10

Neural rhetorical structure parsing with pseudo-labeled data

Revealing hidden structures behind sentences

Abstract

This poster presents a method to identify the hidden structures of documents. Each document has a rhetorical structure, which expresses the relations among clauses. Since building a rhetorical structure parser is based on supervised learning, it requires large amounts of manually annotated training data for accurate parsing. However, conventional methods suffer from a lack of training data, resulting in poor performance because manual annotation is quite labor intensive. To tackle this problem, we propose a method that uses silver data: automatically annotated pseudo-labeled data. We pre-trained the parser with silver data and fine-tuned it with gold data: manually annotated data. Our experimental results demonstrated that our method achieved the best performance. The new parser will contribute to various natural language processing applications, such as machine translation and automatic summarization.

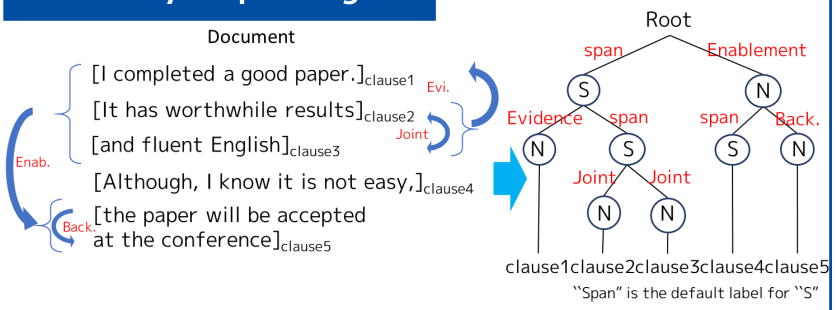
Rhetorical Structure Theory (RST)

RST represents relations among clauses in a document by a labeled binary tree

Leaves of the tree: clauses
Nodes of the tree: span consisting clauses
Labels of the nodes: Role of the span
→Nucleus (N) / Satellite (S)
※An "S" modifies "N"
Labels of the edges: relation label between two spans
→18 classes:
Elaboration, Evidence, etc

Natural language processing applications such as machine translations, automatic summarizations require the trees

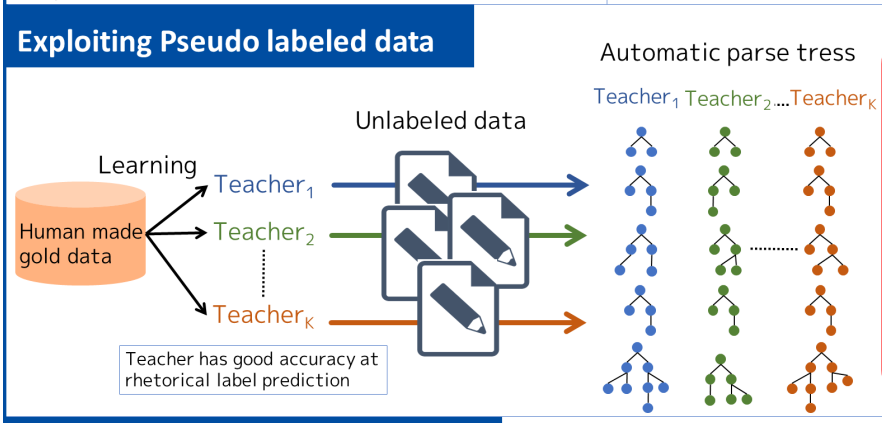
Difficulty of parsing

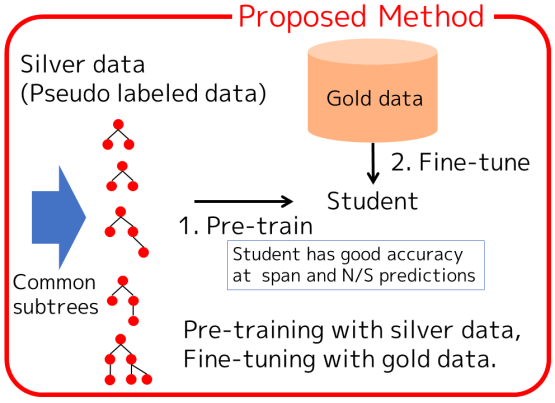


It is hard to annotate large-scale data manually since the annotation is labor-intensive

Only small dataset is available

Exploiting Pseudo labeled data





Effectiveness of our method

F-score (%)

Method＼Metric	Fully-labeled	Span	+N/S label	+Rel. label
Teacher [Wang+17]	58.8	86.0	72.4	59.7
Student [Kobayashi+20]	59.6	**87.1**	74.6	60.0
Proposed method	**62.6** UP!	**87.1**	**75.0**	**63.4** UP!

Fully-labeled score is improved by the gain of relation labeling

This is a collaborative research project between Okumura Lab. at Tokyo Institute of Technology and NTT CS labs.

References

[1] N. Koabayashi, T. Hirao, H. Kamigaito, M. Okumura, M. Nagata, "Improving Neural RST Parsing Model with Silver Agreement Subtrees," in *Proc. 2021 Annual Conference of the Noth American Chapter of the Association for Computational Linguistics*, 2021.

Contact

Tsutomu Hirao / Linguistic Intelligence Research Group, Innovative Communication Laboratory
Email: cs-openhouse-ml@hco.ntt.co.jp