Automatic Semantic Annotation for Abstracts of Scientific Discourses

Lan Huang[1,a], Jinchao Zhu[1,b], Yang Chi[2,c], and Hao Xu[1,d*]

[1] College of Computer Science and Technology, Jilin University, Changchun 130012, China
[2] College of Software, Jilin University, Changchun 130012, China
[a]huanglan@jlu.edu.cn, [b]zhujc14@mails.jlu.edu.cn,
[c]chiyang5512@mails.jlu.edu.cn, [d]xuhao@jlu.edu.cn
*Corresponding Author

Abstract. The abstract of scientific papers has strong semantic structure, which contains abundant meaningful information, such as the background, research problem, solution, and result. Marking it out can help the computer understand and use this underlying information, which can provide great help for searching and scanning papers. In order to annotate the semantics of the paper automatically, we modeled the rhetorical structure of an abstract by linguistic clues and position information.

Keywords: semantic structure, automatic annotation, abstract.

1 Introduction

We can complete the acquisition, dissemination, and communication of knowledge through scientific papers. The exponential growth of electronic scientific papers has made finding and selecting them difficult. Modeling and annotating the rhetorical structure of scientific articles can improve the efficiency of searching and reading [10]. On one hand, it can help search engines to quickly retrieve insight into the core of scientific article [9]; on the other hand, it can help the reader quickly browse and understand those articles. The study of the rhetorical structure and annotation of discourse has long-standing traditions. Automatic annotation of full articles under the existing state-of-the-art is difficult to achieve [3]. The abstract of scientific articles is briefer than full articles, which makes the annotation possible.

There are numbers of well-known approaches to modeling the rhetorical structure of publications, such as Harmsze [8], ABCDE (Annotations,Background, Contribution, Discussion and Entities)[13], and SALT(Semantically Annotated LaTeX)[7]. Based on the previous approaches, we provide a new model for abstract semantic analysis that includes the background, research problem, solution, and result. The model includes a preliminary verification by means of data crawling from the Internet for testing purposes. According to the methods for machine annotation of scientific discourse, Argumentative zoning[11], XIP(Xerox incremental parser)[1], and SemTag[4], this paper use linguistic clues and location information to annotate the scientific publication's abstract automatically.

2 State of the Art

The predecessors of semantic technology have made many contributions for semantic annotated structure, which can improve the efficiency of publications' searching and reading [10]. There are well-known approaches to modeling the rhetorical structure of publications. Harmsze proposed one of the first and the most comprehensive models for extracting the rhetoric and argumentation within scientific papers. This model focused on developing a modular representation for the creation and evaluation of scientific publications [8].De Waard and Tel introduced a different model for representation of discourse called ABCDE, which developed a LaTeX style sheet to identify five components in a discourse [13]. They proposed finer-grained annotation to complement these structures and relationship types [3]. A semantic authoring framework to enrich scientific publications with semantic metadata was called SALT, offering an improved coarse-grained rhetorical structure and a fine-grained semantic network [7].

After modeling the rhetorical structure of publications, we proceed to the automatic annotating of the scientific discourse. The first attempt to automatically annotate rhetorical expressions in research papers is called argumentative zoning [11]. XIP detect rhetorical expressions from language uses of the authors, targeting salient sentences within scientific articles [1]. Both models use clear linguistic clues to annotate the scientific articles. SemTag is a system offering an automatic ontology of semantic information, which identifies the candidate instance's keywords needed annotation. The system is based on TAP, a knowledge base from Stanford University which constructs two text vectors -context (before and after each 10 words) and candidate instance- calculating similarity and selecting best matches [4]. The automatic annotation of research papers should capture and represent the evolution of ideas and findings that authors described in the articles [12]. The main line of the above research aims at extracting factual information from the texts of the articles and transforming them into structured data [2][6]. Semantic structure models and machine annotation models are used for full scientific publications, but the abstracts of papers contain more standardized semantic structures and rhetoric [5]. Some journals like Nature provide a constant structure for their papers, which makes automatic annotation of abstracts possible.

3 Design

Based on the models for full publications mentioned in Section 2-Harmsze, ABCDE, SALT-we provide a new model for abstracts that includes background, research problem, solution, and result. We use linguistic clues extracted from abstract annotation and position information to annotate it automatically.

The flowchart for our system is in Figure 1.

3.1 Framework of Process Model of System

The system is divided into the following seven main phases:

Phase 1: Data Acquisition Module

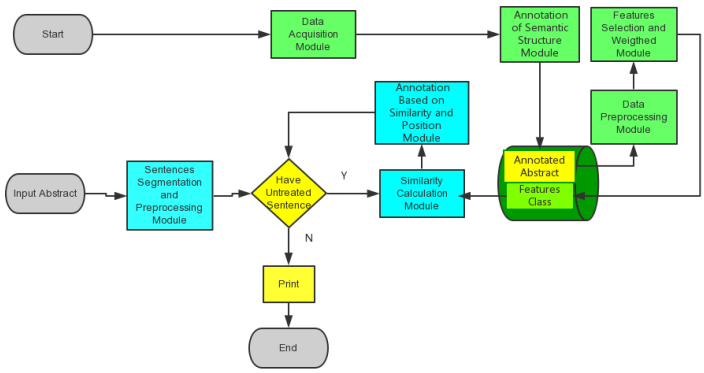

Fig. 1. system's flowchart

This system developed a web crawler with Python to obtain data from DBLP DB. We obtained 208 articles from two different sources, of which 88 were from the journal Data Mining and Knowledge Discovery (DMKD), and the rest were from the European Conference on Principles of Data Mining and Knowledge Discovery (PKDD).

Phase 2: Annotation of Semantic Structure Module

The four semantic tags were added manually to every sentence of scientific papers' abstracts, which includes background, research problem, solution, and result.

Phase 3: Data Preprocessing Module

This module cleaned the text of publications' abstracts, such as removing numbers and punctuation, reducing word roots, converting all words to lower case, and so on.

Phase 4: Features Selection and Weighted Module

In this module, TF-IDF algorithm was used to manage the numerous and complicated information of texts, which helped us discover key words that are more important and appeared more frequently in one text.

Phase 5: Sentences Segmentation and Preprocessing Module

This module was used to segmented the abstract text by period, cleaned the sentences as in Phase 3, and stored them as feature vectors with an ID, which described the sentences' positions in the paper's abstract.

Phase 6: Similarity Calculation Module

This module acquired the similarity level of two feature vector quantities by cosine similarity algorithm, which calculated the value of the included angle of two vectors.

Phase 7: Annotation Based on Similarity and Position Module

The automatic annotation the semantics of the paper,by linguistic clues(They **propose**...) and position information(the position of the sentence).

3.2 Feature Selection and Weight

The results of TF-IDF revealed the information of words as well as the texts in the form of a matrix, and every word had a TF-IDF weight value in each text, which presented the significance of the word in that text. The calculation of the weighted value became the core of the arithmetic.

TF-IDF algorithm was divided into the concepts TF and IDF. TF(Term frequency) reflects that the words with higher rates will attain a higher TF value. The IDF is the

acronym for inverse document frequency. Compared with these words in common use, the ones with a high frequency acquired a higher IDF value, which did not often appear in other texts. As a result, the TF-IDF weight value was calculated with the formula TF*IDF. The key words with high values were chosen according to the integrated conditions of their TF and IDF.

The weighted feature words of the research problem are in Table 1.

Table 1. Weighted feature words of research problem.

Word	Value	Word	Value	Word	Value
generate	0.11142059	discrimination	0.08177003	output	0.06733483
use	0.11142059	answer	0.06733483	association	0.06733483
process	0.1090267	attempt	0.06733483	category	0.06733483
success	0.10100225	classic	0.06733483	choice	0.06733483
neighbor	0.10100225	criterion	0.06733483	design	0.06733483
cluster	0.08913647	move	0.06733483	lack	0.06733483
match	0.0854057	outcome	0.06733483	publish	0.06733483
require	0.08177003	space	0.06733483	warp	0.06733483
become	0.08177003	condition	0.06733483	subsequence	0.06733483
change	0.08177003	critic	0.06733483	warp	0.06733483

3.3 Similarity Calculation

We calculated the similarity using the cosine formula as follows:

$$(A=[A1,A2,...,An],B=[B1,B2,...,Bn])$$

$$\cos\theta = \frac{A \cdot B}{|A| \times |B|} = \frac{\sum_{i=1}^{n}(A_i + B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

Coupled with the growth of the value, the similarity level of two vectors was reduced. In this way, we were able to learn whether the two vectors were similar. In our system, every text in the TF-IDF matrix was considered a word vector, and the test word-frequency vector was another. We used this arithmetic to reveal the similarity level of the two texts. Table 2 revealed the similarity between the sentence and tags that we proposed in the system.

Table 2. Similarity between the sentences and tags.

Sentence	Background	Research problem	Solution	Result
1. It defines the possible world model with probability intervals, and proves that probability intervals of all possible worlds are feasible.	0.1007	0.0401	0	0.0427
2. The experiments on synthetic and real data sets show that the algorithms are effective and significant.	0	0.0420	0.0574	0.1727
3. It gives two lemmas for optimizing the computation of prevalence point probability of a candidate co-location.	0	0.0384	0.0686	0.0409

3.4 Annotation Based on Similarity and Position

Similarity was calculated by featuring words in Phase 6. We found that the difficult point was how to select the sentences from the abstract, which should add the research problem or solution tag. We discovered that the label of the sentence was relevant to their position. We processed information of abstract location using 88 articles from DMKD. The first one to two sentences were the background, the next one to two sentences were research questions, and then the solution was discussed in the next two to six sentences, and the final one to two sentences were the result. Different tags were labeled according to the location of the sentence. For example, for the third sentence mentioned in Phase 6, the system annotated the solution label to it because it is the fifth sentence of the paper's abstract and the sentence in front of it has been marked as the research problem.

4 Model Application

This paper studies how to annotate semantic structure to the abstract of scientific papers automatically, based on real data from DBLP DB. We developed a small experimental system in which the abstract text is the system's input and the abstract with four labels is the output. The system's interface is as shown in Figure 2.

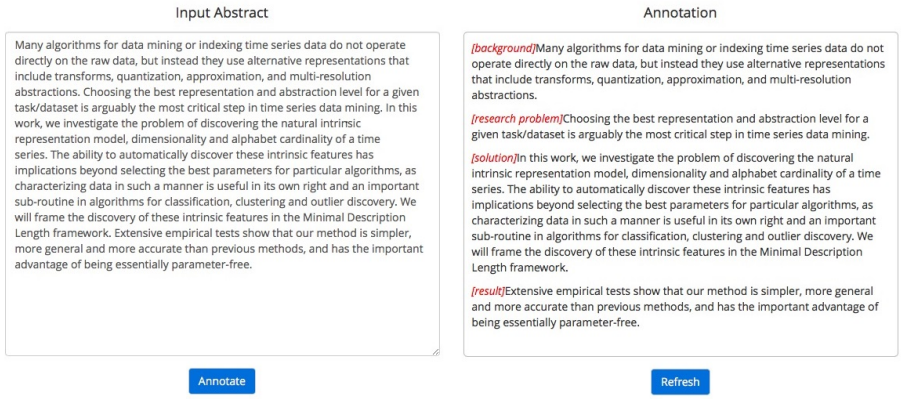

Fig. 2. System interface

Eighty articles were used to test the model proposed, the test results were as Table 3:

Table 3. Test results.

Tags	Background	Research problem	Solution	Result
Correct number	55	52	44	65
Error number	25	28	36	15
Correct rate	68.75%	65%	55%	81.25%

5 Conclusion

This paper established a model to annotate the semantic structure for scientific papers' abstracts automatically. The first step of the process is to build a rhetorical structure that includes background, research problem, solution, and result. Then, through feature extraction and calculating weight for every segment of an abstract, we attain the feature vector. Next, according to the cosine algorithm, the system computes the score of the similarity. Last, the abstract was annotated by similarity and sentence positions.

The correct rate the system proposed in this paper could only reach about 67.5% of the automatic annotation of the semantic structure, and the effect will be better if the ontology system is introduced in the future, and the rhetorical words of scientific articles in different journals and different areas may be different; therefore, the next focus is to improve the accuracy and create a cross-domain model.

6 Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 61300147), China Postdoctoral Science Foundation (No. 2014M551185), and Science and Technology Program of Changchun (No. 14GH014).

References

1. Ait-Mokthar, S., Chanod, J.P.: Robustness beyond shallowness: incremental dependency parsing. Special issue of NLE Journal (2002)
2. Corney, D.P., Buxton, B.F., Langdon, W.B., Jones, D.T.: Biorat: extracting biological information from full-length papers. Bioinformatics 20(17), 3206–3213 (2004)
3. De Waard, A., Kircz, J.: Modeling scientific research articles-shifting perspectives and persistent issues. In: Proc. ELPUB2008 Conference on Electronic Publishing. pp. 234–245 (2008)
4. Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J.A., et al.: Semtag and seeker: Bootstrapping the semantic web via automated semantic annotation. In: Proceedings of the 12th international conference on World Wide Web. pp. 178–186. ACM (2003)
5. Dor, K.: The rhetoric structure of research article abstracts in english studies journals. Prague Journal of English Studies 2, 119–139 (2013)
6. Garten, Y., Altman, R.B.: Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. BMC bioinformatics 10(Suppl 2), S6 (2009)
7. Groza, T., Handschuh, S., Möller, K., Decker, S.: Salt-semantically annotated\ mbox {\ LaTeX} for scientific publications. In: The Semantic Web: Research and Applications, pp. 518–532. Springer (2007)
8. Harmsze, F.A.P.: A modular structure for scientific articles in an electronic environment. University of Amsterdam (2000)
9. Shiffrin, R.M., B02rner, K.: Mapping knowledge domains. Proceedings of the National Academy of Sciences of the United States of America 101 suppl 1(Suppl 1), 5183–5185 (2004)
10. Shum, S.B., Clark, T., Groza, T., Handschuh, S., 09gnes Sndor: Scientific discourse on the semantic web: A survey of models and enabling technologies. Semantic Web Journal Interoperability Usability Applicability (2010)

11. Teufel, S., et al.: Argumentative zoning: Information extraction from scientific text. Ph.D. thesis, Citeseer (2000)
12. de Waard, A., Buitelaar, P., Eigner, T.: Identifying the epistemic value of discourse segments in biology texts. In: Proceedings of the Eighth International Conference on Computational Semantics. pp. 351–354. Association for Computational Linguistics (2009)
13. de Waard, A., Tel, G.: The abcde format enabling semantic conference proceedings. In: SemWiki (2006)