/01

Презентацию подготовила Виктория Фирсанова
СПбГУ, кафедра математической лингвистики

BERT,

Devlin et al.

Google

Дискуссионный клуб NLP| 2020

Обзор

SOTA

Question Answering (SQuAD v1.1)

Natural Language Inference (тестирование понимания)

ПРЕДПОССЫЛКА	МЕТКА (ЛЕЙБЛ)	ГИПОТЕЗА
A man inspects the uniform of a figure in some East Asian country.	**contradiction** C C C C C *ручная разметка*	The man is sleeping **TRUE**
An older and younger man smiling.	**neutral** N N E N N	Two men are smiling and laughing at the cats playing on the floor. **UNDETERMINED**
A black race car starts up in front of a crowd of people.	**contradiction** C C C C C *авторская метка*	A man is driving down a lonely road. **FALSE**
A soccer game with multiple males playing.	**entailment** E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	**neutral** N N E C N	A happy woman in a fairy costume holds an umbrella.

Bowman et al. (2015)

Обзор

SOTA

Question Answering (SQuAD v1.1)

Natural Language Inference (тестирование понимания)

The Stanford Question Answering Dataset

"Проводник" по SQuAD: https://rajpurkar.github.io/SQuAD-explorer/

Версия 1: Rajpurkar et al. *SQuAD: 100,000+ Questions for Machine Comprehension of Text* (2016)

Версия 2: Rajpurkar et al. *Know What You Don't Know: Unanswerable Questions for SQuAD* (2018)

Обзор

SQuAD

- вопросы к статьям Википедии;
- ответы - отрывки текста из соответствующего параграфа;
- есть и вопросы, на которые ответить невозможно

SQuAD 1.1

- более 100,000 вопросов вопросно-ответных пар
- в основе более 500 статей

SQuAD 2.0

- + более 50,000 неотвечаемых вопросов
- задача системы: выявлять такие вопросы

Обзор

SQuAD 2.0


```
{
  "version": "v2.0",
  "data": [
    {
      "title": "Beyoncé",
      "paragraphs": [
        {
          "qas": [
            {
              "question": "When did Beyonce start becoming popular?",
              "id": "56be85543aeaaa14008c9063",
              "answers": [
                {
                  "text": "in the late 1990s",
                  "answer_start": 269
                }
              ],
              "is_impossible": false
            },
```

```
"context": "Following the disbandment of Destiny's Child in
June 2005, she released her second solo album, B'Day (2006), which co
tained hits \"Déjà Vu\", \"Irreplaceable\", and \"Beautiful Liar\". Be
yoncé also ventured into acting, with a Golden Globe-nominated perfor
ance in Dreamgirls (2006), and starring roles in The Pink Panther (20
6) and Obsessed (2009). Her marriage to rapper Jay Z and portrayal of
Etta James in Cadillac Records (2008) influenced her third album, I A
m... Sasha Fierce (2008), which saw the birth of her alter-ego Sasha
ierce and earned a record-setting six Grammy Awards in 2010, includin
Song of the Year for \"Single Ladies (Put a Ring on It)\". Beyoncé to
k a hiatus from music in 2010 and took over management of her career;
her fourth album 4 (2011) was subsequently mellower in tone, explorin
1970s funk, 1980s pop, and 1990s soul. Her critically acclaimed fifth
studio album, Beyoncé (2013), was distinguished from previous release
by its experimental production and exploration of darker themes."
```

Идея BERT

Обычно модели учитывают при анализе только правый или левый контекст, а **в основе** BERT - двусторонний анализ.

Результат: более глубокий анализ контекста и связей внутри высказываний.

Что еще нового? Техника Masked LM (MLM).

Трансферное обучение:
предобученная модель -> настройка под конкретную задачу
(расширенная модель: базовая LM + верхний слой под вашу задачу).

Обучение BERT

Механизм внимания (учитывает контекстуальные связи между далеко стоящими словами и позволяет строить целые тексты)

+ MLM (Mask Language Modelling)

- 15% слов заменяются на фиктивный токен [MASK]
  - *собака лает во [MASK], я сижу на [MASK]*
- задача: предсказать слова "*дворе*" и "*диване*"

+ NSP (Next Sentence Prediction)

- 50% входных данных: последовательная пара из корпуса
- 50%: предложение + случайное предложение из корпуса
- задача: установить, является ли второе предложение в паре последующим предложением в исходном документе

Недостатки BERT

- **За[MASK]ированные слова не зависят друг от друга** в модели, но в реальных данных это не так;
- фиктивные токены вносят **дополнительный шум** во входные данные, что не соответствует фактическим распределениям;
- можно сказать, что BERT не работает с текстами, которые встречаются в реальной жизни.

И все же он классный

- гибкий, легко настроить под совершенно разные задачи;
- работает на 104 языках (Википедия - это сила... ну, одна из сил)

Доступные модели здесь: https://github.com/google-research/bert

Пример использования

Question: Why was Tesla returned to Gospic?
Context paragraph: On 24 March 1879, Tesla was returned to Gospic under police guard for not having a residence permit. On 17 April 1879, Milutin Tesla died at the age of 60 after contracting an unspecified illness (although some sources say that he died of a stroke). During that year, Tesla taught a large class of students in his old school, Higher Real Gymnasium, in Gospic.
Answer: not having a residence permit

Пример из CS 224N Default Final Project: Question Answering on SQuAD 2.0
URL: http://web.stanford.edu/class/cs224n/project/default-final-project-handout.pdf

Список источников:

1. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
2. Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*. Association for Computational Linguistics.

Подготовила Виктория Фирсанова,
СПбГУ, кафедра математической лингвистики
E-mail: vifirsanova@gmail.com