Федеральное государственное автономное
образовательное учреждение высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт математики и фундаментальной информатики
Кафедра высшей и прикладной математики

УТВЕРЖДАЮ

Заведующий кафедрой

_____ / С.Г. Мысливец

«____» _____ 2021 г.

БАКАЛАВРСКАЯ РАБОТА

Направление 01.03.02 Прикладная математика и информатика

Задача распознавания именованных сущностей для клинических текстов

Научный руководитель
кандидат физико-математических наук,
доцент _____ /Д.В. Семенова

Выпускник _____ /А.И. Собенин

Красноярск 2021

РЕФЕРАТ

Выпускная квалификационная работа по теме «Задача распознавания именованных сущностей для клинических текстов» содержит 43 страницы текста, 6 рисунков, 10 таблиц, 0 приложений и 25 использованных источников.

РАСПОЗНАВАНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ, ПРЕДОБРАБОТКА КЛИНИЧЕСКИХ ТЕКСТОВ, СКРЫТЫЕ МАРКОВСКИЕ МОДЕЛИ, АЛГОРИТМ ВИТЕРБИ, РАЗМЕТКА КЛИНИЧЕСКИХ ТЕКСТОВ.

Цель работы — разработка программы для распознавания именованных сущностей в клинических текстах.

В результате работы были изучены основы обработки естественного языка, изучены основные трудности обработки клинических текстов, рассмотрены существующие методы решения задачи. Разработана программа для формирования словаря сущностей в неразмеченных клинических текстах на основе системы правил. Разработана программа для полуавтоматической разметки клинических текстов с использованием регулярных выражений. Разработана программа по решению задачи распознавания именованных сущностей с использованием скрытых марковских моделей.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Система поддержки принятия врачебных решений (СППВР) - информационная система, предназначенная для помощи медицинским специалистам в работе с задачами, связанными с принятием клинических решений. Системы поддержки принятия врачебных решений связывают результаты клинических исследований с данными, имеющимися в отношении конкретного пациента, влияя на выбор врачебного решения для более эффективного оказания медицинской помощи. Разработка и внедрение СППВР в практику принадлежат к самым главным направлениям развития искусственного интеллекта в медицине.

Для создания системы поддержки принятия врачебных решений необходим размеченный корпус данных. Для того чтобы получить такой корпус из коллекции неразмеченных клинических текстов, нужно решить задачу распознавания именованных сущностей в клиническом тексте.

Цели и задачи работы. Целью работы является разработка программных средств распознавания именованных сущностей для клинических текстов (выписки из историй болезни пациентов с панкреатитом). Для достижения цели были поставлены и решены следующие задачи:

1. Исследовать особенности обработки клинических текстов.
2. Исследовать методы и алгоритмы решения задачи распознавания именованных сущностей (NER).
3. Разработать программу для формирования словаря сущностей в неразмеченных клинических текстах на основе системы правил.
4. Разработать программу для полуавтоматической разметки клинических текстов с использованием регулярных выражений.
5. Разработать программу по решению задачи распознавания именованных сущностей с использованием скрытых марковских моделей.

ЗАКЛЮЧЕНИЕ

Результаты проведенной работы представлены ниже.

1. Проведен обзор необходимой литературы, изучены основы обработки естественного языка и задачи распознавания именованных сущностей.
2. Изучены методы предобработки клинических текстов.
3. Изучены методы решения задачи распознавания именованных сущностей.
4. Изучен метод скрытых марковских моделей с применением алгоритма Витерби для определения наиболее вероятной последовательности тэгов для данной последовательности слов.
5. Разработана программа для формирования словаря сущностей в неразмеченных клинических текстах на основе системы правил.
6. Разработана программа на основе системы правил, составленных регулярными выражениями, для полуавтоматической разметки клинических текстов.
7. Проведены эксперименты по разметке текста с использованием разработанной программы на образцах клинических текстов коллекции «Панкреатит».
8. Разработана программа по решению задачи распознавания именованных сущностей с использованием скрытых марковских моделей.

Результаты работы докладывались на научных семинарах кафедры высшей и прикладной математики ИМиФИ СФУ.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие для студентов и аспирантов высших учебных заведений, работающих в области обработки текстов на естественном языке / Е.И. Большакова, Э.С. Клышинский, Д.В. Ландэ, А.А. Носков, О.В. Пескова, Е.В. Ягунова — Москва: Московский институт электроники и математики, 2011. — 272 С.
2. Антонова, А.Г. Использование метода условных случайных полей для обработки текстов на русском языке. Компьютерная лингвистика и интеллектуальные технологии / А.Г. Антонова, А.Н. Соловьев // «Диалог-2013»: сб. науч. статей / Вып. 12 (19).— Москва: Изд-во РГГУ, 2013.— С. 27—44.
3. Ахо, А. В. Структуры данных и алгоритмы. — Издательский дом Вильямс, 2000.
4. Ветров, Д.П. Лекция 3. Скрытые марковские модели. / Д.П. Ветров.// [Электронный ресурс]:Курс лекций «Графические модели» — Москва: МГУ, ВМиК, каф. ММП,—2012. — 55 С. Режим доступа: http://www.machinelearning.ru/wiki/images/8/83/gm12_3.pdf
5. Гефке, Д.А. Применение скрытых марковских моделей для распознавания звуковых последовательностей. / Д.А. Гефке, П.М. Зацепин // Известия Алтайского государственного университета. — 2012. — № 1(93). — 5 С.
6. Маслов, В. П. О законе Ципфа и ранговых распределениях в лингвистике и семиотике / В.П. Маслов, В.В. Маслова// Матем. заметки.—2006.—№5.— С. 718—732.
7. Морфологический анализатор для русского языка pymorphy2, написанный на языке Python. [Электронный ресурс]:сайт с исходным кодом. — Режим доступа: https://github.com/kmike/pymorphy2.
8. Морфологический анализатор MyStem [Электронный ресурс]: сайт содержит реализованный инструмент морфологического анализа. — Режим доступа: https://yandex.ru/dev/mystem.
9. Ненаусников, К.В. Алгоритм автоматического выделения коллокаций из текста. : учеб. пособие / К.В. Ненаусников, С.В. Кулешов — Санкт-

Петербург: Санкт-Петербургский институт информатики и автоматизации РАН, 2019. — 6 С.

10. Николенко, С.И. Глубокое обучение. Погружение в мир нейронных сетей. / С.И. Николенко, А. А. Кадурин – СПб. Питер, 2018. — 480 С.
11. Пакет библиотек и программ для символьной и статистической обработки естественного языка, написанных на языке программирования Python. [Электронный ресурс]. — Режим доступа: https://www.nltk.org .
12. Пентус, А. Е. Математическая теория формальных языков. / А. Е. Пентус, М. Р. Пентус — БИНОМ. Лаборатория знаний Интернет-Университет Информационных Технологий (ИНТУИТ), 2006.
13. Седунов, А. А. Модель графематического анализа в системе обработки естественного языка / А. А. Седунов // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. — 2007. — №. 2. — С. 69—77.
14. Семенова, Д.В. Теория автоматов, языков и вычислений / Д.В. Семенова, В.В. Быкова [Электронный ресурс]: учебное пособие [для студентов-математиков по магистерской программе 01.04.02.06 «Прикладная математика и информатика в гуманитарных и социально-экономических науках» напр. 01.04.02 «Прикладная математика и информатика»] / Сиб. федер. ун-т, Ин-т математики и фундамент. информатики. — Красноярск: СФУ, 2017. — 210 С. Режим доступа: http://Lib3.sfu-kras.ru/ft/LIB2/ELIB/b22/i-298479451.pdf.
15. Сокирко, А. В. Семантические словари в автоматической обработке текста (по материалам системы ДИАЛИНГ) / А. В. Сокирко //Москва: МГПИИЯ. — 2001.
16. Beakcheol, J. Bi-LSTM Model to Increase Accuracy in Text Classification: Combining Word2vec CNN and Attention Mechanism. / J. Beakcheol, K. Myeonghwi — // Department of Computer Science, Sangmyung University, Seoul, Korea, 2020. — P. 13.
17. Dalianis, H. Clinical Text Mining: Secondary Use of Electronic Patient Records / H. Dalianis. — // DSV-Stockholm University, Kista, Sweden, Springer

International Publishing, 2018. — P. 55.
18. Grishman, R. COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics / R. Grishman, B. Sundheim // Association for Computational Linguistics. Eight Street, Stroudsburg, PA, United States — 1996. — P. 466–471.
19. Hochreiter, S. Long Short-Term Memory / S. Hochreiter, J. Schmidhuber — : Neural Computation, 1997. — P. 46.
20. Jurafsky, D. Speech and Language Processing / D. Jurafsky, J. Martin — // Prentice Hall, 2008. — P. 532–568.
21. Lafferty, J. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. / J. Lafferty, A. McCallum — // Department of Computer and Information Science, 2001. — P. 10.
22. Laurene, V. Fundamentals of Neural Networks: Architectures, Algorithms And Applications / V. Laurene — // Pearson Education, 1994. — P. 561.
23. Pustejovsky, J. Natural Language Annotation for Machine Learning / D. Jurafsky, J. Martin — // O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, 2013. — P. 327.
24. Schmid, H. Probabilistic Part-of-Speech Tagging Using Decision Trees / H. Schmid. — // Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, 1994. — P. 9.
25. The spell checker Hunspell. [Электронный ресурс]. Режим доступа: http://hunspell.github.io.

Федеральное государственное автономное
образовательное учреждение высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт математики и фундаментальной информатики
Кафедра высшей и прикладной математики

УТВЕРЖДАЮ

Заведующий кафедрой

_____ / С.Г. Мысливец

«17» июня 2021 г.

БАКАЛАВРСКАЯ РАБОТА

Направление 01.03.02 Прикладная математика и информатика

Задача распознавания именованных сущностей для клинических текстов

Научный руководитель
кандидат физико-математических наук,
доцент _____ /Д.В. Семенова
17 июня 2021

Выпускник _____ /А.И. Собенин
17 июня 2021

Красноярск 2021