



大模型时代句法语义研究何去何从

李正华

<http://web.suda.edu.cn/zhli13>

语析LAGroup研究方向

- 句法语义分析（2007年-）
- 文本纠错（2020年-）
- Text2SQL（2020年-）[构建数据集DuSQL、SeSQL]
- 语音文本融合处理（2022年-）
- 欢迎关注我们的微信公众号“语析LAGroup”

语析LAGroup团队情况（博士后、在读学生）

合作青年教师：[龚晨](#)（AI研究院师资博士后）

在读学生 (字母序)

2020级博士(1): 刘亚慧（山东农大考研18硕，直博）

2021级博士(2): 张宇（苏大考研18硕）、周厚全（矿大保研19硕，直博）

2022级博士(2): 侯洋（苏大保研20硕、直博）、周仕林（苏大20硕，直博）

2020级硕士(4): 侯洋（苏大保研、直博）、李嘉诚（燕山大学、专硕）、李帅克（苏大保研）、李扬（苏大）、刘泽洋（华北电力、专硕）、周仕林（苏大、专硕直博）

2021级硕士(6): 崔秀莲（苏大保研）、窦晨晖（苏大）、黄赛豪（苏大保研）、严福康（苏大）、张磊（文正）、章岳（苏大保研）

2022级硕士(4): 辜仰淦（广东工业保研）、蒋浩辰（苏大）、刘雨萌（江南大学保研）、马溪（太原理工）

团队情况（毕业生）

2017级博士(1): 李英（昆明理工考博，昆明理工任教）[2022.06]

2018级博士(2): 龚晨（苏大保研16硕、直博，苏大任教）、夏庆荣（苏大保研16硕、直博，华为）

2019级硕士(3): 沈嘉钰（苏大）、杨浩苹（苏大保研、唯品会实习、思必驰）、周厚全（矿大保研、直博）、周明月（苏大保研）

2018级硕士(4): 蒋炜（苏大、华为Cloud&AI）、陆凯华（苏大、专硕、上海银行）、吴锬（浙江理工、百度NLP）、张宇（苏大、读博）、刘亚慧（山东农大，直博）

2017级硕士(3): 黄德朋（苏科技，华为杭州实习，小红书上海）、江心舟（苏大保研，百度北京实习）、彭雪（山东农大，华为杭州实习、移动苏州）、章波（苏大，阿里巴巴杭州达摩院实习、转正）

2016级硕士(3): 郭丽娟（江西财经保研，科沃斯实习，狗尾草公司工作）、孙佳伟（北航，搜狗北京实习，微软苏州工作）、朱运（山西大学，搜狗北京工作）、龚晨（苏大保研、直博）、夏庆荣（苏大保研、直博）

2015级硕士(3): 陈伟（南阳理工、爱奇艺北京实习、爱奇艺北京工作）、凡子威（滁州学院、科大讯飞北京实习、搜狗北京工作）、张月（苏大保研、阿里巴巴杭州实习、阿里巴巴杭州工作）

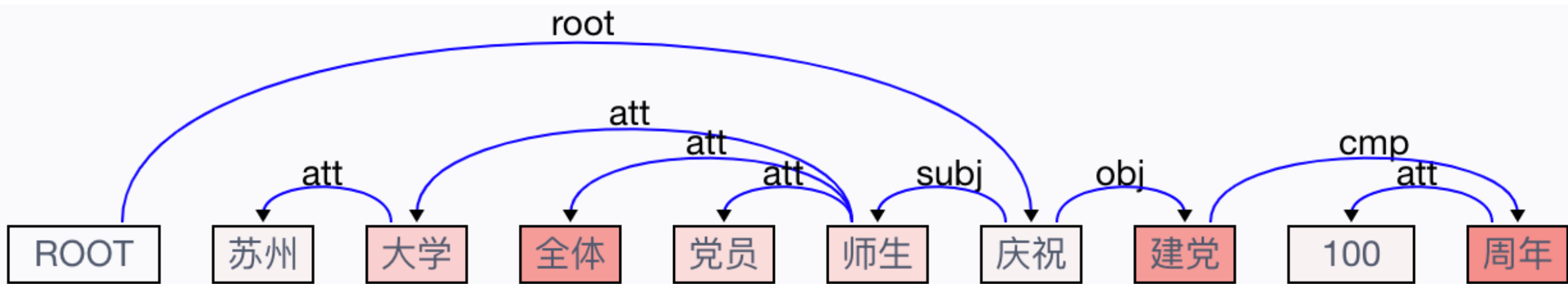
2014级硕士(1)[2017春毕业]: 巢佳媛（苏大、微软北京实习、阿里巴巴杭州工作）

2020年的几个工作

依存句法分析 (<http://hlt-la.suda.edu.cn>)

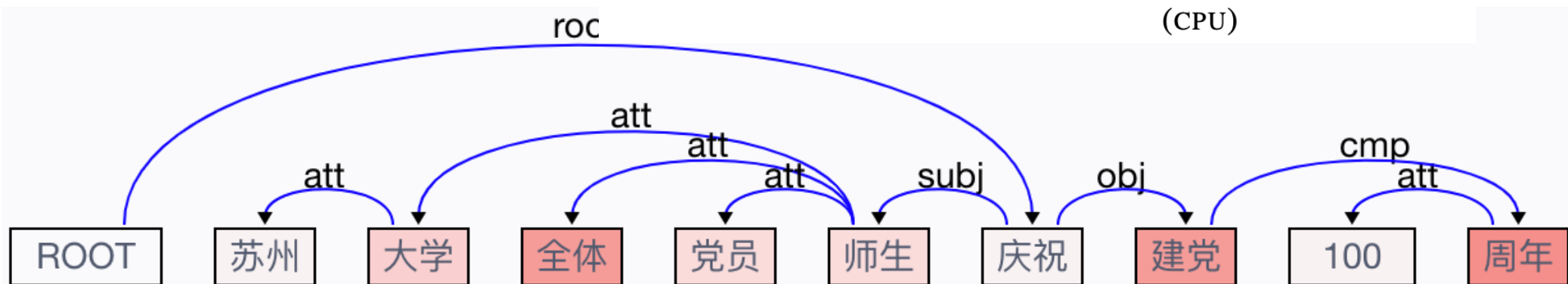
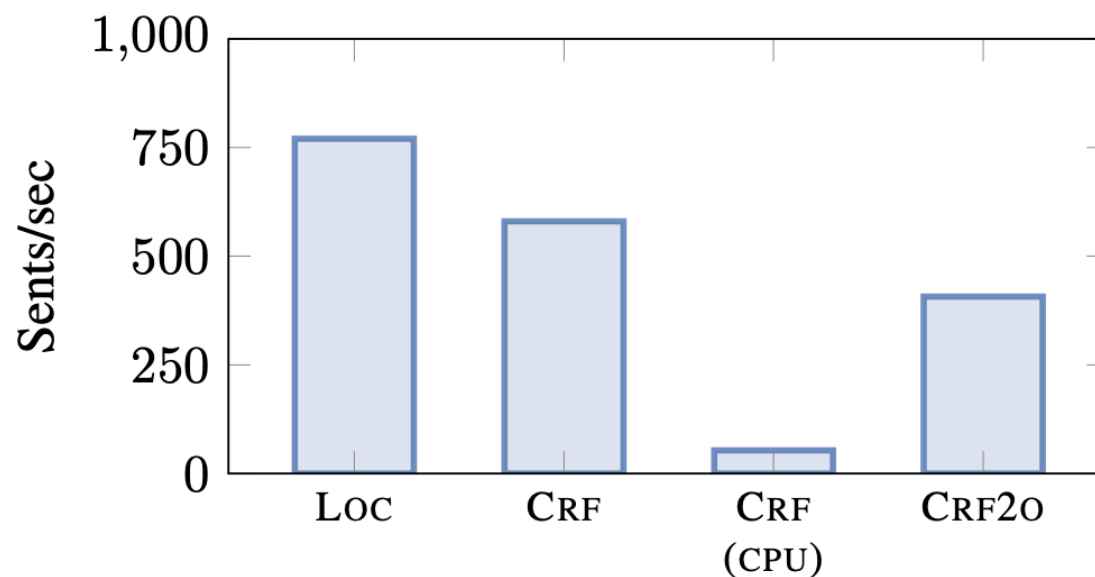
- 快速高精度TreeCRF Parser (ACL-2020)

	bg	ca	cs	de	en	es	fr	it	nl	no	ro	ru	Avg.
UD2.2													
using raw text													
Ji19	88.28	89.90	89.85	77.09	81.16	88.93	83.73	88.91	84.82	86.33	84.44	86.62	85.83
CRF2O	89.72	91.27	90.94	78.26	82.88	90.79	86.33	91.02	87.92	90.17	85.71	92.49	88.13



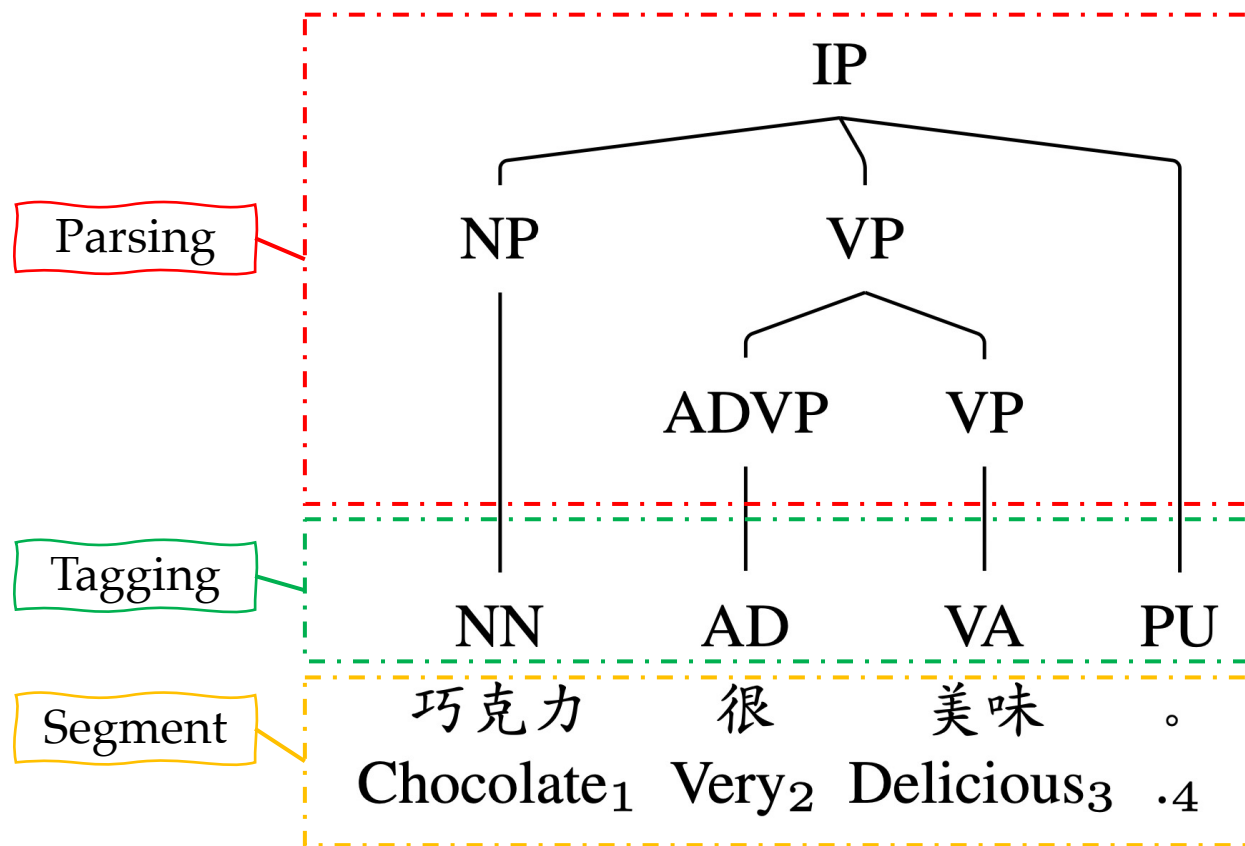
依存句法分析 (<http://hlt-la.suda.edu.cn>)

- 快速高精度TreeCRF Parser (ACL-2020)



短语句法分析

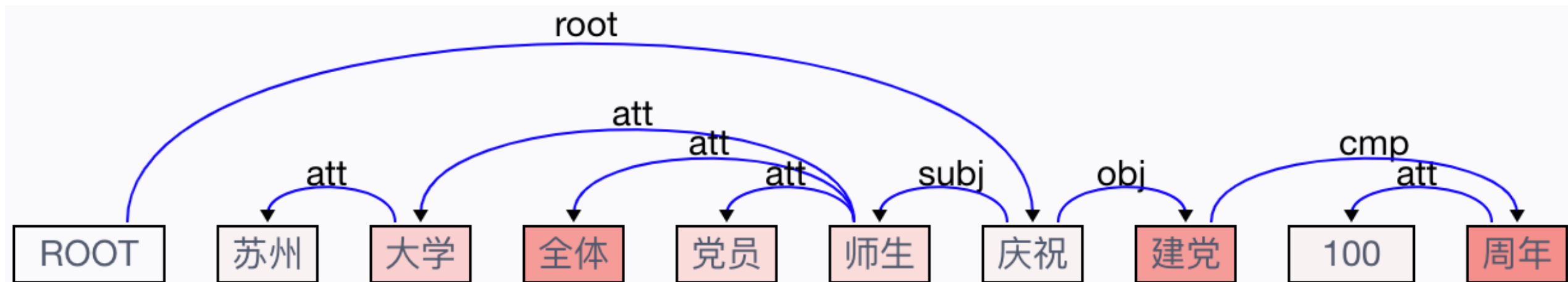
- 快速高精度TreeCRF Parser，速度达到1000句/秒（IJCAI-2020）



Word-level tree

依存句法分析 (<http://hlt-la.suda.edu.cn>)

- 构建多领域汉语开放依存树库CODT（16万句、覆盖21个领域）
- 举办两届跨领域句法分析评测（ NLPCC-2019、CCL-2020），发布6个领域、7万句数据。
- 做了一些跨领域句法分析的尝试（ACL-2019等）



开源句法语义工具包SuPar（星标600+，下载2万+次）

github.com/yzhangcs/parser

yzhangcs / parser Public

Notifications Fork 123 Star 671

<> Code Issues Pull requests Discussions Actions Projects Security Insights

main 2 branches 7 tags Go to file Code

yzhangcs Bump pkg versions d8f2e2e yesterday 427 commits

.github/workflows	Make the generated files readable by others	2 months ago
docs	Update index.md	3 weeks ago
supar	Add naive impls of cumsum/prod for semirings	last month
tests	Constituency CRF with labels	last month
.flake8	SuPar v1.0.0	2 years ago
.gitignore	Ignore yaml configs/other logs	8 months ago
CITATION.cff	Create CITATION.cff	8 months ago
EXAMPLES.md	Many pretrained embeddings available	7 months ago
LICENSE	Update LICENSE	7 months ago
README.md	Update badges	last month
setup.py	Bump pkg versions	yesterday

About

🚀 State-of-the-art syntactic/semantic parsers with pretrained models for many languages.

🔗 parser.yzhang.site

transformers pytorch semiring

dependency-parsing

structured-prediction state-of-the-art

constituency-parsing

semantic-dependency-parsing

📖 Readme

📄 MIT license

🔗 Cite this repository

☆ 671 stars

👁 16 watching

🍴 123 forks

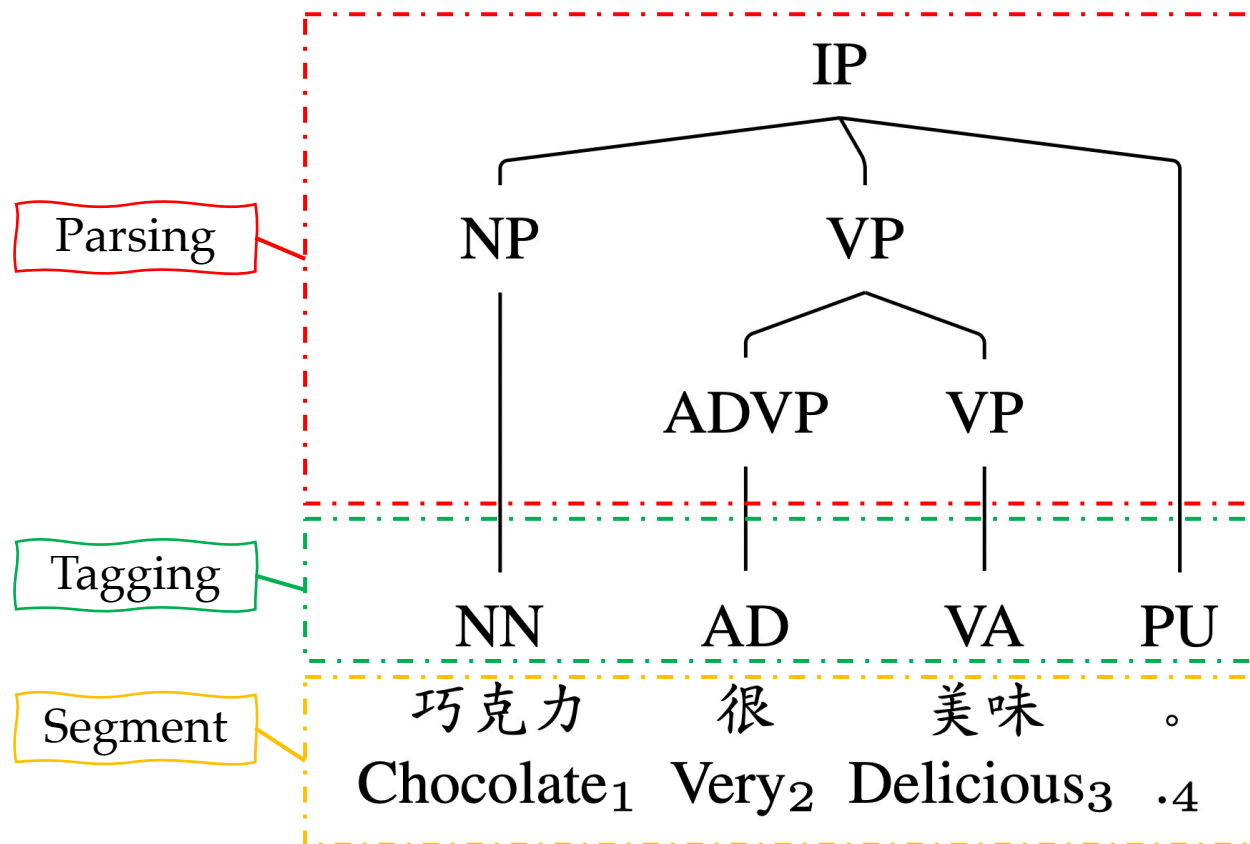
2021年的两个工作

A Coarse-to-Fine Labeling Framework for Joint Word Segmentation, POS Tagging, and Constituent Parsing (CoNLL-2021)

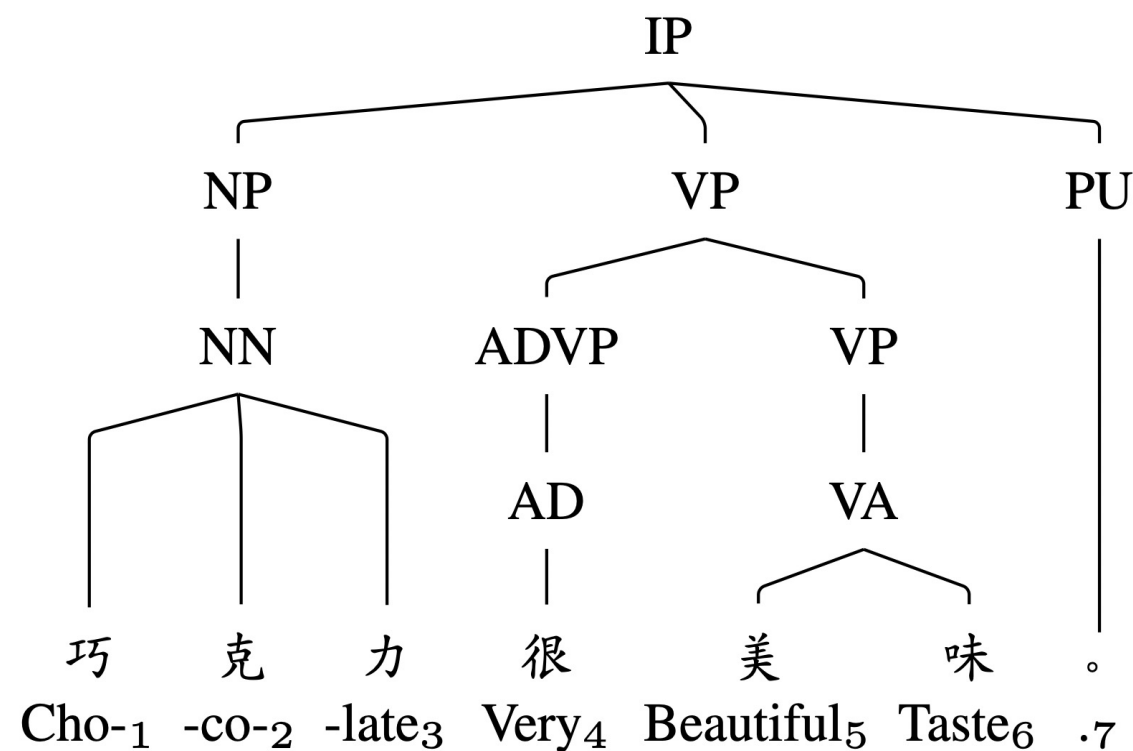
Yang Hou, Houquan Zhou, Zhenghua Li*, Yu Zhang,
Min Zhang, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan



联合建模：char-level parsing

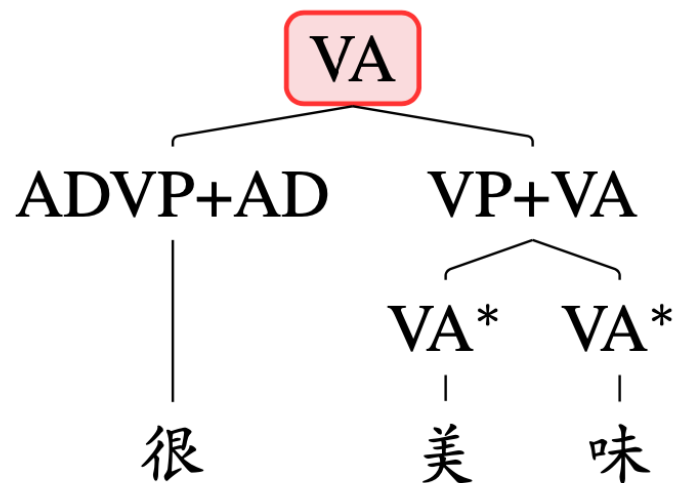


Word-level tree

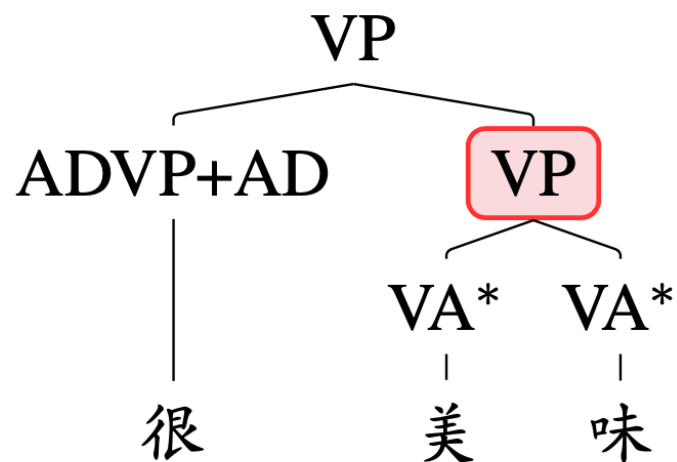


Char-level tree

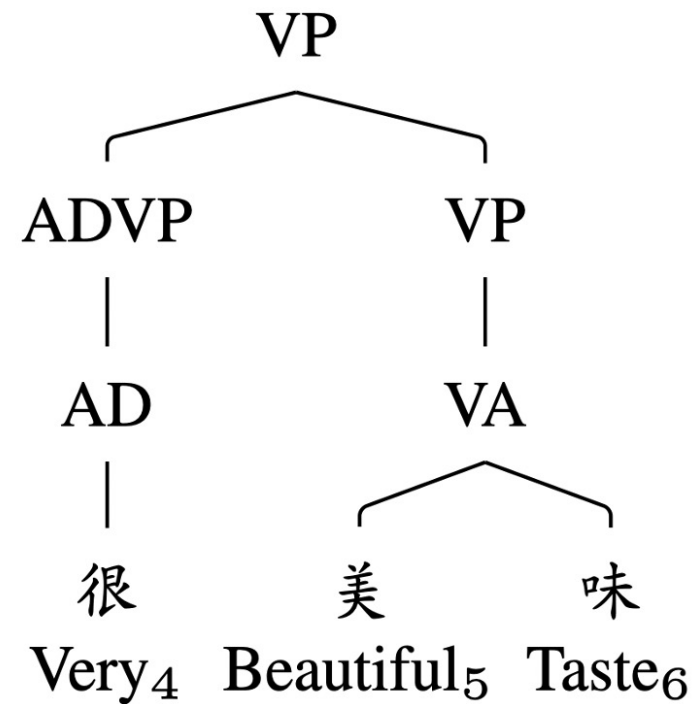
挑战：rule conflict



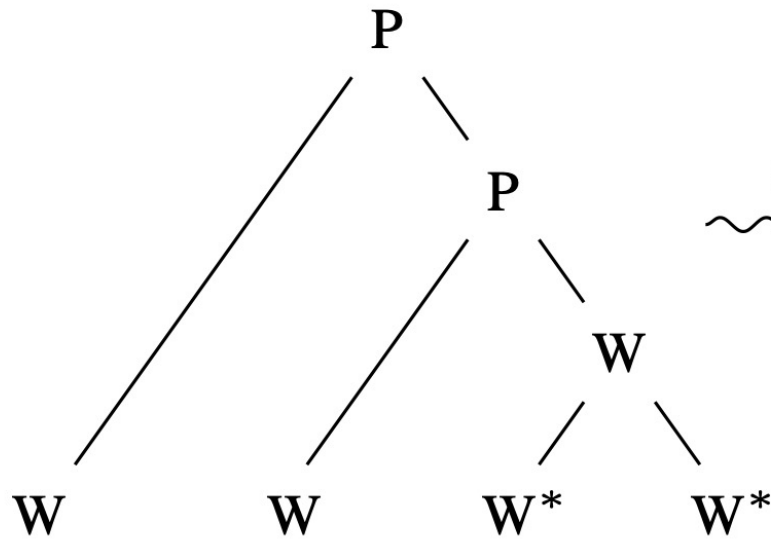
(b) POS tags above phrase tags



(c) Missing a POS tag

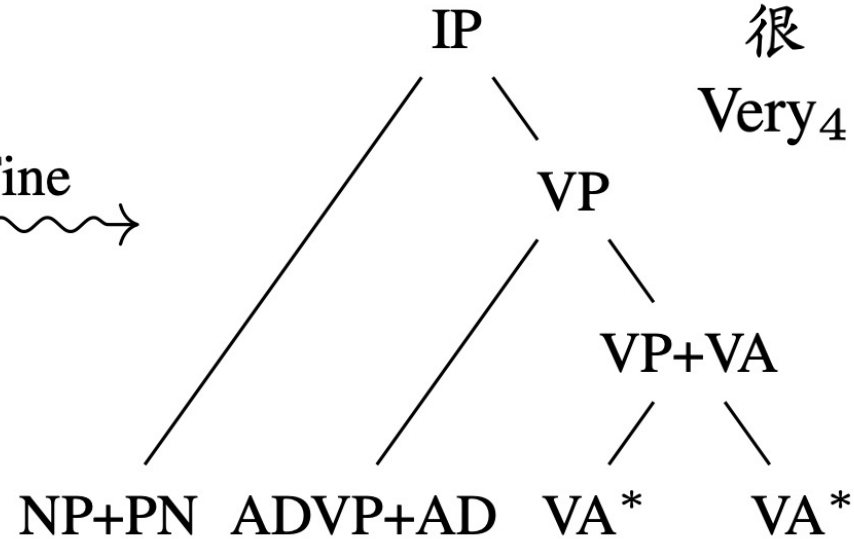


Two-stage Coarse-to-Fine

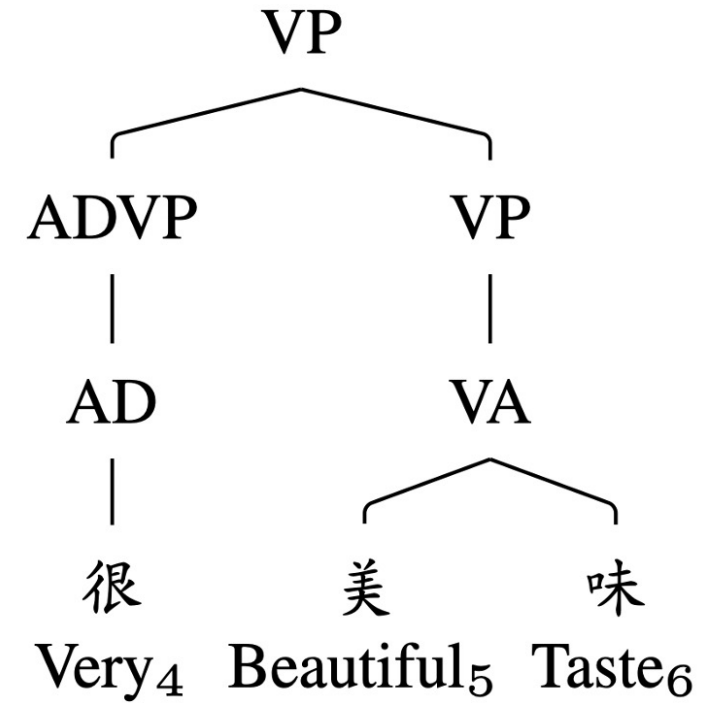


Coarse-grained tree

Coarse-to-Fine
~~~~~>



Fine-grained tree



# 贡献

- 针对联合分词-词性-短语句法分析任务，提出一种快速准确的二阶段解析方法
- 第一阶段粗糙解析时，通过约束解码来解决label conflict问题
- 联合模型 **vs.** 级联方法
  - 性能更高
  - 速度相当
  - 参数量减少一半

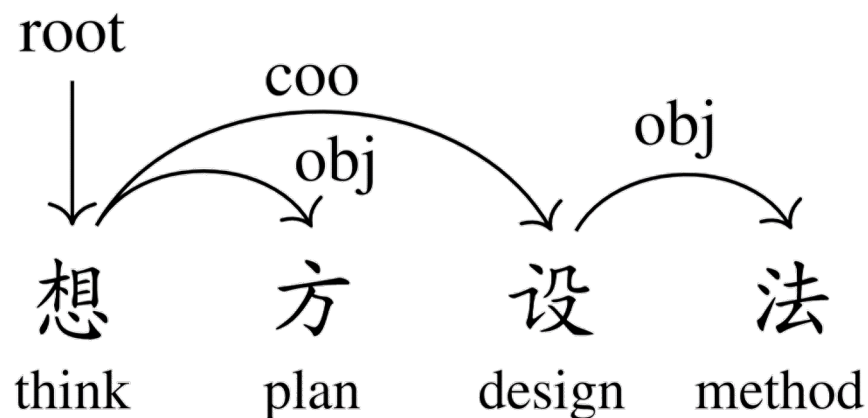
# An In-depth Study on Internal Structure of Chinese Words (ACL-2021)

Chen Gong, Saihao Huang, Houquan Zhou, Zhenghua Li\*,  
Min Zhang, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan



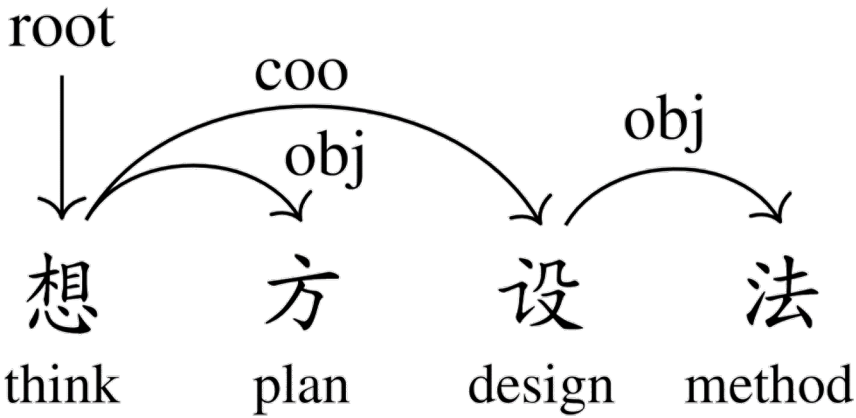
# 动机

- 我们可否像句子级句法分析一样，去分析词语内部搭配结构？
  - 进而尝试理解构词法（分析构词规律）
- 词内部搭配结构对句子级句法分析是否会有帮助？（罕见词理解）
- 词内部搭配结构对词语表示是否会有帮助？（罕见词理解）



# 我们的工作

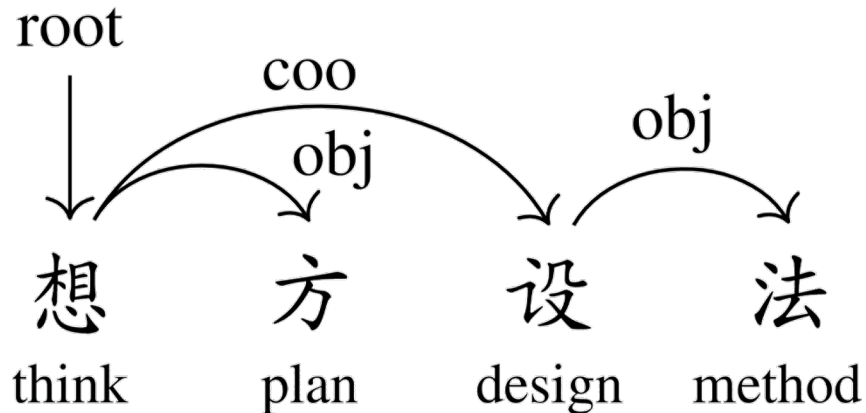
- 标注了CTB5中3万多个字词的词内部搭配结构



| Label | Meaning             | Example | Annotation                                                    |
|-------|---------------------|---------|---------------------------------------------------------------|
| root  | word root           | 登场      | \$ $\xrightarrow{\text{root}}$ 登 $\xrightarrow{\text{obj}}$ 场 |
| subj  | subject             | 年轻      | 年 $\xleftarrow{\text{subj}}$ 轻                                |
| obj   | object              | 下雨      | 下 $\xrightarrow{\text{obj}}$ 雨                                |
| att   | attribute modifier  | 大衣      | 大 $\xleftarrow{\text{att}}$ 衣                                 |
| adv   | adverbial modifier  | 不同      | 不 $\xleftarrow{\text{adv}}$ 同                                 |
| cmp   | complement modifier | 放下      | 放 $\xrightarrow{\text{cmp}}$ 下                                |
| coo   | coordination        | 上下文     | 上 $\xrightarrow{\text{coo}}$ 下                                |
| pobj  | preposition object  | 到期      | 到 $\xrightarrow{\text{pobj}}$ 期                               |
| adjct | adjunct             | 走过      | 走 $\xrightarrow{\text{adjct}}$ 过                              |
| frag  | fragment            | 沙发      | 沙 $\xrightarrow{\text{frag}}$ 发                               |
| repet | repetition          | 常常      | 常 $\xrightarrow{\text{repet}}$ 常                              |

# 我们的工作

- 作为一个新任务，做了一些基本实验（准确率不是很高，人工标注一致率83% dep, 76% word）



|            | Dev   |       | Test  |       |        |
|------------|-------|-------|-------|-------|--------|
|            | UAS   | LAS   | UAS   | LAS   | CM     |
| Random     | 81.18 | 76.15 | 80.63 | 75.58 | 65.13  |
| Pretrained | 82.42 | 77.30 | 81.64 | 76.98 | 67.09  |
|            | +1.24 | +1.15 | +1.01 | +1.40 | +1.96  |
| BERT       | 88.27 | 85.18 | 88.33 | 84.98 | 77.72  |
|            | +5.85 | +7.88 | +6.69 | +8.00 | +10.63 |

# 我们的工作

- 词内部搭配信息增强的词表示，对句法分析的帮助（BERT?）

|                | UAS          | LAS          | CM           |
|----------------|--------------|--------------|--------------|
| Basic CharLSTM | 88.31        | 85.96        | 32.04        |
| LabelCharLSTM  | 88.78        | 86.51        | <b>33.19</b> |
| LabelGCN       | <b>89.02</b> | <b>86.76</b> | 32.93        |
| w/o label      | 88.66        | 86.28        | 32.20        |

|                | all   | $> 2$ | $\leq 2$ | unk.  |
|----------------|-------|-------|----------|-------|
| Basic CharLSTM | 85.96 | 86.42 | 82.03    | 81.73 |
| LabelGCN       | 86.76 | 87.10 | 83.79    | 84.30 |
|                | +0.80 | +0.68 | +1.76    | +2.57 |

# 论文贡献

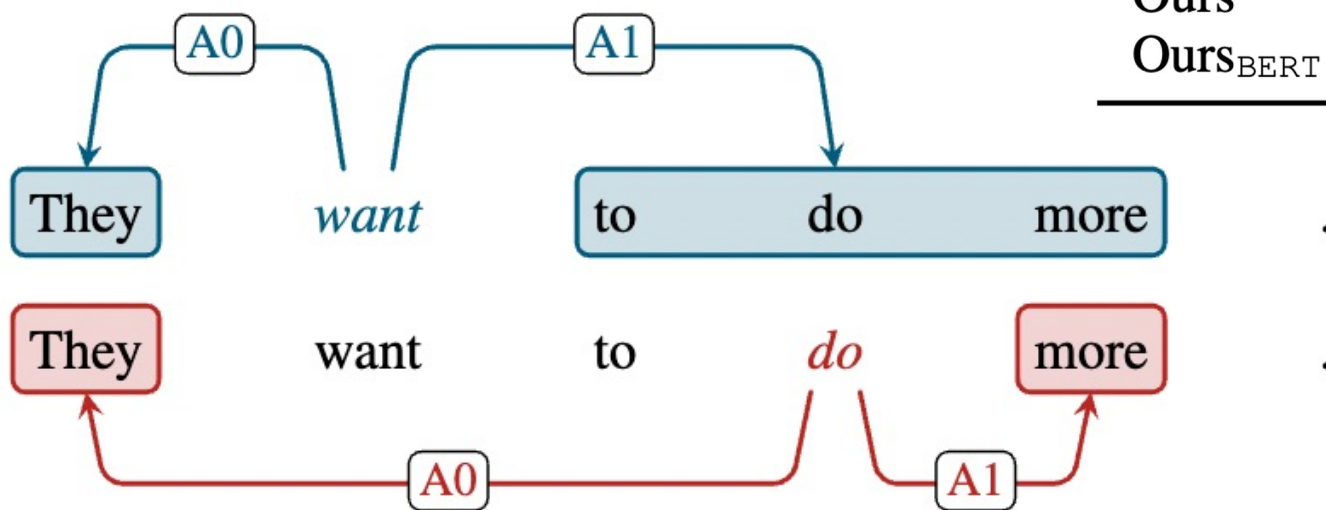
- 把词语内部搭配结构这个小方向向前推进了一步，深入扎实
  - 标注规范、标注数据、数据分析、基准实验
- 做了一些初步的实验，初步验证了词语内部搭配结构可以帮助句法分析
- 但是，这个工作还很粗浅 (**in-depth?**)，还有很多问题没有回答清楚。

# 2022年的两个工作

# 语义分析（语义角色标注SRL）

- 提出一种新的语义分析方法，性能高，速度快（COLING-2022 Best Long Paper）

| Model                               | Type      | Sents/sec  |
|-------------------------------------|-----------|------------|
| He et al. (2018)                    | SGP       | 44         |
| Strubell et al. (2018)              | BIO-based | 45         |
| Zhang et al. (2022)                 | TP        | 214        |
| Zhang et al. (2022) <sub>BERT</sub> |           | 113        |
| Ours                                | WGP       | <b>669</b> |
| Ours <sub>BERT</sub>                |           | 252        |





蘇州大學



# SynGEC: Syntax-Enhanced Grammatical Error Correction with a Tailored GEC-Oriented Parser

Yue Zhang<sup>1</sup>, Bo Zhang<sup>2</sup>, Zhenghua Li<sup>1</sup>, Zuyi Bao<sup>2</sup>, Chen Li<sup>2</sup>, Min Zhang<sup>1</sup>



**1 Soochow University**



**2 Alibaba Group**

# 动机

- GEC和句法分析紧密相关
- 如何利用句法信息来帮助GEC（源端句法）

---

|               |                                                        |
|---------------|--------------------------------------------------------|
| <b>Source</b> | 我不知道他何时返回回来。<br>I don't know when he will return back. |
| <b>Ref. 1</b> | 我不知道他何时返回回来。<br>I don't know when he will return.      |
| <b>Ref. 2</b> | 我不知道他何时返回回来。<br>I don't know when he will be back.     |

---

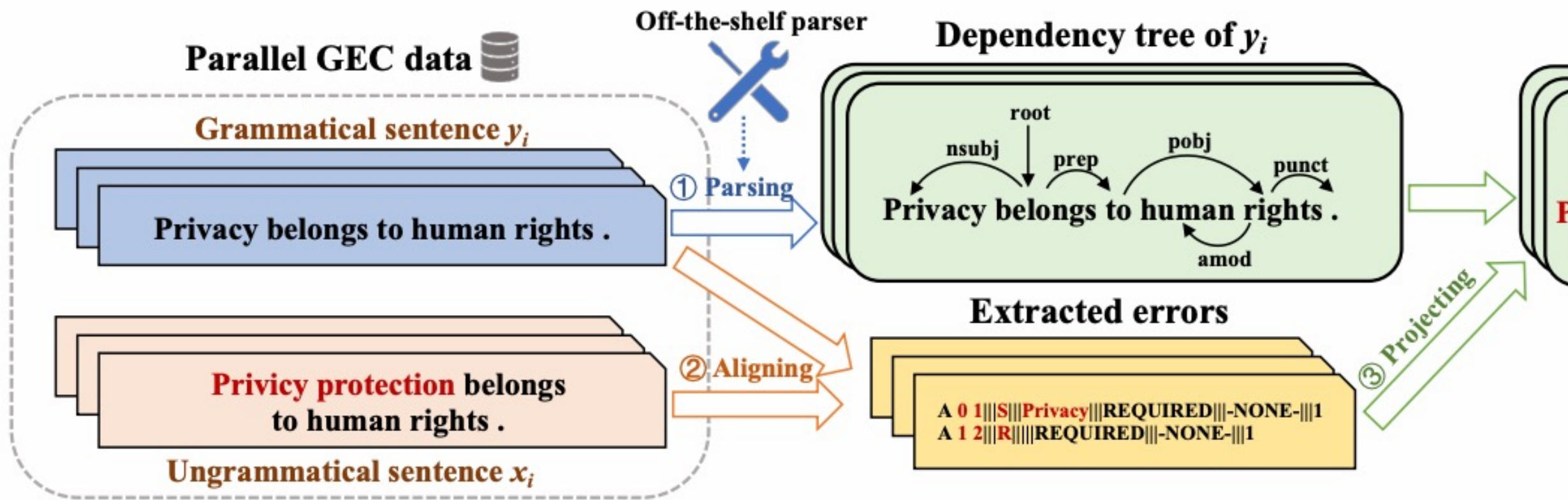
# 挑战

- 句法分析模型处理病句时性能不佳

|               |                                                        |
|---------------|--------------------------------------------------------|
| <b>Source</b> | 我不知道他何时返回回来。<br>I don't know when he will return back. |
| <b>Ref. 1</b> | 我不知道他何时返回回来。<br>I don't know when he will return.      |
| <b>Ref. 2</b> | 我不知道他何时返回回来。<br>I don't know when he will be back.     |

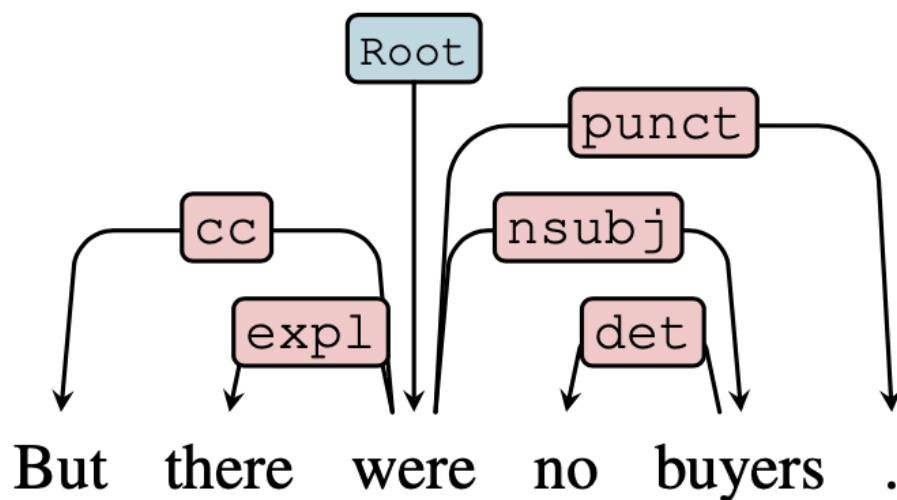
# 本文方法 ( parallel GEC training data as pivot )

- 句法分析器先自动分析目标端句子
- 然后将句法树映射到源端 ( 适配、扩展scheme )

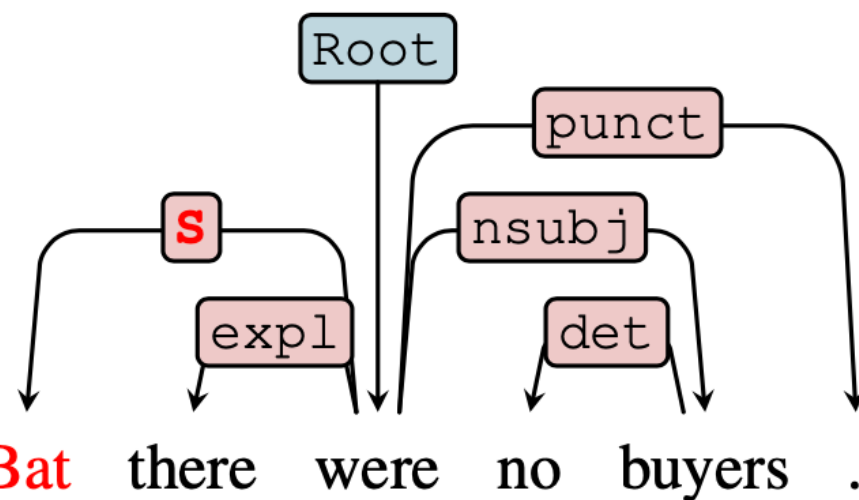


# 本文方法 ( parallel GEC training data as pivot )

- Scheme扩展



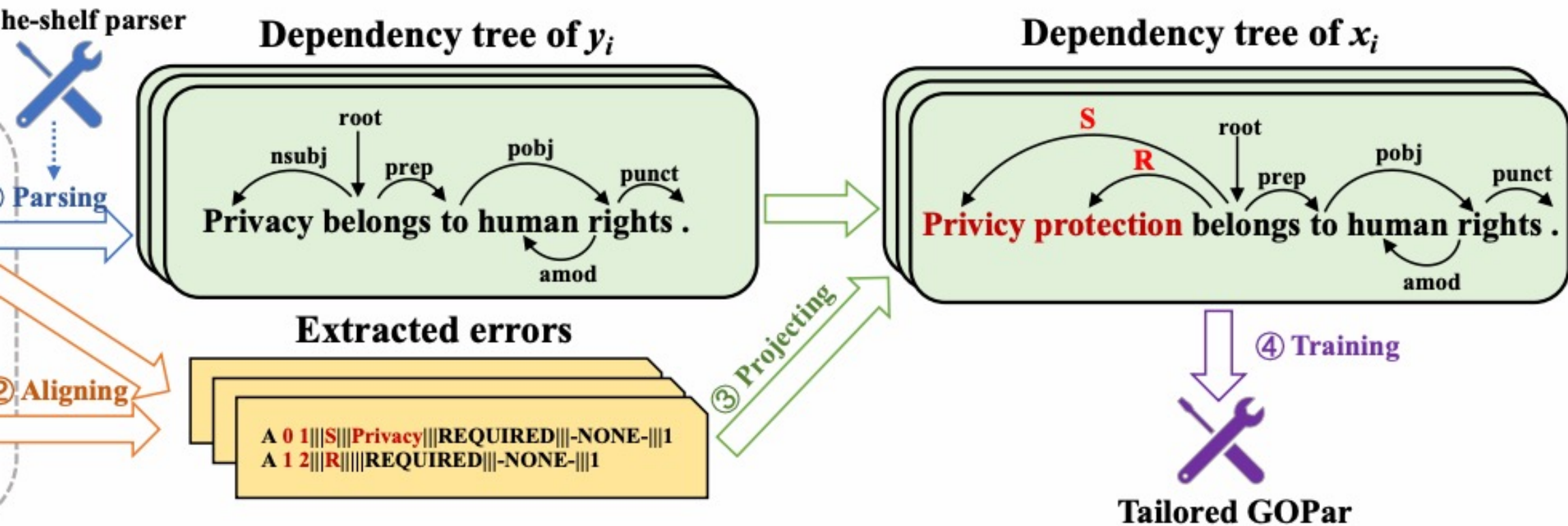
(a) The correct sentence.



(b) Substituted errors.

# 本文方法 ( parallel GEC training data as pivot )

- 在“源端句子-适配句法树”上训练句法分析器



# 实验结果（性能稳定提升）

|         |                                           | CoNLL-14- <i>test</i> |      |                  | BEA-19- <i>test</i> |      |                  |
|---------|-------------------------------------------|-----------------------|------|------------------|---------------------|------|------------------|
| System  |                                           | P                     | R    | F <sub>0.5</sub> | P                   | R    | F <sub>0.5</sub> |
| w/o PLM | w/o syntax                                |                       |      |                  |                     |      |                  |
|         | Kiyono et al. (2019) <sup>◦</sup>         | 67.9                  | 44.1 | 61.3             | 65.5                | 59.4 | 64.2             |
|         | Lichtarge et al. (2020) <sup>△▲</sup>     | 69.4                  | 43.9 | 62.1             | 67.6                | 62.5 | 66.5             |
|         | Stahlberg and Kumar (2021) <sup>△▲□</sup> | 72.8                  | 49.5 | <b>66.6</b>      | 72.1                | 64.4 | <b>70.4</b>      |
|         | <b>Our Baseline</b> <sup>♡</sup>          | 66.9                  | 40.3 | 59.1             | 66.8                | 55.5 | 64.2             |
|         | w/ syntax                                 |                       |      |                  |                     |      |                  |
|         | Wan and Wan (2021) <sup>◆</sup>           | 74.4                  | 39.5 | 63.2             | 74.5                | 48.6 | 67.3             |
|         | Li et al. (2022) <sup>♠</sup>             | 66.7                  | 38.3 | 58.1             | -                   | -    | -                |
|         | <b>SynGEC</b> <sup>♡</sup>                | 70.0                  | 46.2 | <b>63.5</b>      | 70.9                | 59.9 | <b>68.4</b>      |

# 核心贡献

- 提出一种更有效的自动的利用源端句法知识的GEC方法SynGEC
  - Parallel GEC training as pivot
  - 树映射
  - Scheme扩展

# 2023年我想做什么

# 正在做的事情

- 依存句法分析和短语分析之间有什么关系
- 词语内部结构（搭配、层次）有什么规律（构词法）
- 词语内部结构对于句法分析有什么帮助
- 短语句法树中的**N-ary**可否搞成**binary**，且符合人的认知
- 如何更好地利用句法来帮助**GEC**等任务（目标端？可控生成？）

# 正在思考的事情

- 如何利用语言学知识帮助句法语义分析
  - 如利用**Frame**的定义，来帮助**SRL**
  - 用**wordnet**中的知识来帮助句法语义分析
- 更深层的句法分析
  - wh-movement trace none 省略...
  - PTB和CTB的标注质量如何

# 长期思考的问题

- 领域移植问题
- 知识驱动方法（推理能力）
- 知识库构建
  - 一直在小规模摸索

# ChatGPT可能带来的积极的改变

- 基于大模型的知识库构建（人机协同，模型参与顶层设计、推荐）
- 大模型帮着搜集数据、标数据
- 提一个新任务时，只需聚焦Low-resource (Few-shot) 场景
  - Only need to annotate small-scale training data and enough evaluation data
- 强化学习的思路在NLP各个任务中的推广（包括句法语义分析）
  - 浅标注 -> 深标注（chain of thoughts, CoT）
  - 很深标注 -> 较深标注（只标注所关注的，类似于partial annotation）
- 我们可以重新思考：我们想要什么样的句法语义表示？（建模思路变得更广了）

# 大模型的两个问题

- 看起来很好是不够的（科学研究需要严格评价）
- 光靠prompt和对话，仍然无法按照需求控制

| MuCGEC Data | F0.5  | Precision    | Recall      |
|-------------|-------|--------------|-------------|
| ChatGPT     | 22.72 | 20.83        | <u>35.6</u> |
| Our SynGEC  | 46.51 | <u>54.69</u> | 29.10       |

- 后面可做的
  - 面向特定任务定制
  - 面向特定领域定制

# 基于大模型我们近期可能尝试的点

- 从大模型中萃取特定任务相关知识（prompt工程）
- 强化学习技术的利用，对现有任务的一些思考
- 大模型真的理解语言了吗？什么是理解语言？需要理解语言吗？  
【很难】

# 关于NLP研究生的大模型焦虑

- 做大模型的过程对于大部分人可能是
  - 无趣的（搜集数据、摸索prompt，调参）
  - 无能为力的（没有足够GPU资源）
- 读研究生的目的是什么？【考虑写一篇文章发到公众号上】
- 有多少博士在读书期间做出了很有影响力的工作？
- 感觉很热爱科研，但是看起来做科研做得并不快乐
  - 做具体的事情，人更有控制感，更容易快乐。
  - 对事情本身感兴趣
  - Aim high, but not too high.

# lacorrect.com lacorrect.cn演示

← → ↻ ⚠ Not Secure | 192.168.134.31/#/submit



欢迎! zhenghuali

论文提交

纠错结果

关于

点此上传

上传文件需为docx, pdf格式

📁 周楚越-毕业论文\_定稿.docx ✓

## 提交记录

请点击记录跳转到详情页面

×

| 文件ID | 提交日期                | 文件名                 | 状态    |
|------|---------------------|---------------------|-------|
| 361  | 2023-03-27 10:08:24 | 周楚越-毕业论文_定稿.docx    | 模型纠错中 |
| 359  | 2023-03-27 10:05:44 | 06-22李正华_本科毕业设计.doc | 纠错完成  |

Id:60 Page:0

---

原始文本：在整个毕业设计的过程，他会及时了解我的进度，有问必答，认真挑出我的设计中的问题，并时常督促我，贴心地跟我交流他的意见与想法。

纠错参考：在整个毕业设计的过程~~中~~，他会及时了解我的进度，有问必答，认真挑出我的设计中的问题，并时常督促我，贴心地跟我交流他的意见与想法。

---

Id:61 Page:0

---

原始文本：同时，我还要感谢大学相识的优秀可爱的室友们以及所有的同学朋友们，你们的存在让我的大学生活变得丰富多彩。

纠错参考：同时，我还要感谢大学~~相~~认识的优秀可爱的室友们以及所有的同学朋友们，你们的存在让我的大学生活变得丰富多彩。

---

Id:62 Page:0

---

原始文本：当我为本篇论文敲上最后一个字符，大学四年时光也悄然结束。

纠错参考：当我为本篇论文敲上最后一个字符~~时~~，大学四年时光也悄然结束~~了~~。

---



欢迎！zhenghuali

论文提交

纠错结果

关于

点此上传

上传文件需为docx， pdf格式

📄 06-22李正华\_本科毕业... ✔

提交记录

请点击记录跳转到详情页面 ×

| 文件ID | 提交日期                | 文件名                 | 状态    |
|------|---------------------|---------------------|-------|
| 359  | 2023-03-27 10:05:31 | 06-22李正华_本科毕业设计.doc | 模型纠错中 |

ighuali

退

Id:4 Page:0



原始文本：指代消解是自然语言处理研究的一项重要而且复杂的内容，随着文本总结，机器翻译，多语言信息处理，信息抽取，自动文摘的发展，涉及到指代消解的问题越来越多，因此，指代消解研究已经成为自然语言处理研究中的热点问题。

纠错参考：指代消解是自然语言处理研究的一项重要而且复杂的内容，随着文本总结~~，~~机器翻译~~，~~多语言信息处理~~，~~信息抽取~~，~~自动文摘的发展，涉及到指代消解的问题越来越多，因此，指代消解研究已经成为自然语言处理研究中的热点问题。

Id:5 Page:0



原始文本：MUC-1到MUC-5都是针对信息抽取的比赛，拟抽取的信息都预先经过语言学家预先抽取，通过对比IE系统的抽取结果和语言学家抽取结果的对比来评测各个系统。

纠错参考：MUC-1到MUC-5都是针对信息抽取的比赛，拟抽取的信息都预先经过语言学家预先抽取，通过对~~比~~IE系统的抽取结果和语言学家抽取结果~~的~~对比来评测各个系统。

Id:6 Page:0



原始文本：由于历次会议中都发现IE系统抽取的结果中存在许多前言不搭后语的代词或者名词短语，因此，指代消解被认为是信息抽取任务的关键技术之一，所以MUC-6开始专门成立了指代消解评测机构，单独进行了指代消解的评测。

纠错参考：由于历次会议中都发现~~IE~~系统抽取的结果中存在许多前言不搭后语的代词或者名词短语，因此，指代消解被认为是信息抽取任务的关键技术之一，所以~~MUC-6~~开始专门成立了指代消解评测机构，单独进行~~子~~指代消解的评测。

谢谢！