Towards Building a Discourse-annotated Corpus of Russian

Speakers: Dina Pisarevskaya and Margarita Ananyeva
Background: Discourse analysis

Can be useful in **Natural Language Processing tasks**:

- machine translation evaluation,
- sentiment analysis,
- information retrieval,
- information extraction,
- text summarization,
- anaphora resolution,
- question-answering systems,
- text classification.

**Discourse parsers** for English:
RASTA [Corston-Oliver, Corston-Oliver, 1998],
SPADE [Soricut, Marcu, 2003],
HILDA [Hernault et al., 2010],
CODRA [Joty et al., 2015].
Two parsers [Surdeanu et al., 2015]
Discourse analysis approaches

- PDTB: Connective-led annotation (Penn Discourse Treebank) or Punctuation-led annotation (Chinese Discourse TreeBank). Example: PDTB (2008): 43 relations;

- Cohesive relations (Discourse Graphbank);

- Low-level relations (within/between adjacent sentences);

- Focus on discourse connectives;

- Relations have two (and only two) arguments.

- 3 levels of relation labels: class (4 major semantic classes), type (emphasizes the semantics of the class levels), subtype (emphasizes semantic contribution of each argument)

- When an annotator is uncertain of subtype, it is possible to choose higher level (type), it is good for inter-annotator agreement.
Penn Discourse Treebank: Relations
Penn Discourse Treebank: Corpora

Original corpus:

English: Penn Discourse Treebank (newspaper texts, million words).

Related corpora:

Chinese Discourse Treebank (newspaper texts, 70,000 words);

Czech: Prague Discourse Treebank (newspaper texts, 50,000 sentences);

6 languages: Eng, Tur, Deu, Por, Pol, Rus: TED-MDB (TED talks, work in progress);

Hindi: Discourse Relation Bank (newspaper texts, 400,000 words);

Arabic: Leeds Arabic DTB (newspaper texts, 166,000 words);

Turkish: METU-TDB Corpus (different genres, 500,000 words)
Discourse analysis approaches

- PDTB: Connective-led annotation (Penn Discourse Treebank) or Punctuation-led annotation (Chinese Discourse TreeBank). Example: PDTB (2008): 43 relations;

- Cohesive relations (Discourse Graphbank);

Rhetorical Structure Theory

[Mann, Thompson, 1988]

Examples of schema types

“Classic” relations set

- Circumstance
- Solutionhood
- Elaboration
- Background
- Enablement and Motivation
  - Enablement
  - Motivation
- Evidence and Justify
  - Evidence
  - Justify
- Relations of Cause
  - Volitional Cause
  - Non-Volitional Cause
  - Volitional Result
  - Non-Volitional Result
  - Purpose
- Antithesis and Concession
  - Antithesis
  - Concession
- Condition and Otherwise
  - Condition
  - Otherwise
- Interpretation and Evaluation
  - Interpretation
  - Evaluation
- Restatement and Summary
  - Restatement
  - Summary
- Other Relations
  - Sequence
  - Contrast
RST-corpora for different languages

- **English:** RST Discourse Treebank [Carlson et al., 2003], 385 newspaper articles, 176 383 tokens

- **German:** Potsdam Commentary Corpus [Stede, Neumann, 2014], 2 900 sentences from 175 newspaper articles, 32 000 tokens

- **Portuguese:** CorpusTCC [Pardo et al., 2004], 1 350 sentences from 100 scientific texts, 53 000 tokens

- **Portuguese:** Rhetalho [Pardo et al., 2004], 50 texts (30 from scientific papers and 20 from newspaper), approximately 5 000 tokens

- **Spanish:** RST Spanish Treebank [da Cunha et al., 2011], 2 256 sentences from 267 documents of several genres

- **Japanese:** [Kawahara et al., 2014], 30 000 sentences from 10 000 documents from the web, variety of domains
Discourse-annotated corpus of Russian

Texts of 4 genres:
- science;
- popular science;
- news stories;
- analytic journalism.

The project:
- 3 years;
- > 100 texts;
- > 100 000 tokens.
Open-source annotation tool rstWeb [https://corpling.uis.georgetown.edu/rstweb/info/]

I have read a few reviews on this car over at www.cars.com, and despite many reviewers saying this car has the interior set-up of the 1980s Chrysler K cars, but the interior does leave a lot to be desired. It is nowhere near that bad.
Background

D. Pisarevskaya,
“Rhetorical Structure Theory as a Feature for Deception Detection in News Reports in the Russian Language”
- Master thesis in Higher School of Economics, Computational Linguistics (the results were presented on 1st June).
Background (2)

The Laboratory for Computer Linguistics and Intelligent Information Processing (Institute for Systems Analysis FRC CSC RAS).

- manual (21 relations);
- 10 texts (1200 units and 1484 relations) from SynTagRus;
- discourse markers.

Kobozeva M.
“Developing the corpus of Russian texts with markup based on the Rhetorical Structure Theory”
- Master thesis in Russian State University for Humanities, Computational Linguistics
Current research

New corpus - 60 news stories have already been annotated.

User manual has been updated.

Segmentation of Russian texts into clauses: http://gree-gorey.github.io/
Inter-annotator agreement

- Accuracy
- Cohen’s kappa coefficient [Cohen, J., 1960; Cohen, J., 1968]
- Scott’s Pi [Scott, W. A., 1955]
- Token-based Fleiss’ kappa [Fleiss, J. L., 1971]
- Krippendorff’s unitized alpha [Krippendorff K., 2007]
# Relations

<table>
<thead>
<tr>
<th><strong>Mononuclear</strong></th>
<th><strong>Multinuclear</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Background</td>
<td>1. Contrast</td>
</tr>
<tr>
<td>2. Volitional and Non-Volitional Cause</td>
<td>2. Restatement</td>
</tr>
<tr>
<td>3. Evidence</td>
<td>3. Sequence</td>
</tr>
<tr>
<td>4. Volitional and Non-Volitional Effect</td>
<td>4. Joint</td>
</tr>
<tr>
<td>5. Condition</td>
<td>5. Comparison</td>
</tr>
<tr>
<td>6. Purpose</td>
<td>6. Same-unit</td>
</tr>
<tr>
<td>7. Concession</td>
<td></td>
</tr>
<tr>
<td>8. Preparation</td>
<td></td>
</tr>
<tr>
<td>9. Conclusion</td>
<td></td>
</tr>
<tr>
<td>10. Elaboration</td>
<td></td>
</tr>
<tr>
<td>11. Antithesis</td>
<td></td>
</tr>
<tr>
<td>12. Solutionhood</td>
<td></td>
</tr>
<tr>
<td>13. Motivation</td>
<td></td>
</tr>
<tr>
<td>14. Evaluation</td>
<td></td>
</tr>
<tr>
<td>15. Interpretation</td>
<td></td>
</tr>
<tr>
<td>16. Attribution1 and Attribution2</td>
<td></td>
</tr>
</tbody>
</table>
Evolution of relations

Volitional Cause + Non-volitional Cause = Cause

Volitional Effect + Non-volitional Effect = Effect

Interpretation + Evaluation

Attribution1 + Attribution2 = Attribution

Antithesis

Conclusion

Motivation
## New RST relations tree

<table>
<thead>
<tr>
<th>Coherence</th>
<th>Casual-argumentative</th>
<th>Structural</th>
<th>Attribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>Contrastive</td>
<td>Sequence</td>
<td>Attribution</td>
</tr>
<tr>
<td>Elaboration</td>
<td>Concession</td>
<td>Joint</td>
<td></td>
</tr>
<tr>
<td>Restatement</td>
<td>Contrast</td>
<td>Same-unit</td>
<td></td>
</tr>
<tr>
<td>Interpretation -</td>
<td>Causal</td>
<td>Comparison</td>
<td></td>
</tr>
<tr>
<td>Evaluation</td>
<td>Purpose</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preparation</td>
<td>Evidence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solutionhood</td>
<td>Cause-Effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Condition</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Inter-annotator agreement

0.2792  0.7768
0.3173  0.691
0.4965  0.7615

The code used for IAA calculation can be accessed via GitHub [https://github.com/nasedkinav/rst_corpus_rus/blob/master/krippendorffs_alpha.py].
Future work

User-friendly interface: visualisation, search and statistics, file upload mechanism.

Analysis of “marker potential”.

Discourse parser.
References


References


Thank you for attention

Dina Pisarevskaya

dinabpr@gmail.com

https://github.com/nasedkinav/rst_corpus_rus