УДК 004.912

ИЗВЛЕЧЕНИЕ СВЯЗАННЫХ С ЗАБОЛЕВАНИЯМИ ВЫРАЖЕНИЙ НА ОСНОВЕ МОДЕЛИ УСЛОВНЫХ СЛУЧАЙНЫХ ПОЛЕЙ[1]

З.Ш. Мифтахутдинов[1, a], **Е.В. Тутубалина**[1,b]
[1]Казанский Федеральный Университет
[a]zulfatmi@gmail.com, [b]tutubalinaev@gmail.com

Аннотация. В данной работе рассматриваются методы машинного обучения и глубокого обучения в задаче извлечения упоминаний о болезнях и симптомах из отзывов с форумов медицинской тематики. Для извлечения информации были использованы следующие методы – это линейные условные случайные поля (linear chain conditional random fields), рекуррентная нейронная сеть “долгая краткосрочная память” (long short-term memory), рекуррентная нейронная сеть “управляемые рекуррентные нейроны” (gated recurrent unit) и словарный метод. Также в данной работе были изучены различные признаки для линейных условных случайных полей: локальные, словарные, контекстные и распределенные представления слов. Эксперименты проводились на собранной и размеченной для этих целей коллекции данных CADEC. Было проведено сравнения указанных методов и признаков и показано, что линейные условные случайные поля превосходят другие подходы достигая при этом 69.1% и 79.4% F-меры по точному и частичному совпадению соответственно. Также были составлены словари и получены распределенные представления слов.

Ключевые слова*:* извлечение упоминаний о болезнях, условные случайные поля, извлечение информации, рекуррентные нейронные сети

EXTRACTING DISEASE RELATED EXPRESSIONS USING CONDITIONAL RANDOM FIELDS

Z.Sh. Miftahutdinov[1, a], **E.V. Tutubalina**[1,b]
[1]Kazan Federal University
[a]zulfatmi@gmail.com, [b]tutubalinaev@gmail.com

Abstract. In the recent work we considered a machine learning algorithms and deep neural networks in the task of extracting disease related expressions from social media reports. We have applied dictionary based approach, bidirectional Recurrent Neural Networks (RNNs) and Conditional Random Field (CRF) for extracting the disease-related entities and medical events. Specifically, we have employed Long Short Term Memory (LSTM) and Gated Recurring Units (GRU). We explored following features for CRF: local features, dictionary features, contextual features and word embeddings. Experiments were performed on a benchmark corpus. We established that CRF outperformed the RNN models as well as achieved the best exact and partial matching results (69.1% and 79.4% F_1-measure, respectively). As a part of work, we created dictionaries and obtained new word embeddings.

Keywords*:* disease named entity extraction, conditional random field, information extraction, recurrent neural networks

[1] *Работа выполнена при поддержке гранта РНФ (проект 15-11-10019).*

1. Введение

Автоматическое извлечение информации является важным шагом в построении базы знаний (Solovyev V. and Ivanov V. 2016). В последнее время в области анализа биомедицинских текстов наблюдается тенденция направленная на извлечение информации из форумов медицинской направленности и различного рода социальных медиа-ресурсов. Обусловлено это прежде всего тем, что, помимо доступности для проведения исследований, данные с подобного рода ресурсов часто содержит релевантную информацию. К примеру, в работе (A. Nikfarjam et al. 2015) были использованы посты из Twitter'а для построения системы отслеживания побочных эффектов лекарственных средств.

Данная работа посвящена первому этапу построения системы, позволяющей выдвигать гипотезы к перепрофилированию основываясь на отзывах о лекарственных средствах на медицинских форумах и в различных социальных сетях. Первый этап данной системы заключается в извлечении упоминаний о болезнях и симптомах, а также названий лекарственных средств из отзыва. Однако, извлечение названий лекарственных средств в данной работе не рассматривается, так как довольно хорошо решается простым поиском по словарю и не представляет никакого научного интереса.

В работе исследуется два различных метода к извлечению упоминаний о болезнях: линейные условные случайные поля (CRF) (J.Lafferty et al. 2001), рекуррентные нейронные сети (RNN) (Elman et al. 1990). Также исследуются различные признаковые наборы для CRF.

Статья состоит из следующих разделов: в разделе 2 обсуждается современное состояние исследований по задаче извлечения сущностей. Раздел 3 посвящен описанию используемых словарей и предложенным методам, основанных на моделях машинного обучения. В разделе 4 анализируются результаты экспериментов. В разделе 5 обсуждаются выводы, сделанные на основе экспериментов, и направления дальнейшей работы.

2. Текущее состояние исследований

Как уже было отмечено в последнее время проявляется огромный интерес к анализу биомедицинских текстов, в частности существует множество работ направленных на извлечение упоминаний о болезнях из различного рода источников, которые можно объединить в две группы: академические тексты и тексты из социальных медиа-ресурсов. В мировой науке подавляющее большинство работ посвящено английскому языку.

Большинство современных моделей, извлекающих упоминания о болезнях из текстов научной литературы используют линейные условные случайные поля (Lee Hsin-Chun et al. 2015, Dingcheng Li et al. 2015, Yanan Lu et al. 2015, Chih-Hsuan Wei et al. 2015, Qikang Wei et al. 2016) и скрытые марковские модели (Wong T. L. et al. 2011). Наиболее часто используются следующие признаки: само слово, часть речи слова, форма слова, синтаксические признаки и словарные признаки. Следует отметить, что академический текст имеет более формальный стиль и вполне определенную структуру в сравнении с текстами из социальных медиа-ресурсов. В связи с этим при извлечении упоминаний о болезнях из академических текстов достигаются хорошие показатели по F-мере (более 85%).

Довольно популярной прикладной задачей в сфере анализа биомедицинских текстов из социальных медиа-ресурсов является поиск побочных эффектов лекарств. Впервые данная проблема была рассмотрена в работе (Robert Leaman et al. 2010). Также данная задача была рассмотрена другими авторами в работах (Nikfarjam et al. 2015, Adrian Benton et al. 2011, Clark C Freifeld et al. 2014, Metke-Jimenez Alejandro et al. 2015, Miftahutdinov ZSh et al. 2017). Стоит заметить, что во всех работах авторы использовали

ограниченный набор базовых признаков при обучении модели CRF. Данная работа ставит целью провести более детальное исследование.

3. Предложенный метод

Задачу извлечения упоминаний о болезнях можно рассмотреть как задачу тегирования последовательности и решать используя IOB схему разметки, в которой первый токен искомой подстроки помечается как *Begin*, хвостовая часть – *In*, а не релевантная часть текста помечается как *Out*.

Распространенным методом решения подобных задач является использование модели условных случайных полей. Объясняется это тем, что CRF моделирует вероятностное распределение сразу над всей последовательностью и в связи с этим обычно имеет лучшее качество в сравнении с другими моделями. В качестве входного вектора для CRF были использованы следующие признаки:

1. w - лемма слова;
2. pos - часть речи слова;
3. sp - суффиксы и префиксы слова (2-6 последних и первых символов слова);
4. context - группа признаков (w, pos, dict) для 2 слов перед и после текущего слова;
5. wtype - два бинарных признака: написано ли слово в верхнем регистре; является ли слово отрицательной частицей;
6. dict – словарные признаки, т.е. если в тексте было найдено выражение содержащееся в одном из словарей, то все токены данного выражения помечаются с использованием IOB схемы;
7. b - кластерное представление слова, основанное на кластеризации Брауна;
8. emb - распределенное представление слова.

Еще одним методом получившим широкое распространение в последнее время является использование рекуррентных нейронных сетей, в частности рекуррентной нейронной сети “долгая краткосрочная память” (long short-term memory, далее LSTM) (Hochreiter Sepp and Schmidhuber Jürgen 1997) и рекуррентной нейронной сети “управляемые рекуррентные нейроны” (gated recurrent unit, далее GRU) (Kyunghyun Cho et al. 2014).

3.1. Словари индикативных конструкций

Было составлено пять словарей. В качестве источника терминов был взят мета-тезаурус Unified Medical Language System (Donald AB Lindberg et al. 1993), который объединяет в себя другие источники медицинской информации: словари, системы классификации, антологии.

Первый словарь был составлен из всех медицинских терминов из UMLS. Для этого из UMLS были извлечены термины следующих семантических типов:

- *Disease Or Syndrome;*
- *Neoplastic Process;*
- *Sign Or Symptom;*
- *Congenital Abnormality;*
- *Mental or Behavioral Dysfunction;*
- *Anatomical Abnormality*.

Получившийся список был очищен от терминов содержащих не англоязычные слова, стоп-слова, термины обозначающие части тела. Итоговый размер словаря составил 333,905 терминов.

Для составления второго словаря использовался словарь Consumer Health Vocabulary, который также является составной частью UMLS. Однако, в CHV используются названия болезней и симптомов используемых людьми не имеющими профильного медицинского образования. Словарь также был отфильтрован по частоте

встречаемости терминов в собранной коллекции данных. В конечный словарь попало 6,608 терминов.

Третий словарь, названный авторами ADR lexicon (Nikfarjam et al. 2015), был взят из работы, в которой решалась задача поиска побочных эффектов из отзывов.

Четвертый словарь составлен вручную из терминов начинающихся со слов *feel*, *able* или *ability*.

Пятый словарь представляет собой список всех лекарственных средств и действующих веществ. В качестве источника использовался drugbank (Vivian Law et al. 2013).

3.2. Распределенные представления слов

В данной работе для получения распределенного представления слов использовалась технология word2vec (Tomas Mikolov et al. 2013a, Tomas Mikolov et al. 2013b), реализованная в библиотеке gensim (Rehurek Radim et al. 2010). Была обучена модель CBOW со следующими параметрами: размер окна локального контекста равен 10, слова с частотой встречаемости выше или равной 10, размерность вектора равна 200, 5 примеров для негативного сэмплирования (negative sampling). Модель была обучена на собранной текстовой коллекции, источники и статистика по которой приведена в табл. 1. Итоговый размер словаря модели составил 93 526 токенов.

Таблица 1

Статистика по собранным данным

Источник данных	**Количество отзывов**	**Количество токенов**	**Количество уникальных токенов**
webmd.com	284 055	20 794 273	103 935
askapatient.com	113 836	13 649 150	79 036
patient.info	1 472 273	160 750 980	720 380
dailystrength.org	214 489	13 880 025	76 384
drugs.com	93 845	9 191 434	51 530
amazon health reviews	428 777	36 499 681	135 523
Итого	2 607 275	254 765 543	-

4. Результаты

Оценка качества работы описанных подходов была проведена на корпусе CADEC (Karimi S. et al. 2015), который состоит из 1250 отзывов с форума askapatient.com. В табл. 2 и 3 приведены результаты экспериментов проведенных на размеченных отзывах с различными наборами признаков для CRF и различными конфигурациями рекуррентных нейронных сетей. В качестве метрики считались F-мера по точному (Tjong Kim Sang Erik F et al. 2003) и частичному совпадению (Natalia Loukachevitch et al. 2015).

Первое, что можно отметить - это важность контекстных признаков для точного определения сущностей. Вторая важная особенность заключается в том, что совокупность словарей и векторных представлений слов оказывает более значимое влияние, чем эти признаки по отдельности. Двунаправленные рекуррентные нейронные сети показывают сравнимые результаты в частичном извлечении сущностей, однако, значительно уступают в точном совпадении.

Таблица 2

Частичное совпадение

Метод	**P**	**R**	**F1**
Dictionary-based method	**.885**	.442	.551
CRF, features: w	.793	.757	.745
CRF, features: w, sp	.803	.739	.740
CRF, features: w, sp, pos	.804	.739	.740
CRF, features: w, sp, pos, context	.806	.742	.747
CRF, features: w, sp, pos, context, b	.810	.763	.758
CRF, features: w, pos, context, b, dict	.834	.787	.768
CRF, features: w, sp, pos, context, PubMedVec	.810	.769	.764
CRF, features: w, sp, pos, context, HealthVec	.839	.734	.757
CRF, features: w, sp, dict, pos, context, b	.829	.774	.776
CRF, features: w, sp, dict, pos, context, b, PubMedVec	.830	.787	.784
CRF, features: w, sp, dict, pos, context, b, HealthVec	.864	.760	784
1-layer GRU, HealthVec, 40 epochs	.790	.739	.736
4-layer GRU, HealthVec, 40 epochs	.819	**.831**	**.802**
4-layer LSTM, HealthVec, 40 epochs	.825	.808	.793
4-layer LSTM, PubMedVec, 40 epochs	.780	.762	.739

Таблица 3

Точное совпадение

Метод	**P**	**R**	**F1**
Dictionary-based method	.580	.445	.489
CRF, features: w	.597	.588	.593
CRF, features: w, sp	.598	.577	.587
CRF, features: w, sp, pos	.604	.583	.593
CRF, features: w, sp, pos, context	.642	.608	.625
CRF, features: w, sp, pos, context, b	.635	.623	.629
CRF, features: w, pos, context, b, dict	.671	.667	.669
CRF, features: w, sp, pos, context, PubMedVec	.650	.630	.637
CRF, features: w, sp, pos, context, HealthVec	.661	.627	.643
CRF, features: w, sp, dict, pos, context, b	.666	.653	.659
CRF, features: w, sp, dict, pos, context, b, PubMedVec	.665	.658	.661
CRF, features: w, sp, dict, pos, context, b, HealthVec	**.705**	**.667**	**.685**
1-layer GRU, HealthVec, 40 epochs	.568	.486	.524

4-layer GRU, HealthVec, 40 epochs	.674	.609	.640
4-layer LSTM, HealthVec, 40 epochs	.659	.631	.644
4-layer LSTM, PubMedVec, 40 epochs	.586	.596	.591

Заключение

В данной работе была создана система для извлечения упоминаний о болезнях из текстов на английском языке. Были протестированы такие методы машинного обучения и глубокого обучения, как условные случайные поля, двунаправленные модели LSTM и GRU. Для них были подобраны различные признаки, включая локальные признаки слова, контекстные признаки, словарные признаки, а также распределенные представления слов, основанные на word2vec, и кластерные представления слов. Была собрана неразмеченная коллекция отзывов объемом 2.6 млн. для получения векторных представлений слов. Помимо векторов слов, были составлены словари медицинских терминов. Система была протестирована на собранной и размеченной для этих целей коллекции данных. Результаты, описанные в статье, подтверждают необходимость использования различных признаков, включая контекстные и словарные признаки, для точного извлечения сущностей в дальнейших работах. Все дополнительные материалы и код доступны на github.com/dartrevan/ChemTextMining.

Литература

Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor et al. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features // Journal of the American Medical Informatics Association. — 2015. — P. 671–681.

John Lafferty, Andrew McCallum, Fernando Pereira et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data //Proceedings of the eighteenth international conference on machine learning, ICML. — Vol. 1. — 2001. — Pp. 282–289.

Elman, J.L. Finding structure in time. Cognitive science 14 (2) (1990) 179-211

Lee Hsin-Chun, Hsu Yi-Yu, Kao Hung-Yu. An enhanced CRF-based system for disease name entity recognition and normalization on BioCreative V DNER Task // Proceedings of the Fifth BioCreative Challenge Evaluation Workshop. — 2015. — Pp. 226–233.

Dingcheng Li, Naveed Afzal, Majid Rastegar Mojarad et al. Resolution of chemical disease relations with diverse features and rules // The fifth BioCreative challenge evaluation workshop. — 2015. — Pp. 280–285.

Yanan Lu, Donghong Ji, Xiaoyuan Yao et al. CHEMDNER system with mixed conditional random fields and multi-scale word clustering // Journal of cheminformatics. — 2015. — Vol. 7, no. 1. — P. S4.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman et al. Overview of the BioCreative V chemical disease relation (CDR) task // Proceedings of the fifth BioCreative challenge evaluation workshop. — 2015. — Pp. 154–166.

Qikang Wei, Tao Chen, Ruifeng Xu et al. Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks // Database. — 2016. — Vol. 2016.

Wong T. L., Bing L., Lam W. (2011), Normalizing web product attributes and discovering domain ontology with minimal effort, Proceedings of the fourth ACM international conference on Web search and data mining, pp. 805-814.

Robert Leaman, Laura Wojtulewicz, Ryan Sullivan et al. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social

networks // Proceedings of the 2010 workshop on biomedical natural language processing / Association for Computational Linguistics. — 2010. — Pp. 117–125.
Adrian Benton, Lyle Ungar, Shawndra Hill et al. Identifying potential adverse effects using the web: A new approach to medical hypothesis generation //Journal of biomedical informatics. — 2011. — Vol. 44, no. 6. — Pp. 989–996.43
Clark C Freifeld, John S Brownstein, Christopher M Menone et al. Digital drug safety surveillance: monitoring pharmaceutical products in twitter //Drug safety. — 2014. — Vol. 37, no. 5. — Pp. 343–350.
Metke-Jimenez Alejandro, Karimi Sarvnaz. Concept extraction to identify adverse drug reactions in medical forums: A comparison of algorithms //arXiv preprint arXiv:1504.06936. — 2015.
Hochreiter Sepp, Schmidhuber Jürgen. Long short-term memory // Neural computation. — 1997. — Vol. 9, no. 8. — Pp. 1735–1780.
Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches // arXiv preprint arXiv:1409.1259. — 2014.
Donald AB Lindberg, Betsy L Humphreys, Alexa T McCray et al. The unified medical language system // IMIA Yearbook. — 1993. — Pp. 41–51.
Vivian Law, Craig Knox, Yannick Djoumbou et al. DrugBank 4.0: shedding new light on drug metabolism // Nucleic acids research. — 2013. — Vol. 42, no. D1. — Pp. D1091–D1097.
Tomas Mikolov, Ilya Sutskever, Kai Chen et al. Distributed representations of words and phrases and their compositionality // Advances in neural information processing systems. — 2013a. — Pp. 3111–3119.
Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. Efficient estimation of word representations in vector space // arXiv preprint arXiv:1301.3781. — 2013b.
*Rehurek Radim, Sojka Pet*r. Software framework for topic modelling with large corpora // In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks / Citeseer. — 2010.
Karimi, S., Metke-Jimenez, A., Kemp, M., & Wang, C. (2015). Cadec: A corpus of adverse drug event annotations. Journal of biomedical informatics, *55*, 73-81.
Tjong Kim Sang Erik F, De Meulder Fien. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition // Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4 / Association for Computational Linguistics. — 2003. — Pp. 142–147.
Natalia Loukachevitch, Pavel Blinov, Evgeny Kotelnikov et al. SentiRuEval: testing object-oriented sentiment analysis systems in Russian // Proceedings of International Conference Dialog. — Vol. 2. — 2015. — Pp. 12–24.
Miftahutdinov ZSh, Tutubalina EV, A Tropsha. Identifying Disease-Related Expressions in Reviews Using Conditional Random Fields // Computational Linguistics and Intellectual Technologies Papers from the Annual International Conference “Dialogue” (2017). — Vol. 1. — 2017. — Pp. 155–167.
Solovyev, V., & Ivanov, V. (2016). Knowledge-driven event extraction in Russian: corpus-based linguistic resources. Computational intelligence and neuroscience, 2016, 16.

Информация об авторах:

Мифтахутдинов Зульфат Шайхинурович, младший научный сотрудник, Казанский федеральный университет, г. Казань, РТ, zulfatmi@gmail.com
Тутубалина Елена Викторовна, старший научный сотрудник, Казанский федеральный университет, г. Казань, РТ, tutubalinaev@gmail.com