УДК 004.021+004.42

РУССКО-АНГЛИЙСКИЙ НАБОР ДАННЫХ И ВЫРАВНИВАНИЕ СУЩНОСТЕЙ В ГРАФАХ ЗНАНИЙ С НЕСОПОСТАВИМЫМИ СУЩНОСТЯМИ

З. В. Апанович[1] [0000-0002-5767-284X], Д. Г. Керного[2] [0009-0008-4551-7958]

[1]*Институт систем информатики им. А. П. Ершова Сибирского отделения РАН, г. Новосибирск, Россия*

[1, 2]*Новосибирский государственный университет, г. Новосибирск, Россия*

[1]apanovich_09@mail.ru, [2]d.kernogo@alumni.nsu.ru

Аннотация

В последние годы кратно возрос интерес к графам знаний (ГЗ) как в научном, так и в промышленном сообществах. Интеграция различных графов знаний является одной из актуальнейших задач и используется, например, для разработки сложных цифровых двойников промышленных систем. Интеграция графов знаний также необходима при объединении графов знаний, извлеченных из текстов на естественном языке при помощи больших языковых моделей. Одной из компонент решения задачи интеграции ГЗ является задача выравнивания сущностей, пытающаяся идентифицировать в разных ГЗ сущности, описывающие один и тот же объект реального мира. К сожалению, в реальных графах знаний многие сущности не имеют эквивалентов в других графах знаний. В частности, каждый фрагмент графа знаний, извлеченный из отдельной публикации, может иметь свою собственную структуру имен сущностей и идентификаторов, что существенно усложняет задачу идентификации сущностей. В работе описаны эксперименты по выравниванию сущностей при наличии несопоставимых сущностей на примере русско-английского набора данных

Ключевые слова*: графы знаний, выравнивание сущностей, несопоставимые сущности, двусторонний поиск ближайшего соседа с порогом.*

ВВЕДЕНИЕ

Графы знаний (ГЗ) хранят факты об объектах реального мира в виде реляционных и литеральных триплет. Реляционные триплеты отображают отношение между двумя сущностями (т. е. 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*).

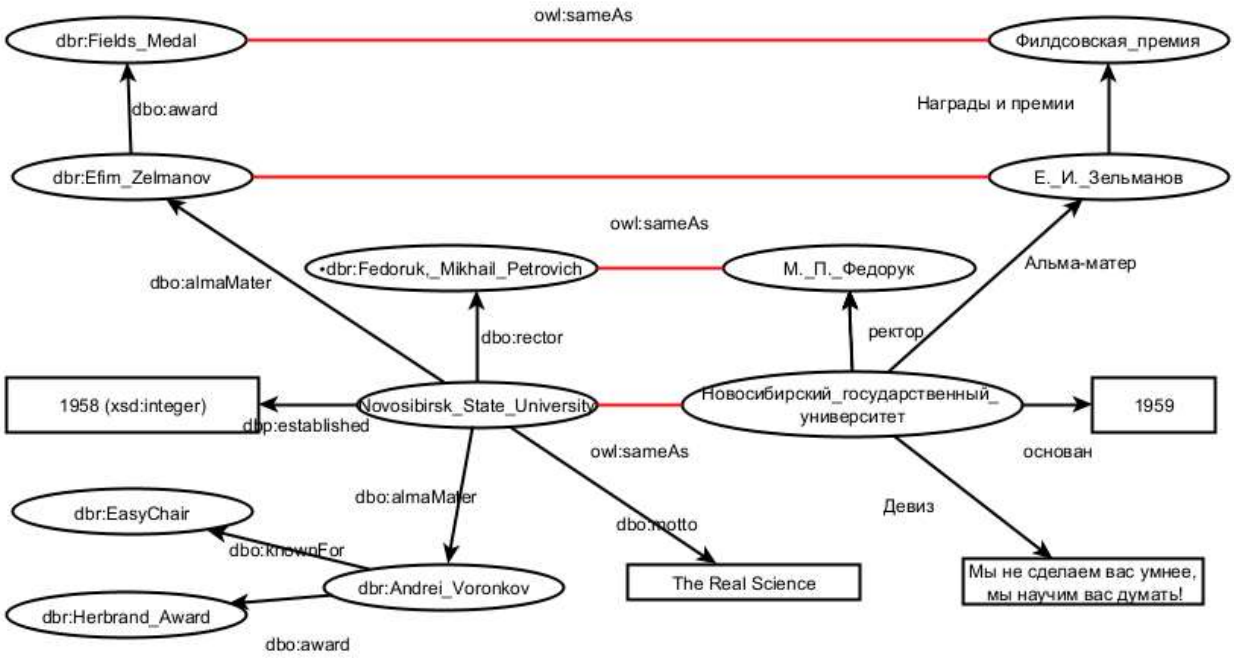

Рис. 1. Соответствие между сущностями в англоязычном и русскоязычном графах знаний.

На рис. 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 – триплета (*Новосибирский_государственный_университет*, *основан*, *1959*). На рисунке красными линиями отмечены отношения эквивалентности *owl:sameAs*, имеющиеся между сущностями в русскоязычном и англоязычном ГЗ. Можно видеть, что англоязычной сущности *dbr:Fedoruk,_Mikhail_Petrovich* в

англоязычном ГЗ 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 государственного университета (НГУ) должны бы совпадать в различных ГЗ. 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 ГЗ. В англоязычной и русскоязычной версиях также указаны разные годы основания университета. Кроме этого, в англоязычной и русскоязычной версиях ГЗ приведены разные списки людей, которые связаны с НГУ (то есть учились или работали). Например, в обеих версиях описан выпускник НГУ Ефим Зельманов, получивший Филдсовскую премию по математике. Но только в англоязычной версии есть информация о том, 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 настоящей 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. Такая постановка задачи известна достаточно давно применительно к реляционным базам данных под такими названиями, как задача устранения неоднозначностей, задача дедупликации и др. К сожалению, методы, наработанные для реляционных баз данных, оказались непригодными

применительно к графам знаний из-за большой неоднородности и разреженности графов знаний.

Современные методы выравнивания сущностей (ВС) основаны на предположении, что эквивалентные сущности имеют похожие окружения. Поэтому активно используются методы *representation learning*, 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. Векторное представление сущности *e* будем обозначать как **e**. 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. Характеристики этого набора данных представлены в табл. 1.

Табл. 1. Характеристики русско-английского набора данных.

Язык ГЗ	Сущности	Отношения	Атрибуты	Количество реляционных триплет	Количество атрибутных триплет
Ru	15000	66	15018	30489	54499
En	15000	163	15106	43796	76852

В работе [2] показано, что качество выравнивания сущностей можно значительно улучшить, повышая качество построения эмбеддингов для имен сущностей. Кроме того, были найдены наилучшие комбинации методов построения эмбеддингов для имен сущностей.

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 выполняется 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. Расстояния между парами выровненных сущностей вычисляются при помощи таких функций, как косинусная близость, манхэттеново или евклидово расстояние и др.

До последнего времени при ВС предполагалось, что каждая сущность в исходном ГЗ имеет эквивалентную сущность в целевом ГЗ. Поэтому эта эквивалентная сущность отыскивалась как ближайший сосед целевой сущности в пространстве вложения.

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.

С использованием ближайшего соседа целевой сущности в качестве эквивалентной сущности тоже обнаружились проблемы. Во-первых, результаты выравнивания сущностей могут зависеть от того, какой из ГЗ считается исходным, а какой – целевым. Во-вторых, методы на основе ближайшего соседа могут привести к тому, что двудольный граф выравнивания сущностей будет иметь вершины-хабы, у которых окажется слишком много «эквивалентных» сущностей, а также появятся изолированные вершины, не имеющие эквивалентных сущностей. Поэтому появились такие методы, как [3, 4], использующие двунаправленное выравнивание, которое интегрирует два направления выравнивания: от исходного ГЗ к целевому и наоборот. Стали также использовать методы установле-

ния соответствия между эмбеддингами сущностей, такие как венгерский алгоритм и алгоритм построения стабильного паросочетания [5–7], которые требуют установления взаимнооднозначного соответствия между сопоставляемыми сущностями.

Кроме того, 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 создают два подграфа, в которых каждая сущность имеет эквивалент во втором графе, а затем случайным образом удаляют два непересекающихся подмножества сущностей из двух ГЗ. Эквиваленты удаленных сущностей становятся несопоставимыми сущностями. Структура набора данных с несопоставимыми сущностями показана на рис. 2. В качестве стартовых наборов триплет был использован русско-английский набор данных [1].

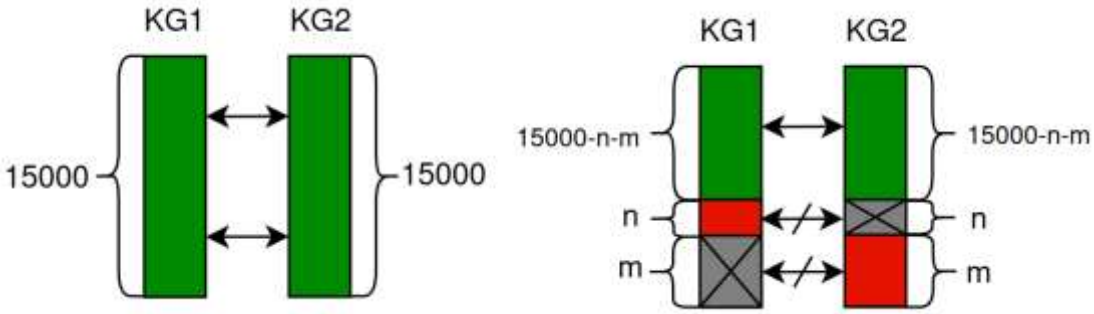

Рис. 2. Построение набора данных с несопоставимыми сущностями

АЛГОРИТМЫ, ИСПОЛЬЗОВАННЫЕ В ЭКСПЕРИМЕНТАХ ПО ВЫРАВНИВАНИЮ СУЩНОСТЕЙ

Алгоритм RREA

В качестве алгоритма построения векторных представлений для сущностей был использован RREA, один из недавно предложенных алгоритмов [4], поскольку он демонстрирует качество выравнивания сущностей, превосходящее, например, качество 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:

– *дифференциация отношений.* Для любых разных отношений r_1 и r_2 и сущности *e* векторное представление сущности должно переводиться в разные векторные пространства;

– *пространственная изометрия.* Нормы векторных представлений сущностей не должны изменяться при применении операции трансформации. 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.

Венгерский алгоритм и данные для его применения

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 соблюдает требование взаимнооднозначного соответствия между сущностями и поэтому сопоставляет все сущности со всеми сущностями, даже в случае наличия несопоставимых сущностей. Для его применения нужна квадратная форма 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.

Алгоритмы TBNNS и C-TBNNS

Для идентификации несопоставимых сущностей был использован 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 заданного порога θ.

Сущности, не удовлетворяющие этому условию, считаются несопоставимыми. Такое ограничение является достаточно жестким и требует тщательной настройки порогового значения. Большое значение порога может породить много выровненных сущностей, которые не будут соответствовать одному и тому же объекту реального мира. С другой стороны, малое значение порога будет давать небольшое множество выровненных сущностей, большинство из них будет соответствовать одному и тому же объекту реального мира, но при этом некоторые эквивалентные сущности могут остаться невыровненными.

Метод C-TBNNS [10] пытается оценить, насколько можно быть уверенным в том, что заданная пара сущностей выровнена правильно. Для этого подсчитывается так называемая мера уверенности $C(u,v)$ (confidence score) для каждой выровненной пары сущностей u и v. Эта мера оценивается как

$$C(u,v) = \mathbf{Dist}(\mathbf{u}, \mathbf{v}') - \mathbf{Dist}(\mathbf{u}, \mathbf{v}) + \mathbf{Dist}(\mathbf{v}, \mathbf{u}') - \mathbf{Dist}(\mathbf{v}, \mathbf{u}'),$$

где сущность u' является второй по близости к сущности v, а сущность v' – второй по близости к сущности u. Идея состоит в том, что если расстояние между эмбеддингами сущностей u и v меньше суммы расстояний до их двух ближайших соседей, то можно с большей уверенностью полагать, что сущности u и v действительно эквивалентны. Эта мера уверенности затем используется при подсчете функции потерь.

Фреймворки EntMatcher и CUEA

Для экспериментов с несопоставимыми сущностями использовались два фреймворка: EntMatcher [3] и CUEA [4]. Достоинствами фреймворка EntMatcher являются 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. Поэтому фреймворк EntMatcher был расширен возможностью работать с несопоставимыми сущностями добавлением в него реализации метода TBNNS [4]. Архитектура этого фреймворка показана на рис. 3.

Рис. 3. Архитектура фреймворка EntMatcher

Специфической особенностью метода CUEA (Confidence-based Unsupervised Entity Alignment) является его способность работать при отсутствии множества предварительно выровненных сущностей. Он может использовать стороннюю информацию, такую как, например, метки сущностей, значения атрибутов или описания сущностей для построения предварительного выравнивания. В наших экспериментах мы использовали эмбеддинги имен сущностей для построения предварительного выравнивания (seed alignments).

Алгоритм CUEA получает на вход начальные структурные и текстовые эмбеддинги сущностей, суммирует их пропорционально в соответствии с коэффициентом α и сохраняет их. Структурные эмбеддинги вычислялись методом RREA, а текстовые эмбеддинги – с помощью языковой модели LaBSE [11].

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

К полученным новым эмбеддингам применяется C-TBNNS с некоторым порогом расстояния выравнивания θ. TBNNS получает на вход текущие эмбеддинги и набор еще не выровненных сущностей в Г31 и Г32 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 небольшое значение и снова применяется C-TBNNS, и так по циклу, пока алгоритм не перестанет добавлять новые сущности во множество выровненных пар. Итоговый список выровненных сущностей используется для оценки полноты, точности и F1-меры всего алгоритма. Архитектура фреймворка CUEA показана на рис. 4.

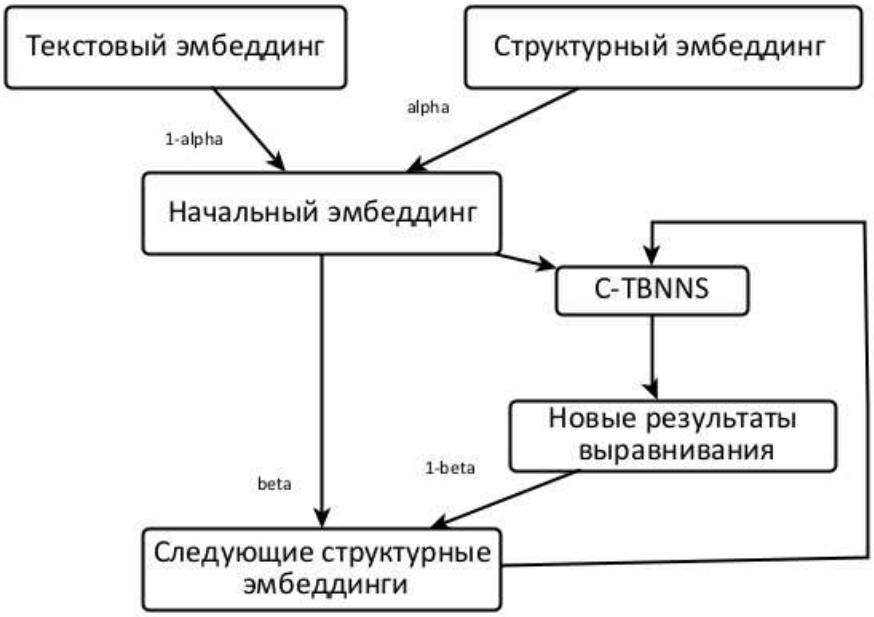

Рис. 4. Архитектура фреймворка CUEA

Оба фрейворка были использованы для проведения экспериментов на русско-английском наборе данных с несопоставимыми сущностями. Тестирование осуществлялось для широкого диапазона параметров. Были проведены три группы экспериментов.

1. Вычисление структурных эмбеддингов с применением фреймворка EntMatcher и сопоставление сущностей с помощью венгерского алгоритма [6].
2. Вычисление структурных эмбеддингов с применением фреймворка EntMatcher и сопоставление сущностей с помощью метода TBNNS.
3. Вычисление структурных эмбеддингов с применением фреймворка CUEA и сопоставление сущностей с помощью C-TBNNS.

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ С НЕСОПОСТАВИМЫМИ СУЩНОСТЯМИ

Был применен модифицированный EntMatcher с возможностью выравнивать графы знаний с различным количеством несопоставимых сущностей. В качестве набора данных был выбран 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 находились 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 соответсвенно.

Сопоставление расстояний венгерским алгоритмом

Результаты применения венгерского алгоритма для выравнивания сущностей представлены в табл. 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.

Наличие несопоставимых сущностей только в левом или только в правом графах знаний дает примерно один и тот же результат.

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

Табл. 2. Сопоставление расстояний между сущностями венгерским алгоритмом с различным 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	0.483
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

Сопоставление векторных представлений сущностей методом TBNNS

Результаты запуска с разными значениями порога расстояния между сущностями, включая бесконечный порог θ, обозначенный в табл. как INF, представлены в табл. 3.

Табл. 3. Сопоставление сущностей алгоритмом TBNNS с использованием порога дистанции и с различным количеством сопоставимых сущностей в Г31 и Г32.

Г31	Г32	НГ31	НГ32	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 слишком большое или малое значения порога θ дают небольшое улучшение результата по тем метрикам по сравнению с бесконечным значением порога. Слишком малый порог ухудшает полноту R и F1-score, но увеличивает точность P.

Эксперименты с несопоставимыми сущностями в CUEA

На вход CUEA подавались только начальные структурные эмбеддинги, обученные на тренировочных парах (использовались только реляционные триплеты). Результаты запуска приведены в табл. 4. 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 соответственно. Как обычно, с увеличением количества несопоставимых сущностей падают все метрики.

Табл. 4. Результаты запуска CUEA на несопоставимых сущностях.

ГЗ1	ГЗ2	НГЗ1	НГЗ2	Пред.	Соп.	Корр.-соп.	P	R	F1
10500	8500	2000	0	8188	7029	3931	0.462	0.480	0.471
10500	7500	3000	0	7074	5576	3230	0.431	0.457	0.443
10500	6500	4000	0	5978	4305	2501	0.385	0.418	0.401
10500	5500	5000	0	4928	3265	1956	0.356	0.397	0.375
10500	4500	6000	0	3857	2297	1453	0.323	0.377	0.348

Сравнение результатов всех методов

В сводной табл. 5 представлены результаты сравнения данных запуска трех методов на RREA и наборе данных en_ru с несопоставимыми сущностями. Был использован бесконечный параметр θ для TBNNS. Результаты показывают, что наилучшей F1-меры и полноты достиг итеративный алгоритм CUEA, а наилучшую точность с низкой полнотой дает EntMatcher совместно с TBNNS.

Табл. 5. Сводная таблица результатов различных алгоритмов сопоставления сущностей.

Г31	Г32	НГ31	НГ32	Метод	P	R	F1
10500	8500	2000	0	CUEA	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	CUEA	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	CUEA	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	CUEA	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	CUEA	0.323	0.377	0.348
				EntMat+TBNNS	0.534	0.236	0.328
				EntMat+Hungarian	0.285	0.285	0.285

Наконец, в табл. 6 показано сравнение результатов выравнивания сущностей, полученных только при помощи построения структурных эмбеддингов сущностей, и результатов, полученных при комбинировании структурных эмбеддингов сущностей и эмбеддингов имен сущностей. Эмбеддинги имен сущностей вычислялись при помощи модели LaBSE. Можно видеть значительное улучшение всех результатов.

Табл. 6. Сравнение результатов, полученных методом CUEA.

Г31	Г32	НГ31	НГ32	Метод	P	R	F1
				Структ. CUEA	0.462	0.480	0.471
10500	8500	2000	0	Комб. CUEA	**0.972**	**0.610**	**0.750**
				Структ. CUEA	0.431	0.457	0.443
10500	7500	3000	0	Комб. CUEA	**0.954**	**0.475**	**0.634**
				Структ. CUEA	0.385	0.418	0.401
10500	6500	4000	0	Комб. CUEA	**0.946**	**0.421**	**0.582**
				Структ. CUEA	0.356	0.397	0.375
10500	5500	5000	0	Комб. CUEA	**0.930**	**0.340**	**0.497**
				Структ. CUEA	0.323	0.377	0.348
10500	4500	6000	0	Комб. CUEA	**0.911**	**0.281**	**0.430**

ЗАКЛЮЧЕНИЕ

Задача выравнивания графов знаний при наличии несопоставимых сущностей соответствует ситуациям в реальной жизни, и поэтому весьма актуальна. В работе представлены результаты экспериментов по выравниванию сущностей в русско-английском наборе данных при наличии висячих сущностей. Весьма хорошо зарекомендовал себя метод выравнивания сущностей без учителя (CUEA), позволяющий прогрессивно создавать пары выровненных сущностей. В дальнейшем будут рассмотрены варианты этой методики, использующие дополнительные атрибуты сущности, в частности мультимодальные графы знаний.

Отметим, что результаты методов выравнивания сущностей отличаются 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].

СПИСОК ЛИТЕРАТУРЫ

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.
https://doi.org/10.1088/1742-6596/2099/1/012023

2. *Gusev D., Apanovich Z.* Methods of processing textual information in entity alignment algorithms // Bulletin of the Novosibirsk Computing Center. Series: Computer Science. 2021. No. 45. P. 49–58.
https://doi.org/10.31144/bncc.cs.2542-1972.2021.n45.p49-58

3. *Lample G., Conneau A., Ranzato M., Denoyer L., Jégou, H.* Word translation without parallel data // ICLR. OpenReview.net. 2018.
https://doi.org/10.48550/arXiv.1710.04087

4. *W. Zeng, X. Zhao, X. Li, J. Tang, W. Wang.* On entity alignment at scale // The VLDB Journal. 2022. Vol. 31. Issue 5. P. 1009–1033.
https://doi.org/10.1007/s00778-021-00703-3

5. *Kuhn H.W.* The hungarian method for the assignment problem. URL: https://web.eecs.umich.edu/~pettie/matching/Kuhn-hungarian-assignment.pdf. https://doi.org/10.1002/nav.3800020109

6. *W. Zeng, X. Zhao, J. Tang, X. Lin*. Collective entity alignment via adaptive features // ICDE. 2020. P. 1870–1873. https://doi.org/10.48550/arXiv.1912.08404

7. *R. Zhu, M. Ma, P. Wang.* RAGA: relation-aware graph attention networks for global entity alignment // PAKDD. 2021. Vol. 12712. P. 501–513. https://doi.org/10.48550/arXiv.2103.00791

8. *Xin M., Wenting W., Huimin X. et al.* Relational Reflection Entity Alignment. // arXiv.org. 2020. https://doi.org/10.48550/arXiv.2008.07962

9. *Zhao X., Zeng W., Tang J*., Recent Advance of Alignment Inference Stage // Entity Alignment. Springer Nature, Singapore, 2023. P. 207–227. https://doi.org/10.1007/s41019-022-00178-4

10. *Zhao X., Zeng W., Tang J. et al.* Toward Entity Alignment in the Open World: An Unsupervised Approach with Confidence Modeling // Data Sci. Eng. 2022. No. 7. P. 16–29. https://doi.org/10.1007/s41019-022-00178-4

11. *Feng F., Yang Y., Cer D., Arivazhagan N., Wang W.* Language-agnostic BERT Sentence Embedding // ArXiv. 2020. https://doi.org/10.48550/arXiv.2007.01852

RUSSIAN-ENGLISH DATASET AND ENTITY ALIGNMENT IN KNOWLEDGE GRAPHS WITH UNMATCHABLE ENTITIES

Z. V. Apanovich[1] [0000-0002-5767-284X], D. G. Kernogo[2] [0009-0008-4551-7958]

[1]*A. P. Ershov Institute of Informatics Systems, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia*

[1, 2]*Novosibirsk State University, Novosibirsk, Russia*

[1]apanovich_09@mail.ru, [2]d.kernogo@alumni.nsu.ru

Abstract

In recent years, interest in knowledge graphs (KGs) has increased exponentially in both the scientific and industrial communities. Integration of various KGs is a pressing problem and is used, for example, to develop complex digital twins of industrial systems. Knowledge graph integration is also necessary when combining KGs extracted from natural language texts using large language models. One component of solving the KG integration problem is entity alignment (EA), which attempts to identify entities in different KGs that describe the same real-world object. In reality, many entities in real KGs have no equivalents in other KGs. In particular, each knowledge graph fragment extracted from a single publication may have its own structure of entity names and identifiers, which significantly complicates the task of identifying entities. This paper describes experiments on entity alignment in the presence of unmatchable entities using a Russian-English dataset as an example.

Keywords*: knowledge graph, entity alignment, unmatchable entities, thresholded bidirectional nearest neighbor search.*

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.
https://doi.org/10.1088/1742-6596/2099/1/012023

2. *Gusev D., Apanovich Z.* Methods of processing textual information in entity alignment algorithms // Bulletin of the Novosibirsk Computing Center. Series: Computer Science. 2021. No. 45. P. 49–58.

https://doi.org/10.31144/bncc.cs.2542-1972.2021.n45.p49-58

3. *Lample G., Conneau A., Ranzato M., Denoyer L., Jégou H.* Word translation without parallel data // ICLR. OpenReview.net. 2018. https://doi.org/10.48550/arXiv.1710.04087

4. *W. Zeng, X. Zhao, X. Li, J. Tang, W. Wang.* On entity alignment at scale // The VLDB Journal. 2022. Vol. 31. Issue 5. P. 1009–1033. https://doi.org/10.1007/s00778-021-00703-3

5. *Kuhn H.W.* The hungarian method for the assignment problem. URL: https://web.eecs.umich.edu/~pettie/matching/Kuhn-hungarian-assignment.pdf. https://doi.org/10.1002/nav.3800020109

6. *W. Zeng, X. Zhao, J. Tang, X. Lin*. Collective entity alignment via adaptive features // ICDE. 2020. P. 1870–1873. https://doi.org/10.48550/arXiv.1912.08404

7. *R. Zhu, M. Ma, P. Wang.* RAGA: relation-aware graph attention networks for global entity alignment // PAKDD. 2021. Vol. 12712. P. 501–513. https://doi.org/10.48550/arXiv.2103.00791

8. *Xin M., Wenting W., Huimin X. et al.* Relational Reflection Entity Alignment. // arXiv.org. 2020. https://doi.org/10.48550/arXiv.2008.07962

9. *Zhao X., Zeng W., Tang J*., Recent Advance of Alignment Inference Stage // Entity Alignment. Springer Nature, Singapore, 2023. P. 207–227. https://doi.org/10.1007/978-981-99-4250-3

10. *Zhao X., Zeng W., Tang J. et al.* Toward Entity Alignment in the Open World: An Unsupervised Approach with Confidence Modeling // Data Sci. Eng. 2022. No. 7. P. 16–29. https://doi.org/10.1007/s41019-022-00178-4

11. *Feng F., Yang Y., Cer D., Arivazhagan N., Wang W.* Language-agnostic BERT Sentence Embedding // ArXiv. 2020. https://doi.org/10.48550/arXiv.2007.01852

СВЕДЕНИЯ ОБ АВТОРАХ

АПАНОВИЧ Зинаида Владимировна– старший научный сотрудник Института систем информатики СО РАН, доцент Новосибирского государственного университета. Сфера научных интересов – графы знаний, выравнивание сущностей, визуализация информации, визуализация графов.

Zinaida Vladimirovna APANOVICH – senior researcher at the Institute of Informatics Systems of SB RAS, associate professor at Novosibirsk State University. Research interests include knowledge graphs, entity alignment, information visualization, graph visualization.

email: apanovich@iis.nsk.su

ORCID: 0000-0002-5767-284X

КЕРНОГО Даниил Георгиевич – магистрант Новосибирского государственного университета. Сфера научных интересов – графы знаний, выравнивание сущностей.

Daniil Georgievich KERNOGO – master's student at Novosibirsk State University. Research interests include knowledge graphs, entity alignment.

email: d.kernogo@alumni.nsu.ru

ORCID: 0009-0008-4551-7958

Материал поступил в редакцию 20 января 2026 года