

# DiscHPO@BC8 Track 3: Recognising and Normalising Continuous and Discontinuous Genetic Phenotypes Using T5 Variants and Sentence-Transformers Models

Areej Alhassan<sup>1,2\*</sup>, Viktor Schlegel<sup>1,3</sup>, Monira Aloud<sup>2</sup>, Riza Batista-Navarro<sup>1</sup>, Goran Nenadic<sup>1</sup>

<sup>1</sup>University of Manchester, United Kingdom, <sup>2</sup>King Saud University, Saudi Arabia,

<sup>3</sup>ASUS Intelligent Cloud Services, Singapore

\*Corresponding author: E-mail: aralhassan@ksu.edu.sa

## Abstract

This paper describes our participation in Track 3 of the BioCreative VIII shared task focused on extracting and normalising genetic phenotypes from dysmorphology physical examination reports. We focus on disjoint entity spans which make up around 14% of the mentions. We developed an approach, DiscHPO, that extracts and normalises both continuous and discontinuous spans. The system consists of two components: a sequence-to-sequence named entity recognition model and an entity normaliser based on a Sentence-Transformer and a Cross-Encoder re-ranker. The best performing model for entity normalisation obtained an F1 score of 0.7229 on the test data, whilst the best model for span extraction achieved an F1 score of 0.6647.

## Introduction

The eighth edition of BioCreative challenges was organised in four tracks. In this paper, we present our participation in the third track, which is focused on extracting and normalising genetic phenotypes in clinical reports (1). In this task, participants were provided with patient physical examination entries that may include both, neither or either of normal findings (e.g., intact palate) or key findings (e.g., wide nasal bridge). Participants have to perform two tasks:

- Extract key findings and neglect normal findings.
- Map the key findings to their Human Phenotype Ontology (HPO) identifier (2).

One of the main challenges in this task is the presence of discontinuous spans in text (also referred to as “disjoint entities” or “disjoint mentions”). We noted that around 14% of the mentions in the training and validation sets are disjoint. The organisers emphasised the importance of solving this problem and going beyond the sequence labelling approaches which are effective in identifying only entity spans with consecutive tokens and fail to extract discontinuous mentions (1). To address this task, we developed DiscHPO, which is a pipeline with two components: [1] detecting continuous and discontinuous named entity spans, [2] normalising the extracted spans to their associated HPO identifiers.

## Dataset

The dataset contains de-identified short reports of clinical observations noted during dysmorphology physical examinations. The corpus contains 3,136 observations, split into a training set (1716 observations), a validation set (454), and a test set (966 + 2427 decoy observations). The annotations include the entity spans' character offsets and their associated HPO ontology identifier. We analysed the prevalence of disjoint mentions and found that they constitute

around 14% of all entity spans. There were several forms and levels of complexity in these mentions. For instance, either one entity is interrupted by one or more non-entity tokens, or two or more entities share a common head, and one entity interrupts the other.

We frame the NER problem as a sequence-to-sequence problem (3), whereby a system will take as input a sequence of tokens and generate a target sequence containing the recognised spans which can be any of the following entity types: Normal Finding, Key Finding, or Not Applicable. As opposed to sequence labelling, which requires models to learn a tagging scheme that assigns tags to each token, sequence-to-sequence models learn to generate a new sequence of tokens which is a transformation of the input sequence. Thus, in our case, the only preprocessing step that was required prior to model training is the conversion of the numerical offsets of the gold standard target spans into the actual word spans. Each span is then prefixed with the entity type. Lastly, all entity spans belonging to one observation were combined into a single sequence separated by a semi-colon. An example of a preprocessed target is "KEYF: Sparse eyebrow; NORMF: Normal lashes", where KEYF and NORMF mean Key Finding and Normal Finding. After the prediction of the spans, a post-processing step is performed to convert the spans back to numerical offsets.

## Methodology

DiscHPO has two components: a sequence-to-sequence NER model that is able to recognise both continuous and discontinuous mentions at the sentence level, and an entity normaliser underpinned by a sentence transformer bidirectional encoder (bi-encoder) for candidate generation and a cross-encoder re-ranker. The DiscHPO pipeline is illustrated in Figure 1.

### NER model

We employed a sentence-level NER based on fine-tuning a pre-trained sequence-to-sequence encoder-decoder language model. Specifically, we utilised several variants of the Text-to-Text Transfer Transformer (T5) (4). Due to T5's generative nature, it can conveniently be fine-tuned to extract spans and predict their target class simultaneously.

**Model fine-tuning** - We examined the following T5 architectures: The original T5 (4), FLAN-T5 (5) and SCIFIVE (6). The models were trained using a single NVIDIA A100 GPU with 40 GB memory, over 20 epochs, learning rate of  $3e-4$  and a batch size of 8.

**Low-Rank Adaptation (LoRA)** - A parameter-efficient fine-tuning (PEFT) approach such as LoRa eliminates the efficiency and cost challenges associated with large language models (LLMs) by using only a much smaller number (around 1.1%) of trainable parameters

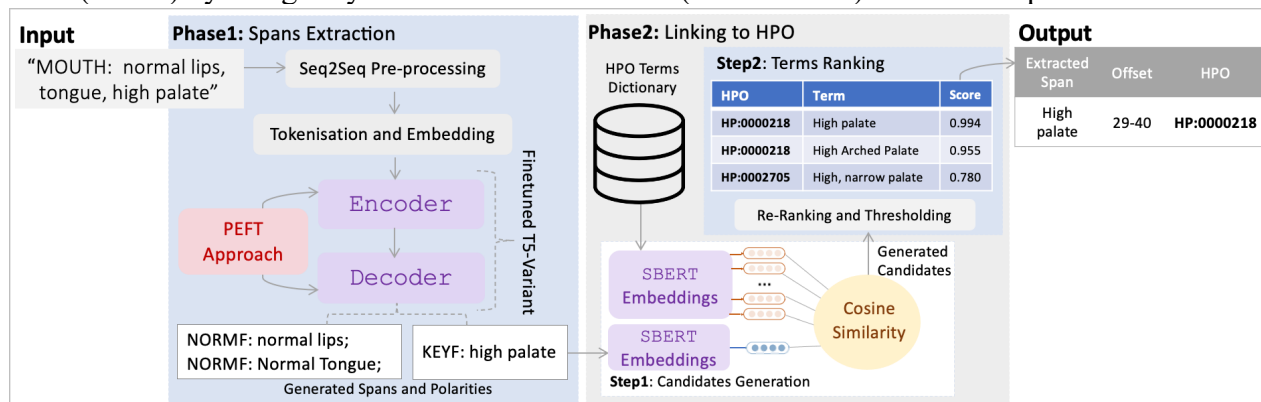


Figure 1: DiscHPO pipeline architecture

(7, 8). Trainable parameters are set using two hyperparameters of LoRa: rank ( $r$ ) and alpha ( $\alpha$ ). Using LoRa, we trained Flan-T5-XL at  $r$  of 16 and  $\alpha$  of 512 and 1024.

### Normalisation to HPO

The normalisation should exclude normal findings since the HPO does not record those findings. A version of the HPO dictionary has been provided by the task organisers, which contains around 17K terms. Another list containing 5K unobservable terms was also provided, which has been helpful in narrowing down the original list.

**Candidate generation** - We utilised sentence transformer bi-encoders (9). Specifically, we fine-tuned the `all-roberta-large-v1` model<sup>1</sup> to create embeddings for both HPO terms and the extracted entity spans. For fine-tuning, we used pairs of spans from the training set and their corresponding HPO terms to train on a semantic similarity objective. Then, semantic matching based on cosine similarity is performed to identify relevant candidates by comparing the embeddings representing the spans with those representing the HPO terms.

**Re-ranking** - The top 30 candidates obtained by semantic matching are then passed on to a sentence transformer cross-encoder model, specifically `ms-marco-electra-base`<sup>2</sup>, which reranks the generated candidates by calculating the scores for each span-candidate term combination. Its results are sorted to produce the final reordered HPO terms. We consider only the top-scoring match as the final output.

## Results and Discussion

Table 1 presents DiscHPO validation results: the extracted findings and their normalised HPO IDs. To evaluate the former, metrics were reported based on exact matching and partial matching (which considered an extracted span to be correct as long as it overlaps with the gold standard span). Our evaluation was conducted using the official script provided by the task organisers.

Model		Normalisation			Exact Extraction and Normalisation			Partial Extraction and Normalisation		
		P	R	F1	P	R	F1	P	R	F1
Scifive-Base		0.722	0.726	0.724	0.695	0.637	0.665	0.719	0.719	0.719
T5-Large		0.725	0.743	0.734	0.696	0.644	0.669	0.724	0.739	0.732
FlanT5 Large		0.708	0.731	0.719	0.678	0.637	0.658	0.708	0.729	0.718
FlanT5 XL-LoRa	512	0.735	0.739	0.737	0.710	0.654	<b>0.681</b>	0.734	0.738	0.736
	1024	0.743	0.739	<b>0.742</b>	0.716	0.642	0.677	0.742	0.735	<b>0.738</b>

Table 1: DiscHPO results on the validation set

In the testing phase, we applied the same normalisation component in all runs while employing different NER settings:

- **Run1:** FlanT5-XL and LoRa,  $\alpha = 1024$  trained on the training set only.
- **Run2:** FlanT5-XL and LoRa,  $\alpha = 1024$  trained on both the training and validation sets.
- **Run3:** FlanT5-XL and LoRa with  $\alpha = 512$  trained on the training set only.

As shown in Table 2, the best normalisation score is achieved using FlanT5-XL with LoRa  $\alpha$  of 512. For span extraction, increasing  $\alpha$  to 1024 yielded the best results.

<sup>1</sup> <https://huggingface.co/sentence-transformers/all-roberta-large-v1>

<sup>2</sup> <https://huggingface.co/cross-encoder/ms-marco-electra-base>

Runs	Normalisation			Exact Extraction and Normalisation			Partial Extraction and Normalisation		
	P	R	F1	P	R	F1	P	R	F1
1	0.7205	0.7234	0.7219	0.6944	0.6375	<b>0.6647</b>	0.7198	0.7210	0.7204
2	0.7069	0.7305	0.7185	0.6735	0.6248	0.6482	0.7058	0.7265	0.7160
3	0.7179	0.7281	<b>0.7229</b>	0.6899	0.6367	0.6623	0.7172	0.7258	<b>0.7214</b>

Table 2: DiscHPO results on the test set

Overall, the highest gain was obtained by using FlanT5-XL with LoRa. It is apparent that the scores for partial extraction and normalisation are very close to the normalisation ones, which implies that normalisation performance is not substantially degraded by partially extracted spans. As shown in Table 3, our NER models can identify discontinuous spans. As can be seen, FlanT5-XL with LoRa  $\alpha$  512 is superior in detecting disjoint mentions of findings with an F1 score of 62.4%. We observed that in some cases, partial extraction might be sufficient for normalisation when one span is interrupted by non-entity tokens. For instance, in the example: "EYES: normal brows, mild hooding", the exact span should be "EYES: hooding", but extracting the partial span "Hooding" leads also to correct normalisation, since the word hooding is conventionally related to an eye condition. In contrast, when disjoint spans share a common head, each span must be resolved to its own head and its constituent tokens to avoid missing normalisations.

Model	Exact Extraction			Partial Extraction		
	P	R	F1	P	R	F1
<b>T5-Large</b>	0.5455	0.5753	0.5699	0.8725	0.93684	0.9035
<b>Scifive</b>	0.6	0.5783	0.5889	0.9108	0.8679	0.8888
<b>FlanT5 LoRa 512</b>	0.6975	0.5638	<b>0.6235</b>	0.79	0.8229	0.80612
<b>FlanT5 LoRa 1024</b>	0.6575	0.5783	0.6154	0.9270	0.8811	<b>0.9036</b>

Table 3: Results of discontinuous span extraction in the validation set

We also observed that the models easily identified simple coordinated ellipses, such as "NOSE: broad nasal bridge and tip". However, when it comes to sharing two heads, there are some complex cases where the models failed to resolve, such as: "Contractures of 3rd and 4th digits", where they should be resolved as "Contractures of 3rd digit" and "Contractures of 4th digit". The error can be attributed to the fact that the word "digit" appeared in the plural form, which might have confused the model in determining whether or not to resolve it as two separate entities.

## Conclusion

This paper provides an overview of our DiscHPO pipeline in the context of the BioCreative VIII - Track 3 challenge. Several variants of the T5 model were investigated for the extraction of the key findings. It was found that by optimising Flan-T5-XL's parameters using LoRa, better extraction results were obtained for both continuous and discontinuous mentions of findings, thereby improving normalisation performance. Moreover, normalisation results can be improved when identifying all disjoint mentions that share a common head.

## Acknowledgements

The corresponding author is supported by a PhD scholarship from King Saud University. This work has been partially supported by the project "Integrating hospital outpatient letters into the healthcare data space" (grant EP/V047949/1; funder: UKRI/EP SRC).

## References

1. Gonzalez-Hernandez, G., Campbell, I. M., Weissenbacher, D., and Zhao, X. (2023) Track 3: Genetic Phenotype Extraction and Normalization from Dysmorphology Physical Examination Entries. *BioCreative VIII*
2. Köhler, S., Gargano, M., Matentzoglou, N., et al. (2021) The human phenotype ontology in 2021. *Nucleic Acids Res.* **49**, D1207–D1217
3. Sutskever, I., Vinyals, O., and Le, Q. V (2014) Sequence to sequence learning with neural networks. *Adv Neural Inf Process Syst*
4. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020) *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*
5. Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., and others (2022) Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*
6. Phan, L. N., Anibal, J. T., Tran, H., Chanana, S., Bahadroglu, E., Peltekian, A., and Altan-Bonnet, G. (2021) Scifive: a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv:2106.03598*
7. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021) Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*
8. Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S., and Bossan, B. (2022) PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods
9. Reimers, N., and Gurevych, I. (2019) Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*