ISSN 1870-4069

Investigating Citation Linkage as an Information Retrieval Task

Hospice Houngbo, Robert E. Mercer

The University of Western Ontario,
Department of Computer Science,
Canada

hhoungbo@uwo.ca, mercer@csd.uwo.ca

Abstract. Scientists have a deluge of literature to consult in their research work. To navigate this overabundance of information, tools such as citation indexes help but do not indicate the precise passage in the paper that is being cited. In this study, we report our method for finding those sentences in a cited article that are the focus of a citation in a citing paper, a task we have called *citation linkage*. We first provide our guidelines for building a corpus annotated by a domain expert. The corpus consists of citing sentences and their cited articles. For this study the citing sentences deal with biochemistry methodology. All sentences in the cited article are annotated with six levels of relevance ranging from 0 (no relevance), to 5 (the annotator had the highest confidence that the sentence is relevant). We hypothesize that citation sentences when used as queries in a retrieval model should point to relevant sentences in the cited paper. To evaluate this hypothesis, a number of established retrieval algorithms using various document ranking methods are compared using information retrieval evaluation metrics. For each citation-paper linkage task, we compute Precision@k and Normalized Discounted Cumulative Gain, NDCG@k, where k is the number of sentences given non-zero relevance scores by the annotator. We found that 18 out of 22 citation linkage tasks have at least one sentence in the top k positions. The best Average NDCG is 49% and the best Average Precision is 50%.

Keywords: Linkage citations, syntagma, citations recovery algorithms.

1 Introduction

The writer of a research paper is required to place its contribution in its research context. This is often done by means of *citations*, which are instruments for connecting ideas in the research literature. The importance that citations play in the research literature has led to a variety of tools to assist researchers.

For instance, citation indexes, an idea conceived in 1964 [9], contain a subset of all of the citations in research articles. More recently, methods have been proposed to classify the purpose of a citation [12, 24]. Citation analysis-based bibliometrics are used to assess research and researcher importance [10, 11]. Potential uses of citations include multiple article summarization [6, 5, 21] and the tracking of scientific argumentation [18, 20] across multiple papers.

The purpose of a citation is typically to highlight a given aspect of the work described in the cited paper. However, citations in scientific writing refer to papers rather than the much smaller text span that is being referred to. Reducing citation targets from papers to a briefer text span such as a paragraph, a sentence, or a set of sentences, would be beneficial for some of the more complex applications mentioned above. This work, therefore, aims at investigating how to determine those sentences in a cited article that are the focus of a citation in a citing paper. We call this operation *citation linkage* and investigate it as an information retrieval problem.

We thus hypothesize that matching a citation sentence to relevant sentences in the cited paper can be compared to how search engines rank documents based on a user's specific query. In this study, each article is segmented into sentences and each sentence is compared to the citation sentence using a retrieval algorithm. The position of each candidate sentence is determined by ranking the similarity scores with regard to the citation sentence. The justification for such an hypothesis can be drawn from the necessary requirements for text retrieval experiments that the linkage task satisfies, such as:

- Information need: Retrieval from a given article, the sentences that match a specific citation sentence[1].
- Test collections: 22 articles, each containing relevant and non relevant sentences with multi level judgments.
- Evaluation methods: We can use ranked retrieval evaluation measures, such as Precision and Normalized Discounted Cumulative Gain (NDCG).

The rest of this paper is organized as follow. The next section reviews some related works. In Section 3, a detailed methodology of the citation linkage task is presented. In Section 4, the results are described. We conclude the paper with a summary and directions for future work.

2 Related Work

Finding the best linkage candidates (i.e., the sentences having more or less the same "contentful" meaning as a given citation sentence) will amount to computing the degree of similarity between the citation sentence and each sentence in the citing paper. Previous works on text similarity detection include paraphrase detection [8] and textual entailment [7]. Paraphrasing methods recognize, generate, or extract (e.g., from corpora) paraphrases, phrases, sentences or longer units of text that convey the same, or "almost" the same information.

Textual entailment methods, recognize, generate, or extract pairs (T, H) of natural language expressions, such that a human who reads (and trusts) T would infer that H is most likely also true [7].

[1] Text (or information) retrieval normally speaks of retrieving documents from a collection of documents. Our purpose here is to retrieve sentences from articles. Hence, we will simply substitute "sentence" for "document" and "article" for "collection" whenever the latter words are typically used in the text (or information) retrieval literature.

Both paraphrase detection and textual entailment have been combined in the SemEval2012 Semantic Textual Similarity shared task [2], consisting of finding similarity between sentences in a text pair (T1 and T2) and returning a similarity score and an optional confidence score. The authors participating in the tasks used a combination of lexical and syntactical approaches as well as machine learning approaches.

Nakov et al., [19] proposed the use of the text surrounding citations as tools for semantic interpretation of bioscience text. This work emphasized several different uses of citation sentences and showed that citation sentences are rich in domain specific concepts and terminology. The work in [19] aims at automatically extracting paraphrases of facts about a cited paper from multiple citations to it, with the eventual goal of using these sentences to automatically create summaries of the cited paper.

This is in line with Small's [23] notion of cited works as concept symbols, whereby a work may come to be repeatedly and consistently cited to represent a specific idea or topic, using descriptions that converge on an almost fixed terminology for that topic, such that the cited work eventually becomes synonymous with the topic. In [5], an attempt to match citation text and cited spans in biomedical literature proved to be a difficult task with limited performance.

3 Methodology

3.1 Building of a Citation Linkage Corpus

The text in which a citation occurs can span one or more sentences in the paper. In this study, this span is limited to one sentence, and the linkage task is assumed to be a sentence-level matching operation. An annotation guideline was defined to match a given citation sentence with candidate cited sentences based on the following criteria:

- To what extent can the person who reads a citation sentence taken in isolation be able to determine the candidate sentences that have been cited in a reference paper.
- Possible candidate sentences are chosen from the full article and presented chronologically as they appear in the article, thereby providing some context for the candidate sentence.
- For each sentence in the cited paper, a score will be given. This score will indicate the confidence that the annotator had in making his/her choice of candidate sentences. For those sentences not chosen as candidate sentences, a score of 0 is given indicating that the annotator is confident that there is no similarity in content with the citing sentence. For those sentences chosen as candidates, a score is given ranging from 1 (low confidence that similarity in content exists) to 5 (the annotator is confident that there is strong similarity between the candidate and citing sentences).

We have chosen the papers to annotate in such a way that they belong to a citation network that shows the links between cited and citing papers. We limit the current research to the biomedical domain and our final corpus is curated with papers from this domain. For this purpose, we use the BioMed Central's research articles corpus which is an open access corpus ideally suited for data mining research.

Table 1. Example of an annotation.

Citation Sentence	
We have been able to amplify 200 bp fragments of DNA obtained from Bouin's fixed and paraffin wax embedded tissues only after a specific restoration method to produce longer reconstructed DNA fragments.	
Candidate Sentences	**Rating**
To obtain longer stretches of DNA, a pre-PCR restoration treatment was required, by filling single strand breaks, followed by a vigorous denaturation step.	4
The development of this simple treatment allowed the analysis of longer fragments of DNA obtained from archival postmortem paraffin wax embedded tissues.	3
A partial restoration and reconstruction of DNA length in these cases is possible.	1
Here , we show that it is possible to analyze human postmortem paraffin wax embedded tissues amplifying a 287 bp sequence of apolipoprotein E (ApoE) and 291 bp of the prealbumin gene (TTR).	3
DNA was extracted from 6 m sections of paraffin wax embedded tissues.	1
The final sample of DNA was obtained by precipitation with ethanol using glycogen as the carrier.	1
DNA samples were incubated for one hour at 55C in 100 l of solution containing 10mM Tris/HCl (pH 8.3), 1.5 mM MgCl2, 2% Triton X-100, and 200M of each dNTP.	2
After this incubation, 1 U Taq DNA polymerase (Amersham) was added and DNA polymerisation was performed at 72C for 20 minutes.	2
The polymerase reaction restores the nicks after DNA rehybridisation, using the other strand as the template.	3
We have developed a method for amplifying longer DNA sequences, ranging up to 300 bases, from postmortem formalin fixed and paraffin wax embedded tissues, with no modification to the usual DNA extraction procedures.	5
Our restoration method is based on the fact that DNA degradation results from random single strand breaks and polymerase reaction restores the nicks, using the other DNA strand as a template.	2
Our restoration method is based on the fact that DNA degradation results from random single strand breaks and PCR restores the nicks, using the other DNA strand as a template.	2
The method proposed can be used to obtain longer amplification fragments of around 300 bp of DNA from normally extracted postmortem paraffin wax embedded tissues.	4

The richness of BioMed Central's XML format also makes the content especially suitable for information extraction and textual analysis. The citation sentences are limited to sentences that contain specific terminology that may represent a set of methods, tools or techniques used in scientific experiments and refer, we believe, to sentences of a similar kind in the cited papers.

Annotators' feedback has also been collected and points to the fact that candidate sentences were chosen by the annotators based on surface level similarity as well as non explicit factors such as background domain knowledge, and inferential deduction. Table 1 presents an example of an annotation produced by a human annotator (only the non-zero rated candidate sentences and their rating scores are presented).

Table 2. Number of candidate sentences per paper.

Paper	Score					Paper	Score				
#	**1**	**2**	**3**	**4**	**5**	#	**1**	**2**	**3**	**4**	**5**
1	4	3	6	4	2	12	3	0	1	1	0
2	0	4	4	3	5	13	1	0	0	1	0
3	0	0	7	1	4	14	8	0	0	3	1
4	1	0	1	1	0	15	4	2	2	5	1
5	6	1	1	9	1	16	2	0	3	2	1
6	0	0	1	1	1	17	1	6	1	5	2
7	1	0	1	0	2	18	2	1	2	0	1
8	3	0	3	1	0	19	2	2	0	0	0
9	1	1	1	0	0	20	0	0	0	2	1
10	0	0	3	2	3	21	0	0	1	0	2
11	2	17	4	3	5	22	3	5	0	4	1

The citation sentence is from the article: *DNA and RNA obtained from Bouin's fixed tissues*; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1770606/. The candidate sentences are from the referenced article: *PCR analysis in archival postmortem tissues*; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1187316/. All 22 annotations are available at https://github.com/hospice/linkagefiles. The annotation guidelines can be found in the first author's PhD thesis [16]. Table 2 shows the number of candidate sentences per paper.

3.2 Evaluation Metrics

When the sentences retrieved by the search techniques are evaluated on a binary relevance basis, each sentence is treated as being either relevant or irrelevant. *Precision* measures how many sentences are relevant to the query among the returned sentences.

In the computation of the output of these search operations, we are interested in the proportion of relevant results among the first k retrieved results. If the search technique returns r relevant sentences in the first k sentences that it finds, then *Precision*@k is:

$$\textit{Precision}@k = \frac{r}{k}.$$

Because each paper in our corpus has a different number of candidate sentences chosen to be relevant by the annotators, we have chosen k to be this number for each paper. In some cases, the degree to which a sentence satisfies the query needs to be taken into account during the evaluation process. When the relevance of sentences in the article can be captured with more than two classes, new measures are needed to capture these degrees of relevance of each retrieved sentence.

The overall score is obtained by combining "relevance" values and the position of the sentence in a ranked search result. Such measures include the Normalized Discounted Cumulative Gain (NDCG)[17]. The NDCG evaluation measure is computed to reflect the ideal position of the very relevant, the marginally relevant, as well as the non-relevant sentences in the ranking. The Normalized Discounted Cumulative Gain (NDCG) principle can be summarized as follows:

- It is applicable to multi-level judgments in a scale of $[1, r]$, $r \geq 2$.
- It measures the total utility of the top k sentences.
- The utility of a lower ranked sentence is discounted.
- The score is normalized to assure comparability across queries.

$$DCG@k = rel_1 + \sum_{i=2}^{k} \frac{rel_i}{\log_2 i}. \quad (1)$$

rel_1 is the graded relevance of the sentence at position 1.
rel_i is the graded relevance of the sentence at position i.

$$NDCG@k = \frac{DCG@k}{IDCG@k}, \quad (2)$$

where $IDCG@k$ is the Ideal Discounted Cumulative Gain for the k ordered results.

3.3 Ranking Models

We experiment with different ranking techniques such as BM25[22], Divergence From Randomness (DFR), Vector Space Model (VSM) [3], Information Based Similarity (IBS) [13], Language Model with Jelinek-Mercer smoothing (LMJ) [25] and Language Model with Dirichlet priors smoothing (LMD)[26]. The Lucene framework [4], which has the implementation of these techniques, is used for this purpose.

3.4 Index Creation

During the indexing process, each article is segmented into sentences, as the linkage is done at the individual sentence level. Sentences are transformed into fields of content (words). An analysis step is then performed to remove stop-words which are not required for the search operations. An index writer creates indexes as required.

3.5 Search Operation

A query expression is derived either from the unmodified citation text or a reformulation of it. Here we use two types of queries: the full unmodified citation sentence and the noun phrases found in the citation sentence. A query parser operation is performed after the query terms are analyzed for stop word removal. The query expression is then passed to the searcher module which returns ranked sentences based on a ranking model.

3.6 Experiments

Experiments can be divided into two main types:

1. Linkage operations using all sentences as potential candidate cited sentences.

2. Linkage operations using a reduced set of sentences as potential candidate cited sentences (for our experiments, here, these are sentences that have been automatically marked rhetorically as Method sentences [1, 14]).

For each type, we set up two experiments: In the first set of experiments, we use the words in the citation sentences as query terms. In the second set of experiments, we reduce the query terms to the noun phrases in the citation sentences. This second set is based on an hypothesis that noun phrases may contain the important query terms (the same hypothesis is found in [5]).

4 Results and Discussion

We use two evaluation metrics for our experiments, namely: *NDCG@*k and *Precision@*k. *NDCG@*k takes into account the multi-level utility score of each sentence, whereas *Precision@*k only takes into account the binary relevance of each sentence. For a given citation–article pair, the evaluation score will depend on the number of candidate sentences, which varies depending on the paper being considered.

Therefore the computation of an overall average score for all the papers is done after we compute the individual score for each paper, taking into account different numbers of candidate sentences. For a total number of N papers, if paper$_i$ has k_i candidates, we first compute *Precision@*k_i and *NDCG@*k_i for each paper$_i$, before computing the Average Precision, *Avg. Precision*, and the Average Normalized Discounted Cumulative Gain, *Avg. NDCG*:

$$\textit{Avg. Precision} = \frac{\sum_{i=1}^{N} \textit{Precision@}k_i}{N}, \tag{3}$$

And,

$$\textit{Avg. NDCG} = \frac{\sum_{i=1}^{N} \textit{NDCG@}k_i}{N}. \tag{4}$$

Table 3 shows the results for the experiments using the full citation sentence as query input and Table 4 shows the results of the experiments using noun phases extracted from the citation sentence as query input. Besides the average scores (Mean), three summary statistics for each retrieval model are presented: minimum, maximum, and median over all the papers. Four retrieval models, VSM, IBS, DFR, LMJ and LMD, show similar performance, with BM25 having much lower Mean values.

The Mean values vary between 0.2506 and 0.3412 for all sentences as possible candidates, and between 0.4064 and 0.4976 for Method sentences as possible candidates for *Precision@k*, and between 0.2514 and 0.3247 for all sentences as possible candidates and between 0.3958 and 0.4913 for Method sentences as possible candidates for the *NDCG@k* for the experiments using full citation sentences as query input. Similarly, the Mean values vary between 0.2362 and 0.3557 and between 0.3962 and 0.4988 for *Precision@k*, and between 0.2411 and 0.3590 and between 0.3702 and 0.4886 for the *NDCG@k* for the experiments using noun phrases as query input. Therefore, models with query reduction using noun phrases have slightly better performances.

Table 3. Evaluation: queries: citations.

Possible Candidates: All sentences in cited paper					Possible Candidates: Method sentences in cited paper				
Precision@k					***Precision@k***				
	Min.	**Median**	**Mean**	**Max.**		**Min.**	**Median**	**Mean**	**Max.**
IBS	0.0000	0.3229	0.3212	0.6667	IBS	0.0000	0.4808	0.4550	1.0000
VSM	0.0000	0.3542	0.3395	0.6667	VSM	0.0000	0.5000	0.4826	1.0000
BM25	0.0000	0.2283	0.2506	0.6667	BM25	0.0000	0.4476	0.4064	1.0000
DFR	0.0000	0.3333	0.2972	0.6667	DFR	0.0000	0.4727	0.4741	1.0000
LMJ	0.0000	0.3542	0.3412	0.6667	LMJ	0.0000	0.5000	0.4854	1.0000
LMD	0.0000	0.3205	0.3309	0.6667	LMD	0.0000	0.4919	0.4976	1.0000
NDCG@k					***NDCG@k***				
	Min.	**Median**	**Mean**	**Max.**		**Min.**	**Median**	**Mean**	**Max.**
IBS	0.0000	0.3293	0.3168	0.7751	IBS	0.0000	0.4830	0.4569	1.0000
VSM	0.0000	0.3271	0.3237	0.7751	VSM	0.0000	0.5197	0.4906	1.0000
BM25	0.0000	0.2984	0.2514	0.7654	BM25	0.0000	0.4644	0.3958	1.0000
DFR	0.0000	0.3222	0.3042	0.7751	DFR	0.0000	0.4906	0.4722	1.0000
LMJ	0.0000	0.3412	0.3247	0.7751	LMJ	0.0000	0.5074	0.4798	1.0000
LMD	0.0000	0.2961	0.3227	0.7751	LMD	0.0000	0.5163	0.4913	1.0000

Table 4. Evaluation: queries: noun phrases.

Possible Candidates: All sentences in cited paper					Possible Candidates: Method sentences in cited paper				
Precision@k					***Precision@k***				
	Min.	**Median**	**Mean**	**Max.**		**Min.**	**Median**	**Mean**	**Max.**
IBS	0.0000	0.3542	0.3557	1.0000	IBS	0.0000	0.4722	0.4451	1.0000
VSM	0.0000	0.3333	0.3476	1.0000	VSM	0.0000	0.5000	0.4799	1.0000
BM25	0.0000	0.1905	0.2362	0.6667	BM25	0.0000	0.4446	0.3962	1.0000
DFR	0.0000	0.3333	0.3386	1.0000	DFR	0.0000	0.4365	0.4540	1.0000
LMJ	0.0000	0.3333	0.3518	1.0000	LMJ	0.0000	0.4643	0.4704	1.0000
LMD	0.0000	0.3095	0.3114	0.6842	LMD	0.0000	0.5000	0.4988	1.0000
NDCG@k					***NDCG@k***				
	Min.	**Median**	**Mean**	**Max.**		**Min.**	**Median**	**Mean**	**Max.**
IBS	0.0000	0.3120	0.3375	1.0000	IBS	0.0000	0.4464	0.4207	1.0000
VSM	0.0000	0.3040	0.3376	1.0000	VSM	0.0000	0.4748	0.4514	1.0000
BM25	0.0000	0.2224	0.2411	0.7654	BM25	0.0000	0.3986	0.3702	1.0000
DFR	0.0000	0.3024	0.3218	1.0000	DFR	0.0000	0.4450	0.4321	1.0000
LMJ	0.0000	0.3206	0.3590	1.0000	LMJ	0.0000	0.4596	0.4371	1.0000
LMD	0.0000	0.2662	0.2932	0.7654	LMD	0.0000	0.4862	0.4886	1.0000

But this doesn't generalize to every paper as demonstrated in Table 5 which shows, as an example, the per paper results for the four experiment types for the Language Model with Jelinek-Mercer smoothing. The overall statistics of the ranked candidate sentences in the top k are presented in Table 6.

We can notice that fewer 5s and 4s have been ranked as a 0 compared to 3s, 2s and 1s. We think that the 1s, 2s, and 3s may require biochemical knowledge and more background information for the linkage to be effective.

Table 5. Evaluation per paper using language model with Jelinek-Mercer smoothing.

Paper #	Numb. of sents.	Numb. of Method sents.	k	Sentences as Candidates				Method Sentences as Candidates			
				Citation as query		NP as query		Citation as query		NP as query	
				Precision @k	NDCG @k	Precision @k	NDCG @k	Precision @k	NDCG @k	Prec. @k	NDCG @k
1	126	59	19	0.5263	0.5027	0.6842	0.7135	0.6316	0.6741	0.6842	0.7339
2	166	65	16	0.375	0.3547	0.3125	0.2716	0.6875	0.5847	0.625	0.5004
3	150	59	12	0.5833	0.4874	0.4167	0.3792	0.6667	0.5676	0.6667	0.5676
4	162	24	3	0.6667	0.7654	1	1	1	1	1	1
5	194	60	18	0.0556	0.0506	0.2222	0.2828	0.5556	0.576	0.5	0.5344
6	185	44	3	0	0	0.3333	0.2346	0	0	0.3333	0.2346
7	169	53	4	0.5	0.4144	0.25	0.1952	0.5	0.6367	0.25	0.1952
8	291	92	7	0.4286	0.2042	0.4286	0.4791	0.5714	0.4791	0.5714	0.4791
9	233	97	3	0.3333	0.2961	0.3333	0.2961	0.3333	0.2961	0.3333	0.2961
10	224	65	8	0.625	0.7031	0.625	0.7134	0.625	0.4785	0.625	0.4974
11	315	93	31	0.2581	0.3525	0.2258	0.2911	0.4516	0.5278	0.3548	0.4487
12	89	32	5	0.4	0.3452	0.4	0.3452	0.4	0.4704	0.4	0.4704
13	239	47	2	0	0	0	0	0	0	0	0
14	236	103	12	0.25	0.2361	0.25	0.2149	0.4167	0.3857	0.3333	0.2618
15	249	40	14	0.1429	0.1132	0.1429	0.1116	0.4286	0.379	0.4286	0.3774
16	189	99	8	0.25	0.3373	0.375	0.4171	0.5	0.5869	0.625	0.5591
17	112	53	15	0.4	0.4839	0.6	0.6498	0.5333	0.5908	0.7333	0.7721
18	143	65	6	0.6667	0.7751	0.5	0.4593	0.6667	0.7619	0.5	0.429
19	185	62	4	0	0	0	0	0.25	0.2463	0	0
20	165	39	3	0.3333	0.2346	0.3333	0.4693	0.6667	0.5307	0.6667	0.5307
21	266	119	3	0	0	0	0	0.3333	0.2961	0.3333	0.2961
22	170	83	13	0.4615	0.4862	0.3077	0.3733	0.4615	0.4869	0.3846	0.4321

Table 6. Statistics of the ranked candidate sentences over all the papers.

Score	Number Ranked	Number Expected	Percent
5	20	33	(60 %)
4	21	48	(44 %)
3	18	42	(43 %)
2	6	42	(14 %)
1	8	44	(18 %)

60% and 44% of the 5s and 4s are ranked respectively in the top positions. Paper 13 did not yield a candidate sentence for *Precision*@k and *NDCG*@k. This can be due to the following reasons:

1. The number of relevant sentences is very low: 2, whereas the average number of candidate sentences for a paper is 8.
2. The choice of most candidate sentences may involve the use of some inferential information that is not easily translated into the retrieval models at this stage of the study.
3. It is difficult to find the best match for some citations when the matching operation involves external domain specific resources that are not yet available.

Despite these shortcomings, our results show that it is possible to find the set of sentences a citation refers to in a cited paper with reasonable performance.

4.1 Comparison with Other Work

Compared to previous attempts to devise a framework for matching citation text and cited spans in research literature, this current work is novel in certain aspects. The dataset used in [6, 5] is divided into twenty topics, each of which comprises a reference paper and a set of related citing papers. In our work, the citation linkage objective is the same, but we looked at a single citing sentence per paper. Also, we focussed on one particular type of citation, the Method rhetorical category.

This allowed an assessment of a candidate sentence reduction method based on this rhetorical label. The linkage task in this work is a sentence-level matching operation which is intended to avoid the reference text span boundary problem noted in [6] and [5], which is a non-trivial task on its own. And we have assessed the linkage relation on a graded basis rather than a binary classification.

The authors of [6] and [5] reported promising results with query reduction to noun phrases and UMLS expansion. We also have found some, but not general, improvement with the reduction of queries to noun phrases, but we abandoned (possibly prematurely) including associated terminology from thesauri-like sources such as UMLS. We also note our machine learning view of the citation linkage task [15].

5 Conclusion

The citation linkage task aims at finding those sentences in a cited article that are the focus of a citation in a citing paper. We have shown that one way to achieve this goal is to treat the problem as an information retrieval task. While most retrieval techniques usually apply to a large collection of documents, they all involve text matching based on document content similarity. The same matching operation can be achieved at the sentence level as is in the case of the linkage between a citation sentence and its cited sentences in a cited article.

Our results show that the information need that the linkage task tends to achieve can be realized using search engine-like retrieval techniques. We notice that most of the retrieval algorithms provide some good results for the majority of the linkage tasks. However, we need to further investigate why some papers don't yield any relevant sentences. Our hypothesis is that these papers might require domain specific information for the linkage to be achieved. We intend to investigate how to include these resources in the ranking models in future studies.

References

1. Agarwal, S., Yu, H.: Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. Bioinformatics, vol. 25, no. 23, pp. 3174–3180 (2009) doi: 10.1093/bioinformatics/btp548

2. Agirre, E., Diab, M., Cer, D., Gonzalez-Agirre, A.: SemEval-2012 task 6: A pilot on semantic textual similarity. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, pp. 385–393 (2012)
3. Amati, G., Van Rijsbergen, C. J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems, vol. 20, no. 4, pp. 357–389 (2002) doi: 10.1145/582415.582416
4. Białecki, A., Muir, R., Ingersoll, G.: Apache Lucene 4. In: Proceedings of the SIGIR 2012 Workshop on Open Source Information Retrieval, pp. 17–24 (2012)
5. Cohan, A., Soldaini, L., Goharian, N.: Matching citation text and cited spans in biomedical literature: A search-oriented approach. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1042–1048 (2015) doi: 10.3115/v1/N15-1110
6. Cohan, A., Soldaini, L., Mengle, S. S., Goharian, N.: Towards citation-based summarization of biomedical literature. In: Proceedings of the Text Analysis Conference (2014)
7. Dagan, I., Glickman, O., Magnini, B.: The PASCAL recognising textual entailment challenge. In: Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment, vol. 3944, pp. 177–190 (2006) doi: 10.1007/11736790_9
8. Dolan, B., Quirk, C., Brockett, C.: Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In: Proceedings of the 20th International Conference on Computational Linguistics, pp. 350–356 (2004) doi: 10.3115/1220355.1220406
9. Garfield, E.: Science citation index–A new dimension in indexing: This unique approach underlies versatile bibliographic systems for communicating and evaluating information. Science, vol. 144, no. 3619, pp. 649–654 (1964) doi: 10.1126/science.144.3619.649
10. Garfield, E.: Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. Science, vol. 178, no. 4060, pp. 471–479 (1972) doi: 10.1126/science.178.4060.471
11. Garfield, E.: Is citation analysis a legitimate evaluation tool? Scientometrics, vol. 1, no. 4, pp. 359–375 (1979) doi: 10.1007/BF02019306
12. Garzone, M., Mercer, R. E.: Towards an automated citation classifier. In: Conference of the Canadian Society for Computational Studies of Intelligence, vol. 1822, pp. 337–346 (2000) doi: 10.1007/3-540-45486-1_28
13. Harter, S. P.: A probabilistic approach to automatic keyword indexing. Part II. An algorithm for probabilistic indexing. Journal of the American Society for Information Science, vol. 26, no. 5, pp. 280–289 (1975) doi: 10.1002/asi.4630260504
14. Houngbo, H., Mercer, R. E.: An automated method to build a corpus of rhetorically-classified sentences in biomedical texts. In: Proceedings of the First Workshop on Argumentation Mining, pp. 19–23 (2014) doi: 10.3115/v1/W14-2103
15. Houngbo, H., Mercer, R. E.: Investigating citation linkage with machine learning. In: Proceedings of the 30th Canadian Conference on Artificial Intelligence, pp. 78–83 (2017) doi: 10.1007/978-3-319-57351-9_10
16. Houngbo, H. K.: Investigating citation linkage between research articles. Ph.D. thesis, Electronic Thesis and Dissertation Repository, The University of Western Ontario (2017)
17. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems, vol. 20, no. 4, pp. 422–446 (2002) doi: 10.1145/582415.582418
18. Mercer, R.: Locating and extracting key components of argumentation from scholarly scientific writing. vol. 6, no. 4, pp. 3–15 (2016)

19. Nakov, P., Schwartz, A., Hearst, M. A.: Citances: Citation sentences for semantic analysis of bioscience text. In: Proceedings of the SIGIR'04 Workshop on Search and Discovery in Bioinformatics (2004)
20. Palau, R. M., Moens, M. F.: Argumentation mining: The detection, classification and structure of arguments in text. In: Proceedings of the 12th International Conference on Artificial Intelligence and Law, pp. 98–107 (2009) doi: 10.1145/1568234.1568246
21. Radev, D. R., Jing, H., Budzikowska, M.: Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In: Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization, vol. 4, pp. 21–30 (2000) doi: 10.3115/1117575.1117578
22. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. vol. 3, no. 4, pp. 333–389 (2009) doi: 10.1561/1500000019
23. Small, H.: Cited documents as concept symbols. Social Studies of Science, vol. 8, no. 3, pp. 327–340 (1978) doi: 10.1177/0306312778008003
24. Teufel, S., Siddharthan, A., Tidhar, D.: Automatic classification of citation function. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp. 103–110 (2006)
25. Zhai, C.: Statistical language models for information retrieval. Synthesis Lectures on Human Language Technologies (2009) doi: 10.1007/978-3-031-02130-5
26. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to Ad Hoc information retrieval. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 334–342 (2001) doi: 10.1145/383952.384019