Biomedical named entity recognition using BioBERT models fine-tuned on the HunNer corpora

Ola Olde & Adam Barvesten

Aim

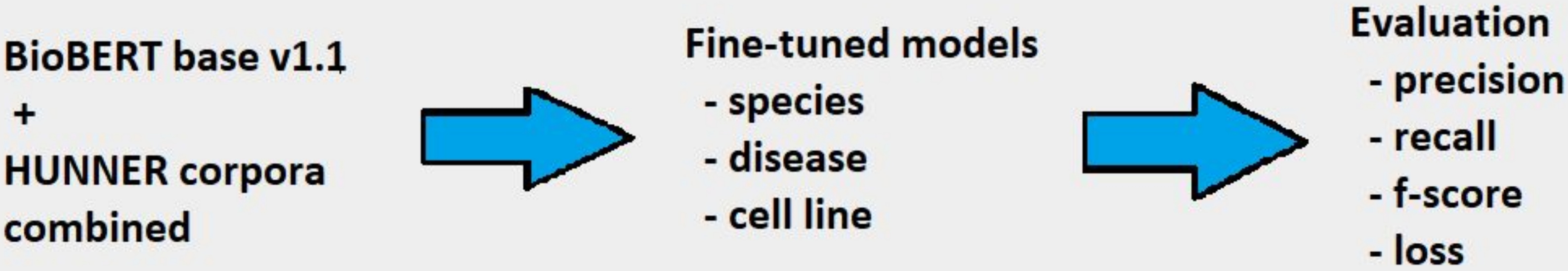
Overall goal

- Finding information in =>30 million biomedical articles => Named entity recognition (NER)

Project task

- Fine-tune BioBERT NER models using the HunNer corpora
    - Gene/protein, species, disease, cell line and chemical

Tools and methodology



BioBERT base v1.1
+
HUNNER corpora
combined

Fine-tuned models

- species
- disease
- cell line

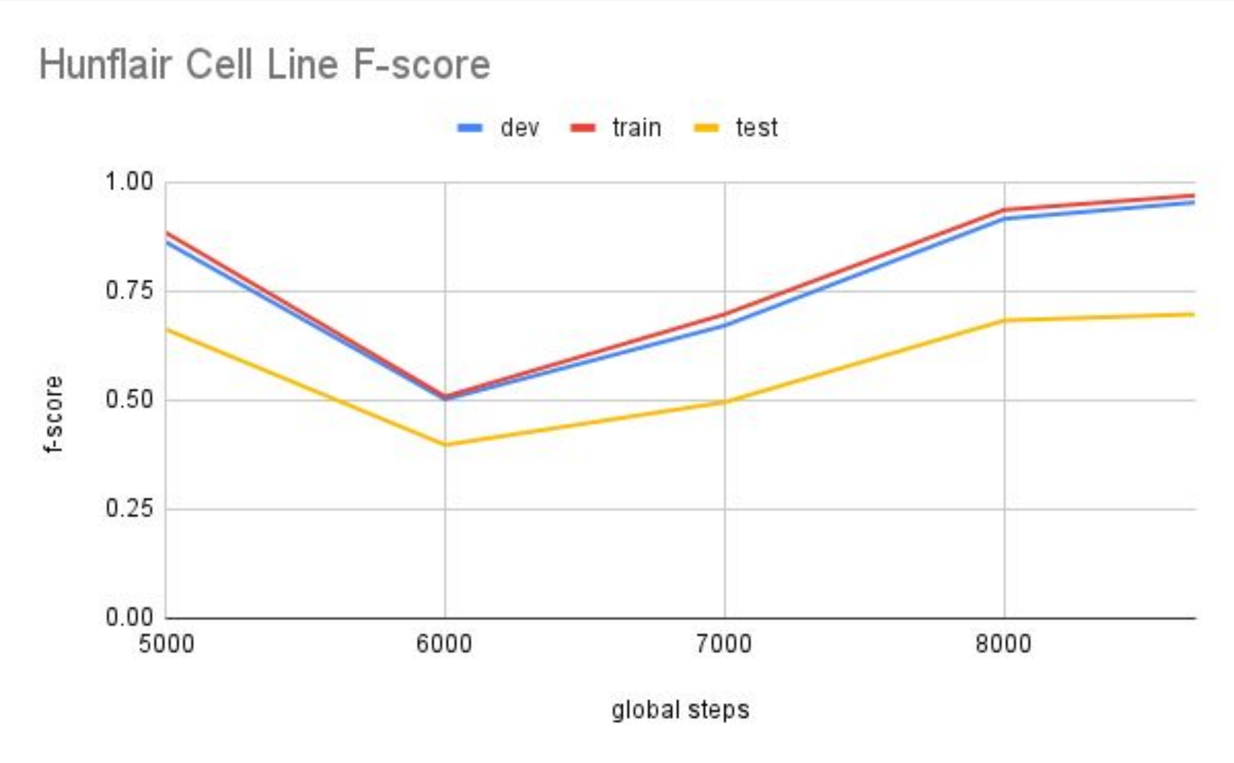
Evaluation

- precision
- recall
- f-score
- loss

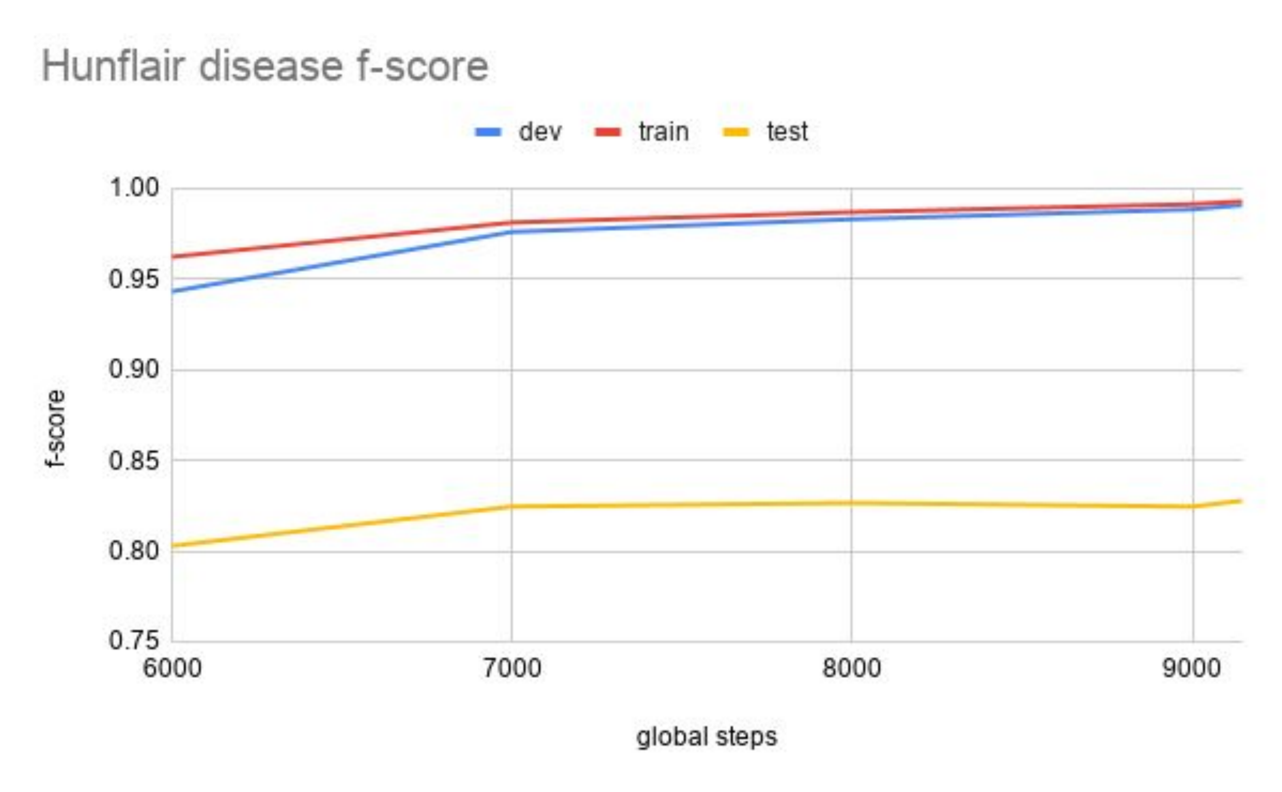
Evaluation metrics

- Precision
- Recall
- F-score
- Loss

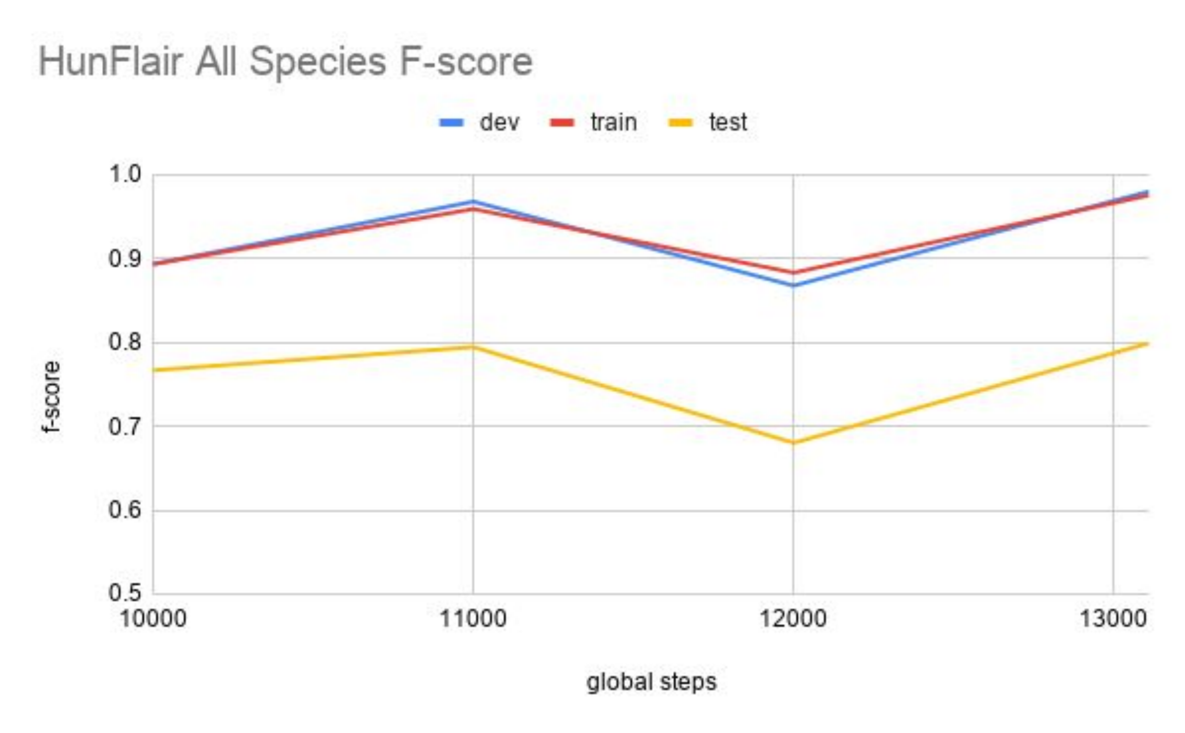
Results: Cell Line model



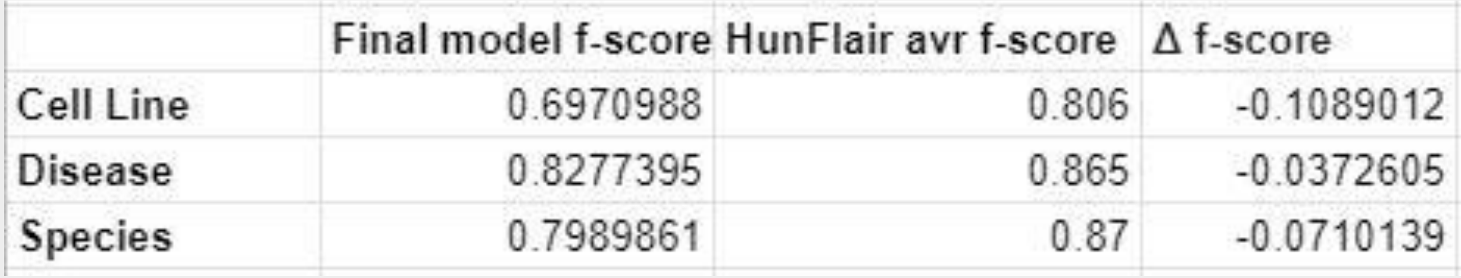
Results Disease model



Results Species model



Comparison of BioBERT models and HunFlair



	Final model f-score	HunFlair avr f-score	Δ f-score
Cell Line	0.6970988	0.806	-0.1089012
Disease	0.8277395	0.865	-0.0372605
Species	0.7989861	0.87	-0.0710139

Conclusion

- Successfully trained BioBERT models that recognize species, disease and cell lines
- BioBERT might be able to outperform HunFlair when trained on same corpora

Future steps

- Use BioBERT-Large v1.1 (+ PubMed 1M)
- Train on the chemical and gene corpora
- Train longer, more epochs
- Try different hyperparameters, learning rate etc.
- Compare directly against Hunflair
- Fit the model into the group’s text mining pipeline