Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 4 (April) (2025)

Taxonomy of Citation Contexts: A Framework for Systematic Analysis of Reference Text Extraction in Computational Linguistic

Afsheen Khalid (Corresponding Author)
Center for Excellence in IT, Institute of Management Sciences, Peshawar, Khyber Pakhtunkhwa, Pakistan. Email: afsheen.khalid@imsciences.edu.pk

Dilawar Khan
Computer Science & IT Department, University of Engineering and Technology, Peshawar, Khyber Pakhtunkhwa, Pakistan.
Email: dilawarkhan@uetpeshawar.edu.pk

Fazal Malik
Department of Computer Science Iqra National University Peshawar, Khyber Pakhtunkhwa, Pakistan. Email: fazal.malik@inu.edu.pk
https://orcid.org/0009-0009-6104-1651

Ashraf Ullah
Institute of Computer Science and Information Technology, University of Science and Technology Bannu, Khyber Pakhtunkhwa, Pakistan.
Email: ashrafbth@gmail.com

Abstract

Citation contexts (CCs)—text near citation marks—is helpful summaries of referenced materials but challenging for machine analysis because they vary in structure and contain inherent vagueness. Current methods largely apply fixed-window extraction, where unnecessary information tends to be obtained or key points go unexamined. The necessity for more formal CC analysis is a result of the weakness in the current strategy. The previous approaches lack a comprehensive framework to categorize CCs based on syntactic scope, information completeness, and ambiguity, thus being less effective for computational linguistics applications such as reference extraction and sentiment analysis. In response to this deficiency, our research constructs an in-depth taxonomy to classify CCs by their positional, syntactic, and contextual properties. We examined 100 ACL Anthology Network research papers, manually classifying CCs into four major dimensions: citation position (head, mid, tail), syntactic units (phrase, clause, and sentence), missing information, and ambiguity. Our results show that tail-position citations usually refer to whole statements, whereas head and mid citations need accurate scope identification. Interestingly, 95% of CCs are single-sentence, with phrases and clauses being most frequent in mid-position citations. Moreover, 15% of CCs showed ambiguity that challenged even human annotators. This taxonomy facilitates CC processing in applications such as reference extraction and opens up future directions of research in multi-sentence CCs and machine learning-based analysis.

Keywords: Citation context, reference extraction, computational linguistics, text categorization, fixed-window extraction, NLP.

Introduction

A Citation Context (CC) is the surrounding text of a citation marker pointing to another work, playing two primary roles: source identification and description of its applicability. It tends to encapsulate the contribution of the cited work so that readers can understand its meaning without access to the original. However, CCs are hard to understand as their structure relies on writer, field, and purpose. Even though they typically contain explicit, descriptive text about the referenced work, their ambiguity renders them problematic to automatically extract [1].

Challenges in CC Interpretation

Standardization is missing in CC construction, resulting in inconsistency. A primary problem arises with citations found at the ends of sentences, such that one is uncertain if they support a clause, a phrase, or a whole sentence. This makes reference text extraction, sentiment analysis, and citation classification challenging. Current approaches tend to oversimplify CCs with the use of whole sentences or fixed-size windows of text [2][3][4], which can incorporate extraneous material or leave essential information out.

Ambiguity and Data Quality Issues

Certain CCs are so ambiguous that even human annotators have difficulty identifying the text referred to. Noise due to unclear syntax, implicit purpose, or very little context contributes to such ambiguity. Such troublesome CCs hurt performance on downstream tasks (e.g., summarization, sentiment analysis) and should be excluded from datasets or flagged so that results aren't skewed.

Proposed Framework for CC Categorization

This paper introduces taxonomy to classify CCs systematically, filling gaps in existing research. Our framework specifies four main dimensions: 1) Citation Mention Position: Sentence location (head, middle, tail), which influences referential scope. 2) Syntactic Units: Grammatical scope (phrase, clause, sentence, or multi-sentence).3) Missing Information: CCs that need external context for proper interpretation.4) Ambiguous CCs: References with ambiguous boundaries due to unclear wording or implicit assumptions.

This categorization allows for stricter dataset building and guides citation-sensitive NLP algorithms, enhancing such tasks as reference extraction and citation classification.

The primary contributions offered by the proposed work are: 1) Presenting the first full CC categorization framework. 2) Disambiguating informative, incomplete, and ambiguous contexts. 3) Enabling higher-quality dataset generation and computational analysis.

The rest of the paper is organized: Section 2 presents related work; Section 3 outlines proposed approach; Section 4 presents the results and discussion; and Section 5 concludes with future directions.

Related Work

This section discusses studies pertinent to reference text extraction, especially those that use citation contexts (CCs). Initial methods were manual, but fixed-window text analysis has been the dominant approach in the last decade's research.

The idea of reference terms evolved in information retrieval (IR) with O'Connor [5], who manually identified the reference terms for cited articles. Bradshaw et al. [2] later used reference-directed indexing, with 100-term windows centered on citations instead of particular reference terms. Ritchie et al. [6] greatly improved cited article indexing by showing CCs' value as sources of novel index terms, while recognizing two primary challenges: (1) linguistic complexity in word-citation association, and (2) noise from fixed windows picking up on irrelevant terms. Their subsequent work [7] fell back on manual reference term extraction since there were not enough automated techniques available.

Liu et al. [8] took a different tack, retrieving CCs from query terms for literature retrieval. Although they employed recurring query terms as citation topics, they did not check if these actually represented citations or were noise. Our research is fundamentally different in that it works not on retrieval, but on how to pick out text that precisely captures citation contexts.

Moro et al. [3] have recently proposed exploratory search via CCs with 100 words on each side of citation points for auto-query refinement. Likewise, [4] made use of 400 characters in proximity to citation points for classifying documents.

Amjad et al. [9] targeted quoting sentences that quote several papers to create summaries, whereas our strategy examines all CC types irrespective of the number of citations. Their objective was the creation of coherent summaries, but we intend to obtain useful text from any CC for varied applications.

For sentiment detection, Athar [10] found optimal context window sizes as single-sentence CCs provided higher F-macro/F-micro scores than multi-sentence contexts (1-4 sentences). They concluded that "jointly detecting sentiment and context is a hard problem" [10]. From this and other researches [11][12], we created a single-sentence CC dataset. On the other hand, Abu-Jabara [13] processed four-sentence windows (citation sentence + one previous + two subsequent sentences) to detect citation purpose and polarity.

He et al. [14] tried to insert citations into a query manuscript wherever the text was considered relevant. They concluded that determining the exact set of words describing why a reference is relevant is a hard task, reserved for future work. Our research question attacks this problem in the opposite direction: rather than seeking possible citation contexts (CCs), we begin with an already known citation position and try to find the snippet of words most closely connected to the known reference.

In a recent experiment by [15] on citation suggestion, the authors employed the CiteSeer dataset, which contains CCs of length 50 on both sides of the mention of the citation. But they didn't try to find out the corresponding reference text itself; they simply used the pre-annotated CCs as provided by the CiteSeer digital library.

A newer study [12] criticizes previous methods on citation recommendation with symmetric window strategies for extracting in-text citation data, stating that the strategies are suboptimal. The authors experimentally illustrate that sentence-level strategies are better than symmetric windows for citation recommendation tasks. The experiments, however, did not seek automation. Through manual annotation of the AAN dataset, they illustrated that sentences from within a five-sentence context offer superior performance.

A more immediately relevant study is that of Kang et al. [11], who manually processed many citing sentences to determine the characteristics of citing behavior and scope of citation. Their main conclusion was that a mere 5% of citing sentences constituted multi-sentence CCs. Considering the dominance of single-sentence CCs, our study accordingly targets single-sentence CCs.

Also, Caragea et al. [16] utilized Citation Contexts (CCs) to extract informative features for a key-phrase extraction task. They clearly indicated that using advanced methods to identify text relevant to citations would be useful and left it as future work. Even if Caragea et al. were working on extracting key-phrases from CCs, they did not treat these key-phrases as reference terms. Although CCs can include numerous key-phrases, not all of them are reference terms. Our work does not try to extract all key-phrases from a CC, but rather those that directly describe citations.

To evaluate the significance of a research paper, Zhu et al. [17] tried to automatically detect references with a core academic impact on the cited paper. A CC was defined as ten words around the citation to craft their features. While this context window is limited—because one or two words preceding the citation usually have author names in most citation formats—their system nonetheless attained robust performance.

Chakraborty et al. [18] proposed the reference intensity concept, which quantifies how strong an influence a citation has on the cited publication. Their method utilized three sentences as the reference context: the citing sentence and the preceding and succeeding two sentences.

Citation function classification is one of the most important topics in citation analysis. Research in this area [19][20][21] tends to examine text within proximity of the citation, hence the rationale behind the use of single-citing sentences in taxonomies covered in Section 3.2. Likewise, the classification of citation importance, as investigated in [22], is an equally important undertaking. This kind of research tends to make use of full-text publications and relies on sets such as the one created by [23]. Another recent study [23] also sought to construct a classifier to identify whether citations are informative. They considered CCs as the noun phrase before the citation and a verb-noun pair after it. Nevertheless, they realized that not all noun phrases are describing the citations and stressed the importance of strong heuristics to correctly identify relevant nouns and verbs.

Methodology

Dataset

The dataset construction process started with articles sourced from the ACL (Association of Computational Linguistics) Anthology Network (ANN) [24]. The search engine provides complete information about citations and research article summaries together with collaboration data and citation contexts for Computational Linguistics conference and journal papers. We used this processing tool to get direct access to CCs because both our research and Computational Linguistics category match and therefore we chose to use the tool for dataset preparation.

A total of 100 research papers comprised our selection because each contained at least fifty outgoing citations. The chosen criterion secured a sufficient number of CCs for our evaluation process. We processed the acquired CCs after collecting them from the 100 papers. The most critical preprocessing step after text cleaning involved modifying citation markers through which citations pointing to the target paper received "(key Citation)" while references to other papers were marked as "[]".

Taxonomy Development and Categorization Framework

This section showcases an overview of various Citation Contexts (CCs) categories which emerged through our examination of CCs from diverse classifications in the dataset. The figure represents how these categories appear (Figure 1) while Section 3 contains a detailed description of each category. The target citation requires the term *key citation* for its designation.

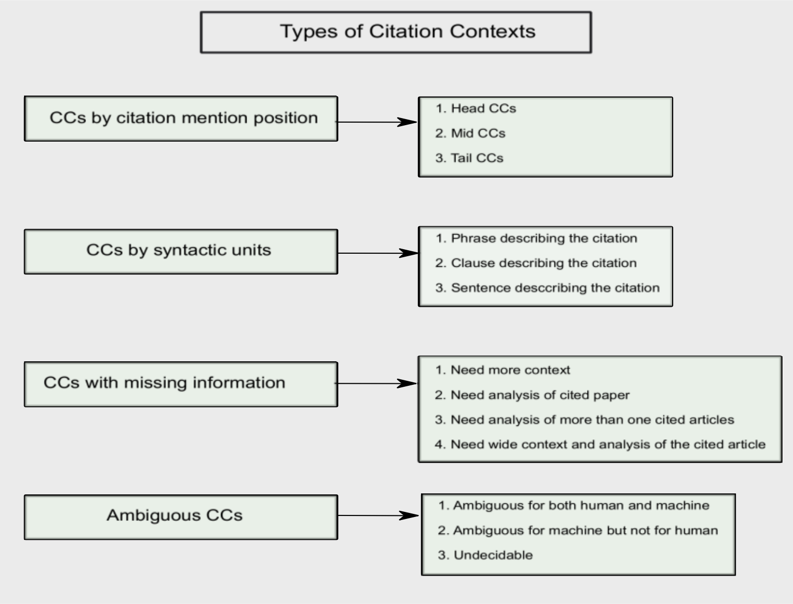
Types of Citation Contexts
CCs by citation mention position
1. Head CCs
2. Mid CCs
3. Tail CCs
CCs by syntactic units
1. Phrase describing the citation
2. Clause describing the citation
3. Sentence desccribing the citation
CCs with missing information
1. Need more context
2. Need analysis of cited paper
3. Need analysis of more than one cited articles
4. Need wide context and analysis of the cited article
Ambiguous CCs
1. Ambiguous for both human and machine
2. Ambiguous for machine but not for human
3. Undecidable

Figure 1. Categorization of Citation Contexts

Citation Contexts (CCs) by Citation Mention Positions

The position of a citation mention refers to the location of the key citation within the citation context (CC). A citation placed at the initial section of the CC becomes a head CC. The position of the key citation at the end of the CC identifies it as a tail CC. A citation mention located inside the CC section calls for using the term mid CC.

Head CCs

The standard design of these CCs starts with a key Citation followed by an appropriate functional verb that demonstrates the purpose of the citing author. The significant information regarding the source material appears behind the functional verb in the fragment. A reference text may fill an entire short sentence when located after the verb within a citation. Long sentences benefit from cutting away unnecessary elements starting from the comma in order to discover the core content. Examples are provided in List 1. The reference text extends throughout all content which follows the note in (1). The portion of text which precedes the comma in (2) establishes what is needed for the citation while the remaining parts can be omitted. The text from "that" until the end of the sentence functions as the reference text per Example (3) whereas all text following the key Citation becomes the relevant fragment in Example (4).

1) “(key Citation) note that the correct decision ***depends on all four lexical events (the verb, the object, the preposition, and the prepositional object)***.” [25]

2) “(key Citation) ***reported on 80 percent attachment accuracy***, an improvement of 13 percent over the baseline (i.e. guessing noun attachment in all 81 cases).” [26]

3) “(key Citation) exploit the fact that ***in sentence- initial NP PP sequences the PP unambiguously attaches to the noun***.” et al. [27]

4) “(key Citation) ***did not have access to a large Treebank***.” [26]

List 1. Head contexts of citations

Mid-CCs

Special attention is needed for finding reference text within this sentence format. The citation reference text either stays within the identical section of the key Citation text box or expands across both sections. The reference texts which support key Citations exist after the Citation in (1) and (2) of List 2. The reference text extends across all parts of (3) but (4) shows that key Citation is flanked by reference text on both left and right sides.

1) “We rely on Gsearch to provide moderately accu- rate information about verb frames in the same waythat (key Citation) ***relied on Fiddich to provide moderately accurate information about syntactic structure, and.*** ” [28]

2) “Clearly, although both types of PPs are well iden- tified, arguments are better identified than adjuncts, an observation already made by several other au- thors, especially (key Citation) ***in their detailed discussion of the errors in a noun or verb PP-attachment task.***” [29]

3) “The problem of attaching prepositional phrases as sister nodes of VP or as adjuncts to its object nouns is a classic case of syntactic ambiguity that causes trouble for parsers (key Citation), e.g., the difference between I ate a fish with a fork and I atea fish with bones 12, i.e. denoting the instrumentor an attribute of the fish.” [30]

4) “***For instance, the results of*** (key Citation) ***indi- cate that their PP attachment system works better for cores than for adjuncts.***” [31]

List 2. Mid-contexts of citations

Tail CCs

The majority of textual references in tail Citation Contexts (CCs) relate to full statements but there are exceptions where shortened phrases are included. Key

Citation mentions a single phrase in (1) of List 3 yet the complete statements receive reference through Citation in (2), (3), and (4).

1) “Recent work has turned to corpus-based or statis- tical approaches (key Citation).” [32]
2) “Attention has mostly been limited to selectional preferences of verbs, which have been used for a variety of tasks (key Citation). [33]
3) “Collocational relations between the words in a sentence proved very helpful in selecting the most plausible among all the possible parse trees for a sentence (key Citation).” [34]
4) “One of the earliest corpus-based approaches to prepositional phrase attachment used lexical prefer- ence by computing co-occurrence frequencies (lex-ical associations) of verbs and nouns with preposi- tions (key Citation).” [35]

List 3. Tail contexts of citations

CCs by Syntactic Units

The author uses citation (CC) to describe phrases as well as clauses while also employing it to quote entire statements from articles alongside descriptions of multiple-sentence content [1]. Two researchers developed four different categories for CCs by studying their syntactic units [2]. Three citation types are discussed while omitting the multi-sentence type because our database holds only single citation sentences [3].

Phrase describing the Citation

The primary usage of citations stems from authors who want to explain technical terms or concepts for reader reference when seeking detailed information. The author uses *probabilistic Earley parser* as the subject of the citation in List 4 to allow readers to find detailed definitions about *probabilistic Earley parsers* in the cited work. The phrases contained in CCs function at any location either from the beginning or middle or ending point of the sentence.

“For SCFGs, a probabilistic Earley parser (key Citation) provides the basic quantities we need to compute...” [36]

List 4. Phrase describing the citation

Clause describing the Citation

Rather than a phrase, a clause can also define a citation. It should be mentioned that in this thesis, the words "clause" and "text fragment" are applied interchangeably. A clause gives some information of the general idea found in the cited article or can be the author's remark on the target citation. Also, the words in the clause do not have to appear consecutively, but must be semantically consistent. For example, as can be seen from the example below in List 5, the clause for the key citation is "an algorithm for calculating exact string prefix probabilities given a PCFG," where the word "algorithm" comes before the key citation, and the other words in the clause come after the key citation. In addition, clauses may occur in any location of the citation context (CC).

“There are efficient algorithms in the literature (key Citation) for calculating exact string prefix probabilities given a PCFG.” [37]

List 5. Clause describing the citation

Sentence describing the Citation

It is typical for authors to employ citations as a way to explain entire sentences, which could either be plagiarized directly from the cited document or are highlights in a given field. Although the main use of such citations is acknowledgment, they also give the reader explanations of key points and ideas. In example (1) of List 6, the entire sentence is defined by the key Citation, and where this kind of citation appears is at the end of the sentence. Another situation, as in (2) of List 6, illustrates where the key Citation explains the entire statement. In this case, the writer describes the method of the article being cited to the reader, using entire sentences. Here, the focus Citation is placed at the beginning of the sentence, i.e., the head citation.

1) “An Earley chart is used for keeping track of all derivations that are consistent with the input (key Citation).” [38]

2) “(key Citation) applies this approach to the so- called IBM Candide system to build context- dependent models, compute automatic sentence splitting and to improve word reordering in translation.” [39]

List 6. Sentence describing the citation

CCs with Missing Information

The Citation Contexts (CCs) employed in this research are single-citation sentences. Nevertheless, there are situations, though less typical than single-sentence CCs, in which a single citation sentence is not adequate to supply information on the referential text and needs a wider context and/or extra information from the cited article. Recent studies [25] proved that, out of the total amount of citing sentences within the analyzed set, only 5% are multi-sentence CCs. We found the same in our data set, though we did not perform statistics for single or multi-sentence context counts. However, there are indeed CCs with missing data, and here below we present the situations when a single citation sentence needs extra processing in order to receive the reference text mentioned in the cited article.

Need more context

In (1) of List 7, the sentence starts with "this problem," indicating the problem discussed in the previous sentence. It needs clarification of what type of problem the target reference has studied. Likewise, in (2) of List 7, "this contrasts" gives incomplete information and needs to be resolved from the previous sentence(s).

1) “This problem has been studied by (Citation).”

2) “This contrasts with the techniques proposed by (Citation), which are extensions of parsing al- gorithms for probabilistic context-free grammars, and require considerably more involved proofs of correctness.” [40]

List 7. Need wide context

Need analysis of cited paper

It is also the practice of citing authors to mention a citation for an in-depth study without specifying particular concepts. In these instances, it is impossible to pull out the needed key phrases or terms without seeing the cited work. For example, in (1) of List 8, the authors mention the experiments but fail to specify the topic or concept of the experiment. In the same way, in (2) of the current list, the method variant is cited, and it is not possible to identify the type of discussed method for the user unless accessing the article and gaining the needed information. Solving the generalized information indicated in such Citation Contexts (CCs) is more complex and demands more processing than in the former case.

1) “They also duplicated the experiment of (key Cita- tion), which scored around 5% less than the rule- based approach.” [41]
2) “We used a variant of the method described in (key Citation) the main difference being that we applied... ” [42]

List 8. Need analysis of cited article

Need knowledge of more than one cited articles

There are also situations where authors compare more than one citation but the object being compared is not clearly stated for any of the citations. That is, a citation can refer to another citation for some purpose, but the subject of both citations is not declared in the sentence of citation and is implied through the cited articles. In List 9 below, the citing author indicates that the problem discussed in one cited article is not the same as the problem in another cited article, but does not indicate what the problem is. To solve this, the content and processing of both cited articles are required. Such a solution is more complicated and requires more processing than in previous cases since multiple cited articles are involved. But such cases of Citation Contexts (CCs) are uncommon in a collection of CCs for a cited article.

But it makes obvious that [] were tackling a problem different from (key citation) given the fact that their baselinewas at 59 percent guessing noun attachment (rather than 67 percent in the Hindle and Rooth experiments).” [43]

List 9. Need knowledge of more than one cited articles

Need wide context and knowledge of the cited article

In some cases, a CC asks for information from the broader context as well as the article being cited. For example, in (1) of List 10, this method, and in (2) of the list, our process asks for information from earlier sentences of the sentence of citation. Similarly, (1) and (2) of this list contrast their respective approaches with the methods cited in the article(s). Therefore, data from the articles quoted are also needed to duly settle the missing data in such CCs. Processing these kinds of CCs adds to the complexity of the applications where they are used.

1) “This technique is similar to the one in (key Cita- tion), and interpolates between the tendencies ...” [44]
2) “Our procedure differs critically from (key Cita- tion) in that we do not iterate, ...” [44]

List 10. Need wide context and knowledge of the cited article

Ambiguous CCs

Ambiguous CCs indicate that it is not clear—to humans, machines, or both—whether the primary Citation (CC) points to a specific phrase or an entire statement..

Ambiguous for both human and machine

In this case, humans and computers do not know what is being referred to by the prominent Citation. In list examples (1) and (2) of List 11, the bold represents one option for referential text while the bold and italic is another option for text relating to the citation. Both options are appropriate for a reader, but only the citing author will know if the referential purpose of the citation is the bold text or the bold and italic text fragment. In these situations, it is hard to precisely determine which text forms the referential text, making it very hard to automatically extract.

1) “One of the earliest corpus-based approaches to prepositional phrase attachment used lexical prefer- ence by computing co-occurrence frequencies (lex-ical associations) of verbs and nouns with preposi- tions (key Citation).” [35]
2) “Since we have a choice between two outcomes, we will use a likelihood ratio to compare the two relation probabilities (key Citation).” [45]

List 11. Ambiguous CCs for human and machine both

***Ambiguous for machine but not for human*:**

A tail citation can define a whole sentence or a sub-sentence, as illustrated in the last section. At times, humans can simply tell whether the citation is for a whole sentence or a sub-sentence. This is hard to automate since defining rules to separate the two is not easy considering the diversity of sentence structures. In list 12, examples (1) and (2), it is simpler for humans to recognize that the citation does not refer to the entire sentence. The bold text is the clear reference text in both examples, but programming this reasoning into a computer is still a challenging task. In the same way, example (3) within the list demonstrates that humans can simply conclude that the citation refers to the whole statement.

1) “Several approaches have statistically addressed the problem of prepositional phrase ambiguity, with comparable results (key Citation).” [42]
2) “As we do not propose long distance attachments, our method cannot be compared with other stan- dard corpus-based approaches to attachment reso- lution (key Citation).” [46]
3) “It is well-known that while lexicalization is useful, lexical parameters determined from the tree bank are poorly estimated because of the sparseness of tree bank data for particular words (key Citation).” [47]

List 12. Ambiguous CCs for machine but not for human

Undecidable

Other than the above-stated categories, there are Citation Contexts (CCs) which do not belong to any particular class. It is then difficult to identify what text needs to be extracted. Some of the examples of these CCs are given below in List 13.

1) “With the exception of (key Citation), most un- supervised work on PP attachment is based on superficial analysis of the unlabeled corpus withoutthe use of partial parsing [].” [48]
2) “Neither (key Citation) with 67% nor [] with 59% noun attachment were anywhere close to this fig- ure.” [26]
3) “They correctly notice that approaches such as theirs, inspired by (key Citation), are based on the assumption that high 371 Computational Linguis- tics Volume 32, Number 3 co-occurrence between words is an indication of a lexical argument hood relation.” [29]

List 13. CCs undecidable

Results and Discussion

We analyzed 100 ACL Anthology articles to discover important structure types along with difficulties which exist in Citation Contexts (CCs).

Positional Characteristics:

The majority of tail-positioned references (65%) extracted entire statements for simplified extraction although it created the risk of making general statements without context. Boundary detection needs refinement when dealing with head or mid-position citations. Head CCs adopted the formula of “citation + verb + reference text” like “(key Citation) note that…” while mid CCs tended to extend over several phrases or clauses starting at a citation marker for “…results of (key Citation) indicate…” The research findings validate rules which take into account the position of textual references when extracting information.

Syntactic Scope

The sample showed that 95% of CCs contained only one sentence while phrases or clauses primarily appeared in mid-position citations (e.g., “probabilistic Earley parser (key Citation)”). Among the 5% multi-sentence CCs researchers needed help from context from neighboring sections to understand the citations (such as "This problem has been studied by (Citation)"). The analysis requires sentence-based heuristics for most citation cases yet requires multi-sentence processing for CCs spanning across different sentences.

Ambiguity and Noise

Human readers would find 15% of the CCs ambiguous because the reference spans remained unclear to them. The CC “corpus-based approaches… (key Citation)” either refers to one complete sentence or just the selected few words within that sentence. Tail citations contained machine-specific references like “comparable results (key Citation)” which made them difficult for automation processes. Researcher intervention should mark these CCs as a quality control measure for better data improvement.

Incomplete Information

About 10% of CCs did not include all necessary information in their references which demanded readers to consult external documents or reviewed publications (example: "They duplicated the experiment of (key Citation)."). Most of these rare cases carried out multiple citation comparisons but failed to specify the underlying methodology which required generalizing the findings.

Comparative Insights and Future Work

Our study confirms previous research results about single-sentence CC dominance [11][25] while challenging the previous fixed-window extraction methods [2][3][4]. The precise alternative to reference classification comes from position-based categorization methods. Future research will explore the potential use of transformer models like BERT for automated taxonomy classification and will test this approach in biomedical and social science fields and others.

Conclusion

This research focused on the fundamental problem of enhancing reference text extraction by systematic classification of citation contexts (CCs). Drawing on 100 ACL Anthology papers, we established a taxonomy categorizing CCs according to their positional, syntactic, and contextual properties. The results show that tail-position citations (65%) are often citing full statements, whereas head/mid citations demand exact boundary specification. A majority of CCs consisted of single sentences (95%), with mid-position citations often comprising phrases or clauses. Surprisingly, 15% of the CCs revealed ambiguity that confounded even human annotators and highlighted the vulnerabilities of one-size-fits-all extraction approaches. The new system facilitates more effective reference extraction based on position-aware rules and cleaner dataset building via the filtering of ambiguous or insufficient CCs. These improvements pay dividends directly back to NLP applications such as citation sentiment analysis and classification. However, focusing on single-sentence CCs in computational linguistics literature limits the study's generalizability to multiple-sentence contexts as well as other fields.

Follow-up research must generalize this taxonomy to multi-sentence CCs and various research areas while constructing automated classification based on transformer models such as BERT. Syntactic and semantic features combined could further enhance hybrid extraction mechanisms. This research provides a foundation for making a transition from heuristic to systematic CC analysis, with broad implications for bettering scholarly NLP systems—from literature mining to smart citation recommendation systems.

References

[1] B. Aljaber, N. Stokes, J. Bailey, and J. Pei, “Document clustering of scientific texts using citation contexts,” *Information Retrieval*, vol. 13, pp. 101–131, 2010.

[2] S. Bradshaw, “Reference directed indexing: Redeeming relevance for subject search in citation indexes,” in *International conference on theory and practice of digital libraries.* Springer, 2003, pp. 499–510.

[3] R. Moro, M. Vangel, and M. Bielikova, “Identification of navigation lead candidates using citation and co-citation analysis,” in *SOFSEM 2016: Theory and Practice of Computer Science: 42nd International Conference*

on Current Trends in Theory and Practice of Computer Science, Harrachov, Czech Republic, January 23-28, 2016, Proceedings 42. Springer, 2016, pp. 556–568.

[4] O. H. Bingol and M. Doslu, "Content sensitive document ranking method by analyzing the citation contexts," Dec. 18 2018, uS Patent 10,157,225.

[5] J. O'Connor, "Citing statements: Computer recognition and use to improve retrieval," *Information Processing & Management*, vol. 18, no. 3, pp. 125–131, 1982.

[6] A. Ritchie, S. Teufel, and S. Robertson, "How to find better indexterms through citations," in *Proceedings of the workshop on how can computational linguistics improve information retrieval?*, 2006, pp. 25–32.

[7] A. Ritchie, S. Robertson, and S. Teufel, "Comparing citation contexts for information retrieval," in *Proceedings of the 17th ACM conference on Information and knowledge management*, 2008, pp. 213–222.

[8] S. Liu, C. Chen, K. Ding, B. Wang, K. Xu, and Y. Lin, "Literature retrieval based on citation context," *Scientometrics*, vol. 101, pp. 1293– 1307, 2014.

[9] A. Abu-Jbara and D. Radev, "Reference scope identification in citing sentences," in *Proceedings of the 2012 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, 2012, pp. 80–90.

[10] A. Athar and S. Teufel, "Context-enhanced citation sentiment detection," in *Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, 2012, pp. 597–601.

[11] I.-S. Kang and B.-K. Kim, "Characteristics of citation scopes: A preliminary study to detect citing sentences," in *Conference on Education and Learning*. Springer, 2012, pp. 80–85.

[12] D. Duma, C. Sutton, and E. Klein, "Context matters: Towards extracting a citation's context using linguistic features," in *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, 2016, pp. 201–202.

[13] A. Abu-Jbara, J. Ezra, and D. Radev, "Purpose and polarity of citation: Towards nlp-based bibliometrics," in *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 2013, pp. 596–606.

[14] Q. He, D. Kifer, J. Pei, P. Mitra, and C. L. Giles, "Citation recommendation without author supervision," in *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011, pp. 755–764.

[15] W. Huang, Z. Wu, C. Liang, P. Mitra, and C. Giles, "A neural probabilis- tic model for context based citation recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015.

[16] C. Caragea, F. Bulgarov, A. Godea, and S. D. Gollapalli, "Citation-enhanced keyphrase extraction from research papers: A supervised approach," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1435–1446.

[17] X. Zhu, P. Turney, D. Lemire, and A. Vellino, "Measuring academic influence: Not all citations are equal," *Journal of the Association for*

Information Science and Technology, vol. 66, no. 2, pp. 408–427, 2015.

[18] T. Chakraborty and R. Narayanam, "All fingers are not equal: Intensity of references in scientific articles," *arXiv preprint arXiv:1609.00081*, 2016.

[19] S. Tuarob, P. Mitra, and C. L. Giles, "A classification scheme for algorithm citation function in scholarly works," in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital Libraries*, 2013, pp. 367–368.

[20] S. Teufel, A. Siddharthan, and D. Tidhar, "Automatic classification of citation function," in *Proceedings of the 2006 conference on empirical methods in natural language processing*, 2006, pp. 103–110.

[21] X. Li, Y. He, A. Meyers, and R. Grishman, "Towards fine-grained citation function classification," in *Proceedings of the international conference recent advances in natural language processing RANLP 2013*, 2013, pp. 402–407.

[22] D. Pride and P. Knoth, "Incidental or influential?-challenges in automatically detecting citation importance using publication full texts," in *Research and Advanced Technology for Digital Libraries: 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessaloniki, Greece, September 18-21, 2017, Proceedings 21*. Springer, 2017, pp. 572–578.

[23] M. Valenzuela, V. Ha, and O. Etzioni, "Identifying meaningful citations." in *AAAI workshop: Scholarly big data*, vol. 15, 2015, p. 13.

[24] D. R. Radev, P. Muthukrishnan, V. Qazvinian, and A. Abu-Jbara, "The acl anthology network corpus," *Language Resources and Evaluation*, vol. 47, pp. 919–944, 2013.

[25] J. Weeds and D. Weir, "Co-occurrence retrieval: A flexible framework for lexical distributional similarity," *Computational Linguistics*, vol. 31, no. 4, pp. 439–475, 2005.

[26] M. Volk, "How bad is the problem of pp-attachment? a comparison of english, german and swedish," *University of Zurich*, 2006.

[27] G. Schneider, K. Kaljurand, F. Rinaldi, and T. Kuhn, "Pro3gres parser in the conll domain adaptation shared task," *University of Zurich*, 2007.

[28] M. Lapata and C. Brew, "Using subcategorization to resolve verb class ambiguity," in *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[29] P. Merlo and E. E. Ferrer, "The notion of argument in prepositional phrase attachment," *Computational Linguistics*, vol. 32, no. 3, pp. 341– 378, 2006.

[30] A. Khalid, F. A. Khan, M. Imran, M. Alharbi, M. Khan, A. Ahmad, and G. Jeon, "Reference terms identification of cited articles as topics from citation contexts," *Computers & Electrical Engineering*, vol. 74, pp. 569–580, 2019.

[31] O. Abend and A. Rappoport, "Fully unsupervised core-adjunct argument classification," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 226–236.

[32] H. Wu and T. Furugori, "Prepositional phrase attachment through a hybrid disambiguation model," in *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*, 1996.

[33] K. Erk and S. Padó , "Paraphrase assessment in structured vector space: Exploring parameters and datasets," in *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, 2009, pp. 57–65.

[34] E. Wehrli, V. Seretan, and L. Nerima, "Sentence analysis and collocation identification," in *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, 2010, pp. 28–36.

[35] P. Pantel and D. Lin, "An unsupervised approach to prepositional phrase attachment using contextually similar words," in *Proceedings of the 38th annual meeting of the Association for Computational Linguistics*, 2000, pp. 101–108.

[36] K. Bicknell and R. Levy, "A model of local coherence effects in human sentence processing as consequences of updates from bottom-up prior to posterior beliefs," in *Proceedings of human language technologies: The 2009 annual conference of the north American chapter of the association for computational linguistics*, 2009, pp. 665–673.

[37] B. Roark, "Probabilistic top-down parsing and language modeling," *Computational linguistics*, vol. 27, no. 2, pp. 249–276, 2001.

[38] Y. W. Wong and R. Mooney, "Learning for semantic parsing with statistical machine translation," in *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, 2006, pp. 439–446.

[39] H. Guo and X. Wang, "Maximum entropy based chinese-japanese word alignment," in *2005 International Conference on Natural Language Processing and Knowledge Engineering*. IEEE, 2005, pp. 437–442.

[40] M.-J. Nederhof and G. Satta, "Prefix probability for probabilistic synchronous context-free grammars," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 460–469.

[41] M. Kayaalp, T. Pedersen, and R. Bruce, "A statistical decision making method: A case study on prepositional phrase attachment," in *CoNLL97: Computational Natural Language Learning*, 1997.

[42] M. Lapata, "Acquiring lexical generalizations from corpora: A case study for diathesis alternations," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, 1999, pp. 397–404.

[43] M. Volk and F. Tidstro¨m, "Comparing french pp-attachment to english, german and swedish," in *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*, 2007, pp. 191–198.

[44] A. Ratnaparkhi, "Statistical models for unsupervised prepositional phrase attachment," *arXiv preprint cmp-lg/9807011*, 1998.

[45] M. Lapata, "The disambiguation of nominalizations," *Computational Linguistics*, vol. 28, no. 3, pp. 357–388, 2002.

[46] P. Gamallo, A. Agustini, and G. Lopes, "Using co-composition for acquiring syntactic and semantic subcategorisation," in *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*, 2002, pp. 34–41.

[47] T. Deoskar, "Re-estimation of lexical parameters for treebank pcfgs," in *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 2008, pp. 193–200.

[48] M. Atterer and H. Schu¨tze, "The effect of corpus size in combining supervised and unsupervised training for disambiguation," in *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, 2006, pp. 25–32.