

4.04.Лаборатория анализа информационных ресурсов

№ госрегистрации
АААА-А20-120121690111-5

УТВЕРЖДАЮ
Директор/декан

«__» _____ Г.

УДК
004.8 Искусственный интеллект
519.767.6 Способы выделения семантической информации из текста

ОТЧЕТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

Фундаментальные проблемы построения систем информатизации,
методология, технология и безопасность крупных информационных систем
по теме:

Методы структурирования трудноформализуемых предметных областей на
основе автоматизированного формирования больших графов знаний и
онтологий по разнородным потокам текстовых данных
(промежуточный)

Зам. директора/декана
по научной работе

«__» _____ Г.

Руководитель темы
Добров Б.В.


«20» 12 2022 Г.

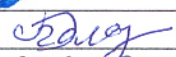
СПИСОК ИСПОЛНИТЕЛЕЙ

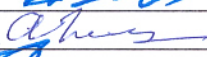
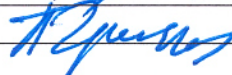
Руководитель темы:
заведующий лабораторией,
кандидат физико-математических наук

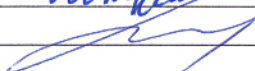
 (Добров Б.В.)

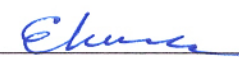
Исполнители темы:

специалист
специалист
специалист
специалист , кандидат технических наук
техник
младший научный сотрудник
заведующий лабораторией,
кандидат филологических наук,
доктор филологических наук,
доцент по кафедре
специалист
ведущий научный сотрудник,
кандидат физико-математических наук,
доктор технических наук
ведущий научный сотрудник,
кандидат филологических наук,
доктор филологических наук,
доцент/с.н.с. по специальности
специалист
специалист , доктор технических наук,
кандидат физико-математических наук
специалист
аспирант
специалист , кандидат технических наук
специалист
специалист
младший научный сотрудник
стажер
специалист

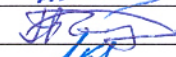
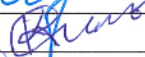
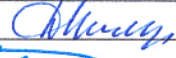
 (Архипенко К.В.)
 (Болдырев И.А.)
уволен 31.10.2022г. (Быстрицкий Д.К.)
уволен 29.09.2022г. (Быстрицкий Н.Д.)

 (Волков Е.В.)
 (Герасимова А.А.)
 (Гращенко П.В.)

 (Кремнева М.Д.)
 (Лукашевич Н.В.)

 (Лютикова Е.А.)

 (Майоров В.Д.)
учер окт. 2022г. (Макаров-Землянский Н.В.)

 (Мячина А.В.)
 (Подшивалов Д.Д.)
 (Рубцова Ю.В.)
 (Сидоров А.В.)
 (Скворцов Н.А.)
 (Тихомиров М.М.)
 (Чернышев Д.И.)
 (Штернов С.В.)

РЕФЕРАТ

Ключевые слова:

предобученные языковые модели, большие данные, обработка естественного языка, графы знаний, глубокое обучение, онтологии, представление знаний, текстовые данные, информационно-аналитические системы, лингвистические онтологии

Ключевые слова по-английски:

bigdata, information-analytical system, text data, knowledge graph, linguistic ontology, knowledge representation, pretrained language models, natural language processing, ontology

Объектом исследования являются графы знаний – структуры, обобщающие онтологии, путем учета более широкой номенклатуры объектов – именованных сущностей различных типов, текстовых фрагментов, отражающих знания, содержащиеся в коллекции документов. Целью работы является разработка методов автоматизированного формирования и сопровождения графов знаний большого размера с использованием методов глубокого обучения на основе содержательной обработки больших массивов текстов, и на основе ранее созданных больших онтологических ресурсов. А также исследование методов использования больших графов знаний для поддержки решения информационно-аналитических задач в реальных социально-экономических и научно-технических предметных областях.

В течение 2022 года при выполнении 3го этапа «Разработка методов автоматизированного формирования больших лингвистических онтологий предметной области, методов извлечения сложных типов именованных сущностей, исследование методов автоматизированного формирования графов знаний» решались следующие задачи:

- исследование методов автоматического извлечения неизвестных отношений с использованием методов машинного обучения и ранее созданных онтологических ресурсов;
- разработка методов автоматизированного формирования больших онтологий предметной области с развитым набором отношений;
- разработка методов связывания различных текстовых вариантов извлеченных именованных сущностей на основе результатов обработки больших текстовых коллекций;
- разработка методов разрешения многозначности текстового выражения именованных сущностей в разных документах.

Практическая значимость полученных результатов заключается в снижении трудоемкости для формирования больших графов знаний, прежде всего в части подключения именованных сущностей и текстовых объектов в виде экстрактивных и абстрактивных аннотаций. По теме опубликовано 11 статей (1 Q1, 6 Scopus, 3 РИНЦ)

ВВЕДЕНИЕ

Целью работы является разработка методов автоматизированного формирования и сопровождения графов знаний большого размера с использованием методов глубокого обучения на основе содержательной обработки больших массивов текстов, и на основе ранее созданных больших онтологических ресурсов. А также исследование методов использования больших графов знаний для поддержки решения информационно-аналитических задач в реальных социально-экономических и научно-технических предметных областях.

В течение 2022 года при выполнении 3го этапа «Разработка методов автоматизированного формирования больших лингвистических онтологий предметной области, методов извлечения сложных типов именованных сущностей, исследование методов автоматизированного формирования графов знаний» решались следующие задачи:

- исследовать методы автоматического извлечения неизвестных отношений с использованием методов машинного обучения и ранее созданных онтологических ресурсов;
- разработать методы автоматизированного формирования больших онтологий предметной области с развитым набором отношений;
- разработать методы связывания различных текстовых вариантов извлеченных именованных сущностей на основе результатов обработки больших текстовых коллекций;
- разработать методы разрешения многозначности текстового выражения именованных сущностей в разных документах.

ОСНОВНАЯ ЧАСТЬ

ОСНОВНЫЕ НАУЧНО-ТЕХНИЧЕСКИЕ РЕЗУЛЬТАТЫ

В течение 2022 года 3го этапа темы «Разработка методов автоматизированного формирования больших лингвистических онтологий предметной области, методов извлечения сложных типов именованных сущностей, исследование методов автоматизированного формирования графов знаний» были получены следующие результаты:

1) По направлению исследования методов автоматического извлечения неизвестных отношений рассматривались задачи применения извлеченных отношений для улучшения решения задачи аннотирования.

Предложен метод построения обучающего набора данных для задачи абстрактивного реферирования новостей и разработан процесс тонкой настройки для правильного обучения. Экспериментальные результаты по набору данных показывают, что новая модель превосходит простые экстрактивные и абстрактивные методы, прежде всего по введенной метрике фактологической достоверности, которая рассчитывается на основе воспроизведения в аннотации именованных сущностей и извлекаемых отношений с именованными сущностями. Лучшие результаты в автоматическом абстрактивном аннотировании в настоящее время показывают подходы на основе специализированных языковых моделей. Эти модели предварительно обучены на задаче построения псевдо-аннотаций, что позволяет им эффективнее отбирать информации в задаче абстрактивного аннотирования.

Был разработан новый метод построения псевдо-аннотаций на основе кластеров – ClusterVote. Метод ClusterVote способен порождать экстрактивные и абстрактивные аннотации различного уровня детализации. На основе метода была создана новая коллекция для задачи абстрактивного аннотирования – Telegram News. Согласно анализу, эталонные аннотации коллекции обладают такими же экстрактивными характеристиками, что и популярные наборы данных для задачи, но при этом имеют более высокие показатели корректности примеров. Коллекция была апробирована для обучения предобученных языковых моделей, специализированных для задачи абстрактивного аннотирования. Результаты экспериментов показали, что аннотации, полученные методом ClusterVote, способствуют повышению фактологической достоверности генерируемых текстов. Метод ClusterVote также был апробирован для обучения русскоязычных предобученных генеративных моделей общего назначения: mBART, ruT5. С помощью метода была собрана самая большая коллекция для аннотирования русскоязычных новостей – Telegram News*CV(RU). Эксперименты показали, что эталонные аннотации этой коллекции также повышают фактологическую достоверность генерируемых аннотаций, однако в отличие от варианта для английского языка, они также способствуют большей абстрактивности текста. Такой результат позволяет полагать, что метод ClusterVote будет эффективным для предобучения специализированной языковой модели для абстрактивного аннотирования текстов на русском языке.

2) По направлению разработки методов автоматизированного формирования больших онтологий предметной области с развитым набором отношений были проведены эксперименты по извлечению отношений (49 типов) на датасете NEREL. Особенностью датасета является то, что он размечен вложенными именованными сущностями, что позволяет увеличивать полноту

извлечения отношений из текстов. Поскольку на текущий момент времени в мире имеется очень мало наборов данных, которые размечены вложенными именованными сущностями и отношениями между ними, то некоторые из современных моделей извлечения отношений из текстов не учитывают особенности извлечения отношений, связывающих верхнюю внешнюю сущность с ее внутренней сущностью (вложенные отношения) или отношения, которые идут из отдельно стоящей сущности во внутреннюю сущность (пересекающиеся отношения). Были проанализированы входные форматы современных моделей извлечения отношений с точки зрения возможности работы с вложенными и пересекающимися отношениями. В частности, было выявлено, что не могут использоваться модели типа SpanBERT, которые основаны на маскировании сущностей для выявления отношений. При анализе входного формата применения пакета для извлечения отношений OpenNRE было выявлено излишнее дублирование вложенных сущностей. Была проведена коррекция входного формата, после чего качество извлечения отношений внутри предложения с помощью этого пакета с использованием контекстуализированных эмбедингов RuBERT, достигло 80.5% F-меры.

Продолжились исследования по автоматическому установлению таксономических отношений при формировании онтологии. Таксономические отношения (также называемые отношения гипоним-гипероним, отношения класс-подкласс) являются наиболее важными для организации знаний в подавляющем большинстве предметных областей. Для исследования извлечения этих отношений из текстовых коллекций в рамках проекта был создан датасет Diachronic wordnets. Данный датасет основан на разных версиях лексико-семантических ресурсов: англоязычного WordNet и русскоязычного RuWordNet, то есть в датасете собраны слова и выражения, которые появились в более поздних версиях ресурсов с теми гиперонимами из предыдущих версий, к которым они приписаны. Таким образом, различные подходы к извлечению гиперонимов можно тестировать на естественных данных развивающихся ресурсов. Дополнительно датасет снабжен межъязыковыми соответствиями между единицами английского и русского языков, что дает дополнительные возможности для исследования кросс-языковых вопросов извлечения таксономических отношений.

На созданном датасете были проверены различные подходы к извлечению таксономических отношений: включая подходы на основе векторных представлений слов, графовых представлений ресурсов, словарных статей Викисловаря, поисковых выдач глобальных поисковых систем, а также новый предложенный подход к извлечению таксономических отношений, основанный на комбинировании различных представлений с помощью так называемых мета-эмбедингов. Суть подхода на основе мета-эмбедингов состоит в том, что используется нейросетевой автокодировщик, кодирующий разнородные представления в единое представление так, чтобы исходные представления были восстановимы из единого представления, а также с дополнительным контролем функции потерь на основе подхода triplet loss, при котором созданное представление слова должно быть похоже на гиперонимы и не похоже на другие слова. В результате проведенных экспериментов было показано, что подход на основе мета-эмбедингов с функцией потерь триплет-лосс, комбинирующий векторные представления слов (word2vec, glove, fasttext) и графовые представления, получает наилучшие результаты извлечения гиперонимов для существительных во всех вариантах датасета.

Было показано, что нахождение гиперонимов для глаголов более сложно, что планируется исследовать в дальнейших экспериментах.

3) По направлению разработки методов связывания различных текстовых вариантов извлеченных именованных сущностей на основе результатов обработки больших текстовых коллекций были проведены эксперименты по связыванию упоминаний именованных сущностей из набора данных NEREL с объектами графа знаний Викиданные. Было выявлено, что существенное влияние на качество связывания сущностей текста с графом знаний имеет правильное предсказание сущностей, отсутствующих в графе знаний (обозначаются в датасете пометкой NULL). Эта проблема имеет большое значение и с практической точки зрения. Поэтому были исследованы различные методы, которые бы позволили улучшить качество определения отсутствующих в графе знаний сущностей.

Все исследования проводились на основе датасета NEREL. Базовой моделью для решения задачи Entity Linking была выбрана модель mGENRE. В качестве способов оценки неопределенности предсказания были рассмотрены различные подходы применимые для детерминированных моделей и для ансамблей: а) оценка на основе максимального правдоподобия (score-based). б) разница между правдоподобием наиболее вероятной и второй по вероятности гипотез (delta), в) энтропия предиктивного распределения ансамбля (predictive entropy): - средняя энтропия предиктивных распределений моделей в ансамбле (expected entropy), - взаимная информация между параметрами модели и предиктивным распределением (BALD/Mutual Information). В силу высокой вычислительной сложности обучения ансамблей независимых моделей, в качестве способа ансамблирования был выбран широко используемый подход MC Dropout Ensembles.

В результате проведенных экспериментов по связыванию сущностей, было показано, что наиболее эффективным из рассмотренных способов оценки неопределенности является score-based подход. Несмотря на показанную в целом ряде работ эффективность более сложных методов оценки неопределенности, основанных на ансамблировании, для рассмотренной задачи простой в реализации и вычислительно нетребовательный метод, основанный на подборе порога правдоподобия наиболее вероятного предсказания, показывает также и наилучшую точность (микро 0.674, макро 0.694, площадь под ROC-кривой 0.833) детекции отсутствующих в графе сущностей. Тем не менее, для ряда категорий рассматриваемого набора данных, более высокую эффективность показывают методы, основанные на ансамблях моделей. Так, наилучшим методом для категории AWARD является predictive entropy (0.684), для категорий CITY и LOCATION - expected entropy (0.867 и 0.705 соответственно).

4) По направлению разработки методов разрешения многозначности текстового выражения именованных сущностей в разных документах был изучен подход, учитывающий априорную многозначность именованных сущностей при связывании сущностей с Викиданными. Для этого по массиву статей русского раздела Викиданных для именованных сущностей был вычислено, сколько сущностей с такими же наименованиями имеются в Викиданных. Эта операция была выполнена с помощью поисковой машины elastic search. Был получен относительный коэффициент неоднозначности. В результате комбинирования score-based оценки с предложенным методом удалось увеличить точность предсказания правильной ссылки сущности в

Викиданных (микро точность повысилась с 67.4% 67.6%).

5) Также проводились исследования специальных лингвистических задач, возникающих при подготовке текстовых данных для автоматической обработки. Изучены и детально описаны принципы построения системы автоматической транскрипции для русского языка. Транскрибировать текст в фонетическое представление необходимо для целого ряда задач, связанных с анализом и синтезом речи. Определяются [9] основные блоки, из которых должна состоять система автоматической транскрипции. Описан функционал каждой части системы и указано, какие части этого функционала не могут быть реализованы с должны уровнем качества на данном этапе развития технологий обработки текста.

В последнее время при сборе эмпирических данных о языковых структурах исследователи все чаще обращаются к методу извлечения суждений о приемлемости языкового выражения. Характерная особенность этого метода – неоднородность суждений носителей языка. Рассмотрены [11] возможные источники неоднородности и предложены способы ее измерения и интерпретации. Результаты работы призваны сделать более корректными представления о языковой структуре, предлагаемые формальной лингвистикой.

СПИСОК ПУБЛИКАЦИЙ

1. Chernyshev D. ClusterVote: Automatic Summarization Dataset Construction with Document Clusters /Daniil Chernyshev, Boris Dobrov //International Conference on Speech and Computer, Lecture Notes in Computer Science, Springer International Publishing AG (Cham, Switzerland), том 13721, с. 99-113. – DOI: http://dx.doi.org/10.1007/978-3-031-20980-2_10 (Scopus)
2. Chernyshev D. Improving Neural Abstractive Summarization with Reliable Sentence Sampling /D. Chernyshev, B. Dobrov //International Conference on Data Analytics and Management in Data Intensive Domains, Communications in Computer and Information Science, Springer Cham, том 1620, с. 246-261 DOI http://dx.doi.org/10.1007/978-3-031-12285-9_16 (Scopus)
3. Dimov I. Methods for Automatic Argumentation Structure Prediction /I. Dimov, B. Dobrov //International Conference on Data Analytics and Management in Data Intensive Domains, Communications in Computer and Information Science, Springer Cham, том 1620, с. 223-233. – DOI: http://dx.doi.org/10.1007/978-3-031-12285-9_14 (Scopus)
4. Kirillovich A. Sense-Annotated Corpus for Russian /A. Kirillovich, N. Loukachevitch, M. Kulaev, A. Bolshina, D. Ilvovsky //Proceedings of CLIB-2022 conference, с. 130-136.
5. Kotelnikov E. RuArg-2022: Argument Mining Evaluation /Evgeny Kotelnikov, Natalia Loukachevitch, Irina Nikishina, Alexander Panchenko //Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2022” Moscow, June 15-18, 2022. - Москва: ПГТУ, том 21, с. 333-348 DOI: <http://dx.doi.org/10.28995/2075-7182-2022-21-333-348> (Scopus)
6. Loukachevitch N. Entity Linking over Nested Named Entities for Russian /Natalia Loukachevitch, Pavel Braslavski, Vladimir Ivanov, et. al. //Proceedings of the Language Resources and Evaluation Conference, European Language Resources Association Marseille, France, с. 4458-4466. (Scopus)
7. Nikishina I. Taxonomy enrichment with text and graph vector representations /Irina Nikishina, Mikhail Tikhomirov, Varvara Logacheva, et. al. //SEMANTIC WEB, том 13, № 3, с. 441-475 DOI <http://dx.doi.org/10.3233/sw-212955>. (Q1)
8. Rozhkov I. Machine Reading Comprehension Model in RuNNE Competition

/Igor Rozhkov, Natalia Loukachevitch //Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2022" Moscow, June 15-18, 2022/ - Москва: РГГУ, том 21, с. 488-496. DOI: <http://dx.doi.org/10.7182-2022-21-488-496> (Scopus)

9. Гращенков П.В. Системы автоматической транскрипции русского текста: общая организация и решенные и нерешенные проблемы /П.В. Гращенков //Вестник Московского университета. Серия 9: Филология, издательство Изд-во Моск. ун-та (М.), № 3, с. 9-21. -URL: <https://elibrary.ru/item.asp?id=48866769> (РИНЦ)

10. Лукашевич Н.В. Автоматический анализ тональности текстов. Проблемы и методы /Н.В. Лукашевич //Интеллектуальные системы. Теория и приложения (ранее: Интеллектуальные системы по 2014, № 2, ISSN 2075-9460), М., том 26, № 1. (РИНЦ)

11. Лютикова Е.А. Три аспекта непротиворечивости языковых данных: оценка и интерпретация /Е.А. Лютикова, А.А. Герасимова, Д.Д. Белова, и др. //Интеллектуальные системы. Теория и приложения. - Том 26, Номер 1, 2022, С.196-202. - URL: <https://elibrary.ru/item.asp?id=49337593> (РИНЦ)

ЗАКЛЮЧЕНИЕ

В течение 2022 года при выполнении 3го этапа «Разработка методов автоматизированного формирования больших лингвистических онтологий предметной области, методов извлечения сложных типов именованных сущностей, исследование методов автоматизированного формирования графов знаний» получены следующие результаты:

1) По направлению исследования методов автоматического извлечения неизвестных отношений показано, что применение выделяемых с использованием нейросетевых методов именованных сущностей и отношений с ними позволяет ввести метрики фактологической достоверности оценки качества экстрактивных и абстрактивных аннотаций. Разработан новый метод построения псевдо-аннотаций на основе кластеров – ClusterVote. Метод апробирован для обучения русскоязычных предобученных генеративных моделей общего назначения: mBART, ruT5. С помощью метода собрана самая большая коллекция для аннотирования русскоязычных новостей – Telegram News*CV(RU).

2) По направлению разработки методов автоматизированного формирования больших онтологий предметной области с развитым набором отношений были проведены эксперименты по извлечению отношений (49 типов) на датасете NEREL. Особенностью датасета является то, что он размечен вложенными именованными сущностями, что позволяет увеличивать полноту извлечения отношений из текстов. Была проведена коррекция входного формата данных, после чего качество извлечения отношений внутри предложения с помощью пакета OpenNRE с использованием контекстуализированных эмбедингов RuBERT, достигло 80.5% F-меры. Для исследования извлечения таксономических отношений из текстовых коллекций в рамках проекта был создан датасет Diachronic wordnets. Был исследован подход на основе мета-эмбедингов с функцией потерь триплет-лосс, комбинирующий векторные представления слов (word2vec, glove, fasttext) и графовые представления, с помощью которого получены лучшие результаты извлечения гиперонимов для существительных во всех вариантах датасета.

3) По направлению разработки методов связывания различных текстовых вариантов извлеченных именованных сущностей на основе результатов обработки больших текстовых коллекций были проведены эксперименты по связыванию упоминаний именованных сущностей из набора данных NEREL с объектами графа знаний Викиданные. Показано, что наиболее эффективным из рассмотренных способов оценки неопределенности является score-based подход. Для ряда категорий рассматриваемого набора данных, более высокую эффективность показывают методы, основанные на ансамблях моделей.

4) По направлению разработки методов разрешения многозначности текстового выражения именованных сущностей в разных документах был изучен подход, учитывающий априорную многозначность именованных сущностей при связывании сущностей с Викиданными. В результате комбинирования score-based оценки с предложенным методом удалось увеличить точность предсказания правильной ссылки сущности в Викиданных.

Практическая значимость результатов заключается в снижении трудоемкости для формирования больших графов знаний в части подключения именованных сущностей, а также текстовых объектов в виде аннотаций.

ПРИЛОЖЕНИЕ А
Объем финансирования темы в 2022 году
Таблица А.1

Источник финанси- рования	Объем (руб.)	
	Получено	Освоено собственными силами