Finding Keystone Citations for Constructing Validity Chains among Research Papers

Yuanxi Fu
School of Information Sciences, University of Illinois at Urbana-Champaign, United States
fu5@illinois.edu

Jodi Schneider
School of Information Sciences, University of Illinois at Urbana-Champaign, United States
jodi@illinois.edu

Catherine Blake
School of Information Sciences and Department of Computer Science, University of Illinois at Urbana-Champaign, United States
clblake@illinois.edu

ABSTRACT

New discoveries in science are often built upon previous knowledge. Ideally, such dependency information should be made explicit in a scientific knowledge graph. The Keystone Framework was proposed for tracking the validity dependency among papers. A keystone citation indicates that the validity of a given paper depends on a previously published paper it cites. In this paper, we propose and evaluate a strategy that repurposes rhetorical category classifiers for the novel application of extracting keystone citations that relate to research methods. Five binary rhetorical category classifiers were constructed to identify Background, Objective, Methods, Results, and Conclusions sentences in biomedical papers. The resulting classifiers were used to test the strategy against two datasets. The initial strategy assumed that only citations contained in Methods sentences were methods keystone citations, but our analysis revealed that citations contained in sentences classified as either Methods or Results had a high likelihood to be methods keystone citations. Future work will focus on fine tuning the rhetorical category classifiers, experimenting with multiclass classifiers, evaluating the revised strategy with more data, and constructing a larger gold standard citation context sentence dataset for model training.

CCS CONCEPTS

• **Information systems** → Information retrieval; Document representation; • **Computing methodologies** → Machine learning; Learning paradigms; Supervised learning.

KEYWORDS

Knowledge dependency, validity, citation context classification, argumentation, methods

ACM Reference Format:
Yuanxi Fu, Jodi Schneider, and Catherine Blake. 2021. Finding Keystone Citations for Constructing Validity Chains among Research Papers. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3442442.3451368

WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3451368

1 INTRODUCTION

New discoveries in science are often built upon previous knowledge. For example, Watson and Crick's discovery of the double helix structure of DNA depends, fundamentally, on Erwin Chargaff's discovery of the A-T and C-G pairings and Rosalind Franklin and Maurice Wilkins' X-ray crystallography work [14]. Ideally, such dependency information should be made explicit in a scientific knowledge graph. Graphs that incorporate dependency information have the potential to reveal the flow of information among researchers and fields; to generate data that can support better research impact assessment; and to track what else in the knowledge graph is affected when a paper loses its validity. This work is motivated by the last case.

Our previous work proposed a framework for tracking validity dependencies among research papers, named the Keystone Framework [6]. A keystone citation indicates that the validity of a given paper depends on a previously published paper it cites. The name is inspired by masonry, where damage to the keystone can threaten the arch it supports. One challenge is that, in general, finding keystone citations requires a global understanding of a scientific paper, which may limit automated approaches. However, a subset of keystone citations is more feasible to automatically detect: Keystone citations that support research methods and materials, as their keystone status can be determined only by using the citation context (i.e., the text surrounding a citation). Thus, for the remainder of the paper, we focus on how to use supervised machine learning to detect this subset of keystone citations.

2 RELATED WORK

2.1 Representing Scientific Evidence

The Keystone Framework is a part of a broader research effort to formalize the knowledge representation of a scientific publication so that its validity can be examined and re-assessed by human and machine readers. The Keystone Framework guides users through a process to find citations that are a "keystone" to the citing paper's arguments. In the first step, a paper's claims and supporting arguments are modeled into graph-like argument diagrams. In the second step, users try to match citations to components in the diagram using the citation contexts. Through a checklist provided in [6], users can determine whether a citation is a keystone citation, and if it is, what type of keystone citation it is.

A few existing semantic models can be used in the first step of document modeling: the Micropublication Ontology [4], the

Scientific Evidence and Provenance Information Ontology (SEPIO) [3], and the Reasoning and Discourse Ontology (RDO) [2].

The Micropublication Ontology was proposed to transform text-bound and linear-format scientific publications into web-friendly and machine-tractable digital objects [4]. In its minimal form, a micropublication has a statement and its attribution. In a more expanded form, a micropublication can be supported by a support-graph, which encompasses many elements critical to the creation of scientific arguments, such as data, methods, materials, and references, allowing more detailed examination.

SEPIO was initially designed to aid data integration across various model organism and clinical genetics databases, but it is also a domain-independent conceptual model capable of representing diverse evidence and provenance information [3]. It consists of four core informational entities: Assertions, propositions, supporting data items, and evidence lines, and two provenance-related entities: Assertion process and data generation process. In particular, the data generation process entity is further supported by entities such as technique (i.e., methods), resources (i.e., materials), date-time, and agents.

RDO is a part of the Scientific EvidencE (SEE) approach, which aims to represent arguments as they are presented in the source [2]. RDO has five core entity classes: Assertion, proposition, text, report, and agent. One key property, "is inferred from," relates one assertion to another and can be infinitely chained, thus creating an evidence trail for a specific claim.

The contribution of the Keystone Framework is that it focuses on citation relationships and the transmission of validity. Moreover, despite the different constructs of the three semantic models, one commonality is that they all considered research methods and materials as an indispensable part of the model, either being explicit entity classes as in the Micropublication Ontology and SEPIO, or as assertions in RDO. Therefore, under any of these three models, citations that support methods and materials will always be keystone citations, backing our assumption that citations that support research methods and materials (referred to as "methods keystone citations" hereafter) can be extracted as keystone citations without a global understanding of a paper.

2.2 Classifying Citation Context Sentences into Methods/non-methods

Citation context sentences can be used to classify citation into "Incidental" and "Important" citations [8, 11, 17]. "Important" citations are cited work being used or extended by the citing papers, which has some overlap with our classification task. The difference is that methods keystone citations provide justifications for the use of methods or materials, which is broader than simply "being used."

Citation context sentences can also differentiate method and non-method papers. Here, a "method paper" refers to a paper whose main contribution to science is the development of a method. Method papers are cited with less hedging [16], and they enjoy more citations than non-method papers [15], since the latter are more likely to receive decreasing number of citations due to a phenomenon called "obliteration by incorporation" [9, 10], which means when a paper's discovery becomes established knowledge, authors no longer feel the need to cite the source paper. Utility

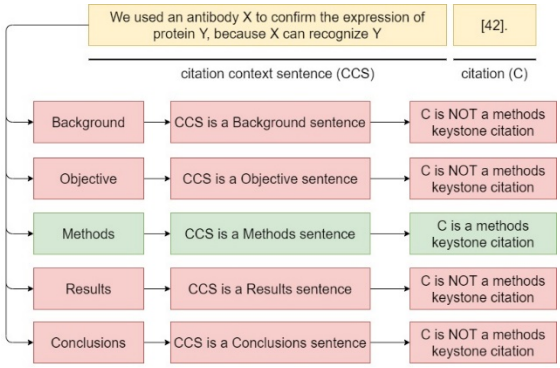

Figure 1: The concept of the strategy

words, such as "use", "used", "using", and "based" in the citation context were found to be strong indicators of method papers [15]. However, as we will show later, method papers may not be directly "used" in the papers citing them. And non-method papers, such as reviews, can also be used to support methods [6].

3 STRATEGY

The proposed strategy to extract keystone citations is depicted in Figure 1. First, we repurposed rhetorical category (RC) classifiers. They are used to assign IMRAD labels (e.g., Introduction, Methods, Results, and Discussion) to sentences in unstructured biomedical abstracts [7, 12, 19]. In particular, a Methods sentence describes "the way of doing research" [7]. One advantage of using RC classifiers is that training data are relatively easy to obtain. They can be constructed using biomedical abstracts with IMRAD labels. Moreover, we were able to obtain a "cleaner" training dataset that was manually labeled at the sentence level to one of the following categories: Background, Objective, Methods, Results, and Conclusions. This dataset allows us to "cold start" the project without labeling our own dataset. One limitation of this dataset is that it is from abstracts, whose language styles may differ from that of the full text of a research paper [5].

As depicted in Figure 1, according to this strategy, a citation context sentence (CCS) is passed through the RC classifiers. If the CCS is classified as Methods, the citation is a methods keystone citation. Otherwise, it is not. One underlying assumption is that the reason authors include a citation in a Methods sentence is to provide support to the research method or material used. In the example sentence shown in Figure 1, a method, the use of antibody X to confirm the expression of protein Y, is followed by a citation "[42]". Unless incorrectly cited, the paper [42] should provide some justification for the method, such as a prior usage of the method or evidence that antibody X can recognize protein Y.

4 METHODS

4.1 Datasets

An unpublished dataset (5,517 sentences) was used to train the models. Each sentence was manually labeled to one of the following

Table 1: Best classifiers by F1 scores obtained from 10-fold cross-validation on the training dataset

Class	No. of Features	Accuracy	Precision	Recall	F1
Background	100	0.858	0.671	0.278	0.392
Objective	100	0.934	0.826	0.339	0.477
Methods	700	0.865	0.820	0.542	0.652
Results	800	0.814	0.835	0.585	0.688
Conclusions	100	0.858	0.684	0.216	0.327

rhetorical categories: Background (16.5%), Objective (8.9%), Methods (23.5%), Results (35.1%), and Conclusions (16.0%). To construct this dataset, 500 abstracts were randomly selected from PubMed without sub-field specifications to maximize the generalizability of the dataset. All sentences in the 500 abstracts were included, except 34 sentences that were not part of the narrative, such as publication information or funding information. Three experts in biomedical informatics annotated the dataset. They first annotated 10 abstracts to develop guidelines, then, all three annotators annotated 50 more abstracts together. The inter-annotator agreement was found to be high (Fleiss' kappa = 0.92) for the 50 abstracts, so the rest 440 abstracts were split among the three.

Two more datasets were used to test our strategy. The first is a gold-standard keystone citation context data set: the JCDL dataset contains nine keystone citation context sentences collected by the authors YF and JS for [6], all supporting methods and materials (Table 2). The second dataset was chosen as a larger testbed: the Willoughby-Hoye dataset is a collection of 99 citation context sentences citing the Willoughby-Hoye protocol [18] downloaded from scite.ai[1] on Dec 30, 2020. This paper was chosen since it was found to contain a code glitch [1] and was a subject of our previous study [6].

4.2 Building classifiers

Five binary classifiers were built, one for each rhetorical category. The standard "bag-of-words" representation was used that is known to work well for text in general [20, 21] and in previous studies of rhetorical category classifiers [7, 12, 19]. Preprocessing included lowing cases and removing of stop words, and features were selected based on information gain [21].

The Support Vector Machines (SVM) algorithm (Scikit-learn version 0.24.0 [13]) was chosen based on a pilot study where this model performed better than the Naïve Bayes and Decision Tree classifiers. The configuration used was C-support vector classification with rbf kernel, using all default settings of sklearn.svm.svc method without fine tuning of the parameters. Comparison between the three classification algorithms (i.e., SVM, NB, and decision tree) can be found in Doc1 in a GitHub repository (https://github.com/yuanxiesa/Scik2021..

The number of features was varied from 100 to 1000, with an increment of 100. The best model for each rhetorical category was identified by the average F1 score obtained through 10-fold cross-validation.

5 RESULTS

Performance metrics for the five best classifiers are listed in Table 1. Accuracy scores for all rhetorical classes were above 0.8. The performance suggests that the predictive performance was likely limited by the training set size, because the two classes with the most instances, Methods and Results, achieved better F1 scores than the other three classes.

Results on the JCDL dataset are shown in Table 2. Four sentences were captured by the Methods classifier. On the other hand, sentence 3 was captured by the Results classifier. Close examination shows that it is a hybrid: It describes both a method, the use of a monoclonal antibody to confirm the expression of tau protein, and a result, the confirmation of the strong expression of tau protein. Among the four sentences that were missed, sentence 1, 2, and 8 describe "ways of doing research" but were not captured, a failure of the Methods classifier. Sentence 4 is special because it provides a justification for a method (i.e., the use of synaptic marker to measure neuron damage [6]), and the relation between sentence 4 and methods used in the paper is not explicit in sentence 4.

When applying the rhetorical category classifiers to the Willoughby-Hoye dataset, 43 of the 99 instances received a positive classification. One of the authors, YF, examined those 43 sentences and determined whether the Willoughby-Hoye protocol is a methods keystone citation in those cases, drawing on experience from the previous analysis [6]. The citation context sentences, their rhetorical category classifications, and keystone citation annotation can be found in Doc 2 of the GitHub repository (link provided in section 4.2). The results are summarized in Table 3, including the number of instances where Willoughby-Hoye protocol is a methods keystone citation, the total number of instances identified by each classifier, and the ratio between the two.

Table 3 shows that our premise that only citations contained in Methods sentences are methods keystone citations (Figure 1) needs revision. Citations contained in Methods and Results sentences both have a high likelihood of being methods keystone citations (95% and 100%, respectively). While we did not expect the Results classifier to be a keystone citation capture device, two factors altered this view. The first is the existence of Results-Methods hybrids. Second, some Results sentences describe "the way of doing research" and contain phrases that give a sense of closure, such as "were calculated" or "were carried out," making them classified as Results.

Sentences classified as Background and Conclusion sentences have a low likelihood of containing methods keystone citations. Background sentences situate the Willoughby-Hoye protocol to a research landscape. While we expected no methods keystone

[1] https://scite.ai/, a proprietary platform

Table 2: Classification results of the JCDL dataset

Keystone citation context sentences	Annotation from [6][a]	Classifier results
(1) We took advantage of a mouse line in which expression of a tet transactivator transgene is under control of the neuropsin gene promoter (Yasuda and Mayford, 2006).	Material	No hit
(2) This line was crossed with the Tg(tetO tauP301L)4510 line that only expresses human tau carrying the P301L frontotemporal dementia mutation in the presence of a tet transactivator (Santacruz et al., 2005).	Material	No hit
(3) Immunohistochemistry using the 5A6 antibody (courtesy of Dr.G.V. Johnson, University of Rochester), a monoclonal antibody raised against the longest form of recombinant human tau which recognizes an epitope between amino acids 19 and 46 (Johnson et al., 1997), confirmed strong expression of tau protein in superficial layers of the MEC and parasubiculum in rTgTauEC mice at 3 months of age compared to a control brain (Figure 1D).	Material	Results
(4) In AD, early hallmarks include the loss of synapses, and comparison of AD patients to age-matched control individuals showed that the density of synapses correlated strongly with cognitive impairment, suggesting that loss of connections is associated with the progression of the disease (DeKosky and Scheff, 1990; Scheff and Price, 2006; Terry et al., 1991).	Methods	No hit
(5) Therefore, we assessed two synaptic markers in the perforant pathway terminal zone of rTgTauEC mice: synapsin-I, a marker of synaptic vesicles, and PSD-95, a postsynaptic marker that has been reported to decrease early in neurodegeneration (Zhao et al., 2006).	Material	Methods
(6) The evaluation of Boltzmann-averaged 13C and 1H magnetic shielding tensors and isotropic chemical shifts from density functional theory (DFT) followed Hoye's protocol[25] adapted as follows.	Methods	Methods
(7) Therefore, we turned to a protocol that relies on density functional theory-based computations of 1H and 13C NMR chemical shifts and the use of statistical tools to assign the experimental data to the correct isomer of a compound[28].	Methods	Methods
(8) The applied procedure is in principle analogous to the one described by Willoughby[43], with slight modifications and different software packages used.	Methods	No hit
(9) To resolve this ambiguity, we conducted NMR prediction calculations (Figure 1 B)[13,14].	Methods	Methods

[a] S1-S5 are from Ref 14 of [6], S6 is from Ref 40 of [6], S7 is from Ref 33 of [6], S8 is from Ref 28 of [6], and S9 is from Ref 17 of [6].

Table 3: Classification results of the Willoughby-Hoye dataset and keystone citation annotation

Class	No. of instances where Willoughby-Hoye protocol is a methods keystone citation	Total No. of Instances	Percentage
Background	1	10	10%
Objective	0	0	-
Methods	21	22	95%
Results	10	10	100%
Conclusions	0	2	0%
No hit	-	56	-
Total[a]	31	99	-

[a] One instance was classified as both Methods and Results, and therefore the total number is 99, not 100.

citations to be classified as Background sentences, we found one: A sentence that described a method in a non-characteristic way ("The entire process begins with DFT prediction. . ."). Likewise, in the two Conclusion sentences, the protocol played an auxiliary role (i.e., reinforcing or contrasting the findings), and neither was a methods keystone citation. And since no Objective sentence were captured, whether citations contained in Objective sentences can be methods keystone citations remains unknown.

This exploratory study resulted in revising our strategy for detecting keystone citations. Our revised strategy, depicted in Figure 2, considers citations contained in Methods or Results sentences to have a high likelihood of being methods keystone citations,

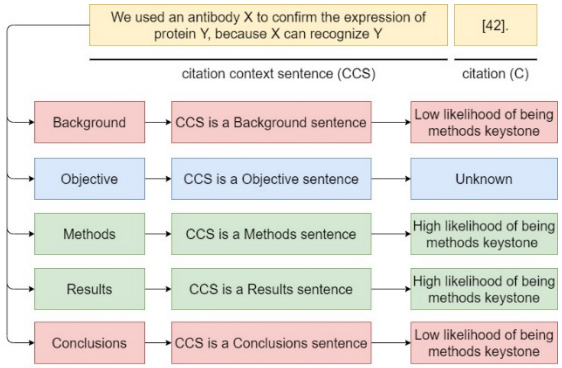

Figure 2: A revised strategy based on two tests

while sentences classified as Background or Conclusions have a low likelihood of containing methods keystone citations. Still, the Willoughby-Hoye dataset is small, and the revised strategy needs to be verified using more data.

Ultimately, a sizable gold-standard keystone citation context dataset is needed, and the rhetorical category classifiers may serve as a useful screening tool for constructing such a dataset. Methods and Results can be quickly scanned to verify that they contain keystone citations; Background and Conclusions sentences can be quickly scanned to ensure that they do not contain keystone citations. Most attention can then be focused on sentences that do not receive a classification.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed and evaluated a strategy that repurposed rhetorical category classifiers for the novel application of extracting keystone citations that relate to research methods. Five binary rhetorical category classifiers were constructed to identify Background, Objective, Methods, Results, and Conclusions sentences in biomedical papers. The resulting classifiers were evaluated using two datasets. The initial strategy assumed that only citations contained in Methods sentences were methods keystone citations, but our analysis revealed that citations contained in sentences classified as either Methods or Results had a high likelihood to be methods keystone citations. Future work will focus on fine-tuning the rhetorical category classifiers, experimenting with multiclass classifiers, evaluating the revised strategy with more data, and constructing a larger gold-standard citation context sentences dataset for model training.

ACKNOWLEDGMENTS

The authors thank Prof. Halil Kilicoglu for providing the training dataset. Jodi Schneider' work was supported by Alfred P. Sloan Foundation G-2020-12623.

REFERENCES

[1] Jayanti Bhandari Neupane, Ram P. Neupane, Yuheng Luo, Wesley Y. Yoshida, Rui Sun, and Philip G. Williams. 2019. Characterization of Leptazolines A–D, polar oxazolines from the cyanobacterium *Leptolyngbya* sp., reveals a glitch with the "Willoughby–Hoye" scripts for calculating NMR chemical shifts. *Org. Lett.* 21, 20 (October 2019), 8449–8453. DOI:https://doi.org/10.1021/acs.orglett.9b03216

[2] Christian Bölling, Michael Weidlich, and Hermann-Georg Holzhütter. 2014. SEE: structured representation of scientific evidence in the biomedical domain using Semantic Web techniques. *J. Biomed. Semant.* 5, 1 (June 2014), S1. DOI:https://doi.org/10.1186/2041-1480-5-S1-S1

[3] Matthew H. Brush, Kent Shefchek, and Melissa Haendel. 2016. SEPIO: A semantic model for the integration and analysis of scientific evidence. In *Proceedings of the Joint International Conference on Biological Ontology and BioCreative*. Retrieved from http://ceur-ws.org/Vol-1747/IT605_ICBO2016.pdf

[4] Tim Clark, Paolo N Ciccarese, and Carole A Goble. 2014. Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications. *J. Biomed. Semant.* 5, 1 (2014), 28. DOI:https://doi.org/10.1186/2041-1480-5-28

[5] K. Bretonnel Cohen, Helen L. Johnson, Karin Verspoor, Christophe Roeder, and Lawrence E. Hunter. 2010. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinform* 11, 1 (September 2010), 492. DOI:https://doi.org/10.1186/1471-2105-11-492

[6] Yuanxi Fu and Jodi Schneider. 2020. Towards knowledge maintenance in scientific digital libraries with the keystone framework. In *2020 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, IEEE, Virtual Event, China, 217–226. DOI:https://doi.org/10.1145/3383583.3398514

[7] 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* (BioNLP '10), Association for Computational Linguistics, USA, 99–107. https://dl.acm.org/doi/10.5555/1869961.1869974

[8] Saeed-Ul Hassan, Anam Akram, and Peter Haddawy. 2017. Identifying important citations using contextual information from full text. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, IEEE, Toronto, ON, Canada, 1–8. DOI:https://doi.org/10.1109/JCDL.2017.7991558

[9] Katherine W. McCain. 2011. Eponymy and Obliteration by Incorporation: The case of the "Nash Equilibrium." *J. Am. Soc. Inf. Sci. Technol.* 62, 7 (2011), 1412–1424. DOI:https://doi.org/10.1002/asi.21536

[10] Robert King Merton. 1965. *On the shoulders of giants; a Shandean postscript*. Free Press, New York.

[11] Suchetha N. Kunnath, David Pride, Bikash Gyawali, and Knoth Petr. 2020. Overview of the 2020 WOSP 3C Citation Context Classification Task. In *Proceedings of The 8th International Workshop on Mining Scientific Publications*, Wuhan, China. Retrieved from https://www.aclweb.org/anthology/2020.wosp-1.12.pdf

[12] Sejin Nam, Senator Jeong, Sang-Kyun Kim, Hong-Gee Kim, Victoria Ngo, and Nansu Zong. 2016. Structuralizing biomedical abstracts with discriminative linguistic features. *Comput. Biol. Med.* 79, (December 2016), 276–285. DOI:https://doi.org/10.1016/j.compbiomed.2016.10.026

[13] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Gärdenfors, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, (2011), 2825–2830. https://dl.acm.org/doi/10.5555/1953048.2078195

[14] Leslie A Pray. 2008. Discovery of DNA Double Helix: Watson and Crick. *Nat. Educ.* 1, 1 (2008), 100.

[15] Henry Small. 2018. Characterizing highly cited method and non-method papers using citation contexts: The role of uncertainty. *J. Informetr.* 12, 2 (May 2018), 461–480. DOI:https://doi.org/10.1016/j.joi.2018.03.007

[16] Henry Small, Kevin W. Boyack, and Richard Klavans. 2019. Citations and certainty: a new interpretation of citation counts. *Scientometrics* 118, 3 (March 2019), 1079–1092. DOI:https://doi.org/10.1007/s11192-019-03016-z

[17] Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *Scholarly Big Data: AI Perspectives, Challenges, and Ideas: Papers from the 2015 AAAI Workshop*, 21–26. Retrieved December 7, 2019 from https://www.aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10185

[18] Patrick H. Willoughby, Matthew J. Jansma, and Thomas R. Hoye. 2014. A guide to small-molecule structure assignment through computation of (1 H and 13 C) NMR chemical shifts. *Nat. Protoc.* 9, 3 (March 2014), 643–660. DOI:https://doi.org/10.1038/nprot.2014.042

[19] Yasunori Yamamoto and Toshihisa Takagi. 2005. A Sentence Classification System for Multi Biomedical Literature Summarization. In *21st International Conference on Data Engineering Workshops (ICDEW'05)*, 1163–1163. DOI:https://doi.org/10.1109/ICDE.2005.170

[20] Yiming Yang. 1999. An Evaluation of Statistical Approaches to Text Categorization. *Inf. Retr.* 1, 1 (April 1999), 69–90. DOI:https://doi.org/10.1023/A:1009982220290

[21] Yiming Yang and J. Pedersen. 1997. A Comparative Study on Feature Selection in Text Categorization. In *ICML*. https://dl.acm.org/doi/10.5555/645526.657137