Е.Д. ИЗОТОВА[1], Д.С. ТАРАСОВ[2]

Казанский (Приволжский) федеральный университет[1]
Отдел исследований компании «Meanotek»[2]
izotova.e.d@gmail.com[1], dtarasov@meanotek.com[2]

ИЗВЛЕЧЕНИЕ ФАРМАЦЕВТИЧЕСКИ ЗНАЧИМЫХ АСПЕКТНЫХ ТЕРМИНОВ МОДЕЛЬЮ РЕКУРРЕНТНОЙ НЕЙРОННОЙ СЕТИ ИЗ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ ПРИ МАЛЫХ ВЫБОРКАХ

Рядом авторов отмечается важность анализа эффективности мониторинга отзывов потребителей о лекарственных средствах. В работе использовалась модель глубоких рекуррентных нейронных сетей. Продемонстрировано хорошее качество классификации для некоторых фармацевтически-значимых аспектных терминов, при небольшом размеченном корпусе (порядка 50 тыс. слов). Обсуждается стратегия улучшения качества результатов и потенциального приложения в области фармаконадзора.

Ключевые слова: аспектные термины, отзыв потребителей, лекарственные средства, рекуррентная нейронная сеть, естественный язык, русский язык

Оценка эффективности лекарственных средств (ЛС) является значимым для всех участников фармацевтического рынка. Большая часть информации об эффективности выпущенных препаратов на рынок поступают из постклинических исследований. Целью проводимых исследований является изучение показаний к применению ЛС, усовершенствование режимов назначения и схем лечения, а также длительное наблюдение с целью выявления влияния на плод и на уровень жизни пациента [1].

Для сбора данных об эффективности лекарственных препаратов на пострегистрационном периоде в РФ применяются: когортные исследования, активное мониторирование стационаров, метод спонтанных сообщений, рецептурный мониторинг, литературный мета-анализ, анализ единичных случаев, описанных в литературе, периодические отчеты–резюме по безопасности. Совокупное использование данных подходов не позволяет создать комплексую картину действия ЛС [2]. Так наиболее

эффективный подход — метод спонтанных сообщений. Однако вклад его от общего объема поступающих данных для стран с отлаженной системой фармаконадзора (Австралия, Новая Зеландия, Великобритания, Швеция, Канада) [3] — лишь от 2% до 10%. При этом получение данных сопряжено с рядом трудностей: недостаток времени, плохое знание системы и, особенно, трудности в установлении причинной связи между реакцией и приемом ЛС [4], отсутствие данных о частоте встречаемости побочной реакции [5], отмечается наличие личного предубеждения сообщающего [6], юридической ответственности в следствии врачебной ошибки [7].

Так же следует отметить, что по разным источникам частота развития нежелательных побочных реакций (НПР) варьирует от 2-3% [8], до 29% [2, 9] для различных ЛС, но только 4-6% [2] больных обращаются по этому поводу к врачу. При этом информация об используемых ЛС населением размещается на форумах, на страницах отзывов о препаратах, в блогах.

По ранее проведенным исследованиям отзывов потребителей о ЛС, размещенных в свободном доступе в сети Интернет, выяснилось хорошее согласование количества и диапазона НПР в текстах отзывов с данными, указанными в инструкции по медицинскому применению препарата [10]. Ручное извлечение информации является весьма трудоемким. При автоматическом извлечении информации широко используются методы построения деревьев - решений [11], позволяющие достаточно хорошо (F1-score = 82.4%) выявлять поставленный диагноз из медицинских карточек пациентов.

Извлечение структурированной информации из отзывов потребителей является сложной задачей для методов компьютерной лингвистики, т. к. отзывы написаны неформальным языком, не содержат структуры изложения, содержат грамматические ошибки, упоминаемые в них термины часто не соответствуют общепринятым упоминаниям ЛС, болезней, профиле врача, объектом приема ЛС и т. д., к примеру *«на моего МЧ эти таблетки не подействовали»*, *«средство вымывает всю бактерию»*.

В извлечении терминов из неструктурированных текстов на естественном языке хорошо зарекомендовали себя модели на основе глубоких рекуррентных нейронных сетей [12, 13].

В связи с выше изложенным, ставится задача разработки интеллектуального алгоритма извлечения фармакологически значимой

информации из данных отзывов потребителей о применяемых ЛС с использованием метода глубоких рекуррентных нейронных сетей.

Материалы и алгоритмы

Наборы данных

В качестве источников информации выступали отзывы потребителей ЛС, размещенные в свободном доступе. На основании разработанной формы-извещения о НПР сотрудниками ИДКЭЛ и инструкции по ее заполнению [14], а так же на основании структуры отзывов были сформированы аспекты, которые являются значимыми для оценки эффективности ЛС.

Для первичного анализа было выделено 14 аспектных терминов: «эффект», «нежелательная реакция», «объект приема», «прием» - самостоятельный прием или назначение врача, «профиль врача», «форма принимаемого ЛС», «порядок приема», «доза», «способ приема», «канцелярит» (в эту группу отнесены устойчивые фразы, рассуждения, выдержки из инструкций), «симптомы», «диагноз», «название ЛС», «форма ЛС».

Был составлен размеченный корпус отзывов, состоящий из 1296 единиц - отзывов о различных лекарственных препаратов (или порядка 50482 размеченных слов), российских и зарубежных производителей. Все отзывы коллекции уникальны.

Важно отметить что встречаемость аспектных терминов неоднородна и варьирует в широких пределах, от 120 («профиль врача») до 2225 («канцелярит») на 50482 размеченной выборки, рис. 1.

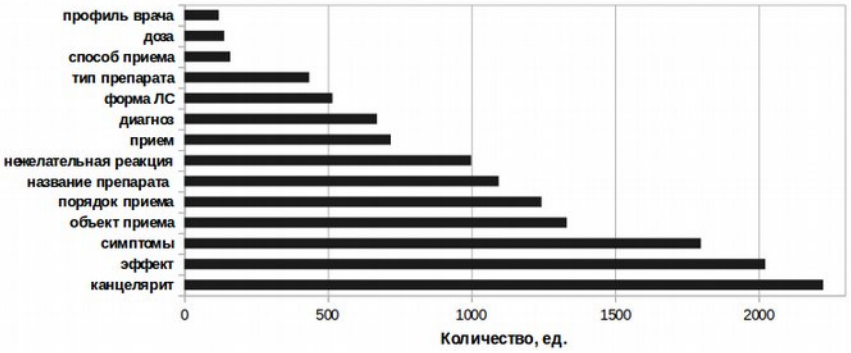

Рис. 1. Количественное распределение встречаемости аспектов

Архитектура использованной рекуррентной нейронной сети

Рекуррентная нейронная сеть [15] представляет собой тип нейронной сети, который имеет рекуррентные соединения (рис. 2а). Это делает такие нейронные сети применимыми для задач прогнозирования последовательностей, в том числе задач обработки естественного языка. В этой работе использовались рекуррентные слои типа сети Эльмана. В данной архитектуре, активации нейронов скрытого слоя на момент времени *t* вычисляются путем преобразования активаций текущего входного слоя *x(t)* и активаций скрытого слоя на предыдущем шаге *h(t-1)*.

Если *{x (m)}*, это последовательность векторов где *t = 1..T*, то активации нейронов в сети Элмана вычисляются как:

$$h(t) = f\left(W_x(t) + V_h(t-1) + b\right) \quad (1)$$

$$y(t) = g\left(U_h(t) + c\right) \quad (2)$$

где *f* это нелинейная функция активации, такая как сигмоид или гиперболический тангенс, а *g* – функция активации нейронов выходного слоя. *W* и *V* это матрицы весов между входным слоем и скрытым слоем, и между нейронами скрытого слоя. *U* – это матрица весов выходного слоя, *b* и *c* – вектора нейронов смещения, соединяющихся со скрытым слоем выходным слоем. *h(0)* в уравнении (1) может быть установлен в константное значение или обучен путем обратного распространения ошибки.

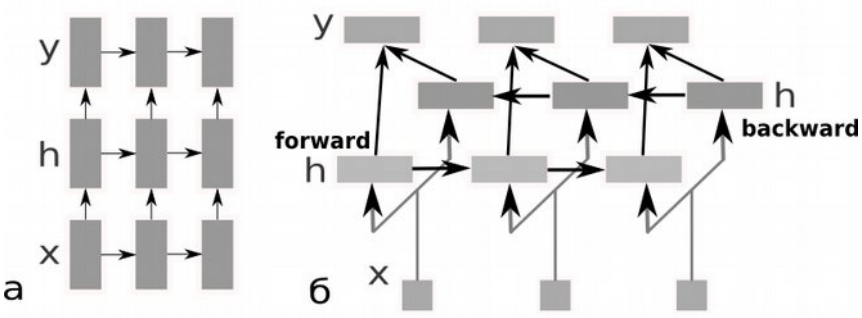

Рис. 2. а – простая рекуррентная нейронная сеть, б - двунаправленная рекуррентная нейронная сеть

Глубокие рекуррентные нейронные сети могут быть получены различными способами [16]. В данной работе использовалось последовательное соединение двух рекуррентных слоев, описанных выше.

Поскольку в задачах выделения терминов из последовательностей слов значение имеет информация не только о предыдущих, но и о последующих словах, мы использовали двунаправленные рекуррентные нейронные сети [17] (рис. 2б). В двунаправленных рекуррентных нейронных сетях состояния нейронов поделены на две части — часть отвечающую за положительное временное направление (прямые состояния) и часть отвечающую за отрицательное временное направление (обратные состояния):

$$h(t)^{forward} = f\left(W^{forward} x(t) + V^{forward} h^{forward}(t-1) + b^{forward}\right) \quad (3)$$

$$h(t)^{backward} = f\left(W^{backward} x(t) + V^{backward} h^{backward}(t+1) + b^{backward}\right) \quad (4)$$

$$y(t) = g\left(U^{forward} h^{forward} + U^{backward} h^{backward} + c\right) \quad (5)$$

Обучение нейронной сети

При каждом запуске случайным образом формировались две взаимно не пересекающихся выборки: тренировочная и тестовая, в соотношении 70% к 30%.

Нейронные сети обучались с помощью алгоритма обратного распространения ошибки через время [18] с использованием стохастического градиентного спуска с размером пакета в одно предложение, как предложено в работе [19]. Для предотвращения переобучения белый гауссовский шум был добавлен ко всем входным значениям нейронной сети.

После обучения для каждого изучаемого аспектного термина рассчитывалась полнота (R), точность (P), и пропорциональная F1 - мера.

Вектора слов были получены путем обучения рекуррентной языковой модели [20] на тексте рускоязычной части Википедии. Входные тексты были предварительно обработаны путем замены всех чисел на специальный токен #number, а все редко встречающиеся слова были заменены на специальный токен #unk.

Результаты

Анализ качества классификатора производился поэтапно с накоплением выборки, с целью оценить динамику обучения. Поэтому с увеличением выборки было произведено построение и расчет качества нейросетевой модели. Было произведено три запуска: первый с объемом размеченного корпуса в 17006 слов; второй - 33476 слова и третий - 50482 слов, табл. 1. По результатам классификационной модели аспекты можно разделить на три группы (по третьему запуску) :

(++) (0,59 до 0,86 по F1) — «объект приема», «профиль врача», «название препарата», «тип препарата», «форма ЛС», «диагноз», «порядок приема»;

(±) (от 0,26 до 0,45 по F1) — «прием», «симптомы»;

(– –) (от 0,00 до 0,18 по F1) — «эффект препарата», «нежелательная реакция », «доза» , «способ приема» , «канцелярит».

Таблица 1

Сводная таблица оценки качества нейросетевых моделей для 14 рассматриваемых аспектных терминов

Тип	№	Аспектов	P	R	F1	Тип	№	Аспектов	P	R	F1
эффект	1	771	0,23	0,03	0,05	форма ЛС	1	155	0,76	0,66	0,71
	2	1252	1,00	0,01	0,01		2	361	0,72	0,67	0,70
	3	2023	0,52	0,11	0,18		3	516	0,81	0,65	0,72
нежелате-льная реакция	1	303	0,00	0,00	0,00	симп-томы	1	678	0,43	0,16	0,23
	2	696	1,00	0,01	0,03		2	1121	0,66	0,20	0,31
	3	999	1,00	0,00	0,01		3	1799	0,57	0,16	0,26
объект	1	441	0,59	0,64	0,61	диагноз	1	249	0,68	0,26	0,38
	2	891	0,56	0,76	0,64		2	422	0,92	0,49	0,64
	3	1332	0,56	0,73	0,63		3	671	0,95	0,46	0,62
прием	1	195	0,47	0,13	0,2	порядок приема	1	510	0,56	0,46	0,51
	2	524	0,67	0,29	0,41		2	734	0,80	0,46	0,58
	3	719	0,65	0,34	0,45		3	1244	0,76	0,49	0,59
профиль врача	1	41	1,0	0,34	0,52	доза	1	49	0,00	0,00	0,00
	2	79	0,83	0,91	0,87		2	90	0,00	0,00	0,00
	3	120	0,85	0,80	0,83		3	139	1,00	0,04	0,08

название препарата	1	382	0,51	0,45	0,48	способ приема	1	91	0,00	0,00	0,00
	2	713	0,66	0,46	0,54		2	69	0,00	0,00	0,00
	3	1095	0,51	0,79	0,62		3	160	0,00	0,00	0,00
тип препарата	1	170	0,97	0,52	0,67	канце-лярит	1	1071	0,83	0,08	0,15
	2	265	0,93	0,82	0,87		2	1154	0,00	0,00	0,00
	3	435	0,90	0,79	0,86		3	2225	0,59	0,06	0,12

Примечание: Где № 1, 2, 3 — количество размеченных единиц текста — 17006, 33476 и 50482, соответственно.

Качество работы классификатора обратно пропорционально сложности аспекта. Так более определенные или лучше структурированные (что характерно для - «порядок приема» ЛС) термины дают лучший показатель по F1, группа (++). При этом количество встречаемых терминов в этой категории варьирует от 120 до 1332 единиц (на третий запуск). Чаще представители данной группы выступаю отдельные слова или словосочетания, к примеру: «тип препарата» - *«антибиотики»*, *«нестероидные - противовоспалительные»* или *«препараты для восстановления микрофлоры»*.

Для самой плохой группы (– –) по F1, количество встречаемых терминов варьирует от 139 до 2225 единиц (третий запуск). В этой группе собраны сложные описательные термины. В которых за аспектом стоит группа описательных слов, к примеру «эффект» - «*не чувствую что легчает*» или «нежелательная реакция» - «*снес мой организм напрочь*».

Для таких аспектов, как «способ приема», «доза», «нежелательная реакция» — нейронная модель не справляется, возможно в связи с недостаточным объемом данных и сложностью самой структуры.

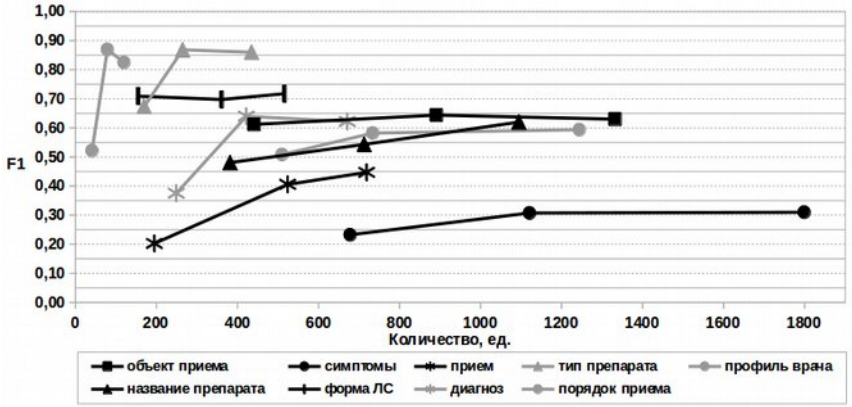

Рис. 3. Динамика качества классификатора с увеличением количества размеченной выборки

Для большинства рассматриваемых аспектов, качество классификации возрастает с увеличением объема размеченных данных, рис. 3. Однако для некоторых аспектов можно заметить отсутствие данной динамики при трех точках запуска. Так для «объект приема» и «форма ЛС» качество классификации остается неизменных и колеблется по F1 около 0,63 и 0,71, соответственно.

Заключение

В работе рассмотрен подход использования глубоких рекуррентных сетей с целью извлечения фармацевтически-значимых аспектных терминов из отзывов потребителей. Отзывы размещены в свободном доступе в русскоязычной сети Интернет. Построение модели произведено на небольшом размеченном корпусе (около 50 тыс. слов) отзывов. Для семи из четырнадцати изучаемых терминов: "объект приема", «профиль врача», «название препарата», «тип препарата», «форма ЛС», «диагноз», «порядок приема» - получено значение F1 от 0,59 до 0,86, причем количество терминов в каждой из этих категорий варьирует от 120 до 1332. Данный разброс говорит об отсутствии явной закономерности в качестве классификации с количеством присутствующих в корпусе терминов. В то же время построение модели классификации сложных,

описательных терминов дают низкие показатели по F1 («прием», «симптомы», «эффект препарата», «нежелательная реакция», «доза» , «способ приема» , «канцелярит»).

Для терминов «объект приема» и «форма ЛС», отсутствует корреляция увеличения качества классификации с увеличением объема обучаемой и тестовой выборок. Для них F1 - мера для всех трех измерениях существенно не изменяется и колеблется около 0,63 и 0,72, соответственно.

Можно предположить, что для некоторых терминов из группы плохо распознаваемых (это «прием», «канцелярит», «эффект»), объем размеченного корпуса недостаточен для построения качественной классификационной модели.

Таким образом увеличив объем корпуса можно улучшить качество классификации и расширить количество распознаваемых классов. В целом примененный подход позволит дополнить имеющиеся способы мониторинга качества ЛС после выхода их на потребительский рынок.

Список литературы

1. Лукьянова Е.А., Ляпунова Т.В., Ольшанская Е.В. Методы и практические навыки управления данными в клинических исследованиях // уч. пос. Москва. 2008. РУДН. 137с.
2. Кукес В.Г. Клиническая фармакология // ГЭОТАР-Медиа. 2009. 1052с.
3. Fletcher A. Spontaneous adverse drug reaction reporting vs. Event reporting vs. event monitoring: a comparison // J. Roy. Soc. Med . 1991. No. 84. P. 341-344.
4. Зырянов С.К. Организация и развитие службы фармаконадзора // Фарматека. 2005 . No. 16 http://www.pharmateca.ru/ru/archive/article/6317
5. Астахова А.В., Лепахин В.К., Романов Б.К. Фармаконадзор: состояние и перспективы развития//Новая аптека Эффективное управление. 2012.-No. 12. С.28-30.
6. Фитилев С.Б. Служба безопасности лекарств в российской Федерации // Новая аптека. 1998. С. 13-19.
7. Королев И.И., Хоробров М.А. Врачебная ошибка: правовые аспекты // XIX всероссийская конф. "Социально-гигиенический мониторинг здоровья населения". 2015. С. 290-299
8. Shear N., Shear N., Sullivan J., Wolverton S. Drug Actions, Interactions, Reactions. Program of the American Academy of Dermatolog // Academy.

2000. August 5.
9. Косенко В.В., Трапкова А.А., Тарасова С.А. Организация государственного контроля качества лекарственных средств на базе федеральных лабораторных комплексов// вестник росздравнадзора. 2012. № 6. С. 17-27
10. Leaman R., Wojtulewicz L, Sullivan R., Skariah A., Yang J., Gonzalez G. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks //BioNLP '10 Proceedings of the 2010 Workshop on Biomedical Natural Language Processing. 2010. pp. 117-125.
11. Shelmanov A.O., Smirnov I.V., Vishneva E.A. Tarasov D.S. Deep Recurrent Neural Networks for Multiple Language Aspect-based Sentiment Analysis of User Reviews // Dialogue. 2015. T.1. pp. 560-572.
12. Tarasov D.S. Deep Recurrent Neural Networks for Multiple Language Aspect-based Sentiment Analysis of User Reviews // Dialogue. 2015. T. 2. pp. 54-64.
13. Wenge R., Baolin P., Yuanxin O., Chao Li, Zhang X. Structural information aware deep semi supervised recurrent neural network for sentiment analysis// Frontiers of Computer Science. 2015. T. 9. V. 2. pp. 171-184.
14. Письмом Росздравнадзора № 01И-518/08 от 15.08.2008 http://minzdrav.tatarstan.ru/rus/file/pub/pub_24138.doc
15. Elman J. Finding structure in time // Cognitive science. 1990. V. 14. T. 2. pp. 179–211.
16. Pascanu, R., Gulcehre, C., Cho, K., Bengio, Y. How to construct deep recurrent neural networks // arXiv preprint arXiv:1312.6026. 2013.
17. Schuster M., Kuldip K. P. Bidirectional recurrent neural networks // IEEE Transactions on Signal Processing. 1997. V. 45. T. 11. pp. 2673–2681.
18. Werbos, P. J. Backpropagation through time: what it does and how to do it// Proceedings of the IEEE. 1990. V. 78. T. 10. pp. 1550-1560.
19. Mesnil G., He X., Deng L., Bengio Y. Investigation of recurrent neural network architectures and learning methods for spoken language understanding // In INTERSPEECH. 2013. pp. 3771-3775.
20. Mikolov T., Karafi´at M., Burget L., Cernock`y J., Khudanpur S. Recurrent neural network based language model // In INTERSPEECH. 2010. pp. 1045–1048.