http://dx.doi.org/10.35596/1729-7648-2020-18-6-49-56

Оригинальная статья
Original paper

УДК 004.912

ОНТОЛОГИЧЕСКИЙ ПОДХОД К ОБРАБОТКЕ ТЕКСТОВ КИТАЙСКОГО ЯЗЫКА

ЦЯНЬ ЛУНВЭЙ

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

Поступила в редакцию 18 июня 2020

Аннотация. Для реализации естественно-языкового пользовательского интерфейса и интеллектуального ответа на вопросы на основе знаний в работе предлагается модель обработки китайского языка, основанная на знаниях. В статье рассматриваются существующие методы обработки естественного языка и различные базы знаний, связанные с обработкой естественного языка. На основе анализа данных методов был сделан вывод о том, что в обработке естественного языка база знаний является самой основной и важной частью. База знаний позволяет обеспечить обработку естественного языка, основываясь на изначально описанных знаниях, а также объяснить процесс обработки. На основании анализа различных методов построения баз знаний об английском и китайском языках был предложен онтологический подход к обработке китайского языка. В модели обработки китайского языка можно выделить два основных аспекта исследования: построение базы знаний о китайском языке и разработка решателя задач на основе онтологии. Предложенный подход направлен на разработку семантической модели знаний о китайском языке. Как один из этапов реализации подхода была построена онтология китайского языка, которую можно использовать в дальнейшем для обработки китайского языка. В данной работе рассмотрены первая версия указанной онтологии и принцип построения базы знаний о китайском языке. Для построения онтологии на данном этапе нет единых стандартов и системы оценки. Расширение и улучшение онтологии и оценка ее качества требуют дальнейших исследований.

Ключевые слова: технология OSTIS, онтология, обработка китайского языка, база знаний.

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Для цитирования. Цянь Лунвэй. Онтологический подход к обработке текстов китайского языка. Доклады БГУИР. 2020; 18(6): 49-56.

ONTOLOGICAL APPROACH TO CHINESE TEXT PROCESSING

QIAN LONGWEI

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

Submitted 18 June 2020

Abstract. To implement natural language user interface and an intelligent answer to questions, the knowledge-based semantic model for Chinese language processing is proposed. The article gives careful consideration to the existing methods and various knowledge bases for natural language processing. The analysis of these methods has led to the conclusion that in natural language processing, the knowledge base is the most fundamental and crucial part. The knowledge base makes it possible to ensure processing of a natural language based on initially described knowledge and to explain the processing operations. By virtue of the analysis of various methods for constructing knowledge bases about the English and Chinese languages, an ontological approach to the Chinese language processing was proposed. The Chinese language processing model has two main aspects: the design of knowledge base about the Chinese language and the development of ontology-based knowledge processing machine. The proposed approach is aimed at developing a semantic model of knowledge on the Chinese language. As a stage in the implementation of the approach, I designed the ontology of the Chinese language that can be applied for further processing of the language. This paper considers the preliminary version of the ontology and the principle of building a knowledge base about the Chinese language. There are no uniform standards and evaluation system for designing an ontology. Expansion, refinement and evaluation of the ontology require further research.

Keywords: OSTIS technology, ontology, Chinese language processing, knowledge base.

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

For citation. Qian Longwei. Ontological approach to Chinese text processing. Doklady BGUIR. 2020; 18(6): 49-56.

Введение

Обработка естественного языка – один из основных компонентов интеллектуальных систем. Целью обработки естественного языка в таких системах является понимание машиной текстов естественного языка и возврат точной информации пользователям на основе результатов понимания. Из-за разнообразия и открытости естественного языка компьютерная обработка различных естественных языков имеет свои особенности и трудности.

В обработке китайского языка в основном существует две основные особенности [1]:

– в европейских языках, таких как английский, русский, слова в основном пишутся с пробелами. Однако в текстах китайского языка нет пробелов для разделения слов;

– одна и та же часть речи может служить несколькими синтаксическими компонентами без морфологических изменений, то есть в текстах китайского языка независимо от любого синтаксического компонента, обслуживаемого каждой частью речи, ее морфология не изменяется.

Среди современных методов обработки естественного языка выделяют следующие направления:

– методы построения системы логических рассуждений на основе правил;

– методы машинного обучения, основанные на математической статистике и теории информации.

Однако, будь то основанные на правилах или статистические методы, методы и стандарты оценки, используемые при обработке китайского языка, почти полностью заимствованы из методов обработки европейских языков, таких как английский. В процессе обработки не учитываются характерные черты текстов китайского языка [1].

Для реализации глубокого семантического понимания текстов необходима поддержка представления различных сложных видов знаний. Например, лексических баз знаний, которые обеспечивают лексический анализ в текстах естественного языка. Основной целью баз знаний является описание различных сущностей и отношений, существующих в реальном мире. Метод обработки естественного языка на основе базы знаний заключается в следующем: при помощи разработанной базы знаний описываются различные понятия и отношения, существующие в текстах естественного языка. Далее, при помощи решателя задач реализуются глубокие рассуждения и семантическое понимание текстов естественного языка. Для обработки естественного языка основным видом знаний является лингвистическое знание со своими собственными языковыми характеристиками. Кроме лингвистических знаний, важную роль в глубоком семантическом понимании текстов естественного языка также играют «повседневные знания» (англ. commonsense knowledge), или знания о предметной области.

В данной работе рассмотрен онтологический подход к обработке текстов китайского языка, то есть подход, основанный на онтологии, лежащей в основе базы знаний и решателя задач, которые соответствуют лингвистическим характеристикам китайского языка. Онтологический подход предполагает обработку текста с учетом базы знаний, описывающей лингвистические характеристики китайского языка вместо обработки отдельных независимых китайских иероглифов, что может эффективно решить проблему отсутствия морфологических изменений в текстах китайского языка.

Базы знаний для обработки китайского языка

Для решения проблем обработки китайского языка Лю Чиюань из Университета Цинхуа предложил добавить знания в модель обработки китайского языка, основанную на данных, и исследовал такую модель обработки [2]. Кроме баз знаний в области обработки естественного языка, таких как WordNet, VerbNet, ConceptNet, существуют также известные базы знаний для обработки именно китайского языка, такие как GKB [3], Chinese ontology FrameNet [4], HowNet [5] и т. д.

База знаний, построенная для конкретного языка, может содержать характеристики конкретного языка, и обработка текста на основе такой базы знаний позволит получить результаты, существенно лучшие, чем в системах, не использующих базы знаний. Интеллектуальные системы должны эффективно организовывать и управлять различными знаниями и эффективно их обрабатывать. Онтология – это концептуальная модель, которая описывает понятия и их отношения, а также описывает семантику понятий через отношения между ними. Эта статья базируется на базах знаний о китайском языке, упомянутых выше, и предлагает онтологию для обработки китайского языка на основе открытой семантической технологии проектирования интеллектуальных систем (технология OSTIS). Технология OSTIS предназначена для представления и обработки различных видов знаний и ориентирована на разработку компьютерных систем, основанных на знаниях [6].

В рамках технологии OSTIS в качестве основы представления знаний используется SC-код. Это внутренний язык для кодирования знаний в памяти, который обеспечивает унифицированное семантически совместимое представление различных предметных областей и соответствующих им онтологий. SC-код позволяет представлять любые виды знаний, позволяет избежать избыточности и дублирования знаний.

В состав модели предметной области (ПО), представленной в SC-коде, входят постоянно существующие объекты исследования, постоянно существующие связи и структуры. Онтология же трактуется как спецификация предметной области, которая определяет понятия предметной области, отношения между ними и описывает взаимосвязи предметной области с другими сущностями, в том числе с другими предметными областями.

База знаний в рамках технологии OSTIS строится на основе онтологического подхода. Сущность проектирования базы знаний заключается в построении иерархической системы предметных областей и соответствующих им онтологий. В технологии OSTIS формализацию знаний можно рассматривать как формализацию и спецификацию предметной области. Такая модель проектирования базы знаний может минимизировать зависимость компонентов базы

знаний между собой и минимизировать необходимость взаимодействия между разработчиками различных компонентов.

Основанные на онтологическом подходе технологии OSTIS знания китайского языка структурированы для обработки текстов китайского языка. Целью построения базы знаний на основе онтологии является обеспечение общего понимания лингвистических знаний китайского языка, определение общепризнанных понятий. Онтология также содержит основные теории и основные принципы китайского языка, а также методы и правила обработки знаний китайского языка. В технологии OSTIS методы и правила обработки знаний могут быть описаны как sc-агенты. SC-агент – это некоторый объект, который может выполнять действия в семантической памяти, в которой хранится база знаний [7]. Разные правила и методы для обработки китайского языка могут быть реализованы в виде sc-агентов.

Глубокая обработка китайского языка выполняется на основе базы знаний о лингвистике и существующей базы знаний конкретной предметной области. С помощью базы знаний о лингвистике решатель задач преобразует текст китайского языка в семантически эквивалентный фрагмент в базе знаний конкретной предметной области. В сочетании с лингвистическим знанием китайского языка и знанием конкретной предметной области, информация из текстов китайского языка интегрируется в существующую базу знаний конкретной предметной области.

При вводе текстов китайского языка в систему, как показано на рис. 1, процесс используется для глубокой обработки китайского языка.

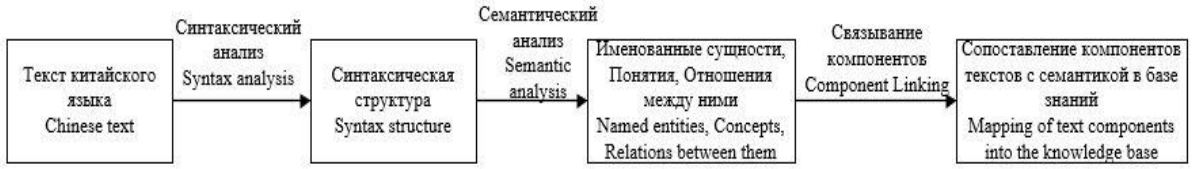

Рис. 1. Процесс для обработки китайского языка
Fig. 1. Operations for the Chinese language processing

Для реализации процесса обработки и понимания китайского языка можно выделить следующие SC-агенты:

– SC-агент анализа синтаксических структур. На основе синтаксических знаний о китайском языке в виде логических правил данный SC-агент анализирует синтаксические компоненты текстов китайского языка и выделяет такие компоненты, как подлежащее, сказуемое и др.;

– SC-агент анализа семантических структур. На основе семантических знаний о китайском языке данный SC-агент анализирует семантические компоненты текстов китайского языка. В результате обработки выделяются именованные сущности, понятия и отношения между ними. Если текст содержит некоторое действие, то агент также анализирует объект и субъект действия;

– SC-агент сопоставления компонентов текстов китайского языка с семантикой в базе знаний в конкретной предметной области. Задачей этого SC-агента является сопоставление полученных именованных сущностей, понятий и отношений из текстов китайского языка с узлами в базе знаний посредством вычисления семантического сходства.

Таким образом, путем сопоставления текста китайского языка с семантикой в базе знаний выполняется предварительное понимание текстов китайского языка.

Общая структура базы знаний для обработки китайского языка

Предложенный подход заключается в разработке предметных областей китайского языка и обработке специфических знаний, основанных на общей структуре предметных областей о лингвистике. Согласно принципу построения базы знаний, используемому в технологии OSTIS, база знаний о лингвистике китайского языка представляет собой иерархическую систему различных предметных областей лингвистических знаний китайского языка и соответствующих им онтологий.

Различные предметные области описывают различные аспекты соответствующего лингвистического знания китайского языка, а также правила и методы обработки знаний. Онтология предметной области о лингвистике китайского языка строится с помощью повторного использования существующих онтологий и лингвистических знаний, ориентированных на обработку китайского языка. Существующие онтологии имеют важное справочное значение для построения новой лингвистической онтологии. Ниже приведена общая структура предметной области китайского языка, представленной на языке SCn (одном из вариантов внешнего отображения текстов SC-кода) [8].

ПО текстов китайского языка
=> частная ПО:*
- *ПО синтаксиса китайского языка*
- *ПО семантики китайского языка*

Предметная область синтаксиса китайского языка описывает характеристики синтаксиса китайского языка, функциональные характеристики синтаксических компонентов. Предметная область семантики китайского языка описывает семантические характеристики, семантические отношения и семантическую структуру в текстах китайского языка. Для глубокой обработки китайского языка необходимо построить семантическую онтологию китайского языка. Предметная область синтаксиса, построенная на основе лингвистических знаний, отражает характеристики синтаксиса китайского языка и может использоваться для анализа синтаксической структуры текстов китайского языка. Предметная область семантики китайского языка используется для поверхностного семантического анализа текста китайского языка. Основное внимание в данном исследовании уделяется предметной области синтаксиса китайского языка, которая будет подробно описана ниже.

В предметной области синтаксиса китайского языка необходимо учитывать синтаксическую информацию о предложениях, словосочетаниях и словах китайского языка. Ниже приведен структурный фрагмент предметной области синтаксиса китайского языка, представленный на языке SCn.

ПО синтаксиса китайского языка
=> частная ПО:*
- *ПО предложений китайского языка*
- *ПО словосочетаний китайского языка*
- *ПО слов китайского языка*
- *ПО частей слов*
- *ПО сокращений*
- *ПО идиом*

Предложения всегда рассматривались в качестве самой маленькой единицы исследования в области обработки естественного языка. Анализ предложений является важной промежуточной стадией, соединяющей анализ всего текста и анализ отдельных слов. Подробное описание знаний о предложениях является важным базовым этапом для обработки естественного языка. Предметная область предложений китайского языка исследует различные типы предложений китайского языка, компоненты предложений и отношения между ними. Семантически полный текст можно разделить на ряд сложных и простых предложений. Простые предложения могут быть составлены из ряда словосочетаний, основанных на различных отношениях. Различные словосочетания, в свою очередь, также имеют свои точные синтаксические и семантические функции в предложении.

Ниже приведена структурная спецификация предметной области предложений китайского языка, представленная на языке SCn.

ПО предложений китайского языка

∋ *максимальный класс объектов исследования':*
предложение китайского языка
∋ *немаксимальный класс объектов исследования':*
• *простое предложение*
• *сложное предложение*
∋ *исследуемое отношение':*
• *подлежащее**
• *сказуемое**
• *дополнение**
• *определение**
• *обстоятельство**
• *числительное**
• *квантификатор**

простое предложение
=> включение:*
• *предложение с подлежащим и сказуемым*
• *предложение без подлежащего и сказуемого*
• *особенное предложение*

В предметной области предложений китайского языка описываются исследуемые абсолютные понятия (классы сущностей), исследуемые отношения и другие лингвистические знания о предложениях. На основе этих базовых лингвистических знаний далее могут описываться правила, задающие структуру предложений, и другие знания, в частности, некоторые специфические характеристики предложений китайского языка. Например, в китайском языке числительное и квантификатор строго различаются. Квантификаторы указывают единицу измерения. У китайского языка есть особенные слова для квантификаторов. В предметной области управление знаниями реализуется через иерархическое проектирование. Такой подход к проектированию может уменьшить область поиска путей решения задачи. При обработке конкретных текстов китайского языка необходимые знания для обработки текстов могут быть быстро найдены в рамках соответствующей предметной области.

Структура простого предложения с подлежащим и сказуемым может быть описана в логической онтологии. Например, на рис. 2 указано на языке SCg (одном из вариантов внешнего отображения текстов SC-кода), что предложение с подлежащим и сказуемым может состоять из *существительного словосочетания* и *глагольного словосочетания*, выполняющих соответственно роли *подлежащего* и *сказуемого*. Как показано на рис. 2, предметная область предложений китайского языка содержит описания разных понятий и отношений, таких как «*предложение с подлежащим и сказуемым*», «*подлежащее'*», «*сказуемое'*» и др. Данный фрагмент базы знаний описывает, что предложение с подлежащим и сказуемым можно рассматривать как последовательность существительного словосочетания и глагольного словосочетания с соответствующими атрибутами. В процессе генерации текстов предложение с подлежащим и сказуемым может генерироваться согласно заданной структуре.

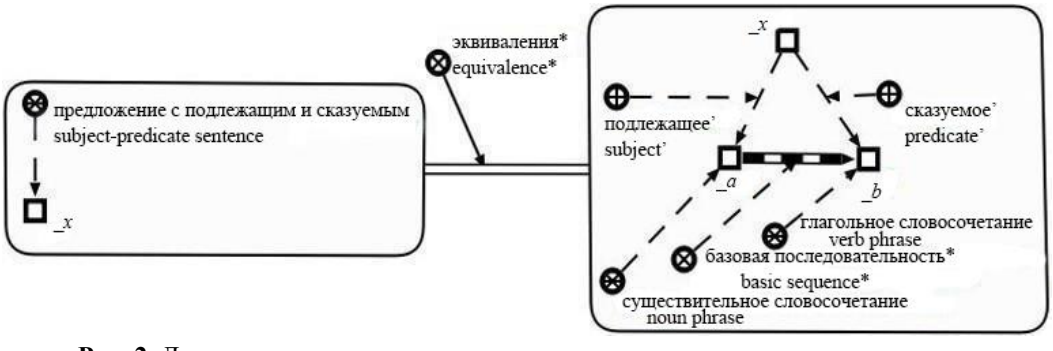

Рис. 2. Логическое утверждение про предложение с подлежащим и сказуемым
Fig. 2. Logical statement about a subject predicate sentence

Анализ словосочетаний может решить большинство проблем, связанных с неоднозначностью обработки отдельно взятых слов. В качестве промежуточного результата анализа предложений, анализ словосочетаний также является основой для более глубокого анализа фрагментов и полного синтаксического анализа. Предметная область словосочетаний исследует типы словосочетаний китайского языка и отношения между внутренними структурами словосочетаний.

Структуры различных основных словосочетаний могут быть описаны онтологией согласно синтаксическим отношениям. Например, на рис. 3 указано на языке SCg, что глагольное словосочетание может состоять из глагола и любого другого базового словосочетания или глагола и существительного в соответствии с отношением сказуемого-дополнения.

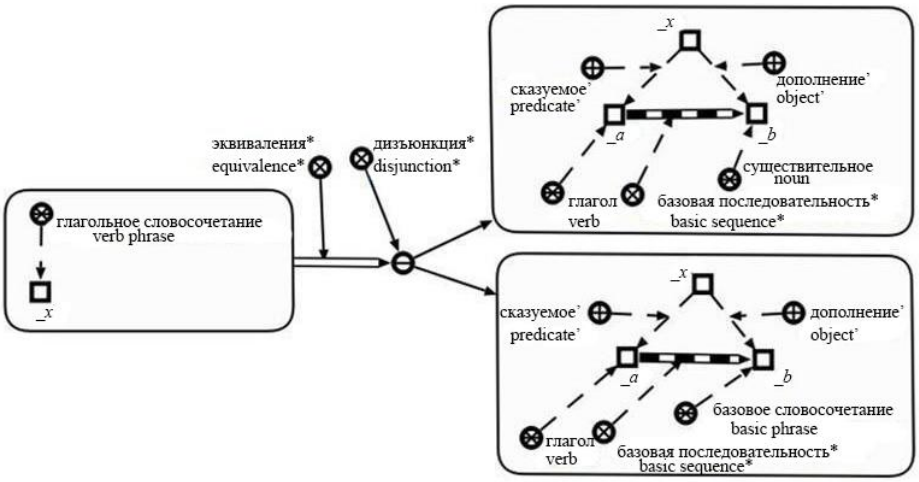

Рис. 3. Логическое утверждение про глагольное словосочетание
Fig. 3. Logical statement about a verb phrase

Из-за проблемы китайских правил письма, определение (выделение) слов в текстах китайского языка является важной теоретической проблемой. В области обработки китайского языка был предложен «Стандарт сегментации слов современного китайского языка, используемый для обработки информации» [9]. В этом стандарте слово представлено «единицей сегментации». Точное определение: «базовая единица для обработки китайского языка с определенными семантическими или грамматическими функциями». В предметной области слов китайского языка исследуются типы, синтаксические и семантические функции китайских слов. В данной предметной области описание слов ориентировано на компьютерную обработку китайского языка и не полностью совпадает с описанием слов в китайской лингвистике.

На основе построенной онтологии синтаксиса китайского языка может быть осуществлен синтаксический анализ текстов китайского языка и генерация текстов китайского языка, которые соответствуют онтологии синтаксиса китайского языка в базе знаний о лингвистике. База знаний делит лингвистические знания китайского языка на различные предметные области, что полезно для управления и применения синтаксических и семантических знаний китайского языка. База знаний о китайском языке может значительно повысить эффективность обработки китайского языка.

Заключение

Онтология используется для построения единой семантической модели лингвистических знаний китайского языка и эффективной организации предметных знаний для обработки текстов китайского языка. Предлагаемый подход выгоден для решения проблемы отсутствия интерпретации лингвистических знаний при обработке китайского языка. Предложенный в данной статье подход является предварительным результатом исследовательской работы по обработке китайского языка на основе онтологий. Эффективная интеграция между семантически эквивалентными фрагментами и базой знаний конкретной предметной области требует более глубокого исследования.

Список литературы / References

1. Zong C.Q., Cao Y.Q., Yu S.W. Sixty Years of Chinese Information Processing. *Applied Linguistics.* 2009;01 (04):53-61. DOI: 10.16499/j.cnki.1003-5397.2009.04.007.
2. Liu Z.Y. Knowledge guided natural language understanding. *Seventh China Conference on Data Mining*. 2018;01:199-206.
3. YU S.W. The Basic Processing of Contemporary Chinese Corpus at Peking University Specification. *Journal of Chinese information processing.* 2002;16(05):49-64.
4. Jia J.Z., Dong G. The Study on Integration of CFN and VerbNet, WordNet. *New Technology of Library and Information Service.* 2008;01(06):06-10.
5. Dong Z.D., Dong Qiang. Theoretical Findings of Hownet. *Journal of Chinese information processing.* 2007;21(04):03-09.
6. Golenkov V.V. Ontology-based Design of Intelligent Systems. *Open semantic technology for intelligent systems.* 2017;02:37-56.
7. Shunkevich D.V. Ontology-based design of knowledge processing machines. *Open semantic technology for intelligent systems.* 2017;02:73-94.
8. Davydenko I.T. Ontology-based knowledge base design. *Open semantic technology for intelligent systems.* 2017;02:57-72.
9. Jie C.Y. Some Key Issues upon Contemporary Chinese Language Word Segmentation Standard Used for Information Processing. *Journal of Chinese information processing.* 1989;03(04): 3-41.

Сведения об авторах

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

Information about the authors

Qian Longwei, PG Student of the Department of Intelligent Information Technologies of Belarusian State University of Informatics and Radioelectronics.

Адрес для корреспонденции

220037, Республика Беларусь,
г. Минск, ул. Платонова, 39,
Белорусский государственный университет
информатики и радиоэлектроники
тел. +375-29-721-60-63;
e-mail: qianlw1226@gmail.com
Цянь Лунвэй

Address for correspondence

220037, Republic of Belarus,
Minsk, Platonava str., 39,
Belarusian State University
of Informatics and Radioelectronics
tel. +375-29-721-60-63;
e-mail: qianlw1226@gmail.com
Qian Longwei