RULE-BASED МЕТОД АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ В ТЕКСТОВЫХ ДОКУМЕНТАХ СЕМАНТИЧЕСКИХ ОТНОШЕНИЙ, ОРИЕНТИРОВАННЫХ НА ИСПОЛЬЗОВАНИЕ В ВОПРОСНО-ОТВЕТНОЙ СИСТЕМЕ

П. Ю. Довнар

ООО IHS Markit
Минск, Беларусь
e-mail: Polina.Dovnar@ihs.com, pdovnar72@gmail.com

Рассматривается опыт разработки лингвистических паттернов для выделения из текста на китайском языке информации о свойствах объектов. Предлагаемый подход апробирован в рамках лингвистического процессора экспертной системы Goldfire.

Ключевые слова: обработка естественного языка; вопросно-ответная система; китайский язык; rule-based метод.

RULE-BASED METHOD OF SEMANTIC RELATIONS RECOGNITION FOR QUESTION-ANSWERING SYSTEM

P. Y. Dounar

IHS Markit
Minsk, Belarus

This paper describes our experience of the development of linguistic patterns for Chinese question-answering system. The proposed approach is used in Goldfire expert system.

Keywords: natural language processing; question-answering system; Chinese language; rule-based approach.

Эффективность инновационной деятельности в большой степени определяется уровнем развития ее информационной поддержки, основной задачей которой является обеспечение изобретателя необходимыми знаниями на различных этапах его работы: во время фундаментальных и прикладных исследований, опытно-конструкторских работ, а также при промышленном внедрении изобретений.

Качественное решение указанной выше задачи может быть достигнуто за счет использования вопросно-ответных систем. Они представляют собой информационно-программные комплексы, которые способны обрабатывать пользовательские вопросы, представленные на естественном языке (ЕЯ), и давать на них краткие и точные ответы [1], что, собственно, и отличает их от систем традиционного информационного поиска. В качестве поисковой базы для таких систем обычно выступают большие коллекции текстовых документов, в том числе техническая документация той или иной компании, патенты, научные статьи и т. д.

В данной работе мы опираемся на опыт разработки вопросно-ответного модуля системы автоматизации инженерии знаний и решения инновационных задач Goldfire (http://goldfire.com).

Прежде всего нами была поставлена и решена задача формирования перечня основных типов знаний, наиболее актуальных для инновационной деятельности. С этой целью была проанализирована выборка из 60 000 уникальных англоязычных ЕЯ-запросов пользователей системы Goldfire. В результате анализа были выделены 50 основных классов запросов в зависимости от выраженной в них информационной потребности (так называемые Question Types, или QT). Эти классы включают такие типы информации, как, например, определение объекта (QT_Definition), значение параметров объекта (QT_Parameter), местоположение объекта (QT_Location), его применение (QT_Application), преимущества и недостатки (QT_Advantage/QT_Disadvantage), функции (QT_Function) и т. д. [2].

Полученная классификация используется при распознавании типов знаний, извлекаемых из текста, на этапе индексации текстовых документов с применением семантического анализа с целью их последующего соотнесения с классом запроса. Используемый семантический анализ текста в нашем случае подразумевает создание лингвистических правил (паттернов) для выделения того или иного типа информации. Мы ориентируемся на лингвистические методы обработки текста, которые в свою очередь опираются на знания, выявленные в процессе качественного экспертного анализа ЕЯ, и как раз формируются в виде указанных паттернов.

Здесь рассматривается опыт разработки множества лингвистических паттернов для выделения из текста на китайском языке информации о свойствах объектов (QT_Property).

Согласно определению, свойство – это «качество, признак, составляющий отличительную особенность чего-либо» [3]; «существенный признак, отличающий один предмет, одно явление от другого; особенная черта, качественная определенность чего-либо» [4].

При разработке лингвистических паттернов были выработаны следующие критерии распознавания QT_Property в дополнение к данному выше определению:

• Property обозначает *качественное* свойство объекта, а не приобретенную характеристику (например, слова «холодный», «теплый», «охлажденный» и т. п. в нашей концепции не обозначают свойство объекта в силу непостоянства признака, а слова «горючий», «хрупкий», «тугоплавкий» и др. – обозначают).

• Property имеет способность к *измерению* (например, *пористость* может быть измерена в процентах, также она может иметь степень выражения признака – «высокая пористость», а, например, *стерильность* не обладает такими характеристиками, поэтому прилагательное *«пористый»*, кит. 多孔 в нашей концепции обозначает свойство объекта, а *«стерильный»*, кит. 无菌 – нет).

• если свойство выражено прилагательным, то оно должно иметь возможность употребляться с *наречиями степени* («высоко», «очень», «сильно» и т. д.):

按照本发明方法形成的膜都是高度多孔的.

Мембрана, сформированная с помощью данного метода, является очень пористой.

Выделение QT_Property в отдельный класс знаний позволяет экспертам, работающим с системой Goldfire, а это инженеры, технологи и т. п., получать необходимую информацию, например о свойствах материалов и их комбинаций, прогнозиро-

вать их совместимость и поведение в различных ситуациях, предотвращать возможные неисправности и т. д.

Придерживаясь описанных выше критериев, мы проанализировали большое количество различных по содержанию и стилевой направленности текстов: энциклопедических статей, патентов, инструкций. На основании проведенного исследования было составлено более 250 тесткейсов (лингвистических шаблонов) с описанием основных способов выражения в текстах свойств различных объектов. На базе данных шаблонов позже был составлен блок из 60 паттернов.

Задача лингвиста-разработчика заключалась в нахождении и изучении конкретных случаев выражения в текстах информации о свойствах объектов, максимальном обобщении этих случаев и составлении паттернов. Мощности используемого лингвистического процессора, который включает графематический (разбиение текста на слова), лексико-грамматический (выделение лексико-грамматических классов слов), синтаксический (выделение грамматической структуры) виды анализа, было достаточно, чтобы осуществить это обобщение вплоть до семантического уровня (выделения смысловых отношений между словами).

Лингвистическое правило, или паттерн, представляет собой шаблон, задающий определенную языковую закономерность, т. е. спецификацию свойства набора примеров, определенную в терминах некоторого формального языка. В нашем случае это специальный язык расширенных регулярных выражений WRE (word-based regular expressions). WRE предоставляет возможность обобщения разрабатываемых правил путем оперирования не только символами алфавита, морфемами, отдельными словами и их совокупностями, но и лексико-грамматическими, синтаксическими и семантическими классами лексических единиц [6]. Таким образом, при написании правил лингвисты-разработчики могут обращаться к семантическим словарям, созданным ранее или в процессе работы над данным QT, объединять определенные слова в группы по их лексическим или формальным признакам, описывать их последовательности в максимально обобщенном виде, тем самым обеспечивая высокие показатели точности и полноты паттернов.

Так, например, в группу слов со значением «свойства» мы в числе прочих выделили слова, которые начинаются с иероглифа 可, придающего слову значение «способный, способность к чему-либо» (это аналогично английским суффиксам – -(a)ble, -(a)bility, как в словах transmissibility – *излучательная способность*, extendible – *растяжимый*). В этой группе оказались слова 可分性 – *делимость*, 可溶性 – *растворимость*, 可焊性 – *пригодность к пайке*, 可燃性 – *горючесть* и др. Но при этом нам пришлось исключить слова, внешне подходящие под описание, но не соответствующие заданным лексическим критериям, например, 可见 – *видимый*, 可怜 – *жалкий*, 可爱 – *симпатичный*, 可笑 – *смешной* и т. д. Описание данной группы слов на языке WRE будет представлено в следующем виде:

define WORD_PROPERTY, '可.+' \- ("可见"|"可怜"|"可爱|"可笑").

Приведем примеры одного из правил QT_Property, формализованного в виде паттерна с помощью языка WRE (табл. 1).

Таблица 1

Формальное представление правила	Расшифровка правила
Condition Subject {NN|NP|FW} Action { Positive } A.MainWord{VERB_LINK_PROPERTY} Adjective {@adj_property } **Result** Focus= = { Subject } AnswerAdjective = {Adjective} QType = { QTR_Property }	Если в предложении на этапе его семантического анализа распознаны как минимум такие компоненты, как Subject (подлежащее), Action (сказуемое), Adjective (атрибут-прилагательное), при этом первый является существительным или именем собственным, второй не содержит отрицаний и является глаголом-связкой из семантической группы VERB_LINK_PROPERTY (которая включает слова 是, 显, 为,成,呈 – *быть, являться*), а третий – слово из семантического словаря слов со значением Property, то последний будет представлять свойства первого. Фокус в данном случае указывает на то, что к любому конкретному значению *X* компонента Subject может быть впоследствии задан вопрос типа «каковы свойства *X*?»

Пример предложения, которое удовлетворяет условию приведенного паттерна.

两栖动物的皮肤是透水的. *Кожа земноводных является водопроницаемой.*

Графематический анализ (деление текста на слова – слова разделены пробелами):

两栖 动物 的 皮肤 是 透水 的 .

Лексико-грамматический анализ (распознавание частей речи):

两栖_NN 动物_NN 的_DE 皮肤_NN 是_VSHI 透水_VA 的_DE .

Семантико-синтаксический анализ (определение базовых семантических ролей):

r__Core_SAO

Subject	两栖_NN 动物_NN 的_DE 皮肤_NN	*Кожа земноводных*
Action	是_VSHI	*является*
Adjective	透水_VA 的_DE	*водопроницаемой*

Слово 透水 – *водопроницаемый* занесено в семантический словарь слов со значением свойства (в WRE обозначено как @adj_property). Как видим, данное предложение полностью соответствует условию представленного выше правила, поэтому в нем фиксируются знания типа Property для «кожи земноводных»: [она] «водопроницаемая».

Благодаря описанному выше инструментарию нам удалось извлекать информацию о свойствах объектов и из более сложных синтаксических конструкций. Приведем примеры некоторых из них (табл. 2).

Таблица 2

Предложение	Извлекаемая информация
金属是一种具有良好传热性质的物质. *Металл – это вещество, которое обладает хорошей теплопроводностью.*	Фокус: 金属 металл Свойство: 良好传热性质 хорошая теплопроводность
因此KClO3具有产生O2的化学性质. *KClO3 обладает химическим свойством производить кислород.*	Фокус: KClO3 Свойство: 产生O2 производит кислород
纳米粒子具有许多优异特性, 如高阻隔性. *Наночастицы имеют много выдающихся характеристик, например, высокие барьерные свойства.*	Фокус: 纳米粒子наночастицы Свойство: 高阻隔性 высокие барьерные свойства
固体酸是有机合成中能够代替硫酸的良好催化剂. *Твердая кислота – это хороший катализатор, который может заменить серную кислоту в органических соединениях.*	Фокус: 固体酸的性质твердая кислота Свойство: 良好催化剂 хороший катализатор

Тестирование точности и полноты паттернов QT_Property проходило в несколько этапов. Для первого этапа мы взяли 400 случайных результатов данного QT из сбалансированного корпуса, на втором этапе в качестве материала для тестирования были выбраны статьи из китайской Wikipedia (https://zh.wikipedia.org/), посвященные химическим элементам, различного рода веществам, программным продуктам, приборам и технологиям. Средний показатель точности извлечения данного типа информации составил 91 %, полноты (без учета расшифровки местоимений) – 50 %, что характеризует rule-based метод как достаточно эффективный способ решения поставленной задачи.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Monz C. From document retrieval to question answering. Amsterdam : Institute for Logic, Language and Computation, Universiteit van Amsterdam, 2003.
2. Постаногов Д. Ю. Классификация основных типов запросов пользователя в системе информационной поддержки инновационной деятельности // Международный конгресс по информатике: информационные системы и технологии : материалы междунар. науч. конгр. Минск : БГУ, 2011.
3. Ожегов С. И., Шведова Н. Ю. Толковый словарь русского языка. М. : Рус. яз., 1992.
4. Большой толковый словарь русского языка / гл. ред. С. А. Кузнецов. СПб. : Норинт, 2014.
5. Воронцов А. В. Лингвостатический метод автоматического лексико-грамматического анализа англоязычных текстов : дис. ... канд. филол. наук : 10.02.21. Минск : МГЛУ, 2008.
6. Чеусов А. В. Разработка алгоритмов и технологии построения многоязычного базового лингвистического процессора : автореф. дис. ... канд. техн. наук : 05.13.17. Минск : БГУ, 2013.
7. Постаногов Д. Ю. Разработка эффективных алгоритмов автоматического анализа текста на основе расширенных регулярных выражений // Информационные системы и технологии (IST'2002) : материалы I Междунар. конф. Минск : БГУ, 2002.