EWT). The domains are shown in Table 1. We call our test set MCTB (Multi-domain constituent Treebank) and provide detailed descriptions in Section 4.

4 Dataset

4.1 Annotation

Our new MCTB testset is composed of texts from 5 genres, including dialogue, forum, law, literature and review. For the dialogue domain, we randomly sample dialogue utterances from Wizard of Wikipedia (Dinan et al., 2019), which is a chit-chat dialogue benchmark produced by humans. For the forum domain, we use users' communication records from Reddit, crawled and released by Völske et al. (2017). For the law domain, we sample text from European Court of Human Rights Database (Stiansen and Voeten, 2019), which includes detailing judicial decision patterns. For the literature domain, we download literary fictions from Project Gutenberg[5]. For the review domain, we use plain text across a variety of product genres, released by SNAP Amazon Review Dataset (He and McAuley, 2016).

We follow PTB's annotation guideline and paradigm (Marcus et al., 1993) to design our annotation guideline, hiring a group of senior undergraduate and master students whose majors are linguistics as our annotators. The annotators are asked to read the guideline, practice and correct the errors of the predicted parse tree, which is produced by a SOTA chart-based parser that is developed based on Berkeley Neural Parser. For annotation clarity, we develop a web-based visualization annotation toolkit, which accepts bracketed format lines and visualizes parse tree structures. The annotation tool

[1] https://github.com/BLLIP/bllip-parser

[2] https://github.com/LeonCrashCode/InOrderParser

[3] https://github.com/dpfried/rnng-bert

[4] https://github.com/nikitakit/self-attentive-parser

[5] https://www.gutenberg.org/

allows adding/deleting constituents in the tree structure. We release our annotation toolkit at https://github.com/Nealcly/AnnoCons.

Annotators are first required to annotate 100 instances from the PTB test set repeatedly, until their labeling is sufficiently accurate to provide useful annotation. To further control annotation quality, the annotators are assigned workloads in batches, with the batch size being 100. For each batch, we randomly select 10 instances (10%), and the main authors check the sampled instances with a side-by-side annotation. If the F-1 scores between the annotator annotated and the inspector annotated for these 10 instance is less than 95%, the corresponding batch will be rejected and assigned to a new annotator. The annotators get their salaries no matter their annotations are rejected or not.

4.2 Data Statistics

We report dataset statistics in Table 1, including the total numbers of instances, of tokens and of constituents, the averaged numbers of tokens within sentences and within constituents and the maximum and minimum numbers of tokens among all sentences. We also report the averaged number of tokens in NP, VP and PP, because they are the most prevalent across all datasets.

From the table, we can see that the dialogue, review and Switchboard domains have the smallest averaged numbers of tokens per sentence, about half of that of PTB. The dialogue, review and Switchboard domains also have the smallest averaged constituent lengths, around 30% shorter than that of PTB. Though the averaged lengths of sentences and of constituents of the literature domain are rather close to those of PTB, the averaged lengths of labeled constituents (especially for NP and PP) are smaller. Among all domains, law shares the most similarities of averaged constituent lengths (both unlabeled and labeled) with PTB. All datasets have similar lengths for shortest sentences, while the literature domain has the largest number of tokens within one sentence.

4.3 Comparison between Features

We report the differences between the PTB training set and various test sets[6] in Table 2, by adopting a list of linguistic features from previous work (Collins and Koo, 2005; Charniak and Johnson, 2005). Each cell in the table represents the Jensen-Shannon divergence between the distribution of a specific feature of the PTB training set and that distribution of a specific test set. Given the distributions P and Q, the Jensen-Shannon divergence is calculated as:

$$JS(P||Q) = \frac{1}{2}(KL(P||M) + KL(Q||M)) \qquad (1)$$

where $KL(P||Q) = \sum_{x \in \chi} P(x) \log(\frac{P(x)}{Q(x)})$ is the Kullback-Leibler divergence, and $M = \frac{1}{2}(P + Q)$. Each value ranges from $0 \sim 1$ and a higher value reflects less correlation on that feature between the PTB training set and the corresponding test set.

In the table, the columns Uni, Bi and Tri denotes unigram, bigram, trigram and fourgram tokens and constituent labels, respectively; GR, HGT and GP denotes grammar rule, headed grammar rule and a chain of (grandparent, parent, child) constituents, respectively. We do not calculate token fourgrams because they are sparse and the OOV rate is over 95% on each domain. Constituent n-grams are calculated within each grammar rule. Grammar rules are unbinarized rules, and examples of headed lexicalized grammar rules include VP [eat] –> VB NP and NP [tomato] –> DT ADJ NN. The OOV rates of token ngrams are also shown in Table 2.

From the table, we can see that the biomedical and review domains have the largest token ngram differences from the PTB training data, while the English Web Treebank is lexically the most similar to PTB-train. Compared to lexical patterns, (unlexicalized) grammatical patterns are relatively more consistent across different domains. Among the different domains, switchboard, dialogues and review have the largest difference in grammar rule patterns as compared to PTB, and the Brown-test, EWT-test and law test sets are relatively the closest to the PTB data. Genia-test, forum, law and literature have a similar level of grammar-feature difference from PTB-train, with brown-test being the closest among the four. From the table, we can see that individual statistics vary across domains, which reflects large domain differences.

5 Experiments

5.1 Overall Results

The performances of the parsers on each domain are shown in Table 3. On PTB-test, all the BERT-based parsers achieve labeled bracket F-scores

[6] For simplicity, we regard Brown and EWT as two whole test sets, respectively. The feature correlations and parser performances including all 25 test sets and subsets are shown in Appendix A.1.

Dataset	N-gram Token (OOV Rate)			GR	HGR	GP	N-gram Constituent			
	Uni	Bi	Tri				Uni	Bi	Tri	Four
PTB-test	0.09 (0.03)	0.41 (0.33)	0.61 (0.72)	0.03	0.19	0.04	0.00	0.01	0.04	0.11
Genia-test	0.38 (0.26)	0.61 (0.71)	0.68 (0.94)	0.16	0.41	0.20	0.05	0.11	0.24	0.37
Brown-test	0.21 (0.06)	0.52 (0.48)	0.67 (0.87)	0.09	0.28	0.11	0.02	0.06	0.16	0.31
Brown-all	0.18 (0.07)	0.45 (0.48)	0.63 (0.87)	0.07	0.24	0.09	0.02	0.05	0.13	0.26
EWT-All-test	0.19 (0.09)	0.49 (0.49)	0.65 (0.86)	0.10	0.29	0.13	0.02	0.06	0.15	0.28
BNC	0.22 (0.11)	0.54 (0.54)	0.67 (0.89)	0.08	0.30	0.10	0.02	0.05	0.12	0.25
Switchboard	0.26 (0.04)	0.49 (0.35)	0.63 (0.78)	0.20	0.39	0.24	0.09	0.16	0.31	0.47
Dialogue	0.28 (0.06)	0.58 (0.46)	0.68 (0.86)	0.16	0.39	0.21	0.03	0.09	0.23	0.41
Forum	0.25 (0.06)	0.55 (0.44)	0.67 (0.84)	0.14	0.36	0.18	0.03	0.09	0.23	0.41
Law	0.27 (0.07)	0.57 (0.51)	0.68 (0.86)	0.12	0.33	0.16	0.01	0.08	0.19	0.34
Literature	0.28 (0.11)	0.57 (0.53)	0.68 (0.90)	0.15	0.36	0.19	0.03	0.09	0.23	0.38
Review	0.30 (0.07)	0.59 (0.51)	0.68 (0.88)	0.16	0.39	0.21	0.03	0.10	0.26	0.45

Table 2: Dataset difference statistics between PTB training set and various test sets. We report Jensen–Shannon divergence of features. Out-of-vocabulary rate (OOV) are also shown for unigram/bigram/trigram tokens. GR, HGR and GP refer to grammar rules, headed lexicalized grammar rules and grandparent rules.

Model Dataset	BLLIP	In-Order	Berkeley	With BERT (Δ Err.)	
				In-Order	Berkeley
PTB-test	91.48	91.53	93.05	95.65 (-48.6%)	95.73 (-38.6%)
Genia-test	78.42	81.06	81.39	86.33 (-27.8%)	86.61 (-28.0%)
Brown-test	85.78	85.74	87.72	93.68 (-55.7%)	93.38 (-46.1%)
Brown-all	85.89	86.55	87.37	93.55 (-52.0%)	93.31 (-47.0%)
EWT-All-test	78.78	81.19	81.98	89.39 (-43.6%)	89.09 (-39.5%)
BNC	84.15	84.55	85.30	92.16 (-49.3%)	91.92 (-45.0%)
Switchboard	77.56	77.44	76.12	84.42 (-30.9%)	84.49 (-35.1%)
Dialogue	77.68	78.40	79.14	85.56 (-33.1%)	86.30 (-34.3%)
Forum	75.25	77.29	78.63	86.33 (-39.8%)	87.04 (-39.4%)
Law	80.67	82.83	84.06	91.50 (-50.5%)	92.06 (-50.2%)
Literature	70.32	76.44	75.98	84.96 (-36.2%)	86.26 (-42.8%)
Review	74.18	75.91	76.15	83.89 (-33.1%)	84.34 (-34.3%)

Table 3: Results (F1 scores) on various test sets. Δ Err. means error reduction rates when using BERT.

above 95%. In comparison, the performances on Genia, BNC, Brown, Switchboard and EWT fall to a range between 84.42% and 93.68%, with relative error increases of 45% to 258%. According to Table 2, these cross-domain test data are relatively close to the PTB data in the distribution of lexical and syntactic patterns. In contrast, on Switchboard, dialogue, forum, literature and review, the results can drop to 83%, with a relative error increase of over 370% (i.e., 95.65% versus 83.89% F-score). This shows that open-domain constituent parsing is still a challenging task to solve.

Among the domains, we find that the review and switchboard domains are the most difficult, with F-scores of around 84% by the BERT-based parsers. The dialogue, forum and literature domains are relatively easier, with F-scores of around 86%. The law domain is the easiest, where the parsers give F-scores of over 90%. Intuitively, the parser performance differences arise from the differences in the text genre between the test domain and PTB: while the review and switchboard domains can contain a fraction of oral and informal English, the law domain is the closest to the newswire domain in style. We give more detailed feature statistics in Section 5.3.

5.2 Comparison between Different Parsers

Among parsers without making use of BERT, the performance drop of In-order parser is relatively the smallest when comparing PTB-test with the domains. As observed by Fried et al. (2019), the relatively larger cross-domain robustness as compared with Berkeley parser may be attributed to the modeling of output structural dependencies by the shift-reduce parser. BLLIP gives a similar cross-domain performance drop as compared with Berkeley parser, which shows that a discrete parser does not necessarily show weaker cross-domain robustness than a neural parser, which again is consistent with findings of Fried et al. (2019).

BERT improves the performances of all neural parser models, with 48.6% and 38.6% error reduction rates for the In-order and Berkeley parsers on

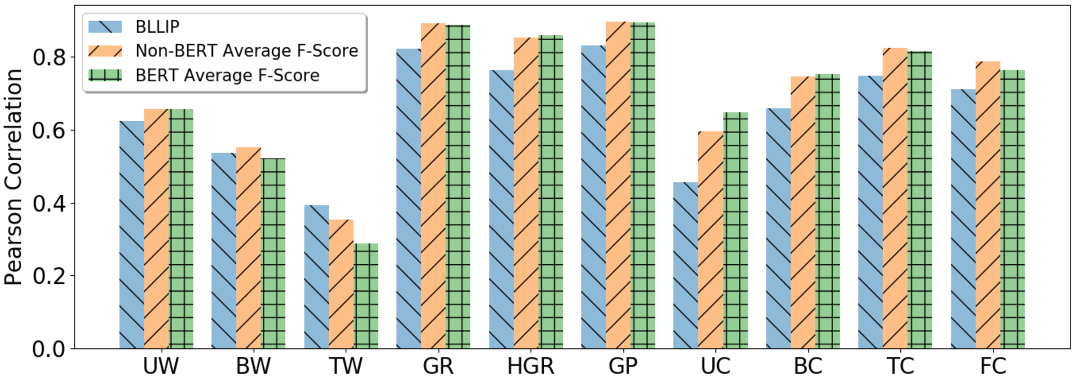


Figure 3: Pearson correlation between feature divergence and parser performance. Because all values are smaller than 0, we simply multiply all values with -1 to make them easier to understand. A higher value represents more reliance on that feature. “Non-BERT Average” refers to the averaged F1 scores of In-order and Berkeley, while “BERT Average” refers to the BERT-augmented version. UW / BW / TW — input token uni- / bi- / tri-gram. GR / HGR / GP — grammar rule / headed grammar rule / grandparent rule. UC / BC / TC / FC — constituent uni- / bi- / tri- / four-gram .

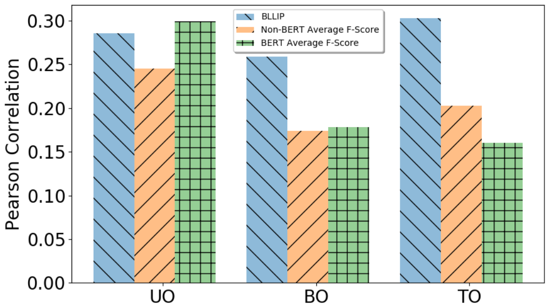


Figure 4: Pearson correlation between OOV rates and parser performance, following the caption of Figure 3. UO / BO / TO — uni- / bi- /tri-gram token OOV.

PTB-test, respectively. For cross-domain test sets, the error reduction rates are 34.3%, 39.4%, 50.2%, 42.8% and 34.3%, respectively for the dialogue, forum, law, literature and review domains with Berkeley neural parser. The reason that a relatively larger error reduction rate is found for the law and literature domains is likely that BERT is trained on Wikipedia and Brown Corpus (i.e., encyclopedia and literature), which has largely similar text genres compared to these datasets. In contrast, the styles of the biomedical (Genia), dialogue and review domains are relatively different from BERT’s training data.

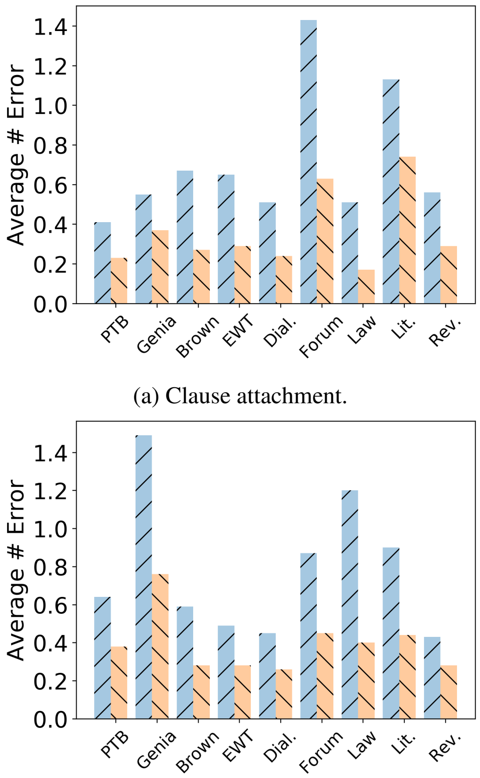
5.3 Key Factors to Cross-domain Challenge

Figure 3 shows the Pearson correlation between parser performances (in Table 3) and feature JS divergences (in Table 2) for all five parsers[7]. In particular, we take the performances of each parser over all the domains in Table 3 (i.e., each column in the table) as a vector, and the JS-divergence values for each feature in Table 2 (i.e. each column in the table) as a vector, calculating the statistical correlation between the two vectors, which reflects the influence of domain shift in each feature on the parser performance. In the figure, each column shows the Pearson correlation of a specific parser with a specific feature, where a longer bar reflects more reliance on the feature.

From Figure 3, we make the following observations. First, overall all the parsers are more influenced by larger grammatical structures such as the whole grammar rule (GR), the grandparent chain (GP) and n-gram sub constituents (BC, TC and FC), while being less influenced by word-level ngram features (BW and TW) and simple constituent label features (UC). This shows that the cross-domain challenge arises mostly from more complex structural variations, instead of cross-domain word and ngram distribution differences.

Second, the traditional BLLIP parser is about as sensitive to word and ngram variations as neural parsers, but less sensitive to syntactic pattern variations such as GR and UC. This shows that the strong representation power of neural models allows them to learn more abstract syntactic structure patterns more accurately. Third, after BERT is used,

[7] In practice, we use Tables 4 and 5, because the domain differences among the sub-genres of Brown or EWT would be eliminated by only using Tables 2 and 3.



(a) Clause attachment.

(b) PP attachment.

Figure 5: Average number of bracket errors per sentence on each dataset using the parser of Liu and Zhang (2017). The errors are classified with Kummerfeld et al. (2012)'s method. Blue bars with slash "/" are without BERT, while orange bars with backslash "\" are with BERT. "SWB", "Dial. ", "Lit. " and "Rev. " are in short for Switchboard, dialogue, literature and review, respectively.

neural parsers show stronger dependence to UC and BC, and weaker dependence to BW, TW, TC and FC features, compared with randomly initialized versions. In fact, as Table 2 shows, the former features are relatively more stable across domains, with less JS-divergence scores between domain test data and PTB training data. This shows that BERT effectively improves parser domain robustness by providing a level of cross-domain knowledge.

Figure 4 shows the Pearson correlation between OOV rates and cross-domain results. Interestingly, the influence of OOV on all parsers are in the range of 16.0% to 30.3%, which is saliently smaller than that of ngram distributions. This shows that the cross-domain challenge arises not simply from unknown tokens, but is more distribution-sensitive. With regard to different parsers performances, the BLLIP parser shows stronger subjectivity to the influence of OOV as compared with the neural parsers, especially for tri-gram OOV, which demonstrates the advantage of dense word representations over sparse one-hot encoding (Bengio et al., 2013). Finally, by further adding BERT, the relative sensitivity of the neural parsers to OOV uni-grams and bi-grams sees increases, while that to OOV tri-grams decreases. This shows that the effect of BERT on cross-domain parsing is more contextualized, in the sense that simply addressing unknown unigram token representations does not necessarily lead to stronger results, but BERT gives the parsers stronger power in representing context distributions.

5.4 Error Characteristics

Figure 5 shows the error distributions of the in-order parser with and without BERT according to the classification of Kummerfeld et al. (2012). In particular, two error types, clause attachment and PP attachment, are shown in the figure, and the charts for more error types are shown in Appendix A.2. As can be seen from Figure 5, the parser makes different types of error across different domains, which reflects different challenges. In the following, we give an example of MCTB-literature. Due to page limitation, figures of the full parse trees and more case studies are shown in Appendix A.2.

It can be seen from Figures 5a and 5b that the literature domain suffers from clause attachment and PP attachment errors, which may result from the fact that sentence structures of the literature domain are more complicated than the stereotype writing style of the newswire domain and there are many rare words in literary works. For example, given a sentence in literature test set: "*The bulldog growls , his scruff standing , a gobbet of pig 's knuckle between his molars through which rabid scumspittle dribbles .*", the gold bracketed-format annotation is

```
...(NP
    (NP (NP (DT a) (NN gobbet))
        (PP
          (IN of)
          (NP
            (NP (NN pig) (POS 's))
            (NN knuckle))))
    (PP (IN between)
      (NP
        (NP (PRP$ his) (NNS molars))
        (SBAR
          (WHPP (IN through) (WHNP (WDT which)))
          (S
            (NP (JJ rabid) (RB scumspittle))
            (VP (NNS dribbles))))))))...
```

and the predicted bracketed-format tree is

```
...(NP
    (NP (DT a) (NN gobbet))
    (PP
  (IN of)
  (NP
    (NP (NN pig) (POS 's)) (NN knuckle)))
(PP
  (IN between)
  (NP (PRP$ his) (NNS molars)))
(SBAR
  (WHPP (IN through) (WHNP (WDT which)))
  (S
  (NP (JJ rabid) (RB scumspittle))
  (VP (NNS dribbles)))))...
```

The clause phrase "*through which rabid scumspittle dribbles*" is supposed to attach to the noun phrase"*his molars*". However, a clause attachment error is produced by the in-order parser, which assigns the clause phrase to the noun phrase "*a gobbet*". In addition, in the predicted tree structure, the PP phrase "*between his molars ...... dribbles*" shares the same parent node with the noun phrase "*a gobbet*" and with the PP phrase "*pig 's knuckle*", which is incorrect. Instead, the PP phrase "*between his molars ...... dribbles*" should be attached to a higher level. This results in a PP attachment error.

6 Conclusion

We investigated the challenges of cross-domain constituent parsing by making use of a large number of test domains, which include newswire, biomedicine, prose, web-text, conversational speeches, as well as give new test domains including dialogue, forum, law, literature and review, for each of which we construct a test set of 1,000 sentences. Results show that the dominant parsers can achieve 83% to 93% accuracies for different domains, and cross-domain parsing is still a challenge, where different domains exhibit varying types of difficulty. We further find that the difficulty for cross-domain parsing lies more in comprehensive distribution differences involving multiple factors such as grammar rules and patterns, as compared to single factors such as OOV rate and token ngram distribution variations. In addition, BERT helps neural parsers improve cross-domain performance by reducing their sensitivity to domain-variant features. Our results show that toward robust open-domain constituent parsing, more work should be done on addressing out-of-distribution generalization in representation learning.

Acknowledgements

We appreciate the insightful comments from the anonymous reviewers. We gratefully acknowledge funding from the National Natural Science Foundation of China (NSFC No.61976180).

References

Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2018. Twitter Universal Dependency parsing for African-American and mainstream American English. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Melbourne, Australia. Association for Computational Linguistics.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

Eugene Charniak. 1997. Statistical parsing with a context-free grammar and word statistics. In *AAAI/IAAI*.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 173–180, Ann Arbor, Michigan. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Michael Collins and Terry Koo. 2005. Discriminative reranking for natural language parsing. *Computational Linguistics*, 31(1):25–70.

James Cross and Liang Huang. 2016. Span-based constituency parsing with a structure-label system and provably optimal dynamic oracles. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1–11, Austin, Texas. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Jennifer Foster and Josef van Genabith. 2008. Parser evaluation and the BNC: Evaluating 4 constituency parsers with 3 metrics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

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. Association for Computational Linguistics.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. *Proceedings of the 25th International Conference on World Wide Web*.

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-domain targeted sentiment analysis via span-based extraction and classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 537–546, Florence, Italy. 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 (Volume 1: Long Papers)*, pages 1190–1199, Melbourne, Australia. Association for Computational Linguistics.

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.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan. Association for Computational Linguistics.

Jonathan K. Kummerfeld, David Hall, James R. Curran, and Dan Klein. 2012. Parser showdown at the Wall Street corral: An empirical investigation of error types in parser output. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1048–1059, Jeju Island, Korea. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Jiangming Liu and Yue Zhang. 2017. In-order transition-based constituent parsing. *Transactions of the Association for Computational Linguistics*, 5:413–424.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 28–36, Los Angeles, California. Association for Computational Linguistics.

Nadav Oved and Ran Levy. 2021. PASS: Perturb-and-select summarizer for product reviews. 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 351–365, Online. Association for Computational Linguistics.

Guy Rotman and Roi Reichart. 2019. Deep Contextualized Self-training for Low Resource Dependency Parsing. *Transactions of the Association for Computational Linguistics*, 7:695–713.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Mitchell Stern, Jacob Andreas, and Dan Klein. 2017. A minimal span-based neural constituency parser. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 818–827, Vancouver, Canada. Association for Computational Linguistics.

Øyvind Stiansen and Erik Voeten. 2019. ECtHR judgments.

Yuka Tateisi, Akane Yakushiji, Tomoko Ohta, and Jun'ichi Tsujii. 2005. Syntax annotation for the GENIA corpus. In *Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts*.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. TL;DR: Mining Reddit to Learn Automatic Summarization. In *Workshop on New Frontiers in Summarization at EMNLP 2017*, pages 59–63. Association for Computational Linguistics.

Xinyi Wang, Hieu Pham, Pengcheng Yin, and Graham Neubig. 2018. A tree-based decoder for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4772–4777, Brussels, Belgium. Association for Computational Linguistics.

Jiacheng Xu and Greg Durrett. 2019. Neural extractive text summarization with syntactic compression. 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 3292–3303, Hong Kong, China. Association for Computational Linguistics.

Yue Zhang and Stephen Clark. 2008. A tale of two parsers: Investigating and combining graph-based and transition-based dependency parsing. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 562–571, Honolulu, Hawaii. Association for Computational Linguistics.

Zixuan Zhang, Nikolaus Parulian, Heng Ji, Ahmed Elsayed, Skatje Myers, and Martha Palmer. 2021. Fine-grained information extraction from biomedical literature based on knowledge-enriched Abstract Meaning Representation. 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 6261–6270, Online. Association for Computational Linguistics.

A Appendix

A.1 Detailed Feature Correlation and Parser Results

Feature correlations and parser performances with all 25 datasets are shown in Tables 4 and 5.

A.2 Error Characteristics

Figure 6 shows nine types of errors made by In-order parser on all test sets. The errors are classified using the method of Kummerfeld et al. (2012). Figures 7 and 8 show the tree structures of the case study in Section 5.4. The tree figures are produced using an open-source visualization toolkit[8].

In Figure 6c, the number of NP internal structure errors of Genia is saliently larger compared to the other domains, which can be because the biomedical domain has a relatively larger amount of special nominal terminologies, which cannot be easily identified using newswire knowledge. Take an instance from Genia test set for example, the gold annotation is

```
...(NP
       (DT a)
       (ADJP
          (NN HLA)
          (NN class)
          (CD II)
          (JJ DR11-restricted))
       (NN fashion))...
```

where "*HLA class II DR11-restricted*" is an adjective phrase modifying the noun "*fashion*". However, the in-order parser prediction is

```
...(NP
       (DT a)
       (NN HLA)
       (NN class)
       (CD II)
       (JJ DR11-restricted)
       (NN fashion))...
```

which does not recognize the sub-structures under the noun phrase "*a HLA class II DR11-restricted fashion*".

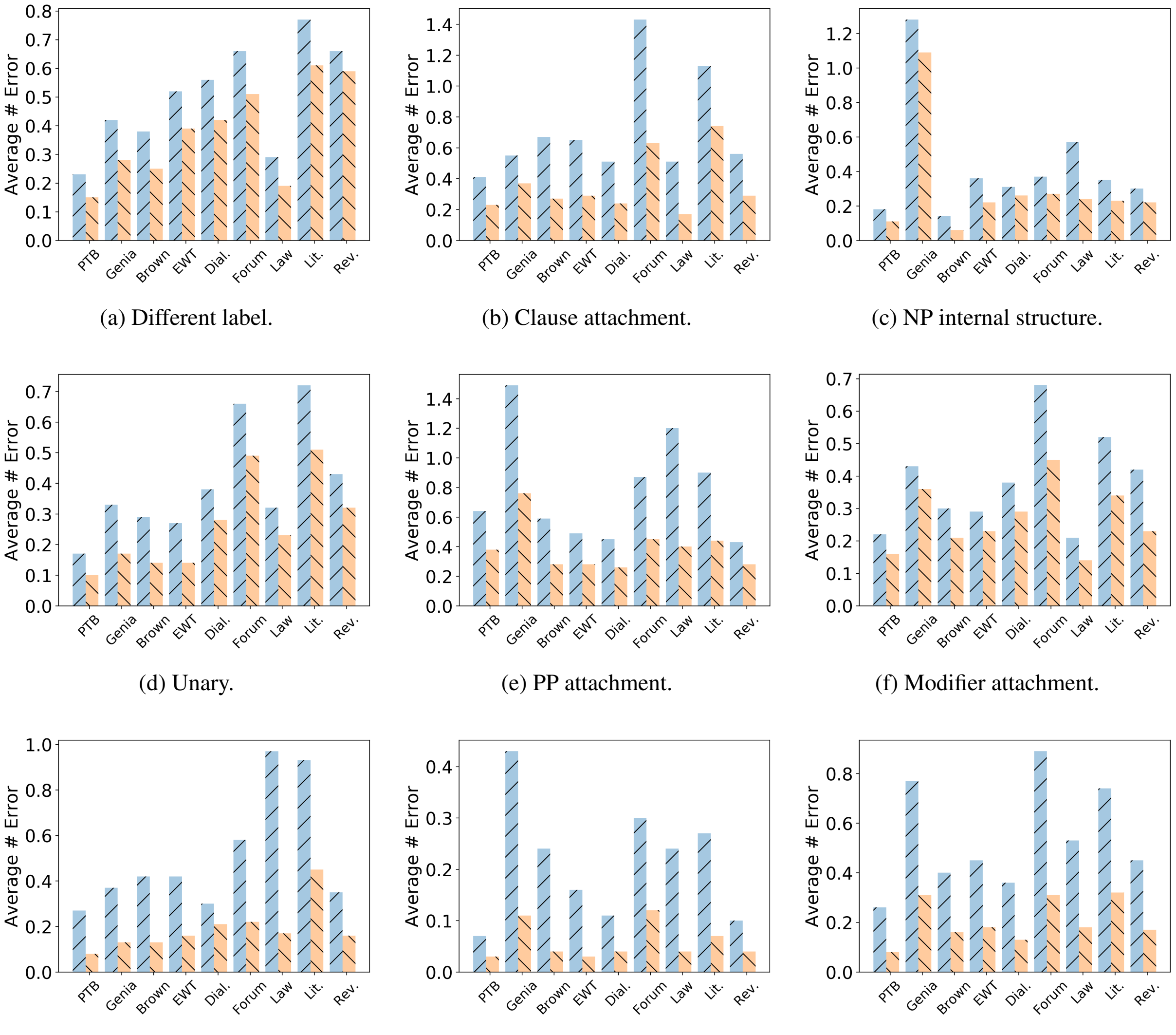
[8] https://github.com/brendano/parseviz

Dataset	N-gram Token (OOV Rate)			GR	HGR	GP	N-gram Constituent			
	Uni	Bi	Tri				Uni	Bi	Tri	Four
PTB-test	0.09 (0.03)	0.41 (0.33)	0.61 (0.72)	0.03	0.19	0.04	0.00	0.01	0.04	0.11
Genia-test	0.38 (0.26)	0.61 (0.71)	0.68 (0.94)	0.16	0.41	0.20	0.05	0.11	0.24	0.37
Brown-test	0.21 (0.06)	0.52 (0.48)	0.67 (0.87)	0.09	0.28	0.11	0.02	0.06	0.16	0.31
Brown-all	0.18 (0.07)	0.45 (0.48)	0.63 (0.87)	0.07	0.24	0.09	0.02	0.05	0.13	0.26
Brown-cf	0.18 (0.07)	0.49 (0.49)	0.66 (0.87)	0.06	0.24	0.08	0.01	0.03	0.11	0.24
Brown-cg	0.19 (0.06)	0.50 (0.48)	0.66 (0.87)	0.07	0.25	0.09	0.01	0.04	0.13	0.27
Brown-ck	0.24 (0.07)	0.53 (0.49)	0.67 (0.87)	0.10	0.30	0.13	0.03	0.07	0.17	0.33
Brown-cl	0.24 (0.06)	0.53 (0.45)	0.66 (0.85)	0.10	0.31	0.14	0.03	0.08	0.19	0.35
Brown-cm	0.27 (0.08)	0.57 (0.49)	0.68 (0.87)	0.11	0.33	0.14	0.03	0.08	0.20	0.37
Brown-cn	0.25 (0.07)	0.54 (0.50)	0.67 (0.88)	0.10	0.31	0.13	0.03	0.08	0.19	0.34
Brown-cp	0.24 (0.06)	0.53 (0.46)	0.66 (0.86)	0.10	0.31	0.13	0.03	0.08	0.18	0.34
Brown-cr	0.24 (0.08)	0.55 (0.49)	0.67 (0.87)	0.09	0.30	0.12	0.02	0.06	0.17	0.33
EWT-All-test	0.19 (0.09)	0.49 (0.49)	0.65 (0.86)	0.10	0.29	0.13	0.02	0.06	0.15	0.28
EWT-answers-test	0.27 (0.07)	0.56 (0.47)	0.67 (0.86)	0.13	0.36	0.17	0.04	0.10	0.22	0.39
EWT-email-test	0.27 (0.11)	0.56 (0.51)	0.67 (0.86)	0.12	0.37	0.17	0.03	0.08	0.22	0.39
EWT-newsgroup-test	0.22 (0.08)	0.55 (0.49)	0.67 (0.85)	0.09	0.32	0.12	0.02	0.05	0.15	0.28
EWT-reviews-test	0.27 (0.08)	0.56 (0.47)	0.67 (0.86)	0.12	0.36	0.16	0.03	0.09	0.21	0.37
EWT-weblog-test	0.23 (0.09)	0.55 (0.49)	0.67 (0.85)	0.09	0.31	0.11	0.02	0.05	0.15	0.30
BNC	0.22 (0.11)	0.54 (0.54)	0.67 (0.89)	0.08	0.30	0.10	0.02	0.05	0.12	0.25
Switchboard	0.26 (0.04)	0.49 (0.35)	0.63 (0.78)	0.20	0.39	0.24	0.09	0.16	0.31	0.47
Dialogue	0.28 (0.06)	0.58 (0.46)	0.68 (0.86)	0.16	0.39	0.21	0.03	0.09	0.23	0.41
Forum	0.25 (0.06)	0.55 (0.44)	0.67 (0.84)	0.14	0.36	0.18	0.03	0.09	0.23	0.41
Law	0.27 (0.07)	0.57 (0.51)	0.68 (0.86)	0.12	0.33	0.16	0.01	0.08	0.19	0.34
Literature	0.28 (0.11)	0.57 (0.53)	0.68 (0.90)	0.15	0.36	0.19	0.03	0.09	0.23	0.38
Review	0.30 (0.07)	0.59 (0.51)	0.68 (0.88)	0.16	0.39	0.21	0.03	0.10	0.26	0.45

Table 4: Dataset difference statistics between PTB training set and various test sets. We report Jensen–Shannon divergence of a list of linguistic features' distributions. These features are adopted from previous work (Collins and Koo, 2005; Charniak and Johnson, 2005). We report both divergence and out-of-vocabulary rate (OOV) for unigram/bigram/trigram input tokens. GR, HGR and GP refer to grammar rules, headed lexicalized grammar rules and grandparent rules.

Model	BLLIP	In-Order	Berkeley	With BERT (Δ Err.)	
Dataset				In-Order	Berkeley
PTB-test	91.48	91.53	93.05	95.65 (-48.6%)	95.73 (-38.6%)
Genia-test	78.42	81.06	81.39	86.33 (-27.8%)	86.61 (-28.0%)
Brown-test	85.78	85.74	87.72	93.68 (-55.7%)	93.38 (-46.1%)
Brown-all	85.89	86.55	87.37	93.55 (-52.0%)	93.31 (-47.0%)
Brown-cf	87.03	87.15	89.06	94.38 (-56.3%)	94.21 (-47.1%)
Brown-cg	85.41	85.86	87.79	93.48 (-53.9%)	93.33 (-45.4%)
Brown-ck	85.49	85.57	86.95	93.17 (-52.7%)	92.26 (-40.7%)
Brown-cl	85.51	85.78	87.15	92.76 (-49.1%)	92.49 (-41.6%)
Brown-cm	87.27	86.33	87.72	93.99 (-56.0%)	93.64 (-48.2%)
Brown-cn	86.85	86.59	88.24	94.19 (-56.7%)	93.88 (-48.0%)
Brown-cp	85.23	85.36	87.18	93.08 (-52.7%)	92.87 (-44.4%)
Brown-cr	84.34	85.23	87.23	93.44 (-55.6%)	92.98 (-45.0%)
EWT-All-test	78.78	81.19	81.98	89.39 (-43.6%)	89.09 (-39.5%)
EWT-answers-test	80.68	80.95	80.83	88.78 (-41.1%)	88.36 (-39.3%)
EWT-email-test	79.86	79.52	80.75	87.69 (-39.9%)	87.42 (-34.6%)
EWT-newsgroup-test	84.58	84.33	83.84	90.22 (-37.6%)	89.99 (-38.1%)
EWT-reviews-test	82.13	81.64	81.96	89.40 (-42.3%)	89.32 (-40.8%)
EWT-weblog-test	85.48	85.28	83.65	90.84 (-37.8%)	91.18 (-46.1%)
BNC	84.15	84.55	85.30	92.16 (-49.3%)	91.92 (-45.0%)
Switchboard	77.56	77.44	76.12	84.42 (-30.9%)	84.49 (-35.1%)
Dialogue	77.68	78.40	79.14	85.56 (-33.1%)	86.30 (-34.3%)
Forum	75.25	77.29	78.63	86.33 (-39.8%)	87.04 (-39.4%)
Law	80.67	82.83	84.06	91.50 (-50.5%)	92.06 (-50.2%)
Literature	70.32	76.44	75.98	84.96 (-36.2%)	86.26 (-42.8%)
Review	74.18	75.91	76.15	83.89 (-33.1%)	84.34 (-34.3%)

Table 5: Results (F1 scores) on various test sets. Δ Err. means error reduction rates when using BERT.



(a) Different label.

(b) Clause attachment.

(c) NP internal structure.

(d) Unary.

(e) PP attachment.

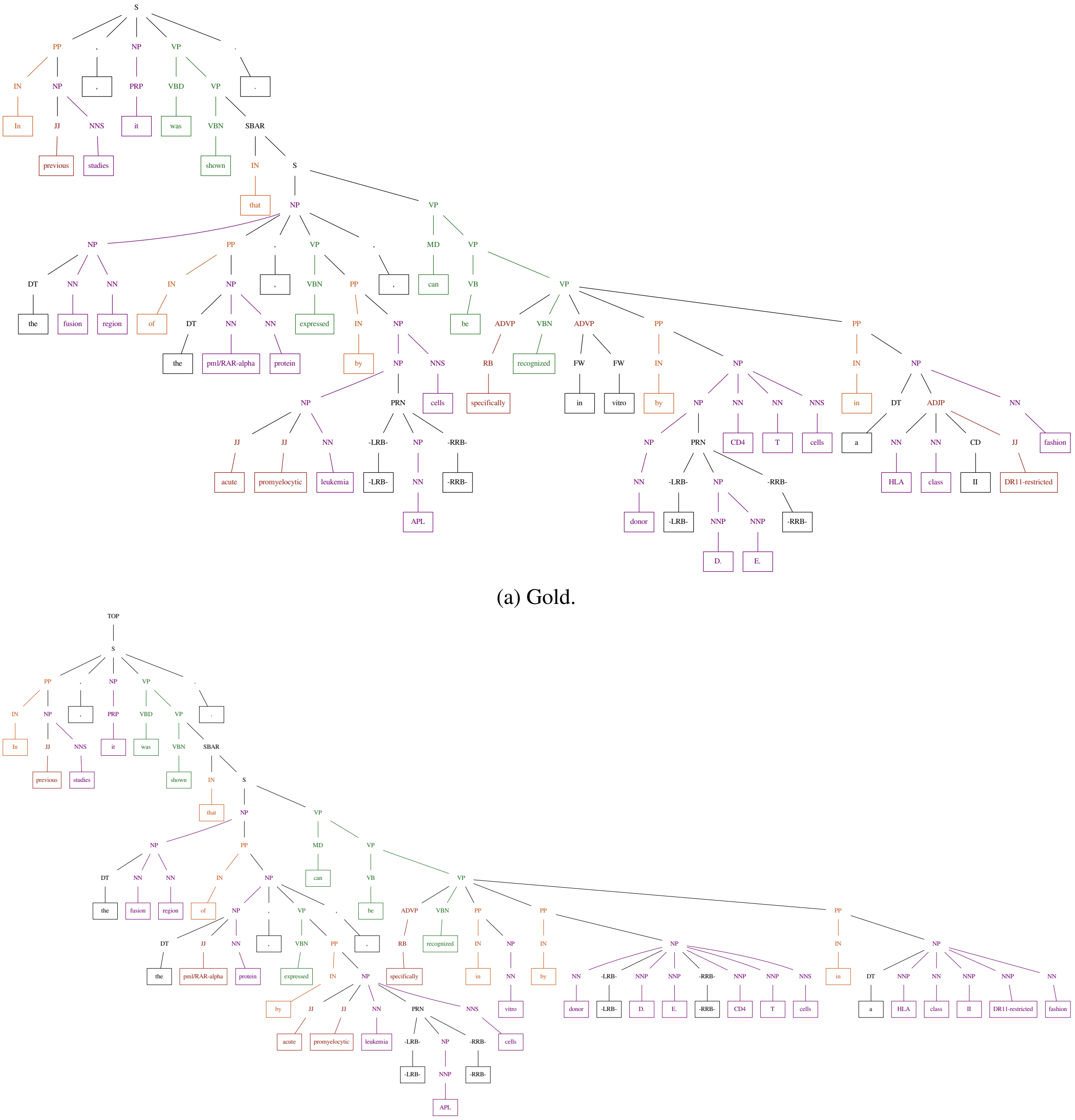
(f) Modifier attachment.

(g) NP attachment.

(h) VP attachment.

(i) Coordination.

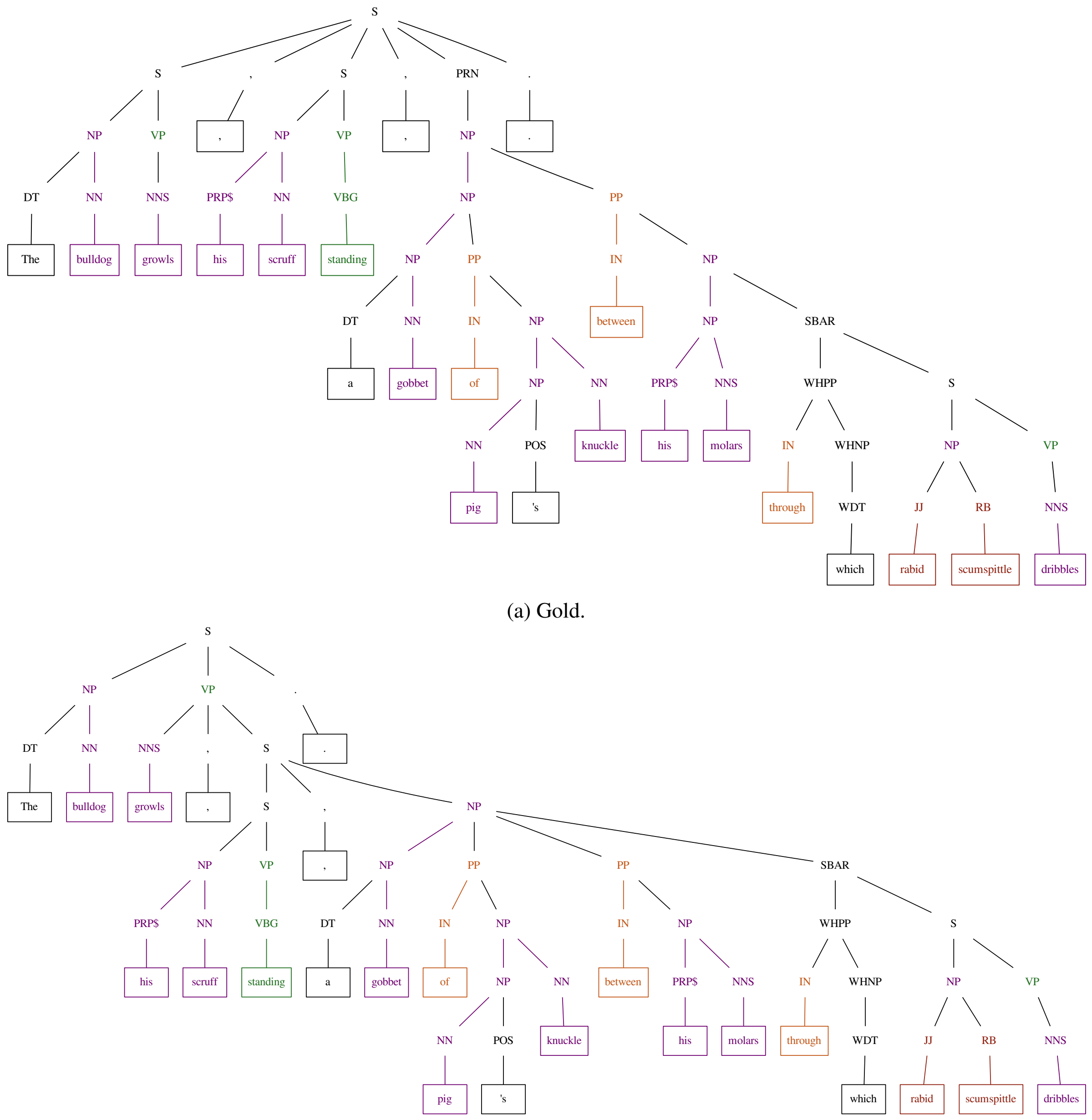
Figure 6: Average number of bracket errors per sentence on each dataset using the parser of Liu and Zhang (2017). The errors are classified with Kummerfeld et al. (2012)'s method. Blue bars with slash "/" are without BERT, while orange bars with backslash "\" are with BERT. "Dial. ", "Lit. " and "Rev. " are in short for dialogue, literature and review, respectively.


S
PP , NP VP .
IN NP , PRP VBD VP .
In JJ NNS it was VBN SBAR
previous studies shown IN S
that NP VP
NP PP , VP , MD VP
DT NN NN IN NP , VBN PP , can VB VP
the fusion region of DT NN NN expressed IN NP be ADVP VBN ADVP PP PP
the pml/RAR-alpha protein by NP NNS RB recognized FW FW IN NP IN NP
NP PRN cells specifically in vitro by NP NN NN NNS in DT ADJP NN
JJ JJ NN -LRB- NP -RRB- NP PRN CD4 T cells a NN NN CD JJ fashion
acute promyelocytic leukemia -LRB- NN -RRB- NN -LRB- NP -RRB- HLA class II DR11-restricted
APL donor -LRB- NNP NNP -RRB-
D. E.
TOP
S
PP , NP VP .
IN NP , PRP VBD VP .
In JJ NNS it was VBN SBAR
previous studies shown IN S
that NP VP
NP PP MD VP
DT NN NN IN NP can VB VP
the fusion region of NP , VP , be ADVP VBN PP PP PP
DT JJ NN , VBN PP , RB recognized IN NP IN NP IN NP
the pml/RAR-alpha protein expressed IN NP specifically in NN by NN -LRB- NNP NNP -RRB- NNP NNP NNS in DT NNP NN NNP NNP NN
by JJ JJ NN PRN NNS vitro donor -LRB- D. E. -RRB- CD4 T cells a HLA class II DR11-restricted fashion
acute promyelocytic leukemia -LRB- NP -RRB- cells
-LRB- NNP -RRB-
APL

(a) Gold.

(b) Predicted by in-order parser with BERT.

Figure 7: Genia NP internal structure error within the noun phrase “*a HLA class II DR11-restricted fashion*”. The in-order parser uses POS-tag information. We adopt a SOTA POS-tagger to predict POS-tags for the in-order parser. But the tagger is not able to generalize well to Genia, so that *DR11-restricted* is mistaken as NNP, which results in the in-order parser to make a wrong prediction (not identify the adjective phrase “*HLA class II DR11-restricted*”).


S
S , S , PRN .
NP VP , NP VP , NP .
DT NN NNS PRP$ NN VBG NP PP
The bulldog growls his scruff standing NP PP IN NP
DT NN IN NP between NP SBAR
a gobbet of NP NN NP WHPP S
NN POS knuckle PRP$ NNS IN WHNP NP VP
pig 's his molars through WDT JJ RB NNS
which rabid scumspittle dribbles
S
NP VP .
DT NN NNS , S .
The bulldog growls , S , NP
NP VP , NP PP PP SBAR
PRP$ NN VBG DT NN IN NP IN NP WHPP S
his scruff standing a gobbet of NP NN between PRP$ NNS IN WHNP NP VP
NN POS knuckle his molars through WDT JJ RB NNS
pig 's which rabid scumspittle dribbles

(a) Gold.

(b) Predicted by in-order parser with BERT.

Figure 8: An example from literature domain, including 2 Clause Attachment errors, 1 PP Attachment error and several other errors.