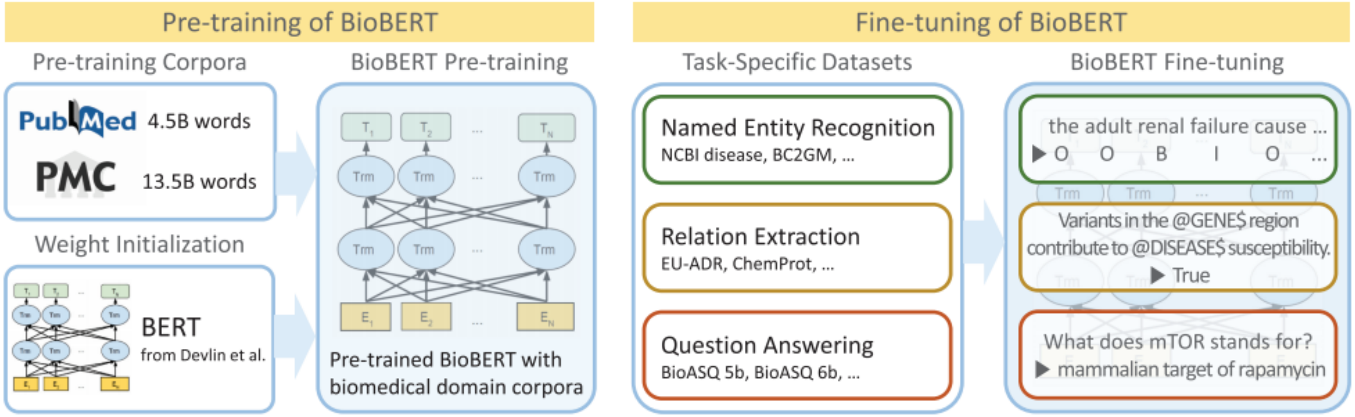
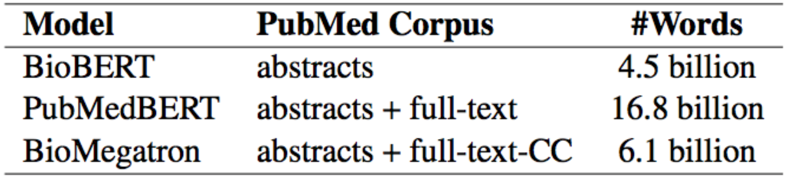
Domain-specific BERT Finetuning Signal

Masked LM

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





Model	PubMed Corpus	#Words
BioBERT	abstracts	4.5 billion
PubMedBERT	abstracts + full-text	16.8 billion
BioMegatron	abstracts + full-text-CC	6.1 billion

Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing

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

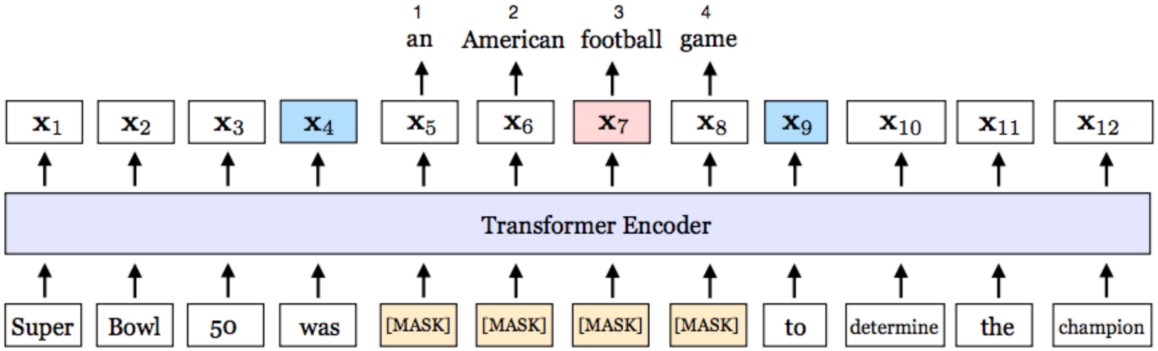
Multitask Learning

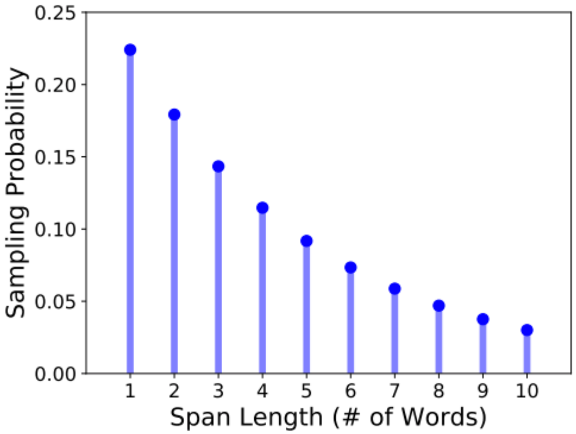
- **NER**
- **Sentiment classification**
- **Question Answering**
- **Relation Extraction**
- **Information Extraction**
- **Textual Entailment**

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**


1
an
2
American
3
football
4
game
$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$ $\mathbf{x}_4$ $\mathbf{x}_5$ $\mathbf{x}_6$ $\mathbf{x}_7$ $\mathbf{x}_8$ $\mathbf{x}_9$ $\mathbf{x}_{10}$ $\mathbf{x}_{11}$ $\mathbf{x}_{12}$
Transformer Encoder
Super
Bowl
50
was
[MASK]
[MASK]
[MASK]
[MASK]
to
determine
the
champion



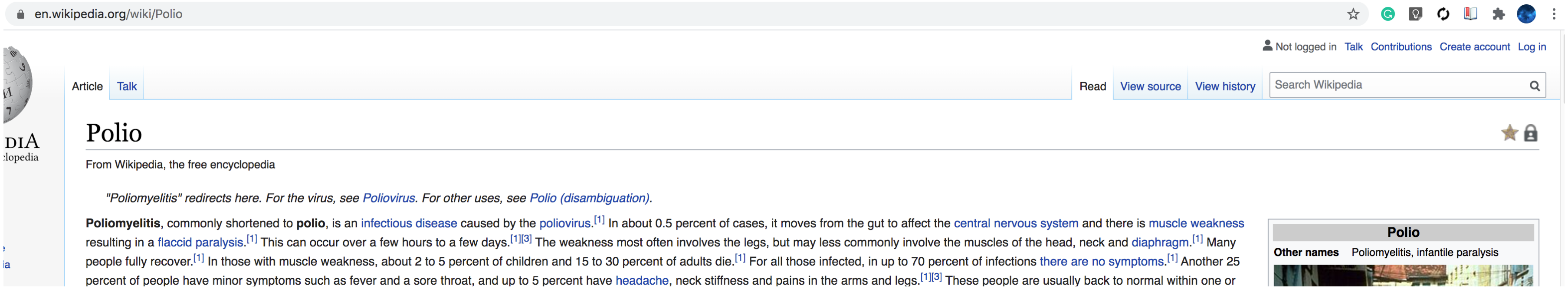
SpanBERT: Improving Pre-training by Representing and Predicting Spans

Masked LM

- **Cloze form question**
- What is [**aspect**] of [**disease**]? <**Paragraph of text**>

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** ….



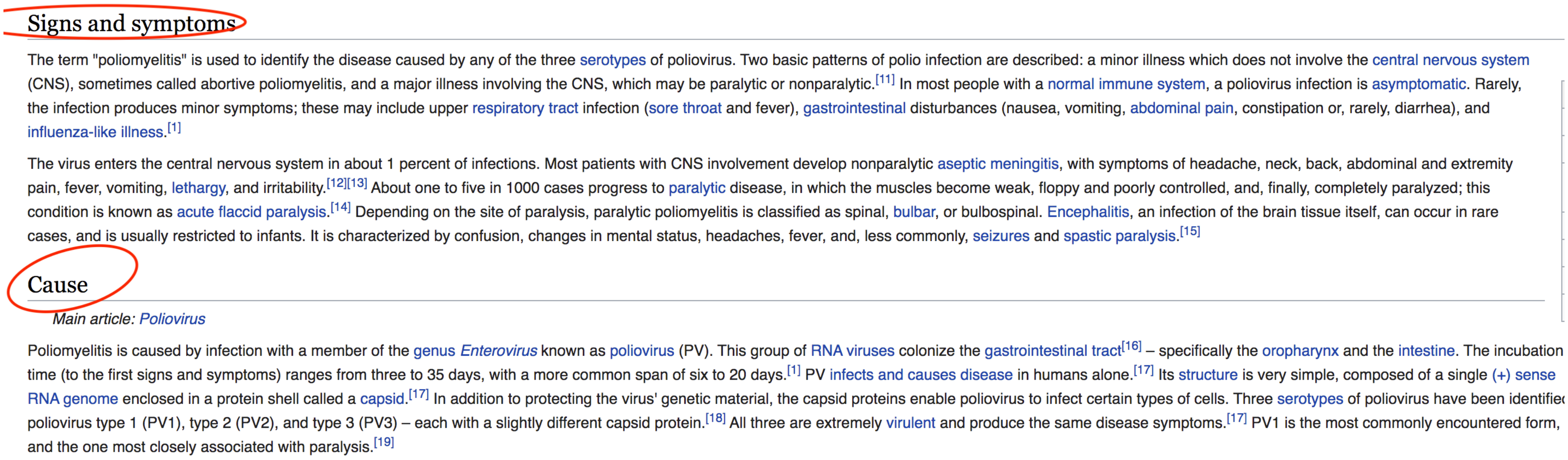
Polio

From Wikipedia, the free encyclopedia

"Poliomyelitis" redirects here. For the virus, see Poliovirus. For other uses, see Polio (disambiguation).

Poliomyelitis, commonly shortened to **polio**, is an infectious disease caused by the poliovirus.[1] In about 0.5 percent of cases, it moves from the gut to affect the central nervous system and there is muscle weakness resulting in a flaccid paralysis.[1] This can occur over a few hours to a few days.[1][3] The weakness most often involves the legs, but may less commonly involve the muscles of the head, neck and diaphragm.[1] Many people fully recover.[1] In those with muscle weakness, about 2 to 5 percent of children and 15 to 30 percent of adults die.[1] For all those infected, in up to 70 percent of infections there are no symptoms.[1] Another 25 percent of people have minor symptoms such as fever and a sore throat, and up to 5 percent have headache, neck stiffness and pains in the arms and legs.[1][3] These people are usually back to normal within one or

Polio	
Other names	Poliomyelitis, infantile paralysis



Signs and symptoms

The term "poliomyelitis" is used to identify the disease caused by any of the three serotypes of poliovirus. Two basic patterns of polio infection are described: a minor illness which does not involve the central nervous system (CNS), sometimes called abortive poliomyelitis, and a major illness involving the CNS, which may be paralytic or nonparalytic.[11] In most people with a normal immune system, a poliovirus infection is asymptomatic. Rarely, the infection produces minor symptoms; these may include upper respiratory tract infection (sore throat and fever), gastrointestinal disturbances (nausea, vomiting, abdominal pain, constipation or, rarely, diarrhea), and influenza-like illness.[1]

The virus enters the central nervous system in about 1 percent of infections. Most patients with CNS involvement develop nonparalytic aseptic meningitis, with symptoms of headache, neck, back, abdominal and extremity pain, fever, vomiting, lethargy, and irritability.[12][13] About one to five in 1000 cases progress to paralytic disease, in which the muscles become weak, floppy and poorly controlled, and, finally, completely paralyzed; this condition is known as acute flaccid paralysis.[14] Depending on the site of paralysis, paralytic poliomyelitis is classified as spinal, bulbar, or bulbospinal. Encephalitis, an infection of the brain tissue itself, can occur in rare cases, and is usually restricted to infants. It is characterized by confusion, changes in mental status, headaches, fever, and, less commonly, seizures and spastic paralysis.[15]

Cause

Main article: Poliovirus

Poliomyelitis is caused by infection with a member of the genus *Enterovirus* known as poliovirus (PV). This group of RNA viruses colonize the gastrointestinal tract[16] – specifically the oropharynx and the intestine. The incubation time (to the first signs and symptoms) ranges from three to 35 days, with a more common span of six to 20 days.[1] PV infects and causes disease in humans alone.[17] Its structure is very simple, composed of a single (+) sense RNA genome enclosed in a protein shell called a capsid.[17] In addition to protecting the virus' genetic material, the capsid proteins enable poliovirus to infect certain types of cells. Three serotypes of poliovirus have been identifie poliovirus type 1 (PV1), type 2 (PV2), and type 3 (PV3) – each with a slightly different capsid protein.[18] All three are extremely virulent and produce the same disease symptoms.[17] PV1 is the most commonly encountered form, and the one most closely associated with paralysis.[19]

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.

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**

$$X_{\text{duet}} = \begin{cases} \{w_1, ..., w_i, ..., w_T\} & \text{Word Sequence;} \\ \{e_1, ..., e_i, ..., e_T\} & \text{Entity Sequence.} \end{cases}$$

- **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**

$$\vec{t_i} = \vec{w_i} + \texttt{Linear}_t(\vec{e_i}), \; \texttt{Linear}_t \in \mathbb{R}^{d_e \times d_w}.$$

- **Next entity prediction**

$$l_e(e_i|t_{<i}) = \texttt{max}(0, \texttt{s}(\vec{h_i}^L, \vec{e_i}) - \texttt{s}(\vec{h_i}^L, \vec{e_-}) + \lambda),$$
$$\texttt{s}(\vec{h_i}^L, \vec{e_j}) = \texttt{cos}(\texttt{Linear}(\vec{h_i}^L), \vec{e_j}),$$
$$\vec{h_i}^L = \texttt{transformer}^L(t_{<i}).$$

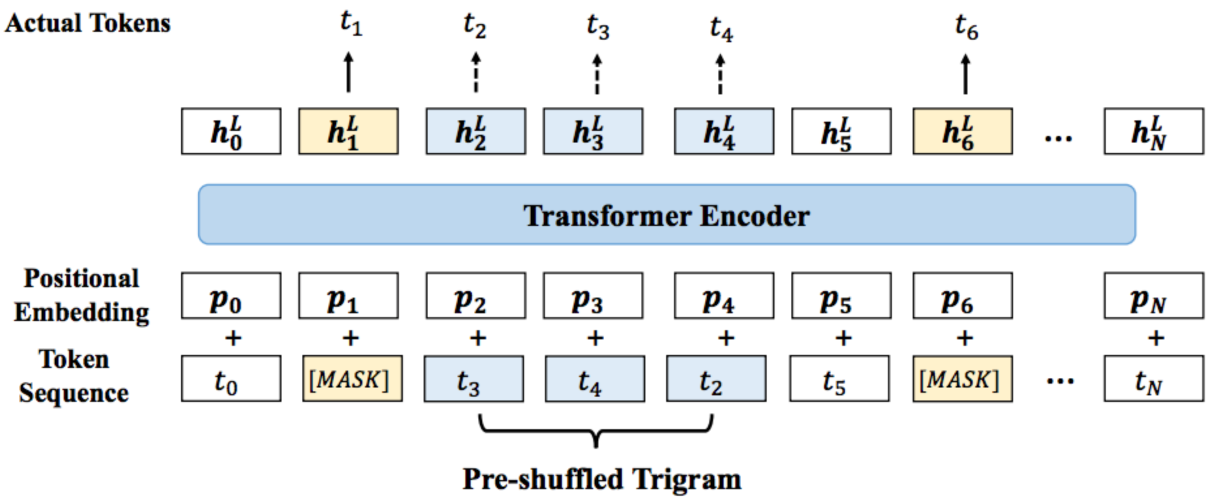
- **Joint train NWP and NEP**

$$l_{\text{KALM}}(X_{\text{duet}}) = \sum_i l_w(p(w_i|t_{<i})) + \alpha l_e(e_i|t_{<i}).$$

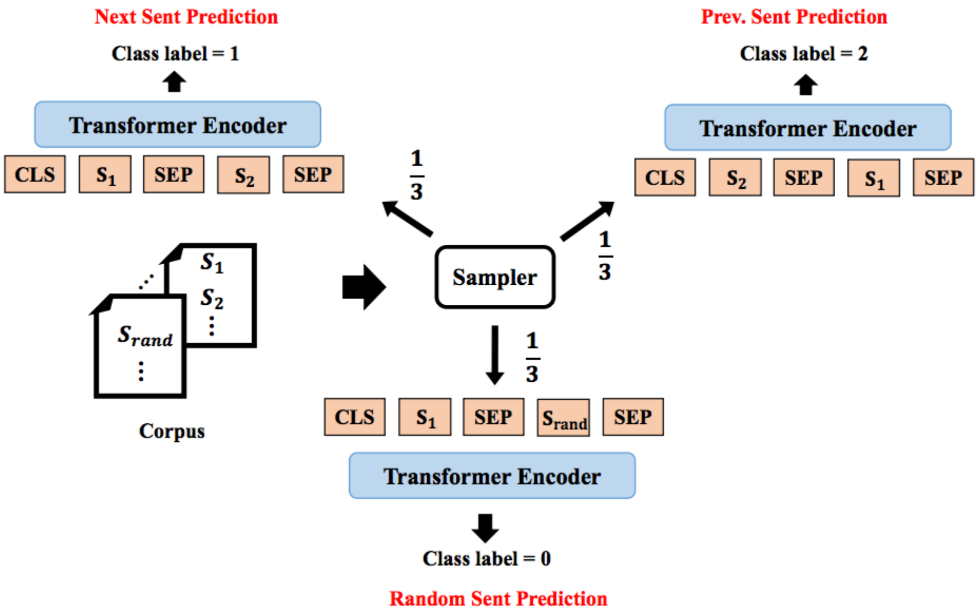
Knowledge-Aware Language Model Pretraining

Structure prediction

- **Word reordering**



- **Sentence order 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

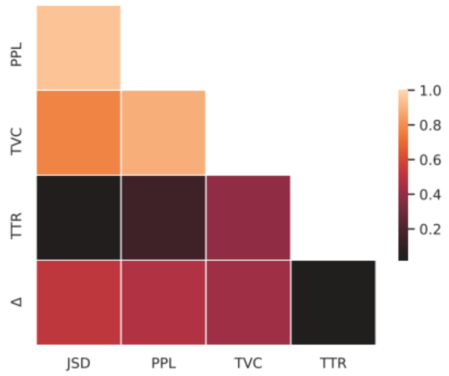


Figure 3: Correlation between different similarity measures and diversity measure and the improvement (Δ) due to domain-specific BERT models.

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