Scientific Article Search System Based on Discourse Facet Representation

Yuta Kobayashi,[1] **Hiroyuki Shindo,**[1,2] **Yuji Matsumoto**[1,2]
[1]Nara Institute of Science and Technology
[2]RIKEN Center for Advanced Intelligence Project (AIP)
{kobayashi.yuta.kp1, shindo, matsu}@is.naist.jp

Abstract

We present a browser-based scientific article search system with graphical visualization. This system is based on triples of distributed representations of articles, each triple representing a scientific discourse facet (Objective, Method, or Result) using both text and citation information. Because each facet of an article is encoded as a separate vector, the similarity between articles can be measured by considering the articles not only in their entirety but also on a facet-by-facet basis. Our system provides three search options: a similarity ranking search, a citation graph with facet-labeled edges, and a scatter plot visualization with facets as the axes.

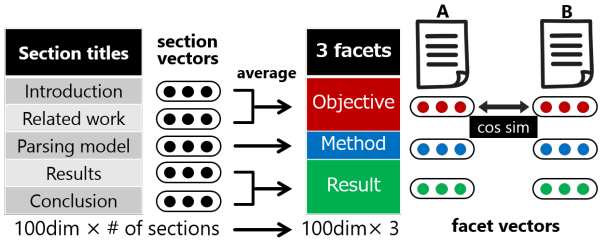

Figure 1: Facet-based article similarity calculation

Introduction

Finding relevant articles can be a challenge for scientists, because they are faced with a flood of digital publications. Therefore, there has been increasing interest in applying natural language processing technologies to scholarly document analysis. In search systems of scientific articles such as the ACL Anthology Searchbench[1] (Schäfer et al. 2011), once a search has been performed, the only option for expanding it is a citation browser that displays a citation graph.

One technology that can be used for expanding a search is discourse structure analysis of scientific literature (Teufel, Siddharthan, and Tidhar 2006), which seeks to automatically classify the body texts and citation contexts by their scientific discourse facets. Previous studies have identified common discourse facets in scientific articles, of which we focus on three: Objective, Method, and Result. If a system can recognize the discourse facets of articles, the similarity between articles can be measured not only based on their full text but also on a facet-by-facet basis. In literature retrieval systems, this opens the possibility for answering queries such as: "find an article with a different objective than the one at hand but with a similar methodology"; It is probable that such queries cannot be answered by a mere keyword search.

As the first step toward our goal, we focus on the steps after selecting an anchor article and provide three faceted search options: a similarity ranking search, a facet-labeled citation graph, and a scatter plot visualization. These search options are based on vector representations that encode not the full article, but each individual discourse facet. Our system enables users to search for scientific articles based on the similarity of discourse facets (Objective, Method, and Result), represented by multi-vector representations of both the text and citation graphs (Figs.1 and 2).

Steps for Building Article Facet Vectors

Step 1: Preprocessing dataset We crawled PDF files from the ACL Anthology and built a retrieval dataset. The body text was extracted from the PDFs as XHTML by our customized Poppler, and some collapsed articles were eliminated. To build the citation graph, we extract citations from the XHTML using regular expressions. This preprocessing resulted in 20,796 articles and 303,767 citation links. Next, we carried out unsupervised learning of the word vector representations using fastText (Bojanowski et al. 2016) with the original corpus composed of the text in English Wikipedia and the ACL Anthology.

Step 2: Section facet classification and learning facet vectors We classified sections of the articles using an annotation dataset for structured abstracts, which are summaries of articles comprising labeled sections for rapid comprehension. We used the National Library of Medicine Category Mappings file[2], a dataset attached to the medical article database MEDLINE. The file contains a list of 3,032 translation rules for canonicalizing various section titles appearing in structured abstracts into one of the facets. First,

[1]http://aclasb.dfki.de

[2]https://structuredabstracts.nlm.nih.gov/

we applied these rules to the section titles of the articles to obtain the facets of these sections. To simplify the model, we merged the original labels into three broad labels by following the categorization. If none of the rules applied to a section title, the corresponding section was labeled by the classifier fastText trained on the sections labeled by the rules.

We estimated the discourse facets of the unlabeled sections using a classifier, which received as input the body text of a section and output a section facet (Objective, Method, or Result). For each article, the sections corresponding to each facet were represented by 100-dimensional vectors. If more than one section corresponded to a facet, the vectors of those sections were averaged. Each article was assigned a 300-dimensional vector by concatenating its three facet vectors.

Step 3: Citation facet classification and citation graph augmentation In the next step, we augmented the citation graph by adding a citation facet to each citation edge using a supervised *fastText* classifier (Fig.2). This citation facet was determined by the textual context of the corresponding citation in the citing article. The edges were divided into the following three facets: Objective, Method, and Result. We used a dataset that included 1,618 citation contexts by combining the Citation Function Corpus (Teufel, Siddharthan, and Tidhar 2006) and CL-SciSumm-2017 SharedTask Corpus (Jaidka, Jain, and Kan 2017). We chose several examples from the combined dataset that could be classified into the three types of facets. The input of the classifier was a citation context, and the output was one of the three facets.

Step 4: Update text-based vectors with citation graph Using the graph to which citation facets were attached, an update was created for each facet using the facet vectors obtained in Step 2 as the initial values. Specifically, we used the graph embedding method LINE (Tang et al. 2015) to integrate the text and graph information. For each vertex k and a facet f, let $\mathbf{v}_{k,f}$ denote the facet vector with vertex k representing facet f. Updating the facet vectors was done by maximizing the following objective function for each edge (i, j) between vertices i and j on the citation graph:

$$\log \sigma\left(\langle \mathbf{v}'_{j,f}, \mathbf{v}_{i,f} \rangle\right) + \sum_{k=1}^{N} \log \sigma\left(-\langle \mathbf{v}'_{n(k),f}, \mathbf{v}_{i,f} \rangle\right),$$

where f is the facet with edge (i, j) determined in Step 3, $\sigma(\cdot)$ is the sigmoid function, $\mathbf{v}'_{j,f}$ is the context facet vector representing the outlink from vertex j. N is the number of "negative" vertex samples, and $n(k)$ $(k = 1, \ldots, N)$ are the negative vertices chosen from among the vertices in the citation graph according to a noise distribution $P_{\text{noise}}(v)$, proportional to the out-degree of vertex v. In this step, we only updated the facet vectors corresponding to an edge's citation facet. Therefore, the model optimizes each facet vector to maximize the inner product of the facet vectors.

By following these procedures, we obtained the facet vectors of the articles and used them for the search. Because the search goal was different, it was difficult to compare this faceted search to the keyword-based literature search.

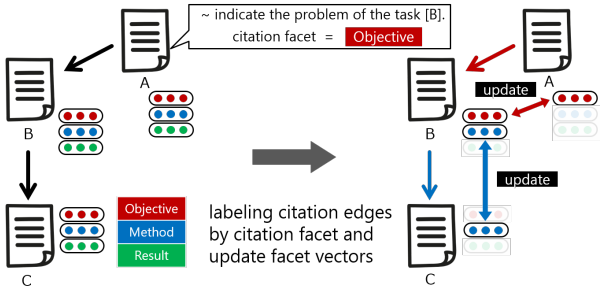

Figure 2: Citation graph augmentation and facet vectors

Demonstration: Faceted Literature Search

Ranking search with discourse facet By inputting a query, the system shows a list of articles matching the query. After selecting an article as an anchor article, the user can switch to a facet tab (Overall, Objective, Method, Result). The system then displays a new ranking based on the cosine similarity of the vectors of the chosen facet.

Citation graph based on citation facet By clicking on the citation button, the system displays a visualization of the citation graphs whose edges are color-coded according to facet, allowing the user to trace citations and understand research trends easily. The user can also adjust the number of vertices using sliders for minimum similarity or minimum degree. By clicking on a node, the user sees the emphasized citation edges closely related to both the anchor paper and the anchor paper's bibliographic information.

Scatter plot of articles with facet axes By clicking the view type button, users see a scatter plot based on t-SNE and can select two elements from the four options (Overall, Objective, Method, and Result) as the x- and y-axis. For example, the x-axis could represent method similarity and the y-axis objective similarity. This plot would help users to find articles using similar methods but different objectives.

Acknowledgments This work was partly supported by JST CREST Grant Number JPMJCR1513, Japan.

References

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Jaidka, K.; Jain, D.; and Kan, M.-Y. 2017. The CL-SciSumm shared task 2017: results and key insights. In *CL-SciSumm Shared Task 2017*, 1–15.

Schäfer, U.; Kiefer, B.; Spurk, C.; Steffen, J.; and Wang, R. 2011. The acl anthology searchbench. In *ACL*, 7–13.

Tang, J.; Qu, M.; Wang, M.; Zhang, M.; Yan, J.; and Mei, Q. 2015. Line: Large-scale information network embedding. In *WWW*.

Teufel, S.; Siddharthan, A.; and Tidhar, D. 2006. Automatic classification of citation function. In *EMNLP*, 103–110.