Citation Recommendation Using Distributed Representation of Discourse Facets in Scientific Articles

Yuta Kobayashi
Nara Institute of Science and Technology
Ikoma, Nara, Japan
kobayashi.yuta.kp1@is.naist.jp

Masashi Shimbo
Nara Institute of Science and Technology
Ikoma, Nara, Japan
shimbo@is.naist.jp

Yuji Matsumoto
Nara Institute of Science and Technology
Ikoma, Nara, Japan
matsu@is.naist.jp

ABSTRACT

Scientific articles usually follow a common pattern of discourse, and their contents can be divided into several facets, such as objective, method, and result. We examine the efficacy of using these discourse facets for citation recommendation. A method for learning multi-vector representations of scientific articles is proposed, in which each vector encodes a discourse facet present in an article. With each facet represented as a separate vector, the similarity of articles can be measured not in their entirety, but facet by facet. The proposed representation method is tested on a new citation recommendation task called *context-based co-citation recommendation*. This task calls for the evaluation of article similarity in terms of citation contexts, wherein facets help to abstract and generalize the diversity of contexts. The experimental results show that the facet-based representation outperforms the standard monolithic representation of articles.

CCS CONCEPTS

• **Information systems** → **Document representation**; **Document structure**; **Recommender systems**; *Link and co-citation analysis*; Information retrieval; • **Computing methodologies** → *Natural language processing*;

KEYWORDS

Scientific article; representation learning; natural language processing; discourse facet; co-citation analysis

ACM Reference Format:
Yuta Kobayashi, Masashi Shimbo, and Yuji Matsumoto. 2018. Citation Recommendation Using Distributed Representation of Discourse Facets in Scientific Articles. In *JCDL '18: The 18th ACM/IEEE Joint Conference on Digital Libraries, June 3–7, 2018, Fort Worth, TX, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3197026.3197059

JCDL '18, June 3–7, 2018, Fort Worth, TX, USA

ACM ISBN 978-1-4503-5178-2/18/06…$15.00
https://doi.org/10.1145/3197026.3197059

1 INTRODUCTION

1.1 Background

With the recent surge in the number of scientific publications, researchers are forced to spend a large amount of time finding relevant literature for their research. To reduce this workload, literature retrieval systems need to be enhanced beyond simple keyword search, in a way that enables more goal-oriented searches.

One key technology to this end is the discourse structure analysis of scientific literature [20, 25, 32, 33], which seeks to automatically classify citation contexts and the parts of body text (i.e., sentences, sections, etc.) by their functions in typical scientific discourse structure, such as Objective, Method, and Result. We call these functions *discourse facets* in this paper.

If the discourse facets of articles are correctly identified, the similarity of articles can be measured not based on their full text but on a facet-by-facet basis. For example, the similarity of two articles can be measured only in terms of methods studied, while ignoring objectives. In literature retrieval systems, this opens the possibility of answering queries such as "find an article with a different objective from the one at hand but with a similar methodology"; most likely, such queries cannot be answered by mere keyword search.

Likewise, if individual citation contexts can be classified into facets (e.g., "Is this citation made for the sake of a methodological argument?"), citation recommendation systems could take advantage of this information to make a recommendation that is consistent with the user-specified citation context.

1.2 Contributions

In this paper, we propose using the discourse facets associated with citation contexts and sections of textfor a new type of citation recommendation task, which we call *context-based co-citation recommendation* (CBCCR): Given a context in which citations should be made, as well as an article that is to be cited in that context, the system must recommend more articles to be co-cited with the given article in that context. This task requires the measurement of article similarity in light of a specific citation context. Another unique feature of CBCCR is that an evaluation dataset can be created nearly automatically from a corpus of scientific articles if we exploit the *enumerated co-citations* in them.

For CBCCR, we propose a technique for learning multi-vector distributed representations of scientific articles that takes into account the discourse facets of both text and citation graphs.

After automatically classifying all sections of the articles into three facets (Objective, Method, and Result), we train a distributed vector representation [3, 22] for each facet present in an article.

Thus, each article is represented by a set of vectors, each of which represents a single facet in that article. Since facets are represented as separate vectors, we can measure the similarity of two articles using only the vectors of a specific facet of interest: for example, how similar two articles are in terms of Method, while ignoring their similarity in Objective and Result.

Also, in CBCCR, these facets help to abstract and generalize the diversity of citation contexts. A CBCCR experiment using the constructed dataset shows that our facet-based article representations are superior to monolithic representations of articles.

1.3 Terminology

As mentioned earlier, we use the term *discourse facet* to refer to a functional role in typical discourse patterns observed in scientific articles, such as Objective, Method, and Result. To our knowledge, this term was originally introduced for the classification of citation contexts [15, 34]. Similar concepts have been referred to by various names in the literature, although the focus, granularity, and scope may vary: for example, *rhetorical status* [32] (for general sentences), or *section function* [20] (for sections), and *citation function* [33] (for citation contexts). This concept also roughly corresponds to the sections used in *structured abstracts* [13, 23, 26].

An *enumerated co-citation* is a set of citations made as a group, typically quoted in a single pair of parentheses or brackets. A *citation context* is the context in which a citation (or an enumerated co-citation) occurs. Here, the "context" merely refers to some string of words adjacent to the location of the citation. The concrete span of a context is task/method-dependent: for example, the span could be a sentence, a paragraph, or words in a fixed window around the citation.

2 RELATED WORK

2.1 Vector representation of scientific articles

In document retrieval, documents are represented as vectors of word frequency–based values, and the similarity of documents is measured by suitable operations (such as cosine) over their vectors. This "bag-of-words" vector representation has also been applied to scientific articles to capture the similarity of their contents. For example, Sugiyama and Kan [30] represented articles as vectors using a tf-idf-based method, and constructed a system that recommends articles on the basis of a list of user's past publications.

In recent years, distributed (dense) word representation [3, 22] has been attracting NLP researchers' attention due to its compatibility with neural networks. These real-valued word vectors can be used as features in a variety of NLP applications, such as machine translation [2], summarization [28], question answering [14], and text classification [18]. It has also been applied to the classification of citations in scientific articles [24].

2.2 Vector representation of citation graphs

Tracking citations provides a way to analyze the relationship between articles. Recently, the distributed representation of graphs (also called *graph embedding*) [11, 27, 31] has been rigorously pursued as a promising approach to modeling graph data including citations. It assumes that vertices capture shared similarities in local graph structure such as co-citation similarity, and vertices with similar neighborhoods will acquire similar representations. For vector representation of scientific articles, Tang et al. [31] computed the vectors of articles not from the text but from a citation graph and classified them into seven computer science conferences to which the articles were submitted: SIGIR, KDD, AAAI, CIKM, ICML, NIPS, and WWW.

In the existing methods for graph-based representation learning, the quality of vectors of the vertices and the learning time are in a trade-off relationship. In this study, we use LINE [31], which is one of the fastest graph embedding methods, as a building block in learning the facet-based representation of articles.

2.3 Capturing discourse facets in text

There have been studies on the classification of text and citation in scientific literature into discourse facets.

Kafkas et al. [20] proposed a rule-based section tagger to assign a function label (rhetorical status) to each section, i.e., Introduction & Background, Materials & Methods, Discussion, and Conclusion & Future Work. Although their rules were handcrafted, they stated that a machine-learning method was more desirable to deal with sections that are not covered by their rules.

Argumentative zoning is the task of automatically classifying sentences in scientific documents by their rhetorical status in scientific discourse [6, 12, 32], such as Aim, Background, Own, Contrast, and Basis. Teufel and Moens [32] used a naive Bayes classifier, Guo et al. [12] used naive Bayes and Support Vector Machine classifiers, and Contractor et al. [6] used a positive-unlabeled (PU) learning technique [8].

In library sciences, the use of information from citation contexts has received considerable attention. Multiple schemes have been proposed for how to classify contexts from many different points of view: by citation function [25, 33] that shows the reason for citation, by polarity (positive/negative) [5], or by importance (important/not important) [35], among others.

The facets of citation contexts have been recognized as useful information for summarizing scientific documents. A series of shared tasks on scientific document summarization continue to provide a subtask (Task 1b) for classification of citation sentences ("citances") into discourse facets: TAC 2014 [34] (CL-SciSumm Pilot 2014), CL-SciSumm 2016 [15], and CL-SciSumm 2017 [16]. Distributed word representation has also been used for facet classification recently. Munkhdalai et al. [24] used neural attention models and distributed word representation to classify citation contexts.

2.4 Co-citation analysis

Co-citation analysis [29] measures the relatedness between two articles by the number of co-citations (joint citations) they receive from articles authored by peer researchers. It is often considered a more objective (and hence reliable) indicator of relatedness than direct citations, which are made subjectively by the article authors themselves.

In co-citation analysis, the strength of individual co-citations is neglected; one joint citation from an article uniformly provides one unit of contribution to the co-citation index between cited articles. However, recent studies have suggested the need to distinguish the

amount of contribution provided by individual co-citations. Gipp and Beel [10] analyzed the strength of the relationship between co-cited articles in terms of the proximity of the location of citations made in the text of an article and indicated that the strongest relationship belongs to co-citations made in the same sentence. They proposed the *co-citation proximity index* as a measure of relatedness between articles by taking the relative location of co-citations. Eto [9] pointed out that relatedness of co-cited articles is highest for those in enumerated co-citations. These findings provide the foundation of the CBCCR task we propose; the task is based on the assumption that articles in an enumerated co-citation bear high similarity, at least from a certain point of view as expressed in the citation context.

2.5 Citation recommendation

Citations are made with different motivations and in different contexts. If such a context is available, a citation recommendation system must output a list of articles that are similar to the input article not in a generic sense but in light of the specific context of citation.

Duma and Klein [7] defined a context-based citation recommendation system as one that assists the author's writing a document by suggesting other documents which are relevant to a particular context in the draft. They proposed *citation resolution* as a pilot task of context-based recommendation. In citation resolution, the system is given a citation context extracted from an article (with the actual citation hidden), as well as the list of articles collected from the References section of the same article. The system then has to choose the actual article (i.e., the one hidden from the system) cited in the given citation context from the list of articles in References.

By definition, our proposed CBCCR task (which is described in Section 3) falls into a task of context-based citation recommendation, and also bears a certain similarity to the citation resolution task. However, CBCCR departs from citation resolution in three ways. First, CBCCR deals only with enumerated co-citations but not singletons. Second, CBCCR extends the input from only a regular citation context to an enumerated co-citation context and the first article in an enumerated co-citation. Third, in CBCCR, the system has to choose articles from a large database of articles, whereas in citation resolution, the candidates are a small set of articles found in the References section of the article from which the citation context is collected. Indeed, citation resolution is not exactly a citation recommendation task, because the candidates are limited to those in References.

Of the three, the presence of a co-cited article as input is a crucial difference between CBCCR and citation resolution. As discussed in Section 2.4, co-cited articles are assumed to share high similarity from the point of view stated in the context of citation. Thus in CBCCR, the context-dependent similarity between the given article and the candidates makes a key factor in determining a recommended article. Such similarity is not a factor in citation resolution, because no co-cited articles are given to the system.

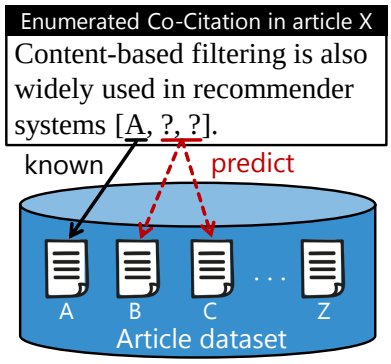

Figure 1: Context-based co-citation recommendation.

3 CONTEXT-BASED CO-CITATION RECOMMENDATION

In this section, we describe a new citation recommendation task, which we call *context-based co-citation recommendation* (CBCCR), and a dataset for this task. As mentioned earlier, CBCCR is intended as a simulation of the following scenario: A user inputs to a system a textual context of citation and a single article that she thinks is suitable to be cited in that context, and then the system suggests additional citations that are to be cited along with the input article.

3.1 Task definition

The task of CBCCR is defined as follows. The system is given a set $\mathcal{D}$ of articles (an article database), a context c of an enumerated co-citation, and the first article a_1 of the multiple articles cited in context c. The system is then required to choose from $\mathcal{D}$ the hidden articles $a_2, \ldots, a_n$ that are cited alongside a_1 in the context c. In practice, the system outputs the ranking of candidates for the hidden articles. Thus, the system is not required to determine the number n of hidden co-citations.

Figure 1 shows a schematic illustration of CBCCR. In this figure, articles A, B and C are cited by article X as an enumerated co-citation, and the article database $\mathcal{D} = \{B, C, \cdots, Z\}$ contains all articles except for A. The system is presented with A ($= a_1$) and its citation context c (= *"Content-based filtering is also widely used in recommender systems"*), and must predict B and C (a_2 and a_3, respectively) as the most suitable citations in this context, over the remaining articles $D, \ldots, Z$ in the database. The system can explore two sources of information: (1) full texts of the articles in the dataset (except for the references in the article from which the enumerated co-citation was collected[1]; i.e., article X in Figure 1), and (2) citation graphs (from which the edge $X \rightarrow B$ and $X \rightarrow C$ are excluded) to make a prediction.

CBCCR has the following advantages:

[1] References are removed from the citing article so as not to make the problem too easy. If references were present, the system would only need to choose articles from them, not from the entire repository of articles. This removal is in line with the intended scenario of the CBCCR task described at the beginning of Section 3: We assume that the user is not yet aware of the articles to cite, except for the one she has already input to the recommendation system.

- Since enumerated co-citations can be extracted automatically (e.g., by pattern matching with regular expressions over text), a large collection of sets of articles can be created without human intervention.
- Articles that are enumeratively co-cited share a citation context, which means that they are similar from a distinct point of view given by the context. This makes CBCCR as a good benchmark for context-based similarity evaluation.

3.2 Dataset

We created a dataset for CBCCR, which consists of (i) the set of the article IDs defined in the ACL Anthology[2], an open-access repository of journal, conference, and workshop articles in the field of computational linguistics/natural language processing, (ii) its citation graph, and (iii) the contexts of 1,000 enumerated co-citations in these articles coupled with the IDs of cited articles. Table 1 shows the statistics of the dataset.

To built the dataset, we first crawled the ACL Anthology for the PDF files of articles that are published in or before 2016. These PDF files were then converted to XHTML, with a customized version of the PDF rendering library Poppler[3]. From those that were successfully processed by Poppler, we finally obtained 20,496 conference/journal/workshop articles that satisfy all of the following conditions:

- The section structure of the article is correctly extracted.
- The article has a reference section at the end.
- The article has three or more sections.

These rules were designed to filter out non-conference/journal/workshop papers, such as the preface of a special issue of a journal, or the title page and the table of contents of proceedings, of which the ACL Anthology also contains a small amount. These types of articles are typically not sectioned or have only a few sections, or do not have a reference section.

In the collected articles, citation and enumerated co-citations were detected by pattern matching. Some of the patterns matched are as follows:

- `(AUTHOR1 YEAR1; AUTHOR2 YEAR2 ...)`
- `[REFERENCE_NUMBER1, REFERENCE_NUMBER2 ...]`

In addition, a citation graph was also constructed by rule-based reference matching. This citation graph is self-contained, in the sense that it only contains edges (i.e., citations) between articles included in the constructed dataset, and those of external articles are removed.

In the evaluation of CBCCR systems, edges in the citation graph from each citing article to its hidden enumeratively co-cited articles (second and subsequent citations) should be removed; otherwise, the answer can often be found merely by looking into the citation graph. However, removing the edges (citations) in all the enumerated co-citations would greatly reduce the size of the citation graph. To avoid this large collapse of citation graph, we randomly sampled 1,000 contexts out of entire 34,760 enumerated co-citation contexts in which an enumerated co-citation appears at only one place in one sentence, and used them for evaluation.

[2]http://aclweb.org/anthology/
[3]https://poppler.freedesktop.org/

Table 2 shows the histogram of the number of cited articles in 1,000 enumerated co-citation contexts in the evaluation dataset. The average number of citations in an enumerated co-citation is 2.38. Thus, for each context, the number of hidden articles is 1.38 on average.

4 FACET-BASED METHOD FOR CO-CITATION RECOMMENDATION

In this section, we describe how to learn the facet-based vector representations of articles, and how they can be used for CBCCR.

4.1 Learning article representations

Figure 2 shows the overview of learning the representations of articles. The inputs of learning model are the texts of articles in ACL Anthology, the corresponding citation graph and the text of Wikipedia. In Steps 1 to 3, the discourse facets of articles are extracted from the body text and vectorized. In Step 4, we label the edges of citation graphs according to each facet predicted by a trained classifier. Finally, in Step 5, the vector representations of the text and the graph are integrated.

4.1.1 *Step 1: Learning word representations.*

In the first step, the representations of words are trained, so that they can be used in the subsequent steps. To learn word representations, we use the representation learning library fastText, which provides modules for unsupervised text representation learning [4]. The vectors produced by fastText encode not only word-level information but also word n-gram and character n-gram (subword) information. We chose fastText because the use of word n-grams is advantageous to capture discriminating lexical structures [1], and the n-gram information of word and character should help learn vector representations of technical terms.

Similarly to Mikolov's skip-gram model [22], fastText learns vector representations of words in vocabulary $\mathcal{W}$ given a large training corpus. To be precise, for each sentence $\mathcal{S} = w_1 \dots w_T$ in the corpus with $w_j \in \mathcal{W}$ for $j = 1, \dots, T$, it tries to find the vector representation for each word $w \in \mathcal{W}$ that maximizes the log-likelihood expressed by the following formula:

$$\sum_{t=1}^{T} \sum_{c \in C_t} \log p\,(c \mid w_t),$$

where C_t is the *context* of word w_t, which is the set of words surrounding w_t in the training sequence S, and $p\,(c \mid w_t)$ is the probability of observing a context word c given w_t. It is defined by

$$p\,(c \mid w_t) = \frac{\exp(s\,(w_t, c))}{\sum_{c' \in \mathcal{W}} \exp(s\,(w_t, c'))},$$

where $s\,(w, c)$ is a score function which maps a pair of words w and c to a score in $\mathbb{R}$, and is expressed by the following expression:

$$s\,(w, c) = \sum_{g \in \mathcal{G}_w} \langle \mathbf{z}_g, \mathbf{v}_c \rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes inner product, $\mathcal{G}_w$ is the set of character n-grams appearing in w and word w itself, and $\mathbf{v}_c$ is the vector of context word c. Associating a vector representation $\mathbf{z}_g$ to each character n-gram g in word w, we represent a word by the sum of the vector representations of its n-grams.

Table 1: Statistical information of CBCCR dataset extracted from ACL Anthology.

Data	#
Articles (vertices in the citation graph)	20496
Citations (edges)	259743
Average degree of vertices	29.6
Total enumerated co-citation contexts	34760
Randomly sampled enumerated co-citation contexts (test data)	1000

Table 2: The distribution of the number of cited articles in 1,000 enumerated co-citations.

# cited articles	frequency
2	722
3	210
4	53
5	7
6	1
7+	2
Total	1000

The facet-based vector representations of articles are built through following steps. First, we extract the title and the text of each section from the dataset we built (see Section 3.2). Then, we build a corpus that combines the English Wikipedia corpus[4] and the text of the articles extracted from the ACL Anthology. We carry out unsupervised learning of word vector representations using fastText [4] with this corpus. These pre-trained word representations are used for the following classification step.

4.1.2 Step 2: Classifying sections by facet using NLMCM rules and a machine learning–based classifier.

In this step, we classify each section of the NLP articles using an annotation dataset for structured abstracts. In a preliminary classification experiment, we found that the word vectors produced by fastText were more accurate than tf-idf or Paragraph Vectors [21].

In classification, we thus use the single layer neural network classifier for text classification provided by the fastText library [18] which uses the pre-trained word vectors obtained in Step 1. We use the resource developed for structured abstracts to train the fastText classifier. A structured abstract is a summary of an article that consists of labeled sections to make it easy to comprehend. As a dataset dealing with structured abstracts, we use the National Library of Medicine Category Mappings (NLMCM) file [26], which is a dataset attached to the medical article database PubMed. The NLMCM file holds a list of 3,032 translation rules for canonicalizing various section titles appearing in structured abstracts into one of Background, Objective, Method, Results, or Conclusions.

First, we applied these rules to the section titles of the articles collected in Step 1 to obtain the facet of these sections. To simplify the model, we use only three facets. The labels of Background and Objective are merged into Objective, and Results and Conclusions are merged into Result. This simplification roughly follows the definition of citation function scheme suggested by Nanba and Okumura [25], and we use these 3 facets (Objective, Method, and Result) as the labels of section facets instead of the original five labels in NLMCM. For example, when the section title is "Introduction," it is first classified into Background based on the NLMCM rules and then classified into Objective in the three label scheme.

If none of the NLMCM rules applies to the section title, that section is left unlabeled; they are subsequently labeled by the classifier fastText [18] trained on the sections labeled by the NLMCM rules. The classifier receives as input the body text of a section, and outputs a section facet (Objective, Method or Result). In this time, each word representation is updated by the information of facet label.

Then, we estimate section facets of the 11,118 unlabeled sections by the classifier with the section titles and texts as input.

4.1.3 Step 3: Building facet vectors for each article.

In this step, we first compute the vector representations of each section in an article, which is simply the average of the vector representations of all words in the section. The word representation vectors used here are those initially trained in Step 1 and then refined for facet classification in Step 2.

We then compute the facet vectors of an article for Objective, Method, and Result by taking the average of all sections deemed as each of these facets in that article.

In the next step, we augment the citation graph by adding a discourse facet to each citation edge using a supervised classifier in the fastText library. In the final step, each facet vector of the article is updated by the graph embedding method LINE [31]. These two steps are optional for using the information of citation graphs and citation function. However, the learning classifier for citation context is necessary for predicting the citation facet of enumerated co-citation context in the CBCCR task.

4.1.4 Step 4: Classifying citations by facet (optional).

In this step, we classify citation contexts in the articles into discourse facets, and label corresponding edges in the citation graph (see Section 3.2) by the facets.

To make a training data for the citation context classifier, we combined two datasets, the Citation Function Corpus [33] annotated with 12 citation function labels and the corpus of CL-SciSumm-2017 SharedTask [16] annotated with 5 discourse facet labels. To be precise, 1,618 examples were extracted from the combined corpora,

[4]https://dumps.wikimedia.org/backup-index.html

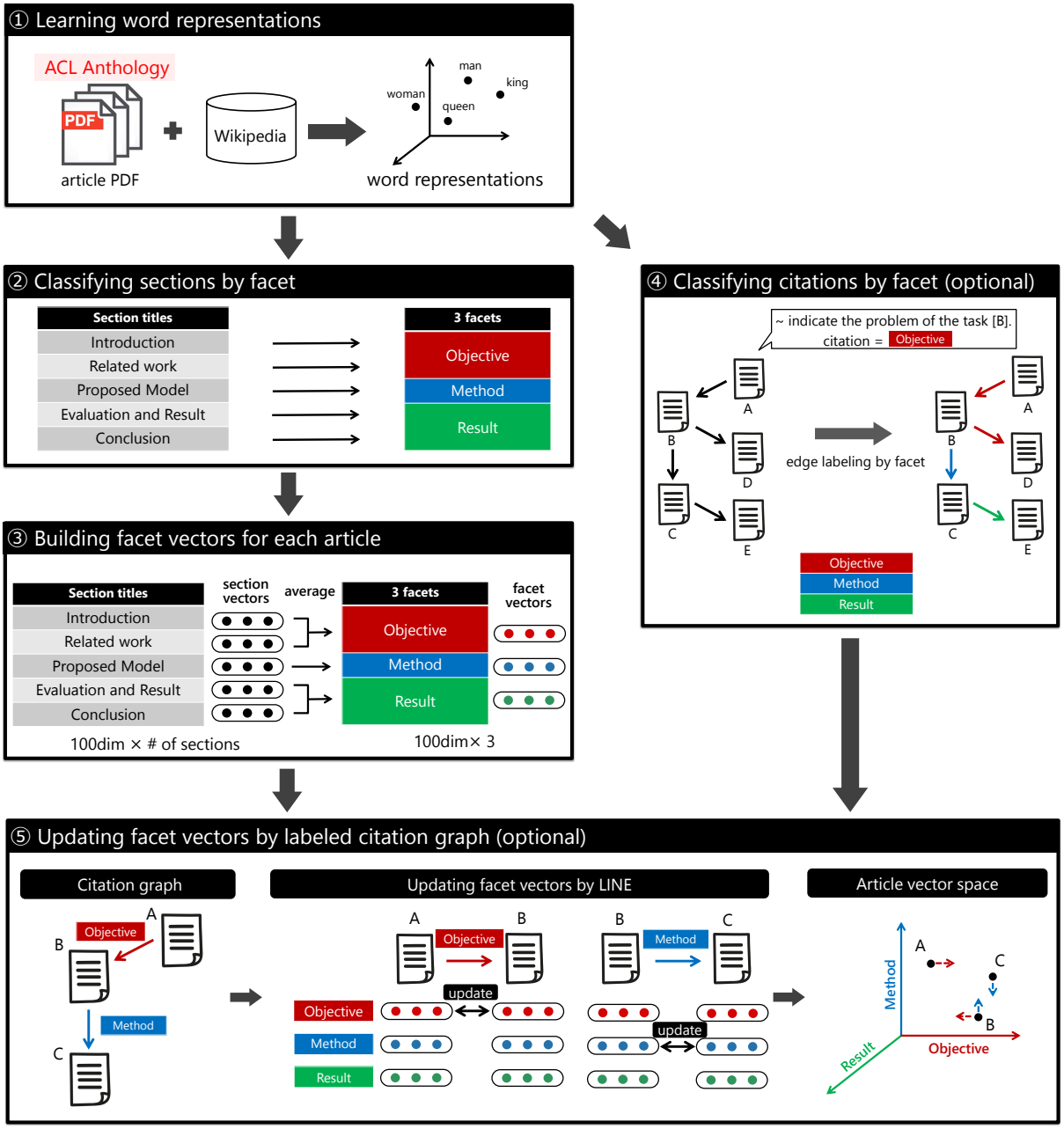

Figure 2: Overview of learning multi-vector representations of articles based on discourse facet.

such that their original function/facet labels straightforwardly correspond to the three facets used in this study. Table 4 shows the distribution of citation facets in the constructed annotation dataset.

To classify contexts into citation facets, we use the supervised classifier module of the fastText library, and train a classifier on the dataset constructed above. The input of the classifier is a citation context, and the output is one of the three citation facets, Objective, Method, and Result.

This discourse facet annotated to the citation graph is used for refining the facet vectors of articles, as we explain shortly in Step 5, where both the text information and graph information are integrated. Further, this context classifier is used subsequently in co-citation recommendation; see Section 4.2.

4.1.5 Step 5: Updating text-based facet vectors by labeled citation graphs (optional).

After the edges in the citation graph is labeled with facets in Step 4, we can refine the facet vectors obtained in Step 3 using this citation graph. Specifically, we use the graph embedding method LINE [31], which performs edge-sampling on the graph to reflect the local and global graph structure in the vector embedding of vertices.

Let us take the example depicted in Step 5 of Figure 2. The algorithm checks if there is an edge between vertices (articles) A and B, and finds that there is one and is labeled by facet "Objective." It then updates the "Object" vectors of articles A and B so that their inner product is increased. Likewise, since the edge from article B to C has the Method facet, the facet vector for Method is updated for B and C.

For each vertex k and a facet f, let $\mathbf{v}_{k,f}$ denote the facet vector of vertex k representing facet f. Updating the facet vectors is 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 of edge (i, j) determined in Step 4, $\sigma(\cdot)$ is the sigmoid function, $\mathbf{v}'_{j,f}$ is the complementary facet vector[5] 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 sampled from all the vertices in the citation graph according to a noise distribution $P_{\text{noise}}(v)$, i.e., $n(k) \sim P_{\text{noise}}(v)$. Following the original LINE procedure, we set $P_{\text{noise}}(v) \propto d(v)^{\frac{3}{4}}$, where $d(v)$ is the out-degree of vertex v.

By following these procedures, we finally obtain a vector representations of an article that accounts for discourse facets.

4.2 Co-citation recommendation using facet-based representation

Figure 3 shows how our proposed model ranks the candidate articles using enumerated co-citation context. First, we classify the input enumerated co-citation context into a discourse facet, using the classifier learned in Section 4.1.4. Then, the similarity between the first cited article and each candidate article is calculated by their cosine similarity of the facet vector corresponding to the facet (Objective, Method, or Result) predicted by citation context classifier.

Finally, all articles are sorted in decreasing order of this similarity, and the resulting ranking list is output.

5 EXPERIMENT

5.1 Evaluation of section classification

To verify the accuracy of section facet classification by fastText, a 5-fold cross-validation was carried out using 72,721 section titles labeled by the NLMCM rules in Section 4.1.2. The number of dimensions of the word vector is 100, and the window size and the parameter of the negative sampling are both 5. The parameter of word n-gram was one of the uni-gram, bi-grams, tri-grams.

[5]This complementary vector is an analogue of the "context" word vectors (as opposed to the "central" word vectors) in Mikolov's distributed word representation models [22]. Complimentary vectors are used only during training, but not for co-citation recommendation; they are simply discarded after training.

Table 3: Result of section facet classification by fastText.

fastText n-gram	Accuracy (%)
uni-gram	89.6
bi-gram	94.3
tri-gram	**96.1**

Table 4: Distribution of citation facets.

Citation facet	#
Objective	200
Method	1214
Result	204
Total	1618

Table 3 shows the cross-validation results of section facet classification with each n-gram. We confirmed that the section can be classified with a high accuracy of 96.1 percent in case of tri-grams. Hereafter, we use the word vectors and the classifier based on tri-grams.

5.2 Evaluation of citation facet classification

We performed the 5-fold cross-validation of citation facet classification by fastText with the annotated citation context dataset (Table 4) and obtained an accuracy of 90.1 percent.

5.3 Evaluation of article representations on a CBCCR task

We compared the performance of the proposed facet-based vector representations with monolithic representations (fastText and LINE) on CBCCR, using the dataset explained in Section 3.2 (see also Table 1 for the statistics of the dataset).

As the evaluation metric, we use the normalized discounted cumulative gain (nDCG) [17]. nDCG is based on the rank in the top k of the true cited article in the list of recommended articles by a system. with its score decreasing logarithmically. We set $k = 100$ in this experiment.

5.3.1 Baseline. We set two types of baseline article vectors. We used the average vector of all words in the article calculated by unsupervised fastText [18] obtained in the first step of the proposed method (see Section 4.1.1) as the text-based baseline. In addition, we compare LINE [31] as a graph-based representation vector (see Section 4.1.5). The value of negative sampling parameter is five and the window size is five. This setting is the same as that of the proposed methods, but since it does not have each component of the article, the dimension of both vectors is 300.

5.3.2 Proposed methods. We prepared facet-based vector representations in two ways. The first method, which we denote "proposed method 1: fastText + section facet", uses the representation

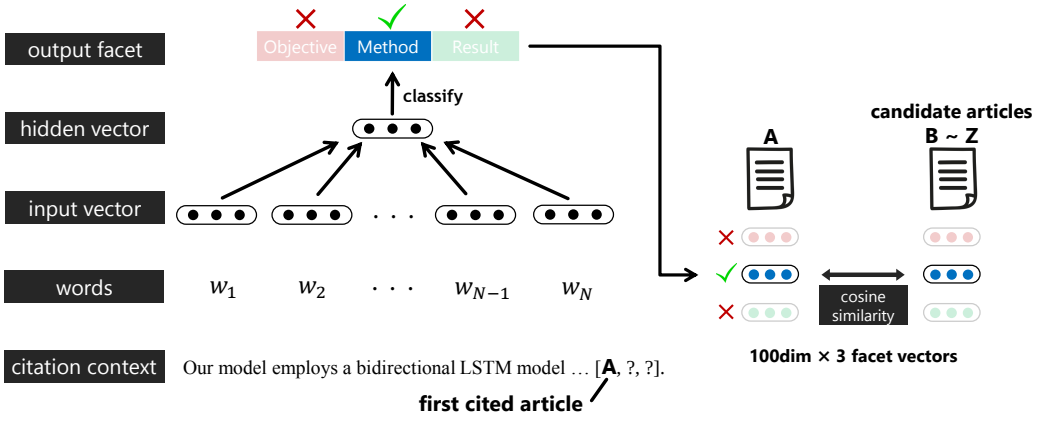

Figure 3: Co-citation recommendation using the citation context classifier and the facet-based representation of articles.

vectors obtained in Section 4.1.3. That is, it only uses Steps 1–3 to train facet vectors.

The second method, which we denote "proposed method 2: fastText + section facet + LINE + citation facet", is obtained through Steps 1–5. Thus, we further refine the vectors of the proposed method 1 with the faceted citation graph obtained in Step 4, using the graph representation learning method LINE. In Step 5 (see Section 4.1.5), the number N of negative samples per edge is set to 5.

The facet vectors in both proposed methods are set to 100 dimensions. Since we have three facets (Objective, Method, Result), an article is represented by a total of 300 dimensions; i.e., the same dimension as the baseline monolithic representations.

5.3.3 Experimental results. Table 5 shows the results of CBCCR using vector representations of the articles. We see that the accuracy of the proposed method 1 (fastText + section facet) using the information of the facet of the article improved the nDCG score compared to the baseline methods (fastText and LINE). We also observe further improvement of accuracy of proposed method 2 (fastText + section facet + LINE + citation facet), which uses not only text information but also the information of citation graphs. These results indicate that the combination of the discourse facet information of texts and citation graphs is effective for calculating the similarity between scientific articles.

6 CONCLUSION

In this study, we presented a technique for learning multi-vector representations of scientific articles, which combines the distributed representations of text and citation graphs, with each vector capturing a discourse facet (Objective, Method, and Result) within an article. In addition, we proposed a new task called *context-based co-citation recommendation* (CBCCR), which is unique in that it requires the evaluation of context-dependent similarity between articles.

We conducted experiments on our original CBCCR dataset and found that our discourse facet–based representation of articles achieves an improvement over monolithic representations.

Our future work is to utilize information from wider contexts, such as the entire section of an article in which a citation resides. We see this direction to be promising, as previous work [19] indicates that the distribution of the facets of citations differs depending on the sections in which the citations appear.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] Awais Athar. 2011. Sentiment Analysis of Citations using Sentence Structure-Based Features. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics 2011 Student Session*. 81–87.
[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
[3] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research* 3 (2003), 1137–1155.
[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* (2016).
[5] Donald O Case and Georgeann M Higgins. 2000. How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the Association for Information Science and Technology* 51, 7 (2000), 635–645.
[6] Danish Contractor, Yufan Guo, and Anna Korhonen. 2012. Using Argumentative Zones for Extractive Summarization of Scientific Articles. In *Proceedings of COLING 2012*. 663–678.
[7] Daniel Duma and Ewan Klein. 2014. Citation Resolution: A method for evaluating context-based citation recommendation systems. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 358–363.
[8] Charles Elkan and Keith Noto. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 213–220.
[9] Masaki Eto. 2012. Spread co-citation relationship as a measure for document retrieval. In *Proceedings of the fifth ACM workshop on Research advances in large digital book repositories and complementary media*. 7–8.
[10] Bela Gipp and Joeran Beel. 2009. Citation Proximity Analysis (CPA) : A New Approach for Identifying Related Work Based on Co-Citation Analysis. In *Proceedings of the 12th International Conference on Scientometrics and Informetrics, vol. 1*. 571–575.
[11] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 855–864.
[12] Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins, Lin Sun, and Ulla Stenius. 2010. Identifying the Information Structure of Scientific Abstracts: An

Table 5: Result of context-based co-citation recommendation using multi-vector representations of articles.

Type of article vector representations	nDCG@100
fastText	0.44
LINE	0.46
Proposed method 1 (fastText + section facet)	0.51
Proposed method 2 (fastText + section facet + LINE + citation facet)	**0.58**

Investigation of Three Different Schemes. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*. 99–107.

[13] R. Brian Haynes, Cynthia D. Mulrow, Edward J. Huth, Douglas G. Altman, and Martin J. Gardner. 1990. More informative abstracts revisited. *Annals of Internal Medicine* 113, 1 (1990), 69–76.

[14] Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III. 2014. A neural network for factoid question answering over paragraphs. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 633–644.

[15] Kokil Jaidka, Muthu Kumar Chandrasekaran, Sajal Rustagi, and Min-Yen Kan. 2018. Insights from CL-SciSumm 2016: the faceted scientific document summarization shared task. *International Journal on Digital Libraries* (2018). To appear. Online version available at https://doi.org/10.1007/s00799-017-0221-y.

[16] Kokil Jaidka, Devanshu Jain, and Min-Yen Kan. 2017. The CL-SciSumm shared task 2017: results and key insights. In *Proceedings of the Computational Linguistics Scientific Summarization Shared Task (CL-SciSumm 2017), organized as a part of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2017)*. 1–15.

[17] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[18] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759* (2016).

[19] David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2016. Citation classification for behavioral analysis of a scientific field. *arXiv preprint arXiv:1609.00435* (2016).

[20] Şenay Kafkas, Xingjun Pi, Nikos Marinos, Andrew Morrison, Johanna R McEntyre, et al. 2015. Section level search functionality in Europe PMC. *Journal of biomedical semantics* 6, 1 (2015), 7.

[21] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. 1188–1196.

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[23] Cynthia D. Mulrow, Stephen B. Thacker, and Jacqueline A. Pugh. 1988. A proposal for more informative abstracts of review articles. *Annals of Internal Medicine* 108, 4 (1988), 613–615.

[24] Tsendsuren Munkhdalai, John Lalor, and Hong Yu. 2016. Citation analysis with neural attention models. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*. 69–77.

[25] Hidetsugu Nanba and Manabu Okumura. 1999. Towards Multi-paper Summarization Using Reference Information. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*. 926–931.

[26] U.S. National Library of Medicine. Oct 26, 2015. Structured Abstracts in MEDLINE. (Oct 26, 2015). https://structuredabstracts.nlm.nih.gov/

[27] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 701–710.

[28] Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685* (2015).

[29] Henry Small. 1973. Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of American Society for Information Science* 24 (1973), 265–269.

[30] Kazunari Sugiyama and Min-Yen Kan. 2015. A Comprehensive Evaluation of Scholarly Paper Recommendation Using Potential Citation Papers. *International Journal on Digital Libraries* 16, 2 (2015), 91–109.

[31] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*. 1067–1077.

[32] Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics* 28, 4 (2002), 410–445.

[33] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing*. 103–110.

[34] Text Analysis Conference 2014. Text Analysis Conference 2014 Biomedical Summarization Task. https://tac.nist.gov/2014/BiomedSumm/index.html. (2014).

[35] Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying Meaningful Citations. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.