ITTC at SemEval 2023-Task 7: Document Retrieval and Sentence Similarity for Evidence Retrieval in Clinical Trial Data

Rahmad Mahendra and **Damiano Spina** and **Karin Verspoor**
RMIT University
Melbourne, Australia
rahmad.mahendra@student.rmit.edu.au
{damiano.spina,karin.verspoor}@rmit.edu.au

Abstract

This paper describes the submissions of the Natural Language Processing (NLP) team from the Australian Research Council Industrial Transformation Training Centre (ITTC) for Cognitive Computing in Medical Technologies to the SemEval 2023 Task 7, i.e., multi-evidence natural language inference for clinical trial data (NLI4CT). More specifically, we were working on subtask 2 whose objective is to identify the relevant parts of the premise from clinical trial report that justify the truth of information in the statement. We approach the evidence retrieval problem as a document retrieval and sentence similarity task. Our results show that the task poses some challenges which involve dealing with complex sentences and implicit evidences.

1 Introduction

Clinical Trial Reports (CTRs) are an important source of evidence for clinical practitioners, which provide information on the safety and efficacy of new treatments and interventions (Akobeng, 2005). However, in recent years, the number of CTRs published has increased significantly, making it difficult for clinicians to stay up-to-date with the latest research (Bastian et al., 2010). This challenge is further exacerbated by the growing number of patient cases that require customized treatment plans. As the result, it has become infeasible for practitioners to manually sift through all the current literature to provide personalized evidence-based care to their patients.

To overcome this issue, there is a need for technologies that can help clinicians efficiently review and interpret the latest medical literature. One such tool is Natural Language Inference (NLI), a task in natural language processing (NLP) to determine the semantic relation between two sentences (Bowman et al., 2015). NLI-powered systems offer a promising solution by extracting relevant information from CTRs and enabling automated evidence synthesis (DeYoung et al., 2020). By leveraging NLI, we could significantly improve the way we link the most current evidence to offer individualized evidence-based care (Sutton et al., 2020).

Task. SemEval 2023 Task 7: Multi-evidence Natural Language Inference for Clinical Trial Data (NLI4CT) consists of two subtasks, i.e., textual entailment and evidence retrieval (Jullien et al., 2023). In the first subtask, the system is challenged to decide whether a statement making claims about a single CTR (or comparing 2 CTRs) is true or not. In other words, the objective of Subtask 1 is to determine the inference relation (entailment vs. contradiction) between CTR and statement pairs. Subtask 2 involves a CTR premise and statement, and requires the system to select relevant facts extracted from the premise to support the statement. These facts are necessary to justify the inference label of Subtask 1.

Dataset. The NLI4CT shared task uses a collection of breast cancer CTRs. A team of clinicians generated a dataset of 2,400 statements from separate section, annotated the gold standard label for the pairs of statement and corresponding CTR, and provided evidence explaining the inference.

CTRs are divided into four sections: "Eligibility Criteria" describe the conditions necessary for patients to be included or excluded in the clinical trial; "Intervention" provides information on the treatments being studied; "Results" show the number of participants and outcome measures; and "Adverse Events" highlight signs and symptoms observed in patients during the clinical trial. While "Eligibility Criteria" contain free text, "Results" and "Adverse Events" are originally tabular data that contains numerical values. On the other hand, "Intervention" is a k-column table (k is number of interventions); the cells in the table contain natural sentences. Each statement in NLI4CT dataset is based on specific section of CTR.

Table 1: List of additional stopwords extracted from clinical trial data

achieved	*administered*	*adverse*	*better*	*both*	*between*	*cannot*	*case*	*cases*	*cohort*
cohorts	*common*	*confirmed*	*depending*	*diagnosed*	*did*	*difference*	*different*	*do*	*does*
each	*either*	*eligible*	*event*	*events*	*every*	*excluded*	*experienced*	*group*	*groups*
having	*higher*	*included*	*ineligible*	*intervention*	*interventions*	*least*	*less*	*lower*	*many*
may	*minimum*	*more*	*most*	*must*	*none*	*number*	*observed*	*occurred*	*one*
outcome	*over*	*part*	*participant*	*participants*	*participate*	*participating*	*participation*	*patient*	*patients*
primary	*prior*	*receive*	*received*	*receiving*	*recently*	*record*	*recorded*	*reported*	*require*
required	*requires*	*response*	*results*	*same*	*secondary*	*several*	*still*	*study*	*subjects*
suffered	*take*	*treated*	*treatment*	*trial*	*types*	*undergo*	*unit*	*use*	*used*
uses	*whereas*	*who*	*will*	*would*					

2 System Description

We participated in Subtask 2, evidence retrieval, for this shared task. We approach the problem in evidence retrieval subtask in two different formulations, that are document retrieval and sentence similarity. Suppose, we have a statement S and chunks $c_1, c_2, ..., c_k$ of CTR (a chunk can be single or several sentences, or table cell).

Document Retrieval We treat S as a query and chunks c_i as documents to be indexed and retrieved. Our retrieval model is based on BM25 algorithm (Robertson et al., 1994) using `rank_bm25` implementation (Brown, 2020). We use the configuration $b = 0.75$, $k_1 = 1.5$ for Okapi's BM25 and retrieve all documents c_i with positive BM25 score for each query as the evidence. We pre-process the query by removing stopwords. In addition to NLTK stopwords (Bird and Loper, 2004), we also exclude frequent terms found in CTR from the query (e.g., "patient", "trial"). Full list of CTR frequent terms are presented in Table 1.

Several statements are complex sentences. We perform sentence decomposition to those statements. Before decomposing the sentence, we apply NLP pre-processing, i.e., syntactic parsing and coreference resolution using AllenNLP implementation (Gardner et al., 2018). As an example, we decompose the statement "*Female cancer patients over the age of 18 can participate in the primary trial, regardless of race, ethnic origin or cancer type, however for the secondary trial, they must have BRCA+ breast cancer.*" into two sentences: (i) "*Female cancer patients over the age of 18 can participate in the primary trial, regardless of race, ethnic origin or cancer type.*" and (ii) "*For the secondary trial, female cancer patients over the age of 18 must have BRCA+ breast cancer.*". We first need to resolve nominal mention "*female cancer patients over the age of 18*" as the antecedent for pronoun "*they*". Then, we decompose the complex sentence into atomic sentences using rule-based approach based on sentence structure. We identify coordinating conjunction and its scope and separate two clauses with subordinating conjunction.[1]

Sentence Similarity We compute similarity between all $\langle S, c_i \rangle$ pairs. For $\langle S, c_i \rangle$ with the similarity greater than a given threshold, we conclude that c_i is the evidence to justify the truth of the statement S. We use Word Mover Distance (WMD) (Kusner et al., 2015) as the sentence similarity measurement in our experiment.

3 Results and Discussion

We submit six runs using BM25 and Word Mover Distance algorithms. The results of the submitted runs compared with a baseline are listed in Table 2. As a baseline, we use a perfect Recall setting, in which the system retrieves all contents from CTR as the evidence. We report results using the effectiveness evaluation measures officially used in the shared task, i.e., Precision, Recall, and F1 score, for both development (dev) and test sets.

The baseline achieves perfect Recall (by design) and an F1 score of 0.622 and 0.716 for the dev and test sets, respectively. The rest of our submitted runs were not able to outperform these scores due to a substantially lower Recall (no greater than 0.65). However, some of our runs have a 0.2 improvement in Precision w.r.t. the baseline; while the baseline obtains 0.558 for Precision in the test set, our runs utilizing BM25 with stopword removal and WMD score Precision 0.755 and 0.763, respectively.

Some queries do not return any relevant document – penalizing Recall – even using semantic sentence similarity. In order to improve Recall, we apply the closure strategy. Specifically, for statements without any documents retrieved when using

[1]We initially explored ABCD framework (Gao et al., 2021) for sentence decomposition task. However it failed to generate correct and grammatical sentences for majority of samples we examined.

Table 2: Subtask 2 results for the dev and test sets.

	Dev			Test		
	Precision	Recall	F1	Precision	Recall	F1
Retrieve all	0.451	**1.000**	**0.622**	0.558	**1.000**	**0.716**
BM25	0.441	0.382	0.410	0.509	0.400	0.448
BM25 + closure	0.504	0.537	0.520	0.557	0.511	0.533
BM25 + stopw	**0.667**	0.189	0.295	0.755	0.195	0.310
BM25 + stopw + closure	0.640	0.518	0.573	0.686	0.565	0.619
WMD	0.666	0.147	0.241	**0.763**	0.153	0.254
WMD + closure	0.427	0.651	0.516	0.510	0.645	0.570

Table 3: Results for Subtask 2 on dev set (breakdown by number of CTR used).

	Single (dev)			Comparison (dev)		
	Precision	Recall	F1	Precision	Recall	F1
Retrieve all	0.428	**1.000**	**0.599**	0.482	**1.000**	**0.651**
BM25	0.442	0.426	0.434	0.439	0.330	0.377
BM25 + closure	0.471	0.499	0.485	0.542	0.583	0.562
BM25 + stopw	**0.744**	0.242	0.365	0.541	0.127	0.206
BM25 + stopw + closure	0.635	0.495	0.557	0.646	0.546	0.592
WMD	0.599	0.132	0.216	**0.744**	0.166	0.272
WMD + closure	0.349	0.565	0.431	0.534	0.752	0.625

either algorithm, we consider all candidate chunks as the evidence. In other words, we apply a perfect Recall strategy only for the queries failing to return any document. Recall increases 3–4 times by employing the closure. For example, WMD on test set has Recall improvement from 0.153 to 0.645. However, the closure strategy penalizes Precision. The best results in our experiments are achieved by BM25 with stopword removal and using closure (dev set, F1 0.573 and test set, F1 0.619).

Statements in the dataset belong to one of two types. The "Single" type comprises of only one Primary trial, where the assertion made in the statement will solely pertain to this trial. On the other hand, the "Comparison" type involves two CTRs: a Primary trial and a Secondary trial, and the evidence for claims made in the statements have to be retrieved from both trials. By analyzing the evidence retrieval by type of CTR, we find that higher performance is obtained by the baseline and the system variations applying the closure difference, as shown in Table 3.

A detailed look at the evidence retrieval performance breakdown by CTR section used as the statement reveal that there are difference in the distribution of evidence among section types. From the baseline performance reported in Table 4, we can see that all instances in dev set with "Intervention" type need to use all chunks in CTR premise as the evidence. For this case, all documents must be relevant and should be retrieved. On the other hand, only a tiny portion of sentences are relevant for the "Eligibility" type (fewer than 20% of sentences in the CTR are relevant as evidence). It is not sufficient to retrieve the most relevant documents based on lexical or semantic similarity from eligibility criteria in CTR (Truong et al., 2022).

By analyzing the experimental result of evidence retrieval, we identify several cases causing low Recall in our system. The retriever fails to retrieve domain-specific vocabulary, especially for BM25 algorithm that relies on term-matching. Query expansion may address this issue. Other instances containing numbers or requiring numerical interpretation (Thawani et al., 2021) may also not be retrieved. For example, the chunk "*Age 18 years*" from Eligibility section in CTR should be the evidence for the statement "*Only adults can take part in the primary trial*", but our current system fails to retrieve it.

Evidence retrieval for supporting natural language inference tasks also poses other challenges

Table 4: Results for Subtask 2 on dev set (breakdown by CTR section).

	Eligibility		Results		Intervention		Adverse Events	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Retrieve all	0.195	**1.000**	0.542	**1.000**	1.000	**1.000**	0.666	**1.000**
BM25	0.226	0.577	0.615	0.520	1.000	0.463	0.766	0.122
BM25 + closure	0.226	0.577	0.626	0.546	1.000	0.476	0.765	0.531
BM25 + stopw	**0.410**	0.229	**0.768**	0.223	1.000	0.321	0.732	0.089
BM25 + stopw + closure	0.380	0.389	0.539	0.420	1.000	0.565	**0.809**	0.662
WMD	0.339	0.109	0.725	0.124	1.000	0.354	0.675	0.114
WMD + closure	0.204	0.860	0.572	0.675	1.000	0.541	0.698	0.559

not typically found in popular document retrieval task. First, the system needs to retrieve implicit evidence that is not explicitly mentioned in the query. This case happens when the statement involving comparison. For example, the statement "*Diarrhoea is the most common adverse event in the primary trial and the secondary trial*" does not only require the sentences containing keyword "Diarrhoea", but also other sentences containing keyword related to other adverse events. Second, the system should deal with negative evidence. For example, the statement "*the primary trial does not report the PFS or objective response rate of its patient cohort*" with entailment label have to retrieve all sentences from CTR to guarantee there is no evidence to support negation of this statement. So, when there is no evidence exists to support the statement, the system must retrieve all sentences as the evidence.

4 Conclusion and Future Work

In this paper, we described our system submitted to SemEval-2023 Task 7: multi-evidence natural language inference for clinical trial data (NLI4CT). We tested document retrieval and sentence similarity method for subtask 2: evidence retrieval. Despite of having low recall, our system improve the precision over the baseline. BM25 retriever and WMD for sentence similarity cannot handle medical vocabulary and reason with numerical information using term matching only. We find that these approaches particularly struggle to retrieve implicit or negative evidence.

Future directions of this work include incorporating deep architectures using language model, e.g., ColBERT (Khattab and Zaharia, 2020) and Dense Passage Retrieval (Karpukhin et al., 2020), for this task. On the other hand, we are interested in exploring whether injecting linguistics knowledge (such as negation and coordinating conjunction) and reasoning capabilities (such as numerical understanding and dealing with comparison and superlative) may help model to retrieve more accurate and complete evidence. We are also willing to examine the connection between two subtasks in this shared task by answering the question whether natural language inference task necessarily need complete evidence in context of clinical trial data.

Acknowledgments

The authors would like to acknowledge Country. This research has been carried out on the unceded lands of the Woi Wurrung and Boon Wurrung language groups of the eastern Kulin nation. We pay our respects to their Ancestors and Elders, past, present, and emerging. We respectfully acknowledge First Nations' connection to land, waters, sky, and community. This research was funded by the Australian Research Council through an Industrial Transformation Training Center Grant (grant IC170100030).

References

Al K Akobeng. 2005. Understanding randomised controlled trials. *Archives of disease in childhood*, 90(8):840–844.

Hilda Bastian, Paul Glasziou, and Iain Chalmers. 2010. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Medicine*, 7(9):e1000326.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Dorian Brown. 2020. Rank-BM25: A Collection of BM25 Algorithms in Python.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. 2020. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132, Online. Association for Computational Linguistics.

Yanjun Gao, Ting-Hao Huang, and Rebecca J. Passonneau. 2021. ABCD: A graph framework to convert complex sentences to a covering set of simple sentences. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3919–3931, Online. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Donal Landers, and André Freitas. 2023. SemEval-2023 task 7: Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 957–966. JMLR.org.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST).

Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine*, 3(1):17.

Avijit Thawani, Jay Pujara, Filip Ilievski, and Pedro Szekely. 2021. Representing numbers in NLP: a survey and a vision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–656, Online. Association for Computational Linguistics.

Thinh Hung Truong, Yulia Otmakhova, Rahmad Mahendra, Timothy Baldwin, Jey Han Lau, Trevor Cohn, Lawrence Cavedon, Damiano Spina, and Karin Verspoor. 2022. ITTC @ TREC 2021 Clinical Trials Track. In *The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings*.