中文跨领域成分句法分析研究

第一作者
工作关系/ 地址一
工作关系/ 地址二
工作关系/ 地址三
email@domain

第二作者
工作关系/ 地址一
工作关系/ 地址二
工作关系/ 地址三
email@domain

摘要

成分句法树为机器翻译等诸多下游任务提供有用的句法分析。为了进一步探究中文成分句法分析器在公开领域上的泛化性能，我们在生物医药、法律文书、传统中医药、网络小说、商品描述这五个实用领域标注了高质量的成分句法分析数据集。经过定量分析，我们发现1）在大规模生语料上进行预训练的语言模型可以显著提升模型的跨领域能力；2) 与英文类似，中文成分句法分析领域泛化的难点更多在于复杂的句法结构，但同时一定程度上也与单一的字词特征相关；3）由于中文字、词的分布相比英文更为稀疏，中文成分句法分析与英文成分句法分析在跨领域上的难点存在差异。

关键词： 中文成分句法分析 ； 领域迁移

Research on Chinese Cross-domain Constituency Parsing

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

Constituency trees provide useful structural information for downstream tasks. To explore the cross-domain generalization ability of current Chinese constituency parsers, we annotate high-quality test treebanks on five diverse domains, including biomedical, law, traditional Chinese medicine, web fiction and item description texts. with quantitative analysis, we find that 1) large pretrained language models can significantly improve the cross-domain generalization ability of existing parsers. 2) Similar to English, the challenges to Chinese cross-domain constituency parsing lies more in complicated syntactic features, as well as single factors such as character and word-level features. 3) Challenges to Chinese cross-domain constituency parsing are somehow different from their English counterparts, because Chinese texts are composed of thousands of different characters and require word segmentation, which results in the sparsity of character and word distributions.

Keywords: Chinese Constituency Parsing , Domain Adaptation

1 引言

成分句法分析是自然语言处理的基础任务之一 (Klein and Manning, 2004; Charniak and Johnson, 2005; Cross and Huang, 2016; Stern et al., 2017)，长期以来广受学界的关注。给定输入文本，成分句法分析旨在根据语法规则将文本分割为层次化的短语结构，例如名词短语、动词短语、介词短语。一个准确率高、鲁棒性强的成分句法分析器可以帮助下游任务提供有用的结构信息，因此，研究跨领域场景下成分句法分析器的泛化性能成为了一个重要课题。Fried et al. (2019)等人的工作表明，尽管中英文的成分句法分析器在新闻领域内均有优异表现（超过95%的F1分数），但在中英文跨领域场景下进行测试时，成分句法分析器的性能均有显著下降；Yang et al. (2022)等人在英文数据集上系统分析了开放域中成分句法分析所面临的挑战，然而中文的开放域成分句法分析泛化性研究却相对较少。

相较于英文，由于不同的语言特性，中文的成分句法分析面临着独有的挑战，例如：句法分析中的基本单位为词，英文天然具有词与词之间的分隔符，而中文则需要对连续的字进行额外的分词预处理后再进行句法分析。因此，为了研究中文成分句法分析在跨领域时面临的独有挑战，我们在这篇论文中开展了一系列的探究。

为了深入进行探究，我们首先在生物医药、商品描述、法律、传统中医药、网络小说这五个多样化的领域标注了高质量的成分句法数据，这五个新数据集以及CTB7&8的多领域数据组成了我们评估分析的基础。首先，我们定量分析了不同领域与CTB-5.1训练集之间所存在的差异，我们使用Jensen–Shannon divergence来计算某个特征在跨领域数据上和在CTB-5.1训练集上的两个分布之间的差异；随后我们在多领域的测试集上评估了三个有代表性的句法分析器：基于统计的句法分析器ZPar (Zhang and Clark, 2011)，基于中序遍历转移的In-order (Liu and Zhang, 2017)句法分析器和基于图解码器的Berkeley Self-attentive句法分析器 (Kitaev and Klein, 2018; Kitaev et al., 2019)；最终，为了在模型的跨领域泛化性能和数据集的特征分布之间建立联系，我们计算了各类特征的分布差异和各个模型的泛化性能之间的皮尔逊相关系数，更大的相关系数表明该特征更能影响到该模型在跨领域场景中的泛化性能，利用这一相关性信息，我们分析了模型在跨领域场景下所面临的挑战。我们的发现包括：

首先，比起正式的新闻文本，中文成分句法里与新闻领域相似度最低的领域包括了口语对话和生物医药，这主要是因为对话文本中包含了更多口语化的简短成分句法短语，而生物医药文本则常有新闻文本中少见的专业学术短语。

其次，通过对比上述几个成分句法分析器在各领域数据集上的表现，我们发现基于神经网络的方法比基于统计的方法有小幅提升，而预训练BERT带来了极大的帮助。同时，尽管基于转移的方法和基于图解码的方法在领域内的表现相似，但跨领域场景下基于图解码的方法更有优势。

最后，我们发现，相较于英文中跨领域场景下统计方法和神经网络方法均更依赖于句法结构特征、而非单一的词特征，中文中跨领域句法分析的挑战既包括字词特征也包括句法结构特征，其中句法特征的重要性略微高于字词特征。在英文所不具备的字特征上，由于中文的常用字有约3000到4000个，分布较为稀疏，因此即便是先进的预训练语言模型也面临着难以有效建模不熟悉单字的问题。

为了帮助后续工作展开对中文成分句法分析领域泛化性的研究，我们的多领域测试集和代码将会公布在`https://anonymous.com`。

2 模型和设置

2.1 模型

我们使用基于统计的句法分析器（ZPar）、基于转移的句法分析器（In-order）、基于图的句法分析器（Self-attentive）作为模型进行分析实验。

基于统计方法的转移句法分析器 ZPar (Zhang and Clark, 2011)是一个基于统计的句法分析器，它采用了一个全局化的感知机模型和一个基于集束搜索（beam search）的解码器来预测一系列动作转移序列。在感知机端，它采用了多种全局和局部的特征向量；在解码端，它采用了增量式的解码，预测一系列的Shift-Reduce动作转移序列，并随后从中复原出成分句法树结构。

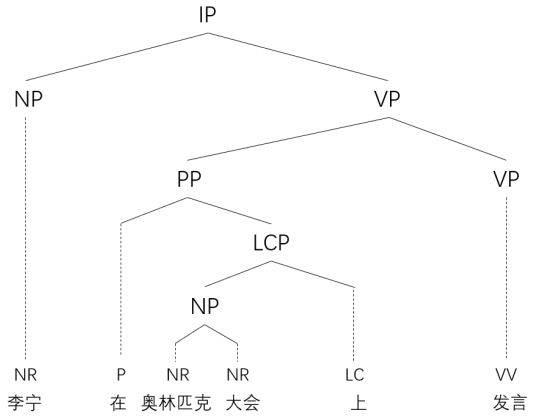


Figure 1: 一个成分句法树的例子。

基于神经网络的转移句法分析器 基于神经网络的状态转移序列句法分析器通过预测一系列的转移行动来预测成分句法树。Liu and Zhang (2017)等人使用中序遍历的方法来编码成分句法树。模型使用两个基于栈的LSTM编码分别编码预测的成分句法子树和缓存信息，并额外使用一个单向的LSTM编码已预测的转移行动序列。这样做模型可以直接对输出的树结构信息进行编码。

基于神经网络的图解码句法分析器 基于神经网络的图解码句法分析器通过神经网络编码器对输入文本的所有范围进行打分，然后使用例如CKY的动态规划算法寻找最可能的树。在本文中，我们使用自注意力网络（Self-attentive Parser）作为编码器的句法分析器 (Kitaev and Klein, 2018)进行实验。

2.2 实验设置

对于ZPar句法分析器，我们采用了网络上的开源版本以及已训练好的模型文件[0]。对于Self-attentive句法分析器，我们采用了开源的代码[1]，并在CTB-5.1上训练了不带BERT和带BERT (Devlin et al., 2019)的两个模型，选取了在CTB-5.1开发集上表现最好的模型。对于In-order句法分析器，我们仅选取了带BERT的开源版本[2]以及随代码公布的模型文件来与Self-attentive进行对比探究。

我们共在11个测试集上进行了实验，其中包括了一个领域内测试集（CTB-5.1的测试集 (Xue et al., 2005b)）和10个跨领域测试集，这10个跨领域测试集中有5个来自于CTB-7&8（广播对话、广播新闻、口语对话、网络论坛、网络博客），另有5个来自于我们新标注的多领域测试集（生物医药、商品描述、法律文书、传统中医药、网络小说）。

3 数据集

3.1 数据标注

我们选取生物、商品描述、法律、中医药、网络小说这五个领域的生文本进行成分句法树的标注。我们使用《新英格兰医学杂志[3]的文本作为生物领域文本。我们使用Chen et al. (2019)收集的淘宝电商标题[4]作为商品描述领域文本。我们使用https://www.cnopendata.com/爬取的《中国裁判文书网》作为法律领域文本。我们使用天池竞赛公布的问题生成数据集https://tianchi.aliyun.com/dataset/dataDetail?dataId=86895作为中医药领域文本。我们使用中文分词数据集 (Qiu and Zhang, 2015)作为网络小说领域文本。对于没有分词的文本，我们使用Tian et al. (2020)开源的中文分词器进行分词。

[0] https://github.com/frcchang/zpar/releases/download/v0.7.5/chinese-models.zip
[1] https://github.com/nikitakit/self-attentive-parser
[2] https://github.com/dpfried/rnng-bert
[3] https://github.com/boxiangliu/ParaMed
[4] https://www.taobao.com/

数据集	样本数量	Avg # char per sent	Avg # word per sent	Avg # cons per sent	Avg # char per cons	Avg # word per cons	# Total words	# Total cons	Max # word of sent	Avg # word of NP	Avg # word of VP	Avg # word of PP
CTB5.1-训练集	17,544	44.43	27.32	31.12	8.59	5.14	479,361	546,003	240	2.62	6.67	5.26
CTB5.1-开发集	352	32.79	19.38	20.89	8.27	4.79	6,821	7,352	101	2.70	7.09	5.04
CTB5.1-测试集	348	39.48	23.01	25.06	10.25	5.82	8,008	8,720	159	2.82	9.11	5.88
CTB7&8 广播对话	1,204	24.83	16.52	18.41	7.33	4.63	19,889	22,165	75	2.41	6.59	6.65
CTB7&8 广播新闻	1,008	55.52	32.94	38.45	9.55	5.50	33,204	38,753	209	2.66	8.03	5.37
CTB7&8 口语对话	1,674	14.46	10.01	10.81	5.01	3.32	16,764	18,088	50	1.51	3.22	3.81
CTB7&8 网络论坛	1,992	34.75	22.65	27.37	7.37	4.61	45,113	54,530	238	2.26	5.45	4.85
CTB7&8 网络博客	1,017	28.83	17.47	19.92	7.23	4.46	17,770	20,257	141	2.30	5.28	4.90
生物医药	1,000	57.99	29.53	30.86	10.26	5.10	29,28	30,865	153	3.09	6.45	6.64
商品描述	1,000	53.71	33.43	40.24	6.73	4.02	33,428	40,242	88	2.29	4.18	3.77
法律文书	1,000	45.01	25.78	29.37	9.06	4.98	25,784	29,366	75	2.53	6.60	5.33
传统中医药	1,000	45.19	29.75	34.58	7.60	4.81	29,751	34,577	111	2.35	5.19	5.05
网络小说	1,000	38.13	26.37	31.00	6.62	4.43	26,368	30,997	102	1.86	5.33	3.64

Table 1: 数据集的统计，其中“Avg”表示平均值（average），“# char”、“# word”和“# cons”分别表示字（character）、词（word）或成分句法短语（constituent）的数量，“Sent”表示句子（sentence）。

我们遵循CTB的标注规范 (Xue et al., 2005a)设计标注手册。我们邀请了语言学专业高年级本科生和研究生作为数据集的标注人员。标注人员被要求使用标注平台[5]修正模型解码的成分句法树中的错误。我们遵循以下步骤进行数据集的标注：

1. 标注人员被要求仔细阅读标注规范，并练习修正使用模型(Cui et al., 2022)解码的成分句法树。

2. 标注人员并要求在100条模型解码的CTB测试集上进行联系，直到修改的结果与真实结果的F1值超过95%。只有通过该测试的标注人员才被允许进行正式的标注。

3. 正式标注中，我们将数据分为每100条一组，每组中随机选择10%的数据进行抽检。为了对标注质量进行监控，本文的主要作者将重新标注这10%的样本。如果作者标注的成分句法树与标注人员标注的是F1值小于95%，这组数据将被丢弃，并分配给另一位标注人员重新进行标注。

3.2 数据统计

表1展示了现有数据集和我们标注的测试集的数据统计。我们使用CTB-7&8的官方划分。具体的数据统计包括样本个数、每句话的平均词数、每句话的平均成分句法数、每个成分句法范围的平均词数、总词数量、总成分句法数、每个训练样本包含的最长和最短词数、名词短语中包含的平均词数、动词短语中包含的平均词数、介词短语中包含的平均词数。

我们发现CTB-5.1的开发集和测试集与CTB-5.1的训练集有显著差异。例如CTB-5.1训练集每句的平均长度为27.32而开发集仅为19.38，训练集中动词的平均词数仅为6.67而测试集中动词的平均词数为9.11。

CTB-7&8广播新闻领域的数据和我们标注的商品描述领域数据的平均句子长度最长。虽然CTB-7&8广播新闻领域的数据的平均长度只比CTB 5.1多5个，但它每句话的平均成分句法数比CTB-5.1平均多8个。CTB-7&8的口语对话领域的每个成分句法包含的词数最少，这是因为人们在日常口语中经常使用缩略语。同理，口语对话领域中动词短语、名词短语、介词短语中包含的词数也最少。我们标注的商品描述领域数据也包含很多简短的动词短语，因为商品描述主要由名词短语、形容词短语和缩写的动词短语组成。

最重要的是，我们新标注的五个领域的测试集与已有数据集显著不同。

3.3 特征相关性

表 2和3展示了CTB-5.1训练集合和多个测试集在不同语言学特征(Collins and Koo, 2005; Charniak and Johnson, 2005)上的差异。表中每个值为该特征在CTB-5.1训练集和该领域测试集的JS散度（Jensen-Shannon divergence）。JS散度的计算公式如下：

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

[5] https://github.com/Nealcly/AnnoCons

Dataset	n元字(未登录词比率)			n元词(未登录词比率)		
	Uni	Bi	Tri	Uni	Bi	Tri
CTB5.1-开发集	0.15 (0.00)	0.49 (0.19)	0.64 (0.57)	0.30 (0.08)	0.60 (0.50)	0.68 (0.85)
CTB5.1-测试集	0.10 (0.00)	0.43 (0.12)	0.60 (0.43)	0.25 (0.04)	0.57 (0.38)	0.67 (0.77)
CTB7&8 广播对话	0.16 (0.01)	0.45 (0.20)	0.63 (0.62)	0.29 (0.07)	0.58 (0.51)	0.68 (0.89)
CTB7&8 广播新闻	0.05 (0.00)	0.33 (0.19)	0.57 (0.59)	0.17 (0.08)	0.51 (0.51)	0.66 (0.88)
CTB7&8 口语对话	0.35 (0.01)	0.59 (0.46)	0.68 (0.84)	0.44 (0.12)	0.64 (0.67)	0.69 (0.95)
CTB7&8 网络论坛	0.08 (0.00)	0.37 (0.23)	0.60 (0.67)	0.20 (0.08)	0.52 (0.54)	0.67 (0.90)
CTB7&8 网络博客	0.16 (0.00)	0.48 (0.27)	0.65 (0.71)	0.27 (0.12)	0.58 (0.56)	0.68 (0.90)
Ours 生物医药	0.28 (0.01)	0.56 (0.47)	0.67 (0.86)	0.43 (0.29)	0.66 (0.84)	0.69 (0.99)
Ours 商品描述	0.18 (0.00)	0.51 (0.35)	0.66 (0.82)	0.34 (0.18)	0.61 (0.70)	0.68 (0.96)
Ours 法律文书	0.18 (0.00)	0.52 (0.30)	0.66 (0.77)	0.35 (0.16)	0.63 (0.70)	0.69 (0.95)
Ours 传统中医药	0.17 (0.01)	0.50 (0.37)	0.65 (0.80)	0.31 (0.18)	0.59 (0.67)	0.68 (0.94)
Ours 网络小说	0.17 (0.01)	0.50 (0.34)	0.66 (0.80)	0.31 (0.17)	0.60 (0.64)	0.68 (0.94)

Table 2: 输入端上的跨领域测试集与源领域（即CTB-5.1训练集）在多个语言学特征上的差异，其中包括了n元字、词、成分句法短语的JS散度和未登录字词比率。这些语言学特征中的一部分取自于前人在基于统计的句法分析上的工作 (Collins and Koo, 2005; Charniak and Johnson, 2005)。更多细节可见于章节 3.3。

Dataset	GR	HGR	GP	n元成分句法短语			
				Uni	Bi	Tri	Four
CTB5.1-开发集	0.05	0.45	0.08	0.01	0.04	0.10	0.19
CTB5.1-测试集	0.04	0.41	0.06	0.01	0.03	0.09	0.17
CTB7&8 广播对话	0.10	0.45	0.13	0.04	0.11	0.25	0.40
CTB7&8 广播新闻	0.02	0.33	0.04	0.00	0.01	0.05	0.12
CTB7&8 口语对话	0.20	0.52	0.24	0.09	0.19	0.36	0.49
CTB7&8 网络论坛	0.04	0.34	0.05	0.01	0.03	0.07	0.14
CTB7&8 网络博客	0.06	0.43	0.08	0.01	0.05	0.14	0.25
Ours 生物医药	0.10	0.52	0.13	0.02	0.08	0.21	0.33
Ours 商品描述	0.08	0.49	0.10	0.02	0.06	0.15	0.23
Ours 法律文书	0.06	0.52	0.09	0.01	0.03	0.10	0.18
Ours 传统中医药	0.06	0.46	0.08	0.02	0.04	0.10	0.17
Ours 网络小说	0.09	0.47	0.11	0.02	0.08	0.15	0.23

Table 3: 输出端上的跨领域测试集与源领域（即CTB-5.1训练集）在多个语言学特征上的差异，GR、HGR和GP分别指代了语法规则（grammar rule）、词法指示的语法规则（headed lexicalized grammar rule）和串联成分句法节点规则（a chain of [grandparent, parent, child] constituents）。其余细节见于表 2标题部分和章节 3.3。

其中P和Q为两个概率分布，$KL(P||Q) = \sum_{x\in\chi} P(x)\log(\frac{P(x)}{Q(x)})$为分布$P$对分布$Q$的KL散度，$M = \frac{1}{2}(P+Q)$为分布$P$和分布$Q$的平均值。表中的所有值为0到1之间，更高的值表示该特征在PTB训练集和测试集差异越大。

表 2主要囊括了输入文字的特征，包括n元中文字和中文词；表 3主要囊括了输出句法结构的特征，例如一些语法规则和n元成分句法结构。在表中，Uni、Bi和Tri和Four分别代表了一元的、二元的、三元的和四元的字（character）、词（word）或成分句法短语（constituent）；GR、HGR和GP分别代表了语法规则（grammar rule）、词法指示的语法规则（headed lexicalized grammar rule）和串联成分句法节点规则（a chain of [grandparent, parent, child] constituents）。以图 1为例，IP → NP VP 即为一个语法规则（GR），PP [在] → P LCP为一个词法指示的语法规则（HGR），IP → VP → PP VP为一个串联成分句法节点规则（GP）。成分句法节点的n元组（n-gram）是在每个语法规则内部形成的。语法规则是非二叉化的规则。表 2同时还包括了n元的字和词的未登录字词比例。

CTB-7&8的口语对话和我们的生物医药这两个领域与CTB-5.1的训练集在一元和二元的字和词的分布上有着最大的散度值，说明了在日常口语和生物医药领域文本中有许多模型不常见到的词汇。CTB-7&8的口语对话领域同样在其他许多语言学特征上与CTB-5.1的训练集有着极大的差异，例如语法规则（GR）、串联成分句法节点规则（GP）、一元/二元/三元/四元的成分句法节点特征等。在词法指示的语法规则（HGR）上，CTB-7&8的口语对话领

数据集 \ 模型	ZPar	Self-attentive	使用BERT	
			In-Order	Self-attentive
CTB5.1-测试集	81.06	84.36	91.08	91.98
CTB7&8 广播对话	62.55	67.10	71.52	73.16
CTB7&8 广播新闻	74.15	78.51	86.76	87.31
CTB7&8 口语对话	51.38	52.03	64.07	65.15
CTB7&8 网络论坛	72.97	76.41	86.56	87.22
CTB7&8 网络博客	66.92	72.34	79.33	80.89
Ours 生物医药	48.19	55.07	72.50	78.10
Ours 商品描述	64.30	68.05	82.95	84.70
Ours 法律文书	68.79	73.36	85.45	88.04
Ours 传统中医药	68.10	74.11	85.98	87.74
Ours 网络小说	66.38	71.40	81.72	84.24

Table 4: 不同句法分析器在多个测试集上的结果（F1分数）。

域和我们的生物医药领域同时具有最大的散度值，即0.52。CTB-7&8的广播新闻领域和网络论坛领域是与CTB-5.1的训练集最为相近的两个领域，甚至比CTB-5.1的开发集和测试集更为接近CTB-5.1的训练集，例如在n元词的特征上，CTB-7&8的广播新闻领域与CTB-5.1训练集的JS散度为0.17，而CTB-5.1测试集与其训练集的JS散度则为0.25。

现有工作 (Yang et al., 2022)在英语上探讨了这些语言学特征在跨领域上与源领域的差异性。通过对比这些特征的跨领域差异性在英语和汉语上的不同表达，我们发现，在领域内测试集（即CTB-5.1测试集或PTB测试集）与训练集（即CTB-5.1训练集或PTB训练集）之间的差异性上，汉语在词相关的特征上明显大于英语，例如在一元词分布的JS散度上，英语的十余个数据集中最高的为0.30，而汉语的CTB-7&8口语对话和我们的生物医药这两个领域上则均超过了0.40。这可能是由于汉语字序列需要经过分词这道工序才能组成词，因此汉语的词分布远比英语词分布更加多样化。之前工作在英语上概括了如下结论：跨领域成分句法的难点不在于单一的词特征，例如未登录词比率或是n元词分布的变化等。然而，基于如上观察，我们对于这一结论是否对汉语同样成立产生了疑问，在本文随后的章节里我们将解答这一疑问。

4 实验与分析

4.1 初步结果

表 4给出了四个句法分析器在多个不同领域的测试集上的各自表现。在领域内的测试集（即CTB-5.1测试集）上，基于统计方法的ZPar和基于神经网络的Self-attentive Parser分别取得了约81%和84%的F1分数，而在使用了BERT之后，Self-attentive Parser的表现得到了较大提升，取得了约91%的F1分数，这表明了大规模的预训练对于成分句法分析任务有巨大的帮助。在跨领域的测试集中，最佳的句法分析器（即，使用了BERT的Self-attentive Parser）的表现在65.51%到88.04%的F1分数区间内。在所有领域中，较为简单的是CTB-7&8的广播新闻领域和网络论坛领域、我们的数据集的法律领域和传统中医药领域，使用了BERT的句法分析器能达到85%以上的F1分数；而CTB-7&8的口语对话领域、广播对话领域以及我们的生物医药领域是最为困难的，使用了BERT的句法分析器的表现最高也仅在75%的F1分数左右。相比与在领域内的良好表现，跨领域成分句法的场景下，最佳的句法分析器的相对错误率增长也高至49%—330%。上述结果表明了，对于中文成分句法分析来说，跨领域的场景仍是一个亟待解决的难题。

比较Yang et al. (2022)等人在英文跨领域成分句法分析的结果，可以发现：无论是中文或是英文，对话相关领域都是最为困难的。这主要归因于PTB和CTB-5.1的源领域是新闻文本，行文较为正式，而口语化的文本行文更为随便，常常伴有句子结构缺失、使用缩略词等现象，同时，对比之下中文和英文中法律文本领域的跨领域表现都较高，这主要是因为法律文本和新闻文本都属于正式写作的范畴。

4.2 成分句法分析器之间的对比

在领域内，基于统计方法的ZPar和基于神经网络的Self-attentive Parser有一定差距，体现在错误率上为17.4%的相对下降；而在跨领域场景下，错误率的相对下降幅度出现了一定波

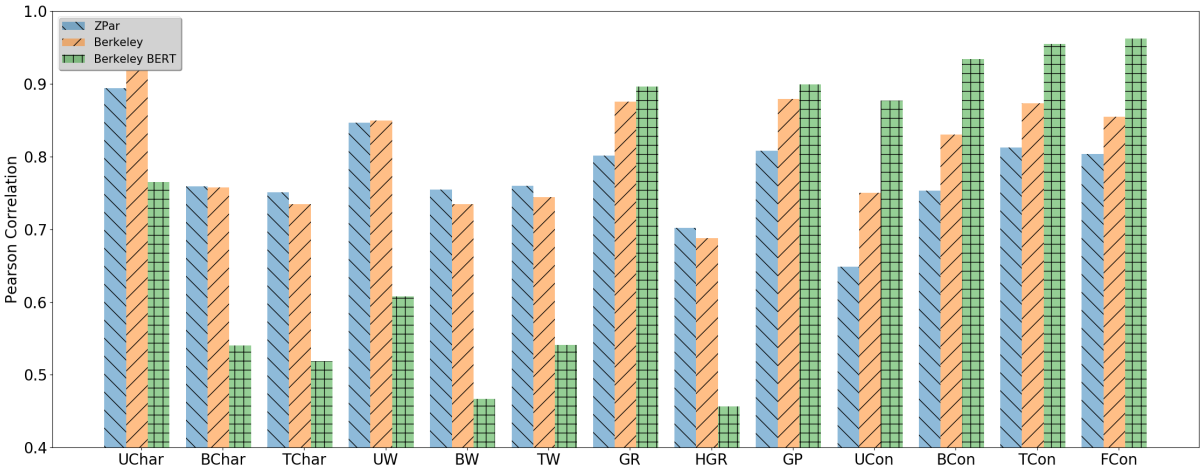


Figure 2: 特征的JS散度和句法分析器的表现之间的皮尔逊相关系数。由于所有值均小于0（代表了负相关），为了让可视化的数据更加易于理解，我们将所有皮尔逊相关系数乘上了−1，图中更高的值代表了与该语言学特征具有更高的相关性。UChar、BChar、TChar分别代表了一元、二元、三元的中文的字（character）特征；UW、BW、TW分别代表了一元、二元、三元的中文的词（word）特征；GR、HGR和GP分别代表了语法规则（grammar rule）、词法指示的语法规则（headed lexicalized grammar rule）和串联成分句法节点规则（a chain of [grandparent, parent, child] constituents）；UCon、BCon、TCon、FCon分别代表了一元、二元、三元、四元的成分句法短语（constituent）特征。

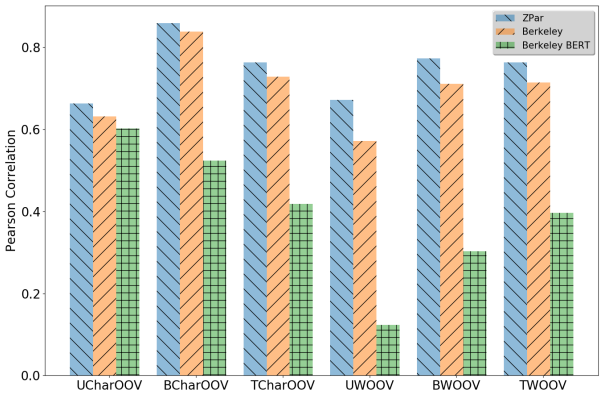


Figure 3: 中文字和词的未登录比率和句法分析器的表现之间的皮尔逊相关系数，其中UChar、BChar、TChar分别加上OOV代表了一元、二元、三元的中文字的未登录比率，UW、BW、TW分别加上OOV代表了一元、二元、三元的中文词的未登录比率。具体细节可见于图 2的标题。

动，在1.34%到18.84%之间，其中错误率下降最小的领域为CTB-7&8的口语对话领域，这说明引入神经网络的稠密化表征并不能帮助正式写作的语言风格向口语化风格进行迁移。

通过对比不使用和使用了BERT的Self-attentive Parser，可以发现：在领域内，BERT带来的提升较为显著，F1分数从84.36%升至91.98%，错误率的相对下降达到了48.72%；同时在跨领域场景下，其带来的提升更是尤其显著，例如，其带来的错误率相对下降幅度在法律领域为55.11%、在生物医药领域为51.26%。这些提升主要来源于两方面：一方面中文BERT的预训练语料主要爬取自维基百科，而维基百科中文本的行文较为偏向书面语，契合了法律领域的语言风格，同时维基百科内有大量生物医药、化学等自然科学类条目，能够为生物医药领域的成分句法分析提供预训练的背景知识；另一方面，表 2显示生物医药领域的未登录词问题最为严重，而BERT使用了基于上下文的表征，使得模型能够通过上下文中的常见词来推断未登录词的信息，因此BERT能够一定程度上缓解生物医药等领域的未登录词问题。

接下来对比In-order和Self-attentive，两者都使用了基于BERT微调的编码器。区别在于，In-order在解码的时候，会建模已经部分解码出来的树结构以帮助后续的预测；而Self-attentive使用了一个不建模树结构的、无参数化的解码器。观察表 4的后两列可以发现：在领域内，In-order比Self-attentive的错误率相对高了约10%，绝对差距为0.90的F1分数；而在跨领域上，前者的表现普遍弱于后者，F1分数的绝对差距在0.55 到5.60之间，其中差距最小的是CTB-7&8的广播新闻领域，这可能是因为广播新闻领域与CTB5.1的领域最为接近，因此两者差距较小。

4.3 跨领域成分句法分析中的关键点

图 2和图 3展示了成分句法分析器的F1分数（见表 4）与语言学特征的JS散度（见表 2和3）之间的皮尔逊相关系数（Pearson Correlation）。具体来说，表 4中每一列作为一个向量，代表了该句法分析器在各个领域上的F1分数，向量中每个元素即为在某领域上的F1分数；表 2或3中每一列也作为一个向量，代表了以某种语言学特征作为指标的情况下各个领域与CTB5.1的源领域之间的JS散度，向量中每个元素即为某领域与CTB5.1的源领域在该特征下的JS散度。通过计算这两个向量之间的皮尔逊相关系数，我们得到图 2，柱状图中每一个柱子都代表着该特定的成分句法分析器与该特定的语言学特征之间的皮尔逊相关系数，即：柱子越高，代表着这个成分句法分析器越依赖于该语言学特征。

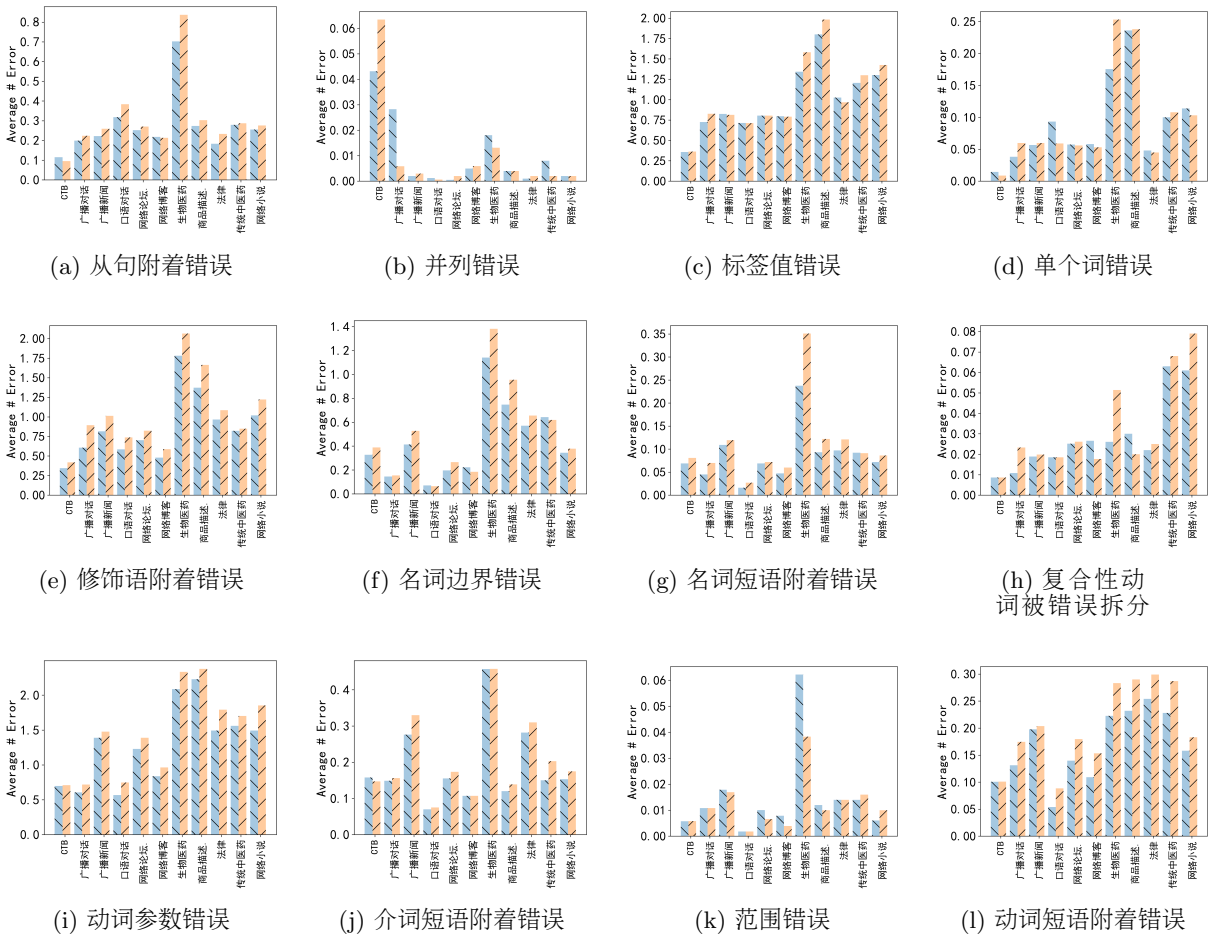
观察图 2和3可以得到如下发现：

1. 基于统计方法的ZPar句法分析器在各语言学特征上的依赖程度较为一致，而基于神经网络的Self-attentive句法分析器、以及其使用了BERT的版本则都更倾向于依赖语法结构特征，而非词级别的特征。Yang et al. (2022)等人在英语上分析发现统计方法会更倾向于语法特征而非词级别特征，可见他们这一结论在汉语上是不成立的，这可能是由于汉语字词有非常多样的组合方式，导致了词特征更加稀疏，因此汉语中的统计方法能够更好利用到稀疏而有辨别度的词信息。

2. 对比基于统计的句法分析器（即ZPar）和基于神经网络的句法分析器（即Self-attentive）可以发现，两者对词级别相关特征的依赖程度非常相似，例如UW、BW、HGR；但对于句法结构特征来说，其与神经网络方法的相关性明显高于与传统统计方法的相关性，这说明了：在运用汉语词级别信息的能力上，传统统计方法与神经网络方法表现相似；但是在处理句法结构特征的能力上，神经网络方法明显优于基于统计的方法，这可能是由于神经网络模型能够建模连续而稠密的词表征，增强了其对于信息的编码能力。

3. 对比使用了BERT前后的Self-attentive句法分析器，可以看到，使用了BERT的模型对词级别的信息的相关性系数下降非常明显，例如UW、BW、UO和HGR等，与此同时对句法结构相关特征的相关性系数有大幅上升，例如UCon、BCon、TCon、FCon和GR等。这说明了BERT极大地缓解了跨领域成分句法分析中的词分布迁移问题，并且更好地建模了句法短语结构之间的相互依赖关系。这可能来源于BERT的编码表征是上下文相关的，通过将常见词与长尾词（即少见词）相关联进行建模，使得模型能通过上下文来增强词信息表征。Yang et al. (2022)等人在英语的跨领域成分句法分析上发现：BERT能通过建模跨领域场景下更鲁棒信息（如句法结构信息）来提升跨领域的成分句法分析性能。参照表 2和3我们发现，这一结论对于汉语也是同样成立的：在表 3中，UCon、BCon等句法结构信息的跨领域JS散度较小，普遍在0.01到0.10之间，而表 2中UW、BW等词级别信息的跨领域JS散度要大上许多，普遍在0.20到0.70之间。

4. 不同于英语中输入文本一般只有词级别的特征，中文有字、词两个细粒度的特征。对比这两个细粒度的特征可以发现：对于ZPar和Self-attentive来说，它们对于字的依赖程度和对于词的依赖程度相似，这一点无论是在字、词的分布（图 2）以及未登录比率（图 3）上均有所体现；然而，当使用了BERT之后，Self-attentive对于词的依赖程度大幅下降，但对于字的依赖程度降幅却不那么明显，尤其是在一元字的未登录比率（UCharOOV）上，加与不加BERT的Self-attentive的相关系数差距很小，约在0.03左右。



(a) 从句附着错误 (b) 并列错误 (c) 标签值错误 (d) 单个词错误

(e) 修饰语附着错误 (f) 名词边界错误 (g) 名词短语附着错误 (h) 复合性动词被错误拆分

(i) 动词参数错误 (j) 介词短语附着错误 (k) 范围错误 (l) 动词短语附着错误

Figure 4: 在Self-attentive和In-order句法分析器的预测结果上的错误分析，错误分析工具来自于Kummerfeld et al. (2013)等人公开的代码[6]。蓝色右斜线“\”图案表示Self-attentive而橘色左斜线“/” 图案表示In-order。

4.4 错误分析

在图 4中，我们展示了分别使用Self-attentive和In-order句法分析器的错误分析情况，两个句法分析器都使用了BERT。首先可以看到，在多数错误类型下Self-attentive句法分析器产生的错误比In-order的要更少，而一个例外为范围错误。其次，平均错误数量比较多的错误类型包括了标签值错误、修饰语附着错误、动词参数错误、动词短语附着错误等。其中最严重的动词参数错误，其所产生的主要原因是动词所对应的对象含有歧义，例如：

一个错误的例子为

```
(CP
    (IP
        (VP
            (VV 提高)
            (NP
                (NN 队员)
            )
        )
        (DEG 的)
    )
        (NP
            (NN 心肺) (NN 功能) (CC 和) (NN 身体) (NN 素质)))
```

而正确的成分句法结构则为

```
(VP
    (VV 提高)
    (NP
        (DNP
```

```
        (NP
            (NN 队员)
        )
        (DEG 的)
    )
    (NP
        (NN 心肺) (NN 功能) (CC 和) (NN 身体) (NN 素质))))
```

5 相关工作

现有的中文成分句法工作主要在宾州中文树库（Penn Chinese Treebank）上进行实验，其中，CTB-5.1为单一领域的树库，其源文本来自于新闻领域（包括了新华社的新闻报道、香港新闻处的新闻、台湾光华杂志社的文章），而CTB-7&8的源文本来自多个领域，包括广播新闻、广播对话、口语对话、网络论坛以及网络博客。在CTB-7&8的基础上，我们的工作增添了几个新领域的测试树库，包括生物医药、商品描述、法律文书、传统中医药和网络小说，这些新领域为中文跨领域成分句法分析带来了新的挑战。

现有工作亦探讨过跨领域场景下的成分句法分析。McClosky et al. (2010)等人研究了多模型融合的领域泛化，具体来说：首先在多个不同领域上各自训练一个成分句法分析器，其次训练一个线性回归模型来预测各个模型的权重分配。Joshi et al. (2018)等人探究了在目标领域有少量部分标注样本的场景下，成分句法模型如何进行泛化，他们同时发现，上下文相关的表征 (Peters et al., 2018; Devlin et al., 2019)有助于成分句法分析器进行跨领域泛化。Fried et al. (2019)在英文和中文上对跨领域成分句法分析展开了系统分析，他们发现：首先，基于统计方法的成分句法分析器和基于神经网络的成分句法分析器的跨领域泛化能力相似；其次，大规模预训练 (Devlin et al., 2019)能够帮助成分句法分析模型更好进行领域泛化；最后，显式建模了结构的In-order句法分析器在跨领域泛化上优于不显式建模结构的Self-attentive句法分析器。Yang et al. (2022)等人在英文上对跨领域成分句法分析所面临的难题进行了系统性分析，他们发现：英文跨领域成分句法分析的挑战更多在于对于复杂句法结构的建模，而非对输入词级别信息的编码。他们同时还发布了一个英文的多领域成分句法树库用于测试模型的跨领域泛化性能。

6 结论

我们系统性地探究了影响中文成分句法分析器跨领域泛化性能的关键因素。具体来说，我们利用现有的多领域树库、以及标注新的跨领域测试树库，测试了几个具有代表性的句法分析器。同时我们设计实验来发掘各类特征与句法分析器的跨领域泛化性能之间的联系。我们的实验结果表明：1) 中文成分句法分析的跨领域泛化仍旧是一个亟待解决的问题；2) 与英文类似，中文成分句法分析领域泛化的难点也更多在于复杂的句法结构；3) 不同于英文的是，中文成分句法分析器在跨领域场景下对句法结构特征和字词特征都表现出了一定程度的相关性，尤其是对于一元字这样分布稀疏且不均衡的特征。

参考文献

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, June. Association for Computational Linguistics.

Qibin Chen, Junyang Lin, Yichang Zhang, Hongxia Yang, Jingren Zhou, and Jie Tang. 2019. Towards knowledge-based personalized product description generation in e-commerce. *arXiv preprint arXiv:1903.12457*.

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, November. Association for Computational Linguistics.

Leyang Cui, Sen Yang, and Yue Zhang. 2022. Investigating non-local features for neural constituency parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2065–2075, Dublin, Ireland, May. 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, June. Association for Computational Linguistics.

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, July. 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, July. 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, July. 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, July. Association for Computational Linguistics.

Dan Klein and Christopher Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 478–485, Barcelona, Spain, July.

Jonathan K. Kummerfeld, Daniel Tse, James R. Curran, and Dan Klein. 2013. An empirical examination of challenges in chinese parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 98–103, Sofia, Bulgaria, August.

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

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, June. Association for Computational Linguistics.

Matthew E. 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, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Likun Qiu and Yue Zhang. 2015. Word segmentation for chinese novels. In Blai Bonet and Sven Koenig, editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2440–2446. AAAI Press.

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, July. Association for Computational Linguistics.

Yuanhe Tian, Yan Song, Xiang Ao, Fei Xia, Xiaojun Quan, Tong Zhang, and Yonggang Wang. 2020. Joint chinese word segmentation and part-of-speech tagging via two-way attentions of auto-analyzed knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8286–8296, Online, July.

Naiwen Xue, Fei Xia, Fu-dong Chiou, and Marta Palmer. 2005a. The penn chinese treebank: Phrase structure annotation of a large corpus. *Nat. Lang. Eng.*, 11(2):207–238, June.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005b. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207–238.

Sen Yang, Leyang Cui, Ruoxi Ning, Di Wu, and Yue Zhang. 2022. Challenges to open-domain constituency parsing. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 112–127, Dublin, Ireland, May. Association for Computational Linguistics.

Yue Zhang and Stephen Clark. 2011. Syntactic processing using the generalized perceptron and beam search. *Comput. Linguist.*, 37(1):105–151, March.

Figure 5: 标注界面的截图。

A 附录

A.1 标注工具

我们标注所使用的标注工具[7]界面如图 5所示，其中左边为可视化的成分句法树结构，右上角为成分句法树中的非叶子节点，右下角为叶子节点，标注人员的目标即为：通过修改右上角部分的“（成分句法标签，起始位置，终止位置）”三元组来纠正成分句法树结构中的错误。

[7]该标注工具来自于开源代码库https://github.com/Nealcly/AnnoCons。