SciDTB: Discourse Dependency Treebank for Scientific Abstracts

An Yang, Sujian Li
Peking University

ACL 2018

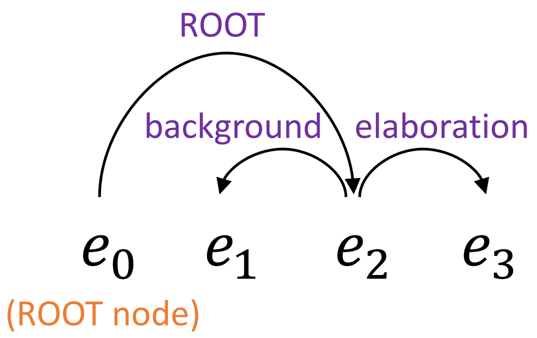
Outline

- Background: discourse dependency structure & treebanks
- Main work: details about SciDTB
  - Annotation framework
  - Corpus construction
  - Statistical analysis
  - SciDTB as evaluation benchmark
- Conclusion & summary

Discourse Dependency Structure & Treebanks

Example text: *[Syntactic parsing is useful in NLP.]$_{e1}$ [We present a parsing algorithm,]$_{e2}$ [which improves classical transition-based approach.]$_{e3}$*

Discourse dependency tree:
[Li. 2014; Yoshida. 2014]



Advantage:
flexible, simple, support non-projection

Discourse dependency treebanks:

- **Conversion based** dependency treebanks from RST or SDRT representations [Li. 2014; Stede. 2016]
- Limitations: **conversion errors** and **not support non-projection**
- Build a dependency treebank from scratch
- Scientific abstracts: short with strong logics

Annotation Framework: Discourse Segmentation

Discourse segmentation: Segment abstracts into **elementary discourse units (EDUs)**

Guidelines:

- Generally treats **clauses** as EDUs [Polanyi. 1988, Mann and Thompson. 1988]
- Subjective and some objective clauses are not segmented [Carlson and Marcu. 2001]

> Example 1: *[The challenge of copying mechanism in Seq2Seq is that new machinery is needed]$_{e1}$ [to decide when to perform the operation.]$_{e2}$*

- Strong **discourse cues** always starts a new EDU

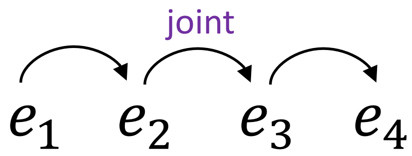
> Example 2: *[Despite bilingual embedding's success,]$_{e1}$ [the contextual information]$_{e2}$ [which is important to translation quality,]$_{e3}$ [was ignored in previous work.]$_{e4}$*

Annotation Framework: Obtain Tree Structure

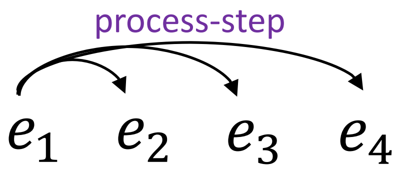
- A tree is composed of relations $< e_h, r, e_d >$
  - e_h: the EDU with essential information
  - e_d: the EDU with supportive content
  - r: relation type (17 coarse-grained and 26 fine-grained types)

- Each EDU has one and only one head
  - One EDU is dominated by ROOT node

- Polynary relations

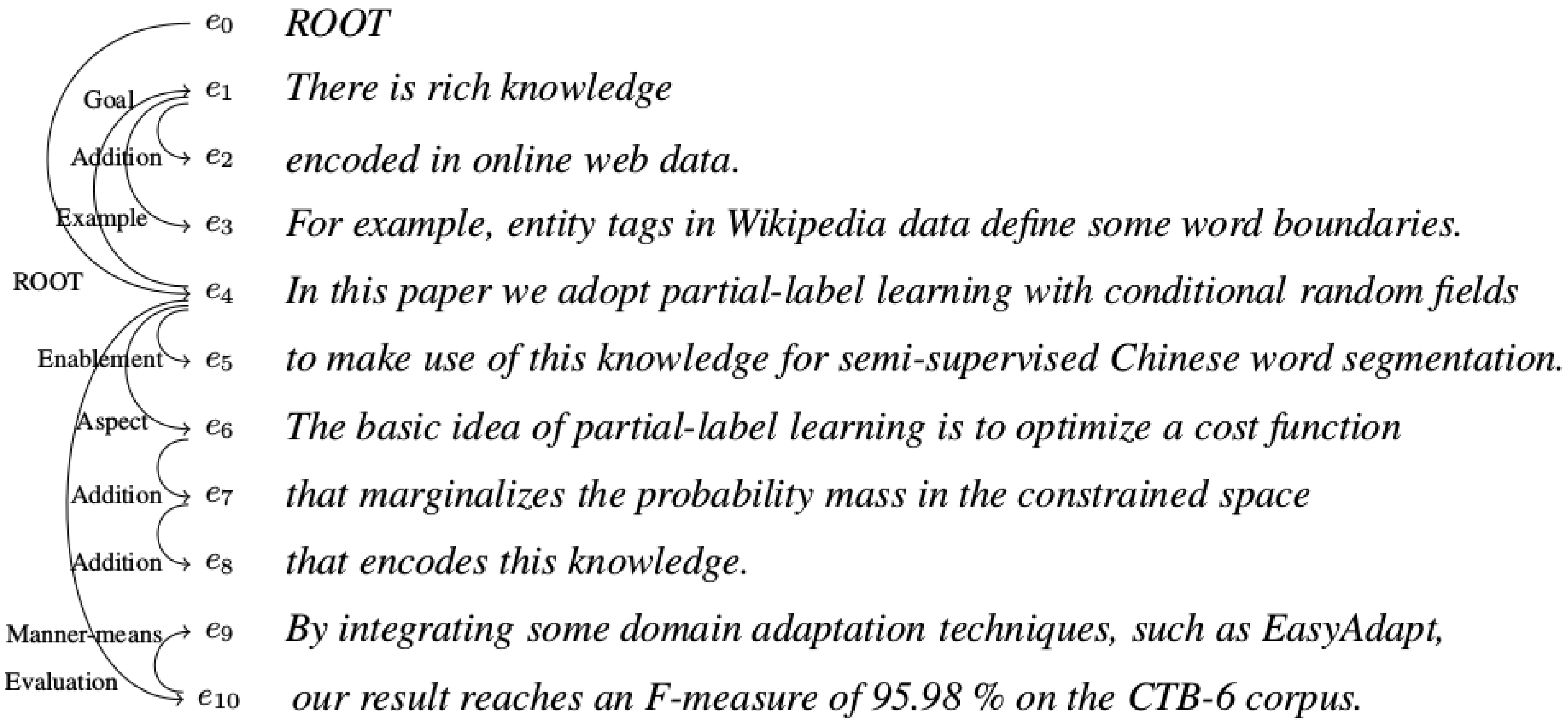


Multi-coordination



One-dominates-many

Annotation Example in SciDTB

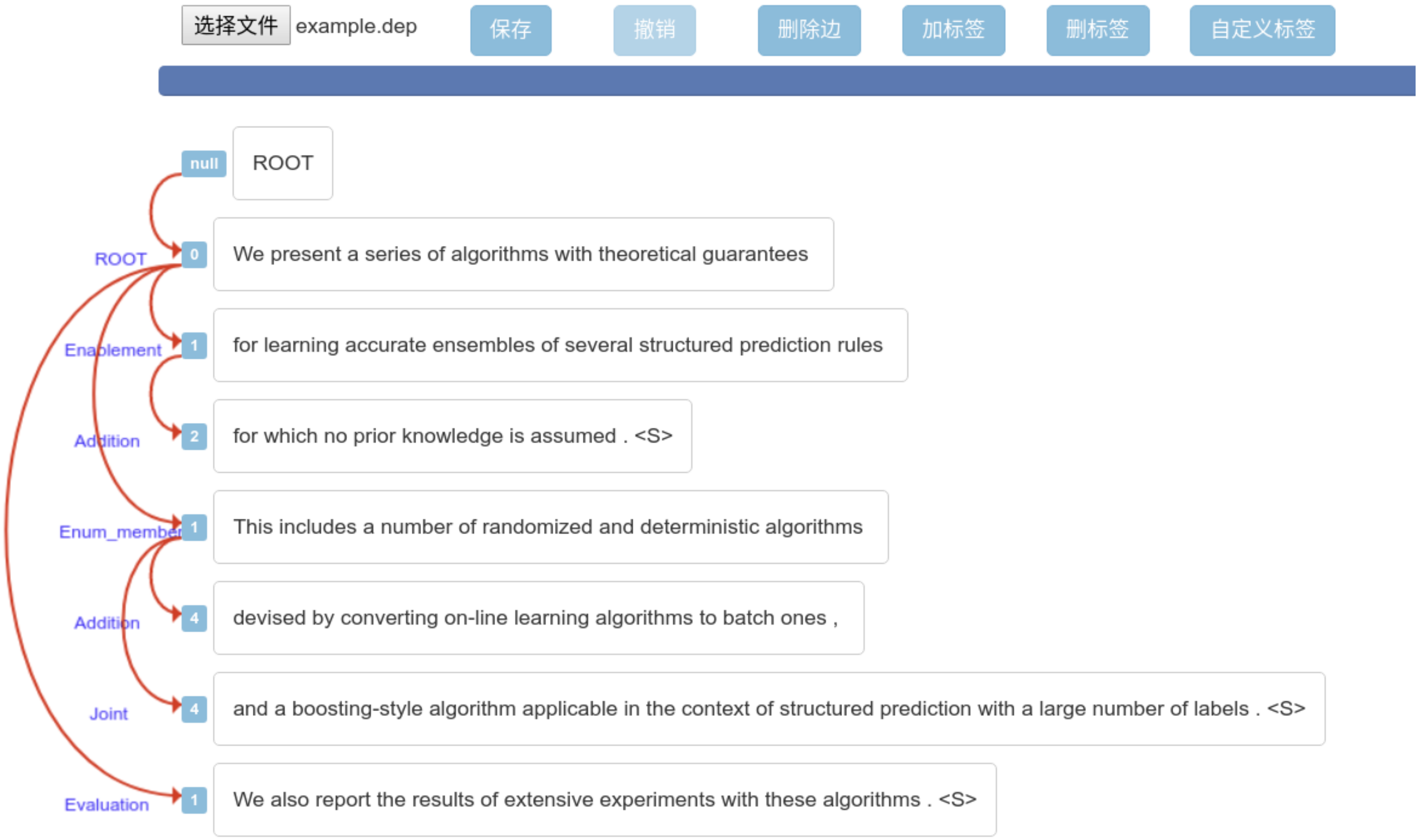


Abstract from http://www.aclweb.org/anthology/

Corpus Construction

- Annotator Recruitment:
  - 5 annotators were selected after test annotation
- EDU Segmentation:
  - Semi-automatic: pre-trained **SPADE** [Soricut. 2003] + **Manual proofreading**
- Tree Annotation:
  - The annotation lasted 6 months
  - 63% abstracts were annotated more than twice
  - An online tool was developed for annotating and visualizing DT trees

Online Annotation Tool



Website: http://123.56.88.210/demo/depannotate/

Reliability: Annotation Consistency

- The consistency of tree annotation is analyzed by 3 metrics:
  - **Unlabeled accuracy score**: structural consistency
  - **Labeled accuracy score**: overall consistency
  - **Cohen's Kappa**: consistency on relation label conditioned on same structure

Annotators	#Doc.	UAS	LAS	Kappa score
Annotator 1 & 2	93	0.811	0.644	0.763
Annotator 1 & 3	147	0.800	0.628	0.761
Annotator 1 & 4	42	0.772	0.609	0.767
Annotator 3 & 4	46	0.806	0.639	0.772
Annotator 4 & 5	44	0.753	0.550	0.699

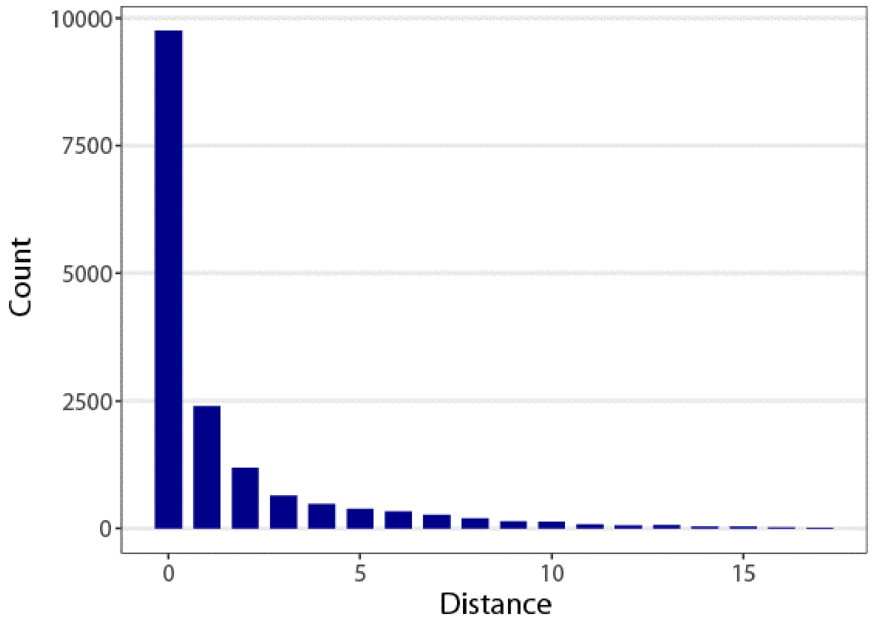
Annotation Scale

- SciDTB is
  - comparable with PDTB and RST-DT considering size of units and relations
  - much larger than existing domain-specific discourse treebanks

Corpus	#Doc.	#Doc. (unique)	#Text unit	#Relation	Source	Annotation form
SciDTB	1355	798	18978	18978	Scientific abstracts	Dependency trees
RST-DT	438	385	24828	23611	Wall Street Journal	RST trees
PDTB v2.0	2159	2159	38994	40600	Wall Street Journal	Relation pairs
BioDRB	24	24	5097	5859	Biomedical articles	Relation pairs

Structural Characteristics

- Dependency distance
  - Most relations (61.6%) occur between neighboring EDUs
  - The distance of 8.8% relations is greater than 5



- Non-projection: 3% of the whole corpus

SciDTB as Benchmark

- We make SciDTB as a benchmark for evaluating discourse dependency parsers
- Data partition: 492/154/152 abstracts for train/dev/test set
- 3 baselines are implemented:
  - Vanilla transition based parser
  - Two-stage transition based parser a simpler version of [Wang, 2017]
  - Graph based parser

Model	Dev set		Test set	
	UAS	LAS	UAS	LAS
Vanilla transition	**0.730**	0.557	**0.702**	0.535
Two-stage transition	**0.730**	**0.577**	**0.702**	**0.545**
Graph-based	0.577	0.455	0.576	0.425
Human	0.806	0.627	0.802	0.622

Conclusions

- Summary:
  - We propose a discourse dependency treebank with following features:
    - constructed from scratch
    - Scientific abstracts
    - comparable with existing treebanks in size
  - We further make SciDTB as a benchmark

- Future work:
  - Consider longer scientific articles
  - Develop effective parsers on SciDTB

Thank you!

Contact: yangan@pku.edu.cn

SciDTB is available:
https://github.com/PKU-TANGENT/SciDTB