Влияние методов построения векторных представлений имен сущностей на качество выравнивания сущностей

Д.И. Гусев[1], З.В. Апанович[1,2]

[1] Новосибирский государственный университет
[2] Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук

Аннотация. Проблема слияния разноязычных графов знаний (Knowledge Graph, KG, КГ), становится все более актуальной. Основным этапом для ее решения является идентификация эквивалентных сущностей и их описаний. Она также известна как проблема выравнивания сущностей (Entity Alignment, EA). В последние годы активно исследуются методы EA на основе векторных представлений сущностей. Недавние исследования показывают, что качество этих подходов зависит от того, каким образом используется информация о структуре графов знаний и методов построения векторных представлений имен сущностей. В данной статье представлены эксперименты, целью которых является улучшение выравнивания сущностей на англо-русском наборе данных.

Ключевые слова: многоязычные графы знаний, выравнивание сущностей, векторное представление, языковые модели,.

Impact of entity names embeddings on the quality of entity alignment

D.I. Gusev[1], Z.V. Apanovich[1,2]

[1] Novosibirsk state university
[2] Institute of Informatics Systems SBRAS

Abstract. The problem of merging multilingual knowledge graphs (KG) is becoming more and more relevant. The main step for its solution is the identification of equivalent entities and their descriptions. It is also known as the entity alignment (EA) problem. In recent years, EA methods based on embeddings of entities have been actively studied. Recent studies show that the quality of these approaches depends on how information about the structure of knowledge graphs and methods for constructing embeddings of entity names are used. This article presents experiments, the purpose of which is to improve the alignment of entities on the English-Russian dataset.

Keywords: multilingual knowledge graphs, entity alignment, embeddings, language model.

1. Введение

Графы знаний являются в настоящее время основой таких приложений как рекомендательные системы, системы принятия решений, вопросно-ответные системы, семантический поиск и др. Чем мощнее граф знаний, тем выше качество приложений, на них базирующихся. Поэтому весьма актуальной является задача интеграции различных графов знаний, а в основе такой интеграции находится решение задачи слияния информации из разных графов знаний об одном и том же объекте реального мира. Данная задача известна под такими названиями как *сопоставление сущностей, выравнивание сущностей, идентификация сущностей* и др. В последние несколько лет возрос интерес к интеграции *разноязычных* графов знаний, поэтому весьма актуальной является задача связывания информации об одних и тех же объектах реального мира, описанных в разноязычных графах знаний. Разные языковые версии графов знаний обладают, с одной стороны, свойством взаимодополнительности, а с другой стороны, каждая языковая версия содержит более точную и полную информацию об объектах, характерных для конкретного языка. Например, русскоязычная версия DBpedia содержит более полную и корректную информацию про объекты, расположенные на территории России.

Графы знаний хранят факты в виде реляционных и литеральных триплет. Реляционные триплеты изображают отношение между двумя объектами реального мира и имеют формат *tr_r* = (*субъектная сущность, отношение, объектная сущность*). Литеральные триплеты хранят

информацию об атрибутах объектов реального мира и имеют формат *tr_l* = (*субъектная сущность, атрибут, литеральное значение*).

В последние несколько лет получили распространение методы установления соответствия между сущностями различных графов знаний, использующие так называемые «эмбеддинги» (embeddings), векторные представления заданной размерности для сущностей и отношений графов знаний. Достоинством подхода на основе эмбеддингов является высокая масштабируемость и небольшие усилия при подготовке обучающих выборок.

Следует сказать, что создание новых методов основано на интуиции разработчиков, эвристиках и экспериментах проб и ошибок. Поэтому весьма важным является создание общей основы для понимания разнообразных методов. В настоящее время такую общую основу составляют результаты тестирования различных алгоритмов на едином наборе данных. В работе [1] представлена библиотека OpenEA, содержащая несколько десятков алгоритмов EA на основе различных стратегий построения векторных представлений, а также результаты экспериментов с этими векторными представлениями на тестовой выборке, содержащей англо-немецкие, англо-французские и англо-китайские данные.

Понятно, что русскоязычному пользователю интересны, прежде всего, эксперименты, использующие русскоязычные данные. Во-первых, такие данные проще интерпретировать, а во-вторых, известно, что различные языковые версии графов знаний обладают свойством «смещенности», то есть, одни и те же алгоритмы могут давать разные результаты на разных версиях графов знаний из-за различной структуры этих графов.

В работе [2] описан русско-английский набор данных для экспериментов с алгоритмами кросс-языкового выравнивания сущностей. К удивлению авторов, алгоритмы, выдававшие наилучшие результаты на стандартных разноязычных наборах данных, выдавали весьма посредственные результаты на русско-английском наборе данных. Этот вопрос потребовал дополнительного изучения, и в данной работе представлены эксперименты с алгоритмами выравнивания сущностей разного типа на англо-русской обучающей выборке. Рассмотрены различные способы построения векторных представлений имен сущностей, а также возможные комбинации этих методов с методами построения векторных представлений сущностей на основе реляционных триплет.

2. Группы алгоритмов сопоставления сущностей на основе векторных представлений (embeddings)

Большинство методов выравнивания сущностей на основе векторных представлений сводятся к двум шагам:

1. Генерация векторных представлений для сущностей и отношений

2. Отображение этих векторных представлений в единое векторное пространство, при помощи предварительно выровненных сущностей (seed alignments) или в различные векторные пространства.

В первом случае, вопрос, являются ли две сущности из разных графов эквивалентными (соответствующими одному и тому же объекту реального мира) решается при помощи сравнения их векторов, например вычислением евклидова расстояния или косинусной близости. При отображении сущностей двух графов знаний в разные векторные пространства нужно также находить матрицу соответствия между векторами этих двух пространств.

Современные решения EA опираются в основном на структурную информацию в графах знаний, то есть реляционные триплеты. Основу этих методов составляет предположение о том, что эквивалентные сущности должны иметь сходные графовые окрестности. Первоначально преобладал так называемый *триплетно-трансляционный подход*, который рассматривал вектор, представляющий отношение между двумя сущностями, как вектор сдвига вектора одной сущности относительно вектора второй сущности. Одним из лучших представителей *триплетно-трансляционного подхода является* **MultiKE (**Multi-view Knowledge Graph Embedding) [3]. MultiKE строит три типа векторных представлений для каждой сущности, используя разные «виды»: вид, зависящий от названия сущности, реляционный вид и атрибутный вид. Каждый из «видов» строится по собственному алгоритму. Например, для каждого слова из названия сущности находится вектор, полученный с помощью word2vec, а если такого не существует, то вектор слова получается с помощью суммирования векторов символов, полученных с помощью алгоритма character embedding. Векторы слов суммируются и получается вектор названия, который непосредственно участвует в обучении модели. Окончательное векторное представление сущности может быть получено при помощи разных способов комбинирования упомянутых трех видов.

В последние годы чрезвычайно популярными стали подходы построения векторных представлений сущностей на основе графовых сверточных сетей. Эти методы выдают очень неплохие результаты, но их основным недостатком является чрезвычайная сложность, значительное время вычислений и плохая интерпретируемость. Представителем этого подхода является RDGCN (Relation-aware Dual-Graph Convolutional Network) [4]. Подход RDGCN использует для построения векторных представлений не только структуру исходных графов знаний (primal entity graph), но и вспомогательные графы, двойственные по отношению к исходным графам (dual relation graph), вершинами которых являются ребра исходных графов. Для осуществления взаимодействия между исходными графами знаний и двойственными реляционными графами используется механизм графовых сетей внимания (Graph Attention Networks, GAT).

Результирующие векторные представления исходных графов затем подаются в графовые сверточные сети (Graph Convolutional networks, GCN), для извлечения информации о структуре окружений вершин.

Совсем недавно появился чрезвычайно простой подход к выравниванию сущностей под названием **SEU** (Simple but Effective Unsupervised EA method) [5], не использующий нейронных сетей. Основная идея SEU состоит в сведении задачи EA к давно известной задаче назначения, для которой существует хорошо известный венгерский алгоритм решения. Основным предположением этого подхода является то, что матрицы смежностей двух графов знаний являются изоморфными. В этом случае матрица смежности исходного графа может быть преобразована в матрицу смежности второго графа посредством переупорядочения строк или столбцов.

Тем не менее, большинство недавних исследований указывают на то, что современные методы EA не способны выдавать удовлетворительные результаты только на основании реляционных триплет, если набор данных имеет распределение степеней сущностей, близкое к реальным КГ. В частности, известно, что примерно половина сущностей в реальных КГ связана с менее чем тремя другими сущностями [6].

Это наблюдение делает важным использование дополнительной информации, такой как *имена сущностей* и комбинирование информации об именах сущностей со структурной информацией. Названия сущностей необходимо привести к общему языку, а затем сравнить. Возможны два базовых подхода для сравнения имен сущностей: подход на основе строкового сходства и подход на основе семантического сходства. Методы семантического сходства можно разбить на две группы: генерация векторных представлений на основе отдельных слов (модели word2vec, glove [7]). В силу ограниченности используемых словарей, часто возникает ситуация, что нужное слово отсутствует в используемом словаре и в этом случае, векторное представление слова строится на основе литер, входящих в его состав (модели fastText, name-BERT).

3. Русско-английский набор данных и метрики для оценки качества алгоритмов EA

Современные графы знаний имеют значительные размеры, поэтому вместо полномасштабных экспериментов по установлению соответствия между сущностями из разных графов знаний осуществляются эксперименты на выборках ограниченного размера. В настоящее время принято экспериментировать с выборками, содержащими 15 000 и 100 000 соответствий между сущностями из двух графов знаний. Наибольшее распространение получил набор данных DBP15K [1], который содержит по 15 000 пар сущностей, связанных отношениями *owl:sameAs* из разных языковых версий DBpedia, для таких пар языков как англо-китайский,

англо-французский и англо-немецкий. В [1] также описан итеративный алгоритм построения разноязычной выборки на основе степеней сущностей IDS (Iterative Degree-based Sampling), в которой распределение степеней сущностей близко к распределениям степеней в реальных графах знаний.

Принимая во внимание то, что каждая языковая версия графа знаний имеет свою собственную структуру, отличную от других графов знаний, а также то, что данные, полученные для русскоязычного графа знаний проще интерпретировать, нами был сгенерирован русско-английский набор тестовых данных на основе русскоязычной и англоязычной версий DBpedia [2]. Использовался набор данных англоязычной и русскоязычной DBpedia за 2016 год. (DBpedia 2016-10, https://wiki.dbpedia.org/downloads-2016-10).

Набор DBP-15K EN-RU (V1, V2) сгенерирован на основе алгоритма IDS и доступен для свободного скачивания (https://www.dropbox.com/sh/4oh3nkzwdr1w4dv/AACZ4v8jCdR7Y4mDtS654Bega?dl=0).

Для анализа результатов алгоритма используют метрики *hits@k u* среднеобратный ранг *(*Mean reciprocal rank, MRR*)*. Метрика *hits@k=n*% означает, что для *n* процентов объектов из одного графа знаний эквивалентный объект из второго графа знаний находится среди ближайший *k* соседей в векторном пространстве. Очевидно, самой показательной считается метрика *hits*@1 - так как эта метрика соответствует алгоритму, который самостоятельно строит правильные отношения *owl:sameAs* между сущностями. *Среднеобратный ранг* определяется как среднее значение обратных рангов по всем запросам. Обратный ранг, в данном случае, означает обратное число номера (ранга) первого правильного ответа в списке откликов.

4. Влияние перевода имен сущностей на результаты методов EA.

Рассмотренные ранее подходы выравнивания сущностей используют предобученные языковые модели слов. Это значительно упрощает объединение схожих значений в единое семантическое пространство. При этом нередки случаи, когда искомые слова не содержатся в модели. Данная проблема малозаметна для языков со схожей морфологией, например английского и французского. Для помещения английских и русских слов в единое векторное представление было решено применить машинный перевод. Для решения указанной проблемы нами был разработан инструмент автоматического перевода на основе Google Translate API. На вход подается язык, с которого будет осуществлен перевод, имена сущностей и литералы. Результат передается в метод формирования векторного представления.

Для экспериментов с русско-английским набором данных были выбраны алгоритмы MultiKE [3] и RDGCN [4], как алгоритмы, выдававшие наилучшие результаты на англо-французском наборе данных, и весьма

посредственные результаты на русско-английском наборе данных. Также были проведены эксперименты с методом SEU [5].

Наше предположение состояло в том, что причиной этих неудовлетворительных результатов было недостаточное использование информации об именах сущностей при построении векторных представлений. Поэтому были рассмотрены переводы имен сущностей с русского на английский язык, а также возможные комбинации различных стратегий построения векторных представлений на основе реляционных триплет с различными вариантами построения векторных представлений для имен сущностей.

В Таблице 1 показаны результаты применения перевода имен сущностей на качество алгоритмов EA. В столбце «Перевод» знаком плюс или минус обозначен факт наличия или отсутствия перевода имен сущностей.

Таблица 1. Влияние перевода имен сущностей на качество алгоритмов EA

Подход	Данные	Перевод	Hits@1	Hits@10	Hits@50	MRR	Разница
MultiKE	EN-FR-15K	-	0,741	0,836	0,889	0,774	
MultiKE	EN-FR-15K	+	0,806	0,885	0,926	0,835	0,065
MultiKE	EN-FR-15K	-	0,855	0,921	0,953	0,878	
MultiKE	EN-FR-15K	+	0,893	0,956	0,978	0,915	0,038
MultiKE	EN-RU-15K	-	0,315	0,457	0,599	0,364	
MultiKE	EN-RU-15K	+	0,520	0,666	0,769	0,570	0,205
MultiKE	EN-RU-15K	-	0,453	0,623	0,742	0,510	
MultiKE	EN-RU-15K	+	0,617	0,770	0,856	0,670	0,164
RDGCN	EN-FR-15K	-	0,770	0,892	0,924	0,813	
RDGCN	EN-FR-15K	+	0,771	0,893	0,924	0,813	0,001
RDGCN	EN-FR-15K	-	0,862	0,948	0,971	0,895	
RDGCN	EN-FR-15K	+	0,871	0,951	0,973	0,903	0,009
RDGCN	EN-RU-15K	-	0,396	0,597	0,712	0,460	
RDGCN	EN-RU-15K	+	0,744	0,882	0,923	0,792	0,347
RDGCN	EN-RU-15K	-	0,537	0,717	0,803	0,599	
RDGCN	EN-RU-15K	+	0,844	0,923	0,967	0,882	0,307
SEU	EN-FR-15K	-	0,989	0,998	0,999	0,992	
SEU	EN-FR-15K	+	0,995	1,000	1,000	0,997	0,006
SEU	EN-FR-15K	-	0,992	0,999	1,000	0,994	
SEU	EN-FR-15K	+	0,996	1,000	1,000	0,997	0,004
SEU	EN-RU-15K	-	0,301	0,348	0,385	0,318	
SEU	EN-RU-15K	+	0,972	0,995	0,998	0,981	0,672
SEU	EN-RU-15K	-	0,424	0,483	0,532	0,445	
SEU	EN-RU-15K	+	0,990	0,998	1,000	0,993	0,566

5. Визуализация для сравнения результатов методов генерации представлений имен сущностей

Для сравнения методов генерации векторных представлений имен сущностей на основе EN-RU-15K (V1) и с применением предварительного

перевода были получены визуализации результатов. В качестве инструмента снижения размерности использован t-SNE.

На представленных изображениях английские имена сущностей имеют синий цвет, русские – красный. Это позволяет оценить эффективность метода генерации векторных представлений. Высокая степень наложения цветов говорит о том, что семантически связанные данные, представленные на разных языках, имеют сходные векторные представления. Наличие одноцветных зон свидетельствует о том, что метод не смог установить кросс-языковое соответствие между именами сущностей.

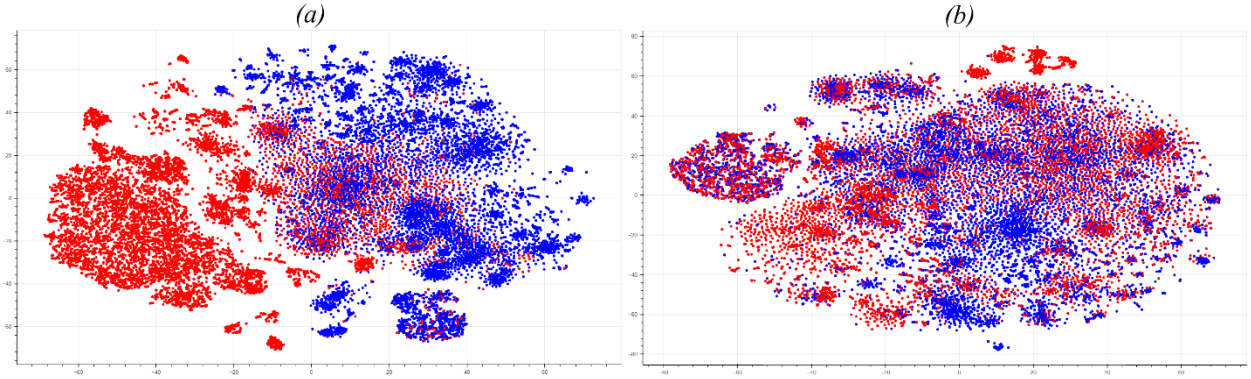

Рис. 1. Векторные представления имён сущностей из MultiKE:
a – без перевода; *b* – с переводом

По результату стандартного генератора MultiKE (рис. 1, *a*) видно, что русские имена сущностей находятся в отдалении от английских. Машинный перевод частично решает данную проблему (рис. 1, *b*). Однако результат метода генерации векторных представлений MultiKE имеет выраженные языковые кластеры, что говорит о невысокой точности.

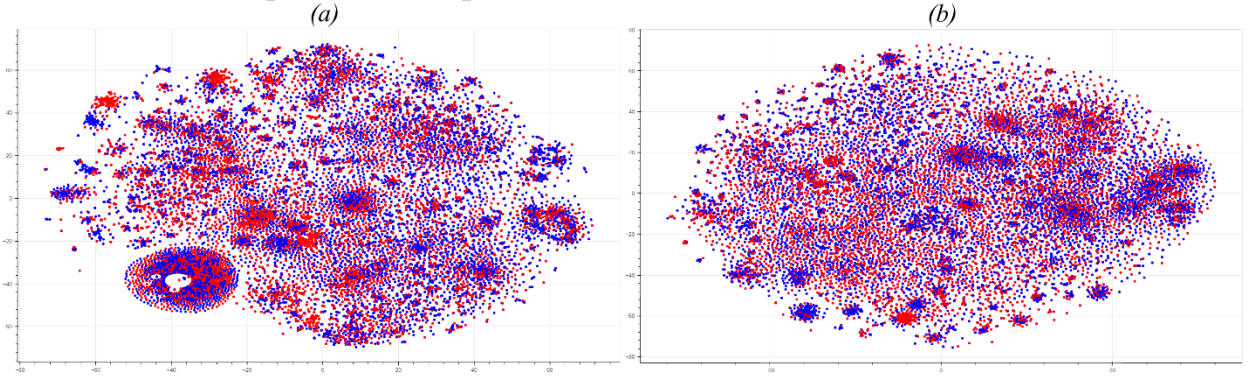

Рис. 2. Векторные представления имён сущностей из подходов EA:
a – RDGCN; *b* – SEU

В векторном представлении имён сущностей из RDGCN (рис. 2, *a*) в левом нижнем углу имеется кластер эллипсоидной формы. Он возник из-за зануления векторов слов, для которых алгоритм из RGDCN не нашел значений в предобученной модели. В остальном же, данное векторное представление имеет большую степень наложения по сравнению с MultiKE. Наименьшее количество языковых кластеров наблюдается у результата, полученного при помощи метода генерации из SEU (рис. 2, *b*).

6. Влияние разных способов построения векторных представлений имен сущностей на результаты разных методов EA

Помимо перевода русскоязычных имен сущностей на английский язык были проведены эксперименты с несколькими моделями генерации векторных представлений имен сущностей. Таким образом рассматриваемые нами методы используют следующие методы построения начальных представлений для имен сущностей:

1 MultiKE – генерация векторных представлений уровня слов (word2vec) и уровня литер (fastText).

2. RDGCN – перевод имен сущностей на английский язык, генерация векторных представлений на уровне слов (glove.840B.300d).

3. SEU – За основу берется предположение, что не только информация о структуре окрестностей, но и текстовая информация эквивалентных сущностей, обладают свойством изоморфизма. Построение векторного представления имен сущностей состоит из следующих этапов: перевод входных данных на английский язык, чтение предобученной модели предобработка входных данных, токенизация по словам, формирование биграмм, формирование векторных представлений, снижение размерности. В качестве предобученной модели использовалась glove.6B.300d.

Также, в качестве альтернативных методов генерации векторных представлений имен сущностей были выбраны современные модели обработки естественных языков XLNet [8] и LaBSE [9]. Спецификой этих моделей является возможность строить векторные представления для наборов слов, таких как предложения.

XLNet. Целью модели является изучение распределений для всех перестановок слов в заданной последовательности [8]. Векторные представления формируются в рамках только одного языка, поэтому для решения нашей задачи потребовалось предварительно применить машинный перевод.

LaBSE. Генерирует независимые от языка векторные представления предложений на основе BERT. Достигается это путём объединения возможностей маскированного и кросс-языкового моделирования [9].

Таблица 2. Результаты методов EA в зависимости от способа генерации векторных представлений имен сущностей на наборе данных EN-RU-15K (V1).

Подход	Генератор	Hits@1	Hits@5	Hits@10	Hits@50	MRR
MultiKE	Стандартный	0,520	0,621	0,666	0,769	0,570
MultiKE	Из RDGCN	0,699	0,781	0,813	0,878	0,737
MultiKE	Из SEU	**0,812**	**0,875**	**0,891**	**0,932**	**0,841**

RDGCN	Стандартный	0,744	0,847	0,882	0,923	0,792
RDGCN	Из MultiKE	0,680	0,796	0,828	0,884	0,733
RDGCN	Из SEU	**0,848**	**0,921**	**0,935**	**0,956**	**0,881**
RDGCN	XLNet	0,434	0,500	0,530	0,605	0,467
RDGCN	LaBSE	0,754	0,837	0,859	0,897	0,792
SEU	Стандартный	**0,972**	**0,991**	**0,995**	**0,998**	**0,981**
SEU	Из MultiKE	0,881	0,935	0,948	0,975	0,905
SEU	Из RDGCN	0,874	0,931	0,954	0,986	0,905
SEU	XLNet	0,325	0,413	0,455	0,549	0,369
SEU	LaBSE	0,949	0,976	0,984	0,993	0,962

По данным таблицы 2 видно, что хорошо себя показал генератор векторных представлений имен сущностей LaBSE. Он оказался эффективнее генераторов MultiKE и RDGCN. Тем не менее, генератор из SEU оказался наиболее эффективным. MultiKE и RDGCN на его основе превысили исходные значения точности.

Результаты применения моделей XLNet и LaBSE к MultiKE не указаны в связи с нехваткой вычислительных ресурсов для построения векторных представлений литералов. Выводы об их эффективности сделаны на основе значений из других подходов. Модель XLNet оказалась непригодной для формирования векторных представлений. Результаты подходов на ее основе близки к значениям, полученным без перевода.

7. Заключение

В данной работе мы изучили влияние методов построения векторных представлений для имён сущностей и литералов на качество результатов различных методов EA. Был исследован вклад применения перевода и современных моделей обработки естественных языков. Следует заметить, что полученные результаты имеют достаточно «идеальный» характер, так как проводились на наборе данных, в котором у каждой сущности англоязычного КГ соответствует эквивалентная сущность русскоязычного КГ. Эта ситуация весьма далека от реального положения вещей и требует дальнейшего исследования.

Литература

1. Zequn Sun, Qingheng Zhang, Wei Hu A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs // arXiv:2003.07743 – 2020

2. Gnezdilova V. A., Apanovich Z. V. Russian-English dataset and comparative analysis of algorithms for cross-language embedding-based entity alignment // J. Phys.: Conf. Ser. 2099 012023 – 2021.
3. Qingheng Zhang, Zequn Sun, Wei Hu, Muhao Chen, Lingbing Guo, Yuzhong Qu Multi-view Knowledge Graph Embedding for Entity Alignment // https://arxiv.org/abs/1906.02390
4. Yuting Wu, Xiao Liu, Yansong Feng, Relation-Aware Entity Alignment for Heterogeneous Knowledge Graphs // Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19) rXiv:1908.08210v1 – 2019.
5. Xin Mao, Wenting Wang, Yuanbin Wu, From Alignment to Assignment: Frustratingly Simple Unsupervised Entity Alignment// https://arxiv.org/abs/2109.02363 – 2021.
6. Lingbing Guo, Zequn Sun, Wei Hu, Learning to Exploit Long-term Relational Dependencies in Knowledge Graphs// Proceedings of the 36 th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019
7. Efficient Estimation of Word Representations in Vector Space / Tomas Mikolov, Kai Chen, Greg Corrado [и др.] // arXiv:1301.3781 – 2013.
8. Zhilin Yang, Zihang Dai, Yiming Yang, XLNet: Generalized Autoregressive Pretraining for Language Understanding// arXiv:1906.08237v2 – 2020.
9. Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, Wei Wang Language-agnostic BERT Sentence Embedding, // arXiv:2007.01852v2 – 2022.

References

1. Zequn Sun, Qingheng Zhang, Wei Hu A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs // arXiv:2003.07743 – 2020
2. Gnezdilova V. A., Apanovich Z. V. Russian-English dataset and comparative analysis of algorithms for cross-language embedding-based entity alignment // J. Phys.: Conf. Ser. 2099 012023 – 2021.
3. Qingheng Zhang, Zequn Sun, Wei Hu, Muhao Chen, Lingbing Guo, Yuzhong Qu Multi-view Knowledge Graph Embedding for Entity Alignment // https://arxiv.org/abs/1906.02390
4. Yuting Wu, Xiao Liu, Yansong Feng, Relation-Aware Entity Alignment for Heterogeneous Knowledge Graphs // Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19) rXiv:1908.08210v1 – 2019.
5. Xin Mao, Wenting Wang, Yuanbin Wu, From Alignment to Assignment: Frustratingly Simple Unsupervised Entity Alignment// https://arxiv.org/abs/2109.02363 – 2021.

6. Lingbing Guo, Zequn Sun, Wei Hu, Learning to Exploit Long-term Relational Dependencies in Knowledge Graphs// Proceedings of the 36 th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019
7. Efficient Estimation of Word Representations in Vector Space / Tomas Mikolov, Kai Chen, Greg Corrado [и др.] // arXiv:1301.3781 – 2013.
8. Zhilin Yang, Zihang Dai, Yiming Yang, XLNet: Generalized Autoregressive Pretraining for Language Understanding// arXiv:1906.08237v2 – 2020.
9. Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, Wei Wang Language-agnostic BERT Sentence Embedding, // arXiv:2007.01852v2 – 2022.