ИЗДАТЕЛЬСТВО ГРАМОТА

Филологические науки. Вопросы теории и практики
Philology. Theory & Practice

ISSN 2782-4543 (online)
ISSN 1997-2911 (print)

2024. Том 17. Выпуск 9 | 2024. Volume 17. Issue 9
Материалы журнала доступны на сайте (articles and issues available at): philology-journal.ru

RU

Проблемы извлечения слабоструктурированной текстовой информации на основе технологии Text Mining (на материале русского и чувашского языков)

Губанов А. Р., Данилов А. А., Исаев Ю. Н., Губанова Г. Ф.

Аннотация. Цель исследования – выявить модели и алгоритмы обработки текстовой информации, связанные с модальной коррекцией схем интенциональных отношений в разноструктурных языках на основе технологии Text Mining. Рост потоков разнородной текстовой информации в Интернете, состоящей из сложноорганизованных документов, ставит перед аналитиками проблемы, связанные с дифференцированным извлечением знаний (в интеллектуальном анализе разнородной текстовой информации используется технология Text Mining). В статье предложен подход к анализу информации модальной коррекции схем интенциональных смысловых отношений (ИСО) в разноструктурных языках на основе методов компьютерной лингвистики и Text Mining. При помощи библиотеки Language Resources проведен анализ русских и чувашских корпусов в БД Datastores (перенос информации на основе анализа проблем интеграции и совместимости данных с различными типами документов из разных источников). На основе предложенного концептуального подхода осуществляется кластеризация (кластеров документов, текстового корпуса). Научная новизна исследования состоит в разработке комплекса моделей и алгоритмов для анализа интенциональных отношений в разноструктурных языках – русском и чувашском, обеспечивающих точность и полноту в извлечении информации в поисковых запросах. Акцентируется внимание на контенте лингвистических ресурсов, проводится классификация лингвистических ресурсов по классам-модусам ИСО, определяется подход к формализации лексико-синтаксических шаблонов, на их основе решается задача построения таксономии концепта ИСО. В результате исследования установлено, что предлагаемый метод эффективен для решения задач интеллектуального анализа текстов и интерпретации его результатов.

EN

Problems of extracting semi-structured textual information based on Text Mining technology (using the material of the Russian and Chuvash languages)

A. R. Gubanov, A. A. Danilov, Y. N. Isaev, G. F. Gubanova

Abstract. The study aims to identify models and algorithms for processing textual information related to modal correction of intentional relationship schemes in languages with different structures based on Text Mining technology. The growth of diverse textual information flows on the internet, consisting of complexly organized documents, poses challenges for analysts. Such challenges are related to differentiated knowledge extraction (Text Mining technology is used in the analysis of diverse textual information). The paper proposes an approach to analyzing information related to modal correction of schemes of intentional semantic relations in languages with different structures involving methods of computational linguistics and Text Mining. Using the Language Resources library, an analysis of Russian and Chuvash corpora in the Datastores database was conducted (transferring information based on an analysis of the problems of integration and compatibility of data with various types of documents from different sources). Based on the proposed conceptual approach, clustering is performed (of document clusters, of the text corpus). The scientific novelty of the study lies in developing a complex of models and algorithms for analyzing intentional relations in languages with different structures, in particular, in Russian and Chuvash, ensuring accuracy and completeness in extracting information in search queries. Attention is focused on the content of linguistic resources; a classification of linguistic resources is conducted according to class-modes of intentional semantic relations. An approach to formalizing lexico-syntactic templates is determined, and on their basis,

Научная статья (original research article) | https://doi.org/10.30853/phil20240437

the task of constructing a taxonomy for the concept of intentional semantic relations is solved. As a result of the study, it has been found that the proposed method is effective for solving problems of Text Mining and interpreting its results.

Введение

Актуальность исследования обусловлена тем, что интеллектуальный анализ текста (англ. Text Mining) является одной из наиболее значимых областей научных исследований, ибо большой рост объемов текстовой информации со сложной структурой естественно-языковых текстов требует дифференцированного извлечения сведений коммуникативной интенции и интенциональности речевого поведения говорящего для лингвистов-аналитиков. В связи с этим становится актуальной задача помочь пользователю на основе содержательной обработки текстов извлечь знания о коммуникативных целях говорящего, а также понять взаимосвязь семантических функций грамматических форм с намерениями говорящего, что важно и для современного прикладного языкознания (и поэтому в настоящее время технология Text Mining широко используется для решения различных задач, связанных с обработкой и интеллектуальным анализом текста (Белоногов, Гиляревский, Хорошилов, 2012; Большакова, Носков, 2010; Ермаков, Плешко, 2009; Лукашевич, 2011; Макаревич, 2019; Мусаев, Григорьев, 2021; Смирнов, 2023a; 2023b; Тихомиров, Смирнов, 2009)).

В связи с этим в исследовании решаются следующие задачи:

- при помощи Text Mining разработать алгоритм структурирования естественно-языкового текста для его формализации в информационных системах в соответствии с кластерами выбранных для анализа документов;
- разработать методы и алгоритм интеллектуальной обработки текста на основе контента лингвистических ресурсов, связанных с классификацией лингвистических ресурсов по классам-модусам ИСО;
- исследовать и разработать архитектуру автоматизированной системы смысловой обработки метадискурса интенциональности, модальной коррекции схем интенциональности со свойственными им когнитивами;
- выявить особенности структурной организации текстов-аннотаций, на основе которых похожие тексты будут объединены (классифицированы) в соответствующие категории;
- создать базу данных (перенос информации в структуры БД Datastores) на основе анализа проблем интеграции и совместимости данных с различными типами документов из разных источников.

При решении задач, поставленных в работе, использовались следующие методы исследования: 1) методы теоретического анализа для систематизации фактического материала и выведения определенных закономерностей в процессе описания; 2) методы иерархической и бинарной кластеризации, позволяющие структурирование естественно-языкового текста для его формализации в информационных системах в соответствии с кластерами документов, семантически связанных между собой (для репрезентативности выборки – использовать весь имеющийся корпус); 3) методы синтаксического и первичного семантического анализа естественно-языковых текстов, компонентного анализа, применявшиеся для интеллектуальной обработки текста на основе контента лингвистических ресурсов, связанных с классификацией лингвистических ресурсов по классам-модусам ИСО; 4) метод словарных дефиниций и концептуального анализа для смысловой обработки метадискурса интенциональности, модальной коррекции схем интенциональности со свойственными им когнитивами; 5) когнитивно-прагматический метод, позволивший выявить особенности структурной организации текстов-аннотаций, на основе которых похожие тексты будут объединены (классифицированы) в соответствующие категории; 6) методы теории проектирования баз данных для переноса информации об ИСО в структуры БД Datastores на основе анализа проблем интеграции и совместимости данных с различными типами документов из разнородных источников.

Материалом для исследования послужила авторская картотека, содержащая тексты с ИСО, извлеченные из произведений русских и чувашских писателей XIX-XX вв.

Теоретическую базу исследования составляют работы Г. Г. Белоногова, Р. С. Гиляревского, А. А. Хорошилова (2012), Е. И. Большаковой, А. А. Носкова (2010), А. Е. Ермакова, В. В. Плешко (2009), А. А. Мусаева, Д. А. Григорьева (2021), посвященные автоматической обработке текста. Кроме того, учитывались работы Г. С. Осипова, И. В. Смирнова (2016), А. М. Чеповского (2014), связанные с современными технологиями извлечения знаний из текстовых сообщений. Важным аспектом исследования явилось определение механизмов выявления смыслового содержания текста и кодирования его в структурированном виде на основе технологии Text Mining, в частности системы GATE, что стало возможным благодаря трудам Е. И. Большаковой, Н. В. Баевой, Е. А. Бордаченковой, Н. Э. Васильевой, С. С. Морозова (2007), М. В. Каменского (2014), Т. И. Макаревич (2019), А. А. Мусаева, Д. А. Григорьева (2021). Кроме того, учитывались работы Е. В. Заюковой (2004), Н. И. Клушиной (2012), которые представляют поле интенциональности в современной лингвистической парадигме.

Практическая значимость исследования состоит в возможности использования его материалов в прикладной лингвистике при проведении интеллектуального анализа текстов, в учебных целях при сравнительно-сопоставительном изучении русского и чувашского языков, в частности в процессе преподавания таких дисциплин, как «Сопоставительная грамматика русского и чувашского языков», «Практикум по художественному переводу», «Современный русский язык», «Современный чувашский язык», «Практический курс чувашского языка», «Введение в теорию коммуникации», «Лингвокультурология», «Прикладная лингвистика».

Обсуждение и результаты

Технология Text Mining в научных исследованиях выступает как один из методов интеллектуального анализа текстовой информации (информационный поиск, классификация и кластеризация текстов, резюмирование на основе методов компьютерной лингвистики) (Тузов, 2004; Чеповский, 2014; Швец, 2015).

В текстах содержатся, как известно, сведения, использующиеся в принятии решений в объективной действительности (Губанов, 1992; 2013; 2015a; 2015b; Губанов, Губанова, Свеклова, 2017; Губанов, Кожемякова, Губанова, 2023; Ермаков, Плешко, 2009; Шелманов, 2015). Для интеллектуального анализа текстов (ИАТ) к универсальным относятся инструменты, решающие частотные проблемы при анализе текстов: GATE (General Architecture for Text Engineering), Knime Analytics Platform, Orangesoftware, Rapid Miner (Губанов, 1992; Тузов, 2004; Чеповский, 2014; Швец, 2015).

Для исследователей-лингвистов, проводящих эксперименты с языком и вычислениями, система GATE по сравнению с другими платформами имеет следующие преимущества: а) количественная оценка (включает встроенную систему для сравнения данных аннотаций к документам и создания количественных показателей); б) совместная работа с другими платформами; в) имеет отношение к компьютерной лингвистике.

В двухфазной технологии аналитической обработки текста первая фаза (ETL) связана с автоматизированным анализом документов (структура их контента, хранилище исходной и аналитической информации и т. д.), а вторая фаза (OLAP, Text Mining, Data Mining) – с извлечением знаний из хранилища. Система GATE как библиотека (Language Resources (LR) – документы, корпуса и графы аннотаций как лингвистические ресурсы, Processing Resources (PR) – ресурсы, связанные с обработкой документов, Visual Resources (VR) – визуальные ресурсы) или как отдельное приложение позволяет структурировать текст и добавлять аннотации к фрагментам текста.

При помощи Text Mining мы осуществляем кластеризацию (формирование кластеров документов) текстового корпуса в выбранной нами программе. Для репрезентативности выборки мы использовали весь имеющийся корпус: корпус «книги» представили 3 кластерами (соответствующая информация чаще всего в данном формате предоставлена в Интернете): 1) набор учебных пособий и монографий; 2) авторефераты (корпус «Авторефераты» взят с сайта http://dissers.ru, а также с сайта Санкт-Петербургского государственного электротехнического университета «ЛЭТИ» им. В. И. Ульянова (Ленина) (http://www.eltech.ru/ru/)); 3) художественная литература (2 подкластера – русская и чувашская литература). В данной работе используются два корпуса текстов, которые представлены на Рис. 1.

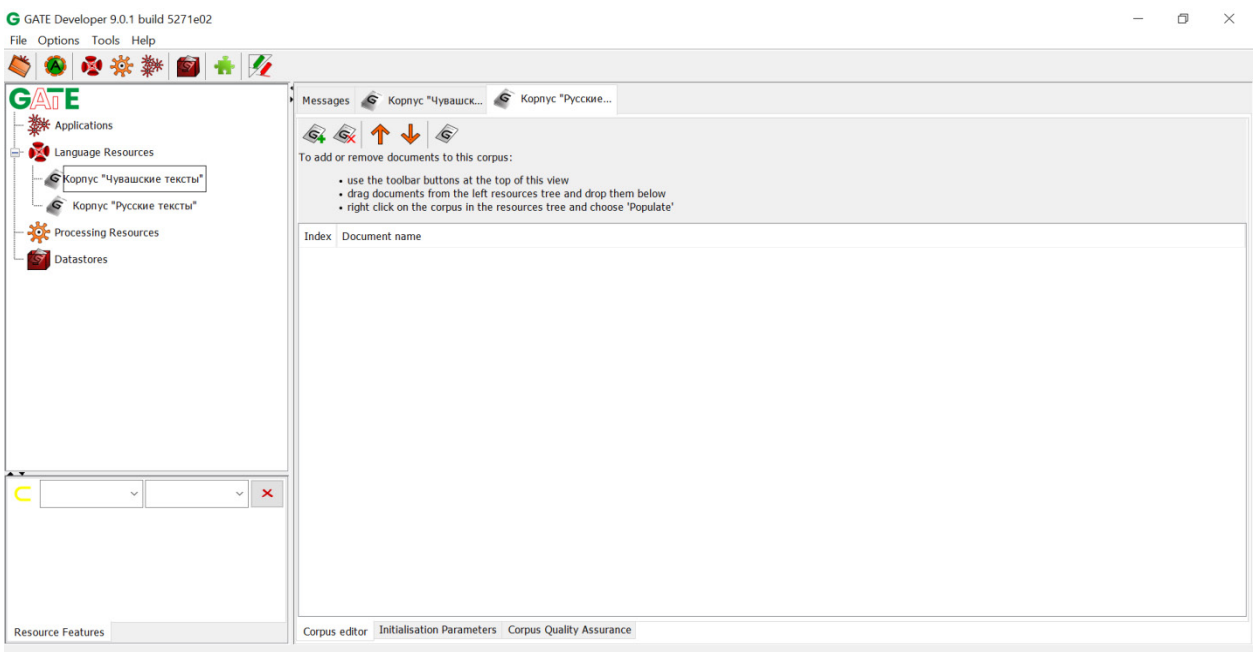

Рисунок 1. *Корпуса текстов в системе GATE*

Следующий этап ИАТ связан с контентом лингвистических ресурсов, с классификацией лингвистических ресурсов по классам (модусам ИСО): выявление в текстах сведений, необходимых для решения тех или иных задач, связанных с ИСО, в частности, в GATE соответствующие классы (потенциальная возможность ИСО) реализуются в редакторе документов (Рис. 2).

В метадискурсе интенциональности (модальная коррекция схем интенциональности), как видно из Рис. 2, компоненты интеллектуальной деятельности идентифицируются и квалифицируются по следующим уровням: 1) эпистемический модус (знание/предположение); 2) онтологический модус (возможность/необходимость/желательность); 3) аксиологический модус (информация с точки зрения ценностной системы объективной действительности).

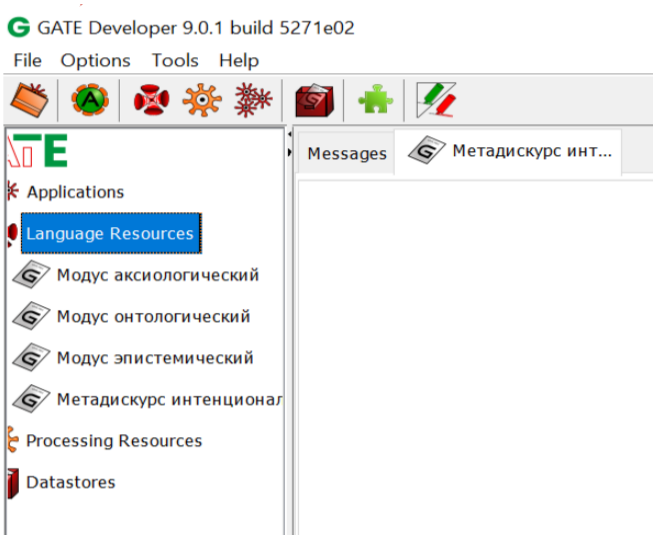

Рисунок 2. *Классификация лингвистических ресурсов*

В документах, содержащих соответствующий метадискурс, можно организовать поиск нужной информации в тексте (retrieval information), например, сведений о маркерах-шаблонах, поиск и извлечение данных шаблона «очевидно» эпистемического модуса (Рис. 3).

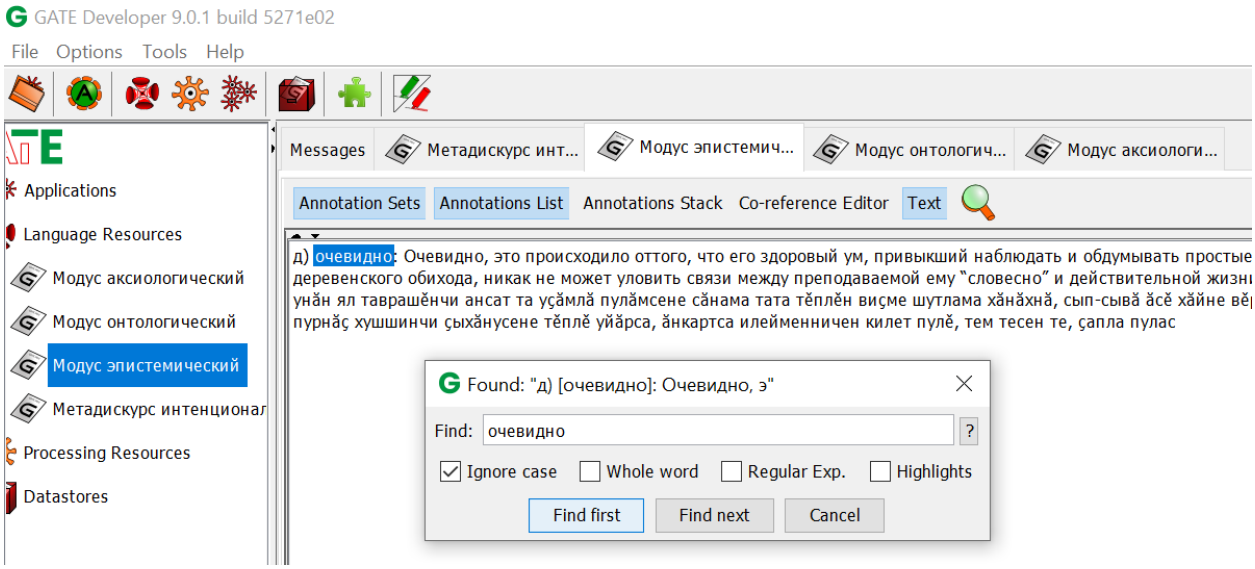

Рисунок 3. *Маркеры-шаблоны эпистемического модуса*

Основные проявления информации в системе GATE позволяют построить систему координат эпистемического модуса ИСО, в частности, по следующим градационным параметрам: а) параметр допустимости – количественная оценка знаний об объективной действительности (типы информации – общая и условная); б) параметр убежденности – субъективная оценка пропозиции.

Модусная коррекция компонентов ИСО по параметру «общая информация» в русском языке оформляется такими когнитивами, как *видимо*, *по-видимому*, *очевидно*, *наверное*, *видно*, *как видно*, *должно быть*. Система GATE выдает следующие модальные коррекции по частотности, связанные с когнитивами: А вот теперь, *видно*, за грехи мои, и ее пришлось пережить (Толстой). / *чув.:* Халь ак вăт, хамăн çылăхсене пула, *ахăр*, аннӳнтен те вилмесĕр юлтăм; *Очевидно*, это происходит оттого, что его здоровый ум, привыкший наблюдать и обдумывать простые и ясные явления деревенского обихода, никак не может уловить связи между преподаваемой ему «словесностью» и действительной жизнью (Куприн). / *чув.:* Ку, *ахăртнех*, унăн ял таврашĕнчи ансат та уçăмлă пулăмсене сăнама тата тĕплĕн виçме, шутлама хăнăхнă, сып-сывă ăсĕ хăйне вĕрентекен «сăмахлăхпа» чăн пурнăç хушшинчи çыхăнусене тĕплĕ уйăрса, ăнкартса илейменничен килет пулĕ, тем тесен те, *çапла пулас*; Отчего это? *Должно быть*, оттого, что я сбыл груз души в письмо (Гончаров). / *чув.:* Мĕне пула ку? Ку эпĕ хамăн чĕрери пĕтĕм хурлăхлă шухăшсене çырура çырса пĕлтернĕрен *пуль*; Спотыкаясь, он шел к ней и видел, как мать качнулась от его крика, словно от удара, – хотела, *видно*, бежать, но силы изменили, и она пошла толчками, будто преодолевая сопротивление ветра (Шолохов). / *чув.:* Вăл такăна-такăна ун патнелле утса кайрĕ, амăшĕ, хăйне кăшкăрса

чĕннине илтсе такам чышса янă пек сулăнса кайнине курчĕ: ахăртнех, чупасшăн пулчĕ вăл, анчах, вăйĕ пĕтсе килнипе, тайкаланса утса кайрĕ; *Наверное*, сначала оцепенеет от ужаса, потом задрожит от бешенства, а потом выпалит как из мортиры (Куприн). / *чув.:* Малтан, хăраса ӳкнипе, шанкас хытса тăрĕ тен, унтан урмăшнăран чĕтреме пуçлĕ, кайран вара, мортирăран пенĕ пекех пĕрĕхтерĕ; Проснулся он оттого, что по лбу его что-то ползало – *наверное*, паучок или какой-нибудь червяк (Шолохов). / *чув.:* Вăл хăйĕн çамки тăрăх темскер шунипе вăранчĕ – эрешмен е мĕнле те пулин хурт пулас.

Суммаризацию (summarization) или аннотирование (краткую справку) обозначим на вкладках Annotation Sets и Annotations List, на основе которых похожие тексты будут объединены (классифицированы) в соответствующие категории (Рис. 4).

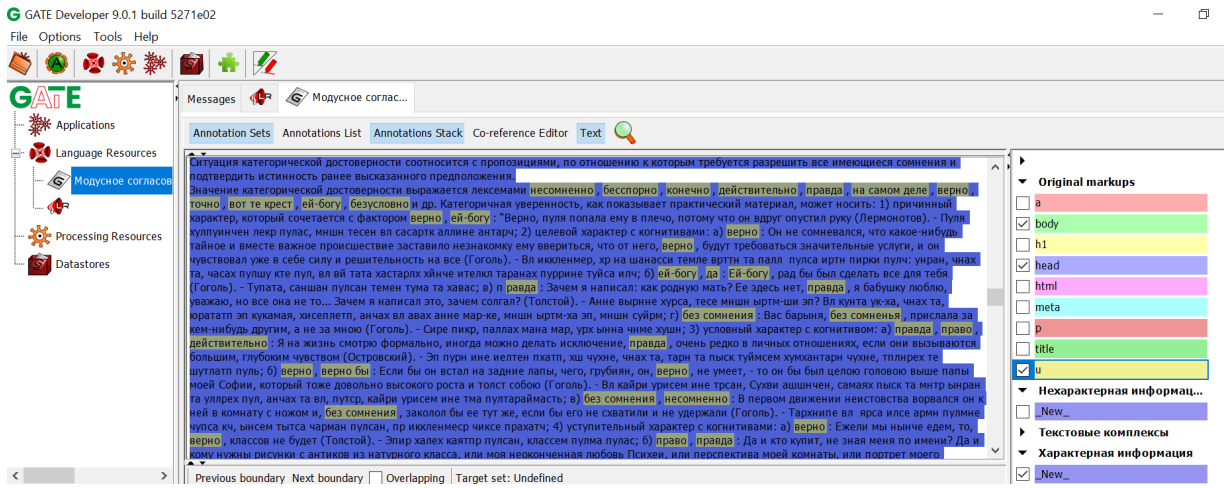

Рисунок 4. *Аннотация*

Аннотации группируются в Annotation Sets, где хранятся несколько вариантов разметки для одного документа.

Следующий этап ИАТ связан с интеллектуальным анализом данных (ИАД, англ. Data Mining), с хранилищами и базами данных (с различными типами документов из разных источников), так как технология Data Mining извлекает востребованные знания в базах данных. Для хранения документов/корпусов и процессов для дальнейшей работы используем хранилище данных (Datastores). Следует отметить, что частотным является хранилище Serial Datastore (Рис. 5).

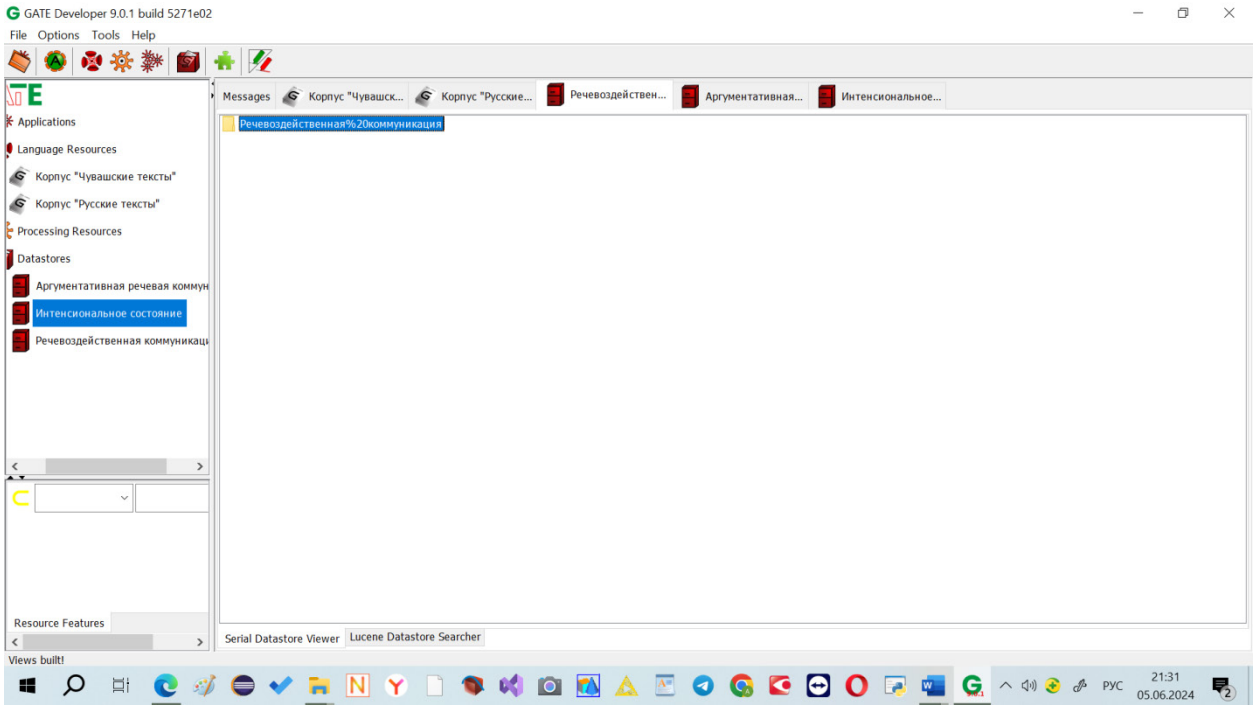

Рисунок 5. *Хранилище данных (Datastores)*

Предлагаемый подход помогает совершенствовать технологии смыслового анализа естественно-языкового текста.

Заключение

Выбранный в работе подход, таким образом, способствует совершенствованию технологий смыслового анализа естественно-языкового текста. Формализованное описание на основе выбранной архитектуры автоматизированной системы смыслового анализа анализируемых текстов в системе GATE позволило решать задачи конкретного прикладного характера (логические, структурированные описания ИСО, категоризация, поиск информации и анализ данных). В данном исследовании предложен подход, позволяющий формализовать структуру ИСО для использования в разработанных процедурах автоматического извлечения из текста выбранной информации. Для формализации и определения смыслового содержания неструктурированных текстов в системе обработки языка GATE были использованы основные компоненты, такие как Applications (для процесса обработки отдельных документов текста, корпусов текстов), лингвистические ресурсы (документы, корпуса) (для релевантной информации), процессинговые ресурсы (для аннотирования, определения смыслового содержания текста), а также компонент Datastores (для сохранения документов/корпусов и процессов для дальнейшего использования).

В GATE, следует отметить, имеются готовые инструменты, связанные с тестированием результатов обработки текста. В перспективе можно использовать для сравнения пар аннотаций вкладку Annotation Diff Tool, а для проверки качества корпуса в окне выбранного корпуса – вкладку Corpus Quality Assurance (CQA). Анализируемый в статье подход к извлечению текстовой информации (интенциональных смысловых отношений) и предложенные в его рамках процедуры извлечения могут быть использованы при решении задачи прикладной лингвистики – интеллектуального анализа текста.

Источники | References

1. Белоногов Г. Г., Гиляревский Р. С., Хорошилов А. А. Проблемы автоматической смысловой обработки текстовой информации // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2012. № 11.
2. Большакова Е. И., Баева Н. В., Бордаченкова Е. А., Васильева Н. Э., Морозов С. С. Лексико-синтаксические шаблоны в задачах автоматической обработки // Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции «Диалог 2007». М.: РГГУ, 2007.
3. Большакова Е. И., Носков А. А. Программные средства анализа текстов на основе лексико-синтаксических шаблонов языка LSPL // Программные системы и инструменты: тематический сборник. 2010. № 11.
4. Губанов А. Р. Машинный фонд чувашского языка и его компоненты // Актуальные вопросы истории и культуры чувашского народа: сборник. Чебоксары: ЧГИГН, 2013.
5. Губанов А. Р. Морфологический стандарт для систем автоматической обработки текстов на чувашском языке и архитектура грамматического словаря // Актуальные вопросы истории и культуры чувашского народа: сборник статей. Чебоксары: ЧГИГН, 2015a. Вып. 3.
6. Губанов А. Р. Национальный корпус чувашского языка: создание лексического поисковика в системе Java // Актуальные вопросы истории и культуры чувашского народа: сборник статей. Чебоксары: ЧГИГН, 2015b. Вып. 3.
7. Губанов А. Р. Семантико-синтаксические особенности предложений с предикатами интенционального состояния в русском и чувашском языках // Высшая школа – народному хозяйству Чувашии. Гуманитарные науки: тез. докл. / Чуваш. гос. ун-т им. И. Н. Ульянова. Чебоксары, 1992.
8. Губанов А. Р., Губанова Г. Ф., Свеклова О. В. Тезаурус чувашского языка (чăваш пĕлĕвĕн мулĕ) как языковая система знаний // Вестник Чувашского университета. Гуманитарные науки. 2017. № 2.
9. Губанов А. Р., Кожемякова Е. А., Губанова Г. Ф. Онтологические модели пословиц как прецедентных текстов (на материале разноструктурных моделей в русском и чувашском языках) // Этническая культура. 2023. Т. 5. № 2.
10. Ермаков А. Е., Плешко В. В. Семантическая интерпретация в системах компьютерного анализа текста // Информационные технологии. 2009. Т. 6.
11. Заюкова Е. В. Семантические и прагматические особенности лексических средств выражения интенциональности // Актуальные проблемы гуманитарного знания: материалы региональной научно-практической конференции молодых ученых. Барнаул, 2004.
12. Каменский М. В. Лингвистическая платформа GATE как среда автоматизированного анализа когнитивно-функциональных свойств дискурсных маркеров // Вестник Северо-Кавказского федерального университета. 2014. № 3 (42).
13. Клушина Н. И. Интенциональный метод в современной лингвистической парадигме // Медиастилистика. 2012. Вып. 4.
14. Лукашевич Н. В. Тезаурусы в задачах информационного поиска. М.: Изд-во Московского ун-та, 2011.
15. Макаревич Т. И. Интеллектуальный анализ текстовой информации в специализированных областях в системе электронного правительства // Цифровая трансформация. 2019. № 2 (7).
16. Мусаев А. А., Григорьев Д. А. Обзор современных технологий извлечения знаний из текстовых сообщений // Компьютерные исследования и моделирование. 2021. Т. 13. № 6.

17. Осипов Г. С., Смирнов И. В. Семантический анализ научных текстов и их больших массивов // Системы высокой доступности. 2016. № 1.
18. Смирнов И. В. Интеллектуальный анализ текстов на основе методов разноуровневой обработки естественного языка: монография. М.: ФИЦ ИУ РАН, 2023a.
19. Смирнов И. В. Разноуровневая обработка естественного языка для интеллектуального поиска и анализа текстов // Искусственный интеллект и принятие решений. 2023b. № 1.
20. Тихомиров И. А., Смирнов И. В. Применение методов лингвистической семантики и машинного обучения для повышения точности и полноты поиска в поисковой машине Exactus // Труды международной конференции «Диалог 2009». М., 2009.
21. Тузов В. А. Компьютерная семантика русского языка. СПб.: Изд-во С.-Петерб. ун-та, 2004.
22. Чеповский А. М. Информационные модели в задачах обработки текстов на естественных языках. М.: Национальный открытый университет «Интуит», 2014.
23. Швец А. В. Взаимодействие информационных и лингвистических методов в задачах анализа качества научных текстов: дисс. … к. техн. н. М., 2015.
24. Шелманов А. О. Исследование методов автоматического анализа текстов и разработка интегрированной системы семантико-синтаксического анализа: дисс. … к. техн. н. М., 2015.

Информация об авторах | Author information

RU

Губанов Алексей Рафаилович[1], д. филол. н., проф.
Данилов Андрей Анатольевич[2], д. ист. н., доц.
Исаев Юрий Николаевич[3], д. филол. н., доц.
Губанова Галина Федоровна[4], к. филол. н.
[1, 2, 4] Чувашский государственный университет имени И. Н. Ульянова, г. Чебоксары
[3] Чувашский государственный институт гуманитарных наук, г. Чебоксары

EN

Aleksey Rafailovich Gubanov[1], Dr
Andrey Anatolyevich Danilov[2], Dr
Yuri Nikolaevich Isaev[3], Dr
Galina Fedorovna Gubanova[4], PhD
[1, 2, 4] Chuvash State University, Cheboksary
[3] Chuvash State Institute of Humanities, Cheboksary

[1] *alexgubm@gmail.com,* [2] *danilov.andrey@mail.ru,* [3] *isaev2828@yandex.ru,* [4] *rggalina@gmail.com*

Информация о статье | About this article

Дата поступления рукописи (received): 10.07.2024; опубликовано online (published online): 09.09.2024.

Ключевые слова (keywords): искусственный интеллект; Text Mining; GATE; Data Mining; разноструктурные языки; интенциональные смысловые отношения (ИСО); artificial intelligence; languages with different structures; intentional semantic relations.