УДК 004.912

РАЗРАБОТКА МОДУЛЯ СТАНДАРТИЗАЦИИ ИМЕНОВАНИЙ ДЛЯ ОБРАБОТКИ МЕДИЦИНСКИХ ДАННЫХ

Муштак О.И., Лимановская О.В., Краглик И.Д.

ФГАОУ ВО УрФУ им. Первого президента России Б.Н. Ельцина, г. Екатеринбург.

Анализ ряда существующих систем для работы с медицинскими базами данных и распространенных методик стандартизации текстовой информации, показал, что нет готовых решений для работы с нестандартизированными медицинскими базами данных, имеющими множество сокращений. В работе предлагается алгоритм стандартизации медицинских терминов, который имеет сходство с алгоритмом хэширования, однако, является более универсальным и показывает высокую точность при работе с сокращениями, чем традиционные алгоритмы хэширования. Алгоритм опробован на содержащей множество сокращений базе данных по смертности больных туберкулезом.

Ключевые слова: обработка естественного языка, медицинская информатика, интеллектуальные системы.

Development of the module for the standardization of naming for the processing of medical data

Mushtak O.I., Limanovskaya O.V., Kraglik I.D.

Urals Federal University, the Institute of Fundamental Education, intelligent information technology department;Ekaterinburg

The analysis of number of present medicine system that works with patient databases and popular standardization methodologies of text information showed that there is not ready solution for working with non-standardization medicine databases that contain a large number of acronyms. The algorithm of medicine terms standardization is considered. It has some similarities with hash algorithm, however, it is more universal and has higher accuracy than traditional hash algorithm. This algorithm was checked by means of module which works with database of tuberculosis patient contained a large number of acronyms.

Keywords: natural language processing, medical informatics, intelligent systems.

Введение

В настоящее время благодаря внедрению информационных технологий в медицину, идет активное развитие медицинских баз данных и экспертных систем на их основе. Объем медицинских данных, хранящихся в электронном виде, стремительно растет, что вызывает

сложности в их обработке. Поэтому возникает необходимость в создании программного обеспечения для обработки больших объемов медицинских данных. В ряде работ [1-3] представлены программные средства для работы с медицинскими базами данных.

Так, в работе [1] авторы представляют систему автоматизированной обработки документов по лабораторных медицинским анализам. Система построена на анализе изображения документа, содержащего лабораторные данные. После анализа изображения данные извлекаются в xml файл и отправляются в базу знаний. Данная система работает только с различными жестко заданными вариантами одного слова из стандартизированного набора медицинских терминов, использующихся в лабораторных анализах, то есть отсутствует интеллектуальный алгоритм приведения термина к необходимой форме, что может вызвать затруднения при больших объемах баз знаний с высокой степенью вариативности написания терминов.

В работах [2, 3] парсеры разрабатывались изначально для работы с большим объемом текста и подразумевали использование корректных стандартизированных терминов во всем тексте, когда одно понятие обозначено одним термином и не имеет синонимов. Использование единых медицинских баз данных позволяет стандартизировать медицинскую информацию и исключить использование синонимов. Нужно отметить, что в работе [2] для обработки текста использовались готовые решения (например, Stanford parser), не предусматривающие поддержку русского языка, что не позволяет подходу быть в полной мере универсальным. В работе [3] рассматривается процесс классификации предложений, содержащих медицинскую информацию, на основе наличия в них отдельных выражений, относящихся к определенным семантическим классам из фиксированного набора данных.

Тем не менее, несмотря на развитие медицинских баз данных, в медицинской практике информация часто хранится либо в бумажном виде, либо в виде электронных таблиц, куда данные заносятся вручную, что неизбежно влечет за собой использование синонимов. Кроме того, в электронных таблицах, которые в российской действительности используются практикующими врачами намного чаще единых баз данных, имеются опечатки и не согласованные единым образом сокращения, что существенно затрудняет автоматизирование процессов обработки информации. Поэтому использование уже известных парсеров, применяемых для единых медицинских баз данных неприемлемо в повседневной медицинской практике.

Цель работы - разработка модуля стандартизации именований для обработки медицинских баз данных

Материалы и методы исследования

Материалы

В качестве исходных данных была использована база данных по смертности больных туберкулезом (Тюменская и Кировская области, ХМАО, ЯНАО, Удмуртия) за 2015-2016гг. объемом 430 записей. В базе была выявлена описанная выше проблема — отсутствие стандартизированности медицинских терминов. Её использование в «сыром виде» затрудняет как непосредственно взаимодействие с самой базой данных, так и построение интеллектуальных систем на её основе.

Методика

Для поставленной задачи были опробованы такие алгоритмы как метод шинглов, где каждая строчка разбивалась на зашифрованные CRC32 участки, а затем сравнивались их контрольные суммы, однако, точность данного метода для данной задачи оказалась неудовлетворительной. Стоит заметить, что подобные способы чаще используются для поисков нечетких дубликатов текстов в целом, а не для работы с небольшими строками, что может объяснять недостаточную точность.

Так как большую часть исходных записей в базе данных составляют не стандартизированные сокращения, одной из важных задач было снизить чувствительность к длине строк в процессе сравнения. В связи с этим был разработан собственный алгоритм, схожий с классическим алгоритмом хэширования, но имеющий дополнительные коэффициенты, повышающие точность при работе с сокращениями.

На входе подается словарь образцов написания терминов и массив сырых данных в виде таблицы формата CSV, где из нужных столбцов выделяется необработанные выражения.

Каждое выражение разбивается по условным (таким как пробел или точка) знакам на массив подслов. Так, например, выражение «дис.туберк.лег.» будет разбито на массив подслов [«дис», «туберк», «лег»]. Далее следует сопоставление каждого подслова из массива необработанных выражений со каждым подсловом из каждого словарного выражения и вычисление коэффициентов схожести. Коэффициенты схожести для пар терминов рассчитываются по формуле:

$$S = \sum_{i=0}^{n} w_i + L, \qquad (1)$$

где w — коэффициент схожести пар подслов, n — количество подслов в исходном выражении, L — коэффициент длин.

Коэффициент w вычисляется по формуле:

$$w = c * 1 + d \quad , \qquad (2)$$

где c — коэффициент вхождений, d — коэффициент идентичности.

Для расчета данных коэффициентов необходимо представить каждое подслово в виде одномерного массива. При итерации по подсловам с целью сопоставления, коэффициент с увеличивается на 1.1 в случае, если элемент первого массива (в данном случае — это буква) включен во второй массив (порядок следования неважен), коэффициент d увеличивается на 1.2 в случае, если соответствующие элементы совпадают и по порядку следования.

Коэффициент L рассчитывается по формуле:

$$L = \frac{l_s}{l_b} * 40 \quad , \qquad (3)$$

где из пары терминов l_s – длина термина, имеющего наименьшую длину, а l_b – длина термина, имеющего наибольшую длину.

Тот термин из словаря, который имеет максимальный коэффициент схожести с необработанным выражением, выбирается в качестве наиболее релевантного и предлагается пользователю для замены первоначального «сырого» выражения.

Отдельное правило введено для работы с аббревиатурами. Если коэффициенты схожести не достигают предельного минимального порога, выполняется проверка на схожесть первых букв. При положительном результате проверки выполняется замена на развернутый медицинский термин.

Результаты исследования и их обсуждение

Алгоритм был опробован на задаче по обработки базы данных по смертности больных туберкулезом. В качестве базы данных использовался документ формата Excel, содержащий данные по смертности больных туберкулезом легких. Фрагмент таблицы из базы приведен на рис.

ИТЛ	ИТЛ		ХОБЛ.Сахар.диабет	И
ИТЛ	ИТЛ	Кахексия.Кровохаркание.Туб.интоксикация		И
Инф.туб-з прав.лег. В фазе рубцевания	Вторичный ФКТ	Хр.панкреотит		Ф
Инф.туб-з лев.л-го	ФКТ левого легкогов форме распада, состояние после оперирования правого легкого	ДН 2ст	хр.бронхит	Д
ИТЛ в/д прав.легкого в фазе обсемен.	ФКТ обоих легких	ХЛС, отек легких	Глубокая паренхиматозная дистрофия внутрен.органов.Кахексия	Л
ИТЛ	ФКТ		ХОБЛ	в н н Л
1 ИТЛ об.легк. С распадом слева		ДН	ХОБЛ	п л

Рис. Пример сырых данных

В результате работы программного обеспечения в приведенном примере сырых данных была выполнена замена выражений «ИТЛ», «Инф.туб-з.лев.л-го» на заданный термин «Инфильтративный туберкулез легких».

Программный модуль разработан в двух вариантах: на языках программирования высокого уровня — Python и Kotlin, что делает программное обеспечение мультиплатформенным. Архитектура системы построена с учетом необходимости интеграции с внешней средой, исходя из чего, имеется возможность использовать модуль в качестве компонента интеллектуальных информационных систем и приложения к базам данных.

Заключение

Программный модуль, имеющий в основе данный алгоритм, позволяет привести медицинскую терминологию к общему виду, позволяя при этом задавать каноничные именования, а также контролировать процесс замены терминов, что сокращает вероятность появления недопустимых ошибок. Модуль имеет универсальный характер, применение не ограничивается отраслью медицины и может использоваться везде, где требуется повышенная точность при работе с опечатками и сокращениями, если применение

классических алгоритмов обработки текста оказывается недостаточно точным, например, по причине чувствительности к длине исходных строк.

Список литературы

1. F. Adamo An automatic document processing system for medical data extraction / F. Attivissimo, A. Di Nisio, M. Spadavecchia // Measurement. №61.2015.C.88-99

2. P. Mukherjee NegAIT: A new parser for medical text simplification using morphological, sentential and double negation / G. Leroy, D. Kauchak, S. Rajanarayanan, D. Y. Romero Diaz, N. P. Yuan, T. Gail Pritchard, S. Colina // Journal of Biomedical Informatics. №69.2017.C.55-62

3. H. Xu Applying semantic-based probabilistic context-free grammar to medical language processing – A preliminary study on parsing medication sentences / S. AbdelRahman, Y. Lu, J. C. Denny, S. Doan // Journal of Biomedical Informatics. №44.2011.C.1068-1075

Муштак Оксана Игоревна – студентка бакалавриата, департамент философии, Уральский Гуманитарный Институт, ФГАОУ ВО УрФУ им. Первого президента России Б.Н. Ельцина, 620002, Екатеринбург, ул. Мира, 19,тел +7 (343) 375-41-51, mushtak.oksana@yandex.ru