Post-print of a Study on Linguistic Features and Distribution Patterns of Research Highlights in Academic Papers

Authors: Suo Chuanjun, Yu Guoxin

Date: 2023-04-01T16:15:53+00:00

Abstract

[Purpose/Significance] During the publication process of academic papers, the reasonable and effective presentation of core viewpoints can not only substantially reduce the time researchers expend on literature search and screening, but also facilitate reading and comprehension. [Method/Process] A research corpus was constructed by annotating 385 XML-format journal papers. Keyword analysis was subsequently employed to analyze the linguistic features of highlights, and natural language processing algorithms were utilized to explore their distribution characteristics. [Results/Conclusion] Highlights constitute a collection of standardized short sentences with explicit semantics, representing the manifestation of novel viewpoints, perspectives, methods, ideas, results, and conclusions that distinguish an academic paper from others. Highlights exhibit characteristics of novelty, conciseness, readability, and "promotional" nature. They can be classified into research innovation highlights, research method highlights, research process highlights, and research conclusion highlights. This study reveals the distribution patterns of highlights within the main text and across different sections.

Full Text

Abstract

[Purpose/Significance] In the process of academic publishing, effectively presenting the core viewpoints of a research paper can not only significantly reduce the time researchers spend searching and screening literature but also facilitate reading and comprehension. **[Method/Process]** By annotating 385 XML-formatted journal papers, we constructed a research corpus, analyzed the linguistic features of highlights using keyword analysis, and explored their distribution characteristics through natural language processing algorithms. **[Re-**

sults/Conclusions] Highlights constitute a collection of standardized, semantically explicit short sentences that embody a paper's novel viewpoints, perspectives, methods, approaches, results, and conclusions compared to other works. Highlights exhibit characteristics of novelty, conciseness, readability, and "promotional" quality. We classify highlights into four types: research innovation highlights, research method highlights, research process highlights, and research conclusion highlights, and identify their distribution patterns within the main text and various sections.

Keywords: academic papers; research highlights; highlight value; linguistic characteristics; distribution features

2. Significance and Value of Automatic Highlight Extraction

2.1 Significance of Automatic Highlight Extraction

Typically, the significance and value of an academic paper can only be assessed after reading it. However, the sheer volume of papers means readers lack sufficient time to read and select them, often causing them to miss valuable literature. For years, the field of knowledge engineering has studied "automatic abstract" generation using computer natural language processing techniques, but without satisfactory results. To address this contradiction, Elsevier proposed that authors annotate research highlights. While manual annotation of research highlights achieves high accuracy, it is costly and inefficient, making it unable to meet the needs of annotating highlights for the massive backlog of existing papers. Therefore, by exploring the linguistic features of highlights and their distribution patterns within papers, we can achieve automatic identification and extraction of paper highlights. Marking the research highlights of each paper can both save readers' time and promote rapid dissemination, holding substantial significance and value.

2.2 Definition of Highlights

From a semantic content perspective, highlights belong to academic papers; from a formal perspective, they are text. Academic papers contain abstracts, main text, references, etc., while text includes long and short forms. In comparison, highlights are similar to abstracts in content but more novel and concise; they resemble short texts in form but are more brief, standardized, and semantically explicit. Based on this analysis, highlights are defined as a collection of standardized, semantically explicit short sentences with five key implications: (1) highlights must conform to grammatical norms with complete semantics; (2) highlights should be as brief as possible while expressing content adequately; (3) highlights explain the novelty of a specific aspect of an academic paper; (4) highlights express important content and reflect a paper's unique characteristics; (5) highlights provide readers with an overview of the paper's innovative content.

2.3 Characteristics of Highlights

Our analysis reveals that highlights possess four main characteristics: novelty, conciseness, readability, and "promotional" quality. Novelty is the most fundamental characteristic, as highlights embody a paper's new viewpoints, perspectives, methods, approaches, results, and conclusions. Highlight content must be independently created by the author and represent an advancement over existing research, whether through improvement, correction, or disruption. Conciseness is crucial since highlights appear before the abstract. Elsevier stipulates that each highlight should not exceed 85 characters, requiring authors to express important content using minimal characters without compromising comprehension. Readability and promotional quality enable highlights to attract potential readers. Readability manifests in accessible expression that differs from cold listings of formulas and data—highlights are vivid representations of academic viewpoints. The promotional quality is reflected in authors' frequent use of adverbs or adjectives as "intensifiers" to strengthen their arguments when writing highlights.

2.4 Value of Highlights

Research highlights serve important functions for editors, readers, publishers, and authors. First, they facilitate more efficient peer review and academic supervision. Highlights help editors and reviewers make preliminary judgments about a paper's academic value, accelerating the review process and improving efficiency. Since experts often struggle to grasp and extract core viewpoints when authors present them unclearly, having authors explicitly mark core research points can expedite preliminary assessments of originality, innovation, and value. Second, they enhance publishers' competitiveness and create added value. Since the "publish or perish" doctrine emerged, academic publishing has become highly competitive, with journals striving to attract authors and expand readership. The conciseness and convenience of highlights make them a powerful tool for improving competitiveness and attracting potential readers to purchase full access. Third, they help readers evaluate paper value and improve reading efficiency. While papers aim to enable readers to learn from and build upon innovative 成果, understanding core viewpoints requires reading extensive text. Highlights significantly save time and effort while increasing paper "visibility," helping readers select papers faster. Fourth, they help authors publicize their work and disseminate academic viewpoints. Directly presenting main findings and 成果 helps readers choose papers, accelerates dissemination, and enhances authors' academic influence.

3. Research Corpus Construction

3.1 Data Sources

We obtained 385 full-text papers from *International Journal of Information Management* published between 2016 and 2018 through Elsevier's ScienceDi-

rect database. Using Oxygene XML Editor software, we annotated the papers according to specific rules to construct our research data source.

3.2 XML Text Markup Rules

Based on the structural characteristics of complete journal papers and the extensible features of XML markup language, we created custom markup rules (see) that include bibliographic information, highlights, abstracts, keywords, and main text, enabling complete markup of an academic paper. Authors, keywords, and highlights require individual markup, with highlights numbered for distinction. The "<[CDATA[]]>" notation indicates special symbols. Main text markup starting from the introduction requires both "name" and "category" attributes: "name" represents the author's title, while "category" represents the standardized name. For example, the tag `<section name="literature review" category="background">` indicates that while authors use "literature review" to describe that section, our unified standardized structure uses "background" for annotation.

3.3 Highlight Markup Rules

Highlight markup is a key issue in this study. We identified each highlight's location in the full text based on sentence matching, phrase relevance, and content relevance principles. Since many highlights do not appear in their original language—that is, highlights are author summaries of specific content—simple string matching is inaccurate and requires sentence-by-sentence verification. Our markup method follows these rules: (1) Add numbering attributes "highlightid=1,2,3…" to highlight tags. The "target" tag corresponds to a specific highlight, while the "match" tag indicates matching status. Special symbols "]]>" and "<![CDATA[" must be added before and after markup locations (see [Figure 1: see original paper]) to prevent markup symbols from being interpreted as regular characters. (2) If a sentence matches only part of a highlight or multiple sentences describe one highlight, the "match" tag is marked as "part" for partial matching; if it describes the complete highlight content, it is marked as "full" for complete matching. (3) If most of a paragraph describes one highlight, the entire paragraph should be marked. If a paragraph describes one highlight but a sentence S1 within it describes another, S1 is marked separately to avoid nesting, with "match" marked as "part" for both parts. (4) One highlight may appear in multiple locations or have multiple matching statements/paragraphs, all of which should be marked (see [Figure 2: see original paper]). The following cases are not marked: highlights appearing as questions, as hypotheses, or citation sentences consistent with highlight viewpoints.

4. Linguistic Feature Analysis of Highlights

The linguistic features of academic paper highlights mainly manifest in characteristic words (keywords) and common expressions. Our approach involves first

conducting keyword analysis to classify highlights by meaning, then summarizing common expressions for each type.

4.1 Keyword Analysis

Keyword analysis, as a form of qualitative analysis, helps identify vocabulary importance in academic discourse and establishes clear understanding of collocation and composition relationships. Keywords are words that appear more frequently than in reference corpora [12]. High-frequency words divide into "high-frequency general words" and "high-frequency characteristic words." Our keywords are "high-frequency characteristic words" that reflect specific content features and style. Speech act verbs constitute an important part of verb vocabulary. From a cognitive perspective, speech act verb sentences involve two participants: the sayer and the object. The object can be concrete or abstract, such as viewpoints presented in highlights. Our survey found that in highlight sentences, the sayer is primarily the paper author or the paper itself, as in "The study identifies the risk of BDT," where "the study" is the sayer, "the risk of BDT" is the object, and the speech act verb "identifies" expresses the core semantic meaning. This validates speech act verbs as keywords.

4.2 Keyword Frequency Statistics

We used WordSmith Tools' keyword retrieval program to identify highlight text keywords and their positions. First, we created two word lists using WordList: one for highlight texts and a larger reference corpus of similar texts for background comparison. Based on this analysis, we selected 154 speech act verbs for further study. Using WordSmith Tools' concordance function to display complete highlight entries, we confirmed valid highlights. Frequency statistics yielded a partial high-frequency speech act verb list (see).

4.3 Classification of Highlights

We classified highlights based on keyword meanings, then verified classifications through tense, voice, or adjectives, ensuring scientific validity. From perspectives of expression form and content, we divided academic paper highlights into four types: research innovation highlights, research method highlights, research process highlights, and research conclusion highlights. This classification demonstrates good inclusiveness and inheritance. lists partial high-frequency keywords and examples for each type.

(1) **Research Innovation Highlights** describe researchers' new viewpoints or discoveries regarding research problems, representing significant differences and substantive progress over existing 成果. These are the most valuable content in a paper. Innovation is the soul of a paper [15], so every paper with research 成果 should contain such highlights. Common keywords include "develop," "explore," "suggest," "devise," "find," "propose," "present," "argue," "advance," and nouns like "finding" and "perspective."

(2) **Research Method Highlights** provide brief introductions to specific methods authors propose to solve research problems. These methods possess novelty and innovation for addressing particular problems and are concrete implementation approaches, not general scientific methods (e.g., observation, empirical research, surveys, expert interviews) or problem-solving approaches (e.g., measurement, co-occurrence, clustering). Common keywords include "use," "through," "employ," "utilize," and nouns like "method," "approaches," and "methodology."

(3) **Research Process Highlights** describe 成果 obtained during the research process. While less innovative than research innovation highlights, they can still drive theoretical improvement and development. Since process descriptions occupy the largest proportion of academic papers, this highlight type appears most frequently. Common keywords include "compare," "introduce," "describe," "outline," and nouns like "description" and "review."

(4) **Research Conclusion Highlights** articulate valuable research conclusions. The basic logic of academic papers involves using certain methods to conduct research and obtain conclusions, making this type an inheritance and summary of method and process highlights. Common keywords include "demonstrate," "validate," "identify," "enter," "drivers," "analysis," "issue," "technique," "application," "process," and nouns like "result," "trend," and "explanation."

Additionally, from a discourse semantics perspective, we can reconfirm highlight classifications through tense, voice, or adjectives (see). Research conclusion descriptions are not necessarily highly innovative; they simply state the final step of the research process.

5. Distribution Pattern Analysis of Highlights Within Papers

Highlight distribution patterns refer to how highlights are distributed throughout the full text and within article sections. Since different sections of academic papers have distinct logical structures and functions, highlights in different parts serve different purposes. Our approach involves first analyzing paper structure, then statistically analyzing highlight types and quantities in each section based on previous analyses.

5.1 Paper Structure Analysis

Academic papers generally adopt the IMRAD standard structure: Introduction (presenting background and research questions), Materials and Methods, Results, and Discussion. This structure has many variants across disciplines—for example, in data-driven fields, "Materials and Methods" becomes "Data and Methods." We surveyed the structure of papers in *International Journal of Information Management* to standardize our corpus structure.

5.1.1 Full-text Section Distribution Our survey of 385 journal papers found 264 with four to six sections, accounting for nearly 70% of the sample. Five-section papers were most common (110 papers, 28.6%), followed by four-section papers (79 papers, 20.5%) and six-section papers (75 papers, 19.5%). Other structural types accounted for about 30%, with many not being full research papers (see [Figure 3: see original paper]).

5.1.2 Section Title Content Section titles showed high diversity. Since multiple sections often share structural functions, we manually judged, distinguished, merged, and categorized titles to form a relatively unified structure for interpreting highlight distribution. Based on frequency statistics, we created a section name structure diagram (see [Figure 4: see original paper]), showing four-, five-, and six-section papers from left to right, with section titles ordered top to bottom. Rectangle size indicates title frequency, colored sequentially from most to least frequent.

5.1.3 Unified Paper Structure Due to structural diversity, we established a unified five-section structure: "Introduction-Research-Method/Methodology-Results-Conclusion." For non-standard papers like case studies, reviews, and commentaries, we applied a three-section structure: "Introduction-Research-Conclusion," merging multiple middle chapters into "Research."

5.2 XML Text Data Parsing

We used Python to parse the corpus. Among common XML parsing methods in Python, the "xml.etree.ElementTree" module (ET) offers a convenient, friendly API with reusable code, fast speed, and low memory consumption. We therefore selected this method. illustrates the XML file processing procedure using document 17 as an example. The parsing approach treats XML content as a tree structure composed of layered nodes. In our study, the root node is the "" tag, with first-level child nodes including "," "," "

," "," "," "," "," "," and "

." The second-level child nodes of "" and "

" are "" tags describing highlight-text matching, requiring sequential traversal to access node values. Obtaining second-level node tags, attributes, and text values clearly reveals highlight matching status and specific content, enabling manual statistics for exploring highlight location distribution.

5.3 Distribution Patterns of Highlights Within Papers

5.3.1 Overall Distribution in Main Text Our 385 papers contained 1,649 highlights. Since some highlights matched multiple locations, they appeared in over 4,000 places throughout the texts (see [Figure 5: see original paper]). Highlights appeared 602 times in Introduction, 325 times in Research, 873 times in Method/Methodology, 1,472 times in Results, and 810 times in Conclusion sections.

5.3.2 Distribution Characteristics by Section Different highlight types show distinct distribution patterns across sections (see [Figure 6: see original paper]). First, the Introduction provides an overview, typically describing research questions, methods, important 成果, and conclusions concisely, thus matching all highlight types. Second, the Research section describes implementation processes, primarily containing research process highlights. Third, the Method section describes specific methods and their implementation, mainly containing method and process highlights. Fourth, the Results section represents a paper's core and innovation, primarily containing research innovation highlights. Fifth, the Conclusion section mainly contains research conclusion highlights but also includes innovation highlights and process highlights when 升华 existing theories or 复述 research processes.

5.3.3 Distribution Within Sections To explore internal distribution patterns, we divided each section into front, middle, and rear portions. We found random distribution without significant clustering in front, middle, or rear sections, except in the Introduction. Since the rear portion typically states research objectives and approaches, highlights appear less frequently there. We merged background and literature review into the "Introduction," so highlights mostly distribute in the front and middle portions.

Conclusion

Rapidly and efficiently discovering valuable content fragments in academic papers to accelerate knowledge innovation has long been a scientific challenge in library science and academic publishing. Current research primarily adopts text mining perspectives, as seen in studies [18,19,20]. Elsevier's highlight proposal has advanced this research. This paper further defines the highlight concept and analyzes its linguistic features and distribution patterns. Our findings are: (1) Academic paper highlights are collections of standardized, semantically explicit short sentences with novelty, conciseness, readability, and promotional characteristics. (2) Highlights hold important value for readers, reviewers, publishers, and authors by facilitating efficient peer review, enhancing publisher competitiveness, improving reader evaluation efficiency, and helping authors publicize their work. (3) Highlights can be classified into research innovation, method, process, and conclusion types. (4) Highlights distribute across all paper sections, primarily in Results and Methods, with random distribution within sections.

Due to corpus limitations, this study only analyzed 385 English papers from one library and information science journal, yielding somewhat limited conclusions. Future research will expand to more disciplines, journals, and languages (including Chinese) to enrich the corpus and further explore highlight features, providing more scientific foundations for automatic highlight extraction rules.

References

[1] Elsevier. Research highlights [EB/OL]. [2018-11-18]. https://www.elsevier.com/authors/journal-

authors/highlights. [2] Hyl and K, Guinda CS. Stance and voice in written academic genres [M]. England: Palgrave Macmillan UK, 2012. [3] Yang W, Wen Hsien. Evaluative language and interactive discourse in journal article highlights [J]. English for specific purposes, 2016, 42: 89-103. [4] Ronzano F, Saggion H. Knowledge extraction and modeling from scientific publications [J]. International workshop on semantic, analytics, visualization, 2016(9792): 11-25. [5] Dahl T. Contributing to the academic conversation: a study of new knowledge claims in economics and linguistics [J]. Journal of pragmatics, 2008, 40(7): 0-1201. [6] Fisas B, Saggion H, Ronzano F. On the discursive structure of computer graphics research papers [C]//Proceedings of the 9th linguistic annotation workshop. Denver: Association for Computational Linguistics, 2015: 42-51. [7] Wen Youkui, Wu Guangyin. Research on dynamic mining of fragmented scientific innovation points [J]. Digital Library Forum, 2014(7): 25-32. [8] Le Xiaoqiu. Analysis of linguistic description characteristics of new discoveries in Chinese scientific literature within a domain [J]. New Technology of Library and Information Service, 2016(5): 47-55. [9] Wen Youkui, Wu Guangyin. Research on dynamic mining of fragmented scientific innovation points [J]. Digital Library Forum, 2014(7): 25-32. [10] Mao Chenyu, Le Xiaoqiu. Analysis of linguistic description characteristics of new discoveries in Chinese scientific literature within a domain [J]. New Technology of Library and Information Service, 2016, 32(5): 47-55. [11] Li Ying, Zhou Li. The value and ideal model of reasonably presenting innovation points in scientific journal papers [J]. Chinese Journal of Scientific and Technical Periodicals, 2018, 29(10): 993-999. [12] Scott M, Tribble C. Textual patterns: keywords and corpus analysis in language education [M]. Philadelphia: John Benjamins, 2006. [13] Zhong Shouman, Zhang Weihua. Classification and semantic cognitive explanation of English and Chinese speech act verbs [J]. Journal of Shangrao Normal University, 2004(5): 88-91. [14] Chen Changlai. Research on syntactic and semantic attributes of modern Chinese verbs [M]. Shanghai: Xuelin Publishing House, 2002. [15] Li Huaizu. Discussion on writing doctoral dissertations in management disciplines [J]. Academic Degrees & Graduate Education, 2000(3): 21-27.

Author Contributions

Suo Chuanjun: Responsible for research topic selection, framework development, and guidance on writing, revision, and improvement.
Yu Guoxin: Responsible for data processing and analysis related to the research topic and initial draft writing.

Note: Figure translations are in progress. See original paper for figures.