Domain-specific BERT Finetuning Signal
Masked LM

- **Large In-domain dataset**
  - Patent
  - Clinical/Biomedical articles (Pubmed)
  - Scientific articles

### Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing

- **BioMegatron**: Larger Biomedical Domain Language Model
- **Patent Classification by Fine-Tuning BERT Language Model**
- **BERT**: A Pretrained Language Model for Scientific Text
Multitask Learning

- NER
- Sentiment classification
- Question Answering
- Relation Extraction
- Information Extraction
- Textual Entailment

An Empirical Study of Multi-Task Learning on BERT for Biomedical Text Mining
MT-Clinical BERT: Scaling Clinical Information Extraction with Multitask
FinBERT: A pretrained LM for Financial Communication
Masked LM

- **Span MLM**
  - Span length is sampled from a Geometric distribution
  - Span are randomly selected

SpanBERT: Improving Pre-training by Representing and Predicting Spans
What is the [cause] of [polio]? Poliomyelitis is caused by infection with a member ....

What is the [signs and symptoms] of [polio]? The term "poliomyelitis" is used to identify ....
Masked LM

- Special span selections
  - Noun phrase
  - Low frequent phrase
  - Keywords

SpanBERT: Improving Pre-training by Representing and Predicting Spans

We present SpanBERT, a pre-training method that is designed to better represent and predict spans of text. Our approach extends BERT by (1) masking contiguous random spans, rather than random tokens, and (2) training the span boundary representations to predict the entire content of the masked span, without relying on the individual token representations within it.
Knowledge-aware LM

• Sentence representation
  - Entities are collected from n-gram dictionary
    \[ X_{\text{duet}} = \begin{cases} \{w_1, \ldots, w_i, \ldots, w_T\} & \text{Word Sequence;} \\ \{e_1, \ldots, e_i, \ldots, e_T\} & \text{Entity Sequence.} \end{cases} \]

• Embeddings
  \[ \vec{e}_i = \text{Embedding}_e(e_i) \in \mathbb{R}^{d_e}, \]
  \[ \vec{w}_i = \text{Embedding}_w(w_i) \in \mathbb{R}^{d_w}. \]

• Knowledge-aware input
  \[ \vec{t}_i = \vec{w}_i + \text{Linear}_t(\vec{e}_i), \text{ Linear}_t \in \mathbb{R}^{d_e \times d_w}. \]

• Next entity prediction
  \[ l_e(e_i|t_{<i}) = \max(0, s(\vec{h}_i^L, \vec{e}_i^L) - s(\vec{h}_i^L, \vec{e}_-^L) + \lambda), \]
  \[ s(\vec{h}_i^L, \vec{e}_j^L) = \cos(\text{Linear}(\vec{h}_i^L), \vec{e}_j^L), \]
  \[ \vec{h}_i^L = \text{transformer}^L(t_{<i}). \]

• Joint train NWP and NEP
  \[ l_{\text{KALM}}(X_{\text{duet}}) = \sum_i l_w(p(w_i|t_{<i})) + \alpha l_e(e_i|t_{<i}). \]

Knowledge-Aware Language Model Pretraining
Structure prediction

- **Word reordering**

  ![Diagram of word reordering](image)

  - Actual Tokens
  - Transformer Encoder
  - Positional Embedding
  - Token Sequence
  - Pre-shuffled Trigram

- **Sentence order prediction**

  ![Diagram of sentence order prediction](image)

  - Next Sent Prediction
  - Prev. Sent Prediction
  - Sampler
  - Transformer Encoder
  - Corpus
  - Random Sent Prediction

---

*StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding*
Data selection

Language model perplexity (PPL)
Jensen-Shannon divergence (JSD)
Target vocabulary covered (TVC)
Type token ratio (TTR) = \#unique-tokens/\#tokens

Figure 3: Correlation between different similarity measures and diversity measure and the improvement (Δ) due to domain-specific BERT models.

Cost-effective Selection of Pretraining Data: A Case Study of Pretraining BERT on Social Media