Подход к автоматическому извлечению информации о назначениях и отставках лиц (на материале новостных сообщений).

© Н.А.Власова
Институт Программных Систем РАН имени А.К.Айламазяна,
г.Переславль-Залесский
nathalie.vlassova@gmail.com

Аннотация

Настоящая работа посвящена описанию подхода к извлечению структурированной информации из новостных текстов в рамках проекта извлечения информации из текстов ИСИДА-Т, разрабатываемого в ИЦИИ ИПС имени А.К.Айламазяна РАН. В предлагаемом подходе к извлечению фактов в качестве основы используются результаты обработки текста в программной системе ИСИДА-Т — морфологический, частично синтаксический анализ текста, извлеченные именованные сущности (имена людей, географические названия, организации, адреса, названия должностей и званий), а также отношения, которые могут быть установлены между ними на уровне анализа связей в именной группе. Извлечение фактов основано на выделении целевых слов, описывающих ситуации, с последующим нахождением именных групп-участников, расположенных непосредственно рядом с целевыми словами.

1 Введение

Большинство современных систем извлечения информации из текстов построены для решения определенных задач, сформулированных заранее. Нет системы, которая была бы рассчитана на максимальное решение задачи извлечения — полный синтаксический и семантический анализ произвольного текста и на получение всей информации, которая содержится в этом тексте. Любая задача для системы извлечения данных из текста может быть сформулирована в диапазоне от минимальной до максимальной. Задача-минимум — извлечь из текста именованные сущности (географические названия, имена людей, должности, звания, названия организаций и т.д.).

Задача-минимум успешно решается во многих современных системах извлечения информации из текста[4],[10]. Следующий шаг – извлечение из текста сложных структур, более сложных, чем просто объекты. То есть необходимо связать извлеченные уже на первом этапе анализа объекты отношениями, информация о которых тоже должна быть извлечена из текста. Например, сбор информации об организациях и учреждениях — кто возглавляет, где находятся филиалы, когда они были открыты, кто ими заведует, информация об изданных указах и распоряжениях, о принятых законах, результатах выборов и референдумов и т.п. В настоящей работе описывается подход к извлечению ситуаций из новостных текстов на примере ситуаций назначения, увольнения, пребывания и смены в должности. Такие ситуации представляют интерес с нескольких точек зрения. Во-первых, участниками этих ситуаций являются объекты, которые чаще всего извлекаются из текста, — это имена людей, названия должностей и организаций. Во-вторых, участники ситуации могут быть выражены однотипными именными группами — например, в ситуации назначения первый и второй участник (кто назначил и кого назначил) могут быть выражены именными группами, обозначающими имя лица, что усложняет задачу и, соответственно, представляет больший интерес для исследователя. В-третьих, такие ситуации часто встречаются в новостных текстах, что позволяет легко собирать тексты для тестовых коллекций.

2 Способы извлечения фактов из неструктурированного текста. Постановка задачи

В настоящей работе речь будет идти только о методах извлечения информации, основанных на правилах. Методы, использующие машинное обучение и статистические методы, не рассматриваются. Правилами на формальном языке обычно задаются шаблоны. Далее в тексте происходит поиск фрагментов, соответствующих шаблону. Достоинства и недостатки метода шаблонов очевидны. К достоинствам можно отнести точность настройки на конкретную задачу, обозримость и ясность правил, отсутствие необходимости создавать большой размеченный корпус текстов для обучающего множества. Недостатки – скорость работы системы

Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г.

часто бывает неоправданно низкой, правил получается достаточно много, перенастроить систему на другую задачу практически невозможно, приходится переписывать всю систему правил. Кроме того, метод шаблонов больше подходит для языков с фиксированным порядком слов (таких, как большинство западноевропейских языков), а для русского, например, этот метод получается неоправданно "дорогим".

В настоящей работе в рамках метода извлечения информации, основанного на правилах (система ИСИДА-Т), предлагается подход, который учитывает недостатки метода шаблонов и позволяет ускорить и оптимизировать работу. Основная идея – опора не на шаблон целиком, а на ключевое слово, описывающее ситуацию (главное слово группы – глагол, отглагольное существительное или причастие, деепричастие). Вокруг ключевого слова с помощью правил собираются именные группы – участники ситуации. Следующий шаг – приписывание ролей найденным участникам в зависимости от синтаксических характеристик главных слов найденных именных групп

Что же будет пониматься под фактом, подлежащим извлечению? Факт может быть задан в тексте по-разному. Границы факта могут быть в пределах именной группы (например, *директор предприятия Виктор Петров, действующий президент*), в пределах глагольной группы (*Иван Петров назначен директором*), а также вообще в нескольких предложениях или даже целиком в тексте. Например, факт назначения Вероники Скворцовой министром задаётся текстом:

Владимир Путин, вступив в должность президента Российской Федерации, произвел ряд кадровых перестановок. Новый министр здравоохранения Вероника Скворцова, по его мнению, сможет справиться с накопившимися проблемами.

Факт назначения Вероники Скворцовой на должность министра здравоохранения автоматически извлечь из такого текста пока не представляется возможным.

Итак, под фактом, подлежащим извлечению, мы будем понимать факт, который задаётся в рамках глагольной группы (глагол и зависимые от него именные группы).

3 Пример реализации предлагаемого подхода в системе ИСИДА-Т

Рассмотрим реализацию описываемого подхода на примере ситуаций назначения, увольнения и пребывания в должности. Исходные данные для извлечения фактов — результаты обработки текста в системе ИСИДА-Т [1], [6], [7]. Извлечение информации из текстов в системе ИСИДА-Т основано на предварительной лингвистической обработке текста. Результатом первичной обработки является полный морфологический разбор всех слов, входящих в текст. Далее с помощью модуля правил на специальном формальном языке производится синтаксический анализ (не всего предложения, а именных групп, включая предложные). Результаты морфологического и синтаксического анализа сохраняются в **аннотациях** — специальных структурах, которые сопоставляются фрагментам текста (при морфологическом анализе этот фрагмент соответствует слову, а при синтаксическом - словосочетанию). При этом специальными аннотациями помечаются группы, главные слова в которых — имена людей, названия должностей или званий, названия организаций, географические названия.

Рассмотрим алгоритм извлечения фактов из произвольного неструктурированного текста (новостного) на простом примере ситуации назначения. Первый этап — поиск предложений, содержащих ключевые слова, описывающие ситуацию назначения. В данном примере найдено предложение:

Президент России Дмитрий Медведев назначил полковника полиции Николая Васильева министром внутренних дел по Карачаево-Черкесской Республике.

4 Исходные данные для анализа

В ресурсе знаний системы ИСИДА-Т [8],[9] есть элемент знаний @назначение. В словаре ему соответствует словарная статья, в которую входят слова, описывающие в тексте ситуацию назначения — **назначить, поставить, назначение, переназначить.** У элемента знаний в ресурсе есть атрибуты по числу участников ситуации: 1-ый участник, 2-ой участник, 3-ий участник. В значениях атрибутов записаны значения ролей участников. В данном случае с ситуацией @назначение это кто_назначил, кого_назначил, кем_назначил.

В результате первичной обработки текста именным группам, входящим в состав рассматриваемого предложения, сопоставлены аннотации, связывающие главное и зависимые слова в группе. Кроме того, в атрибутах этих аннотаций уже прописаны отношения, связывающие главное и зависимое слово. Так, в данном предложении 3 большие именные группы:

Президент России Дмитрий Медведев (между группами *президент России* и *Дмитрий Медведев* установлено отношение $роль-лицо),

Полковника полиции Николая Васильева (между группами *полковник полиции* и *Николай Васильев* установлено отношение $звание-лицо),

Министром внутренних дел по Карачаево-Черкесской Республике (между группой *министром* и группой *внутренних дел* установлено отношение $ограничение, между группой *министром* и группой *по Карачаево-Черкесской республике* — также $ограничение).

На следующем этапе работы алгоритма создаются аннотации для именных групп — потенциальных участников ситуации. С помощью правил на формальном языке PSL (Pattern Specification Language) задается область сопоставления справа и слева от слова, помеченного аннотацией со ссылкой на ситуацию. Таким образом, создаются аннотации сначала для групп, расположенных непосредственно справа и слева от слова *назначил*. Следующий этап поиска — проверка более удаленных позиций, расположенных непосредственно рядом с уже найденными именными группами. Так, в рассматриваемом примере на первом этапе будут отмечены группы

(Президент Дмитрий Медведев) (назначил)

(назначил) (полковника полиции Николая Васильева),

а на втором этапе поиска - группа

(назначил) (полковника полиции Николая Васильева министром внутренних дел по Карачаево-Черкесской республике)

В каждой группе выделено главное слово. Следует обратить внимание на то, что более удаленная именная группа (министром внутренних дел по Карачаево-Черкесской Республике) помечается аннотацией, которая содержит и более близко расположенную группу: *полковника полиции Николая Васильева*. Таким образом, в результате разметки предложения три именные группы оказались отмеченными как потенциальные участники ситуации назначения, и аннотации, которыми они помечены, расположены непосредственно рядом с аннотацией, которой отмечено слово-ситуация. Как можно увидеть из описания алгоритма, предполагается, что именные группы, расположенные непосредственно справа и слева от слова, обозначающего ситуацию, относятся именно к этой ситуации и между ними не может оказаться группы, относящейся к другому глаголу (отглагольному существительному, причастию, деепричастию). При поиске таких именных групп между словом-ситуацией и группой потенциального участника допускается наличие наречий, частиц, указаний на время (временные указатели собираются на более ранней стадии анализа текста). Для каждой именной группы-участника создается столько аннотаций, сколько значений падежа может быть у главного слова группы. В рассматриваемом примере две аннотации будет только у группы *полковника полиции Николая Васильева* (соответственно родительный и винительный падеж).

Следующий этап анализа — запись информации об именных группах, найденных вокруг слова-ситуации, в аннотацию, соответствующую названию ситуации. Информация об участниках записывается посредством создания в аннотации-ситуации атрибутов, которые называются значениями падежа главного слова именной группы, а значением атрибута является ссылка на аннотацию, соответствующую группе с главным словом в данном падеже. Так, в рассматриваемом примере в аннотацию при глаголе *назначил* будет добавлено три атрибута — *Part_I* (участник в именительном падеже, значение – ссылка на именную группу *президент Дмитрий Медведев*), *Part_V* (участник в винительном падеже, значение — ссылка на именную группу *полковника полиции Николая Васильева*) и *Part_T* (участник в творительном падеже, значение – ссылка на именную группу *министром внутренних дел по Карачаево-Черкесской республике*).

Завершающая стадия работы алгоритма — приписывание ролей найденным участникам. На этом этапе создаются специальные аннотации, посредством которых моделируется синтактико-семантическая связь между предикатом и его актантом, так как именно на этой стадии работы алгоритма происходит интерпретация синтаксической связи в семантическое отношение, прописанное в ресурсе знаний у соответствующего элемента. У вновь образующихся аннотаций строятся следующие атрибуты:

Атрибут *Master* — ссылка на главное слово (в нашем примере - назначил),

Атрибут *Slave* — ссылка на зависимое слово (главное слово зависимой именной группы),

Атрибут *DomRel* — название отношения, которое связывает *Master* и *Slave*.

Как же происходит интерпретация? Здесь учитывается сразу несколько факторов:

- Морфологические характеристики слова, называющего ситуацию;
- Наличие других найденных и зафиксированных участников и их морфологических характеристик;
- Прописанные в ресурсе знаний значения отношений между словом, называющим ситуацию, и именной группой-участником.

В рассматриваемом примере глагол *назначил* стоит в личной форме, поэтому участник в именительном падеже однозначно интерпретируется как 1-ый участник ситуации, участник в винительном падеже — как 2-ой участник, а участник в творительном падеже — как 3-ий участник (согласно значению атрибутов у элемента знаний @назначение — кто_назначил — *президент России Дмитрий Медведев*, кого_назначил — *полковника полиции Николая Васильева*, кем_назначил — *министром внутренних дел по Карачаево-Черкесской Республике*). Особенностью предлагаемого подхода является возможность унифицированного описания синтаксических конструкций. Припи-

сывание ролей (1-ый участник, 2-ой участник, 3-ий участник) именным группам происходит только в зависимости от морфологических характеристик слова, называющего ситуацию, а конкретную смысловую интерпретацию этого отношения можно установить по ссылке на элемент знаний, соответствующий анализируемой ситуации. Это делает систему гибкой, легко перенастраиваемой на анализ других ситуаций. Достаточно внести в ресурс знаний элементы, описывающие нужные ситуации и заполнить у этих элементов атрибуты, с помощью которых будут интерпретироваться отношения между словом-ситуацией и именной группой, ― анализ будет производиться точно так же, как и для ранее введенных в рассмотрение ситуаций.

Помимо этого, при подобном устройстве системы извлечения фактов появляется возможность перед приписыванием ролей и окончательным оформлением результата извлечения провести сопоставление найденных участников и, возможно, в некоторых случаях, отказаться от построения аннотаций, интерпретирующих синтаксическую связь в отношение, которое будет записано в итоговый результат анализа. Так, например, в предложении:

На сегодняшнем совещании речь шла о назначениях губернаторов,

у слова, описывающего ситуацию назначения, обнаруживается только один потенциальный участник ― именная группа *губернаторов*. В данном случае речь не идет о конкретной ситуации назначения, другие участники не определены, поэтому из данного предложения факт назначения извлечен не будет. Построение аннотации, соответствующей построению связи между словом-ситуацией и именной группой, будет запрещено на уровне правил. В общем случае, если один из участников ситуации выражен местоимением или если это эллиптическая конструкция, будет помечено слово-ситуация, обозначены потенциальные участники, но аннотации, которые строят отношения между словом-ситуацией и именной группой, построены не будут. Зато эти предварительные результаты можно будет использовать при дальнейшем анализе текста, отождествлении объектов и ситуаций, извлеченных на более ранних стадиях анализа.

5 Заключение

Предлагаемый подход извлечения фактов из текстов основывается на частичном синтаксическом анализе выделенных фрагментов текста под контролем ресурса знаний. Тестирование предлагаемого алгоритма извлечения фактов из неструктурированного текста показало значительное ускорение работы программы по сравнению с подходом, где извлечение было основано на поиске фрагментов текста, удовлетворяющих записанным на языке правил шаблонным конструкциям. Кроме того, процесс разработки и отладки алгоритма демонстрирует большую гибкость нового подхода, легкость дополнения и настраивания под новые задачи. Результаты анализа сохраняются в удобной форме, их можно использовать как основу для дальнейшей обработки текста.

Литература

[1] Александровский Д.А., Кормалев Д.А., Кормалева М.С., Куршев Е.П., Сулейманова Е.А., Трофимов И.В. Развитие средств аналитической обработки текста в системе ИСИДА-Т // Тр. Десятой нац. конф. по искусственному интеллекту с междунар. участием КИИ-2006, Обнинск, 25-28 сентября 2006 г.: В 3 т. — М.: Физматлит, 2006. — Т. 2. — С. 555—563.

[2] Гершензон Л. М., Ножов И. М., Панкратов Д. В. Система извлечения и поиска структурированной информации из больших текстовых массивов СМИ. Архитектурные и лингвистические особенности. Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог'2005» (Звенигород, 1–6 июня, 2005 г.)/ Под ред. И. М. Кобозевой,А. С. Нариньяни, В. П. Селегея. — М.: Наука, 2005.

[3] Ермаков А.Е. Автоматическое извлечение фактов из текстов досье: опыт установления анафорических связей. Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции "Диалог 2007". – Москва, Наука, 2007

[4] Ермаков А.Е. Извлечение знаний из текста и их обработка: состояние и перспективы. Информационные технологии 2009, № 7

[5] Киселев С.Л., Ермаков А.Е., Плешко В.В. Поиск фактов в тексте естественного языка на основе сетевых описаний. Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2004. — М.: Наука, 2004.

[6] Кормалев Д.А., Куршев Е.П., Сулейманова Е.А., Трофимов И.В. Архитектура инструментальных средств систем извлечения информации из текстов. Труды международной конференции "Программные системы: теория и приложения", Переславль-Залесский, М.: Физматлит, 2004, т.2, с.49—70

[7] Кормалев Д.А., Куршев Е.П., Сулейманова Е.А., Трофимов И.В. Извлечение данных из текста. Анализ ситуаций ньюсмейкинга. Труды Восьмой национальной конференции по искусственному интеллекту с международным участием КИИ-2002. Москва, Физматлит, 2002, с. 199-206

[8] Кормалев Д.А., Куршев Е.П., Сулейманова Е.А., Трофимов И.В.. Технология извлечения информации, из текстов, основанная на знаниях. Программные продукты и системы, 2009, №2

[9] Куршев Е.П., Сулейманова Е.А. Ресурсы предметных знаний в системах интеллектуального анализа текста // Тр. междунар. конф. «Программные системы: теория и приложения», ИПС РАН, Переславль-Залесский, октябрь 2006 г.: В 2 т. — М.: Физматлит, 2006. — Т.1. — С. 379—390.

[10] http://www.mlg.ru/

An approach to the automatic fact extraction from news texts on appointments and dismissals in texts

Natalia Vlasova

This work proposes an approach to the fact extraction in the rule-based system of information extraction ISIDA-T. The main idea is a search of “keywords” that describe the fact with the subsequent gathering of suitable noun groups around the founded “keywords”.