NLPCC-2019 Shared Task on Cross-domain Dependency Parsing: Call for Participant

Soochow University and Alibaba Inc.

March, 2019

With the surge of web data (or user generated content), cross-domain parsing has become the major challenge for applying syntactic analysis in realistic NLP systems. To meet the challenge of the lack of labeled data, we have manually annotated large-scale high-quality domain-aware datasets with a lot of effort (http://hlt.suda.edu.cn/index.php/SUCDT) in the past few years.

We provide about 17K sentences from a balanced corpus as the source domain (BC), and as three target domains 10K sentences from product comments (PC), 8K sentences from product blogs (PB), and 3K sentences from the web fiction named ``Zhuxian'' (ZX). We setup four subtasks with two cross-domain scenarios, i.e., unsupervised domain adaptation (no target-domain training data) and semi-supervised (with target-domain training data), and two settings, i.e., closed and open.

1 Task Settings

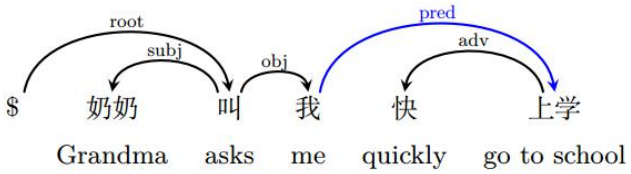

Figure 1: An example of dependency parse tree, where a dependency arc points to the modifier word from the head word, and ``pred'' means the ``predicate'' relation.

Task Description: Given an input sentence, the goal of dependency parsing is to predict the optimal dependency tree, as depicted in Figure 1. This evaluation task consists of four subtasks:

Subtask 1 (un-closed): Unsupervised domain adaptation (closed)

Subtask 2 (semi-closed): Semi-supervised domain adaptation (closed)

Subtask 3 (un-open): Unsupervised domain adaptation (open)

Subtask 4 (semi-open): Semi-supervised domain adaptation (open)

- **Unsupervised** domain adaptation assumes that there is no labeled training data for the target domain. For example, when the target domain is PC, in the unsupervised domain adaptation scenario, you cannot use PC-Train. However, PC-Dev/Unlabeled are allowed to use.

- **Semi-supervised** domain adaptation means that there exists a labeled training dataset for the target domain. For example, when the target domain is PC, PC-Train/Dev/Unlabeled all can be used in semi-supervised domain adaptation scenario.

- **Closed** means:

1) You can only use our provided data and information. We will provide word segmentation, automatic part-of-speech (POS) tags. We will also provide pre-trained word embeddings, which are obtained by training word2vec on the Chinese Gigaword 3 and all the target-domain unlabeled data).

2) Do not use other open-source tools to obtain new information, such as part of speech, word meaning and other features.

3) Do not use other resources, such as dictionaries, syntactic and semantic treebank.

4) Do not use ELMo/Bert and other pre-trained word representations.

- **Open**: Without any restriction, you can use any resource. However, it is strongly recommended that participants clearly describe in their system reports what external resources are used and how the parsing performance is affected.

Special Notice 1: It is not allowed to use training data from other target domains. For example, when the target domain is PC, PB-Train and ZX-Train cannot be used in all four subtasks.

Special Notice 2: It is not allowed to add dev data into training data. Dev data can only be used for parameter and model selection.

If you have any doubts about the rules, please do not hesitate to contact us.

2 Submission and Evaluation

For submission, please package the results into an zip/tar.gz file and send email to Xue Peng (see contact information below). Please organize and name the directories and files as the following format (omit the nonparticipating subtasks), to facilitate our processing. Please refer to the Dev and Readme files and make sure that each submitted file strictly follows the CoNLL format.

```
subtask1-un-closed/PC-Test.out.conll
subtask1-un-closed/PB-Test.out.conll
subtask1-un-closed/ZX-Test.out.conll
subtask2-semi-closed/PC-Test.out.conll
...
subtask3-un-open/PC-Test.out.conll
...
subtask4-semi-open/PC-Test.out.conll
...
```

We use the standard labeled attachment score (LAS, percent of words that receive correct heads and labels). For any subtask, a participant team must submit results on all **three** target domains, so that we can obtain three LAS values. We average the three LAS directly to determine the final ranking.

3 Data Setup

Annotation guideline: After an in-depth study on the HLT-CDT and UD annotation guidelines, we have developed a detailed annotation guideline that on the one hand aims to fully capture Chinese syntax, and on the other hand tries to guarantee inter-annotator consistency and facilitate model learning. Our guideline includes 20 dependency labels (Guo et al., 2018). Please download Annotation guidelines.pdf for the latest guideline.

Annotation process: We provide one source-domain and three target-domain datasets. All sentences are labeled by two persons, and inconsistent submissions are discussed by senior annotators to determine the final answer. In addition, in order to reduce annotation cost, we adopt the active learning procedure based on partial annotation (Jiang et al., 2018). However, all training datasets are automatically complemented into high-quality full trees (Zhang et al., 2017). Table 1 shows data distribution for this shared task.

Table 1: Data statistics (in sentence number)

		Train	Dev	Test	Unlabeled
Source Domain	Balanced Corpus (BC, text selected from HLT-CDT and PennCTB treebanks)	16.3K	1K	2K	0
Target Domains	Comments Products (PC, from Taobao)	6.2K	1.3K	2.6K	350K
	Product Blogs (PB, from Taobao headlines)	5.1K	1.3K	2.6K	300K
	The web fiction ``Zhuxian'' (ZX)	1.6K	0.5K	1.1K	30K

4 Website

More information and further arrangements will be published in http://hlt.suda.edu.cn/index.php/nlpcc-2019-shared-task. For example, how to sign the data license; how to download the data, etc.

Organizers: Zhenghua Li (Soochow University), Rui Wang (Alibaba Inc.)

Contacts: Xue Peng (Graduate student of Soochow University, xpeng1117@qq.com)
Yue Zhang (Alibaba Inc., shiyu.zy@alibaba-inc.com)

References

Lijuan Guo, Zhenghua Li, Xue Peng, Min Zhang. 2018. Annotation Guideline of Chinese Dependency Treebank from Multi-domain and Multi-source Texts. Journal of Chinese Information Processing. 2018, 32 (10): 28-35-52 (In Chinese) (pdf)

Xinzhou Jiang, Bo Zhang, Zhenghua Li, Min Zhang, Sheng Li, Luo Si. 2018. Supervised Treebank Conversion: Data and Approaches. Proceedings of ACL, pp. 2706-2716. Melbourne, Australia. 15-20 Jul. 2018 (pdf)

Yue Zhang, Zhenghua Li, Jun Lang, Qingrong Xia, Min Zhang. 2017. Dependency Parsing with Partial Annotations: An Empirical Comparison. Proceedings of IJCNLP, pp. 49-58. Taiwan, Nov. 27 - Dec. 1, 2017 (pdf)

下一页为中文版邀请函

NLPCC-2019 依存句法分析领域移植评测参赛邀请函

苏州大学、阿里巴巴

2019 年 3 月

近年来，随着网络数据的大幅增加，句法领域移植的跨域分析已成为自然语言处理实际应用的主要挑战。为了应对标注数据的缺乏，我们在过去几年中花费了大量精力人工标注了大规模、高质量、具有领域代表性的数据 (http://hlt.suda.edu.cn/index.php/SUCDT)。

在本次评测任务中，我们提供了约 17K (句子数) 平衡语料(BC)数据作为源领域数据，10K 句产品评论(PC)、8K 句产品博客(PB)和 3K 句网络小说(ZX)数据作为三个目标领域数据；另外，对于目标领域的数据，我们还给出了大规模的无标注数据。针对半监督和无监督的领域移植场景，和封闭评测和开放评测两种设置，我们将评测划分为 4 个子任务：无监督领域移植（封闭或开放）、半监督领域移植 (封闭或开放)。

一、任务设置

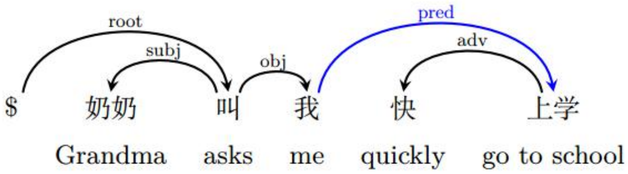

图 1：依存句法分析示例。其中依存弧从核心词指向修饰词， “pred” 表示 “谓词” 关系。

任务描述： 输入一个自然语句，依存句法分析任务的目标是预测最优的依存句法树，如图 1 所示。本评测包含 4 个子任务：

子任务一 (un-closed)：无监督领域移植(封闭)

子任务二 (semi-closed)：半监督领域移植(封闭)

子任务三 (un-open)：无监督领域移植(开放)

子任务四 (semi-open)：半监督领域移植(开放)

含义如下：

- **无监督领域移植：**不能使用对应目标领域的训练数据。例如，当目标领域为 PC 时，那么无监督领域移植场景下， PC-Train 不存在，但是 PC-Dev/unlabeled 是可以使用的。
- **半监督领域移植：**可以使用目标领域的训练数据。例如，当目标领域为 PC 时，那么半监督领域移植场景下， PC-Train/Dev/Unlabeled 都可以使用。
- Closed 的限制：

1） 只能使用我们提供的数据和信息，包括：所有数据我们都提供分词、自动词性；我们会提供 pre-trained word embedding (在 Chinese Gigaword 3 和无标注数据上训练 word2vec 得到)。

2） 不可以使用其他开源工具得到新的信息，如词性、词义等特征。

3） 不可以使用其他的资源，如词典、句法语义树库等。

4）不可以使用 ELMo/Bert 或其他预训练词表示。

- Open：不作任何限制，可以使用任何资源。但是建议参赛者在系统报告中明确说明使用的外部资源以及这些外部资源对性能的影响。

注意事项 1：不允许使用其他目标领域的训练数据。例如，当目标领域为 PC 时，对于任何子任务，都不可以使用 PB-Train 和 ZX-Train。

注意事项 2：不允许将 Dev 加入到训练数据中，Dev 数据只可以用来调参和选择模型。

如果您对规则有疑问，请随时联系我们。

二、结果提交和评价

提交结果时，请将测试文件打包（zip/tar.gz）发送给彭雪同学（联系方式如下）。压缩包内请按照如下目录/文件命名和组织形式（没有参加的子任务忽略即可），方便我们后续处理。请参考 Dev 数据及 Readme，确保每个文件严格采用 CoNLL 格式。

subtask1-un-closed/PC-Test.out.conll
subtask1-un-closed/PB-Test.out.conll
subtask1-un-closed/ZX-Test.out.conll
subtask2-semi-closed/PC-Test.out.conll
...
subtask3-un-open/PC-Test.out.conll
...
subtask4-semi-open/PC-Test.out.conll
...

我们使用 labeled attachment score（LAS，即核心词及其依存关系标签都预测正确的词语百分比)。对于任何一个子任务，所有参赛队伍必须提交 3 个目标领域的测试结果，从而得到 3 个 LAS 值。我们直接对 3 个 LAS 求平均，确定最终排名。

三、数据设置

标注规范：我们充分参考了哈工大依存树库标注规范和 UD 标注规范，针对汉语的特点，同时考虑标注一致性和可计算性，制定了一个详细的标注规范，包含 20 个依存标签(郭丽娟等,2018)。最新的标注规范请查看：依存句法数据标注规范.pdf

标注过程：我们提供一个源领域、三个目标领域数据集，所有的数据均采用严格的双人标注，不一致时由资深标注员讨论确定答案。另外，为了尽量降低人工标注的代价，我们采用基于局部标注的主动学习的方法 (Jiang et al., 2018)，但是所有的训练数据集都自动补全为高质量的完整句法树 (Zhang et al., 2017)。

表 1 中记录了本次测评使用数据的分布情况。

表 1：数据分布情况（句子数）

		训练集 Train	开发集 Dev	测试集 Test	无标注 Unlabeled
源领域	规范平衡语料库 BC，文本选自哈工大依存树库和宾州树库	16.3K	1K	2K	0
目标领域	产品评论 PC（淘宝）	6.2K	1.3K	2.6K	350K
	产品博客 PB（淘宝头条）	5.1K	1.3K	2.6K	300K
	网络小说《诛仙》ZX	1.6K	0.5K	1.1K	30K

四、评测网站

更多信息和进一步的安排将在 http://hlt.suda.edu.cn/index.php/nlpcc-2019-shared-task 发布。如数据使用协议的签署及数据的发布。

组织者：李正华（苏州大学）、王睿（阿里巴巴）

联系人：彭雪（苏大研究生、xpeng1117@qq.com）

张月（阿里巴巴、shiyu.zy@alibaba-inc.com）

参考文献

郭丽娟，李正华，彭雪，张民. 2018.适应多领域多来源文本的汉语依存句法数据标注规范. 中文信息学报. 2018，32(10):28-35-52 (pdf)

Xinzhou Jiang, Bo Zhang, Zhenghua Li, Min Zhang, Sheng Li, Luo Si. 2018. Supervised Treebank Conversion: Data and Approaches. Proceedings of ACL, pp. 2706-2716. Melbourne, Australia. 15-20 Jul. 2018 (pdf)

Yue Zhang, Zhenghua Li, Jun Lang, Qingrong Xia, Min Zhang. 2017. Dependency Parsing with Partial Annotations: An Empirical Comparison. Proceedings of IJCNLP, pp. 49-58. Taiwan, Nov. 27 - Dec. 1, 2017 (pdf)