Tiny But Mighty: A Crowdsourced Benchmark Dataset for Triple Extraction from Unstructured Text

Muhammad Salman[1,*], Armin Haller[1], Sergio J. Rodríguez Méndez[1], Usman Naseem[2]
[1]School of Computing - CECC, The Australian National University, ACT, 2601, Australia
[2]School of Computing, Macquarie University, NSW, 2113, Australia
{Muhammad.Salman, Armin.Haller, Sergio.RodriguezMendez}@anu.edu.au, usman.naseem@mq.edu.au
* Corresponding Author

Abstract

In the context of Natural Language Processing (NLP) and Semantic Web applications, constructing Knowledge Graphs (KGs) from unstructured text plays a vital role. Several techniques have been developed for KG construction from text, but the lack of standardized datasets hinders the evaluation of triple extraction methods. The evaluation of existing KG construction approaches is based on structured data or manual investigations. To overcome this limitation, this work introduces a novel dataset specifically designed to evaluate KG construction techniques from unstructured text. Our dataset consists of a diverse collection of compound and complex sentences meticulously annotated by human annotators with potential triples (subject, predicate, object). The annotations underwent further scrutiny by expert ontologists to ensure accuracy and consistency. For evaluation purposes, the proposed F-measure criterion offers a robust approach to quantify the relatedness and assess the alignment between extracted triples and the ground-truth triples, providing a valuable tool for evaluating the performance of triple extraction systems. By providing a diverse collection of high-quality triples, our proposed benchmark dataset offers a comprehensive training and evaluation set for refining the performance of state-of-the-art language models on a triple extraction task. Furthermore, this dataset encompasses various KG-related tasks, such as named entity recognition, relation extraction, and entity linking.

Knowledge Graph (KG), Natural Language Processing (NLP), Text Annotation, Triple, Large Language Models (LLMs)

1. Introduction

Knowledge Graphs (KGs) have gained significant importance in a wide range of natural language processing (NLP) applications (Hogan et al., 2021). They serve as a valuable tool for organising information and extracting structured knowledge from unstructured data, such as plain text. Information in KGs is stored in a structured form, i.e., in the form of triples (subject, predicate, object), and the main source of information extraction is 'text', which is approximately 80% unstructured (Zong et al., 2021). Constructing KGs from unstructured text poses a challenge as KG requires the extraction of complete and accurate facts (triples) from the text. Many state-of-the-art KG construction methods have been developed, but they lack comparative analysis due to the unavailability of a benchmark dataset (Ji et al., 2021). To address this, a high-quality annotated dataset is essential for the evaluation of a model with competing techniques.

This paper introduces a novel dataset designed for triple extraction and validation from unstructured text. The dataset has been annotated by human annotators and verified by expert ontologists. It offers comprehensive coverage across various general domains and is enriched with high-quality annotations i.e., 96% verified. The dataset and evaluation criteria are publicly available[1] and can be leveraged by the research community.

To construct our dataset, we used an open-source dataset (Zhang et al., 2020), which is mainly based on Wikipedia text and used for *Split and Rephrase* benchmarking. The general benchmark of `'Small But Mighty'` serves our purpose because it contains compound and complex sentences and covers a wide range of textual domains. Before the annotation phase, we applied controlled sentence simplification techniques to the complex sentences contained within this dataset. This step ensured that the sentences were easily comprehensible for annotators, minimising the chances of missing any crucial fact during the labelling process. In the initial phase, a team of volunteer human annotators underwent training to label the text sentences according to our carefully developed annotation schema, which adhered to widely accepted standards in the field (Hogan, 2020). Subsequently, expert ontologists performed rigorous verification of the annotations in the second phase to maintain high quality and consistency. The workflow is shown in Figure-1.

Our dataset encompasses various KG construction tasks, such as entity recognition, relation extraction, and entity linking, all derived from unstructured text. We have created a valuable resource for researchers in NLP, information extraction, and related domains by employing a meticulous annotation process involving human annotators and expert

[1] `https://w3id.org/salmon/TinyButMighty`

ISA-20 Proceedings @LREC-COLING-2024, pages 71–81
20 May, 2024.

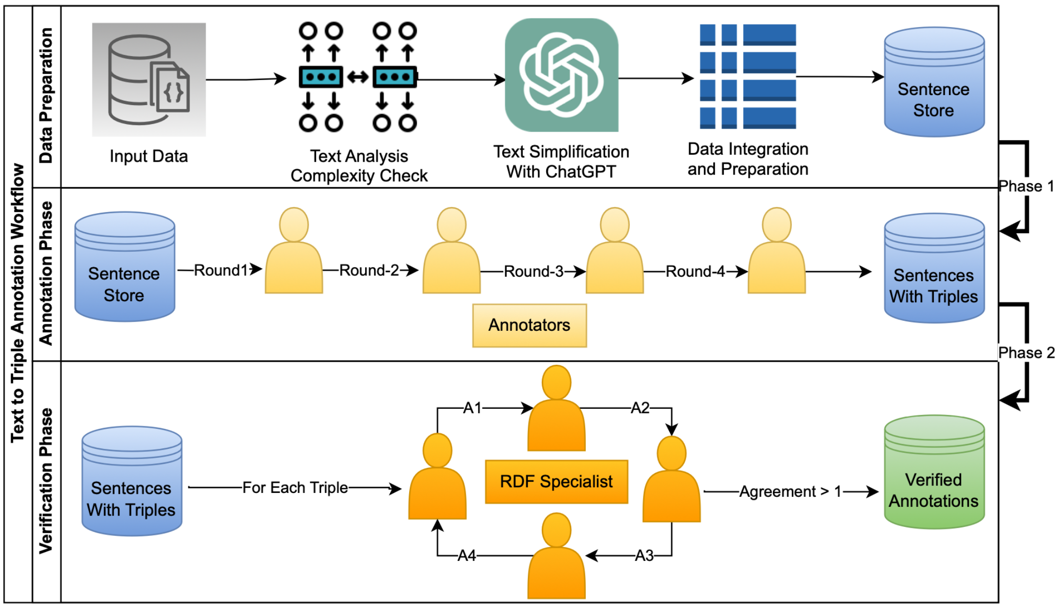


Figure 1: Workflow of Annotation Process

ontologists. The availability of this standardized dataset for KG construction from unstructured text, annotated and verified by experts, aims to stimulate further research and advancements in these critical areas of NLP. This paper has following **contributions**:

Web-based Annotation Tool: We have implemented a crowd-sourcing annotation system which can be used for multiple NLP annotation tasks.

Refinement of Existing Resource: We identified that the original simplified dataset still contains syntactic complexity and reduced that with OpenAI Generative Pre-trained Transformer (GPT), constructing a more robust simplified dataset.

Text-2-Triple Dataset: We were involved in a rigorous annotation process to create a novel dataset for triple extraction from unstructured text.

Evaluation Criterion: We have also proposed a triple-similarity evaluation criteria when the output triples are from unstructured text and cannot be identical to ground-truth triples.

2. Background and Related Work

Constructing KGs from unstructured text is a crucial task with wide-ranging applications in information retrieval, question answering, and other domains (Niklaus et al., 2018). KG construction from unstructured text has been a vibrant area of research in NLP (Zong et al., 2021; Heist et al., 2020; Gutiérrez and Sequeda, 2021), with various approaches proposed, including rule-based, machine learning-based, and hybrid methods that combine both (Hogan et al., 2021; Paulheim, 2017). These approaches typically involve identifying entities (Delpeuch, 2019) and relations (Sakor et al., 2020) in the text and constructing a graph that represents the relationships between the identified entities (Wang and Yang, 2019). While several datasets have been utilised in KG construction approaches (Al-Moslmi et al., 2020), most of them were manually crafted for specific tasks or curated and selected from structured data sources such as Wikipedia or DBpedia (Kertkeidkachorn and Ichise, 2017; Liu et al., 2018). Such datasets pose challenges in training and evaluating KG construction models on unstructured text data, which is typically more diverse and noisy.

In Table 1, we reviewed triple extraction techniques and investigated the evaluation method. It shows that the developed techniques provide no state-of-the-art evaluation and rely on their own investigation. In the task of RDF triple extraction from structured or semi-structured text, the *WebNLG* (Gardent et al., 2017) dataset is being widely adopted for evaluation. To evaluate an RDF triple extractor, *WebNLG* can be used in reverse, i.e., instead of `RDF_Triple-Generated_Text`, `Generated_Text` can be used as input to identify the `RDF_Triples`. For relation extraction, GraphRel (Fu et al., 2019) applied the WebNLG and New York Times (NYT) (Riedel et al., 2010) datasets. The *NYT* dataset is generated from news articles and is also a widely adopted dataset for relation extraction (RE) tasks, but its limited relations (three entity types and 24 relations) restrict it from evaluating the KG construction system from unstructured text. Recently, REBEL (Cabot and Navigli, 2021) has been trained on four differ-

Technique / Dataset	Target	Text Type	Evaluation
SEQ2RDF (Liu et al., 2018)	RDF Triple	Unstructured	✗
T2KG (Kertkeidkachorn and Ichise, 2017)	Triples	Natural Language	✗
FRED (Draicchio et al., 2013)	RDF, OWL	Natural Language	✗
Real Time RDF (Gerber et al., 2013)	RDF Triple	NEWS, Unstructured	✗
Exner System (Exner and Nugues, 2012)	RDF Triple	Wikipedia Articles	✗
UT2KB (Salim and Mustafa, 2021)	Relations	Unstructured	✗
Relation Extraction (Uddin et al., 2014)	Relations	Ebooks	✗
T2R (Hassanzadeh et al., 2013)	RDF Triple	Documents	✗
Seq2KG (Stewart and Liu, 2020)	RDF Triple	Domain Specific	✗
ER Extraction (Prasojo et al., 2016)	Entity, relations	Wikipedia Articles	✗
WebNLG (Gardent et al., 2017)	Text	RDF Triple	✗
NYT (Riedel et al., 2010)	RDF Triple	News	✗
GraphRel (Fu et al., 2019)	Relations	NEWS, Structured	✗
REBEL (Cabot and Navigli, 2021)	Relations	Semi-Structured	✗
CaRB (Bhardwaj et al., 2019)	Triple	Natural Language	✗
Our Benchmark	Triples	Unstructured Complex	✓

Table 1: State-of-the-Art methods For Triple Extraction and Evaluation Type

ent RE datasets and created a RE and classification dataset after fine-tuning and training on the BART-Large (Lewis et al., 2020) framework. CaRB (Bhardwaj et al., 2019) is also an addition to the community, but it seems limited in text domains and has not been qualitatively analysed.

Despite the success of information extraction approaches in different domains (Liu et al., 2023), there is still a need for high-quality annotated benchmarks for KG construction from unstructured text. This is particularly important as more complex and diverse data is becoming available in data lakes. We present a novel dataset for KG evaluation, providing a valuable resource for advancing the state-of-the-art in KG construction from unstructured text.

3. Dataset Creation and Annotation

The dataset creation and annotation workflow is shown in Figure-1, which involves the following steps.

3.1. Data Sources

To create our dataset, we started with an existing benchmark of complex sentences from IBM's `Split and Rephrase` corpus (Zhang et al., 2020). This benchmark comprises over 720 compound and complex sentences from general text domains. We selected the general domain dataset for annotation so that it could best evaluate the models for unstructured text.

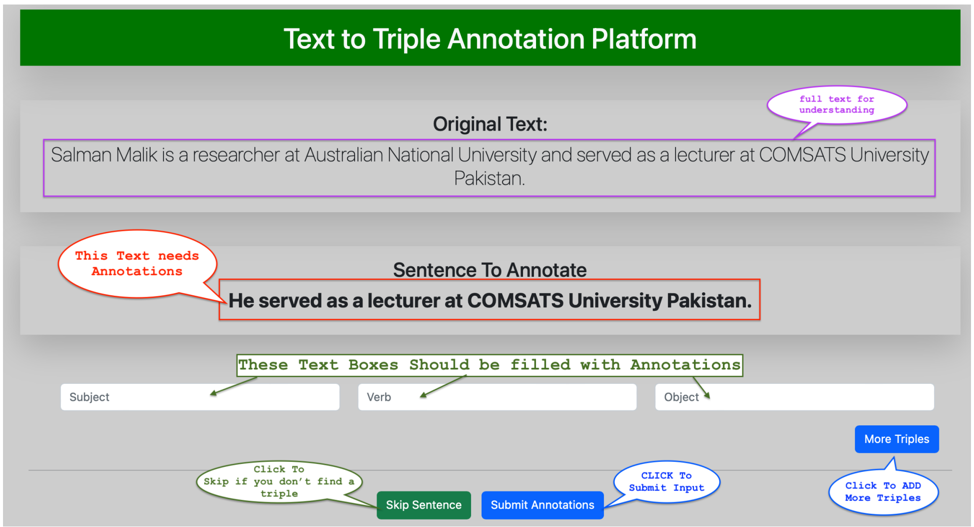
3.2. Data Refinement to Assist Phase-1 Annotators

The authors of the existing benchmark performed the "split and rephrase" function to transform complex sentences into simple sentences. However, our investigations noted that the existing simplification annotations are not robust enough for our purposes (shown in Table 2), as they often contain compound or complex sentences. After reviewing the potential limitation of the original corpus, we applied a method (Salman et al., 2023) to identify the syntactic complexity of the simplified text.

Based on our investigation, we developed a new "split and rewriting" module using GPT-3.5 (Floridi and Chiriatti, 2020), which enabled us to generate more accurate and meaningful simple sentences from the complex sentences in the dataset. In assessment, our pre-processed and re-annotated dataset has a more balanced distribution of complexity, as shown in Table 2. In this work, the sole purpose of sentence simplification is to assist Phase-1 annotators to get complete number of triple annotations. The sentence simplification in not part of Phase-2 and any further evaluation framework.

Description		Value
Complex Sentences		720
Simplified Sentences	IBM Corpus	3,565
	GPT Annotated	2,277
Simplified / Sentence	IBM Corpus	4.95
	GPT Annotated	3.16
Performance	IBM Corpus	**90.15%**
	GPT Annotated	**94.34%**

Table 2: Statistics of Complexity Measurement



Text to Triple Annotation Platform

Original Text:

Salman Malik is a researcher at Australian National University and served as a lecturer at COMSATS University Pakistan.

Sentence To Annotate

He served as a lecturer at COMSATS University Pakistan.

full text for understanding
This Text needs Annotations
These Text Boxes Should be filled with Annotations
Subject
Verb
Object
More Triples
Click To Skip if you don't find a triple
Skip Sentence
Submit Annotations
CLICK To Submit Input
Click To ADD More Triples

Figure 2: Platform Guidelines for Phase 1 Annotators

3.3. Annotation process: Phase 1

We invited participants through flyers in the department and ensured that no personal information was required in this process and that they could leave anytime. We recruited volunteer users to participate in the annotation process through a password-less Website, that does not require any personal information for sign-up. We enforced an `Exclusion Criteria` as well in which participants must have a command of the English language and the sentence structure of the language. We also described the annotation task to give the participants a brief understanding. These task-oriented briefings trained the participant for the annotation task. In the first phase of this annotation process, 127 participants voluntarily contributed to the task.

To ensure the quality and consistency of annotations, we followed a rigorous annotation process that involved four rounds of annotation by human volunteers. Each sentence went through four rounds of annotations, i.e., the round-1 annotations were evaluated by three participants in the following rounds. The annotators of each round were provided with clear guidelines and examples to ensure consistency in the annotation process and directed to a web-based system for annotations, as illustrated in the following section.

Web-based Annotation Tool For this task, we implemented a web-based tool to receive the annotations from participants. Ensuring the consistency of annotations, users are shown unique sentences in real-time with `'Concurrency Control'` while multi-user interaction with the platform. We recorded the elapsed time for each sentence's annotation. We deployed our web tool with Amazon Web Services (AWS) to ensure data safety and service reliability. The users were asked to access the annotation tool through a website link and presented with a simple sentence and the original complex sentence. We asked users to mark/identify the maximum possible triples (subject, predicate, object) in each simple sentence.

During the start of each annotation session, participants were briefed about the annotation tool, and guidelines were supplied to get accurate annotations. On the website, we also provided them with a mock annotation exercise on how to use the tool's different features and provide annotations as shown in Figure-2. Following are some of the features of the website.

`TEXT BOXES`: Input for subject, verb, and object is taken separately in text boxes as shown in Figure-3

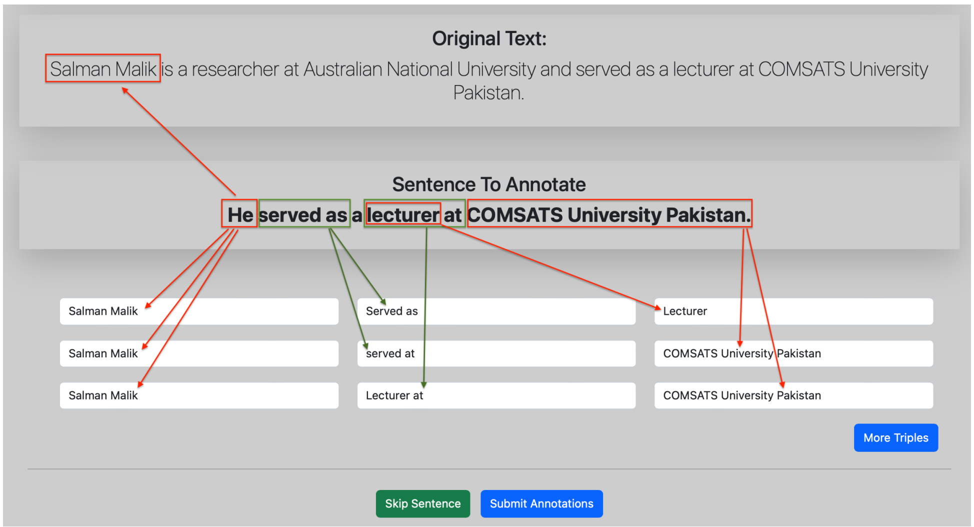
`MORE TRIPLES`: This button will add more text boxes if there is more than one triple in text.

`SKIP SENTENCE`: This button allows an annotator to skip a sentence for which there are no triples, as per the understanding of the annotator.

`Annotation with Multiple Triples`: A sentence is labelled with multiple triples contained in the sentence. Moreover, we encouraged the annotator to analyse *'Original Text'* to replace pronouns (he, she, and it, etc.) with the actual entity/noun label.

3.4. Challenges

During the annotation process, we analysed that participants were unfamiliar with KG concepts even though they were briefed, specifically about the notion of triples. To make it simple for the first round, we asked participants to identify the subject, verb, and object in the sample sentence. The resultant triples of the first annotation round were not of high quality and contained some "nonsensical facts". In the annotations, predicates were verbs only instead



Original Text:

Salman Malik is a researcher at Australian National University and served as a lecturer at COMSATS University Pakistan.

Sentence To Annotate

He served as a lecturer at COMSATS University Pakistan.

Salman Malik	Served as	Lecturer
Salman Malik	served at	COMSATS University Pakistan
Salman Malik	Lecturer at	COMSATS University Pakistan

Figure 3: Entity Co-reference Resolution Handling

of proper prepositions, which sometimes drew confusion among the participants, e.g., `work` instead of `work in`, `work at`, `work from`, etc. Furthermore, we also identified that the participants were not incorporating the entity co-reference resolution in the annotations, i.e., the participants were reporting the pronouns `(He, She, It, and They, etc.)` from the simple sentence instead of referring to proper noun (name or tile) from the original text.

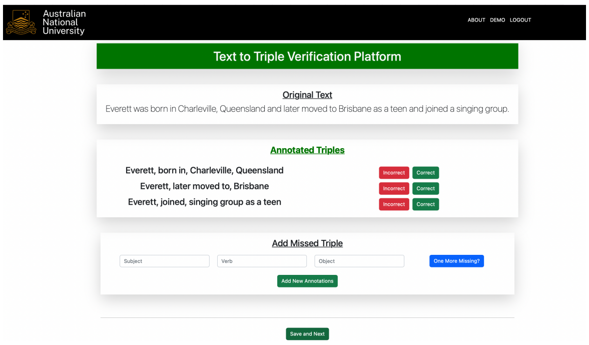
Therefore, we provided some further guidelines to get an improved version of annotations in the following rounds (the guideline for entity co-reference is shown in Figure-3). From the third round of annotation, we observed consistency in annotations between participants, which became more prominent in the fourth round.

3.5. Quality Assurance: Phase 2

To ensure the high quality of the annotations, we had each sentence annotated at least four times by different users in each round. We also provided clear guidelines and examples to the users to ensure consistency in the annotation process. After each round of annotation, we reviewed and refined the annotations to improve their quality and accuracy. We then finalised the annotations from the last round and performed a final entity co-reference resolution task to ensure consistency in the annotations across multiple sentences.

Finally, we involved expert ontologists in verifying each annotated triple's correctness and overall annotations based on the original sentence. This process ensured that the resulting dataset was accurate, reliable and suitable for training and evaluating KG construction models from unstructured text. For each annotated triple in Phase-1, we asked the participants of Phase-2 to verify the quality of triples based on the following questions.

1. Is the annotated triple 'Correct'?
2. Is the annotated triple 'Partially Incorrect'?
3. Is the annotated triple nonsensical or vague?



Text to Triple Verification Platform

Original Text

Everett was born in Charleville, Queensland and later moved to Brisbane as a teen and joined a singing group.

Annotated Triples

Everett, born in, Charleville, Queensland	Incorrect	Correct
Everett, later moved to, Brisbane	Incorrect	Correct
Everett, joined, singing group as a teen	Incorrect	Correct

Add Missed Triple

Figure 4: Annotation Verification Platform

4. Dataset Characteristics

The dataset comprises a collection of complex sentences, which are rewritten into more straightforward simplified sentences using OpenAI [2] as mentioned in Section 3.2. The original complex sentences were obtained from an already published benchmark dataset for text simplification (Zhang et al., 2020). The dataset comprises of 720 complex and 2,277 simple sentences, annotated with all possible triples (subject, predicate, object) by human volunteers, as explained previously.

[2] `https://openai.com`

Phase-1		Phase-2	
# of Complex Sentences	720	**Data Sample**	10%
# of Simplified Sentences	2,277	**# of Triples in Data Sample**	317
# of Annotators	127	**# of Experts**	5
# of Annotation Rounds	4	**# of Correct Triples**	242 (76.34%)
# of Total Annotated Triples	11,425	**# of Partially Correct Triples**	63 (19.87%)
# of Triples in Final Round	3,736	**# of Nonsensical Triples**	12 (3.78%)
Avg. Triples Per Simplified Sentence	1.64	**Verified Triples After Phase-2**	**305**
Avg. Triples Per Original Sentence	5.18	**Success Rate of Annotations**	**96.22%**

Table 3: Statistics of Annotated Dataset for Each Phase

Each simple sentence is part of the original complex sentence, and the annotations include all possible triples that could have been extracted from the simple sentence. The dataset covers a wide range of topics, including science, technology, history, and literature. The complexity of the sentences varies, ranging from moderately complex to compound and complex sentences, making it suitable for evaluating KG construction models from unstructured text.

4.1. Statistics

In this section, we provide the statistics of the annotation process and the contribution of volunteer participants. In phase-1, 127 volunteer annotators participated in the annotation process. We conducted four rounds of annotations; in each round, more than 2,200 annotations were recorded. A total of 11,425 triple annotation hits (add, edit, delete) were recorded. We observed refinement in the quality of annotations in each round and got 3,736 triple annotations for 720 complex sentences from the final round. The statistical summary of the dataset is presented in Table 3. We recorded an average of 5.18 triples per original sentence, depicting the complex nature of sentences. Moreover, each simplified sentence has an average of 1.64 triples, proving the meaning-preserving and fair dispersal of complexity after applying sentence simplification.

To verify the quality of phase-1 annotations, we invited expert ontologists to mark the sampled data as discussed in section 3.5. We removed the simplified sentence layer, and the participants were presented with the original sentence and its annotations only. We chose 80 unique complex sentences in our sample data that were labelled with `317` triple annotations in Phase-1. 76.34% of annotated triples are verified as `'Correct'` while 19.87% were marked as `'Partially Incorrect'` and edited with correct entity/predicate mentions. These updated triples were looped in the verification phase to take the agreement from other specialists and are marked as `'Correct'`. Combining the verified triples, we have an agreement of verification on 96.22% by expert ontologists while 3.78% are marked as `'Incorrect or Nonsensical'`. The relative distribution frequency of the correctness rating by expert ontologists is shown in Figure-5, and the statistics of Phase-2 sampled data are shown in Table 3.

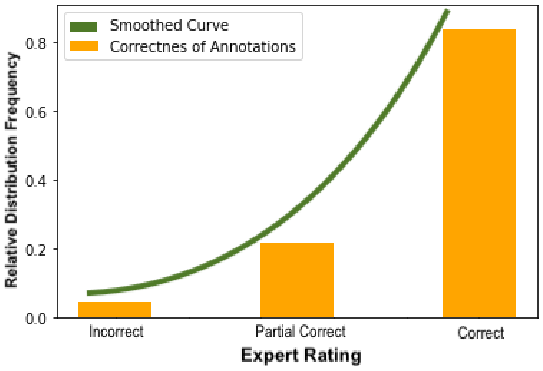


Figure 5: Relative Distribution of Correctness from Expert Ratings

4.2. Inter-Annotation Agreement (IAA)

The calculated Cohen's Kappa (κ) coefficients for Phase 1 revealed inter-rater reliability scores of 0.63, 0.78, and 0.84 for the rounds of annotation R1-R2, R2-R3, and R3-R4, respectively. The initial lower agreement in the R1-R2 round was attributed to inconsistencies observed in the use of the "is-a" relation and challenges in entity co-reference resolution, as analysed from the annotations in Phase 1 (Round #1). In response to these challenges, we introduced additional guidelines detailed in Section 3.4. The subsequent annotation rounds demonstrated improved consistency, evident from the enhanced κ scores observed between R2-R3 and R3-R4.

In Phase 2, the κ scores consistently exceeded 0.97, an authentication of the annotators' expertise within the research domain. The qualitative analysis phase agreed that an annotation must acquire at least two agreements to be deemed 'correct'. The overall score was thus established on annota-

Algorithm 1 F-Measure/Similarity of Triple-sets Calculation Criteria

```
Input: Triples $T$ and Ground Truth $GT$
Output: F-Measure
Coefficient: COSINE, JACCARD
procedure CALCULATEFMEASURE(T, GT)
    Initialize Relatedness($T$) to 0
    Initialize Similarity($t_{ij}$) for each $t_i$ in $T$ and $gt_j$ in $GT$ to an empty list
    Initialize $Adjustment$ as $\max(\text{LEN}(T), \text{LEN}(GT))$
    for all $t_i$ in $T$ do
        Initialize $Similarity_{t_i}$ as an empty list
        for all $gt_j$ in $GT$ do
            $Similarity(t_{ij}) \leftarrow \texttt{Coefficient}(t_i, gt_j)$
            Append $Similarity(t_{ij})$ to $Similarity_{t_i}$
        end for
        $Score_{t_i} \leftarrow \max(Similarity_{t_i})$
        Relatedness($T$) += $Score(t_i)$
        Relatedness($T$) $\leftarrow$ Penalty|Amnesty                ▷ Based on Threshold
    end for
    $F$-Measure (T) $\leftarrow \frac{Relatedness(T)}{Adjustment}$
    return $F$-Measure
end procedure
```

tions that attained consensus among the reviewers. This rigorous consensus requirement supports the reliability and validity of the annotation process, contributing to the exceptionally high agreement rates observed in Phase 2.

5. Experiments

5.1. Evaluation Criterion

Algorithm 1 calculates the F-Measure based on the given triples (T) and ground truth (GT) triples. T and GT sets are comprised of the model's output and verified annotations by experts, respectively. The algorithm iterates over each triple extracted by the model and calculates the similarity between each triple (t_i) and each ground truth (gt_j). The maximum similarity value is awarded as the score for a given t_i. The similarity of each triple is also awarded a penalty or amnesty based on thresholds. F-Measure is computed by taking the average of total similarity of T w.r.t. the maximum number of triples in T or GT to penalise the model for less/more generated triples. Finally, a penalty and amnesty are awarded again for final computation. In the penalising scheme, a triple is considered Incorrect if the relatedness is less than 50% of the ground truth triple. Conversely, in an amnesty scheme, an extracted triple is considered Correct if it has relatedness more than 80% with the ground truth triple.

In summary, our algorithm iterates over the triples and ground truths, computes similarity scores, accumulates the relatedness, and calculates the F-Measure. By incorporating penalties or amnesty based on threshold values, the algorithm allows for flexible adjustment and evaluation of the relatedness and F-Measure.

$$\texttt{SCORE (t)} = \begin{cases} \texttt{1}, & \text{if } Similarity_{t,gt} > 0.80 \\ Sim_{t,gt} & \text{otherwise} \\ \texttt{0}, & \text{if } Similarity_{t,gt} < 0.50 \end{cases}$$

5.2. Baseline Methods

We applied some baseline techniques to our newly annotated dataset for preliminary evaluation. For this purpose, we selected two well-known language processing libraries and one generative pre-trained model (GPT-4).

SpaCy's SVO Model *Textacy* is designed on a high-performance and widely-used NLP library (SpaCy) for text processing. Specifically, we leverage *Textacy's* Subject-Verb-Object extraction feature for preliminary probes of our evaluation criterion and benchmark dataset.

CoreNLP OpenIE Triple Extractor Stanford's CoreNLP OpenIE triple extractor model was tuned to retrieve the maximum triples from a given sentence without any restrictions. While investigating the resulting triples, the Stanford OpenIE model generated non-informative triples for some sentences, such as [Airport, is, Located]. The implementation of Stanford OpenIE is publicly available[3] and can be accessed and used in multiple ways, including as a Python library and wrapper.

[3] https://github.com/philipperemy/stanford-openie-python

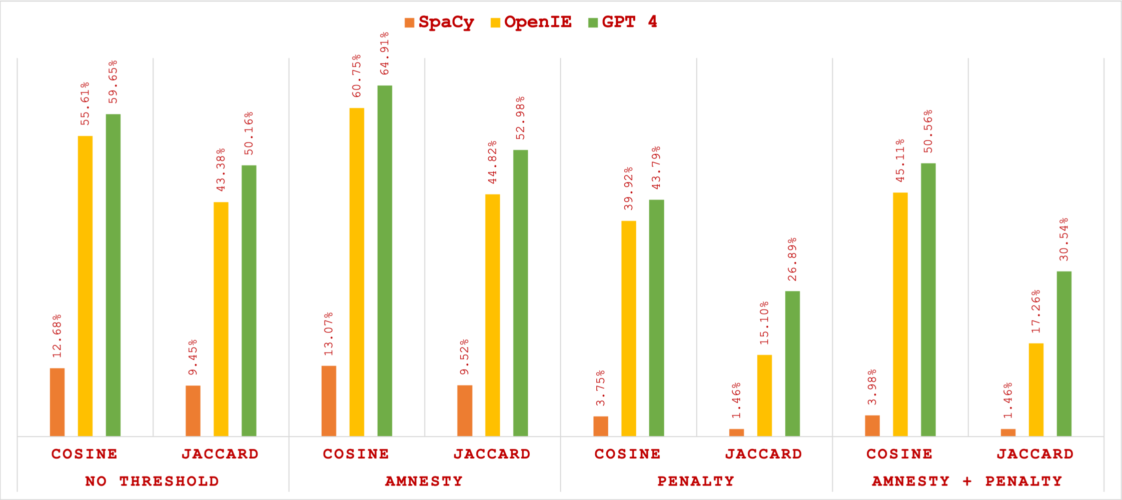


Figure 6: Performance of Models on Verified Sampled Data

GPT-4 Through OpenAI API For the purpose of evaluating triple extraction, we employed a GPT prompt-based approach. Specifically, we utilized the GPT-4 model as a triple extraction tool through *OpenAI* API that elicits relevant triples from the input text. After trying multiple query prompts, we settled on the best-resulting prompt for triples identification.

5.3. Results and Discussion

In this section, we will discuss the performance of our preliminary baselines w.r.t. evaluation framework illustrated in the prior section. As shown in Fig 6, GPT-4 is leading the baseline methods in both similarity coefficients. On the other hand, the performance of SpaCy's SVO model is relatively low because of its incapability to deal with sentences of complex structure. CoreNLP OpenIE also performed very well and competed with the GPT-4 in all measures; however, we investigated that OpenIE generated 834 triples while the ground truth dataset contains 326 triples for sampled data. SpaCy and GPT-4 triple extractor models generated 59 and 261 triples, respectively. In our evaluation framework, we have taken care of the number of triples identified by the model and penalised a model's output with the normalization of the overall relatedness score. We have also applied `Fuzzy Similarity` to evaluate the quality of extracted triples. In fuzzy ratio along with amnesty and penalty, SpaCy, OpenIE and GPT achieved `13.9%`, `56.68%`, and `60.81%` respectively in triple qualitative analysis.

GPT-4 outperformed other baselines in all aspects of the evaluation. The quality of extracted triples from GPT-4 is also of high quality because it generated fewer `(261 VS 326)` triples than the ground truth data but still managed to lead the performance table in all measures. Although GPT is the best-performing model but there is still a huge margin of improvement. The inclusion of this resource in the fine-tuning process can enhance the large language models' (LLMs) understanding and generation capabilities, enabling them to generate more accurate and contextually appropriate triples in KG construction tasks. Furthermore, with its wide range of domain-specific and general knowledge triples, the dataset presents an opportunity to improve the accuracy, reliability, and contextual awareness of LLMs, ultimately benefiting a variety of downstream applications, including question-answering systems, information retrieval, and KG construction.

6. Conclusion

This work presents a novel dataset to evaluate the KG construction tasks from unstructured text. The dataset comprises a collection of compound and complex sentences, which have been annotated with possible triples (subject, verb, object) by human volunteers. Expert ontologists have verified the annotations to ensure their correctness and consistency. We have demonstrated the potential of this dataset by using it to evaluate KG construction models and tools. We also proposed an algorithm that offers a robust approach to quantifying the relatedness and assessing the alignment between extracted triples and ground truth data, providing a valuable tool for evaluating the performance of triple extraction systems. Similar to the relatedness score, we subjected the F-Measure to penalty and amnesty based on threshold values to account for the quality of the matches. The results show that our dataset can improve the performance of KG construction models, especially in terms of extracting complete and accurate triples from unstructured text.

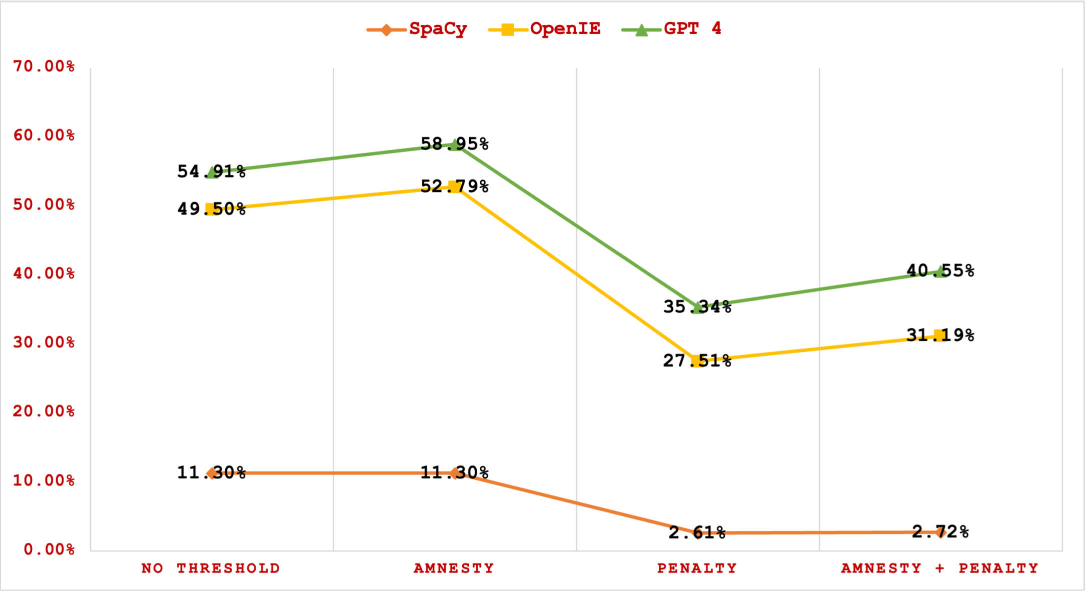


Figure 7: Average Performance of Models

The inclusion of this dataset in the fine-tuning process of an LLM can enhance its understanding and generation capabilities, enabling it to generate more accurate and contextually appropriate triples from unstructured text to construct a KG. Furthermore, the dataset facilitates the evaluation of triple extraction systems and contributes to advancing the research and development of NLP tasks related to knowledge graph construction and information extraction. With its wide range of domain-specific and general knowledge triples, the dataset presents an opportunity to improve the accuracy, reliability, and contextual awareness of LLMs, ultimately benefiting a variety of downstream applications, including question-answering systems, information retrieval, and knowledge graph generation.

Limitations

We acknowledge that there may be a need for larger datasets; however, as discussed in our evaluation of state-of-the-art models, the proposed benchmark dataset is sufficiently large to distinguish significant differences in accuracy for benchmark algorithms. The dataset is relatively small as compared to other text-related corpora and currently stands at 720 Complex and 2,277 simple sentences with high-quality annotations. However, this aspect may affect the dataset generalization, especially when training KG construction models that require a large amount of training data. Therefore, researchers should keep in mind the size of the dataset when using it and may need to supplement it with additional data if necessary. To address this limitation, this dataset will be used to fine-tune LLMs to make them capable of annotating large amounts of text.

This dataset is purely designed to evaluate the triple extraction system from unstructured text. However, it does not deal with RDF triples at this stage and we intend to transform the predicated as per RDF standards in our next release. We also intend to investigate the following research question while extending our dataset with Wikidata Mappings.

“Can KG construction from unstructured text data be improved by incorporating external knowledge sources such as a domain-specific ontology or open-domain knowledge bases?"

Ethics Statement

Since this study required human subjects to annotate and review the textual data (a "human-in-the-loop approach"), following a proper procedure to obtain ethical approval (protocol) was compulsory. A total of 127 participants were recruited to fulfil this purpose in Phase-1. Firstly, we obtained ethical approval for the annotation protocol from our University’s research ethics committee. Under the approved research ethics (`ANU Ethics Protocol 2022/464`), we ensured the privacy and safety of participants. In other aspects of the protocol, volunteer participation to annotate the data is enforced, meaning there is no pressure or workload assignment from any course or program. Participants are also provided with the option to withdraw from the annotation process at any time. We also conveyed to the participants that the dataset would be publicly available to the research community.

References

Tareq Al-Moslmi, Marc Gallofré Ocaña, Andreas L Opdahl, and Csaba Veres. 2020. Named entity extraction for knowledge graphs: A literature overview. *IEEE Access*, 8:32862–32881.

Sangnie Bhardwaj, Samarth Aggarwal, and Mausam Mausam. 2019. CaRB: A crowdsourced benchmark for open IE. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6262–6267, Hong Kong, China. Association for Computational Linguistics.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. Rebel: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381.

Antonin Delpeuch. 2019. Opentapioca: Lightweight entity linking for wikidata. *arXiv preprint arXiv:1904.09131*.

Francesco Draicchio, Aldo Gangemi, Valentina Presutti, and Andrea Giovanni Nuzzolese. 2013. Fred: From natural language text to rdf and owl in one click. In *The Semantic Web: ESWC 2013 Satellite Events: ESWC 2013 Satellite Events, Montpellier, France, May 26-30, 2013, Revised Selected Papers 10*, pages 263–267. Springer.

Peter Exner and Pierre Nugues. 2012. Entity extraction: From unstructured text to dbpedia rdf triples. In *WoLE@ ISWC*, pages 58–69.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. 2019. Graphrel: Modeling text as relational graphs for joint entity and relation extraction. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1409–1418.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for nlg micro-planning. In *55th annual meeting of the Association for Computational Linguistics (ACL)*.

Daniel Gerber, Sebastian Hellmann, Lorenz Bühmann, Tommaso Soru, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. 2013. Real-time rdf extraction from unstructured data streams. In *The Semantic Web–ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I 12*, pages 135–150. Springer.

Claudio Gutiérrez and Juan F Sequeda. 2021. Knowledge graphs. *Communications of the ACM*, 64(3):96–104.

Kimia Hassanzadeh, Marek Reformat, Witold Pedrycz, Iqbal Jamal, and John Berezowski. 2013. T2r: System for converting textual documents into rdf triples. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 3, pages 221–228. IEEE.

Nicolas Heist, Sven Hertling, Daniel Ringler, and Heiko Paulheim. 2020. Knowledge graphs on the web-an overview.

Aidan Hogan. 2020. Resource description framework. *The Web of Data*, pages 59–109.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. 2021. Knowledge graphs. *ACM Computing Surveys (CSUR)*, 54(4):1–37.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*.

Natthawut Kertkeidkachorn and Ryutaro Ichise. 2017. T2kg: An end-to-end system for creating knowledge graph from unstructured text. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yue Liu, Tongtao Zhang, Zhicheng Liang, Heng Ji, and Deborah L McGuinness. 2018. Seq2rdf: An end-to-end application for deriving triples from natural language text. In *CEUR Workshop Proceedings*, volume 2180. CEUR-WS.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2018. A survey on open information extraction. *arXiv preprint arXiv:1806.05599*.

Heiko Paulheim. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 8(3):489–508.

Radityo Eko Prasojo et al. 2016. Entity-relationship extraction from wikipedia unstructured text. In *DC@ ISWC*, pages 74–81.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III 21*, pages 148–163. Springer.

Ahmad Sakor, Kuldeep Singh, Anery Patel, and Maria-Esther Vidal. 2020. Falcon 2.0: An entity and relation linking tool over wikidata. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3141–3148.

Mustafa Nabeel Salim and Ban Shareef Mustafa. 2021. Uttokb: a model for semantic relation extraction from unstructured text. In *2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 591–595. IEEE.

Muhammad Salman, Armin Haller, and Sergio J Rodríguez Méndez. 2023. Syntactic complexity identification, measurement, and reduction through controlled syntactic simplification. *arXiv preprint arXiv:2304.07774*.

Michael Stewart and Wei Liu. 2020. Seq2kg: an end-to-end neural model for domain agnostic knowledge graph (not text graph) construction from text. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, volume 17, pages 748–757.

Ashraf Uddin, Rajesh Piryani, and Vivek Kumar Singh. 2014. Information and relation extraction for semantic annotation of ebook texts. In *Recent Advances in Intelligent Informatics: Proceedings of the Second International Symposium on Intelligent Informatics (ISI'13), August 23-24 2013, Mysore, India*, pages 215–226. Springer.

XiuQing Wang and ShunKun Yang. 2019. A tutorial and survey on fault knowledge graph. *Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health*, pages 256–271.

Li Zhang, Huaiyu Zhu, Siddhartha Brahma, and Yunyao Li. 2020. Small but mighty: New benchmarks for split and rephrase. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1198–1205. Association for Computational Linguistics.

Chengqing Zong, Rui Xia, and Jiajun Zhang. 2021. Information extraction. In *Text Data Mining*, pages 227–283. Springer.