ИНТЕГРАЦИЯ ЛЕКСИЧЕСКИХ БАЗ ДАННЫХ ДЛЯ РЕШЕНИЯ ПРОБЛЕМЫ СИНОНИМИИ В ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ИНТЕРФЕЙСАХ

Гойло Артём Андреевич[1], Садовский Михаил Ефимович[2], Никифоров Сергей Александрович[2]

[1]Минский государственный лингвистический университет, преподаватель кафедры теории и практики перевода №1,

[2]Белорусский государственный университет информатики и радиоэлектроники, кафедра интеллектуальных информационных техологий

https://doi.org/10.5281/zenodo.7856117

Аннотация. *В статье описан подход к интеграции лексических баз данных для устранения проблем синонимии в естественно-языковых интерфейсах построенных с использованием Технологии OSTIS.*

Ключевые слова. *Обработка данных на естественном языке (Natural Language Processing, NLP), понимание естественного языка (Natural Language Understanding, NLU), онтология, семантическая сеть, Открытая семантическая технология для проектирвоания интеллектуальных систем (Open Semantic Technology for Intelligent Systems, OSTIS), естественно-языковой интерфейс, синонимия.*

ВВЕДЕНИЕ

Организация взаимодействия пользователей с компьютерными системами (в том числе и с интеллектуальными компьютерными системами) оказывает существенное влияние на эффективность автоматизации человеческой деятельности, пользовательский опыт и уровень удовлетворенности пользователей.

Пользовательский интерфейс — один из наиболее важных компонентов компьютерной системы и представляет собой совокупность аппаратных и программных средств, обеспечивающих обмен информацией между пользователем и компьютерной системой [1].

Естественно-языковой интерфейс — SILK-интерфейс (Speech, Image, Language, Knowledge, Речь, Образ, Язык, Знание), в котором обмен информацией между компьютерной системой и пользователем происходит за счёт диалога. Диалог ведётся на одном из естественных языков (ЕЯ).

Т.к. ключевой задачей естественно-языкового интерфейса является обеспечение взаимодействия между пользователем и компьютерной системой, для поддержания диалога с пользователем системе необходимо обладать средствами для понимания текстов на ЕЯ.

Под пониманием сообщения мы подразумеваем переход к смысловому представлению данного сообщения в компьютерной системе. Смысловое представление является информационной конструкцией с явным (формальным) представлением того, какие сущности описывает исходная информационная конструкция ЕЯ и как эти сущности связаны между собой. Синтаксическая структура данного смыслового представления близка описываемой конфигурации связей между описываемыми сущностями. Примером такого представления может служить семантическая сеть [2].

Одной из ключевых проблем для реализции данного перехода является разрешение синонимии. Т.к. в формальной онтологии, используемой в базе знаний интеллектуальной

системы, не должны присутствовать синонимичные знаки, при погружении смысла естественно-языковых сообщений в базу знаний необходимо сопоставить синонимичные знаки ЕЯ с общим для них смысловым представлением в базе знаний.

Данная работа основывается на средствах описания синтаксиса и семантики естественного языка, описанных в [3]. Предложенный подход к обработке естественного языка основывается на его формальной модели в виде набора онтологий, сформированных с использованием универсальных средств представления знаний, что способствует интероперабельности как компонента по обработке естественного языка в целом с другими компонентами системы, так и между составляющими самого данного компонента.

Данные средства реализуются на основе Технологии OSTIS, которая нацелена на создание интероперабельных интеллектуальных систем нового поколения и может служить основой для интеграции разнородных знаний и данных, т.к. используемый данной технологией язык — SC-код — обладает достаточной экспрессивностью для описания знаний любого вида [2].

В настоящее время в естественно-языковых интерфейсах, спроектированных с применением данной технологии (описание естественно-языковых интерфейсов ostis-систем см. в [4]), проблема разрешения синонимии все еще остается актуальной. В данной работе мы предложим подход к обработке синонимов ЕЯ с использованием WordNet [5].

ПРЕДЛАГАЕМЫЙ ПОДХОД

В качестве решения проблемы синонимии предлагается использование средств оценки схожести лексем, предоставляемых WordNet.

WordNet представляет собой базу данных лексики английского языка, в которой лексемы с эквивалентной семантикой объединяются во множества, называемые *синсетами* (англ. *synset*). Между синсетами устанавливаются некоторые связи, определенное количество типов которых задается WordNet (например, кодируются гипер-гипонимические связи, связи часть-целое и т.п.). Для каждого синсета приводится дефиниция и пример использования конкретной лексемы из синсета в предложении английского языка.

Таким образом, WordNet предоставляет не только информацию о том, какие лексемы являются синонимами, но также позволяет построить граф знаний с использованием базовых отношений, заданных в этой базе данных.

Все это может быть использовано естественно-языковым интерфейсом системы, спроектированной на базе технологии OSTIS (*ostis-системы*) для разрешения синонимии в процессе понимания текстов ЕЯ.

Входом для модуля понимания ЕЯ является текст ЕЯ, который разбивается на последовательность токенов. Для сопоставления данных токенов с понятиями, находящимися в базе знаний системы, посредством WordNet необходимо выполнение следующих шагов:

1. На первом этапе осуществляется приведение полученного токена к начальной форме (лемматизация) для возможности обработки его с использованием WordNet.
2. Далее осуществляется поиск в WordNet полученного токена.
3. После этого для всех понятий, входящих в текущий тематический контекст диалога [4], осуществляется поиск множества лексем, смыслом которых они являются.

4. В случае, если в полученных на предыдущем этапе множествах лексем найдена лексема, эталон которой совпадает с токеном, полученным от WordNet — генерируем в данном множестве новую лексему, эталоном которой является лемматизированная форма обрабатываемого токена. Данный поиск осуществляется с целью снижения вероятности ошибочного сопоставвления лексемы с ее смыслом.

5. Если среди множеств лексем, являющихся смыслом принадлежащих контексту диалога понятий, отсутствуют множества, содержащие лексему, имеющую в качестве эталона какой-либо из токенов, полученных в результате поиска в WordNet; то в базе знаний создается новое понятие с соответствующим множеством лексем, обозначающих его. Данное множество лексем на данный момент будет содержать только 1 лексему, эталоном которой указывается лемматизированя форма обрабатываемого токена.

В результате выполнения данного алгоритма для всех токенов происходит их сопоставление с понятиями в базе знаний для формирования смыслового представления полученного сообщения с целью последующей его обработки.

На рисунке 1 представлен пример множества синонимичных лексем в базе знаний ostis-системы, связанных с обозначаемым данным множеством понятием.

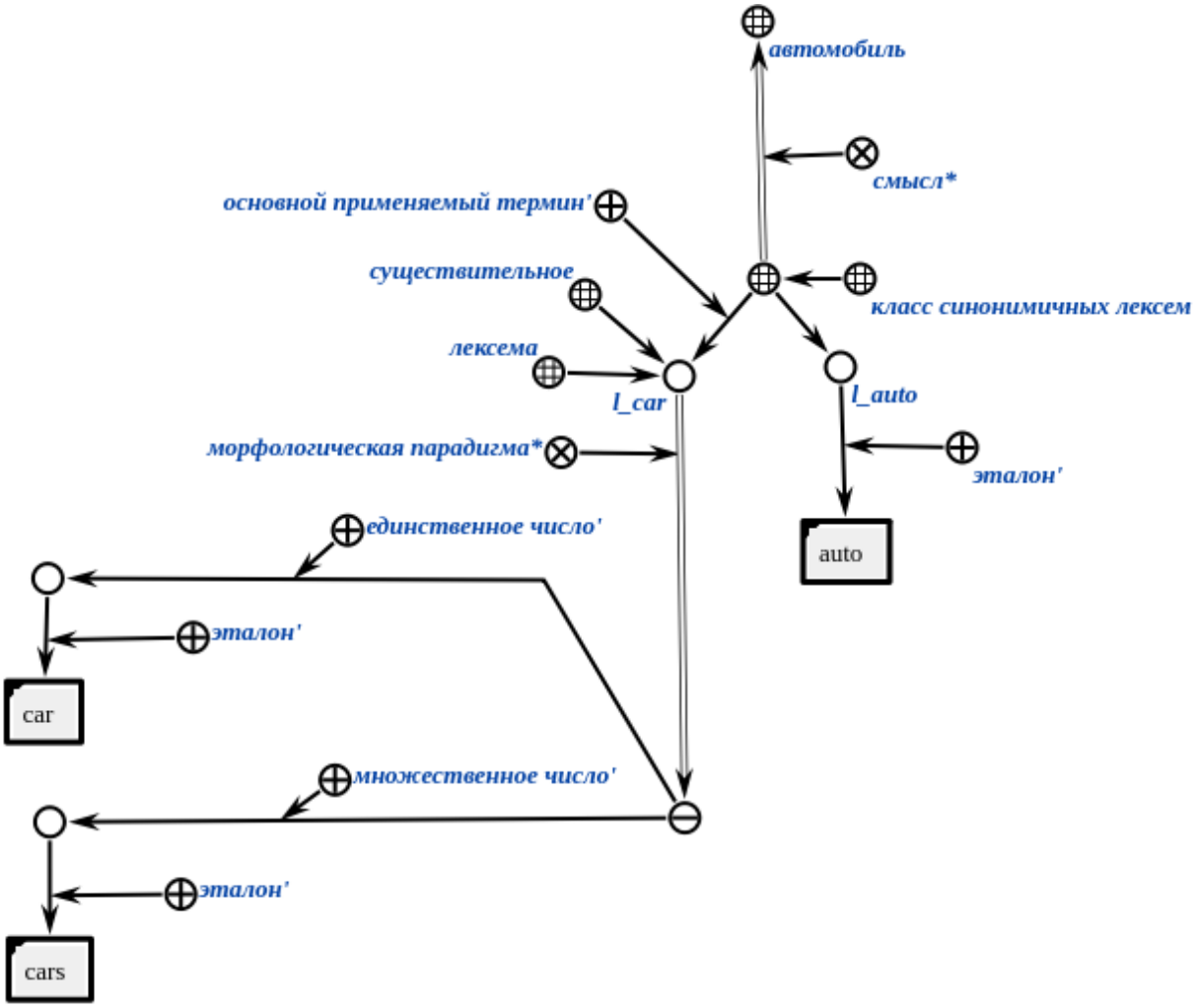

Рисунок 1. Пример множества синонимичных лексем (синсета) в базе знаний ostis-системы

Используемое в данном примере отношение *морфологическая парадигма* — бинарное ориентированное отношение, связывающее лексему и множество ее словоформ. При этом, данное отношения является подмножеством отношения *разбиение*.

Словоформа — подмножество лексемы, которому принадлежат все ее вхождения с определенными грамматическими значениями. Необходмо отметить, что в рамках применяемого подхода словоформа понимается несколько иначе, чем принято в лингвистике, так как все вхождения лексемы в технологии OSTIS являются файлами.

ЗАКЛЮЧЕНИЕ

Лингвистические базы данных, такие как WordNet, могут быть использованы как вспомогательное средство в естественно-языковых интерфейсах движимых онтологиями интеллектуальных систем для решения задач по пониманию текстов ЕЯ. В частности, WordNet особенно уместно использовать в качестве словаря (или части словаря) определенного ЕЯ, интегрированного в базу знаний интеллектуальной системы, с помощью которого можно потенциально решать проблему синонимичности и омонимичности знаков ЕЯ.

В данной статье был приведен подход к сопоставлению синонимичных знаков ЕЯ с общим для них смысловым представлением в базе знаний ostis-системы в рамках лексического анализа, проводимого естественно-языковым интерфейсом такой системы. Результатом анализа является информационная конструкция, описывающая множество синонимичных лексем, смыслом которой является некоторое понятие в базе знаний. Данный подход позволит избежать использования синонимичных знаков в базе знаний интеллектуальной системы в процессе понимания смысла исходного сообщения на ЕЯ.

REFERENCES

1. Sadouski, M. The structure of next-generation intelligent computer system interfaces = Структура интерфейсов интеллектуальных компьютерных систем нового поколения / M. Sadouski // Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2022) : сборник научных трудов / Белорусский государственный университет информатики и радиоэлектроники ; редкол.: В. В. Голенков [и др.]. – Минск, 2022. – Вып. 6. – С. 199–208.
2. Голенков, В. В. Открытая технология онтологического проектирования, производства и эксплуатации семантически совместимых гибридных интеллектуальных компьютерных систем / В. В. Голенков, Н. А. Гулякина, Д. В. Шункевич. – Минск : Бестпринт, 2021. – 690 с.
3. Goylo, A. Means of formal description of syntax and denotational semantics of various languages in next-generation intelligent computer systems = Средства формального описания синтаксиса и денотационной семантики различных языков в интеллектуальных компьютерных системах нового поколения / A. Goylo, S. Nikiforov // Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2022) : сборник научных трудов / Белорусский государственный университет информатики и радиоэлектроники ; редкол.: В. В. Голенков [и др.]. – Минск, 2022. – Вып. 6. – С. 99–118.

4. Goylo, A. Natural language interfaces of next-generation intelligent computer systems = Естественно-языковые интерфейсы интеллектуальных компьютерных систем нового поколения / A. Goylo, S. Nikiforov // Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2022) : сборник научных трудов / Белорусский государственный университет информатики и радиоэлектроники ; редкол.: В. В. Голенков [и др.]. – Минск, 2022. – Вып. 6. – С. 209–216.
5. Fellbaum, Christiane (2005). WordNet and wordnets. In: Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, 665-670.