ИМЕНОВАННЫЕ СУЩНОСТИ В СФЕРЕ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ: РАЗМЕТКА И МЕТОДЫ ИЗВЛЕЧЕНИЯ[1]

Сиротина А.Ю. (overnastuhed@yandex.ru)
Лукашевич Н.В. (louk_nat@mail.ru)
Московский государственный университет им. М.В. Ломоносова (Россия, Москва)

Аннотация. В данной работе рассматривается проблема извлечения именованных сущностей (ИС) из неструктурированных русскоязычных текстов в сфере информационной безопасности (ИБ). Наша первая задача – создание корпуса размеченных текстов, посвященных проблемам ИБ. Был произведен анализ текстов данной предметной области и по его результатам представлена подробная инструкция для аннотаторов. Полученный корпус используется далее для обучения и тестирования систем извлечения ИС, основанных на нейронных сетях различных архитектур.

Ключевые слова. Информационная безопасность, извлечение именованных сущностей, нейронные сети, корпусная лингвистика

NAMED ENTITES IN CYBERSECURITY: ANNOTATION AND EXTRACTION

Sirotina A.Yu. (overnastuhed@yandex.ru)
Loukachevich N.V. (louk_nat@mail.ru)
Lomonosov Moscow State University (Moscow, Russia)

Abstract. This paper is devoted to named entity recognition task for Russian texts concerned with cybersecurity problems. Our first step is construction of a labeled corpus of texts from information security domain. For this purpose we conduct a thorough analysis of the texts in question and present a detailed instruction for human annotators. The labeled corpus is then used to test several methods of named entities extraction based on artificial neural networks.

Keywords. Cybersecurity, named entity extraction, corpus linguistics, neural networks

1.Введение

Основная задача специалистов в сфере информационной безопасности (ИБ) – оперативное предотвращение утечки и потери данных. Для ее решения специалистам необходимо максимально быстро получать актуальные сведения о хакерской активности, вирусах и уязвимостях. Для оптимизации процесса сбора таких сведений специалистами в сфере ИБ может быть использована **система автоматического извлечения информации** в данной предметной области. Для обучения и тестирования такой системы необходим размеченный корпус текстов, посвященных проблемам ИБ.

Источником наиболее актуальной информации об уязвимостях являются специализированные интернет-ресурсы (например, форумы и блоги). Публикации этих ресурсов представляют собой **неструктурированные** тексты. Следовательно, наиболее эффективной для сферы ИБ является система извлечения информации из неструктурированных текстов, и именно такие тексты должны составлять значительную долю от общего объема размеченного корпуса.

В данной работе рассматривается проблема извлечения именованных сущностей (ИС) из неструктурированных русскоязычных текстов в сфере ИБ. Одной из основных задач является разработка и тестирование систем извлечения ИС, основанных на нейронных сетях

[1] Работа частично поддержана РФФИ (проект 16-29-09606)

различных архитектур. Другая задача – создание размеченного корпуса релевантных текстов на базе коллекции Sec_col.

2.Обзор близких работ

Проблема извлечения информации в сфере ИБ не раз затрагивалась исследователями в области прикладной лингвистики ([1, 2, 5, 9, 12]). Однако в подавляющем большинстве работ данная проблема рассматривается только для **частично структурированных** текстов на **английском** языке. Так, обучающие корпуса в работах [1, 2, 12] содержат бюллетени по безопасности Майкрософт и статьи из Национальной базы данных уязвимостей США (NVD, National Vulnerability Database). Обучающий корпус в [5], помимо частично структурированных текстов, включает в себя и неструктурированные, но доля последних в корпусе составляет менее 10%.

При разметке корпусов в работах [1, 5, 12] авторы независимо друг от друга принимают схожие решения при формировании набора классов размечаемых ИС. Во всех трех работах разметке подлежат такие типы сущностей, как, например, название ПО, версия ПО, названия файлов, идентификатор/тип уязвимости. Как следствие, между наборами тегов, предложенными в данных работах, можно достаточно легко установить соответствие, то есть разметки легко переводятся одна в другую, практически без потери информации.

Системы извлечения сущностей, предложенные в работах [1, 2, 5, 12] основаны на различных методах: метод максимальной энтропии в [1], CRF в [5, 12], а также нейронная сеть с архитектурой LSTM-CRF в [2].

На данный момент нам известна единственная работа, посвященная проблеме извлечения информации в сфере ИБ на русском языке ([9]). Система извлечения ИС в [9] обучалась на предварительной версии коллекции текстов Sec_col, содержащей 2000 неструктурированных текстов по ИБ. значительная часть которых была размечена автоматически, переносом классификатора, обученного на новостных текстах [10].

3. Создание размеченного корпуса текстов в сфере ИБ

В рамках данной работы основой для создания корпуса послужила коллекция текстов Sec_col. Sec_col содержит 2000 публикаций и форумов сайта SecurityLab.ru[2], которые, в отличие от статей NVD и бюллетеней по безопасности, представляют собой неструктурированные тексты.

Следует отметить следующие особенности текстов, вошедших в коллекцию:

- Стиль текстов преимущественно неформальный, разговорный, реже – публицистический.
- Большое количество ошибок разных типов: орфографические, пунктуационные, синтаксические, лексико-стилистические, а также опечатки.
- Большое количество иноязычных слов как написанных латиницей (*Mac OS*), так и транслитерированных (*макос*).
- Большое число жаргонизмов, разговорных вариантов названий (*гмыло*).
- Большое количество слов, содержащих небуквенные символы.

[2] https://www.securitylab.ru/

Тексты размечались четырьмя независимыми разметчиками, при этом не все разметчики являлись специалистами в сфере ИБ. Разметка производилась при помощи онлайн-инструмента для аннотации BRAT[3].

Для разметки был сформирован следующий набор тегов: (а) **Person** – имена персон; (б) **Loc** – локации; (в) **Org** – организации; (г) **Hacker** – отдельные хакеры; (д) **Hacker_Group** – группы хакеров; (е) **Program** – программы, в том числе сайты, функции, части программ; (ж) **Device** – электронное оборудование; (з) **Tech** – технологии, написанные с большой буквы; (и) **Virus** – зловредное ПО разной природы; (к) **Event** – различные события и мероприятия.

Всего было размечено 1124 публикаций, из которых в корпус вошли только те, которые содержат хотя бы один из следующих тегов: **Hacker**, **Hacker_Group**, **Program**, **Device**, **Tech**, **Virus**. В результате объем корпуса составил 861 текст.

В целях устранения возможных неточностей разметки был произведен вторичный анализ размеченных текстов. В ходе анализа было установлено, что разметчики склонны принимать разные решения при разметке одинаковых контекстов, что привело к большому количеству ошибок и неточностей и общей непоследовательности разметки. Так, например, программе *ICQ* в 9 случаях была приписана неверная метка **Tech** и в 18 случаях – верная метка **Program**; аббревиатура *СКЗИ* (средство криптографической защиты информации) в 13 случаях была неверно размечена как **Program**, в 43 случаях получила верный тег **Tech.**

Также в результате анализа размеченных текстов были выявлены случаи непоследовательной разметки, в числе которых:

- Выделение или отсутствие ИС на некоторых транслитерированных иностранных названиях: *микрософт*, *виндовс*;
- Выделение или отсутствие ИС на номерах версий продукта (номерах моделей устройств) в случае, если они перечислены вслед за названием продукта (устройства): *Toughbook CF-53 и CF-31;*
- Выделение одной или двух ИС в контекстах, где за именем персоны (названием организации) в скобках следует то же имя (название), но записанное буквами латинского алфавита: *Университет Вирджинии (University of Virginia)*.

Такое многообразие ошибок и неточностей разметки позволяет утверждать, что ручная разметка корпуса текстов в сфере ИБ представляет собой нетривиальную задачу. Для создания точной и последовательной разметки необходима подробная инструкция для аннотаторов, включающая в себя подробное описание используемых тегов и правила выбора тега для ИС различных типов.

3.1. Инструкция для разметчиков

В целях обеспечения точности и последовательности разметки, нами была разработана инструкция, учитывающая все сложные для разметки контексты, обнаруженные в ходе анализа текстов. Например, были приняты следующие решения:

- На транслитерированных иностранных названиях ИС выделяются вне зависимости от регистра;
- На версиях продукта, перечисленных после названия продукта (устройства), отдельные ИС выделяются, если названия версий (номера моделей) содержат буквенные символы*;*

[3] http://brat.nlplab.org/

• Если за именем персоны (названием организации) в скобках следует то же имя (название), но записанное буквами латинского алфавита, то вся такая последовательность выделяется как единая именованная сущность.

Кроме того, инструкция содержит подробное описание каждого тега. Для каждого тега были перечислены типы ИС, которым он должен быть приписан. Так, например, тег **Device** присваивается ИС следующих типов: компьютеры, смартфоны, карты памяти, жесткие диски, материнские платы, видеокарты, модемы, роутеры.

В соответствии с новой инструкцией были исправлены ошибки и неточности разметки корпуса.

Оценка эффективности новой инструкции может быть произведена путем подсчета коэффициента согласованности между разметчиками до и после введения инструкции. В рамках данной работы произвести такую оценку не представляется возможным, так как отсутствуют тексты, для которых были бы получены несколько разметок от различных аннотаторов. В дальнейшем планируется проведение эксперимента, по результатам которого будет определено изменение коэффициента согласованности между разметчиками после введения инструкции.

4. Модели извлечения ИС в сфере ИБ

Итоговый вариант корпуса был использован для обучения нескольких систем извлечения ИС, основанных на нейронных сетях следующих архитектур:

(A) BiLSTM [3, 4]
(B) BiLSTM-CRF [4]
(C) $\text{BiLSTM}_{\text{CHAR}}$-BiLSTM [7]
(D) $\text{BiLSTM}_{\text{CHAR}}$-BiLSTM-CRF [7]
(E) CNN_{CHAR}-BiLSTM [8]
(F) CNN_{CHAR}-BiLSTM-CRF [8]

Основной слой во всех моделях – двунаправленная LSTM (BiLSTM). BiLSTM способна обучаться долговременным зависимостям и учитывает как правый, так и левый контекст каждого слова. Все модели используют предобученные векторные представления слов модели araneum_none_fasttextskipgram_300_5_2018[4] проекта RusVectōrēs [6], словарь которой содержит более 98% уникальных токенов нашего корпуса.

Последний слой моделей (B), (D) и (F) – CRF-классификатор, который максимизирует вероятность цепочки тегов для целого предложения. Поскольку NER-задачи отличаются высокой степенью взаимной зависимости последовательных тегов друг от друга, такой вариант выходного слоя позволяет значительно улучшить качество работы NER-систем (см. [6, 9]).

Модели (C)-(F) помимо векторного представления слов используют обучаемые посимвольные векторные представления слов (символьные эмбеддинги), что, согласно результатам, представленным в работах [7, 8, 11, 13], должно обеспечить улучшение качества работы системы. Для построения символьных эмбеддингов в моделях (C)-(D) используется BiLSTM-сеть (см. [7]), в моделях (E)-(F) – CNN-сеть (см. [8]). Согласно результатам исследований [11, 13], BiLSTM-char и CNN-char демонстрируют примерно одинаковый прирост F-меры, однако CNN-сеть значительно превосходит BiLSTM-сеть по скорости обучения.

[4] http://rusvectores.org/ru/models/

Все модели, основанные на нейронных сетях, были обучены и протестированы на нашем корпусе с использованием кросс-валидации с соотношением обучающей и тестовой выборки 3:1.

Для оценки качества работы систем использовались такие метрики, как точность, полнота, F-мера. Оценка качества работы систем производилась методом полного соответствия (full matching, exact matching), то есть ИС считалась верно размеченной только в том случае, если системой были верно установлены 1) границы ИС (начальный и конечный токены) и 2) метка (тег) ИС. Результаты работы систем представлены в Таблице 1.

При сравнении систем были исключены из рассмотрения результаты извлечения ИС с тегами **Hacker**, **Hacker_Group**, так как в силу крайне малого числа сущностей данных типов в нашем корпусе (16 и 45 вхождений соответственно), ни одна из систем не показала точность и/или полноту, отличную от нуля.

Наибольшую макро F-меру и наибольшую F-меру для большинства отдельных классов показала модель $BiLSTM_{CHAR}$-BiLSTM-CRF. Наиболее близкие к ней результаты отмечены у модели CNN_{CHAR}-BiLSTM-CRF. При этом, как и ожидалось, время обучения модели CNN_{CHAR}-BiLSTM-CRF оказалось значительно меньше времени обучения $BiLSTM_{CHAR}$-BiLSTM-CRF.

	BiLSTM			BiLSTM-CRF			$BiLSTM_{CHAR}$-BiLSTM			$BiLSTM_{CHAR}$-BiLSTM-CRF			CNN_{CHAR}-BiLSTM			CNN_{CHAR}-BiLSTM-CRF		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Org	68.7	30.3	42.0	73.0	38.3	50.2	75.3	62.1	68.1	78.1	69.1	**73.3**	78.3	48.6	59.9	76.4	67.5	71.6
Loc	90.2	39.4	54.8	88.1	53.5	66.6	92.7	70.0	79.8	92.9	82.3	**87.3**	95.5	52.5	67.6	94.6	73.5	82.7
Person	28.9	8.9	13.5	61.2	30.0	40.3	79.1	46.9	58.9	85.7	54.7	**66.8**	72.8	35.0	47.2	79.2	49.1	60.6
Program	56.6	29.0	38.4	65.1	40.4	49.9	77.6	51.3	61.8	85.8	60.0	**70.6**	71.4	57.1	63.4	78.5	58.2	66.6
Device	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	11.1	0.8	1.5	18.8	2.5	**4.3**	11.9	0.8	1.3
Tech	63.0	4.1	13.3	67.2	16.8	26.9	71.8	55.5	62.6	77.4	41.9	54.4	70.2	48.0	57.0	76.6	53.7	**63.1**
Virus	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	37.5	5.1	**9.0**	3.0	0.4	0.7	23.8	3.8	6.6
Event	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	71.4	5.9	**10.9**	0.0	0.0	0.0	37.6	7.2	12.0
Macro	38.4	13.9	20.2	44.3	22.4	29.2	49.6	35.7	41.4	67.5	40.0	**46.7**	51.2	30.5	37.5	59.8	39.2	45.6

Таб. 1 Результаты работы систем

Низкие показатели для класса **Virus** объясняются их семантической близостью с сущностями класса **Tech** и **Program**. Кроме того, сам класс **Virus** оказался разнородным, так как соответствующий тег использовался для разметки как зловредного ПО, так и технологий, которые используются хакерами. Решением данной проблемы могло бы быть особое представление данного класса: в виде двух подклассов **Malware_Tech** и **Malware_Program**.

Причина низкого качества извлечения ИС класса **Event** также может быть связана с разнородностью данного класса: тег **Event** приписывался как названиям различных мероприятий (конференций, соревнований, форумов), так и названиям различных событий (например, названия праздников, войн).

Причины низкого качества извлечения ИС класса **Device** на данный момент не установлены и требуют более пристального изучения.

К сожалению, результаты, показанные моделями извлечения ИС в данной работе, не могут быть сопоставлены с результатами работы моделей в [9] по двум причинам. Во-первых, несмотря на то, что в качестве обучающей и тестовой коллекции в обеих работах используется Sec_col, разметка коллекции значительно отличается. В [9] из 2000 текстов обучающего корпуса лишь порядка 300 были размечены вручную и содержали сущности,

релевантные для ИБ, для остальных 1700 текстов была получена автоматическая разметка низкого качества с помощью CRF-классификатора, обученного на новостных документах. В нашей работе модели тестировали и обучались на корпусе, содержащем 861 текст коллекции Sec_col, при этом все тексты нашего корпуса были размечены вручную согласно инструкции, созданной в рамках данной работы. Во-вторых, в рамках нашей работы оценка качества работы систем производилась методом полного соответствия, в то время, как в [9] используется метод неполного соответствия.

5. Заключение

В данной работе рассмотрена проблема извлечения ИС из русскоязычных текстов в сфере ИБ. Для обучения и тестирования систем извлечения ИС был создан корпус размеченных текстов в данной предметной области. В ходе разметки была также разработана инструкция для аннотаторов, включающая в себя подробное описание тегов и правила выделения и аннотации ИС в неоднозначных контекстах. Данная инструкция может быть в дальнейшем использована при создании новых корпусов размеченных текстов в сфере ИБ или при разметке текстов коллекции Sec_col в целях увеличения объема нашего корпуса.

На полученном корпусе были обучены и протестированы различные системы извлечения ИС, основанных на нейронных сетях следующих архитектур. Наибольшую F-меру показала модель $\text{BiLSTM}_{\text{CHAR}}$-BiLSTM-CRF, использующая векторные представления слов модели araneum_none_fasttextskipgram_300_5_2018. В дальнейшем качество работы системы может быть улучшено посредством использования в качестве отдельного признака токена его наличие/отсутствие в специальных списках релевантных в сфере ИБ сущностей (см. также [9]). Кроме того, отдельной темой дальнейших исследований может стать поиск наиболее эффективных гиперпараметров модели извлечения ИС в сфере ИБ (см. также [11]).

6. Библиография

[1] *Bridges, R. A., Jones, C. L., Iannacone, M. D., Testa, K. M., & Goodall, J. R.* (2013). Automatic labeling for entity extraction in cyber security. arXiv preprint arXiv:1308.4941.

[2] *Gasmi, H., Bouras, A., & Laval, J.* (2018). LSTM Recurrent Neural Networks for Cybersecurity Named Entity Recognition. ICSEA 2018, 11.

[3] *Graves, A., Mohamed, A. R., & Hinton, G.* (2013, May). Speech recognition with deep recurrent neural networks. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 6645-6649). IEEE.

[4] *Huang, Z., Xu, W., & Yu, K.* (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.

[5] *Joshi, A., Lal, R., Finin, T., & Joshi, A.* (2013, September). Extracting cybersecurity related linked data from text. In 2013 IEEE Seventh International Conference on Semantic Computing (pp. 252-259). IEEE.

[6] *Kutuzov, A., & Kuzmenko, E.* (2016, April). WebVectors: a toolkit for building web interfaces for vector semantic models. In International Conference on Analysis of Images, Social Networks and Texts (pp. 155-161). Springer, Cham.

[7] *Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C.* (2016). Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.

[8] *Ma, X., & Hovy, E.* (2016). End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. arXiv preprint arXiv:1603.01354.

[9] *Mazharov, I. A., & Dobrov, B. V.* (2018). Named Entity Recognition for Information Security Domain.

[10] *Mozharova, V. A., & Loukachevitch, N. V.* (2016, April). Combining knowledge and CRF-based approach to named entity recognition in Russian. In International Conference on Analysis of Images, Social Networks and Texts (pp. 185-195). Springer, Cham.

[11] *Reimers, N., & Gurevych, I.* (2017). Optimal hyperparameters for deep lstm-networks for sequence labeling tasks. arXiv preprint arXiv:1707.06799.

[12] *Weerawardhana, S., Mukherjee, S., Ray, I., & Howe, A.* (2014, November). Automated extraction of vulnerability information for home computer security. In International Symposium on Foundations and Practice of Security (pp. 356-366). Springer, Cham.

[13] *Zhai, Z., Nguyen, D. Q., & Verspoor, K.* (2018). Comparing CNN and LSTM character-level embeddings in BiLSTM-CRF models for chemical and disease named entity recognition. arXiv preprint arXiv:1808.08450.