ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ №1(28), 2016, с. 153–170

УДК 004.89:004.912

Н. А. Власова, А. В. Подобряев

К вопросу об определении границ именных групп при решении задач автоматического извлечения информации из текстов на русском языке

Аннотация. Рассматривается задача выделения сложных именных групп в русскоязычных публицистических текстах в приложении к задачам автоматического извлечения информации. Под сложными именными группами понимаются длинные именные группы, содержащие генитивные, предложные конструкции, а также имена собственные. Предлагается схема поиска границ именных групп, начиная с фрагмента текста, заведомо содержащего именную группу. Разработан алгоритм выделения таких фрагментов. Произведена их классификация на основании частоты встречаемости типов фрагментов, количества слов фрагмента, их частеречного состава, наличия уже выделенных именованных сущностей разных видов, информации о вхождении частей фрагментов в списки сложных предлогов и устойчивых сочетаний. Приводится первоначальная система признаков для построения алгоритма автоматического выделения именных групп в границах построенных на первом этапе анализа фрагментов. В экспериментальной части исследования произведено выделение фрагментов (58032 фрагмента) из коллекции текстов общественно-политической тематики (1000 документов), произведен анализ сложных случаев.

Ключевые слова и фразы: автоматическое извлечение информации, выделение именованных сущностей, извлечение ситуаций, выделение именных групп, синтаксический анализ.

Введение

Задача автоматического извлечения информации заключается в отображении некоторых фрагментов текста на естественном языке в определенную структуру (модель), состоящую из объектов и связей между ними. В качестве примеров можно привести задачи извлечения именованных сущностей, терминов, ключевых слов, фактов, автоматическое составление пресс-портретов, авторефератов и др. Любая задача по извлечению информации предполагает:

Работа выполнена при финансовой поддержке РФФИ, проект 15-07-05277_а.

(1) наличие определенной структуры (модели), с элементами которой будут сопоставляться найденные фрагменты текста;
(2) решение задачи сегментации, то есть выделения текстовых фрагментов, соответствующих элементам заданной структуры;
(3) наличие правил сопоставления или интерпретации, то есть способа интерпретации выделенных фрагментов текста в знания, представляемые в определенной модели. Сюда же относится задача нормализации, то есть приведения текстовых фрагментов к стандартному виду, например, к канонической форме, к стандартному представлению дат и т.п.

Воспринимающий текст человек динамически обрабатывает информацию, вычленяя нужную и дополняя ее вновь получаемыми о некотором предмете сведениями. В условиях недостаточного теоретического исследования вопросов понимания текстов человеком, современные системы обычно опираются на информацию, извлекаемую из контактно расположенных фрагментов текста, не производя полного синтаксического и семантического разбора текста. Поэтому критически важными являются правила определения границ релевантных фрагментов текста для каждой конкретной задачи извлечения информации.

В прикладных исследованиях по автоматическому извлечению информации чаще всего задача формулируется в терминах «знаний» — того, что нужно найти. Таковы практические задачи по поиску именованных сущностей [1–4], терминов [5, 6], ключевых слов [7, 8], фактов [9] и т. д. Однако, полный или хотя бы частичный синтаксический анализ оказывается полезен для автоматического извлечения информации. Для решения прикладных задач обычно используются методы, основанные на частичном синтаксическом анализе (shallow parsing) [10] как более быстродействующие и легче настраиваемые на конкретную практическую задачу.

Обычно в задачах на поиск именованных сущностей, терминов, ключевых слов исследователи ограничиваются двух- или трехсловными (максимум — четырехсловными сочетаниями) [5], поиск организован с использованием словарной информации, списков, обращений к тезаурусам и базам знаний. Таким образом, границы искомого фрагмента текста задаются жестко перед началом работы программы. Традиционно используются методы, основанные на правилах, методы машинного обучения и гибридные подходы, использующие возможности как правил, так и методов машинного обучения. Первые обычно опираются на регулярные выражения и списки, используют отсылки

к элементам системы знаний. Методы машинного обучения используют графематические и морфологические признаки, а также данные о последовательности слов в контексте.

Трудности при автоматическом извлечении информации возникают в следующих случаях:

(1) при извлечении упоминаний лиц и названий организаций, не входящих в списки или состоящих более, чем из трех слов, или имеющих в своем составе предложные группы;
(2) при извлечении многословных терминов и ключевых словосочетаний;
(3) при выделении фрагментов текста, обозначающих участников ситуации, в задаче автоматического извлечения фактов;
(4) при извлечении информации об отношениях, например, «роль + лицо», «организация + название», «геополитическая единица + название» (информация об этих отношениях обычно выражается в рамках одной именной группы).

Во всех этих случаях искомые фрагменты текста могут иметь сложную структуру: с сочинением, с обособленными оборотами, с приложениями, в том числе в скобках.

Приведем примеры, иллюстрирующие указанные выше трудности (здесь и далее рассматриваемые именные группы выделены угловыми скобками):

⟨*старший помощник руководителя регионального СКР по взаимодействию со СМИ Валерия Павлова*⟩;

⟨*В конце января*⟩ ⟨*президент России*⟩ *досрочно отправил* ⟨*в отставку*⟩ ⟨*главу Дагестана Магомедсалама Магомедова*⟩ *и назначил* ⟨*его*⟩ ⟨*заместителем главы администрации Кремля*⟩;

⟨*Места в министерстве*⟩ *сохранили* ⟨*статс-секретарь генерал армии Николай Панков*⟩, ⟨*начальник тыла Вооруженных сил Дмитрий Булгаков*⟩, *а также* ⟨*заместитель по международному военному сотрудничеству Анатолий Антонов*⟩;

⟨*зампредседателя Комитета Госдумы по безопасности и противодействию коррупции Александр Хинштейн*⟩;

⟨*руководитель отдела по разведке нефтяных месторождений Иранской национальной нефтяной компании (NIOC) Хормоз Калаванд*⟩;

Об этом заявил ⟨*даже аккуратный в выражениях Александр Новак, глава российского министерства энергетики*⟩;

⟨*Кабинет министров Украины*⟩ *уволил* ⟨*Владимира Козака*⟩, ⟨*который*⟩ *занял пост* ⟨*министра инфраструктуры*⟩, ⟨*с должности гендиректора государственной администрации железнодорожного транспорта Украины*⟩.

Очевидно, что основные трудности при автоматическом выявлении искомой информации связаны с явлением морфологической и синтаксической омонимии, а также (в подходах без синтаксического анализа или с частичным синтаксическим анализом) с отсутствием знаний о синтаксической структуре предложения. Поскольку чаще всего практические задачи по извлечению информации опираются на фрагменты текста, представляющие собой именные группы, перспективным представляется путь выделения синтаксически связных фрагментов — именных групп. Для русского языка такая работа с применением методов машинного обучения проводилась для именных групп определенного вида [**10**], а именно для нерекурсивных именных групп с зависимыми двух типов: согласованными определениями и именами существительными в родительном падеже. В этой же работе отмечены трудности применения машинного обучения с простыми морфологическими и графематическими признаками для именных групп с более сложным устройством, а именно с входящими в состав групп предложных конструкций (например, «*Международный суд по правам человека в Гааге*»).

1. Постановка задачи

В приложении к задаче автоматического извлечения информации настоящая работа посвящена вопросу сегментации, то есть поиску границ релевантного фрагмента текста. Минимальный такой фрагмент — именная группа, обозначающая объект внеязыковой действительности. При этом для решения большого количества практических задач извлечения требуется найти границы максимальных именных групп, заполняющих валентности предикатного слова, поэтому на первом этапе мы будем рассматривать задачу выделения именных групп, которые зависят от предикатного слова. Необходимо сделать замечание, что такие предикатные слова, как отглагольные существительные и причастия в полной форме в настоящей работе не рассматриваются как границы при выделении фрагментов. Вопрос о выделении именных групп, зависящих от предикатных слов этих типов, будет рассматриваться отдельно.

⟨*Заместителем председателя комитета по архитектуре и градостроительству города Москвы*⟩ ***назначена*** ⟨*Татьяна Гук*⟩, ***сообщил*** ⟨*РИА Новости*⟩ ⟨*во вторник*⟩ ⟨*источник в городской администрации*⟩.

В данном примере участники ситуации назначения — это именные группы ⟨*Татьяна Гук*⟩ и ⟨*Заместителем председателя комитета по архитектуре и градостроительству города Москвы*⟩. Кроме того, в этом предложении упоминается и ситуация сообщения и ее участники: ⟨*РИА Новости*⟩, ⟨*источник в городской администрации*⟩ и ⟨*во вторник*⟩. При автоматическом извлечении информации о ситуации увольнения необходимо определить границы всех этих групп.

Таким образом, нашей задачей является определение границ именных групп, заполняющих валентности глаголов в личной форме, а также других предикатных слов, играющих в предложении роль сказуемого.

При использовании методов, основанных на правилах, структуры именных групп задаются регулярными выражениями, при использовании методов машинного обучения возможные типы конструкций также задаются заранее. В обоих случаях типы анализируемых словосочетаний описываются до работы системы извлечения информации, и при этом не учитывается все многообразие именных сочетаний как внутри синтаксической группы, так и встречающиеся в текстах случаи соположения именных групп, между которыми нужно провести границы. Чтобы иметь возможность учитывать все многообразие структур, которые встречаются в текстах, мы предлагаем опираться на реально встречающиеся в документах именные группы. В настоящей работе рассматривается такой способ классификации именных групп, который позволяет сделать наблюдения за статистикой их употребления, определить наиболее часто встречающиеся конструкции. Предлагается следующий алгоритм: сначала из текста выделяются фрагменты, заведомо содержащие искомые именные группы, а затем внутри найденных фрагментов проводятся границы. Таким образом, мы будем рассматривать задачу выделения именных групп как задачу определения границ внутри ранее определенных фрагментов текста.

Максимальный размер фрагмента — это предложение. Однако, если исходить из предположения о проективной структуре каждого предложения в рассматриваемом тексте (а для современных публицистических и научных текстов это, конечно, так), то в предложении можно выделить более мелкие фрагменты, целиком содержащие именные группы. То есть перед выделением непосредственно именных

групп можно простыми правилами выделить фрагменты, в которых они заведомо содержатся целиком.

Предлагается следующий план решения задачи выделения сложных именных групп:

(1) выделение фрагментов текста, содержащих именные группы;
(2) классификация выделенных фрагментов;
(3) анализ выделенных фрагментов, построение системы признаков;
(4) использование построенных признаков для машинного обучения.

Настоящая работа посвящена пунктам 1–3 этого плана.

2. Выделение фрагментов, содержащих именные группы

Для выделения фрагментов, содержащих именные группы, мы используем такие процедуры лингвистического анализа, при автоматическом проведении которых большинство современных систем демонстрируют хорошие результаты. К таким процедурам относятся:

(1) разделение текста на предложения и слова;
(2) определение части речи для каждого слова;
(3) определение морфологических характеристик каждого слова с точностью до омонимии.

Исходя из предположения о проективности структуры каждого предложения в рассматриваемых текстах, легко определить правила выделения фрагментов текста, содержащих именные группы целиком с точностью до групп, в состав которых входят обособленные обороты и сочиненные конструкции. Границы фрагментов внутри предложения проводятся в следующих случаях:

(1) в начале и в конце предложения;
(2) перед знаками препинания, кроме кавычек;
(3) перед союзом «*и*»;
(4) перед и после глагола в личной форме, а также в форме деепричастия (частица «*не*» также относится к глагольной словоформе);
(5) перед и после предикатных слов, таких, как «*нет*», «*нельзя*», «*можно*», «*нужно*» и т.д.;
(6) перед и после прилагательных и причастий в краткой форме;
(7) перед наречием, если за ним сразу идет глагол в личной форме или деепричастие;
(8) после наречия, союза, частицы, если они стоят в начале предложения;
(9) после глагола в неопределенной форме, если он стоит в начале предложения, после запятой, союза, глагола в личной форме.

Необходимо сказать, что на данном этапе исследования не рассматриваются как единая именная группа сочиненные и однородные именные группы, а также входящие в состав сложных именных групп обособленные запятыми обороты. В дальнейшем планируется использовать полученную при выделении фрагментов информацию и признаки для выделения и таких групп.

Автоматическое определение приведенных выше лингвистических характеристик происходило в рамках системы извлечения информации ИСИДА-Т [11, 12]. Эта система имеет модульную структуру и содержит независимые модули последовательного лингвистического анализа: токенизации, выделения основы слова, определения части речи и др.

Первые два модуля стандартны. Определение части речи выполнено на основе алгоритма TnT [13] (метода машинного обучения с использованием скрытой марковской модели). Его обучение производилось на коллекции национального корпуса русского языка [14], содержащей порядка миллиона словоупотреблений. Использовалась номенклатура частей речи, принятая в национальном корпусе русского языка. Если слово не содержится в текстах обучающего множества, то происходит анализ его окончания (набора 1–5 последних букв). Чтобы определить длину окончания слова, которое определяет часть речи, используется следующая процедура. Для вычисления априорной вероятности части речи при условии данного слова, доли данной части речи в обучающем множестве для каждого из возможных окончаний (длины от 1 до 5) суммируются с весами. Веса являются несмещенными оценками дисперсии долей частей речи с данным окончанием. Таким образом, вклад окончания, для которого доли частей речи в обучающем множестве примерно одинаковы, будет меньше вклада окончания с большим разбросом долей частей речи в обучающем множестве.

3. Анализ выделенных фрагментов

Эксперименты по выделению фрагментов, содержащих именные группы, были проведены на коллекции новостных текстов, состоящей из 1000 специально подобранных документов. Эти тексты включают в себя сложные случаи — сложные по структуре именные группы и сложные случаи соположения именных групп. В результате обработки этих документов с помощью описанного выше алгоритма было выделено 58032 фрагмента. Для каждого фрагмента были определены характеристики:

Таблица 1. Структура фрагментов, содержащих именные группы

Длина фрагмента	Количество фрагментов	Доля фрагментов, %, ≈
1	12855	22
2	12739	22
3	10000	17
4	7179	12
5	5119	8
$\geqslant 6$	10135	17

(1) длина фрагмента (число словоформ, составляющих данный фрагмент);
(2) часть речи для каждого слова;
(3) информация о словах или знаках препинания, являющихся границами фрагмента.

Проведенная процедура позволяет сделать статистические наблюдения над структурой фрагментов, полученных из новостных текстов (таблица 1).

Проанализировав полученные данные, можно говорить, что почти половина (44%) фрагментов состоит из 1 или 2 слов, а фрагментов длиной более 6 слов – меньше одной пятой части всех фрагментов. Полученное множество фрагментов удобно анализировать. Фрагменты можно сортировать по длине, составу, что позволяет делать выводы об устройстве именных групп и их соположениях в реально встречающихся текстах. Таким образом, получено разбиение текстов на фрагменты, что позволяет создать классификацию встречающихся в текстах именных конструкций.

Для дальнейшей работы используются именованные сущности, автоматически выделенные с помощью модулей системы ИСИДА-Т [**2**, **3**, **11**, **15**, **16**]. А именно, при определении длины фрагмента найденные имена лиц, названия организаций, геополитических единиц и временные выражения считаются за одну словоформу. После этой процедуры получаются результаты, приведенные в таблице 2.

Выделение внутри фрагментов именованных сущностей позволяет изменить статистику: фрагментов длины 1 стало на 5% больше (примерно на 3000 фрагментов). Значит, можно утверждать, что выделение в тексте именованных сущностей улучшает результаты выделения именных групп даже при простом разбиении текста на фрагменты.

Таблица 2. Структура фрагментов с учетом именованных сущностей

Длина фрагмента	Количество фрагментов	Доля фрагментов, %, ≈
1	15773	27
2	11004	18
3	8173	14
4	6113	11
5	4359	8
$\geqslant 6$	12610	22

4. Обсуждение системы признаков

Используя выделенные именованные сущности, а также проанализировав структуру полученных фрагментов, можно найти некоторые закономерности, которые помогают проводить границы именных групп внутри многословных фрагментов. В настоящем разделе приводятся некоторые наблюдения, которые служат мотивацией введения признаков для последующего машинного обучения. Машинное обучение планируется проводить на множестве фрагментов, выделяемых при помощи описанного выше алгоритма.

В нашем анализе используются такие именованные сущности, которые в большинстве систем автоматического анализа выделяются с высокой точностью и полнотой. Это имена лиц и указания на время. В системе ИСИДА-Т для выделения имен лиц используется специальный алгоритм [15], а выделение указаний на время производится с помощью системы правил в рамках одного из модулей лингвистического анализа текста [16]. F-мера для выделения обоих типов именованных сущностей превышает 95%.

Приведем некоторые наблюдения о границах именных групп внутри фрагментов.

4.1. Использование выделенных именованных сущностей

Одним из признаков для машинного обучения будет являться порядковый номер выделенной именованной сущности во фрагменте. Так, наблюдения над множеством уже полученных фрагментов позволяют заметить, что если фрагмент начинается с определенной именованной сущности (имени лица или указания на время), то граница проводится сразу после этой именованной сущности. В примерах ниже выделенные фрагменты предложений набраны нежирным курсивом,

а символом «|» обозначены искомые границы именных групп внутри фрагментов.

Украинская власть ***позволит*** *Юлии Тимошенко | выехать на лечение в Германию;*

Николай Угаслов | в предвыборном манифесте ***предложил*** *прагматичную идею;*

Временно исполняющим обязанности гендиректора ***назначен*** *Евгений Бахтеев, который месяц назад* ***был уволен*** *Беляевым | с поста заместителя гендиректора за нарушение служебных обязанностей;*

Сегодня | депутат Государственной Думы от КПРФ Владимир Бессонов ***обратился*** *в Думу с просьбой отказать в передаче в суд его уголовного дела;*

Днем в субботу | вооруженные люди в форме, похожей на российскую, ***оцепили*** *парламент Крыма в Симферополе;*

В 2003 году | контрольный пакет акций РЗБ ***консолидировала*** *Елена Батурина, супруга бывшего мэра Москвы Юрия Лужкова.*

Подсчеты показывают, что на рассматриваемом материале данное правило соблюдается в 98% случаев. Исключения — конструкции типа «*Безделов вместе с соучастниками*», причастные обороты типа «*вчера заявивший о своей отставке*».

Кроме того, важным признаком является *последовательность именованных сущностей внутри фрагмента.* Так, если во фрагменте подряд встречаются имя лица и указание на время, то между ними проходит граница именных групп.

Бывшего заместителя министра Носова | осенью 2012 года ***приговорили*** *к четырем годам колонии-поселения за вымогательство;*

Президент России Владимир Путин | 4 января 2014 года ***подписал*** *указ №5, которым назначил генерал-майора полиции Игоря Зиновьева заместителем начальника полиции — начальником управления уголовного розыска ГУ МВД по Москве.*

4.2. Использование знаний

Принадлежность слов к определенному классу в ресурсе знаний (онтологии, базе знаний, тезаурусе) можно рассматривать как один из признаков для машинного обучения. Некоторые последовательности приписанных классов словоформ внутри фрагмента отвечают именным группы, а другие — нет. Например, последовательность (должность/роль, прилагательное, руководящий орган, организация) реализуется именной группой:

Председателем попечительского совета Михайловского театра ***стал*** *Владимир Мединский,*

а в последовательности («геополитическая единица, должность/роль, геополитическая единица, имя лица») необходимо провести границу между двумя именными группами:

Во вторник президент Обама ***принял*** *в Белом доме | избранного президента Мексики Энрике Пенья Ньето.*

4.3. Специальные классы слов

Перспективным представляется выделение специальных классов слов как отдельного признака для последующего машинного обучения.

Например, после прилагательного «*который*», использующегося как союзное слово, всегда проводится граница именной группы.

Фабрика мороженого ***должна начать работать*** *в Брянске следующей весной,* ***что обеспечит*** *гарантированный сбыт молока, производство которого | в регионе* ***планируют наращивать***.

В отдельный класс слов предполагается выделить словоформы «*его*», «*ее*», «*их*», а также притяжательные и указательные местоимения. Конструкции с этими словами часто вызывают затруднения при проведении границ именных групп. Эти трудности возможно преодолеть, если использовать комбинированные признаки с этими словами. Например, их порядковый номер в последовательности словоформ внутри фрагмента вместе с данными о частеречной принадлежности этих словоформ, информацию о типе ограничивающих фрагмент элементов.

Сергей Собянин ***назначил*** *его | своим замом по транспорту 26 октября 2010 года.*

Соположение словоформы «*его*» и притяжательного местоимения «*своим*» однозначно предсказывает границу именных групп.

Фактически *он* ***являлся*** *не только коллегой бывшего директора ФСИН Александра Реймера,* ***но и*** *его земляком по работе в Оренбурге.*

Фрагмент «*его земляком по работе в Оренбурге*» ограничен союзами с одной стороны и точкой — с другой. В таком случае, при заданной последовательности элементов фрагмента, проведение границы именных групп между «*его*» и «*земляком*» невозможно.

Чаплин ***добавил, что*** *все обвинения Кураева* ***строятся*** *на словах его анонимных осведомителей.*

При такой последовательности слов во фрагменте после глагола в личной форме («*строятся*») проведение границы ни перед, ни после словоформы «*его*» невозможно.

4.4. Тип фрагмента

Одним из признаков является длина фрагмента и тип ограничивающих фрагмент элементов. Важность этих признаков в комбинации с другими признаками можно увидеть по приведенным в предыдущем пункте примерам.

4.5. Специальные конструкции

В качестве одного из признаков для машинного обучения предлагается использовать задаваемые списком конструкции, выделяемые запятыми и представляющие собой единую именную группу. К таким конструкциям относятся, например, следующие: «*по словам*», «*по информации*», «*по данным*», «*по мнению*», «*в том числе*», «*согласно*». Например (фрагменты представлены по порядку увеличения длины):

По словам Билялетдинова;

По словам собеседника агентства;

По словам отстраненного главы государства;

По словам одного из собеседников издания;

По словам нового премьер-министра Крыма Сергея Аксенова;

По словам помощника президента РТ по социальным вопросам Татьяны Ларионовой.

4.6. Конструкции с предлогом

Для выделения именных групп, содержащих предложные группы, в качестве признака можно использовать статистические данные о сочетаниях имени с предлогами: «*вопрос о*», «*война с*», «*борьба с*» и т.д.

Вопрос об отставке главы РКС Юрия Урличича ***решится*** *в течение 2-3 дней;*

Его нынешний пресс-секретарь Ирина Зубарева около десяти лет ***отвечала*** *за связи с общественностью в управлении «К», в обязанности которого* ***входит*** *борьба с преступностью в сфере информационных технологий.*

В рассматриваемых примерах имена существительные «*связи*» и «*борьба*» управляют предлогом «*с*», а существительное «*вопрос*» — предлогом «*о (об)*». Статистика употребления существительных с данными предлогами показывает, что граница именных групп перед предлогами не проводится.

Затем *депутат* ***пригласил*** *бизнесмена | в свой офис для продолжения разговора.*

В этом примере сочетание существительного «*бизнесмена*» с предлогом «*в*» не относится к частотным сочетаниям, поэтому между этими словами проводится граница именных групп.

4.7. Подфрагменты в именительном падеже

Если внутри фрагмента есть подфрагмент, в котором словоформы стоят в именительном падеже (и это единственная морфологическая характеристика данной последовательности словоформ), то можно сказать однозначно, что перед ними расположена граница именной группы. Таких примеров много в построенном множестве фрагментов.

Одной из причин революции | Путин ***назвал*** *коррупцию на Украине;*

К новым обязанностям | Максим Солюс ***приступит*** *1 апреля.*

Есть и случаи, когда *внутри фрагмента есть две последовательности словоформ в именительном падеже*, тогда граница именной группы проводится перед первой словоформой в именительном падеже, а между последовательностями точно границы нет. В примере ниже подчеркнуты последовательности словоформ в именительном падеже.

Пресс-секретарь президента России Дмитрий Песков ***отказался комментировать*** *информацию о местонахождении Владимира Путина*

Приведенные признаки, выделенные на основе наблюдений над фрагментами, содержащими именные группы, планируется использовать для машинного обучения наряду с традиционными графематическими и морфологическими признаками для улучшения результатов работы алгоритма выделения именных групп.

Заключение

В настоящей работе рассмотрена задача выделения сложных именных групп, содержащих генетивные, предложные конструкции и именованные сущности. Эта проблема ограничивается приложениями к задачам автоматического извлечения информации из русскоязычных публицистических текстов (в частности, предполагается проективность предложений).

Предложено свести задачу к поиску границ именных групп внутри фрагментов предложения, заведомо содержащих именные группы. Приведен алгоритм выделения таких фрагментов. В дальнейшей работе предполагается использовать графовые методы машинного обучения [17], позволяющие произвести разметку найденного фрагмента целиком (с учетом всего контекста), а не классификацию каждого слова в отдельности. Для этого необходим набор содержательных лингвистических признаков (помимо лексических признаков и признаков контекста). Для построения системы таких признаков был проведен анализ и классификация выделенных фрагментов из коллекции новостных русскоязычных текстов на основании частоты встречаемости типов фрагментов, количества слов фрагмента, их частеречного состава, наличия уже выделенных именованных сущностей разных типов, информации о вхождении частей фрагментов в списки сложных предлогов и устойчивых сочетаний. На основании наблюдений построены некоторые признаки для дальнейшего автоматического выделения именных групп в границах, полученных на первом этапе анализа фрагментов.

Список литературы

[1] М. М. Брыкина, А. В. Файнвейц, С. Ю. Толдова. «Извлечение и идентификация именованных сущностей с использованием словарей в русском языке», *Актуальные инновационные исследования: наука и практика*, 2013, №1, URL: http://www.hse.ru/pubs/share/direct/document/118232483 ↑ 154

[2] А. В. Подобряев. «Региональный классификатор текстов для поиска упоминаний лиц в новостных текстах», *Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции»*, RCDL 2014 (Дубна, Россия, 13–16 октября 2014), с. 214–216, URL: http://rcdl.ru/doc/2014/paper/RCDL2014_214-216.pdf ↑ 154,160

[3] И. В. Трофимов. «Выявление личных имен в новостных текстах на материале коллекций Persons-1000/1111-F», *Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции»*, RCDL 2014 (Дубна, Россия, 13–16 октября 2014), с. 217–221, URL: http://rcdl.ru/doc/2014/paper/RCDL2014_217-221.pdf ↑ 154,160

[4] Л. Г. Крейдлин. «Программа выделения русских индивидуализированных именных групп TagLite», *Компьютерная лингвистика и интеллектуальные технологии*, Сборник трудов ежегодной международной конференции «Диалог», Диалог 2005, URL: http://www.dialog-21.ru/Archive/2005/Kreidlin%20LG/KreydlinL.pdf ↑ 154

[5] П. И. Браславский, Е. А. Соколов. «Автоматическое извлечение терминологии с использованием поисковых машин интернета», *Компьютерная лингвистика и интеллектуальные технологии*, Сборник трудов ежегодной международной конференции «Диалог», Диалог 2007, URL: http://www.kansas.ru/pb/paper/dialog2007.pdf ↑154

[6] П. И. Браславский, Е. А. Соколов. «Сравнение четырех методов автоматического извлечения двухсловных терминов из текста», *Компьютерная лингвистика и интеллектуальные технологии*, Сборник трудов ежегодной международной конференции «Диалог», Диалог 2006, с. 88–94, URL: http://www.dialog-21.ru/digests/dialog2006/materials/html/Braslavski.htm ↑ 154

[7] Н. В. Лукашевич, Ю. М. Логачев. «Комбинирование признаков для автоматического извлечения терминов», *Вычислительные методы и программирование*, **11** (2010), с. 108–116, URL: http://num-meth.srcc.msu.ru/zhurnal/tom_2010/pdf/v11r211.pdf ↑ 154

[8] С. О. Шереметьева, П. Г. Осминин. «Методы и модели автоматического извлечения ключевых слов», *Вестник ЮУрГУ. Серия Лингвистика*, **12**:1 (2015), с. 76–81, URL: http://vestnik.susu.ru/linguistics/article/download/3420/3157 ↑ 154

[9] Н. А. Власова. «Извлечение информации о ситуациях отставок-назначений в новостных текстах. Опыт разметки коллекции. Результаты тестирования», *Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции»*, RCDL 2013 (Ярославль, Россия, 14–17 октября 2013), с. 145–154, URL: http://rcdl2013.uniyar.ac.ru/doc/full_text/s4_2.pdf ↑ 154

[10] M. S. Kudinov, A. A. Romanenko, I. I. Piontkovskaja. “Conditional Random Field in segmentation and Noun Phrase inclination tasks for Russian”, Dialogue 2014 (Bekasovo, Russia, 4–8 June 2014), *Computational Linguistics and Intellectual Technologies*, no.13(20), Papers from the Annual International Conference “Dialogue”, pp. 297–306, URL: http://www.dialog-21.ru/digests/dialog2014/materials/pdf/KudinovMS.pdf ↑ 154,156

[11] 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», *Труды 10-й национальной конференции по искусственному интеллекту с международным участием*. Т. 2, КИИ 2006 (Обнинск, Россия, 25–28 сентября 2006), с. 555–563. ↑ 159,160

[12] 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(86), с. 62–66. ↑ 159

[13] T. Brants. “TnT — A Statistical Part-of-Speech Tagger”, *Foundations of Statistical Natural Language Processing*, eds. Ch. D. Manning, H. Schutze, MIT Press, Stanford, May 1999, URL: http://www.hlt.utdallas.edu/~sanda/courses/NLP/Brants.pdf ↑ 159

[14] Национальный корпус русского языка, Ruscorpora, 2015, URL: http://www.ruscorpora.ru/ ↑ 159

[15] Е. А. Сулейманова, К. А. Константинов. «Морфологический анализ незнакомых фамилий в русскоязычном тексте», *Программные продукты и системы*, 2009, №2(86), с. 66–71. ↑ 160,161

[16] Н. А. Власова. «Об одной проблеме автоматического извлечения временной информации из русскоязычных текстов», *Программные системы: теория и приложения*, 2014, №4(22), с. 231–242, URL: http://psta.psiras.ru/read/psta2014_4_231-242.pdf ↑ 160,161

[17] D. Koller, N. Friedman. *Probabilistic Graphical Models*, MIT Press, 2009. ↑ 166

Рекомендовал к публикации *к.т.н. Е. П. Куршев*

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.com

Пример ссылки на эту публикацию:

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*, 2016, **7**:1(28), с. 153–170.

URL: http://psta.psiras.ru/read/psta2016_1_153-170.pdf

Natalia Vlasova, Alexey Podobryaev. *To the noun phrase recognition problem in application to automatic information extraction from Russian texts.*

Abstract. We consider the problem of complex noun phrase recognition in Russian news texts with application to automatic information extraction. By complex noun phrases we mean long noun phrases that contain genitive or/and prepositional constructions and named entities. We describe a plan of noun phrase recognition that begins with a selection of the sentence fragments that undoubtedly contain noun phrases. The fragments selection algorithm is developed. The fragments are classified by frequency of their types, number of words in the fragment, part of speech structure, presence of extracted named entities, some complex prepositions and stable expressions. We introduce a feature system to make automatic noun phrase recognition inside selected fragments. In experiments we have selected 58032 fragments from 1000 documents collection of Russian news. We consider some complex cases. (*In Russian*).

Key words and phrases: information extraction, named entities recognition, noun phrase chunking.

References

[1] M. M. Brykina, A. V. Fainveitz, S. Ju. Toldova. "Dictionary based extraction and identification of names entities in Russian", *Aktualnye innovazionnye issledovaniya: nauka i prilozheniya*, 2013, no.1 (in Russian), URL: http://www.hse.ru/pubs/share/direct/document/118232483

[2] A. V. Podobryaev. "Regional classification of Russian news texts for person recognition", *Proceedings of 16-th Russian conference RCDL*, RCDL 2014 (Dubna, Russian, October 13–16, 2014), pp. 214–216 (in Russian), URL: http://rcdl.ru/doc/2014/paper/RCDL2014_214-216.pdf

[3] I. V. Trofimov. "Person name recognition in news articles based on the Persons-1000/1111-F collections", *Proceedings of 16-th Russian conference RCDL*, RCDL 2014 (Dubna, Russian, October 13–16, 2014), pp. 217–221 (in Russian), URL: http://rcdl.ru/doc/2014/paper/RCDL2014_217-221.pdf

[4] L. G. Kreidlin. "TagLite: The Program of Identification of Russian Individualized NPs", *Computational Linguistics and Intellectual Technologies, Papers from the Annual International Conference "Dialogue"*, Dialog 2005 (in Russian), URL: http://www.dialog-21.ru/Archive/2005/Kreidlin%20LG/KreydlinL.pdf

[5] P. I. Braslavskiy, E. A. Sokolov. "Automatic terms extraction using web search engine", *Computational Linguistics and Intellectual Technologies, Papers from the Annual International Conference "Dialogue"*, Dialog 2007 (in Russian), URL: http://www.kansas.ru/pb/paper/dialog2007.pdf

[6] P. I. Braslavskiy, E. A. Sokolov. "Comparison of four methods of automatic extraction of two-words terms", *Computational Linguistics and Intellectual Technologies, Papers from the Annual International Conference "Dialogue"*, Dialog 2006, pp. 88–94 (in Russian), URL: http://www.dialog-21.ru/digests/dialog2006/materials/html/Braslavski.htm

[7] N. V. Lukashevich, Ju. M. Logachev. “Combining features for automatic terms extraction”, *Vychislitelnye metody i programmirovanie*, **11** (2010), pp. 108–116 (in Russian), URL: http://num-meth.srcc.msu.ru/zhurnal/tom_2010/pdf/v11r211.pdf

[8] S. O. Sheremetyeva, P. G. Osminin. “On methods and models of keyword automatic extraction”, *Bulletin of the South Ural University. Ser. Linguistics*, **12**:1 (2015), pp. 76–81 (in Russian), URL: http://vestnik.susu.ru/linguistics/article/download/3420/3157

[9] N. A. Vlasova. “Extracting information on appointments and dismissals from news texts. An experience in developing an annotated corpus. Testing results”, *Proceedings of 15-th Russian conference RCDL*, RCDL 2013 (Yaroslavl, Russia, October 14–17, 2013), pp. 145–154 (in Russian), URL: http://rcdl2013.uniyar.ac.ru/doc/full_text/s4_2.pdf

[10] M. S. Kudinov, A. A. Romanenko, I. I. Piontkovskaja. “Conditional Random Field in segmentation and Noun Phrase inclination tasks for Russian”, Dialogue 2014 (Bekasovo, Russia, 4–8 June 2014), *Computational Linguistics and Intellectual Technologies*, no.13(20), Papers from the Annual International Conference “Dialogue”, pp. 297–306, URL: http://www.dialog-21.ru/digests/dialog2014/materials/pdf/KudinovMS.pdf

[11] D. A. Aleksandrovskiy, D. A. Kormalev, M. S. Kormaleva, E. P. Kurshev, E. A. Suleymanova, I. V. Trofimov. “Development of analytical tools of text processing in the system ISIDA-T”, *Proceedings of 10-th Russian conference in artificial intelligence*. V. 2, KII 2006 (Obninsk, Russia, September 25–28, 2006), pp. 555–563 (in Russian).

[12] D. A. Kormalev, E. P. Kurshev, E. A. Suleymanova, I. V. Trofimov. “Knoledge-based information extraction technology”, *Programnye producty i sistemy*, 2009, no.2, pp. 62–66 (in Russian).

[13] T. Brants. “TnT — A Statistical Part-of-Speech Tagger”, *Foundations of Statistical Natural Language Processing*, MIT Press, Stanford, May 1999, eds. Ch. D. Manning, H. Schutze, URL: http://www.hlt.utdallas.edu/~sanda/courses/NLP/Brants.pdf

[14] *National Russian language corpora*, Ruscorpora, 2015, URL: http://www.ruscorpora.ru/

[15] E. A. Suleymanova, K. A. Konstantinov. “Morphological analysis of unknown surnames in Russian text”, *Programnye producty i sistemy*, 2009, no.2(86), pp. 66–71 (in Russian).

[16] N. A. Vlasova. “On one problem of automatic information extraction from Russian texts”, *Programmnye sistemy: teoriya i prilozheniya*, 2014, no.4(22), pp. 231–242 (in Russian), URL: http://psta.psiras.ru/read/psta2014_4_231-242.pdf

[17] D. Koller, N. Friedman. *Probabilistic Graphical Models*, MIT Press, 2009.

Sample citation of this publication:

Natalia Vlasova, Alexey Podobryaev. “To the noun phrase recognition problem in application to automatic information extraction from Russian texts”, *Program systems: theory and applications*, 2016, **7**:1(28), pp. 153–170. (*In Russian*). URL: http://psta.psiras.ru/read/psta2016_1_153-170.pdf