Do Discourse Indicators Reflect the Main Arguments in Scientific Papers?

Do Discourse Indicators Reflect the Main Arguments in Scientific Papers?

Yingqiang Gao[†], **Nianlong Gu**[†], **Jessica Lam**[†]
Richard H.R. Hahnloser[†]

[†]Institute of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland
{ yingqiang.gao, nianlong, lamjessica, rich }@ini.ethz.ch

Abstract

In scientific papers, arguments are essential for explaining authors' findings. As substrates of the reasoning process, arguments are often decorated with discourse indicators such as "which shows that" or "suggesting that". However, it remains understudied whether discourse indicators by themselves can be used as effective markers of the local argument components (LACs) in the body text that support the main claim in the abstract, i.e., the global argument. In this work, we investigate whether discourse indicators reflect the global premise and conclusion. We construct a set of regular expressions for over 100 word- and phrase-level discourse indicators and measure the alignment of LACs extracted by discourse indicators with the global arguments. We find a positive correlation between the alignment of local premises and local conclusions. However, compared to a simple textual intersection baseline, discourse indicators achieve lower ROUGE scores and have limited capability of extracting LACs relevant to the global argument; thus their role in scientific reasoning is less salient as expected.[1]

1 Introduction

Arguments are made by presenting cascades of argument components (ACs) called *premises* and *conclusions*, where the premises are intentional justifications that lend credibility to the conclusions (Wyatt, 2001; Stede and Schneider, 2018). In scientific papers, arguments aim to make claims supported by evidences taken from experiments, observations, and references (Al Khatib et al., 2021), and are usually presented as premise-conclusion pairs that are linked via an argumentative relation (Prasad et al., 2008; Lee et al., 2016). In scientific papers, the main claim or *global argument* of a paper is drawn in the abstract and several *local argument components* (LACs) are formulated throughout the entire body text. However, extracting LACs that support the global argument is hard because of the difficulties in finding premise-conclusion pairs.

Example LAC in our dataset
Assuming that [gene duplications primarily evolve under purifying selection]premise, [the observed acceleration of evolution may be explained by epistatic interaction between gene copies]conclusion.
regex rule: *Assuming that* P, C

Table 1: An example of discoure indicator *Assuming that* which links the premise and conclusion together. P represents the premise and C the conclusion. Best viewed under color printing.

It has been claimed that discourse indicators can be used to extract ACs in unstructured text, such as news articles (Sardianos et al., 2015) and student essays (Stab and Gurevych, 2014; Persing and Ng, 2016). However, the alignment between premises and conclusions in scientific papers is often implicit, especially when several premises correspond to one particular conclusion. Moreover, the extraction rules for ACs strongly depend on the pre-defined argumentation scheme and often do not generalize well (Walton et al., 2008; Prakken, 2010). Kirschner et al. (2015) have annotated a small corpus of 24 scientific papers, but the argumentative relation scheme is only binary (attack or support) and thus cannot represent more complex argumentative relations. Finally, the relation between arguments in the abstract and the body text remains understudied. Therefore, although a lot of progress in mining arguments from unstructured texts (Reed and Rowe, 2004; Van Gelder, 2007; Bex et al., 2014; Ong et al., 2014; Persing and Ng, 2015) has been made, it remains unclear whether discourse indicators can extract LACs that support the global argument in structured texts such as sci-

[1]Data and code are available at https://github.com/CharizardAcademy/discourse-indicator

Proceedings of the 9th Workshop on Argument Mining, co-located with the 29th International Conference on Computational Linguistics, pages 34–50
October 17, 2022.

entific papers.

In this work, we create a sizeable scientific paper dataset consisting of biomedical papers with well-structured abstracts, which enables us to easily extract the global argument of papers. On this dataset, we propose an efficient discourse indicator-based LAC extraction pipeline. We first construct a set of regular expressions of argument-associated discourse indicators; then, for each regular expression, we define how the local premise and the local conclusion are organized either in the sentence or in two consecutive sentences that are linked by this discourse indicator. With this pre-defined set of rules, we extract and disentangle the local premise from the local conclusion, which serve as LACs (see Table 1). To evaluate the effectiveness of our discourse indicator-based LAC extraction pipeline for scientific papers in terms of reflecting the global argument, we first compute the ROUGE-N scores of the union of all LACs extracted by our pipeline with respect to the global argument, and further qualitatively evaluate the extracted LACs and compare with the baselines via human evaluation.

Our main **contributions** are: 1) We propose a set of regular expressions for over 100 word- and phrase-level discourse indicators for extracting LACs from the body text of scientific papers; 2) We show that counter-intuitively, LACs extracted by discourse indicators only poorly reflect the global argument, by the fact that LACs extracted with discourse indicators achieved lower ROUGE-N scores than a simple baseline approach; 3) Human evaluation results suggest that LACs extracted by discourse indicators are precise in the exact wordings, but do not have a high information coverage of the global argument.

2 Related Works

The task of extracting LACs is most similar to argument mining (Lawrence and Reed, 2015, 2017, 2020), which typically classifies sentences into argumentative and non-argumentative text according to their rhetorical and syntactic role. Argument mining usually depends on a carefully designed argumentation scheme, which is, in general, a predefined type of connection between premise and conclusion. Teufel et al. (1999) proposed the first argumentative scheme which was later expanded to 14 categories of ACs (e.g. AIM, SUPPORT, USE, etc.) in scientific texts (Teufel et al., 2009). In our work, we consolidate the argumentation scheme simply as premise-conclusion pairs.

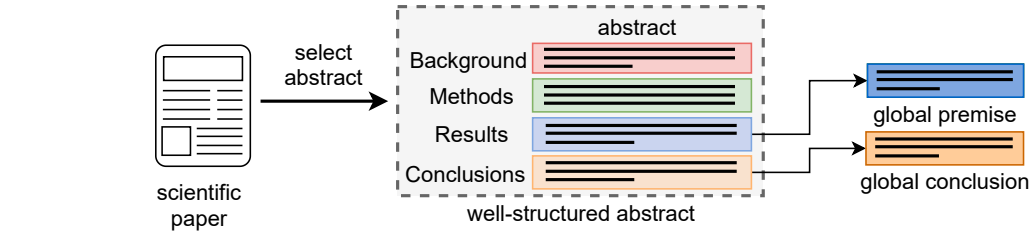
Discourse indicators have been used as rhetorical features to determine the credibility of claimed premises in support of a conclusion (Freeman, 2000). As a milestone, Wyner et al. (2012) showed that premise-conclusion pairs could be located by discourse indicators. Eckle-Kohler et al. (2015) annotated a corpus including 88 German language documents of premise-conclusion pairs and found that particular discourse indicators are more closely linked to either premises or conclusions. Lawrence and Reed (2015) used a small set of discourse indicators to extract premise-conclusion pairs and achieved high precision in recognizing the connections between propositional segments. In their later work (Lawrence and Reed, 2017), they further leveraged contextual knowledge such as topic information by constructing an inferential matrix that indicated the propositional relations, including premise-conclusion pairs. Argument mining has also been studied in series of works of Moens et al. (Moens et al., 2007; Palau and Moens, 2009; Mochales and Moens, 2011), where sentences are classified into *Arguments* and *Non-arguments* in an unsupervised manner using syntactic and semantic features. In these studies, the extraction of ACs is mainly done on the sentence level.

Nevertheless, in these works discussed above, the contribution of discourse indicators alone is not clear, and often the power of discourse indicators are only partially studied for news articles (e.g. Eckle-Kohler et al. (2015)). Unlike news articles, which are often written in plain language and are easy to understand, the readability of scientific papers decreases over time (Plavén-Sigray et al., 2017). In this work, we focus on understanding the role of discourse indicators in scientific papers particularly, mainly on how they contribute to extracting LACs in the body text supporting the global argument in the abstract.

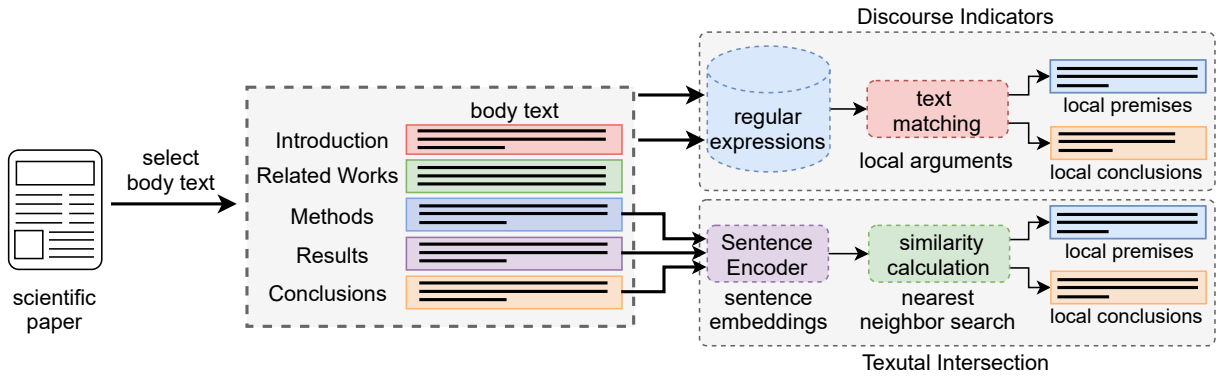
3 Methodology

This section outlines our approaches to extracting global arguments and LACs from scientific papers (see Figure 1 for the proposed pipelines). In this work, we use the term *global* and *local* to refer to argument components located in the abstract and the body text of a paper separately.

We make the following assumptions: 1) Every scientific paper has one global argument and several paired LACs. The global argument expresses



(a) Global argument extraction using well-structured abstracts.



(b) LAC extraction using discourse indicators and textual intersection.

Figure 1: Our argument mining pipelines for biomedical papers: (a) global arguments are extracted from the well-structured abstracts with headers; (b) LACs are extracted from the body text. The textual intersection approach only makes use of *method*, *result* and *conclusion* sections, while the discourse indicator approach leverages the whole body text. We use well-structured abstracts to get the best human labeled global arguments we can find.

the paper's central claim, whereas LACs are individual statements that support the global argument from diverse perspectives; 2) The global argument locates in the paper's abstract, whereas LACs are distributed across the entire body text of the paper.

3.1 Mining Global Argument Components

In order to measure how well extracted LACs reflect the global argument, we first need to extract the global arguments from the abstracts because raw abstracts might also contain non-argumentative text. To ensure we have pure argumentative text extracted as the global argument, we use well-structured abstracts that contain both *result* and *conclusion* headers.

Since the naming convention of headers across different papers can vary greatly, we categorize headers such as "result" and "outcome" as *result* headers and headers such as "conclusion" or "concluding" as *conclusion* headers. A complete list of critical strings for *result*/*conclusion* headers is provided in appendix B. The text after the recognized headers are identified as the global argument: the text after the "result" header was extracted as the global premise and the text after the "conclusion" header as the global conclusion.

3.2 Mining Local Argument Components

Inspired by the work of Lawrence and Reed (2015, 2017), we use a broad set of over 100 discourse indicators both on the word level (e.g. *because*) and the phrase level (e.g. *assuming that*). Each discourse indicator extracts one local premise p_i^{local} and one local conclusion c_i^{local} on either the sub-sentence or the sentence level (see appendix C). The assessments were defined based on the mutual agreement of five experienced experts.

We concatenate the extracted local premises p_i^{local} $(i = 1, ..., .n)$ of all n matched discourse indicators to form the set of local premises P_{local}; similarly, we form the set of local conclusions C_{local} by concatenating all extracted local conclusions $c_i^{local} (i = 1, \ldots, n)$.

Textual Intersection Baseline As a baseline for LAC extraction, we propose an embedding-based approach to extract LACs solely from the *result* and *conclusion* sections. The idea is that sentences in the *result* section that are similar to sentences in the *method* section serve as local premises P_{local}.

The baseline extraction of LACs works as follows:

1. Similar to our definition of global argument, we used the same set of critical strings to parse the section names in the body text of scientific papers and recognize the *method*, *result*, and *conclusion* sections.

2. We remove stopwords, special symbols as well as extra blanks from the section paragraphs, then we tokenize the paragraphs into sentences using the NLTK[2] package (version 3.6.2).

3. For the i_{th} sentence s_m^i in the *method* section S_m and the j_{th} sentence s_r^j in the *result* section S_r, we compute 600-dimensional sentence embeddings e_m^i and e_r^j using a pre-trained universal text encoder, Sent2vec[3] (Pagliardini et al., 2018).

$$e_m^i = \text{Sent2vec}(s_m^i)$$
$$e_r^j = \text{Sent2vec}(s_r^j).$$

We form the set of local premises P_{local} as a collection of *result* sentences that have similarity higher than a threshold value θ against any *method* sentence. Here sentence similarity is measured with the cosine similarity between the sentence embeddings:

$$P_{local} = \left\{ s_r^j \in S_r : \max_{s_m^i \in S_m} \text{sim}(e_m^i, e_r^j) \geqslant \theta \right\}$$

$$\text{where} \quad \text{sim}(e_m^i, e_r^j) = \frac{e_m^i \cdot e_r^j}{||e_m^i|| \cdot ||e_r^j||}.$$

4. We perform the same textual intersection step using the *result* and *conclusion* sentences. The set of local conclusions C_{local} is therefore a collection of *conclusion* sentences whose maximum cosine similarity against *result* sentences is greater than the threshold ϵ:

$$C_{local} = \left\{ s_c^k \in S_c : \max_{s_r^j \in S_r} \text{sim}(e_r^j, e_c^k) \geqslant \epsilon \right\}$$

$$\text{where} \quad \text{sim}(e_r^j, e_c^k) = \frac{e_r^j \cdot e_c^k}{||e_r^j|| \cdot ||e_c^k||}.$$

Both premise threshold θ and conclusion threshold ϵ were set to 0.1 to encourage extracting diverse LACs of rich semantics.

[2]Apache License 2.0, available at https://github.com/nltk/nltk

[3]BSD License, available at: https://github.com/epfml/sent2vec

4 Dataset

Our proposed argument mining pipelines are applied to the Semantic Scholar Open Research Corpus, i.e., S2ORC[4] (Lo et al., 2020), which is an extensive collection of 81.1M well-parsed peer-reviewed English papers, among which around 12.7M are complete with full text.

From the S2ORC corpus, we create a subset of nearly 28k papers in the biomedical domain with full text and structured abstracts available. We use papers with well-structured abstracts of biomedical papers to extract the global arguments due to the following reasons: 1) well-structured abstracts are the best massive human annotated source of global arguments we can get since these papers are peer-reviewed and usually multi-round editor-revised, therefore the quality of the argumentative text is ensured; 2) many journals specialized for biomedicine research naturally require the authors to construct the abstract in a structured manner, where the argumentative text is purposely decomposed into different units; 3) a previous study (Shieh et al., 2019) demonstrates the success of generating global conclusions from global premises using the well-structured abstracts of PubMed papers, thus enlightening the usefulness of well-structured abstracts for mining argument components.

For each paper in our dataset, we extract the LACs using both discourse indicators and textual intersection approaches. We also compute the upper bound of the ROUGE f-measure performance using the greedy strategy of Gu et al. (2022) that iteratively selects sentences to approximately maximize the sum of ROUGE-1 and ROUGE-2 f-measure scores. Table 2 shows the statistics of our proposed dataset *scinf-biomed*. Notice that for LACs extracted with discourse indicators, one local conclusion corresponds to one local premise due to our assessments of discourse indicators, whereas for the textual intersection approach, there is no one-to-one mapping between local conclusions and local premises. In Table 11 of appendix D we demonstrate the LACs extracted by the two proposed approaches.

5 Evaluation

To evaluate the performance of our proposed approaches, we perform the local-to-global comparison between the LACs and the global argument us-

[4]CC BY-NC 2.0 License, available at https://github.com/allenai/s2orc

Dataset	size	global args		local args (*d*)		local args (*t*)		local args (greedy)	
	#papers	#*con*	#*pre*	#*con*	#*pre*	#*con*	#*pre*	#*con*	#*pre*
scinf-biomed	27,924	61,809	133,480	71,245	75,379	179,654	319,272	63,245	136,282

Table 2: Statistics of the dataset of the extracted arguments. #papers represents the number of papers being selected, #*con* and #*pre* denote number of extracted local conclusions and local premises, *d* and *t* denote discourse indicators approach and textual intersection baseline. For LACs extracted using discourse indicators, #*con* and #*pre* are counted for non-empty local conclusions and local premises.

ing the summarization metric ROUGE scores as the automatic evaluation (Lin, 2004). Inspired by the pilot study on argument sufficiency of Gurcke et al. (2021), which showed that conclusion sentences generated from sufficient premises share more word-level commonalities, we choose ROUGE as the evaluation metric to measure the lexical relevance of the extracted LACs, based on the intuition that global arguments in the abstract can only be inferred from local arguments in the body text if they contain sufficient lexical information.

We first concatenate the LACs within their original order of occurrence in the body text, then we average the ROUGE-1, ROUGE-2, and ROUGE-Lsum scores[5] for precision, recall, and f-measure. All evaluations are performed separately for 1) the extracted local conclusions (against the global conclusions); 2) the extracted local premises (against global premises). Discourse indicators themselves are excluded from LACs while computing the ROUGE scores.

In addition, we are particularly interested in the n-gram precisions of the LACs compared to the global argument, since they provide information about whether n-grams in the global argument are favored in local conclusions or local premises. Therefore, we use the ROUGE-N precision as the metric to evaluate the lexical preferences of LACs.

6 Results and Discussion

6.1 Local-to-global comparison

In Table 3, we compare average ROUGE f-measures of the global argument against LACs (both local conclusions and local premises) extracted either with discourse indicators or with our baseline textual-intersection approach. The greedy oracle serves as the theoretical upper bound of the average ROUGE f-measures. In Table 4, we indicate how LACs extracted by the greedy oracle are distributed across sections.

approach	ROUGE-1	ROUGE-2	ROUGE-Lsum
greedy-*con*	62.10	43.75	56.93
indicator-*con*	23.76	5.88	21.72
intersection-*con*	**40.27**	**25.51**	**36.47**
greedy-*pre*	58.00	35.97	53.45
indicator-*pre*	23.76	4.85	20.92
intersection-*pre*	**38.09**	**20.08**	**33.81**

Table 3: Averaged ROUGE f-measures for local-to-global comparison of local conclusions (*con*) and local premises (*pre*) using discourse indicators and textual intersection with similarity thresholds $\theta = 0.1, \epsilon = 0.1$.

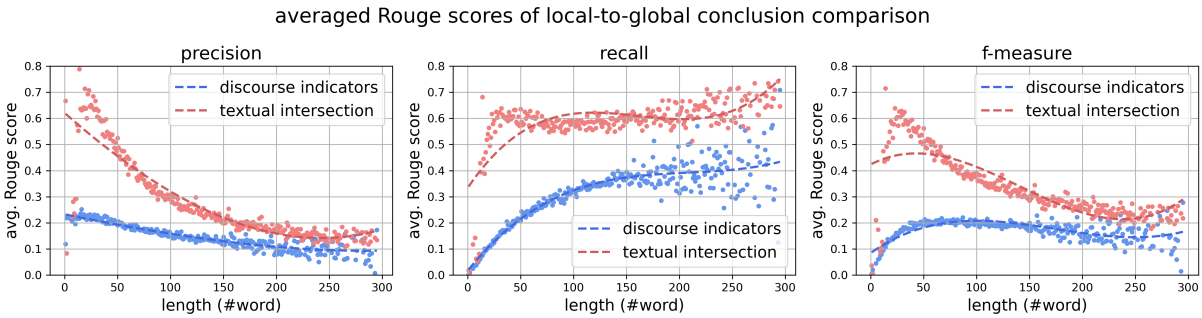
approach	sections	#sent	ratio
greedy-*con*	*conclusion*	33,036	52.2 %
	result	3,999	6.3 %
	method	2,247	3.6 %
greedy-*pre*	*result*	68,759	50.5 %
	method	11,163	8.2 %
	conclusion	6,332	4.7 %

Table 4: Statistics of the extracted LACs using the greedy approach. #sent means the number of sentences extracted from different sections, where ratio is the percentage to all greedily extracted LACs.

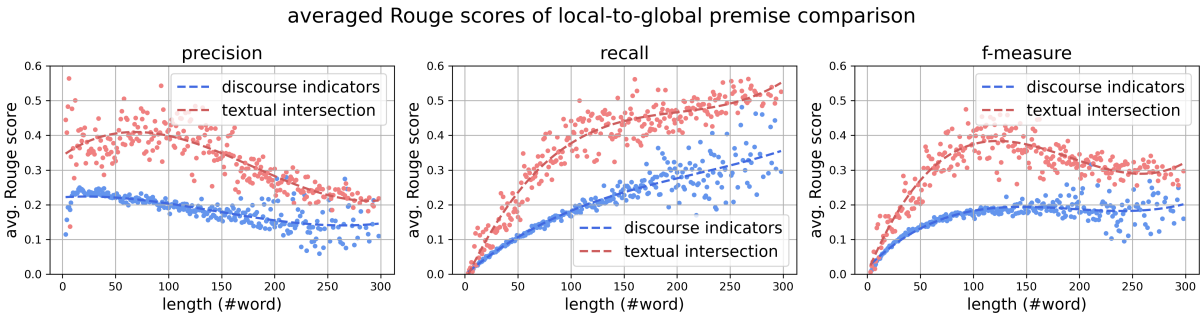
We found that local conclusions and local premises extracted with textual intersection achieve higher average ROUGE scores than those extracted by discourse indicators. This finding suggests that LACs retrieved with discourse indicators are not as well-aligned with the global argument as compared to LACs extracted by the textual intersection baseline. Thus, LACs linked by discourse indicators share less textual commonality with the global argument than those extracted by the textual intersection baseline.

LACs extracted by the two approaches tend to

[5]We use the python package rouge_score (version 0.0.4) to compute the ROUGE measures (Apache 2.0 License). https://pypi.org/project/rouge-score/



(a) Local conclusions compare against global conclusions.



(b) Local premises compare against global premises.

Figure 2: Averaged Rouge scores for local-to-global comparison of premises and conclusions. We choose small similarity thresholds for the textual intersection ($\theta = 0.1$, $\epsilon = 0.1$) to encourage LACs of diverse semantics being extracted. The extracted local premises and local conclusions are limited to the first 300 words for a fair comparison. Best viewed under color printing.

have different lengths. To eliminate the influence of LAC length on ROUGE performance, we compared LACs extracted by the two approaches for a given length. Figure 2 illustrates the average ROUGE scores as a function of the length (number of unigrams) of concatenated LACs. To better visualize the overall trend, for each average ROUGE score, we fit the data with a third-order polynomial (dashed lines in Figure 2).

max. *pr*@10	ROUGE-1	ROUGE-2	ROUGE-Lsum
indicator-*con*	49.06	23.91	47.07
indicator-*pre*	37.02	9.34	35.15
max. *pr*@30	ROUGE-1	ROUGE-2	ROUGE-Lsum
indicator-*con*	34.86	9.60	33.46
indicator-*pre*	34.66	6.83	33.54
max. *pr*@60	ROUGE-1	ROUGE-2	ROUGE-Lsum
indicator-*con*	28.43	6.78	26.57
indicator-*pre*	31.52	5.61	29.97

Table 5: ROUGE-N precisions for local-to-global comparison of local conclusions (*con*) and local premises (*pre*) using top 10, 30, and 60 discourse indicators ranked by averaged ROUGE-N precisions.

We observed that regardless of LAC length, discourse indicators consistently achieved lower performance than the textual intersection baseline. This suggests that LACs linked by discourse indicators do not reflect the global argument well.

6.2 Analysis

We hypothesize that the inferior performance of discourse indicators can be attributed to two aspects: 1) not all discourse indicators are equally useful for the task; 2) discourse indicators are not exclusively used for constructing arguments.

To verify the first hypothesis, we first score each discourse indicator by the average ROUGE-N precision of LACs it extracts. Table 10 of appendix C shows that some discourse indicators like *wherefore* and *on this account* have high scores, whereas other discourse indicators such as *indicating that* and *this is shown by* have much lower scores. In Table 12 of appendix C, we provide an example of LACs extracted by these two discourse indicators.

We evaluated the LACs extracted by the top-k ($k = 10, 30, 60$) discourse indicators in terms of their average ROUGE-N precisions compared to

(a) Top 20 discourse indicators ranked by number of hits.

indicator	#hits	indicator	#hits	indicator	#hits	indicator	#hits
therefore	12,659	*results from*	3,567	*indeed*	2,120	*in conclusion*	1,612
thus	7,194	*resulting in*	3,005	*hence*	2,076	*indicating*	1,612
suggested that	5,324	*is based on*	2,736	*accordingly*	1,918	*demonstrates that*	1,223
because	4,730	*indicates that*	2,628	*in fact*	1,846	*can cause*	1,164
if	4,030	*since*	2,532	*due to*	1,821	*is supported by*	968

(b) Location of indicators (#hits) by sections.

sections	#hits total local conclusions	#hits total local premises	#hits/1k words local conclusions	#hits/1k words local premises
method	6,844	6,948	4.88	4.88
result	7,601	6,929	4.64	4.60
conclusion	5,860	4,434	5.94	5.43
other	50,940	57,068	4.14	4.10
$\sum$ sections	71,245	75,379	4.32	4.26

(c) Average n-gram precision per section.

section	avg. unigram precision		avg. bigram precision		avg. trigram precision	
	premise$_g$	conclusion$_g$	premise$_g$	conclusion$_g$	premise$_g$	conclusion$_g$
method	11.45±6.35	8.11±5.23	3.04±2.92	2.16±2.70	1.10±2.08	0.89±2.21
result	**12.83**±7.14	8.51±5.47	**3.59**±3.43	2.33±2.94	**1.39**±2.67	1.01±2.51
conclusion	12.22±6.91	**10.44**±6.67	3.32±3.61	**3.03**±3.06	1.35±2.92	**1.69**±3.35
other	11.57±6.19	8.62±5.15	2.96±2.80	2.20±2.53	1.05±2.00	0.93±2.10

Table 6: Precision of discourse indicators: (a) discourse indicators ranked by number of hits in the body text of papers; (b) number of discourse indicators in the sections, and the corresponding percentage indicator densities, for local conclusions and local premises within the same section; (c) average n-gram precision with standard deviation, reported for each section. Local premises in the *result* sections achieve higher precision than local premises in the *method* sections, ANOVA test for all n-grams are with $p < 0.01$. Local conclusions in the *conclusion* sections achieve higher precision than local conclusions in other sections. The subscript g denotes global argument.

the global argument. The more discourse indicators we include (the larger k), the lower the average ROUGE-N precision (see Table 5). We also see the average ROUGE-N scores of local conclusions decrease more than the scores of local premises. This suggests that the relevance of discourse indicators varies greatly, i.e., LACs linked by certain discourse indicators are much better aligned with the global argument than others.

To verify the second hypothesis, we compute the overall number of appearances of discourse indicators and the hit rate per 1000 words for different types of sections (see Table 6). We found that regardless of the section type, the hit rate is around 4 to 5, which reveals no distinct section preference of discourse indicators. This may be because scientific papers can contain arguments all through the body text, or because discourse indicators may be overused in non-argumentative occasions for decorative purposes where no scientific reasoning is needed.

As pointed out earlier, we are particularly interested in analyzing the n-gram precision of each LAC with the global argument, to detect re-uses of global-argument n-grams in the LACs.

In Table 6, we show the average n-gram precision in different sections. We see that unigram precision of both local premises and local conclusions are similarly distributed in *method* and *result* sections (see Figure 4 in appendix A), revealing no strong preference for either these section types. Nevertheless, the local premises extracted from the

result sections achieve significantly higher precision with respect to the global premises than from the *method* and *conclusion* sections, revealing a preference for local premises to occur in the *result* sections. Similarly, the local conclusions extracted from the *conclusion* sections are better aligned with the global conclusions than the local conclusions from *method* and *result* sections, revealing a preference for local conclusions to be drawn in the *conclusion* section, as expected.

In addition, we studied correlations between the precisions of local premises and conclusions. We expected that when either the premise or conclusion of a local argument is well aligned with the global counterpart, then so will be the other component of the local argument. We therefore calculated the Pearson correlation coefficients between unigram precisions of local premises and of local conclusions in *method*, *result*, and *conclusion* sections. We find significant correlation coefficients in the range 0.3-0.4 (see Figure 4 in Appendix A), revealing a weak positive correlation between local premises and conclusions.

To depict the relation between local premises and local conclusions as a contour plot, we first meshed the unigram precisions in Figure 4 of appendix A into square cells of size 0.01x0.01. We then smoothed the unigram precisions using a 2D Gaussian kernel with $\sigma = 1$ and summed the values within each cell. Finally, we performed brute force computation to find the levels corresponding to the first one-third and the two-thirds of the summation of the mesh.

In Figure 3 we show the superimposed contours of the unigram precisions in *method*, *result*, and *conclusion* sections. We see that the 2/3 contour associated with *result* sections extends to larger premise precisions than the contours associated with other sections, in agreement with our finding that local premises located in *result* sections are best aligned with global premises.

7 Human Evaluation

Following the evaluation setups proposed by (Gu et al., 2022; Dong et al., 2018), we conducted a human evaluation on how well LACs extracted with the two proposed approaches reflect the global argument. The human evaluation is designed as a text comparison task where we asked the evaluators to choose between the LACs extracted by the two approaches in an interactive UI interface setting (see Figure 5 in appendix E), by carefully reading the text displayed on the interface.

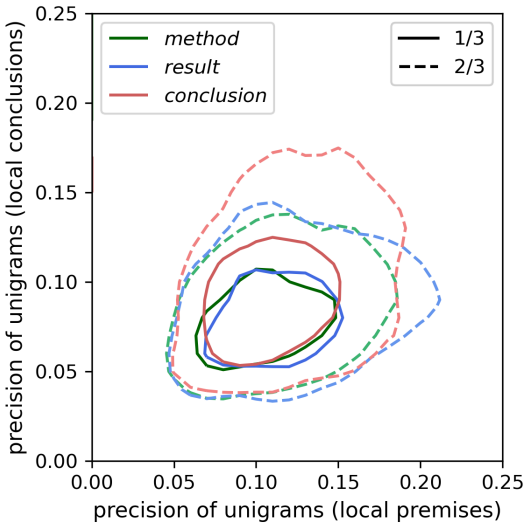


Figure 3: Superimposed contours of the unigram precisions of local premises and local conclusions in *method* (green), *result* (blue), and *conclusion* (red) sections. A solid contour delimits the first one-third of a given (summed) density and a dashed contour the first two-thirds. Best viewed under color printing.

We recruited 6 human evaluators with strong biology/neuroscience backgrounds. Each evaluator was asked to evaluate 25 randomly picked samples from our proposed scinf-biomed dataset. LACs extracted by discourse indicators and textual intersection were randomly displayed in separate text wrappers (Extractor A and Extractor B). In order to prevent the evaluators from inferring the LACs extraction method, we presented the LACs extracted with discourse indicators as complete sentences. To discount for LAC length (as in Figure 2), we truncated LACs to the first 100, 200, and 300 unigrams, respectively. The evaluators were asked to choose the better extractor (value of #1) for each of the following criteria:

- Coverage (Recall): how many different aspects/perspectives of the global argument are mentioned in the LACs;
- Non-redundancy (Precision): how precisely are those aspects/perspectives mentioned in the LACs;
- Overall: the better extractor based on subjective criteria including non-redundancy and coverage.

#unigram@100	Overall	Recall	Precision
indicator	**1.46**	1.54	**1.40**
intersection	1.54	**1.46**	1.60
#unigram@200	Overall	Recall	Precision
indicator	1.60	1.60	1.64
intersection	**1.40**	**1.40**	**1.36**$\star$
#unigram@300	Overall	Recall	Precision
indicator	1.52	1.54	**1.42**
intersection	**1.48**	**1.46**	1.58

Table 7: Average rank of two approaches in human evaluation. Smaller rank corresponds to better performance. $\star$ indicates statistical significance ($p < 0.05$, Wilcoxon signed-rank test).

Table 7 shows the results of the human evaluation. On the overall score, textual intersection achieves better performance on longer LACs (up to 200 and 300 words), whereas the discourse indicator approach ranks higher on shorter LACs (up to 100 words). On coverage, textual intersection is also better, but on non-redundancy results are more mixed. Overall, we see that textual intersection has a slight advantage but that discourse indicators can be useful for retrieving shorter argument components.

8 Conclusion

In this work, we investigate the effectiveness of discourse indicators for retrieving LACs relevant to the global argument of scientific papers. We develop a set of regular expressions for over 100 word- and phrase-level discourse indicators and test the performance of extracting the LACs of scientific papers. Our preliminary results show that discourse indicators have a limited capability of capturing LACs that are well-aligned with the global argument and thus cannot be solely used to extract arguments from scientific papers.

In future works, we will explore the effectiveness of discourse indicators in different types of scientific paper, such as research article, case report, and technical notes, etc. At the moment a notable weakness of our work is the oversimplifying use of regular expressions to disentangle premises from conclusions, thus we believe that the extraction of LACs using discourse indicators may be improved using more sophisticated (hierarchical) parsing techniques. In addition, we will work on a gold standard dataset that consists human annotated premise-conclusion pairs for argument generation, at the same time investigate the power of other more advanced contextualized sentence encoders.

Acknowledgements

We acknowledge the support from Swiss National Science Foundation NCCR Evolving Language, Agreement No.51NF40_180888. We also thank the anonymous reviewers for their constructive comments and feedback.

References

Khalid Al Khatib, Tirthankar Ghosal, Yufang Hou, Anita de Waard, and Dayne Freitag. 2021. Argument mining for scholarly document processing: Taking stock and looking ahead. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 56–65.

Floris Bex, John Lawrence, and Chris Reed. 2014. Generalising argument dialogue with the dialogue game execution platform. In *COMMA*, pages 141–152.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. BanditSum: Extractive summarization as a contextual bandit. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3739–3748, Brussels, Belgium. Association for Computational Linguistics.

Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. 2015. On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2242.

James B Freeman. 2000. What types of statements are there? *Argumentation*, 14(2):135–157.

Nianlong Gu, Elliott Ash, and Richard Hahnloser. 2022. MemSum: Extractive summarization of long documents using multi-step episodic Markov decision processes. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6507–6522, Dublin, Ireland. Association for Computational Linguistics.

Timon Gurcke, Milad Alshomary, and Henning Wachsmuth. 2021. Assessing the sufficiency of arguments through conclusion generation. In *Proceedings of the 8th Workshop on Argument Mining*, pages 67–77.

Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2015. Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 1–11.

John Lawrence and Chris Reed. 2015. Combining argument mining techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 127–136.

John Lawrence and Chris Reed. 2017. Mining argumentative structure from natural language text using automatically generated premise-conclusion topic models. In *EMNLP 2017: Conference on Empirical Methods in Natural Language Processing*, pages 39–48. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Alan Lee, Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2016. Annotating discourse relations with the pdtb annotator. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 121–125.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. 2020. S2orc: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230.

Nathan Ong, Diane Litman, and Alexandra Brusilovsky. 2014. Ontology-based argument mining and automatic essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 24–28.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 528–540.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394.

Pontus Plavén-Sigray, Granville James Matheson, Björn Christian Schiffler, and William Hedley Thompson. 2017. Research: The readability of scientific texts is decreasing over time. *eLife*, 6:e27725.

Henry Prakken. 2010. On the nature of argument schemes. *Dialectics, dialogue and argumentation. An examination of Douglas Walton's theories of reasoning and argument*, pages 167–185.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*. Citeseer.

Chris Reed and Glenn Rowe. 2004. Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 13(04):961–979.

Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument extraction from news. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 56–66.

Alexander Te-Wei Shieh, Yung-Sung Chuang, Shang-Yu Su, and Yun-Nung Chen. 2019. Towards understanding of medical randomized controlled trials by conclusion generation. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 108–117.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56.

Manfred Stede and Jodi Schneider. 2018. Argumentation mining. *Synthesis Lectures on Human Language Technologies*, 11(2):1–191.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1493–1502.

Tim Van Gelder. 2007. The rationale for rationale™. *Law, probability and risk*, 6(1-4):23–42.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*.

Nicole Wyatt. 2001. Ralph h. johnson, manifest rationality: A pragmatic theory of argument. *Philosophy in Review*, 21(3).

Adam Wyner, Jodi Schneider, Katie Atkinson, and Trevor Bench-Capon. 2012. Semi-automated argumentative analysis of online product reviews. In *Computational Models of Argument*, pages 43–50. IOS Press.

A Distribution of Unigram Precisions

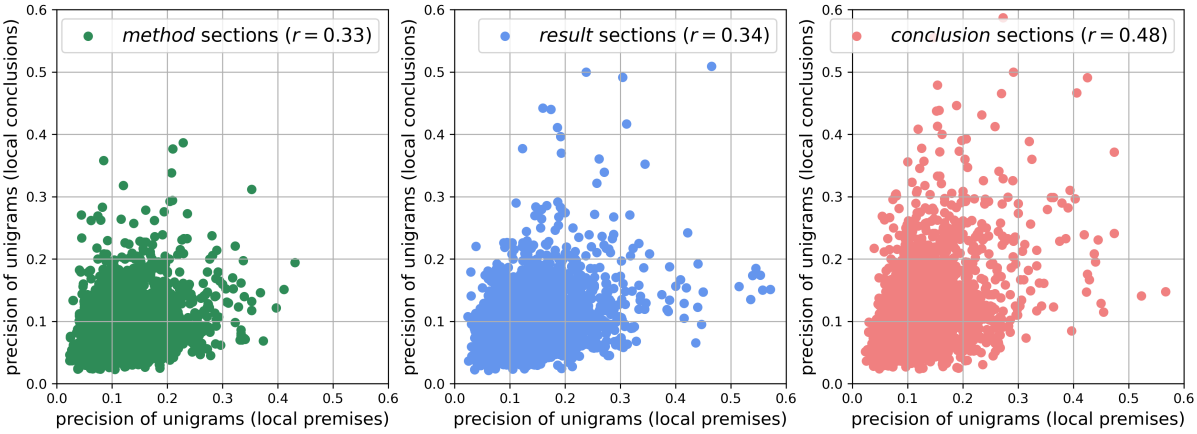


Figure 4: Distribution of unigram precisions of individual local conclusion and local premise occurring in the global conclusion and global premise. Each point in the figure represents a local premise (P) and a local conclusion (C) extracted by the same discourse indicator. r delimits the Pearson correlation coefficient of comparing unigram precisions for P to C. For all 3 type of sections, $p < 10^{-3}$ is observed.

B Sections

To detect *method*, *result*, and *conclusion* sections, we use the following anchors (critical strings for candidate section names) in Table 8. For instance, a section is considered to be a *method* section when its section name contains at least one of these section anchors.

Notice that to ensure no risk of having concluding text treated as local premises, all sections must be exclusive of the string *discussion*.

section	section anchors
method	method, procedure, data, theory, implementation
result	result, outcome, analysis, measure, evaluation
conclusion	conclusion, concluding, summary, remark, key point

Table 8: Critical strings for selecting related sections used in Table 6

C Discourse Indicators

(a) Discourse indicators part A

P. In view of that, C.	P. One can deduce that C.
P. One can infer that C.	P. One can conclude that C.
C. Its proof is that P.	P. As a result, C.
P, resulting in C.	P, in that case C.
C. This comes from P.	P. For this reason, C.
P. In consequence, C.	P. As conclusion, C.
P suggested that C.	P can cause C.

(b) Discourse indicators part B

C, since P.	Granted that P, C.
P, therefore C.	Supposing that P, C.
P. Therefore, C.	C, supposing that P.
P, wherefore C.	Assuming that P, C.
P, so that C.	C, assuming that P.
P, consequently C.	Because P, C.
P, entails that C.	C, because P.
As shown from P, C.	Here is why C: P.
C, if P.	P implies that C.
P, shows that C.	As indicated by P, C.
C, follows from P.	C, as indicated by P.
C, giving that P.	P, indicating that C.
Due to the reason that P, C.	On account of the reason that P, C.
C, due to the reason that P.	C, on account of the reason that P.
In view of the fact that P, C.	C may be deduced from P.
C, in view of the fact that P.	C may be inferred from P.
P, thereby showing that C.	C may be derived from P.
P, thus C.	C can be derived from P.
P establishes that C.	P proves that C.
P justifies that C.	C is supported by P.
In support of C, P.	P, which leads credence to C.
Inasmuch as P, C.	On the hypothesis that P, C.
P demonstrates that C.	C, on the hypothesis that P.
P indicates that C.	P signifies that C.
P, indicating that C.	P guarantees that C.
C is based on P.	On the basis of P, C.
In light of the fact that P, C.	C, on the basis of P.
P. In fact, C.	Convinced by the fact that P, C.
In fact that P, C.	Seeing that P, C.
C, for the reason that P.	C, seeing that P.
P, from which it follows C.	Owing to P, C.
Due to P, C.	C, owing to P.
C, due to P.	C, on the grounds that P.
C, considering P.	On the grounds that P, C.
P, which leads to C.	On account of the fact P, C.
P, which shows that C.	C, on account of the fact P.
P, which allows us to infer C.	P, means that C.
P, which implies C.	P, which points to C.
C. The reason is that P.	P. Accordingly, C.
P. From this we can deduce that C.	P. From this it follows that C.
P. This proves that C.	P. Hence, C.
P. Obviously, C.	P. Evidently, C.
P. In conclusion, C.	P. On this account, C.
C. This is shown by P.	P. This is being so C.
P. Indeed, C	*C, insofar as P.*
P. In short, C.	*P. In sum, C.*
P, in other words, C.	*Now that P, C.*

Table 9: Discourse indicators used in this work.

Table 9 lists all word- and phrase-level discourse indicators used in our work for LAC extraction. For each discourse indicator, P denotes the local premise and C the local conclusion. Based on linguistic facts and experience, the assessment was guided by five qualified scholars. Discourse indicators adapted exclusively from the Penn Discourse Treebank 3.0 (Webber et al., 2019) are marked in italic font.

Table 10 presents statistics of these discourse indicators ranked by: a) averaged length of extracted LACs; b) and c) average ROUGE-N scores. For local premises (P) and local conclusions (C) extracted by each discourse indicator, the averaged ROUGE-N scores are computed against the corresponding global premises and global conclusions, respectively.

(a) Top 5 discourse indicators that have at least 100 appearances (#hits), ranked by the average length of LACs (as number of words). P as local premises and C as local conclusions.

indicator	avg. length of P	#hits total	indicator	avg. length of C	#hits total
indicating that	29.30	741	in short	28.00	161
for these reasons	28.75	297	assuming that	27.63	105
so that	28.63	602	indeed	27.53	2,120
indeed	28.48	2,120	in conclusion	27.35	1,612
as a consequence	27.93	398	in fact	25.87	1,846

(b) Top 10 discourse indicators for P (local premises), ranked by the average Rouge-N score metrics.

indicator	pr	indicator	rc	indicator	fm
wherefore	39.86	which proves that	18.14	which proves that	19.91
in that case	35.30	which can be derived from	9.32	which can be derived from	12.81
one may infer that	34.81	means that	8.72	means that	12.38
in light of the fact that	31.56	in view of that	8.70	in that case	12.19
as indicated by*	29.91	which shows that	8.21	in view of that	11.91
indicating that*	29.75	indicating that*	8.12	indicating that*	11.20
this is shown by	28.20	in that case	7.37	this is shown by	10.45
may be inferred from	27.91	this is shown by	7.33	which proves that	10.19
which proves that	27.78	from this we can deduce that	7.16	wherefore	10.07
inasmuch as	27.76	consequently*	6.87	this proves that	10.06

(c) Top 10 discourse indicators for C (local conclusions), ranked by the average Rouge-N score metrics

indicator	pr	indicator	rc	indicator	fm
on this account	44.41	in conclusion*	26.35	in conclusion*	30.62
in view of that	43.57	one can conclude that	25.16	one can conclude that	28.78
in conclusion*	42.87	on this account	20.47	on this account	28.02
which proves that	36.31	in light of the fact that	18.79	in view of that	21.30
one can conclude that	33.62	demonstrates that*	15.97	demonstrates that*	19.95
demonstrates that*	33.02	in view of that	14.41	this is shown by	17.10
might be derived from	30.09	this is shown by	13.36	might be inferred from	15.84
wherefore	28.42	proves that	11.70	in sum	15.81
granted that	28.36	might be inferred from	10.97	wherefore	15.34
this is shown by	27.45	justifies that	10.73	which proves that	14.77

Table 10: Discourse indicators ranked by the Rouge-N scores: (a) top 5 discourse indicators that extract the longest LACs (length counted as the number of words) (b) top 10 discourse indicators in which local premises (P) have the highest Rouge-N scores to the global premises (c) top 10 discourse indicators which local conclusions (C) have the highest Rouge-N scores to the global conclusions. pr, rc, and fm stand for precision, recall, and f-measure, respectively. * in (b) and (c) denotes discourse indicators that have more than 100 appearances (# > 100).

D Dataset Example

LACs extracted using discourse indicators
The SRT estimated using the CPhT test was significantly higher (worse) for NAL-NL1 than for DSL [i/o] or DSL V, **indicating that** the NAL-NL1 prescription is less effective than the DSL prescriptions in making low level sounds intelligible.
High compression ratios, combined with high amounts of low-frequency gain, may also increase the audibility of background noise, and this may degrade speech understanding in noise via the upward spread of masking. *Thus*, as compression ratios are increased, the potential benefits of increased audibility of speech may be offset by a variety of deleterious effects.
The lower gains **may help to** preserve the relative levels of the first and second formants, which may lead to improved vowel identification.
It is not feasible to restore the audibility of low-level sounds completely to normal for hearing-impaired children or adults, **due to** factors such as the internal noise of hearing aids (especially microphone noise), limitations in the gain that can be achieved without acoustic feedback, and the need to avoid excessive amounts of compression.
A problem with the use of questionnaires is that the outcomes may be influenced by the personality and attitude of the adult or child performing the evaluation. **Hence**, questionnaires may be useful for comparing results across groups, but are not so effective in evaluating the performance of individual children.
avg. ROUGE-N f-measures: 16.05 for local conclusions C, 26.89 for local premises P.
LACs extracted using textual intersection
A few children with moderate hearing loss scored close to ceiling for the-dB SPL stimuli. ANOVAs were conducted separately on the RAU-transformed scores for the presentation levels of and dBA with prescription as a within-subjects factor and severity of hearing loss as a between subjects factor. CAWL scores were derived from the number of phonemes correct for each of the target words. Figure shows the average levels in dBA required for correct identification of each of the Ling sounds, across all subjects, for each hearing aid prescription. For the level of dBA, there was no significant effect of prescription , but there was an effect of severity of hearing loss . . .
The higher output levels prescribed by the DSL i/o and DSL V prescription methods relative to NAL-NL1 led to significantly better detection and discrimination of lowlevel sounds. Using age-appropriate closed-set and open-set speech tests, designed to avoid floor and ceiling effects, we found significant differences between scores for the different hearing aid prescription methods.
avg. ROUGE-N f-measures: 44.10 for local conclusions C, 31.90 for local premises P.
Global premises
Scores for the Consonant Confusion Test and CAPT consonant discrimination and consonant detection were lower for the NAL-NL1 prescription than for the DSL prescriptions. Scores for the CAPT vowel-in-noise discrimination test were higher for DSL V than for either of the other prescriptions. Scores for the Cambridge Auditory Word Lists did not differ across prescriptions for the level of 65 dBA, but were lower for the NAL-NL1 prescription than for either of the DSL prescriptions for the level of 50 dBA. The speech reception threshold measured using the Common Phrases Test and the levels required for identification of the Ling 5 sounds were higher (worse) for the NAL-NL1 prescription than for the DSL prescriptions.
Global conclusions
The higher gains prescribed by the DSL i/o and DSL V prescription methods relative to NAL-NL1 led to significantly better detection and discrimination of low-level speech sounds.

Table 11: An example biomedial paper in our proposed dataset *scinf-biomed*.

(a) Strong (*which proves that*) and weak (*indicating that*) discourse indicators for local-to-global premise comparison

Local Premise (P)	The circadian curves of cortisol secretion compared the day after the end of magnetotherapy and M3P3 magnetostimulation significantly differ from the M2P2 program -nearly by 100%, **which proves that** this type of magnetotherapy and magnetostimulation shows varied influence on cortisol secretion in men.
Global Premise	. . . Statistically significant difference was demonstrated in the participants after the application of magnetotherapy and magnetostimulation with M3P3 program compared to the men submitted to magnetostimulation, with M2P2 program, at 400 p.m. after 15 applications.
Local Premise (P)	Within the families of bipolar probands there is a higher than average rate of unipolar depressive disorders, **indicating that** bipolar susceptibility genes can be expressed in a broad spectrum of mood phenotypes.
Global Premise	. . . Systematic study of the coding and flanking intronic regions of 25 known genes within this latter region failed to identify any highly penetrant autosomal dominant disease-conferring mutations in these pedigrees.

(b) Strong (*one can conclude that*) and weak (*in sum*) discourse indicators for local-to-global conclusion comparison

Local Conclusion (C)	. . . **One can conclude that** RGCs express RS both developmentally and in the adult retina, indicating that local replenishment of RS protein evidently is desirable for maintaining retinal structure, even after retinal development is completed.
Global Conclusion	All major classes of adult retinal neurons . . . strongly suggesting that retinoschisin in the inner retina is synthesized locally rather than being transported, as earlier proposed, from distal retinal photoreceptors . . .
Local Conclusion (C)	Observations were repeated with the same biological replicate for each tissue. **In sum** this is a factorial arrangement of treatments (Diet by Genotype) laid out on a balanced Completely Randomized Design (CRD) with repeated measures on another treatment (Source of Tissue) amounting to a total of 2n = 40 observations.
Global Conclusion	These studies show that high-throughput metabolomics combined with appropriate statistical modeling and large scale functional approaches can be used to monitor and infer changes and interactions in the metabolome and genome of the host under controlled experimental conditions . . . Based on our results, metabolic signatures and metabolic pathways of polyposis and intestinal carcinoma have been identified, which may serve as useful targets for the development of therapeutic interventions.

Table 12: Alignment of LACs extracted by strong and weak discourse indicators to the global argument.

E User Interface for Human Evaluation

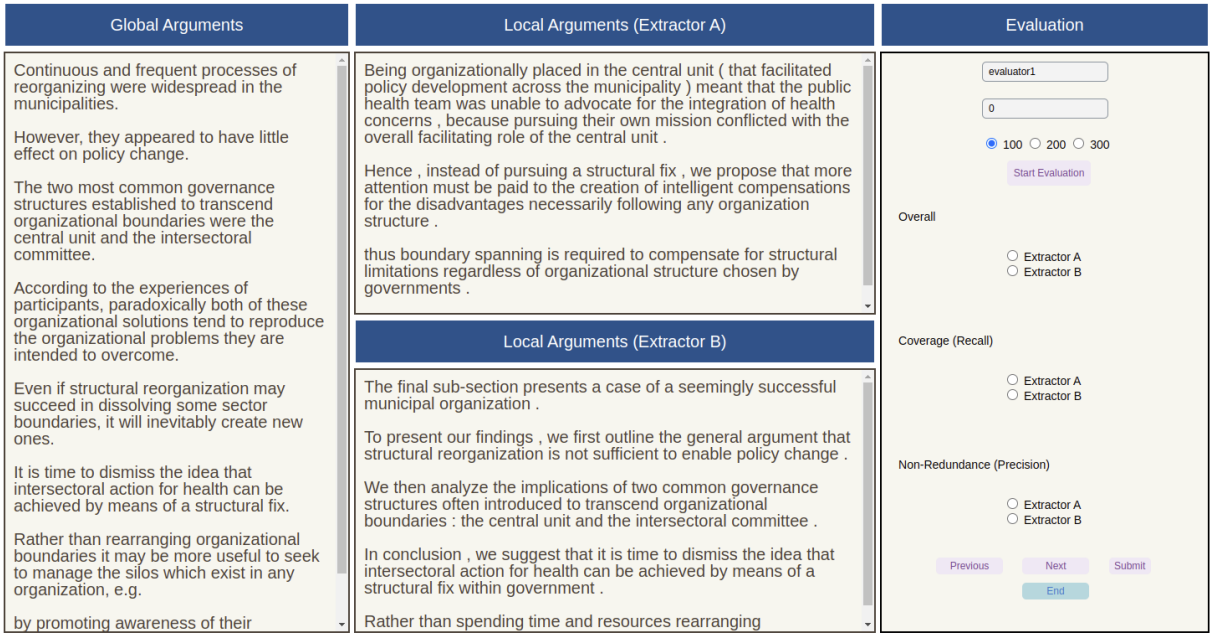


Figure 5: The user interface designed for the human evaluation. The annotators are asked to mark the anonymous extractor which they think is better in terms of overall quality, information coverage, and non-redundancy.