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:**

\[ \text{The challenge of copying mechanism in Seq2Seq is that new machinery is needed} \]_{e1} \[ \text{to decide when to perform the operation.} \]_{e2}

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

**Example 2:**

\[ \text{Despite bilingual embedding’s success,} \]_{e1} \[ \text{the contextual information} \]_{e2} \[ \text{which is important to translation quality,} \]_{e3} \[ \text{was ignored in previous work.} \]_{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

  \[ e_1 \xrightarrow{\text{joint}} e_2 \xrightarrow{\text{process-step}} e_3 \xrightarrow{\text{one-dominates-many}} e_4 \]

  Multi-coordination

  One-dominates-many
Annotation Example in SciDTB

ROOT
There is rich knowledge encoded in online web data.
For example, entity tags in Wikipedia data define some word boundaries.
In this paper we adopt partial-label learning with conditional random fields to make use of this knowledge for semi-supervised Chinese word segmentation.
The basic idea of partial-label learning is to optimize a cost function that marginalizes the probability mass in the constrained space that encodes this knowledge.
By integrating some domain adaptation techniques, such as EasyAdapt, our result reaches an F-measure of 95.98% on the CTB-6 corpus.

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

• Annotator Recruitment:
  • 5 annotators were selected after test annotation

• EDU Segmentation:

• 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

<table>
<thead>
<tr>
<th>Annotators</th>
<th>#Doc.</th>
<th>UAS</th>
<th>LAS</th>
<th>Kappa score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotator 1 &amp; 2</td>
<td>93</td>
<td>0.811</td>
<td>0.644</td>
<td>0.763</td>
</tr>
<tr>
<td>Annotator 1 &amp; 3</td>
<td>147</td>
<td>0.800</td>
<td>0.628</td>
<td>0.761</td>
</tr>
<tr>
<td>Annotator 1 &amp; 4</td>
<td>42</td>
<td>0.772</td>
<td>0.609</td>
<td>0.767</td>
</tr>
<tr>
<td>Annotator 3 &amp; 4</td>
<td>46</td>
<td>0.806</td>
<td>0.639</td>
<td>0.772</td>
</tr>
<tr>
<td>Annotator 4 &amp; 5</td>
<td>44</td>
<td>0.753</td>
<td>0.550</td>
<td>0.699</td>
</tr>
</tbody>
</table>
SciDTB is

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

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UAS</td>
<td>LAS</td>
</tr>
<tr>
<td>Vanilla transition</td>
<td>0.730</td>
<td>0.557</td>
</tr>
<tr>
<td>Two-stage transition</td>
<td>0.730</td>
<td>0.577</td>
</tr>
<tr>
<td>Graph-based</td>
<td>0.577</td>
<td>0.455</td>
</tr>
<tr>
<td>Human</td>
<td>0.806</td>
<td>0.627</td>
</tr>
</tbody>
</table>
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