ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ №4(35), 2017, с. 21–30

УДК 004.89:004.912

Н. А. Власова, А. В. Подобряев

Автоматическое выявление границ именных групп с использованием информации об именованных сущностях

Аннотация. В настоящей работе ставится задача автоматического выявления границ именных групп, заполняющих валентность предиката в предложении. Рассматриваются именные группы любых видов, за исключением сочиненных. Используется предварительная автоматическая сегментация предложений на фрагменты, заведомо содержащие искомые именные группы. Для проведения границ именных групп внутри найденных фрагментов применяется метод машинного обучения. В системе признаков используется информация об извлеченных на предварительном этапе анализа именованных сущностях разных типов, а также данные из базы знаний. Приводятся результаты эксперимента по выявлению границ именных групп.

Ключевые слова и фразы: частичный синтаксический анализ, автоматическое извлечение информации, именованные сущности, машинное обучение.

Введение

Полный синтаксический и семантический анализ текста требует значительных затрат при разработке, и поэтому по-прежнему недоступен для большинства исследователей. Если же текст проанализирован полностью, то встает новая задача извлечения целевой информации из результатов анализа. В системах автоматического анализа текстов все чаще используются подходы, основанные на частичном синтаксическом анализе, который позволяет решать довольно большой спектр практических задач по извлечению и поиску информации [1, 2].

Частичный синтаксический анализ позволяет разобрать текст ровно настолько, насколько требуется для решения конкретной задачи, что значительно сокращает время работы модуля анализа текста и экономит усилия его разработчиков. Одной из наиболее значимых составляющих частичного синтаксического анализа является выделение

Работа выполнена в рамках НИР «Исследование и разработка методов машинного обучения для обнаружения аномалий», номер гос. регистрации 0077-2016-0002.

DOI: 10.25209/2079-3316-2017-8-4-21-30

именных групп (NP-chunking). Выделение именных групп является необходимым при автоматическом выявлении фактов, анализе медицинской документации, при извлечении информации об отношениях и т. д. [3–5].

При выявлении границ именных групп обычно используется два подхода: основанный на правилах, с использованием регулярных выражений, и применение машинного обучения на предварительно размеченной коллекции. В последнее время исследователи все больше склоняются к алгоритмам, использующим машинное обучение как к наиболее быстродействующим и легким в разработке. Появление в открытом доступе размеченных корпусов текстов существенно облегчает подготовку материала для обучения алгоритма.

Для русского языка было предпринято несколько попыток автоматического выявления именных групп [6–8]. В работе [7] алгоритм был основан на правилах, при этом выявлялись только именные группы, содержащие имена персон. В работе [8] использовался алгоритм, основанный на машинном обучении, однако рассматривались только группы с зависимыми прилагательными и существительными в родительном падеже. Множество признаков состояло только из простых признаков, доступных на этапе предварительной сегментации и морфологического анализа текста.

1. Постановка задачи

В настоящей работе ставится задача выявления границ именных групп любых типов (включая предложные и аппозитивные), кроме сочиненных и однородных, разделенных запятыми или другими знаками препинания. См. примеры, иллюстрирующие различные случаи положения именных групп:

(1) *<Последние две недели> <отставки глав регионов>* ***следовали*** *<одна за другой>.*

(2) *<По данным следствия>, <они>* ***получили*** *<70 тысяч рублей> <за то>, чтобы* ***освободить*** *<представителя местной компании> <от административной ответственности за правонарушение в сфере контроля за оборотом алкогольной продукции>.*

(3) *<В конце января> <президент России>* ***досрочно отправил*** *<в отставку> <главу Дагестана Магомедсалама Магомедова>* ***и назначил*** *<его> <заместителем главы администрации Кремля>.*

(4) *<руководитель отдела по разведке нефтяных месторождений Иранской национальной нефтяной компании (NIOC) Хормоз Калаванд>.*

(5) *<Заместителем председателя комитета по архитектуре и градостроительству города Москвы>* ***назначена*** *<Татьяна Гук>,* ***сообщил*** *<РИА Новости> <во вторник> <источник в городской администрации>.*

(6) *<Бывшего заместителя министра Носова> <осенью 2012 года>* ***приговорили*** *<к четырем годам колонии-поселения> <за вымогательство>.*

(7) *<Во вторник> <президент Обама>* ***принял*** *<в Белом доме> < избранного президента Мексики Энрике Пенья Ньето>.*

Как видно из приведенных примеров, в новостных текстах часто встречаются сложные именные группы. Также такие группы могут следовать друг за другом. Именно поэтому проблема определения границ синтаксических групп очень важна при частичном синтаксическом анализе. Если границы именных групп будут выделены правильно, это повысит качество решения задач извлечения фактов и отношений.

Для анализа рассматриваются публицистические тексты новостного содержания. В отличие от художественных текстов они обладают относительно устойчивым синтаксисом, более единообразны с точки зрения лексики и порядка слов. Под именной группой понимается синтаксическая группа, заполняющая валентность предиката (глагола в личной форме, причастия или деепричастия).

Мы предпринимаем попытку использовать при машинном обучении развитую систему признаков, отражающую способ восприятия текста человеком. При восприятии текста глазами или на слух, человек анализирует его динамически, строя представление предложения сразу в процессе восприятия. Поэтому границы между синтаксическими группами читающий человек проводит, даже не дочитав или не дослушав предложение до конца. В устной речи для проведения границ служат паузы и интонация. В письменной речи — другие признаки, и это, конечно, не только информация о токенах и морфологических характеристиках словоформ. На границы разных синтаксических групп указывает комплекс признаков разного уровня. В частности, существуют работы, где для выявления границ именных групп использовалась информация об именованных сущностях, извлеченных из текста на предыдущем этапе анализа [9]. Однако в этом исследовании описывается система, основанная на правилах, а результаты тестирования не

приводятся. Для русского языка, насколько нам известно, подобные эксперименты не проводились.

Мы предполагаем, что для поиска границ синтаксических групп разных типов являются значимыми разные множества признаков. Очевидно, глагол в личной форме, начало или конец предложения, знак препинания являются сильными признаками границы именной группы. Поэтому предлагается двухуровневый алгоритм. Сначала все предложения разбиваются на фрагменты, разделенные сильными ограничителями. Тогда задача выделения именных групп сводится к нахождению их границ внутри фрагментов. Для решения этой задачи предлагается множество признаков, которые одновременно используют информацию разных языковых уровней. Кроме того, используются признаки, характеризующие фрагмент в целом, а не только составляющие его словоформы. Признаки были найдены эвристически на основе анализа множества выделенных фрагментов предложения между «сильными» границами.

2. Подготовка материала для исследования

Предварительная обработка текста проводилась с помощью системы ИСИДА-Т [10]. Эта обработка включает в себя токенизацию, морфологический анализ, а также выявление именованных сущностей нескольких классов [11–13]. После предварительной обработки с помощью системы правил производится разбиение предложений текста на фрагменты, разделенные «сильными» границами.

Правила выделения фрагментов определяются исходя из предположения о проективной структуре предложения (для публицистических текстов это практически всегда верно). Границами фрагментов служат глаголы в личной форме, деепричастия, знаки препинания, кроме кавычек, союзы, предикатные слова, такие, как «нет», «нельзя», «можно» и др., прилагательные и причастия в краткой форме [14].

Система признаков для дальнейшего поиска границ именных групп внутри фрагментов должна отражать реально встречающиеся явления в текстах новостной тематики. Для эвристического построения таких признаков использовался анализ полученных фрагментов на основе следующих характеристик.

(*1*) Длина фрагмента, т.е. количество составляющих фрагмент словоформ. При этом если несколько словоформ представляют собой выявленную именованную сущность, то этой последовательности словоформ присваивается значение длины 1.

(*2*) Типы ограничивающих элементов справа и слева, т.е. начало или конец предложения, глагольная группа, причастие, деепричастие, знак препинания.

(*3*) Последовательность тегов частей речи составляющих фрагмент словоформ.

(*4*) Последовательность тегов именованных сущностей, выявленных на предварительном этапе анализа.

На основе данных характеристик оказалось возможным группировать полученные фрагменты разными способами, что позволило сделать наблюдения над структурой фрагментов и типами их границ. Именно такие наблюдения помогли сформулировать эвристические признаки для алгоритма машинного обучения.

3. Система признаков

Предлагаемая система признаков учитывает различную информацию, полученную на предварительных этапах анализа.

(*1*) Графематическая информация о типе токена (число или слово, латиница или кириллица, с большой или с маленькой буквы, вхождение в выражение в кавычках).

(*2*) Морфологическая информация: частеречный тег для каждой словоформы внутри фрагмента; падежная характеристика каждой словоформы (без снятия падежной омонимии); значение словоизменительной категории «число» для каждой словоформы во фрагменте (без снятия омонимии); принадлежность слова к определенному выделенному классу словоформ (помимо частеречных тегов) — это классы личных, относительных, указательных и определительных местоимений.

(*3*) Информация из базы знаний после обработки текста: вхождение в состав словосочетания, выделенного как именованная сущность определенного типа («имя лица», «название организации», «геополитическая единица», «указание на время»); информация о слове из базы знаний (связь слова с концептом в базе знаний системы ИСИДА-Т).

(*4*) Комбинированные признаки: длина фрагмента (число словоформ, входящих во фрагмент, при этом уже выделенные именованные сущности считаются за одну словоформу); последовательность частеречных тегов внутри фрагмента; типы границ фрагмента (начало предложения, конец предложения, предикатное слово, знак

препинания); последовательность выявленных именованных сущностей и концептов внутри фрагмента с учетом количества слов между ними, не связанных с концептами в базе знаний; порядковый номер словосочетания (словоформы), представляющего именованную сущность определенного вида.

4. Машинное обучение

Назовем наблюдаемой последовательностью последовательность словоформ выделенного на предварительном этапе фрагмента предложения. Набор словоформ, составляющих именованную сущность, считается одним членом последовательности. Членами наблюдаемой последовательности также считаются ограничивающие фрагмент элементы. Выделение внутри фрагмента именных групп эквивалентно расстановке между членами наблюдаемой последовательности меток двух типов: <граница> (именной группы) и <не граница>. Близкие метки не являются независимыми случайными величинами, в то же время достаточно далекие друг от друга метки можно считать независимыми. Такую ситуацию описывает линейная графовая модель CRF (см., например, [**15**]).

Нами использовалась модель с максимальной кликой третьего порядка, то есть каждая вершина графа, соответствующая метке, соединена с соседними и с двумя членами наблюдаемой последовательности. Использовалась l_2-регуляризация.

5. Эксперимент

Для обучения и тестирования алгоритма была размечена коллекция, состоящая из 2000 русскоязычных новостных текстов. Выделено 58763 фрагмента, содержащих именные группы рассматриваемых типов. В каждом фрагменте были вручную размечены границы именных групп. Всего именных групп получилось 75432. Для обучения было взято 2/3 текстов коллекции.

Точность нахождения границ именных групп алгоритмом составила 93,45%, полнота — 89,68%, F-мера — 91,53%.

Список литературы

[1] S. Abney. “Parsing by chunks”, *Principle-based parsing*, Studies in Linguistics and Philosophy, vol. **44**, Kluwer Academic Publishers, 1991, pp. 257–278. ↑ 21

[2] L. Ramshaw, M. Marcus. “Text chunking using transformation-based learning”, 3rd Annual Workshop on Very Large Corpora Proceedings (Boston, Massachussetts, USA, June 1995), Text, Speech and Language Technology, vol. **11**, pp. 82–94. ↑ 21

[3] О. И. Бабина, Т. Ю. Мыларщикова. «Извлечение именных групп из корпуса текстов на испанском языке», *Вестник Южно-Уральского государственного университета. Лингвистика*, **22** (2011), с. 47–53. ↑ 22

[4] P. Jindal, D. Roth. “Extraction of events and temporal expressions from clinical narratives”, *Journal of Biomedical Informatics*, **46**, suppl. (2013), pp. S13–S19. ↑ 22

[5] N. Vazov. “A system for extraction of temporal expressions from French texts”, TALN 2001 (Tours, France, 2–5 juillet 2001), 2001, pp. 313-322, URL: https://www.atala.org/doc/actes_taln/AC_0132.pdf ↑ 22

[6] А. А. Романенко. *Применения условных случайных полей в задачах обработки текстов на естественном языке*, Выпускная квалификационная работа магистра, М., 2014, 27 с., URL: http://www.machinelearning.ru/wiki/images/f/fc/Romanenko2014Application.pdf ↑ 22

[7] Л. Г. Крейдлин. «Программа выделения русских индивидуализированных именных групп TagLite», *Компьютерная лингвистика и интеллектуальные технологии*, Сборник трудов ежегодной международной конференции «Диалог» (Звенигород, Россия, 1–6 июня 2005), с. 292–297. ↑ 22

[8] M. S. Kudinov, A. A. Romanenko, I. I. Piontkovskaja. “Conditional random field in segmentation and noun phrase inclination tasks for Russian”, *Computational Linguistics and Intellectual Technologies*, **13**:20 (2014), pp. 297–306. ↑ 22

[9] P. Osenova, S. Kolkovska. “Combining the named entity recognition task and NP chunking strategy for robust pre-processing”, *Proceedings of 1st Workshop on Treebanks and Linguistic Theories* (Sozopol, Bulgaria, 20–21 September 2002), pp. 167–182. ↑ 23

[10] Д. А. Александровский, Д. А. Кормалев, М. С. Кормалева, Е. П. Куршев, Е. А. Сулейманова, И. В. Трофимов. «Развитие средств аналитической обработки текста в системе ИСИДА-Т», *Тр. Десятой нац. конф. по искусственному интеллекту с междунар. участием КИИ-2006*. Т. 2 (Обнинск, Россия, 25–28 сентября 2006), Физматлит, М., 2006, с. 555–563. ↑ 24

[11] Д. А. Кормалев, Е. П. Куршев, Е. А. Сулейманова, И. В. Трофимов. «Технология извлечения информации из текстов, основанная на знаниях», *Программные продукты и системы*, 2009, №2, с. 62–66. ↑ 24

[12] Н. А. Власова. «Об одной проблеме автоматического извлечения временной информации из русскоязычных текстов», *Программные системы: теория и приложения*, **5**:4(22) (2014), с. 231–242, URL: http://psta.psiras.ru/read/psta2014_4_231-242.pdf ↑ 24

[13] И. В. Трофимов. «Выявление личных имен в новостных текстах на материале коллекций Persons-1000/1111-F», *Электронные библиотеки: перспективные методы и технологии, электронные коллекции*, XVI Всероссийская научная конференция RCDL-2014 (Дубна, Россия, 13–16 октября 2014 г.), 2014, с. 217–221. ↑ 24

[14] Н. А. Власова, А. В. Подобряев. «К вопросу об определении границ именных групп при решении задач автоматического извлечения информации из текстов на русском языке», *Программные системы: теория и приложения*, **7**:1(28) (2016), с. 153–170, URL: http://psta.psiras.ru/read/psta2016_1_153-170.pdf ↑ 24

[15] Ch. Sutton, A. McCallum. “An introduction to conditional random fields”, *Foundations and Trends in Machine Learning*, **4**:4 (2011), pp. 267–373. ↑ 26

Пример ссылки на эту публикацию:

Н. А. Власова, А. В. Подобряев. «Автоматическое выявление границ именных групп с использованием информации об именованных сущностях», *Программные системы: теория и приложения*, 2017, **8**:4(35), с. 21–30.

URL: http://psta.psiras.ru/read/psta2017_4_21-30.pdf

Об авторах:

Наталья Александровна Власова

Младший научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна РАН, один из разработчиков технологии построения систем извлечения информации

e-mail: nathalie.vlassova@gmail.com

Алексей Владимирович Подобряев

Младший научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна РАН, один из разработчиков технологии построения систем извлечения информации

e-mail: alex@alex.botik.ru

Natalia Vlasova, Alexey Podobryaev. *Automatic noun phrases extraction using preliminary segmentation and CRF with semantic features.*

Abstract. We consider the task of finding the borders of noun phrases (NP) that are actants of predicates. First, we make a preliminary segmentation of sentences to fragments that contain NPs. Second, we use CRF to find the borders of NPs inside the fragments. Data from the knowledge base and information about named entities found in the text are used as features for machine learning. We present the results of our experiment and discuss future work. (*in Russian*).

Key words and phrases: shallow parsing, automatic information extraction, named entities, machine learning.

References

[1] S. Abney. "Parsing by chunks", *Principle-based parsing*, Studies in Linguistics and Philosophy, vol. **44**, Kluwer Academic Publishers, 1991, pp. 257–278.

[2] L. Ramshaw, M. Marcus. "Text chunking using transformation-based learning", 3rd Annual Workshop on Very Large Corpora Proceedings (Boston, Massachussetts, USA, June 1995), Text, Speech and Language Technology, vol. **11**, pp. 82–94.

[3] O. I. Babina, T. Yu. Mylarshchikova. "Noun phrase extraction from a Spanish corpus", *Vestnik Yuzhno-Ural'skogo gosudarstvennogo universiteta. Lingvistika*, **22** (2011), pp. 47–53 (in Russian).

[4] P. Jindal, D. Roth. "Extraction of events and temporal expressions from clinical narratives", *Journal of Biomedical Informatics*, **46**, suppl. (2013), pp. S13–S19.

[5] N. Vazov. "A system for extraction of temporal expressions from French texts", TALN 2001 (Tours, France, 2–5 juillet 2001), 2001, pp. 313-322, URL: https://www.atala.org/doc/actes_taln/AC_0132.pdf

[6] A. A. Romanenko. *Applications of conditional random fields in text processing problems in natural language*, Vypusknaya kvalifikatsionnaya rabota magistra, M., 2014 (in Russian), 27 p., URL: http://www.machinelearning.ru/wiki/images/f/fc/Romanenko2014Application.pdf

[7] L. G. Kreydlin. "TagLite: The program of identification of Russian individualized NPs", *Komp'yuternaya lingvistika i intellektual'nyye tekhnologii*, Sbornik trudov yezhegodnoy mezhdunarodnoy konferentsii "Dialog" (Zvenigorod, Rossiya, 1–6 iyunya 2005), pp. 292–297 (in Russian).

[8] M. S. Kudinov, A. A. Romanenko, I. I. Piontkovskaja. "Conditional random field in segmentation and noun phrase inclination tasks for Russian", *Computational Linguistics and Intellectual Technologies*, **13**:20 (2014), pp. 297–306.

[9] P. Osenova, S. Kolkovska. "Combining the named entity recognition task and NP chunking strategy for robust pre-processing", *Proceedings of 1st Workshop on Treebanks and Linguistic Theories* (Sozopol, Bulgaria, 20–21 September 2002), pp. 167–182.

DOI: 10.25209/2079-3316-2017-8-4-21-30

[10] D. A. Aleksandrovskiy, D. A. Kormalev, M. S. Kormaleva, Ye. P. Kurshev, Ye. A. Suleymanova, I. V. Trofimov. "Development of the ISIDA-T system's tools for analytic text processing", *Tr. Desyatoy nats. konf. po iskusstvennomu intellektu s mezhdunar. uchastiyem KII-2006*. V. 2 (Obninsk, Rossiya, 25–28 sentyabrya 2006), Fizmatlit, M., 2006, pp. 555–563 (in Russian).

[11] D. A. Kormalev, Ye. P. Kurshev, Ye. A. Suleymanova, I. V. Trofimov. "Knowledge-based information extraction technology", *Programmnyye produkty i sistemy*, 2009, no.2, pp. 62–66 (in Russian).

[12] N. A. Vlasova. "On one problem of automatic information extraction from Russian texts", *Program Systems: Theory and Applications*, **5**:4(22) (2014), pp. 231–242 (in Russian), URL: http://psta.psiras.ru/read/psta2014_4_231-242.pdf

[13] I. V. Trofimov. "Person name recognition in news articles based on the Persons-1000/1111-F Collections", *Elektronnyye biblioteki: perspektivnyye metody i tekhnologii, elektronnyye kollektsii*, XVI Vserossiyskaya nauchnaya konferentsiya RCDL-2014 (Dubna, Rossiya, 13–16 oktyabrya 2014 g.), 2014, pp. 217–221 (in Russian).

[14] N. A. Vlasova, A. V. Podobryayev. "To the noun phrase recognition problem in application to automatic information extraction from Russian texts", *Program Systems: Theory and Applications*, **7**:1(28) (2016), pp. 153–170 (in Russian), URL: http://psta.psiras.ru/read/psta2016_1_153-170.pdf

[15] Ch. Sutton, A. McCallum. "An introduction to conditional random fields", *Foundations and Trends in Machine Learning*, **4**:4 (2011), pp. 267–373.

Sample citation of this publication:

Natalia Vlasova, Alexey Podobryaev. "Automatic noun phrases extraction using preliminary segmentation and CRF with semantic features", *Program systems: Theory and applications*, 2017, **8**:4(35), pp. 21–30. (*In Russian*).

URL: http://psta.psiras.ru/read/psta2017_4_21-30.pdf