A Survey of Automatic Annotation Models for Discourse Elements in Scientific Literature (Post-print)

Authors: Yu Gaihong, Zhang Zhixiong, Ma Na

Date: 2023-08-27T00:00:00+00:00

Abstract

[目的/意义] To better enhance the semantic enrichment of scientific literature, this paper surveys and summarizes discourse element annotation models, techniques, and methods for scientific literature from both domestic and international sources, providing reference for researchers in text mining, knowledge extraction from scientific papers, and semantic analysis systems.

[方法/过程] Using academic website searches and relevant database search engines, we conducted an in-depth review and investigation of reference literature and research reports related to scientific paper annotation, discourse elements, knowledge extraction, sentence recognition, and automatic article classification, and summarized research progress on automatic discourse element annotation models and related work.

[结果/结论] Discourse element annotation of scientific literature holds significant practical application value. Constructing annotation models requires comprehensive consideration of construction philosophy, annotation domain and granularity, and annotation technical means.

Full Text

Preamble

Yu Gaihong[1,2], Zhang Zhixiong[1,2,3], Ma Na[1,2]

[1]University of Chinese Academy of Sciences, Beijing 100049
[2]National Science Library, Chinese Academy of Sciences, Beijing 100190
[3]Wuhan Documentation and Information Center, Chinese Academy of Sciences, Wuhan 430071

Abstract

[**Purpose/Significance**] To enhance the semantic enrichment of scientific literature, this paper surveys and summarizes annotation models, technologies, and methods for discourse elements in scientific papers both domestically and internationally, providing references for researchers in text mining, knowledge extraction, and semantic analysis systems. [**Method/Process**] Using academic search engines and database search tools, we conducted in-depth reading and investigation of references and research reports related to scientific paper annotation, discourse elements, knowledge extraction, sentence recognition, and automatic article classification, summarizing research progress on automatic annotation models for discourse elements. [**Result/Conclusion**] The annotation of discourse elements in scientific literature holds significant practical application value. Constructing annotation models requires careful consideration of modeling philosophy, annotation domain, granularity, and technical approaches.

Keywords: scientific literature, discourse elements, annotation model, automatic annotation
Classification Number: G251
DOI: 10.13266/j.issn.0252-3116.2018.15.015

Introduction

The proliferation of scientific literature in recent years has promoted the emergence of text mining tools, enabling efficient knowledge extraction from texts. Helping users quickly locate desired documents, identify specific categories of information (such as experimental data), and obtain particular knowledge units has become increasingly meaningful and forward-looking. The ability to rapidly and comprehensively acquire research information of interest has emerged as an urgent problem to solve and represents highly valuable work.

This study defines discourse elements as fragments that can explicitly express functional descriptions of the knowledge value embedded in scientific literature. These may be clauses, complete sentences, paragraphs, or even fragments. We define the annotation of discourse elements as labeling their semantic category information, such as research ideas, theoretical tools and methods, scientific experiments, experimental results, research conclusions, etc. Revealing these valuable semantic knowledge units in papers to enable convenient discovery and utilization has become an important topic in digital library research. In recent years, experts and scholars from digital libraries, knowledge extraction, knowledge organization, and related fields have conducted research from various perspectives, though most have been limited by the rich semantic content hidden in texts. Establishing effective annotation models for discourse elements in scientific literature is fundamental to organizing and revealing this knowledge. Automatic annotation models serve as data processing benchmarks and organizational frameworks for discourse element annotation, with all annotation work built upon these models. Therefore, this paper investigates the work and re-

search progress of internationally renowned researchers and major laboratories in the automatic annotation of discourse element structures in scientific literature, focusing on analyzing and summarizing annotation models for scientific literature as a critical research task.

Based on academic website searches and database engines, we retrieved and read literature using key terms including "discourse annotation scheme," "automatic annotation," "semantic annotation," "sentence classification," "discourse structure," "automatic classification," and "semantic annotation." We then conducted detailed analysis of over 50 research papers from the past 15 years, originating from the UK, US, China, and the EU. Finally, we identified several influential research teams specializing in automatic discourse element annotation models as the focus of this review. The following sections first provide detailed descriptions of typical discourse element annotation models, including work by H. Ribau-pierre et al. [1-6] from Oxford University, M. Liakata et al. [7-19] from the University of Warwick, S. Teufel et al. [20-25] from Oxford University, and F. Ronzano et al. [26-31] from the University of Pula, covering detailed explanations of conceptual layers, metadata layers, article structure layers, rhetorical discourse layers, and relational hierarchies. Second, we analyze and compare differences among these models, summarizing perspectives and aspects to consider when developing multi-level semantic annotation models for scientific literature, such as construction philosophy, task drivers, annotation granularity, and research domains, to help researchers better establish and select models. Finally, we summarize our work and outline future research directions.

2 Typical Annotation Models for Scientific Literature Discourse Elements

2.1 SciAnnoDoc Model

The SciAnnoDoc model was proposed by researchers H. Ribau-pierre et al. from Oxford University. The research aims to improve information retrieval precision and enhance the effectiveness of scientific literature search engines [1-2]. The study assumes that when scientists search for information, they typically have precise retrieval goals. Users are not simply searching for literature "about topic T," but rather trying to answer specific questions, such as finding definitions of concepts, seeking results for specific problems, verifying whether an idea has been proven, or comparing scientific conclusions across papers. Answering these precise or complex queries about scientific papers requires precise modeling and annotation of entire article contents, particularly the discourse types of each paper.

Through repeated questionnaires and expert validation with scientists, the team proposed a user-centered SciAnnoDoc annotation model for scientific literature [3-5] to model discourse elements. This model divides full-text scientific papers into four annotation layers: conceptual layer, metadata layer, rhetorical discourse layer, and citation relation layer, as shown in Figure 1 [Figure 1: see

original paper] [5].

Figure 1 SciAnnoDoc Model

(1) **Conceptual Layer (DomainConcept):** Annotates article ontologies, technical term vocabularies, or concepts within the article.

(2) **Metadata Layer (Metadata):** Describes metadata text information such as authors, publication year, journal or conference information.

(3) **Rhetorical Discourse Layer (DiscourseElement):** The core component of each model, describing the role of elements and their knowledge content attributes, decomposed into five aspects: Findings, Hypothesis, Methodology, Related Work, and Definition.

(4) **Citation Relation Layer (relation):** Describes citation and association relationships between articles.

2.2 CoreSC Model

The development of the CoreSC model underwent two key research phases: the first phase's CISp metadata model (core information about scientific papers) [8] and the second phase's CoreSC model (core scientific concepts) [11].

The first-phase CISp metadata model originated from subclasses describing general scientific concepts in EXPO [7], containing concepts crucial for describing a scientific investigation. Through expert investigation and actual paper annotation analysis, concept categories were refined into 12 final categories as the CISp model classification: goal of investigation, object of investigation, method of investigation, experiment, observation, hypothesis, results, conclusion, motivation, background, problem, and example, with eight core categories being goal, motivation, object, method, experiment, results, observation, and conclusion [8-9].

The second-phase CoreSC model, enriched and improved based on CISp, was formally proposed in 2010. It aims to automatically identify components of a research investigation in articles and is a sentence-level text annotation model. The specific model is described in Table 1 [11-12], containing three annotation layers. Table 1 shows the meanings of 11 categories at the first level and category attributes at the second level (New, Old, Advantage, Disadvantage).

(1) **Rhetorical Category Layer:** The first level contains 11 rhetorical categories: Hypothesis, Motivation, Background, Goal, Object, Method, Experiment, Model, Observation, Result, and Conclusion.

(2) **Concept Attribute Layer:** The second level annotates concept attributes, such as New or Old indicating whether a method is new or old, and Advantage or Disadvantage indicating strengths and weaknesses

of a method.

(3) **Concept Identification Layer:** The third level uses ConceptID to identify sets of related instances of the same concept, linking all sentences belonging to the same method with the same ConceptID.

Table 1 Core Scientific Concepts (CoreSC) Annotation Model

2.3 Argumentative Zoning—AZ Model

S. Teufel's Argumentative Zoning (AZ) model [20, 23] was inspired by the concept of knowledge claims: "The act of writing literature is related to claiming ownership of new knowledge, an act of joining the scientific knowledge base and publishing after peer review." The central idea assumes that scientific literature contains positive and negative statements about other contributors, thus the model focuses more on organizing and revealing knowledge claims.

The AZ model also underwent two key development stages. Initially in 1999, S. Teufel et al. divided literature into seven zones, as described in Figure 2 [Figure 2: see original paper] [20]. OTHER, OWN, and BACKGROUND relate to knowledge ownership attribution of these fragments. BASIS declares the use of other work as the foundation or starting point for current work or support obtained. CONTRAST contains comparisons between different research works (such as pointing out shortcomings of other work). AIM indicates the main knowledge claim of the article. TEXTUAL provides physical location information of the text.

Figure 2 Argumentative Zoning AZ Model

The AZ model has been continuously enriched and improved by academia until 2009 when it developed into the AZ-II model [23], expanding to 15 categories with detailed subdivision of contradictions (ANTISUPP), integration of research limitation comments (GAP_{WEAK}), etc. Compared to the original AZ model, the AZ-II extension changes include:

(1) Category AIM remains unchanged.

(2) Category BACKGROUND is renamed CO_{GRO} or becomes general background.

(3) Category OTHER is subdivided into other people's work (OTHER) and authors' own previous work (PREV_{OWN}).

(4) Category BASIS is subdivided into use (USE) and support (SUPPORT).

(5) Category CONTRAST is subdivided into neutral comparison (CODI), comparison combined with research limitation comments (GAP_{WEAK}).

(6) Category OWN is subdivided into method description (OWN_{MTHD}), results (OWN_{RES}), conclusions (OWN_{CONC}), and authors pointing out correctable error information (OWN_{FAIL}).

(7) Category TEXTUAL is discontinued due to less information compared to other categories.

The model introduces two new categories—novelty advantage of new knowledge claims (NOV_{ADV}) and future work limitation statements (FUT). Specific category meanings are shown in Table 2 [22-23].

Table 2 Argumentative Zoning AZ-II Annotation Model

2.4 Multi-Layer Scientific Discourse Annotation Model

This model was innovatively proposed in 2015 by B. Fisas, F. Ronzano, et al. from the Natural Language Processing team at the University of Favara [27-28], combining the actual situation in the computer graphics field to create a simplified annotation model. Computer graphics is a relatively young discipline with less mature vocabulary annotation compared to biological sciences, and computer graphics researchers typically have technical backgrounds in physics, mechanics, fluid dynamics, mathematics, etc. Therefore, the model focuses more on algorithms, equations, algebra, and mathematical reasoning layers, cross-feature layers, and central relevance layers.

(1) **Discourse Category Layer:** Contains five categories derived from simplified mapping of CoreSC and AZ model categories: Challenge, Background, Approach, Outcome, and FutureWork. Specific category definitions and simplification explanations are shown in Figure 4 [Figure 4: see original paper] [28].

(2) **Citation Purpose Layer:** Primarily provides detailed annotation of citations in literature, mainly adopting the annotation model proposal by A. Abu-Jbara et al. [48]. Specific citation purpose categories are shown in Table 3 [28], including commentary, comparison, use, foundational work, or general research. Each category has different sub-attributes, such as Weakness and Strength containing evaluation polarity, Evaluation aiming to collect sentences with positive and negative comments about a citation, Similarity and Difference for comparative opposition, Use for annotating method, data, or tool citations, etc. Basis class annotates whether authors cite their own work (OwnWork) or others' work. Neutral category contains descriptions of researchers' work, references for more information, or domain-general practices.

(3) **Cross-Feature Layer:** Scientific literature discourse features Advantage and Disadvantage can describe characteristics of authors' own methods and cited literature. Since advantages and disadvantages usually appear

within one sentence, the cross-feature layer includes double-precision categories Advantage-disadvantage and Disadvantage-advantage, innovation (Novelties), and domain-general practice feature annotations. Finally, Limitations features only refer to authors' own work, which is important for comparing different investigations. The five cross-feature categories are ADVANTAGE, DISADVANTAGE, COMMON, NOVELTY, LIMITATION.

(4) **Central Relevance Importance Layer:** Sets five levels of scores according to each sentence's contribution to the central idea, from completely irrelevant (1 point) to very relevant (5 points).

Figure 3 Multi-Layer Scientific Discourse Annotation Model
Figure 4 Multi-Layer Scientific Discourse Annotation Model Category Definition
Table 3 Citation Purpose Categories

2.5 Research Design Fingerprint Description Model

In 2014, Qian Li, Zhang Xiaolin, et al. from the National Science Library of Chinese Academy of Sciences [34] proposed using research design fingerprints to structurally describe scientific literature, enhancing computer recognizability and executability of scientific literature, helping researchers quickly understand research methods, algorithms, tools, and conclusions, and providing publishing specifications for future scientific publishing (i.e., semantic publishing). The specific research design fingerprint description model is shown in Table 4 . The research design fingerprint framework uses research design fingerprints to represent scientific literature research results. The overall structure has two levels: the first level includes research topic, research method, research algorithm, research results, research conclusions, and future research; the second level provides detailed descriptions of scientific literature, mainly including research hypothesis, research scenario, research purpose, research background, research method, research data, research algorithm, research results, research conclusions, future research, and research equipment—11 design fingerprints in total. The two levels are interrelated internally and between levels, effectively supporting correlation calculation and distribution of scientific resources. This model annotates four granularities of full text: title, abstract layer, main text argument zoning layer, sentence layer, and keyword layer.

Table 4 Research Design Fingerprint Description Model

3 Comparative Analysis of Discourse Element Annotation Models and Research Progress

Since the foreign models described above have undergone systematic construction and implementation, this section focuses on analyzing and comparing foreign research progress. First, we compare similarities and differences among

the four models from different perspectives to provide suggestions and references for researchers actually building models, mainly from the angles of model construction philosophy and tasks, model categories and construction methods, annotation domains and corpora, annotation tools and classification algorithms, and final experimental effect analysis. Finally, we summarize problems in model research to provide references for subsequent work.

3.1 Comparative Analysis of Model Construction Philosophy and Tasks

Model construction philosophy, as the researcher's original intention and theoretical foundation supporting the entire model, is critical and determines the model's distinguishing granularity and division angle. The common point is that every researcher aims to better extract and mine discourse element value fragments from literature. The difference is that researchers face different task drivers. For example, the main purpose of the SciAnnoDoc model is to improve retrieval efficiency in search systems, helping users quickly find desired knowledge fragments. The CoreSC model aims to comprehensively explain a research investigation from an ontology research perspective. The AZ model is based on knowledge claim viewpoints, emphasizing authors' contributions and citations of others' work. The Multilayer model adapts to technological development to better solve semantic analysis problems in new domain literature. Therefore, researchers need to establish different research models according to actual research tasks, as shown in Table 5 .

Analysis reveals that if researchers focus on organizing and revealing research content in a disciplinary field, they can adopt the CoreSC ontology-based revelation model. If researchers emphasize discovering intellectual property influences and scholar contributions, they can use the AZ model to conveniently distinguish between others' and authors' own contributions. For practical scenarios involving literature central idea extraction and directional retrieval, researchers can adopt SciAnnoDoc and Multilayer models to help users quickly find desired knowledge.

Table 5 Comparison of Model Construction Philosophy and Tasks

3.2 Comparative Analysis of Discourse Element Categories and Construction Methods

Model construction methods mostly undergo continuous argumentation and improvement. For example, SciAnnoDoc first used questionnaires then invited expert confirmation for final categories. The CoreSC model mainly selected core concept categories based on scientific entity ontology evolution to determine CoreSC categories. Multilayer category confirmation is detailed in Section 2.4 above, also streamlining the 16 categories and concepts from the above models. Most models use sentence-level annotation granularity, such as CoreSC, AZ, and Multilayer models, while SciAnnoDoc selects fragment annotation for

richer retrieval content. See Table 6 for details.

Table 6 Comparison of Discourse Element Categories and Construction Methods

3.2.1 Comparison Between CoreSC and AZ Models M. Liakata and S. Teufel conducted comparative annotation of CoreSC and AZ models [11], pointing out that these two models complement each other in representing scientific literature perspectives. CoreSC's BACKGROUND includes general neutral background knowledge and existing knowledge claims, corresponding to OTHER, PREV_{OWN}, and CO_{GRO} categories in the AZ-II model. The AIM category in the AZ model is a research goal statement, but in the CoreSC model can be decomposed into three categories: GOAL (target state of investigation), HYPOTHESIS (unverified claims), and OBJECT (specific entities related to the investigation or innovative statements about entity properties). OWN_{MTHD} and METHOD both refer to methods used, but CoreSC further distinguishes between experimental methods (EXPERIMENT), methods used in current research (Method-New), and methods mentioned in other work (Method-Old). OWN_{RES} relates to CoreSC's OBSERVATION, representing data or phenomenon records in an investigation. In contrast, CoreSC's RESULT belongs to factual statements originating from OBSERVATION. AZ-II's NOV_{ADV} represents novelty and advantages of methods used in articles, corresponding to CoreSC's ability to annotate METHOD and OBJECT for novelty and advantages. Other categories are completely different: CoreSC's HYPOTHESIS, MOTIVATION, OBJECT, and MODEL are organized entirely according to research investigation ontology, while AZ-II's CODI, GAP_{WEAK}, SUPPORT, ANTISUPP, USE, and FUT are organized according to relationships with other work, and OWN_{FAIL} describes authors' shortcomings.

3.2.2 Comparison Between CoreSC+AZ and Multilayer Models From the Multilayer model definition, we can clearly see that the model's CHALLENGE category represents the research situation faced by current researchers, which can be mapped to CoreSC's HYPOTHESIS, MOTIVATION, and GOAL, as well as AZ model's GAP_{WEAK} and OWN_{FAIL} categories that indicate unresolved problems. BACKGROUND reveals published information useful for understanding the current research subject, mapping to CoreSC's BACKGROUND and AZ model's CO_{GRO} general background knowledge and USE categories. APPROACH maps to CoreSC's possible previous MODEL, EXPERIMENT, OBSERVATION, and METHOD. OUTCOME reveals research findings, including measurable data results (RESULT) or conclusions (CONCLUSION). FUTUREWORK corresponds to AZ model's FUT category.

3.2.3 Comparison Between Multilayer and SciAnnoDoc Models Since the SciAnnoDoc model is completely built from a user perspective using empiri-

cal research methods, it is closer to human retrieval usage angles. For example, DEFINITION is the most basic category users need when wanting to understand domain knowledge. Other categories basically correspond one-to-one with the Multilayer model.

Through comparative analysis, we find that automatic annotation model semantic categories are basically limited to six to meet various work needs: research goal, research background, research method, research findings, research conclusion, and research definition (concept explanation) of key terms in scientific literature. Other categories can be mapped or transformed to these categories, avoiding redundancy from too many categories or excessive broadness from too few, while completely covering semantic value information.

3.3 Overall Research Progress and Results Comparison

Analysis from the perspective of each model's research domain and application projects reflects the specific research value and practical value of this work, also providing rich corpora and project references for subsequent researchers. See Table 7 for details.

Table 7 Comparison of Overall Research Projects and Results

The SciAnnoDoc research team is committed to improving retrieval efficiency. Supported by the Swiss National Science Foundation [5-6], the project selected the humanities gender research field to develop a user-oriented literature retrieval query system (FSAD system). Based on the proposed SciAnnoDoc model, the system provides annotation tools and manually annotated corpora. Rigorous scientific user evaluation proved that compared with traditional keyword-based retrieval systems, the FSAD system significantly improved users' problem-solving accuracy and efficiency.

The CoreSC model's predecessor was the CISp model. Research began in 2007 with the ART project (An ontology-based article preparation tool) funded by the UK Joint Information Systems Committee (JISC) [7-9]. The project produced the manual annotation tool SAPIENT based on the CISp model, which greatly improved manual annotation efficiency and accuracy using a decision tree approach, laying a solid foundation for building reliable training corpora. The project achieved manual annotation of 225 biochemistry domain corpora [8-9]. To achieve machine automatic annotation, the 2010 EU-funded SAPIENTAutomation project continued this research, proposing the CoreSC model and developing the automatic annotation tool SAPIENTA (note the added A for Automation), further enriching and improving the corpora to complete 265 gold-standard annotated datasets [11-12]. SAPIENTA was specifically applied in two systems: first, an automatic summarization system, where experiments proved that CoreSC model-generated summaries were better than Microsoft automatically generated summaries, even surpassing human-written summaries in some cases; second, the CRA project (Cancer Risk Assessment) in life sciences research [14-16] for better annotation of domain articles to facilitate in-depth

research.

The AZ model was initially a Cambridge University research project. In the first phase, the proposer and collaborators annotated 80 computational linguistics articles [20, 22] and achieved machine automatic annotation. The second phase applied the model to the CRA project, led by Cambridge professor A. Korhonen and Karolinska Institute professor U. Stenius, assisting researchers and risk assessors in effectively managing health risks. The project used the AZ-II model to automatically annotate relevant literature, producing 1,000 abstract-annotated corpora and the CRAB 2.0 annotation tool for cancer assessors to read and retrieve annotated scientific literature online [15, 38-39].

The Multilayer model was funded by the EU Seventh Framework Programme, ultimately aiming to promote technological innovation through scientific and technological means, focusing on using scientific literature annotation and knowledge mining to discover potential technological innovation points. The project's greatest contribution is the complete set of annotation frameworks and online tools DRIFramework [26] integrating many open-source tools, facilitating reference and use by relevant researchers, and further application in the SKM Scientific Knowledge Miner project for scientific knowledge mining [29-30].

3.4 Comparative Analysis of Model Annotation Techniques and Classification Effects

The quality of system application effects depends on semantic category annotation effectiveness. This paper compares the correct rate P, recall rate R, and F1 values of automatic annotation implementation, as well as classification methods [43] and comparisons of best, worst, and average effects. Since each researcher works in different domains with different experimental datasets, this paper provides overall experimental effects for each researcher without conducting experimental analysis on the same dataset (experimental comparison based on the same dataset can be a future research task). Although based on different datasets, overall analysis from result data does not affect this review and conclusion analysis.

Specifically, SciAnnoDoc mainly relies on manually written grammar rules for different categories [4], including 20 Finding rules, 34 Definition rules, 11 Hypothesis rules, and 19 Methodology rules. Using 1,400 manually annotated instances for training, it achieved automatic classification of 555 sentences with effects shown in Table 8 .

Table 8 SciAnnoDoc Classification Results

CoreSC model researchers used different features for automatic classification experiments based on Support Vector Machine (SVM), Conditional Random Field (CRF), and linear kernel classifiers. Here we only list better results: using an SVM-based classifier with all feature values achieved classification effects for each category [12] as shown in Table 9 .

Table 9 CoreSC Model Classification Results

Since the AZ-II model is relatively complex and contains some unique categories, its effectiveness is not comparable with other models. Therefore, we selected the relatively simple AZ model for automatic annotation classification comparison. Experiments used 80 manually annotated computational linguistics articles with AZ, conducting cross-validation experiments using a Naive Bayes classifier with effects shown in Table 10 [22].

Table 10 AZ Model Classification Results

The Multilayer model achieved classification effects in the computer graphics field as shown in Figure 6 [Figure 6: see original paper]. Based on 40 manually annotated articles, logistic regression and SVM classifiers were tested, with F1 values recorded and compared. Logistic regression achieved better results (lacking precision and recall statistics).

Figure 6 Multilayer Model Classification Results

Overall comparison of each model is shown in Table 11 .

Table 11 Comparison of Model Annotation Tools and Classification Algorithms

3.5 Analysis of Problems in Discourse Element Automatic Annotation Models

(1) **Difficulty in defining automatic annotation semantic category types and implied meanings.** Too few or too general categories cannot meet user needs (e.g., SciAnnoDoc and Multilayer), but too many categories cause cross-coverage, redundancy, and classification difficulties (e.g., CoreSC and AZ-II models). Therefore, repeated empirical investigation combining actual user needs and research requirements is necessary. Due to different knowledge backgrounds and understanding, investigations are full of subjectivity and human interference factors. Developing categories applicable to more populations to overcome inconsistencies caused by human factors is difficult.

(2) **Manual annotation data is labor-intensive, time-consuming, and tedious.** Every study requires preliminary manual corpus annotation with high precision requirements. The quality of this manually annotated dataset as training data directly affects automatic classification effectiveness. Therefore, domain experts are mostly needed for manual annotation. Only the AZ model established a decision tree-based manual annotation guide [20] to reduce manual annotation difficulties and improve accuracy, enabling non-experts to complete annotation following decision tree instructions, achieving domain independence. However, most models remain domain-dependent.

(3) **Final automatic classification effects are still not ideal, with much room for improvement in classification techniques and methods.** Analysis of model annotation techniques and classification effects reveals that regardless of domain corpora and classification methods, current automatic annotation model effects are not particularly ideal and have room for further improvement. The SciAnnoDoc model differs from other models in using rule-based methods with relatively high precision (average 77%) but low recall (average 35%). More rules are needed to improve recall, but more rules increase the risk of noisy annotation data. The CoreSC model organizes classification according to scientific research investigation and is relatively complete for describing core concepts. However, with too many categories, some categories are easily confused or inherently ambiguous, causing difficulties for manual annotation. After all, having experts annotate sentences with 11 categories is difficult work, so the reliability of the resulting training data itself becomes problematic. Consequently, classification effects are relatively low compared to other models, with F1 values ranging from a maximum of 75% to a minimum of 10%, averaging only 41%. The AZ model's introduction of decision tree-based annotation methods improved training data accuracy to some extent, achieving 81% precision and 91% recall for the OWN category. However, the AZ model itself is relatively primitive compared to AZ-II, with broader classification angles. Using this model generally requires further optimization.

The Multilayer model's research domain and work are relatively new. Based on predecessors' work, its classification effects are significantly better than other models in both model and classification methods. However, the classification process uses some computer graphics domain-specific features for training, so its extensibility to other domains needs further verification. Additionally, with fewer categories, it may not sufficiently meet researchers' needs.

4 Conclusion and Future Work

Research and annotation of discourse elements in scientific literature have important theoretical research and practical application value in the current era of intellectual property development, knowledge publishing business innovation, and knowledge service leadership. They are widely applied in knowledge service processes such as search engines, automatic summarization, scientific innovation point discovery, automatic question-answering systems [17], semantic publishing [47], writing instruction [35], semantic knowledge organization, citation recommendation [19], and Japanese legal article annotation [25]. Additionally, they are widely used in medical and life science research work, such as cancer risk assessment, life science gene function comparison/gene discovery [45-46], evidence-based medicine [23], enabling multidisciplinary and cross-domain research collaboration.

This paper provides detailed research and comparison of several models. Each

team selected different domains for manual annotation according to different research tasks and focuses, producing a series of annotated datasets and method collections. These models are both complementary and distinct, providing valuable references for research in this field. When annotating discourse elements in scientific literature, determining the annotation model is the foundational and core component of this work. First, the annotation task and target domain must be determined. Different research tasks and domains result in vastly different article structures and content due to differences in research content and researcher thinking, leading to obvious differences in model selection and category distinction. Second, the content granularity for annotation must be selected, such as fragment-based, concept zone-based, sentence-based, or event-based annotation. Different granularities are not isolated—full text consists of fragments, fragments consist of sentences, and sentences contain different events. Finally, models and classification categories must be determined flexibly based on actual application scenarios. Typically, a sentence can well express author intent, and computer technology can effectively implement sentence segmentation while avoiding semantic category conflicts in fragment-based annotation. Therefore, sentence-level annotation can be selected for research. Regarding the number of semantic categories, analysis shows that 5-7 categories can typically cover all semantic descriptions. This paper concludes that six semantic categories can cover most discourse element semantic types: research goal, research background, research method, research findings, research conclusion, and research definition.

The final automatic classification effect of scientific discourse elements directly determines application effectiveness and is the research focus and technical difficulty of the entire work. Rule-based classification can achieve high precision but unsatisfactory recall. Machine learning classifiers require large amounts of manually annotated training data, challenging domains lacking manual support, and classification effects for some categories are also not ideal. The rapid development of artificial intelligence has advanced machine learning and deep learning methods, with researchers already exploring weakly supervised [31, 36], unsupervised [40], and deep neural network learning [42] algorithms to solve such problems. Innovative fusion of various algorithms for classification can also further improve classification effects, which will be the focus of our future research.

This paper summarizes and compares the basic theoretical assumptions and construction philosophies of models. Previous comparative studies of models have lacked comparison from this angle, yet theoretical foundations or assumptions are often the starting point for research and have decisive effects on subsequent research and model finalization. However, this survey may have deficiencies and incomplete aspects. We will further strengthen corresponding survey work in the future.

Next, we will combine semantic enrichment work in the physics field to specifically establish and implement a model suitable for this work's needs. Simultaneously, we can conduct experimental verification of different methods based

on the same dataset to make comparative work more valuable for reference, and also hope to propose innovative classification methods to solve the above classification problems.

References

[1] Falquet G. New trends for reading scientific documents[C]//ACM workshop on online books, complementary social media and crowdsourcing. New York: ACM, 2011: 19-24.
[2] Ribau-pierre H, Falquet G. A user-centric model to semantically annotate and retrieve scientific documents[C]//Proceedings of the 14th international conference on knowledge technologies and data-driven business. New York: ACM, 2014: 40.
[3] Ribau-pierre H, Falquet G. User-centric design and evaluation of a semantic annotation model for scientific documents[C]//Proceedings of the 5th international workshop on semantic digital archives. Osaka: International Workshop on Semantic Digital Archives, 2015: 30-41.
[4] Ribau-pierre H. Precise information retrieval in semantic scientific digital libraries[D]//Genève: UNIVERSITÉ DE GENÈVE, 2014.
[5] Ribau-pierre H, Falquet G. An automated annotation process for the SciDocAnnot scientific document model[C]//Proceedings of the 5th international workshop on semantic digital archives. Osaka: International Workshop on Semantic Digital Archives, 2015: 30-41.
[6] Ribau-pierre H, Falquet G. Extracting discourse elements and annotating scientific documents using the SciAnnotDoc model: a use case in gender documents[J]. International journal on digital libraries, 2017, 1(3): 1-16.
[7] Soldatova LN, King RD. An ontology of scientific experiments[J]. Journal of the royal society interface, 2006, 3(11): 795-803.
[8] Soldatova L, Liakata M. An ontology methodology and cisp-the proposed core information about scientific papers[EB/OL]. [2018-05-31]. http://repository.jisc.ac.uk/137/1/Report-CISP.pdf.
[9] Liakata M, Soldatova LN. Guidelines for the annotation of general scientific concepts[J]. Applied & environmental microbiology, 2008, 61(3): 1020-1026.
[10] Liakata M, Claire Q, Soldatova LN. Semantic annotation of papers: interface & enrichment tool[C]//Proceedings of the BioNLP2009 workshop. Boulder: Association for Computational Linguistics, 2009: 193-200.
[11] Liakata M, Teufel S, Siddharthan A, et al. Corpora for the conceptualisation and zoning of scientific papers[C]//International conference on language resources and evaluation. Valletta: European Languages Resources Association (ELRA), 2010: 105-111.
[12] Liakata M, Saha S, Dobnik S, et al. Automatic recognition of conceptualization zones in scientific articles and two life science applications[J]. BMC bioinformatics, 2012, 28(7): 991-1000.
[13] Liakata M, Dobnik S, Saha S, et al. A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task[C]//Proceedings of the 2013 conference on empirical methods in natural

language processing. EMNLP. Seattle: Association for Computational Linguistics, 2013: 747-757.
[14] Korhonen A, Silins I, Lin S, et al. The first step in the development of text mining technology for cancer risk assessment: identifying and organizing scientific evidence in risk assessment literature[J]. BMC bioinformatics, 2009, 10(1): 1-19.
[15] Guo Y, Korhonen A, Liakata M, et al. Identifying the information structure of scientific abstracts: an investigation of three different schemes[C]//Proceedings of the 2010 workshop on biomedical natural language processing. Uppsala: Association for Computational Linguistics, 2010: 99-107.
[16] Guo Y, Korhonen A, Liakata M, et al. A comparison and user-based evaluation of models of textual information structure in the context of cancer risk assessment[J]. BMC bioinformatics, 2011, 12(1): 1-18.
[17] Liakata M, Thompson P, de Waard A, et al. A three-way perspective on scientific discourse annotation for knowledge extraction[C]//Proceedings of the workshop on detecting structure in scholarly discourse. Jeju Island: Association for Computational Linguistics, 2012: 37-46.
[18] Ravenscroft J, Oellrich A, Saha S, et al. Multi-label annotation in scientific articles: the multi-label cancer risk assessment corpus[C]//Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož: European Language Resources Association (ELRA), 2016.
[19] Duma D, Liakata M, Clare A, et al. Applying core scientific concepts to context-based citation recommendation[C]//Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož: European Language Resources Association (ELRA), 2016.
[20] Teufel S, Carletta J, Moens M. An annotation scheme for discourse-level argumentation in research articles[C]//Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 1999: 110-117.
[21] Teufel S. Argumentative zoning: information extraction from scientific text[D]//Edinburgh: University of Edinburgh, 1999.
[22] Teufel S, Moens M. Summarizing scientific articles: experiments with relevance and rhetorical status[J]. Computational linguistics, 2002, 28(4): 409-445.
[23] Teufel S, Batchelor C, Batchelor C. Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics[C]//Conference on empirical methods in natural language processing. Singapore: Association for Computational Linguistics, 2009: 1493-1502.
[24] Heffernan K, Teufel S. Identifying problems and solutions in scientific text[J]. Scientometrics, 2018(1): 1-16.
[25] Yamada H, Teufel S, Tokunaga T. Annotation of argument structure in Japanese legal documents[C]//Proceedings of the 4th workshop on argument mining. Copenhagen: Association for Computational Linguistics, 2017: 22-31.
[26] Ronzano F, Sagion H. Dr. inventor framework: extracting structured information from scientific publications[C]//Japkowicz N, Matwin S. Discovery science. Cham: Springer, 2015: 209-220.

[27] Fisas B, Ronzano F, Sagion H. On the discursive structure of computer graphics research papers[C]//The 9th linguistic annotation workshop held in conjunction with NAACL. Denver: Association for Computational Linguistics, 2015: 42-51.
[28] Fisas B, Ronzano F, Sagion H. A multi-layered annotated corpus of scientific papers[C]//Proceedings of the tenth international conference on language resources and evaluation. Paris: European Language Resources Association, 2016.
[29] Ronzano F, Sagion H. Knowledge extraction and modeling from scientific publications[M]//Osborne: Springer International Publishing, 2016: 11-25.
[30] Ronzano F, Freire A, Saez-Trumper D, et al. Making sense of massive amounts of scientific publications: the scientific knowledge miner project[C]//BIRNDL 2016 joint workshop on bibliometric-enhanced information retrieval and NLP for digital libraries. New York: Digital Libraries. IEEE, 2016: 36-41.
[31] Anke LE, Sagion H, Ronzano F. Weakly supervised definition extraction[C]//Proceedings of the international conference on recent advances in natural language processing. Shoumen: INCOMA Ltd, 2015: 176-185.
[32] Xing Meifeng. Research on sentence-level new information detection methods in scientific literature[D]//Beijing: Graduate University of Chinese Academy of Sciences, 2012.
[33] Bai Guangzu, He Yuanbiao, Ma Jianxia, et al. Automatic recognition of academic abstract structure using small sample machine learning[J]. New technology of library and information service, 2014, 30(7): 34-40.
[34] Qian Li, Zhang Xiaolin, Wang Qian. Research on research design fingerprint description framework based on scientific literature[J]. Journal of academic libraries, 2015(1): 14-20.
[35] Song W, Fu R, Liu L, et al. Discourse element identification in student essays based on global and local cohesion[C]//Proceedings of the 2015 conference on empirical methods in natural language processing. Lisbon: Association for Computational Linguistics, 2015: 2255-2261.
[36] Guo Y, Korhonen A, Poibeau T. A weakly-supervised approach to argumentative zoning of scientific documents[C]//Proceedings of the conference on empirical methods in natural language processing. Edinburgh: Association for Computational Linguistics, 2011: 273-283.
[37] Contractor D, Guo Y, Korhonen A. Using argumentative zones for extractive summarization of scientific articles[C]//Proceedings of International Conference on Computational Linguistics. Mumbai: The COLING 2012 Organizing Committee, 2012: 663-678.
[38] Silins I, Korhonen A, Guo Y, et al. A text-mining approach for chemical risk assessment and cancer research[J]. Toxicology letters, 2014, 229(4): S164-S165.
[39] Guo Y, Saghda Do, Silins I, et al. CRAB 2.0: a text mining tool for supporting literature review in chemical cancer risk assessment[C]//Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations. Dublin: Dublin City University and Association for Computational Linguistics, 2014: 76-80.

[40] Kiela D, Guo Y, Stenius U, et al. Unsupervised discovery of information structure in biomedical documents[J]. BMC bioinformatics, 2014, 31(7): 1084-1092.
[41] Baker S, Silins I, Guo Y, et al. Automatic semantic classification of scientific literature according to the hallmarks of cancer[J]. BMC bioinformatics, 2015, 32(3): 432-440.
[42] Baker S, Korhonen A. Initializing neural networks for hierarchical multi-label text classification[C]//16th Workshop on Biomedical Natural Language Processing. Vancouver: Association for Computational Linguistics, 2017: 307-315.
[43] Kim SN, Martinez D, Cavedon L, et al. Automatic classification of sentences to support evidence based medicine[J]. BMC bioinformatics, 2011, 12(2): S5.
[44] Sollaci LB, Pereira MG. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey[J]. Journal of the medical library association, 2004, 92(3): 364-371.
[45] Gobeill J, Tbahriti I, Ehrler F, et al. Gene ontology density estimation and discourse analysis for automatic GeneRiF extraction[J]. BMC bioinformatics, 2008, 9(3): S9-19.
[46] Jimeno-Yepes AJ, Sticcoo JC, Mork JG, et al. GeneRIF indexing: sentence selection based on machine learning[J]. BMC bioinformatics, 2013, 14(1): 1-10.
[47] Clark T, Ciccarese PN, Goble CA. Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications[J]. Journal of biomedical semantics, 2014, 5(1): 28-61.
[48] Abu-Jbara A, Radev D. Reference scope identification in citing sentences[C]//Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Montreal: Association for Computational Linguistics, 2012: 80-90.

Author Contributions:
Yu Gaihong: Responsible for literature review and article writing;
Zhang Zhixiong: Responsible for article structure and writing guidance;
Ma Na: Participated in paper revision and writing.

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