

LLMを用いた生物医学文献アブストラクトからの細胞摂動影響の自動抽出

Automatic Extraction of Cell Perturbation Effects from Biomedical Abstracts Using LLMs

Masashige Hasumi^{1,2}, Ryosuke Matsuzawa^{1,2}, Haruto Ijiri^{1,2}, Haruka Ozaki^{1,2} (¹RIKEN, ²University of Tsukuba) s2521438@u.tsukuba.ac.jp



Introduction

Background

- Knowledge of how perturbations alter the abundance of specific cell types is important for understanding cell-type functions [1].
- However, knowledge of **the relationships between perturbations and changes in cell numbers** is scattered across a vast biomedical literature. In addition, the wide variety of terms and expressions used or perturbations and cell types makes rule-based information extraction difficult [2].

Objective

This study aimed to develop a large language model (LLM)-based system to **automatically extract information on perturbations and cell count changes from abstracts** and construct a database.

Methods

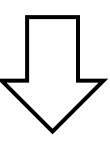
Target abstract corpus

Immunology-related abstracts concerning human, mouse, and rat



Manual annotation of gold standard

177 abstracts



Whether changes in cell numbers are described?

↓ No

↓ Yes

Label: 0

Label: 1

(49 abstracts)

↓ (128 abstracts)

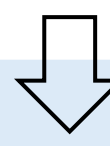
Annotated information

- Perturbation
- Cell_type
- Effect
- Effect_category
- Individual_level
- Anatomical_site_level
- Experimental_condition

LLM system

gpt-4o-2024-08-06 (OpenAI)

Abstracts



Relevance assessment module

Whether changes in cell numbers are described?

↓ No

↓ Yes

Label: 0

Label: 1

Information extraction module

Extracted information

- Perturbation
- Cell_type
- Effect
- Effect_category
- Individual_level
- Anatomical_site_level
- Experimental_condition

Performance evaluation

Relevance assessment module

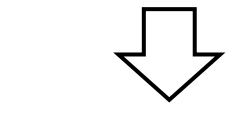
Precision · Recall · F1 score

Information extraction module

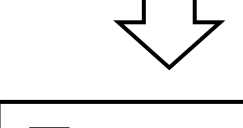
Accuracy

Application of the LLM system to text corpus

Corpus (171,758 abstracts)



LLM system



Extracted information



Normalization

- Ontology mapping
- Semantic embedding-based clustering



Visualization of extracted information characteristics

Ontology mapping by MetaSRA [3]

Target

Perturbation, Cell_type, Individual_level, Anatomical_site_level

Used ontology

Cell_type: CL · CVCL Others: EFO · DOID · UBERON

Semantic embedding by SapBERT [4] and clustering

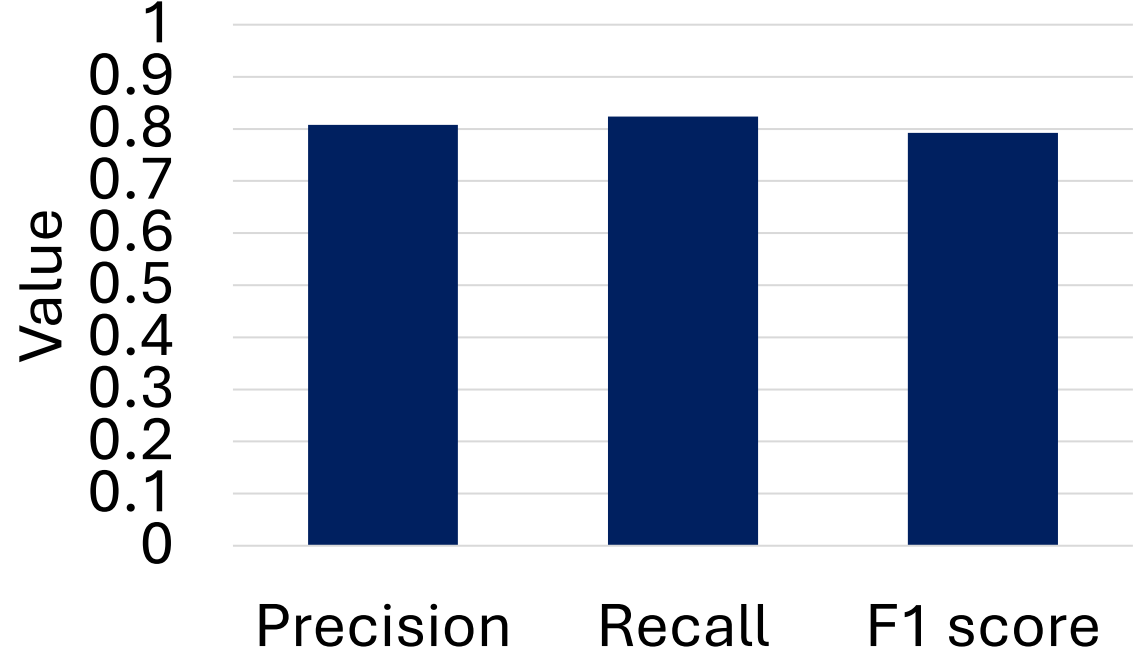
Agglomerative hierarchical clustering using average linkage and cosine distance

This work was supported by the JST NBDC Integrated Database Promotion Program (JPMJND2402).

Results

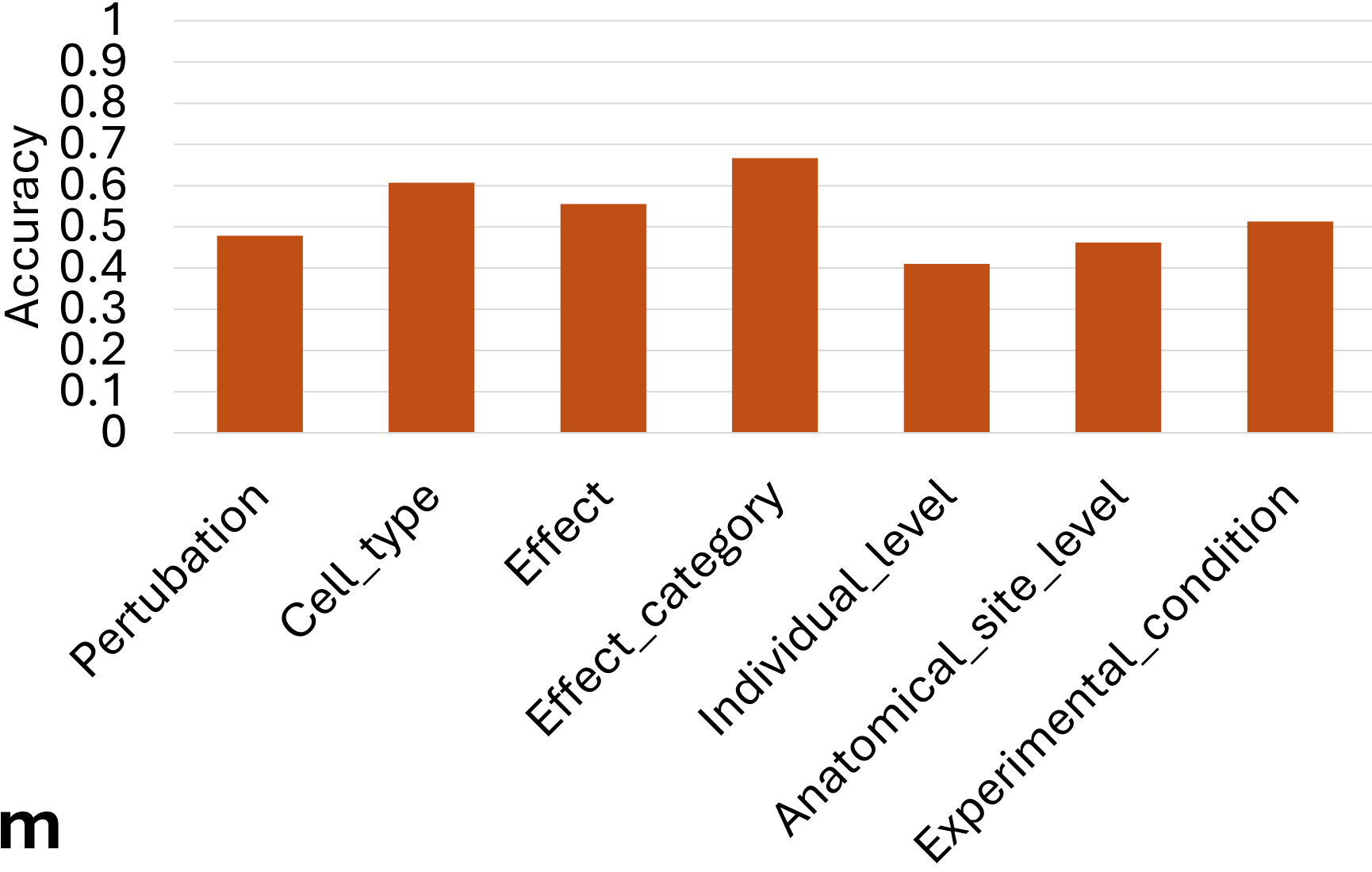
Performance evaluation of each module

Relevance assessment module



Information extraction is feasible.

Information extraction module



Application of the LLM system to the corpus

Number of relevant records

total: 171,758 abstracts

Non-relevant	Relevant
120,691 (70.3%)	51,067 (29.7%)

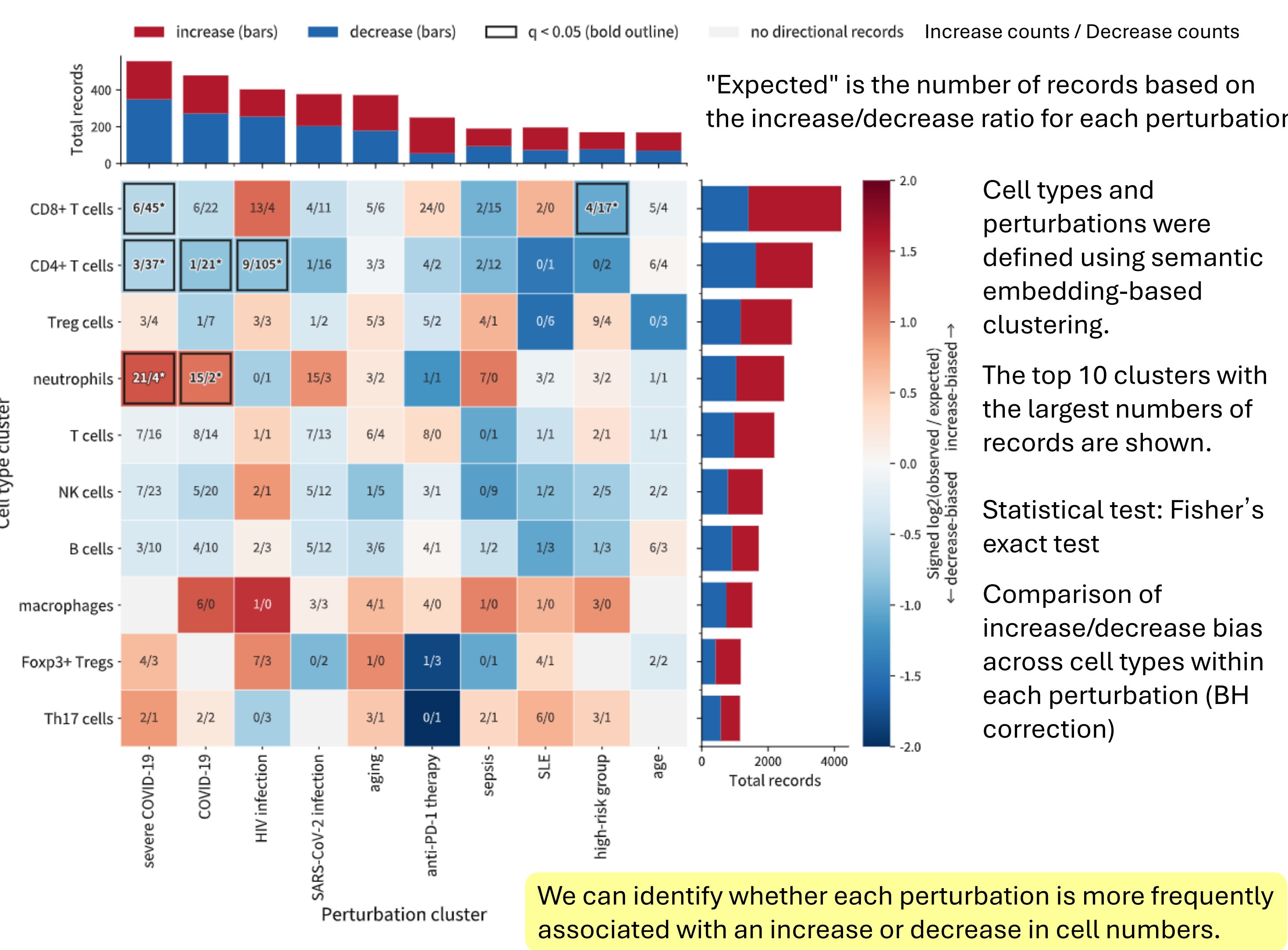
Number of extracted records

total: 80,962 records

	Fill number	Fill rate
Perturbation	80,951	100.0%
Cell_type	80,946	100.0%
Effect	80,961	100.0%
Effect_category	80,866	99.9%
Individual_level	65,827	81.3%
Anatomical_site_level	50,205	62.0%
Experimental_condition	62,229	76.9%

All four required items were correctly extracted.

Enriched cell number changes in the abstracts



Discussion

- We developed and evaluated a system that automatically extracts information on perturbations and cell count changes from research abstracts.
- Applying the system to 171,758 abstracts successfully generated the data required for database construction.

Future work

- We will construct a database recording the extracted information, the ontology mapping results and semantic embedding-based clustering results.
- To make the database useful and user-friendly, we will investigate methods for graph-based visualization.

References

- [1] Roohani et al. Nat Biotechnol. 2023 [2] Zhang et al. ACL-IJCNLP. 2021
[3] MN Bernstein et al. Bioinformatics. 2017 [4] Liu et al. NAACL. 2021