DOI 10.37882/2223–2966.2020.07.25

АВТОМАТИЗИРОВАННОЕ ИЗВЛЕЧЕНИЕ ЗНАНИЙ ИЗ МЕДИЦИНСКИХ ТЕКСТОВ

Погодин Руслан Сергеевич
Новосибирский национальный исследовательский государственный университет
ruspog@gmail.com

Аннотация. Статья посвящена разработке и применению теоретико-модельных методов для извлечения и формального представления знаний из текстов медицинских документов. Цель исследования — автоматизация наполнения базы знаний информацией из текстов, написанных на естественном языке. В качестве гипотезы рассматривается возможность применения теории И. А. Мельчука «Смысл — текст». Для решения задачи извлечения знаний используются методы преобразования предложений естественного языка в бескванторные формулы логики предикатов в виде фрагментов атомарных диаграмм. Затем производится программная формализация формул в машиночитаемый формат JSON. В результате работы была реализована программная система, взаимодействующая с базой знаний облачной платформы IACPaaS. Для тестирования использовались тексты медицинских статей с описанием болезней, их лечения и диагностики.

Ключевые слова: инженерия знаний, теоретико-модельные методы, извлечение знаний, экспертная система, база знаний, атомарная диаграмма.

AUTOMATED KNOWLEDGE RETRIEVAL FROM MEDICAL TEXTS

R. Pogodin

Summary. The article is devoted to the development and application of model theory methods for extracting and formal representing of knowledge from medical documents. The purpose of the research is to automate the filling of the knowledge base with information from texts written in a natural language. The hypothesis is the possibility of applying I. A. Melchuk's theory "Meaning — text". To extract knowledge, we used methods of converting natural language sentences into quantifier free formulas in the form of fragments of atomic diagrams. Then the formulas were programmatically formalized in machine-readable JSON format. As a result, we implemented a software system interacting with the knowledge base of the IACPaaS cloud platform. For testing we used texts of medical articles describing diseases, their treatment and diagnostics.

Keywords: knowledge retrieval, expert systems, knowledge base, atomic diagram.

Идея использовать экспертные системы для постановки диагноза и назначения лечения не нова: ещё в восьмидесятых годах прошлого столетия была разработана MYCIN — экспертная система для диагностики бактериальных инфекций [1]. Однако подобные системы не получили широкого распространения по причине необходимости ручного набора информации.

В данной статье рассматривается способ решения этой проблемы: автоматизированное извлечение информации из медицинских текстов и наполнение базы знаний этой информацией.

За основу представления семантики текстов медицинских статей была взята теория И. А. Мельчука «Смысл текст» [4]. Суть этой теории в рассмотрении глаголов в качестве многоместных предикатов, а других слов предложения в качестве аргументов этих предикатов. Например, предложение «Одним из основных симптомов парагриппа является ларингит» может быть представлено в виде четырёхместного предиката: «Является (Что: ларингит, Чем: симптомом, Каким: одним из основных, Чего: парагриппа)». При этом каждый аргумент предиката имеет свой тип, который определяется вопросом (что?, чем?, каким?, чего?).

На основе данной теории был разработан теоретико-модельный подход к решению задачи извлечения знаний [2]. Для формализации знаний вводятся понятия атомарного предложения и атомарной диаграммы.

Рассмотрим модель $\mathfrak{A} = <A, \sigma> = <A, P_1, \ldots, P_n, c_1, \ldots, c_k>$, где

A — основное множество модели;

σ — сигнатура, включающая в себя символы предикатов ($P_1, \ldots, P_n$) и констант ($c_1, \ldots, c_k$);

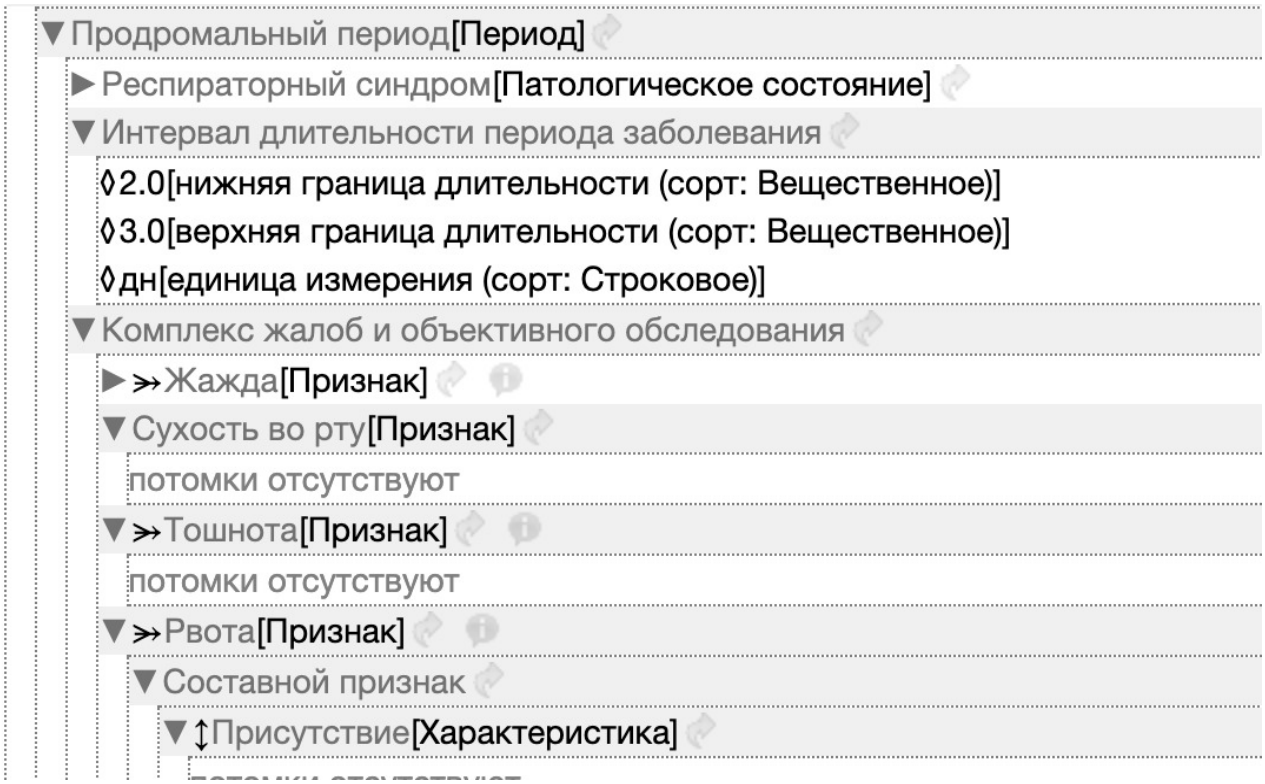

Рис. 1. Пример результата работы программы

S(σ) — множество предложений сигнатуры σ;

Предложение φ — атомарное, если $\varphi = (c_1 = c_2)$, либо $\varphi = (c_1 = c_2)$, либо $\varphi = P(c_1, \ldots, c_n)$, либо $\varphi = P_i(c_1, \ldots, c_n)$, где $P_i, c_1, \ldots, c_n \in \sigma$

$AD(\mathfrak{A}) = \{\varphi \in S(\sigma) \mid \mathfrak{A} \vDash \varphi$ и φ — атомарное$\}$ — атомарная диаграмма модели $\mathfrak{A}$. Любое подмножество $A \subseteq AD(\mathfrak{A})$ — фрагмент атомарной диаграммы.

Мы рассматриваем представление знаний в виде конечных фрагментов атомарных диаграмм, в которых предложения имеют вид $P(c_1, \ldots, c_n)$. Таким образом, согласно теории «Смысл — текст», весь текст мы формализуем в виде конечного набора предикатов с аргументами-константами.

Преобразование текста в машиночитаемый формат состоит из нескольких этапов:

1. Составление словарей.
2. Преобразование предложений во фрагменты атомарных диаграмм.
3. Преобразование многоместных предикатов в двуместные.
4. Разбор однородных членов предложений в двуместных предикатах.
5. Построение смыслового дерева.

На первом этапе составляются словари слов русского языка, глаголов русского языка, медицинских и анатомических терминов.

Медицинские термины необходимы для рассмотрения таких словосочетаний как «дыхательные пути», «головная боль», «половой цикл» в качестве единого целого, то есть, в качестве одной константы, которая будет передаваться аргументом в предикаты.

Отдельный словарь глаголов необходим для определения предикатных символов. Стоит отметить, что некоторые общеупотребимые глаголы и слова/словосочетания/символы объединяются в группу, представляемую определённым предикатным символом. Например, глагол «является» объединяется со словосочетанием «представляет собой», символом «–», а также пропущенным глаголом (глагол считается пропущенным, если его нет в грамматической основе) в предикатный символ "ЯВЛЯЕТСЯ", так как все они несут один и тот же смысл.

На втором этапе предложения текста преобразовываются во фрагменты атомарных диаграмм. Вначале производится первичный анализ предложений, в ходе которого анализируются слова и сопоставляются со словами из словарей. В результате этого анализа получается набор констант и предикатов, при этом каждый преди-

кат должен иметь дополнительный параметр — идентификатор, чтобы в дальнейшем можно было построить взаимосвязи описаний субъектов. Кроме того, в сигнатуру должны быть добавлены уникальные константы, которые будут использоваться как аргументы в качестве этих идентификаторов. Далее, используя введённые константы и предикаты, предложения преобразовываются во фрагменты атомарных диаграмм. Например, предложение «Острый холецистит представляет собой острое воспаление желчного пузыря» преобразовывается в предикат ЯВЛЯЕТСЯ (идентификатор субъекта, Что: Острый холецистит, Чем: воспаление, Чего: желчного пузыря, Каким: острое).

Стоит отметить, что однородные члены, объединённые запятыми, связками «и», «или», «от … до», «больше, чем», «такие как» и т.п., будут являться единой константой. Также в сигнатуру будут добавлены двуместные предикатные символы «ЯВЛЯЕТСЯ_ЧАСТЬЮ», «ОБЪЕДИНЁН_В_И», «ОБЪЕДИНЁН_В_ИЛИ», «ДИАПАЗОН_ОТ», «ДИАПАЗОН_ДО», «ЧАЩЕ», «РЕЖЕ». Это нужно, чтобы в дальнейшем можно было получить информацию об объединениях и диапазонах по определённым шаблонам. Более подробно это будет рассмотрено в описании четвёртого этапа.

В результате второго этапа мы получаем формулы в виде предикатов разной местности. Использовать в автоматической обработке знания, представленные в таком виде, не очень удобно, поэтому проводится третий этап — преобразование многоместных предикатов в двуместные [5]. Для этого используются введённые на предыдущем этапе идентификаторы: поскольку у предиката есть этот идентификатор, мы можем разбить этот предикат на несколько двуместных и ассоциировать каждый двуместный предикат с этой же константой, таким образом не потеряв взаимосвязь. Например, предикат ЯВЛЯЕТСЯ (идентификатор субъекта, Что: Острый холецистит, Чем: Воспаление, Чего: желчного пузыря, Каким: острое) будет преобразован в несколько двуместных предикатов:

- ЯВЛЯЕТСЯ (идентификатор субъекта, Острый холецистит),
- ЧЕМ (идентификатор субъекта, воспаление),
- ЧЕГО (идентификатор субъекта, желчного пузыря),
- КАКИМ (идентификатор субъекта, острое).

На четвёртом этапе производится разбор однородных членов. В сигнатуру модели добавляются дополнительные константы-идентификаторы однородности, которые в дальнейшем займут места констант с однородными членами предложения. В свою очередь константы разобьются на однородные члены и будут построены предикаты объединений и диапазонов, символы которых были добавлены на втором этапе. Первым аргументом таких предикатов будет идентификатор однородности, а вторым — сама константа, полученная при разбиении константы с однородными членами. Например, предикат ДЛИТСЯ (идентификатор субъекта, от 2 до 7 дней) преобразуется в несколько предикатов:

- ДЛИТСЯ (идентификатор субъекта, идентификатор диапазона),
- ДИАПАЗОН_ОТ (идентификатор диапазона, 2 дней),
- ДИАПАЗОН_ДО (идентификатор диапазона, 7 дней).

На финальном этапе строится смысловое дерево, в котором корень дерева — название болезни, ветви — предикаты, не концевые вершины — константы-идентификаторы, а концевые вершины — смысловые константы. Далее это дерево может транслироваться в любой формат данных по заданным правилам.

В рамках данной работы была разработана программная система, состоящая из следующих модулей:

- модуль наполнения базы слов;
- модуль преобразования предложений в многоместные предикаты LogicText [3];
- модуль преобразования многоместных предикатов в двуместные;
- модуль постобработки констант с однородными данными;
- модуль построения дерева знаний, модуль взаимодействия с системой IACPaaS;
- workflow-модуль, оперирующий всем процессом преобразования данных.

Модули написаны на языках Java и C# и взаимодействуют между собой по протоколу HTTP. В качестве базы данных для хранения промежуточных результатов используется PostgreSQL. Экспорт инфоресурсов в IACPaaS производится по протоколу HTTP в формате JSON. На этапе экспорта производится дополнительная обработка дерева знаний для соответствия со схемой данных, ожидаемых облачной платформой.

Таким образом, подход к извлечению знаний из медицинских текстов, основанный на теории «Смысл текст», показал свою состоятельность. Скорость автоматического преобразования текста в машиночитаемый формат намного превышает скорость выполнения данной работы человеком. Недостатком системы является несовпадение падежей, однако это не критическая проблема, так как смысл текста не теряется, и эксперт может внести грамматические корректировки. В дальнейшем систему можно развивать для обработки более сложных оборотов русского языка путём введения специальных предикатов, обрабатывающих такие паттерны.

ЛИТЕРАТУРА

1. Buchanan B.G., Shortliffe E. H. Rule Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project / B. G. Buchanan, E. H. Shortliffe // Addison-Wesley.— 1984.— P. 133–146.
2. Махасоева О.Г., Пальчунов Д. Е. Автоматизированные методы построения атомарной диаграммы модели по тексту естественного языка / О. Г. Махасоева, Д. Е. Пальчунов // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии.— 2014.— Т. 12, вып. 2. С. 64–73.
3. Махасоева О.Г., Пальчунов Д. Е. Программная система построения атомарной диаграммы модели по тексту естественного языка. / О. Г. Махасоева, Д. Е. Пальчунов // Св-во о гос. рег. программы для ЭВМ No 2014619198 от 10.09.2014.
4. Мельчук И. А. Опыт теории лингвистических моделей «Смысл ⇔Текст». / И. А. Мельчук // Школа «Языки русской культуры».— М.— 1999.
5. Ненашева Е.О., Пальчунов Д. Е. Разработка автоматизированных методов преобразования предложений естественного языка в бесквантерные формулы логики предикатов / Е. О. Ненашева, Д. Е. Пальчунов // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии.— 2017.— Т. 15, вып. 3. С. 55–60.

Журнал «Современная наука: актуальные проблемы теории и практики»

Г. Новосибирск