The CLaC Discourse Parser at CoNLL-2015

Majid Laali* Elnaz Davoodi† Leila Kosseim‡

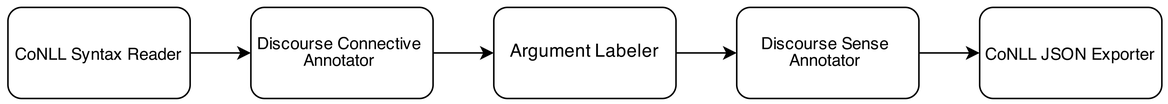
Department of Computer Science and Software Engineering,

Concordia University, Montreal, Canada

I. Summary

- Focus on the treatment of explicit discourse relations.
- Overall F_1 measure of 17.38%, ranking in 6th place out of the 17 parsers submitted to CoNLL 2015.
- Architecture similar to the End-to-End Discourse parser.
- The CLaC Discourse parser is based on the UIMA framework.
- Uses ClearTK to add machine learning functionality.
- Written in Java and its source code is available at "https://github.com/mjlaali/CLaCDiscourseParser.git".

II. Architecture



Parses syntactic information provided by CoNLL and adds this information to the documents in the UIMA framework.

Marks discourse usage or non-discourse usage of discourse connectives.

Labels ARG1 and ARG2.

Disambiguates the sense of the discourse connective.

Reads the output discourse relations annotated in the UIMA documents and generates a JSON file.

III. Discourse Connective Annotator

Algorithm:

- Searches the input texts for terms that match a predefined list of discourse connectives (was built solely from the CoNLL training dataset.
- Checks each match of discourse connective to see if it occurs in discourse usage or not.
  - Uses a binary classifier with six features (i.e. F_1-F_6)

Results:

- F_1 = 90.19%.

IV. Argument Labeler

Algorithm:

- Calculates the *Connective-Root path nodes*, the nodes that appear in the path from the discourse connective to the root of the sentence.
- Labels all constituents that are directly connected to one of the *Connective-Root path nodes* with 'part of ARG1', 'part of ARG2' or 'NON'.
  - Uses a classifier with nine features (i.e. F_1-F_9).
- Merges all constituents which were tagged as part of ARG1 or as part of ARG2 to obtain the actual boundaries of ARG1 and ARG2.
- If no constituent was labeled as a part of ARG1, the whole text of the previous sentence is considered as ARG1.

Results:

- The F_1 score of the *Argument Labeler*:

	Arg1	**Arg2**
Best Result	49.68%	74.29%
CLaC Discourse Parser	**45.18%**	**69.18%**
Average	30.77%	50.91%
Std. deviation	15.31%	20.58%

- Results show that the identification of ARG1 is more difficult than ARG2.

Error Analysis:

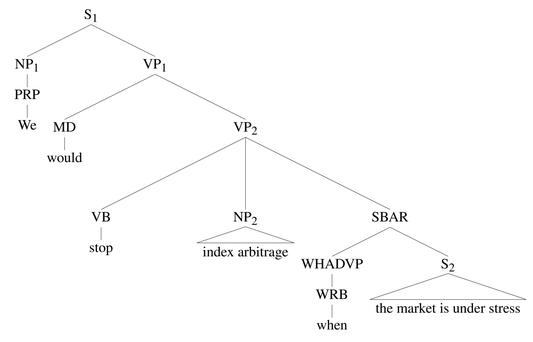
- *Attribute spans*:
  - ~~But the RTC also requires "working" capital~~ *to maintain the bad assets of thrifts that are sold* until **the assets can be sold separately**.
- Subordinate and coordinate clauses:
  - *We would have to wait* until **we have collected on those assets** ~~before we can move forward.~~

V. Features

Category	**Description**
Connective Features	F_1: Discourse connective text in lowercase.
	F_2: Categorization of the case of the connective: *all lowercase*, *all uppercase* and *initial uppercase*
	F_3: Highest node in the parse tree that covers the connective words but nothing more
	F_4: Parent of *SelfCat*
	F_5: Left sibling of *SelfCat*
	F_6: Right sibling of *SelfCat*
Syntactic Node Features	F_7 Path from the node to the *SelfCat* node in the parser tree
	F_8: Context of the node in the parse tree. The context of a node is defined by its label the label of its parent, the label of left and right sibling in the parse tree.
	F_9: Position of the node relative to the *SelfCat* node: *left* or *right*

VI. Example

We would stop index arbitrage when **the market is under stress**.



F_1 = *when*	F_2 = all lowercase
F_3 = WRB	F_4 = WHADVP
F_5 = null	F_6 = S
F_7 = $S \uparrow SBAR \downarrow WHADVP$	F_8 = S-SBAR-WHADVP-null
F_9 = left	

VII. Discourse Sense Annotator

- Uses the naïve approach that labels each discourse connective with its most frequent. The most frequent relation for discourse connectives is mined from the CoNLL training dataset.

VIII. Overall Results

- The F_1 scores of the CLaC discourse parser and the individual performance of its components:

	Discourse Connective Classifier	**Argument Labeler**	**Discourse Parsing (explicit only)**	**Discourse Parsing (explicit and implicit)**
Best Result	91.86%	41.35%	30.58%	24.00%
CLaC Discourse Parser	**90.19%**	**36.60%**	**27.32%**	**17.38%**
Average	74.20%	23.89%	18.28%	13.25%
Standard deviation	23.24%	13.01%	9.93%	6.41%

IX. Conclusion

- CLaC Discourse Parser was developed from scratch for CoNLL 2015.
- 3 person-month effort focused on *Discourse Connective Classification* and *Argument Labeler*.
- Naïve approach for sense labelling and consider only explicit relations.
- Yet, good results.

References

Steven Bethard, Philip Ogren, and Lee Becker. ClearTK 2.0: Design patterns for machine learning in UIMA. LREC, 2014.

David Ferrucci and Adam Lally. UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348, 2004.

Fang Kong, Hwee Tou Ng, and Guodong Zhou. A Constituent-Based Approach to Argument Labeling with Joint Inference in Discourse Parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 68–77, Doha, Qatar, October 2014.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(02):151–184, 2014.

Contact information: Email: *m_laali@encs.concordia.ca †e_davoo@encs.concordia.ca ‡kosseim@encs.concordia.ca; Web: *https://clac.encs.concordia.ca*