An Argument-Annotated Corpus of Scientific Publications

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto
Data and Web Science Research Group
University of Mannheim

Motivation

- Growing number of scientific publications [1] raises the need for computational analysis of the rhetorical aspects of scientific writing (***scitorics***)
- Scientific publications are inherently argumentative [2, 3]

Problem: No publicly available corpus of scientific publications in English annotated with fine-grained argumentative structures for training machine learning-models

Contributions

- An argument annotation-scheme for scientific publications
- Extention of the Dr. Inventor Corpus [7, 8] with argument-annotations
- Statistical and information-theoretic analysis of the corpus

Annotation Process

- 1 expert (computer science) + 3 non-expert annotators (social sciences + humanities)
- Callibration phase with five iterations (IAA measured in F1)

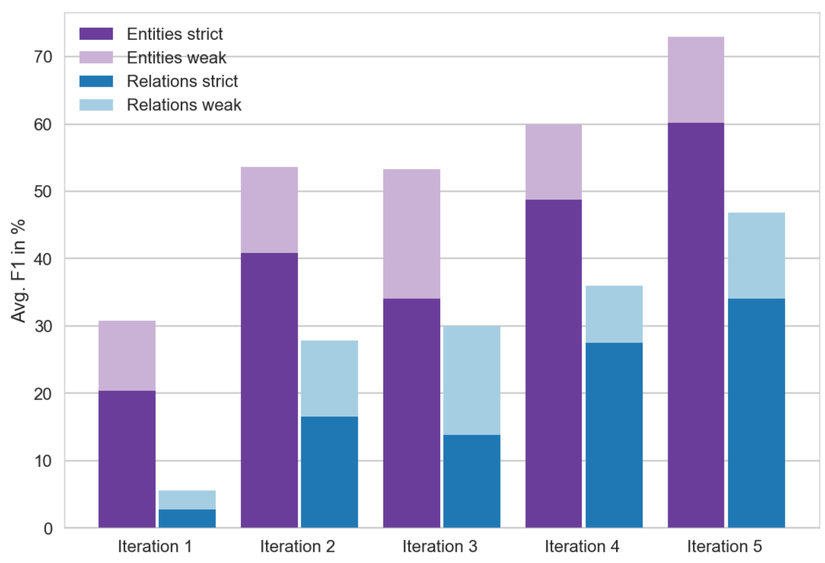

Fig 1: Evolution of the IAA over the 5 callibration phases.

Analysis of the Argument Annotations

Category	Label	Total	Per Publication
Component	Background claim	2,751	68.8 ± 25.2
	Own claim	5,445	136.1 ± 46.0
	Data	4,093	102.3 ± 32.1
Relation	Supports	5,790	144.8 ± 43.1
	Contradicts	696	17.4 ± 9.1
	Semantically same	44	1.1 ± 1.81

Tab 1: Total and per-publication distributions of labels of argumentative components and relations identified.

Label	Min	Max	Avg	Std
Background claim	5	340	87.46	43.74
Own claim	3	500	85.70	44.03
Data	1	244	25.80	27.59

Tab 2: Statistics on the length of argumentative components in the extended Dr. Inventor Corpus (in characters).

Annotation Scheme

- Derived from Toulmin, Bench-Capon, Dung [4, 5, 6, *inter alia*]
- Components
  - Own Claim

    "Furthermore, we show that by simply changing the initialization and target velocity, the same optimization procedure leads to running controllers."
  - Background Claim

    "Despite the efforts, accurate modeling of human motion remains a challenging tasks."
  - Data

    "[...], due to memory and graphics hardware constraints nearly all video game character animation is still done using traditional SSD."
- Relations
  - Supports →
  - Contradicts ↔
  - Semantically same ↔

Data: The Dr. Inventor Corpus [7, 8]

- 40 publications in the domain of computer graphics
- Existing annotation layers:
  - Discourse Roles,
  - Citation Contexts + Citation Purposes,
  - Subjective Aspects, Summarization Relevance

Links to other Rhetorical Aspects

	ArgComp	DiscRoles	SubjAsp	SummRel
ArgComp	-	-	-	-
DiscRoles	**0.22**	-	-	-
SubjAsp	0.08	0.11	-	-
SummRel	0.04	0.10	0.13	-
CitContexts	**0.18**	0.10	0.04	0.01

Tab 3: Normalized mutual information between pairs of label sets.

References

[1] Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology, 66(11):2215–2222.

[2] G Nigel Gilbert. 1976. The transformation of research findings into scientific knowledge. Social Studies of Science, 6(3-4):281–306.

[3] G Nigel Gilbert. 1977. Referencing as persuasion. Social Studies of Science, 7(1):113–122.

[4] Stephen E. Toulmin. 2003. The Uses of Argument, updated edition. Cambridge University Press.

[5] Trevor JM Bench-Capon. 1998. Specification and implementation of toulmin dialogue game. In Proceedings of the 11th Conference on Legal Knowledge Based Systems, pages 5–20, Groningen, Netherlands. Foundation for Legal Knowledge Based Systems.

[6] Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence, 77(2):321–357.

[7] Beatriz Fisas, Francesco Ronzano, and Horacio Saggion. 2016. A multi-layered annotated corpus of scientific papers. In Proceedings of the International Conference on Language Resources and Evaluation, pages 3081–3088, Portoroz, Slovenia. European Language Resources Association.

[8] Beatriz Fisas, Horacio Saggion, and Francesco Ronzano. 2015. On the discoursive structure of computer graphics research papers. In Proceedings of The 9th Linguistic Annotation Workshop, pages 42– 51, Denver, CO, USA. Association for Computational Linguistics.

Code & Data?

http://data.dws.informatik.uni-mannheim.de/sci-arg/
https://github.com/anlausch/sciarg_resource_analysis/