УДК 004.912

ДВУХЭТАПНЫЙ ПОДХОД К ИЗВЛЕЧЕНИЮ ИМЕНОВАННЫХ СУЩНОСТЕЙ

В.А. Можарова (*valerie.mozharova@gmail.com*)
Н.В. Лукашевич (*louk_nat@mail.ru*)
Московский государственный университет имени Ломоносова, Москва

В данной работе рассматривается двухэтапный подход к извлечению именованных сущностей для русского языка. На первом этапе имена извлекаются на основе метода машинного обучения. Далее собирается статистика по полученной разметке токенов, эта статистика преобразуется в набор признаков, который используется для обучения нового классификатора. Было введено три вида новых признаков: история предсказаний, использование статистики внутри документа, использование статистики внутри всей коллекции. Эксперименты проводились на двух открытых текстовых коллекциях. Было показано, что использование двухэтапного анализа улучшает качество извлечения именованных сущностей.

Ключевые слова: извлечение именованных сущностей, CRF, двухэтапный подход.

Введение

Извлечение именованных сущностей представляет собой необходимый этап для большого количества приложений автоматической обработки текстов и информационного поиска. Так, из новостных сообщений важно извлекать имена людей, названия организаций, географических мест и др.

Предложено много подходов к решению данной задачи. В последнее время большинство подходов основано на машинном обучении, в которых используются различные наборы признаков, включая признаки токенов, словари, автоматические порождаемые кластеры слов и т. д.

Важным типом информации, которую полезно учитывать при извлечении именованных сущностей, является информация о том, как часто рассматриваемый токен был классифицирован как именованная сущность. Это помогает при извлечении имен, состоящих из многозначных слов, или в случаях, когда имена находятся в

неопределенных контекстах. Для этого часто используют двухэтапный подход к извлечению именованных сущностей: на первом этапе делается предварительная классификация, затем собирается частотная статистика, которая затем используется для уточнения на втором этапе.

Двухэтапные алгоритмы исследуются в работах ([Krishnan and others, 2006], [Ratinov and others, 2009], [Straková and others, 2013], [Tkachenko and others, 2012]). Однако эти исследования не проводились для русского языка. В данной работе мы рассматриваем несколько вариантов двухэтапного извлечения именованных сущностей и покажем вклад второго этапа в улучшение извлечения на двух текстовых коллекциях.

1. Обзор близких работ

Задача извлечения именованных сущностей для текстов на русском языке рассматривалась в ряде работ. В работах [Antonova and others, 2013], [Gareev and others, 2013], [Подобряев, 2013] использовался метод машинного обучения CRF. В работе [Трофимов, 2014] описывается система, основанная на словарях и правилах.

В системах извлечения имен на основе машинного обучения обычно используются такие признаки именованных сущностей, как признаки токена (написание с большой буквы, конкретная лемма и др.), словари и контекстные признаки [Подобряев, 2013]. Также проводились исследования по использованию кластеризации слов в задаче распознавания именованных сущностей [Gareev and others, 2013]. Однако задача двухэтапного анализа текста для извлечения именованных сущностей не исследовалась для русского языка, и мы решили произвести соответствующие эксперименты, предварительно изучив схожие работы для английского и чешского языков.

В работе [Ratinov and others, 2009] для английского языка рассматриваются два варианта двухэтапного подхода и их комбинирование: история предсказаний и история предсказаний во всем тексте. Используя двухэтапный подход, авторы достигли 90.57% F-меры (из них вклад двухэтапных признаков составил 2.88%) на текстовой коллекции CoNLL03[1] [Tjon Kim Sang and others, 2003], в которой были размечены 4 вида именованных сущностей (персоны, локации, организации, др. именованные сущности). Описанная система также показала 89.19% F-меры (вклад двухэтапных признаков составил 2.33%) на тестовой коллекции MUC7, в которой были размечены 3 вида именованных сущностей (персоны, локации организации).

В работе [Straková and others, 2013] для чешского и английского языков рассматривалась только история предсказаний и авторы достигли 89.16%

[1] http://www.cnts.ua.ac.be/conll2003/ner/

на текстовой коллекции CoNLL03, где двухэтапный проход дал 1.13% увеличения F-меры, и 79.23% на чешской текстовой коллекции Czech Named Entity Corpus 1.0, в которой были размечены 42 класса именованных сущностей.

В работе [Tkachenko and others, 2012] представлена система для английского языка, в которой токены, классифицированные как именованные сущности на первом проходе, использовались как признаки на втором. Авторы не обнаружили улучшений при добавлении двухэтапного подхода и достигли 91.02% F-меры на тестовой коллекции CoNLL03.

2. Текстовые коллекции

Для экспериментов использовались две открытые текстовые коллекции на русском языке. Первая коллекция «Persons-1000»[1] содержит 1000 новостных документов с размеченными именами персон. Эта коллекция была размечена в Исследовательском Центре Искусственного Интеллекта [Власова и др., 2014].

Мы дополнительно разметили данную коллекцию другими типами именованных сущностей:

- Организации (ORG)
- Медиа организации, имеющие функции информирования (MEDIA)
- Географические объекты (LOC)
- Геополитические объекты (GEOPOLIT) – страны и столицы, выступающие в роли правительства (например, «Москва анонсировала»)

При данной разметке использовались те же правила, которые использовались при разметке текстовой коллекции MUC-7: нет вложенных именованных сущностей, они не пересекаются, а каждому токену соответствует лишь один класс Для экспериментов, представленных в этой статье, были использованы только три вида именованных сущностей: персоны, организации и географические объекты.

Вторая коллекция «Persons-1111-F»[2] содержит 1111 новостных документов с размеченными персонами. Эта коллекция была специально собрана из документов, в которых упоминаются сложные для анализа восточные имена такие, как арабские, индийские, китайские и японские, что может негативно повлиять на качество распознавания именованных сущностей.

[1] http://ai-center.botik.ru/Airec/index.php/ru/collections/28-persons-1000

[2] http://ai-center.botik.ru/Airec/index.php/ru/collections/29-persons-1111-f

3. Метод и признаки для распознавания именованных сущностей

Распознавание именованных сущностей рассматривается как задача классификации токена на заданные категории имен или как находящийся вне имени. В качестве метода классификации используется метод машинного обучения CRF [Lafferty and others, 2001], который специально создан для классификации последовательных данных. Характерным отличием этого метода машинного обучения является возможность учета контекста классифицируемого объекта, что является важным для задачи извлечения именованных сущностей. Мы использовали готовую реализацию данного метода машинного обучения - CRF++[1], преимуществом которого является достаточно быстрое обучение и открытый код.

В качестве основных признаков мы использовали признаки токена и словари. Все сформированные признаки вычислялись для каждого токена и его окружения в радиусе 2 токена. Признаки токена включают в себя лемму токена, его длину, регистр букв, наличие окончания фамилии (-ов, -енко, -швили), наличие гласной и т. д. Словари были составлены с помощью телефонных справочников, Википедии и тезауруса Рутез [Loukachevitch and others, 2014]. Словари содержат слова и словосочетания различной длины (Табл. 1), и их общий объем составил 335 тысяч выражений.

Более подробное описание основных используемых признаков можно найти в работе [Mozharova and others, 2016].

4. Двухэтапный подход

Мы предполагаем, что для лучшей классификации каждого токена полезно использовать предыдущий опыт классификатора и запоминать статистику меток для дальнейшего использования.

Таким образом, на первом этапе имена извлекаются на основе метода машинного обучения. Далее собирается статистика по полученной разметке токенов, эта статистика преобразуется в набор признаков, который используется для обучения нового классификатора. В данной работе исследуются следующие варианты таких признаков: история предсказаний, статистика внутри документа, использование статистики текстовой коллекции.

[1] https://taku910.github.io/crfpp/

Табл. 1

Словарь	Размер	Примеры
Известные люди	31482	Владимир Путин, Ангела Меркель
Имена	2773	Василий, Анна, Том
Фамилии	66108	Кузнецов, Грибоедов
Профессии	9935	министр, китаевед
Глаголы информирования	1729	высказать, признаться, отпроситься
Компании	33380	Сбербанк
Типы компаний	6774	организация, ООО, авиафирма
Медиа организации	3909	РИА Новости, Первый канал
Географические объекты	8969	Балтийское море, Владивосток
Географические прилагательные	1739	финский, томский, югославский
Частотные слова	58432	автомобиль, падать, желтый
Устройства	44094	Устройство, телефон

4.1 История предсказаний

При создании этого признака было сделано предположение о том, что в начале текста человеческие имена обычно встречаются в полной форме, и классификатору легче распознать именованную сущность. Например, фамилию в контексте имени и отчества гораздо проще определить как часть именной сущности, чем если она стоит отдельно. Так для текста, в котором встретилось словосочетание «Геннадий Столяр», а потом встретилось просто слово «Столяр», второе упоминание проще определить как именованную сущность, т. к. в начале текста она уже была классифицирована как персона.

Для каждого токена из текста рассматриваются все его предыдущие вхождения в этот текст, и подсчитывается статистика частоты классов, которые классификатор определил для данного токена. Исходя из этой статистики для токена формируется несколько новых признаков (каждый соответствует своему классу), которые могут принимать три значения: *no_one* (токен до этого не встречался в тексте), *best* (больше чем в 50 процентах случаев токен относился именно к этому классу), *rare* (токен редко относился к этому классу).

Например, если очередной токен «Россия» до этого уже встречался в тексте пять раз, из которых два раза классификатор определил его в класс организаций, а три раза в класс географических объектов, то история предсказаний для шестого токена «Россия» будет следующей: PER – *no_one,* ORG – *rare*; LOC – best.

4.2 Использование статистики внутри документа

Этот метод очень похож на предыдущий, разница заключается только в том, что для каждого токена анализируются его вхождения не только в начале текста, а во всем тексте.

4.3 Использование статистики текстовой коллекции

В данном методе, в отличие от предыдущих, рассматривается не отдельный текстовый документ, а вся коллекция для создания статистики предсказаний.

5. Эксперименты

Двухэтапный подход к извлечению именованных сущностей был протестирован на двух текстовых коллекциях: "Persons-1000" и "Persons-1111-F". Мы использовали CRF-модель для классификации именованных сущностей.

В качестве разметки для обучения классификатора использовалось BIO-представление, в котором токены размечаются тремя типами меток для каждой категории извлекаемых имен: начало именованной сущности, продолжение именованной сущности и неименованная сущность. Например, в предложении «Владимир Путин поздравил россиян с праздником» токен «Владимир» получит метку «Begin-Per» (начало именованной сущности), токен «Путин» получит метку «In-Per» (продолжение именованной сущности), а все остальные токены будут помечены меткой «Out» (не принадлежит никакой именованной сущности).

В качестве целевой метрики была выбрана F-мера, которая является сочетанием точности и полноты. Результаты, полученные двухпроходным алгоритмом на данных коллекциях, сравниваются с базовой системой извлечения, основанной на признаках токена и словарях.

Для обучения классификатора на коллекции «Person-1000» была применена кросс-валидация с соотношением обучающей и тестовой частей 3:1. Чтобы проверить переносимость системы, для второй коллекции мы не стали использовать кросс-валидацию, а обучили систему на коллекции «Persons-1000», а потом применили обученную модель на коллекции «Persons-1111-F». Результаты экспериментов представлены в таблицах (Табл. 2) и (Табл. 3).

На коллекции «Persons-1000» для извлечения имен персон наибольший вклад дал признак история предсказаний, для извлечения имен организаций самым значимым оказался признак, основанный на статистике коллекции в целом. Набор всех признаков дал максимальное улучшение для извлечения совокупности заданных типов именованных сущностей.

На второй коллекции «Persons-1111-F»самым эффективным признаком оказалась история предсказаний. Небольшое влияние глобальной статистики можно объяснить разнородностью коллекции: она была составлена из текстов, упоминающих имена разных народов. Но и для "Person-1000", и для "Persons-1111-F" комбинированный подход привел к увеличению F-меры по сравнению с базовой системой.

Табл. 2

Persons-1000, F-мера(%)				
Система	Персоны	География	Организации	Общее
1) Базовая система	96.61	94.94	85.19	92.49
2) (1) + История предсказаний	97.00	94.53	85.32	92.61
3) (1) + Использование статистики внутри документа	96.78	94.51	85.42	92.56
4) (1) + Использование статистики всей коллекции	96.69	94.79	**85.67**	92.77
(1) + (2) + (3) + (4)	**97.21**	**95.21**	85.60	**92.92**
(Трофимов, 2014)	96.62	-	-	-

Можно сравнить полученные результаты с результатами, описанными в работе [Трофимов, 2014]. В этой работе представлена система извлечения имен персон, основанная на словарях и правилах. Данной системой были достигнуты следующие значения F-меры: 96.62 на коллекции «Persons-1000» и 64.43% на коллекции «Persons-1111-F». Наша система показала лучшие результаты на обеих коллекциях, но существенное увеличение F-меры было достигнуто именно на «Persons-1111-F». Из этого можно сделать вывод, что системы, основанные на машинном обучении, лучше переносятся на новые коллекции.

Табл. 3

Person-1111-F, F-мера(%)	
Система	Персоны
1) Базовая система	86.71
2) (1) + История предсказаний	**88.87**
3) (1) + Предсказания во всем тексте	86.78
4) (1) + Предсказания во всей коллекции	86.72
(1) + (2) + (3) + (4)	87.94
(Трофимов, 2014)	64.43

Заключение

В данной работе было рассмотрен двухэтапный подход к извлечению именованных сущностей из текстов на русском языке. На первом этапе имена извлекаются на основе метода машинного обучения CRF. Далее собирается статистика по полученной разметке токенов. Эта статистика преобразуется в набор признаков, который используется для обучения нового классификатора.

Было введено три вида новых признаков: история предсказаний в документе, использование статистики внутри документа, использование статистики всей коллекции. Эксперименты проводились на двух открытых текстовых коллекциях, и был рассмотрен вклад каждого признака. Было показано, что использование двухэтапного анализа улучшает качество извлечения именованных сущностей.

Также мы сравнили нашу систему с системой, основанной на правилах, и показали существенное улучшение F-меры.

Благодарности. Данная работа частично поддержана грантом РФФИ (грант 16-29-09606 офи_м).

Список литературы

[Krishnan and others, 2006] Krishnan V., Manning C. D. An effective two-stage model for exploiting non-local dependencies in named entity recognition. – ACL, 2006.

[Ratinov and others, 2009] Ratinov L., Roth D. Design challenges and misconceptions in named entity recognition // Proc. 13th Conference on Computational Natural Language Learning, CoNLL. − ACL, 2009.

[Tjon Kim Sang and others, 2003] Tjon Kim Sang, Erik, F., Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition //Proc. 7th conference on Natural language learning at HLT-NAACL 2003. − ACL (2003).

[Straková and others, 2013] Straková J., Straka M., Hajič J. A New State-Of-The-Art. Czech Named Entity Recognizer // «Text, Speech, and Dialogue», Proc. 16th International Conference, TSD 2013. − Springer Berlin Heidelberg, 2013.

[Tkachenko and others, 2012] Tkachenko M., Simanovsky A. Named Entity Recognition: Exploring Features // «Empirical Methods in Natural Language Processing», Proc. 11th Conference on Natural Language Processing, KONVENS 2012. pp. 118−127. – Eigenverlag ÖGAI, 2012.

[Antonova and others, 2013] Antonova A. Y., Soloviev A. N. Conditional random field models for the processing of Russian // Proc. the International Conference "Dialog 2013". − RGGU, 2013.

[Gareev and others, 2013] Gareev R., Tkachenko M., Solovyev V., Simanovsky A., Ivanov V. Introducing baselines for Russian named entity recognition // Proc. 14th International Conference, CICLing 2013. − Springer Berlin Heidelberg, 2013.

[Подобряев, 2013] Подобряев А. В. Поиск упоминаний лиц в новостных текстах с использованием модели условных случайных полей // RCDL-2013 – ЯРГУ им. Демидова, 2013.

[Власова и др., 2014] Власова Н. А. К проблеме разметки текстов на русском языке для задачи извлечения фактографической информации // TEL'2014. − Fan, 2014.

[Lafferty and others, 2001] Lafferty J., McCallum A., Pereira F. Conditional random fields: Probabilistic: models for segmenting and labeling sequence data // Proc. of ICML. − 2001

[Loukachevitch and others, 2014] Loukachevitch N., Dobrov B. RuThes linguistic ontology vs. Russian wordnets // Proc. Global WordNet Conference GWC−2014 − Tartu, 2014.

[Mozharova and others, 2016] Mozharova V., Loukachevitch N. Combining Knowledge and CRF-based Approach to Named Entity Recognition in Russian // Proc. the 5th International Conference on Analysis of Images, Social Networks, and Texts, AIST'2016. − 2016.

[Трофимов, 2014] Трофимов И. В. Выявление личных имен в новостных текстах на материале коллекций Persons-1000/1111-F // RCDL-2014. − 2014..

TWO-STAGE NAMED ENTITY RECOGNITION IN RUSSIAN

V.A. Mozharova (*valerie.mozharova@gmail.com*)
N.V. Loukachevitch (*louk_nat@mail.ru*)
Moscow State University, Moscow

The paper represents our approach to Name Entity Recognizing with several methods of two-stage prediction for Russian language using CRF method as a basic classifier. We tested our system on the open Russian collections "Persons-1000", labeled with persons, organizations and locations, and "Persons-1111", labeled with personal names.

Keywords: named entity recognition, two-stage prediction, CRF.