



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
УНИВЕРСИТЕТ

# Компьютерная лингвистика и информационные технологии

Улица сезам, бертология и графы  
(используются материалы Е. Артемовой, ФКН ВШЭ)

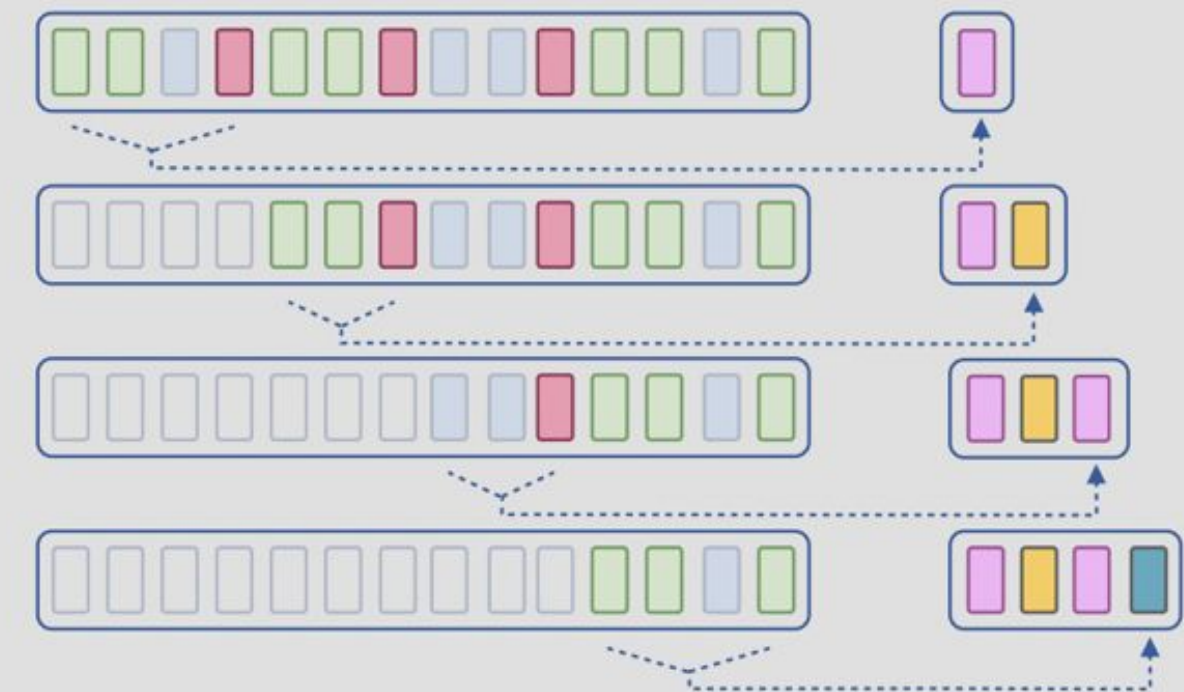


# Byte Pair Encoding Subword Tokenization

- Чего мы хотим от токенизации?
- Представим слова как набор “подслов”
- В чем отличие от fastText?
- Алгоритм сжатия данных

## Попарное битовое кодирование (Byte pair encoding, BPE)

- Считаем частоты пар символов
- Склеиваем самую частую пару символов и превращаем ее в новый символ
- Продолжаем повторять операцию фиксированное число раз



```
text = 'на дворе трава на дворе дрова'
```

```
group_subtokens(text)
```

```
['[CLS]', 'на', 'дворе', 'т', '##рава', 'на', 'дворе', 'др', '##ова', '[SEP]']
```





# BERT внутри

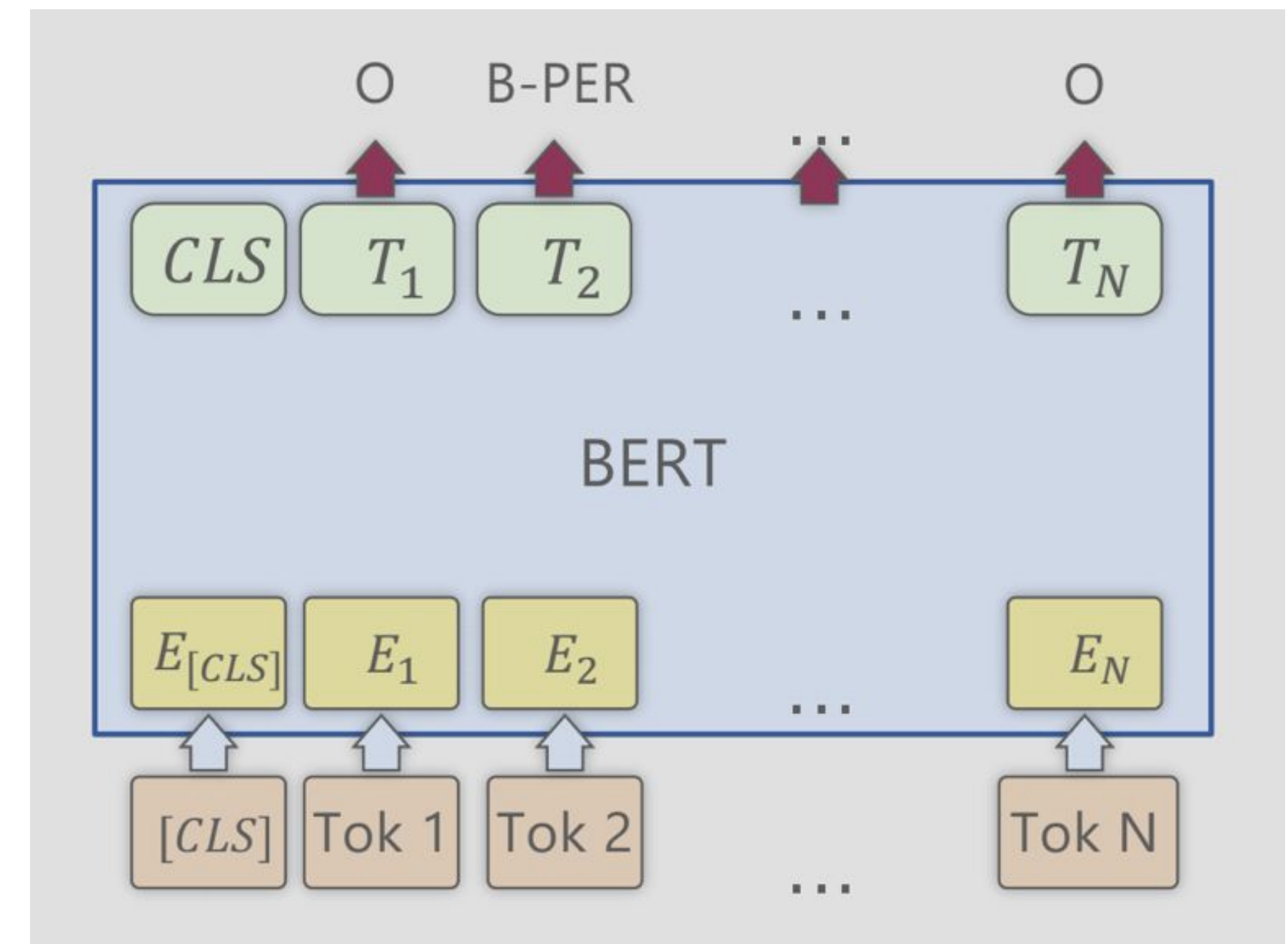
<https://colab.research.google.com/drive/19OFFe9C1D8-5U5uO-WV3anX2jUP1Qfcp?usp=sharing>

```
class BertForTokenClassification(BertPreTrainedModel):

    _keys_to_ignore_on_load_unexpected = [r"pooler"]

    def __init__(self, config):
        super().__init__(config)
        self.num_labels = config.num_labels

        self.bert = BertModel(config, add_pooling_layer=False)
        self.dropout = nn.Dropout(config.hidden_dropout_prob)
        self.classifier = nn.Linear(config.hidden_size, config.num_labels)
```





# BERT внутри

```
class BertForSequenceClassification(BertPreTrainedModel):  
    def __init__(self, config):  
        super().__init__(config)  
        self.num_labels = config.num_labels  
  
        self.bert = BertModel(config)  
        self.dropout = nn.Dropout(config.hidden_dropout_prob)  
        self.classifier = nn.Linear(config.hidden_size, config.num_labels)
```

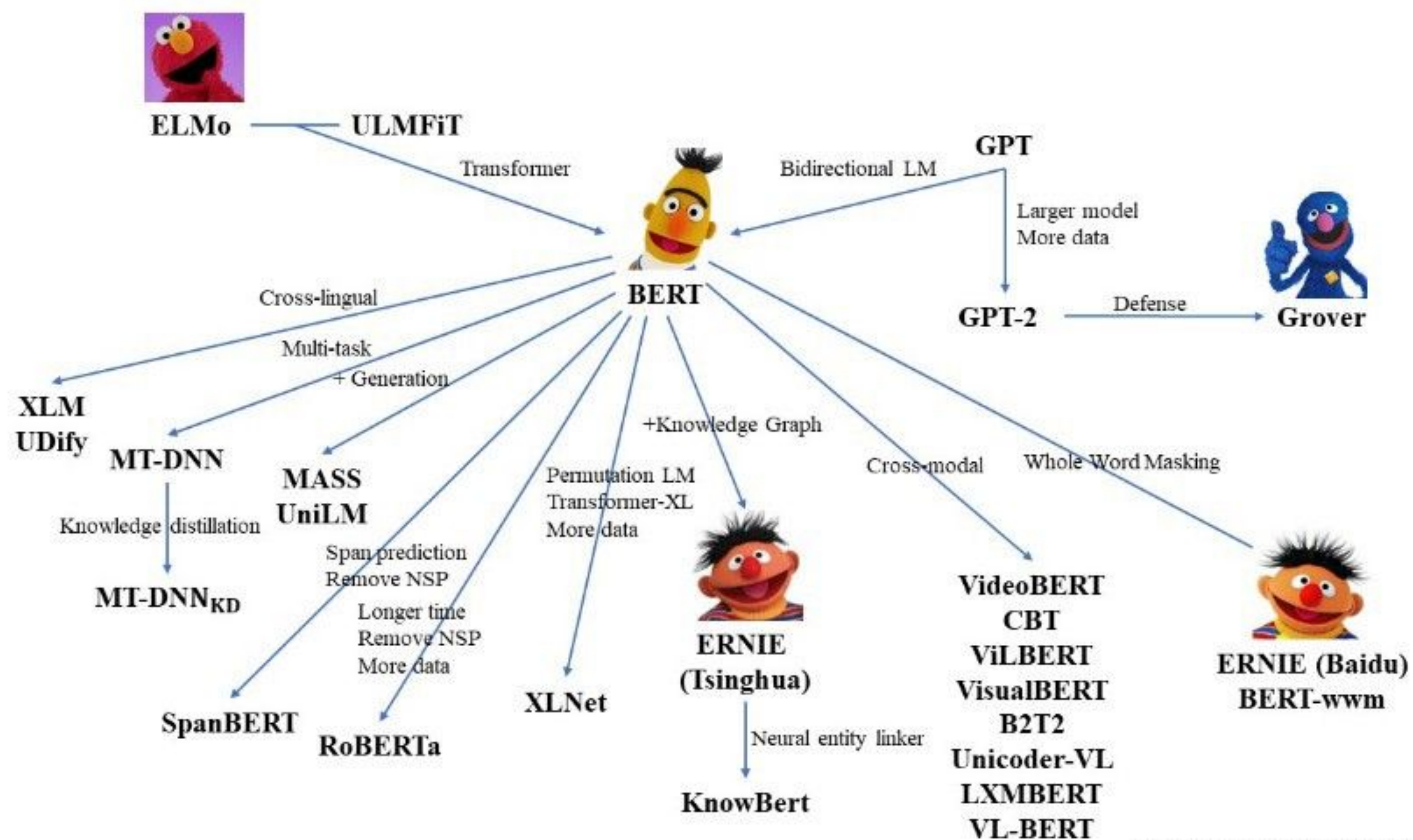


# BERT as Embedder: Pooling

- [CLS]-пулинг – вектор управляющего токена CLS на последнем слое
- MEAN-пулинг – усреднение векторов слов на последнем слое
- MAX-пулинг – покомпонентный максимум векторов слов на последнем слое



# Улица Сезам





# Улица Сезам

<https://arxiv.org/pdf/2003.07278.pdf>

Method	Architecture	Encoder	Decoder	Objective	Dataset
ELMo	LSTM	✗	✓	LM	1B Word Benchmark
GPT	Transformer	✗	✓	LM	BookCorpus
GPT2	Transformer	✗	✓	LM	Web pages starting from Reddit
BERT	Transformer	✓	✗	MLM & NSP	BookCorpus & Wiki
RoBERTa	Transformer	✓	✗	MLM	BookCorpus, Wiki, CC-News, OpenWebText, Stories
ALBERT	Transformer	✓	✗	MLM & SOP	Same as RoBERTa and XLNet
UniLM	Transformer	✓	✗	LM, MLM, seq2seq LM	Same as BERT
ELECTRA	Transformer	✓	✗	Discriminator (o/r)	Same as XLNet
XLNet	Transformer	✗	✓	PLM	BookCorpus, Wiki, Giga5, ClueWeb, Common Crawl
XLM	Transformer	✓	✓	CLM, MLM, TLM	Wiki, parallel corpora (e.g. MultiUN)
MASS	Transformer	✓	✓	Span Mask	WMT News Crawl
T5	Transformer	✓	✓	Text Infilling	Colossal Clean Crawled Corpus
BART	Transformer	✓	✓	Text Infilling & Sent Shuffling	Same as RoBERTa





# Улица Сезам

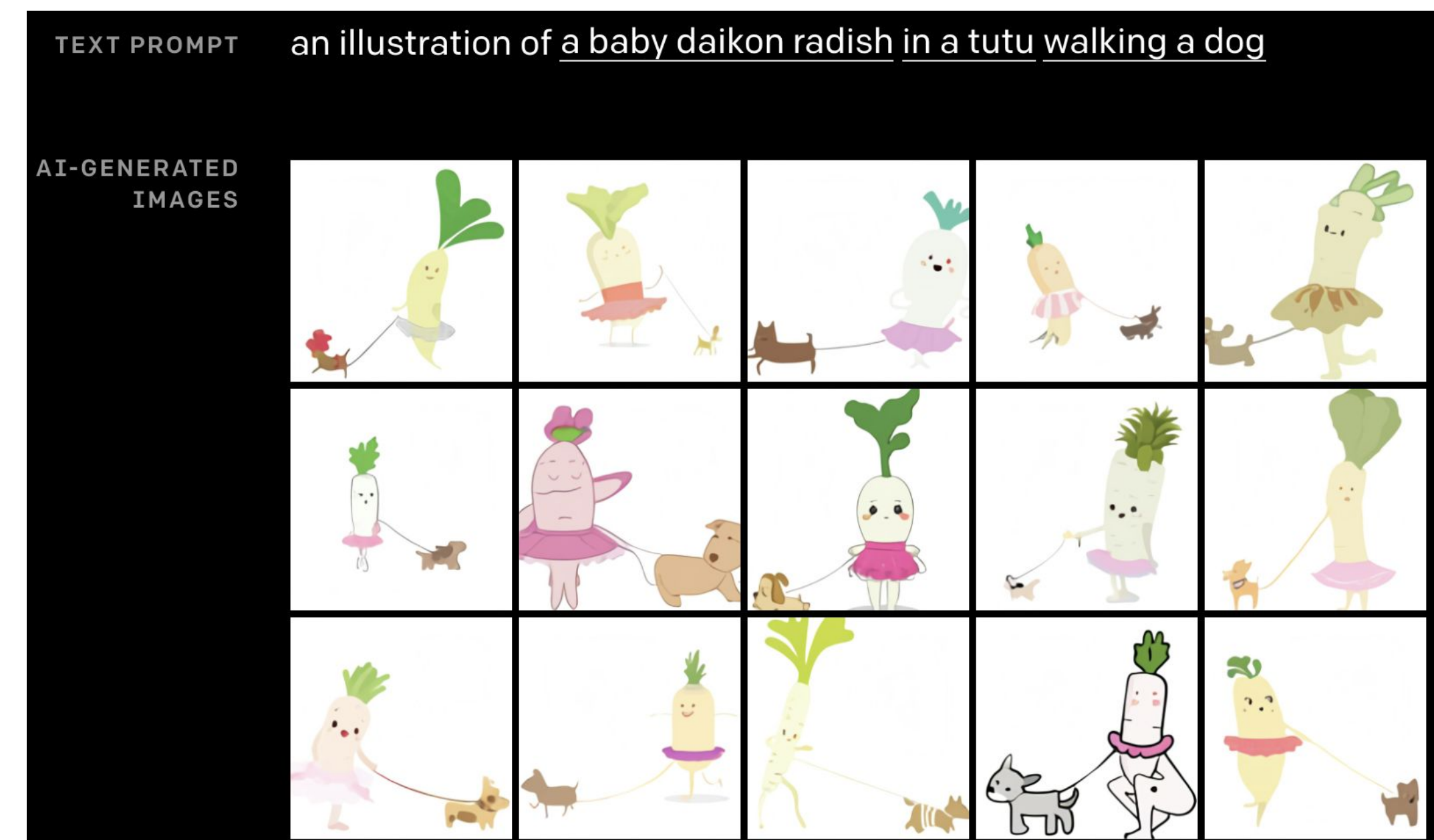
<https://arxiv.org/pdf/2003.07278.pdf>

Objective	Inputs	Targets
LM	[START]	I am happy to join with you today
MLM	I am [MASK] to join with you [MASK]	happy today
NSP	Sent1 [SEP] Next Sent or Sent1 [SEP] Random Sent	Next Sent/Random Sent
SOP	Sent1 [SEP] Sent2 or Sent2 [SEP] Sent1	in order/reversed
Discriminator (o/r)	I am thrilled to study with you today	o o r o r o o o
PLM	happy join with	today am I to you
seq2seq LM	I am happy to	join with you today
Span Mask	I am [MASK] [MASK] [MASK] with you today	happy to join
Text Infilling	I am [MASK] with you today	happy to join
Sent Shuffling	today you am I join with happy to	I am happy to join with you today
TLM	How [MASK] you [SEP] [MASK] vas-tu	are Comment



# Улица Сезам и не только

- Авторегрессионные модели: LM objective (GPT, Reformer)
- Автоэнкодеры: pre-training objectives (BERT, RoBERTa)
- Модели seq2seq: enc-dec architecture (T5, BART)
- Мультимодальные модели: CV & NLP (DALL-E, VQA Models, Speech2Text)





# Как оценить модели?

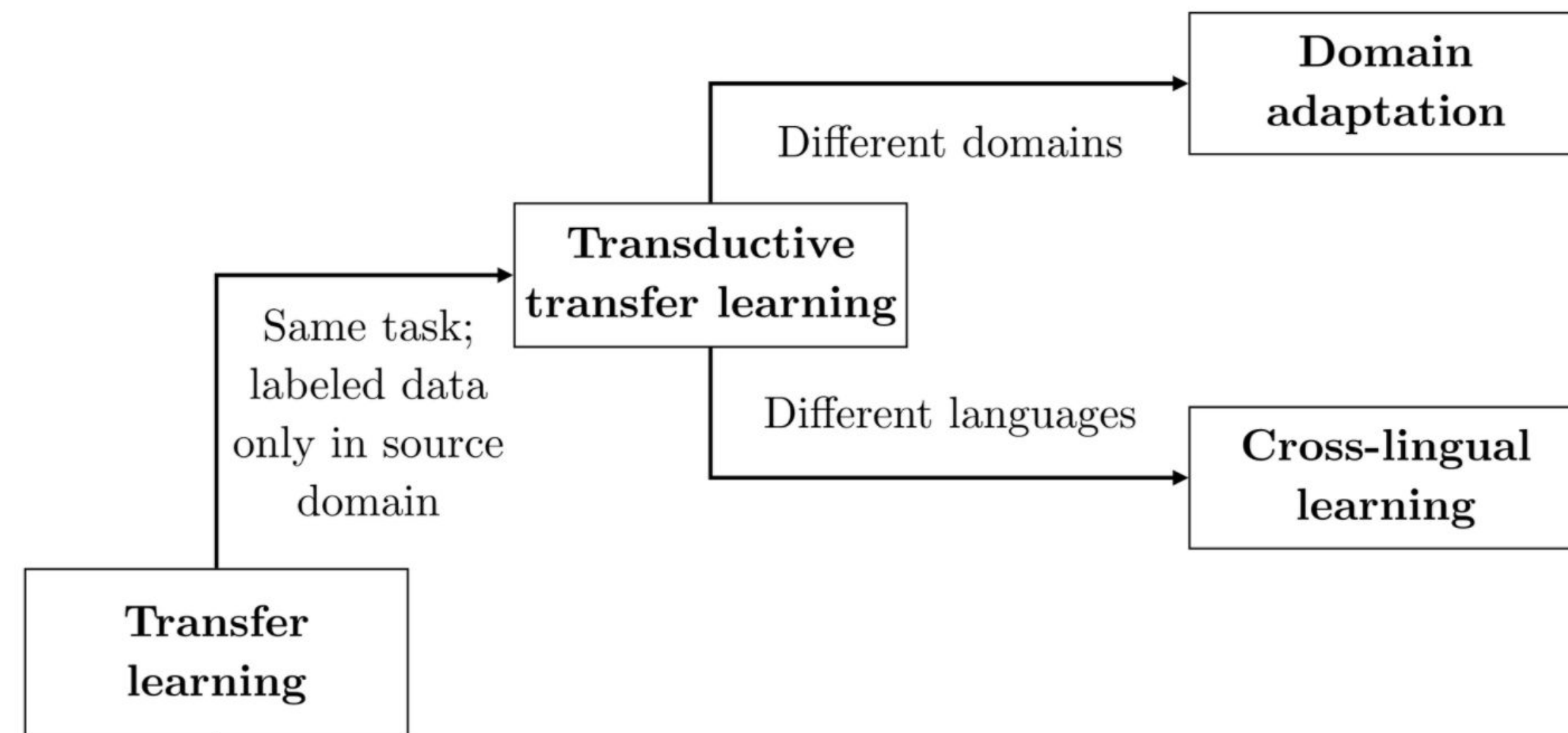
- NLU-бенчмарки GLUE, SuperGLUE, XTREME
- Пробинг моделей
- Производительность моделей





# Бенчмарки

- GLUE: 10 NLU-задач (linguistic acceptability, categories)
- SuperGLUE: 9 еще более сложных задач
- XTREME: перенос знания между языками



Cross-lingual learning in the transfer learning taxonomy (Ruder, 2019)



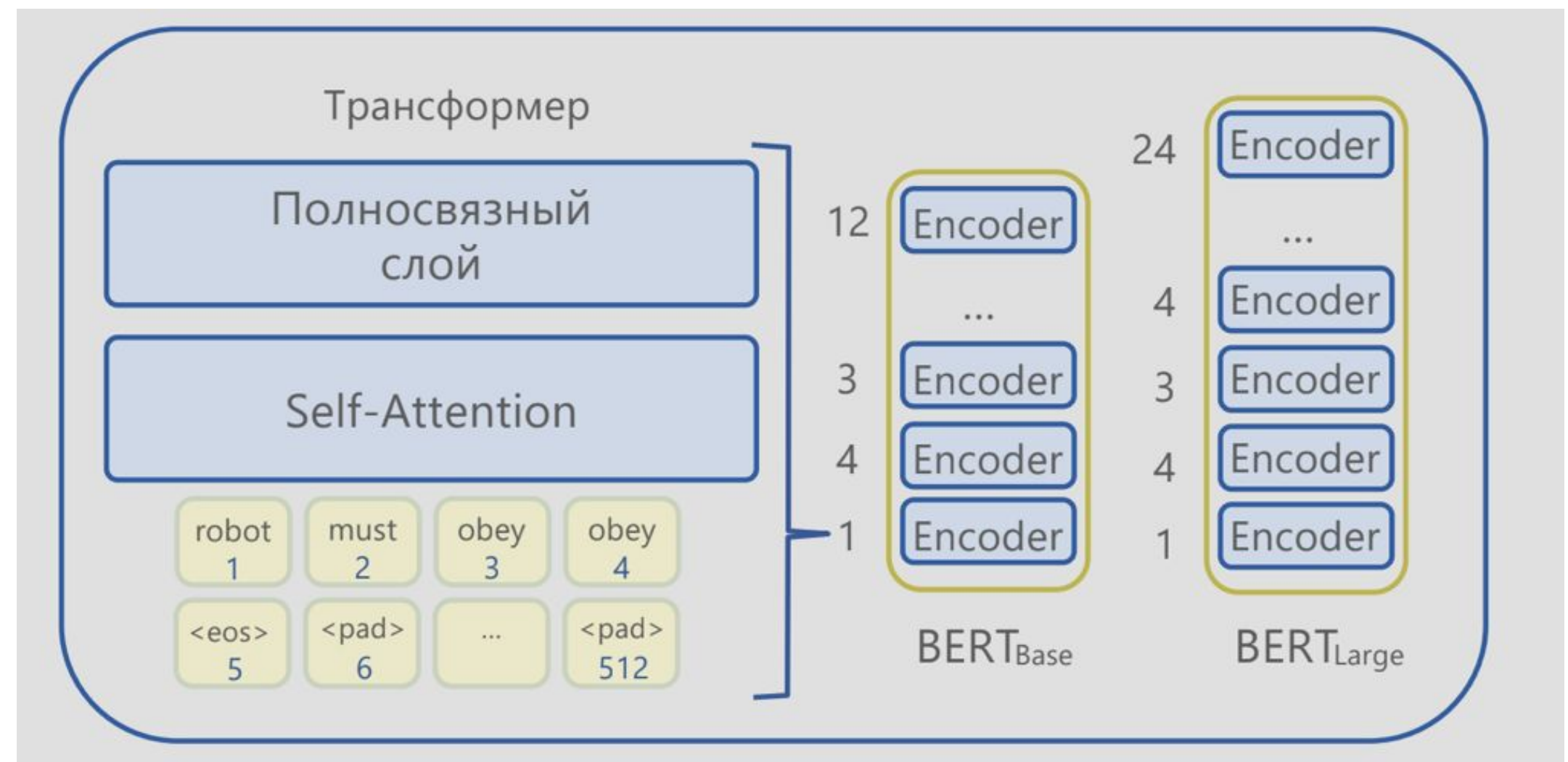
# GLUE, SuperGLUE, XTREME

- Лидерборды позволяют сравнивать между собой языковые модели
- Больше параметров и данных -- языковая модель выше в лидерборде
- Затраты на скорость, память и ресурсы при этом почти не учитываются
- Высокий рейтинг языковой модели не означает хорошие результаты на практике



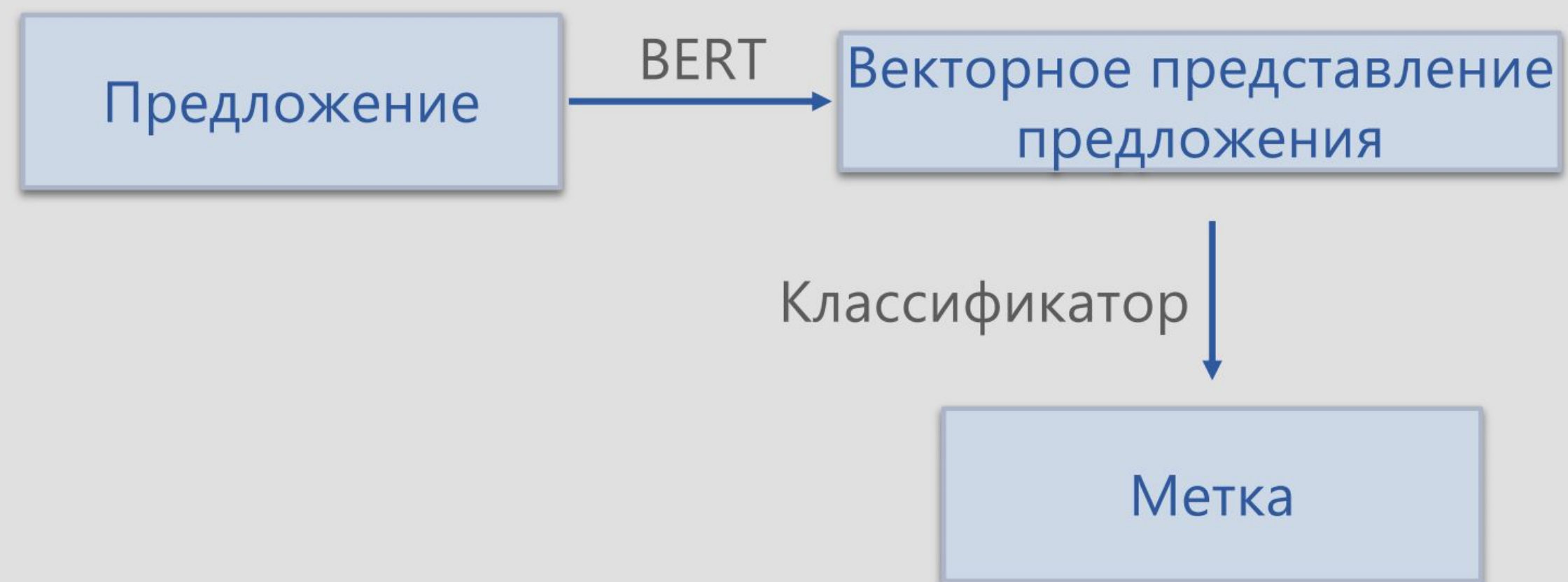
# Бертология

- BERT - модель динамических векторов слов
- Инструменты для изучения и интерпретации:
  - Анализ векторных представлений слов и предложений
  - Анализ весов механизма внимания
  - Диагностические тесты (probing tasks)



# Пробинг

На уровне предложения



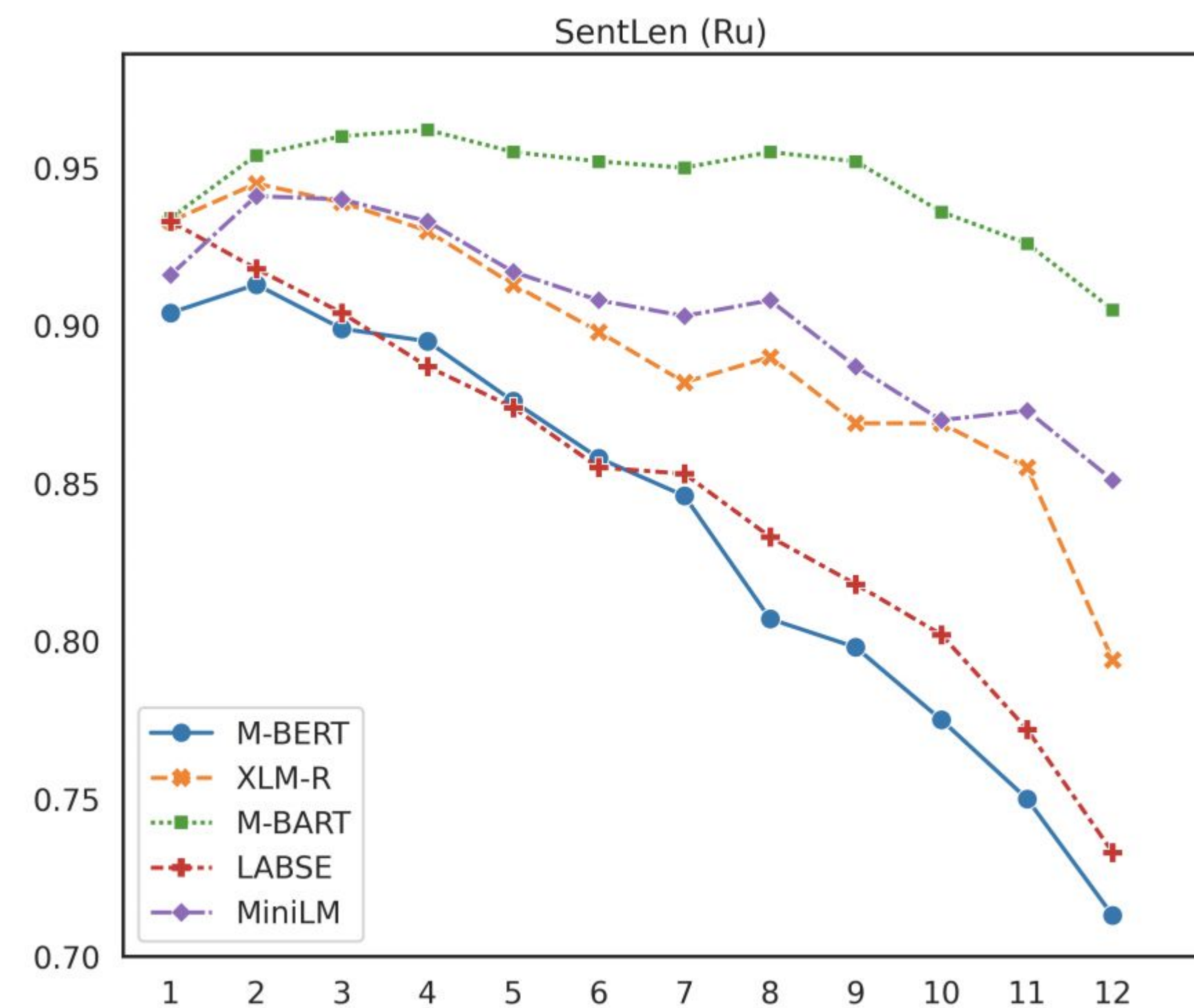
На уровне слова





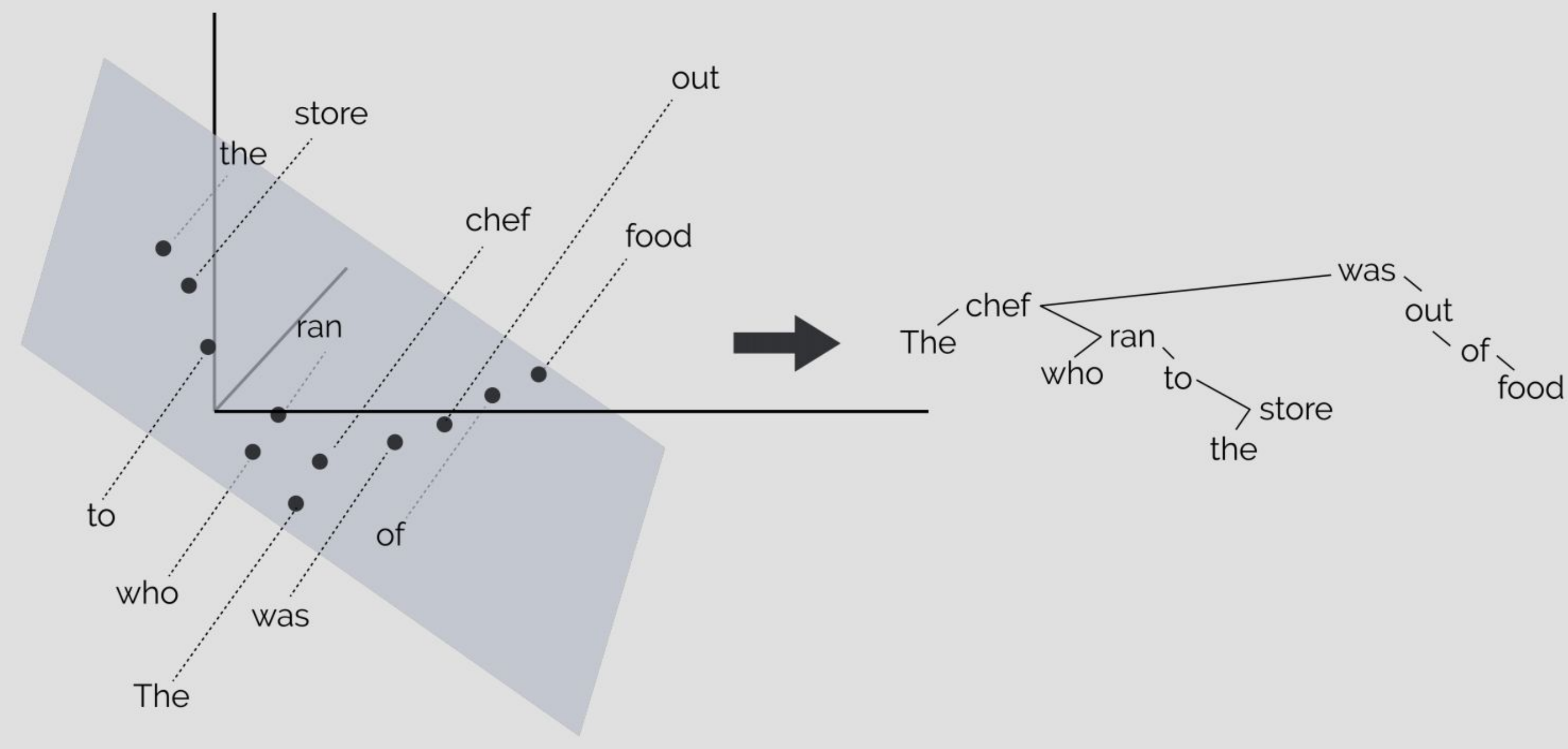
# BERT и уровни языка

- BERT и его компания умеют решать простые диагностические тесты:
- Блины — одно из самых древнейших изделий русской кухни -> есть ли в предложении слово “блины”?
- [Блины] — одно из самых древнейших изделий русской кухни -> NOUN
- Блины — одно из [самых древнейших изделий русской кухни] -> NP



# BERT и синтаксис

Существует преобразование из пространства векторных представлений слов в синтаксическое дерево



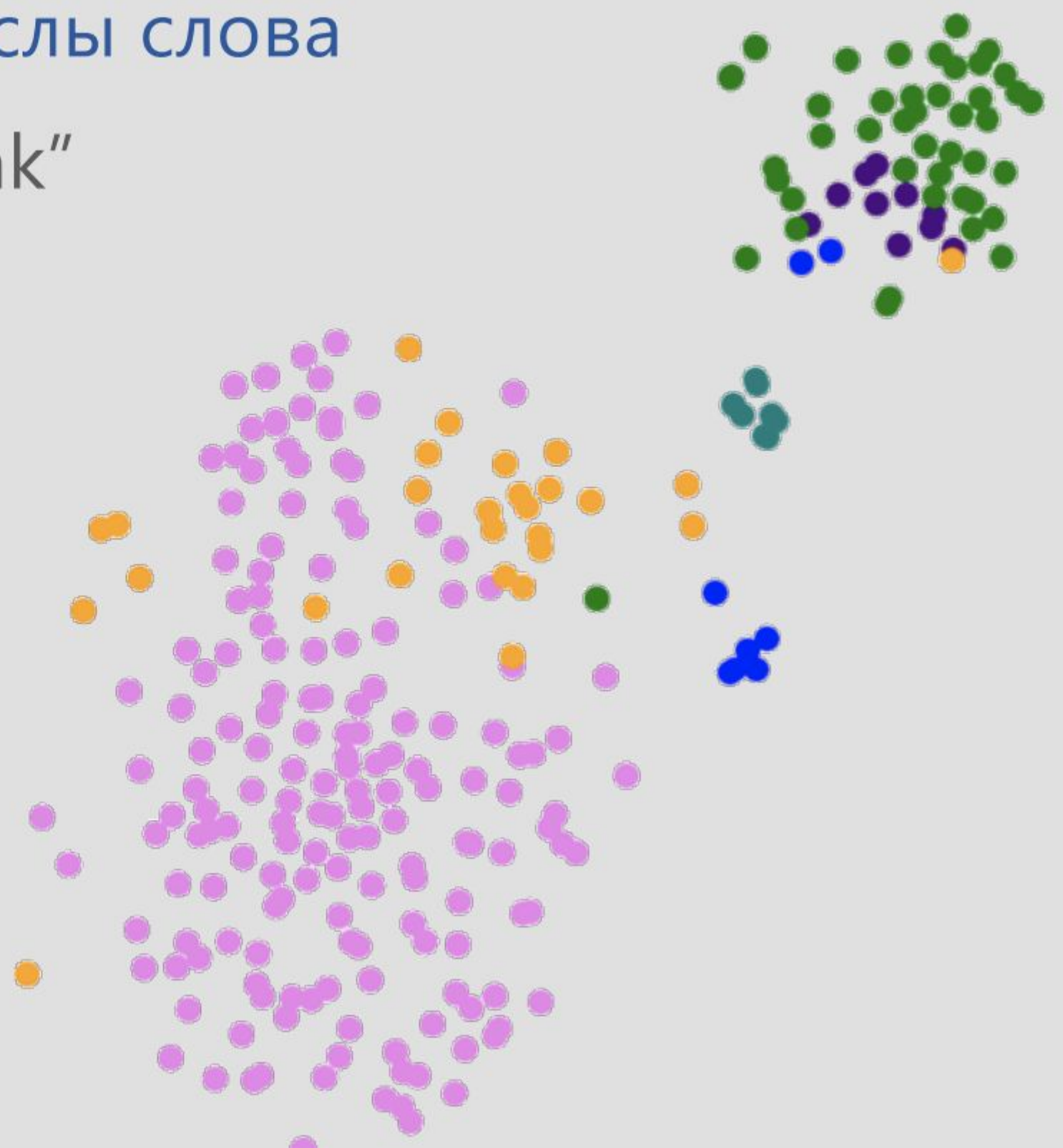


# BERT и семантика

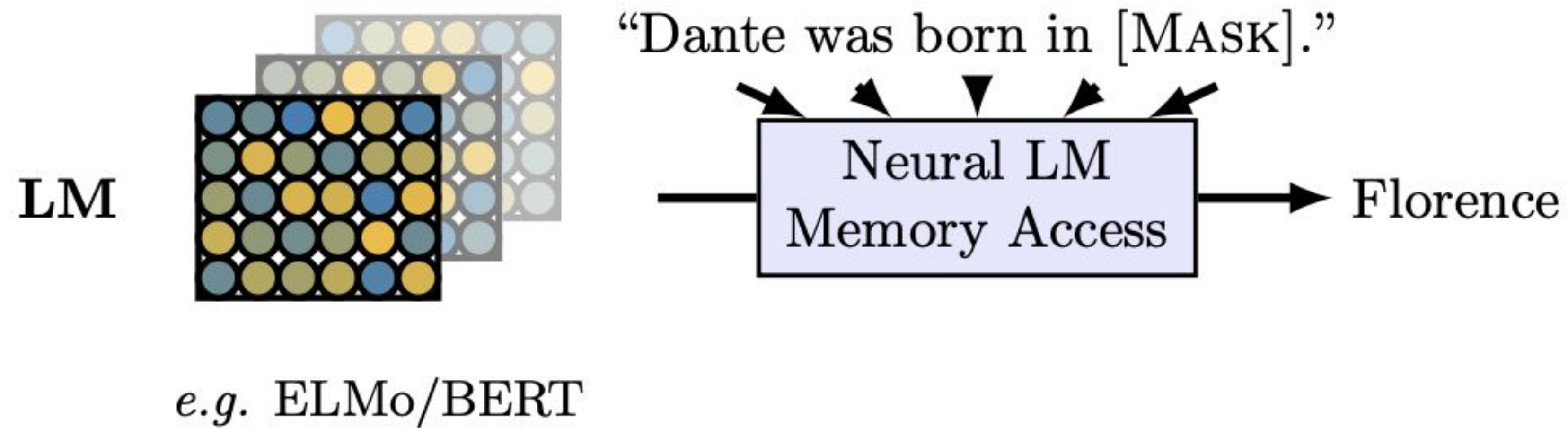
BERT различает разные смыслы слова

- Разные смыслы слова "bank"

- **A Financial Institution:175**
- **Sloping Land:46**
- **A Bank Building:27**
- **A Long Ridge:11**
- **Arrangement of Objects:8**
- **A Flight Maneuver:8**



# BERT и знание о мире





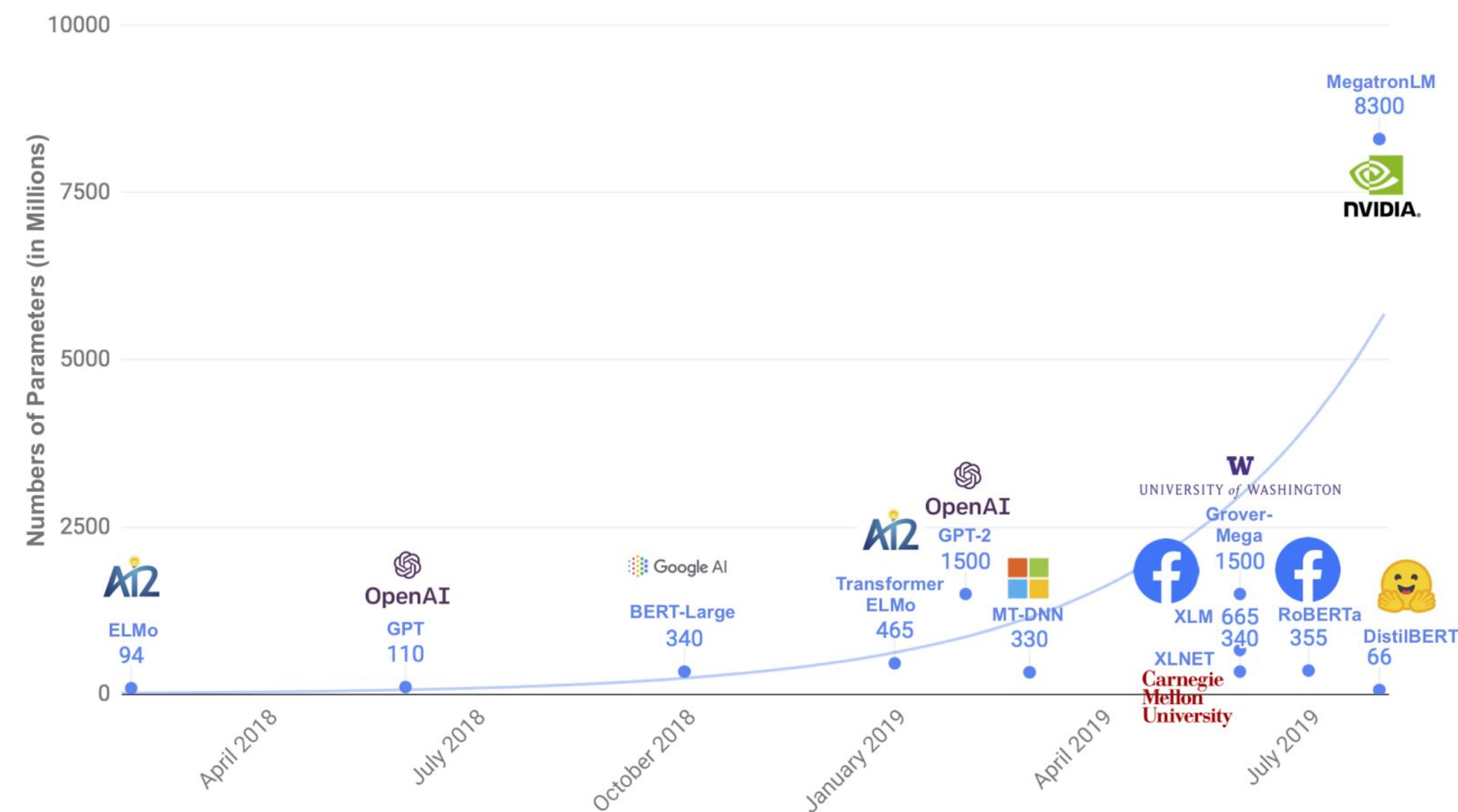
# BERT и уровни языка

- Нижние слои BERT лучше понимают порядок слов и поверхностную информацию
- Средние слои BERT лучше понимают морфологию и синтаксическую структуру предложения
- Верхние слои BERT отвечают за решение конкретной задачи
- За понимание семантики отвечает вся модель



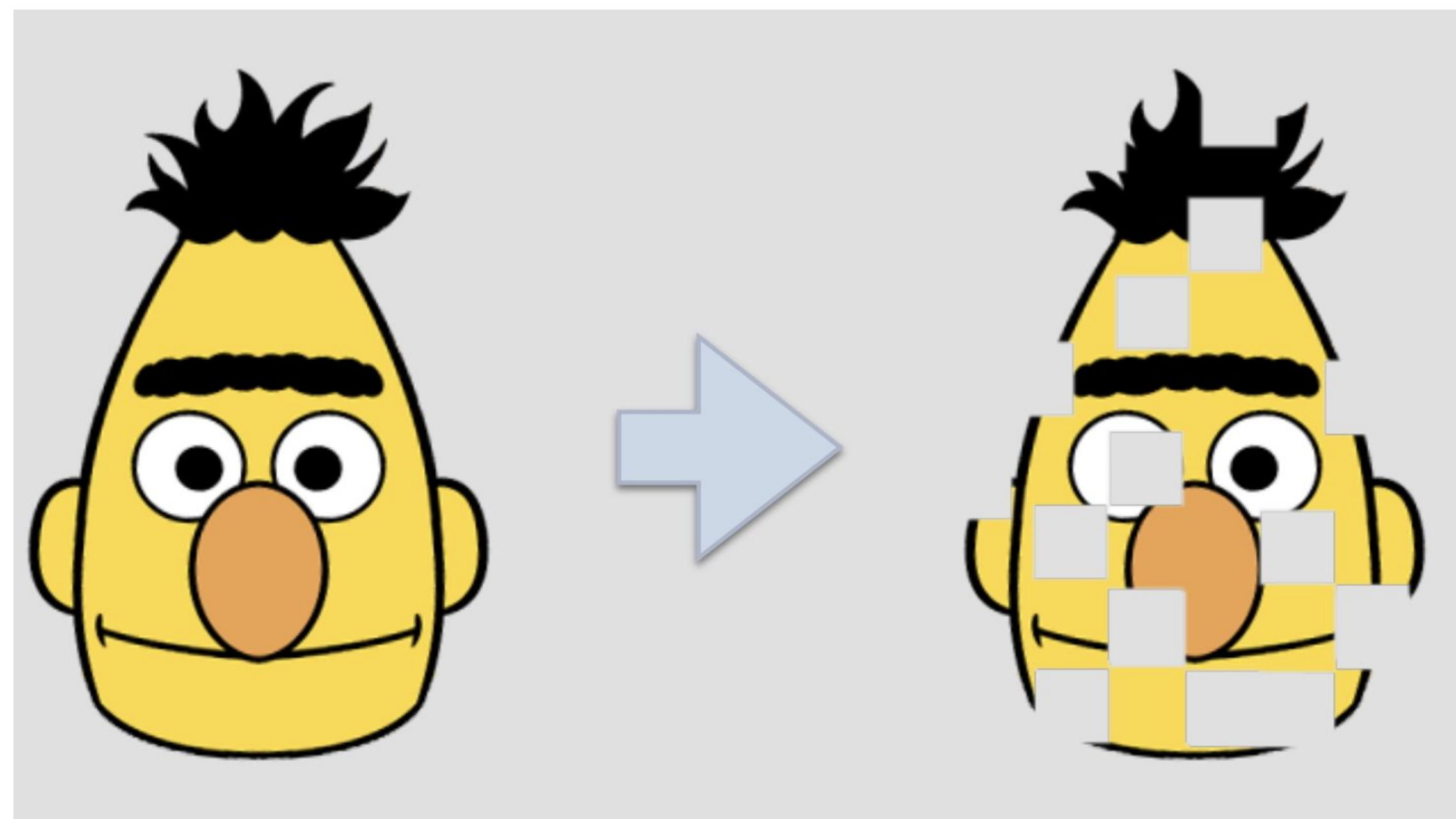
# Малышка BERT

- Техники уменьшения размера моделей:
  - Удаление весов [weights pruning]
  - Дистилляция [distillation]



# Прунинг

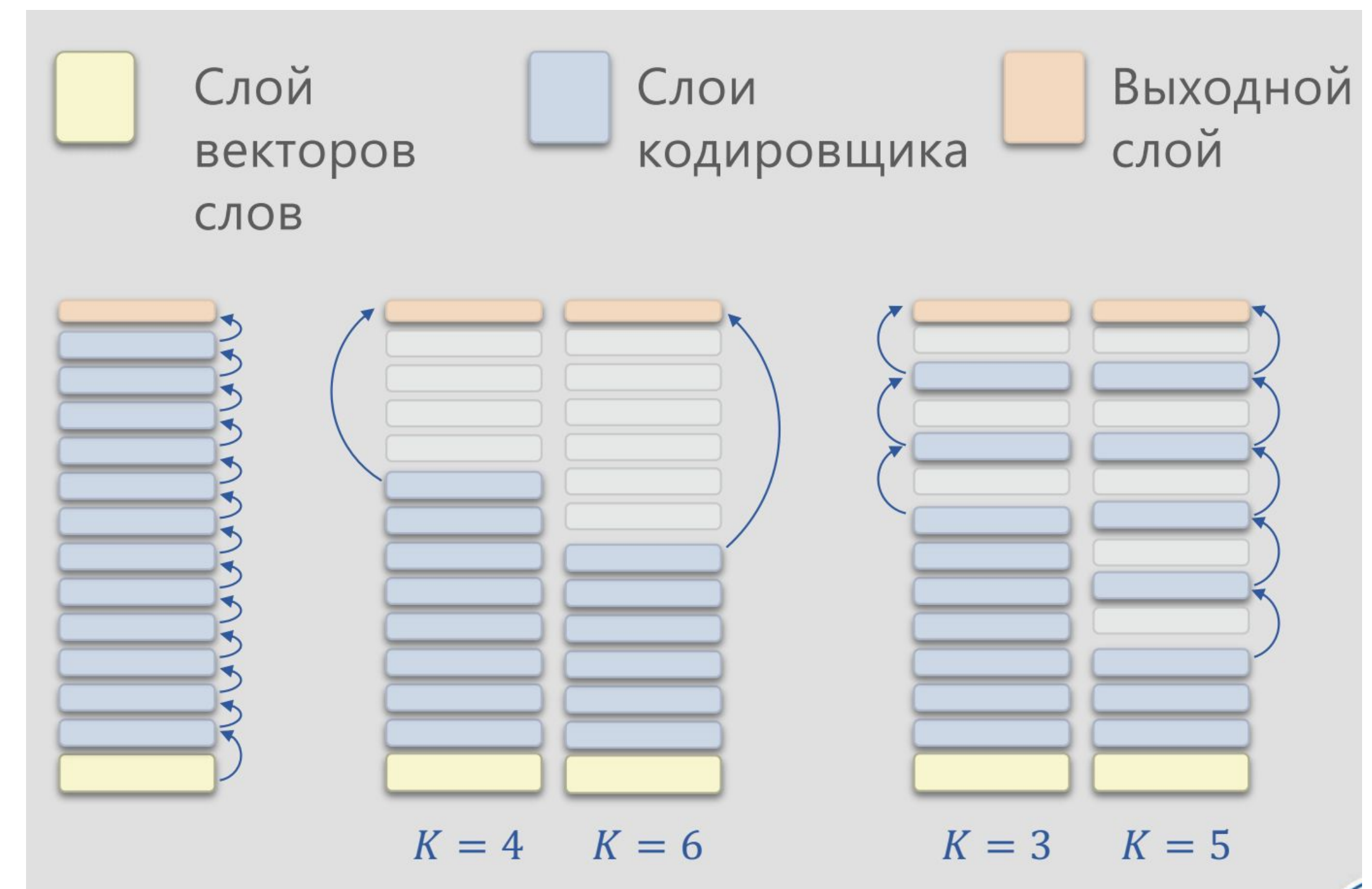
- Poor Man's BERT  
<https://arxiv.org/abs/2004.03844>
- Удаляется порядка 40% весов с использованием разных стратегий
- Модель после удаления весов тюнится на целевую задачу
- Точность сохраняется на уровне исходной модели (98% от показателей на GLUE)





# Прунинг

- Poor Man's BERT
- Удаляется порядка 40% весов с использованием разных стратегий
- Модель после удаления весов дообучается на целевую задачу
- Точность сохраняется на уровне исходной модели (98% от показателей на GLUE)





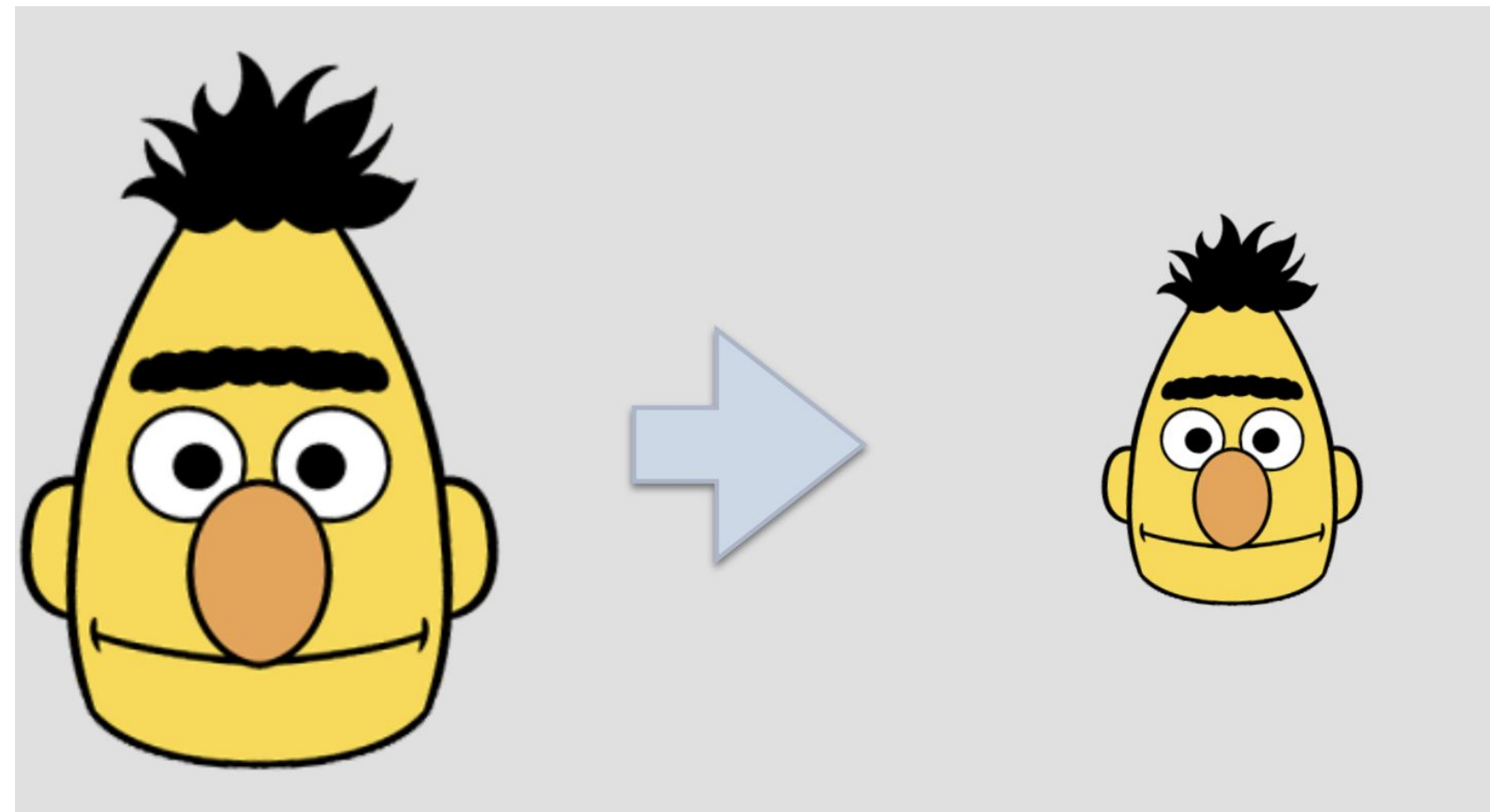
# Одна голова хорошо, а две лучше?

- В режиме тестирования большинство голов механизма внимания избыточны и без них модель не теряет в качестве
- Можно оставить одну голову и практически не потерять в качестве
- Это значит, что на каждом слое существуют головы, которую выполняют всю работу модели
- Удаление 50% голов ускоряет модель на 17%



# Дистилляция

- Скорость, память или качество?
- Можем ли мы сохранить качество, сократив вычислительные затраты?



# Дистилляция

- Шаг 1: Обучаем большую модель (или берем предобученную) – учитель
- Шаг 2: Берем ребенка этой модели (student) и обучаем его воспроизводить поведение учителя
- При обучении используем разные objectives, включая MLM







# Дистиляция

- DistilBERT <https://arxiv.org/pdf/1910.01108.pdf>
- Это работает
- Дистилированные модели сравнимы по качеству с учителями
- Такие модели могут давать нижнюю оценку

Table 2: **DistilBERT yields to comparable performance on downstream tasks.** Comparison on downstream tasks: IMDb (test accuracy) and SQuAD 1.1 (EM/F1 on dev set). D: with a second step of distillation during fine-tuning.

Model	IMDb (acc.)	SQuAD (EM/F1)
BERT-base	93.46	81.2/88.5
DistilBERT	92.82	77.7/85.8
DistilBERT (D)	-	79.1/86.9

Table 3: **DistilBERT is significantly smaller while being constantly faster.** Inference time of a full pass of GLUE task STS-B (sentiment analysis) on CPU with a batch size of 1.

Model	# param. (Millions)	Inf. time (seconds)
ELMo	180	895
BERT-base	110	668
DistilBERT	66	410





НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
УНИВЕРСИТЕТ

# Графовые методы в NLP

# Графы знаний

- Широкое применение
- В области ОЕЯ: алгоритмы и структуры данных, вопросно-ответные системы, датасеты, архитектуры моделей



**Alexander Pushkin (Александр Пушкин)** 

Russian poet

Alexander Sergeyevich Pushkin was a Russian poet, playwright, and novelist of the Romantic era. He is considered by many to be the greatest Russian poet, and the founder of modern Russian literature. Pushkin was born into Russian nobility in Moscow.  
[Wikipedia](#)

**Born:** June 6, 1799, [Moscow, Russia](#)

**Died:** February 10, 1837, [Saint Petersburg, Russia](#)

**Books:** [The Captain's Daughter](#), [The Queen of Spades](#), [MORE](#)

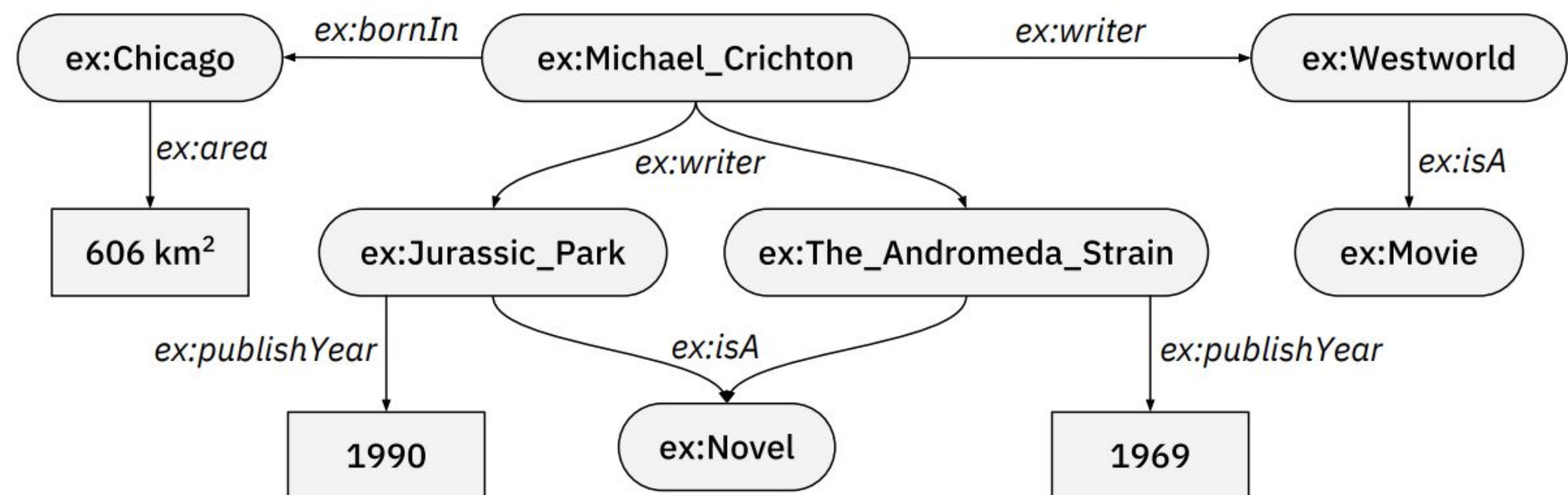
**Plays:** [Eugene Onegin](#), [Boris Godunov](#), [Mozart and Salieri](#), [MORE](#)

**Children:** [Maria Pushkina](#), [Alexander Pushkin](#), [Grigory Pushkin](#), [Natalya Pushkina](#)



# Графы знаний

- Широкое применение
- В области ОЕЯ: алгоритмы и структуры данных, вопросно-ответные системы, датасеты, архитектуры моделей



# ERNIE

- Инкорпорирование графовой информации в контекстуализированные представления
- Архитектура BERT
- Маскирование с использованием графа знаний
- Лучше в задачах извлечения информации





# ERNIE

## Обычное маскирование

Гарри [mask] – серия романов [mask]  
писательницы [mask] К. Роулинг

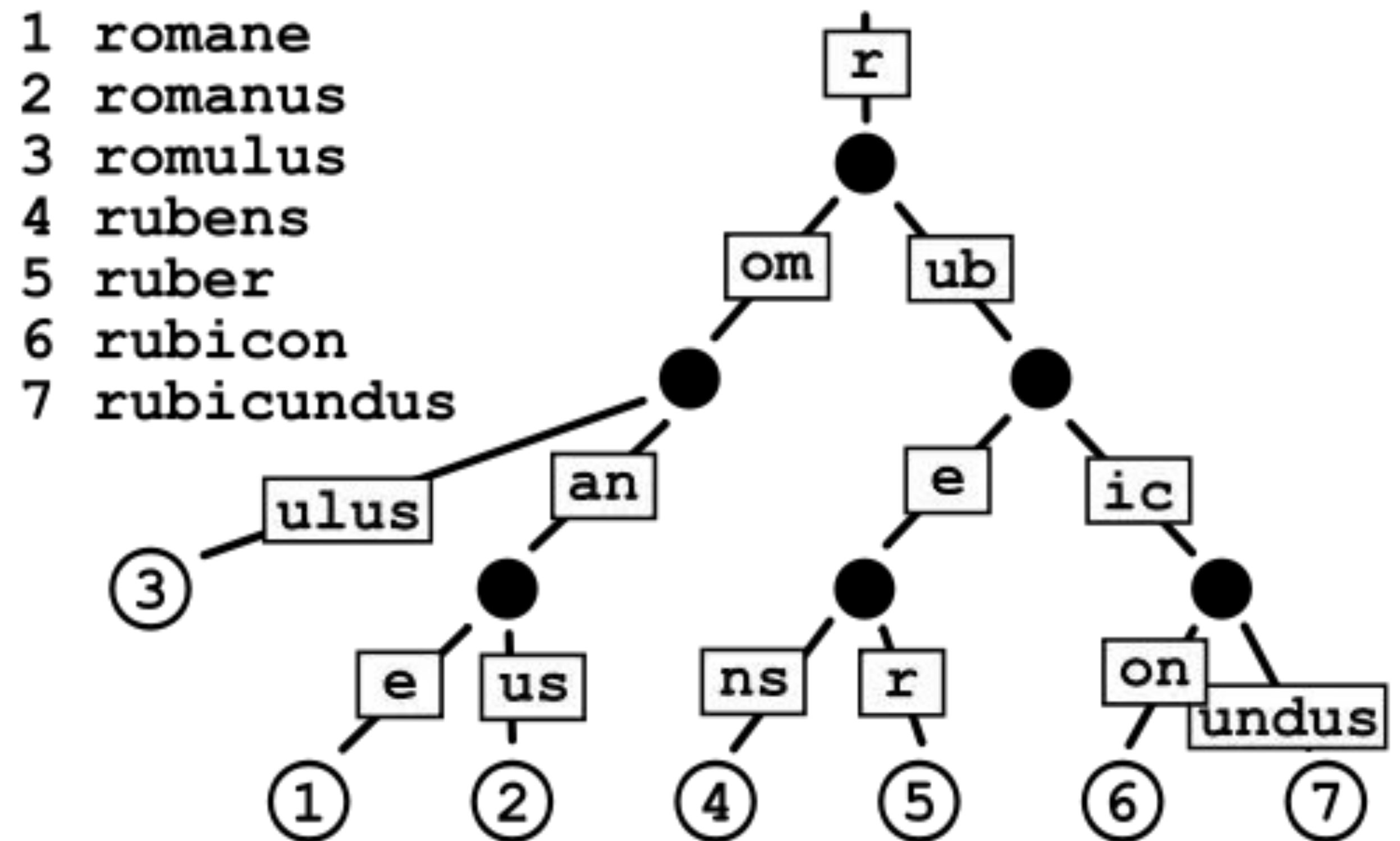
## Маски на месте именованных сущностей и случайных слов

[mask] [mask] – серия [mask] [mask]  
писательницы Дж. К. Роулинг



# Suffix Tree

- Исправление опечаток
- Быстрые алгоритмы
- <https://github.com/wolfgarbe/SymSpell>





# Задача KGQA

- Очень сложная
- Формальный язык
- Многокомпонентная
- Недостатки графов

What is the birthplace of Westworld's writer?

```
SELECT ?uri WHERE {  
  ?x ex:writer ex:Westworld.  
  ?x ex:bornIn ?uri  
}
```

# QA

- Multi-hop QA

<https://arxiv.org/pdf/1710.06481.pdf>

- Hybrid QA

<https://www.aclweb.org/anthology/D18-1455/>

5/

- KGQA <https://arxiv.org/pdf/1907.09361.pdf>

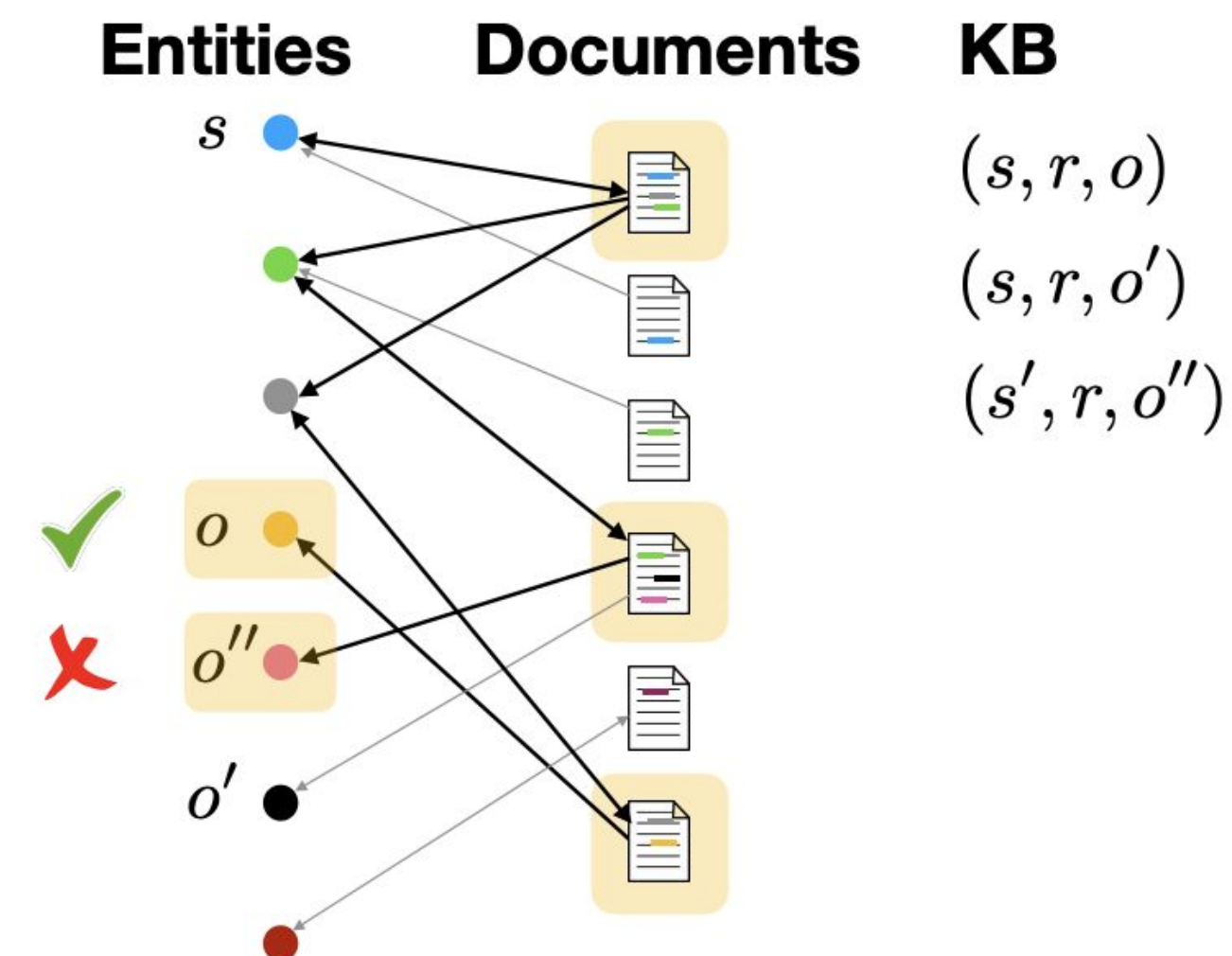


Figure 2: A bipartite graph connecting entities and documents mentioning them. Bold edges are those traversed for the first fact in the small KB on the right; yellow highlighting indicates documents in  $S_q$  and candidates in  $C_q$ .



# QA

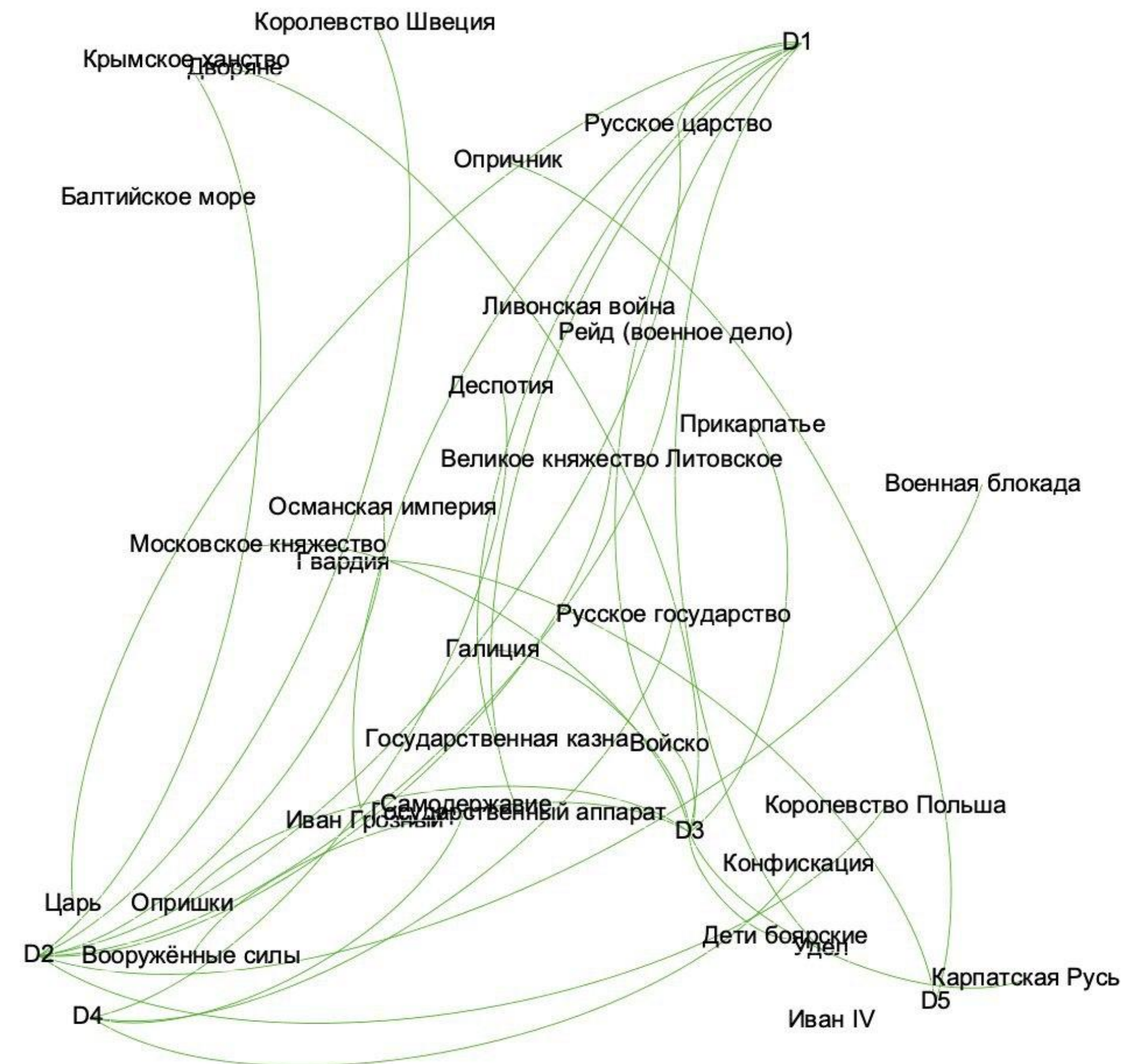
- Multi-hop QA

<https://arxiv.org/pdf/1710.06481.pdf>

- Hybrid QA

<https://www.aclweb.org/anthology/D18-1455/>

- KGQA <https://arxiv.org/pdf/1907.09361.pdf>



# QA

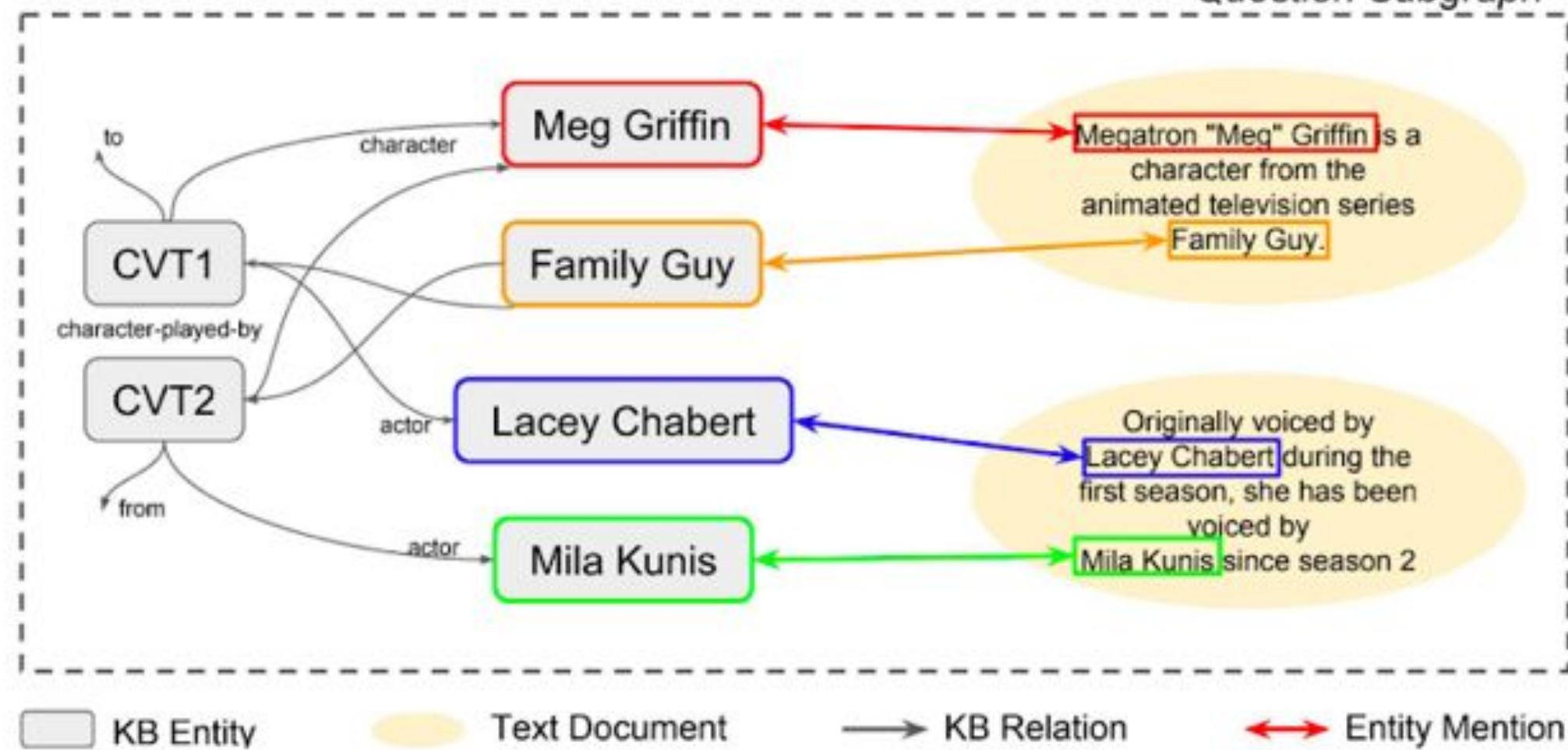
Q. Who voiced Meg in Family Guy?

Freebase

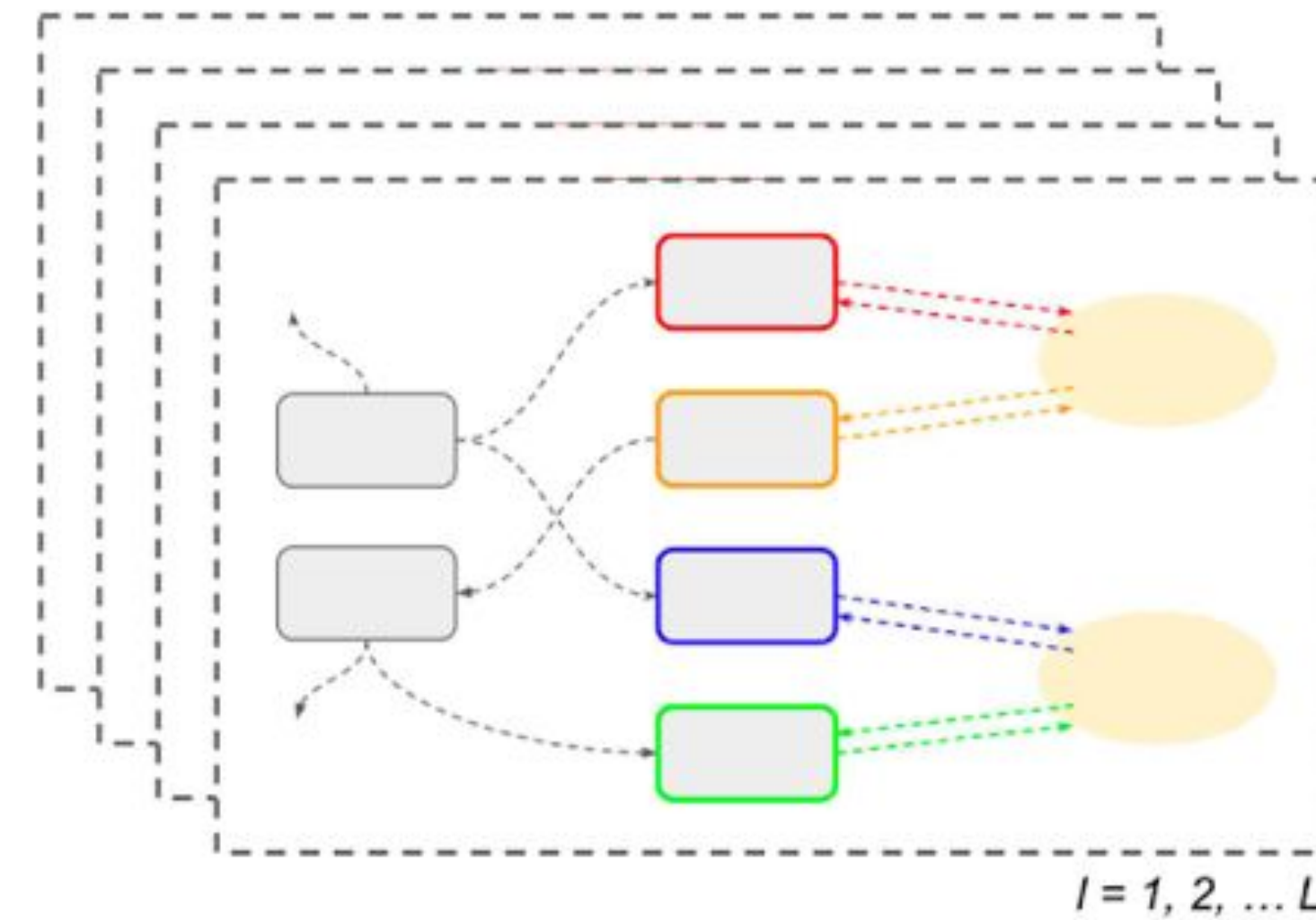


WIKIPEDIA  
The Free Encyclopedia

Question Subgraph



A. Lacey Chabert, Mila Kunis





# Классификация

- Строим граф, вершины которого – документы и слова в документах
- Ребра слово-документ взвешены tf-idf
- Ребра слово-слово взвешены PMI
- 2-layer GCN

