

Finding Pragmatic Differences Between Disciplines

Lee Kezar

University of Southern California
Information Sciences Institute
lkezar@isi.edu

Jay Pujara

University of Southern California
Information Sciences Institute
jpujara@isi.edu

Abstract

Scholarly documents have a great degree of variation, both in terms of content (semantics) and structure (pragmatics). Prior work in scholarly document understanding emphasizes semantics through document summarization and corpus topic modeling but tends to omit pragmatics such as document organization and flow. Using a corpus of scholarly documents across 19 disciplines and state-of-the-art language modeling techniques, we learn a fixed set of domain-agnostic descriptors for document sections and “retrofit” the corpus to these descriptors (also referred to as “normalization”). Then, we analyze the position and ordering of these descriptors across documents to understand the relationship between discipline and structure. We report within-discipline structural archetypes, variability, and between-discipline comparisons, supporting the hypothesis that scholarly communities, despite their size, diversity, and breadth, share similar avenues for expressing their work. Our findings lay the foundation for future work in assessing research quality, domain style transfer, and further pragmatic analysis.

1 Introduction

Disciplines such as art, physics, and political science contain a wide array of ideas, from specific hypotheses to wide-reaching theories. In scholarly research, authors are faced with the challenge of clearly articulating a set of those ideas and relating them to each other, with the ultimate goal of expanding our collective knowledge. In order to understand this work, human readers situate meaning in context (Justin Garten and Deghani, 2019). Similarly, methods for scholarly document processing (SDP) have semantic and pragmatic orientations.

The semantic orientation seeks to understand and evaluate the ideas themselves through information extraction (Singh et al., 2016), summariza-

tion (Chandrasekaran et al., 2020), automatic fact-checking (Sathe et al., 2020), etc. The pragmatic orientation, on the other hand, seeks to understand the context around those ideas through rhetorical and style analysis (August et al., 2020), corpus topic modeling (Paul and Girju, 2009), quality prediction (Maillette de Buy Wenniger et al., 2020), etc. Although both orientations are essential for understanding, the pragmatics of disciplinary writing are very weakly understood.

In this paper, we investigate the structures of disciplinary writing. We claim that a “structural archetype” (defined in Section 3) can succinctly capture how a community of authors choose to organize their ideas for maximum comprehension and persuasion. Analogous to how syntactic analysis deepens our understanding of a given sentence and document structure analysis deepens our understanding of a given document, structural archetypes, we argue, deepen our understanding of domains themselves.

In order to perform this analysis, we classify sections according to their pragmatic intent. We contribute a data-driven method for deriving the types of pragmatic intent, called a “structural vocabulary”, alongside a robust method for this classification. Then, we apply these methods to 19k scholarly documents and analyze the resulting structures.

2 Related Work

We draw from two areas of related work in SDP: interdisciplinary analysis and rhetorical structure prediction.

In interdisciplinary analysis, we are interested in comparing different disciplines, whether by topic modeling between select corpora/disciplines (Paul and Girju, 2009) or by domain-agnostic language modeling (Wang et al., 2020). These comparisons are more than simply interesting; they allow for models that can adapt to different disciplines, helping the generalizability for downstream tasks like

information extraction and summarization.

In rhetorical structure prediction, we are interested in the process of implicature, whether by describing textual patterns in an unsupervised way (Ó Séaghdha and Teufel, 2014) or by classifying text as having a particular strategy like “statistics” (Al-Khatib et al., 2017) or “analogy” (August et al., 2020). These works descend from argumentative zoning (Lawrence and Reed, 2020) and the closely related rhetorical structure theory (Mann and Thompson, 1988), which argue that many rhetorical strategies can be described in terms of *units* and their relations. These works are motivated by downstream applications such as predicting the popularity of a topic (Prabhakaran et al., 2016) and classifying the quality of a paper (Maillette de Buy Wenniger et al., 2020).

Most similar to our work is Arnold et al. (2019). Here, the authors provide a method of describing Wikipedia articles as a series of section-like topics (e.g. `disease.symptom`) by clustering section headings into topics and then labeling words and sentences with these topics. We build on this work by using domain-agnostic descriptors instead of domain-specific ones and by comparing structures across disciplines.

3 Methods

In this section, we define structural archetypes (3.1) and methods for classifying pragmatic intent through a structural vocabulary (3.2).

3.1 Structural Archetypes

We coin the term “structural archetype” to focus and operationalize our pragmatic analysis. Here, a “structure” is defined as *a sequence of domain-agnostic indicators of pragmatic intent*, while an “archetype” refers to *a strong pattern across documents*. In the following paragraphs, we discuss the components of this concept in depth.

Pragmatic Intent In contrast to verifiable propositions, “indicators of pragmatic intent” refer to instances of meta-discourse, comments on the document itself (Ifantidou, 2005). There are many examples, including background (comments on what the reader needs in order to understand the content), discussions (comments on how results should be interpreted), and summaries (comments on what is important). These indicators of pragmatic intent serve the critical role of helping readers “digest”

material; without them, scholarly documents would only contain isolated facts.

We note that the boundary between pragmatic intent and argumentative zones (Lawrence and Reed, 2020) is not clear. Some argumentative zones are more suitable for the sentence- and paragraph-level (e.g. “own claim” vs. “background claim”) while others are interpretative (e.g. “challenge”). This work does not attempt to draw this boundary, and the reader might find overlap between argumentative zoning work and our section types.

Sequences As a sequence, these indicators reflect how the author believes their ideas should best be received in order to remain coherent. For example, many *background* indicators reflect a belief that the framing of the work is very important.

Domain-agnostic archetypes Finally, the specification that indicators must be domain-agnostic and that the structures should be widely-held are included to allow for cross-disciplinary comparisons.

We found that the most straightforward way to implement structural archetypes is through classifying section headings according to their pragmatic intent. With this comes a few challenges: (1) defining a set of domain-agnostic indicators, which we refer to as a “structural vocabulary”; (2) parsing a document to obtain its structure; and (3) finding archetypes from document-level structures. In the proceeding section, we address (1) and (2), and in Section 4 we address (3).

3.2 Deriving a Structural Vocabulary

Although indicators of pragmatic intent can exist on the sentence level, we follow Arnold et al. (2019) and create a small set of types that are loosely related to common section headings (e.g. “Methods”). We call this set a “structural vocabulary” because it functions in an analogous way to a vocabulary of words; any document can be described as a sequence of items that are taken from this vocabulary. There are three properties that the types should satisfy:

- A. **domain independence:** types should be used by different disciplines
- B. **high coverage:** unlabeled instances should be able to be classified as a particular type.
- C. **internal consistency:** types should accurately reflect their instances

Domain Independence As pointed out by Arnold et al. (2019), there exists a “vocabulary mismatch problem” where different disciplines talk about their work in different ways. Indeed, 62% of the sampled headings only appear once and are not good choices for section types. On the other hand, the most frequent headings are a much better choice, especially those that appear in all domains. After merging a few popular variations among the top 20 section headings (e.g. *conclusion* and *summary*, *background* and *related work*), we yield the following types¹: *introduction* (a section which introduces the reader to potentially new concepts; $n = 10916$), *methods* (a section which details how a hypothesis will be tested; $n = 2116$), *results* (a section which presents findings of the method; $n = 3119$), *discussion* (a section which interprets and summarizes the results; $n = 3118$), *conclusion* (a section which summarizes the entire paper; $n = 7738$), *analysis* (a section which adds additional depth and nuance to the results; $n = 951$), and *background* (a section which connects ongoing work to previous related work; $n = 800$). Figure 2 contains discipline-level counts.

High Coverage We can achieve high coverage by classifying any section as one of these section types through language modeling. Specifically, the hidden representation of a neural language model $h(\cdot)$ can act as an embedding of its input. We use the [CLS] tag of SciBERT’s hidden layer, selected for its robust representations of scientific literature (Beltagy et al., 2019).

To classify, we define a distance score $d(\cdot)$ for a section s and a type T as the distance between $h(s)$ and the average embedding across all instances of a type, i.e.

$$d(s, T) = \left| h(s) - \frac{\sum_{t \in T} h(t)}{\|T\|} \right|$$

Note that since the embedding is a vector, addition and division are elementwise. Then, we compute the distance for all types in the vocabulary V and select the minimum, i.e.

$$s_{type} = \arg \min_{T \in V} (d(s, T))$$

Internal Consistency Some sections do not adequately fit any section type, so nearest-neighbor

¹Although *abstract* is extremely common we found it redundant as a section type as it only exists once per paper and in a predictable location.

classification will result in very inconsistent clusters. We address this problem by imposing a threshold on the maximum distance for $d(\cdot)$. Further, since the types have unequal variance (that is, the ground truth for some types are more consistent than other types), we define a type-specific threshold as half of the distance from the center of T to the furthest member of T , i.e.

$$\text{threshold}_T = 0.5 \cdot \max_{t \in T} (d(t, T))$$

The weight of 0.5 was found to remove outliers appropriately and maximize retrofitting performance (Section 4.2).

We also note that some headings, especially brief ones, leave much room for interpretation and make retrofitting challenging. We address this problem by concatenating tokens of each section’s heading and body, up to 25 tokens, as input to the language model. This ensures that brief headings contain enough information to make an accurate representation without including too many details from the body text.

4 Results and Discussion

4.1 Data

We use the Semantic Scholar Open Research Corpus (S2ORC) for all analysis (Lo et al., 2020). This corpus, which is freely available, contains approximately 7.5M PDF-parsed documents from 19 disciplines, including natural sciences, social sciences, arts, and humanities. For our experiments, we randomly sample 1k documents for each discipline, yielding a total of 19k documents.

4.2 Retrofitting Performance

Retrofitting (or normalizing) section headers refers to re-labeling sections with the structural vocabulary. We evaluate retrofitting performance by manually tagging 30 of each section type and comparing the true labels to the predicted values. Our method yields an average F1 performance of 0.76. The breakdown per section type, shown in Table 1, reveals that *conclusion*, *background*, and *analysis* sections were the most difficult to predict. We attribute this to a lack of textual clues in the heading and body, and also a semantic overlap with *introduction* sections. Future work can improve the classifier with more nuanced signals, such as position, length, number of references, etc.

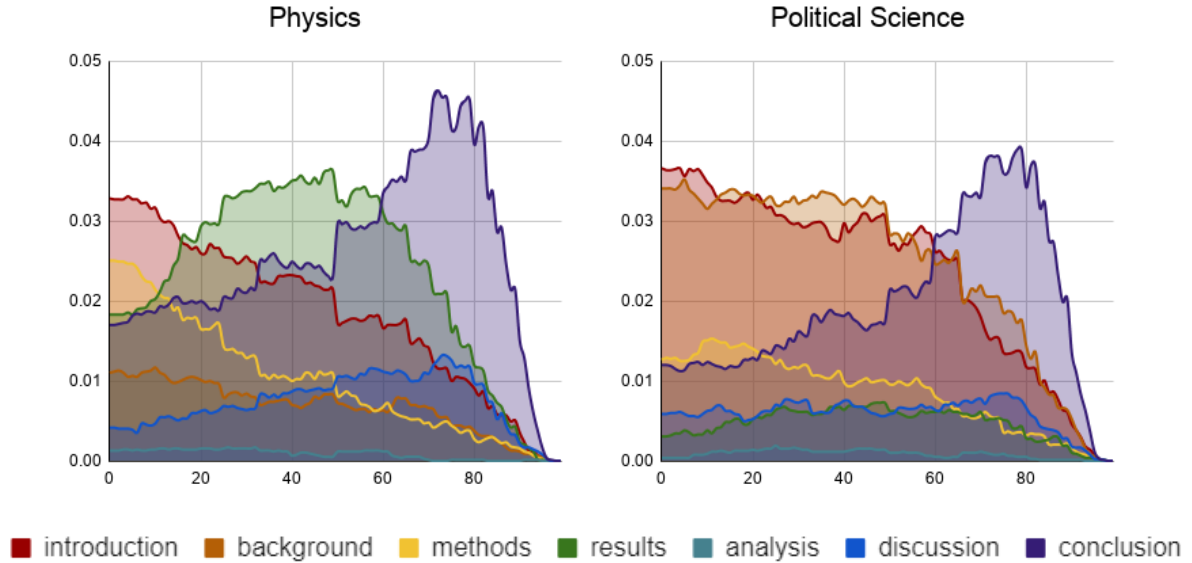


Figure 1: A comparison between the positions (normalized by document length; x axis) and frequencies (y axis) of section types in Physics and Political Science. Comparable distributions of *introduction*, *methods*, *analysis*, *discussion*, and *conclusion*, but different distributions of *background* and *results*.

Type	Precision	Recall	F1
introduction	0.77	0.97	0.85
conclusion	0.67	0.72	0.69
discussion	0.88	0.88	0.88
results	0.80	0.85	0.83
methods	0.83	0.91	0.87
background	0.63	0.77	0.69
analysis	0.50	0.61	0.55
overall	0.72	0.88	0.76

Table 1: Type-level and overall performance for section type retrofitting.

4.3 Analyzing Position with Aggregate Frequency

A simple yet expressive way of showing the structural archetypes of a discipline is to consider the frequency of a particular type at any point in the article (normalized by length). This analysis reveals general trends throughout a discipline’s documents, such as where a section type is most frequent or where there is homogeneity.

To illustrate the practicality of this analysis, consider the hypothesis that Physics articles are more empirically-motivated while Political Science articles are more conceptually-motivated, i.e. that they are on opposing ends of the *concrete* versus *abstract* spectrum. We operationalize this by claiming that Physics articles have more *methods*, *results*,

and *analysis* sections than Political Science. Figure 1 shows the difference between Physics and Political Science at each point in the article. It reveals that not only do Physics articles contain more *methods* and *results*, but also that Physics articles introduce *methods* earlier than Political Science, and that both contain the same amount of *analysis* sections.

4.4 Analyzing Ordering with State Transitions

A more structural analysis of a discipline is to look at the frequency of sequence fragments through computing transition probabilities. As a second example, suppose we have a more nuanced hypothesis: that Psychology papers tend to separate claims and evaluate them sequentially (*methods*, *results*, *discussion*, repeat) whereas Sociology papers tend to evaluate all claims at once. We can operationalize these hypotheses by calculating the transition probability between section s_i and s_{i-1} conditioned on some discipline.

In Table 2, we see evidence that *methods* sections are more likely to be preceded by *results* sections in Psychology than Sociology, implying a new iteration of a cycle. We might conclude that Psychology papers are more likely to have cyclical experiments, but not that Sociology papers conduct multiple experiments in a linear fashion.

Transition Probability	Psych.	Socio.
$P(\text{method} \rightarrow \text{method})$	0.31	0.20
$P(\text{results} \rightarrow \text{results})$	0.22	0.23
$P(\text{disc} \rightarrow \text{disc})$	0.16	0.13
$P(\text{method} \rightarrow \text{results})$	0.21	0.10
$P(\text{results} \rightarrow \text{disc})$	0.15	0.16
$P(\text{disc} \rightarrow \text{method})$	0.23	0.13

Table 2: Transition probabilities for methods, results, and discussion in Psychology and Sociology

5 Conclusion and Future Work

In this paper, we have shown a simple method for constructing and comparing structural archetypes across different disciplines. By classifying the pragmatic intent of section headings, we can visualize structural trends across disciplines. In addition to utilizing a more complex classifier, future directions for this work include (1) further distinguishing between subdisciplines (e.g. abnormal psychology vs. developmental psychology) and document type (e.g. technical report vs. article); (2) learning relationships between structures and measures of research quality, such as reproducibility; (3) learning how to convert one structure into another, with the ultimate goal of normalizing them for easier comprehension or better models; (4) deeper investigations into the selection of a structural vocabulary, such as including common argumentative zoning types or adjusting the scale to the sentence-level; and (5) drawing comparisons, such as by clustering, between different documents based strictly on their structure.

6 Acknowledgements

This work was funded by the Defense Advanced Research Projects Agency with award W911NF-19-20271. The authors would like to thank the reviewers of this paper for their detailed and constructive feedback, and in particular their ideas for future directions.

References

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. 2017. [Patterns of argumentation strategies across topics](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1351–1357, Copenhagen, Denmark. Association for Computational Linguistics.

Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. 2019. [SECTOR: A neural model for coherent topic segmentation and classification](#). *Transactions of the Association for Computational Linguistics*, 7:169–184.

Tal August, Lauren Kim, Katharina Reinecke, and Noah A. Smith. 2020. [Writing strategies for science communication: Data and computational analysis](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5327–5344, Online. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. [SciBERT: A pretrained language model for scientific text](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Muthu Kumar Chandrasekaran, Guy Feigenblat, Eduard Hovy, Abhilasha Ravichander, Michal Shmueli-Scheuer, and Anita de Waard. 2020. [Overview and insights from the shared tasks at scholarly document processing 2020: CL-SciSumm, LaySumm and LongSumm](#). In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 214–224, Online. Association for Computational Linguistics.

Elly Ifantidou. 2005. [The semantics and pragmatics of metadiscourse](#). *Journal of Pragmatics*, 37(9):1325–1353. Focus-on Issue: Discourse and Metadiscourse.

Kenji Sagae Justin Garten, Brendan Kennedy and Morteza Deghani. 2019. [Measuring the importance of context when modeling language comprehension](#). *Behavioral Research Methods*, 51:480–492.

John Lawrence and Chris Reed. 2020. [Argument Mining: A Survey](#). *Computational Linguistics*, 45(4):765–818.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel 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, Online. Association for Computational Linguistics.

Gideon Maillette de Buy Wenniger, Thomas van Dongen, Eleri Aedmaa, Herbert Teun Kruitbosch, Edwin A. Valentijn, and Lambert Schomaker. 2020. [Structure-tags improve text classification for scholarly document quality prediction](#). In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 158–167, Online. Association for Computational Linguistics.

William Mann and Sandra Thompson. 1988. [Rhetorical structure theory: Toward a functional theory of](#)

- text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8:243–281.
- Diarmuid Ó Séaghdha and Simone Teufel. 2014. [Un-supervised learning of rhetorical structure with un-topic models](#). In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2–13, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.
- Michael Paul and Roxana Girju. 2009. [Topic modeling of research fields: An interdisciplinary perspective](#). In *Proceedings of the International Conference RANLP-2009*, pages 337–342, Borovets, Bulgaria. Association for Computational Linguistics.
- Vinodkumar Prabhakaran, William L. Hamilton, Dan McFarland, and Dan Jurafsky. 2016. [Predicting the rise and fall of scientific topics from trends in their rhetorical framing](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1170–1180, Berlin, Germany. Association for Computational Linguistics.
- Aalok Sathe, Salar Ather, Tuan Manh Le, Nathan Perry, and Joonsuk Park. 2020. [Automated fact-checking of claims from Wikipedia](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6874–6882, Marseille, France. European Language Resources Association.
- Mayank Singh, Barnopriyo Barua, Priyank Palod, Manvi Garg, Sidhartha Satapathy, Samuel Bushi, Kumar Ayush, Krishna Sai Rohith, Tulasi Gamidi, Pawan Goyal, and Animesh Mukherjee. 2016. [OCR++: A robust framework for information extraction from scholarly articles](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3390–3400, Osaka, Japan. The COLING 2016 Organizing Committee.
- Chengyu Wang, Minghui Qiu, Jun Huang, and Xiaofeng He. 2020. [Meta fine-tuning neural language models for multi-domain text mining](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3094–3104, Online. Association for Computational Linguistics.

A Section Counts Before and After Retrofitting

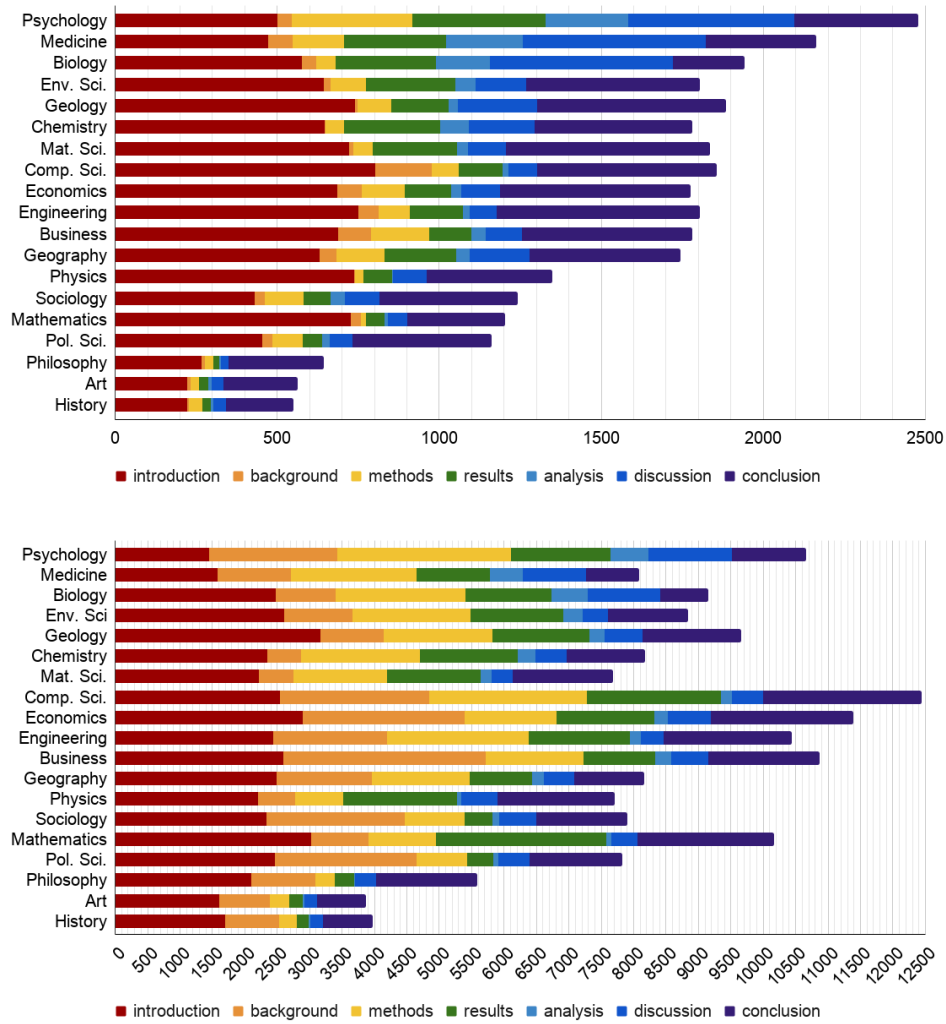


Figure 2: The frequency of the top-7 section headings before (top) and after (bottom) retrofitting.

B Aggregate Frequency for Other Disciplines



Figure 3: Aggregate Frequency for 12 of the 19 disciplines