Научная статья

УДК 004.822
DOI 10.25205/1818-7900-2022-20-2-5-17

Влияние методов построения векторных представлений на подходы выравнивания сущностей

Даниил Иванович Гусев[1], Зинаида Владимировна Апанович[2]

[1, 2] Новосибирский государственный университет

Новосибирск, Россия

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

Новосибирск, Россия

[1]apanovich_09@mail.ru, https://orcid.org/0000-0002-5767-284 X
[2]d.gusev1@g.nsu.ru, https://orcid.org/0000-0001-9636-2783

Аннотация

Проблема слияния графов знаний (KG), представленных на разных языках становится все более актуальной. Основным этапом для ее решения является идентификация эквивалентных сущностей и их описаний. Она также известна как проблема выравнивания сущностей. Недавние исследования показывают, что существующие подходы эффективны не для всех языков. В данной статье представлены эксперименты, целью которых является улучшение выравнивания сущностей на англо-русском наборе данных. Полученные результаты рассмотрены как с точки зрения целого графа, так и с точки зрения отдельных типов сущностей. Произведена оценка влияния количества отношений и атрибутов на точность работы алгоритмов.

Ключевые слова

многоязычные базы знаний, выравнивание сущностей, векторное представление, языковые модели, упорядочивание матриц, графы знаний

Для цитирования

Гусев Д. И., Апанович З. В. Влияние методов построения векторных представлений на подходы выравнивания сущностей // Вестник НГУ. Серия: Информационные технологии. 2022. Т. 20, № 2. С. 5–17. DOI 10.25205/1818-7900-2022-20-2-5-17

Influence of Embeddings Construction Methods on Entity Alignment Approaches

Daniil I. Gusev[1], Zinaida V. Apanovich[2]

[1, 2] Novosibirsk State University
Novosibirsk, Russian Federation

[2] A.P. Ershov Institute of Informatics Systems of the Siberian Branch of the Russian Academy of Sciences
Novosibirsk, Russian Federation

[1]apanovich_09@mail.ru, https://orcid.org/0000-0002-5767-284X
[2]d.gusev1@g.nsu.ru, https://orcid.org/0000-0001-9636-2783

Abstract

The problem of merging knowledge graphs (KG) presented in different languages is becoming more and more relevant. The main stage for its solution is the identification of equivalent entities and their descriptions. It is also known as the entity alignment problem. The recent research shows that existing approaches are not effective for all languages. This article presents the experiments aimed at improving the alignment of entities on an English-Russian dataset. The results obtained are considered from the point of view both of the whole graph and of individual types of entities. The influence of the number of relations and attributes on the accuracy of the algorithms is estimated.

Keywords
multilingual knowledge bases, entity alignment, embedding, language models, matrix ordering, knowledge graphs

For citation
Gusev D. I., Apanovich Z. V. Decision Support Systems Utilization in Forestry: Environmental Aspect. Vestnik NSU. Series: In-formation Technologies, 2022, vol. 20, no. 2, p. 5–17. (in Russ.) DOI 10.25205/1818-7900-2022-20-2-5-17

Введение

Графы знаний являются современной формой представления информации о мире. Они состоят из уникальных сущностей и связей между ними. Для представления фактов используются триплеты (subject entity, relation, object entity) или (subject entity, attribute, literal value). Первый тип называется реляционным и используется для описания отношений между сущностями. Его примером для факта «Новосибирск – часть России» является (dbr:Novosibirsk, dbo:country, dbr:Russia). Второй тип называется атрибутным и используется для описания свойств сущности. Его примером для факта «Новосибирск был основан в 1893 году» является (dbr:Novosibirsk, dbp:establishedDate, “1893”^^xsd:integer). Где «dbp:establishedDate» – атрибут (свойство), а «“1893”^^xsd:integer» – литерал (значение).

К примерам применения графов знаний относят системы рекомендации контента, обнаружение лекарств, анализ инвестиционного рынка, семантический поиск и так далее. При этом чем мощнее базовый граф знаний, тем выше качество основанных на нем приложений.

Дополнить граф знаний можно путём объединения с другими. Достигается это при помощи поиска сущностей в графах знаний, которые ссылаются на один и тот же объект реального мира. Примером является сущность «Austria» в англоязычном графе и «Австрия» в русскоязычном. Данное направление получило название выравнивание сущностей (entity alignment). В ряде литературы может упоминаться как сопоставление сущностей (entity matching).

С недавнего времени широкое распространение получили подходы выравнивания сущностей (EA) на основе векторных представлений (embeddings). Идея состоит в получении символьных описаний KG в виде векторов низкой размерности, таким образом, чтобы семантическая взаимосвязь сущностей захватывалась геометрическими структурами векторного пространства [1]. Потенциально это может смягчить лингвистическую и схематическую неоднородность между независимо созданными графами знаний.

В качестве стратегий выравнивания выделяют следующие: (а) модуль построения векторного представления кодирует два графа знаний в два независимых пространства, в то время как модуль выравнивания изучает сопоставления между ними [2]; (б) модуль выравнивания направляет модуль построения векторного представления для отображения двух графов знаний в одно унифицированное пространство [3].

Для анализа результатов подходов выравнивания сущностей применяют метрики Hits@k и MRR. Hits@k – означает, что сущности из первого графа знаний и эквивалентные сущности из второго графа знаний находятся среди ближайших k соседей. При этом метрика Hits@1 считается наиболее показательной, поскольку эквивалентна точности. MRR (Mean Reciprocal Rank – среднеобратный ранг) представляет среднее обратных значений номеров правильных ответов в списке предполагаемых сущностей. Его можно рассматривать как мягкую версию Hits@1, которая менее чувствительна к выбросам [4]. Для обеих метрик значения лежат от 0 до 1, где большее число говорит о лучшей точности.

Обширное исследование подходов выравнивания сущностей было проведено на англо-французском наборе данных [5]. Хорошие результаты показали MultiKE и RDGCN. Однако на англо-русском наборе данных у них наблюдается значительное снижение точности [6]. Для исследования данной закономерности были изучены подходы выравнивания сущностей и проведен ряд экспериментов. Применение полученных сведений привело к увеличению точности работы алгоритмов.

1. Группы алгоритмов выравнивания сущностей на основе векторных представлений

Большинство алгоритмов 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 или в различные векторные пространства при помощи предварительно выровненных сущностей (seed alignments).

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 в основном опираются на структурную информацию в графах знаний, то есть реляционные триплеты. Основу этих подходов 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 [7]. 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. 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. Данные алгоритмы выдают очень неплохие результаты, но их основным недостатком является чрезвычайная сложность, значительное время вычислений и плохая интерпретируемость. Представителем этого подхода является RDGCN [8]. 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 (GAT). Результирующие векторные представления исходных графов затем подаются в графовые сверточные сети (GCN), для извлечения информации о структуре окружений вершин.

Совсем недавно появился чрезвычайно простой подход к выравниванию сущностей под названием SEU [9] (Simple but Effective Unsupervised EA method), 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 подходы EA не способны выдавать удовлетворительные результаты только на основании реляционных триплет, если набор данных имеет распределение степеней сущностей, близкое к реальным KG. В частности, известно, что примерно половина сущностей в реальных KG связана с менее чем тремя другими сущностями [10].

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 предложений или отдельных слов (word2vec, glove). 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 (fastText, name-BERT).

Ранее упомянутые подходы выравнивания сущностей также имеют свои собственные методы построения векторных представлений имён сущностей. Их основными этапами являются: чтение предобученной модели, токенизация данных и формирование векторных представлений. Однако имеются различия в способах обработки нераспознанных слов и объединения векторов.

Далее приведены основные особенности методов построения векторных представлений имён сущностей из рассматриваемых подходов. Для дальнейшего обозначения они также были пронумерованы.

Метод 1 применяется в MultiKE. В качестве предобученной модели используется wiki-news-300d-1m. Вектора нераспознанных слов формируются путем суммирования векторов символов, полученных при помощи word2vec. Для объединения векторов слов применяется автокодировщик на основе нейронной сети.

Метод 2 применяется в RDGCN. В качестве предобученной модели также используется wiki-news-300d-1m. Входные данные очищаются от специальных символов. Для нераспознанных слов присваиваются нулевые вектора. Объединение векторов слов основано на суммировании.

Метод 3 применяется в SEU. В качестве предобученной модели используется glove.6b.300d. Входные данные приводятся к нижнему регистру. Вектора нераспознанных слов задаются случайным образом. В дополнении к векторам слов применяются вектора биграмм. Объединение векторов слов достигается путем вычисления среднего арифметического.

2. Эксперименты с различными способами построения векторных представлений имен сущностей

2.1. Наборы данных

Из-за сложности запуска подходов на полных графах знаний были отобраны пары сущностей в объеме пятнадцати тысяч. Для их формирования использовался алгоритм IDS [5]. Он одновременно удаляет сущности в двух графах знаний с выравниванием по межъязыковым ссылкам до достижения желаемого размера, сохраняя при этом распределение степеней, аналогичное исходным KG.

Сформированные наборы данных представлены двумя версиями. Результат прямого применения IDS обозначен V1. Вдвое более плотный набор обозначен V2. Для его генерации предварительно случайным образом были удалены сущности с количеством связей меньше пяти, после чего применен IDS.

Источниками являются разноязычные версии DBpedia[1]. В качестве целевых были выбраны англо-французские и англо-русские кросс-языковые наборы данных. Их статистика пред-

[1] URL: https://wiki.dbpedia.org/downloads-2016-10/

ставлена в табл. 1. DBP-15K EN-FR (V1, V2) взят из библиотеки OpenEA. Набор DBP-15K EN-RU (V1, V2) сгенерирован по тем же принципам и доступен для свободного скачивания[2].

Таблица 1

Статистика используемых наборов данных

Table 1

Statistics of used datasets

Набор данных	KG	15K (V1)				15K (V2)			
		Rel.	Att.	Rel tr.	Att tr.	Rel.	Att.	Rel tr.	Att tr.
EN-FR	EN	267	308	47334	73121	193	189	96318	66898
	FR	210	404	40864	67167	166	221	80112	68778
EN-RU	EN	163	173	43796	76959	141	147	76617	75135
	RU	66	52	30489	54517	57	46	56399	56455

2.2. Влияние машинного перевода имен сущностей на результаты 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. В случае объединения английских и русских слов в единое векторное представление имеет смысл 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. В качестве целевого языка выбран английский. Затем происходит разделение входных данных на пакеты по 3500 символов. Это связано с ограничением Google Translate API. Далее каждый пакет конвертируется в строки, а имена отделяются друг от друга для исключения попадания в контекст. После чего за счет обращения к стороннему серверу происходит перевод пакетов и восстановление исходной последовательности данных. Результат передается в метод формирования векторного представления.

Столбец «Разница» в табл. 2 показывает изменение точности по метрике Hits@1 в зависимости от применения машинного перевода. У англо-французского графа знаний наблюдается небольшое изменение во всех подходах.

У англо-русского графа знаний заметно существенное увеличение точности. Наибольшее влияние машинный перевод оказал на SEU. Это говорит о том, что данный подход в значительной степени опирается на текстовые особенности имён сущностей. Для RDGCN удалось достичь показателей, схожих с англо-французским графом знаний. Применение перевода также привело к увеличению точности MultiKE. Однако у данного подхода наблюдается наименьшее изменение среди представленных.

[2] URL: https://www.dropbox.com/sh/4oh3nkzwdr1w4dv/AACZ4v8jCdR7Y4mDtS654Bega?dl=0

Таблица 2

Влияние перевода имен сущностей на качество EA на различных наборах данных

Table 2

The effect of entity name translation on the quality of EA on different datasets

Подход	Данные	Перевод	Hits@1	Hits@10	MRR	Разница
MultiKE	EN-FR-15K (V1)	–	0,741	0,836	0,774	
MultiKE	EN-FR-15K (V1)	+	0,806	0,885	0,835	0,065
MultiKE	EN-FR-15K (V2)	–	0,855	0,921	0,878	
MultiKE	EN-FR-15K (V2)	+	0,893	0,956	0,915	0,038
MultiKE	EN-RU-15K (V1)	–	0,315	0,457	0,364	
MultiKE	EN-RU-15K (V1)	+	0,520	0,666	0,570	0,205
MultiKE	EN-RU-15K (V2)	–	0,453	0,623	0,510	
MultiKE	EN-RU-15K (V2)	+	0,617	0,770	0,670	0,164
RDGCN	EN-FR-15K (V1)	–	0,770	0,892	0,813	
RDGCN	EN-FR-15K (V1)	+	0,771	0,893	0,813	0,001
RDGCN	EN-FR-15K (V2)	–	0,862	0,948	0,895	
RDGCN	EN-FR-15K (V2)	+	0,871	0,951	0,903	0,009
RDGCN	EN-RU-15K (V1)	–	0,396	0,597	0,460	
RDGCN	EN-RU-15K (V1)	+	0,744	0,882	0,792	0,347
RDGCN	EN-RU-15K (V2)	–	0,537	0,717	0,599	
RDGCN	EN-RU-15K (V2)	+	0,844	0,923	0,882	0,307
SEU	EN-FR-15K (V1)	–	0,989	0,998	0,992	
SEU	EN-FR-15K (V1)	+	0,995	1,000	0,997	0,006
SEU	EN-FR-15K (V2)	–	0,992	0,999	0,994	
SEU	EN-FR-15K (V2)	+	0,996	1,000	0,997	0,004
SEU	EN-RU-15K (V1)	–	0,301	0,348	0,318	
SEU	EN-RU-15K (V1)	+	0,972	0,995	0,981	0,672
SEU	EN-RU-15K (V2)	–	0,424	0,483	0,445	
SEU	EN-RU-15K (V2)	+	0,990	0,998	0,993	0,566

2.3. Варианты перевода в MultiKE

В процессе изучения MultiKE было выявлено, что произвести перевод у данного подхода можно следующими способами: до метода генерации векторных представлений, до метода генерации векторных представлений с предварительным отключением автокодировщика и для нераспознанных в модели слов. В последнем случае значения векторов также уточняются при помощи автокодировщика.

При отключении автокодировщика вектора слов, входящие в литерал, объединялись путем вычисления среднего арифметического значения. Согласно табл. 3, это привело к небольшому снижению точности выравнивания сущностей. При этом время обучения сократилось более чем в два раза. Следовательно, в целях упрощения данного метода генерации векторных представлений можно отказаться от применения автокодировщика.

Таблица 3

Эксперименты с переводом в MultiKE на наборе EN-RU-15K (V1)

Table 3

Experiments with translation to MultiKE on EN-RU-15K (V1)

Применение перевода	Hits@1	Hits@5	Hits@10	Hits@50	MRR	Время (с.)
До метода генерации	0,520	0,621	0,666	0,769	0,570	5312
До метода и без автокодировщика	0,470	0,532	0,561	0,648	0,502	2247
Для нераспознанных слов	0,356	0,460	0,511	0,638	0,409	3230

2.4. Визуализация результатов методов построения представлений имен сущностей

Для сравнения методов построения векторных представлений 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, расположены совместно. Наличие одноцветных зон свидетельствует о том, что метод не смог установить кросс-языковое соответствие.

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

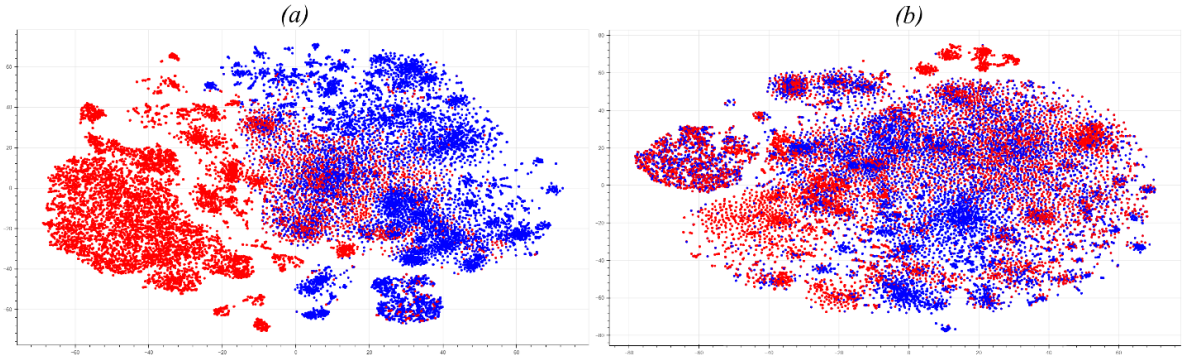

Рис. 1. Векторные представления имён сущностей метода 1: *a* – без перевода; *b* – с переводом
Fig. 1. The embeddings of entity names of method 1: *a* – without translation; *b* – with translation

В векторном представлении имён сущностей из метода 2 (рис. 2, *a*) в левом нижнем углу имеется кластер эллипсоидной формы. Он возник из-за зануления векторов слов, для которых не были найдены значения в предобученной модели. В остальном же данное векторное представление имеет большую степень наложения по сравнению с методом 1. Наименьшее количество языковых кластеров наблюдается у результата, полученного при помощи метода генерации 3 (рис. 2, *b*).

В качестве альтернативных методов генерации векторных представлений были выбраны современные модели обработки естественных языков.

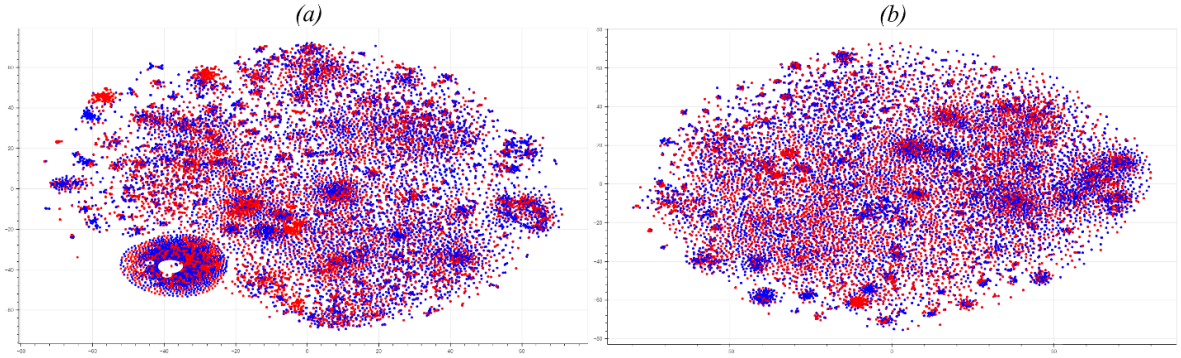

Рис. 2. Векторные представления имён сущностей из различных методов: *a* – метод 2; *b* – метод 3
Fig. 2. The embeddings of entity names from various methods: *a* – method 2; *b* – method 3

XLNet. Целью модели является изучение распределений для всех перестановок слов в заданной последовательности [11]. 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 путем объединения возможностей маскированного и кросс-языкового моделирования [12].

Векторное представление XLNet (рис. 3, *a*) имеет крайне низкую степень наложения. Большинство семантически связанных имён сущностей находятся в отдалении друг от друга. Противоположная картина наблюдается у LaBSE (рис. 3, *b*). Модель смогла сопоставить имена сущностей без применения перевода.

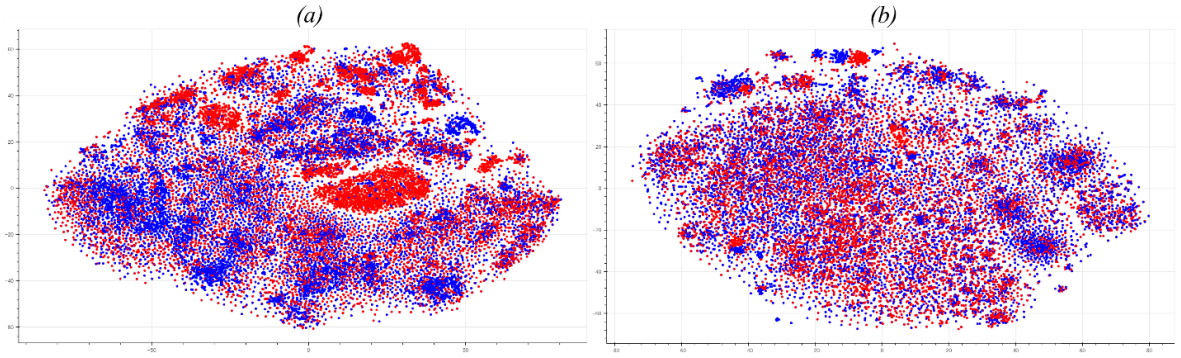

Рис. 3. Векторные представления имён сущностей моделей обработки естественных языков:
a – XLNet; *b* – LaBSE
Fig. 3. The embeddings of entity names of natural language processing models: *a* – XLNet; *b* – LaBSE

2.5. Влияние методов построения представлений на результаты подходов

Помимо перевода русскоязычных имен сущностей на английский язык были проведены эксперименты с несколькими моделями построения векторных представлений имен сущностей.

По данным табл. 4 видно, что метод 3 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. Хорошо себя показал LaBSE. Он оказался эффективнее методов 1 и 2.

Таблица 4

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

Table 4

Results of EA approaches depending on the method of generating embeddings of entity names on the EN-RU-15K (V1)

Подход	Метод	Hits@1	Hits@5	Hits@10	Hits@50	MRR
MultiKE	1	0,520	0,621	0,666	0,769	0,570
MultiKE	2	0,699	0,781	0,813	0,878	0,737
MultiKE	3	**0,812**	**0,875**	**0,891**	**0,932**	**0,841**
RDGCN	1	0,680	0,796	0,828	0,884	0,733
RDGCN	2	0,744	0,847	0,882	0,923	0,792
RDGCN	3	**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	1	0,881	0,935	0,948	0,975	0,905
SEU	2	0,874	0,931	0,954	0,986	0,905
SEU	3	**0,972**	**0,991**	**0,995**	**0,998**	**0,981**
SEU	XLNet	0,325	0,413	0,455	0,549	0,369
SEU	LaBSE	0,949	0,976	0,984	0,993	0,962

3. Изучение результатов

В целях получения более подробной информации о результатах подходов была произведена оценка по типам сущностей, количеству отношений и атрибутов.

Суммарно выделено семьдесят три типа сущностей в наборе данных EN-RU-15K (V1). Для их получения использовались файлы «instance_types». Однако в них, как и в исходных графах знаний, наблюдается неоднородность. В частности, было выявлено, что только 67 % отобранных пар сущностей имеют соответствие по типу. Пример несоответствия выглядит следующим образом: сущность «Эминем» относится к «MusicalArtist», а ее английский эквивалент «Eminem» относится к вышестоящему по иерархии типу «Person».

Для решения данной проблемы была написана программа установления общих типов для сущностей из пары. Модель онтологии получена при помощи SPARQL запроса к англоязычной DBpedia. Парам, в которых одна сущность относится к подтипу другой, назначается подтип. Это делается из соображений о том, что они представляют более уникальную информацию. В случае, когда ни одна сущность из пары не является подтипом другой, но при этом у них имеется общий тип, назначается последний. Например, сущность «Воеводина» относится к «AdministrativeRegion», а ее английский эквивалент «Vojvodina» относится к «Country». Данным сущностям присвоится общий тип «PopulatedPlace». Как результат, для всех парных сущностей удалось получить соответствие по типам.

В качестве исследуемых данных выбраны лучшие результаты подходов. Они получены на основе метода генерации векторных представлений под номером 3. На рис. 4 представлена оценка точности Multike для типов с количеством сущностей больше 100. Значения указа-

ны по метрике Hits@1. На изображении видно, что точность выравнивания отличается у различных типов сущностей. Для изучения данного явления мы вывели количество отношений и атрибутов сущностей.

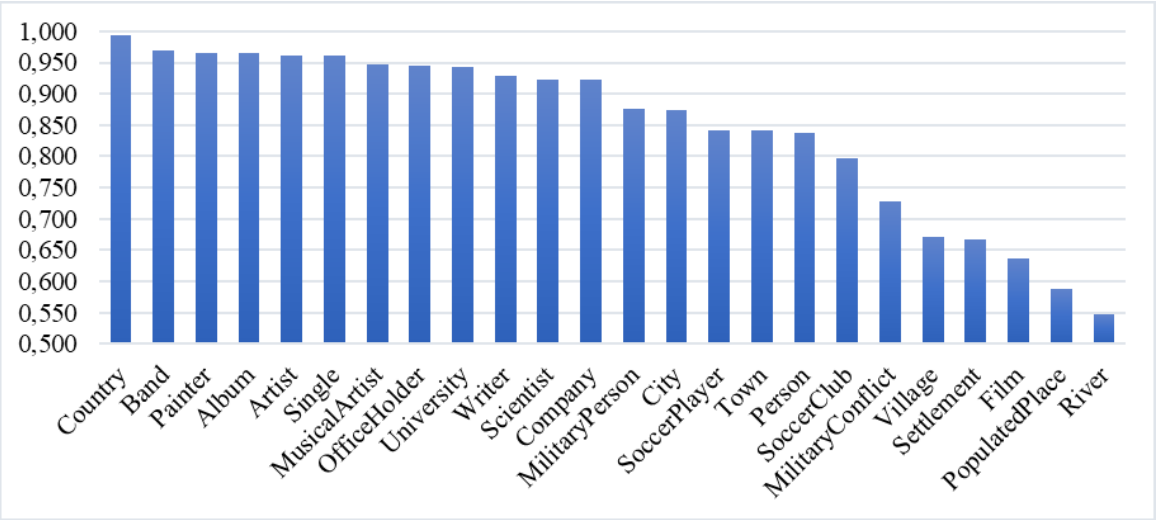

Рис. 4. Оценка точности Multike по типам сущностей
Fig. 4. Evaluation of Multike accuracy by entity types

По данным, представленным в табл. 5, можно сделать вывод о том, что большое количество связей сущностей не гарантирует высокую точность выравнивания. При анализе результатов других подходов была выявлена схожая закономерность. Скорее всего, значение имеют отдельные виды отношений. Данное направление требует дополнительного изучения.

Таблица 5

Значения MultiKE для типа сущностей «Country»

Table 5

MultiKE values for the «Country» entity type

Имя сущности KG-1	Имя сущности KG-2	Расстояние	Отношения KG-1	Отношения KG-2	Атрибуты KG-1	Атрибуты KG-2
Nazi Germany	Третий рейх	0,6655	17	22	8	10
El Salvador	Сальвадор	0,3316	14	13	14	1
Cyprus	Республика Кипр	0,2048	15	10	14	1
Czech Republic	Чехия	0,2048	45	42	13	1
Republic of Ireland	Ирландия	0,1650	36	43	15	1
Turkmenistan	Туркмения	0,1330	22	7	14	1
Iceland	Исландия	0,0030	26	17	13	1
Serbia	Сербия	0,0024	68	19	14	1
Luxembourg	Люксембург	0,0021	29	22	14	1
Roman Empire	Римская империя	0,0018	23	15	6	6
Equatorial Guinea	Экваториальная Гвинея	0,0015	9	1	13	1
Nicaragua	Никарагуа	0,0013	14	12	16	1

Заключение

В данной работе мы изучили влияние методов построения векторных представлений для имён сущностей и литералов на результаты подходов. Был исследован вклад применения перевода и современных моделей обработки естественных языков. Полученные значения метрик превзошли значения, указанные в исходных публикациях. При расширенном анализе результатов не удалось установить прямой зависимости между точностью выравнивания сущностей и количеством связей. Данное направление требует дальнейшего изучения. Следует заметить, что полученные результаты имеют достаточно «идеальный» характер, так как проводились на наборе данных, в котором у каждой сущности англоязычного KG соответствует эквивалентная сущность русскоязычного KG. Эта ситуация весьма далека от реального положения вещей и требует дальнейшего исследования.

Список литературы

1. **Bordes A., Usunier N., Garcia-Durán A., Weston J., Yakhnenko O.** Translating embeddings for modeling multi-relational data. *Proceedings of the 26th International Conference on Neural Information Processing Systems*, 2013, vol. 2, p. 2787–2795. DOI 10.5555/2999792.2999923.
2. **Chen M., Tian Y., Chang K., Skiena S., Zaniolo C.** Co-training Embeddings of Knowledge Graphs and Entity Descriptions for Cross-lingual Entity Alignment. *Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018, p. 3998–4004. DOI 10.24963/ijcai.2018/556.
3. **Xu K., Wang L., Yu M., Feng Y., Song Y., Wang Z., Yu D.** Cross-lingual Knowledge Graph Alignment via Graph Matching Neural Network. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, p. 3156–3161. DOI 10.18653/v1/P19-1304.
4. **Rossi A., Barbosa D., Firmani D., Matinata A., Merialdo P.** Knowledge Graph Embedding for Link Prediction: A Comparative Analysis. *ACM Transactions on Knowledge Discovery from Data*, 2021, vol. 15, p. 1–49. DOI 10.1145/3424672.
5. **Sun Z., Zhang Q., Hu W., Wang C., Chen M., Akrami F., Li C.** A benchmarking study of embedding-based entity alignment for knowledge graphs. *Proceedings of the VLDB Endowment*, 2020, vol. 13, p. 2326–2340. DOI 10.14778/3407790.3407828.
6. **Gnezdilova V. A., Apanovich, Z. V.** Russian-English dataset and comparative analysis of algorithms for cross-language embeddingbased entity alignment. *Journal of Physics: Conference Series*, 2021, vol. 2099. DOI 10.1088/1742-6596/2099/1/012023.
7. **Zhang Q., Sun Z., Hu W., Chen M., Guo L., Qu Y.** Multi-view Knowledge Graph Embedding for Entity Alignment. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019, p. 5429–5435. DOI 10.24963/ijcai.2019/754.
8. **Wu Y., Liu X., Feng Y., Wang Z., Yan R., Zhao D.** Relation-Aware Entity Alignment for Heterogeneous Knowledge Graphs. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019, p. 5278–5284. DOI 10.24963/ijcai.2019/733.
9. **Mao X., Wang W., Wu Y., Lan M.** From Alignment to Assignment: Frustratingly Simple Unsupervised Entity Alignment. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, p. 2843–2853. DOI 10.18653/v1/2021.emnlp-main.226.
10. **Guo L., Sun Z., Hu W.** Learning to Exploit Long-term Relational Dependencies in Knowledge Graphs. Proceedings of the 36th International Conference on Machine Learning, 2019, vol. 57, p. 2505–2514.
11. **Yang Z., Dai Z., Yang Y., Carbonell J., Salakhutdinov R., Le Q.** XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, p. 5753–5763. DOI 10.5555/3454287.3454804.

12. **Feng F., Yang Y., Cer D., Arivazhagan N., Wang W.** Language-agnostic BERT Sentence Embedding. *ArXiv*, 2020. DOI 10.48550/arXiv.2007.01852.

References

1. **Bordes A., Usunier N., Garcia-Durán A., Weston J., Yakhnenko O.** Translating embeddings for modeling multi-relational data. *Proceedings of the 26th International Conference on Neural Information Processing Systems*, 2013, vol. 2, p. 2787–2795. DOI 10.5555/2999792.2999923.
2. **Chen M., Tian Y., Chang K., Skiena S., Zaniolo C.** Co-training Embeddings of Knowledge Graphs and Entity Descriptions for Cross-lingual Entity Alignment. *Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018, p. 3998–4004. DOI 10.24963/ijcai.2018/556.
3. **Xu K., Wang L., Yu M., Feng Y., Song Y., Wang Z., Yu D.** Cross-lingual Knowledge Graph Alignment via Graph Matching Neural Network. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, p. 3156–3161. DOI 10.18653/v1/P19-1304.
4. **Rossi A., Barbosa D., Firmani D., Matinata A., Merialdo P.** Knowledge Graph Embedding for Link Prediction: A Comparative Analysis. *ACM Transactions on Knowledge Discovery from Data*, 2021, vol. 15, p. 1–49. DOI 10.1145/3424672.
5. **Sun Z., Zhang Q., Hu W., Wang C., Chen M., Akrami F., Li C.** A benchmarking study of embedding-based entity alignment for knowledge graphs. *Proceedings of the VLDB Endowment*, 2020, vol. 13, p. 2326–2340. DOI 10.14778/3407790.3407828.
6. **Gnezdilova V. A., Apanovich, Z. V.** Russian-English dataset and comparative analysis of algorithms for cross-language embeddingbased entity alignment. *Journal of Physics: Conference Series*, 2021, vol. 2099. DOI 10.1088/1742-6596/2099/1/012023.
7. **Zhang Q., Sun Z., Hu W., Chen M., Guo L., Qu Y.** Multi-view Knowledge Graph Embedding for Entity Alignment. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019, p. 5429–5435. DOI 10.24963/ijcai.2019/754.
8. **Wu Y., Liu X., Feng Y., Wang Z., Yan R., Zhao D.** Relation-Aware Entity Alignment for Heterogeneous Knowledge Graphs. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019, p. 5278–5284. DOI 10.24963/ijcai.2019/733.
9. **Mao X., Wang W., Wu Y., Lan M.** From Alignment to Assignment: Frustratingly Simple Unsupervised Entity Alignment. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, p. 2843–2853. DOI 10.18653/v1/2021.emnlp-main.226.
10. **Guo L., Sun Z., Hu W.** Learning to Exploit Long-term Relational Dependencies in Knowledge Graphs. Proceedings of the 36th International Conference on Machine Learning, 2019, vol. 57, p. 2505–2514.
11. **Yang Z., Dai Z., Yang Y., Carbonell J., Salakhutdinov R., Le Q.** XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, p. 5753–5763. DOI 10.5555/3454287.3454804.
12. **Feng F., Yang Y., Cer D., Arivazhagan N., Wang W.** Language-agnostic BERT Sentence Embedding. ArXiv, 2020. DOI 10.48550/arXiv.2007.01852.

Сведения об авторах

Гусев Даниил Иванович, студент магистратуры, Новосибирский государственный университет (Новосибирск, Россия)

Апанович Зинаида Владимировна, старший научный сотрудник, Институт систем информатики им. А. П. Ершова СО РАН (Новосибирск, Россия)

Information about the Authors

Daniil I. Gusev, master's student, Novosibirsk State University (Novosibirsk, Russian Federation)
Zinaida V. Apanovich, senior researcher, A. P. Ershov Institute of Informatics Systems, Siberian Branch of the Russian Academy of Sciences (Novosibirsk, Russian Federation)

Статья поступила в редакцию 11.05.2022;
одобрена после рецензирования 09.06.2022; принята к публикации 09.06.2022
The article was submitted 11.05.2022;
approved after reviewing 09.06.2022; accepted for publication 09.06.2022