http://dx.doi.org/10.35596/1729-7648-2023-29-1-57-63

Оригинальная статья
Original paper

УДК 004.912

ОНТОЛОГИЧЕСКИЙ ПОДХОД К ПРИОБРЕТЕНИЮ ЗНАНИЙ ИЗ ТЕКСТОВ ЕСТЕСТВЕННОГО ЯЗЫКА

ЛУНВЭЙ ЦЯНЬ

Белорусский государственный университет информатики и радиоэлектроники
(г. Минск, Республика Беларусь)

Поступила в редакцию 27.10.2022

Аннотация. Главная задача приобретения знаний (также называемая извлечением знаний) из текстов естественного языка – это извлечение знаний из текстов естественного языка в фрагмент базы знаний интеллектуальной системы. С учетом ознакомления с соответствующей литературой о приобретении знаний в стране и за рубежом в статье анализируются преимущества и недостатки классического подхода к извлечению знаний. После тщательного исследования технологии извлечения знаний на основе правил и методов построения онтологий лингвистики предложено решение для реализации извлечения знаний на основе технологии OSTIS. Основной особенностью этого решения является построение единой семантической модели, которая может использовать онтологии лингвистики (в основном синтаксический и семантический аспекты) и интегрировать различные модели решения задач (например, модели на основе правил, модели нейронных сетей) для решения извлечения знаний из текстов естественного языка.

Ключевые слова: онтология, база знаний, обработка естественного языка, извлечение знаний, интеллектуальная система.

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Для цитирования. Лунвэй Цянь. Онтологический подход к приобретению знаний из текстов естественного языка / Лунвэй Цянь // Цифровая трансформация. 2023. Т. 29, № 1. С. 57–63. http://dx.doi.org/10.35596/1729-7648-2023-29-1-57-63.

ONTOLOGY-BASED KNOWLEDGE ACQUISITION METHOD FOR NATURAL LANGUAGE TEXTS

LONGWEI QIAN

Belarusian State University of Informatics and Radioelectronics (Minsk, Republic of Belarus)

Submitted 27.10.2022

Abstract. The main task of knowledge acquisition (also named knowledge extraction) from natural language texts is to extract knowledge from natural language texts into fragment of knowledge base of intelligent system. Through the induction of the related literature about knowledge acquisition at a home country and abroad, this paper analyses the strengths and weaknesses of the classical approach. After emphatically researching the rule-based knowledge extraction technology and the method of building ontology of linguistics, this article proposes a solution to the implementation of knowledge acquisition based on the OSTIS technology. The main feature of this solution is to construct a unified semantic model that is able to utilize ontologies of linguistics (mainly, syntactic and semantic aspect) and integrate various problem-solving models (e. g., rule-based models, neural network models) for solving knowledge extraction process from natural language texts.

Keywords: ontology, knowledge base, natural language processing, knowledge extraction, intelligent system.

Conflict of interests. The author declares no conflict of interests.

For citation. Longwei Qian (2023) Ontology-Based Knowledge Acquisition Method for Natural Language Texts. *Digital Transformation.* 29 (1), 57–63. http://dx.doi.org/10.35596/1729-7648-2023-29-1-57-63 (in Russian).

Введение

В настоящее время приложения интеллектуальных систем, основанные на фактологических знаниях, начали служить людям в разных сферах их жизнедеятельности. Одна из технических трудностей при разработке этого типа приложений заключается в необходимости извлечения структурированных фактологических знаний для построения цифровых баз знаний в интеллектуальных системах. Текст естественного языка – наиболее распространенная форма неструктурированных данных. Использование автоматизированных подходов для обнаружения и извлечения фактологических знаний из неструктурированных данных для различных интеллектуальных систем становится все более актуальной задачей. Цель исследований автора – построение единой семантической модели на основе технологии OSTIS [1] к извлечению фактологических знаний (сущностей и отношений между ними) из разных текстов естественного языка. По сравнению с системами извлечения знаний из закрытых областей, в которых нужно определить конкретные типы для извлечения фактологических знаний, предлагаемый подход предоставляет возможность извлечения фактологических знаний из открытой области без заранее определенных конкретных типов.

Сопутствующие работы

Задачу извлечения фактологических знаний из неструктурированного текста разделяют на два направления: извлечение знаний из закрытых и открытых областей. Данная задача заключается в получении структурированных данных, которые затем выражаются в некоторой форме представления знаний. Таким образом, основная цель этого процесса – извлечение именованных сущностей и отношений между ними из текстов естественного языка, причем извлеченные результаты сохраняются в форме внутреннего языка представления знаний интеллектуальной системы, например, в RDF, SC-коде [1] и др. С точки зрения SC-кода отношения рассматриваются как особенная сущность, которая извлекается из текстов естественного языка и представляется в базе знаний интеллектуальной системы.

Для извлечения фактологических знаний из закрытых областей часто нужны заранее определенные типы именованных сущностей и отношений между ними. В данных системах извлечение фактологических знаний с использованием машинного обучения или модели нейронных сетей всегда требует очень мощного оборудования. Однако цель извлечения фактологических знаний из открытой области состоит в том, чтобы извлекать различные наборы отношений между именованными сущностями из массивных и разнородных текстов естественного языка без требований заранее заданного словаря для определения типа данных отношений.

Извлечение фактологических знаний из открытых областей напрямую определяет относительное словосочетание в тексте, анализируя текст естественного языка (в частности, предложение), чтобы реализовать моделирование классификации именованных сущностей и отношений между ними без необходимости заранее определять категории отношений. Системы ReVerb [2] и OLLIE [3] используют общий синтаксис и лексические ограничения для извлечения фактологических знаний для английских предложений из открытых областей. Однако из-за разнообразия и сложности языков способ сегментации слов и реализация грамматических функций заметно различаются (например, в английском и китайском).

Согласно письменной особенности текстов китайского языка, в предложении иероглифы пишутся один за другим, между ними не существуют естественные пробелы. Синтаксическая структура и категория частей речи для обработки английского языка не полностью применимы к обработке китайского текста. Авторы [4] предложили классическую архитектуру CORE для извлечения именованных сущностей и отношений между ними из открытой области для китайских предложений.

Несмотря на то что большое количество систем извлечения фактологических знаний реализовано и успешно применяется на практике, остаются проблемы, которые не решились ни в одном

из перечисленных методов. При решении однотипной задачи извлечения знаний накладные расходы увеличиваются из-за модификации модели и добавления в систему новых моделей. В существующих системах отсутствует единая основа для представления любых видов знаний в единой базе знаний, используемых для анализа текстов естественного языка, в том числе синтаксических и семантических знаний, а также правил для их извлечения. Таким образом, синтаксические, семантические знания и правила извлечения необходимо многократно разрабатывать в разных системах, что значительно усложняет создание прикладной системы и соответственно накладные расходы.

Онтологический подход

Рассматривается онтологический подход, основанный на технологии OSTIS, для извлечения фактологических знаний (в основном именованные сущности и отношения между ними из открытой области для текстов естественного языка). Он устраняет ограничения ранее принятых онтологических подходов, которые применялись только для извлечения знаний из закрытых областей. Предлагаемый автором подход ориентирован на извлечение фактологических знаний из открытых областей, он выполняет синтаксически-семантический анализ текстов естественного языка с помощью построения онтологии лингвистики, а затем напрямую извлекает фрагменты базы знаний (фактологические знания в виде SC-кода), построенной в логической онтологии. Также обеспечивает единую унифицированную основу для объединения синтаксических и семантических знаний, логических правил для анализа текстов естественного языка и извлечения фактологических знаний в единую базу знаний лингвистики.

Технология OSTIS направлена на разработку класса систем, которые названы управляемыми знаниями компьютерными системами (OSTIS-системами). Компонент, реализующий извлечения знаний из текста естественного языка, обычно разрабатывается как часть естественно-языкового интерфейса OSTIS-системы. OSTIS-система – это интеллектуальная система, основанная на знаниях, состоящая из базы знаний, представленной на внутреннем языке SC-кода, и решателей задач, объединяющих все программные агенты для решения конкретных проблем. SC-код является формой языка семантической сети с базовой теоретико-множественной интерпретацией. Элементы таких семантических сетей называются sc-элементами (sc-узлами и sc-коннекторами, которые, в свою очередь, могут быть sc-дугами или sc-ребрами, в зависимости от направленности). Для внешнего представления абстрактных sc-текстов используются несколько внешних форм отображения, таких как SCg-код, SCn-код, SCs-код.

На основе анализа структуры базы знаний OSTIS-систем, представленной в SC-коде, можно определить цель извлечения знаний. Чтобы преобразовать тексты естественного языка в sc-структуру, необходимо описать синтаксические и семантические знания конкретного естественного языка для анализа текста, а также построение правил извлечения. Более того, в технологии OSTIS база знаний построена как иерархическая система предметных областей и соответствующих онтологий [5]. Каждая онтология представляет собой спецификацию системы понятий, используемых в соответствующей предметной области. В каждой предметной области описаны различные отличительные онтологии, отражающие определенный набор особенностей понятий в предметной области, например, терминологическая онтология, логическая онтология, теоретико-множественная онтология и т. д. Ниже приводится общая иерархия предметной области лингвистики на SCn-коде.

Предметная область лингвистики
=> Частная ПО:*
- *Предметная область текстов китайского языка*
- *Предметная область текстов английского языка*
- *Предметная область текстов русского языка*

Из приведенной общей структуры видно, что для реализации анализа текстов конкретного естественного языка необходимо построить предметную область текстов конкретного естественного языка. В качестве примера используем анализ текста китайского языка. Приведем построенную предметную область текстов китайского языка (более подробное ее объяснение дано в [6]).

Предметная область текстов китайского языка
=> Частная ПО:*
- *Раздел. Предметная область лексического анализа*
- *Раздел. Предметная область синтаксического анализа*
- *Раздел. Предметная область семантического анализа*

Все три раздела описывают спецификацию системы понятий, логических правил (например, правил извлечения) и других знаний с лексического, синтаксического и семантического аспектов китайского языка соответственно. На рис. 1 на SCg-коде указано простое правило извлечения, используемое для извлечения sc-конструкции, которое представлено логическим утверждением в логической онтологии.

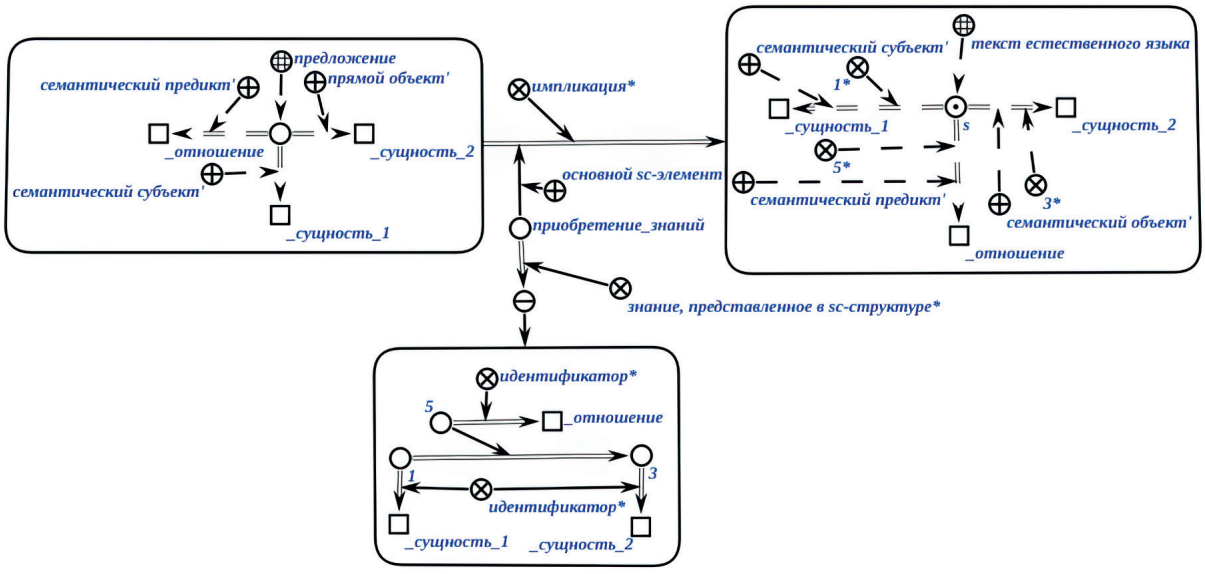

Рис. 1. Логическое утверждение о правиле извлечения в логической онтологии
Fig. 1. Logical statement about extraction rule in the logical ontologies

Используя лингвистические знания в базе знаний лингвистики для реализации автоматического извлечения sc-структур из текстов естественного языка на основе ряда технологий обработки текста, т. е. лексического, синтаксического, семантического анализа и правил извлечения, необходимо разработать решатель задач. В рамках технологии OSTIS решатель задач интерпретируется как иерархическая система агентов (sc-агентов) [7]. Многоагентный подход предоставляет возможность комбинировать разные модели решения задач при выполнении одной и той же сложной задачи, а также добавлять новые их модели в решатель. Агенты могут реализовывать как логические рассуждения на основе иерархии утверждений, так и алгоритмы обучения на основе данных с использованием различных языков программирования. В соответствии с классической архитектурой далее приведена общая структура решателя задач на SCn-коде для перевода текстов внешнего языка во фрагменты базы знаний.

Абстрактный sc-агент трансляции внешних текстов в фрагменты базы знаний
<= Декомпозиция абстрактного sc-агента:*
{
- *Абстрактный sc-агент лексического анализа*
- *Абстрактный sc-агент синтаксического анализа*
- *Абстрактный sc-агент семантического анализа*
- *Абстрактный sc-агент генерации sc-структур*

}

Абстрактные sc-агенты лексического анализа – группы агентов, реализующие механизмы декомпозиции входных внешних текстов на лексические единицы. Компоненты внешних текстов могут быть определены. Абстрактные sc-агенты синтаксического и семантического анализа –

агенты, реализующие механизмы построения синтаксической и семантической структур внешних текстов. Абстрактные sc-агенты генерации sc-структур – агенты, реализующие механизмы интеграции сематического эквивалента sc-текста в базу знаний конкретной OSTIS-системы.

Реализация извлечения фактологических знаний для китайского языка

Общий процесс извлечения структурированных фактологических знаний из разных текстов естественного языка аналогичен. С точки зрения технологии OSTIS любой текст из внешнего ввода в OSTIS-систему представляется в виде файла (то есть sc-узла с содержимым). На рис. 2 приведен пример такого узла, в котором указан конкретный фрагмент текста китайского языка, описывающий конечное множество, тройку и ориентированное множество.

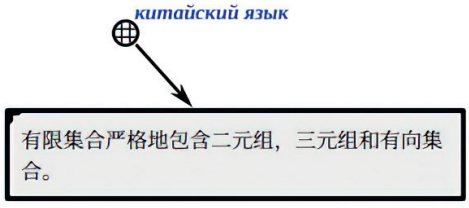

Рис. 2. Представление текста китайского языка в OSTIS-системе
Fig. 2. Representation of Chinese language text in the OSTIS-system

С помощью ранее построенных онтологий лингвистики и решателей задач в OSTIS-системе рассмотрим обработку указанного на рис. 2 фрагмента текста естественного языка, чтобы проиллюстрировать каждый из этапов извлечения знаний из этого текста. Без какого-либо определенного словаря этот текст китайского языка рассматривается как единственный ввод в OSTIS-систему, а его вывод представляет собой извлеченные sc-структуры. Для большинства задач обработки китайского языка основным предварительным этапом является автоматическая сегментация слов в текстах. Для обработки китайского языка был предложен Стандарт сегментации слов современного китайского языка, используемый для обработки информации, в котором слово представлено единицей сегментации. Отличительная черта китайского языка – у китайского слова есть только одна форма и нет никаких изменяемых форм, таких как единственное и множественное число, временная форма, падежи. Извлечение знаний из текста китайского языка представлено следующими этапами.

Этап 1. Лексический анализ текста китайского языка раскладывает предложение, приведенное на рис. 2, на отдельные лексические единицы сегментаций.

Этап 2. Синтаксический анализ выполняет переход от данных отдельных единиц сегментаций текста к его синтаксической структуре, которая переносится в sc-конструкцию (рис. 3).

В приведенном на рис. 3 фрагменте видно, что между текстом китайского языка и отдельными лексическими единицами возникают отношения, характеризующие его синтаксические отношения в предложении. Отношения описываются как онтологии в соответствующей предметной области.

Этап 3. Семантический анализ выполняет анализ семантических взаимоотношений в тексте и переход от обработанного текста к sc-конструкции с помощью соответствующей предметной области (рис. 4). Представленный на рис. 4 фрагмент базы знаний включает в себя семантическую структуру данного текста, установленную отношениями зависимости.

Этап 4. Генерация sc-структур извлекает семантический эквивалент sc-текста в фрагмент базы знаний конкретной OSTIS-системы на основе предыдущего анализа текста и правил извлечения, описанных в соответствующей предметной области. Конечная извлеченная sc-конструкция представлена на рис. 5.

Согласно построенному на рис. 5 правилу извлечения, семантическое подлежащее и прямое дополнение текста извлекаются как именованные сущности в базе знаний. В семантической структуре существует отношение «согласованные аргументы» между тремя прямыми дополнениями, которые извлекаются вместе как именованные сущности. Семантическое сказуемое текста служит отношением между этими именованными сущностями. Процесс также включает связывание для sc-элементов и устранение противоречий.

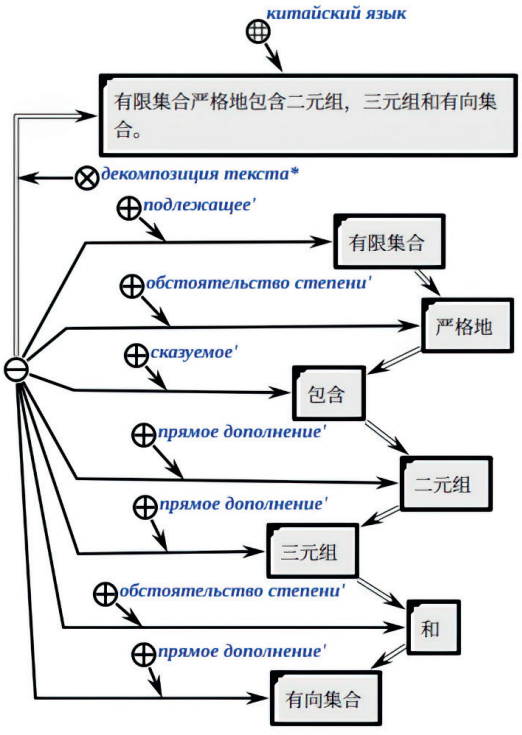

Рис. 3. Результат синтаксического анализа текста китайского языка
Fig. 3. The result of syntactic analysis of Chinese language text

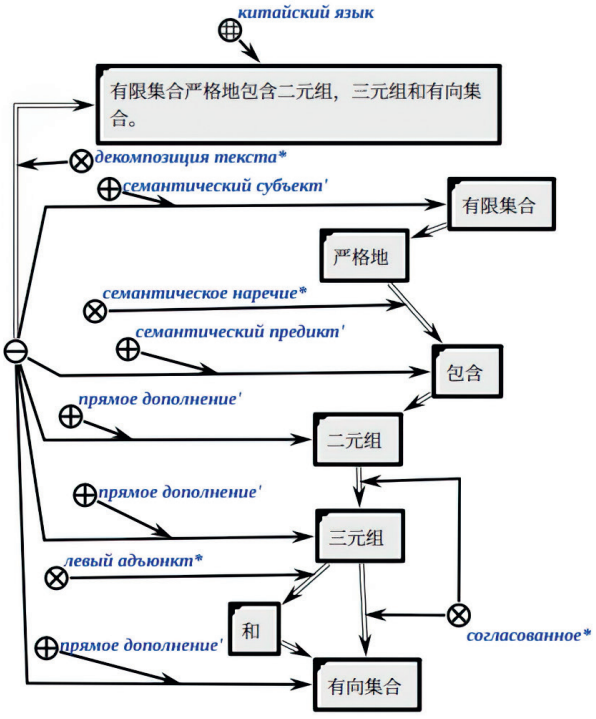

Рис. 4. Результат семантического анализа текста китайского языка
Fig. 4. The result of semantic analysis of Chinese language text

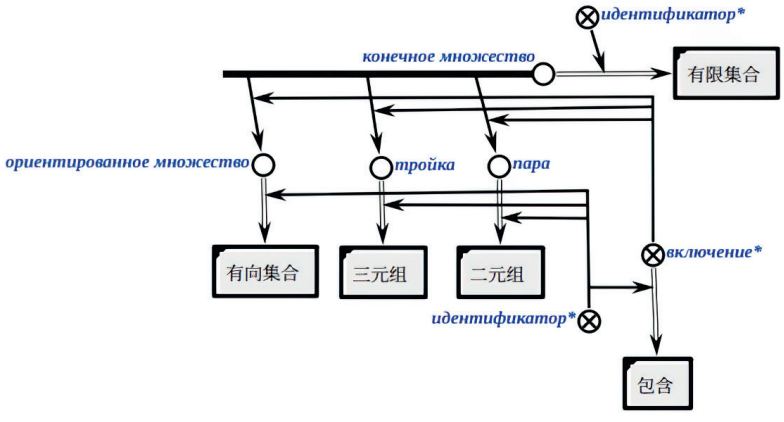

Рис. 5. Извлеченная sc-структура в базе знаний
Fig. 5. The extracted sc-structure in the knowledge base

Заключение

На основе предложенной единой семантической модели может реализоваться автоматическое извлечение фактологических знаний в виде SC-кода (именованные сущности и отношения между ними) без какого-либо ручного вмешательства (т. е. заранее определенные конкретные типы именованных сущностей и отношений) посредством синтаксически-семантического анализа текстов и логических правил из области лингвистики. Модель, построенная на основе технологии OSTIS, реализует унифицированное управление онтологией лингвистики, правилами извлечения фактологических знаний и решателями задач, эффективно сокращает сложность и временные затраты на разработку системы извлечения фактологических знаний.

Список литературы

1. Голенков, В. В. Проект открытой семантической технологии компонентного проектирования интеллектуальных систем. Ч. 1. Принципы создания / В. В. Голенков, Н. А. Гулякина // Онтология проектирования. 2014. № 4. С. 42–64.
2. Fader, A. Identifying Relations for Open Information Extraction / A. Fader, S. Soderland, O. Etzioni // In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Edinburgh, Scotland, UK. 2011. P. 1535–1545.
3. Open Information Extraction: the Second Generation / O. Etzioni [et al.] // International Joint Conference on Artificial Intelligence (IJCAI'11). Barcelona: AAAI Press, 2011. P. 3–10.
4. Tseng, Y. H. Chinese Open Relation Extraction for Knowledge Acquisition / Y. H. Tseng, L. H. Lee // Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, April 26–30, 2014. P. 12–16.
5. Davydenko, I. T. Ontology-based Knowledge Base Design / I. T. Davydenko // Open Semantic Technologies for Intelligent Systems (OSTIS–2017): Materials of the International Scientific and Technical Conference, Minsk, 16–18 Feb. 2017. Minsk: BSUIR, 2017. P. 57–72.
6. Цянь Лунвэй. Онтологический подход к обработке текстов китайского языка / Цянь Лунвэй // Доклады БГУИР. 2020. Т. 18, № 6. С. 49–56. http://dx.doi.org/10.35596/1729-7648-2020-18-6-49-56.
7. Shunkevich, D. V. Ontology-based Design of Knowledge Processing Machines / D. V. Shunkevich // Open Semantic Technologies for Intelligent Systems (OSTIS–2017): Materials of the International Scientific and Technical Conference, Minsk, 16–18 Feb. 2017. Minsk: BSUIR, 2017. С. 73–94.

References

1. Golenkov V. V., Guliakina N. A. (2014) Project of Open Semantic Technology of the Componental Design of Intelligent Systems. Part 1. The Principles of Creation. *Ontology Designing*. (4), 42–64 (in Russian).
2. Fader A., Soderland S., Etzioni O. (2011) Identifying Relations for Open Information Extraction. *In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Edinburgh, Scotland, UK*. 1535–1545.
3. Etzioni O., Fader A., Christensen J., Soderland S., Mausam M. (2011) Open Information Extraction: the Second Generation. *International Joint Conference on Artificial Intelligence (IJCAI'11).* Barcelona, AAAI Press. Publ. 3–10.
4. Tseng Y. H., Lee L. H. (2014) Chinese Open Relation Extraction for Knowledge Acquisition. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, April 26–30*. 12–16.
5. Davydenko I. T. (2017) Ontology-Based Knowledge Base Design. *Open Semantic Technologies for Intelligent Systems (OSTIS–2017): Materials of the International Scientific and Technical Conference, Minsk, 16–18 Feb. 2017.* Minsk, Belarusian State University of Informatics and Radioelectronics. 57–72.
6. Qian Longwei (2020) Ontological Approach to Chinese Text Processing. *Doklady BGUIR*. 18 (6), 49–56. http://dx.doi.org/10.35596/1729-7648-2020-18-6-49-56 (in Russian).
7. Shunkevich D. V. (2017) Ontology-Based Design of Knowledge Processing Machines. *Open Semantic Technologies for Intelligent Systems (OSTIS–2017): Materials of the International Scientific and Technical Conference, Minsk, 16–18 Feb. 2017.* Minsk, Belarusian State University of Informatics and Radioelectronics. 73–94.

Вклад автора / Author's contribution

Лунвэй Цянь осуществил постановку задачи для извлечения знаний из текстов естественного языка, предложил новый подход к решению задачи, подготовил рукопись статьи / Qian Longwei formulated the problem for extracting knowledge from natural language texts, proposed a new approach to solving the problem, and prepared the manuscript of the article.

Сведения об авторе

Лунвэй Цянь, аспирант кафедры интеллектуальных информационных технологий Белорусского государственного университета информатики и радиоэлектроники

Адрес для корреспонденции

220013, Республика Беларусь,
г. Минск, ул. П. Бровки, 6
Белорусский государственный университет
информатики и радиоэлектроники
Тел.: +375 29 721-60-63
E-mail: qianlw1226@gmail.com
Лунвэй Цянь

Information about the author

Longwei Qian, Postgraduate at the Department of Intelligent Information Technologies of the Belarusian State University of Informatics and Radioelectronics

Address for correspondence

220013, Republic of Belarus,
Minsk, P. Brovki St., 6
Belarusian State University
of Informatics and Radioelectronics
Tel.: +375 29 721-60-63
E-mail: qianlw1226@gmail.com
Longwei Qian