Experiments with Pre-Trained Deep Neural Language Models for Clinical NLP

Andriy Mulyar[1], Elliot Schumacher[2], Masoud Rouhizadeh[2], Chris Chute[2] and Mark Dredze[2]
Johns Hopkins University[2], Virginia Commonwealth University[1]

Concept Linking with a Neural Ranker

Our submission to the Concept Normalization task frames UMLS Concept Normalization as a ranking task, in which for a given query (mention) all possible matches (CUIs) are ranked.

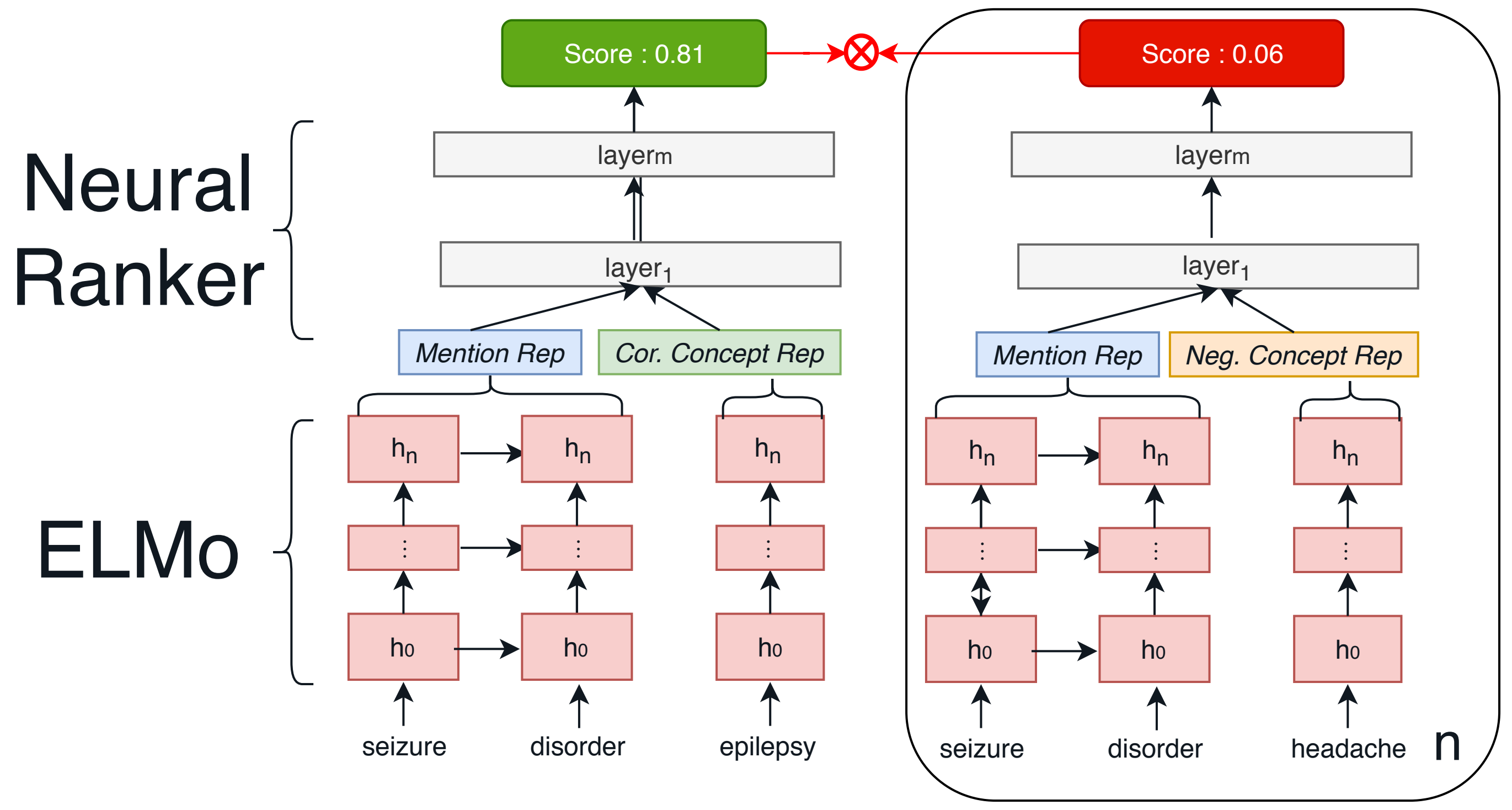


Fig. 1: Our deep learning concept linking system depicted with an n negative concept sampling architecture. During training, we sample $n = 10$ incorrect concepts for every correct concept.

Architecture and Training

- We paramaterize our ranker as a multi-layered, fully-connected neural network trained with a negative sampling scheme.
- The neural network takes as input a representation of the mention and a concept then outputs a compatibility score.
- Mentions and concepts are represented by ELMo embeddings.

Results

Our submission followed similar pre-processing steps to the dataset article. An ablated evaluation of these steps is showcased in Table 1.

System	Accuracy (%)
Rules	71.23
Rules → MetaMap	73.42
Rules → Deep Learning Linker	75.49
Rules → MetaMap → Deep Learning Linker	76.83

Tab. 1: Ablation Experiments on Challenge Submission Results

In our experiments, the deep learning linker achieves the best standalone performance. This means our ranker assigns the top compatibility score to the correct concept 60% of the time from a possible 430k concepts across RxNorm and SNOMED.

System Component	Accuracy (%)	Linked (%)
Training Set Look up	51.83	53.69
UMLS Look up	52.45	66.66
MetaMap	38.37	53.76
Deep Learning Linker	60.98	100.00

Tab. 2: Standalone Component Performance

Shared Task Insights

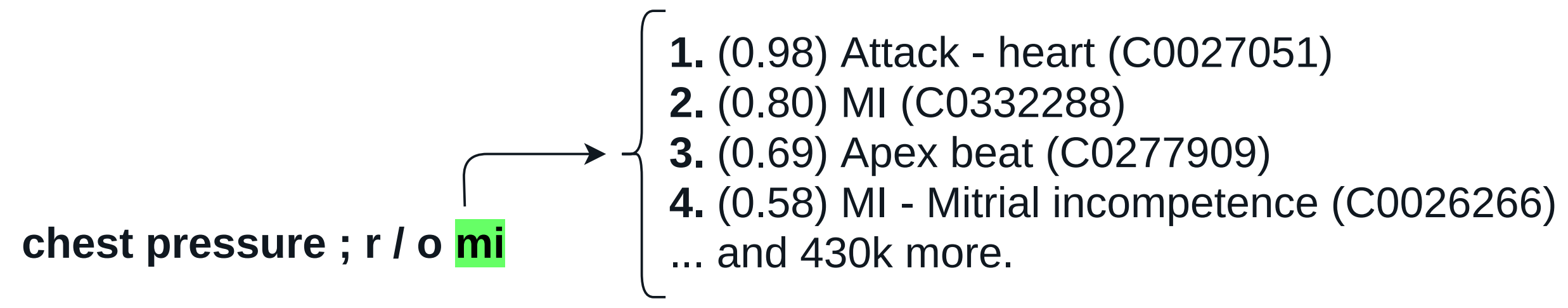
- Representations from Neural Language Models provide a high-performing baseline for tackling clinical NLP tasks.

Clinical STS

- Lexical similarity metrics (such as Jaro-Winkler) achieve high annotator correlation on MedSTS. This suggests that little semantic understanding may be required in an automated system for assessing similarity between sentences in the corpus.
- Bert-base (original Google model trained on non-clinical data) nearly matches the performance of ClinicalBERT when finetuned on MedSTS. This suggests that BERT captures and leverages domain agnostic information when scoring semantic similarity.

Concept Normalization

- The MCN normalization corpus is annotated at very fine granularity. This means that some system errors may not actually matter in practice, eg. *Exercise stress test* (C0430120) vs. *Exercise test* (C0015260).
- Framing normalization as a ranking task requires considering all 430k concepts when attempting to link a mention. It would be useful to prune the space of possible concepts to ease the computational burden.


chest pressure ; r / o mi

1. (0.98) Attack - heart (C0027051)
2. (0.80) MI (C0332288)
3. (0.69) Apex beat (C0277909)
4. (0.58) MI - Mitrial incompetence (C0026266)
... and 430k more.

Fig. 2: Linking the mention **'mi'**

Language Modeling

Recent advances in language modeling have produced systems capable of generating contextualized representations of text for use in downstream clinical NLP tasks such as concept normalization and semantic similarity.

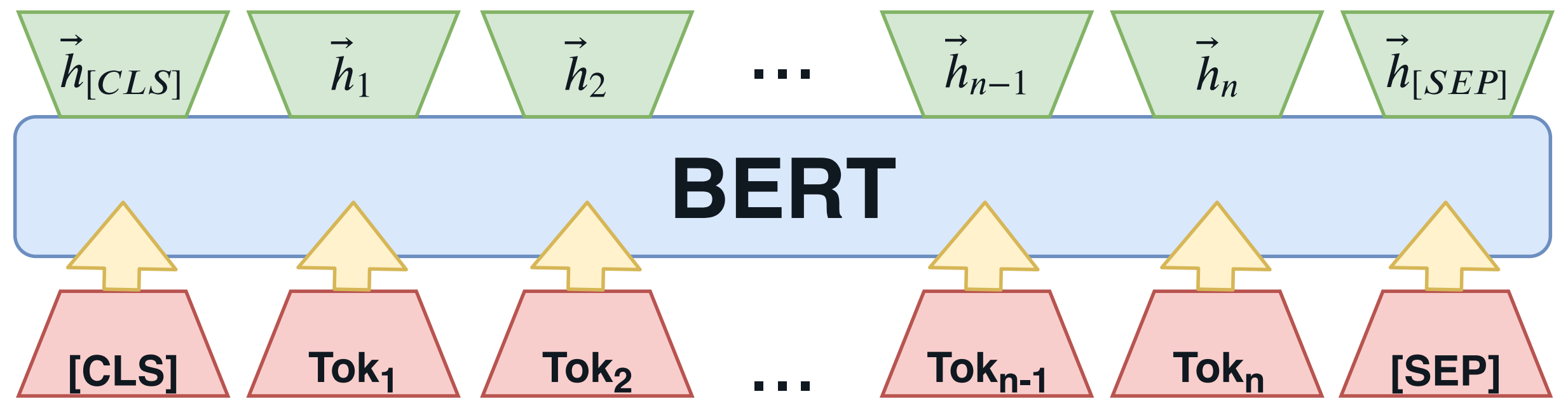


Fig. 3: BERT

Clinical Semantic Similarity with BERT

Our submission to the STS shared task consisted of a MedSTS fine-tuned, data augmented instance of ClinicalBERT. Under the hypothesis that semantic similarity indicative information from general English may transfer to clinical text, we explored augmenting the MedSTS training set with instances from web English.

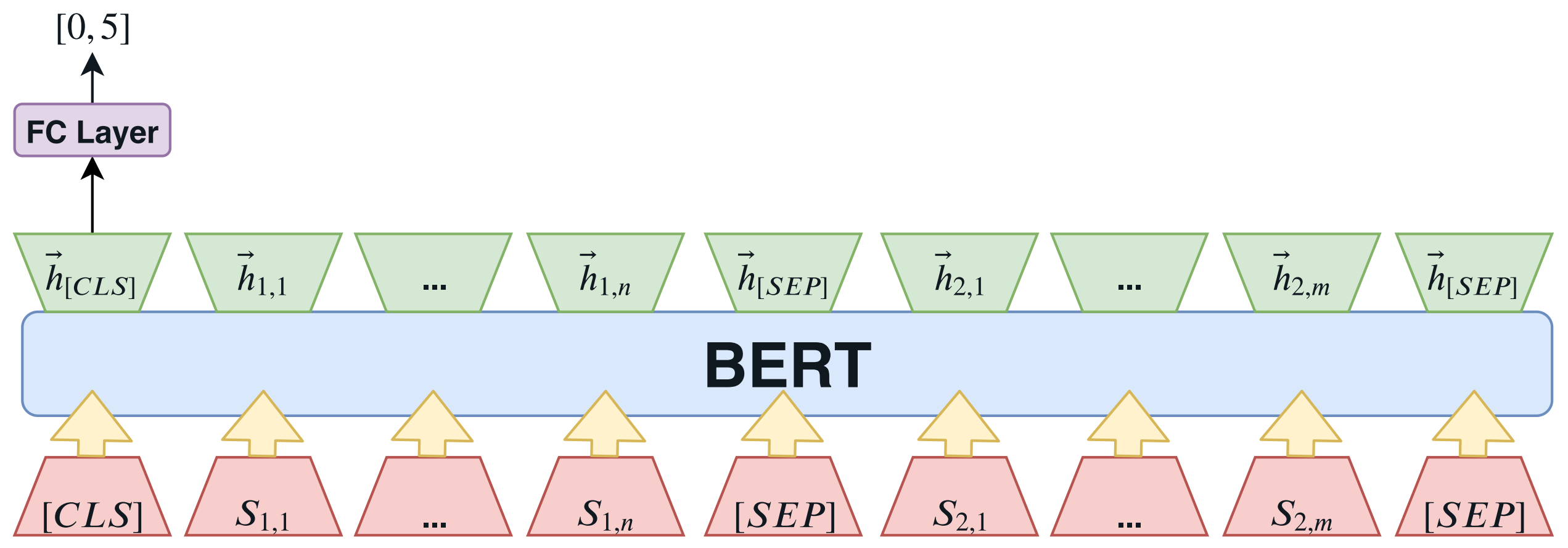


Fig. 5: Finetuning BERT towards clinical semantic similarity. Given two sub-word tokenized sentences, a fully connected linear regression layer finetunes BERT towards ascertaining the semantic similarity of the sentences.

Training Scheme

1. Finetune BERT on STS-B - a dataset of general English sentences annotated for semantic similarity.
2. Finetune parameters from previous step on MedSTS - the shared task training set.

Data Augmentation

Increasing the number of training instances is a great way to improve a model. We augment our training data with the 16,000 instance STS-B dataset.

Dataset	Statistic	Train	Development	Test
MedSTS	Sentence Pair Count	1313	328	410
STS-B	Sentence Pair Count	5749	1500	8628

Results

- BERT finetuned solely on STS-B and evaluated Med-STS performs suprisingly well. Their appears to be a large amount of information transfer between web and clinical STS.
- Finetuning on MedSTS after finetuning on STS-B yields a slight generalization improvement.

Model	Data	ρ
Clinical-Bert	STS-B-train-dev-test	.771
Clinical-Bert-5	MedSTS-train	.838
Clinical-Bert-10	MedSTS-train	.849
Clinical-Bert-Transfer	MedSTS-train	.854

Tab. 4: Correlation of our BERT based models on the MedSTS held-out evaluation set.

Codebase: github.com/AndriyMulyar/semantic-text-similarity