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

EXCELENCIA MARÍA DE MAEZTU

ArgMining 2019
ACL 2019,Florence, Italy
1 August 2019

MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

upf. Universitat Pompeu Fabra Barcelona taln

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

Challenge: Data!

"… 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)

Lippi, M., Torroni, P.: *Argumentation mining: State of the art and emerging trends.* ACM Trans. Internet Technol. 16(2), 10:1-10:25 (2016)

Kirschner, C., Eckle-Kohler, J., Gurevych, I.: *Linking the thoughts: Analysis of argumentation structures in scientific publications.* In: Proceedings of the 2nd Workshop on Argumentation Mining. pp. 1-11 (2015)

Green, N. *Identifying argumentation schemes in genetics research articles*. In Proceedings of the 2nd Workshop on Argumentation Mining (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)

Peldszus, A., Stede, M.: *Rhetorical structure and argumentation structure in monologue text.* In: Proc. of the 3rd Work. on Arg Mining, pp. 103–112 (2016)
Stab, C., Kirschner, C., Eckle-Kohler, J., Gurevych, I.: *Argumentation mining in persuasive essays and scientific articles from the discourse structure perspective.* In: ArgNLP, pp. 21–25 (2014)

Background results

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**

Accuosto, P, Saggion, H.: *Discourse-driven argument mining in scientific abstracts.* In 24th International Conference on Applications of Natural Language to Information Systems, pages 1–13. Springer.

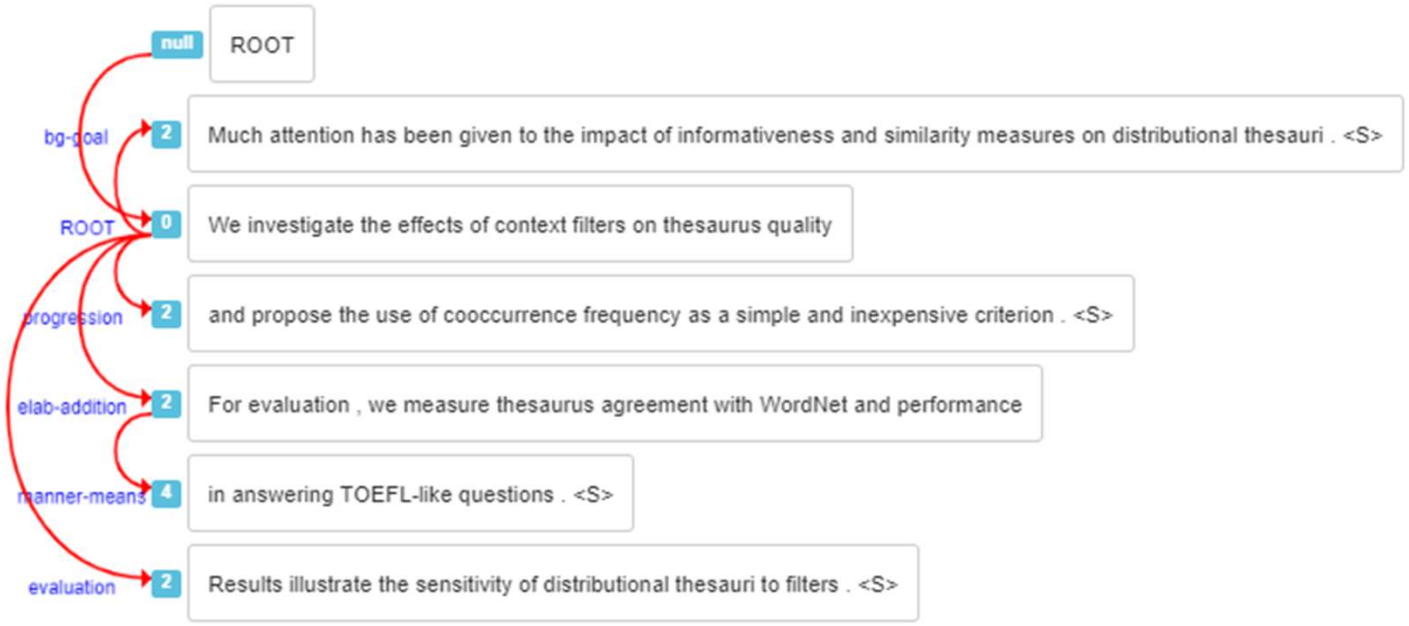
III. SciDTB Corpus

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)



Yang, A., Li, S.: SciDTB: *Discourse dependency treebank for scientific abstracts.* In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Vol. 2, pp. 444-449 (2018)

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

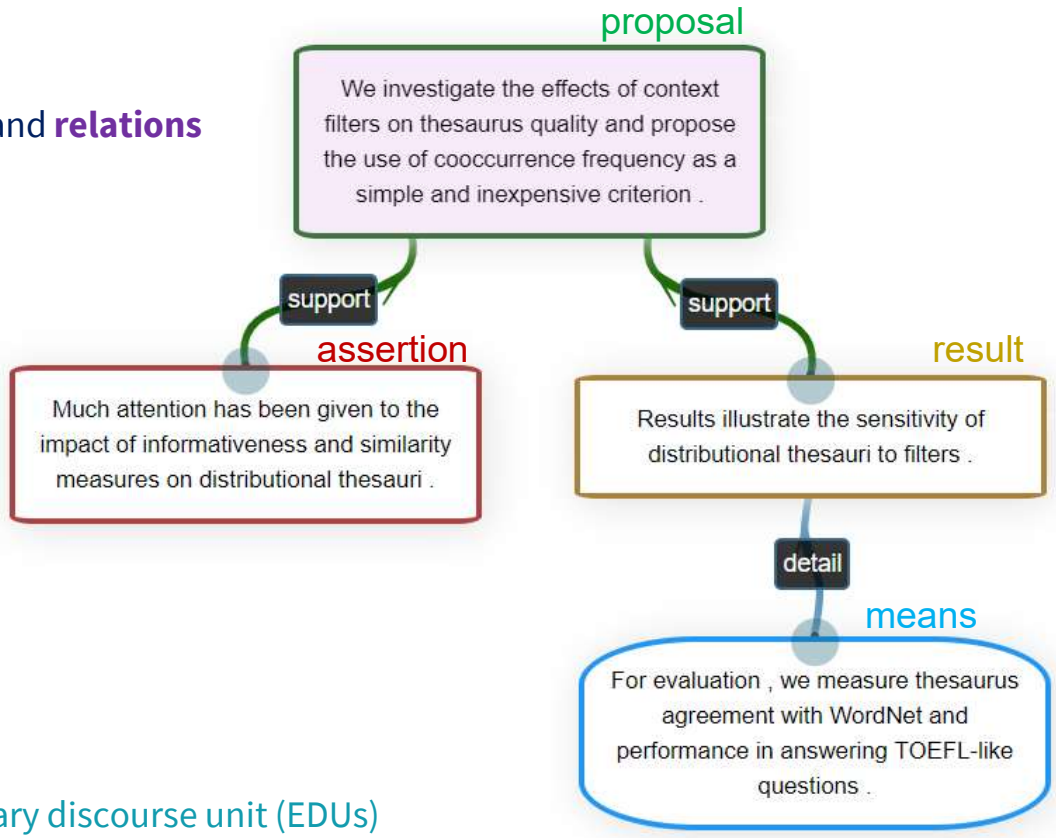
units

claims
- *proposal* (problem or approach)
- *assertion* (conclusion or known fact)

premises
- *result* (interpretation of data)
- *observation* (data)
- *means* (implementation)
- *description* (definitions/other)

relations

- *support*
- *(attack)*
- *detail* (elaboration, means, etc.)
- *sequence* (sequence)
- *additional* (joint)



Argumentative units (AUs): One or more elementary discourse unit (EDUs)

Argumentation layer

Type of unit	%
proposal	31
assertion	25
result	21
means	18
observation	3
description	2

Type of relation	%
detail	45
support	42
additional	9
sequence	4

V. Argument mining experiments

Argument mining tasks

AM Task	Description
ATy	Identify the **type** of argumentative units (e.g.: *proposal*)
AFu	Identify the **function** of the argumentative units (e.g.: *support*)
APa	Identify the relative position of the **parent** argumentative unit (e.g.: *-2*)

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

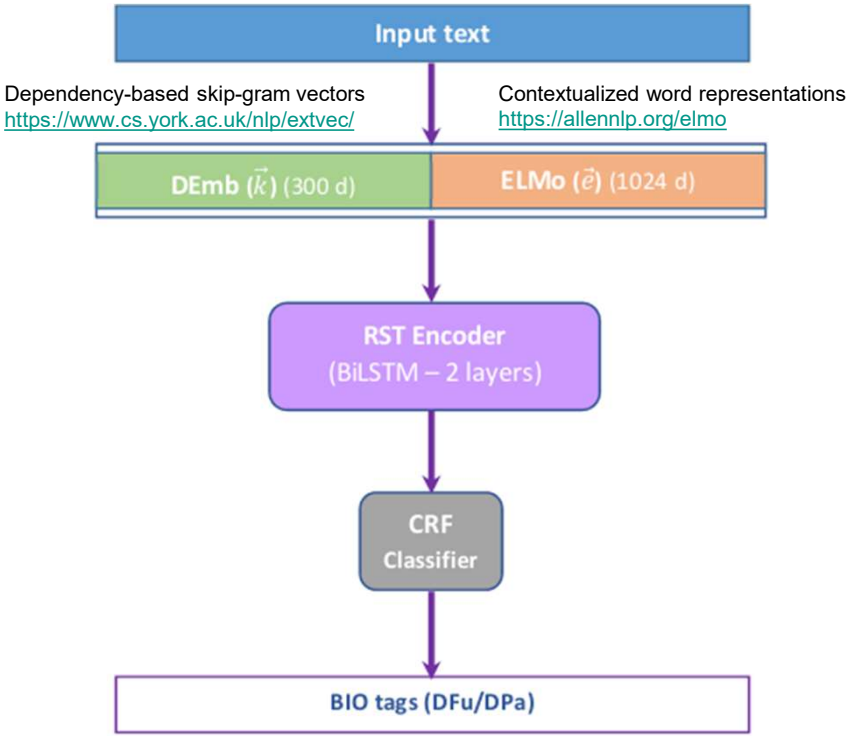
RST Task	Description
DFu	Identify the discourse **roles** of the EDUs (e.g.: *attribution, evaluation*)
DPa	Identify the relative **position** of the parent EDU in the RST tree

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

Experimental settings

Discourse models

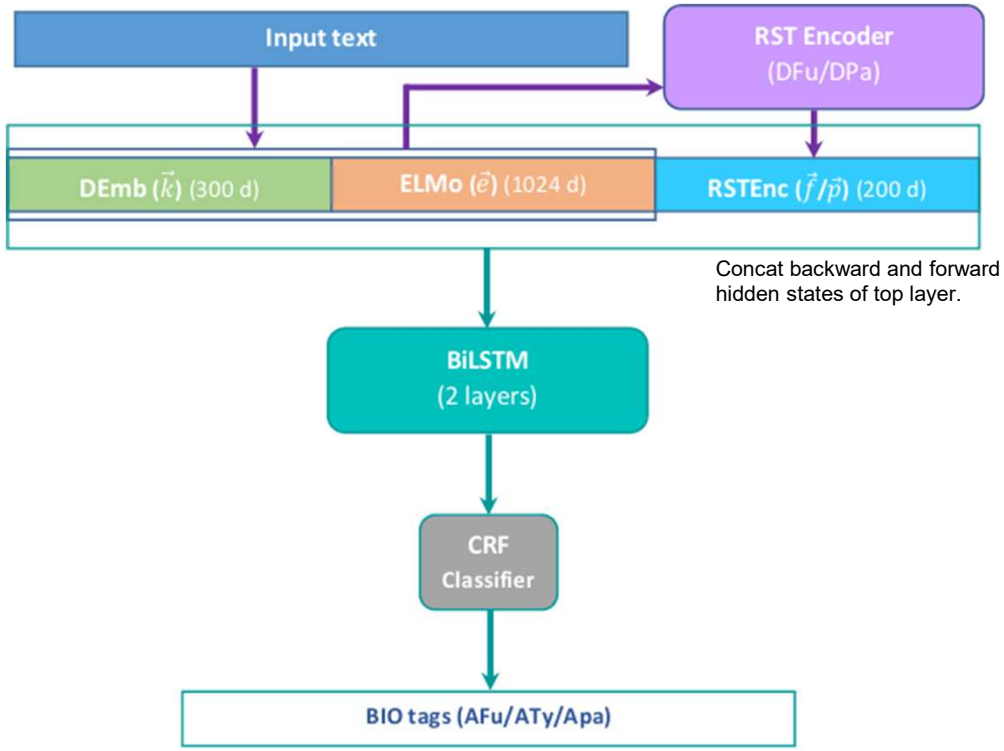
Trained with 738 abstracts:
SciDTB – 60 annotated with arguments



Reimers, N., Gurevych, I.: *Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging*. EMNLP (2017)
https://github.com/UKPLab/emnlp2017-bilstm-cnn-crf/

Experimental settings

Argument mining models



Results

Setting	AFu	ATy	APa
DEmb+ELMo	0.66	0.63	0.38
DEmb+ELMo+RSTEnc	**0.69**	**0.67**	**0.40**

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

Setting	AFu	ATy	APa
DEmb+ELMo	0.66	0.63	0.38
DEmb+ELMo+GloVe	0.65	0.65	0.38
DEmb+ELMo+RSTEnc	**0.69**	**0.67**	**0.40**

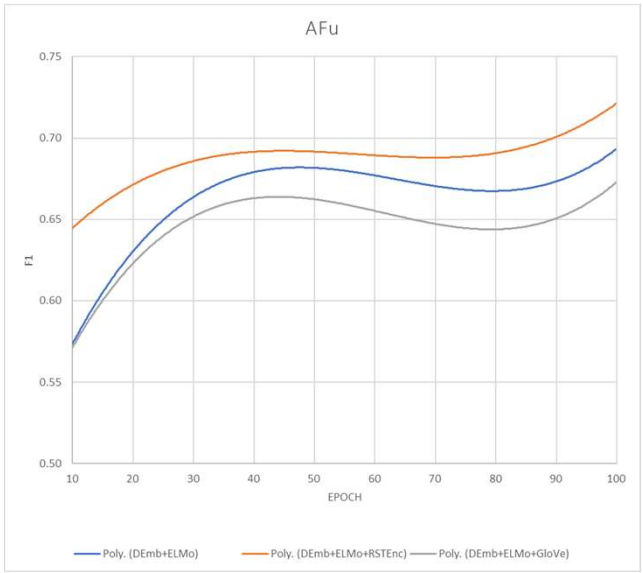
Average F1 scores for epochs 10 to 100

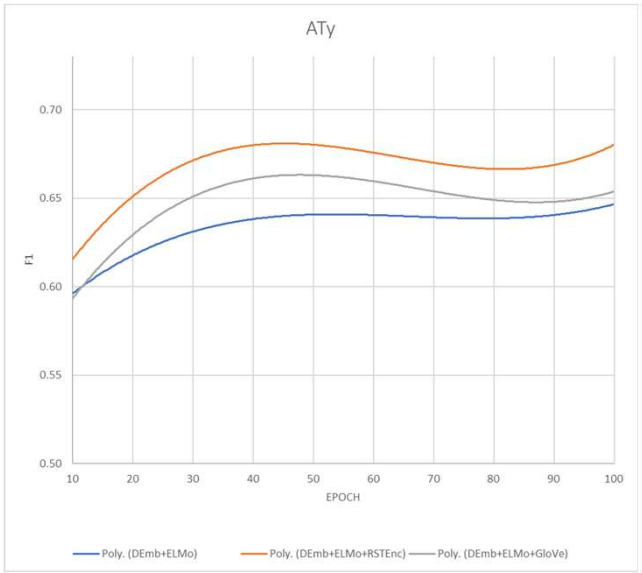
Results

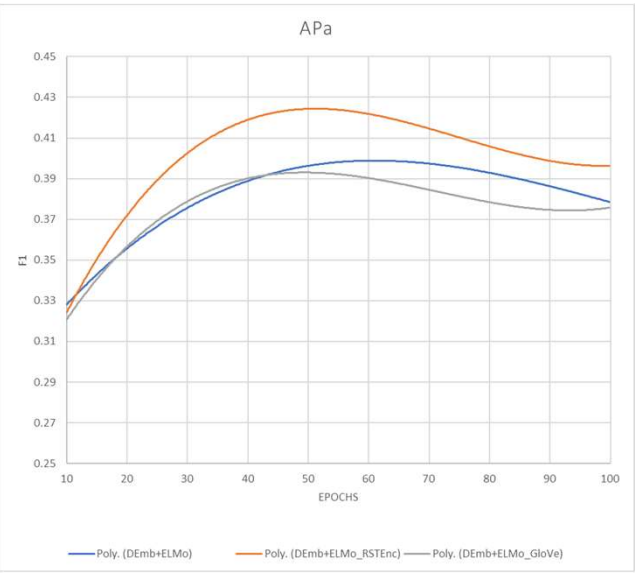
Setting	support	proposal	assertion	result
DEmb+ELMo	0.61	0.67	0.65	0.61
DEmb+ELMo+RSTEnc	**0.63**	**0.71**	**0.67**	**0.63**

Average F1 scores for epochs 10 to 100

Results







Polynomial trend lines for F1 in epochs 10-100 for AFu, ATy, APa

Results

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

Acceptance prediction

As an application, we explore whether the **argumentative structure** of the abstracts can predict **acceptance / rejection** of papers in computer science venues.

Dataset

	Conference	Accepted	Rejected
Training (117)	*CDNNRIA 2018*	35	23
	IRASL 2018	30	29
Test (30)	*ICLR 2018*	15	15

- *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)

AFu				ATy				APa				
none	support	...	support	proposal	result	...	observation	0	1	...	3	REJECT
additional	support	...	–	assertion	assertion	...	–	1	1	...	–	ACCEPT
...	...	...	...	...	...	...	...	...	...	...	...	...
support	none	...	–	assertion	proposal	...	–	1	0	...	–	ACCEPT

Results

Algorithm/parameters set with 20-80 random split of training set

Classifier	P	R	F1
Random	0.50	0.50	0.50
Decision tree	**0.67**	**0.67**	**0.67**

Acceptance classification results

Decision points (and feature analysis) show that all three types of features are relevant for classification.

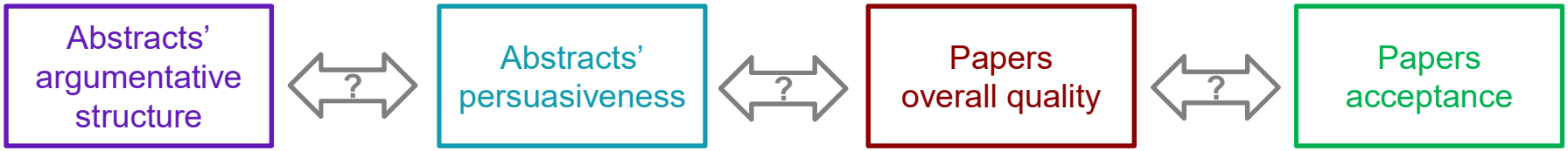
E.g.: The parent of first unit, the functions of the first two units and the type of the first unit are particularly informative.

Acceptance prediction

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 require to know what the potential correlation means.



We are making no claims with respect to these relations.

VII. Conclusions and future work

Conclusions

- 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