Predicting the Rise and Fall of Scientific Topics from Trends in their Rhetorical Framing

Vinodkumar Prabhakaran
Stanford University
vinod@cs.stanford.edu

William L. Hamilton
Stanford University
wleif@stanford.edu

Dan McFarland
Stanford University
dmcfarla@stanford.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

Abstract

Computationally modeling the evolution of science by tracking how scientific topics rise and fall over time has important implications for research funding and public policy. However, little is known about the mechanisms underlying topic growth and decline. We investigate the role of rhetorical framing: whether the rhetorical role or function that authors ascribe to topics (as methods, as goals, as results, etc.) relates to the historical trajectory of the topics. We train topic models and a rhetorical function classifier to map topic models onto their rhetorical roles in 2.4 million abstracts from the Web of Science from 1991-2010. We find that a topic's rhetorical function is highly predictive of its eventual growth or decline. For example, topics that are rhetorically described as results tend to be in decline, while topics that function as methods tend to be in early phases of growth.

1 Introduction

One of the most compelling research questions in the computational analysis of scientific literature is whether the vast collections of scientific text hold important clues about the dynamics involved in the evolution of science; clues that may help predict the rise and fall of scientific ideas, methods and even fields. Being able to predict scientific trends in advance could potentially revolutionize the way science is done, for instance, by enabling funding agencies to optimize allocation of resources towards promising research areas.

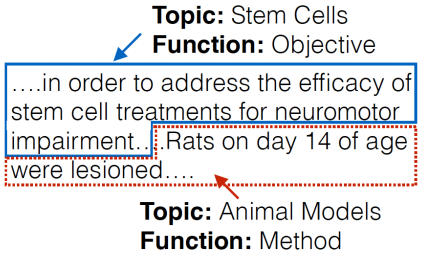


Figure 1: Example abstract snippet. The abstract rhetorically frames the *stem cells* topic as the OBJECTIVE of the research, while the *animal models* topic functions as the research METHOD.

Prior studies have often tracked scientific trends by applying topic modeling (Blei et al., 2003) based techniques to large corpora of scientific texts (Griffiths and Steyvers, 2004; Blei and Lafferty, 2006; Hall et al., 2008). They capture scientific ideas, methods, and fields in terms of *topics*, modeled as distributions over collection of words. These approaches usually adopt a decontextualized view of text and its usage, associating topics to documents based solely on word occurrences, disregarding where or how the words were employed. In reality, however, scientific abstracts often follow narrative structures (Crookes, 1986; Latour, 1987) that signal the specific *rhetorical roles* that different topics play within the research (Figure 1). The rhetorical role of a topic is the purpose or role it plays in the paper: as its background (scientific context), its objective/goal, the data employed, the design or method used (mode of inference), the results (what is found) or the conclusions (what they mean).

RATIONALE: Neonatal ibotenic acid lesion of the ventral hippocampus was proposed as a relevant animal model of schizophrenia reflecting positive as well as negative symptoms of this disease. Before and after reaching maturity, specific alterations in the animals' social behaviour were found. **OBJECTIVE:** In this study, social behaviour of ventral hippocampal lesioned rats was analysed. For comparison, rats lesioned either in the ventral hippocampus or the dorsal hippocampus at the age of 8 weeks were tested. **METHODS:** Rats on day 7 of age were lesioned with ibotenic acid in the ventral hippocampus and social behaviour was tested at the age of 13 weeks. For comparison, adult 8-week-old rats were lesioned either in the ventral or the dorsal hippocampus. Their social behaviour was tested at the age of 18 weeks. **RESULTS:** It was found that neonatal lesion resulted in significantly decreased time spent in social interaction and an enhanced level of aggressive behaviour. This shift is not due to anxiety because we could not find differences between control rats and lesioned rats in the elevated plus-maze. Lesion in the ventral and dorsal hippocampus, respectively, in 8-week-old rats did not affect social behaviour. **CONCLUSIONS:** The results of our study indicate that ibotenic acid-induced hippocampal damage per se is not related to the shift in social behaviour. We favour the hypothesis that these changes are due to lesion-induced impairments in neurodevelopmental processes at an early stage of ontogenesis.

Figure 2: An example of a self-annotated abstract.
Source: `http://www.ncbi.nlm.nih.gov/pubmed/10435405`.

Rhetorical functions that topics take part in could hold important clues about the stage or development of an intellectual movement they stand to represent. For example, a topic that shifts over time from being employed as a method to being mentioned as background may signal an increase in its maturity and perhaps a corresponding decrease in its popularity among new research.

In this paper, we introduce a new algorithm to determine the rhetorical functions of topics associated with an abstract. There is much work on annotating and automatically parsing the rhetorical functions or narrative structure of scientific writing (e.g., Teufel, 2000; Chung, 2009; Gupta and Manning, 2011; de Waard and Maat, 2012). We derive insights from this prior work, but since we desire to apply our analysis to a broad range of domains, we build our narrative structure model based on over 83,000 self-labeled abstracts extracted from a variety of domains in the Web of Science corpus. Figure 2 shows an example of an abstract in which the authors have labeled the different narrative sections explicitly and identified the rhetorical functions. We use our narrative structure model to assign rhetorical function labels to scientific topics and show that these labels offer important clues indicating whether topics will eventually grow or decline.

Contributions: The three main contributions of our paper are: 1) we introduce the notion of the rhetorical scholarly functions of scientific *topics*, extending previous work which tended to focus on the rhetorical functions of individual *sentences*. We present an algorithm to assign rhetorical function labels to a topic as used in an individual paper; 2) we derive a new narrative scheme for scientific abstracts from over 83,000 abstracts that are labeled with narrative structures by their authors themselves, and present a tagger trained on this data that can parse unseen abstracts with 87% accuracy; 3) we show that the rhetorical function distribution of a topic reflects its temporal trajectory, and that it is predictive of whether the topic will eventually rise or fall in popularity.

2 Related Work

Our work builds upon a wealth of previous literature in both topic modeling and scientific discourse analysis, which we discuss in this section. We also discuss how our work relates to prior work on analyzing scientific trends.

2.1 Topic Modeling

Topic modeling has a long history of applications to scientific literature, including studies of temporal scientific trends (Griffiths and Steyvers, 2004; Steyvers et al., 2004; Wang and McCallum, 2006), article recommendation (Wang and Blei, 2011), and impact prediction (Yogatama et al., 2011). For example, Hall et al. (2008) and Anderson et al. (2012) show how tracking topic popularities over time can produce a 'computational history' of a particular scientific field (in their case ACL, where they tracked the rise of statistical NLP, among other dramatic changes).

Technical advancements in these areas usually correspond to modifications or extensions of the topic modeling (i.e., LDA) framework itself, such as by incorporating citation (Nallapati et al., 2008) or co-authorship information (Mei et al., 2008) directly into the topic model; Nallapati et al. (2011) employ such an extension to estimate the temporal "lead" or "lag" of different scientific information outlets. We contribute to this line of work by showing how we can build off of the standard LDA

framework—by overlaying rhetorical roles—and how this allows us to not only detect the growth and decline of scientific topics but also to predict these trends based upon the rhetorical roles being employed. Since our framework is structured as a pipeline (Figure 3) and works with the output of a topic modeling system, it is compatible with the vast majority of these extended topic models.

2.2 Scientific Discourse Analysis

Scientific discourse analysis is an active area of research with many different proposed schema of analysis — Argument Zones (Teufel, 2000), Information Structure (Guo et al., 2010), Core Scientific Concepts (Liakata, 2010), Research Aspects (Gupta and Manning, 2011), Discourse Segments (de Waard and Maat, 2012), Relation Structures (Tateisi et al., 2014), and Rhetorical Roles (Chung, 2009) to name a few. Most studies in this area focus on improving automatic discourse parsing of scientific text, while some works also focus on the linguistic patterns and psychological effects of scientific argumentation (e.g., de Waard and Maat, 2012). A wide range of techniques have been used in prior work to parse scientific abstracts, from fully supervised techniques (Chung, 2009; Guo et al., 2010) to semi-supervised (Guo et al., 2011c; Guo et al., 2013) and unsupervised techniques (Kiela et al., 2015).

Scientific discourse parsing has also been applied to other downstream tasks within the biomedical domain, such as information retrieval from randomized controlled trials in evidence based medicine (Chung, 2009; Kim et al., 2011; Verbeke et al., 2012), cancer risk assessment (Guo et al., 2011b), summarization (Teufel and Moens, 2002; Contractor et al., 2012), and question answering (Guo et al., 2013). Our work also falls in this category in the sense that our goal is to apply the rhetorical function parser to better understand the link between rhetoric and the historical trajectory of scientific ideas.

2.3 Scientific Trends Analysis

There is also a large body of literature in bibliometrics and scientometrics on tracking scientific trends using various citation patterns. Researchers have attempted to detect emerging research fronts using topological measures of citation networks (Shibata et al., 2008) as well as co-citation clusters (Small, 2006; Shibata et al., 2009). Unlike this line of work, our focus is not on citation patterns, but on how scientific trends are reflected in the texts of scientific publications.

Prior studies have also analyzed text to detect scientific trends. Mane and Börner (2004) and Guo et al. (2011a) use word burst detection (Kleinberg, 2003) to map new and emerging scientific fields, while Small (2011) examined sentiments expressed in the text surrounding citations, showing *uncertainty* in interdisciplinary citations contrasted with *utility* in within-discipline citations. In contrast to this previous work, we analyze the rhetorical function of automatically extracted topics from abstract text, without access to the citation context in full text.

3 Corpus

We use the Thomson Reuters Web of Science Core Collection, which contains scientific abstracts from over 8,500 of leading scientific and technical journals across 150 disciplines. We limit our study to the subset of abstracts from 1991 to 2010, which forms the majority of articles. This subset (denoted WoS hereafter) contains over 25 million articles from around 250 fields.

4 Rhetorical Functions of Scientific Topics

We use the term *rhetorical function* to identify the purpose or role a scientific topic plays within a research paper. This function qualifies the association between a topic and a paper.

A topic could represent the general domain of the research or its main objective/goal. It could also correspond to the data used, the way the research is designed, or the methods used. A topic may serve one or more of these roles within the same paper. The same topic may also serve different roles in different papers. We are interested in finding the different rhetorical functions by which topics are associated with the papers in our corpus, as a tool for understanding the growth and decline of topics over time. Our focus is thus on the rhetorical functions that topics play across papers, in order to understand the 'rhetorical structure of science' (Latour, 1987), although these are cued by specific rhetorical structures in individual sentences of individual papers. (Our work on the function of *topics* thus differs somewhat from previous research focusing on the rhetorical role of individual *sentences* or segments in the structure of the paper.)

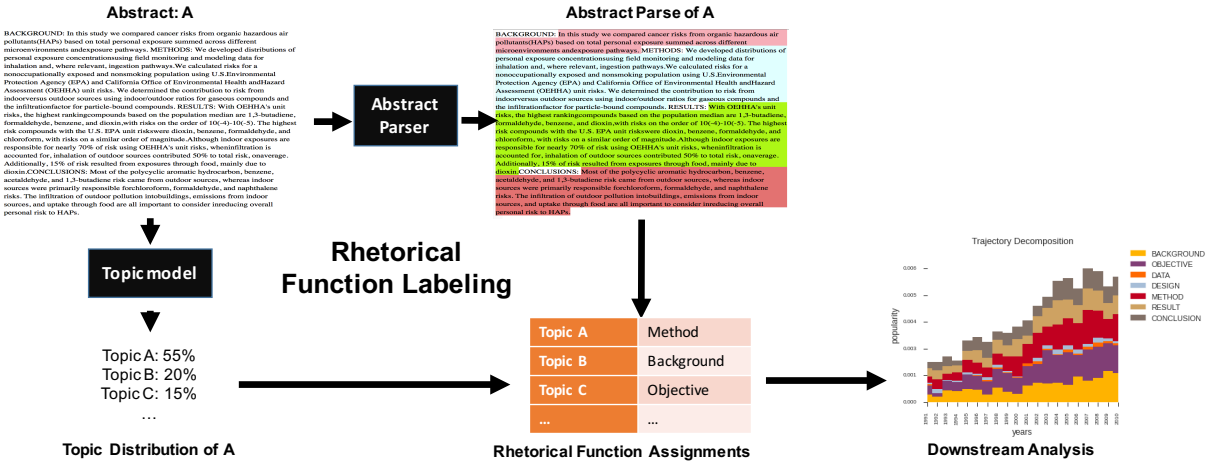


Figure 3: Rhetorical Function Labeling
The topic model (step a) assigns topic distributions to the abstract text (bottom left) and the abstract parser (step b) divides the text into discourse segments (top right). Rhetorical function labeling (step c) combines these two analyses to assign rhetorical functions to topics (bottom middle). These labels enrich the analysis of trends in topic popularity over decades (bottom right).

We follow a three-step process to assign rhetorical function labels to a large corpus of scientific abstracts. Figure 3 presents the pictorial representation of the procedure we follow — 1) obtain topic probabilities for each abstract, 2) parse the narrative structure of the abstracts in order to arrive at segments of text with different discourse intentions, and 3) superimpose the topic assignments and the abstract parse to arrive at the rhetorical function labels that capture how the topics are associated with the work presented in the abstract. We describe each of these steps in detail below.

a. Topic modeling: The core of our approach relies on the popular latent Dirichlet allocation (LDA) (Blei et al., 2003) algorithm. It probabilistically assigns both words and abstracts to different topics, in an unsupervised manner. Let $A = \{a_1, a_2, ...a_{|A|}\}$ be the set of abstracts within the field we are studying, and let $T = \{t_1, t_2, ...t_{|T|}\}$ be the set of different topics in the field. A topic model trained on A assigns $\theta_{a,t}$, the probability of topic t occurring in abstract a for all $a \in A$ and $t \in T$. The topic model also provides the $\phi_{t,w}$, the probability of word w occurring in topic t.

b. Abstract parsing: An abstract parser divides the abstract text into a sequence of discourse segments, with each segment assigned a specific label denoting its rhetorical purpose. Let $S(a) = (s_1, s_2, ...s_{|S(a)|})$ be the sequence of segments identified by the abstract parser, and let L denote the set of labels in the abstract parsing framework. The abstract parser assigns a label $l(s_i) \in L$, for each $s_i \in S(a)$.

c. Rhetorical Function Labeling: We tease apart the abstract-level topic distribution $\theta_{a,t}$ assigned by the topic model (step a) along the segments found by the abstract parser (step b), and find the topic weights on each label $\theta_{l,t}(a)$ by calculating topic weights for segments that are assigned label l, i.e., $\{s_i \in S(a) : l(s_i) = l\}$. We calculate the topic weights for each segment by aggregating the topic weights on each word derived from $\phi_{t,w}$ inferred by the topic model:

$$\theta_{l,t}(a) \propto \sum_{w_i \in s_i : l(s_i) = l} \phi_{t,w} \qquad (1)$$

We first describe the abstract parsing system (step b) we built for this purpose in Section 5, before discussing the execution details of each of the above three steps in Section 6.

5 Abstract Parsing

Datasets with manual annotations for discourse structure of abstracts (e.g., Guo et al., 2010; Gupta and Manning, 2011) are few, small, and limited to specific domains. It is not clear how accurate an abstract parser trained on these datasets will perform on other domains. Since we want to obtain the structure of abstracts in a broad range of domains over different time-periods, a parser trained

on small datasets in specific domains may not be adequate for our purposes. Hence, we exploit the large number of abstracts in the WOS corpus in order to gather a dataset of self-labeled abstracts from a wide range of domains, over a period of two decades. By self-labeled abstracts, we refer to the abstracts where the authors have identified the discourse segments using explicit section labels.

5.1 Extracting Self-labeled Abstracts

In the first step, we extract all the patterns from the WOS corpus that could potentially be a segment label. For this, we look for a pattern that is commonly used by authors to label abstract segments — a capitalized phrase of one or more words occurring at the beginning of the abstract or preceded by a period, and followed by a ":". We obtained 455,972 matches for the above pattern, corresponding to 2,074 unique labels, majority of which were valid discourse segment labels. These include variations of the same labels (e.g., "OBJECTIVE" and "AIM", "CONCLUSION" and "CONCLUSIONS" etc.) and typos (e.g., "RESLUTS"). There were also instances where two common labels were combined (e.g., "DATA AND METHODS"). The extracted matches also contained a long tail of false positives (e.g., "BMI").

One of the challenges in using the set of abstracts we obtained above is that they do not follow a common labeling scheme. Hence, we manually analyzed the top 100 unique labels (which corresponds to labels with more than $\sim$50 instances) and mapped them into a unified labeling scheme, grouping together labels with similar intentions. This resulted in a typology of seven labels:

- BACKGROUND: The scientific context
- OBJECTIVE: The specific goal(s)
- DATA: The empirical dimension used
- DESIGN: The experimental setup
- METHOD: Means used to achieve the goal
- RESULT: What was found
- CONCLUSION: What was inferred

We use this mapping to obtain abstracts that are self-labeled. We exclude the abstracts that had combined labels, since they may add noise to the training data. We also exclude abstracts that contained only false positive matches. This preprocessing resulted in a dataset of 83,559 abstracts. We refer to this dataset as SL, hereafter. We divide the SL dataset into Train/Dev/Test subsets for our experiments (Table 1).

	Train	Dev	Test
# of abstracts	58,600	12,331	12,628
# of labeled segments	243,217	51,111	52,403
# of sentences	681,730	143,792	147,321

Table 1: Statistics of self-labeled abstracts

5.2 Automatic tagging of abstracts

We use the SL dataset to build a supervised learning system that can predict the abstract structure in unseen documents. We perform the prediction at the sentence level. We used the CRF algorithm to train the model, as it has been proven successful in similar tasks in prior work (Hirohata et al., 2008; Merity et al., 2009). In our preliminary experiments, we also tried using SVM, but CRF was faster and outperformed SVM by 4-5% points consistently. We use the following features:

1. Location: location of the sentence from the beginning or end of the abstract
2. Word ngrams: unigrams and bigrams of word lemmas
3. Part-of-speech ngrams: unigrams and bigrams of part-of-speech tags
4. Verb lemmas and part-of-speeches: lemmas of verbs, and their part-of-speech tags in order to identify the tense of verb usage
5. Verb classes: we looked up each verb in the VerbNet (Kipper-Schuler, 2005) index and added the VerbNet class to the feature set if the verb maps to a unique verb class.
6. Concreteness rating: we used the max, min, and mean concreteness ratings of words based on (Brysbaert et al., 2014).

Most of these features are commonly used in similar tasks, while the concreteness features are new (and significantly improved performance). We do not use parse features, however, since our pipeline emphasizes computational efficiency, and parse features showed minimal utility relative to their computational cost in prior work (Guo et al., 2010).

We evaluate the performance of our learned model in terms of overall accuracy as well as per-class precision, recall and F-measure of predicting the segment labels at the sentence level. We performed experiments on the Dev set to choose the best feature configuration (e.g., tuning for word and part-of-speech ngram length). Each feature set described in the previous paragraph were con-

	Precision	Recall	F-measure
BACKGROUND	74.6	77.2	75.8
OBJECTIVE	85.2	81.8	83.5
DATA	82.6	76.8	79.6
DESIGN	68.0	64.8	66.3
METHOD	80.4	80.1	80.2
RESULT	90.8	93.3	92.0
CONCLUSION	93.8	92.0	92.9
Accuracy		86.6	

Table 2: Results of parsing abstract structure

tributing features to the best performance obtained on the Dev set. The concreteness features we introduced significantly improved the overall accuracy by around 2%.

Table 2 shows the results obtained on testing the final classifier system on the Test subset of SL.[1] We obtain an overall high accuracy of 86.6% at the sentence level. While RESULT and CONCLUSION obtained F-measures above 90%, OBJECTIVE and METHOD reported reasonable F-measures above 80%. DESIGN obtained the lowest precision, recall and F-measure. Overall, the performance we obtain is in the range of other reported results in similar tasks (Guo et al., 2013).

6 Analysis Setup

In the rest of this paper, we apply the rhetorical function labeling system described in Section 4 to analyze the growth and decline of scientific topics. We chose four diverse fields from the WOS corpus with large numbers of abstracts for our analysis, which are: *Biochemistry & Molecular Biology* (BIO): 850,394 abstracts, *Applied Physics* (PHY): 558,934 abstracts, *Physical Chemistry* (CHM): 533,871 abstracts, and *Neurosciences* (NEU): 477,197 abstracts. We apply the steps (a), (b), and (c) of rhetorical function labeling as described in Section 4 to these fields as follows:

Topic Modeling: We use the LightLDA implementation (Yuan et al., 2015) of LDA. It employs a highly efficient and parallelized Metropolis-Hastings sampler that allows us to scale our approach to massive datasets (e.g., millions of abstracts in our case). For all our experiments, we ran the algorithm for 1000 iterations, as this was sufficient for convergence. We use 500 topics for all four fields, but otherwise use the default hyperparameter settings from the LightLDA package.[2]

Abstract Parsing We applied the 7-label abstract parsing system described in Section 5 on all abstracts in each of the four disciplines.

Rhetorical Function Labeling Once the steps (a) and (b) are completed, we applied the rhetorical function labeling step (c) from Section 4 in order to obtain topic weights for each segment. In addition, we calculate a rhetorical function label distribution (referred to as *label distribution* hereafter) for each topic associated with an abstract.

7 Dissecting Topic Trajectories

In this section, we investigate whether the label distribution of the topics (i.e., across the rhetorical function labels) sheds light on the kind of trajectories they follow; in particular, whether it can predict the up-trend vs. down-trend of topics. We formalize the problem as follows: given two sets of topics clustered based on their historical trajectories, do the label-distribution based features of a topic have predictive power to classify the topic to be in either set?

7.1 Tracking topic popularities

For each field, we first calculated the number of articles in which each topic occurred in at least one of the rhetorical functions for each year in the 20-year period 1991-2010. Since the fields themselves grow over the years, we divide this number by the number of articles within that field to obtain the popularity of each topic in any year:

$$popularity(t, y) = DocsWithTopic(t)/|A_y|,$$

where A_y denotes the subset of articles in year y.

7.2 Detecting growing vs. declining topics

We are interested in the set of topics in each field that are either growing or declining. Figure 4 shows example popularity trends for two such topics in neuroscience: stem cell research, which sees

[1] We report only the results obtained in unseen abstracts in the Test set due to lack of space. Similar performance was obtained in the Dev set as well.

[2] In preliminary experiments, we used 100, 500 and 1000 topics. Upon manual inspection, we found that 500 topics resulted in a granularity that better captures the scientific intellectual movements within fields.

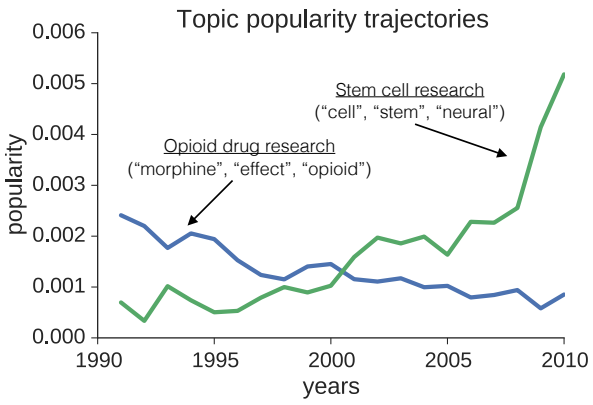


Figure 4: Example topic popularity curves from 1991 till 2010. Stem cell research (green) sees an increase in popularity over this time period, while Opioid research (blue) declines in popularity.

a dramatic increase in popularity from 1991-2010, and opioid[3] drug research, which declines in popularity during the same time-period.

Of course, topics do not always follow trajectories of pure growth or decline. A topic may have risen and subsequently fallen in popularity over the period of 20 years, or may stay more or less the same throughout (Griffiths and Steyvers, 2004). Hence, categorizing topics to be growing or declining solely based on the popularity at the beginning and end of the time period is problematic. We circumvent this issue and avoid manually-defined thresholds by discovering topical growth/decline curves in an unsupervised fashion. We use the K-Spectral Centroid (K-SC) algorithm (Yang and Leskovec, 2011), which groups different time-series into clusters, such that similar shapes get assigned to the same cluster, irrespective of scaling and translation (unlike other popular time series clustering algorithms such as Dynamic Time Warping). We run the clustering algorithm using $K = 3$, and choose the cluster with the upward trending centroid to be the set of growing topics and the cluster with the downward trending centroid to be the set of declining topics.[4] Figure 5 shows example centroids from Physical Chemistry, which clearly exhibit decreasing, increasing, and non-changing trends.

[3]Opioids are a popular class of analgesic (i.e., painkilling) drugs, including morphine.

[4]We also performed experiments using K = 5, 10 and 15 and grouped the clusters that are growing vs. declining. We obtained similar results in those experiments as well.

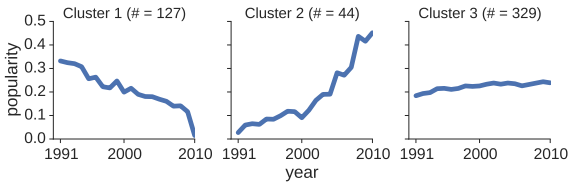


Figure 5: Cluster centroids in Physical Chemistry
Cluster 1: topics that declined in popularity;
Cluster 2: topics that grew in popularity;
Cluster 3: topics that stayed mostly the same.

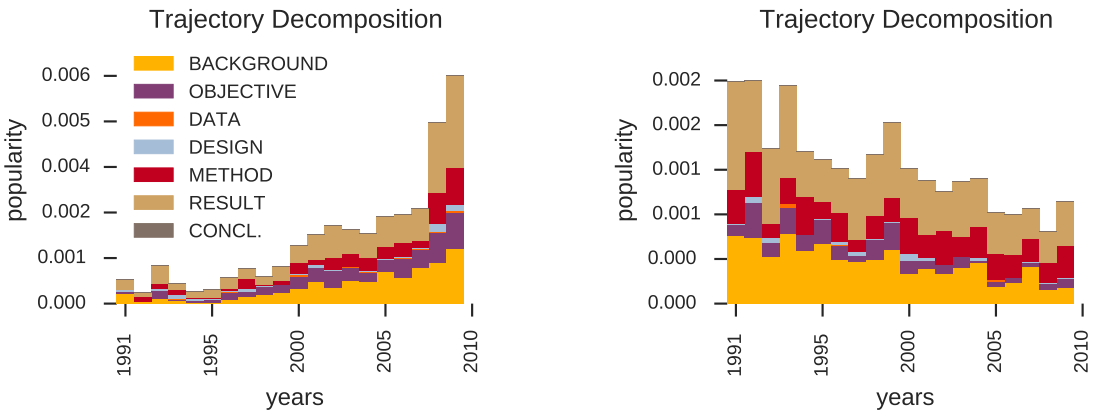
7.3 Characterizing topic growth vs. decline

We not only seek to detect growing vs. declining topics; our goal is to *characterize* these trajectories based upon the rhetorical functions that the topics are fulfilling during different points of their life-cycles. Figure 6 shows how the rhetorical function label distributions of the opioid and stem cell research topics shift from 1991 to 2010. The opioid research topic, which declines in popularity during this time-period, is frequently discussed in the RESULT and BACKGROUND roles during the early years. In contrast, the stem cell research topic, which is dramatically increasing in popularity, only begins to be discussed frequently in these roles towards the end of the time-period. Intuitively, these shifts make sense: topics become results-oriented and mentioned as background as they reach their peak; this peak is seen at the beginning of the time-period for opioid research, while stem cell research appears to be increasing towards a peak by the end of our data (i.e., 2010). These observations indicate that the rhetorical functions which a topic is fulfilling may be indicative of its current growth vs. decline.

7.4 Experiments

We perform two sets of experiments to quantify the qualitative insights noted above, i.e. that a topic's rhetorical function distribution is indicative of its eventual growth vs. decline. First, we show that a topic's rhetorical function distribution over the entire time-period can be used to *classify* the topic as either growing or declining. Next, we show that using only the first five years of data, we can *predict* whether a topic will eventually grow or decline. We use two sets of features:

- **label-distribution-percents (LD-%)**: seven features corresponding to the percentage of topics across the seven rhetorical function labels (e.g., % of time the topic is a *METHOD*)



(a) Stem cell topic.

(b) Opioid topic.

Figure 6: Examples of topics with changing rhetorical functions over time.

- **label-distribution-changes (LD-Δ)**: seven features that aggregate the mean change in this percentage over the years (e.g., is the % of being METHOD going up/down?).

These features are fed into a standard $L2$-regularized logistic regression classifier (with regularization parameter $C = 10$), which predicts whether a topic belongs to the growing or declining cluster.[5] We use random prediction and majority prediction as two uninformed baselines.

Classification task

System	ALL	BIO	PHY	CHM	NEU
Random	50.3	47.2	47.8	50.9	51.2
Majority	56.1	56.3	81.6	74.3	56.6
LR	74.2	81.0	83.3	81.9	74.8
LR - LD-R	71.3	77.7	81.6	73.1	70.5

Table 3: Results on classifying trajectories
LR: Logistic Regression using LD-%, LD-Δ, and LD-R.
LR - LD-R: Logistic Regression without using LD-R.

For the task of classifying a topic as either growing or declining, we compute the LD-% and LD-Δ features separately over the first and last five years of the data. This allows the model to analyze how the rhetorical functions are differentially expressed at the start vs. the end of the period under consideration. We also add a feature **label-distribution-ratio (LD-R)**, which is the ratio of the end to beginning LD-% values, so that the model has access to relative increases/decreases in the label distributions over the entire time-period.

[5]Similar performance was obtained with a linear SVM

Table 3 shows the performance of our model on this task. As expected a topic's label distribution over its entire life-time is very informative with respect to classifying the topic as growing or declining. We achieve a significant improvement over the baselines on the full dataset (32.3% relative improvement over majority prediction), and this trend holds across each field separately. The ratio feature proved to be extremely predictive in this task, i.e. relative increases/decreases in a topic being used in different functional roles are very predictive of the type of its trajectory.

Prediction task

We now turn to the more difficult task of *predicting* whether a topic will grow or decline, given access only to the first five years of its label distribution. The setup here is identical to the classification task, except that we now have access only to the LD-Δ and LD-% features aggregated over the first five years of data.

System	Accuracy on ALL
LD-% + LD-Δ	72.1
LD-% only	71.0
LD-Δ only	60.4

Table 4: Results on predicting trajectory

Table 4 shows the performance of our model on this task. (The baseline performances are the same as in the classification task). These results show that we can accurately predict whether a topic will grow or decline using only a small amount of data.

Label	BKGRND.	OBJ.	DATA	DESIGN	METHOD	RESULT	CONC.
LD-% Weight	-1.21	-0.10	6.38	1.21	3.82	-8.65	1.67
LD-Δ Weight	2.05	-0.01	2.20	-1.08	-1.63	-0.26	-1.27

Table 5: Logistic Regression feature weights for the prediction task on the full (ALL) dataset.

Moreover, we see that both percentage and delta features are necessary for this task.

7.5 Analysis

The feature weights of our learned linear model also provide insights into the dynamics of scientific trends. Table 5 contains the learned feature weights for the prediction task. Overall, these weights reinforce the conclusions drawn from the case study in Section 7.3. The strongest feature is the LD-% feature for the RESULT rhetorical function, which is highly negative. This indicates that topics that are currently being discussed as a result are likely at the peak of their popularity, and are thus likely to decrease in the future. Interestingly, the weights on the LD-% features for the methodological rhetorical functions (METHODS, DATA, and DESIGN) are all significantly positive. This suggests that topics occupying these functions may have reached a level of maturity where they are active areas of research and are being consumed by a large number of researchers, but that they have not yet peaked in their popularity. Finally, we see that the weights for the BACKGROUND and CONCLUSION roles have opposite trends: growing topics are more often mentioned as conclusions whereas dying topics, i.e. topics at the peak of their life-cycles, tend to be mentioned in background, or contextualizing, statements.

8 Conclusion

We introduce a novel framework for assigning rhetorical functions to associations between scientific topics and papers, and we show how these rhetorical functions are predictive of a topic's growth vs. decline.

Our analysis reveals important regularities with respect to how a topic's usage evolves over its life-cycle. We find that topics that are currently discussed as results tend to be in decline, whereas topics that are playing a methodological role tend to be in the early phases of growth. In some ways these results are counter-intuitive; for example, one might expect topics that are being discussed as results to be the focus of current cutting edge research, and that methodological topics might be more mundane and lacking in novelty. Instead our results suggest that results-oriented topics are at the peak of their life-cycle, while methodological topics still have room to grow. This result has important implications for research funding and public policy: the most promising topics—in terms of potential for future growth—are not those that are currently generating the most results, but are instead those that are active areas of methodological inquiry.

Our analysis does suffer from some limitations. Examining only 20 years of scientific progress prevents us from analyzing drastic scientific changes, e.g. paradigm shifts, that are only obvious over longer time-scales (Kuhn, 2012). Access to longer time-spans—along with varying data sources such as grants and patents—would also allow us to more completely model the trajectory of a topic as it moves from being active area of research to potentially impacting commercial industries and economic development. Nonetheless, we hope this work offers another step towards using computational tools to better understand the 'rhetorical structure of science' (Latour, 1987).

Acknowledgments

This work was supported by the NSF Award IIS-1159679, by the Stanford Data Science Initiative, and the Brown Institute for Media Innovation. W.H. was supported by the SAP Stanford Graduate Fellowship. We would also like to thank the ACL anonymous reviewers for their constructive feedback.

References

Ashton Anderson, Dan McFarland, and Dan Jurafsky. 2012. Towards a computational history of the ACL: 1980-2008. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 13–21. Association for Computational Linguistics.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *The Journal of machine Learning research*, 3:993–1022.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Grace Y. Chung. 2009. Sentence retrieval for abstracts of randomized controlled trials. *BMC Medical Informatics and Decision Making*, 9(1):1–13.

Danish Contractor, Yufan Guo, and Anna Korhonen. 2012. Using argumentative zones for extractive summarization of scientific articles. In *COLING*, volume 12, pages 663–678. Citeseer.

Graham Crookes. 1986. Towards a validated analysis of scientific text structure. *Applied linguistics*, 7(1):57–70.

Anita de Waard and Henk Pander Maat. 2012. Verb form indicates discourse segment type in biological research papers: Experimental evidence. *Journal of English for Academic Purposes*, 11(4):357 – 366.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins Karolinska, Lin Sun, and Ulla Stenius. 2010. Identifying the information structure of scientific abstracts: an investigation of three different schemes. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 99–107. Association for Computational Linguistics.

Hanning Guo, Scott Weingart, and Katy Börner. 2011a. Mixed-indicators model for identifying emerging research areas. *Scientometrics*, 89(1):421–435.

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins, Johan Hogberg, and Ulla Stenius. 2011b. A comparison and user-based evaluation of models of textual information structure in the context of cancer risk assessment. *BMC bioinformatics*, 12(1):1.

Yufan Guo, Anna Korhonen, and Thierry Poibeau. 2011c. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 273–283, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Yufan Guo, Ilona Silins, Ulla Stenius, and Anna Korhonen. 2013. Active learning-based information structure analysis of full scientific articles and two applications for biomedical literature review. *Bioinformatics*, 29(11):1440–1447.

Sonal Gupta and Christopher Manning. 2011. Analyzing the dynamics of research by extracting key aspects of scientific papers. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1–9, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 363–371. Association for Computational Linguistics.

Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. 2008. Identifying sections in scientific abstracts using conditional random fields. In *In Proc. of the IJCNLP 2008*.

Douwe Kiela, Yufan Guo, Ulla Stenius, and Anna Korhonen. 2015. Unsupervised discovery of information structure in biomedical documents. *Bioinformatics*, 31(7):1084–1092.

S Nam Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic classification of sentences to support evidence based medicine. *BMC bioinformatics*, 12(2):1.

Karin Kipper-Schuler. 2005. *VerbNet: a broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, Computer and Information Science Department, Universiy of Pennsylvania, Philadelphia, PA.

Jon Kleinberg. 2003. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397.

Thomas S Kuhn. 2012. *The structure of scientific revolutions*. University of Chicago press.

Bruno Latour. 1987. *Science in action: How to follow scientists and engineers through society*. Harvard university press.

Maria Liakata. 2010. Zones of conceptualisation in scientific papers: a window to negative and speculative statements. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 1–4, Uppsala, Sweden, July. University of Antwerp.

Ketan K Mane and Katy Börner. 2004. Mapping topics and topic bursts in pnas. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5287–5290.

Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. 2008. Topic modeling with network regularization. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 101–110, New York, NY, USA. ACM.

Stephen Merity, Tara Murphy, and James R Curran. 2009. Accurate argumentative zoning with maximum entropy models. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 19–26. Association for Computational Linguistics.

Ramesh M. Nallapati, Amr Ahmed, Eric P. Xing, and William W. Cohen. 2008. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 542–550, New York, NY, USA. ACM.

Ramesh Maruthi Nallapati, Xiaolin Shi, Daniel McFarland, Jure Leskovec, and Daniel Jurafsky. 2011. Leadlag lda: Estimating topic specific leads and lags of information outlets. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Naoki Shibata, Yuya Kajikawa, Yoshiyuki Takeda, and Katsumori Matsushima. 2008. Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation*, 28(11):758–775.

Naoki Shibata, Yuya Kajikawa, Yoshiyuki Takeda, and Katsumori Matsushima. 2009. Comparative study on methods of detecting research fronts using different types of citation. *Journal of the American Society for Information Science and Technology*, 60(3):571–580.

Henry Small. 2006. Tracking and predicting growth areas in science. *Scientometrics*, 68(3):595–610.

Henry Small. 2011. Interpreting maps of science using citation context sentiments: A preliminary investigation. *Scientometrics*, 87(2):373–388, May.

Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. 2004. Probabilistic author-topic models for information discovery. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 306–315, New York, NY, USA. ACM.

Yuka Tateisi, Yo Shidahara, Yusuke Miyao, and Akiko Aizawa. 2014. Annotation of computer science papers for semantic relation extrac-tion. In *LREC*, pages 1423–1429.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445.

Simone Teufel. 2000. *Argumentative zoning: information extraction from scientific text*. Ph.D. thesis, University of Edinburgh.

Mathias Verbeke, Vincent Van Asch, Roser Morante, Paolo Frasconi, Walter Daelemans, and Luc De Raedt. 2012. A statistical relational learning approach to identifying evidence based medicine categories. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 579–589. Association for Computational Linguistics.

Chong Wang and David M. Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 448–456, New York, NY, USA. ACM.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433. ACM.

Jaewon Yang and Jure Leskovec. 2011. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186. ACM.

Dani Yogatama, Michael Heilman, Brendan O'Connor, Chris Dyer, Bryan R Routledge, and Noah A Smith. 2011. Predicting a scientific community's response to an article. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 594–604. Association for Computational Linguistics.

Jinhui Yuan, Fei Gao, Qirong Ho, Wei Dai, Jinliang Wei, Xun Zheng, Eric Po Xing, Tie-Yan Liu, and Wei-Ying Ma. 2015. LightLDA: Big Topic Models on Modest Computer Clusters. In *Proceedings of the 24th International Conference on World Wide Web*.