Концепт предобученных языковых моделей в контексте инженерии знаний

Д.И. Понкин

Аннотация **– Статья посвящена исследованию концепта и технологий предобученных языковых моделей в контексте инженерии знаний. Автор обосновывает актуальность вопроса о содержании в предобученных языковых моделях имплицитных, интернализированных знаний, извлеченных из текстовых корпусов, использовавшихся для предобучения (pretraining) или переноса обучения (transfer learning) языковых моделей. В работе дан развернутый обзор существующих подходов к интерпретации указанного концепта – в разных исследовательско-интерпретационных проекциях. Автором рассмотрен ряд новейших исследований, связанных с методами предобучения и переноса обучения языковых моделей. В статье также рассматриваются новейшие исследования на темы аугментации языковых моделей внешними знаниями. Кроме того, рассматриваются исследования на тему использования предобученных языковых моделей для поиска и извлечения неструктурированных знаний, использования таких моделей как вспомогательных инструментов в процессе построения баз знаний, а также их использования в качестве самостоятельных баз знаний. Объясняется содержание понятия «предобученные языковые модели». Автором приводятся референтные примеры реализации предобученных языковых моделей на практике, в том числе обсуждается вопрос применения языковых моделей в качестве баз знаний. Затрагивается суть концепта предобучения языковых моделей без учителя на крупных и неструктурированных текстовых корпусах перед дальнейшим дообучением модели под конкретную задачу (тонкой настройкой) – «transfer learning». Автором рассматривается понятие «граф знаний», ныне широко используемое как в целом, так и в контексте релевантных настоящей статье тем, а также ряд новейших исследований в области предобучения и переноса обучения языковых моделей.**

Ключевые слова **— предобученные языковые модели, трансформеры, базы знаний, обработка естественного языка, инженерия знаний.**

I. Введение

С появлением предобученной языковой модели BERT [1] область обработки текстов на естественном языке претерпела резкие изменения. Предобученные модели последовательно демонстрировали существенно более высокие результаты в решении всевозможных релевантных задач. С. Рудер и др. сравнивают возникновение множества новых методов предобучения и переноса обучения с такими важными событиями в новейшей истории машинного обучения, как появление Word2Vec и ImageNet [2, с. 15].

С 2019 г. активно изучаемым вопросом стало содержание в предобученных языковых моделях интернализированных знаний, извлеченных из текстовых корпусов, использовавшихся для предобучения (pretraining) или переноса обучения (transfer learning). В этом контексте возник значительный исследовательский интерес как к использованию предобученных языковых моделей в задачах поиска, извлечения и формализации знаний, так и к их эксплицитной аугментации дополнительными знаниями для решения определенных категорий задач. Кроме того, как продемонстрировано в настоящем материале, ряд авторов поднимают вопрос об использовании предобученных языковых моделей в качестве альтернативы традиционным базам знаний, исходя из обоснованных предположений о способностях этих моделей, во-первых, к захвату (извлечению) знаний, и, во-вторых, к рассуждениям.

II. Предобученные языковые модели

За последние несколько лет значительно распространилась практика предобучения языковых моделей без учителя на крупных и неструктурированных текстовых корпусах перед дальнейшим дообучением модели под конкретную задачу (тонкой настройкой) – «transfer learning» [3, с. 1; 4, с. 1]. Как утверждают А. Робертс и др., предобучение обеспечивает языковые модели некоторой степенью полезной осведомленности о смысле, синтаксисе и «словесных знаниях», с чем связан эмпирический успех предобучения во многих задачах обработки текста на естественном языке [5, с. 2]. К.Рафл и др., в свою очередь, пишут, что предобучение без учителя приводит к тому, что языковая модель в какой-то степени обретает универсальные, неспециализированные навыки и знания, которые могут быть затем «перенесены» на более узкие задачи [3, с.1]. К. Гуу и др. [6, с. 1] тоже указывают на то, что предобучение языковых моделей позволяет захватывать удивительное количество знаний о мире, что играет значительную роль в решении ряда задач обработки текста на естественном языке.

Статья получена 12 августа 2020 г.

Д.И. Понкин – Институт автоматики и вычислительной техники Национального исследовательского университета «МЭИ» (Московский энергетический институт), аспирант кафедры прикладной математики и искусственного интеллекта (e-mail: PonkinDI@mpei.ru).

В 2018 году М. Питерс и др. предложили понятие глубокого контекстуализированного векторного представления слов [7], которое моделирует как комплексные характеристики использования слова (синтакс и семантика), так и то, как они меняются в зависимости от лингвистического контекста. Для этого использовались векторы, произведенные двунаправленной LSTM – ELMo (Embeddings from Language Models).

З. Бурауи и др. [8] отмечают, что языковые модели, построенные с использованием векторных представлений, при обучении захватывают какую-то меру знаний об отношениях между объектами, несмотря на то, что задачей этого обучения является захват схожести слов.

В 2018 году сотрудники Google AI Дж. Девлин и др. представили BERT (Bidirectional Encoder Representations from Transformers) – модель представления естественного языка. [1] В отличие от других подобных моделей, BERT предназначен для предобучения двунаправленных представлений на неразмеченном корпусе путем обучения на контексте и слева, и справа. Под двунаправленностью подразумевается обучение предсказанию токенов (слов) в зависимости и от префикса, и от суффикса, окружающих маскированное слово [9, с. 2].

Авторы BERT утверждают, что, в отличие от однонаправленных, двунаправленные модели предлагают значительно более широкие возможности эффективного предобучения, что приводит к повышению эффективности языковых моделей в решении таких задач, как извлечение знаний, перефразирование, выделение именованных сущностей и формирование ответов на вопросы. В то время, как М. Питерс и др. в своей работе [7] используют неглубокое объединение независимо обученных слева направо и справа налево языковых моделей, BERT значительно отличается от других подобных подходов применением маскированных (т.е. использующих маску) языковых моделей для достижения двунаправленности.

Модель BERT привлекла значительный научный интерес, что подтверждается существованием множества производных от нее моделей, целый ряд которых упоминается и в настоящей статье.

Согласно И. Лю и др. [10], предобучение языковых моделей вычислительно затратно, такое обучение зачастую осуществляется на частных (приватных) наборах данных, а гиперпараметры оказывают значительное влияние на конечные результаты, из-за чего объективное сравнение различных подходов оказывается весьма затруднительным. Авторы попытались точно воссоздать предобученную языковую модель BERT, тщательно измеряя влияние множества ключевых гиперпараметров и объемов обучающих данных. В результате проведенной работы, И. Лю и др. обнаружили, что модель BERT в ее исходном исполнении в значительной степени недостаточно обучена, и что она еще способна достичь или преодолеть производительность множества подобных моделей, разработанных позже. Авторы предлагают улучшенную методику обучения модели BERT – RoBERTa, способную конкурировать с наиболее современными предобученными языковыми моделями.

Параллельно перечисленным разработкам, К. Рафл и др. [3] предложили рассматривать все задачи обработки текста на естественном языке как задачу «перевода текста в текст», то есть, использование входного текста для вывода (генерирования) какого-то иного текста. Авторы рассмотрели производительность такого подхода на целом ряде задач обработки текста на естественном (в данном случае – английском) языке, включая ответы на вопросы, уплотнение текста, анализ тональности и прочие. В своей работе, авторы предложили специально подготовленный набор данных – Colossal Clean Crawled Corpus (C4) и модель Text-to-Text Transfer Transformer (T5). Авторы отмечают, что их подход предоставляет простой способ обучения единой модели на широком наборе текстовых задач используя единую функцию потерь. Полученная модель показала передовые результаты во множестве задач обработки текста на естественном языке, однако, в ряде случаев такие результаты были достигнуты лишь наиболее масштабным вариантом модели с 11 млрд. параметров.

Как подчеркивают А. Рэдфорд и др. [11], большинство методов глубокого обучения требуют существенного количества вручную размеченных данных, что ограничивает применимость таких методов во множестве сфер, испытывающих недостаток размеченных данных. Авторы считают, что, в таких ситуациях, ценной альтернативой расширению объемов размеченных данных (что представляется дорогостоящим и времязатратным процессом) являются модели, способные извлекать лингвистическую информацию из неразмеченных наборов данных. Тем не менее, авторы утверждают о том, что извлечение более глубокой (за пределами словестного уровня) информации является нетривиальной задачей. Для ее решения А. Рэдфорд и др. предлагают использовать комбинацию генеративного предобучения (generative pre-training – GPT) языковых моделей на разнообразном корпусе неразмеченного текста и последующей дискриминативной тонкой настройки под каждую конкретную задачу. Авторы демонстрируют эффективность предлагаемого подхода в широком ряде испытаний, где ими были достигнуты показатели, существенно превышающие таковые у других подходов.

А. Рэдфорд и др. продолжают развивать свой подход и в последующих публикациях [12], стремясь к универсальной языковой модели, способной выполнять множество задач обработки текста на естественном языке, обучаясь на неразмеченных наборах данных и без учителя. Используя самую крупную из построенных в рамках проведенного исследования языковую модель GPT-2 (1.5 млрд. параметров), А. Рэдфорд и др. достигли передовых результатов в большинстве рассматриваемых задач, отмечая, что успех такого подхода к переносу обучения прямо пропорционален емкости языковой модели. Резюмируя полученные результаты, авторы заявляют о том, что предложенный подход может быть многообещающим путем к построению систем обработки текста на естественном языке, способных обучаться выполнять задачи на их естественно возникающих проявлениях.

Ч. Ян и др. [13] утверждают, что, из-за использования намеренного «повреждения» входных данных масками, предобученная языковая модель BERT

пренебрегает зависимостью между позициями маски. Кроме того, как утверждают авторы, модель BERT использует искусственные специальные символы вроде «[MASK]» для предобучения, однако эти символы не встречаются в реальных данных во время тонкой настройки, что приводит к возникновению расхождений в результатах. В контексте установленных авторами несовершенств предобученной языковой модели BERT, Ч. Ян и др. предлагают XLNet – обобщенный авторегрессивный метод предобучения, способный запоминать двунаправленные контексты. Авторы утверждают, что XLNet за счет своей природы не страдает от недостатков BERT, а также заимствует некоторые идеи из авторегрессивной модели Transformer-XL, что позволяет ему превосходить предобученную языковую модель BERT в 20 задачах обработки текстов на естественном языке и достигать передовых результатов в 18 из них, включая ответы на вопросы, формирование рассуждений на естественном языке, анализ тональности и ранжирование документов.

К. Ричардсон и А. Сабхарвал [14] рассматривают производительность языковых моделей, основанных на трансформере (в первую очередь – BERT), в задачах генерирования ответов на вопросы. В частности, авторы задаются вопросом о том, действительно ли такие модели обладают общими знаниями об определениях слов и действительно ли они способны к общим классификационным рассуждениям. Основываясь на результатах проведенного исследования, К. Ричардсон и А.Сабхарвал утверждают, что основанные на трансформере модели обладают замечательной способностью к распознаванию определенных типов лексических знаний. Однако, авторы также установили, что производительность таких моделей существенно падает с увеличением глубины таксономии (т.е. в случаях с гипонимами).

III. Графы знаний

Как пишут Л. Эрлингер и В. Весс, понятие «граф знаний» вошедшее в обиход с выходом Google Knowledge Graph в 2012 г., ныне широко и часто используется в науке и бизнесе, зачастую в контексте семантической паутины (Semantic Web), связанных данных, анализа больших данных и облачных вычислений. Однако, как в 2016 г. отмечали Л. Эрлингер и В. Весс, понятие «граф знаний» страдало от полного отсутствия какого-либо общепризнанного и исчерпывающего определения. Исследовав массив публикаций, связанных с этим понятием, авторы предложили следующее определение: «граф знаний получает и интегрирует информацию в онтологию и применяет механизм рассуждений для вывода новых знаний» [15, с.3]. С. Ю и О. Чон, ссылаясь на это определение, полагают, что граф знаний можно считать понятием, находящимся на уровень выше понятия базы знаний [16, с. 1].

Н. Чжан и др. считают, что «графы знаний» структурировано организуют факты в триплеты вида «субъект – предикат – объект», обозначаемые как (*s,p,o*), где под «s» и «о» подразумеваются сущности, а «р» строит между ними отношения. Кроме того, авторы отмечают, что за последние годы было построено множество крупномасштабных графов знаний, что привело к их повсеместному прикладному использованию в ряде задач, в число которых входит генерирование ответов на вопросы, понимание естественного языка, анализ связанных данных, разработка рекомендательных систем и др. [17, с. 1]

Согласно Ц. Вэн и др., Л. Яо и др. «граф знаний» – это новый, динамически взаимосвязанный метод представления знаний, выражающий сущности и отношения между сущностями с помощью вершин и ребер [18, с. 25] [19, с. 1].

С. Лю и др. утверждают, что понятие «граф знаний» ныне преимущественно используется для описания баз знаний семантической паутины, то есть, основанных на стандарте схемы описания ресурсов (Resource Description Framework, RDF) представлений некоторых широких предметных областей [20, с. 343]. Авторы также утверждают, что за последние годы было построено (с помощью краудсорсинга или извлечения содержимого из сети Интернет) множество крупномасштабных графов знаний, включая DBpedia, YAGO, Freebase и Wikidata [20, с. 344].

Согласно Л. Яо и др., крупномасштабные графы знаний (FreeBase, YAGO, WordNet и др.) служат эффективным подспорьем для решения целого ряда важных задач в области искусственного интеллекта, включая семантический поиск, поддержку принятия решений и генерирование ответов на вопросы [19, с. 1].

С. Цзи и др., в свою очередь, пишут, что графы знаний используются для представления структурных отношений между сущностями, а само понятие «граф знаний» можно считать синонимом «базы знаний» [21, с. 1].

IV. Аугментация языковых моделей знаниями

Р. Логан и др. [22] утверждают, что, несмотря на то, что языковые модели способны генерировать грамматически корректные предложения, что они обладают какой-то мерой способности к рассуждениям на основе здравого смысла и что они содержат какое-то количество знаний, их способность генерировать фактически корректные предложения является весьма скромной. Авторы полагают, что, в этом контексте, основным ограничением существующих языковых моделей является то, что они, в лучшем случае, могут запоминать только факты, наблюдаемые во время обучения, следовательно, эти языковые модели плохо приспособлены к генерированию фактически корректных предложений и не способны к обобщению на неизвестные или редко встречающиеся объекты. В свете перечисленных проблем, Р. Логан и др. предлагают Knowledge Graph Language Model (KGLM) – нейронную языковую модель, аугментированную механизмами выбора и копирования информации из внешнего графа знаний. Авторы также описывают KGLM как языковую модель, способную обращаться к внешнему структурированному источнику фактов, представленному в виде графа знаний, для генерирования фактически корректного текста. Рассматривая результаты своей работы, авторы заявляют, что языковая модель KGLM, используя внешний источник знаний, способна генерировать

высококачественный, фактически корректный текст на естественном языке, содержащий упоминания редких объектов и специфических токенов, вроде чисел и дат. Тем не менее, стоит отметить, что KGLM, в отличие от других новейших языковых моделей, требует размеченного набора обучающих данных.

К. Гуу и др. [6] пишут, что, несмотря на то, как предобучение языковых моделей позволяет захватывать значительное количество знаний о мире, эти знания имплицитно содержатся в параметрах нейронной сети, что затрудняет установление того, какие знания хранятся в сети и где они хранятся. Кроме того, как подчеркивают авторы, объем «памяти» для хранения таких знаний ограничен размером нейронной сети, из чего следует, что для захвата большего количества знаний о мире, потребуется обучить еще большую сеть, что может оказаться чрезмерно медленным или затратным. Для решения этих проблем, К. Гуу и др. предлагают методику предобучения Retrieval-Augmented Language Model (REALM), аугментирующую алгоритмы предобучения языковых моделей обученной системой поиска текстовых знаний. Как утверждают авторы, в отличие от моделей, которые содержат знания в своих параметрах, их подход эксплицитно выявляет роль знаний, так как модели требуется решить, какие знания ей потребуются для рассуждений. Перед тем, как делать каждое предсказание, языковая модель использует систему поиска для обращения к документам из крупного текстового корпуса (например Wikipedia) и использует их текст в выводе предсказаний. В результате проведенных испытаний, авторы утверждают, что такой подход в задачах ответов на вопросы (Open-QA) позволяет достичь показателей, превосходящих таковые у других современных моделей. В частности, авторы особенно подчеркнули превосходство их подхода над моделью Text-to-Text Transfer Transformer (T5) в наиболее масштабном ее исполнении (11 млрд. параметров), будучи в 30 раз меньше. Авторы выражают интерес в дальнейшем развитии подхода и, кроме всего прочего, указывают на потенциальную пользу от использования REALM в многоязыковом контексте.

Р. Ван и др. [23] утверждают, что, несмотря на значительный успех предобученных языковых моделей в эмпирических исследованиях, такие модели, будучи предобученными без учителя, не справляются с захватом обширного количества знаний. Кроме того, авторы подчеркивают трудности, связанные с «внедрением» многообразных знаний в единую предобученную модель с помощью изменения исходных параметров таких моделей, в частности, риск катастрофической забывчивости. Авторы предлагают K-Adapter – гибкий и простой подход, «внедряющий» знания в крупные предобученные языковые модели, сохраняя их исходные параметры. Это достигается за счет аугментации предобученных моделей т.н. «адаптерами» – подстроенные под определенные знания модели, на входной слой которых подаются состояния промежуточных слоев предобученной модели. Как подчеркивают авторы, K-Adapter обладает рядом привлекательных особенностей, включая поддержку непрерывного внедрения знаний. Демонстрируя предлагаемый подход, Р. Ван и др. рассматривают использование K-Adapter на предобученной модели RoBERTa, внедряя в нее два типа знаний: фактические знания, полученные из Wikipedia и Wikidata, а также лингвистические знания, полученные с использованием готового синтаксического анализатора на текстовом корпусе. В результате проведенных авторами испытаний выяснилось, что в ряде задач, включая классификацию отношений и генерирование ответов на вопросы, каждый из «адаптеров» привел к повышению производительности модели, еще более высокие показатели были достигнуты с применением комбинации из двух адаптеров. Кроме того, авторы установили, что K-Adapter позволяет захватывать более значительные объемы фактических знаний и знаний, основанных на здравом смысле, чем RoBERTa.

Б. Ян и Т. Митчелл [24] пишут о недостатках традиционных методов задействования баз знаний в улучшении производительности рекуррентных нейронных сетей в задачах машинного чтения, ссылаясь на низкую способность к обобщению признаков и на то, что, для достижения высокой производительности в каждой отдельной задаче, требуется конструирование признаков. Авторы предлагают KBLSTM – новая модель искусственной нейронной сети, применяющая непрерывные представления базы знаний для повышения эффективности обучения рекуррентных нейронных сетей в задачах машинного чтения. Описывая предлагаемый подход, Б. Ян и Т. Митчелл пишут об использовании моделью механизма внимания для адаптивного принятия решений об обращении к фоновым знаниям и определении того, что из базы знаний может быть полезным.

Согласно полученным авторами результатам, предлагаемая модель достигает высоких показателей в задачах извлечения объектов и событий на широко используемом наборе данных ACE2005.

М. Остендорф и др. [25] рассматривают задачу классификации текстов книг с применением предобученной языковой модели BERT. В процессе тонкой настройки модели под поставленные задачи, авторы демонстрируют метод комбинирования текстовых представлений с метаданными и эмбеддингами графов знаний, содержащими информацию об авторах. Авторы утверждают, что «обогащение» языковой модели BERT внешними знаниями из Wikidata и ряда других источников повышает производительность модели в задачах классификации в условиях относительной недостаточности (немногочисленности) обучающих данных. Основываясь на результатах ряда испытаний, М. Остендорф и др. заявляют, что аугментированная (или «обогащенная») внешними знаниями языковая модель BERT достигает лучших показателей в сравнении с базовыми (предобученная неаугментированная модель BERT) в задачах классификации текстов книг.

И. Ху и др. [26] утверждают, что предобученная языковая модель BERT, несмотря на ее высокую (даже превышающую таковую у человека) точность в задачах ответов на вопросы, неспособна запоминать знания, основанные на здравом смысле, и, следовательно применять их в таких задачах. Аргументируя этот тезис, авторы экспериментально установили, что

незначительная модификация вопросов способна существенно снизить показатели модели BERT, не влияя на таковые у людей. Для улучшения способностей языковых моделей к запоминанию знаний, основанных на здравом смысле, И. Ху и др. предлагают методику из трёх частей: построение графа повседневных знаний (базы знаний), основанной на датасете SQuAD, conceptNet и wordnet; построение на основе графа датасета повседневных знаний; и тонкая настройка с помощью полученного датасета предобученной языковой модели BERT с расчетом на задачи ответов на вопросы, требующих наличия знаний, основанных на здравом смысле. Производительность своего подхода авторы испытали на части единого экзамена Индии Common Admission Test (CAT) – тесте на понимание прочитанного. И. Ху и др. установили, что языковая модель BERT, дополненная повседневными знаниями, смогла достичь тех же результатов на втором уровне сложности теста, которых исходная модель BERT смогла достичь лишь на первом.

Б. Хэ и др. [27] пишут о том, что комплексные взаимоотношения между вершинами графов знаний содержат дополнительные обширные знания, но, несмотря на это, традиционные методы обучения на представлениях знаний (knowledge representation learning, KRL) ограничиваются триплетами, пренебрегая контекстуализированной информацией, содержащейся в вершинах графов знаний. Авторы же предлагают новый метод обучения на представлениях знаний, способный моделировать произвольные подграфы для использования (после трансформирования в последовательность вершин) в качестве обучающих примеров, что позволяет ему существенно увеличить количество информации, содержащейся в представлениях. Кроме того, Б. Хэ и др. предлагают BERT-MK (BERT-based language model with Medical Knowledge) – предобученную языковую модель BERT, тонко настроенную с помощью обучения на крупномасштабном медицинском корпусе и аугментированную медицинскими знаниями, с помощью представлений, построенных на основе авторского подхода. Испытав полученную модель, авторы приходят к выводу о том, что медицинские знания действительно являются полезными для решения некоторых задач обработки текстов на естественном языке в сфере медицины, что демонстрируется опережающими аналоги результатами, показанными BERT-MK, и выражают глубокую заинтересованность в продолжении работы над комбинированием медицинских знаний и языковых моделей.

Согласно В. Сюн и др. [28], предобученные языковые модели демонстрируют, кроме высокой эффективности в решении синтаксических и семантических задач в области обработки текстов на естественном языке, и положительные результаты в задачах, подразумевающих знания об окружающем мире, что говорит о том, что крупномасштабное моделирование естественных языков может быть имплицитным методом захвата знаний. Авторы развивают этот тезис и предлагают метод предобучения со «слабым» учителем (weakly supervised pretraining), эксплицитно заставляющий модель охватывать знания об объектах реального мира. В. Сюн и др. утверждают, что модели, предобученные с применением их метода, достигают существенно лучших показателей в задаче дополнения предложений о фактах. Исходя из результатов проведенных испытаний, авторы также заявляют, что, в более узких задачах, их модель последовательно превосходит BERT на четырех датасетах вопросов и ответов (WebQuestions, TriviaQA, SearchQA, Quasar-T).

V. Использование языковых моделей в задачах построения и наполнения баз знаний

С. Ван и др. [29] утверждают, что предобученные языковые модели вроде BERT и RoBERTa, несмотря на блестящие результаты в задачах обработки текста на естественном языке и способности к извлечению лингвистических знаний из неразмеченных текстовых корпусов, как правило, недостаточно способны к захвату фактов о мире. Авторы предлагают рассмотреть взаимную аугментацию языковых моделей с графами знаний, предлагая модель Knowledge Embedding and Pre-Trained LanguagE Representation (KEPLER). С. Ван и др. полагают, что, используя текстовые описания объектов как «мост» между методами эмбеддинга знаний и предобученными языковыми моделями, можно получить единую модель, в которой графы знаний обеспечивают предобученные языковые модели фактическими знаниями, а информативные текстовые данные могут повысить качественные характеристики этих знаний. В результате проведенных испытаний, авторы утверждают, что, благодаря такому совмещенному обучению, модель KEPLER лучше (в сравнении с неаугментированными предобученными языковыми моделями, вроде RoBERTa) разбирает связанный с фактическими данными текст и, соответственно, лучше извлекает из него знания. Авторы также подчеркивают, что полученные результаты указывают на необходимость дальнейших исследований в области интегрирования фактических знаний в предобученные языковые модели.

Согласно С. Ю и О. Чон [16], графы знаний имеют определенные недостатки, в частности то, что их масштаб и размер для некоторых языков весьма ограничен, а также то, что они неспособны работать с неологизмами. Для решения этих проблем, авторы предлагают PolarisX (Polaris Expander) – автоматически растущий граф знаний. Как утверждают авторы, автоматический рост обеспечивается сканированием новостных веб-сайтов и социальных сетей, извлечением новых отношений (из предопределенного множества), генерированием подграфов и их объединением с графами знаний. Кроме того, авторы пишут, что такой подход позволяет графу знаний работать с неологизмами. С. Ю и О. Чон для реализации своих предложений используют предобученную многоязыковую модель BERT, в частности, она используется для построения графов знаний, извлечения новых отношений и их интеграции в граф. Полученные хорошие результаты авторы, в частности, атрибутируют предобученной языковой модели BERT.

Согласно А. Босселю и др. [30], базы общеизвестных понятий, в отличие от традиционных баз знаний, содержат лишь слабоструктурированные текстовые

описания знаний, и важным шагом в их развитии является разработка порождающих (генеративных) моделей знаний, основанных на здравом смысле. В этом контексте, авторы предлагают Commonsense Transformers (COMET) – модель, способную обучиться генерированию обширных и разнообразных описаний на естественном языке. В своей работе А. Босселю и др. получили многообещающие результаты при переносе иплицитных знаний из предобученных языковых моделей для генерирования эксплицитных в графах знаний, основанных на здравом смысле. Используя предобученную языковую модель GPT, авторы продемонстрировали, что модель COMET способна генерировать новые высококачественные (с точки зрения человека) знания. Авторы также утверждают, что проведенная ими работа указывает на то, что использование порождающих (генеративных) моделей для автоматического заполнения баз общеизвестных понятий может в ближайшем будущем стать приемлемой альтернативой экстрактивным методам.

Дж. Фелдман и др. [31] утверждают, что, в связи с немногочисленностью размеченных данных для обучения, методы поиска и извлечения знаний, основанных на здравомсмысле, подразумевающие обучение сучителем, демонстрируют низкую производительность на новых данных. Авторы предлагают метод генерирования знаний, основанных на здравом смысле, с использованием масштабной, предобученной двунаправленной языковой модели. Переформулируя триплеты в маскированные предложения на естественном языке, эта модель может использоваться для оценки и ранжирования триплетов. Дж. Фелдман и др. демонстрируют, что, несмотря на низкие результаты на тестовой выборке, их метод превосходит аналоги в задачах поиска и извлечения знаний из новых источников. На основе полученных результатов, авторы делают вывод о том, что методы обучения без учителя позволяют достичь лучшей способности к обобщению.

Л. Яо и др. [19], ссылаясь на то, насколько далеки графы знаний от того, чтобы быть завершенными, рассуждают об актуальности задачи дополнения графов знаний, подразумевающей оценку правдоподобности триплетов, еще не внесенных в граф. В контексте этой задачи, авторы предлагают расценивать триплеты в графах знаний как текстовые последовательности и представляют Knowledge Graph Bidirectional Encoder Representations from Transformer (KG-BERT) – предобученная языковая модель BERT, тонко настроенная под задачу установления степени правдоподобности триплетов или отношения. Методика, разработанная Л. Яо и др., предполагает переформулирование объектов, отношений и триплетов в текстовые последовательности, что превращает задачу дополнения графа знаний в задачу классификации последовательностей. Авторы утверждают, что предлагаемый метод может достигать высокой производительности в ряде задач дополнения графов знаний.

Ч. Ван и др. [32] утверждают, что в поиске и извлечении знаний одной из наиболее сложных задач является точное установление лексических отношений между сущностями, ссылаясь на немногочисленность паттернов, указывающих на существование таких отношений. Авторы предлагают фреймворк Knowledge-Enriched Meta-Learning (KEML) для решения задач в области классификации лексических отношений, использующий, кроме всего прочего, предобученную языковую модель Lexical Knowledge Base-BERT (LKB-BERT), для извлечения представлений из крупных текстовых корпусов с применением внедрения лексических знаний методом обучения со «слабым» учителем (distant supervision). В результате проведения ряда испытаний, Ч. Ван и др. утверждают, что Knowledge-Enriched Meta-Learning (KEML) в задачах различения лексических отношений превосходит другие современные решения и подходы, и выражают заинтересованность в продолжении работы, в частности, в расширении KEML для решения задач в области обработки текста на естественном языке, включая формирование рассуждений.

Ц.Чжан и др.[33] используют предобученную языковую модель BERT для построения семантических эмбеддингов медицинской терминологии, которые применяются в авторском методе дополнения баз знаний о медицинской терминологии для установления относимости между сущностями и терминологией. Авторы отмечают, что их подход впервые использует предобученную языковую модель для «горячего запуска» эмбеддингов базы знаний, что позволят этому подходу опережать аналоги по ряду метрик. Кроме того, Ц. Чжан и др. считают, что в будущем развитии подобных методов следует рассматривать, кроме всего прочего, более продвинутые предобученные языковые модели.

Как утверждают Б.Д. Триседья и др. [34], крупные публично доступные базы знаний, такие, как DBpedia, Wikidata и Yago, содержащие миллионы фактов об объектах, представленных в виде триплетов «субъект – предикат – объект», весьма далеки от завершения и требуют непрерывного дополнения, обогащения и курирования. Используя этот тезис в качестве одного из обоснований, авторы рассматривают вопрос обогащения (дополнения) баз знаний с помощью извлечения отношений из текстовых источников и предлагают комплексную модель для извлечения и нормализации триплетов для обогащения баз знаний, использующую, кроме всего прочего, и векторные представления слов. Рассмотрев полученные результаты, Б. Д. Триседья и др. утверждают, что предложенная ими комплексная модель значительно и последовательно опережает существующие методы извлечения отношений из текстовых данных.

VI. Языковые модели как базы знаний

Как отметили Ф. Петрони и др. [35], получая лингвистические знания, языковые модели, предобученные на больших текстовых корпусах, могут также обретать и знания об отношениях, содержащиеся в обучающей выборке. Благодаря этому, такие модели могут быть способны отвечать на запросы, сформулированные в виде маскированных предложений (связного текста с пробелами). Авторы утверждают, что языковые модели имеют множество преимуществ перед структурированными базами знаний, приводя в пример

то, что они не требуют проектирования структуры, их легко дополнять данными и они не требуют разметки.

Рассматривая эти вопросы, Ф. Петрони и др. предложили оценить с этой точки зрения готовые двунаправленные языковые модели, вроде ELMo [7] и BERT [1]. В частности, авторы задают вопросы о том, сколько же знаний об отношениях такие модели содержат, насколько отличается их производительность от таковой у баз знаний, и насколько разнятся такие показатели для других типов знаний. Для того, чтобы дать ответы на все перечисленные вопросы, авторы предложили испытание LAMA (LAnguage Model Analysis), состоящее из набора источников знаний, каждый из которых представляет из себя набор фактов. Авторы предлагают допускать, что предобученная языковая модель обладает знанием о факте (субъект, отношение, объект, к примеру «Данте, родился-в, Флоренция»), если она успешно предсказывает маскированные объекты в предложениях с пробелами, выражающих этот факт (например, «Данте родился в ___»). В результате осуществленных испытаний, авторы установили, что модель BERT [1] содержит в себе знания об отношениях в количестве, сравнимом с базой знаний, построенной с помощью коммерчески доступных инструментов, и достигает сопоставимой с системой извлечения знаний (DrQA [36]) точности при 10 в ответах на вопросы (датасет был переформулирован с пар «вопрос-ответ» в предложения с пробелами).

З. Брауи [8] и др. также рассмотрели BERT, поставив вопрос об оценке преимущества таких предобученных моделей перед простыми языковыми моделями, построенными с использованием векторных представлений, с точки зрения способностей к хранению знаний об отношениях. В числе прочего, авторы предложили методологию извлечения этих знаний из предобученной языковой модели. Полученные авторами результаты говорят о том, что предобученная языковая модель BERT действительно захватывает знания об отношениях, основанные на здравом смысле и фактах, в большей степени, чем простые векторные представления, и что извлечение таких знаний из этой и иных подобных языковых моделей может осуществляться полностью автоматически.

Н. Каснер и Х. Шутце расширили предложенную Ф. Петрони и др. методику испытаний (LAMA), рассмотрев отрицание в контексте формулирования запросов к языковым моделям как к «базам знаний» [4]. Авторы предлагают «отрицательный» набор данных для LAMA, собранный с помощью добавления элемента отрицания (например, частицы «не») в маскированные предложения, используемые LAMA. Таким образом, этот датасет состоит из утверждений вида «теория относительности была разработана не ___». Исследуя этот подход, авторы осуществляли пары запросов к готовым предобученным языковым моделям – как оригинальные утверждения из набора данных LAMA, так и соответствующие им отрицательные утверждения. Затем, авторы сравнили полученные результаты в части ранжирования и пересечения предлагаемых вариантов и установили, что предсказываемые для заполнения масок слова зачастую в значительной степени пересекаются. Таким образом, авторы сделали вывод, что добавление отрицания в маскированные предложения во многих случаях не влияет на предсказания, что, само собой, расходится с желаемым результатом. В качестве наиболее яркого примера дефектности таких результатов, авторы неоднократно упоминают маскированное предложение «___ умеют летать». Наиболее вероятным предсказанием для этого утверждения оказалось слово «птицы», однако, это же слово было предложено моделью и для варианта с отрицанием, то есть «птицы не умеют летать».

А. Лю и др. [37] также подняли вопрос об использовании BERT вместо баз знаний для поиска ответов на вопросы на естественных языках. Авторы отмечают, что использование баз знаний в этих целях предполагает необходимость ручного проектирования и низкую способность к обобщению. Используя предобученную языковую модель на основе BERT – BB-KBQA, авторы достигли F 1 в 84.12% на датасете NLPCC-ICCPOL 2016 (набор вопросов и ответов на китайском языке), что опережает любые иные существующие на тот момент решения.

Как отмечает А. Робертс и др. [5], языковые модели способны интернализировать нечто подобное неявным «базам знаний» после предобучения.

Авторы отмечают, что эта особенность может представлять из себя значительную пользу в связи с тем, что, во-первых, такие «базы знаний» строятся с помощью предобучения на неразмеченных и неструктурированных текстовых данных, наборов которых в открытом доступе великое множество, и, во-вторых, что извлечение информации из таких баз знаний возможно с помощью запросов на естественном языке.

А. Робертс и др. [5] рассмотрели способность языковых моделей отвечать на вопросы на естественном языке без какого-либо доступа к внешним знаниям или контексту. Авторы утверждают, что предыдущие работы, рассматривающие способность языковых моделей отвечать на вопросы на естественном языке, тем или иным образом привлекали внешние знания или контекст, например, подавая вместе с вопросом текст, содержащий ответ, или позволяя модели осуществлять поиск информации во внешнем источнике знаний. А. Робертс и др., в свою очередь, рассматривают такой подход, при котором модели для ответа на вопрос приходится анализировать запрос на естественном языке и затем «отыскивать информацию», содержащуюся в ее параметрах. В результате проведенного исследования, авторы пришли к выводу, что языковые модели, работающие без доступа к внешним данным, также способны демонстрировать высокие результаты в задачах предоставления ответов на вопросы. Тем не менее, авторы отмечают, что такие показатели были достигнуты лишь на самой крупной из рассмотренных моделей (Text-to-Text Transfer Transformer [3], 11 млрд. параметров). Кроме того, А. Робертс и др. обращают внимание на то, что, отвечая на вопрос на естественном языке, такие модели, в отличие от языковых моделей, имеющих доступ к внешним данным, не предоставляют никакой индикации того, к какой (очевидно, интернализированной в параметрах) информации они обращаются.

Н. Пернер и др. [38] не согласны с утверждениями о том, что языковая модель BERT способна запоминать фактические знания в процессе предобучения. Авторы

утверждают, что впечатляющие результаты, показанные BERT, объясняются поверхностными рассуждениями (выводом) на основе названий объектов, приводя в пример то, как модель предсказывает родной язык людей исходя из этнической принадлежности их имени (то есть, если имя человека «звучит», как что-то итальянское, то BERT предположит, что этот человек говорит на итальянском языке). Кроме того, авторы провели испытание LAMA на модифицированном, «более фактическом» датасете – LAMA-UHN (UnHelpful Names), из которого были исключены запросы (маскированные предложения), ответы на которые легко угадать, исходя из имен объектов. В результате этого испытания выяснилось, что качественные показатели работы языковой модели BERT оказываются существенно ниже на наборах данных без легко угадываемых запросов. Для решения этих проблем, Н. Пернер и др. предлагают E-BERT - расширение модели BERT, заменяющее упоминания объектов символьными эмбеддингами. Ссылаясь на значительное превосходство E-BERT над BERT и ERNIE в испытаниях на наборе данных LAMA-UHN, авторы утверждают, что E-BERT захватывает большие объемы фактических знаний. Кроме того, авторы продемонстрировали, что использование совокупности E-BERT с BERT в различных формах достигает высоких результатов на оригинальном датасете LAMA.

А. Талмор и др. [39] пишут о недостаточной изученности способностей крупных предобученных языковых моделей к рассуждениям. Авторы предлагают ряд задач на рассуждения, решение которых, на концептуальном уровне, требует таких операций, как сравнение, конъюкция и произведение. В результате проведенных испытаний, А. Талмор и др. приходят к ряду выводов, в числе которых то, что различные языковые модели демонстрируют качественно различные способности к рассуждениям – модель RoBERTa справилась с задачами на рассуждение, а модель BERT совершенно не справилась, и то, что языковые модели не способны к абстрактным рассуждениям и они опираются на контекст, например, хоть модель RoBERTa и может сравнивать возраст, но делать это она может только пока значения остаются в типичном для человека диапазоне. Кроме того, как установили авторы, с половиной испытаний все рассматриваемые модели вовсе не справились.

VII. Заключение

В настоящей статье был приведен и рассмотрен ряд новейших исследований, связанных с методами предобучения и переноса обучения языковых моделей, которые, как утверждают С. Рудер и др., вероятно станут повсеместно используемым инструментом в решении задач в области обработки текстов на естественном языке [2, с. 15].

Рассмотренные в рамках настоящей статьи исследования в области взаимной аугментации предобученных языковых моделей и баз знаний (представленных, в том числе, в виде графов знаний) дают основания считать о чрезвычайной перспективности дальнейшего развития этой темы. В частности, как утверждают Ч. Сунь и др., потенциал предобученной языковой модели BERT остается преимущественно нераскрытым, что объясняется немногочисленностью исследований методов улучшения (тонкой настройки) показателей в конкретных узких задачах [40, с. 194].

Также в настоящей статье был рассмотрен ряд исследований, связанных с вопросом применения языковых моделей в качестве баз знаний. Полученные в рамках этих исследований результаты дают основания полагать, что эта тема отнюдь не исчерпана и заслуживает дальнейшего развития. В частности, как утверждают Ф. Петрони и др., «языковые модели, обученные на постоянно пополняющихся корпусах, в будущем могут стать конкурентоспособной альтернативой традиционным базам знаний, извлеченных из текста» [35, с. 9].

Однако, предобучение языковых моделей вроде BERT требует слишком значительных вычислительных ресурсов, таким образом, эти инструменты не всегда могут использоваться (развертываться) непосредственно на местах для решения реальных задач. [41, с. 2] Кроме того, они часто обучаются на частных (приватных) наборах данных различных размеров, что может приводить к противоречивым результатам [10, с. 1].

Также, исходя из рассмотренных в рамках настоящей статьи работ, следует отметить недостаточное развитие вопроса применения таких, в настоящий момент непрерывно развивающихся, методов, моделей и инструментов для решения практических задач в реальных условиях.

Автором настоящего исследования были рассмотрены вопросы применения композитных языковых моделей с использованием векторных представлений (включая BERT) в задачах, чувствительных к актуальности модели и ко времени (в том числе word completion – автодополнение). Существующие на данный момент версии предобученной языковой модели BERT и производные от нее способны осуществлять вывод в «бюджет» виртуального собеседника – 10 мс, но, тем не менее, предобучение (как и тонкая настройка) модели BERT, как упоминается в настоящей статье, является вычислительно затратным. В этом контексте особый интерес представляет вопрос модификации и разработки методов непрерывной тонкой настройки и аугментации внешними данными предобученных языковых моделей для поддержания их актуальности в релевантных задачах с наименьшими временными затратами.

Библиография

[1] *Devlin J., Chang M.-W., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv preprint arXiv:1810.04805. <https://arxiv.org/abs/1810.04805>.

[2]. *Ruder S., Peters M.E., Swayamdipta S., Wolf T.* Transfer learning in natural language processing // The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts. – Minneapolis (Minnesota, USA): Association for

Computational Linguistics (ACL), 2019. – x; 27 p. – P. 15–18.

[3] *Raffel C., Shazeer N., Roberts A. etc.* Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer // arXiv preprint arXiv:1910.10683. <https://arxiv.org/abs/1910.10683>.

[4] *Kassner N., Schütze H.* Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly // arXiv preprint arXiv:1911.03343. <https://arxiv.org/abs/1911.03343>.

[5] *Roberts A., Raffel C., Shazeer N.* How Much Knowledge Can You Pack Into the Parameters of a Language Model? // arXiv preprint arXiv:2002.08910. <https://arxiv.org/abs/2002.08910>.

[6] *Guu K., Lee K., Tung Z. etc.* REALM: Retrieval-Augmented Language Model Pre-Training // arXiv preprint arXiv:2002.08909. <https://arxiv.org/abs/2002.08909>.

[7] *Peters M.E., Neumann M., Iyyer M. etc.* Deep contextualized word representations // arXiv preprint arXiv:1802.05365. <https://arxiv.org/abs/1802.05365>.

[8] *Bouraoui Z., Camacho-Collados J., Schockaert S.* Inducing Relational Knowledge from BERT // arXiv preprint arXiv:1911.12753. <https://arxiv.org/abs/1911.12753>.

[9] *Goldberg Y.* Assessing BERT's Syntactic Abilities // arXiv preprint arXiv:1901.05287. <https://arxiv.org/abs/1901.05287>.

[10] *Liu Y., Ott M., Goyal N. etc.* RoBERTa: A Robustly Optimized BERT Pretraining Approach // arXiv preprint arXiv:1907.11692. <https://arxiv.org/abs/1907.11692>.

[11] *Radford A., Narasimhan K., Salimans T. etc.* Improving Language Understanding by Generative Pre-Training // <https://pdfs.semanticscholar.org/cd18/800a0fe0b668a1cc19f2ec95b5003d0a5035.pdf>.

[12] *Radford A., Wu J., Child R. etc.* Language models are unsupervised multitask learners // <https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf>.

[13] *Yang Z., Dai Z., Yang Y. etc.* XLNet: Generalized Autoregressive Pretraining for Language Understanding // <https://papers.nips.cc/paper/8812-xlnet-generalized-autoregressive-pretraining-for-language-understanding.pdf>.

[14] *Richardson K., Sabharwal A.* What Does My QA Model Know? Devising Controlled Probes using Expert Knowledge // arXiv preprint arXiv:1912.13337. <https://arxiv.org/abs/1912.13337>.

[15] *Ehrlinger L.,Wöß W.* Towards a Definition of Knowledge Graphs // Joint Proceedings of the Posters and Demos Track of 12th International Conference on Semantic Systems – SEMANTiCS2016 and 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS16). – Leipzig (Germany), 2016. (Vol. 1695).

[16] *Yoo S.-Y., Jeong O.-K.* Automating the expansion of a knowledge graph // Expert Systems with Applications. – 2020, March. – Vol. 141.

[17] *Zhang N., Deng S., Sun Z. etc.* Relation Adversarial Network for Low Resource Knowledge Graph Completion // arXiv preprint arXiv:1911.03091. <https://arxiv.org/abs/1911.03091>.

[18] *Weng J., Gao Y., Qiu J. etc.* Construction and Application of Teaching System Based on Crowdsourcing Knowledge Graph // Knowledge Graph and Semantic Computing: Knowledge Computing and Language Understanding: 4th China Conference, CCKS 2019, Hangzhou, China, August 24–27, 2019: Revised Selected Papers. – Singapore: Springer. 2019. – P. 25–37.

[19] *Yao L., Mao C., Luo Y.* KG-BERT: BERT for Knowledge Graph Completion // arXiv preprint arXiv:1909.03193. <https://arxiv.org/abs/1909.03193>.

[20] *Liu S., d'Aquin M., Motta E.* Measuring Accuracy of Triples in Knowledge Graphs // Language, Data, and Knowledge: First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017: Proceedings. – Cham, Switzerland, 2017. – P. 343–357.

[21] *Ji S., Pan S., Cambria E. etc.* A Survey on Knowledge Graphs: Representation, Acquisition and Applications // arXiv preprint arXiv:2002.00388. <https://arxiv.org/abs/2002.00388>.

[22] *Logan R., Liu N.F., Peters M.E. etc.* Barack's Wife Hillary: Using Knowledge Graphs for Fact-Aware Language Modeling // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. – Florence (Italy): Association for Computational Linguistics, 2019. – P. 5962–5971.

[23] *Wang R., Tang D., Duan N. etc.* K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters // arXiv preprint arXiv:2002.01808. <https://arxiv.org/abs/2002.01808>.

[24] *Yang B., Mitchell T.* Leveraging Knowledge Bases in LSTMs for Improving Machine Reading // arXiv preprint arXiv:1902.09091. <https://arxiv.org/abs/1902.09091>.

[25] *Ostendorff M., Bourgonje P., Berger M. etc.* Enriching BERT with Knowledge Graph Embeddings for Document Classification // arXiv preprint arXiv:1909.08402. <https://arxiv.org/abs/1909.08402>.

[26] *Hu Y., Lin G., Miao Y. etc.* Commonsense Knowledge + BERT for Level 2 Reading Comprehension Ability Test // arXiv preprint arXiv:1909.03415. <https://arxiv.org/abs/1909.03415>.

[27] *He B., Zhou D., Xiao J. etc.* Integrating Graph Contextualized Knowledge into Pre-trained Language Models // arXiv preprint arXiv:1912.00147. <https://arxiv.org/abs/1912.00147>.

[28] *Xiong W., Du J., Wang W.Y. etc.* Pretrained Encyclopedia: Weakly Supervised Knowledge-Pretrained Language Model // arXiv preprint arXiv:1912.09637. <https://arxiv.org/abs/1912.09637>.

[29] *Wang X., Gao T., Zhu Z.* KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation // arXiv preprint arXiv:1911.06136. <https://arxiv.org/abs/1911.06136>.

[30] *Bosselut A., Rashkin H., Sap M. etc.* COMET: Commonsense Transformers for Automatic Knowledge Graph Construction // arXiv preprint arXiv:1906.05317. <https://arxiv.org/abs/1906.05317>.

[31] *Davison J., Feldman J., Rush A.M.* Commonsense knowledge mining from pretrained models // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-

IJCNLP). – Hong Kong (China): Association for Computational Linguistics, 2019. – P. 1173–1178.

[32] *Wang C., Qiu M., Huang J. etc.* KEML: A Knowledge-Enriched Meta-Learning Framework for Lexical Relation Classification // arXiv preprint arXiv:2002.10903. <https://arxiv.org/abs/2002.10903>.

[33] *Zhang J., Zhang Z., Zhang H. etc.* Enriching Medical Terminology Knowledge Bases via Pre-trained Language Model and Graph Convolutional Network // arXiv preprint arXiv:1909.00615. <https://arxiv.org/abs/1909.00615>.

[34] *Trisedya B.D., Weikum G., Qi J.* Neural Relation Extraction for Knowledge Base Enrichment // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. – Florence (Italy): Association for Computational Linguistics, 2019. – P. 229–240.

[35] *Petroni F., Rocktaschel T., Lewis P.* Language Models as Knowledge Bases? // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). – Hong Kong (China): Association for Computational Linguistics, 2019. – P. 2463–2473.

[36] *Chen D., Fisch A., Weston J. etc.* Reading Wikipedia to Answer Open-Domain Questions // arXiv preprint arXiv:1704.00051. <https://arxiv.org/abs/1704.00051>.

[37] *Liu A., Huang Z., Lu H.* BB-KBQA: BERT-Based Knowledge Base Question Answering // Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019: Proceedings. – Cham (Switzerland): Springer, 2019. – P. 81–92.

[38] *Poerner N., Waltinger U., Schütze H.* E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT // arXiv preprint arXiv:1911.03681. <https://arxiv.org/abs/1911.03681>.

[39] *Talmor A., Elazar Y., Goldberg Y. etc.* oLMpics – On what Language Model Pre-training Captures / A. Talmor // arXiv preprint arXiv:1912.13283. <https://arxiv.org/abs/1912.13283>.

[40] *Sun C., Qiu X., Xu Y. etc.* How to Fine-Tune BERT for Text Classification? // Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019: Proceedings. – Cham (Switzerland): Springer, 2019. – P. 194—206.

[41] *Lu W., Jiao J., Zhang R.* TwinBERT: Distilling Knowledge to Twin-Structured BERT Models for Efficient Retrieval // arXiv preprint arXiv:2002.06275. <https://arxiv.org/abs/2002.06275>.

The concept of pretrained language models in the context of knowledge engineering

Dmitry Ponkin

***Abstract* – The article studies the concept and technologies of pre-trained language models in the context of knowledge engineering. The author substantiates the relevance of the issue of the existence of internalized and implicit knowledge, extracted from text corpora used for pre-training or transfer learning in pre-trained language models. The article provides a detailed overview of the existing approaches to the interpretation of this concept. The author reviews a number of recent studies related to pre-training and transfer learning methods in regards to language models. This article discusses the latest research on the augmentation of language models with knowledge. Moreover, it studies the current research on the use of pre-trained language models to search and retrieve knowledge, to aid in the process of building knowledge bases, as well as their use as independent knowledge bases. The content of the concept "pretrained language models" is explained. The author provides examples of the implementation of pre-trained language models in practice, including the discussion of the use of language models as knowledge bases. The essence of the concept of unsupervised pre-training of language models using large and unstructured text corpora before further training for a specific task (fine tuning), "transfer learning", is also touched on. The author examines the concept of "knowledge graph", which is now widely used both in general and in the context relevant to this article, as well as a number of recent research in the realm of pre-training and transfer learning in regards to language models.**

***Keywords* – pretrained language models, transformers, knowledgebase, natural language processing, knowledge engineering.**

УДК 519.6; 519.7:004.8; 004.822; 004.85
ББК 22.18; 3

REFERENCES

[1] *Devlin J., Chang M.-W., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv preprint arXiv:1810.04805. <https://arxiv.org/abs/1810.04805>.

[2]. *Ruder S., Peters M.E., Swayamdipta S., Wolf T.* Transfer learning in natural language processing // The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts. – Minneapolis (Minnesota, USA): Association for Computational Linguistics (ACL), 2019. – x; 27 p. – P. 15–18.

[3] *Raffel C., Shazeer N., Roberts A. etc.* Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer // arXiv preprint arXiv:1910.10683. <https://arxiv.org/abs/1910.10683>.

[4] *Kassner N., Schütze H.* Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly // arXiv preprint arXiv:1911.03343. <https://arxiv.org/abs/1911.03343>.

[5] *Roberts A., Raffel C., Shazeer N.* How Much Knowledge Can You Pack Into the Parameters of a Language Model? // arXiv preprint arXiv:2002.08910. <https://arxiv.org/abs/2002.08910>.

[6] *Guu K., Lee K., Tung Z. etc.* REALM: Retrieval-Augmented Language Model Pre-Training // arXiv preprint arXiv:2002.08909. <https://arxiv.org/abs/2002.08909>.

[7] *Peters M.E., Neumann M., Iyyer M. etc.* Deep contextualized word representations // arXiv preprint arXiv:1802.05365. <https://arxiv.org/abs/1802.05365>.

[8] *Bouraoui Z., Camacho-Collados J., Schockaert S.* Inducing Relational Knowledge from BERT // arXiv preprint arXiv:1911.12753. <https://arxiv.org/abs/1911.12753>.

[9] *Goldberg Y.* Assessing BERT's Syntactic Abilities // arXiv preprint arXiv:1901.05287. <https://arxiv.org/abs/1901.05287>.

[10] *Liu Y., Ott M., Goyal N. etc.* RoBERTa: A Robustly Optimized BERT Pretraining Approach // arXiv preprint arXiv:1907.11692. <https://arxiv.org/abs/1907.11692>.

[11] *Radford A., Narasimhan K., Salimans T. etc.* Improving Language Understanding by Generative Pre-Training // <https://pdfs.semanticscholar.org/cd18/800a0fe0b668a1cc19f2ec95b5003d0a5035.pdf>.

[12] *Radford A., Wu J., Child R. etc.* Language models are unsupervised multitask learners // <https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf>.

[13] *Yang Z., Dai Z., Yang Y. etc.* XLNet: Generalized Autoregressive Pretraining for Language Understanding // <https://papers.nips.cc/paper/8812-xlnet-generalized-autoregressive-pretraining-for-language-understanding.pdf>.

[14] *Richardson K., Sabharwal A.* What Does My QA Model Know? Devising Controlled Probes using Expert Knowledge // arXiv preprint arXiv:1912.13337. <https://arxiv.org/abs/1912.13337>.

[15] *Ehrlinger L.,Wöß W.* Towards a Definition of Knowledge Graphs // Joint Proceedings of the Posters and Demos Track of 12th International Conference on Semantic Systems – SEMANTiCS2016 and 1st International Workshop on Semantic Change & Evolving

Semantics (SuCCESS16). – Leipzig (Germany), 2016. (Vol. 1695).

[16] *Yoo S.-Y., Jeong O.-K.* Automating the expansion of a knowledge graph // Expert Systems with Applications. – 2020, March. – Vol. 141.

[17] *Zhang N., Deng S., Sun Z. etc.* Relation Adversarial Network for Low Resource Knowledge Graph Completion // arXiv preprint arXiv:1911.03091. <https://arxiv.org/abs/1911.03091>.

[18] *Weng J., Gao Y., Qiu J. etc.* Construction and Application of Teaching System Based on Crowdsourcing Knowledge Graph // Knowledge Graph and Semantic Computing: Knowledge Computing and Language Understanding: 4th China Conference, CCKS 2019, Hangzhou, China, August 24–27, 2019: Revised Selected Papers. – Singapore: Springer. 2019. – P. 25–37.

[19] *Yao L., Mao C., Luo Y.* KG-BERT: BERT for Knowledge Graph Completion // arXiv preprint arXiv:1909.03193. <https://arxiv.org/abs/1909.03193>.

[20] *Liu S., d'Aquin M., Motta E.* Measuring Accuracy of Triples in Knowledge Graphs // Language, Data, and Knowledge: First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017: Proceedings. – Cham, Switzerland, 2017. – P. 343–357.

[21] *Ji S., Pan S., Cambria E. etc.* A Survey on Knowledge Graphs: Representation, Acquisition and Applications // arXiv preprint arXiv:2002.00388. <https://arxiv.org/abs/2002.00388>.

[22] *Logan R., Liu N.F., Peters M.E. etc.* Barack's Wife Hillary: Using Knowledge Graphs for Fact-Aware Language Modeling // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. – Florence (Italy): Association for Computational Linguistics, 2019. – P. 5962–5971.

[23] *Wang R., Tang D., Duan N. etc.* K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters // arXiv preprint arXiv:2002.01808. <https://arxiv.org/abs/2002.01808>.

[24] *Yang B., Mitchell T.* Leveraging Knowledge Bases in LSTMs for Improving Machine Reading // arXiv preprint arXiv:1902.09091. <https://arxiv.org/abs/1902.09091>.

[25] *Ostendorff M., Bourgonje P., Berger M. etc.* Enriching BERT with Knowledge Graph Embeddings for Document Classification // arXiv preprint arXiv:1909.08402. <https://arxiv.org/abs/1909.08402>.

[26] *Hu Y., Lin G., Miao Y. etc.* Commonsense Knowledge + BERT for Level 2 Reading Comprehension Ability Test // arXiv preprint arXiv:1909.03415. <https://arxiv.org/abs/1909.03415>.

[27] *He B., Zhou D., Xiao J. etc.* Integrating Graph Contextualized Knowledge into Pre-trained Language Models // arXiv preprint arXiv:1912.00147. <https://arxiv.org/abs/1912.00147>.

[28] *Xiong W., Du J., Wang W.Y. etc.* Pretrained Encyclopedia: Weakly Supervised Knowledge-Pretrained Language Model // arXiv preprint arXiv:1912.09637. <https://arxiv.org/abs/1912.09637>.

[29] *Wang X., Gao T., Zhu Z.* KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation // arXiv preprint arXiv:1911.06136. <https://arxiv.org/abs/1911.06136>.

[30] *Bosselut A., Rashkin H., Sap M. etc.* COMET: Commonsense Transformers for Automatic Knowledge Graph Construction // arXiv preprint arXiv:1906.05317. <https://arxiv.org/abs/1906.05317>.

[31] *Davison J., Feldman J., Rush A.M.* Commonsense knowledge mining from pretrained models // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). – Hong Kong (China): Association for Computational Linguistics, 2019. – P. 1173–1178.

[32] *Wang C., Qiu M., Huang J. etc.* KEML: A Knowledge-Enriched Meta-Learning Framework for Lexical Relation Classification // arXiv preprint arXiv:2002.10903. <https://arxiv.org/abs/2002.10903>.

[33] *Zhang J., Zhang Z., Zhang H. etc.* Enriching Medical Terminology Knowledge Bases via Pre-trained Language Model and Graph Convolutional Network // arXiv preprint arXiv:1909.00615. <https://arxiv.org/abs/1909.00615>.

[34] *Trisedya B.D., Weikum G., Qi J.* Neural Relation Extraction for Knowledge Base Enrichment // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. – Florence (Italy): Association for Computational Linguistics, 2019. – P. 229–240.

[35] *Petroni F., Rocktaschel T., Lewis P.* Language Models as Knowledge Bases? // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). – Hong Kong (China): Association for Computational Linguistics, 2019. – P. 2463–2473.

[36] *Chen D., Fisch A., Weston J. etc.* Reading Wikipedia to Answer Open-Domain Questions // arXiv preprint arXiv:1704.00051. <https://arxiv.org/abs/1704.00051>.

[37] *Liu A., Huang Z., Lu H.* BB-KBQA: BERT-Based Knowledge Base Question Answering // Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019: Proceedings. – Cham (Switzerland): Springer, 2019. – P. 81–92.

[38] *Poerner N., Waltinger U., Schütze H.* E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT // arXiv preprint arXiv:1911.03681. <https://arxiv.org/abs/1911.03681>.

[39] *Talmor A., Elazar Y., Goldberg Y. etc.* oLMpics – On what Language Model Pre-training Captures / A. Talmor // arXiv preprint arXiv:1912.13283. <https://arxiv.org/abs/1912.13283>.

[40] *Sun C., Qiu X., Xu Y. etc.* How to Fine-Tune BERT for Text Classification? // Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019: Proceedings. – Cham (Switzerland): Springer, 2019. – P. 194—206.

[41] *Lu W., Jiao J., Zhang R.* TwinBERT: Distilling Knowledge to Twin-Structured BERT Models for Efficient Retrieval // arXiv preprint arXiv:2002.06275. <https://arxiv.org/abs/2002.06275>.