Document-level Claim Extraction and Decontextualisation for Fact-Checking

Zhenyun Deng, Michael Schlichtkrull, Andreas Vlachos
Department of Computer Science and Technology, University of Cambridge
{zd302, mss84, av308}@cam.ac.uk

Abstract

Selecting which claims to check is a time-consuming task for human fact-checkers, especially from documents consisting of multiple sentences and containing multiple claims. However, existing claim extraction approaches focus more on identifying and extracting claims from individual sentences, *e.g.,* identifying whether a sentence contains a claim or the exact boundaries of the claim within a sentence. In this paper, we propose a method for *document-level* claim extraction for fact-checking, which aims to extract check-worthy claims from documents and decontextualise them so that they can be understood out of context. Specifically, we first recast claim extraction as extractive summarization in order to identify central sentences from documents, then rewrite them to include necessary context from the originating document through sentence decontextualisation. Evaluation with both automatic metrics and a fact-checking professional shows that our method is able to extract check-worthy claims from documents more accurately than previous work, while also improving evidence retrieval.

1 Introduction

Human fact-checkers typically select a claim in the beginning of their day to work on for the rest of it. Claim extraction (CE) is an important part of their work, as the overwhelming volume of claims in circulation means the choice of *what* to fact-check greatly affects the fact-checkers' impact (Konstantinovskiy et al., 2021). Automated approaches to this task have been proposed to assist them in selecting check-worthy claims, *i.e.,* claims that the public has an interest in knowing the truth (Hassan et al., 2017a; Guo et al., 2022).

Existing CE methods mainly focus on detecting whether a sentence contains a claim (Reddy et al., 2021; Nakov et al., 2021b) or the boundaries of the claim within a sentence (Wührl and Klinger, 2021; Sundriyal et al., 2022). In real-world scenarios though, claims often need to be extracted from documents consisting of multiple sentences and containing multiple claims, not all of which are relevant to the central idea of the document, and verifying all claims manually or even automatically would be inefficient.

Moving from sentence-level CE to document-level CE is challenging; we illustrate this with the example in Figure 1. Sentences in orange are claims selected by a popular sentence-level CE method, Claimbuster (Hassan et al., 2017b), that are worth checking in principle but do not always relate to the central idea of the document, and multiple sentences with similar claims are selected, which would not all need to be fact-checked (*e.g.,* sentences 1 and 6).

Claims extracted for fact-checking are expected to be unambiguous (Lippi and Torroni, 2015; Wührl and Klinger, 2021), which means that they cannot be misinterpreted or misunderstood when they are considered outside the context of the document they were extracted from, consequently allowing them to be fact-checked more easily (Schlichtkrull et al., 2023). Figure 1 shows an example of claim decontextualisation, where the claim "*Bird is scrapping thousands of e-scooters in the Middle East .....* " requires coreference resolution to be understood out of context, *e.g.,* "*Bird*" refers to "*California scooter sharing start-up Bird*". However, existing CE methods primarily focus on extracting sentence-level claims (*i.e.,* extracting sentences that contain a claim) from the original document (Reddy et al., 2021) and ignore their decontextualisation, resulting in claims that are not unambiguously understood and verified.

To address these issues, we propose a novel method for *document-level* claim extraction and decontexualisation for fact-checking, aiming to extract salient check-worthy claims from documents that can be understood outside the context of the document. Specifically, assuming that salient

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11943–11954

Document (CNBC News)

[1] Between 8,000 and 10,000 e-scooters are being destroyed in the Middle East by California scooter sharing start-up Bird, according to sources. [2] They belong to Circ, an e-scooter company that was acquired by Bird in January. [3] Bird shut down its entire Middle East operation as a result of Covid-19. [4] Bird is scrapping thousands of e-scooters in the Middle East and shutting down its operations in the majority of the region as a result of the coronavirus pandemic, according to five people familiar with the matter. [5]The e-scooters being scrapped belong to Circ, which was acquired by Bird for an undisclosed sum in January. [6] There are between 8,000 and 10,000 Circ scooters across cities in Qatar, Bahrain and United Arab Emirates, according to one former employee and one company source who asked to be kept anonymous as they've signed a confidentiality agreement. ..... [7] But there have been questions about the longevity of their vehicles, with reports suggesting some Bird e-scooters have a life span of just a few months. [8] Last week, it emerged that Uber is scrapping thousands of e-bikes and e-scooters worth millions of dollars after selling its Jump unit to mobility start-up Lime. .....

Document-level Claim Extraction

Sentence-level: 8, 6, 1	**Document-level:** 4, 5, 7

Gold Claim (Fact-checking Organization, Misbar):
Bird e-scooters are shutting down service in the Middle East, and scrapping as many as 10,000 scooters.

Claim extracted by decontextualising the 4th sentence:
California scooter sharing start-up Bird is scrapping thousands of e-scooters in the Middle East and shutting down its operations in the majority of the region as a result of the coronavirus pandemic, according to five people familiar with the matter.

Figure 1: An example of document-level claim extraction. Document[1] is a piece of news from CNBC. Gold Claim[2] is annotated by the fact-checking organization, Misbar. Sentences in orange denote check-worthy claims extracted by sentence-level CE (Claimbuster). Sentences in blue denote salient claims extracted by our document-level CE. The claim in green is a decontextualised claim derived from the 4th sentence obtained by our document-level CE.

claims are derived from central sentences, i) we recast the document-level CE task into the extractive summarization task to extract central sentences and reduce redundancy; ii) we decontextualise central sentences to be understandable out of context by enriching them with the necessary context; iii) we introduce a QA-based framework to obtain the necessary context by resolving ambiguous information units in the extracted sentence.

To evaluate our method we derive a CE dataset[3] containing decontextualised claims from AVeriTeC (Schlichtkrull et al., 2023), a recently proposed benchmark for real-world claim extraction and verification. Our method achieves a Precision@1 score of 47.8 on identifying central sentences, a 10% improvement over Claimbuster. This was verified further by a fact-checking professional, as the sentences returned by our method were deemed central to the document more often, and check-worthy more often than those extracted by Claimbuster. Additionally, our method achieved a character-level F score (chrF) (Popović, 2015) of 26.4 against gold decontextualised claims, outperforming all baselines. When evaluated for evidence retrieval potential, the decontextualised claims obtained by enriching original sentences with the necessary context, are better than the original claim sentences, with an average 1.08 improvement in precision.

[1] https://www.cnbc.com/2020/06/03/bird-circ-scooters-middle-east.html

[2] https://misbar.com/en/factcheck/2020/06/18/are-bird-e-scooters-leaving-the-middle-east

[3] https://github.com/Tswings/AVeriTeC-DCE

2 Related Work

Claim Extraction Claim extraction is typically framed either as a classification task or claim boundary identification task. The former framing focuses on detecting whether a given sentence contains a check-worthy claim. Claimbuster (Hassan et al., 2017b), the most popular method in this paradigm, computes the score of how important a sentence is to be fact-checked. More similar to our work are studies that formulate the task of check-worthy claim detection as a sentence ranking task. For example, Zhou et al. (2021) present a sentence-level classifier by combining a fine-tuned hate-speech model with one dropout layer and one classification layer to rank sentences. However, these methods were not able to handle the challenges of document-level claim extraction, *e.g.,* avoid redundant claim sentences.

The framing of claim extraction as boundary identification focuses on detecting the exact claim boundary within the sentence. Nakov et al. (2021a) propose a BERT-based model to perform claim detection (Levy et al., 2014) by identifying the boundaries of the claim within the sentence. Sundriyal et al. (2022) tackle claim span identification as a token classification task for identifying argument units of claims in the given text. Unlike the above methods, where the claims are extracted from given sentences, our work aims to extract salient check-worthy claims from documents, thus addressing the limitations of sentence-level methods in extracting salient claim sentences and avoiding redundancy.

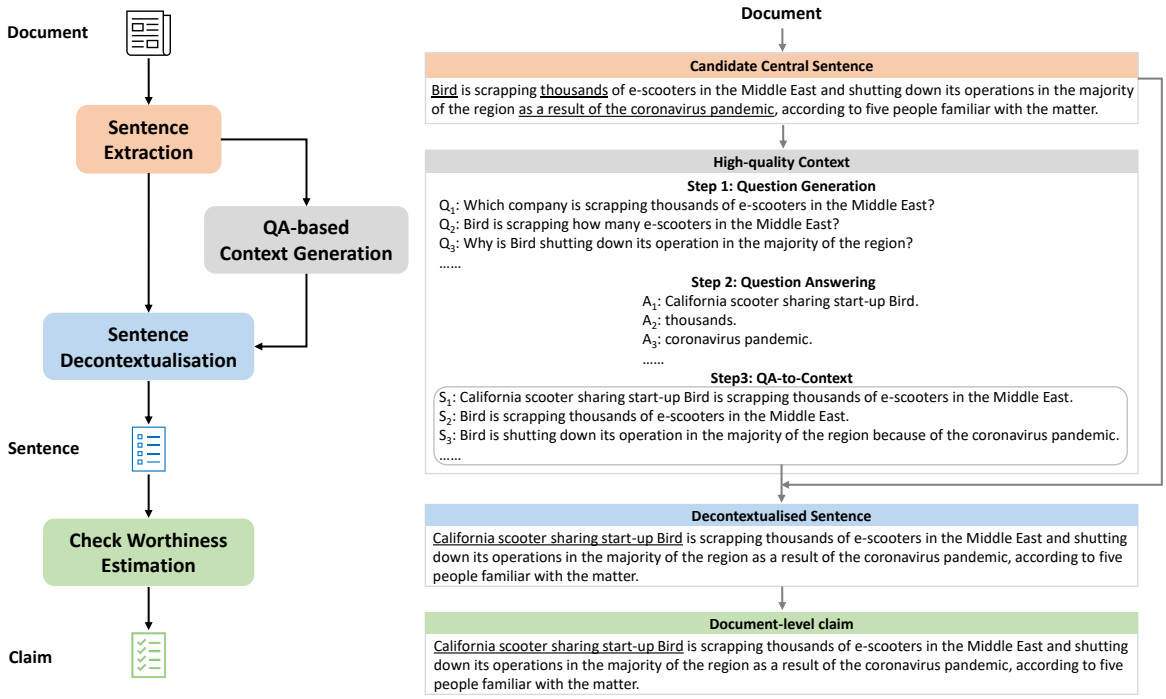

Figure 2: An overview of our document-level claim extraction framework. Given an input document, we first use extractive summarization to rank all sentences and select summary sentences as central sentences. Then, we describe a QA-based framework to generate a specific high-quality context for important information units in the sentence. Next, we use a seq2seq generation model to decontextualise sentences by enriching them with their corresponding context. Finally, a claim check-worthiness classifier is used to select salient check-worthy claim sentences based on the score that reflects the degree to which sentences belong to the check-worthy claim.

Decontextualisation Choi et al. (2021) propose two different methods for decontextualisation, based on either a coreference resolution model or a seq2seq generation model. Both methods use the sentences in the paragraph containing the target sentence as context to rewrite it. Newman et al. (2023) utilize an LLM to generate QA pairs for each sentence by designing specific prompts, and then use an LLM with these QA pairs to rewrite each sentence. Sundriyal et al. (2023) propose to combine chain-of-thought and in-context learning for claim normalization. Unlike the above methods, we generate declarative sentences for potentially ambiguous information units in the target sentence based on the whole document, and combine them into context to rewrite the target sentence.

3 Method

As illustrated in Figure 2, our proposed document-level claim extraction framework consists of four components: i) Sentence extraction (§3.1); extracts the sentences related to the central idea of the document as candidate claim sentences; ii) Context generation (§3.2), extracts context from the document for each candidate sentence; iii) Sentence decontextualisation (§3.3), rewrites each sentence with its corresponding context to be understandable out of context; iv) Check-worthiness estimation (§3.4), selects the final check-worthy claims from candidate decontextualised sentences.

3.1 Sentence Extraction

The claims selected by human fact-checkers are typically related to the central idea of the document considered. Thus we propose to model sentence extraction as extractive summarization. For this purpose, we concatenate all the sentences in the document into an input sequence, which is then fed to BertSum (Liu and Lapata, 2019), a document-level extractive summarization method trained on the CNN/DailyMail dataset. Specifically, given a document consisting of n sentences $\mathcal{D} = \{s_1, s_2, ..., s_n\}$, we first formulate the input sequence C as "[CLS] s_1 [SEP] [CLS] s_2 [SEP] ... [CLS] s_n[SEP]", where [CLS] and [SEP] denote the start and end token for each sentence, respectively, and then feed them into a pre-trained encoder BERT to obtain the sentence representation s. Finally, a linear layer on sentence representations

$\mathbf{S} = \{\mathbf{s}_1, ..., \mathbf{s}_i, ..., \mathbf{s}_n\}$ is used to score sentences.

$$\begin{aligned} \mathbf{S} &= \text{BERT}(C) \\ score_i &= \sigma(W\mathbf{s}_i + b_0) \end{aligned} \tag{1}$$

where σ is a sigmoid function, $\mathbf{s}_i$ denotes the representation of the i-th [CLS] token, *i.e.,* the representation of the i-th sentence, and $score_i$ denotes the score of the i-th sentence. All sentences are constructed into an ordered set $S = \{s'_1, ..., s'_i, ..., s'_n\}$ according to their scores. Since all sentences are ranked by sentence-level scoring, some top-scoring sentences may have the same or similar meaning. To avoid redundancy, we add an entailment model DocNLI (Yin et al., 2021), a more generalizable model trained on five datasets from different benchmarks, on top of the output of BertSum to remove redundant sentences by calculating the entailment scores between sentences, *e.g.,* we first remove the sentences that have an entailment relationship with the top-1 sentence in S, and then repeat this process for the remaining top-2/3/... sentence until we extract k central sentences. Following previous work (Liu and Lapata, 2019), we only select the top-k sentences with the highest scores in Equation 1 as candidate central sentences.

$$S' = \text{DocNLI}(S) \tag{2}$$

where $S' = \{s'_1, s'_2, ..., s'_k\}$ is a set of central sentences that do not contain the same meaning.

3.2 Context Generation

After sentence extraction, the next step is to clarify the (possibly) ambiguous sentences in S' by rewriting them with their necessary context. Unlike Choi et al. (2021) where the context consists of a sequence of sentences in the paragraph containing the ambiguous sentence, we need to consider the whole document, *i.e.,* sentences from *different* paragraphs. We propose a QA-based context generation framework to produce a specific context for each ambiguous sentence, which contains three components: i) Question Generation: extracts potentially ambiguous information units from the sentence and generates questions with them as answers; ii) Question Answering: finds more information about ambiguous information units by answering generated questions with the whole document; iii) QA-to-Context Generation: converts question-answer pairs into declarative sentences and combines them into context specific to the sentence. In the following subsections, we describe each component in detail.

Question Generation. To identify ambiguous information units in candidate central sentences, we first use Spacy[4] to extract named entities, pronouns, nouns, noun phrases, verbs and verb phrases in the sentence i as the potentially ambiguous information units $U_i = \{u_i^1, u_i^2, ..., u_i^j, ..., u_i^m\}, i \in [1, 2, ..., k]$, where u_i^j denotes the j-th information unit of the i-th candidate sentence s'_i.

Once the set of information units for a sentence U_i is identified, we then generate a question for each of them. Specifically, we concatenate u_i^j and s'_i in which u_i^j is located as the input sequence and feed it into QG (Murakhovs'ka et al., 2022), a question generator model trained on nine question generation datasets with different types of answers, to produce the question q_i^j with u_i^j as the answer.

$$Q_i = \{q_i^j\}_{j=1}^m = \{\text{QG}(s'_i, u_i^j)\}_{j=1}^m \tag{3}$$

where Q_i denotes the set of questions corresponding to U_i in the sentence s'_i.

Question Answering. After question generation, our next step is to clarify ambiguous information units by answering corresponding questions with the document $\mathcal{D}$. Specifically, following Schlichtkrull et al. (2023), we first use BM25 (Robertson et al., 2009) to retrieve evidence E related to the question q_i^j from $\mathcal{D}$, and then answer q_i^j with E using an existing QA model (Khashabi et al., 2022) trained on twenty datasets that can answer different types of questions.

$$\begin{aligned} E &= \text{BM25}(\mathcal{D}, q_i^j) \\ a_i^j &= \text{QA}(E, q_i^j) \end{aligned} \tag{4}$$

where a_i^j denotes a more complete information unit corresponding to u_j^i, *e.g.,* a complete coreference. We denote all question-answer pairs of the i-th sentence as $P_i = \{(q_i^1, a_i^1), (q_i^2, a_i^2), ..., (q_i^m, a_i^m)\}$.

QA-to-Context Generation. After question answering, we utilize a seq2seq generation model to convert QA pairs P_i into the corresponding context C'_i. Specifically, we first concatenate the question q_i^j and the answer a_i^j as the input sequence, and then output a sentence using the BART model (Lewis et al., 2019) finetuned on QA2D (Demszky et al., 2018). QA2D is a dataset with over 500k

[4] https://spacy.io

Data	#sample	med.sent	avg.sent	len.claim	len.document
Train	830	9	1.09	17	274
Dev	149	6	1.01	16	120
Test	252	5	1.05	16	63
All	1231	7	1.07	17	17

Table 1: Descriptive statistics for AVeriTeC-DCE. #sample refers to the number of samples in AVeriTeC available for claim extraction, *i.e.,* the total number of accessible *source url*, med.sent refers to the median number of sentences in documents. avg.sent refers to the average number of sentences in claims, len.claim refers to the median length of claims in words, len.document refers to the median length of documents in words.

NLI examples that contains various inference phenomena rarely seen in previous NLI datasets. More formally,

$$\tilde{s}_i^j = \text{BART}(q_i^j, a_i^j) \tag{5}$$

where $\tilde{s}_i^j$ is a declarative sentence corresponding to the information unit u_i^j. Finally, all generated sentences are combined into high-quality context $C_i' = \{\tilde{s}_i^1, \tilde{s}_i^2, ..., \tilde{s}_i^m\}$ corresponding to the information units U_i in sentence s_i', which is then used in the next decontextualisation step to enrich the ambiguous sentences.

3.3 Sentence Decontextualisation

Sentence decontextualisation aims to rewrite sentences to be understandable out of context, while retaining their original meaning. To do this, we use a seq2seq generation model T5 (Raffel et al., 2020) to enrich the target sentence with the context generated in the previous step for it. Specifically, we first formulate the input sequence as "[CLS] $\tilde{s}_i^1$ [SEP] $\tilde{s}_i^2$ ..... [SEP] $\tilde{s}_i^m$ [SEP] s_i'", where s_i' denotes the potential ambiguous sentence and [SEP] is a separator token between the context sentences generated. We then feed the input sequence to D (Choi et al., 2021), a decontextualisation model was trained on the dataset annotated by native speakers of English in the U.S. that handles various linguistic phenomena, to rewrite the sentence. Similarly, we set the output sequence to be [CAT] [SEP] y.

$$y_i = \begin{cases} \text{D}(s_i', C_i'), if\ \text{CAT} = feasible \\ s_i' \quad , if\ \text{CAT} = infeasible \\ s_i' \quad , if\ \text{CAT} = unnecessary \end{cases} \tag{6}$$

where CAT $= feasible$ or $infeasible$ denotes that s_i' can or cannot be decontextualised, CAT $= unnecessary$ denotes that s_i' can be understood without being rewritten, y_i denotes the i-th decontextualised sentence.

3.4 Check-Worthiness Estimation

Unlike existing CE methods that determine whether a sentence is worth checking without considering the context, we estimate the check-worthiness of a sentence after decontextualisation because some sentences may be transformed from not check-worthy into check-worthy ones in this process. Specifically, we use a DeBERTa model trained on the ClaimBuster dataset (Arslan et al., 2020) to classify sentences into three categories: Check-worthy Factual Sentence (CFS), Unimportant Factual Sentence (UFS) and Non-Factual Sentence (NFS). Formally,

$$\begin{gathered} score(y_i) = \text{DeBERTa}(class = \text{CFS} \mid y_i) \\ \text{claim} = \text{argmax}\{score(y_i)\}_{i=1}^{k} \end{gathered} \tag{7}$$

where $score(y_i)$ reflects the degree to which the decontextualised sentence y_i belongs to CFS, and claim denotes that the final salient check-worthy claim that can be understood out of context.

4 Dataset

We convert AVeriTeC, a recently proposed dataset for real-world claim extraction and verification (Schlichtkrull et al., 2023), into AVeriTeC-DCE, a dataset for the document-level CE task. AVeriTeC is collected from 50 different fact-checking organizations and contains 4568 real-world claims. Each claim is associated with attributes such as its type, source and date. In this work, we mainly focus on the following attributes relevant to claim extraction: i) $claim$, the claim as extracted by the fact-checkers and decontextualised by annotators, and ii) $source\ url$: the URL linking to the original web article of the $claim$. The task of this work is to extract the salient check-worthy claims from the $source\ url$. We also consider whether claims need to be decontextualised when extracting them from documents, as this will directly affect the subsequent evidence retrieval and claim verification.

To extract claim-document pairs from AVeriTeC that can be used for document-level CE, we perform the following filtering steps: 1) Since we focus on the extraction of textual claims, we do not include *source urls* containing images, video or audio; 2) To extract the sentences containing the claims from *source urls*, we build a web scraper to extract text data in the *source url* as the document. We found that the attribute *source url* is not always available in samples, thus we only select those samples where the web scraper can return text data from *source urls*. We obtain a dataset AVeriTeC-DCE, containing 1231 available samples, for document-level CE. We do not divide the dataset into train, dev and test sets, as all models we rely on are pre-trained models (*e.g.,* BertSum) and approaches that do not require training (*e.g.,* BM25). Statistics for AVeriTeC-DCE are described in Table 1.

5 Experiments

Our approach consists of four components: sentence extraction, context generation, sentence decontextualisation and check-worthiness estimation. As such, we conduct separate experiments to evaluate them, as well as an overall evaluation for document-level CE.

5.1 Sentence Extraction

We compare our sentence extraction method, the combination of BertSum and DocNLI stated in Section 3.1, against other baselines through automatic evaluation and human evaluation.

Baselines 1) Lead sentence: the lead (first) sentence of most documents is considered to be the most salient, especially in news articles (Narayan et al., 2018); 2) Claimbuster (Hassan et al., 2017b): we use this well-established method to compute the check-worthiness score of each sentence and we rank sentences based on their scores; 3) LSA (Gong and Liu, 2001): a common method of identifying central sentences of the document using the latent semantic analysis technique; 4) TextRank (Mihalcea and Tarau, 2004): a graph-based ranking method for identifying important sentences in the document; 5) BertSum (Liu and Lapata, 2019): a BERT-based document-level extractive summarization method for ranking sentences.

Automatic Evaluation Since the central sentences of documents are not given in AVeriTeC, we cannot evaluate the extracted central sentences by exact matching. Thus, we instead rely on the sentence that has the highest chrF (Popović, 2015) with the *claim*, as the *claim* is the central claim annotated by human fact-checkers. We use Precision@k as the evaluation metric, which denotes the probability that the first k sentences in the extracted sentences contain the central sentence. Table 2 shows the results of different sentence extraction methods. We can see that our method outperforms all baselines in identifying the central sentence, achieving a P@1/P@3/P@5/P@10 score of 47.8/63.1/68.6/73.8, which indicates that the combination of the extractive summarization (BertSum) and entailment model (DocNLI) can better capture the central sentences and avoid redundant ones. We found that the common extractive summarization methods (*e.g.,* Lead Sentence, TextRank and BertSum) are better than Claimbuster on P@1, confirming what we had stated in the introduction, that sentence-level CE methods have limitations when they are applied at the document-level CE. Moreover, we observe that the lead sentence achieves a P@1 score of 42.3, indicating that there is a correlation between the sentences selected for fact-checking and the lead sentence that often served as the summary. We list the source URLs of the samples for claim extraction in Appendix A1.

Method	P@1	P@3	P@5	P@10
Claimbuster	37.8	59.1	65.7	71.4
Lead Sentence	42.3	-	-	-
LSA	38.4	55.3	62.1	70.4
TextRank	42.7	60.6	65.1	71.2
BertSum	43.4	61.6	67.5	72.3
Ours	**47.8**	**63.1**	**68.6**	**73.8**

Table 2: Results with different sentence extraction methods. P@k denotes the probability that the first k sentences in the ranked sentences contain the central sentence.

	Claimbuster	Ours
IsCheckWorthy	0.36	0.44
IsCentralClaim	0.24	0.68

Table 3: Human Evaluation of sentence extraction on two different dimensions.

Human Evaluation To further compare sentences extracted by our method and Claimbuster,

Coreference Resolution
Sentence: He has pubicly stated that he sympathized with their cause and even hinted that he would provide them with American resources should they be in need during his 2008 State of the Union address. **Decontextualised sentence**: President Obama has pubicly stated that he sympathized with their cause and even hinted that he would provide them with American resources should they be in need during his 2008 State of the Union address.
Global Scoping
Sentence: During the attack, Capitol Police made the request again. **Decontextualised sentence**: During the attack on Washington D.C., Capitol Police made the request again.
Bridge Anaphora
Sentence: The government does not have proper storage facilities for stocking such a large amount of excess grain. **Decontextualised sentence**: The government of India does not have proper storage facilities for stocking such a large amount of excess grain.

Figure 3: Case studies of sentence decontextualisation solving linguistic problems, such as coreference resolution, global scoping and bridge anaphora.

we asked a fact-checking professional to evaluate the quality of extracted sentences on the following two dimensions: 1) **IsCheckWorthy**: is the sentence worth checking? 2) **IsCentralClaim**: is the sentence related to the central idea of the article? We randomly select 50 samples, each containing at least 5 sentences. For simplicity, we only select the top-1 sentence returned by each method for comparison. As shown in Table 3, we observed that 68% of the central sentences extracted by our method are related to the central idea of the document compared to 24% of Claimbuster, which further supports our conclusion obtained by automatic evaluation, *i.e.,* the sentences extracted by our method were more often central to the document, and more often check-worthy than those that extracted by Claimbuster. This indicates that when identifying salient check-worthy claims from documents, it is not enough to consider whether a sentence is worth checking at the sentence level, but also whether the sentence is related to the central idea of the document. Thus, we believe that the claims related to the central idea of the document are the ones that the public is more interested in knowing the truth.

5.2 Decontextualisation

To evaluate the effectiveness of decontextualisation on evidence retrieval, for a fair comparison, we select the sentence that has the highest chrF with the $claim$ as the best sentence as we considered in Section 5.1, and conduct a comparison between it and its corresponding decontextualised sentence. The evidence set used for evaluation is retrieved from the Internet using the Google Search API given a claim, each containing gold evidence and additional distractors (Schlichtkrull et al., 2023). We use Precision@k as the evaluation metric.

Baselines 1) Coreference model: decontextualisation by replacing unresolved coreferences in the target sentence, *e.g.,* (Joshi et al., 2020); 2) Seq2seq model: decontextualisation by rewriting the target sentence with necessary context (Choi et al., 2021).

Retrieval-based Evaluation For a given claim different decontextualisations could be considered correct, thus comparing against the single reference in AVeriTeC would be suboptimal. Thus we prefer to conduct the retrieval-based evaluation, assuming that better decontextualisation improves evidence retrieval, as it should provide useful context for fact-checking. Following previous work (Choi et al., 2021), we compare our QA-based decontextualisation method against other baselines through retrieval-based evaluation. We use BM25 as the retriever to find evidence with different sentences as the query. Table 4 shows the results for evidence retrieval with different decontextualised sentences on the dev set of AVeriTeC-DCE (AVeriTeC only publicly released the train and dev sets). We found that our method outperforms all baselines, and improves the P@3/P@5/P@10 score over the original sentence by 1.42/0.82/0.99, achieving an average 1.08 improvement in precision. After further analysis, we found that only 21/149 sentences are decontextualised by our method, and 17/21 of these sentences obtain better evidence retrieval, with an average improvement of 1.21 in precision over the original sentences, proving that decontextualisation enables evidence retrieval more effectively.

Case Study Figure 3 illustrates three case studies of sentence decontextualisation. The first case is an example that requires coreference resolution. To make the sentence understandable out of context, these words (*e.g.,* "*He*", "*their*") need to be rewritten with the context. After decontextualisa-

Method	P@3	P@5	P@10
Sentence	35.45	44.72	61.31
Coreference	36.02	44.98	61.79
Seq2seq(Context)	36.17	45.04	61.99
Ours$_{\text{seq2seq(Context}^\star\text{)}}$	**36.87**	**45.54**	**62.30**

Table 4: Results for evidence retrieval with different decontextualised sentences. Context consists of a sequence of sentences in the paragraph containing the target sentence. Context* consists of declarative sentences generated by our context generation module.

tion, we can see that "*He*" is rewritten to "*President Obama*", which helps us understand the sentence better without context. As for "*their*", we cannot decontextualise it because there is no information about this word in the document. This supports the claim that providing a high-quality context is necessary for better decontextualisation. The second case is an example that requires global scoping, which requires adding a phrase (*e.g.,* prepositional phrase) to the entire sentence to make it better understood. In this case, we add "*on Washington D.C.*" as a modifier to "*During the attack*" to help us understand where the attack took place. The third case is an example that requires a bridge anaphora, where the phrase noun "*The government*" becomes clear by adding a modifier "*India*". In summary, decontextualising the claim is helpful for humans to better understand the claim without context.

5.3 Document-level Claim Extraction

In this section, we put four components together to conduct an overall evaluation for document-level CE, *e.g.,* extracting the claim in green from the document in Figure 1. We first select top-3 sentences returned by different sentence extraction methods as candidate central sentences, and then feed them into our decontextualisation model to obtain decontextualised claim sentences, finally use a claim check-worthiness classifier to select the final claim. We evaluate performance by calculating the similarity between our final decontextualised claim sentence and the $claim$ decontextualised by fact-checkers. We use the chrF as the evaluation metric. The chrF computes the similarity between texts using the character n-gram F-score. Other metrics are reported in Appendix A2.

Statistics for decontextualisation are described in Table 5, we observe that 122 out of 1231 (10%) sentences can be decontextualised (feasible); 122 out of 1231 (10%) sentences cannot be decontextualised (infeasible); and 987 out of 1231 (80%) sentences can be understood without being decontextualised (unnecessary), including the lead sentence. Since the lead sentence of the document is often considered to be the most salient, we do not decontextualise the lead sentence. In Table 6, we show the results of original sentences and decontextualised sentences for claim extraction, and our method achieves a chrF of 26.4 on gold claims decontextualised by the fact-checkers, outperforming all baselines. We observe that the performance of claim extraction and sentence extraction is positively correlated, *i.e.,* the closer the extracted sentence is to the central sentence, the more similar the extracted claim is to the $claim$, which supports our assumption that salient claims are derived from central sentences. For this reason, the performance of our document-level CE is limited by the performance of sentence extraction, *i.e.,* if our sentence extraction method cannot find the gold central sentence, decontextualisation may not improve the performance of CE, and may even lead to a decrease in the performance of CE due to noise caused by decontextualisation.

Moreover, empirically, we found that central sentences led to improved overall performance. This might be a consequence of the dataset – social media sites such as Twitter or Facebook are common

Data	#claim	#fea.	#infea.	#unnec.
All	1231	122	122	987

Table 5: Statistic of decontextualisation. #fea./infea. denotes the number of sentences that can/cannot be decontextualised, #unnec. denotes the number of sentences that can be understood without context.

Method	chrF	
	Sentence*	Dec. Sentence*
Claimbuster	24.3	24.5
Lead Sentence	23.8	-
LSA	24.1	24.3
TextRank	24.5	25.4
BertSum	25.6	25.9
Ours	**25.9**	**26.4**

Table 6: Results of Document-level CE. Sentence* denotes the best sentence returned by different sentence extraction methods. Dec. Sentence* denotes the decontextualised Sentence*.

sources of claims in AVeriTeC (along with more traditional news), and a Twitter or Facebook thread often contains only a few key points. AVeriTeC was collected by reverse engineering claims which fact-checkers from around the world chose to work on. As such, the distribution of source articles represents what journalists found to be check-worthy and chose to work on. Our work as such reflects the contexts wherein real-world misinformation appears (but indeed, may have a bias towards what works well in those contexts).

To further verify the effectiveness of our method, we conduct a comparison on the document-level CE dataset (CLEF-2021, subtask 1B (Shaar et al., 2021)) using our method and Claimbuster. Table 7 shows the results of two different methods for identifying check-worthy claims on the dev set of subtask 1B. We observe that our method outperforms Claimbuster on P@1/3/5/10, indicating that our document-level CE method can better identify check-worthy claims than Claimbuster (sentence-level CE). This supports our conclusion that the document-level check-worthy claims extracted by our method are the claims that the public is more interested in knowing the truth.

Method	P@1	P@3	P@5	P@10
Claimbuster	0.111	0.074	0.156	0.089
Ours	0.222	0.185	0.200	0.144

Table 7: Results of different methods for identifying check-worthy claims on the dev set of subtask 1B.

6 Conclusions and Future work

This paper presented a *document-level* claim extraction framework for fact-checking, aiming to extract salient check-worthy claims from documents that can be understood out of context. To extract salient claims from documents, we recast the claim extraction task as the extractive summarization task to select candidate claim sentences. To make sentences understandable out of context, we introduce a QA-based decontextualisation model to enrich them with the necessary context. The experimental results show the superiority of our method over previous methods, including document-level claim extraction and evidence retrieval, as indicated by human evaluation and automatic evaluation. In future work, we plan to extend our document-level claim extraction method to extract salient check-worthy claims from multimodal web articles.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation programme grant AVeriTeC (Grant agreement No. 865958). We thank David Corney from Full Fact for his help with evaluating the output of our models, and Yulong Chen for his helpful comments and discussions. We would also like to thank the anonymous reviewers for their helpful questions and comments that helped us improve the paper.

Limitations

While our method has demonstrated superiority in extracting salient check-worthy claims and improving evidence retrieval, we recognize that our method is not able to decontextualise all ambiguous sentences, particularly those that lack the necessary context in the $source\ url$. Also, human fact-checkers have different missions, thus check-worthiness claims to one fact-checking organization may not be check-worthiness to another organization (*i.e.,* some organizations check parody claims or claims from satire websites, while others do not). Furthermore, since the documents we use are extracted from the $source\ url$, a powerful web scraper is required when pulling documents from $source\ urls$. Moreover, our method assumes salient claims are derived from central sentences. Although this assumption is true in most cases, it may be inconsistent with central claims collected by human fact-checkers. Besides, we use the chrF metric to calculate the similarity between claims extracted from the $source\ url$ and the gold claim, while gold claims are decontextualised by the fact-checkers with fact-checking articles and may contain information that is not in the original article, thus the metrics used to evaluate document-level claims are worth further exploring.

Ethics Statement

We rely on fact-checks from real-world fact-checkers to develop and evaluate our models. Nevertheless, as any dataset, it is possible that it contains biases which influenced the development of our approach. Given the societal importance of fact-checking, we advise that any automated system is employed with human oversight to ensure that the fact-checkers fact-check appropriate claims.

References

Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. A benchmark dataset of check-worthy factual claims. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 821–829.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*.

Yihong Gong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017a. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1803–1812.

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. 2017b. Claimbuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, 8:64–77.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. Unifiedqa-v2: Stronger generalization via broader cross-format training. *arXiv preprint arXiv:2202.12359*.

Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2021. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital threats: research and practice*, 2(2):1–16.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argument mining. In *IJCAI*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Lidiya Murakhovs'ka, Chien-Sheng Wu, Philippe Laban, Tong Niu, Wenhao Liu, and Caiming Xiong. 2022. MixQG: Neural question generation with mixed answer types. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1486–1497, Seattle, United States. Association for Computational Linguistics.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021a. Automated fact-checking for assisting human fact-checkers. *arXiv preprint arXiv:2103.07769*.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeno, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, et al. 2021b. The clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*, pages 639–649. Springer.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Benjamin Newman, Luca Soldaini, Raymond Fok, Arman Cohan, and Kyle Lo. 2023. A controllable qa-based framework for decontextualization. *arXiv preprint arXiv:2305.14772*.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Revanth Gangi Reddy, Sai Chinthakindi, Zhenhailong Wang, Yi R Fung, Kathryn S Conger, Ahmed S Elsayed, Martha Palmer, and Heng Ji. 2021. Newsclaims: A new benchmark for claim detection from news with background knowledge. *arXiv preprint arXiv:2112.08544*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. *arXiv preprint arXiv:2305.13117*.

Shaden Shaar, Maram Hasanain, Bayan Hamdan, Zien Sheikh Ali, Fatima Haouari, Alex Nikolov, Mücahid Kutlu, Yavuz Selim Kartal, Firoj Alam, Giovanni Da San Martino, et al. 2021. Overview of the clef-2021 checkthat! lab task 1 on check-worthiness estimation in tweets and political debates. In *CLEF (working notes)*, pages 369–392.

Megha Sundriyal, Tanmoy Chakraborty, and Preslav Nakov. 2023. From chaos to clarity: Claim normalization to empower fact-checking. *arXiv preprint arXiv:2310.14338*.

Megha Sundriyal, Atharva Kulkarni, Vaibhav Pulastya, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Empowering the fact-checkers! automatic identification of claim spans on Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7701–7715, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Amelie Wührl and Roman Klinger. 2021. Claim detection in biomedical twitter posts. *arXiv preprint arXiv:2104.11639*.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. Docnli: A large-scale dataset for document-level natural language inference. *arXiv preprint arXiv:2106.09449*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xinrui Zhou, Bohuai Wu, and Pascale Fung. 2021. Fight for 4230 at checkthat! 2021: Domain-specific preprocessing and pretrained model for ranking claims by check-worthiness. In *CLEF (Working Notes)*, pages 681–692.

A1 Statistic of Source URLs

We described the statistic of the source URLs of the samples for document-level CE in Table A1.

URL	#sample
twitter.com	241
facebook.com	235
perma.cc	63
channelstv.com	37
aljazeera.com	34
president.go.ke	27
gov.za	25
instagram.com	18
c-span.org	16
factba.se	13
axios.com	12
youtu.be	12
rumble.com	12
abcnews.go.com	11
rev.com	11
cnn.com	10
news24.com	10
punchng.com	10
washingtonpost.com	10
cbsnews.com	8
foxnews.com	8
misbar.com	7
thegatewaypundit.com	7
politifact.com	7
nypost.com	7
nbcnews.com	7
telegraph.co.uk	7
wisn.com	6
tatersgonnatate.com	6
bustatroll.org	6
dailymail.co.uk	5
whitehouse.gov	5

Table A1: Statistic of the source URLs of the samples for document-level CE. We only list URLs with a total number number greater than 5.

A2 Evaluation Metrics

We use the following metrics to assess the similarity between the claim decontextualised by our method and the $claim$ decontextualised by fact-checkers. **SARI** (Xu et al., 2016) is developed to compare the claim with the reference claim by measuring the goodness of words that are added, deleted and kept. **BERTScore** (Zhang et al., 2019) is utilized to compute the semantic overlap between the claim and the reference claim by sentence representation. Since most $claims$ in AVerTeC are decontextualised by fact-checkers with fact-checking articles, they may contain some information that is not in the $source\ url$, making it challenging for SARI and BERTScore to be used as evaluation metrics in this task. Thus, we use the chrF as our main evaluation metric for claim extraction.

Method	Sentence*			Dec. Sentence*		
	SARI	BERTScore	chrF	SARI	BERTScore	chrF
Claimbuster	6.23	82.7	24.3	6.24	82.8	24.5
Lead Sentence	6.41	83.4	23.8	-	-	-
LSA	5.56	83.2	24.1	5.57	83.2	24.3
TextRank	6.60	83.1	24.5	6.61	83.1	25.4
BertSum	6.54	83.6	25.6	6.55	83.6	25.9
Ours	6.56	83.7	25.9	6.70	83.8	26.4

Table A2: Results of Document-level CE on three different metrics.

A3 Implementation Details

All models we use in this paper are pre-trained models (*e.g.,* BertSum) or approaches that do not require training (*e.g.,* BM25). The hyperparameters of each model can be found in the original paper. To help readers reproduce our method, we have released our code on GitHub[5].

A4 ChatGPT for Decontextualisation

To verify how well ChatGPT would do on decontextualisation, we use ChatGPT to decontextualise three claim sentences in Figure 3. The ChatGPT prompt for decontextualisation is as follows:

ChatGPT Prompt

Claim: [claim]
Context: [context]

To rewrite the Claim to be understandable out of context based on the Context, while retaining its original meaning.

Decontextualised sentences produced by ChatGPT:
• Sentence 1: Barack Obama publicly expressed sympathy for ISIS and hinted at providing them with American resources during his 2008 State of the Union address.
• Sentence 2: During a specific event, there was a delay in obtaining approval from a certain authority for assistance requested by the Capitol Police.
• Sentence 3: The Indian government lacks adequate storage facilities for managing the large surplus of grain it possesses.

From the results, we can see that ChatGPT can produce well-formed claims that can be understood out of context, but it tends to rephrase the claim. In AVeriTeC (Appendix J.3.1), the decontextualised claims are required to be as close as possible to their original form. Our method tends not to change the original claims, but to rewrite only the ambiguous information units in claims, thus our generated claims are closer to the claims decontextualised by annotators than ChatGPT.

[5] https://github.com/Tswings/AVeriTeC-DCE