Т. В. Дзибук

АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ ФАКТОВ ИЗ ТЕКСТОВ ДОСЬЕ

В настоящее время для автоматической обработки больших неструктурированных текстовых массивов активно используются специализированные технологии *Information Extraction, Data Mining* и *Text Mining*. Разработанные на основе статистического и лингвистического анализа, а также методов искусственного интеллекта, указанные технологии предназначены для проведения смыслового анализа, поиска и извлечения различных понятий, фактов, описаний информационных объектов, мнений, отзывов об определенных объектах и т.п. По прогнозам аналитических компаний, спрос на подобные технологии и реализованные на их основе программные продукты существенно возрастет в течение ближайших 4–5 лет.

В данной статье описывается один из возможных подходов к моделированию автоматического извлечения фактов из текстов досье – автобиографий и резюме (заявок на работу). Необходимо отметить, что в основе любого процесса моделирования лежит процедура разработки формальной модели определенного типа. Формальная модель лингвистического явления представляет собой систему правил, описывающих и имитирующих структуру и/или поведение лингвистических объектов и позволяющую хотя бы частично воспроизвести эти объекты либо с помощью человека, либо с помощью компьютера. Для воспроизведения явления с помощью компьютера необходимо составить лингвистическую базу данных/знаний, описывающих определенное лингвистическое явление, и построить, опираясь на нее, алгоритм его функционирования.

Материалом исследования послужили двадцать текстов англоязычных автобиографий и двадцать текстов англоязычных резюме, взятых с разных сайтов сети Интернет. В ходе анализа отобранных текстов были выделены лексические и структурные маркеры, содержащие указание на определенную фактографическую информацию. Выделенные единицы были унифицированы и преобразованы в определенное число лексических и фразовых шаблонов. Поскольку тексты автобиографий и тексты резюме обладают некоторыми лингвистическими отличиями, проанализированные закономерности их организации нашли свое отражение, соответственно, в первой и второй частях лингвистической базы данных системы. С опорой на базу данных были построены два варианта формальной модели, отражающие особенности процесса извлечения фактов из текстов автобиографий и резюме. Несмотря на то, что тексты содержат практически идентичную информацию об их авторе, для извлечения фактов из документов необходимо использовать разные алгоритмы, поскольку, как отмечалось выше, способы представления информации в них не являются полностью аналогичными.

Процесс извлечения фактов из текстов досье заданного типа состоит из следующих этапов: ввода анализируемого текста досье, ввода запроса на поиск и извлечение фактов, просмотра результатов обработки запроса, ввода дополнительных поисковых запросов и просмотр дополнительных резуль-

татов поиска. После завершения процесса ввода текста автобиографии или текста резюме компьютер информирует пользователя о необходимости задания с помощью клавиатуры следующих параметров поиска: *Which information do you need?* (ключевые слова *name, age, birth, marital status, relatives, education, degree, career* для автобиографии и ключевые слова *name, age, birth, marital status, contacts, education, degree, career* для текста резюме). За каждым элементом поискового запроса закреплена определенная последовательность правил поиска фактов в тексте документа. При этом осуществляется поиск строк текста, содержащих соответствующие элементам поискового запроса лексические и фразовые шаблоны. Система поочередно анализирует каждую строку текста. Если в какой-либо строке текста обнаружен заданный шаблон, происходит извлечение конкретной фактографической информации. Такая информация, как имя автора, извлекается один раз, после чего осуществляется автоматический переход к следующему слову поискового запроса. Некоторая информация, например, конкретные факты об учебе и профессиональной деятельности автора текста часто находится в нескольких предложениях, которые могут располагаться в тексте в произвольном порядке. В таких случаях для качественного извлечения фактов необходимо проанализировать весь текст документа. Если все строки текста были проанализированы, но ни один шаблон не был обнаружен, происходит вывод результата поиска типа *Education: no information.* Если результаты поиска и извлечения фактов не соответствуют ожиданиям пользователя, существует возможность ввода дополнительных ключевых слов для организации нового поиска фактов по тексту документа. На данном этапе ключевые слова не закреплены за последовательностью правил, описывающих поиск и извлечение фактов определенного типа. Поэтому поиск строк, которые могут соответствовать введенному ключевому слову или группе слов, осуществляется по специальному алгоритму. По требованию пользователя данная функция может быть выполнена произвольное количество раз.

Одной из общих характеристик принципа поиска фактов в текстах автобиографий и заявок на работу является то, что при указании конкретных шаблонов в ходе составления алгоритма необходимо учитывать пробелы перед словами и после них, так как отдельные слова (особенно служебные части речи) могут совпадать с частями других слов. Допустим, если необходимо извлечь фрагмент строки от предлога *in* до конца строки, необходимо оформить условие, в котором перед предлогом и после него указаны пробелы. В противном случае система извлечет все слова, в которых есть последовательность символов *in*, например, *My name is* ***Jin*** *Flem****in****gs, and I was born* ***in****...* и т.п.

Для извлечения имени автора текста автобиографии производится поиск фразового шаблона *My name is* и выделение следующей за шаблоном части предложения до первого знака препинания. В большинстве случаев таким знаком препинания является точка. Однако автор текста может привести в одном предложении разнородные факты, вследствие чего искомая информация будет располагаться во фрагменте предложения от шаблона до первой запятой. Поэтому из шаблона, представленного в лингвистической базе

данных, была удалена первая буква (в зависимости от контекста она может быть строчной либо прописной). Для извлечения факта о возрасте автора текста используется лексический шаблон *years old*. С учетом того, что данный шаблон может быть обнаружен в предложениях, содержащих информацию не только об авторе текста (например, *My sister is 11 years old*), релевантной может быть лишь та строка, которая содержит местоимение *I* (*I am 20 years old*). Если возраст выражен числом, достаточно извлечь число, стоящее перед шаблоном. Однако возраст может быть указан словами, причем количество символов в словах заранее предугадать невозможно. В обоих случаях извлекаемая информация расположена между личным местоимением, следующим за ним вспомогательным глаголом и указанным выше шаблоном (***I am*** *twenty* ***years old***). Как правило, возраст автора в тексте автобиографии указан в самом начале документа, а последующие строки с аналогичным шаблоном употреблены для указания возраста ближайших родственников. Если пользователь хочет извлечь факт, касающийся возраста кого-либо из родственников автора текста, необходимо либо указать в поисковом запросе ключевое слово *relatives*, либо воспользоваться функцией дополнительного поиска. В алгоритме предусмотрено дополнительное правило поиска и извлечения факта о возрасте автора. Если перед лексическим шаблоном *years old* не было обнаружено личное местоимение, осуществляется поиск фразового шаблона *My age is* (*My age is 20 years old*). Для извлечения факта о дате и месте рождения автора текста используется лексический шаблон *born*. Искомой информацией является фрагмент строки текста от шаблона до конца предложения (в данном случае последовательность и характер используемых знаков препинания может варьироваться). Так как данный шаблон может быть употреблен в контексте, когда речь идет о других лицах, необходимо найти в этой же строке относящееся к нему личное местоимение *I*. Система не разделяет факты о дате и месте рождения и выдает один конечный результат, например, *Born: in 1991 in Berlin.* Как правило, дата рождения вводится предлогом *on*, а место рождения – предлогом *in*. Однако возможны варианты типа *I was born in 1991*, создающие дополнительные трудности при составлении правил поиска и извлечения фактов. Если необходимо извлечь факт о семейном положении автора, система производит поиск строк текста, содержащих определенные шаблоны. Чтобы убедиться, что они относятся непосредственно к автору текста, осуществляется поиск личного местоимения *I*. О факте *not married* свидетельствуют лексические единицы *not married, divorced, single, widowed.* Если в тексте документа встречается лексический шаблон *married*, необходимо проверить наличие перед ним отрицательной частицы *not*, влияющей на смысл извлекаемого факта. В ряде исследованных текстов их авторы указывают информацию о ближайших родственниках. Невозможно заранее предугадать степень детальности данной информации и значения приведенных в тексте фактов. Поэтому если необходимо найти информацию о родственниках автора, осуществляется поиск лексических шаблонов *mother, father, sister, daughter, wife* и т.д. Факт о получении образования представлен в тексте автобиографии соответствующими периодами. Система определяет

факты, относящиеся к разным этапам обучения, опираясь на лексические шаблоны *school, college, university*. Если возникает потребность узнать информацию о факте *kindergarten*, можно воспользоваться функцией дополнительного поиска фактов. Учебное заведение может быть не только местом учебы, но и местом работы. Для того чтобы отнести учебное заведение к месту учебы, необходимо найти рядом с наименованием учебного заведения такие глаголы (маркеры), как *study, go to, attend, graduate*. Заранее невозможно предугадать, на каком этапе обучения находится автор текста: планирует поступать в учебное заведение, является студентом или выпускником. В зависимости от временного интервала процесса обучения, автор текста может употреблять глаголы в различных временных формах, например, *I study at school/I studied at school*. В связи с этим осуществляется поиск не полной глагольной формы, а ее неизменяемой части. В итоге, данный факт имеет следующий вид – *тип учебного заведения:* строка текста, содержащая маркеры. Для извлечения из текста такого факта, как наличие у автора автобиографии научной степени, в поисковом запросе необходимо указать ключевое слово *degree*. Важная информация может находиться как перед аналогичным лексическим шаблоном, так и после него, причем нередко – одновременно в двух позициях. Для наиболее точного извлечения факта о научной степени строка, содержащая данный шаблон, выделяется из текста целиком. Поиск фактов, связанных с профессиональной деятельностью автора текста, осуществляется на основе лексических шаблонов *company, firm, office, job, position*. Место работы автора текста может также являться учебным заведением (*school, college, university)*. Если в строке текста с одним из вышеуказанных слов обнаружена форма глагола *to work* и личное местоимение *I*, значение факта о профессиональной деятельности относится к автору текста. Трудовая деятельность может быть описана несколькими предложениями (в том числе, расположенными дистантно). Для накопления таких предложений используется лексический шаблон *career*. Названия таких шаблонов, как *Professional Experience*, *Education* и т.д. могут варьироваться.

В тексте резюме поиск имени автора документа осуществляется на основе структурного шаблона *Name*, который должен располагаться в начале строки текста, содержащей данную информацию. Такая строка может быть извлечена целиком без изменений. В ряде исследованных текстов шаблон *Name* отсутствует, однако имя автора текста всегда занимает первую строку документа. Контактные данные находятся в строках текста со следующими структурными шаблонами: *Address, Phone, e-mail*. Если шаблон находится в начале строки, достаточно извлечь всю строку. Перед шаблоном *Phone* могут находиться дополнительные лексические шаблоны (*cell/phone, mobile/phone, office/phone*). Подобные шаблоны не являются релевантными при составлении правил извлечения фактов. Поэтому достаточно указать в правиле шаблон *Phone* и при положительном результате поиска извлечь всю строку целиком. Таким образом, дополнительные шаблоны будут сохранены, а правило извлечения факта будет более компактным и универсальным. При извлечении электронного адреса используется дополнительное правило поиска с учетом наличия символа @ в каждой строке текста. Когда контактная

информация не сопровождается соответствующими структурными шаблонами, а представлена в верхней части документа сплошным текстом, при извлечении фактов возникают определенные трудности. При этом невозможно не только отделить факты друг от друга, но и выделить данный фрагмент текста без использования правил его семантического анализа. В такой ситуации из текста можно извлечь все строки, расположенные после структурного шаблона, следующего за личной и контактной информацией. Поиск фактов о профессиональной деятельности автора заявки на работу осуществляется на основе одного из следующих структурных шаблонов: *Professional Experience, Employment History, Work Experience, Construction Career Summary, Relevant Experience.* В шаблоне представлен период работы, наименование организации работодателя, должность автора текста и исполняемые им обязанности. Следовательно, при извлечении факта о карьере автора резюме из текста можно выделить весь фрагмент, соответствующий данному шаблону. Таким образом, чтобы извлечь все факты о профессиональной деятельности автора, необходимо выделить все строки текста, расположенные между шаблоном, указывающим на карьеру, и следующим шаблоном. В большинстве случаев, следующий фрагмент текста содержит факт, указывающий на уровень образования. Факт, отражающий информацию об образовании автора заявки на работу, представлен аналогичным образом. Информация о каждом учебном заведении включает в себя период обучения (полный период обучения либо год окончания учебного заведения), наименование учебного заведения и достигнутые результаты (полученная квалификация, ученая степень и т.д.). В данном случае используются следующие варианты структурного шаблона: *Education, Education and Further Training, Additional Courses* и т.д. Документ может содержать несколько фактов об образовании, например, факт о получении основного образовании и факт о получении дополнительном образовании. Если важным является поиск и извлечение фактов об определенной должности, занимаемой автором, или о конкретном учебном заведении, в котором он обучался, можно воспользоваться функцией дополнительного поиска. Информация о научной степени автора (если такой факт содержится в тексте резюме) с большой вероятностью будет находиться в одной строке с наименованием учебного заведения. Тем не менее, существует возможность извлечь данный факт с опорой на структурный шаблон *Degree* (как и в тексте автобиографии).

Для проведения компьютерного эксперимента на основе двух вариантов формальной модели были написаны два программных кода (язык программирования *Python*). Программные продукты функционируют в командной строке среды *Idle 3.4.3*. Фрагменты результатов работы программ представлены ниже на рис. 1–2.

Анализ результатов работы первой программы (автоматическое извлечение фактов из текстов англоязычных автобиографий) показал, что компьютер может успешно извлекать факты из текста на основе лексических и фразовых шаблонов, представленных в первой части лингвистической базы данных. Результаты работы второй программы (автоматическое извлечение фактов из текстов резюме) также можно признать достаточно успешными.

Исключение составляют случаи, когда некоторые факты, в частности, контактная информация, не представлены в тексте лексическими единицами, совпадающими со структурными шаблонами. Компьютер не может разделить подобные факты без детального семантического анализа текста. По окончании компьютерного эксперимента были сделаны следующие выводы о возможностях совершенствования предложенной формальной модели.

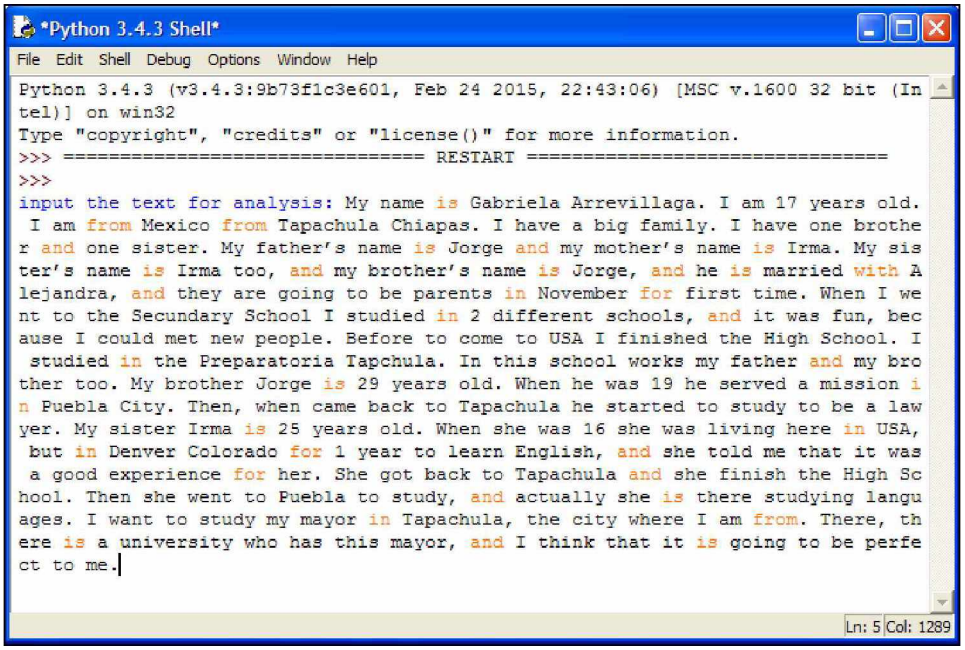

Рис. 1. Рабочее окно программы с текстом автобиографии № 1

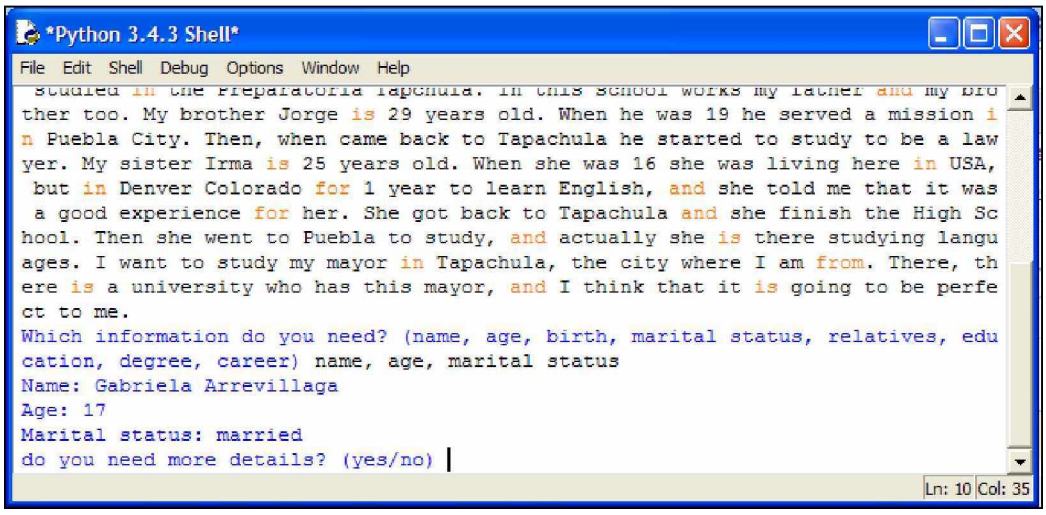

Рис. 2. Поисковый запрос и результаты извлечения фактов из текста автобиографии № 1

1. Для улучшения качества процесса извлечения фактов необходимо разработать правила поиска дат, а также правила, позволяющие определять, к каким фактам относятся найденные даты.

2. Некоторые модули системы должны осуществлять более детальный поиск и извлечение фактов (например, поиск фактов о дате и месте рождения, образовании, профессиональной деятельности). Так, в рамках фор-

мальной модели факт об учебном заведении, которое окончил автор текста, будет извлечен по образцу: *University: From 1990 till 1994 I studied Physics in the University of Physics*. При использовании дополнительных правил извлечения фактов возможно получение искомого результата следующего вида:

University: University of Physics;
Years of studying: 1990 – 1994;
Major: Physics.

3. Если пользователь работает с большим количеством текстов автобиографий и резюме, для сохранения извлеченных фактов можно создать фактографическую базу данных. В дальнейшем по базе данных можно организовать дополнительный автоматический поиск фактов, ее можно постоянно обновлять. Допустим, авторы двух автобиографий работали или работают в одной организации. Если в одном из текстов указан адрес этой организации, данный факт может быть автоматически скопирован в массив фактов о втором лице. Кроме того, если в базе данных содержатся факты, извлеченные из текста автобиографии определенного лица, и затем происходит поиск фактов в тексте заявки на работу, написанного тем же лицом, обнаруженные в первом случае факты могут быть дополнены фактами, извлеченными из второго документа. Используя базу данных, можно проводить поиск и извлечение фактов не только об определенном лице или лицах, но и о конкретной организации, жителях определенного города и т.д.

Представленная в статье формальная модель может быть использована для поиска и извлечения фактов из текстов досье, связанных с жизнью и деятельностью конкретной личности либо группы лиц; автоматизации документооборота в отделах кадров; формирования и ведения тематических досье; проведения информационной разведки; отслеживания тенденций и различных показателей в обществе и т.д.