Evaluating Semantic Search Strategies for Retrieving Sources for a Claim

Pavan Rauch
The Ohio State University
rauch.139@osu.edu

Abstract

Scientific knowledge is stored as natural language within scientific papers. This unstructured format is difficult to search over, a fact which hinders the development of systems that find papers which discuss a given claim. As this task has important applications, including countering misinformation, it is important to find ways to efficiently search for facts within scientific papers. We investigate the application of semantic searches using sentence embedding models to this task and compare variations on this approach to a baseline method of searching keywords. We find that semantic search approaches improve upon the accuracy of keyword searches and that controlling the context present in extracted claims is critical for effective searches.

1 Introduction

The current practice of science relies on research papers as the method of storage for scientific findings. This natural language format is able to accommodate a wide range of possible facts; however, as an unstructured format, it is difficult for algorithms to process. There are several scenarios in which algorithmically searching facts within the corpus of scientific knowledge would be useful. These include locating prior work during the research process, automatically verifying claims made on social media, and building structured databases that link human knowledge. The natural language format of scientific research impedes the development of these technologies.

At present, scientific articles are commonly searched using keywords. While this approach is able to generalize to any document, it is only accurate when the searcher is aware of the terminology that the article is likely to use. This may be true of academics searching for papers in their field but is unlikely for the use case of online fact-checking. To overcome this limitation, recent research has explored the use of semantic search algorithms which embed excerpts of the article text as vectors and then compare a query's embedded vector to the excerpts using a similarity metric.

Within scientific article retrieval tasks, claims are the key unit of linkage between queries and articles. The task can be thought of as such: given a claim, can an algorithm be written to reliably retrieve scientific articles that are relevant to the claim? As claims often contain different wording than scientific writing, semantic search is a promising approach for achieving high accuracy on this task. However, the best practices for embedding an article's claims are not clear, as it is difficult to interpret how meaning is stored within an embedded vector. One important question must be answered: how should article text be preprocessed prior to embedding? Answering this requires an evaluation method that is able to measure the efficacy of different excerpt extraction procedures for the semantic search document retrieval task.

2 Related Work

Prior research on retrieving sources for a claim has largely focused on claim verification. The focus of the verification task is to return relevant documents and mark whether they support or refute a claim's truth value; this last aspect is not relevant to this study. Scientific articles often contain nuanced perspectives on facts which might be misaligned with the way they are presented outside of the scientific domain. Flattening each article to a support-refute dichotomy is less useful than providing the original article, which contains the full nuance of the finding. Work in claim verification is still relevant to this study as it pertains to the challenge of claim extraction and document retrieval.

Claim extraction is the task of identifying well-formatted claims from a section of text. This challenge is nontrivial, as not all sentences contain

claims, and many sentences containing a claim are formatted in such a way that removing them from context would make their meaning ambiguous. The first step of solving this problem is to define what the system will consider to be a claim; existing datasets tend to use different definitions [1]. The AIDA format - standing for atomic, independent, declarative, and absolute - is a useful framework for describing well-formatted claims [2]. If extracted from an article's text, AIDA claims provide necessary and sufficient context for semantic search. While rule-based methods for extracting AIDA claims have been studied [3], recent natural language processing (NLP) developments have yet to be applied to the AIDA format in particular.

Recent claim extraction research utilizes the abilities of Large Language Models (LLMs) to manipulate text according to natural language commands. These systems usually take the form of pipelines of data where the LLM is asked to modify claims it has extracted until they are properly decontextualized. Existing systems decontextualize by prompting the LLM to generate and answer questions about ambiguous claims [4][5] or by prompting the LLM to follow multiple steps to remove defined classes of ambiguity from the claims [6][7].

Document retrieval is central to the efficacy of claim verification systems. Existing systems tend to split retrieval into two steps: first, a keyword-matching approach such as BM25 is used to create a list of documents most likely to match the claim; then, a transformer is used to evaluate whether each document supports, refutes, or is irrelevant to the claim [8][9][10]. This split approach is effective for tasks in which truth evaluation is the central challenge, but does not apply well when the claims and documents share few keywords.

In order to apply semantic search to the document retrieval problem, a language model must be chosen with which to embed claims into vectors. Sentence-BERT [11] was chosen for this study for its ability to encode the meanings of text longer than a single word. Training an embedding model for the scientific domain is possible [12][13], though these approaches are often less generalizable to new data and require significant effort to train [14].

Many existing datasets can be used to evaluate the accuracy of the document retrieval task. Of these, FEVER, which matches claims to multiple Wikipedia articles that can be used to prove or disprove them [15], is the most widely used today. SciFact-Open [16], an extension of the SciFact dataset, contains claims matched to scientific abstracts that support or refute them. Manual inspection revealed that neither of these datasets was satisfactory for the document retrieval task specified in this study. Claims in FEVER largely contain proper nouns such as celebrity names and locations, while claims in SciFact-Open often contain highly technical biomedical terminology. These keywords make it easy to match claims to documents in these datasets based on word matching alone. While this is acceptable for a claim verification task, it makes the datasets inapplicable to a context in which claim-writers' lack knowledge of the scientific terminology relevant to their claim.

3 Dataset

A new dataset was constructed in order to better evaluate the accuracy of semantic search procedures for retrieving documents for claim-writers unaware of relevant scientific terminology. Citations on Wikipedia were used to build this dataset.

A list of Wikipedia articles to search was first acquired by parsing the Wikipedia article "List of common misconceptions", which lists topics that have commonly-spread misconceptions about them. This method was chosen to narrow down the set of topics to those that are commonly discussed outside of any particular field and for whom claims are commonly circulated by the general public. The Wikipedia article pertaining to each of these topics was then parsed using regular expressions for its

[1] https://aclanthology.org/D17-1218.pdf
[2] https://arxiv.org/pdf/1303.2446
[3] https://arxiv.org/pdf/1707.07678
[4] https://arxiv.org/pdf/2406.03239v1
[5] https://arxiv.org/html/2407.18367v1
[6] https://arxiv.org/pdf/2502.10855
[7] https://arxiv.org/pdf/2102.05169
[8] https://dl.acm.org/doi/pdf/10.1145/3485127
[9] https://assets-eu.researchsquare.com/files/rs-3007151/v1_covered_28f6a157-717f-436a-8251-1179e6f31ed7.pdf?c=1722246935
[10] https://arxiv.org/pdf/2210.13777
[11] https://arxiv.org/pdf/1908.10084
[12] https://www.nature.com/articles/s41597-019-0055-0
[13] https://arxiv.org/pdf/1903.10676
[14] https://assets-eu.researchsquare.com/files/rs-3007151/v1_covered_28f6a157-717f-436a-8251-1179e6f31ed7.pdf?c=1722246935
[15] https://arxiv.org/pdf/1803.05355v3
[16] https://arxiv.org/pdf/2210.13777

citations, keeping only citations that include a DOI address. Each of these links was stored alongside the span of text in the Wikipedia article that cited it. Abstracts were then matched to these DOI citations using the CrossRef API. Citations that lacked an accessible abstract - approximately two thirds of all abstracts - were removed. Abstracts were chosen instead of the entire article because abstracts are likely to contain the most important findings of a study and are easier to access via public APIs.

The dataset was finally split into "queries" - spans of Wikipedia articles that ended in one or more citations - and "documents" - abstracts of the papers cited by those spans. Each query could cite one or more documents and each document could be cited by zero or more queries. Queries under 75 characters and over 175 characters, and queries containing hyperlinks and formatting characters, were removed from the dataset, as they are unlikely to contain valid claims. Similarly, abstracts with less than 50 characters were removed from the dataset, as these examples were usually the result of parsing errors. The final dataset contained 2329 queries, each likely to contain a claim, and the 4876 abstracts they were matched to. As shown in 1, most documents had no queries matched to them; only about 41% had a matched query. As shown in 2, the vast majority of queries were matched to exactly one document.

The Wikipedia excerpts in the dataset were parsed by selecting spans that ended in a citation and started either after a different citation or at the beginning of a paragraph. Because citations can occur in the middle of a sentence, many of these excerpts are sentence fragments. The excerpts, when taken out of context, often lack critical context about the topic they represent. Additionally, many sentences are sourced from claims found in the body of the cited paper rather than in the paper's abstract. These limitations increase the difficulty of making correct matches on the dataset. This may not be a flaw; real-world application of source retrieval run into similar limitations.

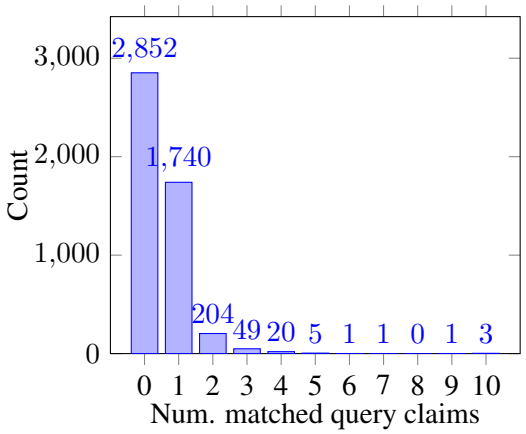

Figure 1: Distribution of document abstracts by number of query claims matched to them

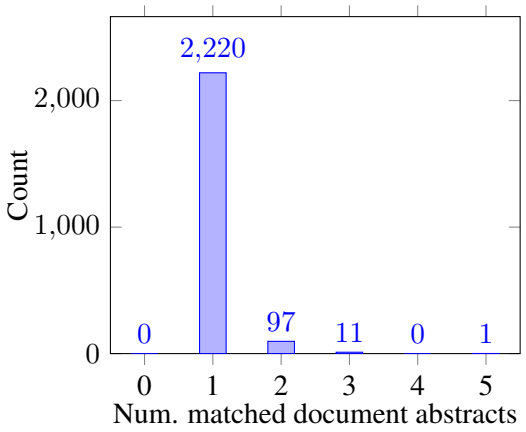

Figure 2: Distribution of query claims by number of document abstracts matched to them

4 Methodology

Two important questions were studied. First, are semantic searches more accurate than keyword searches when retrieving academic documents? Second, how can claims be extracted from documents in order to best match query claims?

To answer these questions, multiple approaches

Approach	Implementation	Search Type
baseline	BM25	keyword
paragraph	SBERT	semantic
sentences	spaCy + SBERT	semantic
generated	gemma + SBERT	semantic

Table 1: Description of each approach tested.

for retrieving documents were performed. The first was BM25, a similarity method that weighs the words common between two documents by their frequency in the whole corpus. As a state-of-the-art keywords search method, BM25 acted as a baseline method for retrieving abstracts most similar to a claim.

The remaining approaches used different methods to extract text from the abstract before embedding that extracted text as a vector. The simplest method used the full text of the abstract for the document embedding. To test if a more granular match was beneficial, the next method first divided the abstract into sentences using the spaCy model `en_core_web_trf` before proceeding with embedding. Embedding individual sentences limits the amount of information held within each vector and may bring each vector closer to representing a single claim. The full-paragraph and individual-sentence methods were then altered by prepending each extraction with the article's title before embedding them. This technique may clarify meaning that is missing from the raw text, especially for sentences in the middle of the abstract that lose context when isolated. The final method involved the use of the large language model `gemma-2-2b-it`. Gemma was prompted to extract the top 3 claims from the abstract and generate a simple title for the article. The extracted claims were tested as a method with and without the generate title prepended to them. This large language model method confers a large cost for generation but results in a more focused set of claims that may produce better semantic matches.

Embedding the extracted text was done using the Sentence-BERT model `all-MiniLM-L6-v2`. The query claims were embedded with this same model. The cosine similarity of each abstract against each claim was computed, and the documents with the highest similarity were ranked as most relevant. The accuracy of the approaches was evaluated using the Mean Average Precision metric, which is the average precision score for each recall value

Prompt
Article: "*<Abstract>*" Write one short, boring header for the article that doesn't use acronyms. Write 3 short claims synthesized from the article after it.

Table 2: The prompt given to gemma. <Abstract> is replaced with the abstract for a paper. The response is parsed for the extracted header and claims for that paper.

Approach	Default	Added Context
baseline	.337	-
paragraph	.538	.568
sentences	.513	**.572**
generated	.513	.507

Table 3: Evaluated score for each approach using the Mean Average Precision calculated over the top five documents. Each approach was run with and without extra context prepended to it. The best result in each column is in bold.

across every query; this metric measures how high up in the ranking the target documents were. The Mean Average Precision was computed using the top five ranked documents per query.

5 Results

All semantic search approaches performed far better than the baseline keyword search approach. Prepending with context was more effective for both paragraph and sentence extracting methods. Paragraph embedding was the most effective method when no context was added and sentence embedding was the most effective method when context was added. Large language model extractions score lower than methods that used the raw text. Overall, prepending the article title to each sentence in the abstract individually resulted in the highest MAP@5 score for retrieval.

6 Discussion

The Wikipedia citation dataset provided a meaningful means of evaluating different academic document retrieval strategies. This is noteworthy, as existing datasets from the fact-checking domain are insufficient for this task. It is also clear that the Wikipedia citation dataset is relatively difficult; the state-of-the-art keyword search method BM25

achieved only a 34% MAP@5 score. This is likely because the dataset was generated by scraping all citations on certain Wikipedia articles, resulting in many abstracts sharing the same topic. Additionally, the parsing process introduced ambiguity to many of the query claims taken from Wikipedia, meaning there was a disconnect between the information in the query and in the document labeled as matching it. This can be seen as an accurate representation of real-world fact checking scenarios, where claims are not always stated unambiguously in their source text.

On the Wikipedia citation dataset, all semantic search methods outperformed the baseline BM25 search. This result indicates that, when both used in isolation, semantic search is more effective at matching claims to sources than keyword search. This result supports the intuition that semantic search is able to compare claims with a similar meaning but dissimilar wording and validates the pursuit of semantic search methods for document retrieval.

Embedding the entire abstract by itself was initially more effective than embedding the abstract's sentences individually. This suggests that isolating sentences removes information necessary for semantic matching. This ambiguity is enough to offset the advantage that sentences might have over paragraphs in their similarity of length and structure to the query claims. This pattern was flipped when the article title was prepended to the full abstract and the individual sentences; both title-prepended strategies were more effective than their counterparts and sentence embedding became a more effective strategy than embedding the full text. From this it can be inferred that the titles contain a sufficient amount of extra context to clarify ambiguities in the isolated sentences from the abstract. Although the score increase from the title-prepended paragraph to the title-prepended sentences strategy is small, this increase indicates that accuracy drops when more information that necessary is included. These results have implications for the design of document retrieval systems, particularly when the queries have a similar format but different wording than the documents they are matched to. Such systems should use semantic search strategies rather than keyword-based methods and they must be careful to balance the amount of information included in each extracted claim. There is no conclusive evidence as to whether the grammatical structure of the extracted claims affects the retrieval accuracy; this question should be investigated in future work.

The final set of strategies used a large language model to generate claims and a header from the papers' abstracts. The retrieval accuracy for the generated claims was lower than that of the raw text methods, as was the accuracy for the concatenations of generated header with generated claims. These results should not be used to discount the possibility that large language models can aid in extraction. The method used for this study was very limited, as it generated a fixed number of claims per article, utilized a model with a low parameter count, and occasionally produced unparseable outputs. Further refinements could improve these results. Nevertheless, these results show that generated claims are not always better for semantic matching. As large language models cost much more to run than other strategies, this method may be less worthy of pursuit.

This study has evaluated numerous strategies for extracting text from scientific articles for the purpose of semantic searching. The evaluations should not be taken as statements about the viability of broad groups of approaches to this task; rather, they demonstrate a way to systematically evaluate approaches. The study also produced results that indicate what properties of each approach should be given special attention. To extend this study, more advanced forms of each approach can be studied. A more advanced form of introducing context to isolated sentences could use NLP techniques to resolve specific types of ambiguity such as ambiguous subjects or undefined acronyms. Similar strategies in the form of a pipelined series of prompts could be used to improve the quality of generated claims. Finally, other embedding models could be tested against the Sentence-BERT results. Fine-tuning on the scientific domain or on a claim-retrieval dataset may introduce a performance benefit by improving the representational accuracy of vector embeddings within the space.

Limitations

This study's findings are limited by the scope of the experiments. The fact that certain approaches were effective under this implementation of the task does not imply that those design decisions would be more effective when in a different retrieval pipeline, when using a different language model, or when applied to another domain.

Additionally, the Wikipedia claims dataset does not represent a real use case for automatic source retrieval. In the real world, claims are written without knowledge of the source; the challenge is to find a source that backs up that claim. However, the examples in this dataset were explicitly written to reflect the text of the article they cited. When certain terms or statistics are carried over from the scientific paper to the Wikipedia article, matching them becomes much easier than it would be in the real-world scenario. Thus, a more robust study of the semantic search document retrieval would require a dataset that better reflects its use case.

It is also probable that the document retrieval task grows more challenging as the size of the dataset increases, as more articles leads to more chances for articles to be falsely labeled as a match. As these tests were run on a relatively small dataset, their results cannot be taken as an accurate measure of performance on a realistically-sized corpus.

References

Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. 2021. A review on fact extraction and verification. *ACM Comput. Surv.*, 55(1).

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *Preprint*, arXiv:1903.10676.

Katarina Boland, Alica Hövelmeyer, Pavlos Fafalios, Konstantin Todorov, Usama Mazhar, and Stefan Dietze. 2023. Robust and efficient claim retrieval for online fact-checking applications.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Preprint*, arXiv:2102.05169.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

Zhenyun Deng, Michael Schlichtkrull, and Andreas Vlachos. 2024. Document-level claim extraction and decontextualisation for fact-checking. *Preprint*, arXiv:2406.03239.

Nazanin Jafari and James Allan. 2024. Robust claim verification through fact detection. *Preprint*, arXiv:2407.18367.

Tom Jansen and Tobias Kuhn. 2017. Extracting core claims from scientific articles. *Preprint*, arXiv:1707.07678.

Tobias Kuhn, Paolo Emilio Barbano, Mate Levente Nagy, and Michael Krauthammer. 2013. *Broadening the Scope of Nanopublications*, page 487–501. Springer Berlin Heidelberg.

Dasha Metropolitansky and Jonathan Larson. 2025. Towards effective extraction and evaluation of factual claims. *Preprint*, arXiv:2502.10855.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *Preprint*, arXiv:1803.05355.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. Scifact-open: Towards open-domain scientific claim verification. *Preprint*, arXiv:2210.13777.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh.