Doctoral Thesis

Citation Knowledge Mining for On-the-fly Recommendations

Yang Zhang

February 2022

Department of Social Informatics
Graduate School of Informatics
Kyoto University

Doctoral Thesis
submitted to Department of Social Informatics,
Graduate School of Informatics,
Kyoto University
in partial fulfillment of the requirements for the degree of
DOCTOR of INFORMATICS

Thesis Committee: Qiang Ma, Associate Professor
Keishi Tajima, Professor
Shinsuke Mori, Professor

Citation Knowledge Mining for On-the-fly Recommendations*

Yang Zhang

Abstract

When writing an academic paper, one of the most frequent questions considered cloud be: "Which paper should I cite at this place?" However, searching and finding suitable articles for referencing from a massive number of publications has become a challenge for researchers. According to the STM scholarly publishing report, up to 2018, there were 150 million articles in total published in the Web of Science databases. In addition, the number of newly published papers is also growing at 5-6% per year in recent years. It could imagine that researchers would not be able to find and read all the potential papers relevant to their studies. How to effectively recommend appropriate papers to scholars have become a non-trivial problem.

Researchers are currently relying on keyword-based systems to search for appropriate sources (such as Google Scholar). The recommendations are made by matching the title of papers with the input keywords from the users. Nevertheless, the input keywords might be over-abbreviated to fully demonstrate the user's searching need, which might lead to two potential drawbacks:

- First, keyword-based systems might be **inefficient** (or time-consuming) to be used. The users often have to try different combinations of words and read the candidate papers extensively to pick the correct paper. Much of the users' time is wasted trying different combinations of keywords and reading necessary papers.

- Second, keyword-based systems often lead to **inaccurate** recommendations. Keyword-based systems merely match the title of candidate papers with the input keyword; the titles of the target papers do not always contain

*Doctoral Thesis, Department of Social Informatics, Graduate School of Informatics, Kyoto University, KU-I-DT6960-30-2528, February 2022.

the input keywords. For example, when users want to find the original paper proposed "Word2Vec" algorithm, they might intuitively use the word "Word2Vec" as the query keyword. However, the original paper comes with the title "Efficient Estimation of Word Representations in Vector Space" does not contain the keyword "Word2Vec"; as a result, Google Scholar could not find the correct paper from our trials.

In this research project, we propose the concept of "on-the-fly" citation recommendations to efficiently and accurately recommend useful candidate papers for citation to assist the writing of academic papers. It is defined "on-the-fly" should come with three attributes:

- The system should detect the citing intent online from manuscripts underwriting;
- The system should match a citing intent from the manuscript with the content semantics of candidate papers, instead of matching the input keywords with words in the titles;
- The system should recommend not only the papers from the database but also the out-of-dataset papers, i.e. the newly published papers.

Technically, we propose citation modelling, which leverages the advantages of the embedding techniques to model the knowledge from academic corpus and citation networks, and adapts them for the recommendation tasks. Citation modelling involves three modules: 1. **source representation** for extracting the citing intents of users, 2. **target representation** for inferring the content semantics of candidate papers from the databases; and 3. **citation relationship mining** to enhance the recommendation by leveraging the patterns from the citation network. The three modules are illustrated in detail as follows:

1. **Source Representation**: source representation focuses on representing the citing intents of users from the input manuscript into a semantic space. The algorithm is designed to detect the core citing intent from query contexts and adaptively detect the topic semantics from the continuous updates of the drafts. This module focuses on two tasks: 1. extracting the core citing intent by capturing deeper semantics from the query context, i.e. the

word-wise relatedness, importance and sectional purposes; and 2. extracting the topic semantics from continuous updates of the incomplete manuscript via manuscript dynamic sampling. Experiments have been implemented to verify the effectiveness of our proposed approaches' effectiveness against the previous methods dependent on leveraging local contexts for extracting citing intents.

2. **Target Representation**: target representation is designed to represent the content semantics of the candidate papers. We construct a "content knowledge modelling" by adapting document-level transformer neural networks, complied with dynamic content sampling strategies focused on essential sentences from papers regarding the topic. The constructed content modelling can be adapted for representing and recommending both in-dataset and out-of-dataset papers (newly published papers).

3. **Citation Relationship Mining**: we further leverage the information mined from citation networks, such as co-citation relations and historical co-citation frequencies complied with reinvented objective functions for retrieving multiple positive candidates to improve recommendation performances.

The proposed methods are verified through experiments simulating real-world applications. For example, three completing stages with different amounts of finished contents for input manuscripts are adapted to test the on-the-fly recommendations. In addition, extensive user tests and explainability studies are implemented to verify the usability and rationality of the approaches. The proposed models are also analyzed in the ablation tests to testify component of the model. Overall, the experiments could verify the framework's effectiveness and rationality from the perspective of accuracy, rationality, and usability.

By conducting this dissertation, we provided the following contributions:

- A novel recommendation concept, i.e. "on-the-fly" recommendation, is proposed, which comes with better usability and potentiality to remit the issues from the current keyword-based search engines, and is applicable for different types of documents. The proposed approach can be utilized for recommending both in-dataset and out-of-dataset papers. Experiments have

verified the significant improvements ranging from 10% to 20% on recall for in-dataset and out-of-dataset papers.

- We propose to capture the deeper information from a query context, such as the sectional purpose of the query context, word-wise relatedness, and word-wise importance via the designed attention mechanisms, for inferring the users' citing intents effectively. Experiments have testified the significance of the performance for improvements ranging from 2% to 8% on recall for different scenarios.

- The content semantics of the candidate papers are captured more efficiently by utilizing the sentences containing essential points regarding their topic semantics. It could provide maximally 10% improvements on recall.

- We mine the information from the citation relation's historical patterns to recommend the frequent co-citations and structural contexts from the historical patterns of citation relations. Recommendations could be further improved for 4% to 6% on recall.

This research project has the potential to be applied for various downstream applications, such as a writing assistant for academic papers, a citation checker for reviewers, a citation quality evaluator, a cross-lingual recommender and other types of applications considering referencing resources. Besides, the proposed approach can also be adapted for different types of documents, such as patents, news, judicial papers, etc. We will focus on deploying the application and extending the application to other types of documents for future work.

Keywords: Citation Recommendation, Document Recommendation, Recommender System, Document Embedding, Information Retrieval

Contents

List of Symbols

W ..... The vocabulary

D ..... The paper ID collection

w ..... A word from the vocabulary

d_i ..... The ID of paper i from D

H ..... An academic paper, consisting of its ID, content words, and structural contexts, i.e. $H := \{d_H\} \cup \hat{W} \cup \hat{D}$

$\hat{D}$ ..... The set of IDs of the cited papers in a source paper H

D_n ..... The set of IDs of the structural contexts in a source paper H

$\mathcal{C}$ ..... Tuple for representing a citation relation $\mathcal{C} := \langle d_H, d_t, D_n, C \rangle$

$\mathbf{W}$ ..... Embedding matrix for the vocabulary

$\mathbf{D}$ ..... Embedding matrix for the paper ID collection

$\mathbf{w}_{w_j}$ ..... The j-th column of $\mathbf{W}$, the embedding vector for word w_j

$\mathbf{d}_{d_i}$ ..... The i-th column of $\mathbf{D}$, the embedding vector for paper d_i

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

This chapter presents the background and overview of this dissertation. First, the social background of this dissertation is provided, including the issues from the overload of papers, the current drawbacks of the search engines, and the motivations to propose "on-the-fly" citation recommendations to alleviate the current issues. Second, it illustrates the study's technical background, including the technological trend of the digital academic libraries and the AI technologies developed for the academic community. Third, the brief descriptions of the three main research modules: **source representation**, **target representation**, and **citation relation mining**, as well as the associating contributions are presented. Last, it presents the overall structure of this dissertation.

1.1 Research Background

1.1.1 Overload of Papers

As a researcher, when writing an academic paper, one of the most frequent questions considered could be: "Which paper should I cite at this place?" However, searching and finding suitable articles for citation from a massive number of publications is often time and energy consuming for researchers.

How many papers exist in the world? Answering this question could help

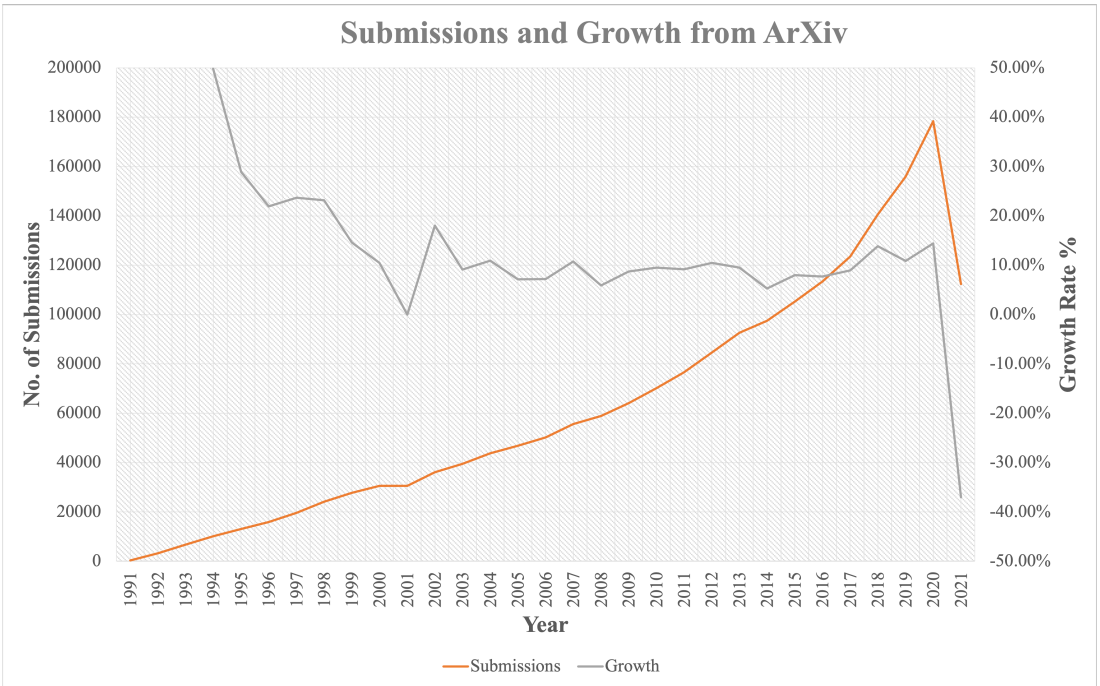


Figure 1.1: No. of submissions and growth rate from ArXiv

us estimate how many challenges the researchers face when searching for references. Herein, it is defined "papers" to include journal and conference papers, dissertation and master theses, books, technical reports, and working papers [1]. Due to the divergence of recorded papers in different databases, the numbers are accounted for differently. According to the STM report [2], the metadata database, CrossRef [1], had recorded over 97 million DOIs from about 60,000 recorded journals in 2018. The "core database" for academic papers, Web of Science [2], has included more than 150 million articles since the year 2018. Google scholar [3], which might be the most dominant the researchers are relying on currently, had been estimated to record nearly 100 million papers in English since 2014 [1]; the other online source, Microsoft Academic Search (MAS), had been estimated to have over 260 million papers since the time of writing this dissertation according to the homepage[4]. One of the most rapid growth fields, i.e. computer science, had been recorded over 160,000 submissions in recent years from ArXiv, as shown in Figure 1.1.

[1]https://search.crossref.org
[2]www.webofknowledge.com
[3]https://scholar.google.com
[4]https://academic.microsoft.com/home

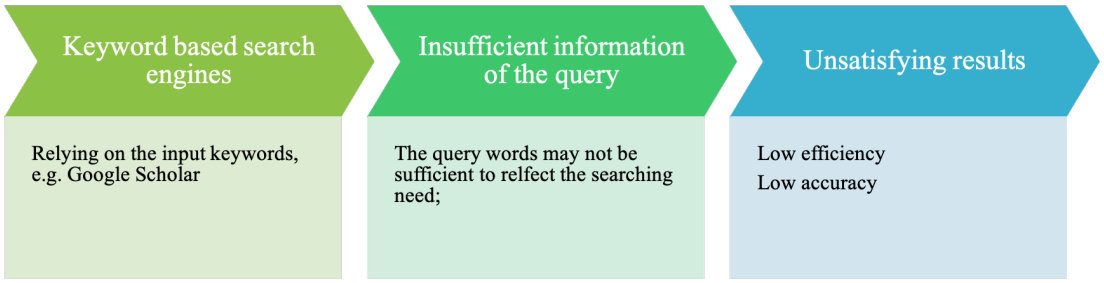


Figure 1.2: Illustration of drawback from current academic search engines

Considering that authors may only look for the papers in their field, it is reported that over 2 million papers were published in computer science, and nearly 25 thousand papers in the sub-field, natural language processing (NLP), since 2021 from the homepage of MAS. It could be drawn that the number of publications is massive in both sub-domains and the overall academic community.

Not only the numbers are massive, but also they are exponentially growing. It could be computed from the MAS statistics, that the number of papers is growing at 5.90% for all fields published since 1990; a growing rate of 8.81% for papers in the field of computer science since 1990; and the number of papers in the sub-filed of NLP is increasing at 7.56% since 1990. ArXiv's growth rates have been kept at about 10% in recent years according to Figure 1.1. It could be drawn that researchers generally face mountains of papers for studying. Especially for the researchers across different fields, they would easily face the challenge to explore tens of thousands of papers.

Therefore, researchers need efficient tools to help exploit the mountains of papers to find the most relevant ones.

1.1.2 Drawbacks of the Current Academic Searching Engines

As researchers, we currently lack efficient tools to deal with the aforementioned problem of "overload of papers". Many scholars are relying on "keyword searches"

on search engines, such as Google Scholar [5] and DBLP [6]. On such systems, the users input query words and then select and read the resulting candidate papers to determine whether to cite them or not. However, the query words are generally too abbreviated to convey adequate information to reflect the searching need of the users, which cloud lead to two drawbacks of such systems, namely: **low-accuracy** and **low-efficiency** as illustrated in Figure 1.3(a).

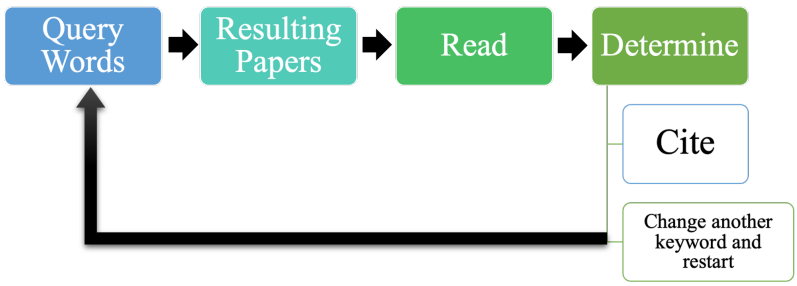
- **Low-accuracy**: due to the over-simplicity of the input keywords and the diversity of the words that might appear in candidate papers' titles, keyword-based systems often lead to inaccurate searching results. For example, suppose a researcher is looking for the paper that proposed the "Word2Vec" algorithm whose title is: "Efficient Estimation of Word Representations in Vector Space"[3]. If the user does not know the exact title, they are trying to search for it by using some query words for the search engine; the word "Word2Vec" might most likely be the adapted keyword, as it indicates the algorithm proposed by the paper, which is also its main point for publication. Nevertheless, suppose we input the keyword "Word2Vec" into Google Scholar. In that case, the correct paper does not appear in the searching results, as the research results are generally the papers with the title including the word "Word2Vec". However, the correct paper's title does not actually have the word "Word2Vec" (as shown in Figure 1.3b), which leads to inaccurate results.

- **Low-efficiency**: the keyword searching systems are inefficient to use, since the users usually have to change different combinations of keywords to find the expected papers. The users also have to read the possible papers for determination, during which much effort is wasted for reading the redundant papers.

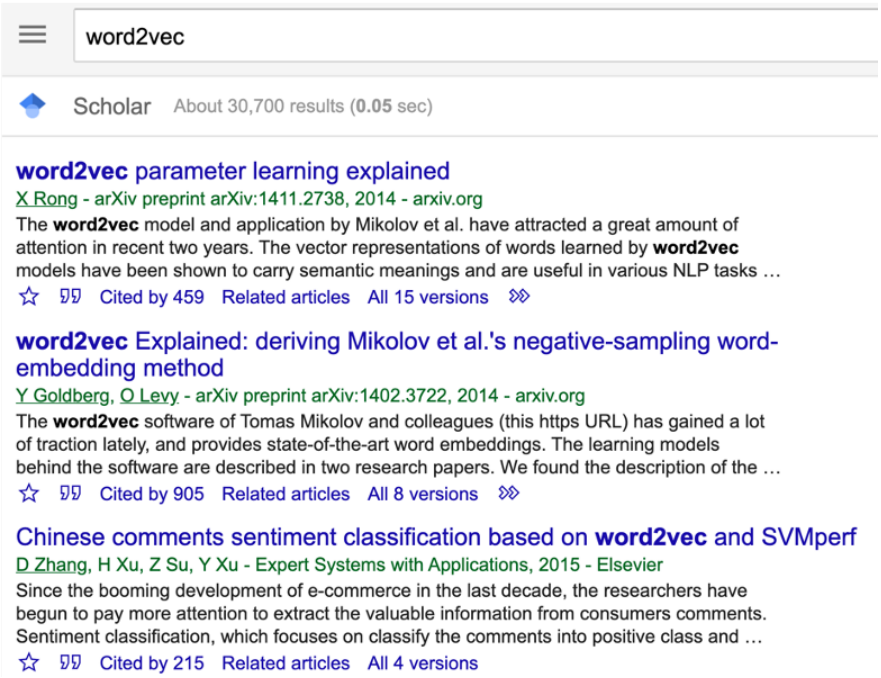
Due to the lack of effective tools to exploit the overload number of papers, researchers rely heavily on cited references in known items, recommendations received from colleagues, or contents of a small number of familiar journals [4].

A side effect of so-called "scientific elites" also arises due to the in-efficiency and in-accuracy of the current searching tools. Some papers are cited unequally

[5] https://scholar.google.com/
[6] https://dblp.uni-trier.de/



(a) Illustration of low-efficiency from current academic search engines



(b) Example of inaccuracy from the current search engines when searching for the original paper of "Word2Vec"

Figure 1.3: Illustration of the inaccuracy and low-efficiency issue from the current academic search engines

higher than others. According to the researches [5, 6, 7, 8], citations tend to approximate the "20/80" phenomenon, which means 20% of papers contributed 80% of citations. The highly cited papers are testified to follow some specific characteristics: (1) published by highly productive scholars; (2) published in top venues; (3) collaborated articles; (4) authors from North American and Western European countries. These researchers are defined as the "scientific elites". Due to the lack of efficient tools for finding citations, researchers tend to cite these

papers published by the scientific elites more frequently, since they came with high appearances. However, young researchers' papers with similar quality and topic relevance might take more effort to be found.

Regarding the drawbacks of the current search systems, which may lead to unsatisfactory experiences to assist writing papers, we aim to propose a novel recommender concept that could deliver better accuracy and efficiency for finding appropriate citations when researchers are wiring their papers.

1.2 Background of Technology Trend in Academia

The recent development of computer technologies in AI, data mining, and databases could drive tremendous changes to academia to improve research efficiency. As stated in [2]:

> "Technology is driving profound changes in the ways research is conducted and communicated, both of which are likely to have impacts on journal publishing ".

This subsection discusses the development of open-source archives and AI techniques associated with scholarly papers.

1.2.1 Open source archives and research data

Open access refers to the making available of published scholarly content (such as journal articles, monographs and conferences proceedings) in online digital copies, free of charge at the point of use, free of most copyright and licensing restrictions, and free of technical or other barriers to access [2]. The open-access may include publisher's platform (such as IEEE Access [7], and open access repositories (such as arXiv [8], and bioRxiv [9]).

Open access comes with a few advantages to facilitate the research process. First, it is faster than the traditional journal editing procedure, which can help researchers exchange the most recent ideas. Second, it is free of charge, which is

[7] https://ieeeaccess.ieee.org
[8] https://arxiv.org
[9] https://www.biorxiv.org

budget-friendly to researchers with limited funds. Third, it allows the development of AI-related algorithms in academia.

Open access might have become a trend in the academic community. It is estimated in 2018 that the proportion of open access articles accounted for 15-20% of papers in all fields, whereas 26-29% of journals are open accessed. The numbers also continue to increase [2]. According to arXiv statistics from the homepage, the submissions to arXiv is only 2,363 at the year of 1992, which had been enlarged about 46 times to 151,084 in 8 years at 2000. Since 2020, there have been 1,930,045 submissions in total to arXiv. From 1992 to 2020, the submissions grow at amazingly 28.84% per year. From 2010 to 2010, the growth rate is 10.56%, significantly larger than the growth rate recorded from MAS for all topics, 3.93%.

Data has become increasingly important to researchers, especially those studying data-intensive topics, such as NLP, CV, and information retrieval in computer science. Sharing of data could facilitate the validation and development of research. According to the FAIR data principles [10], scientific data management should be subject to: **Findable** by leveraging metadata and persistent identifiers, **Accessible** through free and open communications protocols, **Interoperable** by using controlled vocabularies, implementing machine-readability and including references where appropriate and, **resusable** by highlighting clear licence statements that enable the greatest possible reusability.

Currently, conferences and journals in computer science have notified the importance of data in research. Some of them encouraged authors to submit source code and datasets, such as the supplementary requirement for submitting source code and data from ACL 2021 [11].

Data repositories can be of two forms, i.e. **online repositories** and **repositories supplied by publishers**. Online repositories are freely open to the public, which includes GitHub [12], Zenodo [13], re2data [14], FAIRsharing [15], etc. Repositories could also be built and supplied by the publishers, such as IEEE Code Ocean [16],

[10]https://www.force11.org/group/fairgroup/fairprinciples

[11]https://2021.aclweb.org/calls/papers/#optional-supplementary-materials-appendices-software-and-data

[12]https://github.com

[13]https://zenodo.org

[14]https://www.re3data.org

[15]https://fairsharing.org

[16]https://innovate.ieee.org/ieee-code-ocean/

Elsevier's Mendeley Data [17], Nature's Scientific Data [18], etc.

Along with the gradual openness of data, research and applications relying on vast amounts of data, such as our on-the-fly citation recommender, could be implemented in the near future.

1.2.2 AI in academia

Due to the rapid developments of machine learning and AI technology, it appears that certain techniques could be applied to the scholarly community to deliver better cost efficiency to publishers and word efficiency to authors.

The essence of AI algorithms is to extract specific patterns from massive data, including identifying properties or locations, recognizing the trend, etc. The learned knowledge is then applied to deliver output from new data. The main advantage of AI algorithms is to substitute humans for repetitive works on a big volume of searching.

Semantic mining in AI allows the algorithms to learn textual knowledge like a human, such as grammar rules, name entities, content knowledge, etc., which could be potentially adapted for real-world applications. From the publishers' perspective, semantic knowledge could aid the publishing process, such as helping viewers check grammar, completeness of citations, the manuscript's topic to detect whether it suits the conference or journal, etc. From the authors' perspective, AI applications could help the authors search for relevant papers, assist in writing, check grammar, etc.

Our on-the-fly citation recommender is built on semantic techniques. It could potentially be applied for both the publisher's and author's perspectives to help check the completeness of citations and assist the writing of papers by effectively finding citations for the authors.

1.2.3 Network-based and Content-based Approaches in Citation Recommendations

The studies in citation recommendation can be generally categorized by two quadrants from the perspective of technique and concept, i.e. ether based on

[17] https://www.elsevier.com/authors/tools-and-resources/research-data/mendeley-data-for-journals

[18] https://www.nature.com/sdata/policies/repositories

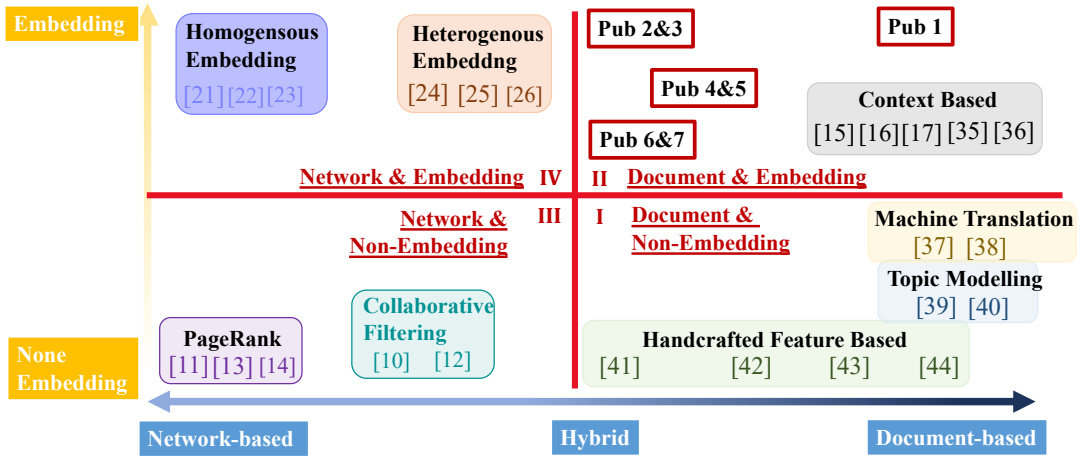


Figure 1.4: Relative position of the publications with the past studies (the shown publications 1-7 are listed in the Chapter 1.3.3)

embedding or non-embedding techniques, based on the concept of network or document, as illustrated in Figure 1.4.

Network-based approaches generally represent each paper as a node, and citations are links, and therefore to construct a citation network for analyses. For example, the PageRank-based techniques [9, 10, 11] recommend un-linked papers by their popularity (ii.e. the number of in-ward links) through non-embedding network techniques, which are positioned at the bottom left of the figure; the studies learn node embeddings by predicting the community that a paper belongs to via embedding based techniques [12, 13, 14], which are placed top left of the figure.

On the other hand, content-based approaches consider extracting content semantics from the texts and therefore recommend suitable candidates by content suitability. For the non-embedding-based techniques, a line of studies [15, 16] considered to construct topic modellings from the term frequencies and recommended topically relevant papers. Another line of text-based but non-embedding-based methods considered to adapt the machine translation techniques to "translate" the query context to the candidate citations [17, 18]. Embedding-based techniques represent words and content through vector spaces reflecting their semantic distances. Text-based approaches are placed to the right of Figure 1.4, within which the embedding-based studies are placed at the top, and the non-embedding-based ones are positioned at the bottom.

Our publications, as presented in the Chapter "Selected List of Publications" are placed in Figure 1.4 to demonstrate their relative positions from the two perspectives. All the publications are primarily based on the document-based concept; hence they are generally placed at the top right quadrant. However, publications 1,2,6 and 7 are considered to also leverage the information from citation networks (structural context and co-citations), so they are placed relative to the middle. Publication 5 is considered to explore deeper content semantics for the recommendation; hence it is placed at the top right.

In summary, this dissertation's research is primarily developed based on NLP algorithms. However, we also considered utilizing useful features from the citation networks for support. In later research, it will be considered to deeply combine the network-based and text-based techniques to improve accuracy and usability.

1.3 "On-the-fly" Citation Recommendation

To improve the usability of the current recommender for better efficiency and accuracy and to alleviate the issue of "scientific elites", a novel approach is proposed to recommend candidate papers on-the-fly while an author is working on their manuscripts. "On-the-fly" recommendation aims to support academic authors, especially younger researchers, to find appropriate candidate papers for their manuscripts during writing according to the citing intents detected from the drafts, reducing the time spent relative to manual searching via keywords; it can also potentially help authors and reviewers to check the completeness of a paper's citations before publication, especially some newly published papers might might not be aware to them. "On-the-fly" approaches include two features: source and candidate "on-the-fly".

- **Source On-the-fly**: instead of letting the users "trial and error" different keywords, it is considered a more efficient and intelligent system should be able to detect the citing intent of the user directly from the incomplete updates of an input manuscript during writing or reviewing. The system should consider both "micro-level" semantics directly from the target sentences needed citations and the topic semantics from the finished content to infer the "macro-level" citing intent. Automatic detection of citing intent from the manuscript's content could ultimately save time for the users to

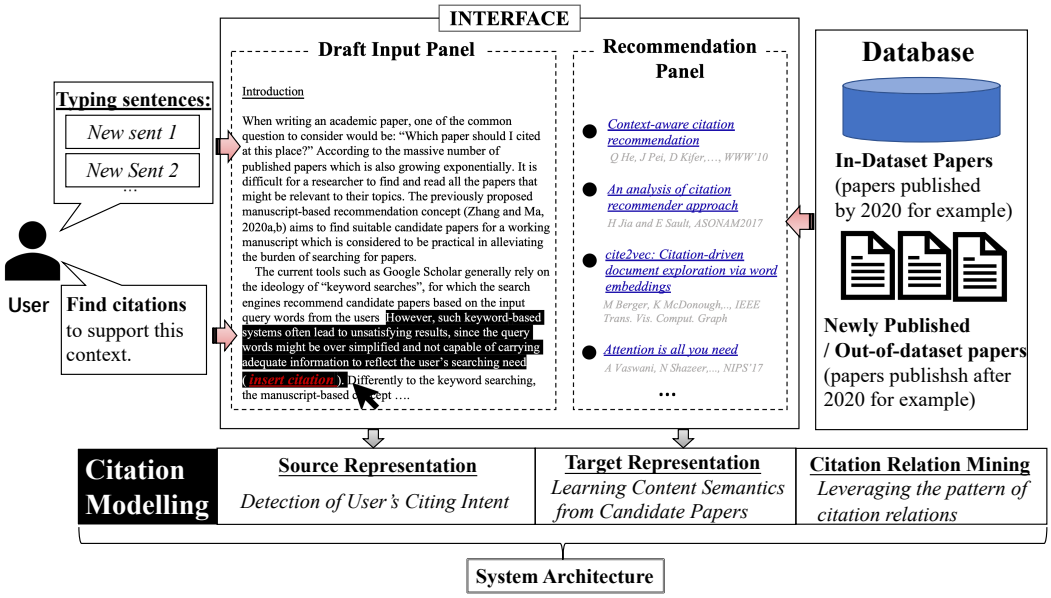


Figure 1.5: A hypothesized system to recommend candidate citations "on-the-fly" for an input manuscript

trail and error keywords and improve the accuracy for matching candidate papers.

- **Candidate On-the-fly**: considering the mismatching issue between the input keywords and the words that appear in the title of candidate papers and the upcoming newly published papers, our proposed approach matches the extracted content knowledge extracted from the in-dataset and out-of-dataset (new papers) candidates to the detected citing intent. Content-dependent matching could effectively find relevant papers based on the content semantics rather than the term appearances, delivering better accuracy in searching.

Figure 1.5 illustrates the idea, where an author is writing the draft or a reviewer is reviewing a paper in the "*Draft Input Panel*"; if the author wants to find citations to support a particular context or check the completeness of citations for a context, they can simply highlight the corresponding sentences using a mouse pointer, and then the system presents the best-matched papers in the "*Recommendation Panel*". The candidates are found by matching the content knowledge of the in-dataset and out-of-dataset papers.

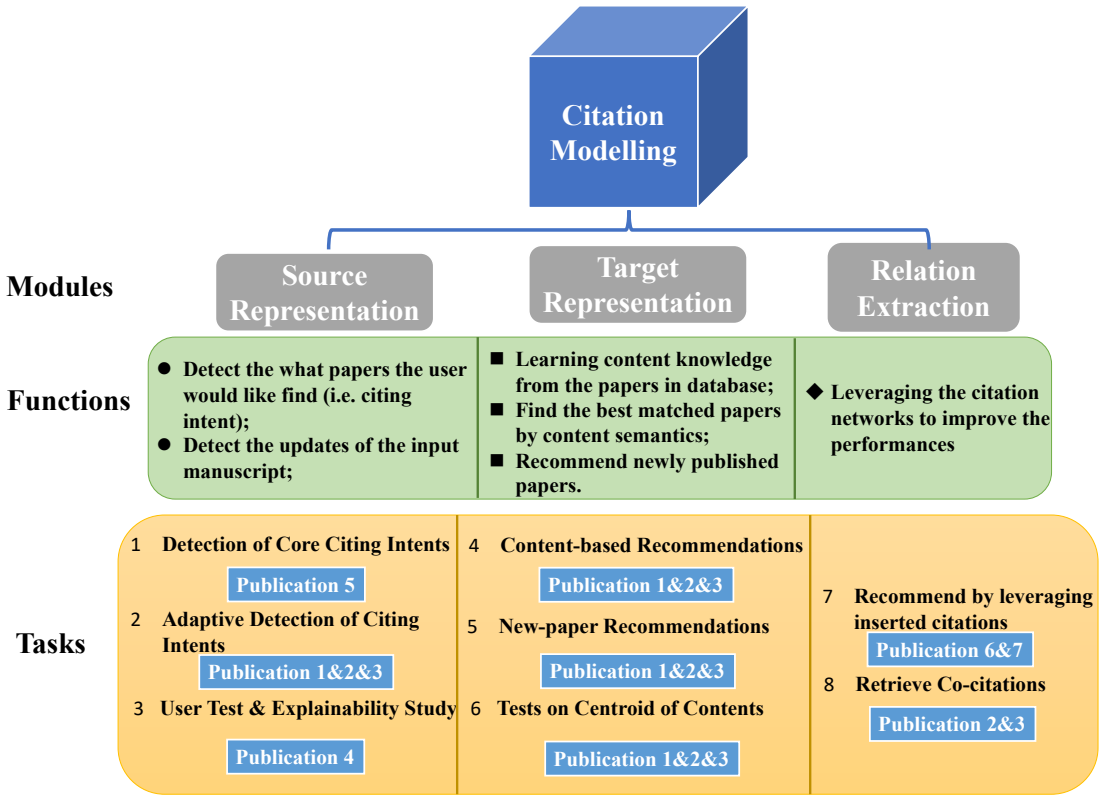


Figure 1.6: Overall Architecture of the Project and Thesis

1.3.1 Major Tasks from "On-the-fly" Citation Recommender

Three modules are designed to implement the hypothesized system for "on-the-fly" citation recommendations, namely: **source representation**, **target representation**, and **citation relationship mining**. The overall view of the designed tasks and associated publications are presented in Figure 1.6.

Source representation is concerned with effectively extracting the citing intent from the input manuscript. It is considered that the citing intent should be captured from two levels of views: **micro** and **macro**-level views. The **micro** view indicates the semantics carried by the local context (the surrounding context around a placeholder for inserting citations), noted as the core citing intent in Figure 1.6; whereas the **macro** view is reflected by the finished content of the input manuscript to infer the topic in a board context, as indicated as adaptive detection of citing intents in Figure 1.6. Micro viewed citing intents are captured by mining semantics from the query context. Specifically, the information word-to-word

relatedness, and word-level importance are learned to infer the semantics. As for the macro-viewed citing intent, a dynamic context sampling strategy is developed to detect the change of the input manuscript, and sample sentences from the finished content to infer the main topic of the manuscript. The combination of the micro and macro viewed semantics has led to superior performance from our research than prior works. Publications 1, 2, 3, 4, and 5, as listed in Section 1.3.3 involved the studies regarding the tasks in this module.

Target representation refers to representing the content semantics of the candidate papers. To effectively capture the content semantics, by leveraging the sentences correlated to the "centrality" of a paper, where the "centrality" denotes the main topic of the article. This module involves three tasks: content-based recommendation, recommendation of new papers (i.e. out-of-dataset papers), and the tests of the centroid of papers. Publications 1, 2, and 3 in Section 1.3.3 mainly address these topics.

Citation relationship mining aims to adapt the knowledge from the citation networks to provide supplementary performances, such as the co-citation information and co-appeared citations in a draft, as illustrated in Figure 1.6. Co-appeared citations (publication 6 and 7 in Section 1.3.3) indicate the citations that appeared in the finished content of the manuscript, which is adapted to further understand the citing intent of the users, since they might carry information on the topics of the papers that the user might be interested. Moreover, co-citations are adopted to improve the recommendation performances. The underlying logic is: if two papers are frequently cited together if one appears as a citation, then there is a higher chance for the other to also appear as a citation. The relevant studies are conducted in publications 2 and 3 listed in Section 1.3.3.

1.3.2 Major Contributions

"On-the-fly" citation recommender aims to supply academic authors with better working efficiency, by innovating the techniques in semantic mining at the cutting edge. It could potentially contribute values in two perspectives: 1) from **application view**, it could provide better usability; 2) from **technical view**, it is enhanced the conventional semantic mining algorithms to be learnable for scientific knowledge, and adaptable for "on-the-fly" scenario.

From the **application view**, it could potentially deliver the following three

contributions to the academic community:

- **Writing support**: Our "on-the-fly" recommender is primarily designed to provide a more efficient solution to find citations. By adopting it, users (authors) could save their time and energy spent on trial and error keyword-based searching to improve writing efficiency. In addition, the citation quality could also be more completed than conventional search engines;
- **Review support**: The "on-the-fly" recommender is adaptable for assist writing papers for authors, but it can also help the reviewers check the completeness of the citations in a manuscript. Right now, the reviewer might primarily rely on their expert knowledge in a particular area to judge the quality of the citations. However, there might exist papers relevant to the reviewed study, but they are not aware by the reviewers, such as the newly published papers. "On-the-fly" recommender could be utilized in such scenarios to help check whether there exit additional references that the authors should cite;

In a brief summary, "on-the-fly" recommender is proposed to simulate the workflow of academic publication from authors' and reviewers' perspectives. It also provides a foundation for other types of academic applications.

From the **technical view**, the following contributions are potentially made:

- **Scientific knowledge modelling**: Implementing "on-the-fly" recommender also provides universal modelling trained from scientific knowledge that could be adapted for other types of tasks, such as scientific name entity recognition, classification of the topic of papers, etc.
- **Recommendation modelling for scientific papers**: It also constructs a recommendation modelling for scientific papers' recommendation by leveraging the knowledge and citation relationships from papers. The recommendation modelling can also be adapted to recommend other types of "connected documents" such as patents, news, technical reports, etc.
- **"In-depth" information mining from scientific papers**: It mines and leverages the "in-depth" information to stimulate recommendation accuracy ultimately. For example, it is considered to utilize the purpose of sections headers where a query context comes from, to recommend citations which

suits the sectional purpose; the word-wise relatedness and importance are adapted to infer the citing intent; as well as utilizing the citation relationships to retrieve positive and negative samples to make efficient training strategies to suit different tasks, such as multiple positive recommendations.

- **Dynamic sampling for "on-the-fly" scenario**: The dynamic sampling strategies are specifically designed for "on-the-fly" recommendation scenario. It involves a manuscript sampling strategy that can detect the citing intent in a "macro-scope" according to the change of the manuscript and a citation sampling strategy to sample essential context from papers and extract content-dependent semantics. Dynamic sampling strategies allow the algorithm to recommend both in-dataset and out-of-dataset papers for incomplete manuscripts in real-time.

1.3.3 Publications

By conducting the project, the following works had been published or submitted:

- Publication 1: "'On-the-fly' Citation Recommendation based on Content-dependent Embeddings", SIGIR 2022 (Under reivew);
- Publication 2: "MP-BERT4CR: Recommending Multiple Positive Citations for Academic Manuscripts via Content-Dependent BERT and Multi-Positive Triplet", IEICE (Under Review);
- Publication 3: "Recommending Multiple Positive Citations for Manuscript via Content-Dependent Modeling and Multi-Positive Triplet", WI 2021;
- Publication 4: "Dual Attention Model for Citation Recommendation with Analyses on Explainability and Qualitative Experiments", Computational Linguistics (Accepted for publication: 04 Jan 2022);
- Publication 5: "Dual Attention Model for Citation Recommendation", COLING 2020;
- Publication 6: "Doccit2vec: Citation recommendation via embedding of content and structural contexts", IEEE Access;
- Publication 7: "Citation Recommendations Considering Content and Structural Context Embedding", BigComp 2020.

1.4 Dissertation Structure

This dissertation is outlined as the following: Chapter 2 presents the review of related works, as well as the relative positive of the researches in this dissertation; Chapter 3 illustrates the research tasks regarding **source representation**; Chapter 4 aims to provide the detail of the research tasks involved in **target representation**; Chapter 5 explains the studies about **mining citation relationships** and their applications. Chapter 7 summarizes the dissertation with key points.

CHAPTER 2

Related Work

This chapter aims to review the related literature from a macro view. First, the discipline is introduced to which our study belongs, i.e. data mining. Then, it provides a review of recommender systems from the application's perspective. Lastly, it provides the technical reviews for other citation recommendation approaches. Citation recommendation refers to finding relevant documents based on an input query.

It aims to provide a literature review from a macro view. The detailed technical illustrations on the previous studies are presented in the chapter for each research task.

2.1 Computation Social Science

Computation Social Science (CSS) is defined as the interdisciplinary investigation of the social universe through the medium of computation approaches, ranging from information extraction algorithms to computer simulation models [19]. In other words, this discipline studies how to tackle the issues individuals or organisations face by adapting computer science techniques. CSS involves the following main areas: information extraction, social networks, social complexity, and social simulation modelling [19]. Our study is closely relevant to information extraction and social networks, which are presented in detail in this section.

2.1.1 Information Extraction

This area refers to computation ideas and methodologies pertaining to the creating of scientifically useful information based on raw data sources [19]. Our studies focus on leveraging content analysis techniques to match the appropriate candidates for a given source paper based on their content knowledge.

Some previous works model citation recommendation as a link prediction problem in networks, where the papers are denoted as nodes, and the citation relations are links between nodes. For example, the studies from [20, 9, 21, 10, 11] proposed to arrange the citation relations into directed networks, and the recommendations are made based on a collection of seed papers. This line of approaches might help find relevant papers during early studying of a field for a researcher; however, it is not directly applicable to the on-the-fly recommendation scenario.

Considering the on-the-fly scenario, this project is considered to extract the citing intents of the users directly from the content of the input papers. The previously context-based approaches [22, 23, 24] leveraged the embedded semantics of the query context as the citing intent of the users for ranking the candidate papers. Following this line of study, it is proposed to additionally adapt the bibliographical couplings and word-wise relatedness and importance by using DNN-based networks with link information [25, 26], and attention mechanisms [27].

2.1.2 Bibliographical Network

Academic papers can also be treated as directed networks, where the nodes are papers, links are citations, and the direction of the links denote the relation of citing.

As discussed above, the previous random walk based approaches [20, 9, 21, 10, 11] rely on the input seed paper to find recommendations, by using collaborative filtering [20, 21] , pagerank-based techniques [9, 10, 11], embedding-based methods [13, 14, 12], heterogeneous network embedding for considering both netowrk and content words [28, 29, 30] Generally, they extracted the connectivity information from the citation networks, and predict the probable links as the recommendations. However, they are hardly adaptable for the on-the-fly scenario, since they do not consider the content information.

In our studies, we proposed to leverage both the word and network information to make the recommendations more effective. Basically, we combine the linking information from the citation networks to our document embedding models. For example, in studies [25, 26], we combine the document embeddings with the bibliographical coupling information so that when a paper is given as the input, the algorithm can predict by content and also the possible bibliographic couplings from the historical citations. In addition, the study [31] combined the information of historical co-citations to the documents semantic embeddings so that the frequently co-cited papers can be effectively found compared to the models solely considering content semantics.

2.2 Natural Language Processing

Natural Language Processing (NLP) Technology converts daily oral or written language into binary code that the machine can recognize to enable the machine to understand the real meaning of human beings [32, 33]. It was primarily designed to study how to extract the information from the input content on the semantic level, and it also has certain common-sense knowledge, and reasoning ability [32]. NLP plays an essential role in various application domains, such as recommender systems, machine translation, speech understanding, dialogue systems, etc. The overview of NLP involved systems are illustrated as the following:

Applications with NLP techniques usually involve a representation model to represent the text into machine-readable data [32, 34] for extracting the semantic information, and an intermediate neural architecture designed for the downstream tasks, such as recommendation, translation, dialogue, etc. Basically, NLP acts as the foundation for applications that require language understating.

2.2.1 Document Representation

Document representation techniques are essential in the NLP field, which transforms the discrete and sparse text information into the form of data that machines can process by preserving their content semantics.

Early approaches represent the input texts according to the term frequencies. They firstly assign one-hot vectors to the words in the corpus and then adapted term-frequency-based methods, such as TF-IDF[35] and BM25[36], to find the

most essential as the representations of the texts. However, frequency-based approaches do not preserve the content semantics from the input texts well. Also, they consume computer memory proportionally to the vocabulary size of the corpus, which is limited in applying for the tasks relying on the understating of content semantics and a large amount of corpus.

Recent representation methods transform texts into continuous vectors by preserving the semantic proximities of words, from which the memory consumption is not dependent on the size of the corpus. This line of methods adapt neural networks to convert each word to an n-dimensional vector, then the model masks words from the sampled sentences of the text and adapts objective functions to predict those masked words. For example, Word2Vec [3, 37] and Doc2Vec [38] adapted DNN-based neural networks and soft-max objective function to predict the masked words from the sampled context with a pre-set length (the window size). However, they suffer from information loss when applied to citation recommendation tasks. Herein, we proposed DocCit2Vec[25, 26], and DACR[27] based on specifically designed fine-tuning models to allow the word embedding combined with information on citation relations, bibliographical couplings, and word-wise relatedness and importance, to improve the recommendation performance. However, they are still limited in applying for on-the-fly scenarios, especially for new paper recommendations and detection of the updates of the incomplete input drafts.

Scholars have leveraged the advantages of the text embedding techniques to extract the content semantics for the task of citation recommendations, such as [22, 23, 24, 39, 40]. Another line of text-based methods considered to adapt the machine translation techniques to "translate" the query context to the candidate citations [17, 18]. Topic models have also been adopted to recommend citations based on the topic similarities [15, 16]. Some studies considered to construct user profiles by handcrafted features, such as searching history and past publications to recommend similar papers [41, 42, 43, 44].

The latest research have adapted the transformer neural network[45] complied with the objective of masked word predictions, such as BERT[46], RoBERTa[47], Sentece-BERT[48]. These models come with more effective abilities in representing the content semantics than the previous DNN-based networks. We leverage the advantage of transformer neural networks to provide universal modelling for on-the-fly recommendations, which could first detect the updates of the input draft and matches candidate papers according to their content semantics.

The network and text-based approaches are summarized in Figure 1.4.

2.3 Recommender Systems

Recommender Systems (RSs) indicate the applications and techniques providing suggestions of items to their users [49, 50, 51, 52]. The suggestions aims to assist the users in various decision-making processes, such as what items to buy, what music to listen to, or what news to read [49]. RSs are primarily designed to support individuals who lack sufficient experiences, knowledge, or time to explore and evaluate the potential candidates in the overwhelming number of alternative items [49].

RSs generally involve three main components: 1. user's need detection, 2. candidate representation, and 3. candidate ranking. RSs may ask the users to input keywords, as adapted in the mainstream search engine, or adopt the historical browsing views, as utilized in Online shopping websites, to detect the searching need of the users. RSs may represent the candidate items from their databases from different perspectives to match the users' detected searching needs. The best-matched items are recommended results. The techniques could be generally categorized as the following: collaborative filtering, content-based, community-based, and hybrid systems.

The collaborative filtering technique is widely implemented in e-commerce systems [53]. This technique detects a user's need by analyzing the historical ratings of viewed items. It makes recommendations by comparing with other users with similar browsing history [54, 55]. Conventional approaches in this sub-field firstly find neighbourhood users to the target user based on their historically viewed items [56, 57] by a technique such as Pearson coefficient [58]. Then, the predicted rating of a candidate item for the target user is determined by its neighbourhood's ratings on the item [59, 60]. Some recent studies train predictive models, such as Singular Value Decomposition (SVD) [61, 62, 63], Support Vector Machines (SVM) [64], Latent Dirichlet Allocation (LDA) [65], Latent Semantic Analysis (LSA) [66], and Bayesian Clustering [67], based on the users' ratings, and then predict the rating for a candidate item to the target user. However, this line of studies have to adopt all the rating data in the database, which might consume unexpectedly large computation time for large-scale datasets; in addition, the approaches do not consider the content knowledge, which could not

directly applied to our application scenario for recommending citations based on the content semantics.

Community-based RSs are based on the preferences of the users. They generally follow the concept that "tell me who your friends are, and I will tell you who you are [68]". In a macro-view, these approaches convert the users to a connected network, where the users are the nodes, and interactions (or trust) between the users are the links and associated weights between the nodes. The nodes are usually the papers in the citation recommendation scenario, and citations relations are the links and weights. The queries are usually a set of seed papers, and the recommendation task is to find the papers that belong to the same community as the seed papers. These systems adapts techniques in network analysis to extract the attributes of the nodes, for example PageRank-based techniques [9, 10, 11], or the network embedding techniques [12, 13, 14]. The recommendations are made by similarity metrics.

Content-based approaches rely on the intrinsic properties of the candidate items relative to that of a query item. For example, if a user has positively rated a movie, the system can learn to recommend similar movies that come with similar stories. The most essential process in content-based recommender systems is to extract features from the input of the users (text, query words, image, etc.) and the candidates resources in the database [69], by relying on the techniques from information retrieval (for example, the techniques introduced in section 2.2). These techniques may include the term-frequency based methods, e.g. [35, 36], or embedding based methods [3, 37, 46], to represent the users' queries and database resources into a set of attributes. Then, the candidates are computed by similarity metrics in vector space. Content-based approaches are considered more applicable to the "on-the-fly" scenario since the learned content models are independent of the user profiles; recommendations based on content knowledge and new items could be recommended.

CHAPTER 3

Source Representation

The main objective of the source representation module is to detect the users' citing intents from the input manuscripts; in simple words, "what papers would the users need"?

Three tasks are considered to be essential to accomplish the objective.

- The first task aims to define the core of the citing intent from the query context (the surrounding words around the target placeholder to insert a reference);
- The second task aims to adaptively extract the topic semantics from the updates of the input manuscripts, to express the citing intent more comprehensively;
- The third task involves a user study to testify whether the proposed approach could be applied for checking and reviewing the completeness of citations for a paper before publication and an explainability study to analyze the learned weights in the attention mechanisms.

Please refer to section 3.3 for the first task, section 3.4 for the second task, and section 3.6 and section 3.7 for detailed discussions.

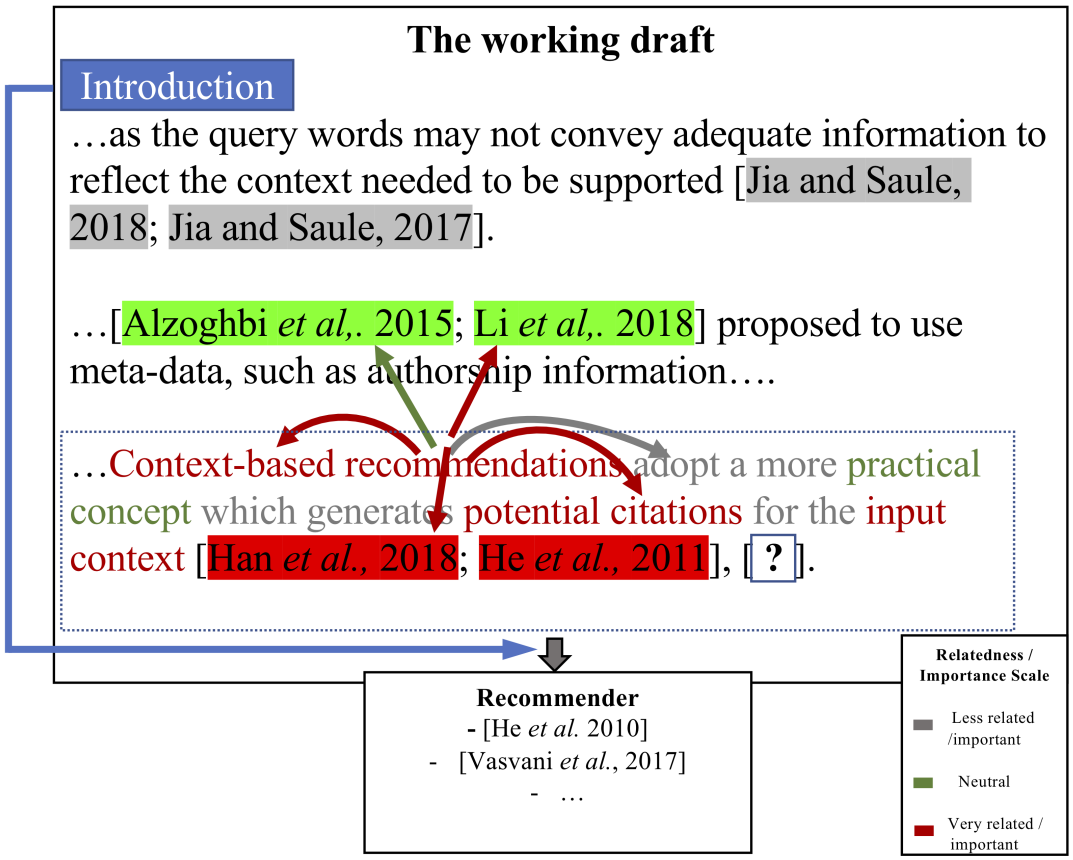


Figure 3.1: Capturing word-wise relatedness, importance, and sectional purpose for inferring citing intents

3.1 Motivation

To detect the citing intents of users, keywords-based searching from Google Scholar has been the predominant method currently. However, keyword-based systems often generate unsatisfying results because query words may not convey adequate information to reflect the context that needs to be supported [11, 70].

Researchers in various fields have proposed various methods to solve this problem. For example, studies in [20, 9, 21, 10, 70] considered recommendations based on a collection of seed papers, and [71, 72] proposed methods using meta-data, such as authorship information, titles, abstracts, keyword lists, and publication years. [16, 73] considered the "local contexts" (surrounding words around a target placeholder for insertion) are essential to understand the citing intents of the users

since they are primarily describing the cited papers. However, solely adaption of local contexts might be limited to fully uncovering the citing intents of the users' since they are generally short in length and do not tell the topic semantics of the manuscript. To comply with the proposed "on-the-fly" scenario, where the users might continuously update the manuscript, it is considered that two kinds of information are essential for the detection: **detection of core citing intents from local contexts** and **adaptive detection of citing intents from updates of the manuscripts**.

3.1.1 Detection of Core Citing Intents

Understanding the local contexts are essential to detect the core of the citing intents of the users. However, the local contexts are usually short in length (about a few sentences). They do not tell much information about the manuscript in a macro-view, for example, the main topic. Hence, additional information might still be essential to uncover the need of the users fully:

1. Scientific papers tend to follow the established "IMRaD" format (introduction, methods, results and discussion, and conclusions) [74], where each section of an article has a specific purpose. For example, the introduction section defines the paper's topic in a broader context, the method section includes information on how the results were produced, and the results and discussion section presents the results. Therefore, citations used in each section should comply with the specific purpose of that section. For example, citations in the introduction section should support the main concepts of the paper, citations in the methods section should provide technical details, and citations in the results and discussion section should aim to compare results to those of other works. Therefore, recommendations of suitable citations for a given context should also consider the purpose of the corresponding section.

2. Certain words and cited articles in a paper are much more closely related than other words and articles in the same paper. Capturing these interactions is essential for understanding a paper. For example, in Figure 3.1, the word "recommendation" is closely related to the words "context-based," "citations," and "context," but has a weak relationship with the words

"adopt," "more," and "input." Additionally, a given word may have strong relatedness with some citations that appear in the paper. For example, the word "recommendation" has a strong relatedness to citations "[72]" and "[24]" because both of these citations focus on recommendation algorithms.

3. Not every word or cited article has the same importance within a given paper. Important words and cited articles are more informative with respect to the topic of a paper. For example, in Figure 3.1, the words "context-based," "recommendations," "citations," and "context" are more informative than the words "adopt," "more," or "generates." The citation, "[24]," may be more essential than "[70]" because the former is related to context-based recommendations, while the latter is related to a different approach.

The core of the proposed approach to capture the aforementioned information is composed of the two attention mechanisms, namely self-attention and additive attention. The former captures the relatedness between contextual words and structural contexts, and the latter learns the importance of contextual words and structural contexts. Additionally, the proposed model embeds sections into an embedding space and utilizes the embedded sections as additional features for recommendation tasks.

3.1.2 Adaptive Detection of Citing Intents

Adaptive detection aims to extract the citing intents from the incomplete papers, including manuscripts still being written, first editions of papers, and papers under review, where the user may continuously update the input manuscript. In addition to the information extracted from the local context, it is considered that the algorithm should also detect the main topic through the incomplete portions of the manuscript, which require updating. The detected citing intent should be extracted adaptively considering the local context, and topic semantics are essential for the on-the-fly scenario.

Thereby, a manuscript sampling strategy is proposed to extract the sentences regarding the core citing intent and the topic semantics. Firstly, the manuscript sampling strategy extracts the base context (the local context) as the "backbone" for citing intent, and the superstructural context (context from the finished content of the draft) as the topic knowledge, and leverages the two extracted information to express citing intents for the on-the-fly scenario.

In addition, qualitative analyses to test whether DACR could recommend additional ground-truth citations. The proposes of these tests are two-folded: 1) test whether DACR could find appropriate recommendations that the conventional keyword-based systems could not find; and 2) test whether DACR could be applied for checking the completeness of citations.

In summary, the following contributions are made:

- First, it verified that the word-wise relatedness, importance, and sectional purpose are effective to detect the core citing intents from the query contexts, and the learned weights from the attention mechanisms can appropriately reflect the three information;
- Second, the topic semantics of the input manuscript is helpful to extract the citing intent comprehensively;
- Third, the proposed approach is testified for checking the completeness of the citations for authors and reviewers.

The remaining sections of this chapter is organized as the follows: Section 3.2 discusses the previous studies in the field; Section 3.3 and Section 3.4 present the proposed approaches; Section 3.5 presents the experimental results and analyses; Section 3.7 and 3.6 illustrate the user tests and explainability study.

3.2 Related Work

This section presents the past studies on context-based methods for extracting citing intents, attention mechanisms, and explainability studies.

Citing intents can be extracted from the input queries. The query could be a collection of seed papers [20, 9, 21, 10, 11], and recommendations are generated via collaborative filtering [20, 21] or PageRank-based methods [9, 10, 11]. Some studies [71, 72] have proposed using meta-data, such as titles, abstracts, keyword lists, and publication years, as query information. However, in real-world applications for supporting the writing of manuscripts, these techniques lack practicability. Context-based methods [16, 73, 24, 25] use a passage requiring support as a query to find the most relevant papers, which can potentially enhance the paper-writing process. However, such methods may suffer from information loss because they do

not consider sections within papers or the importance and relatedness of words. In addition, adaptively detection the topic semantics from the remaining part of the manuscript is also essential to fully uncover the citing intents.

The attention mechanism is commonly applied in the field of computer vision [75] and detects important parts of an image to improve prediction accuracy. This mechanism has also been adopted in recent research in text mining. For example, [76] extended Word2Vec with a simple attention mechanism to improve word classification performance. Google's BERT algorithm [46] uses multi-head attention and provides excellent performance for several natural language processing tasks. The method introduced in [77] uses self-attention and additive attention to improve recommendation accuracy for news sources.

However, attention mechanisms are generally treated as "black-boxes", where the internal functions of the learned weights are not fully uncovered. [78] analyzed the pair-wise weights of self-attention layers in BERT [46], to study the pattern of word-to-word correlations, and linguistic correlations. [79] studied the identifiability of weights and explanatory insight between the weights and input tokens, which demonstrated that self-attention weights were not directly identifiable and explainable. [80] analyzed the most emphasized words from self-attention, which was found that few words are likely to be over-emphasized. In this article, we presume that the pair-wise self-attention weights indicate the "relatedness" between words, and the weights of additive attention correspond to the "importance" of words. The analyses were made in four aspects: 1. correspondence of most emphasized items (high relatedness) with the citing intent of the input context; 2. pattern of weights at different heads of self-attention; 3. correspondence of the highest scored words from additive attention (high importance) and the citing intent of the input context; 4. differences of the most-emphasized items between self-attention (relatedness) and additive attention (importance).

3.3 Detection of Core Citing Intents

As discussed in Section 3.1, three kinds of information are essential to infer the core citing intents from the local contexts, namely word-wise relatedness, importance, and sectional purposes. A dual attention model (DACR) is proposed to capture the three pieces of information, which involves a context encoder for encoding contextual words, sections, and structural contexts into a fixed-length vector and

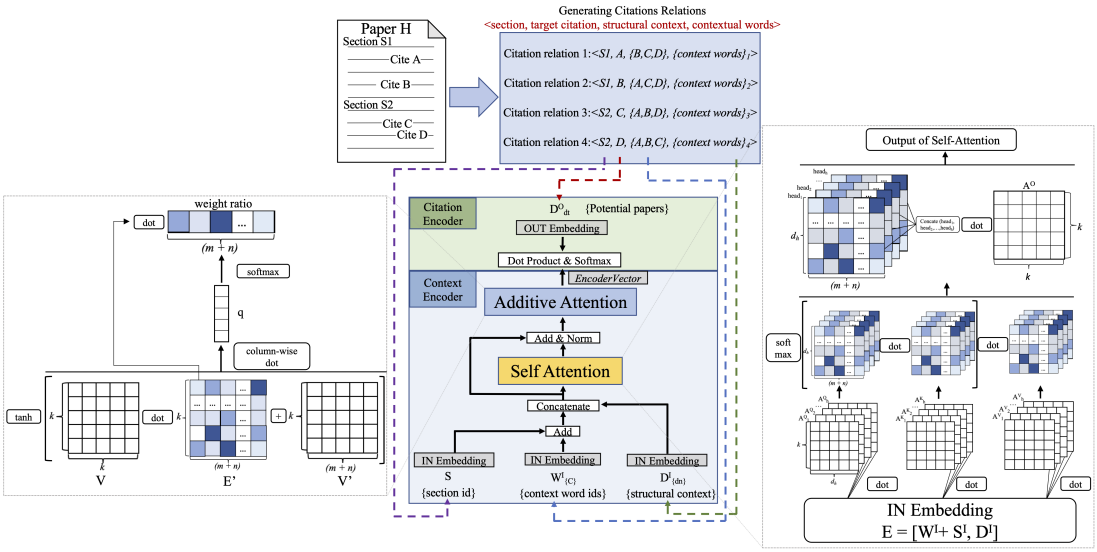


Figure 3.2: Overview of DACR for Capturing Core Citing Intents

a citation encoder for predicting the probability of a target citation. An overview of the model is illustrated in Figure 3.2.

Academic papers can be treated as a type of hyper-document, where citations are identical to hyperlinks. Based on paper modelling with citations, [24], and the modelling of citations with contextual articles [25], we introduce modelling with citations, contextual articles, and sections.

Definition 1 (Academic Paper)**.** *Let $w \in W$ represent a word from a vocabulary W, where $s \in S$ represents a section from a section collection S and $d \in D$ represents a document ID (paper DOI) from an ID collection D. The textual information of a paper H is represented as a sequence of words, sections, and document IDs (i.e., $\hat{W} \cup \hat{S} \cup \hat{D}$, where $\hat{W} \subseteq W$, $\hat{S} \subseteq S$, and $\hat{D} \subseteq D$).*

Definition 2 (Citation Relationships)**.** *The citation relationships $\mathcal{C}$ (see Figure 3.2) in a paper H are expressed by a tuple $\langle s, d_t, D_n, C \rangle$, where $d_t \in \hat{D}$ represents a target citation, $C \subseteq \hat{W}$ is the local context surrounding d_t, and $s \in \hat{S}$ is the title of the section on which the contextual words appear. If other citations exist within the same manuscript, then they are defined as the "contextual articles" and denoted as D_n, where $\{d_n | d_n \in \hat{D}, d_n \neq d_t\}$.*

3.3.1 Context encoder

The context encoder takes from citation relationships, namely, context words, sections, and structural contexts. The encoder contains three layers: an embedding layer for converting words and documents (structural contexts) into vectors, a self-attention layer with an Add&Norm sub-layer [45] for capturing the relatedness between words and structural contexts, and an additive attention layer [77] for recognizing the importance of each word and structural context.

3.3.2 IN Embedding, Add and Concatenation layer

The IN embedding layer initially generates three embedding matrices $\mathbf{D^I}$, $\mathbf{W^I}$, and $\mathbf{S^I}$ for the document collection, word vocabulary and the section header collection. For a given citation relationship, the one-hot vectors of structural contexts, context words, and sections are projected with the three embedding matrices, denoted as $\mathbf{D^I}_{\{D_n\}}$, $\mathbf{W^I}_{\{C\}}$, and $\mathbf{S^I}_s$. The projected section vectors are then added to the word vectors (each word vector is added to a section vector), and the resultant matrix is denoted as $\mathbf{W'}$. $\mathbf{W'}$ and $\mathbf{D^I}_{\{D_n\}}$ are then concatenated column-wise and form one matrix, i.e., $[\mathbf{w'_1}, ..., \mathbf{w'_m}, \mathbf{d^I_1}, ..., \mathbf{d^I_n}]$, and denoted as $\mathbf{E}$, where m is the number of input context words and n is the number of input structural contexts.

Self-attention Mechanism with Add&Norm

Self-attention [45] is utilized to capture the relatedness between input context words and structural contexts. It applies scaled dot-product attention in parallel for a number of heads, to allow the model to jointly consider interactions from different representation sub-spaces at different positions.

The k-dimensional embedding matrix, $\mathbf{E}$, from the last layer is first transposed and projected with three linear projections ($\mathbf{A}_i^Q$,$\mathbf{A}_i^K$, and $\mathbf{A}_i^V$) to a d_h dimensional space, where $d_h = k/h$, $i \in \{1...h\}$, and h denotes the number of heads. The $\mathbf{E}$ matrix is projected h times, and each projection is called a "head". At each projection (i.e., within a "head"), the dot products of the first two projected versions of $\mathbf{E}$ with $\mathbf{A}_i^Q$ and $\mathbf{A}_i^K$ are computed, and divided by $\sqrt{d_h}$. Subsequently, softmax is applied to obtain the resulting weight matrix with dimensions of $(m+n) * (m+n)$, i.e., $softmax(\frac{\mathbf{E}^T\mathbf{A}_i^Q \cdot (\mathbf{E}^T\mathbf{A}_i^K)^T}{\sqrt{d_h}})$, where $(m+n)$ is the total number of input context words and structural contexts. This weight matrix

represents the relatedness between the input words and articles. The dot product of the weight matrix and the third projected version of $\mathbf{E}$, i.e., $\mathbf{E}^T\mathbf{A}_i^V$, is computed as the output matrix of the head, denoted as $\mathbf{head}_i$. The h numbers of the output head matrices are concatenated column-wise and projected again with $\mathbf{A}^O$ to yield the final output matrix. The computation procedure is represented as follows:

$$SelfAttention(\mathbf{E}) = Concat(\mathbf{head}_1, ..., \mathbf{head}_h)\mathbf{A}^O, \tag{3.1}$$

$$\mathbf{head}_i = softmax(\frac{\mathbf{E}^T\mathbf{A}_i^Q \cdot (\mathbf{E}^T\mathbf{A}_i^K)^T}{\sqrt{d_h}}) \cdot (\mathbf{E}^T\mathbf{A}_i^V), \tag{3.2}$$

where $\mathbf{A}^O \in \mathbb{R}^{k\times k}$, $\mathbf{A}_i^Q \in \mathbb{R}^{k\times d_h}$, $\mathbf{A}_i^K \in \mathbb{R}^{k\times d_h}$, and $\mathbf{A}_i^V \in \mathbb{R}^{k\times d_h}$ are projection parameters. d_h is the embedding dimension of the heads, h is the number of heads, and $k = d_h \times h$, where k is the dimension of the embedding vectors. The output matrix of the self-attention mechanism is then transposed and added to the original $\mathbf{E}$ matrix. Next, dropout is applied [81] to avoid over-fitting, and applied with layer normalization [82] to facilitate the convergence of the model during training. The final output matrix is denoted as $\mathbf{E}'$.

Additive Attention Mechanism

The additive attention layer [77] is utilized to recognize informative contextual words and structural contexts. It takes matrix $\mathbf{E}'$ from the last layer as input, whereby each column represents the vector of a word or document. The weight of each item is computed as follows:

$$\mathbf{Weight} = \mathbf{q}^T \cdot \tanh(\mathbf{V} \cdot \mathbf{E}' + \mathbf{V}'), \tag{3.3}$$

where $\mathbf{V} \in \mathbb{R}^{k\times k}$ is the projection parameter matrix, $\mathbf{V}' \in \mathbb{R}^{k\times(n+m)}$ is the bias matrix, and $\mathbf{q}$ (k-dimensional) is a parameter vector. The **Weight** vector is a row vector of dimension $(m + n)$, where each column represents the weight of a corresponding word or document. The **Weight** vector is applied with the dropout technique to avoid over-fitting.

The output, **EncoderVector**, is the dot product of the softmaxed **Weight** vector and input matrix, $\mathbf{E}'$, where all rows of the embedding vectors are weighted and summed, as illustrated below:

$$\mathbf{EncoderVector} = \mathbf{E}' \cdot softmax(\mathbf{Weight}^T). \tag{3.4}$$

3.3.3 Citation Encoder

The citation encoder is designed to predict potential citations by calculating the probability score between an OUT document matrix, $\mathbf{D^O}$, and the **EncoderVector** from the context encoder, which is defined as follows:

$$\hat{\mathbf{y}} = \mathbf{EncoderVector}^T \cdot \mathbf{D^O}. \tag{3.5}$$

The scores are then normalized using the softmax function as follows:

$$\mathbf{p} = softmax(\hat{\mathbf{y}}). \tag{3.6}$$

3.3.4 Model Training and Optimization

We adopted a negative sampling training strategy [37] to speed up the training process for DACR. In each iteration, it generates a positive sample (correctly cited paper) and n negative samples. Therefore, the calculated probability vector, $\mathbf{p}$, is composed of $[p_{positive}, p_{negative-1}, p_{negative-2}, ..., p_{negative-n}]$. The loss function computes the negative log-likelihood of the probability of a positive sample, as follows:

$$\mathcal{L} = -\log(p_{positive}) + \sum_{i=1}^{n} \log(p_{negative-i}). \tag{3.7}$$

Stochastic gradient descent (SGD) [83] is used to optimize the model.

3.4 Adaptive Detection of Citing Intents

However, when applying for the "on-the-fly" recommendation scenario, DACR might still suffer from information loss, for example, it does not consider the topic semantics of the input manuscript. In addition, the recently developed transformer neural networks [45] could also further improve the performances.

A manuscript dynamic sampling strategy and a transformer-based approach are proposed to detect the citing intents from incomplete drafts adaptively. The mechanism (illustrated in Figure 3.3) involves two components: **base context** and **superstructural context**. **Base context** is relatively stable and functions as the "backbone" for inferring citing intent, whereas **superstructural context** aims to provide supplemental knowledge. Specifically, the **base context** of manuscript sampling includes three sentences: one before the predicting citation,

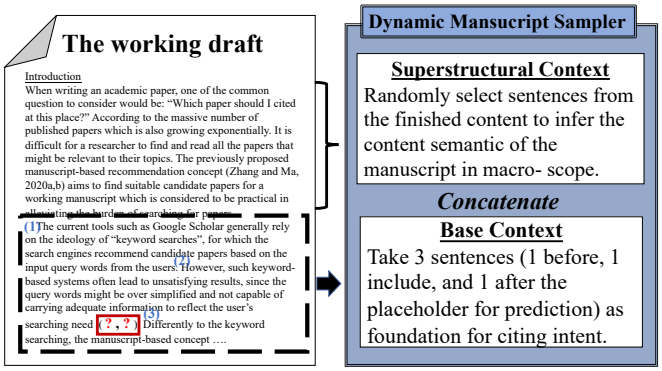


Figure 3.3: Dynamic Manuscript Sampling for Adaptive Detection of Citing Intents

the sentence including the predicting citation, and one after the predicting citation. Base context functions as the "backbone" to infer the citing intent in a micro-scope. The **superstructural context** is defined as a pre-set number of sentences selected from the finished content (the sentences appearing before the predicting citation excluding the base context) to infer the topic semantics of the manuscript. In addition, to simulate the "on-the-fly" application scenario, from which a user is typing sentences continuously to the manuscript, the algorithm is designed to randomly sample sentences from the finished content as the superstructural context.

Based on the hierarchical transformers, a novel approach designed to detect the citing intents adaptively is proposed, namely CBERT4REC, i.e. Content-dependent BERT for Citation Recommendations. The detail of the proposed neural network is presented in Chapter 4.

3.5 Experiments

Two sets of experiments are conducted corresponding for the tasks: recommendation via DACR based on core citing intents, and recommendation via CBERT4REC based on adaptively detected citing intents.

Table 3.1: Statistics of the datasets for DACR

Overview of the Dataset					Count of sections in the Datasset								
		All	Train	Test	Generic Section	Abstract	Background	Introduction	Method	Evaluation	Discussions	Conclusions	Unknown
DBLP	No. of Docs	649,114	630,909	18,205	Train	617,402	9,589	452,430	3,226,521	153,737	19,738	435,514	155,777
	No. of Citations	2,874,303	2,770,712	103,591	Test	5,243	155	6,437	25,956	1,312	200	1875	58,975
ACL	No. of Docs	20,408	14,654	1,563	Train	11,725	114	9,973	42,749	4,186	442	9,456	847
	No. of Citations	108,729	79,932	28,797	Test	3,789	33	3,429	12,625	1,587	159	3,186	0

3.5.1 Recommendation based on core citing intents

Datasets and Preprocessing

DACR is evaluated the recommendation performance of our model and five baseline models on two datasets, namely DBLP and ACL Anthology [24]. The recall, mean average precision (MAP), mean reciprocal rank (MRR), and normalized discounted cumulative gain (nDCG) are reported to compare the models.

The larger dataset, DBLP dataset [24] contains 649,114 full paper texts with 2,874,303 citations (approximately five citations per paper) in the domain of computer science. The ACL Anthology dataset [24] comes with a smaller size, containing 20,408 texts with 108,729 citations. However, it has similar citations per paper (about five per paper) to the DBLP dataset. We split the datasets into a training dataset for training the document, word, and section vectors. We also split a testing dataset with the paper containing more than one citation published in the latest years for recommendation experiments. An experimental overview is provided in Table 3.1.

The texts were preprocessed using ParsCit [84] to recognize citations and sections. In-text citations were replaced with the corresponding unique document ids in the dataset. Section headers often have diverse names. For example, many authors name the "methodology" section using the customized algorithm names. Therefore, we replaced all section headers with fixed generic section headers using ParsLabel [85]. The generic headers from ParsLabel are *abstract, background, introduction, method, evaluation, discussions, and conclusions*. If ParsLabel is not able to recognize a section, we label it as "*unknown*." The detailed counts for each section are listed in Table 3.1.

Implementation Details

DACR was developed using PyTorch 1.2.0 [86]. In our experiments, word and document embeddings were pre-trained using DocCit2Vec with an embedding size

Table 3.2: Citation recommendation results for DACR (** statistically significant at 0.01)

Model	DBLP				ACL			
	Recall@10	**MAP@10**	**MRR@10**	**nDCG@10**	**Recall@10**	**MAP@10**	**MRR@10**	**nDCG@10**
W2V (case 1)	20.47	10.54	10.54	14.71	27.25	13.74	13.74	19.51
W2V (case 2)	20.46	10.55	10.55	14.71	26.54	13.55	13.55	19.19
W2V (case 3)	20.15	10.40	10.40	14.49	26.06	13.21	13.21	18.66
D2V-nc (case 1)	7.90	3.17	3.17	4.96	19.92	9.06	9.06	13.39
D2V-nc (case 2)	7.90	3.17	3.17	4.96	19.89	9.06	9.06	13.38
D2V-nc (case 3)	7.91	3.17	3.17	4.97	19.89	9.07	9.07	13.38
D2V-cac (case 1)	7.91	3.17	3.17	4.97	20.51	9.24	9.24	13.68
D2V-cac (case 2)	7.90	3.17	3.17	4.97	20.29	9.17	9.17	13.58
D2V-cac (case 3)	7.89	3.17	3.17	4.97	20.51	9.24	9.24	13.69
HD2V (case 1)	28.41	14.20	14.20	20.37	37.53	19.64	19.64	27.20
HD2V (case 2)	28.42	14.20	14.20	20.38	36.83	19.62	19.62	27.18
HD2V (case 3)	28.41	14.20	14.20	20.37	36.24	19.32	19.32	26.79
DC2V (case 1)	44.23	21.80	21.80	31.34	36.89	20.44	20.44	27.72
DC2V (case 2)	40.31	20.16	20.16	28.69	33.71	18.47	18.47	25.17
DC2V (case 3)	40.37	19.02	19.02	26.84	31.14	16.97	16.97	23.20
DACR (case 1)	**48.96****	**23.25****	**23.25****	**33.93****	**42.43****	**22.92****	**22.92****	**31.64****
DACR (case 2)	**45.39****	**22.32****	**22.32****	**31.98****	**40.13****	**21.93****	**21.93****	**30.04****
DACR (case 3)	**42.32****	**21.39****	**21.39****	**30.22****	**38.01****	**20.84****	**20.84****	**28.45****

of 100, window size of 50, a negative sampling value of 1000, and 100 iterations (default settings in [25]). The word vectors for the generic headers, such as "introduction" and "method," were selected as the pre-trained vectors for the section headers. DCAR was implemented with 5 heads, 100 dimensions for the query vector, and a negative sampling value of 1000. The SGD optimizer was implemented with a learning rate of 0.025, batch size of 100, and 100 iterations for the DBLP dataset, or 200 iterations for the ACL Anthology dataset. To avoid over-fitting, we applied 20% dropout in the two attention layers.

Word2Vec and Doc2Vec were implemented by using Gensim 2.3.0 [87], and HyperDoc2Vec and DocCit2Vec were developed based on Gensim. All the baseline models were initialized with an embedding size of 100, window size of 50, and default values for the remaining parameters.

Results Analyses

Three usage cases are designed to simulate real-world scenarios:

- Case 1: In this case, we assumed the manuscript was approaching its completion phase, meaning the writer had already inserted the majority of their citations into the manuscript. Based on the leave-one-out approach,

the task was to predict a target citation, by providing the contextual words (50 words before and after the target citation), structural contexts (the other cited papers in the source paper), and section header as input information for DACR.

- Case 2: Here, we assumed that some existing citations were invalid because they were not available in the dataset, i.e., the author had made typographical errors or the manuscript was in an early stage of development. In this case, given a target citation, its local context and section header, we randomly selected structural contexts to predict a target citation. The random selection was implemented using the build-in Python3 *random* function. All case 2 experiments were conducted three times to determine the average results to rule out biases.

- Case 3: It is assumed that the manuscript was in an early phase of development, where the writer has not inserted any citations or all existing citations are invalid. Only context words and section headers were utilized for the prediction of the target citation (no structural contexts were used).

To conduct recommendation via DACR, an encoder vector was initially inferred using the trained model with inputs of cases 1, 2, and 3, and then, the OUT document vectors were ranked based on dot products.

Five baseline models were adapted for comparison with DACR. As the baseline models do not explicitly consider section information, information on the section headers were neglected in the inputs.

1. **Citations as words via Word2Vec (W2V)** This method was presented in [40], where all citations were treated as special words. The recommendation of documents was defined as ranking OUT word vectors of documents relative to the averaged IN vectors of context words, and structural contexts via dot products. The word vectors were trained using the Word2Vec CBOW algorithm.

2. **Citations as words via Doc2Vec (D2V-nc)**[40]. The citations were removed in this method, and the recommendations were made by ranking the IN document vectors via cosine similarity relative to the vector inferred from the learnt model by taking context words and structural contexts as

input (this method results in better performance than the dot product). The word and document vectors were trained using Doc2Vec PV-DM.

3. **Citations as content via Doc2Vec (D2V-cac)** [24]. In this method, all context words around a citation were copied into the cited document as supplemental information. The recommendations were made based on cosine similarity between the IN document vectors and inferred vector from the learnt model. The vectors were trained using Doc2Vec PV-DM.

4. **Citations as links via HyperDoc2Vec (HD2V)** [24]. In this method, citations were treated as links pointing to target documents. The recommendations were made by ranking OUT document vectors relative to the averaged IN vectors of input contextual words based on dot products. The embedding vectors were pre-trained by Doc2Vec PV-DM using default settings.

5. **Citations as links with structural contexts via DocCit2Vec (DC2V)** [88]. The recommendations were made by ranking OUT document vectors relative to the averaged IN vectors of input contextual words and structural contexts based on dot products. The embedding vectors were pre-trained by Doc2Vec PV-DM with default settings.

There are three main conclusions that can be drawn from Table 3.2. First, DACR outperforms all baseline models at 1% significance level across all evaluation scores for all cases and datasets. This implies that the additionally included combined information: namely sections, relatedness, and importance, are essential for predicting useful citations.

Second, performance increases when additional information is preserved in the embedding vectors. When comparing Word2Vec, HyperDoc2Vec, DocCit2Vec, and DACR, Word2Vec only preserves contextual information, HyperDoc2Vec considers citations as links, DocCit2Vec includes structural contexts, and DACR exploits the internal structure of a scientific paper to extract richer information. The evaluation scores increase with the amount of information preserved, indicating that overcoming information loss in embedding algorithms is helpful for recommendation tasks.

Third, DACR is effective for both the large (DBLP) and medium (ACL Anthology) sized datasets. However, we also realized that the smaller dataset requires

higher iterations for the model to produce effective results. It is presumed that more iterations of training can compensate for a lack of diversity in the training data.

The performance of DACR could be further improved by more accurately recognizing section headers. Moreover, we determined that some labels were incorrectly recognized or unable to be recognized by ParsLabel. Therefore, we will work on improving the accuracy of section recognition in future work.

Ablation Tests

Ablation tests are additionally conducted to verify the effectiveness of the three added components: self-attention, additive attention, and section embeddings. Three modified DACR models are run without the corresponding layer, for example, removing the section embedding layer for verifying the effectiveness of section information, removing the self-attention layer for determining the relatedness between contextual words and articles, and removing additive attention for demonstrating the importance of context. The results are illustrated in Figure 3.4.

To conduct in-depth analyses, the citation embeddings of the four models are plotted in Figure 3.6 with the top 10 predicted candidate citations from the full DACR. The dimensions of the citation embeddings were reduced by adapting TSNE [89] implemented via Scikit-learn [90] with default parameters. We aim to inspect the overall distributions of the citation embeddings of the four models, and how locations of the top candidates from the full DACR appearing in the rest of the distribution plots.

Four points could be drawn from Figure 3.4 and Figure 3.6. First, all modified models performed worse than the full model from Figure 3.4, which supports our hypothesis that sections, relatedness, and importance between contextual words and articles are important for recommending useful citations. The relatedness information is more beneficial than section information, which is evident when comparing DACR without section embedding and DACR without self-attention.

Second, DACR without additive attention performed significantly worse with almost zero scores. We consider the primary reason for the 0-close scores of the model without additive attention is that the losses of the model did not converge without the additive attention layer. According to Figure 3.5b, the loss curve of DACR without additive attention has been raised at the beginning of training

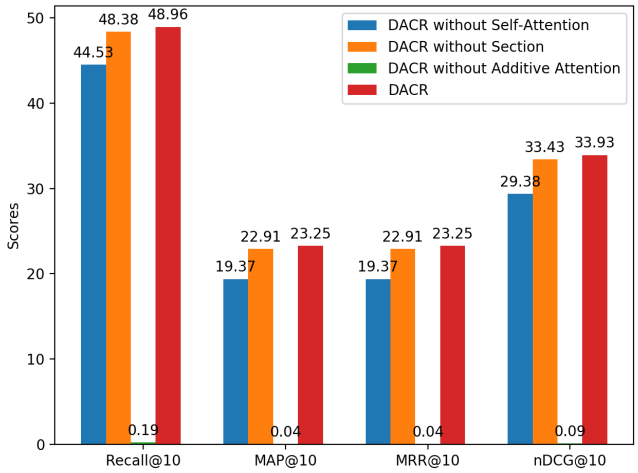
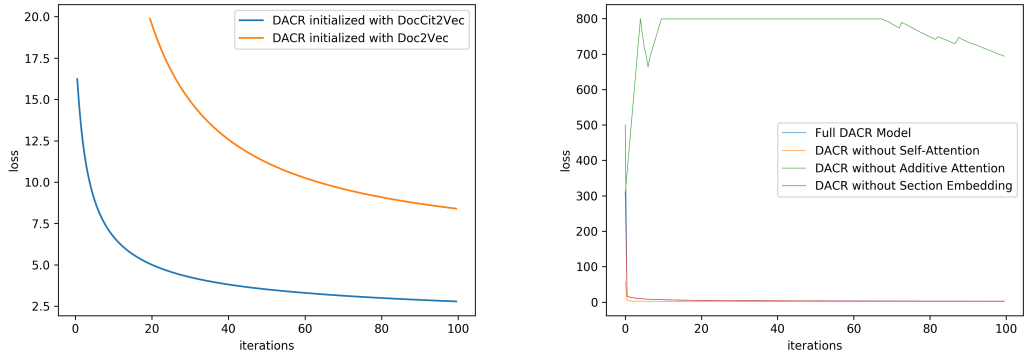


Figure 3.4: Effectiveness of adding sections, relatedness, and importance from DACR

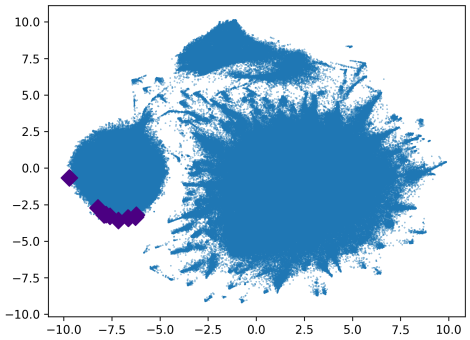


(a) Losses of DACR pre-trained with DocCit2Vec and Doc2Vec on DBLP Dataset

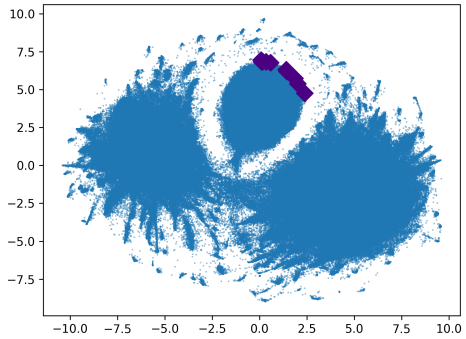
(b) Losses of complete DACR, DACR without Self-Attention, DACR without Additive Attention, and DACR without Section Embedding

Figure 3.5: Plots of Training Losses

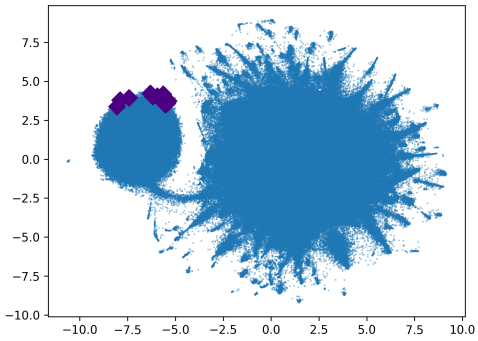
on the DBLP dataset, and maintained at a high level afterwards; whereas the loss curves of the rest of the DACR models (the full DACR, DACR without self-attention, and DACR without section embedding) have been converged at low levels. Therefore, we consider that additive attention has a two-fold purpose: ensuring convergence and learning the importance of context.



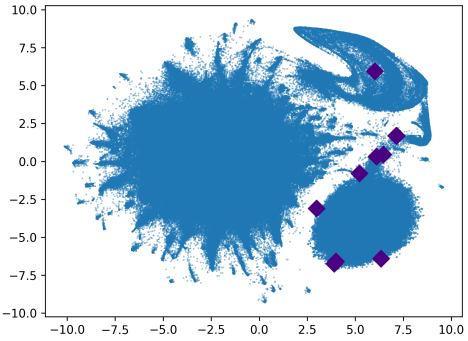
(a) Distribution of Dimension-Reduced Citation Embedding from Full DACR (Diamond dots indicate the top 10 candidates)



(b) Distribution of Dimension-Reduced Citation Embedding from DACR without Section Embedding (Diamond dots indicate the top 10 candidates from the full DACR)



(c) Distribution of Dimension-Reduced Citation Embedding from DACR without Self-Attention (Diamond dots indicate the top 10 candidates from the full DACR)



(d) Distribution of Dimension-Reduced Citation Embedding from DACR without Additive Attention (Diamond dots indicate the top 10 candidates from the full DACR)

Figure 3.6: Distribution of Dimension-Reduced (via TSNE) Citation Embedding from Full DACR, DACR without Section Embedding, DACR without Self-Attention, and DACR without Additive Attention with Top 10 Candidates (diamond dots) via Full DACR for DBLP Sample in Table 3.3

Third, DACR without additive attention did not well preserve the word similarities. Considering Figure 3.6, it is found that the overall distribution of full DACR, DACR without self-attention, and DACR without additive attention are similar. However, the top candidate locations (diamond dots) of DACR without additive attention are widely spread, whereas the candidate locations of full DACR, DACR without section embedding, and DACR without self-attention are closely located. It could be drawn that DACR without additive attention did not preserve the similarity well compare to the rest of the three models. In addition, despite the difference in the overall distribution of the citation embeddings (e.g. DACR without section embedding vs. others), relative positions of the candidates are more important to infer the accurate recommendations.

Lastly, only appropriate combinations of information and neural network layers lead to optimal solutions, as deficits in any of the three types of information (section embedding, relatedness, importance, or attention layers) result in low performance.

3.5.2 Test on adaptive detection of citing intents

This subsection provides the results on whether the topic semantics extracted from the incomplete manuscripts could help to improve the recommendation performances. We adapted the CBERT4REC model as introduced in Chapter 4. We set two experimental scenarios for the testing: 1. use only the query context (base context); and 2. use the query and randomly selected context from the remaining part of the manuscript (superstructural context) for extracting the topic semantics. The number of superstructural contexts is set to 27. The results are shown in Table 4.3, noted as "CB4R (No Dynamic Sampling)" and "CB4R (Dynamic Sampling)". The former indicates the model only considering the query context, whereas the latter denotes the test when considering the topic semantics. It could be drawn that the topic semantics have provided additional improvements.

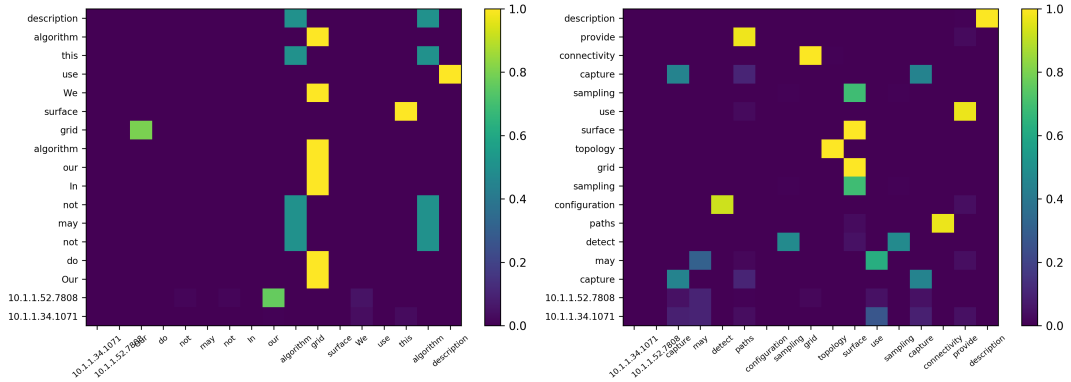
Then, we use the trained CB4R model to conduct tests on different completing stages of manuscripts by limiting the finished content to include 0 (only base context), 7 (few sentences are finished), and 27 sentences (about a paragraph is finished). The results are presented in Table 4.4. It was found that the model performed acceptably during the early development of drafts, which achieved

41.35% on recall@10 when the only base context is available, compared to 18.78% for the best baseline and 31.31% for CB4R without dynamic sampling. However, as the author completes more content of the manuscript, the performances are improved.

3.6 Explainability Study

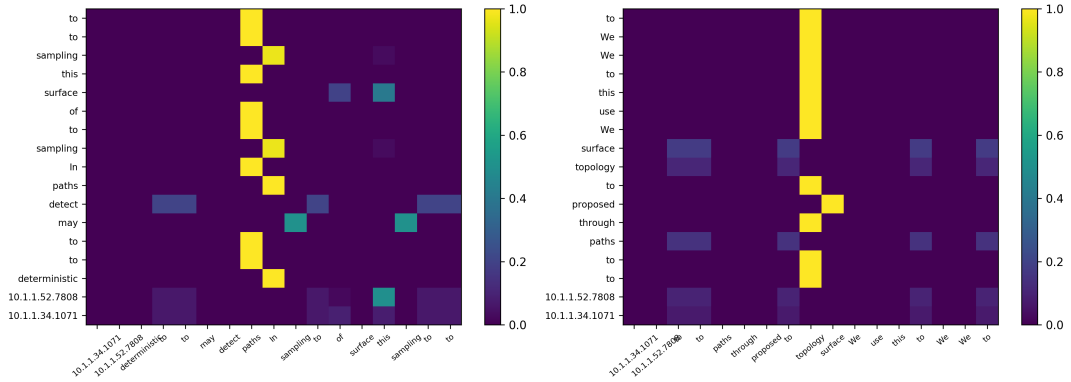
In this section, the weights of self-attention and additive attention in the model are analyzed. The self-attention mechanism generates pair-wise scores for the input words. For example, for every word appearing in a piece of context with n words and m structural contexts, self-attention assigns a $1 \times (m+n)$ weight vector within each head (i.e. a row vector of the resulting matrix $softmax(\frac{\mathbf{E}^T\mathbf{A}_i^Q \cdot (\mathbf{E}^T\mathbf{A}_i^K)^T}{\sqrt{d_h}})$ from Equation 3.2, which sums to 1), where each of the items identifies the weight of correlations between a source word and the target words. The resulting weight matrix $softmax(\frac{\mathbf{E}^T\mathbf{A}_i^Q \cdot (\mathbf{E}^T\mathbf{A}_i^K)^T}{\sqrt{d_h}})$ with $(m+n) \times (m+n)$ dimensions summarizes all the pair-wise word correlation weights, which are presumed to be the "relatedness" between words and structural contexts; whereas the additive attention assigns one score for each item of the input sentence (a $(m+n)$ dimensional vector, i.e. $softmax(\mathbf{Weight})$ from Equation 3.4, and the sum of total scores is 1, where each of the items indicates how much weight it contributes to predicting the final target citation, which is presumed to be the score of "importance" for each item of the input.

Therefore, weights from the two attention mechanisms from the trained models under the case 1 setting (as designed in Section 3.5.1) are fetched and plotted to analyze how the model interprets "relatedness" and "importance" information. Two correctly predicted sample contexts were randomly selected from each of the datasets to illustrate the scores of relatedness and importance for the appearing words and structural contexts. The textual information of the chosen samples is presented in Table 3.3, where the "[=?=]" marker indicates the location for inserting the target citation. For the DBLP sample, we inspect that the citing intent of the authors is to cite the "specific research about a sampling algorithm to generate octree grid by preserving the surface topology"; whereas for the ACL sample, the authors might need to cite a study stating the fact that "their framework was originally developed in NLG to realize deep-syntactic structures".



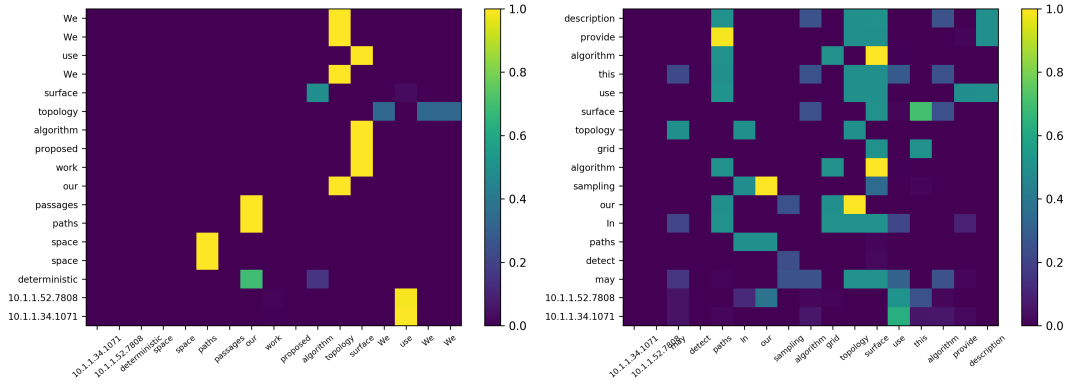
(a) Head 1 Scores

(b) Head 2 Scores



(c) Head 3 Scores

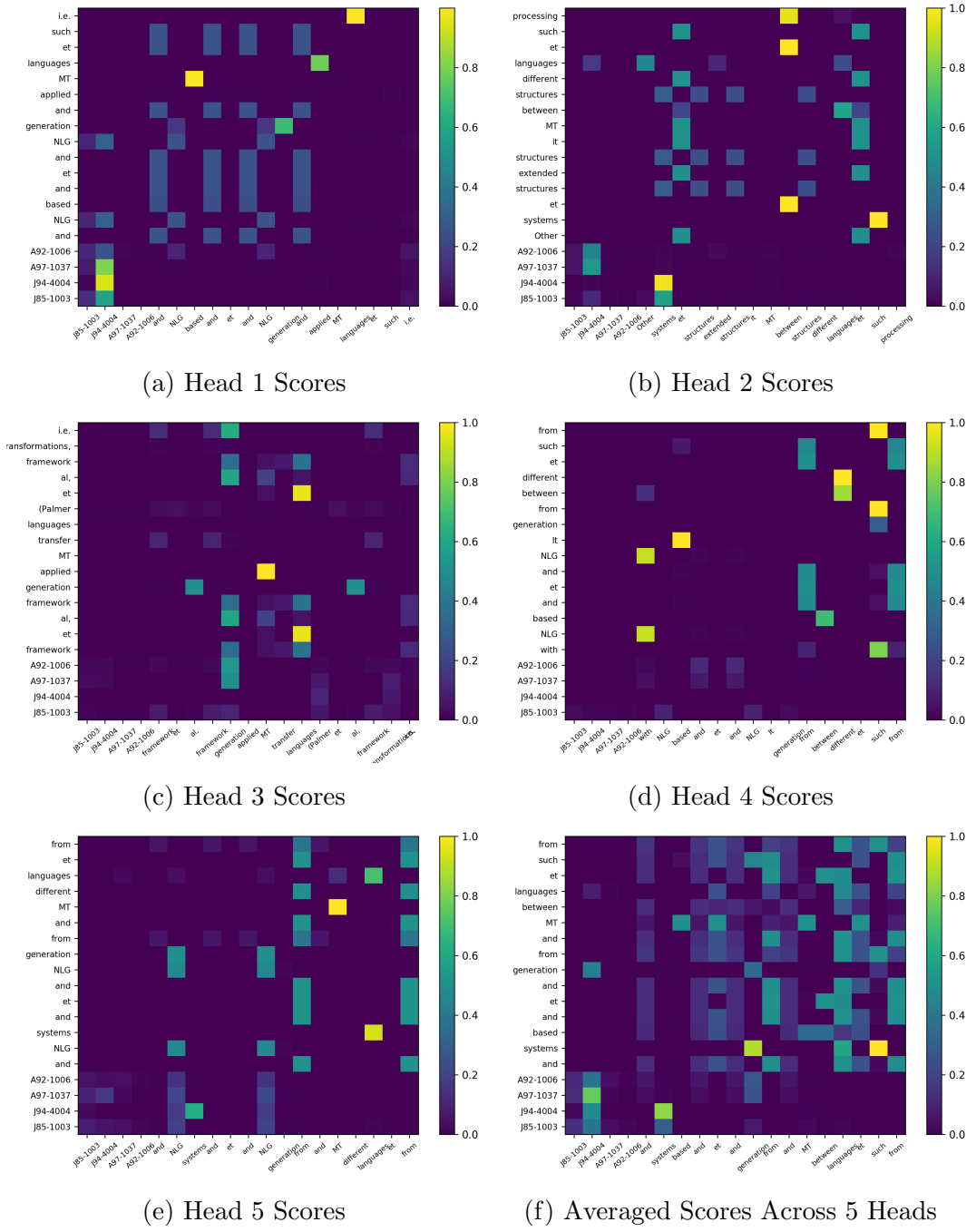
(d) Head 4 Scores



(e) Head 5 Scores

(f) Averaged Scores Across 5 Heads

Figure 3.7: Pair-wise Self-attention Scores (Top 15 Items) for DBLP sample via Complete DACR



(a) Head 1 Scores

(b) Head 2 Scores

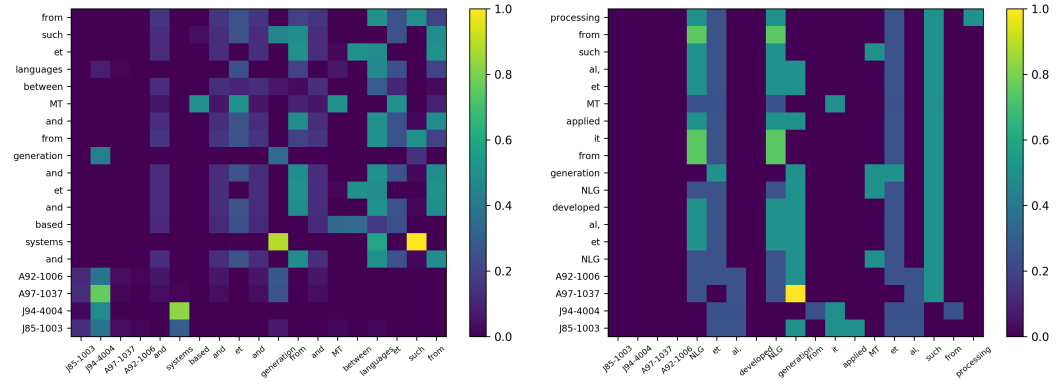
(c) Head 3 Scores

(d) Head 4 Scores

(e) Head 5 Scores

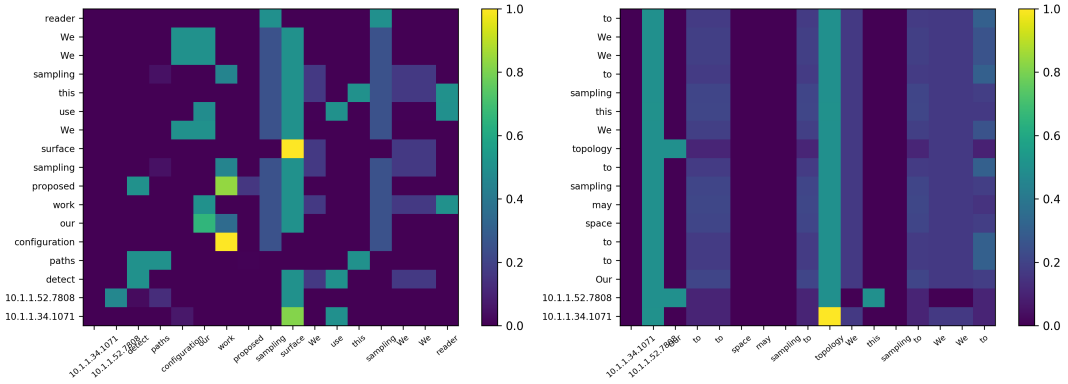
(f) Averaged Scores Across 5 Heads

Figure 3.8: Pair-wise Self-attention Scores (Top 15 Items) for ACL sample via Complete DACR



(a) Self-attention Scores (Averaged from 5 Heads) via Complete DACR for ACL sample

(b) Self-attention Scores (Averaged from 5 Heads) via DACR without Additive Attention for ACL Sample



(c) Self-attention Scores (Averaged from 5 Heads) via Complete DACR for DBLP sample

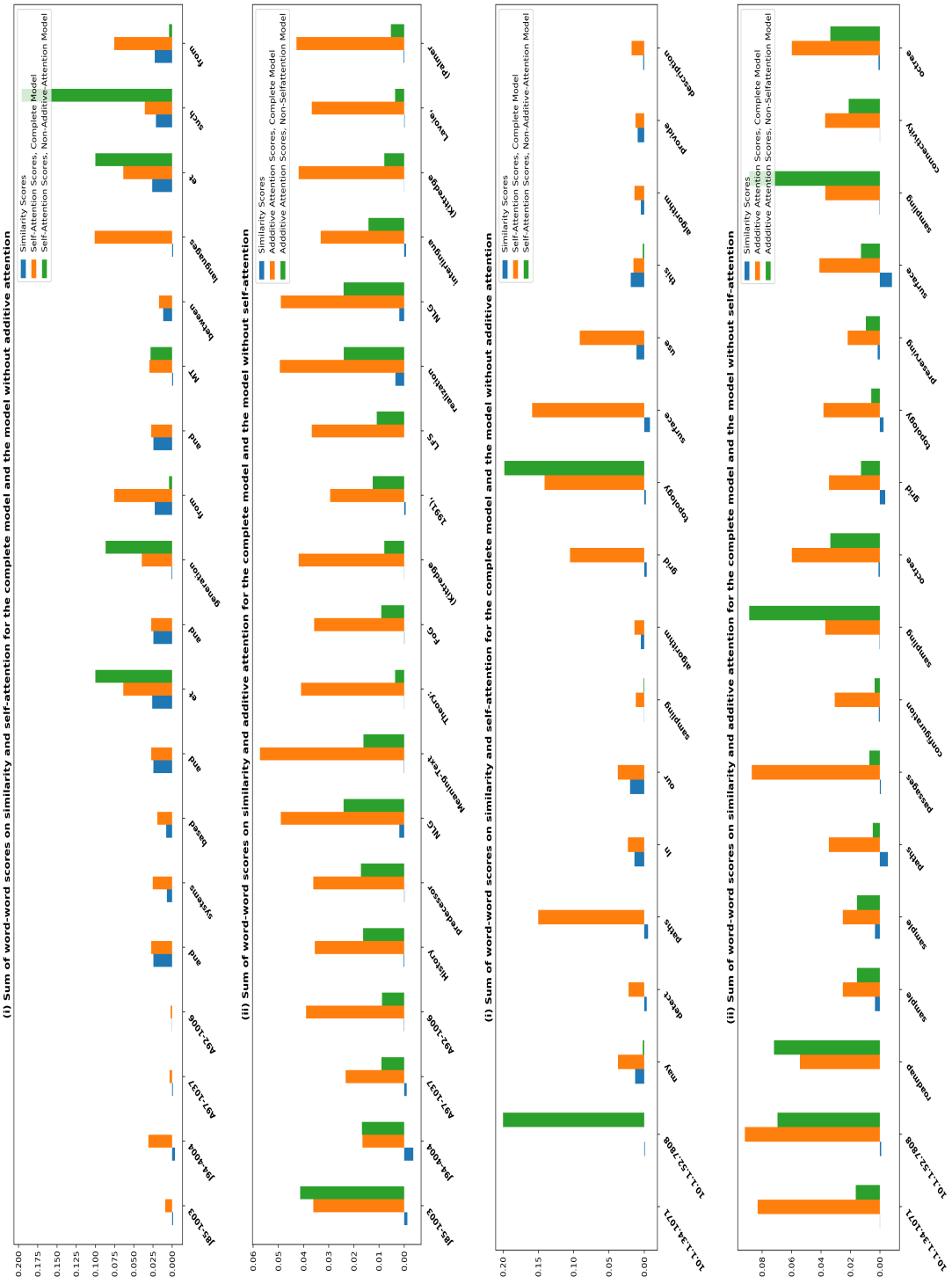
(d) Self-attention Scores (Averaged from 5 Heads) via DACR without Additive Attention for DBLP Sample

Figure 3.9: Comparison of Self-Attention Scores (Averaged from 5 Heads) between the Complete DACR and DACR without Additive Attention

3.6.1 Self-attention Analysis

For the self-attention, it is determined to use the softmaxed pair-wise probabilities as the word-to-word scores of "relatednesses". According to Equation 3.2, within each head, the projected embedding of the context words and structural contexts ($\mathbf{E}^T\mathbf{A}_i^V$) are multiplied by the pair-wise weighted ratios computed by the equation $softmax(\frac{\mathbf{E}^T\mathbf{A}_i^Q \cdot (\mathbf{E}^T\mathbf{A}_i^K)^T}{\sqrt{d_h}})$, where $\mathbf{E}$ is the embedding matrix of the context words and structural contexts, and $\mathbf{A}_i^V$, $\mathbf{A}_i^Q$, and $\mathbf{A}_i^K$ are projection weights. The weight

Figure 3.10: Scores of Additive Attention (Top 15) and Summed Self-attention Against Similarities for the Samples



(a) Sample from ACL dataset

(b) Sample from DBLP dataset

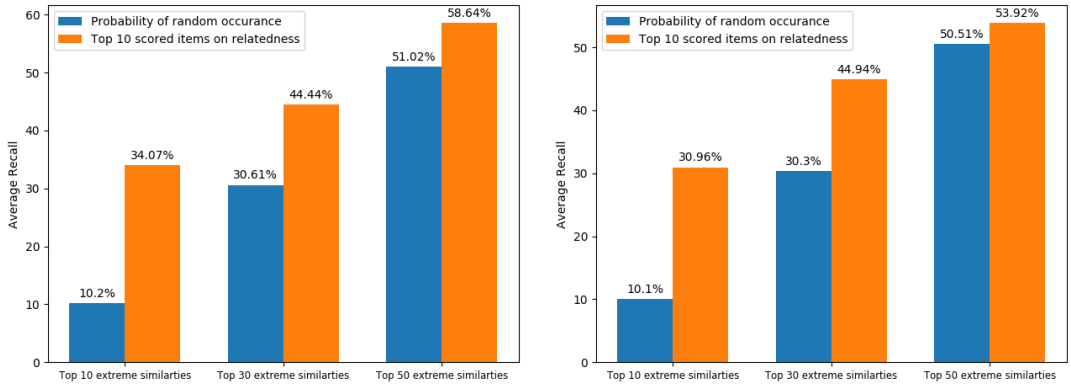
Table 3.3: Textual Information of the Sampled Contexts

Dataset	Source paper ref.	Page	Target paper ref.	Context
DBLP	Varadhan et al. [91]	7	Varadhan et al. [92]	we construct a roadmap in a deterministic fashion. Our goal is to sample the free space sufficiently to capture its connectivity. If we do not sample the free space adequately, we may not detect valid paths that pass through the narrow passages in the configuration space. In our prior work [=?=] we proposed a sampling algorithm to generate an octree grid for the purpose of topology preserving surface extraction. We use this sampling algorithm to capture the connectivity of free space. We provide a brief description of the octree generation algorithm. We refer the reader to [20] for a detailed
ACL	Lavoie et al. [93]	7	Lavoie and Rainbow [94]	History of the Framework and Comparison with Other Systems The framework represents a generalization of several predecessor NLG systems based on Meaning-Text Theory: FoG (Kittredge and 1991), LFS (Iordanskaja et al, 1992), and The framework was originally developed for the realization of deep-syntactic structures in NLG [=?=] It was later extended for generation of deep-syntactic structures from conceptual interlingua (Kittredge and Lavoie, 1998). Finally, it was applied to MT for transfer between deep-syntactic structures of different languages (Palmer et al, 1998). The current framework encompasses the full spectrum of such transformations, i.e. from the processing of

matrix has dimensions $(m + n)$ and $(m + n)$, where m denotes the number of structural contexts and n denotes the number of context words appearing in the sentence. Each row of the weight matrix represents the weight ratios of a word or structural context against all other words and structural contexts from the sentence, which is summed to 1, and presumably treated as the "relatedness" between them. The top 15 pair-wise scores of weight ratios from each head (5 heads in total) and the averaged scores for 5 heads are plotted in Figure 3.7 for the DBLP sample, and Figure 3.8 for the ACL sample.

To make clear explanations, the boldface fonts are used for the items from the horizontal axis in Figure 3.7 and Figure 3.8 (such as "**algorithm**" and "**surface**" at the middle of x-axis in Figure 3.7(a)), and italic font to indicate the items from the vertical axis (such as "*description*" and "*algorithm*" for the top two words in head 1 in Figure 3.7(a)).

Three points could be drawn from Figure 3.7. First, the topic words for inferring the citing intent had received high scores. According to 3.7(f) which pooled all the highly scored words, it is realized that the words such as "**grid**", "**surface**", and "**topology**" had received the highest scores, which are also considered to be highly correlated to the citing intent of the context, i.e. "cite a research about the sampling algorithm by preserving the surface topology". Second, it is realized that



(a) ACL Dataset (b) DBLP Dataset

Figure 3.11: Top 10 Scored Words (or Structural Contexts) on Relatedness vs. Top 10/30/50 Extreme Scored Words (or Structural Contexts) on Similarity

each head had focused on a few words in the sentence, and different heads had focused on different words. For example,"**grid**" in head 1, "**surface**" in head 2, "**paths**" in head 3, "**topology**" in head 4, and "**surface**" in head 5. The averaged scores generally pooled all the highly scored items from each head. Third, it is found that some highly scored words are correlated to almost all the rest of the words in the sentence, such as "**surface**" from head 1; whereas some words are only correlated to a very limited number of words, such as "**description**" which merely correlated to "*provide*" and "*algorithm*" from the averaged head. Generally, DACR trained by the DBLP dataset had testified that the topic words for inferring the citing intent received high scores.

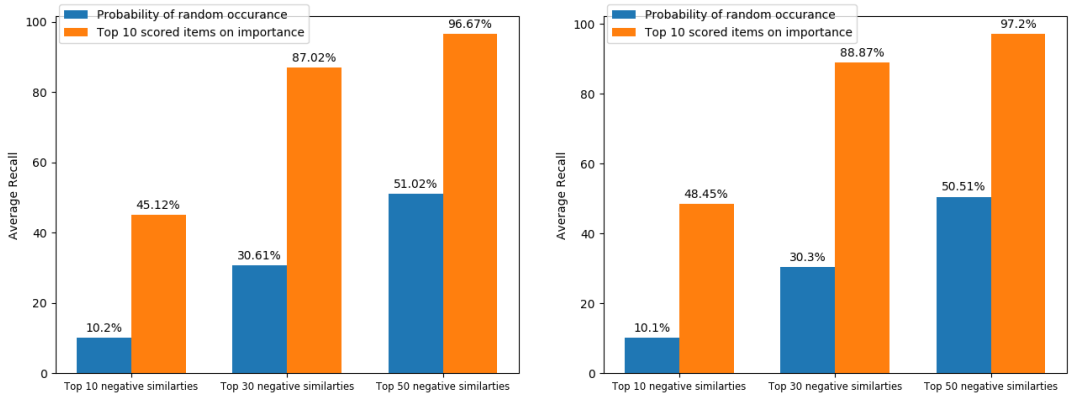
As for the model trained by the ACL dataset shown in Figure 3.8, it is first noticed that the scores from each head are generally scarce than that of the DBLP sample, as the scores are spread across multiple words. Second, it is realized that, not only the topic words (such as "**systems**" from head 2, and "**framework**" from head 3) have received high scores, but also the "connecting words", such as "**and**" from head 1, and "**from**" and "**such**", had also received high scores. Generally, the learned scores from the ACL dataset are less concentrated than the scores learnt from the DBLP dataset. Although the topic words had attracted high weights, however, more connecting words have also been assigned with high weights than the scores learnt from the DBLP dataset.

To further uncover the characteristics of the self-attention scores, it is aimed to

investigate the relationship between the learned scores with the pair-wise similarity of word embeddings, and the relationship between the self-attention scores learned from the complete DACR with the scores from DACR without additive attention. The former aims to study whether the self-attention scores capturing the word semantics; whereas the latter is utilized to further analyze the reason for the failure of the model without additive attention. In Figure 3.10(a) and Figure 3.10(a), we plot the summed self-attention scores along with columns via the complete DACR (orange bars), against the summed pair-wise word embedding similarities (blue bars), and the summed self-attention scores via DACR without additive attention. In addition, the pair-wised scores for the complete model and the model without additive attention are presented in Figure 3.9 for detailed comparisons.

First, it is noticed that some highly scored words for relatedness also yielded high similarity scores (Figure 3.10). For example, the words **"and"** from the ACL sample and **"We"** from the DBLP sample. In addition, some low scored words on similarity, such as "**MT**", and "**languages**" also received high self-attention scores. To make in-depth analyses, we computed the recall of the top 10 highest-scored words or structural contexts on relatedness in the top 10, 30, and 50 words or structural contexts with highest or lowest similarities (extreme similarities), against the probability of random occurrences (number of highest items divided by the total number of words and structural contexts appearing in the input context). Figure 3.11 illustrates the averaged recall for all contexts from the two datasets. According to the figure, the probability of the top 10 scored words or structural contexts on relatedness with extreme similarities is significantly higher than the probabilities of random occurrences for both of the two datasets, especially for the recall among the top 10 and 30 words with extreme similarities.

Second, we compared the relatedness score of the complete model with that of the model without additive attention. It is realized that the self-attention scores are concentrated on few words from the model without additive attention (see Figure 3.9), such as the words "**et**", "**generation**", and "**such**" from the ACL dataset, and "**We**", **"topology"**, and the structural context "**10.1.1.52.7808**" from the DBLP dataset. The rest of the items generally are assigned with close-to-zero scores for the two datasets. In addition, most of the highly scored words are irrelevant to the topic or the citing intent of the context.

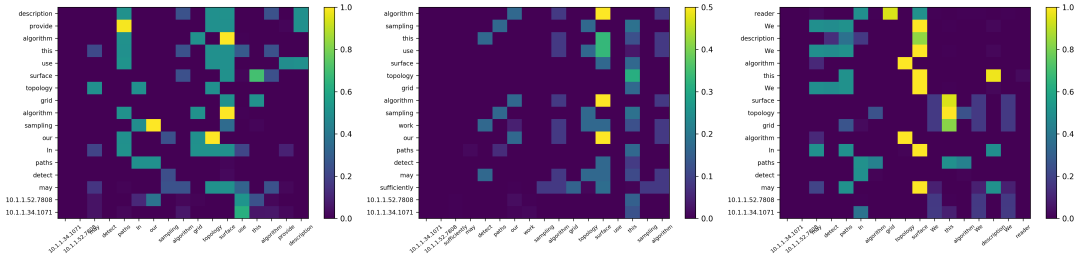


(a) ACL Dataset (b) DBLP Dataset

Figure 3.12: Top 10 Scored Words (or Structural Contexts) on Importance vs. Top 10/30/50 Lowest Scored Words (or Structural Contexts) on Similarity

3.6.2 Additive Attention Analysis

For the additive attention, the importance scores are defined as follows: first, the weight for each embedding is computed according to Equation 3.3, and then the weights are softmaxed by Equation 3.4 to output the weight ratios as the final scores for importance. We plot the top 15 importance scores against the sum of pair-wise similarities of the words and the structural contexts from the sampled sentences in Figure 3.10(a)(ii) and Figure 3.10(b)(ii) for analyses. Two points can be drawn from the plots. First, it is noticed that all of the top 15 scored words (orange bars in Figure 3.10(a)(ii) and 3.10(b)(ii)) from the two samples are basically the unique words from the context (words that are not likely to frequently occur), such as **"NLG"**, and **"Theory"** from the ACL sample, and **"roadmap"**, **"sampling"**, **"surface"**, and **"topology"** form the DBLP sample, which are relevant to the topic of the context. The occurred connecting words from the self-attention mechanism are not assigned with high scores. However, few items are realized to be irrelevant to the topic, such as the words "**1991),**", and"**(Kittredge**" from the ACL sample, which denote a reference from the paper. Adaption of specialized pre-process techniques to filter these words would help to improve the learnt scores on the importance for the context words. Second, most of the highly scored items on importance had the lowest similarity scores (blue bars), such as the words **"History"** and **"Meaning-Text"** from the ACL sample, and **"detect"** and **"surface"** from the DBLP sample are close-to-zero or



(a) Averaged head from DACR with seed 1 (default)

(b) Averaged head from DACR with seed 2

(c) Averaged head from DACR with seed 3

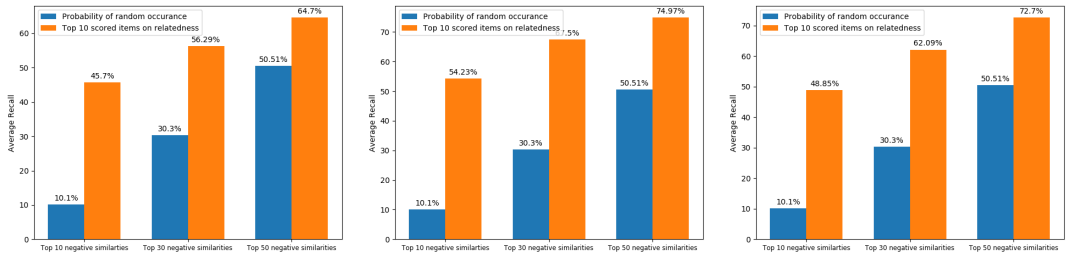
Figure 3.13: Plot of top 15 Self-attention weights of averaged head, and the probabilities of top 10 scored words from self-attention accounted in top 10/30/50 extreme scored words on similarity, from DACR with different seeds

negatively scored on similarity. Figure 3.12 plots the average recall of the top 10 highest scored items on importance in the top 10, 30, and 50 lowest scored items on similarity from all the contexts of the two datasets against the probability of random occurrences. The items with high scores on importance demonstrated superior chances of being scored lower on similarity.

In addition, comparing the scores from the complete DACR (orange bars in Figure 3.10(a)(ii) and 3.10(b)(ii)) with DACR without the self-attention mechanism (green bars), it is realized that the scores are concentrated on few items, such as the words "**sampling**", and "**roadmap**" from the DBLP sample, whereas the scores for the rest of the scores are lowered; similarity for the ACL sample, the scores are concentrated on the words, such as "**NLG**", and "**realization**", however, the intensity is lower than that of the DBLP sample. The reason that the additive attention could prevent the over-concentration of self-attention weights is because of the *softmax* function adopted in the additive attention function, as introduced in Equation 3.4, from which the attention weights are re-scaled according to the uniform distribution.

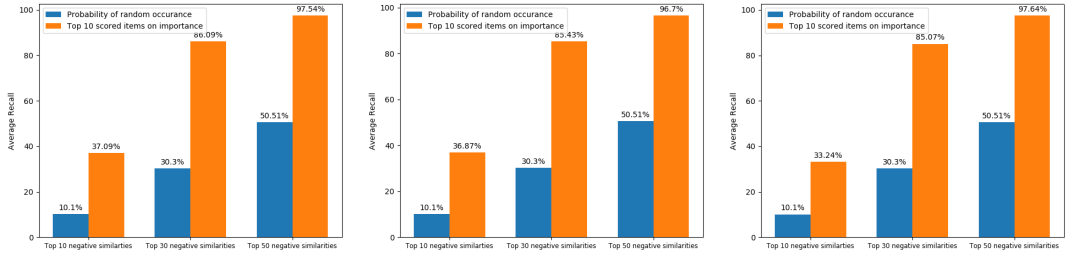
3.6.3 Stability Tests on Different Initialization of Attention Weights

In this subsection, it is aimed to test the stability of the learned weights at self-attention and additive attention. We initialize the weights with three different seeds



(a) DACR with seed 1 (b) DACR with seed 2 (c) DACR with seed 3

Figure 3.14: Probabilities of top 10 scored words from self-attention accounted in top 10/30/50 extreme scored words on similarity, from DACR with different seeds



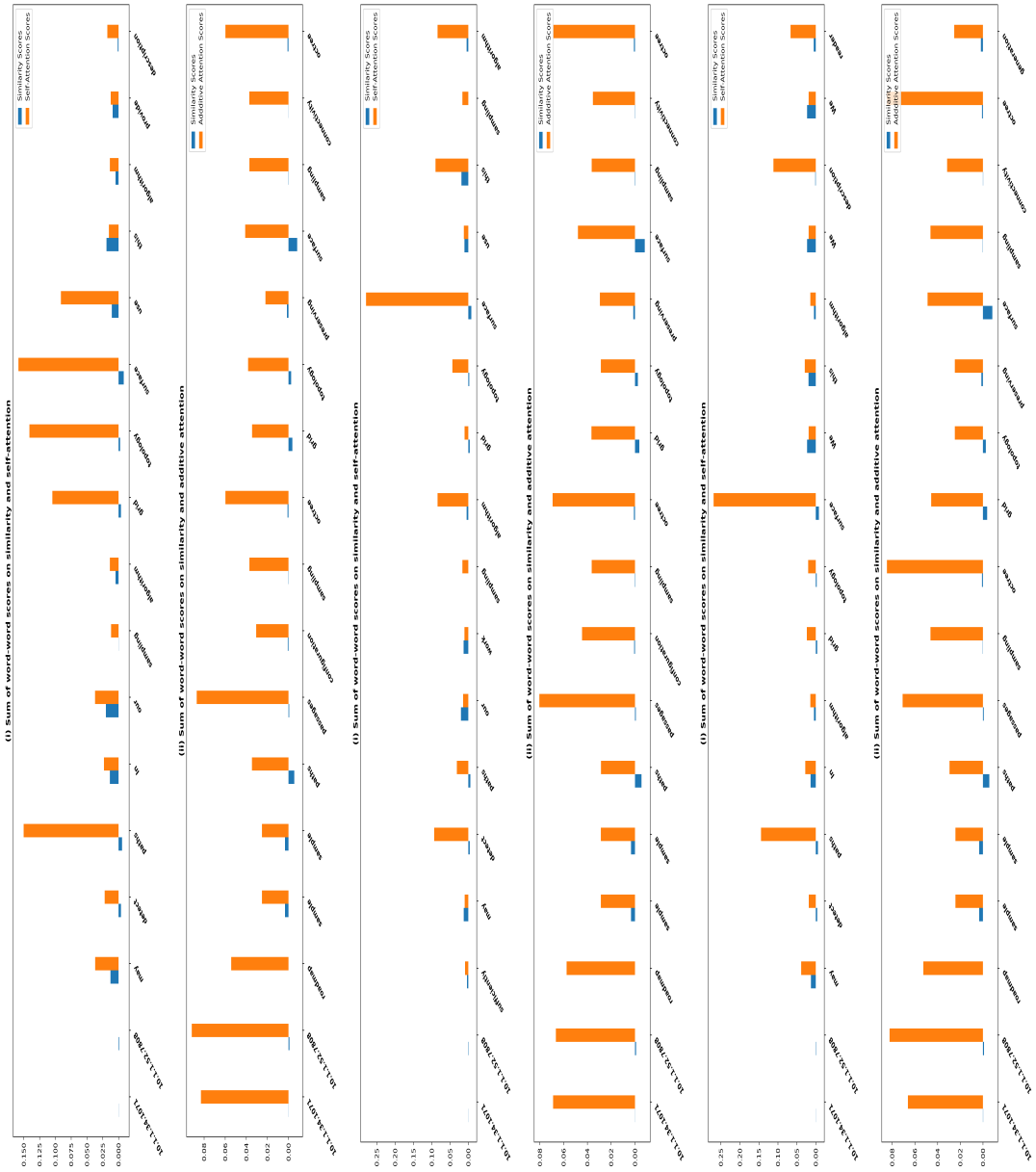
(a) DACR with seed 1 (b) DACR with seed 2 (c) DACR with seed 3

Figure 3.15: Probabilities of top 10 scored words from additive attention accounted in top 10/30/50 negatively scored words on similarity, from DACR with different seeds

at the beginning of the training, so that the weights at self-attention and additive attention were different at the starting point. We report the final recommendation scores from the three runs (Table 3.4), and the plots of the attention weights (Figure 3.13 and Figure 3.16), to inspect whether DACR could produce consistent performances and interpretability through learned attention weights. Three points could be drawn from the table and figures, which are discussed as follows.

First, the recommendation performances are consistent across different seeds. According to the recommendation scores in Table 3.4, it is realized that the differences between the maximum and minimum scores are within 1.50, which result in about 3% maximum percentage change (calculated via $\frac{max-min}{min}$). It is inspected that DACR generally produced consistent performances by initialization from different seeds.

Second, the self-attention weights from the three models initialized with different



(a) DACR initialized with seed 1 (default)

(b) DACR initialized with seed 2

(c) DACR initialized with seed 3

Figure 3.16: Top 15 scored items from sum of Self-attention weights, and additive attention weights, against similarity scores from DACR initialized with different seeds

	Seed 1 (default)	Seed 2	Seed 3	Max Difference	Max %Change
Recall@10	49.51	48.31	49.79	1.48	3.06
MAP@10	23.58	22.95	23.63	0.68	2.96
MRR@10	23.58	22.95	23.63	0.68	2.96
nDCG@10	34.38	33.49	34.49	1	2.99

(a) Recommendation scores from DACR models initialized with three seeds

	Seed 1 & Seed 2	Seed 1 & Seed 3	Seed 2 & Seed 3
Proportion	73.33%	73.33%	100.00%

(b) The proportion of identical items in top 15 words ranked from self-attention weights between the model with three seeds

	Seed 1 & Seed 2	Seed 1 & Seed 3	Seed 2 & Seed 3
Proportion	100%	93.33%	93.33%

(c) The proportion of identical items in top 15 words ranked from additive attention weights between the model with three seeds

Table 3.4: Recommendation scores, proportion of identical items in top 15 words ranked from self-attention, and additive attention

seeds generally extracted similar patterns on "relatedness". According to Figure 3.13, the exact scores for each item are different when the model is initialized with a different seed. However, it is realized that the high scored items from the three models are both correlated with extreme similarities (Figure 3.14). In other words, items scored very high and low on word-wise similarity are gained high scores from self-attention, which is an identical finding to the analysis in subsection 3.6.1. In addition, it is found that most of the highly scored topics are the same from the three seeded models, such as "**paths**", "**topology**", and "**algorithm**" had occurred in both of the three seeded models; and the connecting words, such as "**may**" and "**In**" had also appeared in both models. According to Table 3.4b, the model with seed 1 shared 73.33% same items with the model with seed 2 in the top 15 scored words from self-attention; model with seed 2 also shared 73.33% same items with the model with seed 3; whereas the model with seed 2 shared exactly the same items with model 3 for the top 15 scored words. It could be drawn that, although the exact scores learned from different seeded models are different, however, the weights demonstrated the pattern.

Third, the patter additive attention weights from models with different seeds

also demonstrated even higher consistency. According to Table 3.4c, more than 90% of the items in the top 15 highest scored candidates from additive attention are the same, especially for the model with seed 1 and 2 from which all the highest scored items are the same. In addition, the scores are of the items are very close according to Figure 3.16, which also result in the same pattern, i.e. weight are highly correlated with negative word-wise similarities according to Figure 3.15.

In summary, from the recommendation scores, and pattern of attention weights from the model initialized with three seeds, it could be drawn that, although the exact learned scores can be different, however, the final recommendation performance, and pattern of the weights from two attention mechanisms would stay consistent.

3.6.4 Summary for Attention Mechanisms

To further confirm the patterns of the learned weights, four additional samples are provided (two samples from the DBLP dataset, and two samples from the DBLP dataset) for analyses. In a nutshell, the findings are similar, where the self-attention are relevant to the words with extreme similarity scores, which include the topic-related words, such as "**lexical**", "**alignment**", and "**syntactic**" from supplement sample 1 and 2, and connecting words, such as "**we**", "**by**" from supplement sample 1 and 2; whereas the additive attention emphasizes the words with low similarities, including the topic-related words, such as "**adaptive**", and "**spectral**" from supplement sample 3 and 4, and the unique but irrelevant words, such as "**"the**", and "**"you"**"" from supplement sample 2, which are the wrong words made from the prepossessing procedure, or "**King**" from supplement sample 4, which is unique, but irrelevant to the topic.

In summary, it could be drawn that the "relatedness" captured by the weights of self-attention is correlating to the words with extreme pair-wise similarities, which include both of the topic related words, and connecting words, similarly to the supplemental examples in Appendix A.

Additive attention emphasizes the unique words (with low pair-wise similarities) from the context, which are mostly topic-related words. However, in some occasions where the words are not well pre-processed, they could be mistakenly recognized as unique words.

In addition, according to the stability tests, although the exact learned scores

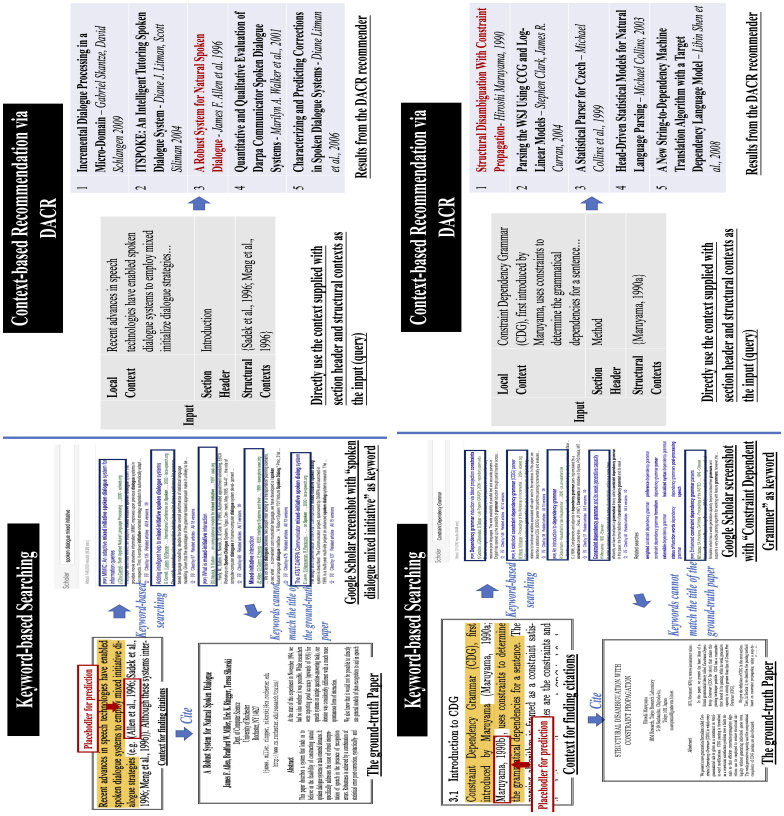
can be different, the final recommendation performance and pattern of the weights from two attention mechanisms would stay consistent.

3.7 User Tests

As discussed in this article's introduction, scholars are generally relying on "keyword-based" search engines to search for citations. However, due to the over-simplicity of the input keywords, which may not carry adequate information to reflect the searching intent of users, they often lead to unsatisfactory searching results, especially when the potential papers' titles do not contain the input keywords.

It is considered that the current keyword-based systems may be limited when applying for two types of scenarios:

1. Scenario 1: when a user would like to find a line of studies in a sub-field, the target papers are difficult to be found by keyword matching with the titles of target papers; whereas the context-based approach matches the semantics of the local context and citations' semantic embeddings which could result in more accurate recommendations. As the example illustrated in Figure 3.17a, a sampled piece of context from [95] in the upper left frame of the left part shows that the author would like to cite a line of studies regarding "dialogue system combined with mixed initiative dialogue strategies". The terms such as "dialogue system", or "mixed initiative strategies" seem to be reasonable as the keywords to be used in Google Scholar for searching. However, since these terms are not fully contained in the title of the target paper which is titled "A Robust System for Natural Spoken Dialogue" [96], so Google Scholar could not effectively find it by matching the keywords with its title. On the other hand, context-based recommender, DACR, directly takes the local context as the input, along with additional inputs, such as the section header and structural contexts, which carry richer information regarding the searching need of the user. Regardless of divergent terms between the titles and input keywords, the candidate citations from context-based systems are found by matching their semantic embeddings and the semantic embedding of the query context. Hence, the target paper was successfully found from our experimental results as shown in the right part of 3.17a.



(a) Scenario 1: keyword-based searches are difficult to find a line studies regarding a sub-field by matching the keywords with the titles; whereas context-based approach matches the semantics of the local context and candidate citations which result in more accurate recommendations.

(b) Scenario 2: keyword-based system could not find the extract paper proposing a specific approach (e.g. Constraint Dependency Grammar) if its title does not contain the name of the approach; whereas the context-based system does not fully rely on the terms of papers' titles, and thus could effectively trace to the target paper by matching the semantics of local context and candidate citations.

Figure 3.17: Two Scenarios where the context-based approach may generate more effective results than keyword-based systems

2. Scenario 2: when a user would like to find the source paper of a specific approach, the keyword-based search engine would not be able to find if the title does not contain the name of the specific approach; whereas the context-based system could successfully find it by matching the semantics of the local context and candidate citations. As the example illustrated in Figure 3.17b, the local context selected from [97] shows that the author would cite the paper which proposed the approach of "Constraint Dependency Grammar". However, the ground-truth paper's title, i.e. "Structural Disambiguation With Constraint Propagation" [98], does not contain the terms "Constraint Dependency Grammar". As a result, Google Scholar could not effectively find the paper in the searching results as shown at the right frame of the left part of the Figure 3.17b. On the other hand, as context-based systems that do not fully rely on the terms of papers' titles, it could effectively trace to the target paper by leveraging the advantage of the semantics of the local context.

It is presumed that the authors of the papers from our datasets also adapted keyword-based systems (or maybe even physical libraries for the early papers) during the writing of the papers. We would like to test whether additional "ground-truth" papers should be cited but not successfully found out due to the limitations of the keyword-based systems.

To this end, in this section, qualitative analyses are conducted to analyze the "wrong predictions" from DACR to test whether there exists "additional ground-truth" papers that the authors should cite; however, they are not successfully found out due to the limitations of the searching tools. The tests are made for two purposes: 1. test the effectiveness of context-based systems on detecting the searching needs of the users; 2. test whether the system can help to check the completeness of the citations for the reviewers of papers.

Specifically, three analyzers are hired to answer a questionnaire designed for evaluation. The ten input context pieces (five from each of the datasets) are selected from eight papers, each of which comes with 5 candidate references recommended from the trained models (please refer to Table 3.5 and Appendix B for the details of the contexts). The three analyzers comprise a third-year doctoral student, second-year doctoral student, and second-year master student majoring in computer science and specialising in natural language processing. For the questionnaire, for each input context, the analyzers are required to answer the

Table 3.5: Summary of questionnaire

Input Contexts (IC)	Citing Intent	Candidates No.	Topic of Candidate	Analyzer's Decision (AD)	Relevancy
IC1	Techniques about sentence alignment	CAN1	Text analysis	AD1: No AD2: No AD3: No	Not Relevant
		CAN2	Machine translation or parameter estimation	AD1: Yes AD2: No AD3: No	Weakly Relevant
		CAN3	English-Chinese alignment	AD1: No AD2: No AD3: Yes	Weakly Relevant
		CAN4	Word correspondence algorithm	AD1: No AD2: No AD3: Yes	Weakly Relevant
		CAN5	Noun phrase alignment	AD1: No AD2: Yes AD3: No	Weakly Relevant
IC2	Noun phrase parsing	CAN1	Part-of-speech tagger	AD1: Yes AD2: No AD3: No	Weakly Relevant
		CAN2	Rule-based parser	AD1: No AD2: No AD3: No	Not Relevant
		CAN3	Anaphora resolution	AD1: No AD2: No AD3: No	Not Relevant
		CAN4	Formalism for parsing grammar statements	AD1: No AD2: No AD3: No	Not Relevant
		CAN5	Analysis of word association norm	AD1: No AD2: No AD3: Yes	Weakly Relevant
IC3	Part-of-speech tagger	CAN1	Part-of-speech tagger	AD1: Yes AD2: Yes AD3: Yes	**Strongly Relevant**
		CAN2	Noun phrase tagger	AD1: No AD2: No AD3: Yes	Weakly Relevant
		CAN3	Rule-based parsesr	AD1: No AD2: No AD3: No	Not Relevant
		CAN4	Rule-based extraction of linguistic knowledge	AD1: No AD2: No AD3: No	Not Relevant
		CAN5	Case study of part-of-speech taggers	AD1: No AD2: No AD3: No	Not Relevant
IC4	Sentence parser	CAN1	Theoretical and empirical study on tree representation	AD1: No AD2: No AD3: No	Not Relevant
		CAN2	Text-chunking	AD1: No AD2: Yes AD3: No	Weakly Relevant
		CAN3	Bilingual alignment	AD1: No AD2: No AD3: No	Not Relevant
		CAN4	Statistical parser	AD1: No AD2: Yes AD3: No	Weakly Relevant
		CAN5	Machine translation	AD1: No AD2: No AD3: No	Not Relevant
IC5	Bilingual alignment	CAN1	Word-sense disambiguation	AD1: No AD2: No AD3: No	Not Relevant
		CAN2	Word-sense disambiguation	AD1: No AD2: No AD3: No	Not Relevant
		CAN3	Word-sense disambiguation	AD1: No AD2: No AD3: No	Not Relevant
		CAN4	Bilingual word coding	AD1: Yes AD2: Yes AD3: Yes	**Strongly Relevant**
		CAN5	Bilingual alignment	AD1: Yes AD2: Yes AD3: No	**Strongly Relevant**

Input Contexts No.	Citing Intent	Candidates (CAN)	Topic of Candidate	Decisions	Relevancy
IC6	Facial modeling and drawbacks	CAN1	Image registration	AD1: No AD2: No AD3: No	Not Relevant
		CAN2	Facial modeling	AD1: Yes AD2: Yes AD3: No	**Strongly Relevant**
		CAN3	Hierarchical motion estimation	AD1: No AD2: No AD3: No	Not Relevant
		CAN4	Optical flow constraint	AD1: Yes AD2: No AD3: No	Weakly Relevant
		CAN5	Facial model	AD1: No AD2: No AD3: No	not Relevant
IC7	Limitation of FACS approach	CAN1	Facial modeling	AD1: No AD2: Yes AD3: No	Weakly Relevant
		CAN2	Facial modeling and limitation of FACS	AD1: Yes AD2: No AD3: Yes	**Strongly Relevant**
		CAN3	Facial modeling	AD1: Yes AD2: No AD3: No	Weakly Relevant
		CAN4	Analysis of facial models	AD1: No AD2: No AD3: Yes	Weakly Relevant
		CAN5	Image motion	AD1: No AD2: No AD3: No	Not Relevant
IC8	Maximum likelihood linear regression (MLLR)	CAN1	MLLR	AD1: Yes AD2: Yes AD3: Yes	**Strongly Relevant**
		CAN2	Maximum aposteriori estimation	AD1: No AD2: No AD3: No	Not Relevant
		CAN3	Hidden Markov model	AD1: No AD2: Yes AD3: No	Weakly Relevant
		CAN4	New covariance matrix	AD1: No AD2: Yes AD3: No	Weakly Relevant
		CAN5	Speech recognition	AD1: No AD2: Yes AD3: No	Weakly Relevant
IC9	Vector quantization	CAN1	Latent dirichlet allocation (LDA)	AD1: No AD2: No AD3: No	Not Relevant
		CAN2	Matrix factorization	AD1: No AD2: Yes AD3: No	Weakly Relevant
		CAN3	Probabilistic latent semantic analysis (PLSA)	AD1: No AD2: No AD3: No	Not Relevant
		CAN4	PLSA	AD1: No AD2: No AD3: No	Not Relevant
		CAN5	Latent variable models	AD1: No AD2: No AD3: No	Not Relevant
IC10	Non-negative matrix factorization (NMF)	CAN1	NMF	AD1: No AD2: Yes AD3: Yes	**Strongly Relevant**
		CAN2	LDA	AD1: No AD2: No AD3: No	Not Relevant
		CAN3	PLSA	AD1: No AD2: No AD3: No	Not Relevant
		CAN4	Auto-encoder with new training technique	AD1: No AD2: No AD3: No	Not Relevant
		CAN5	Matrix decomposition on an over-complete basis	AD1: No AD2: No AD3: No	Not Relevant

question *"What is the ground truth paper about?"*, which aims to evaluate which topics are suitable to be cited in the context. This question is designed to allow the analyzers to perceive the citation intent and hence can be adopted to check whether the analyzers understand the context correctly. For each candidate, they are asked to answer *"Is the candidate paper suitable for use as a citation for the context? Explain reasons, and rate from 0 to 5."*, which is designed to analyze the candidates. The analyzers are expected to provide at least one sentence for each question. The original answers to the questionnaire are provided in Appendix B.

To concisely demonstrate the answers, we summarize the citing intent of the input contexts and the main topic of the associating candidates by using a succinct number of words and the analyzers' decisions according to the original answers from the questionnaire in Table 3.5. If a candidate reference is agreed upon by two or more analyzers to be cited, we indicate the reference to be "strongly relevant." A reference is indicated as "weakly relevant" upon only one analyzer's agreement. The candidate is marked as "not relevant" if no analyzer answered "yes" for the decision. According to Table 3.5, out of the ten input contexts, six of them were detected to have "strongly relevant" candidate(s), that is, input contexts 3, 5, 6, 7, 8, and 10, and eight of them have candidate reference(s) with one agreement, that is, input contexts 1, 2, 3, 4, 6, 7, 8, and 9. In the following subsections, we present the analysis of selected "strongly relevant" and "weakly relevant" samples and evaluate the appropriateness of recommending the structural contexts.

3.7.1 Examination of "strongly relevant" recommendations

To specifically examine the "strongly relevant" candidates made based on two or three agreements, two samples (one from each dataset) with three and two agreements, respectively, from the questionnaire are selected to check the citing intent of the input context and the main topic of the candidates from the original texts and, therefore, to compare with the answers of the analyzers. We select the input context 5 (IC5) from the ACL dataset, for which the forth candidate (CAN4) reference is detected as "strongly relevant," and input context 8 (IC8), for which the first candidate (CAN1) is "strongly relevant." The following shows the text of IC5 from [99] where the "=?=" marker indicates the placeholder for the recommendation.

"*Many corpus-based MT systems require parallel corpora (Brown et al., 1990;*

Brown et al., 1991; =?= ; Resnik, 1999). Kikui (1999) used a word sense disambiguation algorithm and a non-parallel bilingual corpus to resolve translation ambiguity."

Perceptibly, it could be drawn from the context that the authors are citing papers about machine translation that adapts parallel corpora for the placeholder. The fourth candidate article (CAN4) by [100] is considered to propose an algorithm for word correspondence between texts in different languages that could be adapted for machine translation, as stated in their introduction:

"*That is, we would like to know which words in the English text correspond to which words in the French text. The identification of word-level correspondence is the main topic of this paper.*"

Hence, we consider CAN4 could potentially be cited by IC5.

The analyzers' reviews for CAN4 are as the following:

- **Analyzer 1:** Yes. The candidate paper might be appropriate to be cited, as it describes a word correspondence technique to be applied in machine translation based on parallel corpora, which seems to suit the citing purpose. Rate: 4.
- **Analyzer 2:** Yes. This study utilizes parallel corpora and aims to solve the correspondence problem, which can also be applied to MT systems. Rate: 4.
- **Analyzer 3:** Yes. This study focused on identifying words corresponding to parallel corpora, which is a finer-level problem in machine translation tasks. Thus, this agrees with the citing intention. Rate: 4.

It could be inspected that all of the analyzers had correctly detected the citing intent of the input context, and the main topic of the candidate article, and therefore provided the agreements for citing.

Input context 7 (IC7) from the DBLP dataset was selected for examination. The context from [101] states the following:

"*Most of the current systems designed to solve this problem use 'Facial Action Coding System', FACS [10] for describing non-rigid facial motions. Despite its wide use, FACS has the drawback of lacking the expressive power to describe different variations of possible facial expressions =?= .*"

The sentence, including the prediction marker "=?=" indicates that the FACS has a drawback. Hence, it is recognized that the context is looking for papers

describing the drawbacks of the FACS algorithm. The second recommended article (CAN2) for IC7 also addressed the same drawback in their introduction, which is stated as follows:

"*Most such systems attempt to recognize a small set of prototypic emotional expressions, i.e., joy, surprise, anger, sadness, fear, and disgust. This practice may follow from the work of Darwin [9] and more recently Ekman and Friesen [13]... In everyday life, however, such prototypic expressions occur relatively infrequently.*"

The in-text reference "Ekman and Friesen [13]" appearing in the CAN2 context denotes the same paper cited as "FACS [10]" in IC7, which proposed the FACS algorithm. This indicates that the FACS algorithm is insufficient for expressing facial motions that suit the citing intent of CAN2. The reviews from the three analyzers are as follows:

- **Analyzer 1:** Yes. The candidate paper might be suitable to be cited, as it also described the same drawback (lack of expressing facial expressions) in the first paragraph. Rate: 4;
- **Analyzer 2:** No. This paper presents an automatic face analysis (AFA) system to analyze facial expressions based on both permanent facial features (brows, eyes, mouth) and transient facial features (deepening of facial furrows) in a nearly frontal-view face image sequence. It cannot be applied in IC7 because it does not use a realistic parameterized muscle model and focuses on designing features. Rate: 0;
- **Analyzer 3:** Yes. In this study, we developed an automatic face analysis system based on FACS to analyze facial expressions on both permanent and transient facial features. As it is a superior system to FACS, it shows the limitation of FACS and thus becomes proper to be cited. Rate: 4.

According to the reviews, the first and third analyzers recognized the drawback of FACS in CAN2, and therefore made the agreements. The second analyzer detected the main topic of CAN2 correctly; however, he or she missed the point of addressing the drawback. Nevertheless, the two agreements from the first and third analyzers are potentially sufficient for making an appropriate decision.

3.7.2 Examination of "weakly relevant" recommendations

The recommended articles with one agreement are denoted as "weakly relevant" to the input context. It was found that although they would not suit the citing intent of the input context precisely, they might have made points relevant to the main topic of the input context and, therefore, could be additionally cited in a comprehensive manner. Here, we analyze two "weakly relevant" samples, i.e. the input context 1 (IC1) with the second candidate (CAN2) from the ACL dataset and the input context 8 (IC8) with the third candidate (CAN3) from the DBLP dataset.

IC1 is stated as the following [102]:

"*...Aligning English-Chinese parallel texts is already very difficult because of the great differences in the syntactic structures and writing systems of the two languages. A number of alignment techniques have been proposed, varying from statistical methods =?= to lexical methods(Kay and Röscheisen, 1993; Chen, 1993)...*"

The context describes the difficulty of aligning texts in different languages, and it looks for the statistical methods proposed to address this problem at the placeholder. The main topic of the CAN2 article [103] is the proposal of five statistical models for machine translation and methods for estimating the associated parameters. Although proposing statistical methods for text alignment is not the predominant purpose of CAN2, the proposed statistical models can be applied to sentence alignment in different languages for translation according to the context in its abstract [103]:

"*We describe a series of five statistical models of the translation process and present algorithms for estimating the parameters of these models, given a set of pairs of sentences that are translations of one another. We define the concept of word-by-word alignment between such pairs of sentences. For any given pair of such sentences, each of our models assigns a probability to each of the possible word-by-word alignments...*"

The analyzers' reviews are listed as the following:

- **Analyzer 1:** Yes. It might be suitable. The candidate paper proposes a technique for machine translation that involves word-to-word alignment via statistical methods. The paper is also cited in other places for the introduction of machine translation and word alignment. Rate: 4;

- **Analyzer 2:** No. The paper does not propose a new statistical technique for aligning sentences; it details the methods for estimating the parameters of five statistical methods. It is better to use papers that propose these five statistical methods. Rate: 3;

- **Analyzer 3:** No. This paper presents a comparison of a set of statistical models of the translation process and provides algorithms for estimating the parameters of these models. However, it does not involve a text alignment technique itself. Rate: 2.

From among the three reviews, the first analyzer recognized the two-fold purpose of CAN2 and one that suits the citing intent. However, the second and third analyzers merely noticed the most dominant purpose, that is, parameter estimation. Based on the citing intent of IC1, the two-fold purpose of CAN2, and the three reviews, it is argued that although not inevitably necessary, it could be cited in a comprehensive manner or as an extensively related knowledge for the authors to learn.

For the DBLP sample, IC8, the citing context is stated as follows [104]:

"*On each cluster of speech segments, unsupervised acoustic model adaptation is carried out by exploiting the transcriptions generated by a preliminary decoding step. Gaussian components in the system are adapted using the Maximum Likelihood Linear Regression (MLLR) technique (Leggetter & Woodland, 1995; =?=)...*"

It is apparent that IC8 cites articles on the MLLR technique. The associate CAN3 article [105] aims to propose a hidden Markov model (HMM) for speech recognition according to the abstract stated as follows:

"*In this work we formulate a novel approach to estimating the parameters of continuous density HMMs for speaker-independent (SI) continuous speech recognition...*"

It seems that CAN3 had applied a different approach (HMM) to MLLR, which is the citing intent of IC8. However, it should be noted that their HMM approach detailed in section "3. SAT PARAMETER ESTIMATION," is developed based on the MLLR technique, as follows [105]:

"*...In this work we model the speaker specific characteristics using linear regression matrices, motivated by the Maximum Likelihood Linear Regression (MLLR) method [8, 6] that has recently shown to operate effectively in a variety of scenarios of supervised and unsupervised speaker adaptation...*"

The applied HMM also comes with the Gaussian components mentioned in IC8 [104] according to Equation 3 from CAN3 [105]. Hence, it can be concluded that a part of the CAN3's approach is constructed using the same mathematical framework.

The three analyzers' reviews are listed as the following:

- **Analyzer 1:** No. The candidate paper proposes a speech recognition based on HMMs, which is different from the citing purpose. Rate: 0;
- **Analyzer 2:** Yes. This paper proposes an approach to HMM training for speaker-independent continuous speech recognition that integrates normalization as part of the continuous density HMM estimation problem. The proposed method is based on a maximum likelihood formulation that aims to separate the two processes, one being the speaker-specific variation and the other the phonetically relevant variation of the speech signal. In addition, it can be applied for speech recognition. Rate:4;
- **Analyzer 3:** No. This paper presented a novel formulation of the speaker-independent training paradigm in the HMM parameter estimation process. It has a low relevance to the purpose of the citation. Rate: 2.

It can be concluded that although the first and third analyzers detected the main purpose of CAN3 to propose the HMM-based approach, they did not realize the relevance between HMM and MLLR. Nevertheless, the second analyzer notices the technical similarities between the two approaches and provides an agreement. Based on the above analysis of IC8, CAN3, and the reviews, it is argued that although the approach of CAN3 is not strictly based on MLLR, part of its approach contains the same mathematical concepts as MLLR, and, therefore, could be cited in a comprehensive manner to IC8, or as an extensive study by the authors.

3.7.3 Recommendation of structural contexts

Theoretically, DACR carries the information of structural contexts (defined in Definition 2), which is supposed to recommend articles that are frequently cited together. In other words, if a paper is cited by a paper, it may frequently be recommended at other placeholders. Such a recommendation could lead to better accuracy or redundancy. We quantitatively analyze the recommended

structural contexts, out of which, we summarize the useful and redundant articles to determine the effectiveness of the adoption of structural contexts.

According to Table 3.5, out of the 50 candidates in total, 12 candidates are structural contexts (cited in the same paper), which implies that 24% of the recommendations come from the citing paper.

Considering the 12 recommended structural contexts, 5 of them are indicated to be "weakly relevant" and 3 of them are "strongly relevant" which result in 41.56% and 25% respectively, or 66.67% being at least "weakly relevant."

According to the quantitative summaries on the performance of structural contexts, the recommendations are generally effective as 66.67% of the structural contexts is useful. Nevertheless, as these articles are likely to be already known to the users, it is expected that the structural contexts are only adapted for a "remainder" of the users. We subjectively judge that it is slightly redundant for 24% of the recommendation to be from the citing paper. Hence, we will consider designing a penalty mechanism in future work to reduce the ratio of recommending the structural contexts.

Overall, the results show that 6 of 10 sampled contexts have "strongly relevant" candidates, which may imply that these would be the "additional ground-truth" citations that the author did not notice due to the limitations of the searching tools. In addition, although the "weakly relevant" citations might not be strong enough to be used as citations, however, these citations might be helpful to provide supplemental sources for studying the field in a broad view as they are also relevant to some aspects of the field. It is believed that, after further optimizations of the approach (such as adapting larger training datasets, and more sophisticated models), context-based approaches could be applied for assisting the writing of papers and checking the completeness of the citations.

3.8 Summary

In summary, regarding the module of source representation, two main tasks are accomplished: detection of core citing intents via DACR, and adaptive detection of citing intents via CBERT4REC.

They come with the following contributions:

- First, DACR considers three types of essential information: a section for which a user is working and needs to insert citations, relatedness between

the local context words and structural contexts, and their importance, through self-attention and additive attention, which provided significant improvements.

- Second, analyses are conducted to study the correlations between the learned weights and the word semantics. It was found that the highly scored words on "relatedness" by self-attention generally come with extreme similarity scores, whereas the highly scored words on "importance" by additive attention are considered to be unique words relevant to the main topic.

- Third, qualitatively analyses are conducted on the candidates recommended by DACR for selected samples to evaluate whether there exist unnoticed but appropriate citations for the authors. It is believed that, after further optimizations of the approach (such as adapting larger training datasets, and more sophisticated models), context-based approaches could be applied for assisting writing of papers and checking the completeness of the citations.

- Last, the adaptive detection of citing intents can further improve the performances under the on-the-fly scenario.

In future work, I will continue to explore the additional information to detect the citing intent of users more accurately. Also, I will consider adopting more sophisticated neural networks to further improve the performances.

CHAPTER 4

Target Representation

Target representation represents the candidate papers' content semantics and recommends both in-dataset and out-of-dataset papers according to their represented semantic embeddings. Unlike the previous methods adapted the "label-based" embeddings that do not carry the content semantics; and which also limited the ability to recommend newly published papers. This study aims to construct a universal content modelling to represent the content semantics of the existing and newly published papers; hence, the recommendations can be matched by the suitability of the content knowledge and recommend the newly published papers.

This chapter is structured as the follows: Section 4.1 discusses the motivation of creating the content-dependent embeddings, Section 4.2 presents the past studies, Section 4.3 illustrates the approach, and Section 4.4 provides the experimental results and analyses.

4.1 Motivation

Previous works have considered adapting either "label-based" embeddings [24, 27] or "content-based" embeddings [106] generated from abstracts for the embedding of candidate papers. However, they might be constrained to comprehensively represent the content semantics of the candidate papers from in-dataset and out-of-dataset. For example, the label-based embedding is trained from classification

objectives, which do not carry content semantics from the candidate papers. On the other hand, the content-based embedding [106] leverages the abstracts to infer the content semantics. However, the body content could include points that are not contained in the abstract. Hence, embedding content semantics solely from the abstracts would be limited to represent the content semantics comprehensively. Third, owing to the limitations in comprehensively representing the content semantics from the previous label-based and content-based embeddings, as well as the adapted sentence-level neural network [106, 107], the recommendation performances on in-dataset and out-of-dataset papers might also be constrained.

- First, CBERT4REC detects the citing intent directly from the manuscript in a "macro-scope" by leveraging the sentences needing support, and the topic semantics are obtained from incomplete updates of the manuscript. The proposed dynamic context sampling strategy extracts the base context (the local context) as the "backbone" for citing intent, and the superstructural context (context from the finished content of the draft) as the topic knowledge, and leverages the two extracted information to express the purpose for citing of the user comprehensively.

- Second, CBERT4REC can extract the content semantics comprehensively from the in-dataset and out-of-dataset candidate papers by covering the essential points in the papers. The previous content-dependent method, Specter [106] used abstracts for embedding the candidate papers, which may not fully cover the essential points of a paper. We supplement the abstracts with sentences containing essential points not included in the abstracts for comprehensively embedding the content semantics. Inspired by the unsupervised extractive summarization approaches [108, 109], we propose "global centrality" for determining the essential sentences regarding the topic semantics of a paper. We construct biased sampling distributions to draw either sentences with similar points or distant points regarding the main topic to supplement the abstracts for embedding.

- Third, based on the hierarchical transformer [110], we extend the previous content-dependent approach on citation recommendation to be document-level. In combination with the comprehensively captured content semantics from dynamic sampling strategies, the proposed framework can provide

provided superior performances on in-dataset and out-of-dataset recommendations.

The in-dataset and out-of-dataset recommendations as well as the ablation test are conducted, to verify CBERT4REC on four datasets.

4.2 Related Work

This section presents the relative studies regarding the task of creating content-dependent embeddings.

Early embedding models are developed based on DNN-like neural networks, such as Word2Vec [3] and Doc2Vec [38].However, they suffer from information loss when they are applied to citation recommendation tasks. Later developments, such as HyperDoc2Vec [24], and DACR [27], adapted specifically designed fine-tuning models to recover lost information, thus improving the recommendation performance. Recent NLP models adopt the transformer architecture [45] to model universal language knowledge, such as BERT [46], which could be fine-tuned for various downstream tasks, including answering questions and sentence classification. HIBERT [110] extended the BERT model to embed document-level texts for extractive summarization. Sentence-BERT [48] proposed a triplet BERT encoder to conduct sentence-level similarity recognition tasks. This study further extends the existing models for paper-level embedding and recommendation. We modify and compose the hierarchical transformer [110] and triplet-transformer [48] and use them in combination with a dynamic context sampling strategy to construct a universal recommendation model for the task of "on-the-fly" citation recommendations.

The previous context-based approaches [24, 26, 27] take a passage as the input from the manuscript as a query (i.e., the stabilized context sampling) to find the most relevant articles. Although potentially practical, their query vector may be limited in expressing "macro-scoped" citing intents from manuscripts. In addition, their citations vectors are label-based embeddings or content embeddings by encoding the abstracts, which may not effectively represent the paper-level semantics.

Thereby, it is proposed to adopt the transformer neural networks to generate content modelling by learning the contents of academic papers and, hence, creating the content dependent embedding, for better effective matching.

4.3 Content-dependent BERT

4.3.1 Problem Definition

In this section, **academic papers** and **citation relationships** are defined following previous terminologies [24, 27, 26]. Based on these definitions, we define the task of **citation recommendation**.

Definition 3 (Embedding Model)**.** *An embedding model is defined as a function* $\mathbb{E}$ *which represents a paper* $\mathcal{H}$ *into a d-dimensional vector, i.e.* $\mathbf{h} = \mathbb{E}(\mathcal{H})$.

Definition 4 (Citation Recommendation)**.** *Given a source paper* $\mathcal{H}_s$*, a collection of papers* $\mathbb{H}$*, and an embedding model* $\mathbb{E}$*, the task is defined to find top k ranked papers* $\{\mathcal{H}_1, \mathcal{H}_2, ..., \mathcal{H}_k\}$ *based on the geometric distances between embeddings of* $\mathbb{H}$*, i.e.,* $\mathbf{H} = \mathbb{E}(\mathbb{H})$ *and the embedding of* $\mathcal{H}_s$*, i.e.,* $\mathbf{h}_s = \mathbb{E}(\mathcal{H}_s)$.

4.3.2 The Model

CBERT4REC involves three main components: pre-training, dynamic sampling and fine-tuning. Pre-training aims to model the contents; dynamic sampling is proposed to sample the essential context to infer the citing intents from the manuscripts, and content semantics from the candidate papers; and the fine-tuning aims to optimize the embedded distances between the manuscript embedding the ground-truth embedding.

Pre-training

The pre-training task aims to learn universal content knowledge from academic texts. We adapted the hierarchical transformer [110] (see "Pre-training Architecture" in Figure 4.1) with modifications on pooling strategies (we used MEAN pooling instead of the EOS pooling from the original model) for aggregating sentence-level representations, as shown in Equation 4.3.

The model comprises a **sentence encoder** to obtain sentence-level representations, a **document encoder** to encode the document-level vectors and a decoder to decode the masked sentences as the objective training task. The two encoders adapted the same architecture of the transformer encoder [45]. Given a paper $\mathcal{H}$ defined in Definition 1, its content words, $\hat{W}$, is processed into sentences,

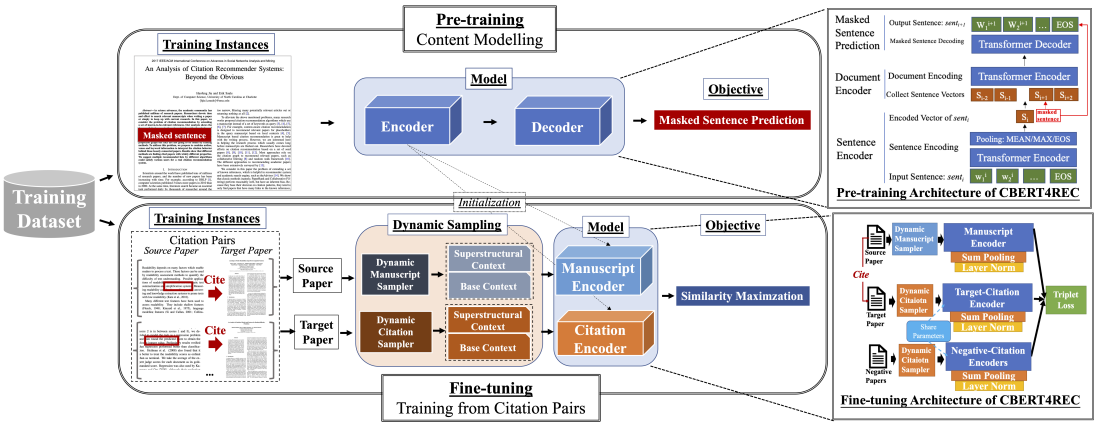


Figure 4.1: The pipeline of CBERT4REC: the pertaining model, dynamic context sampling and fine-tuning model

that is, $\mathcal{H} = (\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_{|\mathcal{H}|})$, where $\mathcal{S}_i = (w_1^i, w_2^i, ..., w_{|\mathcal{S}_i|}^i)$ denotes a sentence with words from $\hat{W}$, where w_j^i represents a word from the sentence $\mathcal{S}_i$. At the end of each sentence, a special EOS (*end-of-sentence*) token is inserted.

The **sentence encoder** first embeds words from $\mathcal{S}_i$ into vectors added its positional embeddings, and then encodes the resulting embeddings using the transformer encoder [45]. The output comprises $|\mathcal{S}_i|$ word embedding vectors with a preset dimension:

$$\mathbf{S}_i^I = \{\mathbf{w}_j^i + \mathbf{p}_j | w_j^i \in S_i\}, \tag{4.1}$$

$$\mathbf{S}_i^O = \mathbf{TransformerEncoder}_{\mathbf{sent}}(\mathbf{S}_i^I), \tag{4.2}$$

where w_j^i is the j-th word from the i-th sentence; $\mathbf{w}_j^i$ is the embedding of w_j^i; and $\mathbf{p}_j$ is the positional embedding for the j-th word. The final sentence embedding $\mathbf{S}_i^O$ composes resultant word vectors $(\mathbf{w'}_1^i, \mathbf{w'}_2^i, ..., \mathbf{w'}_{|\mathcal{S}_i|}^i)$. The MEAN pooling is then adapted to aggregate the sentence vectors and added to its positional embedding:

$$\mathbf{S}_i = \frac{1}{|\mathcal{S}_i|} \sum_{j \in |\mathcal{S}_i|} \mathbf{w'}_j^i + \mathbf{p}_j. \tag{4.3}$$

The gathered sentence vectors $(\mathbf{S'}_1, \mathbf{S'}_2, ..., \mathbf{S'}_{|\mathcal{H}|})$ are then transformed via the **document encoder** , which is another transformer encoder embed sentences with knowledge of neighbor sentences:

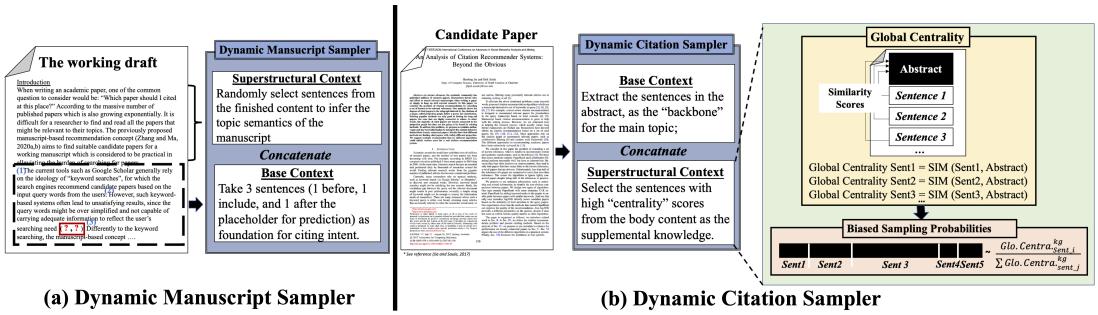


Figure 4.2: Illustration of Dynamic Sampling Strategy: Manuscript Sampler and Citation Sampler based on Global Centrality

$$\mathbf{H} = \mathbf{TransformerEncoder}_{\mathbf{doc}}(\{\mathbf{S}'_i | \mathcal{S}_i \in \mathcal{H}\}), \tag{4.4}$$

where $\mathbf{H}$ composes $(\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_{|\mathcal{H}|})$ are the content-dependent sentence embeddings encoded from the document encoder for the final output.

We adapted the masked sentence task [110] as the training objective, where 15% sentences from the input are masked. The stacked transformer decoders are adopted to decode the masked sentences.

Dynamic Context Sampling Strategy

After obtaining the pre-trained model, the algorithm conducts dynamic sampling to generate the input contexts for the fine-tuning model as illustrated in Figure 4.1. Given a citation relationship, which involves a source paper (the manuscript) and a cited paper. We designed two samplers to sample the contexts from the source and cited paper:

- **Dynamic manuscript sampler**: samples essential context from the manuscript (the source paper) to reflect the citing intent in a macro-scope, and also simulate the "on-the-fly" scenario where an author is inserting new sentences;
- **Dynamic citation sampler**: samples essential context that ultimately covers the essential points discussed in the paper.

Both the **manuscript sampler** and **citation sampler** are composed of two essential components: **base context** and **superstructural context**. **Base context** is relatively stable and functions as the "backbone" for inferring citing

intent or content semantics, whereas **superstructural context** aims to provide supplemental knowledge.

Dynamic Manuscript Sampler (Figure 4.2(a))

The **base context** of manuscript sampling includes three sentences: one before the predicting citation, the sentence including the predicting citation, and one after the predicting citation. Base context functions as the "backbone" to infer the citing intent in a micro-scope. The **superstructural context** is defined as a pre-set number of sentences selected from the finished content (the sentences appearing before the predicting citation excluding the base context) to infer the topic semantics of the manuscript. In addition, to simulate the "on-the-fly" application scenario, from which a user is typing sentences continuously to the manuscript, the algorithm is designed to randomly sample sentences from the finished content as the superstructural context. For a default setting, we set the manuscript sampler to include 30 sentences, which contains the 3-sentences base context, and 27-sentences superstructural context, according to our maximum GPU memory consumption.

Dynamic Citation Sampler (Figure 4.2(b))

The **base context** of citation sampling functions as the backbone for inferring the content semantics. Inspired by the content-dependent approach from [106], we defined the base context to be the abstract of a paper, because they include the essential points from papers. However, the abstracts are still short in length, which may miss stating some points from the body content; therefore, they may lead to information loss for the embedded content semantics.

4.4 Experiments

This section presents the datasets, and analyze the experiments for in-dataset and out-of-dataset test, and tests for dynamic sampling.

Table 4.1: Statistics of the datasets for testing CBERT4REC

	DBLP-1	DBLP-2	DBLP-3	ACL	In-common Papers btw DBLP-1 and DBLP-2
Total Doc. / Cit. No.	50,000 / 96,698	50,000 / 160,797	50,000 / 145,079	20,405 / 103,557	10,000 (20%) / 16,496
Train Doc. / Cit. No.	38,436 / 51,747	38,475 / 99,220	38,475 / 90,256	16,634 / 57,721	6,866 / 8,101
Test Doc. / Cit. No.	4,270 / 13,674	4,274 / 18,944	4,274 / 16,498	1,848 / 18,522	401 / 1,720
New-paper Doc. No.	2,501	2,501	2,501	1,021	235
New-paper Test Doc. / Cit. No.	4,320 / 10,038	4,415 / 9,388	4,052 / 10,011	898 / 3,073	198 / 900
Mixed Test [a] Doc. / Cit. No.	4,793 / 19,966	4,750 / 21,417	4,410 / 21,832	902 / 10,467	245 / 1,970

[a]Mixed test set contains all the papers in "new-paper test set", however we only use the cited docs from the "new-paper set" for conducting the new-paper recommendations, and cited docs from both of the "new-paper set" and the "train set" for the mixed recommendations.

4.4.1 Dataset

Four datasets were adapted, including the ACL Anthology (2013 release) and three datasets generated from the DBLP corpus[1]. The ACL Anthology corpus contains 20,405 full-paper texts, with 107,2418 citations, whereas the larger corpus DBLP has 649,114 papers, with 2,874,303 citations. To cross-evaluate the performances on out-of-dataset papers, we produced three datasets from the DBLP corpus, that is, DBLP-1, DBLP-2, and DBLP-3, each of which contains 50,000 papers. The three DBLP datasets were generated based on a "biased-individuality" strategy, by which DBLP-1 and DBLP-2 share 20 % papers (10,000 papers) in common, whereas the rest were all different. DBLP-3 contains completely different papers to DBLP-1 and DBLP-2. We expect the DBLP-3 to function as an individual dataset, whereas the common papers from DBLP-1 and DBLP-2 might help evaluate the model's stability. The complete ACL corpus was used for the fourth dataset.

The datasets were divided into four parts: a train set for training the models, a test set for conventional (in-dataset) recommendation tasks, a new-paper set, and a test set for new-paper (out-of-dataset) recommendations. The statistics of the dataset are listed in Table 4.1. Five% of the papers published in recent years were chosen as the new-papers, with 10% of the papers chosen from DBLP-1, DBLP-2, and DBLP-3 as test set for new-papers, or 5% from the ACL dataset as the test set for new-papers.

Table 4.2: Parameters of CBERT4REC

Transformer Params	**Block No. (L)**	**Hidden Size (H)**	**Attention No. (A)**
	6	768	12
Optimizer Params	**Update Schedule**	**Warmup Updates**	**Update Frequency**
	Inverse Square Root	1,000 (ACL) / 2,000 (DBLP)	4 (pretrain) / 8 (finetune)
	Warmup Learning Rate	**Learning Rate**	**Weight Decay**
	1e-7 (pretrain) / 1e-9 (finetune)	1e-4 (pretrain) / 2e-5 (finetune)	0.01
	β_1	β_2	**Dropout**
	0.9	0.999 (pretrain) / 0.98 (finetune)	0.1

4.4.2 Implementation Details

CBERT4REC was developed using Fairseq 0.50 [111] and PyTorch 1.2.0 [86]. The baseline models, Word2Vec and Doc2Vec, were implemented using Gensim 2.3.0 [87], and HyperDoc2Vec and DocCit2Vec were developed using Gensim 2.3.0. The DACR was developed using the Gensim 2.3.0, and PyTorch 1.2.0. SciBERT and Specter were implemented based on Huggingface 4.2 [112].

The parameters for pre-training and fine-tuning model of CBERT4REC, and Adam optimizer [113] are presented in Table 4.2. We ran 30 iterations of pre-training and 3 iterations of fine-tuning for DBLP-1, DBLP-2, and DBLP-3 datasets, or 300 iterations of pre-training and 4 iterations of fine-tuning for the ACL dataset, as it was found that the smaller dataset required more iterations for the loss function to converge.

The default parameters of the dynamic context sampling strategy were as follows: For the manuscript, we set the total number of sentences to 30, which includes the base context composed of three sentences (the sentence including the target citation, the sentence before it, and the sentence after). Superstructural contexts include 27 sentences randomly selected from the content before the base context. For a citation paper, we set the total number of sampling contexts to 60, including 30 sentences for the base context (randomly selected from the abstract) and 30 for the superstructural context (randomly selected from the body content). In addition to the default setting, subsection 4.4.6 describes the tests that were performed using different settings of the sampling process for manuscripts. We used the build-in *Random* function from Python3 to draw samples from the generated biased probabilities as explained in Section , the seeds are updated according to the present time of training, to make the random

[1]Both the ACL Anthology and the DBLP corpus were adapted from [24].

numbers randomized ultimately.

Because of the limitations of our facilities, we did not conduct open-source pre-training over a large-scale language dataset, as conducted in [46] and [110]. Instead, we directly began from the in-domain pre-training stage individually on the DBLP datasets or the ACL dataset. We limit the maximum number of words in a sentence to 50 words and the maximum length for a text to be 30 sentences. For pre-training, texts with lengths longer than 30 sentences were sliced into multiple texts. The batch sizes were set to 7 for pre-training and 1 for fine-tuning.

For the baseline models, Word2Vec and Doc2Vec were trained with an embedding size of 100, a window size of 50, and default settings on the remaining parameters. We also experimented with different sets of parameters; however, the variations in the results were negligible. For HyperDoc2Vec, DocCit2Vec, DACR and Specter, we adapted the same parameters from the original papers [24, 26, 27, 106]. For SciBERT [107], we fine-tune the pre-trained model with for 3 iterations on DBLP-1, DBLP-2, and DBLP-3, or 4 iterations on ACL (the same number of fine-tune iterations as DBERT4REC). Specter and SciBERT were fine-tuned with "augmented abstracts", i.e. abstract concatenated with all the base contexts from a paper. The objective is to allow these two algorithms carrying knowledge on both abstract and citing intents, making them adaptable to our application scenario.

For the hardware, we used four GPUs, either Nvidia 1080Ti or T4, according to the availability of servers for running the pre-trained model. Depending on availability, the fine-tuned models were processed on two GPUs, including the Nvidia 1080Ti, 2080Ti, and T4.

We applied ParsCit [84] to analyze the in-text citations that were replaced by unique citation IDs for detection. ParseLabel [85] was applied to recognize the abstracts and the rest of the section headers. CoreNLP [114] was applied to separate the sentences. Rare words were compressed to reduce memory consumption by applying byte pair encoding [115] with the adaption of the learned vocabulary table from [110].

4.4.3 Recommendation Methods for CBERT4REC and Baselines

Baselines

We adapted six baseline models for comparisons: Word2Vec (W2V) [3], Doc2Vec [38], HyperDoc2Vec (HD2V) [24], DACR [27], SciBERT [107], and Specter [106]. W2V and D2V are the two conventional embedding algorithms, which produce fixed word embedding by preserving the information of the context they belong to. HD2V and DACR are the two fine-tuning models based on D2V, to predict the linked documents by their label embeddings. SciBERT and Specter are the two latest models for the recommendation of scientific articles based on the sentence-level transformer.

In-dataset Recommendation

Given a manuscript from the test set, we masked the citations and used the ones from the train set as the ground truth for prediction. For **CBERT4REC** (**CB4R** thereafter), we ran the citation sampler on all the papers from the train set and encoded the sampled contexts as the candidate **citation vectors** by using the averaged word vectors from the last layer. For each paper from the test set, we ran the manuscript sampler on it regarding a masked location and encoded it as a **query vector**. **Citation vectors** were ranked by taking cosine similarity with the **query vector**. It was found that different similarity approaches yielded similar scores; hence, we herein only report the scores computed using cosine similarity.

For **Word2Vec(W2V)**, **Doc2Vec(D2V)**, **HyperDoc2Vec(HD2V)**, **DocCit2Vec(DC2V)**, and **DACR** baselines, we followed the same settings as those provided in the original studies [24, 26, 27], which uses the 50 words before and after a target citation as the base context. The query vectors were computed as the average of the input embeddings of the base context. Candidate citation vectors were produced using various approaches: For **W2V**, we use the output word embedding of the citation IDs as citation vectors; for **D2V**: we use inferred vectors from content words of a paper using the trained model; for **HD2V, DC2V** and **DACR**: we use the output embeddings of citation IDs. Citation vectors are ranked by dot product with the query vector adapted according to original

methods. For **SciBERT** and **Specter**: citation vectors are the averaged word vectors from the last layer of the encoder by encoding the abstracts; the query vector was encoded by the base context (the sentences include, before and after the target citation). The recommendations were ranked by cosine similarities.

Out-of-dataset Recommendation

We use the test papers' citations not included in the train set to run out-of-dataset recommendations and all the available citations (including in-dataset and out-of-dataset citations) to run the mixed tests. The mixed tests could better comply with the actual application scenario.

For **CB4R**, we concatenate the **new-paper vectors** and in-dataset **citation vectors** to form **combined citation embeddings**. Recommendations were made by ranking the cosine similarities between the **manuscript vector** and **combined citation embeddings**.

We adapted different approaches for computing the new-paper vectors using the baselines to yield the best results. For **W2V**: new-paper vectors are computed as averaged word vectors from the content words; for **D2V**: we use the inferred vectors from new-papers by the trained model as new-paper vectors; for **HD2V, DC2V** and **DACR**: no *infer*-like functions available, so we use averaged vectors from the content words of new-papers; for **SciBERT** and **Specter**: mean vectors of the last layer by encoding abstracts via the trained encoder. Recommendations were made by dot product for **W2V**, **D2V**, **HD2V**, **DC2V**, and **DACR**, or cosine similarities for **SciBERT** and **Specter**.

4.4.4 Analysis on Recommendation of in-dataset Papers

This subsection presents the recommendation performances for the in-dataset papers. We report Recall, mean average precision (MAP), mean reciprocal rank (MRR), and normalized discounted cumulative gain (nDCG) at top 10 candidates for evaluations. Five inferences can be made from the results presented in Table 4.3. First, CB4R outperformed all the baselines across all the datasets and metrics by a significant margin, which testifies the approach's effectiveness in the "on-the-fly" scenario. Second, the dynamic sampling strategy could improve CB4R compared with CB4R without dynamic sampling to improve the recall from 2% to 10%. Third, CB4R without dynamic sampling testified the effectiveness

Table 4.3: Results (@top 10 scores) of Citation Recommendations on in-dataset Papers (* $p < 0.05$ for paired t test against best baselines)

Model	DBLP-1				DBLP-2				DBLP-3				ACL			
	Recall	**MAP**	**MRR**	**nDCG**	**Recall**	**MAP**	**MRR**	**nDCG**	**Recall**	**MAP**	**MRR**	**nDCG**	**Recall**	**MAP**	**MRR**	**nDCG**
W2V	12.29	5.67	5.70	7.22	19.94	10.81	10.88	12.95	14.82	7.59	7.66	9.30	14.46	6.51	6.50	8.28
D2V	3.88	1.38	1.42	1.95	1.61	0.39	0.40	0.67	2.58	0.73	0.74	1.16	3.08	0.81	0.81	1.33
HD2V	18.78	5.55	5.56	8.65	30.4	8.99	8.98	14.02	26.51	13.43	13.49	16.49	24.77	11.26	11.24	14.42
DACR	12.31	5.36	5.38	6.96	23.81	12.86	12.82	15.40	17.57	8.87	8.92	10.90	18.85	8.38	8.38	10.82
SciBERT	6.97	3.47	3.47	4.30	6.11	2.98	2.98	3.71	6.12	2.80	2.80	3.58	6.69	3.39	3.39	4.16
Specter	3.60	1.84	1.90	2.24	3.98	2.01	2.03	2.45	5.34	2.55	2.58	3.19	0.72	0.33	0.33	0.42
CB4R (No Dynamic Sampling)	38.31	18.95	18.95	23.51	45.54	21.97	21.97	27.28	51.36	25.78	25.78	31.82	28.54	12.53	12.53	16.28
CB4R[a] (Dynamic Sampling)	**49.76***	**25.00***	**25.00***	**30.85***	**49.81***	**23.87***	**23.87***	**29.96***	**57.23***	**29.17***	**29.17***	**35.81***	**30.68***	**13.19***	**13.19***	**17.27***

[a]We run the tests on DBLP-1 three times to verify the consistency of dynamic context sampling. The results shown in the table are average values. The complete results are as follows: Recall: 49.81 / 49.97 / 49.49; MAP: 24.76 / 25.16 / 25.07; MRR: 24.76 / 25.16 / 25.07; nDCG: 30.68 / 31.03 / 30.84.

Table 4.4: "On-the-fly" Citation Recommendation for Manuscript at Different Stages of Completion

Manuscript at Different Stages of Completion [a]	**Recall@10**
Best Baseline (base context only)	18.78
CB4R without dynamic sampling (base only)	31.31
CB4R with Dynamic Sampling	
Finishing@3Sents (base only)	41.35
Finishing@10Sents (3base+7super)	42.27
Finishing@20Sents (3base+17super)	43.75
Finishing@30Sents(3base + 27super, default)	49.76

[a]tests were run on the model fine-tuned on DBLP-1 with 3 base + 27 superstructural context (default setting)

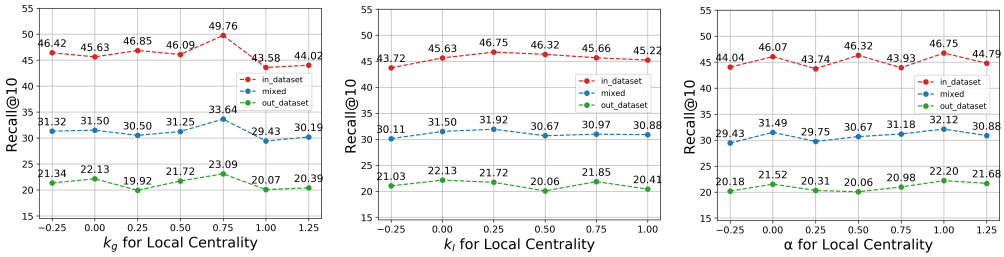
Table 4.5: Recall@10 for tests on the divergent test sets

Model / **Test Set**	**DBLP-1 Model**	**DBLP-2 Model**	**DBLP-3 Model**	**ACL Model**	**Best Baseline**
DBLP-1 Test Set	49.76	40.09	50.53	19.53	18.78
DBLP-2 Test Set	41.39	49.81	44.83	17.79	30.40
DBLP-3 Test Set	42.49	36.50	57.23	17.72	26.51
ACL Test Set	12.00	8.54	13.06	30.68	24.77

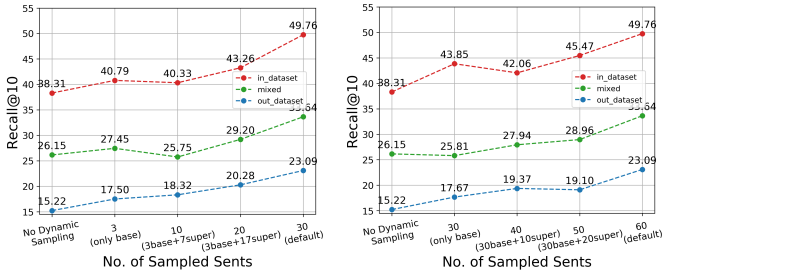
of the hierarchical transformer for producing paper-level embeddings compared to SciBRET and Specter based on the conventional sentence-level transformer. Third, owing to the different settings of the recommendation task of Specter [106], from which they focused on finding the ground-truth paper in a pre-selected 25 candidates; and Specter emphasized to use the negative candidates who are bibliographically coupled with the source paper; hence they might be constrained in our scenario in which the whole collection of in-dataset papers should be considered for optimizing the model. Fifth, the label-embedding-based approaches (HD2V and DACR) provided the best results among baselines.

"On-the-fly" Recommendation Testing

CB4R is tested according to different completion stages of the manuscript. We assumed that the manuscript has three completion stages: 1. The manuscript contains 3 sentences (only the base context); 2. The manuscript contains 10



(a) Test for k_g of Global Centrality (b) Recall for k_l of Local Centrality (c) Recall for α of Local Centrality



(d) Tests of superstructural Context in Manuscript Sampler (e) Tests of superstructural Context in Citation Sampler

Figure 4.3: Tests of k_g of Global Centrality (a), k_l and α (b) of Local Centrality, and Different Levels of superstructural Context in Manuscript Sampler (d), and Citation Sampler (e)

sentences in total (3-sentence base context and 7-sentence superstructural context); and 3. manuscript contains 20 sentences. We generate the incomplete manuscript by using the built-in Python3 *random* function to choose available sentences randomly. Then, we use the trained CB4R model with default settings to encode the generated manuscripts with the aforementioned amounts of context for tests. The results are presented in Table 4.4. Each test was conducted three times to report the average score for consistency. It was found that the model performed acceptably during the early development of drafts, which achieved 41.35% on recall@10 when the only base context is available, compared to 18.78% for the best baseline and 31.31% for CB4R without dynamic sampling. However, as the author is completing more content of the manuscript, the performances are improved.

Table 4.6: Tests on the New (Out-of-Dataset) Papers

(a) Recall@10 for testing out-of-dataset papers from the original datasets (mixed / new-paper results) (* $p < 0.05$)

Model	**DBLP-1**	**DBLP-2**	**DBLP-3**	**ACL**
W2V	7.89 / 10.88	14.69 / 5.74	11.76 / 7.30	10.59 / 3.89
D2V	2.06 / 2.08	1.01 / 1.15	2.22 / 2.94	2.85 / 4.10
HD2V	1.53 / 0.36	3.16 / 0.20	0.90 / 0.51	2.60 / 1.32
DC2V	2.10 / 4.25	2.62 / 6.00	2.84 / 6.14	2.47 / 8.29
DACR	1.03 / 2.08	0.28 / 3.87	2.11 / 4.64	0.63 / 2.12
SciBERT	10.49 / 14.70	6.68 / 9.78	7.30 / 11.85	8.84 / 16.29
Specter	2.25 / 1.34	3.10 / 1.47	4.17 / 1.41	0.35 / 0.50
CB4R(No Dynamic Sampling)	26.25 / 15.22	30.65 / 13.85	37.72 / 22.28	16.07 / 13.02
CB4R (Dynamic Sampling)	**33.64*** / **23.09***	**35.21*** / **19.68***	**45.17*** / **28.25***	**19.51*** / **16.35***

(b) Recall@10 for tests on the divergent new-paper sets (mixed / new-paper results)

Model / **New-paper Set**	**DBLP-1 Model**	**DBLP-2 Model**	**DBLP-3 Model**	**ACL Model**	**Best Baseline**
DBLP-1 New-paper Set	33.64 / 23.09	25.45 / 17.72	**34.74 / 27.20**	11.51 / 9.42	10.49 / 14.70
DBLP-2 New-paper Set	30.09 / **22.90**	**35.21** / 19.68	31.09 / 22.06	9.48 / 6.03	14.69 / 9.78
DBLP-3 New-paper Set	33.19 / 27.58	26.12 / 20.10	**45.17 / 38.25**	11.27 / 9.58	11.76 / 11.85
ACL New-paper Set	9.18 / 10.36	6.54 / 5.06	10.77 / 11.81	**19.51 / 16.35**	10.59 / 16.29

Testing on Divergent Test Sets

Different datasets might came with very different topics. To further evaluate CB4R models' ability on recommendations, we run cross tests of a trained model on a test set from different datasets. The results are presented in Table 4.5. First, the models performed poorly when the test sets were from different fields, when inspecting the results produced from the DBLP models on the ACL test set. Second, the ACL model performed poorer than the baselines on the DBLP test sets. Overall, the results tend to be correlated with the size of the training set; when training size is larger, the results are more effective on test sets from a similar field; in addition, the field knowledge also significantly affects the trained models' ability.

4.4.5 Analysis of the Recommendations of New (*Out-of-Data*) Papers

In this sub-section, the new-paper and "mixed-paper" recommendations are presented. Because the test set for new-papers can contain citations from either the new-paper set or the train set, we predicted the citations from the new-paper set to test new-paper recommendations and predict the citations from both the new-paper and train set for "mixed" recommendations. Table 4.6a presents the results on mixed and new-paper tests. Four points can be drawn from the scores of recall@10, as illustrated in Table 4.6a. First, CB4R outperformed all the baselines across all datasets, which confirmed its effectiveness on the recommendation of new-papers. The dynamic sampling had provided significant improvements. Second, the label-based embeddings (HD2V, DC2V, and DACR) performed the worst, which suggests that content semantics are helpful in the recommendation of new-papers. Third, W2V produced the best baselines scores due to the citation embeddings generated based on the complete content of the candidate papers. Fourth, although SciBERT and Specter preserved content semantics in their citation embeddings, they solely used the abstracts that may not fully contain the essential points of the papers, so they did not outperform W2V and CB4R.

Testing on Divergent New-Paper Sets

New-papers might come with very different knowledge and appearances compared to papers from different fields or published years ago. To further evaluate CB4R models' ability to recommend new-papers that are significantly different from the in-dataset papers, we run cross tests of a trained model on a new-paper set from different datasets. The results are presented in Table 4.6b. First, the models trained from the DBLP datasets outperformed the baselines on other DBLP new-paper sets. Second, all the models trained from the DBLP datasets outperformed the baselines tested on the original new-papers sets. Third, models did not perform well across the DBLP and ACL datasets. Overall, the models might be applicable across datasets if they are related to similar fields; however, they are constrained if they come with a high divergence in knowledge.

4.4.6 Tests on Dynamic Sampling

This section aims to present the tests on different settings of manuscript and citation sampler and the ablation tests.

Test on Dynamic Manuscript Sampler

The effects are tested from different levels of superstructural context in manuscript sampling. Different amounts of context, including 3 (only base context), 10 (3 base context + 7 superstructural context), and likewise for 20 sentences, were compared to the default setting of 30 sentences. In contrast to the tests presented in Table 4.4, we fine-tuned CB4R from scratch with the aforementioned settings on manuscript sampling. According to Figure 4.3d, it is found that the model provided significant improvements when the amount of superstructural context is beyond 17 sentences. Generally, more sampled sentences in the dynamic sampled lead to higher performances. Given the constraint of our GPU memories, 17 sentences of superstructural context provide the optimal solution.

Tests on Dynamic Citation Sampler

This subsection aims to the effectiveness of global and local centrality and the amount of sampled context in citation sampler. We run the tests on the DBLP-1 dataset.

Test on Global Centrality Tests are conducted on the global centrality in the citation sampler by adjusting the k_g values. k_g is the parameter for determining whether to sample sentences containing similar or dissimilar points with the abstracts as explained in Section 4.3.2. According to Figure 4.3a, the trend of the performances across different k_g values peaks at 0.75, which means that the sentences containing similar points to the abstract are down-weighted; whereas the sentences containing dissimilar points to the abstract are up-weighted. The performance drops to about 44% on recall which is closed to the performance when only the abstract is used, according to Figure 4.3e when k_g equals 1.00 and 1.25, since the algorithm mostly selected the sentences similar to the abstract. The performances fluctuate around 46% when k_g is lower than 0.75, where the sentences dissimilar to the abstract are gained higher chances to be selected, implying that a higher portion of dissimilar sentences is not helpful. Overall, the

abstract needs to be compensated for additional information; however, a small proportion of sentences dissimilar to the abstract are also needed.

Test on Local Centrality

Tests are additionally conducted on the sampler with both global and local centrality inspired by [109] in subsection. The local centrality aims to find the most "popular" sentences in the body content by computing the sum of the sentence-wise similarities. However, it is found that CB4R did not provide superior performances by combining the global and local centrality than solely adapting the global centrality. It might be because the local centrality did not consider the abstract. In summary, global centrality suits our abstract-based sampling strategy better and produces the best performances.

Local centrality aims to find the "popular" sentence among the body content. The body content of a paper is arranged as a graph, where all the sentences appearing before a target sentence are treated as in-links, and the sentences appearing after it are taken as the out-links. The in- and out-links are weighted differently but summed to 1. The weighted sum of all the link weights is defined as the local centrality of a sentence. The equation is illustrated as:

$$\begin{aligned} LocalCentrality(S_i) = \sum_{f=1}^{i-1} \sum_{b=i+1}^{|\mathcal{H}|} \alpha \cdot sim[\mathbf{E}(S_f), \mathbf{E}(S_i)] - \\ (1-\alpha) \cdot sim[\mathbf{E}(S_b), \mathbf{E}(S_i)], \end{aligned} \tag{4.5}$$

where S_f defines a front sentence appearing before S_i, and S_b indicates a sentence appearing behind S_i. The k_l parameter adjusts whether to pick the "popular" sentences for higher probabilities, or "unpopular" sentences for higher probabilities to supplement the abstract, similar to k_g in global centrality:

- when $k_l = 1$, then the probability of a sentence to be drawn is proportionally dependent on the proportion of its local centrality score weight accounted for the sum of all sentences;
- when $k_l = 0$, then the sampling is random;
- when $0 < k_l < 1$, then the probability for drawing the popular sentences (high local-centrality) in the body content are down-weighted, the sentences which are relatively unpopular are gained extra probabilities to be selected;

- when $k_l > 1$, then the popular sentences (high local centrality) are gained extra probabilities;
- when $k_l < 0$, then the unpopular sentences are gained extra probabilities.

The front and behind sentences contribute differently to the centrality of S_i [108], so we set hyper-parameter α to define the importance weight of the front sentences, and $(1 - \alpha)$ to be the weight of the sentences appearing behind.

- when $\alpha = 0.5$, then the former and behind sentences are assigned the same weight;
- when $\alpha < 0$, then the behind sentences are gained extra probabilities;
- when $\alpha > 1$, then the former sentences are gained extra probabilities;
- when $0 < \alpha < 1$, then the probabilities for drawing the former sentences are down-weighted.

First, the k_g value is fixed to be 0.75, and the α value to 0.50 (the fore) and adjust the k_l values to test k_l. According to Figure 4.3c and 4.3b, it is found that the scores produced by using both global and local centrality are lower than solely adapting global centrality. The best scores were produced when k_l is 0.50. Then, k_l and test the α values are fixed to test the effects of former and latter sentences. It is found that, the best score is produced when α is 1.0, when the probabilities are proportional to the summed similarity scores of the former sentences. However, adapting both local and global centrality generates lower performances than solely adapting the global centrality, hence results are mainly reported by adapting the global centrality in the main report.

Test on Superstructural Context in Citation Sampler

We tested the effect from different levels of superstructural context in the citation sampler. Different amounts of superstructural context, including 30 (only base context, i.e. abstract), 40 (30 base context + 10 superstructural context), and likewise for 50 sentences, were compared to the default setting of 60 sentences. We fine-tune CB4R with the aforementioned settings to compare with the default setting. The results are presented in Figure 4.3e. The trend shows that, as more superstructure context is sampled, the better the performances. Given the constraint of our GPU memories, 30 sentences of superstructural context provide the optimal solution.

4.5 Summary

This study proposed CBERT4REC, a *content-dependent* embedding model with a dynamic sampling strategy for "on-the-fly" citation recommendations, to assist researchers in writing their academic manuscripts and the reviewer to check the citations. CBERT4REC has the following advantages. First, it can extract citing intents from incomplete manuscripts from the extracted topic semantics and semantics preserved from the based context. Second, it can comprehensively extract the content semantics by leveraging the essential points from a paper to provide effective in-dataset and new-paper recommendations. Third, the neural network is constructed based on paper-level architecture.

CHAPTER 5

Citation Relation Mining

This chapter aims to leverage the information preserved in the citation networks to improve the performances in citation recommendations. It includes two tasks: first, it considers utilizing the structural contexts (previously existing citations in a manuscript) to enhance the performances; second, it proposes novel objective functions to retrieve co-citations more effectively.

This chapter is structured as follows: the motivations for designing the two tasks are presented in Section 5.1, the approaches are illustrated in Section 5.3 and 5.4. The experimental results and analyses are discussed in Section 5.5.

5.1 Motivation

Previous methods generally considered adoption of local contexts to infer citing intents, however, it is argued local contexts may not be effective in the on-the-fly scenario:

- First, when applying for the on-the-fly scenario, the inserted citations in the fleshed-out content could be adopted to explore the citing intents further;
- Second, the previous methods do not consider recommending frequent co-citations from the historical citation patterns;

To this end, we propose two novel methods to enhance the context-based approaches, namely DocCit2Vec and MP-BERT4CR. The former leverages structural contexts, i.e. the inserted citations in the fleshed-out content of the input manuscript, to define the citing intent of the users further and reduce the redundancy of recommendations. The latter approach proposed a novel objective function to retrieve co-citations more effectively. The underlying logic is: "if two papers are frequently co-cited in the past, and one of them is given, then another one should be recommended."

5.2 Related Work

The context-based approaches for citation recommendation, e.g. Word2Vec [3, 37], Doc2Vec [38], and HyperDoc2Vec [24] might be limited to be applied in on-the-fly scenario for citation recommendations, since they do not explicitly consider leveraging the previously existed citations from the user and recommending co-cited papers from the historical patterns.

The co-citation relationship is initially defined as two (or more) prior works are cited together by the later literature [116]. According to the qualitative analyses by [116], the majority of the co-cited papers come with direct citation relations and bibliographic coupling relations. Co-citations also demonstrated strong topic relatedness. [117] testified the strong strength between co-citations through human examiners. [30] conducted recommendation tasks for co-citation recommendations; however, the adapted datasets are relatively small (about 20,000 papers), and they did not comprehensively compare with best-performed baselines and tasks for non-co-citations. In this paper, we leverage the information of strong topic relevance carried by co-citation pairs pointed from the prior works to improve the recommendation performances by building noise distributions and optimizing via multi-positive objective functions.

5.3 Leveraging structural contexts via DocCit2Vec

DocCit2Vec involves two steps: embedding the content and fine-tuning the paper features. The first step aims to represent each paper and content word into the vector space to reflect their semantic meaning. We adapt the *pv-dm* [38] model to accomplish this task. It learns two vectors (IN and OUT) for each word,

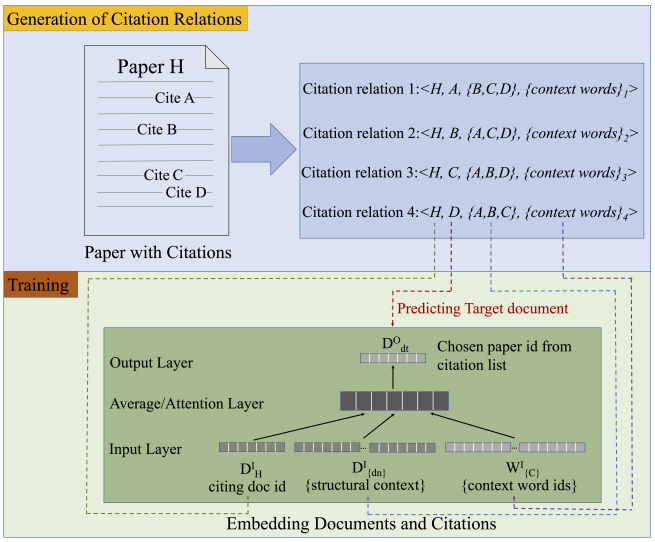


Figure 5.1: Overview of DocCit2Vec

denoted as $\mathbf{w^I}$ and $\mathbf{w^O}$, respectively, and an IN vector for each paper id. The paper and word vectors are trained by predicting the target words selected from a preset-sized context from the documents. The trained paper vectors aim to reflect what content words it includes. The trained word vectors preserve the context information.

The second step aims to fine-tune the learnt paper and word vectors with unique features of academic papers. DocCit2Vec conceptualizes the learning process as the prediction of a target citation such that an embedded document vector carries the information of a target citation. During training, the model generates a series of citation relations as shown in Figure 5.1. For each citation relation, one publication from the citation list of the paper H is picked as the target, and the surrounding context and structural context are used as known information to maximize the occurrence of the target citation by updating the parameters (i.e., the embedding vectors) of the neural network. The model learns two embedding vectors, namely an IN vector $\mathbf{d^O}$ and an OUT vector $\mathbf{d^O}$ for each document, where $\mathbf{d^I}$ characterizes the document as a citing paper and the OUT vector $\mathbf{d^O}$ encodes its role as a cited paper [24]. In addition, the model learns an IN vector $\mathbf{w^I}$ for each word.

To be more specific, we adopt the retrofitting technique as in [24], which initializes a predefined number of iterations based on the *pv-dm* model and then uses the learned vectors as the "base" vectors for training DocCit2Vec as the

fine-tuning process.

Two architecture of DocCit2Vec are proposed: DocCit2Vec-avg, which uses a conventional average hidden layer, and DocCit2Vec-att, which uses an attention hidden layer.

5.3.1 DocCit2Vec-avg: DocCit2Vec with an Average Hidden Layer

The architecture of DocCit2Vec-avg is founded on the *pv-dm* structure of Doc2Vec [38] and HyperDoc2Vec [24]. An overview of the model is shown in Figure 5.1. It involves an Input Layer to initialize an IN document matrix $\mathbf{D^I}$ and an IN word matrix $\mathbf{W^I}$, and an Output Layer to initialize an OUT document matrix $\mathbf{D^O}$. To optimize the embedding vectors, the model take a citation relation $\langle d_H, d_t, \{d_n | d_n \in D_n\}, \{w | w \in C\}\rangle$ as input, and averages over the corresponding IN vectors of d_H, $\{d_n | d_n \in D_n\}$, and $\{w | w \in C\}$. The output layer is computed using a multi-class softmax classifier, and the output value is regarded as the probability of occurrence of d_t.

To learn all the citation relations $\mathcal{C}$, the model is statistically expressed as

$$\max_{\mathbf{D^I},\mathbf{D^O},\mathbf{W^I}} \frac{1}{|\mathcal{C}|} \sum_{\langle d_H, d_t, D_n, C\rangle \in \mathcal{C}} \log P(d_t | d_H, D_n, C). \tag{5.1}$$

The hidden layer of the neural network is expressed as

$$\mathbf{x} = \frac{1}{1 + |D_n| + |C|} \Big(\mathbf{d}^{\mathbf{I}}_H + \sum_{d_n \in D_n} \mathbf{d}^{\mathbf{I}}_{d_n} + \sum_{w \in C} \mathbf{w}^{\mathbf{I}}_w \Big). \tag{5.2}$$

The output layer employs a multi-class softmax function, which is represented as

$$P(d_t | d_H, D_n, C) = \frac{\exp\left(\mathbf{x}^T \mathbf{d}^{\mathbf{O}}_{d_t}\right)}{\sum_{d \in D} \exp\left(\mathbf{x}^T \mathbf{d}^{\mathbf{O}}_{d}\right)}. \tag{5.3}$$

The negative sampling technique [37] is adopted to optimize the efficiency of the training procedure, and the objective function (Equation (5.3)) yields

$$\log \sigma\big(\mathbf{x}^T \mathbf{d}^{\mathbf{O}}_{d_t}\big) + \sum_{d_j \in \mathcal{D}_{\text{neg}}} \log \sigma\big(- \mathbf{x}^T \mathbf{d}^{\mathbf{O}}_{d_j}\big), \tag{5.4}$$

where $\mathbf{d}^{\mathbf{O}}_{d_t}$ represents the OUT embedding vector of the target document d_t and $\mathcal{D}_{neg} = \{d_j | j = 1, ..., n\}$ is the set of negative sampled documents sampled from the noise distribution $P_n(d)$.

The gradient descent optimizer is used to update the parameters. Each parameter is updated based on its gradient and a pre-set learning rate. The derived equation for the OUT vectors is:

$$\mathbf{d}_{d_j}^{\mathbf{O}\,(\text{new})} = \mathbf{d}_{d_j}^{\mathbf{O}\,(\text{old})} - \eta(\sigma(\mathbf{x}^T \mathbf{d}_{d_j}^{\mathbf{O}\,(old)} - t_{tj})) \tag{5.5}$$

where $\mathbf{d}_{d_j}^{\mathbf{O}\,(\text{new})}$ is the updated OUT embedding vector of a document d_j from the set $\{d_t\} \cup \mathcal{D}_{\text{neg}}$, $\mathbf{d}_{d_j}^{\mathbf{O}\,(\text{old})}$ is the embedding vector in the previous iteration, η is the learning rate, and the term t_{ij} equals one if d_j is the target document and zero otherwise. The derived equations for the IN vectors are as follows:

$$\mathbf{d}_{d_i}^{\mathbf{I}\,(\text{new})} = \mathbf{d}_{d_i}^{\mathbf{I}\,(\text{old})} - \frac{1}{|D_n| + |C| + 1} \cdot \eta \cdot \text{EH}, \tag{5.6}$$

$$\mathbf{w}_{w}^{\mathbf{I}\,(\text{new})} = \mathbf{w}_{w}^{\mathbf{I}\,(\text{old})} - \frac{1}{|D_n| + |C| + 1} \cdot \eta \cdot \text{EH}, \tag{5.7}$$

where $\mathbf{d}_{d_i}^{\mathbf{I}\,(\text{new})}$ and $\mathbf{d}_{d_i}^{\mathbf{I}\,(\text{old})}$ are the IN embedding vectors of a document d_i from the set $\{d_t\} \cup D_n$ after and before the update, and $\mathbf{w}_{w}^{\mathbf{I}\,(\text{new})}$ and $\mathbf{w}_{w}^{\mathbf{I}\,(\text{old})}$ are the IN embedding vectors of a word w from the set of contextual words $\{w|w \in C\}$ after and before the update, respectively. Further, EH is the back-propagated gradient, which is represented as

$$\text{EH} = \sum_{d_j \in \{d_t\} \cup \mathcal{D}_{\text{neg}}} (\sigma(\mathbf{x}^T \mathbf{d}_{d_j}^{\mathbf{O}} - t_{tj})) \cdot \mathbf{d}_{d_j}^{\mathbf{O}}. \tag{5.8}$$

5.3.2 DocCit2Vec-att: DocCit2Vec with an Attention Hidden Layer

The architecture of DocCit2Vec-att adopts the same architecture as that of DocCit2Vec-avg as shown on the right-hand side of Figure 5.1, except that the average hidden layer is replaced by an attention layer, inspired by [76]. In addition to the original parameters of DocCit2Vec-avg, a weight vector $\mathbf{K} \in \mathbb{R}^{1 \times (|D|+|W|)}$ is introduced in the attention layer, where each value denotes the importance of a word or document. The model is statistically expressed asThe gradient descent optimizer to update the parameters. Each parameter is updated based on its gradient and a pre-set learning rate. The derived equation for the OUT vectors is:

$$\max_{\mathbf{D^I},\mathbf{D^O},\mathbf{W^I},\mathbf{K}} \frac{1}{|\mathcal{C}|} \sum_{\langle d_H,d_t,D_n,C\rangle\in\mathcal{C}} \log P(d_t|d_H,D_n,C) \tag{5.9}$$

In contrast to the average hidden layer, the attention layer computes a weighted sum of an individual word and a document by multiplying the vector and its weight ratio, which is expressed as follows:

$$\mathbf{x} = \mathbf{d}^{\mathbf{I}}_H \cdot a_H + \sum_{d_n\in D_n} \mathbf{d}^{\mathbf{I}}_{d_n} \cdot a_{d_n} + \sum_{w\in C} \mathbf{w}^{\mathbf{I}} \cdot a_w. \tag{5.10}$$

The terms a_H, a_{d_n}, and a_w are the associated weight ratios for the documents d_H and d_n and the word w. The weight ratios are computed using the matrix $\mathbf{K}$ as follows:

$$a_i = \frac{\exp k_i}{\sum_{k\in\{d_H\}\cup D_n\cup C} \exp k_k}. \tag{5.11}$$

Equation (5.4) is kept the same as the objective function. The update equation is also identical to Equation (5.5) for the OUT vectors. The derived equations for the IN vectors are as follows:

$$\mathbf{d}^{\mathbf{I}\,(\text{new})}_{d_i} = \mathbf{d}^{\mathbf{I}\,(\text{old})}_{d_i} - a_{d_i} \cdot \eta \cdot \text{EH}, \tag{5.12}$$

$$\mathbf{w}^{\mathbf{I}\,(\text{new})}_{w} = \mathbf{w}^{\mathbf{I}\,(\text{old})}_{w} - a_{w} \cdot \eta \cdot \text{EH}, \tag{5.13}$$

where a_i or a_w is the weight ratio for the input document $d_i \in \{d_t\} \cup D_n$ or input word $w \in C$. The gradient term EH is the same as that in Equation (5.8). The update equation for a weight k_j in the attention layer, where $k \in \{d_H\} \cup D_n \cup C$, is derived as

$$k_k^{(\text{new})} = k_k^{(\text{old})} - \eta \cdot \text{EH} \cdot \text{EA}, \tag{5.14}$$

where EA is the gradient of the objective function with respect to the weights of the attention layer:

$$\text{EA} = \sum_{i\in\{d_H\}\cup D_n\cup C} a_i \cdot (t_{ik} - a_k) \cdot \mathbf{e}^{\mathbf{I}}_i, \tag{5.15}$$

where t_{ik} equals one if k is i and zero otherwise. Further, $\mathbf{e}^{\mathbf{I}}_i$ is the embedding IN vector of a document $\mathbf{d}^{\mathbf{I}}_i$ or word $\mathbf{w}^{\mathbf{I}}_i$.

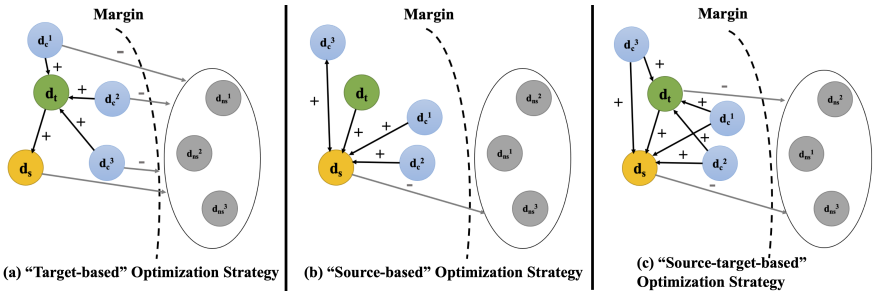


Figure 5.2: Illustration of Optimization Strategies

5.4 Multi-positive optimization for retrieving co-citations

Based on the CBERT4REC model proposed in Section 4.3, we propose a series of optimization strategies to retrieve co-citations more effectively.

For the query context with multiple ground-truth citations, in addition to the target citation, we sample a pre-defined number of citations according to their historical co-citation frequencies with the target citation as the additional positive samples. Noise distribution is built based on the historical frequencies of co-citation pairs, so both high frequent and low-frequent pairs are assigned a probability to be drawn. High frequent pairs are assigned a relatively higher probability since it is likely for a pair to be co-cited again if they are frequently cited before. The probability for low-frequent pairs can be adjusted by the power value, which is set to $\frac{3}{4}$ for a default value in Equation 5.16.

Specially, given $\mathcal{H}_{t*}$ as the target ground-truth citation for prediction, and $\mathbb{H}_p$ for the full list of co-citations, the algorithm samples n number (a pre-defined value) of positive citations from the paper collection $\mathbb{H}_d$ via the noise probability distribution [3]:

$$\mathbf{p}(\mathcal{H}_{ti} \in \mathbb{H}_p) = \frac{\mathbf{frequency}(\mathcal{H}_{\mathbf{t*}}, \mathcal{H}_{\mathbf{ti}})^{\frac{3}{4}}}{\sum_{\mathcal{H}_{tj} \in \mathbb{H}_d} \mathbf{frequency}(\mathcal{H}_{\mathbf{t*}}, \mathcal{H}_{\mathbf{tj}})^{\frac{3}{4}}}, \tag{5.16}$$

where $frequency$ denotes the count of the two input papers being appeared as co-citations from the dataset. The number of positive samples is set to be 3, as it is found that the number of co-citations with greater than 3 items are neglectable small in our datasets.

A negative sampling strategy is adapted to pick the papers which are not cited by the input context as the targets for similarity minimization. The objective is that irrelevant papers should not appear in the top recommendation list.

The negative citations are sampled based on their occurrences as citations. The underlying intuition is that if other papers frequently cite a paper, however, it is not cited by the input context, then it would be drawn more frequently for similarity minimization between it and the input context. Similar to positive sampling, a noise distribution is constructed based on their occurrences as citations:

$$\mathbf{p}(\mathcal{H}_i \in \mathbb{H}) = \frac{\mathbf{count}(\mathcal{H}_\mathbf{i})^{\frac{3}{4}}}{\sum_{\mathcal{H}_j \in \mathbb{H}} \mathbf{count}(\mathcal{H}_\mathbf{j})^{\frac{3}{4}}}, \tag{5.17}$$

where $\mathbb{H}$ denotes all the papers in the dataset except the positive citations; and *count* denotes the number of occurrences as citations of a paper. A pre-defined m number of papers are picked from the distribution as negative samples, noted as $\mathbb{H}_{ns}$. We set m to be 4 in this study.

Multi-Positive Triplet Objectives are designed to optimize the algorithm for recommending multiple positive citations.

Suppose a piece of context in the manuscript,$\mathcal{H}_s$ containing the ground-truth citation, $\mathcal{H}_t$, along with N number of co-citations $\mathcal{H}_t$, i.e., $\mathcal{H}_c^1, ..., \mathcal{H}_c^N$. The algorithm retrieves n positive samples $\{\mathcal{H}_c^n | 1 \leq n \leq N\}$ from Equation 5.16, and m negative samples $\{\mathcal{H}_{ns}^m | 1 \leq m\}$ from Equation 5.17.

We propose multiple positive objectives considering the multiple positive samplings by modifying the original triplet-encoder [48, 118] which originally considers one target and one negative sample:

$$\max(||\mathbf{d}_s - \mathbf{d}_t|| - ||\mathbf{d}_s - \mathbf{d}_{ns}|| + \varepsilon, 0), \tag{5.18}$$

where $||.||$ denotes euclidean distance, and ε is the margin which is normally set to 1. The original triplet loss implies that the embedding of the manuscript is only guaranteed to be geometrically closed to one target citation. When applying for recommending multiple positive citations, this training objective might be limited.

Hence, the three multi-positive triplet objectives are proposed so that the embedding of the manuscript $\mathbf{d}_s$ should not only be geometrically closed to one target citation $\mathbf{d}_t$, but also be closed to the other positive citations $\{\mathbf{d}_c^n | 1 \leq c \leq n\}$.

Meanwhile, the manuscript embedding should be distanced to the m number of negative embeddings $\{\mathbf{d}_{ns}^{m}|1 \leq m\}$.

Three strategies were designed to achieve the objective, which is illustrated in Figure 5.2:

- **"Target-based" optimization strategy**: Minimize the distance between $\mathbf{d}_s$ and $\mathbf{d}_t$, and the distances between $\mathbf{d}_t$ and $\{\mathbf{d}_c^n|1 \leq c \leq n\}$. Hence, when $\mathbf{d}_t$ is found to be similar, embeddings $\{\mathbf{d}_c^n|1 \leq c \leq n\}$ could also be recommended.

- **"Source-based" optimization strategy**: Minimize the distance between $\mathbf{d}_s$ and $\mathbf{d}_t$, and the distances between $\mathbf{d}_s$ and $\{\mathbf{d}_c^n|1 \leq c \leq n\}$, so that given $\mathbf{d}_s$, the both of the target and positive citations could be retrieved.

- **"Source-target-based" optimization strategy**: Combining **"target-based"** and **"source-based** strategies, it is proposed to minimize the distances between $\mathbf{d}_s$, and both of the target and positive embeddings, and the distances between the target and positive embeddings.

Based on the N-tuplet loss function [119], we propose three designs of multi-positive triplet objectives following the aforementioned strategies:

- Multi-positive target-based triplet (***mpt-tgt***):

$$\mathcal{L}_{mpt-tgt} = \log(1 + \sum_{i=1}^{n}\sum_{j=1}^{m} \exp(||\mathbf{d}_s - \mathbf{d}_t|| - ||\mathbf{d}_s - \mathbf{d}_{ns}^{j}||) + \exp(||\mathbf{d}_c^i - \mathbf{d}_t|| - ||\mathbf{d}_c^i - \mathbf{d}_{ns}^{j}||)) \quad (5.19)$$

- Multi-positive manuscript-based triplet (***mpt-src***):

$$\mathcal{L}_{mpt-ms} = \log(1 + \sum_{j=1}^{m} \exp(||\mathbf{d}_s - \mathbf{d}_t|| - ||\mathbf{d}_s - \mathbf{d}_{ns}^{j}||) + \sum_{j=1}^{m}\sum_{i=1}^{n} \exp(||\mathbf{d}_s - \mathbf{d}_c^i|| - ||\mathbf{d}_s^i - \mathbf{d}_{ns}^{j}||)) \quad (5.20)$$

- Multi-positive source-target-based triplet (***mpt-src-tgt***):

$$\mathcal{L}_{mpt-src-tgt} = \log(1 + \sum_{j=1}^{m} \exp(||\mathbf{d}_s - \mathbf{d}_t|| - ||\mathbf{d}_s - \mathbf{d}_{ns}^j||) + \sum_{j=1}^{m}\sum_{i=1}^{n} \exp(||\mathbf{d}_s - \mathbf{d}_c^i|| - ||\mathbf{d}_s - \mathbf{d}_{ns}^j||) + \sum_{j=1}^{m}\sum_{i=1}^{n} \exp(||\mathbf{d}_t - \mathbf{d}_c^i|| - ||\mathbf{d}_t - \mathbf{d}_{ns}^j||)) \quad (5.21)$$

We set the maximum number of positive samples to be 3, since the number of co-citations with greater than 3 pairs is neglectable small from our datasets. **Mpt-src-tgt** is set to be the default objective for experiments since it is testified to be the most effective objective according to Section 5.5.2.

5.5 Experiments

This section presents the experimental results for DocCit2Vec when adapting the structural contexts for citation recommendations and MP-BERT4CR for retrieving co-citations.

5.5.1 Tests by adapting structural contexts via DocCit2Vec

We designed two categories of experiments to validate the model: citation recommendation tasks to verify the recommendation ability and classification tasks to explore the model's strength further.

The experiments for citation recommendation were conducted on two datasets, namely DBLP and ACL Anthology [24] as illustrated in Section 3.5.1. Our model was compared with two categories of baseline approaches, i.e., document-based methods (Word2Vec, Doc2Vec, and HyperDoc2Vec) and network-based methods (DeepWalk, LINE, and Node2Vec). In addition, we adopted a hybrid method that combines DocCit2Vec with a network embedding method to determine whether the combined information can provide supplementary performance. We defined the recommendation tasks as a ranking problem in which a query extracted from a document is converted into vectors based on the learned embedding models, and we then ranked the document vectors by taking the dot product with the converted query vector (the details are presented in Section 5.3).

Two classification-based experiments were conducted: 1) a topic classification experiment and 2) a classification experiment on the functionality of the citations. The first experiment classifies the research field of a given document according to the Cora dataset [1]. Document embedding vectors trained from the DBLP dataset are employed and coupled with a support vector machine (SVM) classifier to classify the topics. This experiment aims to verify the capability of document vectors from the DocCit2Vec model in classification tasks. The second experiment uses word vectors to classify the purpose of citing a paper based on the dataset from [120]. The word vectors are learned from the DBLP dataset and coupled with an SVM classifier. The Cora dataset for the topic classification experiment includes 5,975 papers. Each paper is labelled with the research field it belongs to ("Artificial Intelligence", "Information Retrieval", "Networking", etc.), i.e., a total of 10 classes. The citation functionality dataset [120] contains 2,824 citation contexts, and each context is annotated with a function, e.g., comparing methods (CoCoGM), comparing results (CoCoR0), and neural description of the cited work (Neut).

Implementation and Settings

Three baseline models based on document embedding, namely Word2Vec, Doc2Vec, and HyperDoc2Vec, were implemented using the Gensim package [87], which also served as the foundation for developing DocCit2Vec. We adopted the same hyper-parameter settings as in [24]. For Word2Vec, the embedding size was set to 100 with the *cbow* structure, and the default Gensim settings were followed. For Doc2Vec, the same embedding was adopted with the *pv-dbow* structure, and the rest of the settings were the default ones. For HyperDoc2Vec, the same settings as those in [24] were adopted: embedding size, 100; window size, 50; iterations, 100 and 1000 negative samplings; initialization of Doc2Vec, 5 epochs. DocCit2Vec used the same settings as HyperDoc2Vec.

The network-based methods were implemented with the code from the authors' GitHub repositories [2]. The embedding sizes were set to 100 and the rest of the settings were the default ones.

[1] https://people.cs.umass.edu/~mccallum/data.html

[2] DeepWalk: https://github.com/phanein/deepwalk

LINE: https://github.com/tangjianpku/LINE

Node2Vec: https://github.com/aditya-grover/node2vec

The models were implemented on a Linux server (12-core Intel Xeon E5-1650 CPU and 128 GB memory) installed with Anaconda 5.2.0 and Gensim 2.3.0.

Methodology for Citation Recommendations

For the document-based baseline methods (Word2Vec, Doc2Vec, and Hyper-Doc2Vec), as Word2Vec and Doc2Vec do not explicitly model the citations, we adopted the "citation as word" [40] and "context as content" [24] approaches for the recommendation tasks.

In the "citation as word" approach [40], the citations are treated simply as words. This approach was adopted for Word2Vec (denoted as W2V in Table 5.1 and Table 5.2) and Doc2Vec (denoted as D2V-nc in Table 5.1 and Table 5.2) algorithms. For "citation as word" via Word2Vec, the average vector of the contextual words, i.e., 50 words before and after the citation, are used as a query to rank the IN word embedding vectors of all the IN vectors of "citation words" by the dot product. For Doc2Vec (D2V-nc), we used the learned model to infer a vector based on the input IN vectors initially and then rank the IN vectors of the documents by cosine similarity, as we found that cosine similarity provides higher scores than the dot product.

In the "context as content" approach [24], the citations are removed, and the context words surrounding a cited document are copied into the cited document as the supplementary content. This approach is conducted based on Doc2Vec (denoted as Doc2Vec-cac in Table 5.1 and Table 5.2); the embedded vectors of the "augmented documents" are treated as citation vectors. Recommendations are made by ranking the IN document vectors with cosine similarities to the inferred vectors by the IN vectors of an input context.

The models of HyperDoc2Vec [24] and DocCit2Vec explicitly embed citations using the OUT matrix; hence, we use the query vector to rank the OUT vector of the documents by the dot product. For HyperDoc2Vec, the query vector is the average IN vector of the context words. For DocCit2Vec, the query vector is the average IN vector of the context words and the structural contexts.

For more comprehensive comparisons, we employed three other baseline models based on network embedding: DeepWalk [13], LINE [14], and Node2Vec [12].

A directed citation graph is initially built, in which every node represents a paper, and an edge expresses a citation relation. If paper A cites paper B, then the edge is directed from A to B. The document embedding vectors are learned

Table 5.1: Results of citation recommendation for DocCit2Vec on DBLP dataset

Model		Recall@10	Recall@15	Recall@20	MAP@10	MAP@15	MAP@20	MRR@10	MRR@15	MRR@20	nDCG@10	nDCG@15	nDCG@20
Network-based*	DW (case 1)	3.57	4.39	5.04	1.35	1.41	1.44	1.35	1.4	1.44	2.16	2.38	2.54
	DW (case 2)	3.37	4.18	4.68	1.82	1.88	1.94	N/A	N/A	N/A	2.87	3.09	3.31
Document-based	W2V (case 1)	20.47	24.09	26.34	10.54	11.01	11.13	10.54	11.01	11.13	14.71	15.89	16.43
	W2V (case 2)	20.46	23.51	25.76	10.55	10.78	10.91	10.55	10.78	10.91	14.71	15.54	16.08
	W2V (case 3)	20.15	23.12	25.35	10.40	10.63	10.76	10.40	10.63	10.76	14.49	15.30	15.84
	D2V-nc (case 1)	7.90	9.88	11.53	3.17	3.32	3.41	3.17	3.32	3.41	4.96	5.49	5.90
	D2V-nc (case 2)	7.90	9.89	11.52	3.17	3.32	3.41	3.17	3.32	3.41	4.96	5.50	5.90
	D2V-nc (case 3)	7.91	9.90	11.53	3.17	3.32	3.41	3.17	3.32	3.41	4.97	5.51	5.89
	D2V-cac (case 1)	7.91	9.88	11.52	3.17	3.33	3.42	3.17	3.33	3.42	4.97	5.52	5.90
	D2V-cac (case 2)	7.90	9.88	11.52	3.17	3.33	3.42	3.17	3.33	3.42	4.97	5.51	5.91
	D2V-cac (case 3)	7.89	9.90	11.53	3.17	3.33	3.42	3.17	3.33	3.42	4.97	5.51	5.90
	HD2V (case 1)	28.41	31.52	33.36	14.20	14.43	14.53	14.20	14.43	14.53	20.37	21.22	21.67
	HD2V (case 2)	28.42	31.56	33.42	14.20	14.44	14.55	14.20	14.44	14.55	20.38	21.23	21.68
	HD2V (case 3)	28.41	31.58	33.45	14.20	14.45	14.55	14.20	14.45	14.55	20.37	21.24	21.70
	DC2V-att (case 1)	7.38	9.20	10.72	3.36	3.50	3.59	3.36	3.50	3.59	4.89	5.38	5.75
	DC2V-att (case 2)	6.05	6.06	7.86	2.74	2.87	2.93	2.74	2.87	2.93	3.99	4.42	4.69
	DC2V-att (case 3)	5.20	6.52	7.58	2.36	2.46	2.52	2.36	2.46	2.52	3.43	3.78	4.04
	DC2V-avg (case 1)	**44.23**	**49.45**	**52.76**	**21.80**	**22.21**	**22.40**	**21.80**	**22.21**	**22.40**	**31.34**	**32.77**	**33.57**
	DC2V-avg (case 2)	**40.31**	**45.41**	**48.76**	**20.16**	**20.57**	**20.75**	**20.16**	**20.57**	**20.75**	**28.69**	**30.09**	**30.88**
	DC2V-avg (case 3)	**40.37**	**42.56**	**45.91**	**19.02**	**19.41**	**19.60**	**19.02**	**19.41**	**19.60**	**26.84**	**28.19**	**29.00**
Combined	DC2V-avg+DW (case 1)	44.24	49.41	52.76	21.81	22.22	22.41	21.81	22.22	22.41	31.34	32.76	33.56
	DC2V-avg+DW (case 2)	40.37	45.43	48.77	20.18	20.57	20.77	20.18	20.57	20.77	28.72	30.09	30.90

* Node2Vec and LINE produced scores of 0.00 across all metrics for both case 1 and case 2.

Table 5.2: Results of citation recommendation for DocCit2Vec on ACL dataset

Model		Recall@10	Recall@15	Recall@20	MAP@10	MAP@15	MAP@20	MRR@10	MRR@15	MRR@20	nDCG@10	nDCG@15	nDCG@20
Network-based	DW (case 1)	0.05	0.07	0.09	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.04	0.04
	DW (case 2)	0.03	0.05	0.06	0.02	0.03	0.03	N/A	N/A	N/A	0.04	0.05	0.06
	LINE (case 1)	0.07	0.15	0.17	0.04	0.04	0.05	0.04	0.04	0.05	0.04	0.07	0.07
	LINE (case 2)	0.05	0.13	0.15	0.05	0.04	0.05	N/A	N/A	N/A	0.06	0.10	0.13
	N2V (case 1)	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	N2V (case 2)	0.01	0.03	0.04	0.03	0.05	0.03	N/A	N/A	N/A	0.03	0.06	0.03
Document-based	W2V (case 1)	27.25	31.24	34.10	13.74	14.05	14.21	13.74	14.05	14.21	19.51	20.59	21.28
	W2V (case 2)	26.54	30.41	33.15	13.55	13.75	13.89	13.55	13.75	13.89	19.19	20.08	20.73
	W2V (case 3)	26.06	29.71	32.48	13.21	13.50	13.65	13.21	13.50	13.65	18.66	19.65	20.32
	D2V-nc (case 1)	19.92	24.14	27.46	9.06	9.39	9.57	9.24	9.39	9.57	13.39	14.55	15.34
	D2V-nc (case 2)	19.89	24.15	27.41	9.06	9.39	9.58	9.06	9.39	9.58	13.38	14.53	15.32
	D2V-nc (case 3)	19.89	24.13	27.41	9.07	9.39	9.58	9.07	9.39	9.58	13.38	14.54	15.34
	D2V-cac (case 1)	20.51	24.75	27.93	9.24	9.56	9.75	9.24	9.56	9.75	13.68	14.84	15.60
	D2V-cac (case 2)	20.29	24.74	27.94	9.17	9.56	9.75	9.06	9.56	9.75	13.58	14.83	15.61
	D2V-cac (case 3)	20.51	24.76	27.93	9.24	9.56	9.75	9.24	9.56	9.75	13.69	14.84	15.60
	HD2V (case 1)	**37.53**	**42.09**	**45.23**	19.64	20.32	20.49	19.64	20.32	20.49	27.20	28.95	29.70
	HD2V (case 2)	**36.83**	**41.43**	**44.62**	**19.62**	**19.86**	**19.99**	**19.62**	**19.86**	**19.99**	**27.18**	**28.45**	**29.21**
	HD2V (case 3)	**36.24**	**40.85**	**44.11**	**19.32**	**19.69**	**19.87**	**19.32**	**19.69**	**19.87**	**26.79**	**28.05**	**28.83**
	DC2V-att (case 1)	27.48	32.26	35.57	13.42	13.79	13.98	13.42	13.79	13.98	19.24	20.55	21.34
	DC2V-att (case 2)	25.05	29.15	32.37	12.34	12.70	12.84	12.37	12.70	12.84	17.63	18.77	19.49
	DC2V-att (case 3)	25.01	26.86	29.76	11.48	11.78	11.94	11.48	11.78	11.94	16.27	17.31	18.00
	DC2V-avg (case 1)	36.89	41.22	44.21	**20.44**	**20.78**	**20.95**	**20.44**	**20.78**	**20.95**	**27.72**	**28.90**	**29.62**
	DC2V-avg (case 2)	33.71	37.88	40.74	18.47	18.79	18.96	18.40	18.79	18.96	25.17	26.35	27.04
	DC2V-avg (case 3)	31.14	35.17	38.07	16.97	17.28	17.45	16.97	17.28	17.45	23.20	24.30	25.00
Combined	DC2V+DW (case 1)	36.82	41.25	44.24	20.45	20.80	20.97	20.45	20.80	20.97	27.70	28.91	29.63
	DC2V+DW (case 2)	33.67	38.11	40.71	18.46	18.81	18.95	18.46	18.81	18.95	25.21	26.37	27.02
	DC2V+LINE (case 1)	36.91	41.21	44.35	20.44	20.78	20.95	20.44	20.78	20.95	27.72	28.89	29.62
	DC2V+LINE (case 2)	33.74	37.85	40.75	18.47	18.77	18.94	18.47	18.77	18.94	25.19	26.32	27.01

from the citation graph through the aforementioned methods.

As the network methods do not model the word information, we adopt the "seed papers" approach from [11], which takes a collection of "known to be relevant references" as input and thus extend the list. We consider the case 1 and case 2 scenarios described in Section 3.5.1, in which the structural contexts are used as the "seed papers". Recommendations are made by ranking the node embedding vectors based on the average vector of the seed papers by the dot product.

To compensate for the lack of word information in network-based methods, we consider combining document and network embedding. For each document, we compute the sum vector learned from DocCit2Vec and DeepWalk (denoted as "DC2V-avg+DW" in Table 5.1 and Table 5.2), or from DocCit2Vec and LINE (denoted as "DC2V+LINE" in Table 5.2). We rank the sum vectors based on the average vector of the context words and structural contexts by the dot product.

Methodology for Topic and Functionality Classifications

We adopted three approaches for the topic classification experiment. The first approach uses a concatenation of IN and OUT vectors of the documents ("IN+OUT" in Table 5.3), while the second approach uses only the IN vectors of the documents. In the third approach, the IN or IN+OUT vectors are concatenated with the network embedding vectors from DeepWalk, LINE, and Node2Vec. An SVM classifier with 5-fold cross-validation was employed. For the citation functionality experiment classification, we used the average IN vector of the context words and an SVM classifier with 10-fold cross-validation.

Analyses on recommendation results

We reported the scores for four metrics under the three scenarios introduced in Section 3.5.1, namely recall, mean average precision (MAP), mean reciprocal rank, and normalized discounted cumulative gain (nDCG), for the top 10, 15, and 20 results. The results for the DBLP and ACL datasets are summarized in Table 5.1 and Table 5.2, respectively.

Five observations are made based on the results.

1. The scores of the network-embedding-based methods are significantly lower than those of the document-based methods. The network methods make recommendations based solely on the network structures, which indicates

that a lack of consideration of the word information would lead to inefficient recommendation performance. Among the network-embedding-based methods, DeepWalk is the only model that produced non-zero scores for the DBLP dataset (630,909 nodes and 2,874,303 links), which confirms the finding reported in [121] that DeepWalk can capture the semantic similarities between nodes more effectively than Node2Vec and LINE. However, LINE outperformed DeepWalk for the smaller but denser graph, i.e., the ACL dataset (14,654 nodes and 79,932 links), because, as indicated by [121], LINE preserves the properties of denser graphs more effectively than DeepWalk.

2. The combined methods did not demonstrate significant improvements compared to the purely document-based methods. The differences are negligible. It is inferred that the two types pf information preserved by document embedding and network embedding cannot be effectively combined by addition.

3. DocCit2Vec-avg demonstrated superior performance on the larger dataset, i.e., DBLP, with significant improvement among the document-based methods. The recall scores at different levels were higher by approximately 12% to 20% according to the different cases, compared to the second-best model HyperDoc2Vec, with 4% to 7% improvement for MAP and 6% to 12% improvement for nDCG. Second, all the models exhibited better performance on the medium-sized dataset, i.e., ACL Anthology, except for DocCit2Vec-avg, which implies that the loss of DocCit2Vec-avg requires a larger volume of data to converge. HyperDoc2Vec yielded the best results for this dataset, followed by DocCit2Vec-avg with close scores.

4. All the baseline models yielded similar scores across the three cases, except for DocCit2Vec-avg. It was observed that DocCit2Vec-avg constantly yielded the best scores for the first case, where all the structural contexts were included, and the second-best scores for the second case, where the structural contexts were randomly picked. This indicates that the information on the structural contexts is embedded into the embedding vectors.

5. The performance of DocCit2Vec-att was among the lowest, suggesting that citation recommendation is not a suitable task for this model, as the attention mechanism emphasizes the "differences" between embedding vectors rather than the "similarities".

Table 5.3: Results of classification experiments for DocCit2Vec

Model		Topic Classification								Classification of Functionality	
		Original		with DeepWalk		with LINE		with Node2Vec		F1-micro	F1-macro
		F1-micro	F1-macro	F1-micro	F1-macro	F1-micro	F1-macro	F1-mico	F1-macro		
Network based	DeepWalk	79.53	72.39	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	LINE	17.03	8.49	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Node2Vec	22.82	6.58	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Document based	W2V (IN+OUT)	58.11	37.00	74.62	63.66	56.83	33.18	56.76	33.02	N/A	N/A
	W2V (IN)	57.83	36.90	76.61	66.94	55.75	32.05	55.61	31.49	77.36	56.52
	D2V-nc	78.70	70.91	82.67	76.16	66.05	76.66	76.37	66.16	63.51	6.47
	D2V-cac	78.99	71.11	82.41	75.95	83.47	77.43	76.61	66.12	63.51	6.47
	HD2V (IN+OUT)	80.26	73.72	82.38	76.11	80.08	73.32	80.21	73.35	N/A	N/A
	HD2V (IN)	79.12	72.78	82.38	76.30	80.23	73.68	80.47	73.78	75.63	54.33
	DC2V-att (IN+OUT)	82.70	76.86	84.84	79.50	82.21	76.47	82.54	76.74	N/A	N/A
	DC2V-att (IN)	**83.80**	**78.43**	**85.49**	**80.59**	**83.44**	**77.97**	**83.91**	**78.38**	74.46	54.43
	DC2V-avg (IN+OUT)	77.56	70.43	79.23	72.65	77.47	69.39	78.16	71.43	N/A	N/A
	DC2V-avg (IN)	75.28	68.18	77.28	70.36	76.23	69.00	77.17	70.32	**75.91**	**54.67**

Analyses on classification tasks

The two experiments, namely topic classification [3] and classification of the functionality of citations [4], aim to test the performance of document and word vectors separately; the first experiment employs document vectors while the second experiment employs words vectors. The dataset for topic classification includes 5,975 academic papers and 10 unique fields to which they belong. The dataset for the classification of the functionality of citations includes 2,824 citation contexts, each with a classified functionality, such as "PBas: Cited work used as a basis or starting point,', which represent 12 unique classes [120].

The F1-micro and F1-macro scores were reported for evaluation. For F1-micro, the scores of recall and precision were summed up category by category and Equation (5.22) was then employed with the summed recall and precision to get the F1-micro score [122]. F1-macro is the average value of the F1 scores (Equation (5.22)) computed for each category [122]. F1-micro evaluates the classification performance, whereas F1-macro verifies the distribution of scores over categories. If the distribution over categories is completely balanced, F1-micro and F1-macro should be the same [122].

$$f1 = 2 * \frac{recall * precision}{recall + precision} \tag{5.22}$$

The DocCit2Vec-att model and the concatenation of DocCit2Vec-att with

[3]Cora dataset, https://people.cs.umass.edu/~mccallum/data.html.

[4]Dataset from [120].

network embedding models exhibited the best performance in topic classification (Table 5.3), with F1-macro and F1-micro scores approximately 3% to 5% higher compared to the best baseline model, i.e., HyperDoc2Vec. The concatenation of DocCit2Vec-att and DeepWalk with the IN vector is the best approach for this task.

For the classification of citation functionality, neither DocCit2Vec-att nor DocCit2Vec-avg is ranked at the top. First, the results indicate that DocCit2Ve-att improves the document embedding vectors' classification abilities but not the word embedding vectors. Second, DocCit2Vec-avg exhibits lower performance because it takes multiple documents as input and emphasizes the "similarity" between documents. In summary, the attention mechanism focuses on "difference", whereas the average layer approach emphasizes "similarity".

The F1-macro scores were lower than the F1-micro scores in both the experiments, implying a relatively unbalanced distribution over categories, i.e., the sparse categories have higher scores than the dense categories.

5.5.2 Tests for recommendation via MP-BERT4CR

Dataset

Four datasets including the ACL three datasets generated from the DBLP corpus were adapted for experiments. The ACL Anthology corpus includes 20,405 papers with 108,729 citations, whereas the DBLP corpus contains 649,114 papers with 2,874,303 citations. Three datasets were produced from the DBLP corpus, i.e., DBLP-1, DBLP-2, and DBLP-3, including 50,000 papers. A "biased-individuality" dataset generating strategy was adapted to produce the three DBLP datasets, by which DBLP-2 and DBLP-3 shares 20% papers (10,000) in-common to evaluate the stability on the performance of the model; whereas the DBLP-1 dataset contains completely different papers to DBLP-1 and DBLP-2. The complete ACL corpus was adapted for the fourth dataset. The datasets were divided into a train set for pre-training and fine-tuning the model and a test set with 20% of the total amount randomly selected conducting the recommendation tests. The statistics of the datasets are presented in Table 5.4. ParsCit [84] is applied to parse the in-text citations so that the algorithms can recognize them. ParseLabel [85] was applied to recognize the abstracts, as well as other section headers. The rare words from the vocabulary were compressed to reduce memory consumption by applying byte

Table 5.4: Statistics of Datasets for MP-BERT4CR

	Paper No. (Total / Train / Test)	Train Cit No.	Test Cit for P≥1	Test Cit for P=1	Test Cit for P≥2
DBLP-1[a]	50,000 / 40,000 / 10,000	74,153	21,688	20,537	1,151
DBLP-2	50,000 / 40,000 / 10,000	128,380	26,300	24,643	1,657
DBLP-3	50,000 / 40,000 / 10,000	103,467	28,382	27,066	1,316
ACL	20,406 / 16,325 / 4,081	31,017	21,420	15,783	188

[a] DBLP-1 and DBLP-2 shares 20% papers in common to test the consistency of the performances.

Table 5.5: Parameters of MP-BERT4CR

Encoder / Decoder Params	**Block No. (L)** 6	**Hidden Size (H)** 768	**Attention (A)** 12
Optimizer Params	**Update Schedule** Inverse Square Root	**Warmup Updates** 1,000 (ACL) / 2,000 (DBLP)	**Update Frequency** 4 (pretrain) / 8 (finetune)
	Warmup Learning Rate 1e-7 (pretrain) / 1e-9 (finetune)	**Learning Rate** 1e-4 (pretrain) / 2e-5 (finetune)	**Weight Decay** 0.01
	$\boldsymbol{\beta_1}$ 0.9	$\boldsymbol{\beta_2}$ 0.999 (pretrain) / 0.98 (finetune)	**Dropout** 0.1

pair encoding [115] with the adaption of the learned vocabulary table from [110].

Implementation Details

MP-BERT4CR was developed based on Fairseq 0.4.0 [111], Gensim 2.3.0 [87], and Pytorch 1.6.0 [86]. For the baseline models, Word2Vec and Doc2Vec were implemented using Gensim 2.3.0; HyperDoc2Vec was developed based on Gensim 2.3.0; DACR was developed based on Pytorch 1.6.0 and Gensim 2.3.0; SciBERT and Specter were implemented using Hugginface 4.2.0 [112].

Adam optimizer [113] was adapted to optimize MP-BERT4CR, and the complete parameters are provided in Table 5.5. We immediately began from the in-domain pre-training stage on our datasets and then fine-tuned them. Due to the GPU memory constraints, we limit the maximum words in a sentence to be 50 words and the maximum length for a text to be 30 sentences. The batch sizes are set to be 7 for pre-training and 1 for fine-tuning. We run 10 iterations of pre-training and 2 iterations of fine-tuning on DBLP-1, DBLP-2, and DBLP-3 datasets; or 50 iterations of pre-training and 5 iterations of fine-tuning on the ACL dataset since the smaller dataset required larger iterations of training for the loss function to converge. The number of negative samples is set to 4, and the maximum number of positive samples is set to 3.

Table 5.6: Recommendation Scores for Single and Multiple Positive Citations (* $p < 0.05$, ** $p < 0.01$ for paired t test against best baseline scores)

Datasets	**DBLP-1**						**DBLP-2**						**DBLP-3**						**ACL**					
No. of Pst. Cits	positive = 1		positive ≥ 1		positive ≥ 2		positive = 1		positive ≥ 1		positive ≥ 2		positive = 1		positive ≥ 1		positive ≥ 2		positive = 1		positive ≥ 1		positive ≥ 2	
Metrics	Recall	MAP	Recall	MAP	Recall	MAP	Recall	MAP	Recall	MAP	Recall	MAP	Recall	MAP	Recall	MAP	Recall	MAP	Recall	MAP	Recall	MAP	Recall	MAP
W2V	13.26	6.29	13.26	6.29	13.48	7.10	19.25	10.25	19.35	10.38	20.81	12.41	17.76	9.09	17.76	9.09	17.77	9.15	17.16	8.2	17.13	8.25	14.37	10.37
D2V	2.07	0.48	2.07	0.48	2.17	0.68	1.27	0.29	1.27	0.30	1.26	0.38	2.44	0.78	2.45	0.79	2.48	0.92	2.22	5.73	2.23	5.75	2.69	0.74
HD2V	34.15	17.50	34.15	17.50	41.01	22.49	40.12	21.56	40.53	21.86	46.59	26.22	40.64	21.64	40.67	21.64	41.35	21.55	30.91	16.40	30.83	16.34	23.63	11.97
DACR	22.42	10.34	22.42	10.34	25.06	11.94	27.42	14.07	27.79	14.36	33.21	18.51	27.15	13.48	27.14	13.45	26.89	12.92	24.31	12.21	24.25	12.19	19.49	10.93
SciBERT	8.95	4.03	8.95	4.03	10.23	5.23	7.79	3.40	7.87	3.46	9.06	4.39	10.42	4.64	10.51	4.70	12.45	6.01	0.16	0.03	0.16	0.03	0.21	0.06
Specter	4.51	2.26	4.51	2.26	5.42	3.72	3.42	1.67	3.45	1.74	3.89	2.71	4.51	2.26	4.51	2.26	5.42	3.72	3.42	1.67	3.45	1.74	3.89	2.71
MB4R$_{no\_dynamic}$	38.97	18.98	39.42	19.33	47.29	25.48	39.51	18.71	40.00	19.03	47.21	23.67	43.35	21.69	43.77	21.96	52.30	27.54	30.47	14.31	30.50	14.34	32.78	17.09
MB4R$_{triplet}$	40.07	18.08	40.34	18.33	45.04	22.79	40.44	19.51	40.79	20.01	48.24	27.73	47.52	22.61	48.05	22.98	59.02	30.24	30.82	13.23	30.92	13.29	39.55	18.30
MB4R$_{mpt-src-tgt}$[a]	**44.81****	**21.45***	**45.21****	**21.77***	**52.49****	**27.50***	**47.25****	**23.41***	**47.66****	**23.76***	**53.69****	**29.10***	**48.86****	**23.42***	**49.38****	**23.73***	**60.05****	**30.60***	**36.13****	**16.90***	**36.18****	**16.93***	**40.61****	**19.89***

[a] To confirm the consistency of the *dynamic* sampling strategy, the tests for DBLP-1 were conducted three times to output the averaged scores. The complete results are: 44.81 / 44.81 / 44.81 & 21.69 21.69 21.69 for p=1 cases; 45.21 / 45.21 / 45.21 & 21.77 / 21.77 / 21.77 for p≥1 cases; 52.49 / 52.49 / 52.49 & 27.50 27.50 27.50 for ≥2 cases.

Table 5.7: Comparison on Multi-Positive Triplet Objectives with Conventional Triplet on DBLP-1

No. of Positive Cits	**positive = 1**		**positive ≥ 1**		**positive ≥ 2**	
Metrics	**Recall**	**MAP**	**Recall**	**MAP**	**Recall**	**MAP**
Best Baseline	34.15	17.50	34.15	17.50	41.01	22.49
MB4R$_{triplet}$	40.07	18.08	40.34	18.33	45.04	22.79
MB4R$_{Mpt-src}$	41.86	19.99	42.23	20.22	48.82	24.26
MB4R$_{Mpt-tgt}$	**44.85**	**21.69**	45.10	**21.94**	49.68	26.53
MB4R$_{Mpt-src-tgt}$**(default)**	44.81	21.45	**45.21**	**2**1.77	**52.49**	**27.50**

Recommendation Methodology

We present the methods for generating recommendations for MP-BERT4CR (MB4R thereafter) and the baselines. Recall and MAP at top 10 recommendations were reported for analyses.

For MB4R, we firstly conduct dynamic sampling for the manuscripts (papers in the test set) and citations (papers in the train set) in our test dataset. The sampled manuscript and citation contexts are then encoded via the fine-tuned manuscript encoder or citation encoder to get **query vectors** and **citation vectors**. Recommendations were selected as the top 10 citations by computing the cosine similarities between the **query vector** and **citation vectors**. Recommended candidates are compared with the ground-truth citations (one or multiple) for reporting the scores of Recall and MAP.

Analyses on retrieving single positive citation

Two points could be drawn from Table 5.6 (*positive* = 1 cases). First, MB4R outperformed all the baseline models across all the datasets and metrics by significant margins, from which MB4R$_{triplet}$'s superiority compared with baselines testified the effectiveness of the proposed neural architecture, and MB4R$_{mpt-src-tgt}$ further testified the effectiveness of the proposed multi-positive objectives. Second, the multi-positive objective function does not only help to identify multiple ground-truth citations. The single positive performances are also improved when comparing MB4R$_{mpt-src-tgt}$ to MB4R$_{triplet}$. In addition, owing to the more hierarchical transformer and dynamic sampling strategy, MB4R outperformed Sci-BERT

and Specter.

Analyses on retrieving multiple positive citations

According to the scores with multiple ground-truth citations (cases of $positive \geq 1$ and $positive \geq 2$) in Table 5.6, the scores for multiple positive recommendations from MB4R$_{mpt-src-tgt}$ remained effective comparing with the baselines and MB4R$_{triplet}$. However, as the $positive \geq 2$ test samples are much less than the test samples for $positive \geq 1$, so the ability of *mpt* might not be fully demonstrated, especially on DBLP-3 and ACL datasets. We will produce datasets with a higher number of positive samples in the later stage for further tests.

Comparisons on Different Designs of Multi-Positive Objectives

We compare the three designs of multi-positive objectives in Table 5.7. First, all the three proposed multi-positive objectives (MB4R$_{mpt-src}$, MB4R$_{mpt-tgt}$, and MB4R$_{mpt_src-tgt}$) outperformed the best baseline (HD2V), and the original triplet objective (MB4R$_{triplet}$), since the original triplet objective did not consider multiple targets. Second, the two target-based strategies MB4R$_{mpt-tgt}$ and MB4R$_{mpt_src-tgt}$ are superior to the source-based strategy (MB4R$_{mpt-src}$), which means that the distances between the ground-truth candidate and other positive candidates play the central role for multiple retrieving; in addition, when the distances of the co-citations are larger than the negative citations, co-citations are hard to be retrieved. Third, MB4R$_{mpt_src-tgt}$ produced superior performances comparing with MB4R$_{mpt-tgt}$ for case $positive \geq 2$, but very close performances for cases $positive = 1$ and $positive \geq 1$, since the co-citations might be further than the target citation to the source paper from the "target-based" strategy if they come from the same direction as the target citation. Hence, the "source-target-based" objective was set as the default. We also tested MB4R with or without dynamic sampling for an ablation analysis. It is found that the dynamic sampling mechanism can significantly improve the performances by comparing the results from $MB4R_{mpt-src-tgt}$ and $MB4R_{no\_dynamic}$.

Analysis on Retrieving Full List of Co-citations

This subsection computes the proportion of fully retrieved co-citations in the top 10 results from the recommendation results. The objective is to test how many

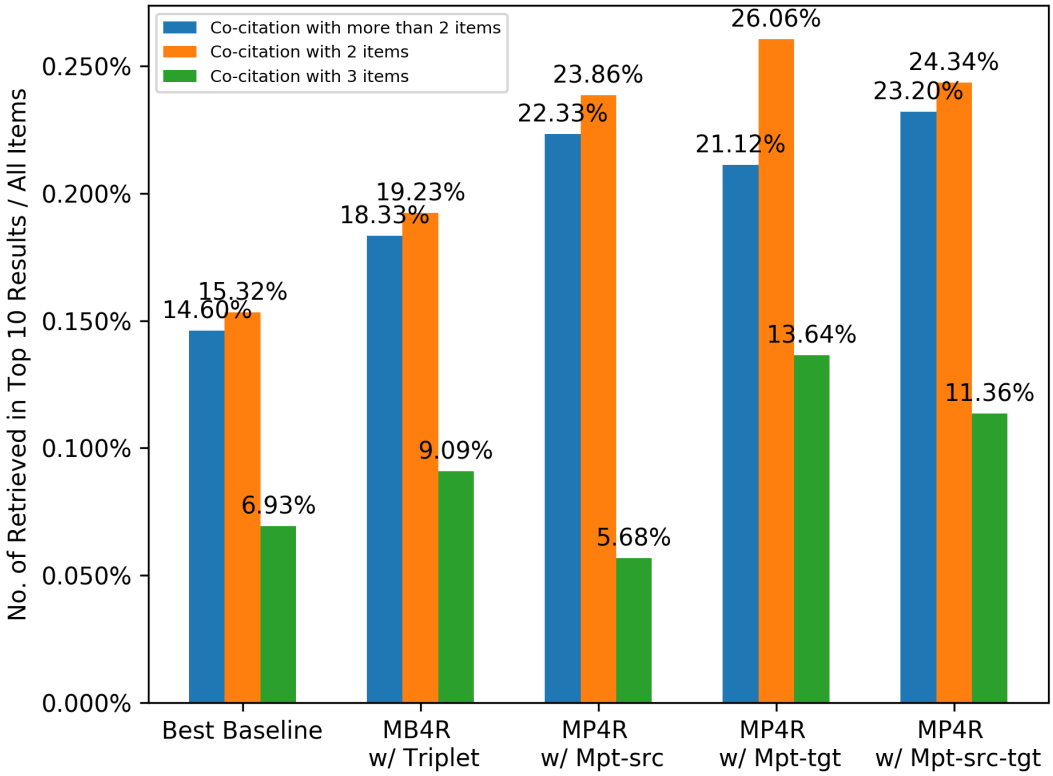


Figure 5.3: Proportion of fully retrieved co-citations in top 10 results on DBLP-1

chances the user can find all the co-citations from our algorithm's top 10 searching results. We report the scores on DBLP-1 from the four MB4R models, as well as the best baseline model (HD2V) in Figure 5.3. Four points could be drawn from the figure. First, our proposed four MB4R models generally outperformed the best baseline model, except that MB4R$_{mpt-src}$ comes with lower scores on citation with 3 items by a small margin. Second, comparing the four triplet-based objectives, *mpt-tgt* and *mpt-src-tgt* have outperformed the original triplet for all kinds of co-citation pairs; whereas *mpt-src* produced superior performances on co-citations with more than 2 items and exactly 2 item, however a lower performance on co-citation with exactly 3 co-citations. Fourth, the scores for co-citations with more than 2 items and with exactly 3 items are lower than the scores for co-citations with exactly 2 items for all models, which implies that when the number of co-citation pairs increases, the accuracy decreases. In future work, we further improve the accuracy for retrieving co-citations with more than

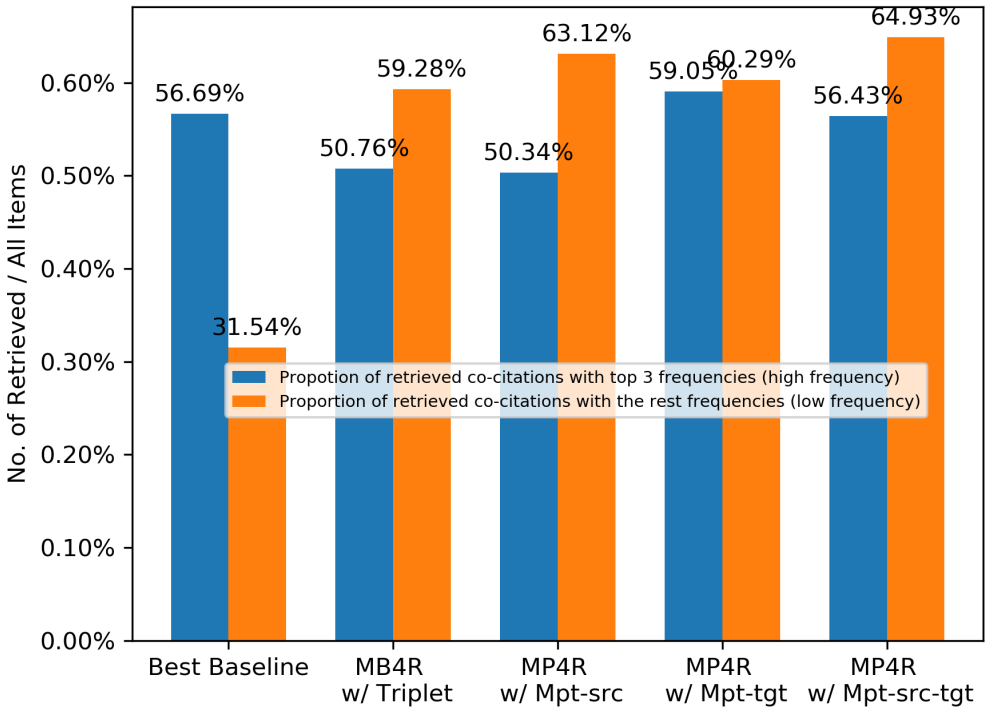


Figure 5.4: Proportion of retrieved top 3 most frequent co-citations in history vs. proportion of the rest retrieved on DBLP-1

2 items.

Analysis the Historical Frequencies of Retrieved Co-citations

One of the objectives of this study is to utilize the historical co-citation occurrences to improve the chances of retrieving the rest of the co-citations when one of them is given. In this subsection, we analyze the historical frequencies of the successfully retrieved co-citations from our model on DBLP-1. We report the proportion of the top 3 most frequently co-cited items in top 10 results when *positive ¿ 1*, against the proportion of the rest of the retrieved co-citations, in Figure 5.4. Three points could be drawn. First, MB4R models are advantageous in retrieving co-citations with lower historical frequencies when comparing the orange bars. Second, the top 3 frequent co-citations are retrieved in a similar proportion from *mpt-tgt* and *mpt-src-tgt* objectives comparing with the baseline; whereas the original

triplet produced lower score for retrieving the top 3 frequent co-citations. Third, the *mpt-src-tgt* performed the best on retrieving the low-frequent co-citations. In summary, *mpt-tgt* and *mpt-src-tgt* are especially superior on retrieving low-frequent co-citations in history, and the effectiveness of retrieving the high-frequent ones remain the same level as the best baseline. We will further improve the performances on high frequent co-citations in future work.

5.6 Summary

Regarding the adaption of citation, relations to enhance the recommendation performances. Two approaches are proposed.

The DocCit2Vec model is proposed considering the "structural context" to improve the recommendation performance for authors of academic papers. Two model implementations were proposed: the first involves an average hidden layer (DocCit2Vec-avg), while the second involves an attention hidden layer (DocCit2Vec-att). We conducted experiments to compare the different categories of approaches for citation recommendation, i.e. network-based, document-based, and combined approaches. The network-based approaches exhibited the poorest performance, as they lack consideration of the word information. DocCit2Vec-avg exhibited superior performance among the document-based methods in citation recommendation tasks compared to the baseline methods and DocCit2Vec-att. The combined methods provided indistinguishable scores for DocCit2Vec-avg. Furthermore, DocCit2Vec-att exhibited effective performance in the classification task using document embedding vectors.

MP-BERT4CR is then proposed for recommending multi-positive citation recommendations, i.e. co-citations. It has the following advantages: first, it comes with a series of multiple positive objectives to optimize the model for multi-positive recommendation; second, it leverages the historical co-citation information so that both the historically high and low frequent co-citation pairs can be effectively found, and the performances on retrieving the full list of co-citations are improved. Third, it uses a dynamic context sampling strategy to extract the citing intent in a macro-scope, empowering the citation embeddings to carry content semantics.

CHAPTER 6

Conclusion and Future Work

6.1 Conclusion

Considering the rapidly increasing number of academic papers published, searching for appropriate references and citing them accurately has become a non-trivial research task, especially for new researchers. Thereby, this dissertation proposes an "On-the-fly" approach for citation recommendations to support researchers in efficiently finding useful citations while writing, and can potentially help both authors and reviewers in checking the completeness of a draft's citations before publication, which could potentially deliver a time and energy saving solution to the users. In this dissertation, three main modules are planned and implemented to accomplish the on-the-fly scenario for citation recommendations, including a module of **source representation**, **target representation**, and **citation relation mining**.

source representation focuses on representing the citing intents of users from the input manuscript into semantic space. The algorithm should be able to detect the core citing intents from the query contexts and adaptively detect the topic semantics from the continuous updates of the drafts. This module focuses on two tasks: 1. extracting the core citing intent by considering deeper semantics from the query context, i.e. the word-wise relatedness, importance and sectional purposes; and 2. extracting the topics semantics from continuous updates of the

incomplete manuscript via manuscript dynamic sampling. Experiments have been implemented to verify the effectiveness of our proposed approaches' effectiveness against the previous methods dependent on leveraging local contexts for extracting citing intents.

Target Representation: target representation is designed to represent the content semantics of the candidate papers. Using current content-dependent models, we construct a "universal content modelling" and comply with a dynamic content sampling strategy designed to sample essential sentences from papers regarding the topic. The constructed content modelling can be adapted for representing and recommending both in-dataset and out-of-dataset papers (newly published papers).

Citation Relationship Mining: we leverage the information mined from citation networks, such as co-citation relations, co-citation frequencies, and structural context, to improve the recommendation performances.

The proposed methods are verified through experiments simulating real-world applications. For example, three completing stages with different amounts of finished contents for input manuscripts are adapted to test the on-the-fly recommendations. In addition, extensive user tests and explainability studies are implemented to verify the usability and rationality of the approaches. The proposed models are also analyzed in the ablation tests to testify each model component. Overall, the experiments could verify the framework's effectiveness and rationality from the perspective of accuracy, rationality, and usability.

6.2 Future Work

This dissertation has proposed a feasible approach for on-the-fly citation recommendations. To realize the application, it is planned to continue the work in the following perspectives.

Technical Progression: from the technical perspective, the current approaches are generally based on the embedding algorithms by matching the content semantics. Firstly, I will focus on innovating the algorithms to provide better accuracy by leveraging mode advanced neural networks, such as GPT3. Second, I will also consider adapting different recommendation concepts to comprehensively make the recommendations, such as constructing the user profiles and combining them with the network or collaborative filtering methods.

Usability Perspective: from the application standpoint, the proposed approaches have to be optimized for running speed and resource-saving purposes. Currently, the algorithms have to be running on servers with GPU facilities with costly time, and source spending. I will consider first compressing the model with a significantly smaller size to run on personal devices in future work. The code will also be re-constructed for speed optimization. An application will be planned for development for testing. I will also leverage the feedback collected from the prototype testing to come up with further improvements, such as popularity-based recommendations, and personalized literature pushing.

Expanding Applications: on-the-fly recommendation could also be applicable for other types of documents, such as patents, news, and financial reports for wiring support, referencing, and advising. I will also focus on the recommendations of cross-lingual literature, such as recommending papers written in Japanese, Chinese, German and other languages, by training the models with datasets in different languages.

Acknowledgements

I would like to send my special thanks to my supervisors, Associate Professor Qiang Ma, Professor Keishi Tajima, and Professor Shinsuke Mori, for their advice and support to conduct this doctoral dissertation and the associated research tasks.

I would like to thank all the teachers at Yoshikawa and Ma Laboratory, Professor Masatoshi Yoshikawa, Associate Professor Yang Cao, and Associate Professor Kazunari Sugiyama, for raising questions and providing valuable comments at the seminar meetings.

I would also like to thank all the Lab members for sharing ideas. Thanks to Mr Junjie Sun, Mr Chengyang Ye in our group, and Mr Jiexin Wang and Shuyuang Zheng from the Yoshikawa group.

Special Thanks to Dr. Zhu for the ultimate tolerances and the kind guidance.

Yang Zhang, February 2022

References

[1] Madian Khabsa and C. Lee Giles. The number of scholarly documents on the public web. *PLOS ONE*, 9(5):1–6, 05 2014.

[2] Rob Johnson, Anthony Watkinson, and Michael Mabe. The stm report: An overview of scientific and scholarly publishing. *International Association of Scientific, Technical and Medical Publishers*, pages 1–214, 2018.

[3] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[4] Michael L. NEWMAN and John SACK. Information workflow of academic researchers in the evolving information environment: an interview study. *Learned Publishing*, 26(2):123–131, 2013.

[5] John N. Parker, Stefano Allesina, and Christopher J. Lortie. Characterizing a scientific elite (B): publication and citation patterns of the most highly cited scientists in environmental science and ecology. *Scientometrics*, 94(2):469–480, 2013.

[6] Per O. Seglen. The skewness of science. *J. Am. Soc. Inf. Sci.*, 43(9):628–638, 1992.

[7] Jeppe Nicolaisen and Birger Hjørland. Practical potentials of bradford's law: a critical examination of the received view. *Journal of Documentation*, 63(3):359–377, 2007.

[8] Eugene Garfield. The History and Meaning of the Journal Impact Factor. *JAMA*, 295(1):90–93, 01 2006.

[9] Marco Gori and Augusto Pucci. Research paper recommender systems: A random-walk based approach. In *WI*, pages 778–781, 2006.

[10] Onur Küçüktunç, Erik Saule, Kamer Kaya, and Ümit V. Çatalyürek. Towards a personalized, scalable, and exploratory academic recommendation service. In *ASONAM*, pages 636–641, 2013.

[11] Haofeng Jia and Erik Saule. An analysis of citation recommender systems: Beyond the obvious. In *ASONAM*, pages 216–223, 2017.

[12] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 855–864. ACM, 2016.

[13] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: online learning of social representations. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 701–710. ACM, 2014.

[14] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. LINE: large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, pages 1067–1077. ACM, 2015.

[15] Saurabh Kataria, Prasenjit Mitra, and Sumit Bhatia. Utilizing context in generative bayesian models for linked corpus. In Maria Fox and David Poole, editors, *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*. AAAI Press, 2010.

[16] Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and C. Lee Giles. Context-aware citation recommendation. In *WWW*, pages 421–430, 2010.

[17] Jing He, Jian-Yun Nie, Yang Lu, and Wayne Xin Zhao. Position-aligned translation model for citation recommendation. In Liliana Calderón-Benavides, Cristina N. González-Caro, Edgar Chávez, and Nivio Ziviani,

editors, *String Processing and Information Retrieval - 19th International Symposium, SPIRE 2012, Cartagena de Indias, Colombia, October 21-25, 2012. Proceedings*, volume 7608 of *Lecture Notes in Computer Science*, pages 251–263. Springer, 2012.

[18] Wenyi Huang, Saurabh Kataria, Cornelia Caragea, Prasenjit Mitra, C. Lee Giles, and Lior Rokach. Recommending citations: translating papers into references. In Xue-wen Chen, Guy Lebanon, Haixun Wang, and Mohammed J. Zaki, editors, *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, pages 1910–1914. ACM, 2012.

[19] Claudio Cioffi-Revilla. *Introduction to Computational Social Science - Principles and Applications, Second Edition*. Texts in Computer Science. Springer, 2017.

[20] Sean M. McNee, István Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K. Lam, Al Mamunur Rashid, Joseph A. Konstan, and John Riedl. On the recommending of citations for research papers. In *CSCW*, pages 116–125, 2002.

[21] Cornelia Caragea, Adrian Silvescu, Prasenjit Mitra, and C. Lee Giles. Can't see the forest for the trees?: a citation recommendation system. In *JCDL*, pages 111–114, 2013.

[22] Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and C. Lee Giles. Context-aware citation recommendation. In Michael Rappa, Paul Jones, Juliana Freire, and Soumen Chakrabarti, editors, *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pages 421–430. ACM, 2010.

[23] Wenyi Huang, Zhaohui Wu, Chen Liang, Prasenjit Mitra, and C. Lee Giles. A neural probabilistic model for context based citation recommendation. In Blai Bonet and Sven Koenig, editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2404–2410. AAAI Press, 2015.

[24] Jialong Han, Yan Song, Wayne Xin Zhao, Shuming Shi, and Haisong Zhang. hyperdoc2vec: Distributed representations of hypertext documents. In *ACL*, pages 2384–2394, 2018.

[25] Yang Zhang and Qiang Ma. Citation recommendations considering content and structural context embedding. In Wookey Lee, Luonan Chen, Yang-Sae Moon, Julien Bourgeois, Mehdi Bennis, Yu-Feng Li, Young-Guk Ha, Hyuk-Yoon Kwon, and Alfredo Cuzzocrea, editors, *2020 IEEE International Conference on Big Data and Smart Computing, BigComp 2020, Busan, Korea (South), February 19-22, 2020*, pages 1–7. IEEE, 2020.

[26] Yang Zhang and Qiang Ma. Doccit2vec: Citation recommendation via embedding of content and structural contexts. *IEEE Access*, 8:115865–115875, 2020.

[27] Yang Zhang and Qiang Ma. Dual attention model for citation recommendation. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3179–3189. International Committee on Computational Linguistics, 2020.

[28] Tao Dai, Li Zhu, Yifan Wang, Hongfei Zhang, Xiaoyan Cai, and Yu Zheng. Joint model feature regression and topic learning for global citation recommendation. *IEEE Access*, 7:1706–1720, 2019.

[29] Libin Yang, Zeqing Zhang, Xiaoyan Cai, and Lantian Guo. Citation recommendation as edge prediction in heterogeneous bibliographic network: A network representation approach. *IEEE Access*, 7:23232–23239, 2019.

[30] Yuta Kobayashi, Masashi Shimbo, and Yuji Matsumoto. Citation recommendation using distributed representation of discourse facets in scientific articles. In Jiangping Chen, Marcos André Gonçalves, Jeff M. Allen, Edward A. Fox, Min-Yen Kan, and Vivien Petras, editors, *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL 2018, Fort Worth, TX, USA, June 03-07, 2018*, pages 243–251. ACM, 2018.

[31] Yang Zhang and Qiang Ma. Recommending multiple positive citations for manuscript via content-dependent modeling and multi-positive triplet. In *WI*, 2021.

[32] Eghbal Ghazizadeh and Pengxiang Zhu. A systematic literature review of natural language processing: Current state, challenges and risks. In *Proceedings of the Future Technologies Conference (FTC) 2020, Volume 1*, pages 634–647, Cham, 2021. Springer International Publishing.

[33] Erik Cambria and Bebo White. Jumping NLP curves: A review of natural language processing research [review article]. *IEEE Comput. Intell. Mag.*, 9(2):48–57, 2014.

[34] Karlo Babic, Sanda Martincic-Ipsic, and Ana Mestrovic. Survey of neural text representation models. *Inf.*, 11(11):511, 2020.

[35] Gerard Salton and Michael McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, 1984.

[36] Stephen E. Robertson and Steve Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In W. Bruce Croft and C. J. van Rijsbergen, editors, *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum)*, pages 232–241. ACM/Springer, 1994.

[37] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[38] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, pages 1188–1196, 2014.

[39] Travis Ebesu and Yi Fang. Neural citation network for context-aware citation recommendation. In Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White, editors, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 1093–1096. ACM, 2017.

[40] Matthew Berger, Katherine McDonough, and Lee M. Seversky. cite2vec: Citation-driven document exploration via word embeddings. *IEEE Trans. Vis. Comput. Graph.*, 23(1):691–700, 2017.

[41] Yaning Liu, Rui Yan, and Hongfei Yan. Guess what you will cite: Personalized citation recommendation based on users' preference. In *Information Retrieval Technology - 9th Asia Information Retrieval Societies Conference, AIRS 2013, Singapore, December 9-11, 2013. Proceedings*, volume 8281 of *Lecture Notes in Computer Science*, pages 428–439. Springer, 2013.

[42] Avishay Livne, Vivek Gokuladas, Jaime Teevan, Susan T. Dumais, and Eytan Adar. Citesight: supporting contextual citation recommendation using differential search. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, pages 807–816. ACM, 2014.

[43] Qi He, Daniel Kifer, Jian Pei, Prasenjit Mitra, and C. Lee Giles. Citation recommendation without author supervision. In Irwin King, Wolfgang Nejdl, and Hang Li, editors, *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011*, pages 755–764. ACM, 2011.

[44] Daniel Duma and Ewan Klein. Citation resolution: A method for evaluating context-based citation recommendation systems. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 358–363. The Association for Computer Linguistics, 2014.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[46] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[47] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[48] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics, 2019.

[49] Francesco Ricci, Lior Rokach, and Bracha Shapira. Introduction to recommender systems handbook. In Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, *Recommender Systems Handbook*, pages 1–35. Springer, 2011.

[50] Tariq Mahmood and Francesco Ricci. Improving recommender systems with adaptive conversational strategies. In Ciro Cattuto, Giancarlo Ruffo, and Filippo Menczer, editors, *HYPERTEXT 2009, Proceedings of the 20th ACM Conference on Hypertext and Hypermedia, Torino, Italy, June 29 - July 1, 2009*, pages 73–82. ACM, 2009.

[51] Paul Resnick and Hal R. Varian. Recommender systems - introduction to the special section. *Commun. ACM*, 40(3):56–58, 1997.

[52] Marlesson R. O. Santana and Anderson Soares. Hybrid model with time modeling for sequential recommender systems. In Catalin-Mihai Barbu, Ludovik Coba, Amra Delic, Dmitri Goldenberg, Tsvi Kuflik, Markus Zanker, and Julia Neidhardt, editors, *Proceedings of the Workshop on Web Tourism co-located with the 14th ACM International WSDM Conference (WSDM 2021), Jerusalem, Israel, March 12, 2021*, volume 2855 of *CEUR Workshop Proceedings*, pages 53–57. CEUR-WS.org, 2021.

[53] Zan Huang, Daniel Zeng, and Hsinchun Chen. A comparison of collaborative-filtering recommendation algorithms for e-commerce. *IEEE Intell. Syst.*, 22(5):68–78, 2007.

[54] Mei-Hua Hsu. A personalized english learning recommender system for ESL students. *Expert Syst. Appl.*, 34(1):683–688, 2008.

[55] Badrul Munir Sarwar, George Karypis, Joseph A. Konstan, and John Riedl. Analysis of recommendation algorithms for e-commerce. In Anant Jhingran, Jeff MacKie-Mason, and Doug J. Tygar, editors, *Proceedings of the 2nd ACM Conference on Electronic Commerce (EC-00), Minneapolis, MN, USA, October 17-20, 2000*, pages 158–167. ACM, 2000.

[56] Filipe Braida do Carmo, Carlos E. Mello, Marden B. Pasinato, and Geraldo Zimbrão. Transforming collaborative filtering into supervised learning. *Expert Syst. Appl.*, 42(10):4733–4742, 2015.

[57] Bingrui Geng, Lingling Li, Licheng Jiao, Maoguo Gong, Qing Cai, and Yue Wu. Nnia-rs: A multi-objective optimization based recommender system. *Physica A: Statistical Mechanics and its Applications*, 424:383–397, 2015.

[58] Guibing Guo, Jie Zhang, and Neil Yorke-Smith. Leveraging multiviews of trust and similarity to enhance clustering-based recommender systems. *Knowl. Based Syst.*, 74:14–27, 2015.

[59] Mukund Deshpande and George Karypis. Item-based top-N recommendation algorithms. *ACM Trans. Inf. Syst.*, 22(1):143–177, 2004.

[60] Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Comput.*, 7(1):76–80, 2003.

[61] Robert M. Bell, Yehuda Koren, and Chris Volinsky. Modeling relationships at multiple scales to improve accuracy of large recommender systems. In Pavel Berkhin, Rich Caruana, and Xindong Wu, editors, *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007*, pages 95–104. ACM, 2007.

[62] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In Ying Li, Bing Liu, and Sunita Sarawagi, editors, *Proceedings of the 14th ACM SIGKDD International Conference on*

Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008, pages 426–434. ACM, 2008.

[63] Gábor Takács, István Pilászy, Bottyán Németh, and Domonkos Tikk. Investigation of various matrix factorization methods for large recommender systems. In *Workshops Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*, pages 553–562. IEEE Computer Society, 2008.

[64] Miha Grcar, Blaz Fortuna, Dunja Mladenic, and Marko Grobelnik. knn versus SVM in the collaborative filtering framework. In Vladimir Batagelj, Hans-Hermann Bock, Anuska Ferligoj, and Ales Ziberna, editors, *Data Science and Classification*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 251–260. Springer, 2006.

[65] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[66] Thomas Hofmann. Collaborative filtering via gaussian probabilistic latent semantic analysis. In Charles L. A. Clarke, Gordon V. Cormack, Jamie Callan, David Hawking, and Alan F. Smeaton, editors, *SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28 - August 1, 2003, Toronto, Canada*, pages 259–266. ACM, 2003.

[67] John S. Breese, David Heckerman, and Carl Myers Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In Gregory F. Cooper and Serafín Moral, editors, *UAI '98: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, University of Wisconsin Business School, Madison, Wisconsin, USA, July 24-26, 1998*, pages 43–52. Morgan Kaufmann, 1998.

[68] David Ben-Shimon, Alexander Tsikinovsky, Lior Rokach, Amnon Meisels, Guy Shani, and Lihi Naamani. Recommender system from personal social networks. In *Advances in Intelligent Web Mastering, Proceedings of the 5th Atlantic Web Intelligence Conference - AWIC 2007, Fontainebleau, France, June 25 - 27, 2007*, volume 43 of *Advances in Soft Computing*, pages 47–55. Springer, 2007.

[69] Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. Content-based recommender systems: State of the art and trends. In Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, *Recommender Systems Handbook*, pages 73–105. Springer, 2011.

[70] Haofeng Jia and Erik Saule. Local is good: A fast citation recommendation approach. In *ECIR*, pages 758–764, 2018.

[71] Anas Alzoghbi, Victor Anthony Arrascue Ayala, Peter M. Fischer, and Georg Lausen. Pubrec: Recommending publications based on publicly available meta-data. In *LWA 2015 Workshops*, pages 11–18, 2015.

[72] Shuchen Li, Peter Brusilovsky, Sen Su, and Xiang Cheng. Conference paper recommendation for academic conferences. *IEEE Access*, 6:17153–17164, 2018.

[73] Qi He, Daniel Kifer, Jian Pei, Prasenjit Mitra, and C. Lee Giles. Citation recommendation without author supervision. In *WSDM*, pages 755–764, 2011.

[74] Chris A. Mack. How to Write a Good Scientific Paper: Structure and Organization. *Journal of Micro/Nanolithography, MEMS, and MOEMS*, 13(4):1 – 3, 2014.

[75] Yichuan Tang, Nitish Srivastava, and Ruslan Salakhutdinov. Learning generative models with visual attention. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 1808–1816, 2014.

[76] Wang Ling, Yulia Tsvetkov, Silvio Amir, Ramon Fermandez, Chris Dyer, Alan W. Black, Isabel Trancoso, and Chu-Cheng Lin. Not all contexts are created equal: Better word representations with variable attention. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1367–1372. The Association for Computational Linguistics, 2015.

[77] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. Neural news recommendation with multi-head self-attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6388–6393. Association for Computational Linguistics, 2019.

[78] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 276–286. Association for Computational Linguistics, 2019.

[79] Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[80] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Self-attention attribution: Interpreting information interactions inside transformer. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12963–12971. AAAI Press, 2021.

[81] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012.

[82] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016.

[83] Ilya Sutskever, James Martens, George E. Dahl, and Geoffrey E. Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 1139–1147. JMLR.org, 2013.

[84] Isaac G. Councill, C. Lee Giles, and Min-Yen Kan. Parscit: an open-source CRF reference string parsing package. In *LREC*, 2008.

[85] Minh-Thang Luong, Thuy Dung Nguyen, and Min-Yen Kan. Logical structure recovery in scholarly articles with rich document features. *IJDLS*, 1(4):1–23, 2010.

[86] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NIPS*, pages 8024–8035, 2019.

[87] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *LREC Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA.

[88] Y. Zhang and Q. Ma. Doccit2vec: Citation recommendation via embedding of content and structural contexts. *IEEE Access*, 8:115865–115875, 2020.

[89] L. V. D. Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[90] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[91] Gokul Varadhan, Shankar Krishnan, T. V. N. Sriram, and Dinesh Manocha. A simple algorithm for complete motion planning of translating polyhedral robots. *Int. J. Robotics Res.*, 25(11):1049–1070, 2006.

[92] Gokul Varadhan, Shankar Krishnan, T. V. N. Sriram, and Dinesh Manocha. Topology preserving surface extraction using adaptive subdivision. In Jean-Daniel Boissonnat and Pierre Alliez, editors, *Second Eurographics Symposium on Geometry Processing, Nice, France, July 8-10, 2004*, volume 71 of *ACM*

International Conference Proceeding Series, pages 235–244. Eurographics Association, 2004.

[93] Benoit Lavoie, Richard I. Kittredge, Tanya Korelsky, and Owen Rambow. A framework for MT and multilingual NLG systems based on uniform lexico-structural processing. In *6th Applied Natural Language Processing Conference, ANLP 2000, Seattle, Washington, USA, April 29 - May 4, 2000*, pages 60–67. ACL, 2000.

[94] Benoit Lavoie and Owen Rainbow. A fast and portable realizer for text generation systems. In *5th Applied Natural Language Processing Conference, ANLP 1997, Marriott Hotel, Washington, USA, March 31 - April 3, 1997*, pages 265–268. ACL, 1997.

[95] Jennifer Chu-Carroll. Evaluating automatic dialogue strategy adaptation for a spoken dialogue system. In *6th Applied Natural Language Processing Conference, ANLP 2000, Seattle, Washington, USA, April 29 - May 4, 2000*, pages 202–209. ACL, 2000.

[96] James F. Allen, Bradford W. Miller, Eric K. Ringger, and Teresa Sikorski. A robust system for natural spoken dialogue. In Aravind K. Joshi and Martha Palmer, editors, *34th Annual Meeting of the Association for Computational Linguistics, 24-27 June 1996, University of California, Santa Cruz, California, USA, Proceedings*, pages 62–70. Morgan Kaufmann Publishers / ACL, 1996.

[97] Mary P. Harper, Christopher M. White, Wen Wang, Michael T. Johnson, and Randall A. Helzerman. The effectiveness of corpus-induced dependency grammars for post-processing speech. In *6th Applied Natural Language Processing Conference, ANLP 2000, Seattle, Washington, USA, April 29 - May 4, 2000*, pages 102–109. ACL, 2000.

[98] Hiroshi Maruyama. Structural disambiguation with constraint propagation. In Robert C. Berwick, editor, *28th Annual Meeting of the Association for Computational Linguistics, 6-9 June 1990, University of Pittsburgh, Pittsburgh, Pennsylvania, USA, Proceedings*, pages 31–38. ACL, 1990.

[99] Patrick Pantel and Dekang Lin. Word-for-word glossing with contextually similar words. In *6th Applied Natural Language Processing Conference,*

ANLP 2000, Seattle, Washington, USA, April 29 - May 4, 2000, pages 78–85. ACL, 2000.

[100] William A. Gale and Kenneth Ward Church. Identifying word correspondences in parallel texts. In *Speech and Natural Language, Proceedings of a Workshop held at Pacific Grove, California, USA, February 19-22. 1991.* Morgan Kaufmann, 1991.

[101] Alper Yilmaz, Khurram Shafique, and Mubarak Shah. Estimation of rigid and non-rigid facial motion using anatomical face model. In *16th International Conference on Pattern Recognition, ICPR 2002, Quebec, Canada, August 11-15, 2002*, pages 377–380. IEEE Computer Society, 2002.

[102] Jiang Chen and Jian-Yun Nie. Automatic construction of parallel english-chinese corpus for cross-language information retrieval. In *6th Applied Natural Language Processing Conference, ANLP 2000, Seattle, Washington, USA, April 29 - May 4, 2000*, pages 21–28. ACL, 2000.

[103] Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguistics*, 19(2):263–311, 1993.

[104] F. Brugnara, M. Cettolo, M. Federico, and D. Giuliani. A system for the segmentation and transcription of italian radio news. In *Content-Based Multimedia Information Access - Volume 1*, RIAO '00, page 364–371, Paris, FRA, 2000. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.

[105] Tasos Anastasakos, John W. McDonough, Richard M. Schwartz, and John Makhoul. A compact model for speaker-adaptive training. In *The 4th International Conference on Spoken Language Processing, Philadelphia, PA, USA, October 3-6, 1996*. ISCA, 1996.

[106] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. SPECTER: document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2270–2282. Association for Computational Linguistics, 2020.

[107] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3613–3618. Association for Computational Linguistics, 2019.

[108] Hao Zheng and Mirella Lapata. Sentence centrality revisited for unsupervised summarization. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6236–6247. Association for Computational Linguistics, 2019.

[109] Ziyi Yang, Chenguang Zhu, Robert Gmyr, Michael Zeng, Xuedong Huang, and Eric Darve. TED: A pretrained unsupervised summarization model with theme modeling and denoising. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1865–1874. Association for Computational Linguistics, 2020.

[110] Xingxing Zhang, Furu Wei, and Ming Zhou. HIBERT: document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5059–5069. Association for Computational Linguistics, 2019.

[111] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR, 2017.

[112] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan

Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *EMNLP*, pages 38–45, 2020.

[113] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[114] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*, pages 55–60. The Association for Computer Linguistics, 2014.

[115] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016.

[116] Henry Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *J. Am. Soc. Inf. Sci.*, 24(4):265–269, 1973.

[117] Bela Gipp and Jöran Beel. Citation proximity analysis (cpa) : A new approach for identifying related work based on co-citation analysis. In *ISSI*, São Paulo, 2009.

[118] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823. IEEE Computer Society, 2015.

[119] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems 29: Annual*

Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, pages 1849–1857, 2016.

[120] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. Automatic classification of citation function. In Dan Jurafsky and Éric Gaussier, editors, *EMNLP 2006, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 22-23 July 2006, Sydney, Australia*, pages 103–110. ACL, 2006.

[121] Ayushi Dalmia, Ganesh J, and Manish Gupta. Towards interpretation of node embeddings. In Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis, editors, *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 945–952. ACM, 2018.

[122] Taeho Jo. *Text Mining: Concepts, Implementation, and Big Data Challenge*. Studies in Big Data. Springer International Publishing, 2018.

[123] Stanley F. Chen. Aligning sentences in bilingual corpora using lexical information. In Lenhart K. Schubert, editor, *31st Annual Meeting of the Association for Computational Linguistics, 22-26 June 1993, Ohio State University, Columbus, Ohio, USA, Proceedings*, pages 9–16. ACL, 1993.

[124] Carolyn Penstein Rosé. A framework for robust semantic interpretation learning. In *6th Applied Natural Language Processing Conference, ANLP 2000, Seattle, Washington, USA, April 29 - May 4, 2000*, pages 311–318. ACL, 2000.

[125] Ralph Grishman, Catherine Macleod, and Adam L. Meyers. Comlex syntax: Building a computational lexicon. In *15th International Conference on Computational Linguistics, COLING 1994, Kyoto, Japan, August 5-9, 1994*, pages 268–272, 1994.

[126] Dianne R. Kumar and Walid A. Najjar. Combining adaptive and deterministic routing: Evaluation of a hybrid router. In Anand Sivasubramaniam and Mario Lauria, editors, *Network-Based Parallel Computing: Communication, Architecture, and Applications, Third International Workshop, CANPC '99, Orlando, Forida, USA, January 9, 1999, Proceedings*, volume 1602 of *Lecture Notes in Computer Science*, pages 150–164. Springer, 1999.

[127] José Duato. A new theory of deadlock-free adaptive routing in wormhole networks. *IEEE Trans. Parallel Distributed Syst.*, 4(12):1320–1331, 1993.

[128] Xiaoming Wei, Wei Li, Klaus Mueller, and Arie E. Kaufman. The lattice-boltzmann method for simulating gaseous phenomena. *IEEE Trans. Vis. Comput. Graph.*, 10(2):164–176, 2004.

[129] Jos Stam and Eugene Fiume. Turbulent wind fields for gaseous phenomena. In Mary C. Whitton, editor, *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1993, Anaheim, CA, USA, August 2-6, 1993*, pages 369–376. ACM, 1993.

[130] Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. Aligning sentences in parallel corpora. In Douglas E. Appelt, editor, *29th Annual Meeting of the Association for Computational Linguistics, 18-21 June 1991, University of California, Berkeley, California, USA, Proceedings*, pages 169–176. ACL, 1991.

[131] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguistics*, 19(1):61–74, 1993.

[132] Dekai Wu. Aligning a parallel english-chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, ACL '94, page 80–87, USA, 1994. Association for Computational Linguistics.

[133] Julian Kupiec. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, ACL '93, page 17–22, USA, 1993. Association for Computational Linguistics.

[134] Thomas C. Rindflesch, Jayant V. Rajan, and Lawrence Hunter. Extracting molecular binding relationships from biomedical text. In *Sixth Applied Natural Language Processing Conference*, April 2000.

[135] Kenneth Ward Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Second Conference on Applied Natural Language Processing*, February 1988.

[136] Douglas R. Cutting, Julian Kupiec, Jan O. Pedersen, and Penelope Sibun. A practical part-of-speech tagger. In *3rd Applied Natural Language Processing Conference, ANLP 1992, Trento, Italy, March 31 - April 3, 1992*, pages 133–140. ACL, 1992.

[137] Eric Brill and Philip Resnik. A rule-based approach to prepositional phrase attachment disambiguation. In *15th International Conference on Computational Linguistics, COLING 1994, Kyoto, Japan, August 5-9, 1994*, pages 1198–1204, 1994.

[138] Elaine Rich and Susann LuperFoy. An architecture for anaphora resolution. In *Proceedings of the Second Conference on Applied Natural Language Processing*, ANLC '88, page 18–24, USA, 1988. Association for Computational Linguistics.

[139] Fred Karlsson. Constraint grammar as a framework for parsing running text. In *Proceedings of the 13th Conference on Computational Linguistics - Volume 3*, COLING '90, page 168–173, 1990.

[140] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, March 1990.

[141] Thorsten Brants. Tnt: A statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, page 224–231, 2000.

[142] Adwait Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In Eric Brill and Kenneth Church, editors, *Conference on Empirical Methods in Natural Language Processing, EMNLP 1996, Philadelphia, PA, USA, May 17-18, 1996*, 1996.

[143] Kenneth Ward Church. A stochastic parts program and noun phrase parser for unrestricted text. In *2nd Applied Natural Language Processing Conference, ANLP 1988, Austin, Texas, USA, February 9-12, 1988*, pages 136–143. ACL, 1988.

[144] Eric Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Comput. Linguistics*, 21(4):543–565, 1995.

[145] Marilyn A. Walker. Evaluating discourse processing algorithms. *CoRR*, abs/cmp-lg/9410006, 1994.

[146] Daniel Marcu, Lynn Carlson, and Maki Watanabe. The automatic translation of discourse structures. In *6th Applied Natural Language Processing Conference, ANLP 2000, Seattle, Washington, USA, April 29 - May 4, 2000*, pages 9–17. ACL, 2000.

[147] Mark Johnson. PCFG models of linguistic tree representations. *Comput. Linguistics*, 24(4):613–632, 1998.

[148] Lance A. Ramshaw and Mitch Marcus. Text chunking using transformation-based learning. In David Yarowsky and Kenneth Church, editors, *Third Workshop on Very Large Corpora, VLC@ACL 1995, Cambridge, Massachusetts, USA, June 30, 1995*, 1995.

[149] Adam Meyers, Roman Yangarber, and Ralph Grishman. Alignment of shared forests for bilingual corpora. In *16th International Conference on Computational Linguistics, Proceedings of the Conference, COLING 1996, Center for Sprogteknologi, Copenhagen, Denmark, August 5-9, 1996*, pages 460–465, 1996.

[150] Adwait Ratnaparkhi. A linear observed time statistical parser based on maximum entropy models. In Claire Cardie and Ralph M. Weischedel, editors, *Second Conference on Empirical Methods in Natural Language Processing, EMNLP 1997, Providence, RI, USA, August 1-2, 1997*. ACL, 1997.

[151] Heidi Fox. Phrasal cohesion and statistical machine translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002, Philadelphia, PA, USA, July 6-7, 2002*, pages 304–3111, 2002.

[152] Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. Word-sense disambiguation using statistical methods. In Douglas E. Appelt, editor, *29th Annual Meeting of the Association for Computational Linguistics, 18-21 June 1991, University of California, Berkeley, California, USA, Proceedings*, pages 264–270. ACL, 1991.

[153] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In Hans Uszkoreit, editor, *33rd Annual Meeting of the Association for Computational Linguistics, 26-30 June 1995, MIT, Cambridge, Massachusetts, USA, Proceedings*, pages 189–196. Morgan Kaufmann Publishers / ACL, 1995.

[154] Ido Dagan and Alon Itai. Word sense disambiguation using a second language monolingual corpus. *Comput. Linguistics*, 20(4):563–596, 1994.

[155] Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Comput. Linguistics*, 16(2):79–85, 1990.

[156] Demetri Terzopoulos and Keith Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(6):569–579, 1993.

[157] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In Patrick J. Hayes, editor, *Proceedings of the 7th International Joint Conference on Artificial Intelligence, IJCAI '81, Vancouver, BC, Canada, August 24-28, 1981*, pages 674–679. William Kaufmann, 1981.

[158] Irfan A. Essa and Alex Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):757–763, 1997.

[159] James R. Bergen, P. Anandan, Keith J. Hanna, and Rajesh Hingorani. Hierarchical model-based motion estimation. In Giulio Sandini, editor, *Computer Vision - ECCV'92, Second European Conference on Computer Vision, Santa Margherita Ligure, Italy, May 19-22, 1992, Proceedings*, volume 588 of *Lecture Notes in Computer Science*, pages 237–252. Springer, 1992.

[160] Douglas DeCarlo and Dimitris N. Metaxas. Optical flow constraints on deformable models with applications to face tracking. *Int. J. Comput. Vis.*, 38(2):99–127, 2000.

[161] Michael J. Black and Yaser Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *Procedings of the Fifth International Conference on Computer Vision (ICCV 95), Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, June 20-23, 1995*, pages 374–381. IEEE Computer Society, 1995.

[162] Ying-li Tian, Takeo Kanade, and Jeffrey F. Cohn. Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(2):97–115, 2001.

[163] Takeo Kanade, Ying-li Tian, and Jeffrey F. Cohn. Comprehensive database for facial expression analysis. In *4th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2000), 26-30 March 2000, Grenoble, France*, pages 46–53. IEEE Computer Society, 2000.

[164] Gianluca Donato, Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski. Classifying facial actions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(10):974–989, 1999.

[165] M. J. F. Gales. Maximum likelihood linear transformations for hmm-based speech recognition. *Comput. Speech Lang.*, 12(2):75–98, 1998.

[166] Mark J. F. Gales and Philip C. Woodland. Mean and variance adaptation within the MLLR framework. *Comput. Speech Lang.*, 10(4):249–264, 1996.

[167] Jean-Luc Gauvain and Chin-Hui Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Trans. Speech Audio Process.*, 2(2):291–298, 1994.

[168] Mark J. F. Gales. Semi-tied covariance matrices for hidden markov models. *IEEE Trans. Speech Audio Process.*, 7(3):272–281, 1999.

[169] Philip C. Woodland, Mark John Francis Gales, and David Pye. Improving environmental robustness in large vocabulary speech recognition. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, ICASSP '96, Atlanta, Georgia, USA, May 7-10, 1996*, pages 65–68. IEEE Computer Society, 1996.

[170] David A. Ross and Richard S. Zemel. Learning parts-based representations of data. *J. Mach. Learn. Res.*, 7:2369–2397, 2006.

[171] Geoffrey E. Hinton and Richard S. Zemel. Autoencoders, minimum description length and helmholtz free energy. In Jack D. Cowan, Gerald Tesauro, and Joshua Alspector, editors, *Advances in Neural Information Processing Systems 6, [7th NIPS Conference, Denver, Colorado, USA, 1993]*, pages 3–10. Morgan Kaufmann, 1993.

[172] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*, pages 556–562. MIT Press, 2000.

[173] Thomas Hofmann. Probabilistic latent semantic analysis. In Kathryn B. Laskey and Henri Prade, editors, *UAI '99: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, July 30 - August 1, 1999*, pages 289–296. Morgan Kaufmann, 1999.

[174] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, 42(1/2):177–196, 2001.

[175] Kobus Barnard, Pinar Duygulu, David A. Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. Matching words and pictures. *J. Mach. Learn. Res.*, 3:1107–1135, 2003.

[176] Patrik O. Hoyer. Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.*, 5:1457–1469, 2004.

[177] Michael S. Lewicki and Terrence J. Sejnowski. Learning overcomplete representations. *Neural Comput.*, 12(2):337–365, 2000.

Selected List of Publications

- **Journals**

  [1] Yang Zhang, and Qiang Ma. MP-BERT4CR: Recommending Multiple Positive Citations for Academic Manuscripts via Content-Dependent BERT and Multi-Positive Triplet. *IEICE Transactions on Information and Systems* (Under review))

  [2] Yang Zhang, and Qiang Ma. Dual Attention Model for Citation Recommendation with Analyses on Explainability of Attention Mechanisms and Qualitative Experiments. *Journal of Computational Linguistics* (Accepted for publication: 04 Jan 2022))

  [3] Yang Zhang and Qiang Ma. Citation Recommendations Considering Content and Structural Context Embedding. *IEEE Access*, vol. 8, pp. 115865-115875, 2020.

- **International Conferences and Workshops**

  [4] Yang Zhang and Qiang Ma. "On-the-fly" Citation Recommendation based on Content-dependent Embeddings, 2022 International Conference on Research and Development in Information Retrieval (Under review). (Submitted)

  [5] Yang Zhang and Qiang Ma. Recommending Multiple Positive Citations for Manuscript via Content-Dependent Modeling and Multi-Positive Triplet. In *The 20th IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Melbourne, Australia (Online), 2021.

[6] Yang Zhang and Qiang Ma. Dual Attention Model for Citation Recommendation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3179—3189, Barcelona, Spain (Online), 2021.

[7] Yang Zhang and Qiang Ma. Citation Recommendations Considering Content and Structural Context Embedding. In *2020 IEEE International Conference on Big Data and Smart Computing*, pages 1–7, Busan, Korea (South), February 19-22, 2020.

Appendix

A Supplementary Samples

A.1 Supplementary Samples (1 & 2) from ACL Dataset

Considering the first sample in Table 6.1, it could be drawn that the author would like to cite studies on alignment techniques based on statistical methods or lexical methods. The study is generally about proposing a language alignment algorithm. According to Figure 6.1, the topic related words, such as such as "**lexical**", "**method**", and "**alignment**" are recognized in the top 15 scored items from self-attention; whereas the connecting words, such as "**we**" and "**et**" are also recognized due to the high pair-wise similarities they have received. Additive attention in Figure 6.3 assigned higher weights to the unique words (low pair-wise similarities) of the context, most of them are relevant to the general topic of the context, such as "**Aligning**", "**parallel**", and "**cognateness**". However, some words which are directly relevant to the citing intent (such as "**lexical**" from self-attention) are not recognized by additive attention. It is also realized that the words detected by additive attention are mostly appeared in the content of the target paper.

For the second sample in Table 6.1, it is realized that the author is citing the paper proposed COMLEX grammar and lexicon, and providing a description on the contribution of it (i.e. assigning syntactic functional roles to constituents). Similar to sample 1, self-attention has recognized topic related words (Figure 6.1), such as "**syntactic**", "**grammar**", and "**lexicon**"; more over, the connecting words with high word similarities, such as "**by**", and "**and**". For the unique words that the additive attention has detected (Figure 6.3), "**COMLEX**" and "**functional**"

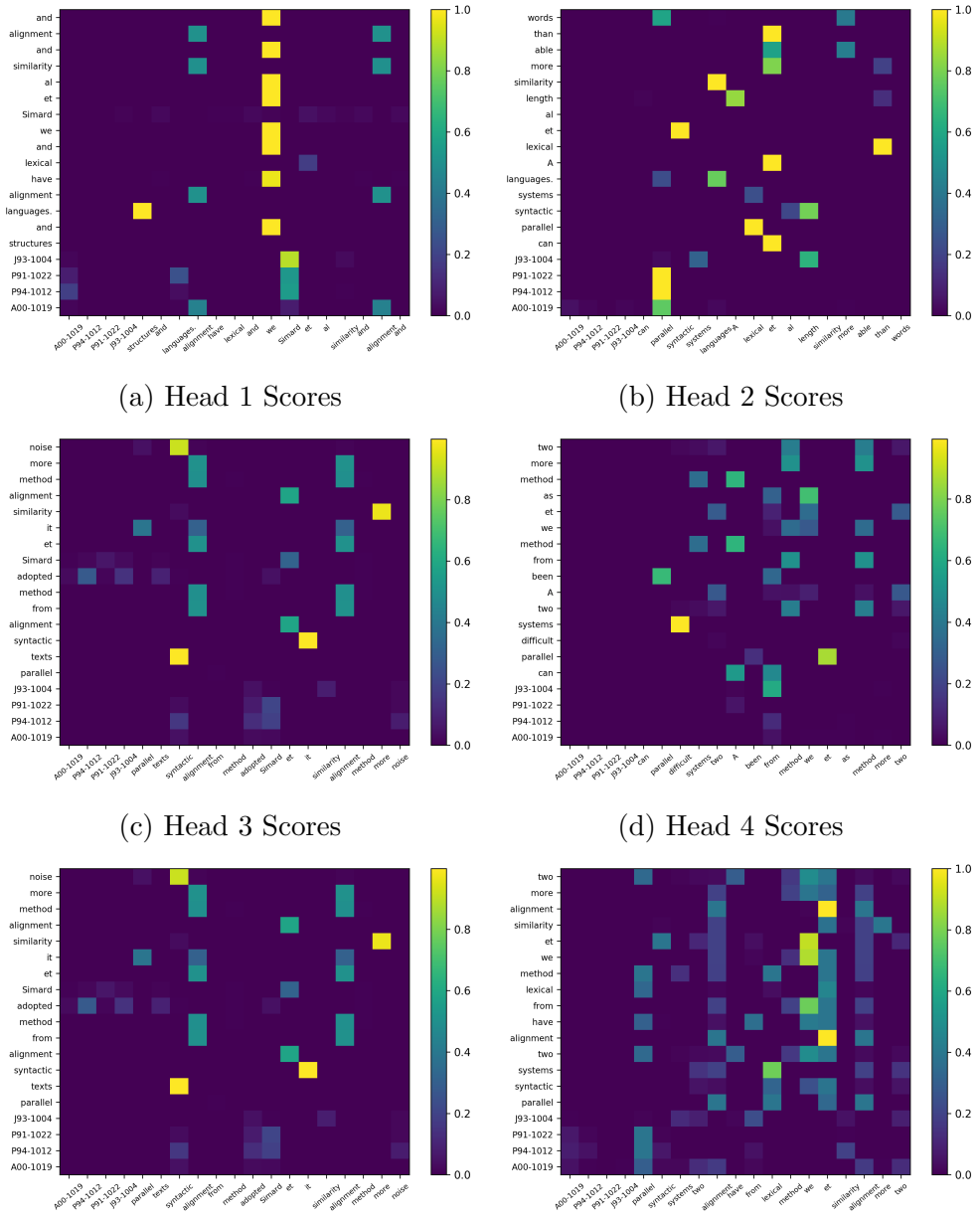
Table 6.1: Textual Information of Supplementary Sample 1 & 2

No.	Dataset	Source paper ref.	Page	Target paper ref.	Context
1	ACL	Chen and Nie [102]	4	Chen [123]	others can be very noisy. Aligning English-Chinese parallel texts is already very difficult because of the great differences in the syntactic structures and writing systems of the two languages. A number of alignment techniques have been proposed, varying from statistical methods to lexical methods (Kay and RSscheisen, 1993 [=?=]; The method we adopted is that of Simard et al (1992). Because it considers both length similarity and cognateness as alignment criteria, the method is more robust and better able to deal with noise than pure length-based methods. Cognates are identical sequences of characters in corresponding words in two
2	ACL	Rosé [124]	3	Grishman, Macleod , and Meyers [125]	into the corresponding slots in the so Otherwise the constructor function fails. Take as an example the sentence "The meeting I had scheduled was canceled by you." as it is processed by using the CARMEL grammar and lexicon, which is built on top of the COMLEX lexicon [=?=] The grammar assigns deep syntactic functional roles to constituents. Thus, "you" is the deep subject of and "the meeting" is the direct object both of and of The detailed subcategorization classes associated with verbs, nouns, and adjectives in COMLEX make it possible to determine what these

are considered to be directly relevant to the citing intent of the context, however the rest of the words are not considered to be relevant to the citing intent, or the general topic of the context. Additive attention over-emphasized the words that are not properly pre-processed, such as "**"you"**", and "**"the**".

Overall, the characteristics of the attention mechanisms of supplementary sample 1 and sample 2 correspond to the main samples in Section 3.6, except that the additive attention over-emphasized the some wrong words. It could be drawn that, the top contributed words to predict the output that the additive attention recognized, could be irrelevant to the citing intent, however should be ultimately

unique (pair-wise low similarity).



(a) Head 1 Scores

(b) Head 2 Scores

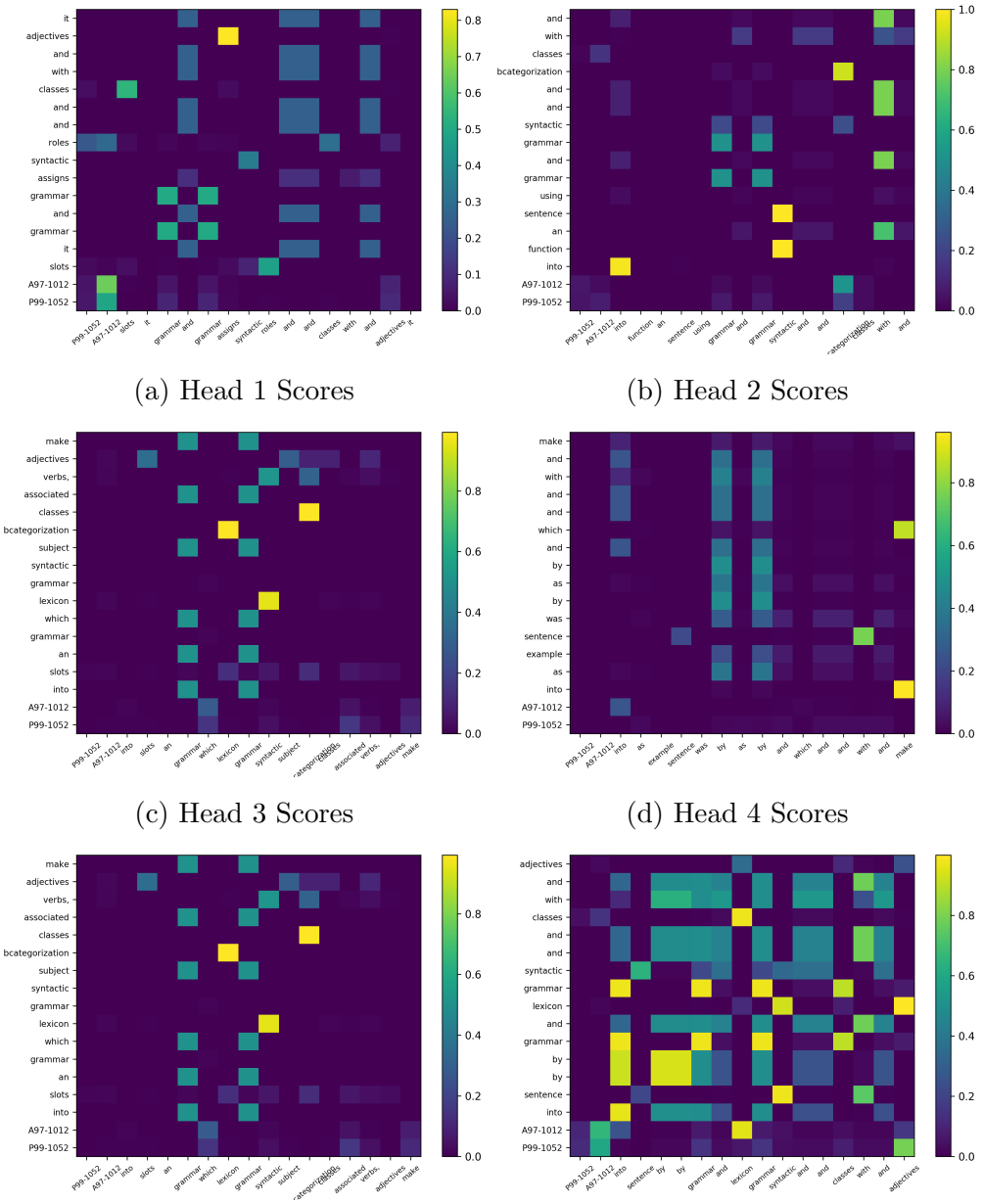
(c) Head 3 Scores

(d) Head 4 Scores

(e) Head 5 Scores

(f) Averaged Scores Across 5 Heads

Figure 6.1: Pair-wise Self-attention Scores (Top 15 Items) for Supplementary Sample 1 via Complete DACR



(a) Head 1 Scores

(b) Head 2 Scores

(c) Head 3 Scores

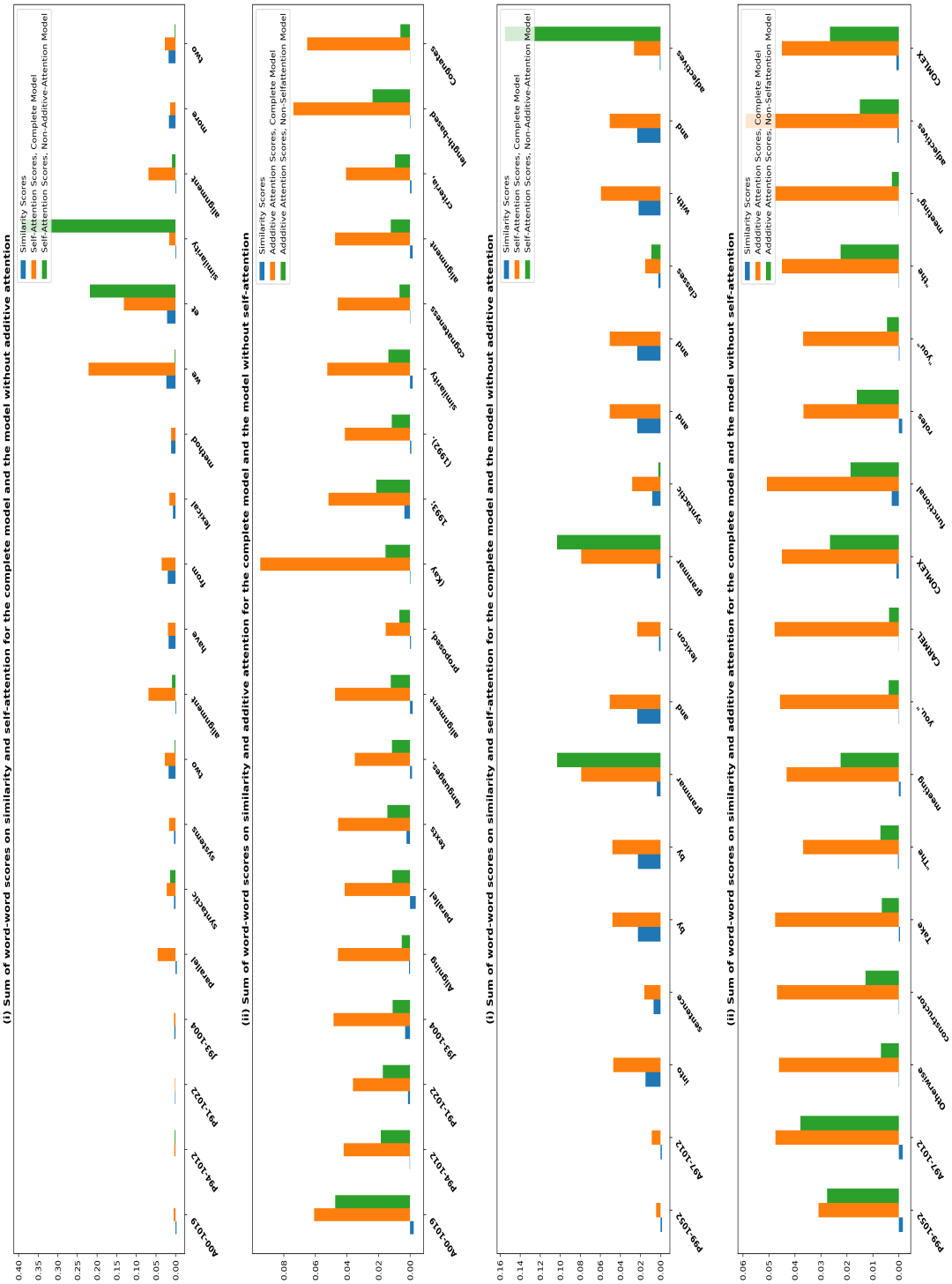
(d) Head 4 Scores

(e) Head 5 Scores

(f) Averaged Scores Across 5 Heads

Figure 6.2: Pair-wise Self-attention Scores (Top 15 Items) for Supplementary Sample 2 via Complete DACR

Figure 6.3: Scores of Additive Attention (Top 15) and Summed Self-attention Against Similarities for Supplementary Sample 1 & 2



(a) Supplementary Sample 1

(b) Supplementary Sample 2

A.2 Supplementary Samples (3 & 4) from DBLP Dataset

Table 6.2: Textual Information of Supplementary Sample 3 & 4

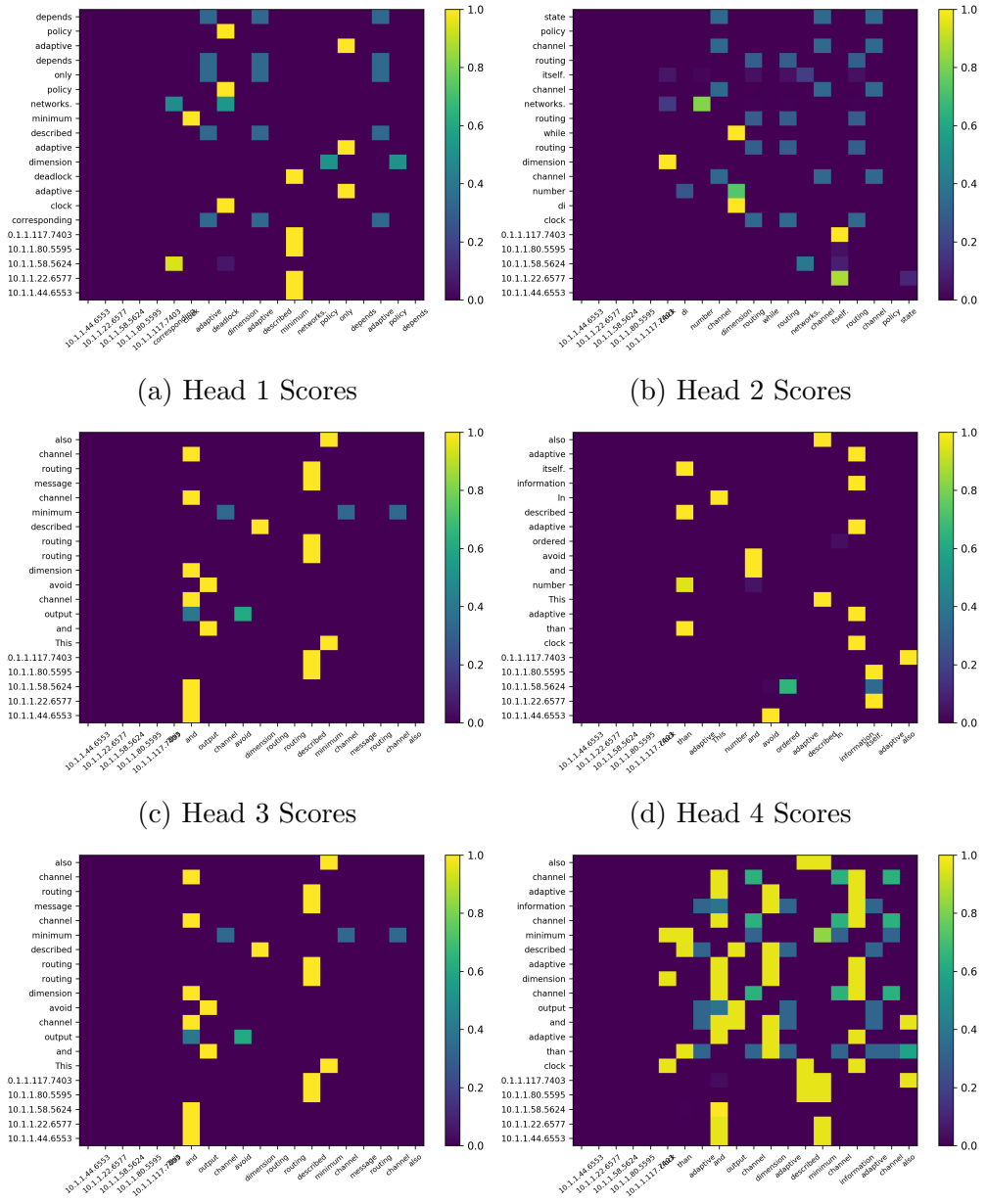
No.	Dataset	Source paper ref.	Page	Target paper ref.	Context
3	DBLP	Kumar and Najjar [126]	1	Duato [127]	corresponding clock cycles, can be significantly lower than adaptive routers This di erence in router delays is due to two main reasons: number of VCs and output (OP) channel selection. Two VCs are su cient to avoid deadlock in dimension ordered routing [6]; while adaptive routing (as described in [=?=] requires a minimum of three VCs in k-ary n-cube networks. In dimension-ordered routing, the OP channel selection policy only depends on information contained in the message header itself. In adaptive routing the OP channel selection policy depends also on the state of the router (i.e the occupancy of various
4	DBLP	Wei et al. [128]	3	Stam and Fiume [129]	the physically correct large-scale behaviors and interactions of the gaseous phenomena, at realtime speeds. What we require now is an equally efficient way to add the small-scale turbulence details into the visual simulation and render these to the screen. One way to model the small-scale turbulence is through spectral analysis [=?=] Turbulent motion is first defined in Fourier space and then it is transformed to give periodic and chaotic vector fields that can be combined with the global motions. Another approach is to take advantage of commodity texture mapping hardware, using textured splats [6] as the rendering primitive. King et

Considering the third sample in Table 6.2, it could be drawn that the author would like to cite a paper on adaptive routing by addressing its technique features, i.e. three VCs were utilized to avoid deadlock. Similar to the previous analyses, self-attention recognized words which are relevant to the citing intent, such as "**adaptive**", "**dimension**", and "**clock**, but also the connecting words with high pair-wise word similarities, such as "**and**", and "**also**". Additive attention (Figure 6.6) mostly recognized the words relevant to the citing intent, such as "**routing**", and "**n-cude**", as the most of the word relevant to the citing intent are not common to appear.

For the fourth sample in Table 6.2, it is realized that the author would like to cite the study about spectral analysis by addressing the characterise of the technique. According to Figure 6.5, similar to sample 1, self-attention has recognized words which relevant to the citing intent, such as "**spectral**", "**analysis**", "**rendering**", etc., but also few connecting words such as "**can**" and "**et**" are recognized.

Additive attention (Figure 6.6) mostly recognized the words relevant to the citing intent, such as "**turbulence**","**spectral**", and "**analysis**". However, some unique but irrelevant words are also recognized, such as "**King**."

Overall, the characteristics of the attention mechanisms of supplementary sample 3 and sample 4 correspond to the main samples in Section 3.6, and supplementary sample 1 and 2. It could be drawn that, both self-attention and additive attention recognize the words which are relevant to the citing intent, however self-attention may also assign high weights to the connecting words, whereas additive attention may assign high weighs to the unique but irrelevant words.



(a) Head 1 Scores

(b) Head 2 Scores

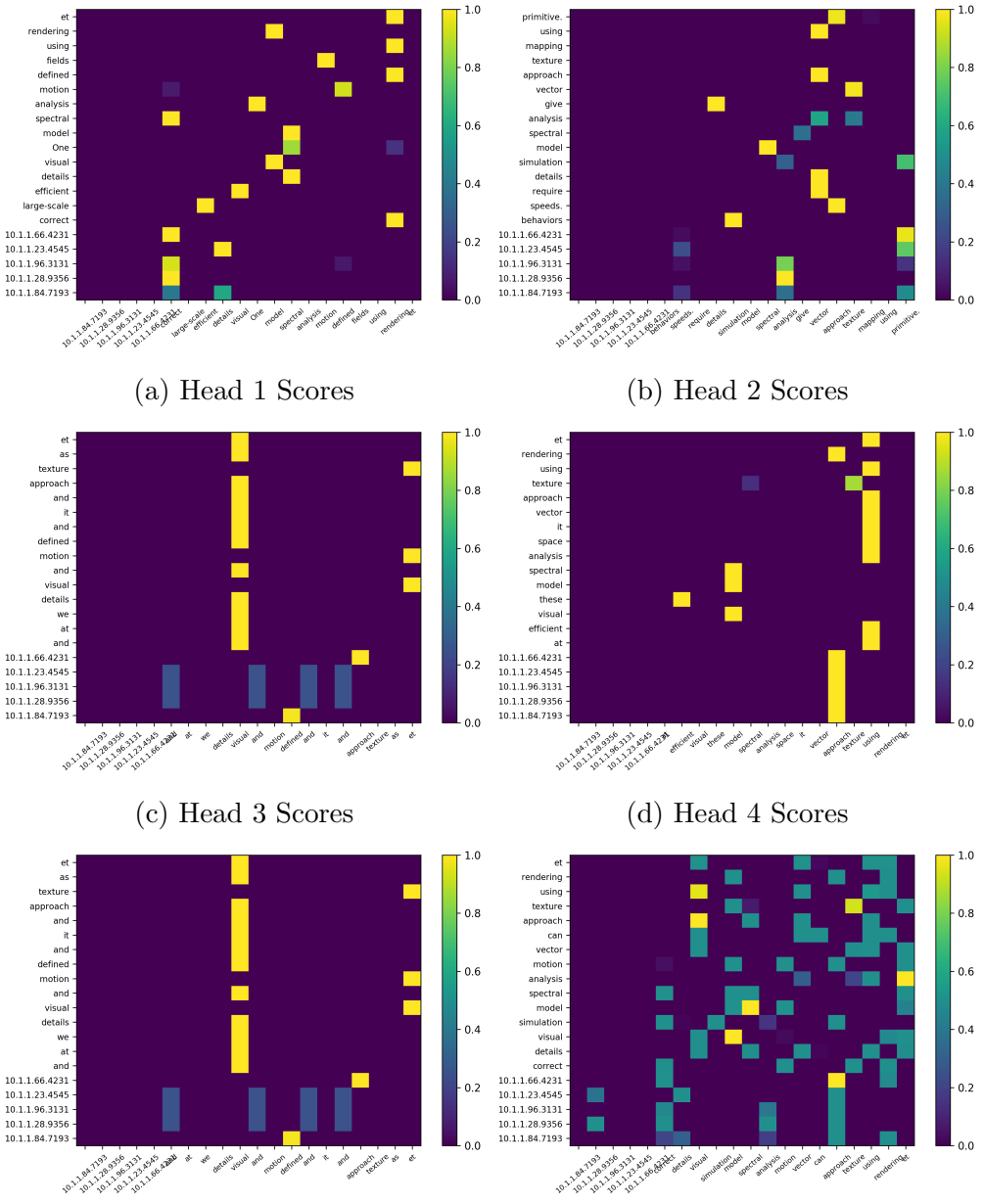
(c) Head 3 Scores

(d) Head 4 Scores

(e) Head 5 Scores

(f) Averaged Scores Across 5 Heads

Figure 6.4: Pair-wise Self-attention Scores (Top 15 Items) for Supplementary Sample 3 via Complete DACR



(a) Head 1 Scores

(b) Head 2 Scores

(c) Head 3 Scores

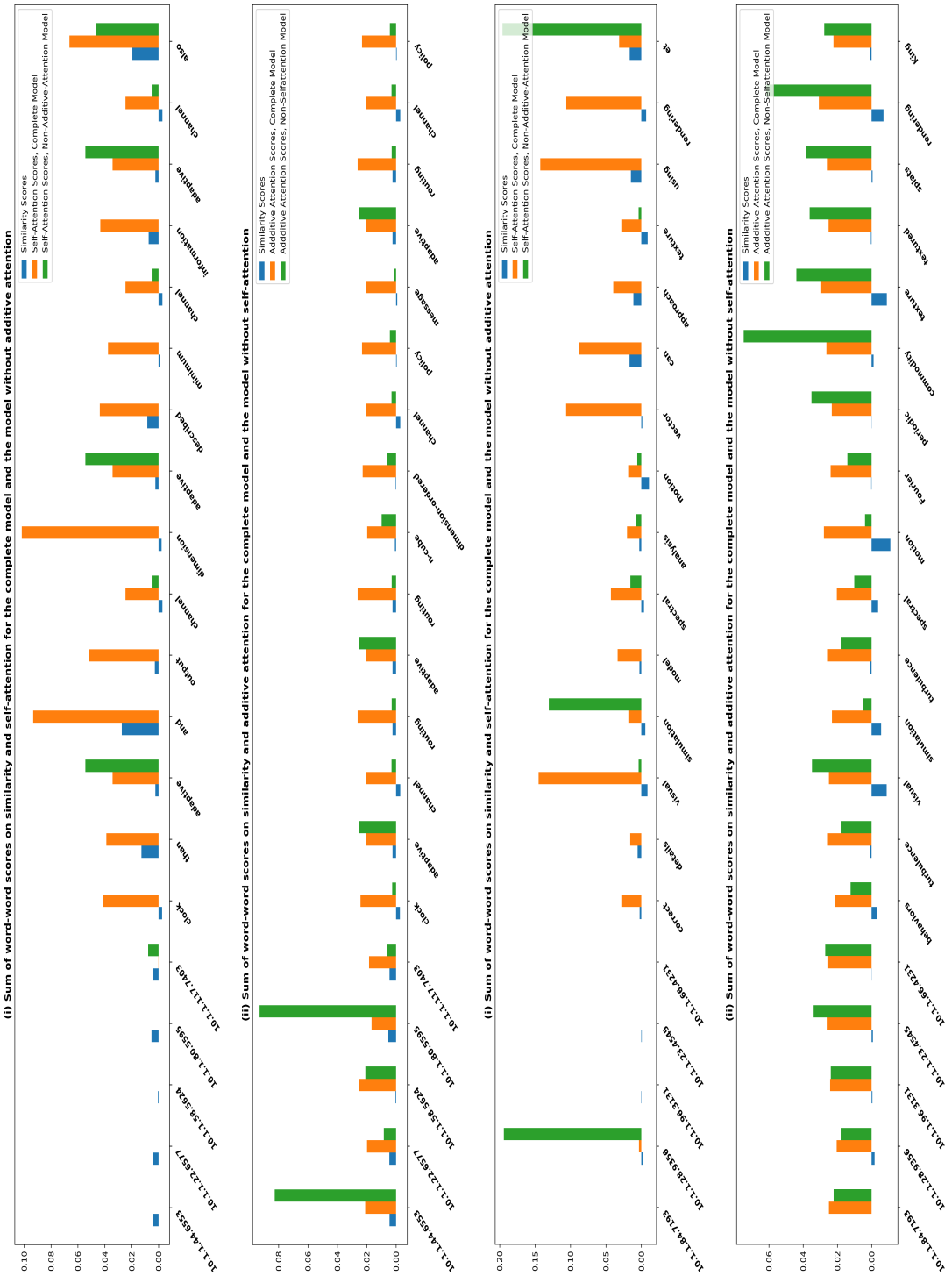
(d) Head 4 Scores

(e) Head 5 Scores

(f) Averaged Scores Across 5 Heads

Figure 6.5: Pair-wise Self-attention Scores (Top 15 Items) for Supplementary Sample 4 via Complete DACR

Figure 6.6: Scores of Additive Attention (Top 15) and Summed Self-attention Against Similarities for Supplementary Sample 3 & 4



(a) Supplementary Sample 3

(b) Supplementary Sample 4

B Questionnaire and Answers

B.1 Answers for Input Context 1 (IC1)

Input Context (IC) 1:

"*...Some are highly parallel and easy to align while others can be very noisy. Aligning English-Chinese parallel texts is already very difficult because of the great differences in the syntactic structures and writing systems of the two languages. A number of alignment techniques have been proposed, varying from statistical methods =?= to lexical methods (Kay and Röscheisen, 1993; Chen, 1993). The method we adopted is that of Simard et al. (1992). Because it considers both length similarity and cognateness as alignment criteria, the method is more robust and better able to deal with noise than pure length-based methods...*" [102]

What is ground truth paper [130] about?

- **Analyzer 1:** Provide the past studies about sentence alignment, especially the ones adapts statistical methods.
- **Analyzer 2:** The paper describes a pure statistical technique rather than lexical methods for aligning sentences.
- **Analyzer 3:** This paper describes statistical methods of parallel corpora alignment techniques.

Is the first candidate (CAN1) by [131] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5

- **Analyzer 1:** No. The candidate paper aims to propose a metric for techniques of text analysis which is different from the purpose of the citing intent. Rate: 0.
- **Analyzer 2:** No. The paper does not focus on the aligning methods of translation. The goal of the paper is to present a practical measure that is motivated by statistical considerations and that can be used in a number of settings. Rate: 0.
- **Analyzer 3:** No. This paper describes the basis of a measure based on likelihood ratios that can be applied to the analysis of text, which is little relevant to the comparative corpora alignment. Rate: 1.

Is the second candidate (CAN2) by [103] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5

- **Analyzer 1:** Yes. It might be suitable. The candidate paper proposes a technique for machine translation which involves word-to-word alignment via statical methods. The paper is also cited in other places for introduction of machine translation and word alignment. Rate: 4.
- **Analyzer 2:** No. The paper does not propose new statistical technique for aligning sentences, it discusses the methods for estimating parameters of five statistical methods. It is better to use the papers proposing these five statistical methods. Rate: 3.
- **Analyzer 3:** No. This paper compares a set of statistical models of the translation process and give algorithms for estimating the parameters of these models. It, however, does not come up with a text alignment technique itself. Rate: 2.

Is the third candidate (CAN3) by [132] suitable to be used a citations for the context? Explain reasons, and rate from 0 to 5

- **Analyzer 1:** No. The candidate paper aims to: 1. propose a dataset for English-Chinese translation 2. experiment one of the previous word alignment approaches, which are different to the purpose of the citing intent. Score : 0.
- **Analyzer 2:** No. The paper does not propose a pure statistical technique for aligning sentences, it combines the statistical technique with lexical cues. Rate:2.
- **Analyzer 3:** Yes. This paper proposes an improved statistical method incorporating domain-specific lexical cues to the task of aligning English with Chinese. Rate: 4.

Is the forth candidate (CAN4) by [100] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5

- **Analyzer 1:** No. The candidate paper describes a technique for detection of word correspondences which is a different task to word alignment. Score: 0.

- **Analyzer 2:** No. Rate:3. Although the method is statistical-based, the paper focuses on the correspondence problem rather than alignment problem.

- **Analyzer 3:** Probably. This paper introduces several novel techniques that find corresponding words in parallel texts given aligned regions. However, it distinguishes the terms alignment and correspondence. For this, it focused more on word correspondence problem than sentence-level alignment. Rate: 3.

Is the fifth candidate (CAN5) by [133] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5

- **Analyzer 1:** No. The candidate paper proposes a word alignment technique based on noun phrases that is different to the citing intent. Score: 0.

- **Analyzer 2:** Yes. The paper aims to solve noun phrase alignment problem, and it focuses on statistics-based techniques. Rate:4.

- **Analyzer 3:** No. The algorithm described in this paper provides a practical way for obtaining correspondences between noun phrases in a bilingual corpus. It differs from statistical method. Rate: 3.

B.2 Answers for IC2

IC2:

"*...The output produced is in the tradition of partial parsing (Hindle 1983, McDonald 1992, Weischedel et al. 1993) and concentrates on the simple noun phrase,what Weischedel et al. (1993) call the "core noun phrase,"that is a noun phrase with no modification to the right of the head. Several approaches provide similar output based on statistics (=?=, Zhai 1997, for example),a finite-state machine(Ait-Mokhtar and Chanod 1997), or a hybrid approach combining statistics and linguistic rules (Voutilainen and Padro 1997)...*"[134]

What is ground truth paper [135] about?

- **Analyzer 1:** The source paper is citing papers about noun phrase parsing based on statistical methods.

- **Analyzer 2:** The paper presents a noun phrase parser and is a statistics-based method.

- **Analyzer 3:** This paper is cited because it proposed a statistical method solving the task of noun-phrase parsing.

Is CAN1 by [136] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5

- **Analyzer 1:** The candidate paper seems to be related, as it proposes a parsing tagger based statistical methods. However, the context asks for a method specially designed for noun phrase parsing, and secondly the candidate paper has been cited at the beginning of the paragraph, which seems to be redundant for a citation here. Rate:1.

- **Analyzer 2:** No. The paper focuses on Part-of-Speech Tagger based on a hidden Markov model. Rate:3.

- **Analyzer 3:** No. This paper presents an implementation of a part-of-speech tagger based on a hidden Markov model. It is not either a statistical method or solving a noun-phrase parsing task. Rate: 2.

Is CAN2 by [137] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5

- **Analyzer 1:** No. The candidate paper is about a rule-based phrase parser which is different from the citing intention. Rate: 0.

- **Analyzer 2:** No. The paper describes a rule-based approach to prepositional phrase attachment, which is not a noun phrase parser and is not a statistics-based method. Rate: 0.

- **Analyzer 3:** No. This paper aims to solve the prepositional phrase attachment disambiguation problem, which is little relevant to the intention of citing place. Rate: 2.

Is CAN3 by [138] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5

- **Analyzer 1:** No. The candidate paper proposes an anaphora resolution model that is different from the citing intention. Rate: 0.

- **Analyzer 2:** No. The paper is about anaphora resolution, which is not a noun phrase parser or POS parser. Rate: 0.

- **Analyzer 3:** No. This paper came up with a novel module of Lucy system that resolves pronominal anaphora, which has little relevance to the task of noun-phrase parsing. Rate: 1.

Is CAN4 by [139] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5

- **Analyzer 1:** No. The candidate paper proposed a parser based grammar rules which is different from the citing intention. Rate: 0.

- **Analyzer 2:** No. The paper is not about a noun phrase parser or POS parser, it presents a formalism to be used for parsing where the grammar statements are closer to real text sentences and more directly address some notorious parsing problems, especially ambiguity. Rate: 0.

- **Analyzer 3:** No. This paper presents a parsing formalism to be used for parsing where the grammar statements are closer to real text sentences and further address ambiguity problems. It is however concentrated on parsing the structure of sentences rather than noun-phrase. Rate: 3.

Is CAN5 by [140] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5

- **Analyzer 1:** No. The candidate paper aims to analyze the word associations rather than proposing a parsing method. Rate: 0.

- **Analyzer 2:** No. The paper is not about a noun phrase parser or POS parser, the authors began this paper with the psycholinguistic notion of word association norm, and extended that concept toward the information theoretic definition of mutual information. Rate: 0.

- **Analyzer 3:** Yes. This paper proposed an objective measure from the perspective of statistics, for estimating word association norms. The proposed measure estimates word association norms directly from corpora, making it possible to estimate norms for words. Rate: 4.

B.3 Answers for IC3

IC3: "...The debate about which paradigm solves the part-of-speech tagging problem best is not finished. Recent comparisons of approaches that can be trained on corpora (van Halteren et al., 1998; Volk and Schneider,1998) have shown that in most cases statistical approaches(Cutting et al., 1992; Schmid, 1995; =?=) yield better results than finite-state,rule-based,or memory-based taggers(Brill, 1993; Daelemans et al., 1996). They are only surpassed by combinations of different systems,forming a "voting tagger"..." [141]

What is ground truth paper [142] about?

- **Analyzer 1:** The cited paper is about part-of-speech tagger based on statistical methods.

- **Analyzer 2:** This paper presents a statistical model which trains from a corpus annotated with Part-Of-Speech tags and achieves the best results at that time.

- **Analyzer 3:** This paper contrasts a novel statistical model with the state-of-the-art methods on Part-Of-Speech tags problem, demonstrating the superiority of statistical approaches in this task.

Is CAN1 by [136] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5

- **Analyzer 1:** Yes. The candidate paper is suits the citing intention. In addition, this paper is already co-cited at the location. Rate: 5.

- **Analyzer 2:** Yes. The paper proposed a Part-of-Speech Tagger, which is based on a hidden Markov model. In addition, it also shows good results. Rate: 5.

- **Analyzer 3:** Yes. It describes that statistical methods have also been used and provide the capability of resolving ambiguity on the basis of most likely interpretation. Rate: 4.

Is CAN2 by [143] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5

- **Analyzer 1:** No. The candidate paper proposed a noun phrase parser which is different from the citing intention. Rate: 0.

- **Analyzer 2:** No. The paper presents a stochastic part of speech program and noun phrase parser, but it mainly focuses on noun phrase parser, and not show the accuracy of the pos tagger. Rate:2.

- **Analyzer 3:** Probably. This paper introduces a program that finds the assignment of parts of speech to words optimizing the produce of both lexical and contextual probability. From this perspective, the program is based on statistics method. Rate: 3.

Is CAN3 by [137] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5

- **Analyzer 1:** No. The candidate paper aims to propose a rule-based part-of-speech tagger which seems to be unsuitable. Rate: 0.

- **Analyzer 2:** No. The paper describes a rule-based approach to prepositional phrase attachment, which does not focus on solving pos problem. Rate: 0.

- **Analyzer 3:** No. This paper describes a novel rule-based approach to prepositional phrase attachment disambiguation problem. Rate: 2.

Is CAN4 by [144] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5

- **Analyzer 1:** No. The candidate paper aims to propose a rule-based technique to extract linguistic knowledge which is different from the citing intention. Rate: 0.

- **Analyzer 2:** No. The paper describes a simple rule-based approach to capture the linguistic information, which is not corpus-based training approach. In addition, it does not focus on pure part-of-speech tagging method but a method to automated learning of linguistic knowledge. Rate: 1.

- **Analyzer 3:** No. This paper described a simple rule-based approach to automated learning of linguistic knowledge and conducted a case study of this method applied to part-of-speech tagging. However, it did not show any relationship to statistical ways. Rate: 3.

Is CAN5 by [145] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5

- **Analyzer 1:** No. The candidate paper conducted cast studies on part-of-speech tagging that is different from the citing intention. Rate: 0.

- **Analyzer 2:** No. The paper is not about POS tagger methods, it focuses on the evaluation of the algorithms. Rate: 0.

- **Analyzer 3:** No. This paper conducted a case study aiming to evaluate two different methods to anaphoric processing in discourse by comparing the measures of accuracy and coverage. Therefore, it has little relevance to the task of Part-Of-Speech. Rate: 2.

B.4 Answers for IC4

IC4:

"*...In order to solve the problem in definition 3.1, we extend the shift-reduce parsing paradigm applied by =?=, Hermjakob and Mooney (1997), and MarcH (1999). In this extended paradigm, the transfer process starts with an empty Stack and an Input List that contains a sequence of elementary discourse trees edts, one edt for each edu in the tree T_s given as input...*" [146]

What is ground truth paper [142] about?

- **Analyzer 1:** The cited paper is about sentence parser based on decision trees.

- **Analyzer 2:** This paper proposes a statistical parser(SPATTER parser) based on decision-tree learning techniques which constructs a complete parse for every sentence. And the main-paper extend this method.

- **Analyzer 3:** This paper is cited because it constructs the shift-reduce parsing paradigm applied to sentence parsing.

Is CAN1 by [147] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5

- **Analyzer 1:** No. The candidate paper aims to compare the empirical results from tree-based methods which is different to the citing intention. Rate: 0.

- **Analyzer 2:** No. The paper presents theoretical and empirical evidence that the choice of tree representation can make a significant difference to the performance of a PCFG-based parsing system. Rate: 3.

- **Analyzer 3:** No. This paper studies the effect of varying the tree structure representation of PP modification based on PCFG models, from both a theoretical and an empirical point of view. Thus, it has low relevance to the citing place. Rate: 1.

Is CAN2 by [148] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The part-of-speech tagger proposed by the candidate paper is based on sentence chunking which is different from the citing intention. Rate: 0.

- **Analyzer 2:** Yes. The paper is focus on text chunking, and is a transformation-based learning method. It does not use the tree architecture and cannot be applied to solve the problem in definition 3.1 in the main paper. The main-paper can also extend this method. Rate: 4.

- **Analyzer 3:** No. This paper applied the transformation-based learning method to tagging problem. It differs from the intention of citing. Rate: 2.

Is CAN3 by [149] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper introduced an alignment algorithm rather than a sentence parser. Rate: 0.

- **Analyzer 2:** No. This paper proposes an efficient algorithm for bilingual tree alignment, which is different from the tree which constructs a complete parse for every sentence. Rate: 1.

- **Analyzer 3:** No. This paper came up with a novel tree-based alignment algorithm for example-based machine translation. Thus, it is not proper to cite this paper. Rate: 2.

Is CAN4 by [150] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper introduced a statistical parser rather than a parser based on decision tress. Rate: 0.

- **Analyzer 2:** Yes. The parser presented in this paper also utilizes tree architecture and outperforms both the bigram parser and the SPATTER parser, and uses different modeling technology and different information to drive its decisions. The main-paper can also extend this method. Rate: 5.

- **Analyzer 3:** No. This paper presents a statistical parser for natural language. However, The parser does not concentrate on shift-reduce paradigm. Rate: 2.

Is CAN5 by [151] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper aims to study machine translation rather than sentence parser. Rate: 0.

- **Analyzer 2:** No. This paper examined the differences in cohesion between Treebank-style parse trees, trees with flattened verb phrases, and dependency structures. However, it focuses on the MT problem and the approach is hard to be applied in the main-paper. Rate: 3.

- **Analyzer 3:** No. This paper explores how well phrases cohere across two languages helps to improve statistical machine translation. It does not coincide with the intention of citing. Rate: 2.

B.5 Answers for IC5

IC5:

"*...In order to solve the problem in definition 3.1, we extend the shift-reduce parsing paradigm applied by =?=, Hermjakob and Mooney (1997), and MarcH (1999). In this extended paradigm, the transfer process starts with an empty Stack and an Input List that contains a sequence of elementary discourse trees edts, one edt for each edu in the tree T_s given as input...*" [146]

What is ground truth paper [99] about?

- **Analyzer 1:** The cited paper is about machine translation algorithms which is based on sentence alignment for parallel corpora.

- **Analyzer 2:** This paper proposes a method for aligning sentences in a bilingual corpus, which requires parallel corpora.

- **Analyzer 3:** This paper is cited because it describes a system for aligning sentences based on a statistical model in bilingual corpora.

Is CAN1 by [152] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper proposed a word-sense disambiguation methods rather than machine translation. rate: 0.
- **Analyzer 2:** No. The paper focuses on solving word-sense disambiguation problem rather than MT problem, and it does not use parallel corpora. Rate: 0.
- **Analyzer 3:** No. This paper does not involve bilingual corpora. Rate: 2.

Is CAN2 by [153] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper proposed a word-sense disambiguation methods rather than machine translation. Rate: 0.
- **Analyzer 2:** No. The paper focuses on solving word-sense disambiguation problem rather than MT problem, and it uses monolingual corpora rather than parallel corpora. Rate: 0.
- **Analyzer 3:** No. This paper comes up with an unsupervised algorithm that disambiguates word senses in a single corpus. From this perspective, it does not coincide with the citation intention of bilingual corpora. Rate: 2.

Is CAN3 by [154] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper proposed a word-sense disambiguation methods rather than machine translation. rate: 0.
- **Analyzer 2:** No. The paper focuses on solving word-sense disambiguation problem rather than MT problem, similarly, it does not use parallel corpora.Rate: 0.
- **Analyzer 3:** No. Though this paper involves using a bilingual corpora, it solves the problem of word sense disambiguation rather than machine translation (MT). Rate: 3.

Is CAN4 by [100] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** Yes. The candidate paper might be appropriate to be cited, as it describes a word correspondence technique to be applied in machine translation based on parallel corpora which seems to suit the citing purpose. Score: 4.
- **Analyzer 2:** Yes. The paper utilizes parallel corpora, and aims to solve the correspondence problem, which can also be applied in MT system. Rate: 4.
- **Analyzer 3:** Yes. This paper focused on identifying word corresponding in parallel corpora, which is a finer-level problem in machine translation task. Thus, it agrees with the citing intention. Rate: 4.

Is CAN5 by [155] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** Yes, the candidate paper is actually a co-citation at the placeholder. Score: 5.
- **Analyzer 2:** Yes. The paper proposes a method for alignment problem which makes use of parallel corpora. Rate: 4.
- **Analyzer 3:** No. This paper introduces a novel statistical translation model applied to a large database of translated text. It does not coincide with the citation requirement for parallel corpora. Rate: 2.

B.6 Answers for IC6

IC6:

"*...In contrast to [5], non-rigid motion parameters are modeled using the affine motion model, which gives them more flexibility to generate different expressions. A synthesis feedback is used to reduce the error accumulated due to motion estimation in tracking. Our approach is partly motivated by the research conducted by =?=, [5] and [9]. In contrast to [1], while utilizing the muscles contraction parameters as our local deformation model, we are using the optical flow constraint similar to [5]. Our model differs from [5] in two ways...*" [101]

What is ground truth paper [156] about?

- **Analyzer 1:** The cited paper is to propose a facial model based on muscle modeling.
- **Analyzer 2:** This paper proposes a method to the analysis of dynamic facial images and also discusses the drawbacks of FACS, which lacks the expressive power to describe different variations of possible facial expressions.
- **Analyzer 3:** This paper is cited because the idea that considering muscle contraction parameters while recognizing dynamic facial images inspires the authors to use it also as their base model.

Is CAN1 by [157] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper aims to propose a image registration technique rather than a facial model. Rate: 0.
- **Analyzer 2:** No. The paper is not even about the facial expressions, it presents a new image registration technique and also does not talk about FACS. Rate: 0.
- **Analyzer 3:** No. This paper present a novel model utilizing the spatial intensity gradient of the images to solve the image registration problem. It has low relevance to the citing intention. Rate: 1.

Is CAN2 by [158] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** Yes. The candidate paper might be suitable for a citation, as the research is about a facial model based on muscle modeling. Rate: 4.
- **Analyzer 2:** Yes. The paper derives a new, more accurate representation of human facial expressions and call it FACS+. And also talks about the disadvantages of FACS. Rate: 5.
- **Analyzer 3:** No. This paper describe also a model for observing facial motion by using an optimal estimation optical flow method, which has somehow related with the citing intention. However, according to the context, the range of cited papers should be very limited. Rate: 2.

Is CAN3 by [159] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper is of a different purpose that is about hierarchical estimation rather than facial model. Rate: 0.
- **Analyzer 2:** No. The paper presents a new hierarchical motion estimation framework. It does not talk about facial expressions or FACS. Rate: 0.
- **Analyzer 3:** No. This paper describes a hierarchical motion estimation framework for computation of diverse representations of motion information. It should not be cited by the original paper. Rate: 2.

Is CAN4 by [160] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** The candidate paper might be suitable to be cited as an extension to the co-citation [5] of the target citation. The candidate paper is describing a optical flow constraint technique which is similar to [5]. Rate: 3.
- **Analyzer 2:** No. The paper presents a method for treating optical flow information as a hard constraint on the motion of a deformable model. Although it makes use of FACS, it does not discuss its drawbacks. Rate: 1.
- **Analyzer 3:** No. This paper applies a system incorporating flow as constraints to the estimation of face shape and motion using a 3D deformable face model. It might be relevant to the original paper but considering the limited context, it is better not to cite this paper. Rate: 2.

Is CAN5 by [161] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper describe a facial model based on parameterized method which is different from the purpose of the citing paper. Rate: 0.
- **Analyzer 2:** No. This paper proposes local parameterized models of image motion that can cope with the rigid and non-rigid facial motions that are an integral part of human behavior. However, it does not talk about FACS or its drawbacks. Rate: 0.
- **Analyzer 3:** No. This paper introduces a method for recognizing human facial expressions in image sequences and is different to the purpose of utilizing muscle contraction constraints when recognizing dynamic facial images. Rate: 2.

B.7 Answers for IC7

IC7:

"*...Most of the current systems designed to solve this problem use "Facial Action Coding System", FACS [10] for describing non-rigid facial motions. Despite its wide use, FACS has the drawback of lacking the expressive power to describe different variations of possible facial expressions =?=. In this paper, we propose a system that can capture both rigid and non-rigid motions of a face. Our approach uses a realistic parameterized muscle model proposed in [1], which overcomes the limitations of the FACS and provides realistic generation of facial expressions as compared to the other physical models...*" [101]

What is ground truth paper [158] about?

- **Analyzer 1:** The source paper aims to indicate the drawback of one of the previous method FACS.
- **Analyzer 2:** The approach proposed IC7 uses a realistic parameterized muscle model proposed in the paper, which overcomes the limitations of the FACS and provides realistic generation of facial expressions as compared to the other physical models.
- **Analyzer 3:** The model proposed by this paper exposes the limitation of Facial Action Coding System (FACS) that it lacks the expression power to describe different variations of possible facial expressions.

Is CAN1 by [156] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No, the candidate paper did not refer to the drawbacks of FACS. Rate: 0.
- **Analyzer 2:** Yes. This paper also presented a new approach to facial image analysis using a realistic facial model. And also incorporates with a set of anatomically motivated facial muscle actuators. Rate: 4.
- **Analyzer 3:** No. This paper comes up with a model to the analysis of dynamic facial images for resynthesizing facial expressions. It has low relevance to FACS or its drawbacks. Rate: 2.

Is CAN2 by [162] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** Yes. The candidate paper might be suitable to be cited, as it also described the same drawback (lack of expressing facial expressions) in the first paragraph. Rate: 4.
- **Analyzer 2:** No. The paper presents the Automatic Face Analysis (AFA) system, to analyze facial expressions based on both permanent facial features (brows, eyes, mouth) and transient facial features (deepening of facial furrows) in a nearly frontal-view face image sequence. It cannot be applied in IC7 because it does not use realistic parameterized muscle model and focus on designing features. Rate: 0.
- **Analyzer 3:** Yes. This paper developed an automatic face analysis system based on FACS to analyze facial expressions on both permanent- and transient- facial features. As it is a superior system to FACS, it shows the limitation of FACS and thus becoming proper to be cited. Rate: 4.

Is CAN3 by [163] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** The candidate paper might be suitable to be cited, as it also mentioned the same drawback of lacking of "emotion-specified expressions" in the second page. Rate: 3.
- **Analyzer 2:** No. The paper presents the CMU-Pittsburgh AU-Coded Face Expression Image Database, and does not focus on developing facial expression recognition model. Rate: 0.
- **Analyzer 3:** No. This paper published a comprehensive dataset for facial expression analysis and does not show the shortcomings of FACS. So it is better not to cite this paper. Rate: 3.

Is CAN4 by [164] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper did not seem to mention the drawback of FACS. Rate: 0.
- **Analyzer 2:** No. This paper explores and compares approaches to face image representation. And it does not focus on the facial muscle models. Rate: 2.
- **Analyzer 3:** Yes. This paper detailly explores and compares various techniques of FACS and summarizes the merits and drawbacks from different perspectives. Thus, it is proper to cite this paper. Rate: 4.

Is CAN5 by [161] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper did not seem to mention the drawback of FACS. Score: 0.
- **Analyzer 2:** No. This paper explores the use of local parametrized models of image motion for recovering and recognizing the non-rigid and articulated motion of human faces. However, the method cannot be applied in main-7 because it does not use muscle model. Rate: 1.
- **Analyzer 3:** No. This paper proposed local parameterized models of image motion that can cope with the rigid and non-rigid facial motions that are an integral part of human behavior. It does not explicitly or implicitly shows the limitations of (FACS). Rate: 2.

B.8 Answers for IC8

IC8:

"*...On each cluster of speech segments, unsupervised acoustic model adaptation is carried out by exploiting the transcriptions generated by a preliminary decoding step. Gaussian components in the system are adapted using the Maximum Likelihood Linear Regression (MLLR) technique (Leggetter & Woodland, 1995; =?=). A global regression class is considered for adapting only the means and both means and variances. Mean vectors are adapted using a full transformation matrix, while a diagonal transformation matrix is used to adapt variances...*" [104]

What is ground truth paper [165] about?

- **Analyzer 1:** The cited paper is about the technique of maximum likelihood linear regression (MLLR).
- **Analyzer 2:** This paper introduces maximum likelihood trained linear transformations and how it can be applied to an HMM-based speech recognition system.
- **Analyzer 3:** This paper is cited because it uses the Maximum Likelihood Linear Regression (MLLR) technique.

Is CAN1 by [166] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** Yes. The candidate proposed an unconstrained method of maximum likelihood linear regression, however this method is also described in the target citation. The author could additional cite this candidate for a comprehensive manner. Rate: 3
- **Analyzer 2:** Yes. This paper examines the Maximum Likelihood Linear Regression (MLLR) adaptation technique and can be applied to speech recognition. Rate: 5.
- **Analyzer 3:** Yes. This paper examines the Maximum Likelihood Linear Regression (MLLR) technique and extend it for variance transforms. So it's highly possible to cite this paper. Rate: 4.

Is CAN2 by [167] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper describes a MAP methods rather than a maximum likelihood linear regression. Rate: 0.
- **Analyzer 2:** No. Rate:1. The paper proposed a theoretical framework for MAP estimation rather than Maximum Likelihood Linear Regression, and can not be applied to speech recognition easily. Rate: 1.
- **Analyzer 3:** No. This paper presented a framework for maximum a posteriori estimation of hidden Markov models, which is different to the MLLR method. Rate: 2.

Is CAN3 by [105] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper aims to propose a speech recognition based on HMMs, which is different from the citing purpose. Rate: 0.
- **Analyzer 2:** Yes. This paper proposes an approach to HMM training for speaker independent continuous speech recognition that integrates the normalization as part of the continuous density HMM estimation problem. The proposed method is based on a maximum likelihood formulation that aims at separating the two processes, one being the speaker specific variation and the other the phonetically relevant variation of the speech signal. And can be applied to speech recognition. Rate: 4.
- **Analyzer 3:** No. This paper came up with a novel formulation of the speaker-independent training paradigm in HMM parameter estimation process. It has low relevance to the purpose of citation. Rate: 2.

Is CAN4 by [168] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper proposed a HMMs method which is different from the citing purpose. Rate: 0.
- **Analyzer 2:** Yes. This paper introduces a new form of covariance matrix which allows a few full covariance matrices to be shared over many distributions and this technique fits within the standard maximum-likelihood criterion used for training HMMs. This method can be applied to speech recognition. Rate: 4.
- **Analyzer 3:** No. This paper introduced a new form of covariance matrix, to choose a compromise between the large number of parameters of the full-covariance matrix and the poor modeling ability of the diagonal case. Though it also derives the maximum likelihood re-estimation formulae, the main focus deviates from the purpose of citing. Rate: 3.

Is CAN5 by [169] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper aims to propose a speech recognition system rather than proposing a MLLR method. Rate: 0.
- **Analyzer 2:** Yes. This paper also introduces Maximum Likelihood Linear Regression and how it can be applied to speech recognition. Rate: 4.
- **Analyzer 3:** No. This paper mainly described the modification and improvement on HMM model, which differs the intention of citing. Rate: 2.

B.9 Answers for IC9

IC9:

"*...MCVQ falls into the expanding class of unsupervised algorithms known as factorial methods, in which the aim of the learning algorithm is to discover multiple independent causes, or factors, that can well characterize the observed data. Its direct ancestor is Cooperative Vector Quantization [32, =?=, 10], which has a very similar generative model to MCVQ, but lacks the stochastic selection of one VQ per data dimension. Instead, a data vector is generated cooperatively - each VQ selects one vector, and these vectors are summed to produce the data (again using a Gaussian noise model)...*" [170]

What is ground truth paper [171] about?

- **Analyzer 1:** The source paper is citing papers about cooperative vector quantization.
- **Analyzer 2:** The paper discusses factorial stochastic vector quantization and proposes a new objective function for training autoencoders that allows them to discover non-linear, factorial representations.
- **Analyzer 3:** This paper is cited because it came up with a new objective function for training auto encoders that allows to discover non-linear, factorial representations, combining the merits of both Principal Components Analysis (PCA) and Vector Quantization (VQ). VQ is directly related to the citing place.

Is CAN1 by [65] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper proposed the latent dirichlet allcatoion method (LDA) rather than a vector quantization method. Rate: 0.
- **Analyzer 2:** No. The paper introduces Latent Dirichlet Allocation, and does not talk about anything about VQ (although sometimes LDA need to be combined with VQ). Rate: 0.
- **Analyzer 3:** No. This paper introduced the Latent Dirichlet Allocation (LDA) model, a generative probabilistic model for topic modeling of a text corpora. It has low relevance to the VQ process. Rate: 2.

Is CAN2 by [172] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper aims to propose a method for factorizing matrix which is different from the citing purpose. Rate: 0.
- **Analyzer 2:** Yes. Rate:4. The paper mainly analyzes PCA and VQ in detail for learning the optimal non-negative factors from data. Rate: 4.
- **Analyzer 3:** No. This paper focused on the method of matrix factorization, which is somehow related to vector quantization. However, the connection between MF and VQ is not clearly shown. Rate: 3

Is CAN3 by [173] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper introduces a probabilistic model rather than a vector quantization method. Rate: 0.
- **Analyzer 2:** No. The paper proposes a widely applicable generalization of maximum likelihood model fitting by tempered EM and called it Probabilistic Latent Semantics Analysis (PLSA). And does not talk about VQ. Rate: 0.
- **Analyzer 3:** No. This paper introduced the Latent Semantic Analysis (LSA) model for the analysis of two-mode and co-occurrence data. It has little relevance to the VQ process. Rate: 2.

Is CAN4 by [174] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper introduces a probabilistic model rather than a vector quantization method. Rate: 0.
- **Analyzer 2:** No. The paper is nearly the same to REF 3. It proposes a widely applicable generalization of maximum likelihood model fitting by tempered EM and called it Probabilistic Latent Semantics Analysis (PLSA). And does not talk about VQ. Rate: 0.
- **Analyzer 3:** No. This paper presents a novel statistical method for factor analysis of binary and count data which is closely related to a technique known as Latent Semantic Analysis. It does really relate to VQ method. Rate: 2

Is CAN5 by [175] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper is about a model for matching words and pictures, which is different from the citing purpose. Rate: 0.
- **Analyzer 2:** No. This paper explores a variety of latent variable models that can be used for auto-illustration, annotation and correspondence. It just mentions VQ but not explain too much about VQ. Rate: 0.
- **Analyzer 3:** No. This paper explores a variety of latent variable models that can be used for auto-illustration, annotation and correspondence. It differs from the purpose of citation. Rate: 2.

B.10 Answers for IC10

IC10:

"*...Unfortunately CVQ can learn unintuitive global features which include both additive and subtractive effects. A related model, non-negative matrix factorization (NMF) [20, =?=, 24], proposes that each data vector is generated by taking a non-negative linear combination of non-negative basis vectors. Since each basis vector contains only nonnegative values, it is unable to 'subtract away' the effects of other basis vectors it is combined with....*" [170]

What is ground truth paper [172] about?

- **Analyzer 1:** The cited paper is about non-negative matrix factorization (NMF).
- **Analyzer 2:** The paper explains Non-negative matrix factorization (NMF) and how it works.
- **Analyzer 3:** This paper is cited because it focuses on the description of non-negative matrix factorization (NMF) algorithm.

Is CAN1 by [176] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** The candidate paper seems to fit the role by topic, however it is published later than the source paper. Rate: 0.
- **Analyzer 2:** Yes. The paper shows how explicitly incorporating the notion of 'sparseness' improves the found decompositions in NMF. It also explains NMF and how it works. Rate: 4.
- **Analyzer 3:** Yes. This paper has relatively high relevance to the keywords. Also, the limitation of citation is not strict by context. So, it's appropriate to cite this paper. Rate: 4.

Is CAN2 by [65] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper proposed the latent dirchlet allocation (LDA) which is different from the purpose. Rate: 0.
- **Analyzer 2:** No. The paper introduces latent Dirichlet allocation (LDA), and does not explain NMF. Rate: 0.
- **Analyzer 3:** No. This paper describes Latent Dirichlet Allocation (LDA), which is different from the purpose of referencing the NMF algorithm. Rate: 2.

Is CAN3 by [173] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper proposed the probabilistic latent semantic analyses (PLSA) rather than a NMF model. Rate: 0.
- **Analyzer 2:** No. The paper introduces probabilistic latent Dirichlet allocation (LDA), and does not explain NMF. Rate: 0.
- **Analyzer 3:** No. This paper describes Latent Semantic Analysis, which is different from the purpose of referencing the NMF algorithm. Rate: 2.

Is CAN4 by [171] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper proposed a vector quantization method based on Boltzmann distribution which is different to the citing purpose. Rate: 0.
- **Analyzer 2:** No. This paper shows that an autoencoder network can learn factorial codes by using non-equilibrium Helmholtz free energy as an objective function. It does not talk about NMF and how it works. Rate: 0.
- **Analyzer 3:** No. This paper came up with a new objective function for training auto encoders that allows to discover non-linear, factorial representations, combining the merits of both Principal Components Analysis (PCA) and Vector Quantization (VQ). Therefore, the relevance to NMF algorithm is very low. Rate: 1.

Is CAN5 by [177] suitable to be used as a citation for the context? Explain reasons, and rate from 0 to 5.

- **Analyzer 1:** No. The candidate paper proposed a matrix decomposition method based on overcomplete basis rather than a NMF method. Rate: 0.
- **Analyzer 2:** No. This paper presents an algorithm for learning an overcomplete basis by viewing it as probabilistic model of the observed data. But it does not talk about NMF and how it works. Rate: 0.
- **Analyzer 3:** No. This paper presents an algorithm for the generalization of independent component analysis and provides a method for identification when more sources exist than mixtures. It has low relevance to the NMF algorithm. Rate: 2.