УДК 004.912

COMPARATIVE ANALYSIS OF ENGLISH AND RUSSIAN-SPEAKING MEDICAL FORUMS AS SOURCES OF INFORMATION ON ADVERSE DRUG REACTIONS

Романовская А.А. (Национальный исследовательский университет ИТМО)
Научный руководитель – к.техн.н., старший преподаватель (квалификационная категория «старший преподаватель») Добренко Н.В.
(Национальный исследовательский университет ИТМО)

This article examines the ways of mining and processing data from online medical forums to extract information on adverse drug reactions. Author proposes dictionary-based approach with corpora of Russian and English texts.

Введение. Data from online medical forums can be an important source of information on adverse drug reactions. However, a feature of such forums is the use of the conversational style of speech and slang, which makes it inefficient to use the vocabulary apparatus extracted from official documents. The purpose of the article was to extract information on adverse drug reactions from comments on online forums discussing medical topics and create a dictionary of terms and phrases indicating adverse drug reactions.

Основная часть. The study was conducted in relation to preparations with various trade names containing the common active substance Bisoprolol. On the basis of popular English-speaking and Russian-speaking medical forums relevant text corpora were created. Corpora processing was carried out in LancsBox and AntConc packages respectively. In order to recognize named entities, the phrase chunking technique was used - segmentation of sentences into collocations tied to a noun, for which a specialized parser was written. The accuracy of the parser in English was 81.6%, while in Russian it turned out to be even higher - 85.25%. Based on N-grams analysis (N = 2, 3, 5), dictionaries of English-language and Russian-language phrases characteristic of the forums and indicating adverse drugs reactions were compiled. In addition, dictionaries of synonymous expressions were compiled, which official sources and users of Internet forums use when describing adverse drug reactions.

Выводы. It is shown that the compiled dictionaries can be used to identify adverse reactions of other drugs, and not just Bisoprolol. All compiled dictionaries are publicly available. A comparison of the constructed dictionaries showed significant differences in the linguistic and semantic structure of messages on forums in two different languages, which are analyzed in detail.

Романовская А.А. (автор)
ariromanovskaya@gmail.com

Добренко Н.В. (научный руководитель)

УДК 004.912

СРАВНИТЕЛЬНЫЙ АНАЛИЗ АНГЛОЯЗЫЧНЫХ И РУССКОЯЗЫЧНЫХ МЕДИЦИНСКИХ ФОРУМОВ КАК ИСТОЧНИКОВ ИНФОРМАЦИИ О ПОБОЧНЫХ ЭФФЕКТАХ ЛЕКАРСТВ

Романовская А.А. (Национальный исследовательский университет ИТМО)
Научный руководитель – к.техн.н., старший преподаватель (квалификационная категория «старший преподаватель») Добренко Н.В.
(Национальный исследовательский университет ИТМО)

В работе рассматриваются способы сбора и обработки данных из медицинских онлайн-форумов с целью получения информации о побочных эффектах лекарств. Для создания базы побочных эффектов препаратов автор предлагает словарный подход, основанный на автоматизированной обработке текстовых данных из корпусов на русском и английском языках.

Введение. Данные из медицинских онлайн-форумов могут стать важным источником информации о побочных эффектах лекарств. Однако особенностью таких форумов является использование разговорного стиля речи и сленга, что делает неэффективным использование словарного аппарата, извлечённого из официальных медицинских документов. Целью статьи было извлечение информации о побочных реакциях на лекарства из комментариев на онлайн-форумах, где обсуждались медицинские темы, и создание словаря терминов и фраз, указывающих на побочные реакции на лекарства.

Основная часть. Исследование проводилось в отношении препаратов с различными торговыми наименованиями, содержащих общее действующее вещество бисопролол. На основе популярных англоязычных и русскоязычных медицинских форумов созданы соответствующие текстовые корпуса. Обработка корпусов проводилась в корпус-менеджерах LancsBox и AntConc соответственно. Для того, чтобы распознать именованные сущности, использовался метод разделения предложений на словосочетания, привязанных к существительному, для которого был написан специальный синтаксический парсер. Точность парсера на английском языке составила 81,6%, а на русском языке оказалась ещё выше – 85,25%. На основе анализа N-грамм (N = 2, 3, 5) были составлены словари англоязычных и русскоязычных фраз, характерных для форумов и указывающих на побочные реакции на лекарства. Кроме того, были составлены словари синонимичных выражений, которые официальные источники и пользователи интернет-форумов используют при описании побочных реакций на лекарства.

Выводы. Таким образом, показано, что составленные словари можно использовать для выявления побочных реакций других препаратов, а не только бисопролола. Сравнение построенных словарей показало существенные различия в лингвистической и семантической структуре сообщений на форумах на двух разных языках, которые подробно анализируются.

Романовская А.А. (автор)
ariromanovskaya@gmail.com

Добренко Н.В. (научный руководитель)