

Dual Attention Model for Citation Recommendation with Analyses on Explainability of Attention Mechanisms and Qualitative Experiments

Yang Zhang

Graduate School of Informatics
Kyoto University
yzha5395@alumni.sydney.edu.au

Qiang Ma

Graduate School of Informatics
Kyoto University
qiang@i.kyoto-u.ac.jp

Based on an exponentially increasing number of academic articles, discovering and citing comprehensive and appropriate resources have become non-trivial tasks. Conventional citation recommendation methods suffer from severe information losses. For example, they do not consider the section header of the paper that the author is writing and for which they need to find a citation, the relatedness between the words in the local context (the text span that describes a citation), or the importance of each word from the local context. These shortcomings make such methods insufficient for recommending adequate citations to academic manuscripts. In this study, we propose a novel embedding-based neural network called dual attention model for citation recommendation (DACR) to recommend citations during manuscript preparation. Our method adapts the embedding of three semantic pieces of information: words in the local context, structural contexts,¹ and the section on which the author is working. A neural network model is designed to maximize the similarity between the embedding of the three inputs (local context words, section headers, and structural contexts) and the target citation appearing in the context. The core of the neural network model comprises self-attention and additive attention; the former aims to capture the relatedness between the contextual words and structural context, and the latter aims to learn their importance. Recommendation experiments on real-world datasets demonstrate the effectiveness of the proposed approach. To seek explainability on DACR, particularly the two attention mechanisms, the learned weights from them are investigated to determine how the attention mechanisms interpret “relatedness” and “importance” through the

¹ Cited papers other than the target citation in a citing paper, which are defined in Zhang and Ma (2020a) and Definition 2 in Section 3.1 in this paper.

Submission received: 31 March 2021; revised version received: 19 December 2021; accepted for publication: 4 January 2022.

<https://doi.org/10.1162/coli.a.00438>

learned weights. In addition, qualitative analyses were conducted to testify that DACR could find necessary citations that were not noticed by the authors in the past due to the limitations of the keyword-based searching.

1. Introduction

When writing an academic paper, one of the most frequent questions considered is: “Which paper should I cite at this place?” Based on the considerable number of papers being published, it is impossible for a researcher to read every article that might be relevant to their study. Thus, recommending a handful of useful citations based on the content of a working draft can significantly alleviate the burden of writing a paper. An example of an application scenario is shown in Figure 1.

Currently, many scholars rely on “keyword searches” on search engines, such as Google Scholar² and DBLP.³ However, keyword-based systems often generate unsatisfactory results because query words may not convey adequate information to reflect the context that needs to be supported (Jia and Saule 2017, 2018). Researchers in various fields have proposed various methods to solve this problem. For example, in some studies (McNee et al. 2002; Gori and Pucci 2006; Caragea et al. 2013; Küçüktunç et al. 2013; Jia and Saule 2018), recommendations based on a collection of seed papers were considered, and in others (Alzoghbi et al. 2015; Li et al. 2018), methods were proposed using metadata such as authorship information, titles, abstracts, keyword lists, and publication years. However, when applying such methods to real-world paper-writing tasks, there is a lack of consideration for the local context of a citation within a draft, potentially leading to suboptimal results. Context-based recommendations adopt a more practical concept that generates potential citations for an input context (He et al. 2010, 2011). Based on the context-based methodology, HyperDoc2Vec (Han et al. 2018) uses an embedding framework that further considers embedding with information of citation links between the local context in a citing paper and the content in a cited paper. In our previous study (Zhang and Ma 2020a), we adapted the structural context in addition to the citation link to further improve the recommendation performance. Context-based approaches could be potentially applicable to real-world paper-writing processes.

However, the aforementioned studies fail to consider several essential characteristics of academic papers, thereby limiting their usefulness.

1. Scientific papers tend to follow the established IMRaD format (Introduction, Methods, Results and Discussion, and Conclusions) (Mack 2014), where each section header has a specific purpose. For example, the Introduction section defines the topic of the paper in a broader context, the Methods section includes information on how the results were produced, and the Results and Discussion section presents the results. Therefore, the 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 with those of other works. Therefore, recommendations of suitable

² <https://scholar.google.com/>.

³ <https://dblp.uni-trier.de/>.

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 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 a strong relationship with some citations that appear in the paper. For example, the word "recommendation" has a strong relationship to citations "(Li et al. 2018)" and "(Han et al. 2018)" 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 the paper. For example, in Figure 1, the words "context-based," "recommendations," "citations," and "context" are more informative than the words "adopt," "more," or "generates." Citation "(Han et al. 2018)" may be more essential than "(Jia and Saule 2018)" because the former is related to context-based recommendations, whereas the latter is related to a different approach.

Adequate recommendations of citations for a manuscript should capture the relatedness and importance of words and cited articles in the context that needs citations, as well as the purpose of the section on which the author is currently working. To this end,

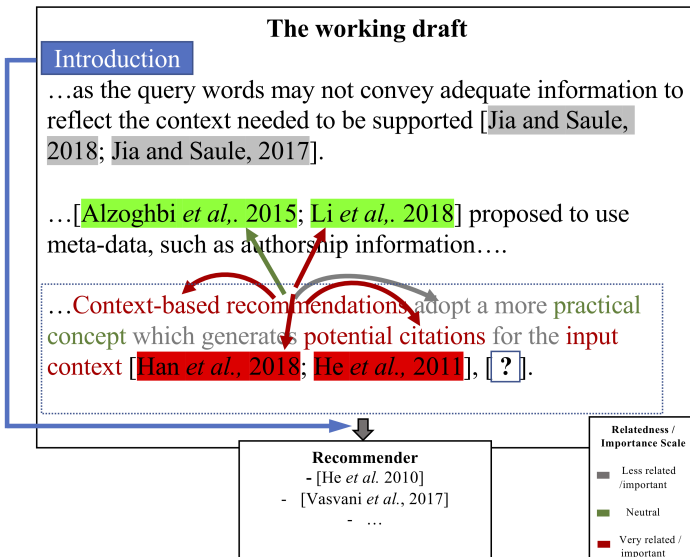


Figure 1 Concept of dual attention model for citation recommendation (DACR). For a context needing citations, DACR makes recommendations by considering the relatedness between contextual words and structural contexts (previously cited papers), the importance of contextual words and structural contexts, and the section where the context appears.

we propose a novel embedding-based neural network called dual attention model for citation recommendation (DACR) to capture the relatedness and importance of words in a context that requires citations and structural contexts in the manuscript, as well as the section for which the author is working. The core of the proposed neural network is composed of two attention mechanisms: 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.

In our previous work (Zhang and Ma 2020b), we introduced the architecture of DACR, experiments on citation recommendations, and ablation studies on the three added features (self-attention, additive attention, and section embedding) to verify its effectiveness. However, it still leaves room for a further study. In this article, we extend the research with the following two additional studies.

First, we aim to parse the internal functions of the adapted attention mechanisms. Attention mechanisms are widely applied recently, such as the the studies from Tang, Srivastava, and Salakhutdinov (2014), Ling et al. (2015), Devlin et al. (2019) and Brunner et al. (2020); however, the internal functions of the learned weights are not yet to be fully understood (Hao et al. 2021). They are generally treated as effective “black boxes.” It is presumed that self-attention captures “relatedness” between words and structural contexts; whereas additive attention extracts “importance” of them in our model. We analyze the patterns of the words with high relatedness and importance scores; and the correlations between them and the semantics of the local context. 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; and (4) differences of the most-emphasized items between self-attention (relatedness) and additive attention (importance). It is found that self-attention assigns high relatedness scores to the items with extreme pairwise similarities (the highest and lowest ones), which includes both topic relevant words and general words (such as “and,” “from,” etc.); whereas the additive attention emphasizes unique words (words with low pairwise similarities) for assigning high importance scores. The analyses are presented in Section 6.

Second, we conduct qualitative analyses to test whether DACR could recommend appropriate citations that could not be found from the keyword-based matching, considering that the keyword-based systems basically find relevant papers by matching the input keywords with the title of the articles. If the title of a potential citation does not contain the input keywords, then the system could not recommend it (Figure 17 demonstrates two scenarios in which the keyword-based search is potentially insufficient). Nevertheless, DACR matches citations based on the semantics of an input context, which does not require the keywords to appear in the title of the potential papers. Because the authors of the papers in our dataset might use keyword-based systems (such as Google Scholar) for writing their papers, there might exist appropriate references that were not found out due to the limitations of the keyword searching. Hence, three annotators with expertise in computer science were hired to inspect whether the recommended citations from DACR should be additionally cited by answering a designed questionnaire (please refer to Section 7 and Appendix B for the details of the questionnaire). Similar to the idea from Bahdanau, Cho, and Bengio (2015), which tests the “completeness” of the translated sentences, we test the completeness of the

in-dataset papers. According to the results, six out of ten selected contexts would require additional citations found by DACR.

The main contributions of this article are as follows:

- First, we provide a neural model, DACR, which leverages the information of the word-level “relatedness” and “importance” of the contextual words (the query context), as well as the sectional purpose of the context, to extract the semantics for recognizing the citing intent of a user. The model is composed of a self-attention (Vaswani et al. 2017) for capturing the word-wise “relatedness,” an additive attention (Wu et al. 2019) for extracting the word-wise “importance,” and a section embedding for learning the sectional purposes, which was testified to be effective compared to the baseline models. Extensive ablation tests were also conducted to test the effectiveness of each of the neural components.
- Second, given that the attention mechanisms were mostly treated as “black boxes” in neural networks, in this work, we would like to conduct qualitative analyses on the learned weights of the attention mechanisms to investigate how they interpret the information of “relatedness” and “importance” through the learned weights of the attention mechanisms. We analyze the patterns of the words with high relatedness and importance scores, the correlations between them and word-wise similarities, and the correlations between them and the semantics of the local context. It is found that self-attention assigns high relatedness scores to the items with extreme pairwise similarities (the highest and lowest ones), which includes both topic relevant words and general words (such as “and,” “from,” etc.), whereas the additive attention emphasizes unique words (words with low pairwise similarities) for assigning high importance scores.
- Third, we conduct qualitative analyses to test whether DACR could recommend additional ground-truth citations. The purpose of these tests are two-fold: (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. Considering that the current keyword-based systems might lead to inaccurate results when the title of the target papers do not contain the input keywords, we would like to test whether our proposed approach, DACR, could provide effective results by utilizing the local context as the query. It is presumed that the authors of the papers from our datasets had used the keyword-based search engines for writing their papers. The experiment could also confirm whether DACR could check the completeness of the citations by qualitatively analyzing the recommended candidate and comparing with the original citation list. We conduct qualitative analyses by hiring three human annotators to parse 10 searching queries, and each comes with 5 searching results from DACR, to confirm whether there exist suitable references in addition to the existing ones. According to the results, six out of ten selected contexts would require additional citations found by DACR.

The remainder of this paper is structured as follows. Section 2 presents a survey of the relevant literature, Section 3 provides the notations and problem definitions, and Section 4 describes the architecture of the proposed model. Section 5 illustrates the experiments, including the experimental results for recommendations, and the results of an ablation study to verify the neural architecture of the model. Section 6 illustrates the analyses on the interpretability of the attention weights. Section 7 presents the qualitative study to test whether DACR could recommend the additional ground-truth citations that the conventional keyword-based systems could not find.

2. Related Work

2.1 Document Embedding

Document embedding refers to the representation of words and documents as continuous vectors. Word2Vec (Mikolov et al. 2013a) was proposed as a shallow neural network for learning word vectors from texts while preserving word similarities. Doc2Vec (Le and Mikolov 2014) is an extension of Word2Vec for embedding documents with content words. However, these two methods generally treat documents as “plain texts,” meaning that when they are applied to scholarly articles, some essential information can be lost (for example, citations and metadata in scientific papers), thereby leading to suboptimal recommendation results. More recent studies have attempted to address this problem. HyperDoc2Vec (Han et al. 2018) is a fine-tuning model for embedding additional citation relations. DocCit2Vec (Zhang and Ma 2020a), proposed in our previous work, considers both structural contexts and citation relations. Nevertheless, some vital information is still not considered, such as the semantics of section headers and the relatedness and importance of words in the context requiring support of citations, which are included in this study.

2.2 Citation Recommendation

Citation recommendation refers to the task of finding relevant documents based on an input query. The query could be a collection of seed papers (McNee et al. 2002; Gori and Pucci 2006; Caragea et al. 2013; Küçükünç et al. 2013; Jia and Saule 2017), and the recommendations are then generated by using collaborative filtering (McNee et al. 2002; Caragea et al. 2013) or PageRank-based methods (Gori and Pucci 2006; Küçükünç et al. 2013; Jia and Saule 2017). Some studies (Alzoghbi et al. 2015; Li et al. 2018) have proposed the use of metadata such as titles, abstracts, keyword lists, and publication years as query information. However, in real-world applications, when providing support for writing manuscripts, these techniques lack practicability. Context-based methods (He et al. 2010, 2011; Han et al. 2018; Zhang and Ma 2020a) use a passage that requires support as a query to determine the most relevant papers, potentially enhancing the paper-writing process. However, such methods may suffer from information loss because they do not consider section headers within papers or the relative importance and relatedness of local context words.

2.3 Attention Mechanisms

Attention mechanism is commonly applied in the field of computer vision (Tang, Srivastava, and Salakhutdinov 2014) to detect important parts of an image and improve the prediction accuracy. This mechanism has also been adopted in recent text-mining

research. For example, in (Ling et al. 2015), Word2Vec was extended with a simple attention mechanism to improve the word classification performance. Google’s BERT algorithm (Devlin et al. 2019) uses multihead attention and provides excellent performance in several natural language processing tasks. The method introduced in Wu et al. (2019) uses self-attention and additive attention to improve the recommendation accuracy for news sources.

2.4 Explainability of Attention Mechanisms

Attention mechanisms have been adapted in multiple neural architectures recently and improved the performances of various tasks, such as pre-training language modeling, BERT (Devlin et al. 2019), or specialized models for specific tasks, such as NRMS (Wu et al. 2019). However, attention mechanisms are generally treated as “black-boxes,” where the internal functions of the learned weights are not fully uncovered. Clark et al. (2019) analyzed the pairwise weights of self-attention layers in BERT (Devlin et al. 2019) to study the pattern of word-to-word correlations, and linguistic correlations. Brunner et al. (2020) 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. Hao et al. (2021) analyzed the most emphasized words from self-attention, and found that some words are likely to be over-emphasized. In this article, we presume that the pairwise 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; and (4) differences of the most-emphasized items between self-attention (relatedness) and additive attention (importance).

3. Preliminary

3.1 Notations and Definitions

Academic papers can be treated as a type of hyperdocument in which citations are equivalent to hyperlinks. Based on paper modeling with citations (Han et al. 2018) and modeling of citations with structural contexts (Zhang and Ma 2020a), we introduce a novel model with citations, structural contexts, and section headers.

Definition 1 (Academic Paper)

Let $w \in W$ represent a word from a vocabulary, W , where $s \in S$ represents a section header from a section header collection, S , and $d \in D$ represents the document ID (paper DOI) from an ID collection, D . The textual information of a paper, H , is represented as a sequence of words, section headers, and IDs of cited documents (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 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, \hat{D} represents the ID of all the cited documents from H , $C \subseteq \hat{W}$ is the local context surrounding d_t , and $s \in \hat{S}$ is the

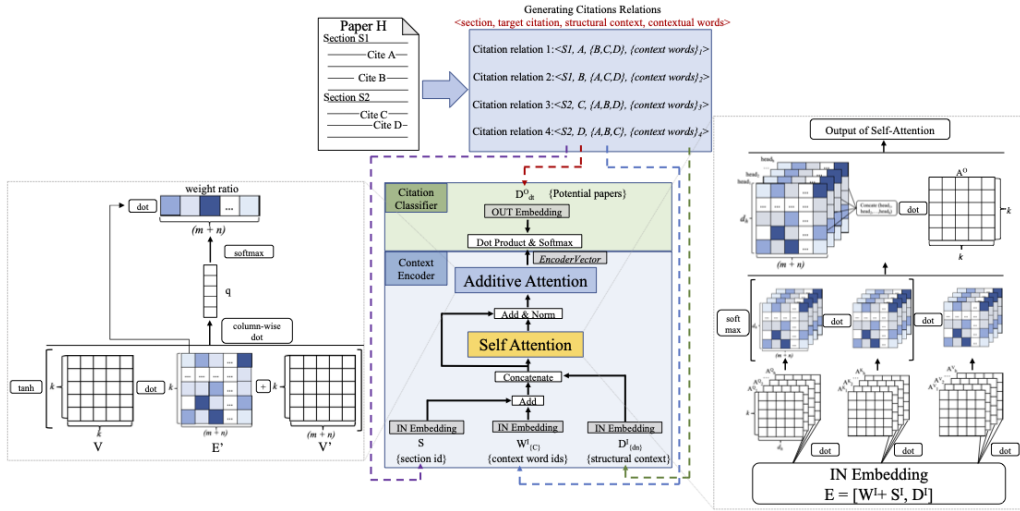


Figure 2 Architecture of DACR.

title of the section in which the contextual words appear. If other citations exist within the same manuscript, then they are defined as **structural contexts** and denoted by D_n , where $\{d_n | d_n \in \hat{D}, d_n \neq d_t\}$.

3.2 Problem Definition

The embedding matrices are denoted as $D \in \mathbb{R}^{k \times |D|}$ for documents, $W \in \mathbb{R}^{k \times |W|}$ for words, and $S \in \mathbb{R}^{k \times |S|}$ for section headers. The i -th column of D , denoted by d_i , is a k -dimensional vector representing document d_i . Additionally, the j -th column of W is a k -dimensional vector for word w_j , and the s -th column of S is a k -dimensional vector for section header s .

The proposed model initializes two embedding matrices (IN and OUT) for documents (i.e., D^I and D^O), a word embedding matrix, W^I , and a section embedding matrix, S^I . A column vector from D^I represents the role of a document as a structural context, and a column vector from D^O represents the role of a document as a citation (the implementation details of the experiment in Section 5.4 explains this in more detail). The word embedding matrix, W^I , and section embedding matrix, S^I , are initialized for all words of the word vocabulary and all sections of the section header collection.

The goal of this model is to optimize the following objective function:

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

4. Dual Attention Model for Citation Recommendation

An overview of the proposed DACR approach is presented in Figure 2. DACR has two main components: a context encoder (Section 4.1) for encoding contextual words,

sections, and structural contexts into a fixed-length vector and a citation classifier (Section 4.2) for predicting the probability of a target citation.

4.1 Context Encoder

The context encoder takes three inputs, namely, context words, sections, and structural contexts, from citation relationships. 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 sublayer (Vaswani et al. 2017) for capturing the relatedness between words and structural contexts, and an additive attention layer (Wu et al. 2019) for recognizing the importance of each word and structural context.

4.1.1 IN Embedding, Add, and Concatenation Layer. The IN embedding layer involves three embedding matrices, \mathbf{D}^I , \mathbf{W}^I , and \mathbf{S}^I , for document collection, word vocabulary, and section header collection, respectively. For a citation relationship defined in Definition 2, that is, $\langle s, d_t, D_n, C \rangle$, the one-hot vectors of structural contexts D_n , context words C , and section headers s are projected with the three embedding matrices, denoted as $\mathbf{D}^I_{\{D_n\}}$, $\mathbf{W}^I_{\{C\}}$, and \mathbf{s}^I_s , respectively. $\mathbf{D}^I_{\{D_n\}}$ is a $k \times |D_n|$ dimensional matrix, where each column indicates the embedding vector of an item from D_n . Likewise, each column of $\mathbf{W}^I_{\{C\}}$ represents the embedding for a word from C . \mathbf{s}^I_s is a k -dimensional embedding vector for the section header s .

The projected section vectors are then added to the word vectors, which is represented as:

$$\mathbf{W}' := [\mathbf{w}_1 + \mathbf{s}^I_s, \mathbf{w}_2 + \mathbf{s}^I_s, \dots, \mathbf{w}_{|C|} + \mathbf{s}^I_s] \tag{2}$$

Then, \mathbf{W}' and $\mathbf{D}^I_{\{D_n\}}$ are concatenated column-wise to form one matrix:

$$\mathbf{E} := [\mathbf{w}'_1, \dots, \mathbf{w}'_{|C|}, \mathbf{d}^I_1, \dots, \mathbf{d}^I_{D_n}] \tag{3}$$

It is expected that contextual words C should reflect two pieces of information: (1) the semantics and (2) the sectional purpose to help determine the citing intent. Hence, in addition to the word embeddings, which indicate the semantics, the section embedding was added to combine the information of the sectional purpose. As a result, the final embedding might reflect the two pieces of information. On the other hand, the structural contexts were based on document embeddings. We hope to use these co-cited documents to infer the other close papers. Hence, they were kept in their original forms.

4.1.2 Self-attention Mechanism with Add&Norm. Self-attention (Vaswani et al. 2017) 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 subspaces 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 $d_h = k/h$, $i \in \{1 \dots h\}$, 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 $(m+n) * (m+n)$, that is, $\text{softmax}\left(\frac{\mathbf{E}^T \mathbf{A}_i^Q \cdot (\mathbf{E}^T \mathbf{A}_i^K)^T}{\sqrt{d_h}}\right)$, 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} , that is, $\mathbf{E}^T \mathbf{A}_i^V$, 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 computational procedure is as follows:

$$\text{SelfAttention}(\mathbf{E}) = \text{Concat}(\mathbf{head}_1, \dots, \mathbf{head}_h) \mathbf{A}^O \quad (4)$$

$$\mathbf{head}_i = \text{softmax}\left(\frac{\mathbf{E}^T \mathbf{A}_i^Q \cdot (\mathbf{E}^T \mathbf{A}_i^K)^T}{\sqrt{d_h}}\right) \cdot (\mathbf{E}^T \mathbf{A}_i^V) \quad (5)$$

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 (Hinton et al. 2012) to avoid overfitting and applied with layer normalization (Ba, Kiros, and Hinton 2016) to facilitate the convergence of the model during training. The final output matrix is denoted as \mathbf{E}' .

4.1.3 Additive Attention Mechanism. The additive attention layer (Wu et al. 2019) is utilized to recognize informative contextual words and structural contexts. It takes matrix \mathbf{E}' from the last layer as input, where each column represents the vector of a word or document. The weight of each item is computed as follows:

$$\mathbf{Weight} = \mathbf{q}^T \cdot \tanh(\mathbf{V} \cdot \mathbf{E}' + \mathbf{V}') \quad (6)$$

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 overfitting.

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

$$\mathbf{EncoderVector} = \mathbf{E}' \cdot \text{softmax}\left(\mathbf{Weight}^T\right) \quad (7)$$

4.2 Citation Classifier

The citation classifier 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 and is defined as follows:

$$\hat{\mathbf{y}} = \mathbf{EncoderVector}^T \cdot \mathbf{D}^O \quad (8)$$

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

$$\mathbf{p} = \text{softmax}(\hat{\mathbf{y}}) \tag{9}$$

4.3 Model Training and Optimization

We adopted a negative sampling training strategy (Mikolov et al. 2013b) to accelerate the training process for DACR. In each iteration, a positive sample (correctly cited paper) and n negative samples are generated. Therefore, the calculated probability vector, \mathbf{p} , is composed of $[p_{\text{positive}}, p_{\text{negative}-1}, p_{\text{negative}-2}, \dots, p_{\text{negative}-n}]$. The loss function computes the negative log-likelihood of the probability of a positive sample as follows:

$$\mathcal{L} = -\log(p_{\text{positive}}) + \sum_{i=1}^n \log(p_{\text{negative}-i}) \tag{10}$$

Stochastic gradient descent (Sutskever et al. 2013) was used to optimize the model.

5. Experiments

We evaluated the recommendation performance of our model and five baseline models on two datasets, namely, DBLP and ACL Anthology (Han et al. 2018). The recall, mean average precision (MAP), mean reciprocal rank (MRR), and normalized discounted cumulative gain (nDCG) were reported for a comparison of the models. The values are summarized in Table 2. Additionally, we proved the effectiveness of adding information about sections, relatedness, and importance, as shown in Figure 4.

5.1 Dataset Overview

The larger dataset, DBLP (Han et al. 2018), contains 649,114 full-paper texts with 2,874,303 citations from the dataset (approximately five citations per paper) in the field of computer science. Originally, as illustrated in Figure 3, the papers in the dataset come with a higher number of citations, out of which five of them come from the dataset—these are the effective ones for training. The citations that are not from the dataset were ignored for training. The ACL Anthology dataset (Han et al. 2018) is smaller and contains 20,408 texts with 108,729 citations from the dataset; however, it has a similar number of citations per paper (approximately five per paper) to the DBLP dataset. We split the datasets into a training dataset to train the document, word, and section

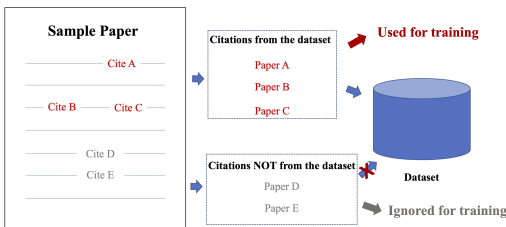


Figure 3 The in-text citations that are the papers from our dataset were recognized for training. The citations that are not from the dataset were ignored.

Table 1
Statistics of the datasets.

Overview of the dataset				Number of sections in the dataset									
	All	Train	Test	Generic Section	Abstract	Background	Introduction	Method	Evaluation	Discussion	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	1,875	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

vectors, and a test dataset with papers containing more than one citation published in the last few years for the recommendation experiments. An experimental overview is provided in Table 1.

5.2 Document Preprocessing

The texts were pre-processed using ParsCit (Councill, Giles, and Kan 2008) to recognize citations and section headers. 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 customized algorithm names. Therefore, we replaced all section headers with fixed generic section headers using ParsLabel (Luong, Nguyen, and Kan 2010). Generic headers from ParsLabel are *abstract*, *background*, *introduction*, *method*, *evaluation*, *discussion*, and *conclusions*. If ParsLabel cannot recognize a section header, we label it as *unknown*. Detailed information for each section header is provided in Table 1.

5.3 Implementation and Settings

DACR was developed using PyTorch 1.2.0 (Paszke et al. 2019). In our experiments, word and document embeddings were pre-trained using two different models: Doc2Vec and DocCit2Vec with default settings, labeled as DACR_{D2V} and DACR_{DC2V}, respectively, in Table 2. For DACR_{D2V}, the citation embeddings were inferred by the trained Doc2Vec model, whereas the word embeddings were directly adopted from Doc2Vec; for DACR_{DC2V}, the word and citation embeddings were directly adopted from the trained DocCit2Vec. The two DACR models were trained with an embedding size of 100, a window size of 50 (also known as the length of the local context, that is, 50 words before and after a citation), a negative sampling value of 1,000, and 100 iterations (default settings in Zhang and Ma [2020a]). The word vectors for generic headers, such as “introduction” and “method,” were selected as pre-trained vectors for the section headers. DACR was implemented with five heads, 100 dimensions for the query vector, and a negative sampling value of 1,000. The stochastic gradient descent optimizer was implemented with a learning rate of 0.0001, batch size of 100, and 100 iterations for the DBLP dataset or 300 iterations for the ACL Anthology dataset. To avoid overfitting, we applied a 20% dropout rate in the two attention layers.

Word2Vec and Doc2Vec were implemented using Gensim 2.3.0 (Řehůřek and Sojka 2010), and HyperDoc2Vec and DocCit2Vec were developed based on Gensim. All baseline models were initialized with an embedding size of 100, a window size of 50, and default values for the remaining parameters.

Table 2

Citation recommendation results (** 0.01 significance level and * 0.05 significance level for paired t test against the best baseline scores for a case).

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
SciBERT (case 1)	4.63	2.13	2.13	2.71	0.01	0.02	0.01	0.03
SciBERT (case 2)	4.63	2.13	2.13	2.71	0.05	0.13	0.05	0.19
SciBERT (case 3)	4.70	2.17	2.17	2.76	0.01	0.02	0.01	0.03
DACR _{D2V} (case 1)	1.04	0.40	0.40	5.50	6.42	2.43	2.43	3.35
DACR _{D2V} (case 2)	1.04	0.40	0.40	5.50	6.64	2.43	2.43	3.36
DACR _{D2V} (case 3)	1.04	0.40	0.40	5.50	6.64	2.43	2.43	3.36
DACR _{DC2V} (case 1)	49.51*	23.58*	23.58*	34.38*	42.43**	22.92**	22.92**	31.64**
DACR _{DC2V} (case 2)	45.39**	22.32**	22.32**	31.98**	40.13**	21.93**	21.93**	30.04**
DACR _{DC2V} (case 3)	42.32**	21.39**	21.39**	30.22**	38.01**	20.84**	20.84**	28.45**

5.4 Recommendation Evaluation

We designed three usage cases to simulate real-world scenarios:

- Case 1: In this case, we assumed the manuscript was approaching its completion phase, meaning the author 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 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, that is, 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. Random selection was implemented using the built-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 is in an early phase of development, where the author has not inserted any citations or all existing citations are invalid. Only context words and section headers were utilized to predict the target citation (no structural contexts were used).

To conduct a recommendation via DACR, an encoder vector was initially inferred using the trained model with inputs of cases 1, 2, and 3, and subsequently 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 was neglected in the inputs.

1. **Citations as words via Word2Vec (W2V).** This method was presented in Berger, McDonough, and Seversky (2017), where all citations were treated as special words. The recommendation of documents was defined as ranking the 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)** (Berger, McDonough, and Seversky 2017). The citations were removed using this method, and the recommendations were made by ranking the IN document vectors via cosine similarity relative to the vector inferred from the learned 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)** (Han et al. 2018). In this method, all context words around a citation were copied into the cited document as Supplementary information. The recommendations were made based on the cosine similarity between the IN document vectors and the inferred vector from the learned model. The vectors were trained using the Doc2Vec PV-DM.
4. **Citations as links via HyperDoc2Vec (HD2V)** (Han et al. 2018). In this method, citations are treated as links pointing to the target documents. The recommendations were made by ranking the OUT document vectors relative to the averaged IN vectors of input contextual words based on dot products. The embedding vectors were pre-trained using Doc2Vec PV-DM using default settings.
5. **Citations as links with structural contexts via DocCit2Vec (DC2V)** (Zhang and Ma 2020a). 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 using Doc2Vec PV-DM with default settings.
6. **Pre-trained model with scientific knowledge via SciBERT** (Beltagy, Lo, and Cohan 2019). In this method, we would like to use the pre-trained SciBERT to retrieve the IN vector by inferring the citing intent from the local context, and the OUT vectors for the citations to infer the content semantics. IN vectors are computed by taking the averaged vector of input contextual words and structural contexts at the last embedding layer from SciBERT. For the citation embeddings (OUT vectors), we retrieve the averaged vectors from the last embedding layer from the content of the papers. However, due to the GPU memory limitation and the large-scaled size of the model, encoding complete contents exceeded our GPU memory.

Hence, we use the title concatenated with abstracts as the “condensed” contents to be encoded as OUT vectors. Recommendations are made by ranking OUT embeddings according to the IN vector via cosine similarity.

Four main conclusions can be drawn from Table 2. First, $\text{DACR}_{\text{DC2V}}$ outperforms all baseline models at the 1% significance level across all evaluation scores for all cases and datasets. This implies that the additional combined information, namely, section headers, relatedness, and importance, is essential for predicting useful citations. The effectiveness of each added information type is presented in Section 5.5.

Second, the 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, $\text{DACR}_{\text{DC2V}}$ is effective for both large-(DBLP) and medium-sized (ACL Anthology) datasets. However, we also realized that a smaller dataset requires more iterations for the model to produce effective results. It is presumed that more iterations of training can compensate for the lack of diversity in the training data.

In addition, DACR_{D2V} and SciBERT produced the lowest performances in the recommendation tests. For the former model, according to the plot of losses from the two models in Figure 5a, we see that the loss of $\text{DACR}_{\text{Doc2Vec}}$ decreases significantly slower than that of $\text{DACR}_{\text{DocCit2Vec}}$, which implies that $\text{DACR}_{\text{Doc2Vec}}$ would require a significantly higher number of iterations to achieve the same performance as $\text{DACR}_{\text{DocCit2Vec}}$. As for SciBERT, we consider that there would need to be a specifically designed training task to fine-tune the pre-trained model for conducting recommendation tasks.

The performance of DACR can be further improved by more accurately recognizing section headers. Moreover, we determined that some labels were incorrectly recognized or could not be recognized by ParsLabel. Therefore, we will work on improving the accuracy of section header recognition in future work.

5.5 Effectiveness of Adding Section Embedding, Relatedness, and Importance

In this section, we explore the effectiveness of adding the following information: section headers, relatedness, and importance. We run three modified DACR models without the corresponding layer; for example, removing the section embedding layer to verify the effectiveness of section information, removing the self-attention layer to determine the relatedness between contextual words and articles, and removing additive attention to demonstrate the importance of context. We present the scores of recall, MAP, MRR, and nDCG at 10 for case 1 on the DBLP dataset for comparison, which are illustrated in Figure 4.

To conduct in-depth analyses, we plot the citation embeddings of the four models in Figure 6 with the top 10 predicted candidate citations from the full DACR. The dimensions of the citation embeddings were reduced by adapting TSNE (Maaten and Hinton 2008) implemented via Scikit-learn (Pedregosa et al. 2011) with default parameters. We aim to inspect the overall distributions of the four models’ citation embeddings and how locations of the top candidates from the full DACR appear in the rest of the distribution plots.

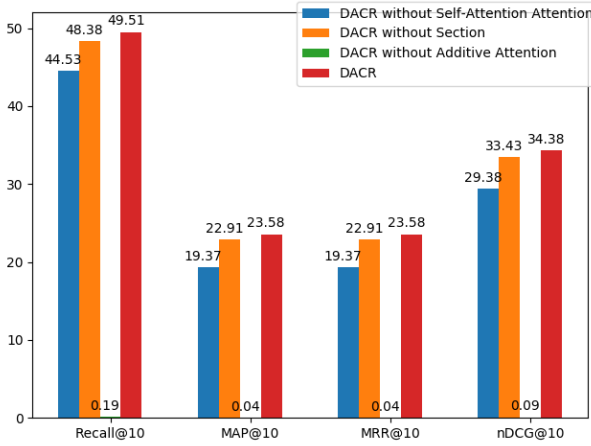
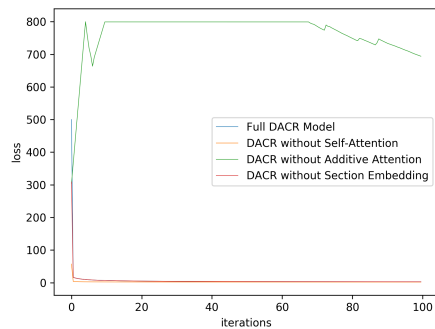
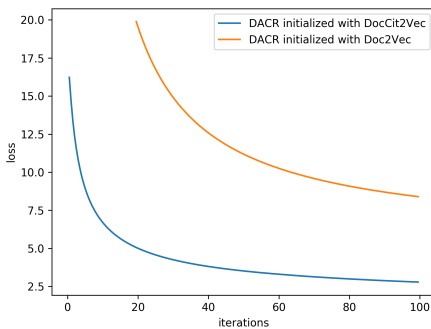


Figure 4
Effectiveness of adding section embedding, relatedness, and importance.

Four points could be drawn from Figure 4 and Figure 6. First, all modified models performed worse than the full model from Figure 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.

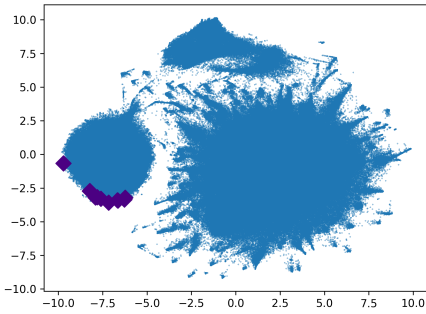
Second, DACR without additive attention performed significantly worse with almost zero scores. We consider the primary reason for the 0-close scores of the model without additive attention is that the losses of the model did not converge without the additive attention layer. According to Figure 5b, the loss curve of DACR without



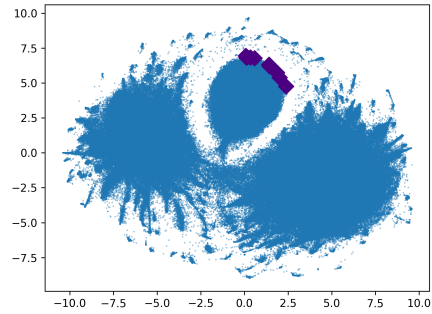
(a) Losses of DACR pre-trained with Doc-Cit2Vec and Doc2Vec on DBLP dataset

(b) Losses of complete DACR, DACR without self-attention, DACR without additive attention, and DACR without section embedding

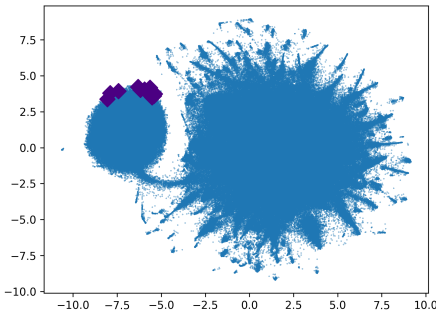
Figure 5
Plots of training losses.



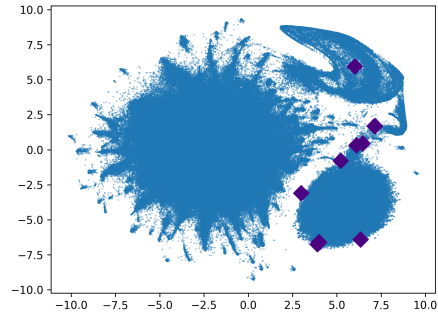
(a) Distribution of dimension-reduced citation embedding from full DACR (diamond dots indicate the top 10 candidates)



(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 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.

additive attention has been raised at the beginning of training 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.

Third, DACR without additive attention did not preserve the word similarities well. Considering Figure 6, we see 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 that DACR

Table 3
Textual information of the sampled contexts.

Dataset	Source paper ref.	Page	Target paper ref.	Context
DBLP	Varadhan et al. (2006)	7	Varadhan et al. (2004)	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. (2000)	7	Lavoie and Rainbow (1997)	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 (lordanskaja 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

without additive attention did not preserve the similarity well compared 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.

6. Analyses of Attention Mechanisms

We analyze the weights of self-attention and additive attention in the model. The self-attention mechanism generates pairwise 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 $\text{softmax}(\frac{\mathbf{E}^T \mathbf{A}_i^Q \cdot (\mathbf{E}^T \mathbf{A}_i^K)^T}{\sqrt{d_h}})$ from Equation 5, 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 $\text{softmax}(\frac{\mathbf{E}^T \mathbf{A}_i^Q \cdot (\mathbf{E}^T \mathbf{A}_i^K)^T}{\sqrt{d_h}})$ with $(m + n) \times (m + n)$ dimensions summarizes all the pairwise 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, namely, $\text{softmax}(\mathbf{Weight})$ from Equation 7, 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, we fetch and plot the weights from the two attention mechanisms from the trained models under the case 1 setting (as designed in Section 5.4) 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, where the “[?=]” marker indicates the location for inserting the target citation. For the DBLP sample, we see 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.”

6.1 Self-attention Analyses

For self-attention, we determined to use the softmaxed pairwise probabilities as the word-to-word scores of “relatednesses.” According to Equation 5, within each head, the projected embedding of the context words and structural contexts ($\mathbf{E}^T \mathbf{A}_i^V$) are multiplied by the pairwise weighted ratios computed by the equation $\text{softmax}(\frac{\mathbf{E}^T \mathbf{A}_i^Q \cdot (\mathbf{E}^T \mathbf{A}_i^K)^T}{\sqrt{d_h}})$, 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 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 pairwise scores of weight ratios from each head (5 heads in total) and the averaged scores for 5 heads are plotted in Figure 7 for the DBLP sample, and Figure 8 for the ACL sample.

To make clear explanations, we use boldface font for the items from the horizontal axis in Figure 7 and Figure 8 (such as “**algorithm**” and “**surface**” at the middle of the x -axis in Figure 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 7(a)).

Three points can be drawn from Figure 7. First, the topic words for inferring the citing intent received high scores. According to Figure 7(f), which pools all the highly scored words, we see that words such as “**grid**,” “**surface**,” and “**topology**” received the highest scores, which are also considered to be highly correlated to the citing intent of the context, that is, “cite a research about the sampling algorithm by preserving the surface topology.” Second, we see that each head focused on a few words in the sentence, and different heads 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, we see that some highly scored words are correlated to almost all 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 shows that the topic words for inferring the citing intent received high scores. In addition, self-attention scores are concentrated on a few words from the model without additive attention (see Figure 9).

As for the model trained by the ACL dataset shown in Figure 8, we notice that the highest weighted items in each head of the ACL, (i.e., Figure 8(a–e), are

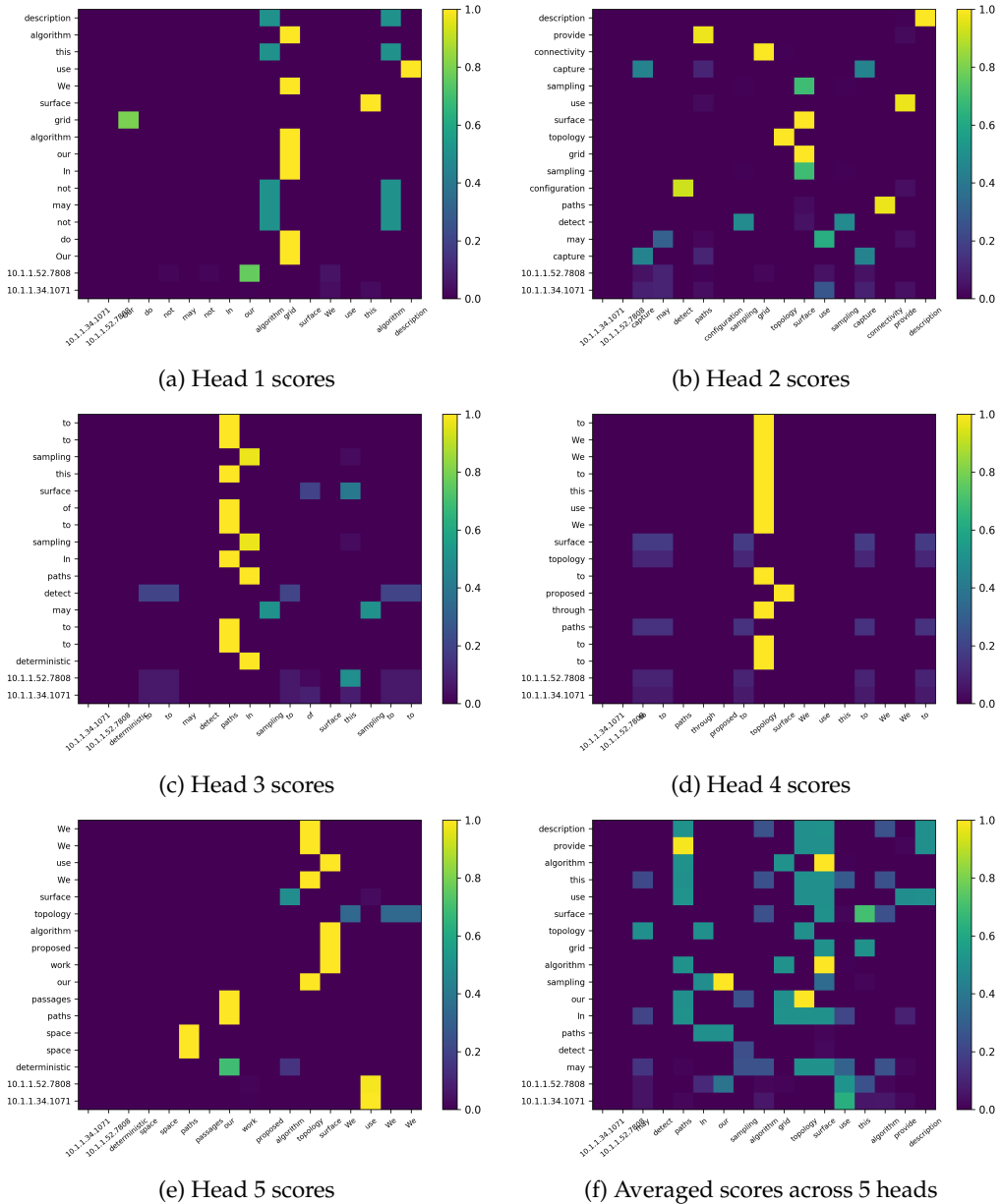


Figure 7 Pairwise self-attention scores (top 15 items) for DBLP sample via complete DACR.

generally lower than the highest weighted items in each head of the DBLP sample (Figure 7(a–e)). Second, we see 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,**” which also received high scores. Generally, the learned scores from the ACL dataset are less concentrated than the scores learned from the DBLP dataset. Although the topic words attracted high weights, more connecting words were also assigned with high weights than the scores learned from the DBLP dataset.

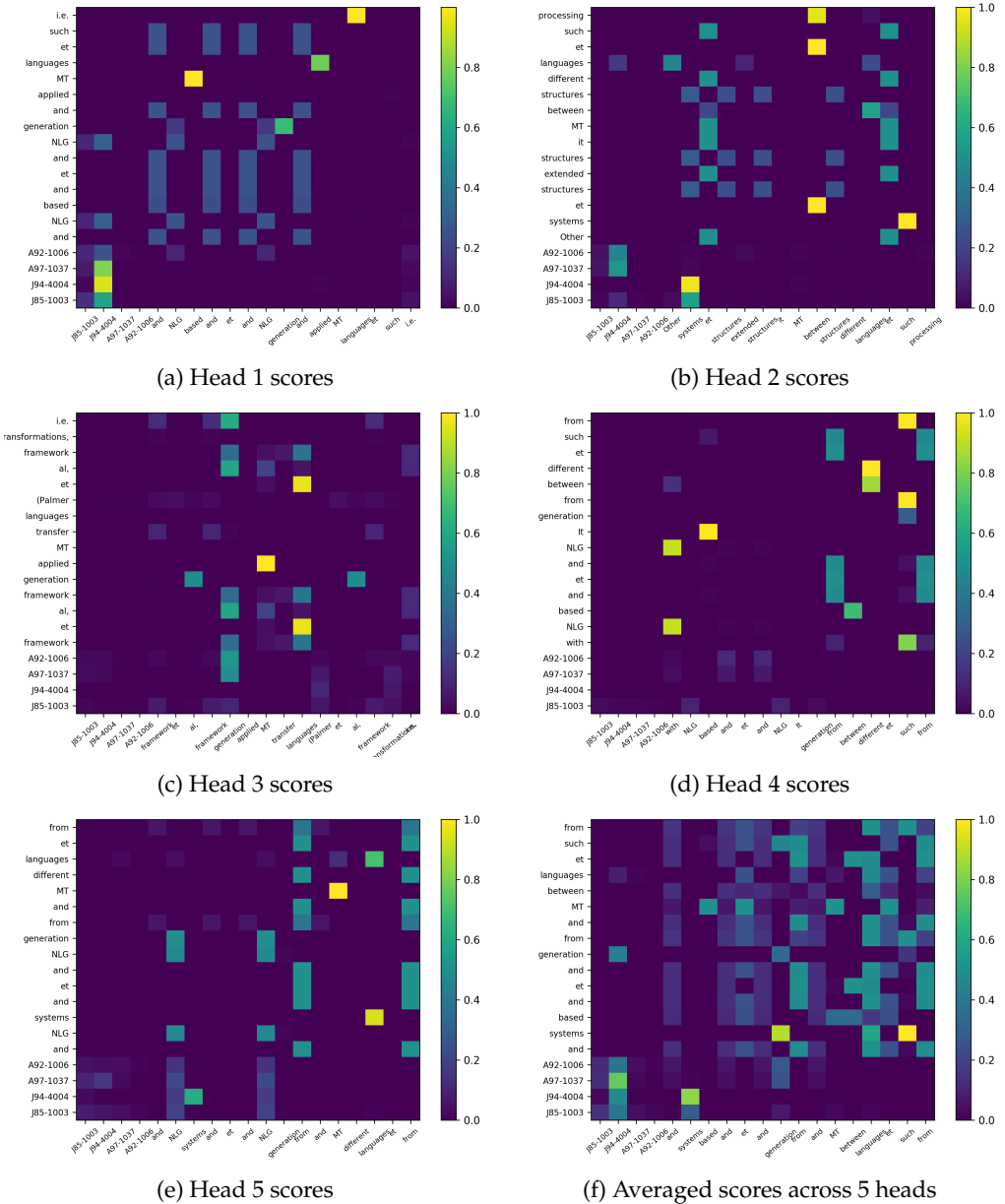


Figure 8 Pairwise self-attention scores (top 15 items) for ACL sample via complete DACR.

6.1.1 Analyses on Correlation between Self-Attention and Similarity Scores. The objective of the word embedding models is that the semantically closed words come with high similarity based on their embedding vectors. In Figure 10(a) and Figure 10(b), we plot the summed self-attention scores along with columns via the complete DACR (orange bars), against the summed pairwise word embedding similarities (blue bars).

It is noticed that some highly scored words for relatedness also yielded high similarity scores (Figure 10). For example, the words “and” from the ACL sample and “We”

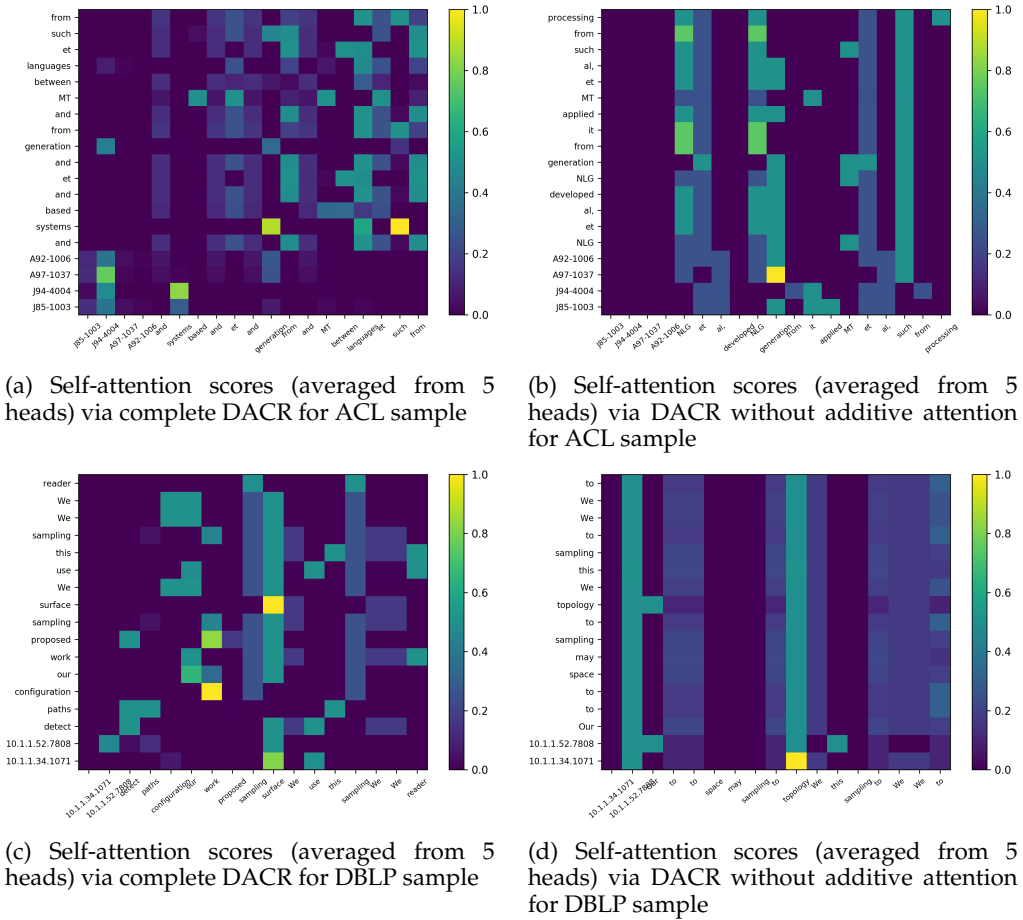


Figure 9 Comparison of self-attention scores (averaged from 5 heads) between the complete DACR and DACR without additive attention.

from the DBLP sample. In addition, some low-scored words on similarity, such as “MT” and “languages,” also received high self-attention scores.

To further confirm the patterns of the learned weights, we provide four additional samples (two samples from the DBLP dataset, and two samples from the ACL dataset) for analyses, which are illustrated in Appendix A. 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 Supplementary sample 1 and 2, and connecting words, such as “we,” “by” from Supplementary sample 1 and 2.

To make in-depth analyses, we conduct quantitative analyses based on 2,000 correctly predicted samples by DACR from each of the datasets to inspect whether the items scored highest for relatedness could also come with extreme (very high or very low) similarity scores. Specifically, in Figure 11(a) and 11(b), we compute 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). The recalls are compared with the probability of random occurrences (number of

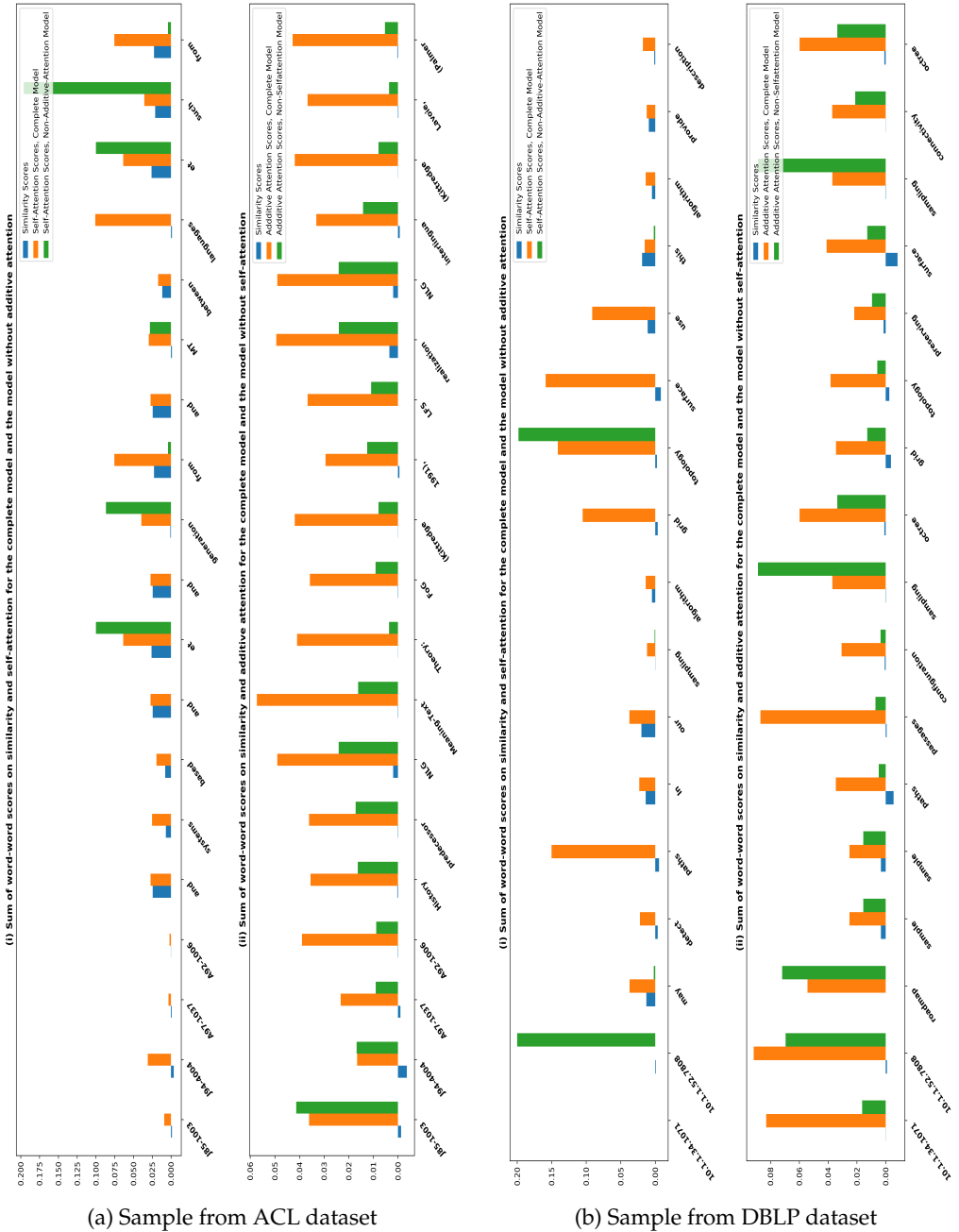
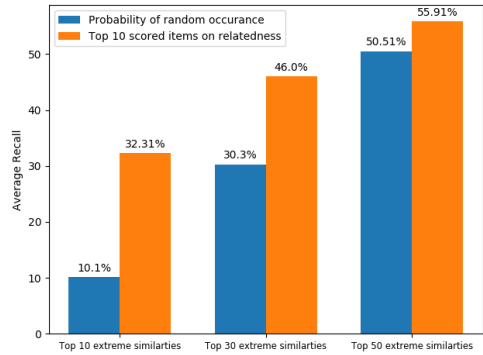
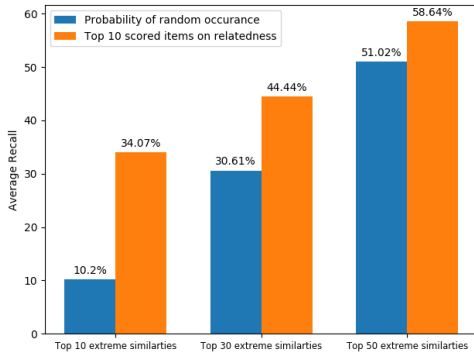


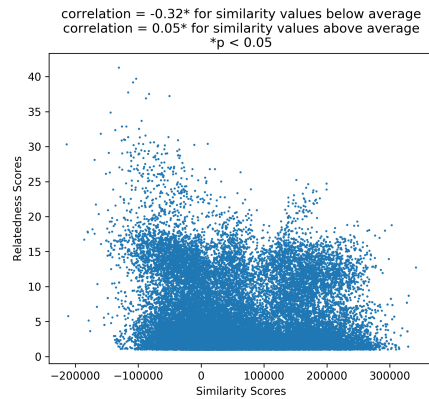
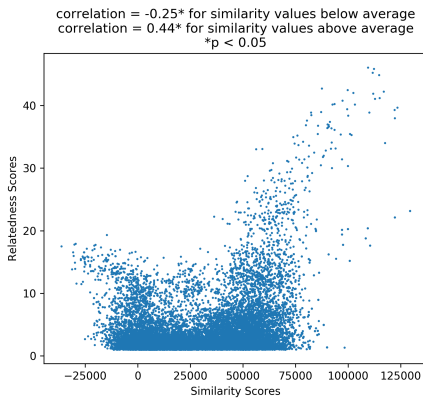
Figure 10 Scores of additive attention (top 15) and summed self-attention against similarities for the samples.

highest items divided by the total number of words and structural contexts appearing in the input context). If the recalls are higher than the natural probabilities, it could imply that the highest-scored items on relatedness are likely to have extreme similarities. Figure 11(a) and 11(b) would have confirmed the positive correlation between high



(a) Recall of items with top 10 highest relatedness scores in top 10/30/50 extreme scores on similarity vs. random probabilities for ACL dataset

(b) Recall of items with top 10 highest relatedness scores in top 10/30/50 extreme scores on similarity vs. random probabilities for DBLP dataset



(c) Correlation between relatedness and similarity scores on ACL dataset

(d) Correlation between relatedness and similarity scores on DBLP dataset

Figure 11

Recall of top 10 highest scored words (or structural contexts) on relatedness in top 10/30/50 extreme scored words (or structural contexts) on similarity against random probabilities (a)(b), and correlation plots between relatedness and similarity scores (c)(d), based on 2,000 correctly prediction samples.

relatedness and extreme similarity scores since 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.

Figure 11(c) and 11(d) further analyze the correlation in detail by plotting the scatters of the words and structural contexts based on their relatedness and similarity scores from the 2,000 samples of each dataset. Each dot represents a word or structural context. We then calculated the Pearson coefficient for the scatters to inspect the trends numerically. It could be concluded that items that come with similarity scores below the average (about 25,000 for the ACL dataset, or 1,000,000 for the DBLP dataset) are negatively correlated with the similarity scores, for a Pearson coefficient of -0.25 (ACL) or -0.32 (DBLP). However, the correlations can also be positive when their

similarity scores are above the average for a coefficient of 0.44 (ACL) or 0.05 (DBLP). The correlations are statistically significant from t tests. Hence, the scatterplots would have confirmed that the relatedness scores are correlated with extreme similarity scores.

6.1.2 Analyses on the Function of Self-Attention in DACR. To inspect the function of the self-attention mechanism in DACR, we compare the complete DACR model with that of the model without the self-attention layer. Specifically, the summed additive attention scores from the complete model and the model without self-attention are plotted in Figure 10 to inspect the effects when removing the self-attention layer.

Comparing the scores from the complete DACR (orange bars in Figure 10(a)(ii) and 10(b)(ii)) with DACR without the self-attention mechanism (green bars), we see that the full models' importance scores are concentrated on a few items, such as the words "sampling," and "roadmap" from the DBLP sample, whereas the scores for the rest of the scores are lowered. Similarly, 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.

Considering that the DACR model without self-attention performed worse than the full model as shown in Section 5.5, it could be concluded that the self-attention mechanism could help the additive attention avoid the importance scores being overweighed, which could improve the model's overall effectiveness.

6.2 Additive Attention Analysis

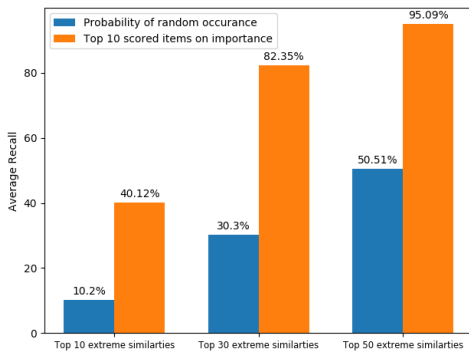
For the additive attention, the importance scores are defined as follows: First, the weight for each embedding is computed according to Equation 6, and then the weights are softmaxed by Equation 7 to output the weight ratios as the final scores for importance. We plot the top 15 importance scores against the sum of pairwise similarities of the words and the structural contexts from the sampled sentences in Figure 10(a)(ii) and Figure 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 10(a)(ii) and 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" from the DBLP sample, which are relevant to the topic of the context. The occurring connecting words from the self-attention mechanism are not assigned with high scores. However, a 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. Adapting specialized pre-process techniques to filter these words would help improve the learned scores on the importance of 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 negatively scored on similarity.

To further confirm the patterns of the learned weights, we provide four additional samples (two samples from the DBLP dataset and two samples from the ACL dataset) for analysis. 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 Supplementary sample 1 and 2, and connecting words, such as "we," "by" from Supplementary 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 the Supplementary

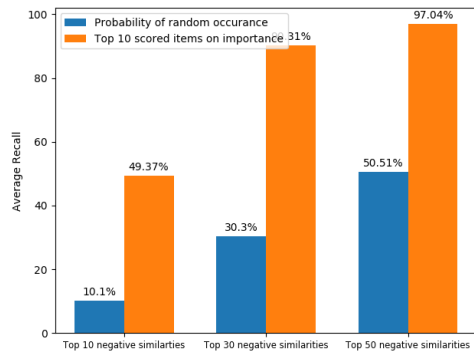
sample 3 and 4, and the unique but irrelevant words, such as “the,” and “you” from Supplementary sample 2, which are the wrong words made from the preprocessing procedure, or “King” from Supplementary sample 4, which is unique, but irrelevant to the topic.

6.2.1 Analyses on Correlation between Additive Attention and Similarity Scores. Similarly to the quantitative analyses on the correlation between self-attention weights and similarity scores, this subsection quantitatively analyzes whether additive attention weights are associated with word similarities.

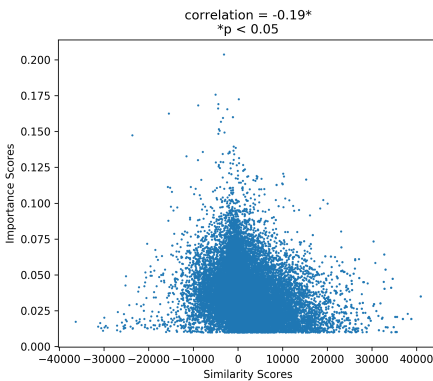
Based on 2,000 correctly predicted samples from DACR from each of the datasets, Figure 12(a) and 12(b) plot the average recall of the top 10 highest scored items on importance in the top 10, 30, and 50 lowest scored items on similarity against the probability of random occurrences. According to the plots, the items with high scores on



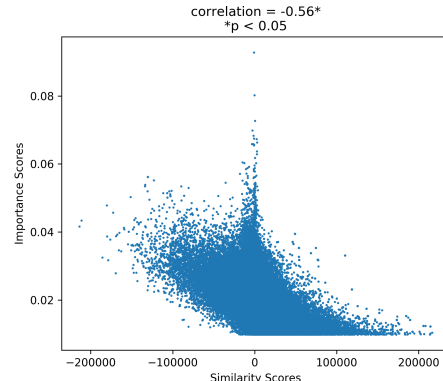
(a) Recall of items with top 10 highest relatedness scores in 10/30/50 lowest scores on similarity vs. random probabilities for ACL dataset



(b) Recall of items with top 10 highest relatedness scores in 10/30/50 lowest scores on similarity vs. random probabilities for DBLP dataset



(c) Correlation between importance and similarity scores on ACL dataset



(d) Correlation between importance and similarity scores on DBLP dataset

Figure 12

Recall of top 10 highest scored words (or structural contexts) on importance in 10/30/50 lowest scored words (or structural contexts) on similarity against random probabilities (a)(b), and correlation plots between importance and similarity scores (c)(d), based on 2,000 correctly prediction samples.

importance demonstrated superior chances of being scored lower on similarity than the random probabilities. It may reveal that the importance scores are negatively correlated with the similarity scores.

To analyze the correlation patterns, we plotted the importance score and similarity score of words and structural contexts in Figure 12(c) and 12(d) based on the 2,000 samples from each dataset. The scatterplots show that the items' importance scores are negatively correlated to the similarity scores by a Pearson coefficient of -0.19 (ACL) and -0.56 (DBLP). The coefficients are statistically significant from t tests. The scatterplots confirm the negative correlation between importance scores and similarity scores.

6.2.2 Analyses on the Function of Additive Attention in DACR. To inspect the function of additive attention, the effects on self-attention weights are investigated by comparing the full DACR model and DACR without additive attention. The pairwised and summed scores for the complete model and the model without self-attention are plotted in Figures 9 and 10 for the inspection.

We see that the self-attention scores are concentrated on a few words from the model without additive attention (see Figure 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. According to Figure 10(a)(ii) and 10(b)(ii), 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. It could be concluded that removing the additive attention cloud leads to biased concentration of self-attention scores on a few items, thus leading to the model's failure, as discussed in Section 5.5.

6.3 Stability Tests on Different Initialization of Attention Weights

In this subsection, we aim to test the stability of the learned weights at self-attention and additive attention. We initialize the weights with three different seeds at the beginning of the training, so that the weights for self-attention and additive attention were different at the starting point. We report the final recommendation scores from the three runs (Table 4a), and the plots of the attention weights (Figure 13 and Figure 16), to inspect whether DACR could produce consistent performance 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 4a, we can see that the differences between the maximum and minimum scores are within 1.50 points, which result in about 3% maximum percentage change (calculated via $\frac{\max - \min}{\min}$). It is observed that DACR generally produced a consistent performance by initializing from different seeds.

Second, the self-attention weights from the three models initialized with different seeds generally extracted similar patterns on “relatedness.” According to Figure 13, the exact scores for each item are different when the model is initialized with a different seed. However, we notice that the high scored items from the three models are both correlated with extreme similarities (Figure 14). In other words, items scored very high and low on wordwise similarity gained high scores from self-attention, which is an identical finding to the analysis in subsection 6.1. In addition, we find that most of the highly scored topics are the same from the three seeded models, such as “**paths**,”

Table 4

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

	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

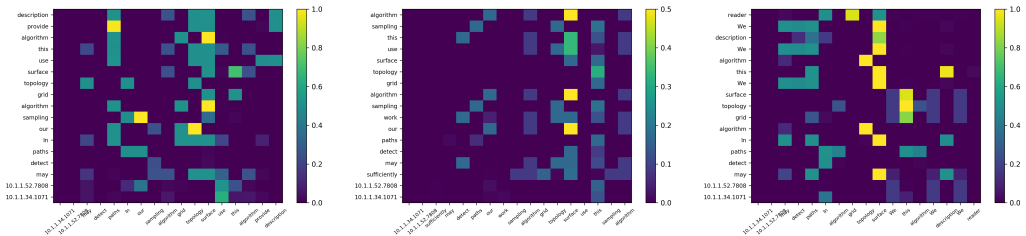
(a) Recommendation scores from DACR models initialized with three seeds

	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



(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 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.

“topology,” and “algorithm,” which occurred in both of the three seeded models. The connecting words, such as “may” and “In” also appeared in both models. According to Table 4b, the model with seed 1 shared 73.33% of the same items with the model with seed 2 in the top 15 scored words from self-attention; the model with seed 2 also shared 73.33% of the same items with the model with seed 3; whereas the model with seed 2 shared the same items with model 3 for the top 15 scored words. It could be concluded that, although the exact scores learned from different seeded models are different, the weights demonstrate the pattern.

Third, the pattern of additive attention weights from models with different seeds also demonstrated even higher consistency. According to Table 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, first, the distribution of scores for each item are similar across

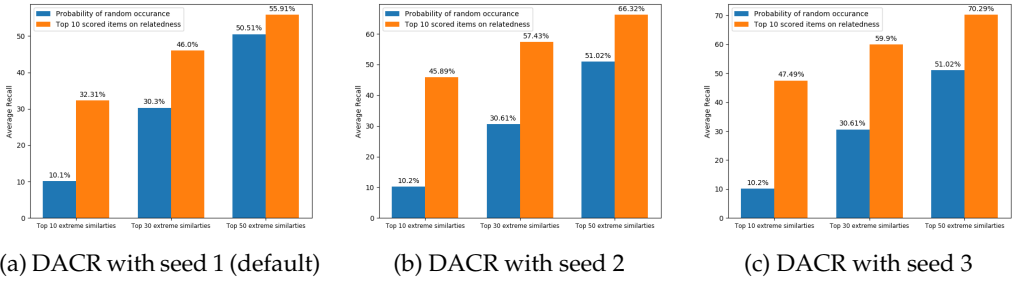


Figure 14 Probabilities of top 10 scored words from self-attention accounted in top 10/30/50 extreme scored words on similarity, from DACR with different seeds.

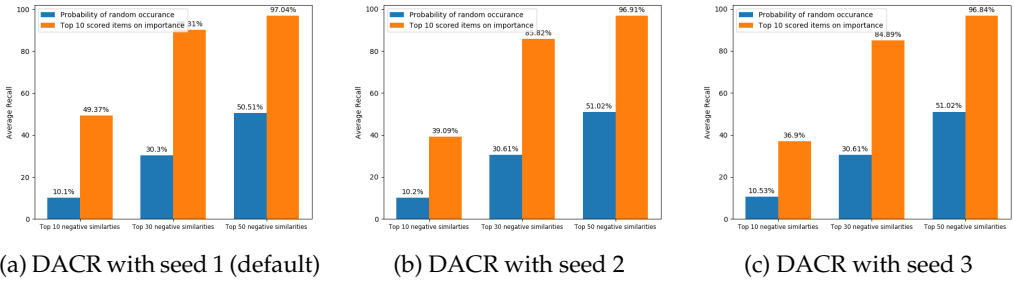


Figure 15 Probabilities of top 10 scored words from additive attention accounted in top 10/30/50 negatively scored words on similarity, from DACR with different seeds.

Figure 16(a)(ii), Figure 16(b)(ii), and Figure 16(c)(ii); second, the scores are negatively correlated to the similarity scores, according to Figure 16(a)(ii), Figure 16(b)(ii), and Figure 16(c)(ii) and Figure 15(a–c).

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

6.4 Summary for Attention Mechanisms

In summary, it could be concluded that the “relatedness” scores captured by the weights of self-attention correlate to the words with extreme pairwise similarities, including both of the topic-related words and connecting words, similarly to the Supplementary examples in Appendix A. The correlation between relatedness scores and extreme similarity scores is quantitatively confirmed by using Pearson correlation analysis, from which the relatedness score of items with similarity scores below the average is negatively correlated with the similarity by a coefficient of -0.25 (ACL) or -0.32 (DBLP). However, the items are positively correlated when the similarity scores are above the average for a coefficient of 0.44 (ACL) or 0.05 (DBLP).

Additive attention emphasizes the unique words (with low pairwise similarities) from the context, mostly topic-related words. However, when the words are not well pre-processed, they could be mistakenly recognized as unique words. From the

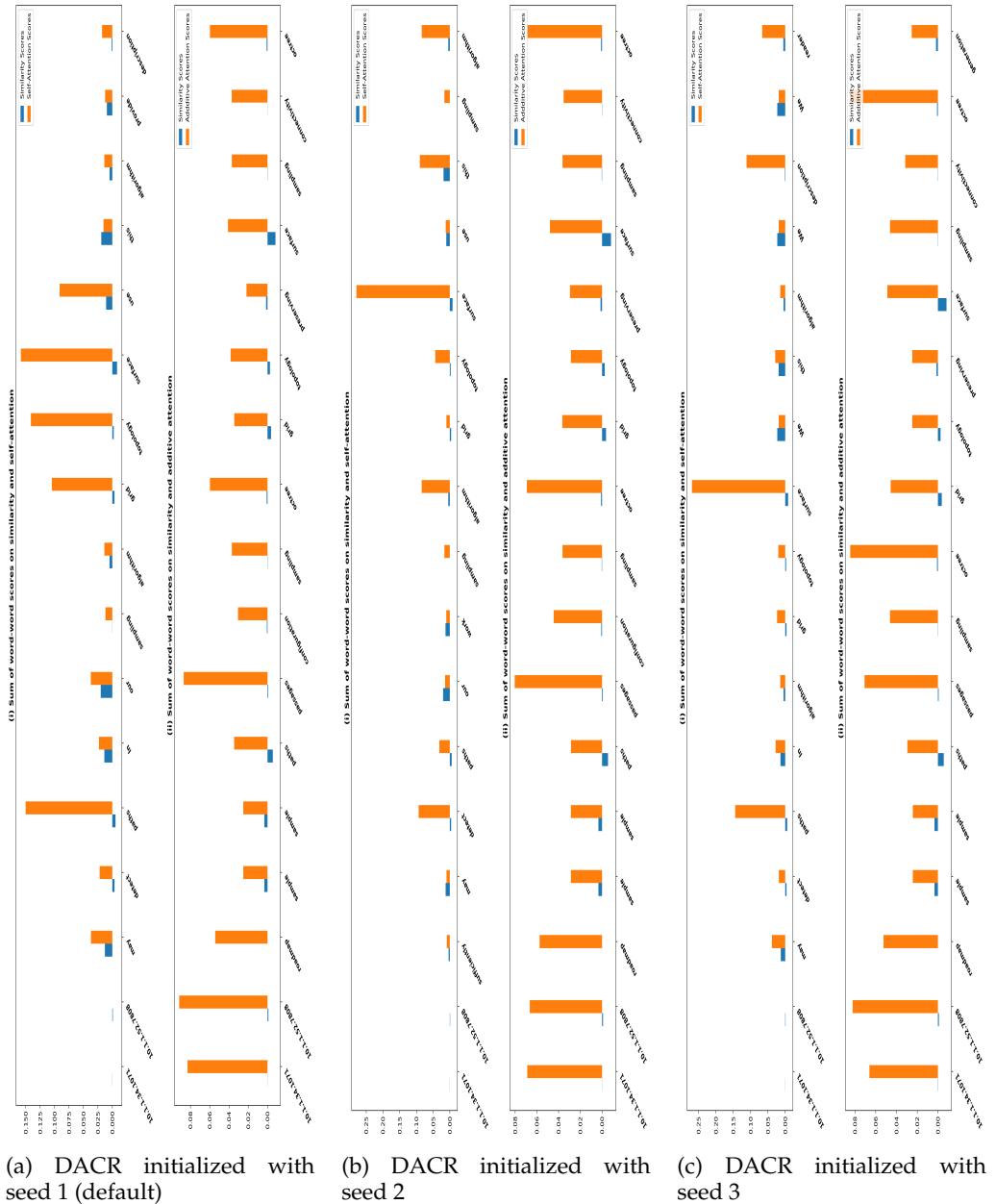


Figure 16 Top 15 scored items from sum of self-attention weights, and additive attention weights, against similarity scores from DACR initialized with different seeds.

quantitative analyses, importance scores are negatively correlated to similarity scores at a coefficient of -0.19 (ACL) or -0.56 (DBLP).

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.

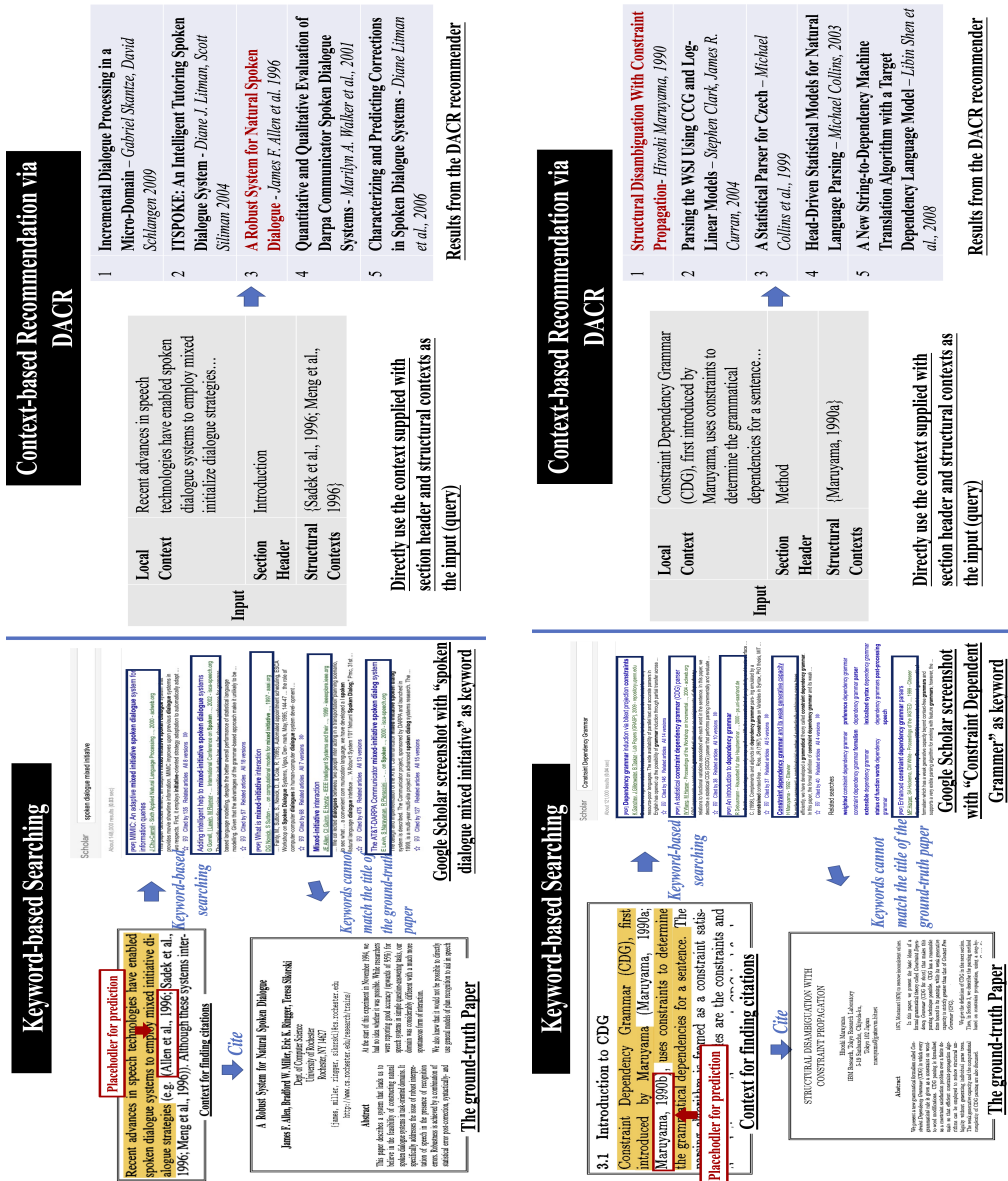
In this study, we have focused on analyzing the correlations between the attention weights and the semantics of citing intents, and word-wise similarities. However, the inner mechanisms of attention layers have not yet been fully uncovered—for example, the theoretical explanations for the reasons that the attention mechanisms could produce these benefits. In future work, we will continue this line of study to seek a deeper understanding of the theoretical basis of the attention mechanisms.

7. Qualitative Analysis for Testing Additional “Ground-truth” Citations

As was discussed in the introduction of this article, scholars are generally relying on “keyword-based” search engines to search for citations. However, due to the oversimplification 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.

We consider that the current keyword-based systems may be limited when applying for two types of scenarios:

1. Scenario 1: In the case where a user would like to find a line of studies in a subfield, target papers are difficult to find 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, and this could result in more accurate recommendations. As the example illustrated in Figure 17a, a sampled piece of context from Chu-Carroll (2000) in the upper left frame of the left side shows that the author would like to cite a line of studies regarding “dialogue system combined with mixed initiative dialogue strategies.” Terms such as “dialogue system,” or “mixed initiative strategies” seem reasonable as the keywords to be used in Google Scholar for searching. However, because these terms are not fully contained in the title of the target paper, titled “A Robust System for Natural Spoken Dialogue” (Allen et al. 1996), Google Scholar could not effectively find it by matching the keywords with its title. On the other hand, our 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 on the right side of Figure 17a.
2. Scenario 2: In the case where 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. In the example illustrated in Figure 17b, the local context selected from Harper et al. (2000) shows that the author would cite the paper that proposed the “Constraint Dependency Grammar” approach. However, the ground-truth paper’s



(a) Scenario 1: Keyword-based searches are difficult to find a line studies regarding a subfield by matching the keywords with the titles; whereas the 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 17 Two scenarios in which the keyword-based search is potentially insufficient.

title, namely, “Structural Disambiguation With Constraint Propagation” (Maruyama 1990), does not contain the terms “Constraint Dependency Grammar.” As a result, Google Scholar could not effectively find the paper in the search results, as shown in the right frame of the left side of Figure 17b. On the other hand, because context-based systems do not fully rely on the terms in a papers’ title, it could effectively trace to the target paper by leveraging the advantage of the semantics of the local context.

We presume 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 there are additional “ground-truth” papers that should be cited but are not successfully identified due to the limitations of the keyword-based systems.

To this end, in this section, we conduct qualitative analysis to analyze the “wrong predictions” from DACR to test whether there exist “additional ground-truth” papers that the authors should cite, but are not successfully found due to the limitations of the searching tools. The tests are made for two purposes: (1) to test the effectiveness of context-based systems on detecting the searching needs of the users; and (2) to test whether the system can help check the completeness of the citations for the reviewers of papers.

Specifically, three analyzers were 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 five candidate references recommended from the trained models (please refer to Table 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 specializing in the field of natural language processing. For the questionnaire, for each input context, the analyzers are required to answer the 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 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 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, as well as evaluate the appropriateness of recommending the structural contexts.

Table 5
Summary of questionnaire.

Input Contexts (IC)	Citing Intent	Candidates No.	Topic of Candidate	Analyzer's Decision (AD)	Relevancy	Input Contexts No.	Citing Intent	Candidates (CAN)	Topic of Candidate	Decisions	Relevancy
IC1	Techniques about sentence alignment	CAN1	Text analysis	AD1: No	Not Relevant	IC6	Facial modeling and drawbacks	CAN1	Image registration	AD1: No	Not Relevant
				AD2: No						AD2: No	
				AD3: No						AD3: No	
		CAN2	Machine translation or parameters estimation	AD1: Yes	Weakly Relevant			CAN2	Facial modeling	AD1: Yes	Strongly Relevant
				AD2: No						AD2: Yes	
AD3: No	AD3: No										
CAN3	English-Chinese alignment	AD1: No	Weakly Relevant	CAN3	Hierarchical motion estimation	AD1: No	Not Relevant				
		AD2: No				AD2: No					
AD3: Yes	AD3: Yes										
CAN4	Word correspondence algorithm	AD1: No	Weakly Relevant	CAN4	Optical flow constraint	AD1: Yes	Weakly Relevant				
		AD2: No				AD2: No					
		AD3: Yes				AD3: No					
CAN5	Noun phrase alignment	AD1: No	Weakly Relevant	CAN5	Facial model	AD1: No	Not Relevant				
		AD2: Yes				AD2: No					
		AD3: No				AD3: No					
IC2	Noun phrase parsing	CAN1	Part-of-speech tagger	AD1: Yes	Weakly Relevant	IC7	Limitation of FACS approach	CAN1	Facial modeling	AD1: No	Weakly Relevant
				AD2: No						AD2: Yes	
				AD3: No						AD3: No	
		CAN2	Rule-based parser	AD1: No	Not Relevant			CAN2	Facial modeling and limitation of FACS	AD1: Yes	Strongly Relevant
				AD2: No						AD2: No	
AD3: No	AD3: Yes										
CAN3	Anaphora resolution	AD1: No	Not Relevant	CAN3	Facial modeling	AD1: Yes	Weakly Relevant				
		AD2: No				AD2: No					
AD3: No	AD3: No										
CAN4	Formalism for parsing grammar statements	AD1: No	Not Relevant	CAN4	Analysis of facial models	AD1: No	Weakly Relevant				
		AD2: No				AD2: No					
AD3: No	AD3: Yes										
CAN5	Analysis of word association norm	AD1: No	Weakly Relevant	CAN5	Image motion	AD1: No	Not Relevant				
		AD2: No				AD2: No					
AD3: Yes	AD3: No										
IC3	Part-of-speech tagger	CAN1	Part-of-speech tagger	AD1: Yes	Strongly Relevant	IC8	Maximum likelihood linear regression (MLLR)	CAN1	MLLR	AD1: Yes	Strongly Relevant
				AD2: Yes						AD2: Yes	
				AD3: Yes						AD3: Yes	
		CAN2	Noun phrase tagger	AD1: No	Weakly Relevant			CAN2	Maximum a posteriori estimation	AD1: No	Not Relevant
				AD2: No						AD2: No	
AD3: Yes	AD3: No										
CAN3	Rule-based parser	AD1: No	Not Relevant	CAN3	Hidden Markov model	AD1: No	Weakly Relevant				
		AD2: No				AD2: Yes					
AD3: No	AD3: No										
CAN4	Rule-based extraction of linguistic knowledge	AD1: No	Not Relevant	CAN4	New covariance matrix	AD1: No	Weakly Relevant				
		AD2: No				AD2: Yes					
AD3: No	AD3: No										
CAN5	Case study of part-of-speech taggers	AD1: No	Not Relevant	CAN5	Speech recognition	AD1: No	Weakly Relevant				
		AD2: No				AD2: Yes					
AD3: No	AD3: No										

Table 5
 Continued

Input Contexts (IC)	Citing Intent	Candidates No.	Topic of Candidate	Analyzer's Decision (AD)	Relevancy	Input Contexts No.	Citing Intent	Candidates (CAN)	Topic of Candidate	Decisions	Relevancy
IC4	Sentence parser	CAN1	Theoretical and empirical study on tree representation	AD1: No	Not Relevant	IC9	Vector quantization	CAN1	Latent dirichlet allocation (LDA)	AD1: No	Not Relevant
				AD2: No						AD2: No	
				AD3: No						AD3: No	
		CAN2	Text-chunking	AD1: No	Weakly Relevant			CAN2	Matrix factorization	AD1: No	Weakly Relevant
				AD2: Yes						AD2: Yes	
				AD3: No						AD3: No	
		CAN3	Bilingual alignment	AD1: No	Not Relevant			CAN3	Probabilistic latent semantic analysis (PLSA)	AD1: No	Not Relevant
				AD2: No						AD2: No	
				AD3: No						AD3: No	
		CAN4	Statistical parser	AD1: No	Weakly Relevant			CAN4	PLSA	AD1: No	Not Relevant
AD2: Yes	AD2: No										
AD3: No	AD3: No										
CAN5	Machine translation	AD1: No	Not Relevant	CAN5	Latent variable models	AD1: No	Not Relevant				
		AD2: No				AD2: No					
		AD3: No				AD3: No					
IC5	Bilingual alignment	CAN1	Word-sense disambiguation	AD1: No	Not Relevant	IC10	Non-negative matrix factorization (NMF)	CAN1	NMF	AD1: No	Strongly Relevant
				AD2: No						AD2: Yes	
				AD3: No						AD3: Yes	
		CAN2	Word-sense disambiguation	AD1: No	Not Relevant			CAN2	LDA	AD1: No	Not Relevant
				AD2: No						AD2: No	
				AD3: No						AD3: No	
		CAN3	Word-sense disambiguation	AD1: No	Not Relevant			CAN3	PLSA	AD1: No	Not Relevant
				AD2: No						AD2: No	
				AD3: No						AD3: No	
		CAN4	Bilingual word coding	AD1: Yes	Strongly Relevant			CAN4	Auto-encoder with new training technique	AD1: No	Not Relevant
AD2: Yes	AD2: No										
AD3: Yes	AD3: No										
CAN5	Bilingual alignment	AD1: Yes	Strongly Relevant	CAN5	Matrix decomposition on an over-complete basis	AD1: No	Not Relevant				
		AD2: Yes				AD2: No					
		AD3: No				AD3: No					

7.1 Examination of “Strongly Relevant” Recommendations

To specially 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 fourth 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 Pantel and Lin (2000) where the “=?=” marker indicates the placeholder for 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 Gale and Church (1991) 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 can be seen that all of the analyzers correctly detected the citing intent of the input context, as well as 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 Yilmaz, Shafique, and Shah (2002) 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, we see 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, they 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.

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, namely, 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 (Chen and Nie 2000):

“...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 (Brown et al. 1993) 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 (Brown et al. 1993):

"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 follows:

- **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 (Brugnara et al. 2000):

"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 (Anastasakos et al. 1996) 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 (Anastasakos et al. 1996):

"...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 (Brugnara et al. 2000), according to Equation 3 from CAN3 (Anastasakos et al. 1996). 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 follows:

- **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.

We can conclude that although the first and third analyzers detected the main purpose of CAN3 to propose the HMM-based approach, they did not realize the relevancy 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, we argue 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.

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 one 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 adoption of structural contexts.

According to Table 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 are 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 recommendations 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, these citations might be helpful to provide Supplementary sources for studying the field in a broad view as they are also relevant to some aspects of the field. We believe 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.

8. Conclusions and Future Work

This study proposed a citation recommendation model with dual attention mechanisms. This model aims to simplify real-world paper-writing tasks by alleviating information loss in existing methods. Our model 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. The core of the proposed model is composed of two attention mechanisms: self-attention for capturing relatedness and additive attention for learning importance. Extensive experiments demonstrated the effectiveness of the proposed model in designed scenarios intended to mimic real-world scenarios, as well as the efficiency of the proposed neural network.

In addition, we conducted an analysis of correlations between the attention weights and the semantics regarding semantics on citing intents, and word-wise similarities. We 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. However, the inner mechanisms of attention layers are not yet fully uncovered—for example, the theoretical explanations on the reasons that the attention mechanisms could produce these benefits.

Furthermore, we qualitatively analyzed the candidates recommended by DACR for selected samples to evaluate whether there exist unnoticed but appropriate citations for the authors. We believe 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.

In future work, first, we will attempt to improve the accuracy of recognizing section headers to improve the usability and performance of the algorithm. Second, we will include additional paper-related information in the model, such as word positions. Third, we will explore more sophisticated neural network architectures to improve the accuracy and reduce the training time of the model. Fourth, we will continue to seek a deeper understanding of the theoretical level of the attention mechanisms. Last but not least, in the next stage, we will also focus on developing a prototype for citation recommendations to help find paper candidates during the writing of papers and reviewing the completeness of citations by optimizing the DACR model and combining it with potentially related approaches.

Appendix A: Supplementary Samples

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

Considering the first sample in Table A.1, it could be concluded 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 A.1, the topic-related words, 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 pairwise similarities they have received. Additive attention in Figure A.3 assigned higher weights to the unique words (low pairwise similarities) of the context; most of them are relevant to the general topic of the context, such as “Aligning,” “parallel,” and “cognateness.”

Table A.1
Textual information of Supplementary samples.

No.	Dataset	Source paper ref.	Page	Target paper ref.	Context
1	ACL	Chen and Nie (2000)	4	Chen (1993)	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 RScheisen, 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 (2000)	3	Grishman, Macleod, and Meyers (1994)	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 COMPLEX 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 COMPLEX make it possible to determine what these

Downloaded from http://direct.mit.edu/colli/article-pdf/48/2/403/2029119/colli_a_00438.pdf by guest on 08 September 2023

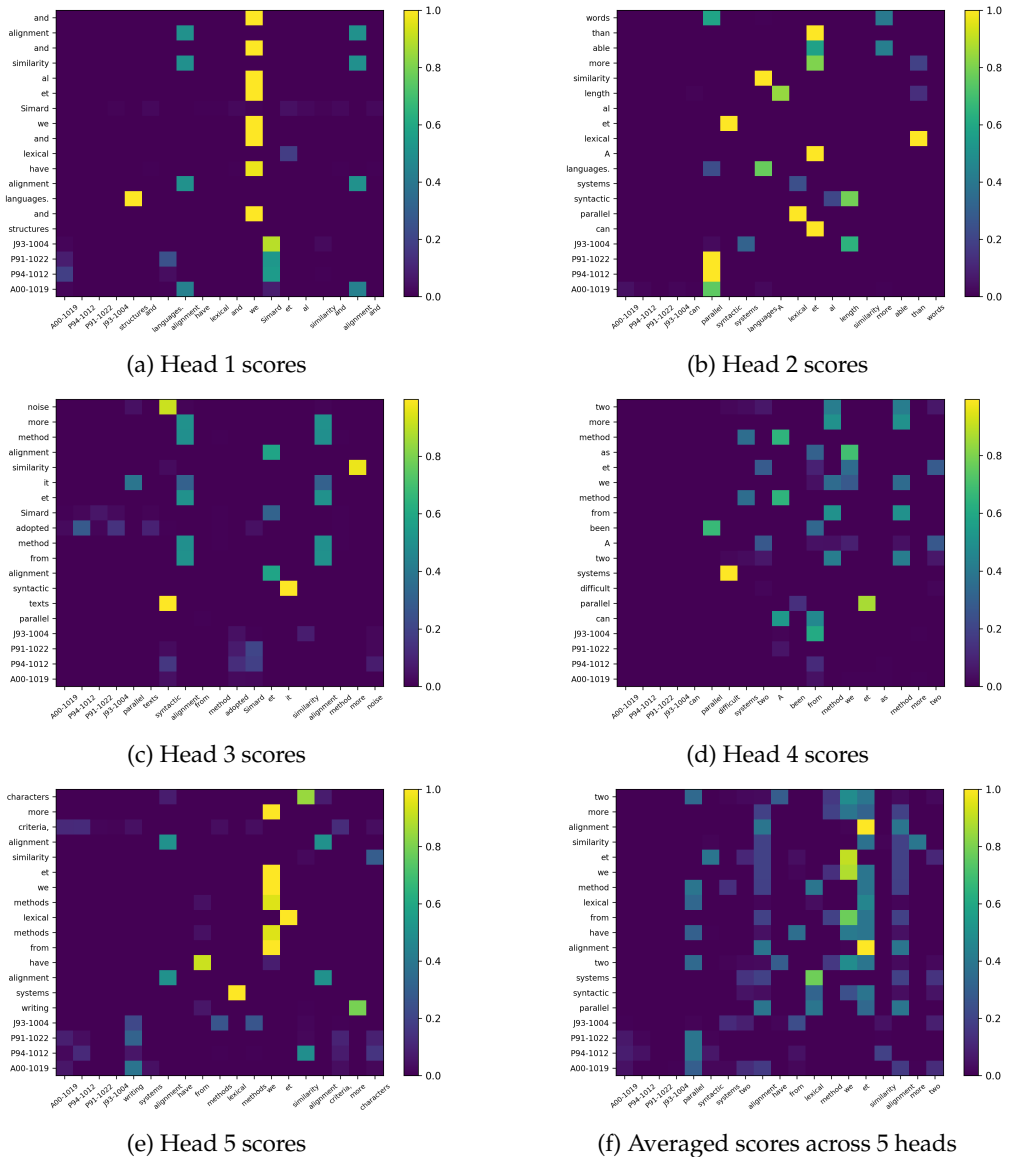


Figure A.1 Pairwise self-attention scores (top 15 items) for Supplementary sample 1 via complete DACR.

However, some words that are directly relevant to the citing intent (such as “lexical” from self-attention) are not recognized by additive attention. Note also that the words detected by additive attention mostly appear in the content of the target paper.

For the second sample in Table A.1, we see that the author is citing the paper that proposed the COMPLEX grammar and lexicon, and providing a description of its contribution of it (i.e., assigning syntactic functional roles to constituents). Similar to sample 1, self-attention has recognized topic-related words (Figure A.1), such as “syntactic,” “grammar,” and “lexicon”; moreover, the connecting words with high word similarities, such as “by,” and “and.” For the unique words that the additive

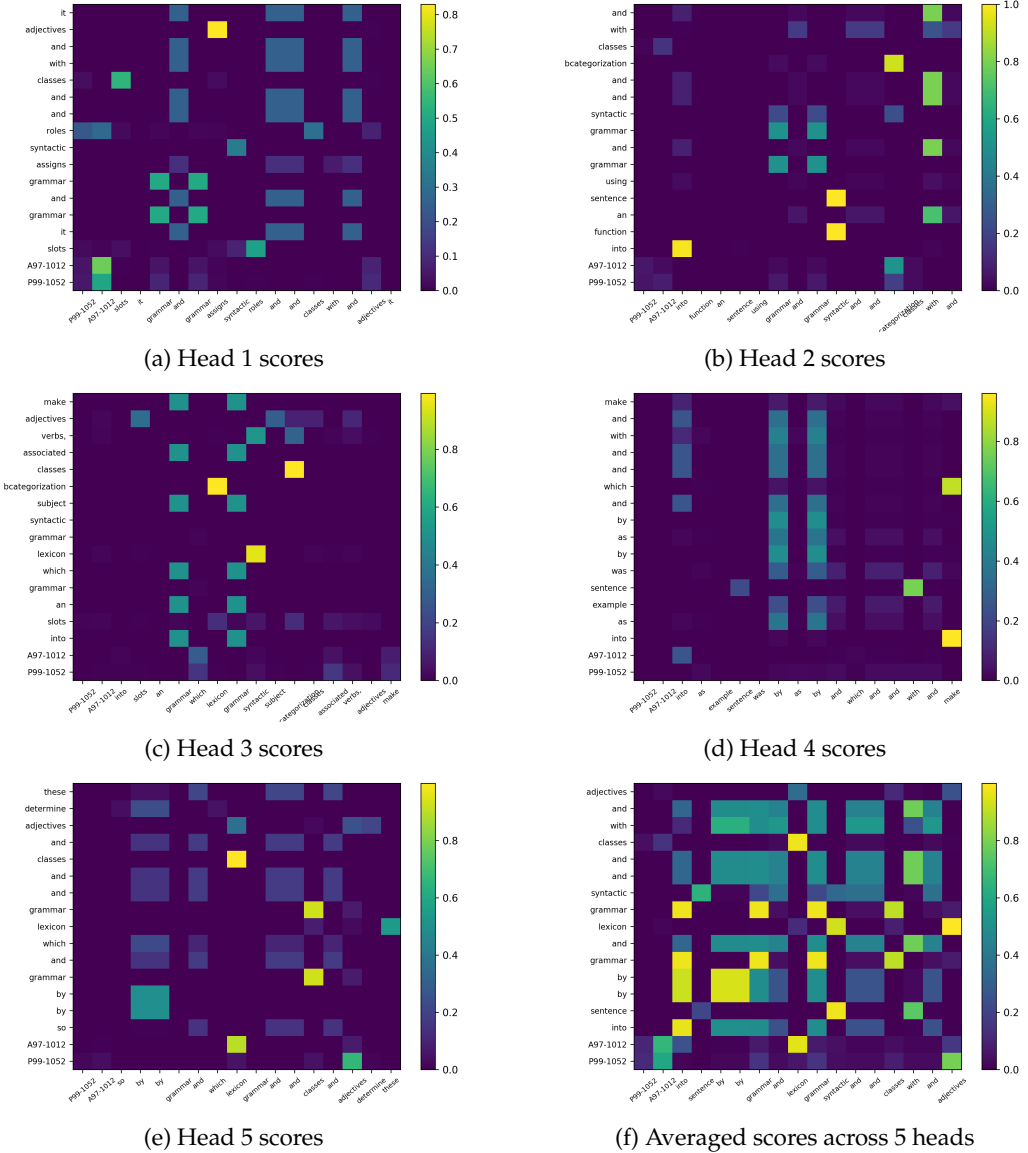


Figure A.2 Pairwise self-attention scores (top 15 items) for Supplementary sample 2 via complete DACR.

attention has detected (Figure A.3), “COMPLEX” and “functional” 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 6, except that the additive attention over-emphasized some wrong words. We could conclude that the top-weighted

words from additive attention could be sometimes irrelevant to the authors citing intent, although they are all unique (i.e., low pairwise similarity).

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

Considering the third sample in Table A.2, we could deduce 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 that are relevant to the citing intent, such as “**adaptive**,” “**dimension**,” and “**clock**,” but also the connecting words with high pairwise word similarities, such

Table A.2
Textual information of Supplementary sample 3 & 4.

No.	Dataset	Source paper ref.	Page	Target paper ref.	Context
3	DBLP	Kumar and Najjar (1999)	1	Duato (1993)	corresponding clock cycles, can be significantly lower than adaptive routers This difference in router delays is due to two main reasons: number of VCs and output (OP) channel selection. Two VCs are sufficient to avoid deadlock in dimension ordered routing [6]; while adaptive routing (as described in [6]) 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. (2004)	3	Stam and Fiume (1993)	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 [6] 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

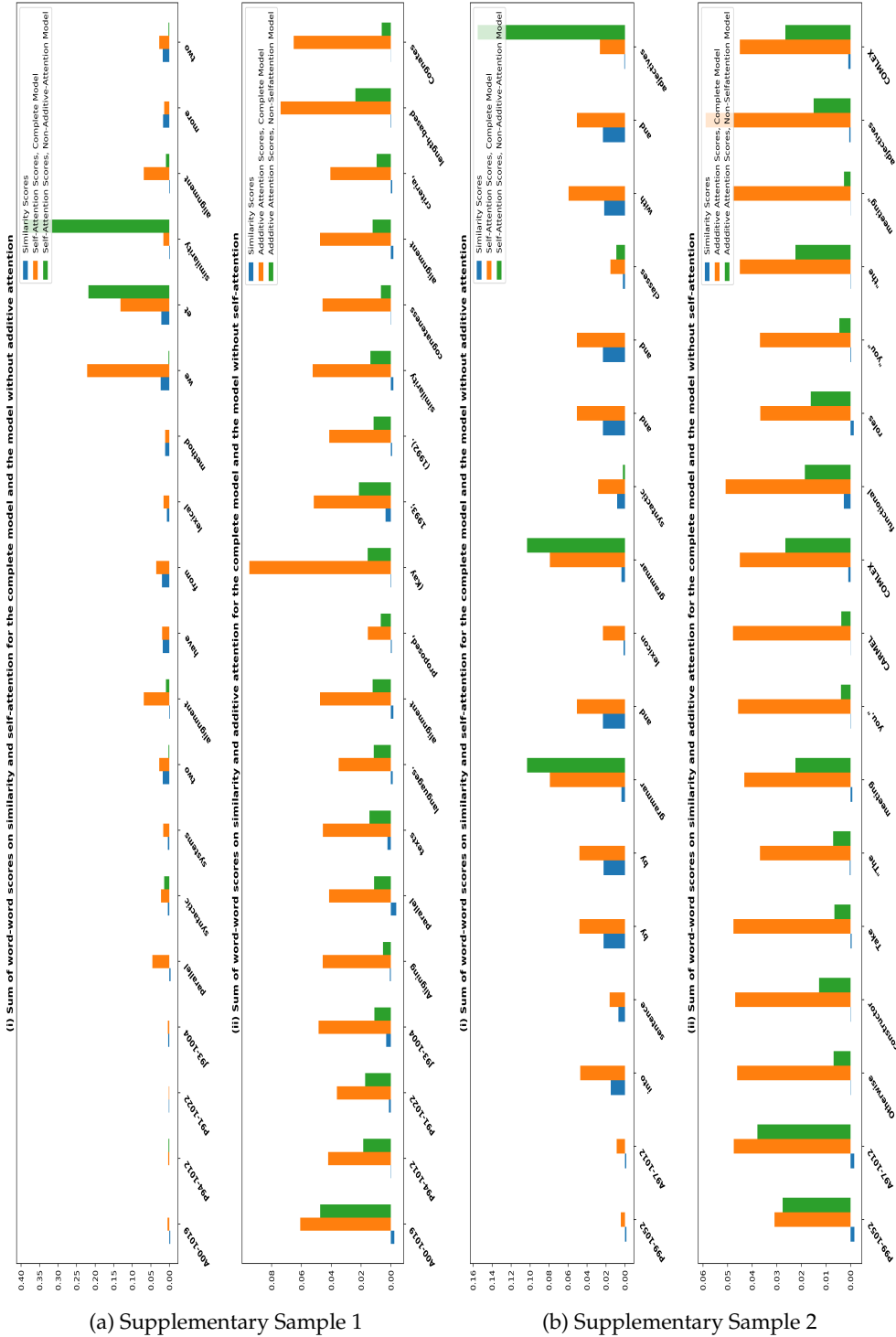


Figure A.3 Scores of additive attention (top 15) and summed self-attention against similarities for Supplementary sample 1 & 2.

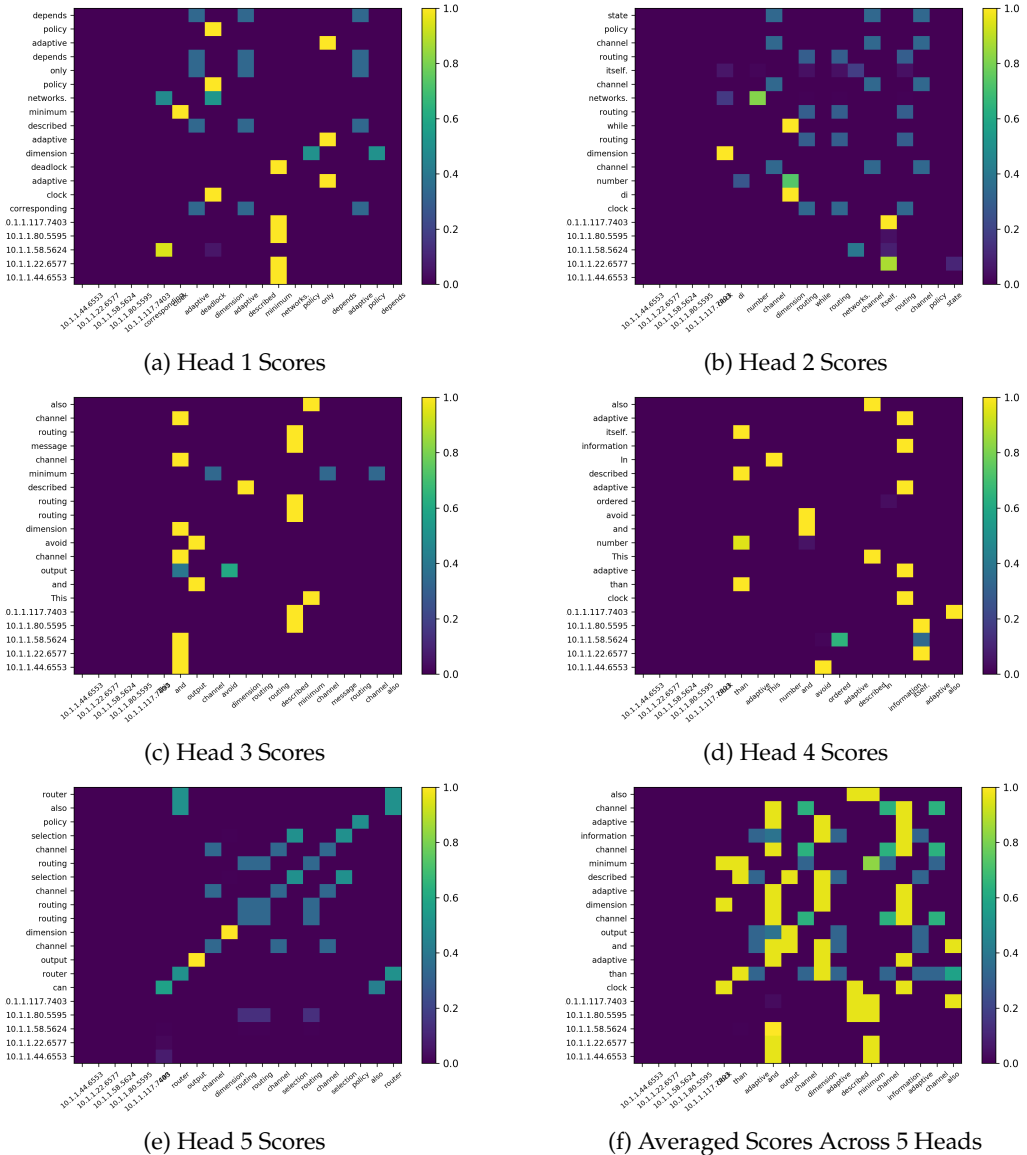


Figure A.4 Pairwise self-attention scores (top 15 items) for Supplementary sample 3 via complete DACR.

as “and,” and “also.” Additive attention (Figure A.6) mostly recognized the words relevant to the citing intent, such as “routing,” and “n-cude,” as most of the words relevant to the citing intent are not likely to appear.

For the fourth sample in Table A.2, we see that the author would like to cite the study about spectral analysis by addressing the characteristics of the technique. According to Figure A.5, similar to sample 1, self-attention has recognized words that are relevant to the citing intent, such as “spectral,” “analysis,” “rendering,” and so forth, but also a few connecting words such as “can” and “et” are recognized. Additive attention (Figure A.6) mostly recognized the words relevant to the citing intent, such as

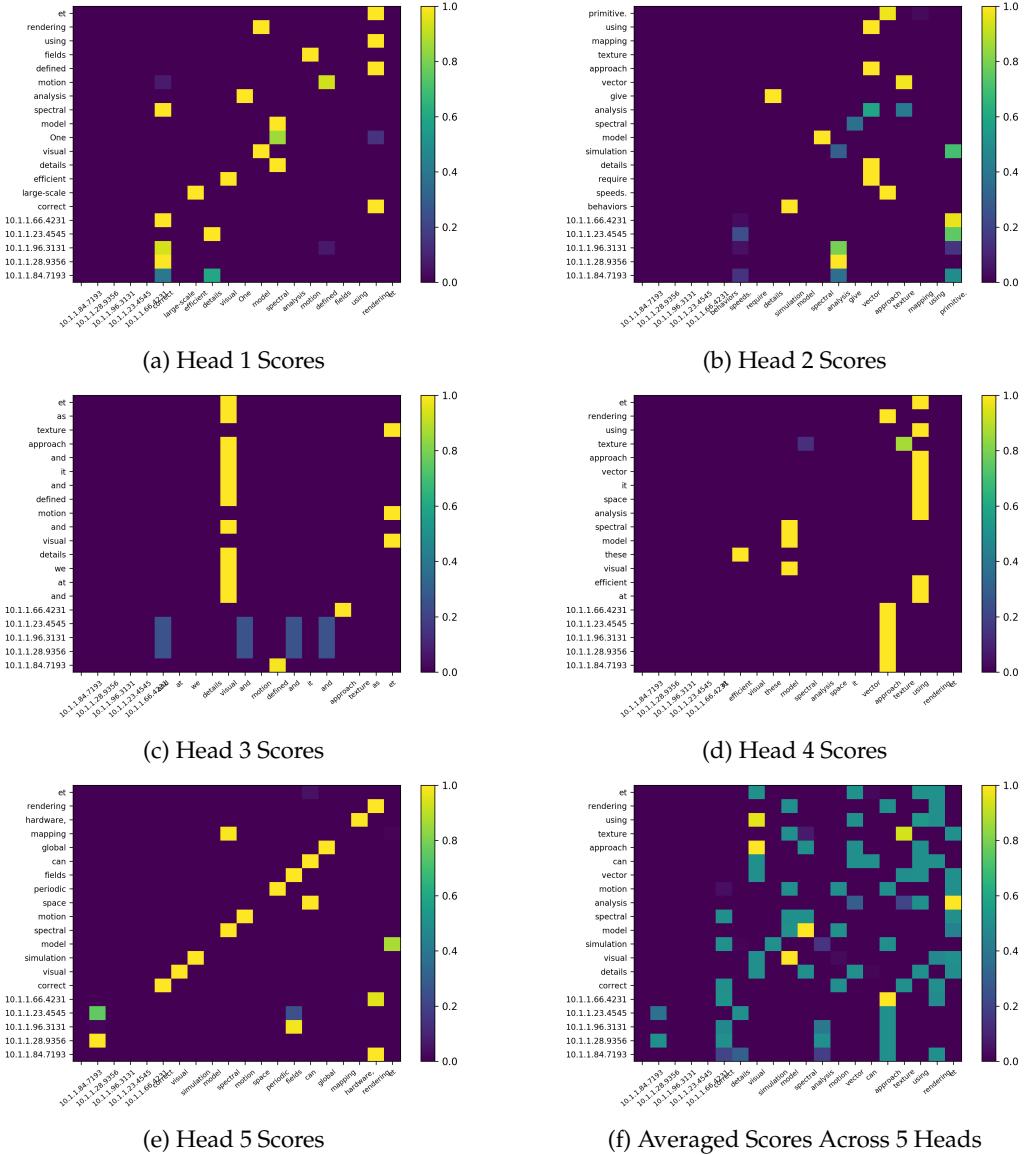


Figure A.5 Pairwise self-attention scores (top 15 items) for Supplementary sample 4 via complete DACR.

“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 6, and Supplementary sample 1 and 2. We could conclude that both self-attention and additive attention recognize the words that are relevant to the citing intent, although self-attention may also assign high weights to connecting words, whereas additive attention may assign high weights to unique but irrelevant words.

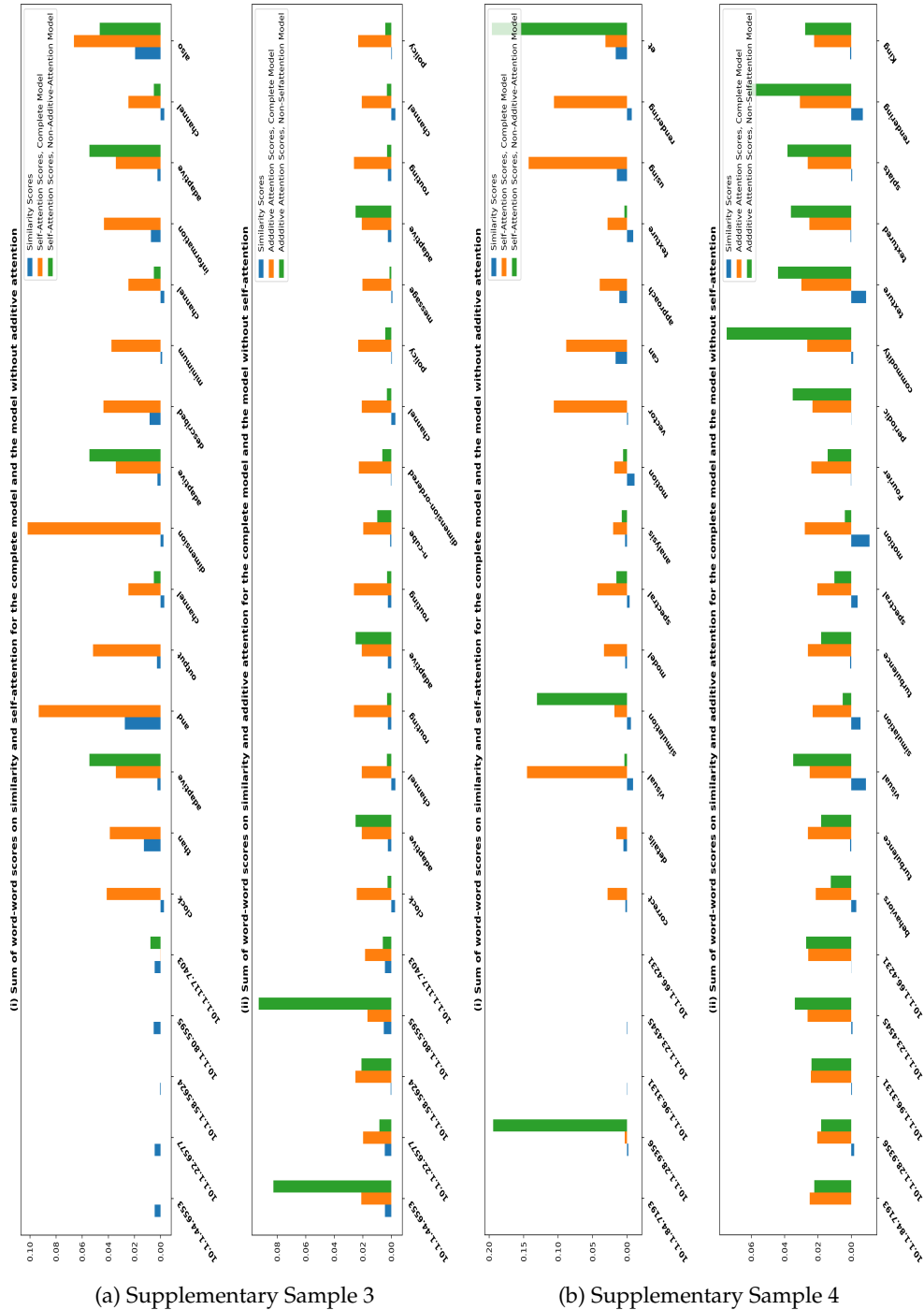


Figure A.6 Scores of additive attention (top 15) and summed self-attention against similarities for Supplementary sample 3 & 4.

Appendix 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..." (Chen and Nie 2000)

What is ground truth paper (Brown, Lai, and Mercer 1991) 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 Dunning (1993) 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 Brown et al. (1993) 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 gives 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 Wu (1994) 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: 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 fourth candidate (CAN4) by Gale and Church (1991) 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. Rate: 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 Kupiec (1993) 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. Rate: 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)..." (Rindfleisch, Rajan, and Hunter. 2000)

What is ground truth paper (Church 1988) 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 Cutting et al. (1992) 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 Brill and Resnik (1994) 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 Rich and LuperFoy (1988) 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 Karlsson (1990) 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 Church and Hanks (1990) 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”...” (Brants 2000)

What is ground truth paper (Ratnaparkhi 1996) 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 Cutting et al. (1992) 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 Church (1988) 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 Brill and Resnik (1994) 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 Brill (1995) 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 Walker (1989) 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..." (Marcu, Carlson, and Watanabe 2000)

What is ground truth paper (Ratnaparkhi 1996) 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 Johnson (1998) 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 Ramshaw and Marcus (1995) 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 Meyers, Yangarber, and Grishman (1996) 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 Ratnaparkhi (1997) 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 tree. 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 Fox (2002) 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 MarC (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..." (Marcu, Carlson, and Watanabe 2000)

What is ground truth paper (Pantel and Lin 2000) 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 Brown et al. (1991) 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 Yarowsky (1995) 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 Dagan and Itai (1994) 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 Gale and Church (1991) 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. Rate: 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 Brown et al. (1990) 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. Rate: 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..." (Yilmaz, Shafique, and Shah 2002)

What is ground truth paper (Terzopoulos and Waters 1993) 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 Lucas and Kanade (1981) 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 Essa and Pentland (1997) 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 Bergen et al. (1992) 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 DeCarlo and Metaxas (2000) 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 Black and Yacoob (1995) 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..." (Yilmaz, Shafique, and Shah 2002)

What is ground truth paper (Essa and Pentland 1997) 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 Terzopoulos and Waters (1993) 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 Tian, Kanade, and Cohn (2001) 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 Kanade, Tian, and Cohn (2000) 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 Donato et al. (1999) 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 Black and Yacoob (1995) 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 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..." (Brugnara et al. 2000)

What is ground truth paper (Gales 1998) 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 Gales and Woodland (1996) 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 Gauvain and Lee (1994) 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 Anastasakos et al. (1996) 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 Gales (1999) 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 Woodland, Gales, and Pye (1996) 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)...” (Ross and Zemel 2006)

What is ground truth paper (Hinton and Zemel 1993) 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 Blei, Ng, and Jordan (2003) 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 allcotoion 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 Lee and Seung (2000) 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 Hofmann (1999) 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 Hofmann (2001) 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 Barnard et al. (2003) 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..." (Ross and Zemel 2006)

What is ground truth paper (Lee and Seung 2000) 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 Hoyer (2004) 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 Blei, Ng, and Jordan (2003) 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 Hofmann (1999) 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 Hinton and Zemel (1993) 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 Lewicki and Sejnowski (2000) 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.

Acknowledgments

This research has been supported in part by JSPS KAKENSHI under grant number 19H04116 and by MIC SCOPE under grant numbers 201607008 and 172307001.

References

- Allen, James F., Bradford W. Miller, Eric K. Ringger, and Teresa Sikorski. 1996. A robust system for natural spoken dialogue. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 62–70. <https://doi.org/10.3115/981863.981872>
- Alzoghbi, Anas, Victor Anthony Arrascue Ayala, Peter M. Fischer, and Georg Lausen. 2015. PubRec: Recommending publications based on publicly available meta-data. In *Proceedings of the LWA 2015 Workshops: KDML, FGWM, IR, and FGDB, Trier*, volume 1458 of *CEUR Workshop Proceedings*, pages 11–18.
- Anastasakos, Tasos, John W. McDonough, Richard M. Schwartz, and John Makhoul. 1996. A compact model for speaker-adaptive training. In the *4th International Conference on Spoken Language Processing*.
- Ba, Lei Jimmy, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.
- Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations*.
- Barnard, Kobus, Pinar Duygulu, David A. Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. 2003. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135.
- Beltagy, Iz, Kyle Lo, and Arman Cohan. 2019. 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*, pages 3613–3618. <https://doi.org/10.18653/v1/D19-1371>
- Bergen, James R., P. Anandan, Keith J. Hanna, and Rajesh Hingorani. 1992. Hierarchical model-based motion estimation. In *Computer Vision - ECCV'92, Second European Conference on Computer Vision*, volume 588 of *Lecture Notes in Computer Science*, pages 237–252. https://doi.org/10.1007/3-540-55426-2_27
- Berger, Matthew, Katherine McDonough, and Lee M. Seversky. 2017. cite2vec: Citation-driven document exploration via word embeddings. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):691–700. <https://doi.org/10.1109/TVCG.2016.2598667>
- Black, Michael J. and Yaser Yacoob. 1995. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *Proceedings of the Fifth International Conference on Computer Vision*, pages 374–381. <https://doi.org/10.1109/ICCV.1995.466915>
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- Brants, Thorsten. 2000. TNT: A statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, pages 224–231.
- Brill, Eric. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565.
- Brill, Eric and Philip Resnik. 1994. A rule-based approach to prepositional phrase attachment disambiguation. In *15th International Conference on Computational Linguistics*, pages 1198–1204. <https://doi.org/10.3115/991250.991346>
- Brown, Peter F., John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer,

- and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.
- Brown, Peter F., Jennifer C. Lai, and Robert L. Mercer. 1991. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 169–176. <https://doi.org/10.3115/981344.981366>
- Brown, Peter F., Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 264–270. <https://doi.org/10.3115/981344.981378>
- Brown, Peter F., Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.
- Brugnarà, F., M. Cettolo, M. Federico, and D. Giuliani. 2000. A system for the segmentation and transcription of Italian radio news. In *RIAO '00: Content-Based Multimedia Information Access - Volume 1*, pages 364–371.
- Brunner, Gino, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On identifiability in transformers. In *International Conference on Learning Representations*.
- Caragea, Cornelia, Adrian Silvescu, Prasenjit Mitra, and C. Lee Giles. 2013. Can't see the forest for the trees?: A citation recommendation system. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 111–114. <https://doi.org/10.1145/2467696.2467743>
- Chen, Jiang and Jian-Yun Nie. 2000. Automatic construction of parallel English-Chinese corpus for cross-language information retrieval. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, pages 21–28. <https://doi.org/10.3115/974147.974151>
- Chen, Stanley F. 1993. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 9–16. <https://doi.org/10.3115/981574.981576>
- Chu-Carroll, Jennifer. 2000. Evaluating automatic dialogue strategy adaptation for a spoken dialogue system. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, pages 202–209.
- Church, Kenneth Ward. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136–143. <https://doi.org/10.3115/974235.974260>
- Church, Kenneth Ward and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.
- Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. 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*, pages 276–286. <https://doi.org/10.18653/v1/W19-4828>
- Councill, Isaac G., C. Lee Giles, and Min-Yen Kan. 2008. ParsCit: An open-source CRF reference string parsing package. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 661–667.
- Cutting, Douglas R., Julian Kupiec, Jan O. Pedersen, and Penelope Sibun. 1992. A practical part-of-speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, pages 133–140. <https://doi.org/10.3115/974499.974523>
- Dagan, Ido and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.
- DeCarlo, Douglas and Dimitris N. Metaxas. 2000. Optical flow constraints on deformable models with applications to face tracking. *International Journal of Computer Vision*, 38(2):99–127. <https://doi.org/10.1023/A:1008122917811>
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- Donato, Gianluca, Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski. 1999. Classifying facial actions. *IEEE Transactions on Pattern*

- Analysis and Machine Intelligence*, 21(10):974–989. <https://doi.org/10.1109/34.799905>
- Duato, José. 1993. A new theory of deadlock-free adaptive routing in wormhole networks. *IEEE Transactions on Parallel Distributed Systems*, 4(12):1320–1331. <https://doi.org/10.1109/71.250114>
- Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.
- Essa, Irfan A. and Alex Pentland. 1997. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–763. <https://doi.org/10.1109/34.598232>
- Fox, Heidi. 2002. Phrasal cohesion and statistical machine translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 304–311. <https://doi.org/10.3115/1118693.1118732>
- Gale, William A. and Kenneth Ward Church. 1991. Identifying word correspondences in parallel texts. In *Proceedings of the Workshop on Speech and Natural Language*, pages 152–157. <https://doi.org/10.3115/112405.112428>
- Gales, M. J. F. 1998. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech & Language*, 12(2):75–98. <https://doi.org/10.1006/cs1a.1998.0043>
- Gales, Mark J. F. 1999. Semi-tied covariance matrices for hidden Markov models. *IEEE Transactions on Speech Audio Processing*, 7(3):272–281. <https://doi.org/10.1109/89.759034>
- Gales, Mark J. F. and Philip C. Woodland. 1996. Mean and variance adaptation within the MLLR framework. *Computer Speech & Language*, 10(4):249–264. <https://doi.org/10.1006/cs1a.1996.0013>
- Gauvain, Jean-Luc and Chin-Hui Lee. 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech Audio Processing*, 2(2):291–298. <https://doi.org/10.1109/89.279278>
- Gori, Marco and Augusto Pucci. 2006. Research paper recommender systems: A random-walk based approach. In *2006 IEEE / WIC / ACM International Conference on Web Intelligence*, pages 778–781. <https://doi.org/10.1109/WI.2006.149>
- Grishman, Ralph, Catherine Macleod, and Adam L. Meyers. 1994. Complex Syntax: Building a computational lexicon. In *Proceedings of the 15th Conference on Computational Linguistics*, pages 268–272. <https://doi.org/10.3115/991886.991931>
- Han, Jialong, Yan Song, Wayne Xin Zhao, Shuming Shi, and Haisong Zhang. 2018. hyperdoc2vec: Distributed representations of hypertext documents. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2384–2394. <https://doi.org/10.18653/v1/P18-1222>
- Hao, Yaru, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, The Eleventh Symposium on Educational Advances in Artificial Intelligence*, pages 12963–12971.
- Harper, Mary P., Christopher M. White, Wen Wang, Michael T. Johnson, and Randall A. Helzerman. 2000. The effectiveness of corpus-induced dependency grammars for post-processing speech. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, pages 102–109.
- He, Qi, Daniel Kifer, Jian Pei, Prasenjit Mitra, and C. Lee Giles. 2011. Citation recommendation without author supervision. In *Proceedings of the Fourth International Conference on Web Search and Web Data Mining*, pages 755–764. <https://doi.org/10.1145/1935826.1935926>
- He, Qi, Jian Pei, Daniel Kifer, Prasenjit Mitra, and C. Lee Giles. 2010. Context-aware citation recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 421–430. <https://doi.org/10.1145/1772690.1772734>
- Hinton, Geoffrey E., Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580.
- Hinton, Geoffrey E. and Richard S. Zemel. 1993. Autoencoders, minimum description length and helmholtz free energy. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, pages 3–10.

- Hofmann, Thomas. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 289–296.
- Hofmann, Thomas. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1/2):177–196. <https://doi.org/10.1023/A:1007617005950>
- Hoyer, Patrik O. 2004. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning*, 5:1457–1469.
- Jia, Haofeng and Erik Saule. 2017. An analysis of citation recommender systems: Beyond the obvious. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 216–223. <https://doi.org/10.1145/3110025.3110150>
- Jia, Haofeng and Erik Saule. 2018. Local is good: A fast citation recommendation approach. *Advances in Information Retrieval*, 10772:758–764. https://doi.org/10.1007/978-3-319-76941-7_73
- Johnson, Mark. 1998. PCFG models of linguistic tree representations. *Computational Linguistics*, 24(4):613–632.
- Kanade, Takeo, Ying-li Tian, and Jeffrey F. Cohn. 2000. Comprehensive database for facial expression analysis. In *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53. <https://doi.org/10.1109/AFGR.2000.840611>
- Karlsson, Fred. 1990. Constraint grammar as a framework for parsing running text. In *Proceedings of the 13th Conference on Computational Linguistics*, pages 168–173. <https://doi.org/10.3115/991146.991176>
- Küçüktunç, Onur, Erik Saule, Kamer Kaya, and Ümit V. Çatalyürek. 2013. Towards a personalized, scalable, and exploratory academic recommendation service. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 636–641. <https://doi.org/10.1145/2492517.2492605>
- Kumar, Dianne R. and Walid A. Najjar. 1999. Combining adaptive and deterministic routing: Evaluation of a hybrid router. In *Proceedings of the Third International Workshop on Network-Based Parallel Computing: Communication, Architecture, and Applications*, volume 1602, pages 150–164. https://doi.org/10.1007/10704826_11
- Kupiec, Julian. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 17–22. <https://doi.org/10.3115/981574.981577>
- Lavoie, Benoit, Richard I. Kittredge, Tanya Korelsky, and Owen Rambow. 2000. A framework for MT and multilingual NLG systems based on uniform lexico-structural processing. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, pages 60–67. <https://doi.org/10.3115/974147.974156>
- Lavoie, Benoit and Owen Rainbow. 1997. A fast and portable realizer for text generation systems. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 265–268. <https://doi.org/10.3115/974557.974596>
- Le, Quoc V. and Tomáš Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, volume 32, pages 1188–1196.
- Lee, Daniel D. and H. Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, pages 556–562.
- Lewicki, Michael S. and Terrence J. Sejnowski. 2000. Learning overcomplete representations. *Neural Computation*, 12(2):337–365. <https://doi.org/10.1162/089976600300015826>
- Li, Shuchen, Peter Brusilovsky, Sen Su, and Xiang Cheng. 2018. Conference paper recommendation for academic conferences. *IEEE Access*, 6:17153–17164. <https://doi.org/10.1109/ACCESS.2018.2817497>
- Ling, Wang, Yulia Tsvetkov, Silvio Amir, Ramon Fernandez, Chris Dyer, Alan W. Black, Isabel Trancoso, and Chu-Cheng Lin. 2015. Not all contexts are created equal: Better word representations with variable attention. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language*, pages 1367–1372. <https://doi.org/10.18653/v1/D15-1161>
- Lucas, Bruce D. and Takeo Kanade. 1981. An iterative image registration technique with an application to stereo vision. In

- Proceedings of the 7th International Joint Conference on Artificial Intelligence*, pages 674–679.
- Luong, Minh-Thang, Thuy Dung Nguyen, and Min-Yen Kan. 2010. Logical structure recovery in scholarly articles with rich document features. *International Journal of Digital Library Systems*, 1(4):1–23. <https://doi.org/10.4018/jdls.2010100101>
- Maaten, L. V. D. and Geoffrey E. Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.
- Mack, Chris A. 2014. How to write a good scientific paper: Structure and organization. *Journal of Micro/Nanolithography, MEMS, and MOEMS*, 13(4):1–3. <https://doi.org/10.1117/1.JMM.13.4.040101>
- Marcu, Daniel, Lynn Carlson, and Maki Watanabe. 2000. The automatic translation of discourse structures. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, pages 9–17.
- Maruyama, Hiroshi. 1990. Structural disambiguation with constraint propagation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 31–38. <https://doi.org/10.3115/981823.981828>
- McNee, Sean M., István Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K. Lam, Al Mamunur Rashid, Joseph A. Konstan, and John Riedl. 2002. On the recommending of citations for research papers. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, pages 116–125. <https://doi.org/10.1145/587078.587096>
- Meyers, Adam, Roman Yangarber, and Ralph Grishman. 1996. Alignment of shared forests for bilingual corpora. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 460–465. <https://doi.org/10.3115/992628.992708>
- Mikolov, Tomáš, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations (Workshop)*.
- Mikolov, Tomáš, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, volume 2, pages 3111–3119.
- Pantel, Patrick and Dekang Lin. 2000. Word-for-word glossing with contextually similar words. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, pages 78–85.
- Paszke, Adam, 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. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference Advances in Neural Information Processing Systems*, volume 32, pages 8024–8035.
- Pedregosa, F., 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. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Ramshaw, Lance A. and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.
- Ratnaparkhi, Adwait. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of the 1st Conference on Empirical Methods in Natural Language Processing*.
- Ratnaparkhi, Adwait. 1997. A linear observed time statistical parser based on maximum entropy models. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*.
- Řehůřek, Radim and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *LREC Workshop on New Challenges for NLP Frameworks*, pages 45–50.
- Rich, Elaine and Susann LuperFoy. 1988. An architecture for anaphora resolution. In *Proceedings of the 2nd Conference on Applied Natural Language Processing*, pages 18–24. <https://doi.org/10.3115/974235.974239>
- Rindflesch, Thomas C., Jayant V. Rajan, and Lawrence Hunter. 2000. Extracting

- molecular binding relationships from biomedical text. In *Sixth Applied Natural Language Processing Conference*, pages 188–195.
- Rosé, Carolyn Penstein. 2000. A framework for robust semantic interpretation learning. In *6th Applied Natural Language Processing Conference, ANLP 2000*, pages 311–318.
- Ross, David A. and Richard S. Zemel. 2006. Learning parts-based representations of data. *Journal of Machine Learning Research*, 7:2369–2397.
- Stam, Jos and Eugene Fiume. 1993. Turbulent wind fields for gaseous phenomena. In *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques*, pages 369–376. <https://doi.org/10.1145/166117.166163>
- Sutskever, Ilya, James Martens, George E. Dahl, and Geoffrey E. Hinton. 2013. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 1139–1147.
- Tang, Yichuan, Nitish Srivastava, and Ruslan Salakhutdinov. 2014. Learning generative models with visual attention. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, volume 1, pages 1808–1816.
- Terzopoulos, Demetri and Keith Waters. 1993. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):569–579. <https://doi.org/10.1109/34.216726>
- Tian, Ying-li, Takeo Kanade, and Jeffrey F. Cohn. 2001. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115. <https://doi.org/10.1109/34.908962>
- Varadhan, Gokul, Shankar Krishnan, T. V. N. Sriram, and Dinesh Manocha. 2004. Topology preserving surface extraction using adaptive subdivision. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*, volume 71, pages 235–244. <https://doi.org/10.1145/1057432.1057464>
- Varadhan, Gokul, Shankar Krishnan, T. V. N. Sriram, and Dinesh Manocha. 2006. A simple algorithm for complete motion planning of translating polyhedral robots. *International Journal of Robotics Research*, 25(11):1049–1070. <https://doi.org/10.1177/0278364906071199>
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pages 5998–6008.
- Walker, Marilyn A. 1989. Evaluating discourse processing algorithms. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 251–261. <https://doi.org/10.3115/981623.981654>
- Wei, Xiaoming, Wei Li, Klaus Mueller, and Arie E. Kaufman. 2004. The lattice-Boltzmann method for simulating gaseous phenomena. *IEEE Transactions on Visualization and Computer Graphics*, 10(2):164–176. <https://doi.org/10.1109/TVCG.2004.1260768>
- Woodland, Philip C., Mark John Francis Gales, and David Pye. 1996. Improving environmental robustness in large vocabulary speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing Conference*, pages 65–68. <https://doi.org/10.1109/ICASSP.1996.540291>
- Wu, Chuhan, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. 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*, pages 6388–6393. <https://doi.org/10.18653/v1/D19-1671>
- Wu, Dekai. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 80–87. <https://doi.org/10.3115/981732.981744>
- Yarowsky, David. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196. <https://doi.org/10.3115/981658.981684>
- Yilmaz, Alper, Khurram Shafique, and Mubarak Shah. 2002. Estimation of rigid and non-rigid facial motion using

- anatomical face model. In *Proceedings of the 16th International Conference on Pattern Recognition*, pages 377–380. <https://doi.org/10.1109/ICPR.2002.1044729>
- Zhang, Yang and Qiang Ma. 2020a. DocCit2Vec: Citation recommendation via embedding of content and structural contexts. *IEEE Access*, 8:115865–115875. <https://doi.org/10.1109/ACCESS.2020.3004599>
- Zhang, Yang and Qiang Ma. 2020b. Dual attention model for citation recommendation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3179–3189. <https://doi.org/10.18653/v1/2020.coling-main.283>