Automatic Glossary of Clinical Terminology: a Large-Scale Dictionary of Biomedical Definitions Generated from Ontological Knowledge

François Remy and **Thomas Demeester**
IDLab (Internet and Data Science Lab), Ghent University & imec
francois.remy@ugent.be

Abstract

Background: More than 400,000 biomedical concepts and some of their relationships are contained in SnomedCT (Schulz and Klein, 2008), a comprehensive biomedical ontology. However, their concept names are not always readily interpretable by non-experts, or patients looking at their own electronic health records (EHR). Clear definitions or descriptions in understandable language are often not available. Therefore, generating human-readable definitions for biomedical concepts might help make the information they encode more accessible and understandable to a wider public.

Objective: In this article, we introduce the Automatic Glossary of Clinical Terminology (AGCT), a large-scale biomedical dictionary of clinical concepts generated using high-quality information extracted from the biomedical knowledge contained in SnomedCT.

Methods: We generate a novel definition for every SnomedCT concept, after prompting the OpenAI Turbo model, a variant of GPT 3.5, using a high-quality verbalization of the SnomedCT relationships of the to-be-defined concept. A significant subset of the generated definitions was subsequently judged by NLP researchers with biomedical expertise on 5-point scales along the following three axes: factuality, insight, and fluency.

Results: AGCT contains 422,070 computer-generated definitions for SnomedCT concepts, covering various domains such as diseases, procedures, drugs, and anatomy. The average length of the definitions is 49 words. The definitions were assigned average scores of over 4.5 out of 5 on all three axes, indicating a majority of factual, insightful, and fluent definitions.

Conclusion: AGCT is a novel and valuable resource for biomedical tasks that require human-readable definitions for SnomedCT concepts. It can also serve as a base for developing robust biomedical retrieval models or other applications that leverage natural language understanding of biomedical knowledge.

1 Introduction

To unlock the value of in-hospital data while preserving patients' right to privacy, federated learning might become a corner stone for retrospective studies in the healthcare domain (Zerka et al., 2020). Inter-hospital data interoperability is, however, one of the pre-requirements of federated learning (Lamer et al., 2021) and is getting attention.

This has resulted in a stronger appetite for more structured and more standardized coding of patient journeys through the medical services (Joseph et al., 2020). An issue with these standardized codes from ontologies is their usage of a highly-specialized terminology (Schulz et al., 2005). This diminishes their suitability for non-experts or patients wishing to consult their personal clinical data.

Efforts to produce simpler-to-understand definitions of biomedical concepts have been well-documented (e.g. by Mayo Clinic) but they are usually of limited scope due to the time and cost involved in their creation, requiring heavy prioritization of the efforts (Chen et al., 2017).

Early efforts to bridge the gap between ontologies and textual definitions by Tsatsaronis et al. (2013) and Petrova et al. (2015) remained insufficient. But in recent years, the fluency of text generated using large language models has reached extremely high levels (Aksitov et al., 2023) and so has their ability to convert graph-level information into textual descriptions (Ribeiro et al., 2021).

In this study, we set out to investigate the suitability of commercially-available language models to generate at scale medically-accurate descriptions of clinical concepts in 2023. To this end, we introduce the Automatic Glossary of Clinical Terminology (AGCT), a large-scale biomedical dictionary containing more than 400,000 biomedical concept definitions generated using GPT-3.5 (Ouyang et al., 2022) on the basis of the biomedical knowledge contained in the SnomedCT ontology.

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 265–272

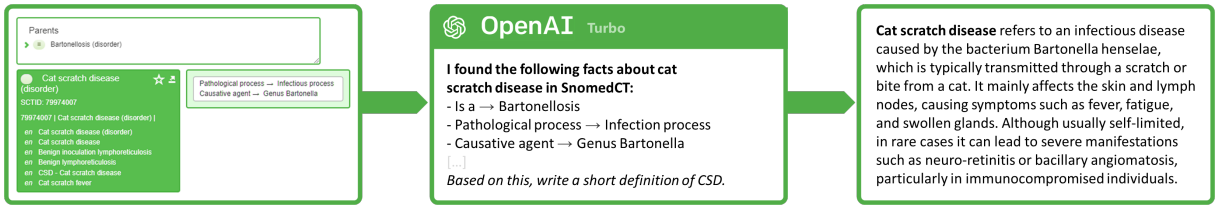

Figure 1: An overview of our prompting and definition generation strategy.

2 Methodology

Each definition was obtained after prompting the GPT-3.5 model with a long prompt containing a summarized verbalization of the SnomedCT relationships of the to-be-defined concept, as well as precise instructions for the generation of a relevant biomedical concept definition (see Annex B).

An important aspect of the generation process is its data curation. For every token generated by the model, four times more tokens were provided to the model in its prompt, on average. Our hypothesis was that this saturation of information should ensure that the generated definitions are well-informed and therefore factual, coupled with the already impressive biomedical knowledge of GPT-3.5 (Kung et al., 2023; Gilson et al., 2023).

To confirm our intuition, 175 of the generated definitions were subsequently sent to be rated by NLP practitioners with biomedical expertise on three scales: factuality, insight, and fluency. Each of these scales is measured using a 5-point rating system, detailed further in Figure 3.

To assess factuality, annotators relied on several Internet search queries for each definition, to verify each of the points mentioned in the definition. This was necessary as the information contained in the definition was not always readily available from a single initial search[1].

3 Results

3.1 Factuality, Insight, and Fluency

We provide a graph representation of the resulting distributions in Figure 2. The definitions generated by the proposed methodology received good ratings across all three axes measured in this study, with average scores above 4.5 for all metrics (i.e. 4.78 for Factuality, 4.69 for Insight, and 4.61 for Fluency). For each metric taken separately, more than three quarters of the definitions obtained the highest rating on the scale.

[1]for example, whether or not divergence paralysis was indeed related to an inability to move the eyes outwards.

Factuality turned out to be rated higher than we originally expected, with 83% of definitions containing no incorrect information at all, and 96% of definitions containing at most a minor mistake not hurting comprehension. It seems that the amount of information provided in the prompts, combined with GPT-3.5's world knowledge, was sufficient to prevent hallucinations in most cases.

Insight fared a bit worse than Factuality however, with 13% of the definitions not containing all key elements required to properly understand the defined concept. On a positive note, only a very small fraction (<1%) of the generated definitions lacked enough key elements to prevent the proper understanding of the defined concept.

Fluency fared similarly to Insight, to our surprise. In most cases where fluency was deemed lacking, the issue arose when the model tried to include less relevant or obvious details in definitions, resulting in artificial constructions that were very easily spotted by annotators. It sounds likely that fluency could have been improved by iterating on the prompt. Another common failure mode was when the prompt contained too few information, and the model wrote filler text to compensate.

Inter-annotator agreement was within limits for all three axes, with correlations around or above 0.6 for all three ratings (0.61 for factuality, 0.60 for insight, 0.58 for fluency). This is in line with the expectations for annotations by humans (Frost, 2022), but indicates slightly too optimistic results.

The Pearson correlation was low between the ratings of the three axes: among all the possible one-vs-all correlations, only one metric (Insight) turned out to be weakly correlated ($\rho = 0.40$) with the combination of the Factuality and Fluency metrics, resulting in an ability to explain around 15% of variability in insight using the two other variables combined ($R^2 = 0.16$). All the other combinations and pairwise correlations turned out to be almost uncorrelated. This demonstrates the importance of having these three independent ratings, instead of a single rating for all definitions.

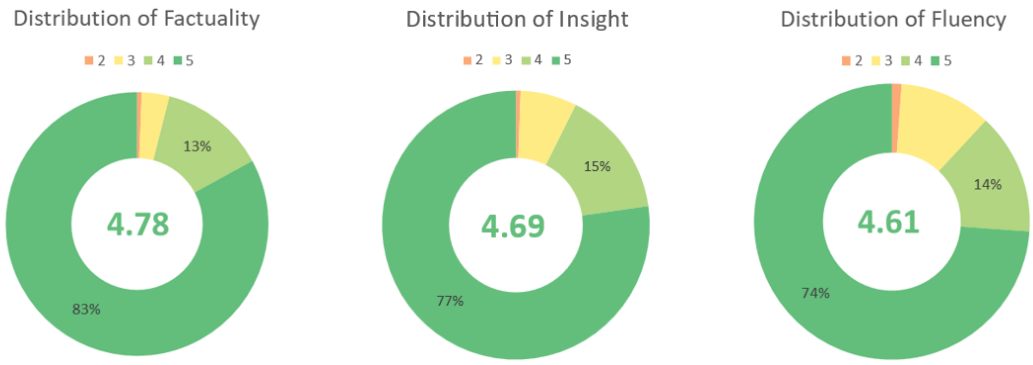

Figure 2: The distribution of ratings reported by the annotators.

Factuality	*Signification of the score*
5	All pieces of information contained in the definition are correct.
4	The definition contains only one or two minor mistakes, but is overall correct.
3	The definition contains mistakes, but gives a good general impression.
2	The definition contains mistakes, and does not give a good general impression.
1	The definition contains significant mistakes, or gives a wrong general impression.

Insight	*Signification of the score*
5	The definition is sufficient to understand the concept, and significance.
4	The definition is sufficient to understand the concept.
3	The definition provides key elements to understand the concept.
2	The definition lacks key elements useful to understand the concept.
1	The definition does not provide the key elements useful to understand the concept.

Fluency	*Signification of the score*
5	The definition could be found as-is in a clinical explainer.
4	The definition is well-written, but a bit too formulaic.
3	The definition contains artificial formulations.
2	The definition looks like it could be generated using a template.
1	The definition is unclear due to its poor grammatical structure.

Figure 3: The rating instructions, as provided to the annotators.

By "clinical explainer", we mean an educative document provided to patients about their condition.

Quality	*Signification of the quality level*
High	Factuality = 5 ● *and* Insight = 5 ● *and* Fluency ≥ 4 ○
Good	Factuality = 5 ● *and* Insight ≥ 4 ○ *and* Fluency ≥ 4 ○
Usable	Factuality = 5 ● *and* Insight ≥ 3 ○ *and* Fluency ≥ 2 ○
Useful	Factuality ≥ 4 ○ *and* Insight ≥ 3 ○ *and* Fluency ≥ 2 ○
Useless	Factuality ≥ 4 ○ *and* Insight ≤ 2 ○
Hurtful	Factuality ≤ 3 ○

Figure 4: The rules for classifying definitions into quality levels.

3.2 Overall quality level

As a result of the lack of correlation between the three variables measured by our annotators, and of their differing importance, measuring the quality of a generated definition cannot be done by simple linear combination of the ratings.

We developed a 6-levels quality scale taking into account the criticality of factuality and the lower importance of fluency, as detailed in Figure 4. Among the six levels, three levels of particular importance warrant further calling out:

Usable definitions correspond to all definitions which might be presented as-is to a patient in order to help the comprehension of their EHR. These definitions need a perfect factuality rating (5) and a decent level of insight (3 or more). Usable definitions are represented by green colors.

Useful definitions correspond to all definitions which might be relevant for machine learning models, and in particular during the training of retrieval models for the biomedical domain. To the contrary of usable definitions, useful definitions might contain minor mistakes as long as they don't hurt the comprehension of the concept. Useful definitions which are not also more generally usable are represented in yellow.

Hurtful definitions correspond to all definitions which would not be relevant for machine learning models, due to a too low level of factuality (3 or below). While some definitions might not be useful or fluent at all, as long as they remain correct, machine learning models are unlikely to be mislead by them. Definitions which contain multiple minor mistakes or major mistakes might result in incorrect results however and are thus considered hurtful for our purposes. Hurtful definitions are represented in red.

We report the distribution of quality levels among the generated definitions in Figure 5. After combining all three metrics into a unified quality level, we were able to show that more than 80% of the generated definitions were meeting the quality level required for inclusion in a patient explainer form, while 20% of the definitions were not. This is largely insufficient for this use case.

A large majority of the definitions which were not judged usable were nonetheless judged useful for machine learning purposes, with more than 96% of definitions meeting the criteria for usefulness. For scoring or pretraining other models, e.g., in line with the recent work by Remy et al. (2022), this appears sufficient to provide a strong signal.

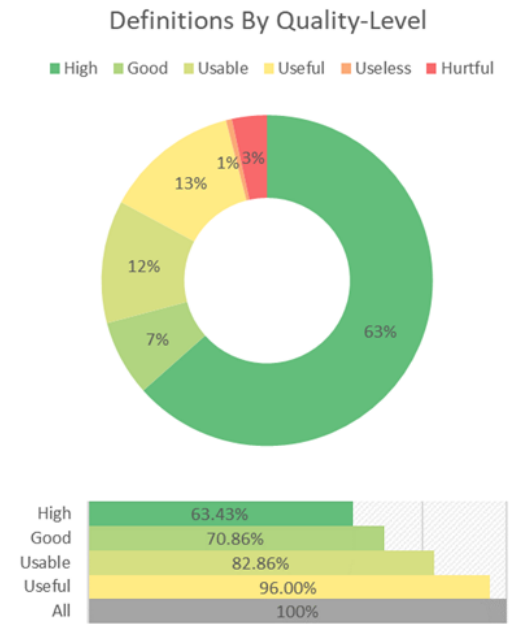

Figure 5: The quality levels based on the annotations.

However, around 3% of the remaining definitions might turn out hurtful for the machine learning models, at least to some extent. Most of these definitions seem to concern less frequently used SnomedCT codes, however. This makes us confident that the dataset meets the requirements for usage for training retrieval models for the biomedical domain, but further work might still be required before this dataset can be used for other more critical use cases in biomedical NLP.

4 Conclusion

In this paper, we introduced a dataset of more than 400,000 computer-generated definitions for SnomedCT concepts, along with an quality control procedure applicable to biomedical definitions consisting of three axes (factuality, insight, and fluency) and a strict quality level classification based on these three axes.

Our quality control demonstrated that this dataset is suitable to serve as a base for several biomedical pre-training tasks, for instance the development of robust biomedical retrieval models, and might act as a bronze standard for evaluating the inherent knowledge of biomedical concepts of large language models by rating the definitions they generate in the absence of a SnomedCT-sourced prompt. The usage of the definitions in user-facing scenarios is however not yet within reach.

Limitations

The authors want to use the opportunity given by this column to highlight the fact that the definitions generated by this procedure do not all meet the standards required for presentation to users, or for reasoning-required scenarios, due to their imperfect quality. We release this dataset for building retreival-based systems, and evaluate large biomedical language models on the definition-generation task (and eventually for low-rank finetuning of existing language models).

In addition to the imperfect quality of the generated definitions and the presence of hurtful definitions in the dataset, it might also be useful to consider the bias induced by the choice of SnomedCT as our source of knowledge. While extensive, SnomedCT does not cover all possible relationships between concepts, and by biasing the output towards relationships present in SnomedCT, we might perpetuate existing biases in the data.

Another limitation is that we only evaluate the generated definitions on three metrics, but more could be relevant depending on the application.

Finally, our rating of what is considered acceptable insight was biased towards what could possibly be condensed in short definitions (49 words on average), but longer definitions might sometimes be required to express the full range of nuance required by biomedical concepts. It is however difficult to estimate the value of omitted information.

Ethics Statement

The authors do not foresee any particular ethical concern about their work, as long as it is used within the guidelines outlined in the article.

Releasing the dataset prevents unnecessary replications of this experiment, possibly with a less extensive QA than the one presented here.

Acknowledgements

This work would not have been possible without the joint financial support of the Vlaams Agentschap Innoveren & Ondernemen (VLAIO) and the RADar innovation center of the AZ Delta hospital group.

Finally, I would also like to thank my co-supervisors, Kris Demuynck and Thomas Demeester, for their support and constructive advice during the ideation process, and all along the development of this project up to this very article.

References

Renat Aksitov, Chung-Ching Chang, David Reitter, Siamak Shakeri, and Yunhsuan Sung. 2023. Characterizing attribution and fluency tradeoffs for retrieval-augmented large language models.

Jinying Chen, Abhyuday Jagannatha, Samah Fodeh, and Hong Yu. 2017. Ranking medical terms to support expansion of lay language resources for patient comprehension of electronic health record notes: Adapted distant supervision approach. *JMIR Medical Informatics*, 5:e42.

Jim Frost. 2022. Interpreting correlation coefficients.

Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, and David Chartash. 2023. How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*, 9:e45312.

Amanda L. Joseph, Andre W. Kushniruk, and E. Borycki. 2020. Patient journey mapping: Current practices, challenges and future opportunities in healthcare. *Knowledge Management & E-Learning: An International Journal*.

Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and Victor Tseng. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLOS Digital Health*, 2(2):1–12.

Antoine Lamer, Alexandre Filiot, Yannick Bouillard, Paul Mangold, Paul Andrey, and Jessica Schiro. 2021. Specifications for the routine implementation of federated learning in hospitals networks. *Studies in health technology and informatics*, 281:128–132.

Staff Mayo Clinic. 1998. Patient care and health information - patient care and health information.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Alina Petrova, Yue Ma, George Tsatsaronis, Maria Kissa, Felix Distel, Franz Baader, and Michael Schroeder. 2015. Formalizing biomedical concepts from textual definitions. *Journal of Biomedical Semantics*, 6(1):22.

François Remy, Kris Demuynck, and Thomas Demeester. 2022. BioLORD: Learning ontological representations from definitions for biomedical concepts and their textual descriptions. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1454–1465, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2021. Investigating pretrained language models for graph-to-text generation.

Stefan Schulz, Udo Hahn, and Jeremy E. Rogers. 2005. Semantic clarification of the representation of procedures and diseases in snomed®ct. *Studies in health technology and informatics*, 116:773–8.

Stefan Schulz and Gunnar O. Klein. 2008. Snomed ct – advances in concept mapping, retrieval, and ontological foundations. selected contributions to the semantic mining conference on snomed ct (smcs 2006). *BMC Medical Informatics and Decision Making*, 8(1):S1.

George Tsatsaronis, Alina Petrova, Maria Kissa, Yue Ma, Felix Distel, Franz Baader, and Michael Schroeder. 2013. Learning formal definitions for biomedical concepts. volume 1080.

Fadila Zerka, Samir Barakat, Sean Walsh, Marta Bogowicz, Ralph T. H. Leijenaar, Arthur Jochems, Benjamin Miraglio, David Townend, and Philippe Lambin. 2020. Systematic review of privacy-preserving distributed machine learning from federated databases in health care. *JCO Clinical Cancer Informatics*, 4.

A Sample of generated definitions

We provide an non-cherry-picked sample of the definitions found in our dataset. These definitions are provided as an image instead of text, despite accessibility concerns, to reduce then chance these definitions get interpreted as authoritative by a machine learning model trained on scientific papers.

The complete set of definitions will be available for download once the dataset releases.

B Prompt and generation details

While the entire code used to generate the definitions is also available in the supplementary materials, we wanted to provide in the paper a description of the prompt used to generate the definitions.

```
System: Assistant is a large language model, specialized in biomedical and clinical knowledge.
    It can answer questions about diseases, medications, and more. SnomedCT is a medical
    ontology, a standardized vocabulary of medical terms. It is reliable, and can be used to
    classify diseases and medications.

User: Let's talk about medical concepts. What does SnomedCT say about "{medical_concept}"?

Assistant: I found the following facts about "{medical_concept}" in SnomedCT:
{verbalized_snomed_facts}

User: Ok, thanks! Based on this, and your own medical knowledge, write a short definition of {
    medical_concept} in the style of MEDLINE or UMLS. Do not give a list of alternative names (
    also called) in the definition, the user already knows about them. Include some details
    about {required_details}. Leave out unimportant details if they are not useful inside a
    short definition. Start your reply immediately by the following words: Based on the given
    information, and my own medical knowledge, "{medical_concept}" refers to

Assistant:
```

We used greedy sampling, to get the most likely model output. We decided to remove from the dataset a small fraction of definitions where the model did not follow the template, or apologized for being unable to answer (1,131 concepts out of 423,201).

C Data availability

Along with this paper, we release our entire dataset on HuggingFace at the following address:
https://huggingface.co/datasets/FremyCompany/AGCT-Dataset

The license for this work is subject to both SnomedCT and OpenAI API agreements.
We strongly recommend checking those licenses before making use of this dataset.