Transferring knowledge from discourse to arguments: A case study with scientific abstracts

Pablo Accuosto    Horacio Saggion
Large-Scale Text Understanding Systems Lab, NLP Group (LaSTUS/TALN)
Universitat Pompeu Fabra
Presentation outline

- Objective
- Motivation
- SciDTB Corpus
- Argumentation layer
- Argument mining experiments
- Pilot application
- Conclusions and future work
I. Objective
Objective

Explore if/how discourse annotations can be exploited to facilitate mining arguments in scientific texts.

Conduct a pilot experiment with scientific abstracts using automatically identified argumentative units and relations.
II. Motivation
“… Constructing annotated corpora is, in general, a complex and time-consuming task.

This is particularly true for argumentation mining, as the identification of argument components, their exact boundaries, and how they relate to each other can be quite complicated (and controversial!) even for humans…”

Lippi and Torroni (2016)

Especially challenging in scientific texts due to their argumentative complexity.
(Kirscher et al. 2015; Green 2015)

Leverage existing resources

Schema / corpora / models developed for related tasks

In particular, discourse annotated corpora and models

- Rhetorical Structure Theory (RST)

This would allow to take advantage of resources (corpora, models) developed for discourse parsing (RST in particular)

Previous works explore relations between discourse analysis and argument mining tasks (Peldszus and Stede 2016)


In previous experiments (Accuosto and Saggion, 2019) we observed that:

• Explicitly incorporating discourse features contributes to improve the performance of argument mining tasks.
• Neural models (BiLSTMs) perform better than traditional sequence labelling algorithms (CRF) even if a low resource setting.

The obtained models can only be applied with texts annotated with discourse.

Alternatives

Pipeline: Discourse parsing + Argument mining

Transfer representations obtained from discourse parsing models

III. SciDTB Corpus
SciDTB Corpus
Discourse Dependency TreeBank for Scientific Abstracts

798 ACL Anthology abstracts annotated with RST-like units and relations
Binary relations between elementary discourse units → discourse dependency trees (simplifies annotation and processing)

IV. Argumentation layer
Pilot experiment
SciDTB Argumentation layer

New argumentative annotation layer
**60 abstracts** annotated with fine-grained **units** and **relations**
327 sentences, 8012 tokens

- **units**
  - **claims**
    - proposal (problem or approach)
    - assertion (conclusion or known fact)
    - result (interpretation of data)
    - observation (data)
    - means (implementation)
    - description (definitions/other)
  - **premises**
    - support (attack)
    - detail (elaboration, means, etc.)
    - sequence (sequence)
    - additional (joint)

- **relations**

**Argumentative units (AUs):** One or more elementary discourse unit (EDUs)
### Argumentation layer

<table>
<thead>
<tr>
<th>Type of unit</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>proposal</td>
<td>31</td>
</tr>
<tr>
<td>assertion</td>
<td>25</td>
</tr>
<tr>
<td>result</td>
<td>21</td>
</tr>
<tr>
<td>means</td>
<td>18</td>
</tr>
<tr>
<td>observation</td>
<td>3</td>
</tr>
<tr>
<td>description</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of relation</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>detail</td>
<td>45</td>
</tr>
<tr>
<td>support</td>
<td>42</td>
</tr>
<tr>
<td>additional</td>
<td>9</td>
</tr>
<tr>
<td>sequence</td>
<td>4</td>
</tr>
</tbody>
</table>
V. Argument mining experiments
## Argument mining tasks

<table>
<thead>
<tr>
<th>AM Task</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ATy</strong></td>
<td>Identify the type of argumentative units (e.g.: proposal)</td>
</tr>
<tr>
<td><strong>AFu</strong></td>
<td>Identify the function of the argumentative units (e.g.: support)</td>
</tr>
<tr>
<td><strong>APa</strong></td>
<td>Identify the relative position of the parent argumentative unit (e.g.: -2)</td>
</tr>
</tbody>
</table>

All the tasks are modeled as sequence tagging problems. Encoded with the beginning-inside-outside (BIO) tagging scheme (e.g.: B-support, I-assertion).
Discourse parsing tasks

<table>
<thead>
<tr>
<th>RST Task</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DFu</strong></td>
<td>Identify the discourse <strong>roles</strong> of the EDUs (e.g.: attribution, evaluation)</td>
</tr>
<tr>
<td><strong>DPa</strong></td>
<td>Identify the relative <strong>position</strong> of the parent EDU in the RST tree</td>
</tr>
</tbody>
</table>

These tasks are also modeled as sequence tagging problems with BIO tagging scheme.
Experimental settings

Discourse models

Trained with 738 abstracts:
SciDTB – 60 annotated with arguments

https://github.com/UKPLab/emnlp2017-bilstm-cnn-crf/
Experimental settings

Argument mining models

Concat backward and forward hidden states of top layer.
Results

<table>
<thead>
<tr>
<th>Setting</th>
<th>AFu</th>
<th>ATy</th>
<th>APa</th>
</tr>
</thead>
<tbody>
<tr>
<td>(DEmb+ELMo)</td>
<td>0.66</td>
<td>0.63</td>
<td>0.38</td>
</tr>
<tr>
<td>(DEmb+ELMo+RSTEnc)</td>
<td>0.69</td>
<td>0.67</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Average F1 scores for epochs 10 to 100

In all cases, the models are evaluated in a [10-fold cross-validation](#) setting with fixed hyperparameters.
## Results

<table>
<thead>
<tr>
<th>Setting</th>
<th>AFu</th>
<th>ATy</th>
<th>APa</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEmb+ELMo</td>
<td>0.66</td>
<td>0.63</td>
<td>0.38</td>
</tr>
<tr>
<td>DEmb+ELMo+GloVe</td>
<td>0.65</td>
<td>0.65</td>
<td>0.38</td>
</tr>
<tr>
<td>DEmb+ELMo+RSTEnc</td>
<td>0.69</td>
<td>0.67</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Average F1 scores for epochs 10 to 100
## Results

<table>
<thead>
<tr>
<th>Setting</th>
<th>support</th>
<th>proposal</th>
<th>assertion</th>
<th>result</th>
</tr>
</thead>
<tbody>
<tr>
<td>$DEmb+ELMo$</td>
<td>0.61</td>
<td>0.67</td>
<td>0.65</td>
<td>0.61</td>
</tr>
<tr>
<td>$DEmb+ELMo+RSTEnc$</td>
<td>0.63</td>
<td>0.71</td>
<td>0.67</td>
<td>0.63</td>
</tr>
</tbody>
</table>

*Average F1 scores for epochs 10 to 100*
Polynomial trend lines for F1 in epochs 10-100 for AFu, ATy, APa
Transferring discourse knowledge by means of representations learned in discourse parsing tasks can contribute to improve the performance of argument mining models.
VI. Pilot application
As an application, we explore whether the argumentative structure of the abstracts can predict acceptance / rejection of papers in computer science venues.
### Dataset

<table>
<thead>
<tr>
<th>Conference</th>
<th>Accepted</th>
<th>Rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDNNRIA 2018</td>
<td>35</td>
<td>23</td>
</tr>
<tr>
<td>IRASL 2018</td>
<td>30</td>
<td>29</td>
</tr>
<tr>
<td>ICLR 2018</td>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>

- *Compact Deep Neural Network Representation with Industrial Applications (CDNNRIA) - NIPS 2018*
- *Interpretability and Robustness for Audio, Speech and Language (IRASL) - NIPS 2018*
- *International Conference on Learning Representations (ICLR) - 2018*

*Retrieved from OpenReviews.net*
### Experimental setting

Features obtained with best AM model (RST encoders)

<table>
<thead>
<tr>
<th>none</th>
<th>support</th>
<th>...</th>
<th>support</th>
<th>proposal</th>
<th>result</th>
<th>...</th>
<th>observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>additional</td>
<td>support</td>
<td>...</td>
<td>—</td>
<td>assertion</td>
<td>assertion</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>support</td>
<td>none</td>
<td>...</td>
<td>—</td>
<td>assertion</td>
<td>proposal</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AFu</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>ATy</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>APa</th>
</tr>
</thead>
</table>

- Features obtained with best AM model (RST encoders)
- 0 1 ... 3
- REJECT
- 1 1 ... —
- ACCEPT
- 1 0 ... —
- ACCEPT
Algorithm/parameters set with 20-80 random split of training set

<table>
<thead>
<tr>
<th>Classifier</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>Decision tree</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Acceptance classification results

Decision points (and feature analysis) show that all three types of features are relevant for classification. E.g.: The parent of first unit, the functions of the first two units and the type of the first unit are particularly informative.

28
More experiments are needed to evaluate how generalizable these results are.

Experiments with ICLR 2017 dataset and compare with AllenNLP’s PeerRead results (F1 = 0.65)
Kang, D et al. A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications. NAACL 2018

…and also more detailed analysis would be required to know what the potential correlation means.

We are making no claims with respect to these relations.
VII. Conclusions and future work
• Confirm previous results - Discourse information contributes to improve the performance of argument mining tasks.

• Transfer learning approaches show potential to leverage available discourse annotated corpora to train argument mining models with limited amount of data.

• Pilot experiment using argumentative structure of abstracts to predict acceptance of papers encourages further research in this line.
Future work

• Increase coverage of annotation layer of SciDTB

• Evaluation of annotations: intrinsic and extrinsic methods
  Current metrics inadequate due to inherent ambiguity (Stab et al., 2014; Kirschner, 2015)

• Model improvement and optimization
  Other architectures/representations: Transformer-based embeddings

• Compare to other approaches
  Discourse parsing
Thank you
pablo.accuosto@upf.edu