Эксперименты по выравниванию сущностей на русско-английском наборе данных с несопоставимыми сущностями

З.В. Апанович[1], Д.Г. Керного[2]

[1] ИСИ СО РАН
[2] НГУ

Аннотация. В последние годы кратно возрос интерес к графам знаний (ГЗ) как в научном, так и в промышленном сообществе. ГЗ играют важную роль в таких приложениях ИИ как обработка текстов на естественном языке, включая вопросно-ответные системы, рекомендательные системы и поисковые движки. Интеграция различных графов знаний является одной из актуальнейших задач и используется, например, для разработки сложных цифровых двойников промышленных систем. Одной из компонент решения задачи интеграции ГЗ является задача выравнивания сущностей (ВС), пытающаяся идентифицировать в разных ГЗ сущности, описывающие один и тот же объект реального мира. Частным случаем этой задачи является задача кросс-языковой идентификации сущностей, решение которой требуется для импортозамещения, например поиска эквивалентных лекарств, запчастей или устройств для Интернета Вещей. К сожалению, в реальных графах знаний многие сущности могут не иметь эквивалентов в других графах знаний. В данной работе описаны эксперименты по выравниванию сущностей при наличии несопоставимых сущностей на примере русско-английского набора данных

Ключевые слова: графы знаний, выравнивание сущностей, несопоставимые сущности

Entity alignment experiments on Russian-English dataset with unmatchable entities

Z.V. Apanovich[1], D.G. Kernogo[2]

[1] IIS SBRAS
[2] NSU

Abstract. In recent years, interest in knowledge graphs (KG) has increased exponentially in both the scientific and industrial communities. KGs play an important role in AI applications such as natural language processing including

question-answering systems, recommender systems, and search engines. Integration of different KGs is one of the most pressing problems and is used, for example, to develop complex digital twins of industrial systems. One of the components of the KG integration problem is the entity alignment problem, which attempts to identify entities in different KGs that describe the same real-world object. A special case of this problem is the problem of cross-language entity alignment (EA), which is closely related to the problem of import substitution, such as finding equivalent drugs, spare parts, or devices for the Internet of Things. Unfortunately, in real KGs, many entities may not have equivalents in another KG. This paper describes entity alignment experiments using the example of a Russian-English dataset with unmatchable entities.

Keywords: knowledge graph, entity alignment, unmatchable entities

1. Введение

Графы знаний хранят факты об объектах реального мира в виде реляционных и литеральных триплет. Реляционные триплеты изображают отношение между двумя сущностями (то есть, объектами реального мира, имеющими уникальный Интернет-идентификатор, IRI) и имеют формат *tr_r* = (*субъектная сущность, отношение, объектная сущность).* Литеральные триплеты хранят информацию об атрибутах сущностей и имеют формат *tr_l* = (*субъектная сущность, атрибут, литеральное значение*).

На рисунке 1 показаны фрагменты из англоязычной и русскоязычной версий DBpedia, описывающие Новосибирский государственный университет. Примером реляционной триплеты является триплета (*Новосибирский_государственный_университет*, *ректор*, *М._П._Федорук*), а примером литеральной триплеты является триплета (*Новосибирский_государственный_университет*, *основан*, *1959*). На рисунке красными линиями изображены отношения эквивалентности *owl:sameAs*, имеющиеся между сущностями в русскоязычном и англоязычном графах знаний. Можно видеть, что англоязычной сущности *dbr:Fedoruk,_Mikhail_Petrovich* в англоязычном графе знаний соответствует русскоязычная сущность *М._П._Федорук,* англоязычному предикату *dbo:rector* – русскоязычный предикат *ректор*, а англоязычной сущности *dbr:Novosibirsk_State_University* – русскоязычная сущность *Новосибирский государственный университет*. Понятно, что в идеале англоязычный и русскоязычный списки сотрудников, ректоров и студентов Новосибирского университета должны бы совпадать в разных графах знаний. Однако на практике эти списки очень неполные и поэтому могут сильно отличаться, а также наблюдаются некоторые различия в описаниях эквивалентных сущностей.

Прежде всего, следует обратить внимание на разное написание имен в русскоязычной и англоязычной версиях графов знаний. Также, в англоязычной и русскоязычной версии указаны разные годы основания университета. Кроме этого, в англоязычной и русскоязычной версии графов

знаний приведены разные списки людей, которые связаны (то есть, учились или работали) с НГУ. Например, в обеих версиях описан выпускник НГУ Ефим Зельманов, получивший Филдсовскую премию по математике. Но только в англоязычной версии есть информация про то, что выпускником НГУ был и Андрей Воронков, не только получивший премию Эрбрана (*dbr:Herbrand_Award*), но и являющийся разработчиком программы EasyChair (*dbr: EasyChair*), которой пользуются научные сотрудники во всем мире для представления статей на научные конференции. Поскольку описания объектов реального мира, таких как Андрей Воронков, и премия Эрбрана в момент написания данной статьи отсутствуют в русскоязычной версии DBpedia, при попытке сопоставления англоязычной и русскоязычной версий возникают так называемые *висячие* или *несопоставимые* сущности *dbr:Andrei_Voronkov* и *dbr:Herbrand_Award*.

Понятно, что наиболее полное описание сущности можно получить объединением всех триплет, описывающих одну и ту же сущность в разных ГЗ. Но для решения этой задачи должны быть правильно установлены соответствия между сущностями.

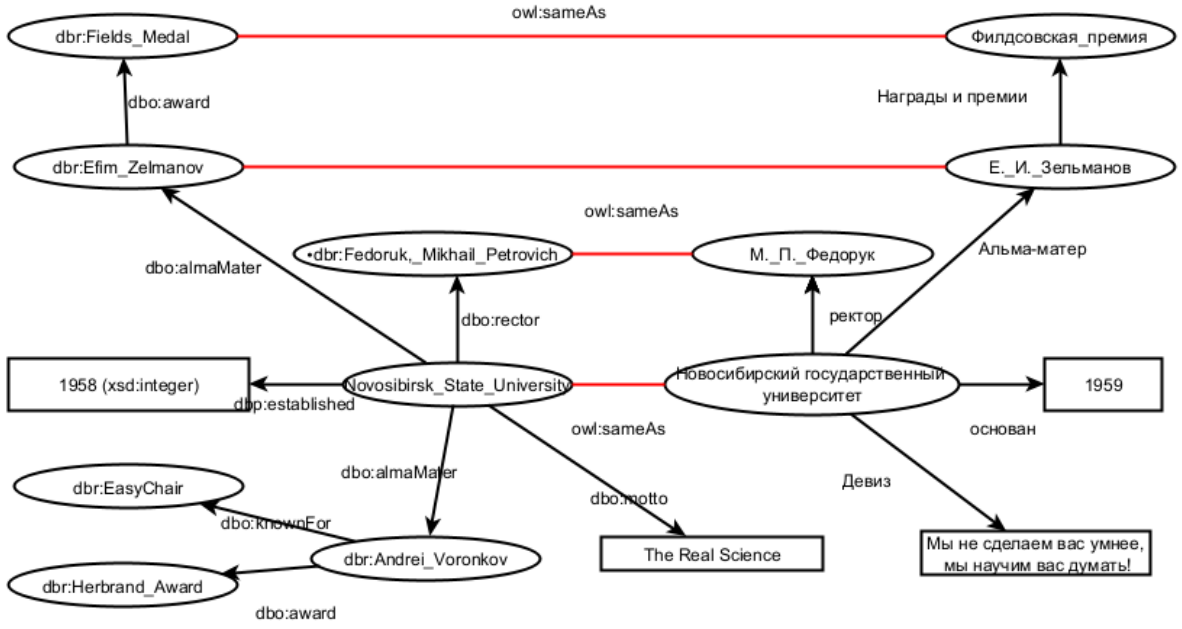

Рис. 1. Соответствие между сущностями в англоязычном и русскоязычном графах знаний.

Современные методы ВС основаны на предположении, что эквивалентные сущности имеют похожие окружения. Поэтому активно используются методы, генерирующие векторные представления заданной размерности для сущностей и отношений графов знаний, так называемые *вложения* или *эмбеддинги* (embeddings). Поскольку в русскоязычной литературе, связанной с графами знаний, пока нет устоявшейся

терминологии, в дальнейшем мы будем использовать такие термины, как *векторное представление, вложение, и эмбеддинг* взаимозаменяемо.

Достоинствами подхода на основе эмбеддингов являются высокая масштабируемость и небольшие усилия при подготовке обучающих выборок.

Понятно, что русскоязычному пользователю интересны, прежде всего, эксперименты, использующие русскоязычные данные. В работе [1] описан русско-английский набор данных для экспериментов с алгоритмами кросс-языкового выравнивания сущностей. В работе [2] показано, что качество выравнивания сущностей можно существенно улучшить, улучшая качество построения эмбеддингов для имен сущностей. Также были найдены наилучшие комбинации методов построения эмбеддингов для имен сущностей.

2. Выравнивание сущностей при наличии несопоставимых сущностей

Метод выравнивания сущностей на основе векторных представлений состоит как правило из двух компонент: вычисления векторных представлений для сущностей, принадлежащих разным графам знаний, и сопоставления этих векторных представлений. Вычисление векторных представлений для сущностей из двух графов знаний осуществляется раздельно, поэтому эти представления могут попасть в разные векторные пространства. Значит, необходимо их собрать в едином векторном пространстве, что делается при помощи так называемых «seed alignments», которые содержат пары эквивалентных сущностей в двух графах знаний. Расстояние между парами выровненных сущностей вычисляются при помощи таких функций как косинусная близость, манхэттеново или евклидово расстояние и др.

До недавнего времени решения ВС предполагали, что каждая сущность в исходном ГЗ имеет эквивалентную сущность в целевом ГЗ. Поэтому эта эквивалентная сущность искалась как ближайший сосед целевой сущности в пространстве вложения.

Таблица 1. Характеристики русско-английского набора данных.

Язык ГЗ	Сущности	Отношения	Атрибуты	Кол-во рел. триплет	Кол-во атриб. триплет
Ru	15000	66	15018	30489	54499
En	15000	163	15106	43796	76852

На практике, всегда существуют несопоставимые сущности. Например, один из самых больших графов знаний wikidata.org содержит эквивалентные сущности из многих других наборов данных, таких как viaf.org, ror.org, imdb.com, и др, в то время как каждый из перечисленных наборов данных имеет сущности, отсутствующие в других ГЗ. Поэтому идеальная система ВС должна быть способна обнаруживать и обрабатывать несопоставимые сущности.

Также необходимы специальные наборы данных для работы с этой проблемой. Основной вызов связан с тем, что надо гарантировать, что «висячие» сущности действительно не имеют эквивалентных сущностей во втором наборе данных. Сначала создается два подграфа, в которых каждая сущность имеет эквивалент во втором графе, а затем случайным образом удаляется два непересекающихся подмножества сущностей из двух графов знаний. Эквиваленты удаленных сущностей становятся несопоставимыми сущностями. Структура набора данных с несопоставимыми сущностями показана на рисунке 2. В качестве стартовых наборов триплет использовался русско-английский набор данных [1].

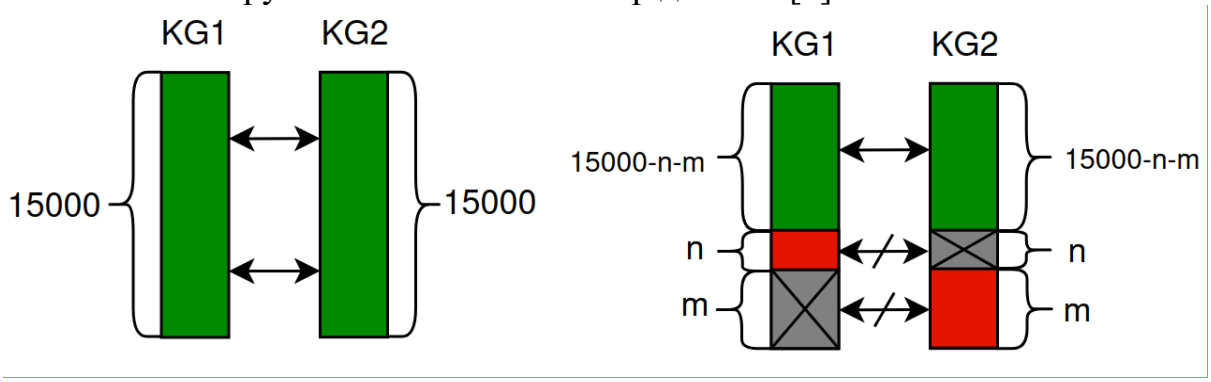

Рис. 2. Построение набора данных с несопоставимыми сущностями

Для экспериментов с несопоставимыми сущностями использовались два фреймворка: EntMatcher [3] и UEA [4]. Достоинством фреймворка EntMatcher является большой набор алгоритмов построения векторных представлений, а также множество стратегий установления соответствия между сущностями из различных графов знаний. Но этот фреймворк не предназначен для работы с несопоставимыми сущностями. С другой стороны, фреймворк UEA изначально создан для работы с несопоставимыми сущностями, но его набор методов построения векторных представлений и способов установления соответствия между сущностями очень ограничен. Поэтому было реализовано расширение обоих фреймворков. Фреймворк EntMatcher был расширен возможностью работать с несопоставимыми сущностями добавлением в него реализации метода TBNNBS [4]. Фреймворк UEA был расширен реализацией метода построения векторных представлений сущностей RREA [5]. Затем оба фреймворка использовались для экспериментов на русско-английском

наборе данных с несопоставимыми сущностями с большим набором различных параметров.

Таким образом было проведено три группы экспериментов:

1. Вычисление эмбеддингов методом RREA с установлением соответствия между сущностями при помощи венгерского алгоритма [6];
2. Вычисление эмбеддингов методом RREA с установлением соответствия между сущностями при помощи алгоритма TBNNBS;
3. Вычисление эмбеддингов методом RREA с установлением соответствия между сущностями методом UEA.

Поскольку во всех трех группах экспериментов для генерации эмбеддингов использовался алгоритм RREA, в дальнейшем эта информация будет опускаться и будем говорить только о разных методах установления соответствия между сущностями.

3. Алгоритмы, использованные в экспериментах по выравниванию сущностей.

3.1 Алгоритм RREA

В качестве алгоритма построения векторных представлений для сущностей использовался алгоритм RREA [4], поскольку он демонстрирует качество выравнивания сущностей, превосходящие, например, такой алгоритм как RDGCN. Этот метод использует подход к выравниванию сущностей на основе графовых нейронных сетей (Graph Neural Networks, GNN), и его основной спецификой является то, что предлагается использовать ортогональную матрицу трансформации. С этой целью вводится *трансформация реляционного отражения*, которая удовлетворяет двум условиям:

1. Дифференциация отношений. Для любых отношений r_1 и r_2, и сущности e векторное представление сущности h_e должно переводиться в разные пространства.
2. Пространственная изометрия. Нормы векторных представлений сущностей не должны меняться при применении операции транформации. Расстояния между векторными представлениями сущностей должны сохраняться.

Эта операция для любого векторного представления отношения h_r строит матрицу отражения и использует эту матрицу в качестве матрицы трансформации.

3.2 Алгоритм TBNNS.

Для идентификации висячих сущностей использовался метод двустороннего ближайшего соседа с пороговым значением TBNNS (Thresholded Bi-directional Nearest Neighbor Search) [5].

Метод TBNNS задает три условия, при которых пара сущностей u из ГЗ1, и v из ГЗ2 считается выровненной:

- v является ближайшим соседом u в ГЗ2 из всех остальных сущностей в ГЗ2;

- u является ближайшим соседом v в ГЗ1 из всех остальных сущностей в ГЗ1;

- расстояние между u и v меньше некоторого порога θ.

Сущности, не удовлетворяющие этому условию, считаются несопоставимыми. Данное ограничение является достаточно жестким и требует тщательной настройки порогового значения.

3.3 Специфика применения венгерского алгоритма.

Задачу сопоставления сущностей можно решать как задачу назначения сущностям из одного ГЗ сущностей из второго ГЗ. Для этого надо найти минимальную сумму попарных расстояний между сущностями Венгерский алгоритм сопоставляет все сущности со всеми сущностями, соблюдая требование взаимно-однозначного соответствия между сущностями.

Венгерский алгоритм использует квадратную форму у матрицы расстояний между сущностями двух графов знаний, т.к. строится паросочетание минимальной стоимости. Поэтому в матрицу расстояний добавляют фиктивные сущности в тот граф знаний, у которого меньше сущностей. Итоговые выровненные сущности получаются удалением пар, в которых участвует фиктивная сущность.

3.4 UEA - алгоритм выравнивания сущностей без учителя.

Специфической особенностью метода UEA (Unsupervised Entity Alignment) является его способность работать в отсутствие множества предварительно выровненных сущностей. Он может использовать стороннюю информацию, такую как, например, метки сущностей для построения предварительного выравнивания. Алгоритм UEA получает на вход начальные структурные и текстовые эмбеддинги сущностей, суммирует их пропорционально в соответствии с коэффициентом α, и сохраняет их. В данном случае, начальные эмбеддинги вычислялись методом RREA.

Текущие эмбеддинги используются в дальнейшем повторном обучении. Также хранится список выровненных пар сущностей, в который добавляются новые выровненные пары сущностей, но не удаляются (это выход всего алгоритма).

Дальше выполняется TBNNS с некоторым порогом расстояния выравнивания θ. TBNNS получает на вход текущие эмбеддинги и набор еще не выровненных сущностей в ГЗ1 и ГЗ2, и добавляет новые пары в список выровненных.

Дальше происходит повторное обучение эмбеддингов на текущем списке выровненных пар. Новые текущие эмбеддинги получаются из

суммирования текущих и получившихся при повторном обучении эмбеддингов с весовым коэффициентом β.

Затем пороговое значение θ увеличивается на некоторое значение, и снова применяется TBNNS, и так по циклу, пока алгоритм не перестанет добавлять новые сущности в выровненные. Итоговый список выровненных сущностей используется для получения полноты точности и F1-меры.

4. Результаты экспериментов с несопоставимыми сущностями

Использовался модифицированный EntMatcher с возможностью выравнивать графы знаний с разным количеством несопоставимых сущностей. В качестве набора данных использован набор данных ru_en. Алгоритм вычисления эмбеддингов обучался на двух полных графах знаний и 4500 известных эквивалентных парах сущностей. Использовались только реляционные триплеты.

Тестирование проводилось на оставшихся парах (максимум 10500 пар сущностей, максимум 70% от полного размера данных). На вход модели подавались сущности из левого графа знаний (ГЗ1). На выходе модели предсказанные выровненные пары. Использовались следующие обозначения:

ГЗ1 и ГЗ2 - количество сущностей в сопоставляемых графах знаний.

НГЗ1 и НГЗ2 - количество несопоставимых сущностей в соответствующих ГЗ.

Корр - количество правильно выровненных пар.

P, R, F1 – точность, полнота и F1-мера.

4.1. Сопоставлением расстояний венгерским алгоритмом

Результаты применения венгерского алгоритма для выравнивания сущностей представлены в таблице 2. Можно заметить, что с увеличением количества несопоставимых сущностей падают все три метрики, что ожидаемо, т.к. венгерский алгоритм выравнивает все сущности друг с другом.

Если несопоставимые сущности находятся только в одном графе знаний, то полнота равна точности.

Нахождение несопоставимых сущностей только в левом, или только в правом графе знаний дает примерно один и тот же результат.

Нахождение несопоставимых сущностей в обоих графах знаний дает чуть хуже F1-score при большом количестве несопоставимых сущностей.

Таблица 2. Сопоставление расстояний между сущностями венгерским алгоритмом с разным количеством сопоставимых сущностей в ГЗ1 и ГЗ2.

ГЗ1	ГЗ2	НГЗ1	НГЗ2	Корр.	P	R	F1
10500	10500	0	0	5424	0.517	0.517	0.517
10250	10250	250	250	4783	0.467	0.478	0.472
10000	10000	500	500	4395	0.440	0.463	0.451

10250	10500	0	250	5134	0.501	0.501	0.501
10000	10500	0	500	4832	0.483	0.483	00.483
9500	10500	0	1000	4344	0.375	0.457	0.457
10500	10250	250	0	5120	0.500	0.500	0.500
10500	10000	500	0	4879	0.488	0.488	00.488
10500	9500	1000	0	4329	0.456	0.456	0.456

4.2. Сопоставление сущностей методом TBNNS

Результаты запуска с разными значениями порога расстояния между сущностями, включая бесконечный порог θ, обозначенный в таблице как INF, представлены в таблице ниже.

Таблица 3. Сопоставление сущностей алгоритмом TBNNS с использованием порога дистанции и с разным количеством сопоставимых сущностей в ГЗ1 и ГЗ2

ГЗ1	ГЗ2	НГЗ1	НГЗ2	Корр.	Порог	P	R	F1
8000	8000	2500	2500	208	0.025	0.756	0.038	0.072
8000	8000	2500	2500	643	0.05	0.703	0.117	0.200
8000	8000	2500	2500	869	0.075	0.675	0.158	0.256
8000	8000	2500	2500	1068	0.1	0.657	0.158	0.256
8000	8000	2500	2500	1260	0.15	0.657	0.194	0.300
8000	8000	2500	2500	1360	0.2	0.622	0.229	0.335
8000	8000	2500	2500	1387	0.25	0.607	0.247	0.351
8000	8000	2500	2500	1408	0.3	0.591	0.252	0.353
8000	8000	2500	2500	1406	0.35	0.576	0.256	0.354
8000	8000	2500	2500	1384	0.4	0.557	0.256	0.350
8000	8000	2500	2500	1414	0.45	0.542	0.257	0.349
8000	8000	2500	2500	1397	0.5	0.535	0.254	0.345
8000	8000	2500	2500	1386	INF	0.539	0.252	0.434

Не сильно большое, и не сильно малое значение порога (θ) дает небольшое улучшение результата по трем метрикам по сравнению с бесконечным значением порога. Слишком малый порог ухудшает полноту (R) и F1-score, но увеличивает точность (P).

4.3. Эксперименты с несопоставимыми сущностями в UEA

На вход UEA подавались только начальные структурные эмбеддинги, обученные на тренировочных парах (использовались только реляционные триплеты). Результаты запуска приведены в таблице ниже. В столбце “Пред.” показано количество предсказанных пар выровненных сущностей. В столбце “Соп.“ показано количество сопоставимых сущностей среди предсказанных. В столбце “Корр.” показано количество правильно сопоставленных сущностей. В столбцах P, R и F1 показаны значения

полноты, точности и F1-меры. Как обычно, с увеличением количества несопоставимых сущностей падают все метрики.

Таблица 4. Результаты запуска UEA на несопоставимых сущностях

ГЗ1	ГЗ2	НГЗ1	НГЗ2	Пред.	Соп.	Корр-соп.	P	R	F1
10500	8500	2000	0	8188	7029	3931	3931	0.462	0.471
10500	7500	3000	0	7074	5576	3230	3230	0.431	0.443
10500	6500	4000	0	5978	4305	2501	2501	0.385	0.401
10500	5500	5000	0	4928	3265	1956	1956	0.356	0.375
10500	4500	6000	0	3857	2297	1453	1453	0.323	0.348

4.4. Сравнение результатов всех методов

Ниже приведена сводная таблица сравнения результатах запуска трех методов на RREA и наборе данных en_ru с несопоставимыми сущностями. Использованный параметр TBNNS θ - бесконечный. Результаты показывают, что наилучшей F1-меры и полноты достиг итеративный алгоритм UEA, а наилучшей точности с низкой полнотой дает EntMatcher совместно с TBNNS.

Таблица 5. Сводная таблица результатов разных алгоритмов сопоставления сущностей.

ГЗ1	ГЗ2	НГЗ1	НГЗ2	Метод	P	R	F1
10500	8500	2000	0	UEA	0.462	0.480	0.471
				EntMat+TBNNS	0.708	0.319	0.439
				EntMat+Hungarian	0.416	0.416	0.416
10500	7500	3000	0	UEA	0.431	0.457	0.443
				EntMat+TBNNS	0.683	0.294	0.411
				EntMat+Hungarian	0.376	0.376	0.376
10500	6500	4000	0	UEA	0.385	0.418	0.401
				EntMat+TBNNS	0.630	0.271	0.379
				EntMat+Hungarian	0.343	0.343	0.343
10500	5500	5000	0	UEA	0.356	0.397	0.375
				EntMat+TBNNS	0.598	0.264	0.366
				EntMat+Hungarian	0.320	0.320	0.320
10500	4500	6000	0	UEA	0.323	0.377	0.348
				EntMat+TBNNS	0.534	0.236	0.328
				EntMat+Hungarian	0.285	0.285	0.285

5. Заключение

Задача выравнивания графов знаний при наличии несопоставимых сущностей соответствует ситуациям в реальной жизни и поэтому весьма актуальна. В данной работе представлены результаты экспериментов по выравниванию сущностей в русско-английском наборе данных при наличии

висячих сущностей. Весьма хорошо себя зарекомендовал метод выравнивания сущностей без учителя (UEA), позволяющий прогрессивно создавать пары выровненных сущностей. В дальнейшем будут рассмотрены варианты этой методики, использующие дополнительные атрибуты сущности и, в частности, мультимодальные графы знаний.

Следует отметить, что результаты методов выравнивания сущностей различаются в зависимости от пар языков, используемых графами знаний. Как правило, наилучшие результаты дают англо-французские и англо-немецкие выравнивания. Ухудшение результата на англо-русском наборе данных связано, с одной стороны, с языковой спецификой. С другой стороны, на качество выравнивания влияет не только качество алгоритма, но и структура самого набора данных, прежде всего, плотность и распределение степеней вершин в графах знаний [7].

Литература

1. Gnezdilova V.A., Apanovich Z.V., Russian-English dataset and comparative analysis of algorithms for cross-language embedding-based entity alignment // Journal of Physics: Conference Series. 2021. Vol. 2099.
2. Gusev D., Apanovich Z. Methods of processing textual information in entity alignment algorithms, Bulletin of the Novosibirsk Computing Center. Series: Computer Science. 2021. № 45. С. 49-58.
3. Zhao X., Zeng W., Tang J. Recent Advance of Alignment Inference Stage // Entity Alignment. Springer Nature, Singapore, 2023. С. 207-227. https://doi.org/10.1007/978-981-99-4250-3_4
4. Zhao, X., Zeng, W., Tang, J. et al. Toward Entity Alignment in the Open World: An Unsupervised Approach with Confidence Modeling. Data Sci. Eng. 7, 16–29 (2022). https://doi.org/10.1007/s41019-022-00178-4
5. Xin M., Wenting W., Huimin X. et al. Relational Reflection Entity Alignment. // arXiv.org. 2020. https://doi.org/10.48550/arXiv.2008.07962
6. H. W. Kuhn. The hungarian method for the assignment problem,URL: https://web.eecs.umich.edu/~pettie/matching/Kuhn-hungarian-assignment.pdf
7. Apanovich Z., Kolganova A., AEVis:A Visualization Method to facilitate Understanding Data and Entity Alignment Results //IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON). 2024, С. 391 – 395.

References

1. Gnezdilova V.A., Apanovich Z.V., Russian-English dataset and comparative analysis of algorithms for cross-language embedding-based entity alignment // Journal of Physics: Conference Series. 2021. Vol. 2099.

2. Gusev D., Apanovich Z. Methods of processing textual information in entity alignment algorithms, Bulletin of the Novosibirsk Computing Center. Series: Computer Science. 2021. № 45. C. 49-58.
3. Zhao X., Zeng W., Tang J., Recent Advance of Alignment Inference Stage // Entity Alignment. Springer Nature, Singapore, 2023. C. 207-227.
4. Zhao X., Zeng W., Tang J. *et al.* Toward Entity Alignment in the Open World: An Unsupervised Approach with Confidence Modeling. *Data Sci. Eng.* **7**, 16–29 (2022). https://doi.org/10.1007/s41019-022-00178-4
5. Xin M., Wenting W., Huimin X. et al. Relational Reflection Entity Alignment. // arXiv.org. 2020. https://doi.org/10.48550/arXiv.2008.07962
6. Kuhn H. W. The hungarian method for the assignment problem URL: https://web.eecs.umich.edu/~pettie/matching/Kuhn-hungarian-assignment.pdf .
7. Apanovich Z., Kolganova A., AEVis:A Visualization Method to facilitate Understanding Data and Entity Alignment Results //IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON). 2024, C. 391 – 395.