Pattern-Based Acquisition of Scientific Entities from Scholarly Article Titles

Jennifer D'Souza[1(✉)] and Sören Auer[1,2]

[1] TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
{jennifer.dsouza,auer}@tib.eu
[2] L3S Research Center at Leibniz University of Hannover, Hannover, Germany

Abstract. We describe a rule-based approach for the automatic acquisition of salient scientific entities from Computational Linguistics (CL) scholarly article titles. Two observations motivated the approach: (i) noting salient aspects of an article's contribution in its title; and (ii) pattern regularities capturing the salient terms that could be expressed in a set of rules. Only those lexico-syntactic patterns were selected that were easily recognizable, occurred frequently, and positionally indicated a scientific entity type. The rules were developed on a collection of 50,237 CL titles covering all articles in the ACL Anthology. In total, 19,799 *research problems*, 18,111 *solutions*, 20,033 *resources*, 1,059 *languages*, 6,878 *tools*, and 21,687 *methods* were extracted at an average precision of 75%.

Keywords: Terminology extraction · Rule-based system · Natural language processing · Scholarly knowledge graphs · Semantic publishing

1 Introduction

Scientists increasingly face the information overload-and-drown problem even in narrow research fields given the ever-increasing flood of scientific publications [19,21]. Recently, solutions are being implemented in the domain of the digital libraries by transforming scholarly articles into "digital-first" applications as machine-interpretable scholarly knowledge graphs (SKGs), thus enabling completely new technological assistance to navigate the massive volumes of data through intelligent search and filter functions, and the integration of diverse analytics tools. There are several directions to this vision focused on representing, managing and linking metadata about articles, people, data and other relevant keyword-centered entities (e.g., Research Graph [3], Scholix [7], SpringerNature's SciGraph or DataCite's PID Graph [9], SemanticScholar [1]). This trend tells us that we are on the cusp of a great change in the digital technology applied to scholarly knowledge. Notably, next-generation scholarly digital library (DL) infrastructures have arrived: the Open Research Knowledge Graph (ORKG) [18]

Supported by TIB Leibniz Information Centre for Science and Technology, the EU H2020 ERC project ScienceGRaph (GA ID: 819536).

H.-R. Ke et al. (Eds.): ICADL 2021, LNCS 13133, pp. 401–410, 2021.
https://doi.org/10.1007/978-3-030-91669-5_31

digital research and innovation infrastructure by TIB and partner institutions, argues for obtaining a semantically rich, interlinked KG representations of the "content" of the scholarly articles, and, specifically, only *research contributions*.[1] With intelligent analytics enabled over such contributions-focused SKGs, researchers can readily track research progress without the cognitive overhead that reading dozens of articles impose. A typical dilemma then with building such an SKG is deciding the type of information to be represented. In other words, what would be the information constituent candidates for an SKG that reflects the overview? While the scope of this question is vast, in this paper, we describe our approach designed with this question as the objective.

"*Surprisingly useful information can be found with only a very simply understanding of the text.*" [14] The quotation is the premise of the "Hearst" system of patterns which is a popular text mining method in the CL field. It implemented discovering lexical relations from a large-scale corpus simply by looking for the relations expressed in well-known ways. This simple but effective strategy was leveraged in supporting the building up of large lexicons for natural language processing [15], e.g., the WordNet lexical project [24]. Our approach is inspired after the "Hearst" methodology but on scholarly article titles content thereby implementing a pattern-based acquisition of scientific entities. Consider the two paper title examples depicted in Table 1. More fluent readers of English can phrase-chunk the titles based on lexico-syntactic patterns such as the colon punctuation in title 1 and prepositional phrase boundary markers (e.g., 'to' in title 2). Following which, with some domain awareness, the terms can be semantically conceptualized or typed (e.g., as *research problem*, *resource*, *method*, *tool*, etc.). Based on such observations and circling back to the overarching objective of this work, we propose and implement a pattern-based acquisition approach to mine contribution-focused, i.e. salient, scientific entities from article titles. While there is no fixed notion of titles written with the purpose of reflecting an article's contribution, however, this is the generally known practice that it contains salient aspects related to the *contribution* as a single-line summary. To the best of our knowledge, a corpus of only article titles remains as yet comprehensively unexplored as a resource for SKG building. Thus, our work sheds a unique and novel light on SKG construction representing *research overviews*.

In this paper, we discuss CL-Titles-Parser – a tool for extracting salient scientific entities based on a set of lexico-syntactic patterns from titles in Computational Linguistics (CL) articles. Six concept types of entities were identified applicable in CL titles, viz. *research problem*, *solution*, *resource*, *language*, *tool*, and *method*. CL-Titles-Parser when evaluated on almost all titles (50,237 of 60,621 total titles) in the ACL Anthology performs at a cumulative average of 75% IE precision for the six concepts. Thus, its resulting high-precision SKG integrated in the ORKG can become a reliable and essential part of the scientist's workbench in visualizing the overview of a field or even as crowdsourcing signals for authors to describe their papers further. CL-Titles-Parser is released as a standalone program https://github.com/jd-coderepos/cl-titles-parser.

[1] The ORKG platform can be accessed online: https://orkg.org/.

Table 1. Two examples of scholarly article titles with their concept-typed scientific terms which constitutes the IE objective of the CL-Titles-Parser

SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News
research_problem: ['SemEval-2017 Task 5']
resource: ['Financial Microblogs and News']
method: ['Fine-Grained Sentiment Analysis']
Adding Pronunciation Information to Wordnets
solution: ['Adding Pronunciation Information']
tool: ['Wordnets']

2 Related Work

Key Summary of Research in Phrasal Granularity. To bolster search technology, the phrasal granularity was used to structure the scholarly record. Thus scientific phrase-based entity annotated datasets in various domains including multidisciplinarily across STEM [4,10,13,22] were released; machine learning systems were also developed for automatic scientific entity extraction [2,5,6,23]. However, none of these resources are clearly indicative of capturing only the salient terms about research contributions which is the aim of our work.

Pattern-Based Scientific Terminology Extraction. Some systems [16] viewed key scholarly information candidates as problem-solution mentions. [8] used the discourse markers "thus, therefore, then, hence" as signals of problem-solution patterns. [12] used semantic extraction patterns learned via bootstrapping to the dependency trees of sentences in Abstracts to mine the research focus, methods used, and domain problems. Houngbo and Mercer [17] extracted the methods and techniques from biomedical papers by leveraging regular expressions for phrase suffixes as "algorithm," "technique," "analysis," "approach," and "method." AppTechMiner [26] used rules to extract application areas and problem solving techniques. The notion of application areas in their model is analogous to research problem in ours, and their techniques are our tool or method. Further, their system extracts research problems from the article titles via rules based on functional keywords, such as, "for," "via," "using" and "with" that act as delimiters for such phrases. CL-Titles-Parser also extracts problems from titles but it does so in conjunction with other information types such as tools or methods. AppTechMiner uses citation information to determine term saliency. In contrast, since we parse titles, our data source itself is indicative of the saliency of the scientific terms therein w.r.t. the article's contribution. Finally, [20], like us, use a system of patterns to extract methods from the titles and Abstracts of articles in Library Science research. We differ in that we extract six different types of scientific entities and we focus only on the article titles data source.

Next, in the article, we describe the CL-Titles-Parser for its pattern-based acquisition of scientific entities from Computational Linguistics article titles.

3 Preliminary Definitions

We define the six scientific concept types handled in this work. The main aim here is not to provide rigorous definitions, but rather just to outline essential features of the concepts to explain the hypotheses concerning their annotation.

i. Research problem. The theme of the investigation. E.g., "Natural language inference." In other words, the answer to the question "which problem does the paper address?" or "On what topic is the investigation?" ***ii. Resource.*** Names of existing data and other references to utilities like the Web, Encyclopedia, etc., used to address the *research problem* or used in the *solution*. E.g., "Using Encyclopedic Knowledge for Automatic Topic Identification." In this sentence, "Encyclopedic Knowledge" is a *resource* used for *research problem* "Automatic Topic Identification." ***iii. Tool.*** Entities arrived at by asking the question "Using what?" or "By which means?" A *tool* can be seen as a type of a *resource* and specifically software. ***iv. Solution.*** A novel contribution of a work that solves the *research problem*. E.g., from the title "PHINC: A Parallel Hinglish Social Media Code-Mixed Corpus for Machine Translation," the terms 'PHINC' and 'A Parallel Hinglish Social Media Code-Mixed Corpus' are solutions for the *problem* 'Machine Translation.' ***v. Language.*** The natural language focus of a work. E.g., Breton, Erzya, Lakota, etc. *Language* is a pertinent concept w.r.t. an overview SKG about NLP solutions. ***vi. Method.*** They refer to existing protocols used to support the *solution*; found by asking "How?"

4 Tool Description

4.1 Formalism

Every CL title T can be expressed as one or more of the following six elements $te_i = \langle rp_i, res_i, tool_i, lang_i, sol_i, meth_i \rangle$, representing the *research problem*, *resource*, *tool*, *language*, *solution*, and *method* concepts, respectively. A title can contain terms for zero or more of any of the concepts. The goal of CL-Titles-Parser is, for every title t_i, to annotate its title expression te_i, involving scientific term extraction and term concept typing.

4.2 Rule-Based Processing Workflow

CL-Titles-Parser operates in a two-step workflow. First, it aggregates titles as eight main template types with a default ninth category for titles that could not be clustered by any of the eight templates. Second, the titles are phrase-chunked and concept-typed based on specific lexico-syntactic patterns that are group-specific. The concept type is selected based on the template type category and some contextual information surrounding the terms such as prepositional and verb phrase boundary markers.

Step 1: Titles clustering based on commonly shared title lexico-syntactic patterns. While our rule-based system implements eight patterns in total, we describe four template patterns as examples.

Template "hasSpecialCaseWord():" applies to titles written in two parts – a one-word solution name, a colon separator, and an elaboration of the solution name. E.g., "SNOPAR: A Grammar Testing System" consisting of the one word 'SNOPAR' solution name and its elaboration 'A Grammar Testing System.' Further, there are other instances of titles belonging to this template type that are complex sentences, i.e. titles with additional prepositional or verb phrases, where mentions of the *research problem*, *tool*, *method*, *language* domain etc. are also included in the latter part of the title. E.g., "GRAFON: A Grapheme-to-Phoneme Conversion System for Dutch" is a complex title with a prepositional phrase triggered by "for" specifying the *language* domain "Dutch."

Template "Using ..." applies to titles that begin with the word "Using" followed by a *resource* or *tool* or *method* and used for the purpose of a *research problem* or a *solution*. E.g., the title "Using WordNet for Building WordNets" with *resource* "WordNet" for *solution* "Building WordNets"; or "Using Multiple Knowledge Sources for Word Sense Discrimination" with *resource* "Multiple Knowledge Sources" for *research problem* "Word Sense Discrimination."

Template "... case study ..." Titles in this category entail splitting the phrase on either side of "case study." The first part is processed by the precedence-ordered rules to determine the concept type. The second part, however, is directly cast as *research problem* or *language* since they were observed as one of the two. The checks for *research problem* or *language* are made by means of regular expressions implemented in helper functions. E.g., the title "Finite-state Description of Semitic Morphology: A Case Study of Ancient Accadian" would be split as "Finite-state Description of Semitic Morphology" and "Ancient Accadian" language domain, where "Ancient Accadian" is typed as *language* based on regex patterns. See Table 2 for examples of some regular expressions employed.

Table 2. Regular expressions in suffix patterns for scientific term concept typing

languages	$reLanguage = (...\|Tigrigna\|Sundanese\|Balinese\|...)$
tool	$reTool = (...\|memory\|controller\|workbench(es)?\|...)$
resource	$reResource = (...\|corp(ora\|us)\|vocabular(ies\|y)\|cloud\|...)$
method	$reMethod = (...\|protocol\|methodolog(ies\|y)\|recipe\|...)$

Template "... : ..." A specialized version of this template is "hasSpecialCaseWord():". Here, titles with two or more words in the phrase preceding the colon are considered. They are split in two parts around the colon. The parts are then further processed to extract the scientific terms. E.g., "Working on the Italian Machine Dictionary: A Semantic Approach" split as "Working on the Italian Machine Dictionary" and "A Semantic Approach" where the second part is a non-scientific information content phrase. By non-scientific information content phrase we mean phrases that cannot be categorized as one of the six concepts.

Step 2: Precedence-ordered scientific term extraction and typing rules. The step is conceptually akin to sieve-based systems that were successfully demonstrated on the coreference resolution [25] and biomedical name normalization [11] tasks. The idea in sieve-based systems is simply that an ordering is imposed on a set of selection functions from the most constrained to the least. Similarly, in this second step of processing titles in our rule-based system, we apply the notion of selection precedence as concept precedence. However, there are various concept precedences in our system that depend on context information seen as the count of the connectors and the type of connectors in any given article title.

In this step, within each template category the titles are generally processed as follows. **Step 1. Counting connectors** – Our connectors are a collection of 11 prepositions and 1 verb defined as: $connectors\_rx = (to|of|on|for|from|with|by|via|through|using|in|as)$. For a given title, its connectors are counted and the titles are phrase chunked as scientific phrase candidates split on the connectors themselves. **Step 2. Concept typing** – This involves selecting the workflow for typing the scientific phrases with concept types among our six, viz. *language*, *tool*, *method*, *resource*, and *research problem*, based on context information. Workflow branches were implemented as a specialized system of rules based on the number of connectors. The next natural question is: after the workflow branch is determined, what are the implementation specifics for typing the scientific terms per our six concepts? We explain this with the following specific case. A phrase with 0 connectors is typed after the following concept precedence order: $language \prec tool \prec method \prec resource \prec research\ problem$ where each of the concepts are implemented as regex checks. Some example regexes were shown earlier in Table 2. And it only applies to five of the six concepts, i.e. *solution* is omitted. On the other hand, if a title has one connector, it enters first into the OneConnectorHeu() branch. There, the first step is determining which connector is in the phrase. Then based on the connector, separate sets of concept type precedence rules apply. The concept typing precedence rules are tailored based on the connector context. For instance, if the connector is 'from,' the title subphrases are typed based on the following pattern: *solution* from *resource*.

This concludes a brief description of the working of CL-Titles-Parser.

5 Evaluation

In this section, some results from CL-Titles-Parser are discussed for scientific term extraction and concept typing in Computational Linguistics article titles.

Evaluation Corpus. We downloaded all the article titles in the ACL anthology as the 'Full Anthology as BibTeX' file dated 1-02-2021. See https://aclanthology.org/anthology.bib.gz. From a total of 60,621 titles, the evaluation corpus comprised 50,237 titles after eliminating duplicates and invalid titles.

When applied to the evaluation corpus, the following total scientific concepts were extracted by the tool: 19,799 *research problem*, 18,111 *solution*, 20,033

resource, 1,059 *language*, 6,878 *tool*, and 21,687 *method*. These scientific concept lists were then evaluated for extraction precision.

5.1 Quantitative Analysis: Scientific Concept Extraction Precision

First, each of the six scientific concept lists were manually curated by a human annotator to create the gold-standard data. The extracted lists and the gold-standard lists were then evaluated w.r.t. the *precision* metric. Table 3 shows the results. We see that CL-Titles-Parser demonstrates a high information extraction precision for all concept types except *research problem*. This can be attributed in part to the long-tailed list phenomenon prevalent in the scientific community as the scholarly knowledge investigations steadily progress. With this in mind, the gold-standard list curation was biased toward already familiar research problems or their derivations. Thus we estimated that at least 20% of the terms were pruned in the gold data because they were relatively new as opposed to being incorrect. Note, recall evaluations were not possible as there is no closed-class gold standard as scientific terms are continuously introduced.

5.2 Qualitative Analysis: Top N Terms

As qualitative analysis, we examine whether the terms extracted by our tool reflect popular research trends. Table 4 shows the top five terms in each of the six concept types. The full scientific concept lists sorted by occurrences are available in our code repository https://github.com/jd-coderepos/cl-titles-parser/tree/master/data-analysis. Considering the *research problem* concept, we see that variants of "machine translation" surfaced to the top accurately reflective of the large NLP subcommunity attempting this problem. As a *tool*, "Word Embeddings" are the most predominantly used. "Machine Translation" itself was the most employed *method*. Note that the concept types are not mutually exclusive. A term that is a *research problem* in one context can be a *method* in a different context. As an expected result, "English" is the predominant *language* researched. "Twitter" showed as the most frequently used *resource*. Finally, predominant *solutions* reflected the nature of the article itself as "overview" or "a study" etc. Then Table 5 shows the *research problem*, *resource*, and *tool* concepts research trends in the 20th vs. 21st century. Contemporarily, we see new predominant neural *research problem* mentions, an increasing use of social media as a *resource*; and various neural networks as *tools*.

Table 3. Precision of CL-Titles-Parser for scientific term extraction and concept typing from 50,237 titles in the ACL anthology

Concept type	*Precision*	Concept type	*Precision*
research problem	58.09%	*method*	77.29%
solution	80.77%	*language*	95.12%
tool	83.40%	*resource*	86.96%

Table 4. Top 5 scientific phrases for the six concepts extracted by CL-Titles-Parser

research-problem	statistical machine translation (267), machine translation (266), neural machine translation (193), sentiment analysis (99), information extraction (85)
tool	word embeddings (77), neural networks (63), conditional random fields (51), convolutional neural networks (41), spoken dialogue systems (32)
method	machine translation (105), domain adaptation (68), sentiment analysis (68), named entity recognition (67), statistical machine translation (66)
language	English (150), Chinese (87), Japanese (87), German (81), Arabic (74)
resource	Twitter (204), text (173), social media (132), the web (115), Wikipedia (98)
solution	overview (39), a study (23), an empirical study (25), a comparison (21), a toolkit (17)

Table 5. Top 5 *research problem*, *resource*, and *tool* phrases from paper titles reflecting research trends in the 20th (7,468 titles) vs. the 21st (63,863 titles) centuries.

	research problem	*resource*	*tool*
20th	machine translation (56)	text (38)	machine translation system (8)
21st	statistical machine translation (258)	text (251)	word embeddings (87)
20th	information extraction (19)	discourse (17)	natural language interfaces (7)
21st	machine translation (210)	Twitter (204)	neural networks (57)
20th	speech recognition (16)	TAGs (9)	neural networks (6)
21st	neural machine translation (193)	social media (132)	conditional random fields (51)
20th	natural language generation (15)	bilingual corpora (9)	WordNet (3)
21st	sentiment analysis (99)	the web (115)	convolutional neural networks (41)
20th	continuous speech recognition (12)	dialogues (9)	semantic networks (3)
21st	question answering (81)	Wikipedia (98)	spoken dialogue systems (31)

6 Conclusion and Future Directions

We have described a low-cost approach for automatic acquisition of contribution-focused scientific terms from unstructured scholarly text, specifically from Computational Linguistics article titles. Work to extend the tool to parse Computer Science titles at large is currently underway. The absence of inter-annotator agreement scores to determine the reliability with which the concepts can be selected will also be addressed in future work. Evaluations on the ACL anthology titles shows that our rules operate at a high precision for extracting *research problem*, *solution*, *resource*, *language*, *tool*, and *method*. We proposed an incremental step toward the larger goal of generating contributions-focused SKGs.

References

1. Ammar, W., et al.: Construction of the literature graph in semantic scholar. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Papers), vol. 3, pp. 84–91 (2018)
2. Ammar, W., Peters, M.E., Bhagavatula, C., Power, R.: The AI2 system at SemeEal-2017 task 10 (ScienceIE): semi-supervised end-to-end entity and relation extraction. In: SemEval@ACL (2017)
3. Aryani, A., et al.: A research graph dataset for connecting research data repositories using RD-switchboard. Sci. Data **5**(1), 1–9 (2018)
4. Augenstein, I., Das, M., Riedel, S., Vikraman, L., McCallum, A.: SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In: SemEval@ACL (2017)
5. Beltagy, I., Lo, K., Cohan, A.: SciBERT: pretrained language model for scientific text. In: EMNLP (2019)
6. Brack, A., D'Souza, J., Hoppe, A., Auer, S., Ewerth, R.: Domain-independent extraction of scientific concepts from research articles. In: Jose, J.M., et al. (eds.) ECIR 2020. LNCS, vol. 12035, pp. 251–266. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45439-5_17
7. Burton, A., et al.: The Scholix framework for interoperability in data-literature information exchange. D-Lib Mag. **23**(1/2) (2017)
8. Charles, M.: Adverbials of result: phraseology and functions in the problem-solution pattern. J. Engl. Acad. Purp. **10**(1), 47–60 (2011)
9. Cousijn, H., et al.: Connected research: the potential of the PID graph. Patterns **2**(1), 100180 (2021)
10. D'Souza, J., Hoppe, A., Brack, A., Jaradeh, M.Y., Auer, S., Ewerth, R.: The STEM-ECR dataset: grounding scientific entity references in stem scholarly content to authoritative encyclopedic and lexicographic sources. In: LREC, Marseille, France, pp. 2192–2203, May 2020
11. D'Souza, J., Ng, V.: Sieve-based entity linking for the biomedical domain. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 297–302 (2015)
12. Gupta, S., Manning, C.D.: Analyzing the dynamics of research by extracting key aspects of scientific papers. In: Proceedings of 5th International Joint Conference on Natural Language Processing, pp. 1–9 (2011)
13. Handschuh, S., QasemiZadeh, B.: The ACL RD-TEC: a dataset for benchmarking terminology extraction and classification in computational linguistics. In: COLING 2014: 4th International Workshop on Computational Terminology (2014)
14. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Coling 1992 Volume 2: The 15th International Conference on Computational Linguistics (1992)
15. Hearst, M.A.: Automated discovery of wordnet relations. WordNet: An Electronic Lexical Database, vol. 2 (1998)
16. Heffernan, K., Teufel, S.: Identifying problems and solutions in scientific text. Scientometrics **116**(2), 1367–1382 (2018)
17. Houngbo, H., Mercer, R.E.: Method mention extraction from scientific research papers. In: Proceedings of COLING 2012, pp. 1211–1222 (2012)

18. Jaradeh, M.Y., et al.: Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge. In: Proceedings of the 10th International Conference on Knowledge Capture, K-CAP 2019, pp. 243–246. Association for Computing Machinery, New York (2019). https://doi.org/10.1145/3360901.3364435
19. Johnson, R., Watkinson, A., Mabe, M.: The STM Report. An Overview of Scientific and Scholarly Publishing. 5th edn., October 2018. https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf
20. Katsurai, M., Joo, S.: Adoption of data mining methods in the discipline of library and information science. J. Libr. Inf. Stud. **19**(1), 1–17 (2021)
21. Landhuis, E.: Scientific literature: information overload. Nature **535**(7612), 457–458 (2016)
22. Luan, Y., He, L., Ostendorf, M., Hajishirzi, H.: Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In: EMNLP (2018)
23. Luan, Y., Ostendorf, M., Hajishirzi, H.: Scientific information extraction with semi-supervised neural tagging. arXiv preprint arXiv:1708.06075 (2017)
24. Miller, G.A.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
25. Raghunathan, K., et al.: A multi-pass sieve for coreference resolution. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 492–501 (2010)
26. Singh, M., Dan, S., Agarwal, S., Goyal, P., Mukherjee, A.: AppTechMiner: mining applications and techniques from scientific articles. In: Proceedings of the 6th International Workshop on Mining Scientific Publications, pp. 1–8 (2017)