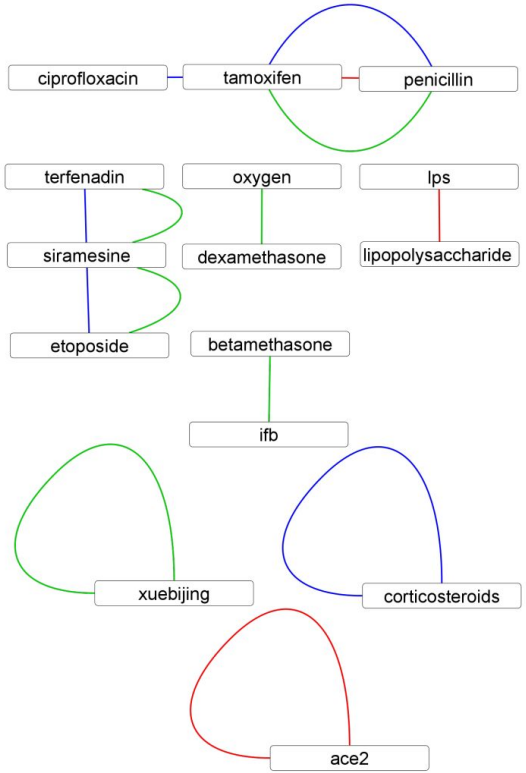
Biomedical Relation Extraction with Deep Neural Networks

Nils Broman

Background

- Studying DNA, protein and chemical interactions key to understand fundamentals of life
- **A lot of information** - PubMed alone has over 20 million articles (as of 2020), with over 1 million new published every year
- Overwhelming - Need simpler and more accessible



Natural Language Processing (NLP)

"the ability of a computer program to understand human language as it is spoken and written"

- Transformer architecture - **Encoder**/decoder using multi-headed attention
- Language model: BERT (Bidirectional Encoder Representations from Transformers)
- Transfer learning - Pretrain with unlabeled data, then finetune for specific task

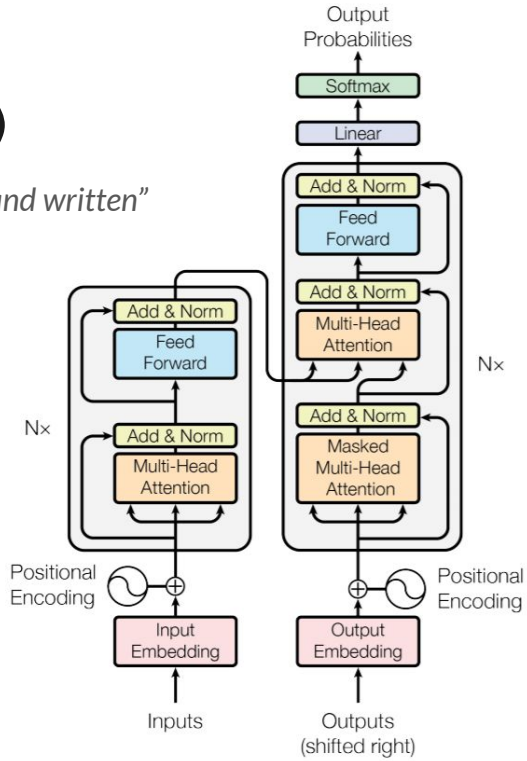
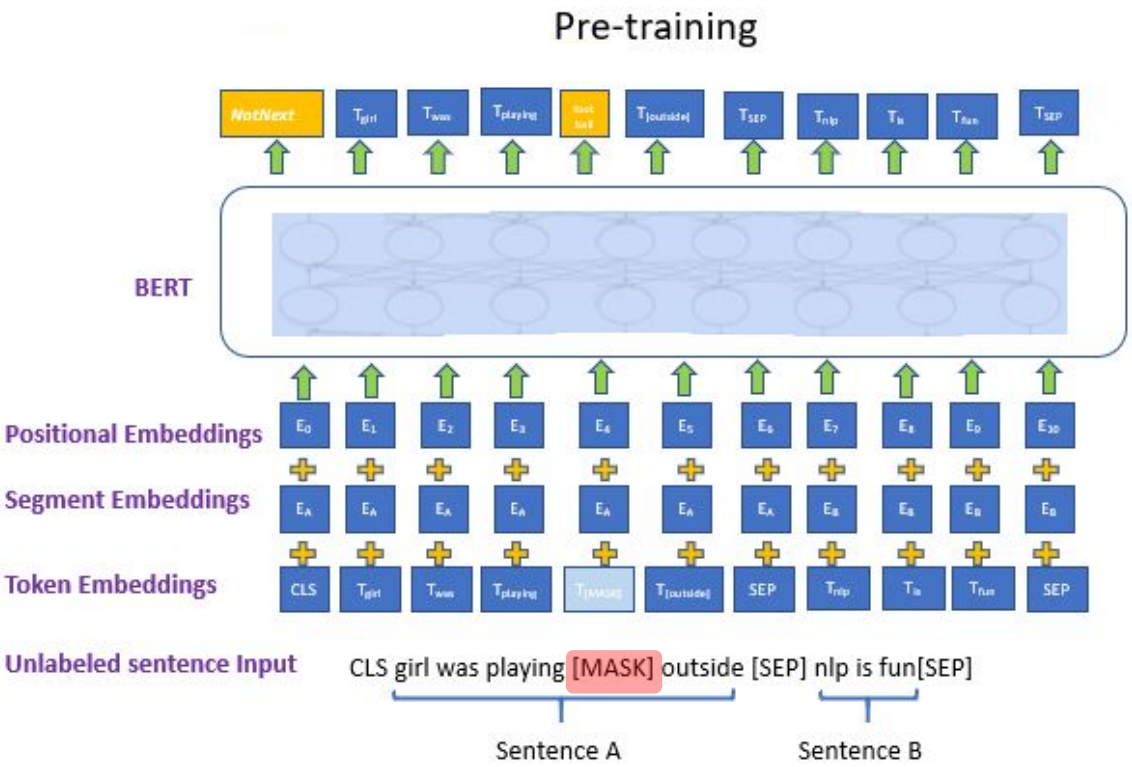


Figure 1: The Transformer - model architecture.

BERT

- Fully connected - Trains to predict masked words
- BERT Base is trained on wikipedia articles (2500M words) and books (800M words)
- SciBERT - Same architecture but trained on 1.14M scientific articles



ChemProt Corpora - example sentence

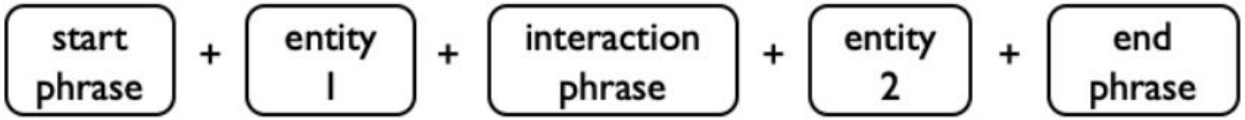
REGULATOR-POSITIVE

"The results showed that administration of << AlCl3 >> resulted in a significant elevation in the levels of [[AchE]] activity, CRP, NF-κB, and MCP-1 accompanied with a significant depletion in the Ach level."

ChemProt Corpora - Class Balance

Class	Train		Dev	
	Count	%	Count	%
INTERACTOR	2583	40.13	1350	37.96
NOT	241	3.74	175	4.92
PART-OF	308	4.79	153	4.30
REGULATOR-NEGATIVE	2505	38.92	1302	36.61
REGULATOR-POSITIVE	799	12.41	576	16.20
Total	6436	100	3556	100

Artificially constructed data



- Phrases from Cell Line Ontology and by supervisor (Sonja Aits) - Replace words with synonyms for more variety

- Entities (proteins) from Uniprot database

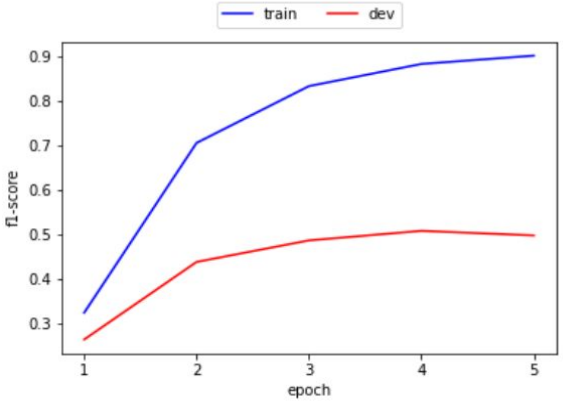
Metrics

$$Precision = \frac{\text{True Positives}}{\text{True Positives} + \text{True Negatives}}$$

$$Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

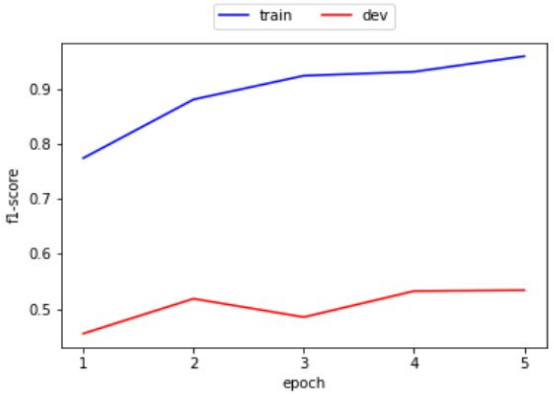
$$\boxed{F1\text{-}score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}}$$

Prior Results



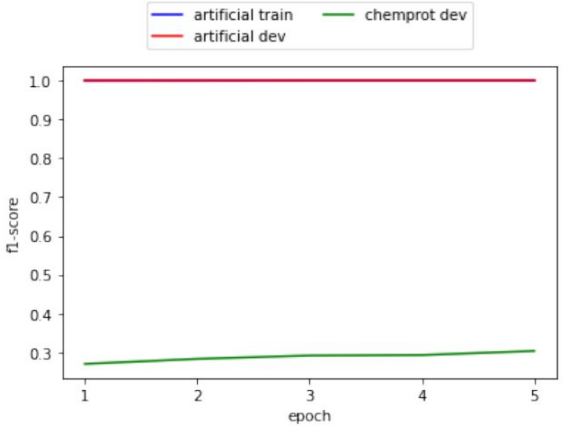
Baseline

Train: 0.88 Dev: 0.51



Oversampled

Train: 0.97 Dev: 0.65 (?)



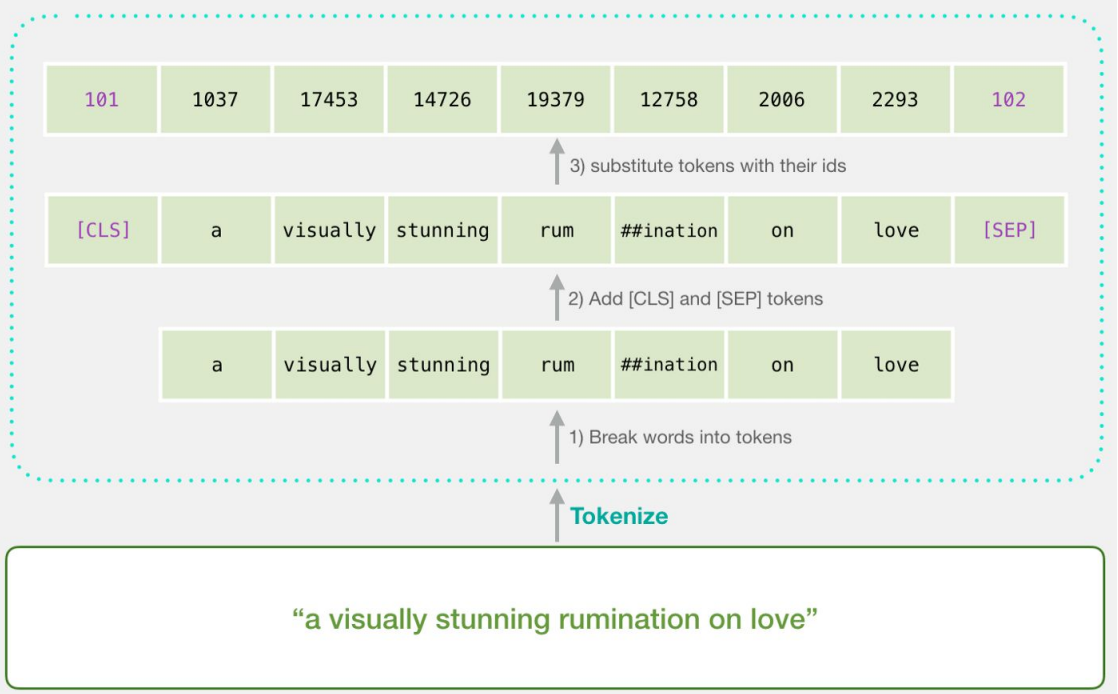
Artificial

Train: 1.00 Dev: 0.30

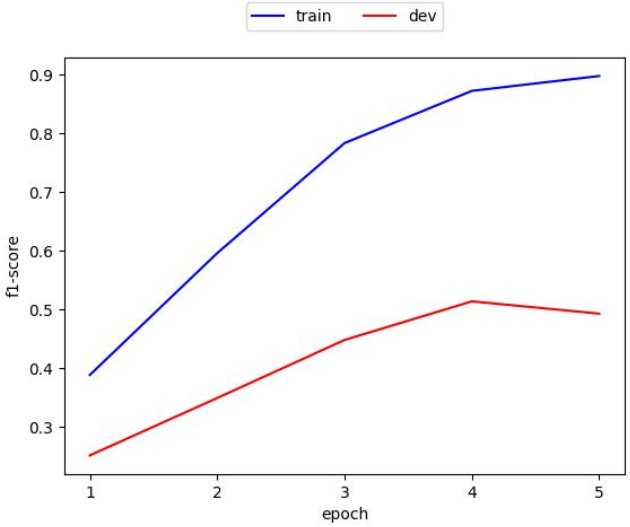
Tokenization

- Wordpiece embedding

- SciBERT trained on different data, hence has different tokenization

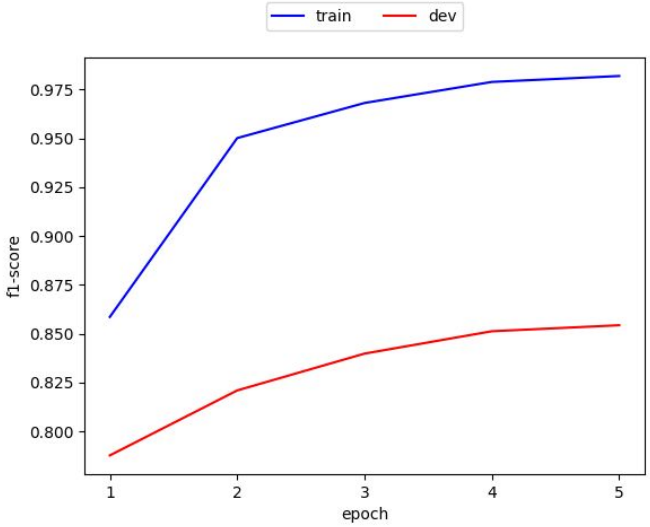


Changing the tokenizer



BERT Base tokenizer

Train: 0.88 Dev: 0.51



SciBERT Tokenizer

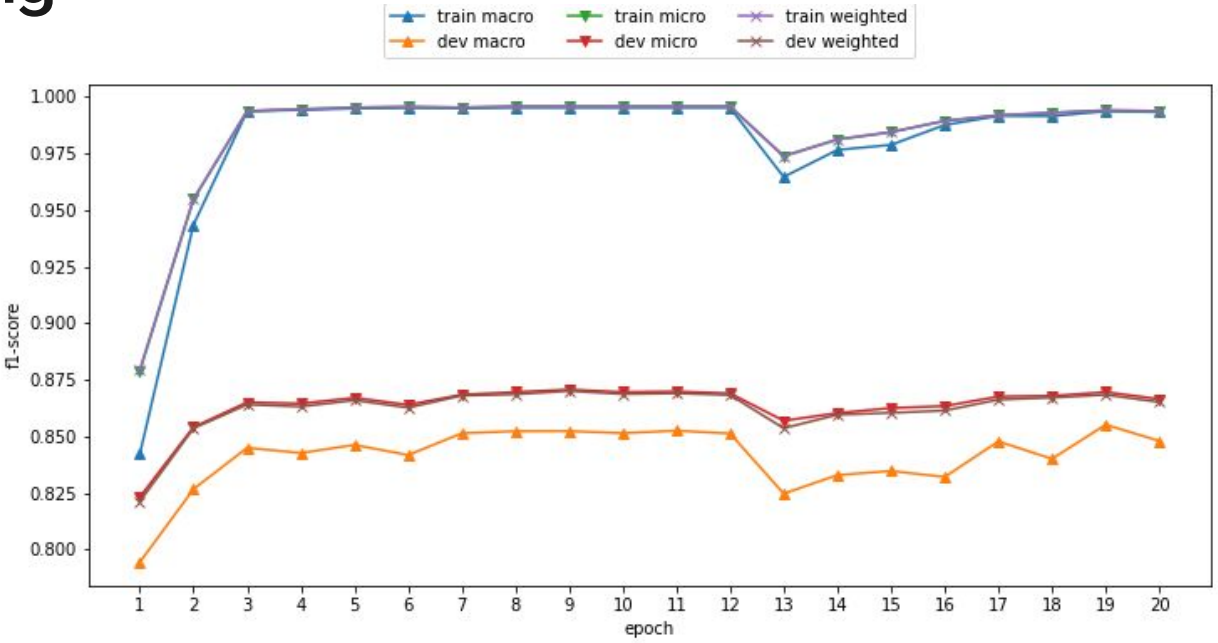
Train: 0.98 Dev: 0.84

Longer fine tuning

- Strange drop in performance after e 12

Top performance:

Epoch	19	11	9
Train	0.993	0.995	0.995
Dev	0.855	0.852	0.852

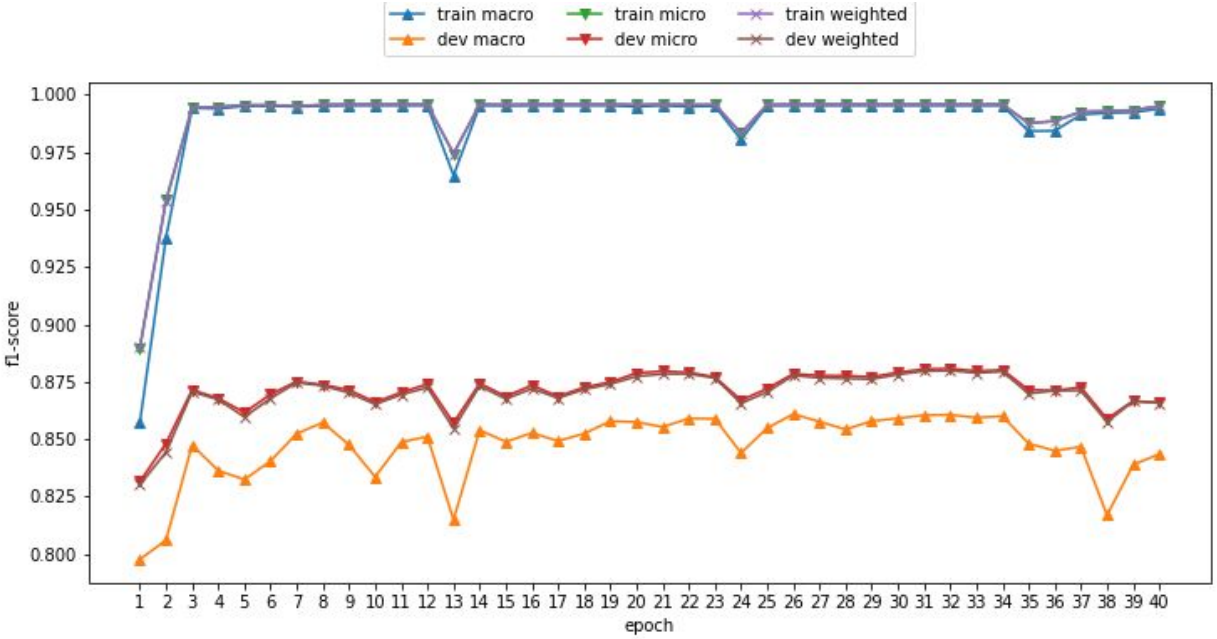


… even longer

- Again, drop at epoch 13, but now also at 24 and 35 (steps of 11?)

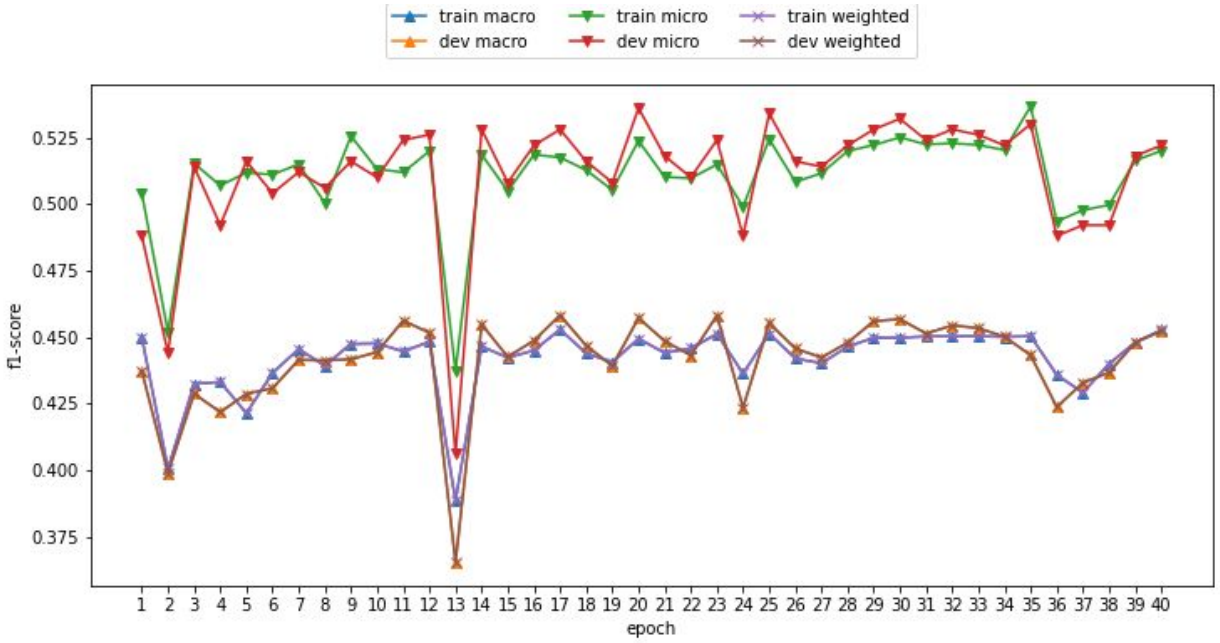
Top performance:

Epoch	26
Train	0.995
Dev	0.861



Performance on artificial data

- Noisy
- F1-score ~ 0.43
- Suggests that there are significant differences between the real and artificial data, not only that the artificial is lacking

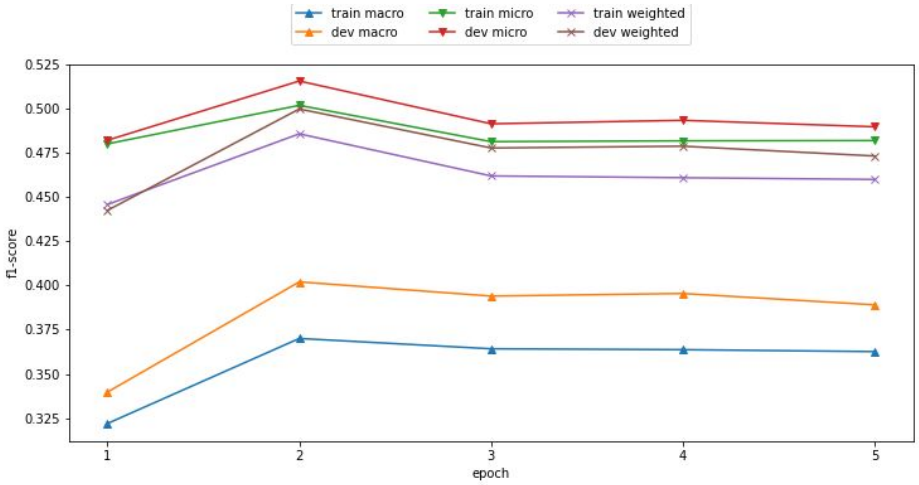


Artificial models

- Artificial data only proteins while ChemProt exclusively chemical and protein/DNA
- Use chemical names from the ChemProt training set
- Models trained on either scored perfect on artificial data and similar on ChemProt
- The one with chemicals scored higher when evaluated on the ChemProt train set, which makes sense due to using the same chemical names

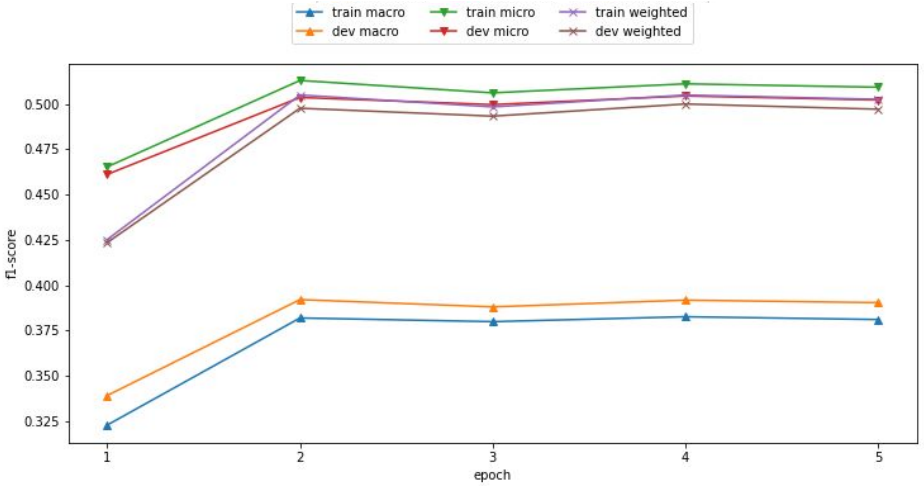
Artificial models - Protein/Protein

- Perfect scores on the artificial data, but poor on ChemProt
- Still large improvement compared to artificial with Base tokenizer



Artificial models - Chemical/Protein

- Better on train set - make sense since more of the same words
- Slightly worse on dev set though

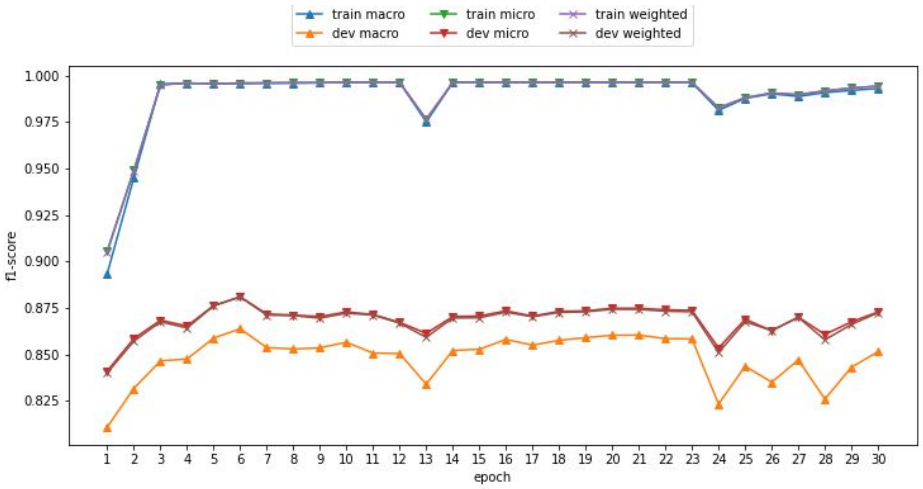


Mixed models - 10% Artificial

- Best results so far
- Trained many models, and averages similar to baseline

Top performance:

Epoch	6
Train	0.995
Dev	0.864

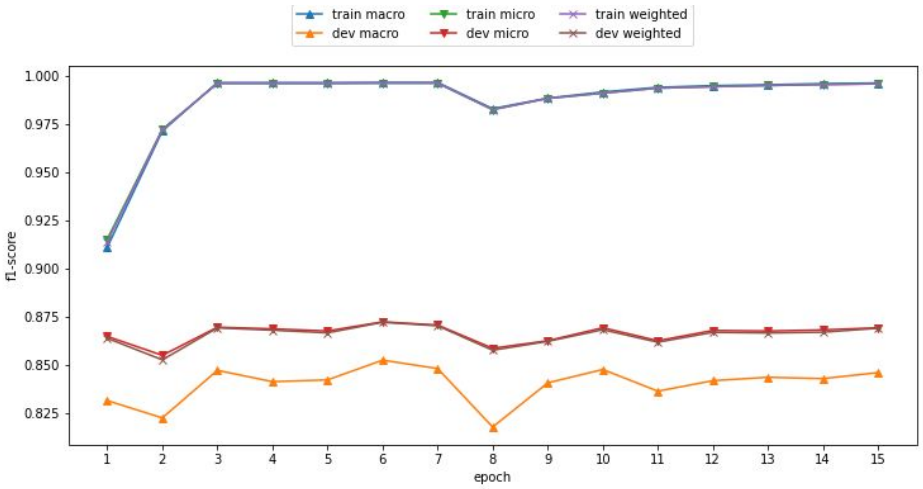


Mixed models - 25% Artificial

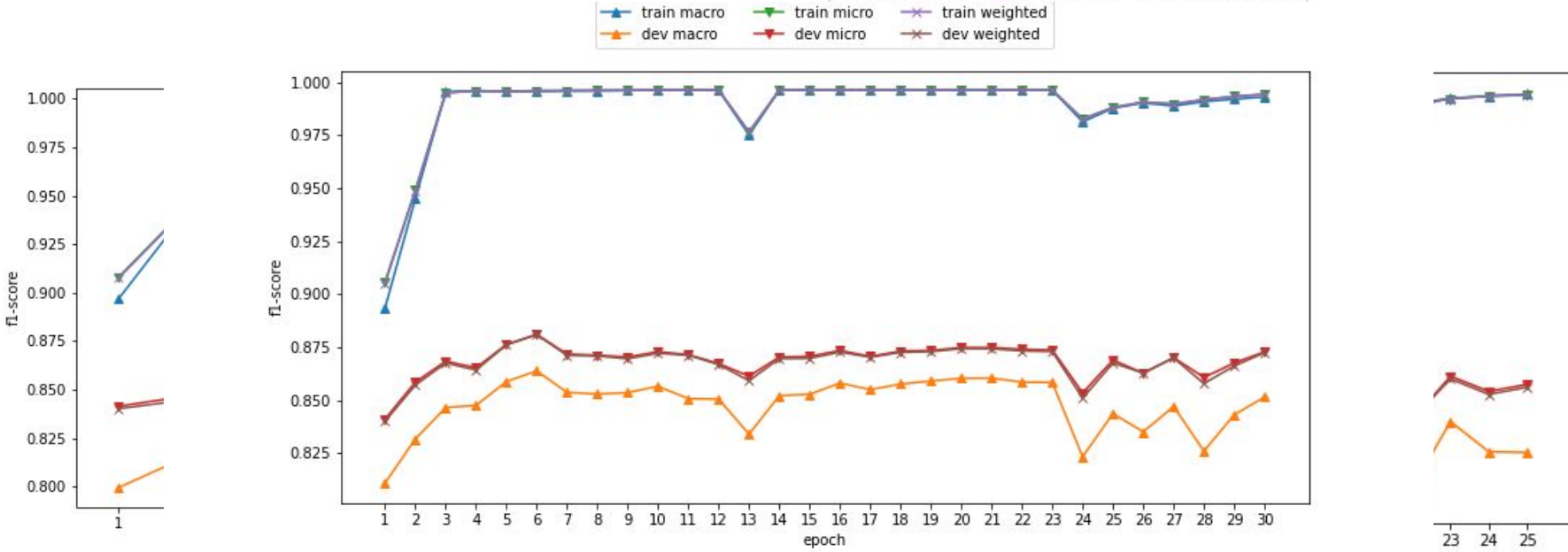
- Similar score to baseline

Top performance:

Epoch	6
Train	0.997
Dev	0.852



Drops depend on max epochs - optimizer?



Conclusions

- Very important to use correct tokenizer for BERT (or whenever using token embeddings)
- Large improvement compared to earlier models
- Subtle differences between baseline and mixed models
- Slight favour towards the mixed, though could be bias due to more models trained.

Earlier Models	**Baseline**	**Oversampled**	**Artificial**
Epochs	4	5	5
F1 Train (Macro)	0.88	0.97	-
F1 Dev (Macro)	0.51	0.65	0.30
New Models	**Baseline**	**Mixed 10**	**Mixed 25**
Epochs	26 (9)	6	6
F1 Train (Macro)	0.995 [+.11]	0.995 [+.02]	0.997 [+-02]
F1 Dev (Macro)	0.861 [+.21] (0.852 [+.20])	0.864 [+.21]	0.852 [+.20]

Limitations

- Single sentences - relations could be described over several
- Artificial sentences have little variation - single type of structure and only one author
- Not enough time to tweak hyperparameters

Future Development

- More diverse artificial building blocks
- Chemical names from some collection rather than just from train set (for more variation)
- Weighted support for the added artificial data
- Investigate what causes the sudden drops during longer training (optimizer?)
- Train a model using both train and dev set and do final evaluation on the test set