ЭКСТРАКЦИЯ ФАКТОВ ИЗ СЛАБОСТРУКТУРИРОВАННОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ

Нина Хайрова, Наталья Шаронова, Аджит Пратап Сингх Гаутам

Аннотация *Извлечение фактов из текстов представляет собой одно из центральных направлений Natural Language Processing. Большинство существующих подходов позволяет извлекать факты из хорошо структурированных текстов узкой тематической направленности, тогда как наибольший интерес представляет возможность автоматизации излечения фактической информации из слабоструктурированных текстов неограниченных предметных областей. Факт, представляющий собой фиксацию некоторого отношения сущностей в предложении, можно записать в виде триплета: Subject - Predicate - Object, в котором предикат представляет отношение, а субъект и объект определяют два предмета или понятия. В работе предлагается строгая модель, связывающая смысловые отношения между сущностями с элементами поверхностной структуры предложений естественного языка. Для формализации и явного представления средствами поверхностной структуры партиципантов триплета факта, называемого предложением английского языка, выделены и описаны предметными переменными конечные множества синтаксических и морфологических категорий. В статье рассмотрены три типа фактов и их атрибутов. Разработана программная имплементация полученной модели, предложена метрика формальной оценки эффективности технологии экстракции фактов из слабоструктурированной текстовой информации и обоснован объем экспериментальной выборки текстов, позволившей подтвердить достоверность полученной точности и полноты извлекаемых фактов.*

Ключевые слова: *извлечение фактов из текста, Natural Language Processing, семантические отношения, алгебра конечных предикатов, полнота и точность.*

ACM Classification Keywords*: H.3.3 .Information Search and Retrieval, I.2.4. Knowledge Representation Formalisms and Methods*

Введение

Извлечение фактов из текстов представляет собой одно из центральных направлений автоматической обработки естественного языка — Natural Language Processing (NLP). Существует достаточное количество исследований в данном направлении, но пока не появились

достаточно надежные систем, извлекающих факты из слабоструктурированных текстов [Хайрова, 2015]. Традиционно используемые подходы: поиск по шаблону, поиск опорного элемента, поиск по онтологии и т.д. имеют как свои преимущества, так и недостатки. Но все они работают только на заранее определенных предметных областях, ограничивающих тематику исследуемых текстов, и требуют хорошо структурированных текстов (патентов, библиографических описаний, авторефератов и т.д.). В то же время подавляющая часть текстовой информации, представленной в компьютерных сетях, – это не структурированные и слабоструктурированные тексты различной тематической направленности.

Общая постановка задачи

Факты представляют собой классифицированные, зафиксированные и произошедшие события. Субъектами фактов, как правило, являются сущности, объекты или темы, которые обладают дополнительными выделенными свойствами (временными, пространственными, качественными, количественными и т.д.) [Andersen, 1994]. При этом факт может быть извлечен из текстовой информации (как слабо структурированной, так и не структурированной) и может определять как свойства объекта, так и связь объекта с другими объектами.

Согласно «Логико-философскому трактату» [Витгенштейн, 2005] мир подразделяется на факты, любой факт — это фиксация некоторого отношения. Все факты фиксируются фразами, а структура любого предложения включает несколько каким-то образом связанных объектов, например, элементарное предложение связывает два объекта. Таким образом, для получения фактической картины мира необходимо построить модель, связывающую лингвистические элементы текста с их содержательной формой. При этом, факты выделяются из предложений, содержащих упоминание сущности или анафорические ссылки на нее.

Структурное описание модели

Для излечения связей между определенными понятиями в тексте необходимо выделить семантические (или понятийные) связи в предложении. Для этого необходимо разработать строгую модель, связывающую информацию, содержащуюся в определении смысловых связей с элементами поверхностной структуры предложений естественного языка.

Для формализации и явного представления средствами поверхностной структуры субъекта и объекта триплета факта *Subject→Predicate→Object*, называемого предложением английского языка, выделены и описаны предметными переменными следующие конечные множества синтаксических и морфологических категорий:

$$z^{to} \vee z^{by} \vee z^{with} \vee z^{about} \vee z^{of} \vee z^{on} \vee z^{at} \vee z^{in} \vee z^{out} = 1,$$

$$y^{ap} \vee y^{aps} \vee y^{out} = 1, \; x^{f} \vee x^{l} \vee x^{kos} = 1,$$

$$m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were} \vee m^{out} = 1,$$

$$p^{III} \vee p^{ed} \vee p^{I} \vee p^{ing} \vee p^{II} = 1,$$

где z — предметная переменная, определяющая синтаксические характеристики наличия (*to*, *by*, *with*, *about*, *of*, *on*, *at*, *in*) или отсутствия (*out*) предлога в английской фразе; y — предметная переменная, определяющая наличие (*ap*, *aps*) или отсутствие (*out*) апострофа в конце слова; x — предметная переменная, определяющая позицию существительного перед (*f*), после (*l*) личным глаголом или после непрямого дополнения (*kos*); m — предметная переменная, определяющая существование любой формы глагола "*to be*" (*is*, *are*, *havb*, *hasb*, *hadb*, *was*, *were*, *out*); p — предметная переменная, определяющая форму основного глагола (*III*, *ed*, *I*, *ing*, *II*).

Семантическое значение участников действия, называемых словами предложения, определяется предикатом [Хайрова, 2015]

$$P(x, y, z, m, p) \rightarrow P(x) \wedge P(y) \wedge P(z) \wedge P(m) \wedge P(p).$$

В конъюнкции предикатов, описывающей взаимосвязь грамматических характеристик слов, предикат γ_k исключает часть связей поверхностной структуры, не присущих сущностям триплета факта

$$P(x, y, z, m, p) = \gamma_k(x, y, z, m, p) \wedge P(x) \wedge P(y) \wedge P(z) \wedge P(m) \wedge P(p).$$

где $k \in [1; h]$, h – число рассматриваемых в системе фактов. Предикат γ_k принимает значение 1, если комплекс выбранных характеристик для n-ой фразы формирует некоторое семантическое значение участника триплета, и значение 0 в противном случае.

Рассмотрено несколько профильных типов фактов: 1) утверждения об обладании (или принадлежности) некоторой сущности субъекта некоторой сущностью объекта (рис. 1); 2) утверждение о перемещении субъектом объекта; 3) утверждение о потере (продаже) некоторого объекта некоторым субъектом; а также факты-атрибуты трех вышеупомянутых типов фактов — времени действия, локации действия и иерархической принадлежности субъекта или объекта действия факта иной сущности.

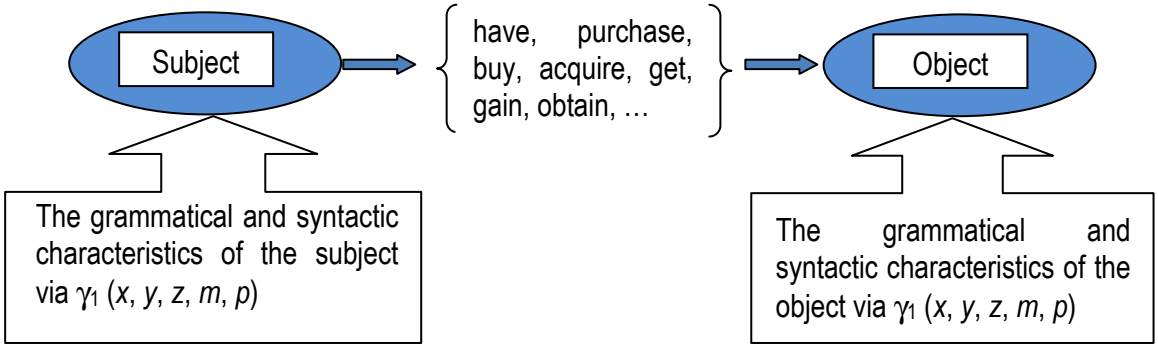

Рис. 1. Структурная схема идентификации факта принадлежности (собственности).

Предикат γ_1 (x, y, z, m, p) определяет грамматические и синтаксические характеристики *Subject* триплета факта:

$$\gamma_1(x, y, z, m, p) = z^{out}\, y^{out}\, x^{f}\, m^{out}\, p^{I} \vee z^{out}\, y^{out}\, x^{f}\, m^{out}\, p^{II} \vee z^{out}\, y^{out}\, x^{f}\, m^{out} p^{ed} \vee z^{by}\, y^{out}\, x^{I}\, p^{ed}\,(m^{is} \vee m^{are} \vee \vee\, m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were}) \vee z^{by}\, y^{out}\, x^{I}\, p^{III}\,(m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were}). \quad (1)$$

Предикат γ_2 (x, y, z, m, p) определяет грамматические и синтаксические характеристики объекта факта:

$$\gamma_2(x, y, z, m, p) = z^{out}\, y^{out}\, x^{I}\, m^{out}\, p^{I} \vee z^{out}\, y^{out}\, x^{I}\, m^{out} p^{ed} \vee z^{out}\, y^{out}\, x^{I}\, m^{out} p^{II} \vee \vee z^{out}\, y^{out}\, x^{f} p^{III}\,(m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were}) \vee z^{out}\, y^{out}\, x^{f} p^{ed}\,(m^{is} \vee \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were}). \quad (2)$$

Предикат γ_3 (x, y, z, m, p) определяет грамматические характеристики атрибута времени действия факта:

$$\gamma_3(x, y, z, m, p) = (z^{on} x^{kos}\, y^{out} \vee z^{in} x^{kos}\, y^{out} \vee z^{at} x^{kos}\,(p^{III} \vee p^{ed} \vee p^{I} \vee p^{ing} \vee \vee p^{II})\ (m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were} \vee m^{out}). \quad (3)$$

Предикат γ_4 (x, y, z, m, p) определяет грамматические характеристики атрибута иерархической принадлежности субъекта или объекта действия факта иной сущности:

$$\gamma_4(x, y, z, m, p) = z^{out}\, x^{f}(y^{ap} \vee y^{aps})\,(m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee \vee m^{was} \vee m^{were} \vee m^{out})\,(p^{III} \vee p^{ed} \vee p^{I} \vee p^{ing} \vee p^{II}). \quad (4)$$

Предикаты γ_1 и γ_2 определяют первый тип фактов – факты описывающие связь двух сущностей, при этом одна из сущностей будет определяться как субъект, а вторая как объект предикатного действия. Например, "*the company had revenue*" (субъект: *company*, объект: *revenue*, предикат: *had*). Предикаты γ_3 и γ_4 определяют второй тип фактов – факты представляет собой триплет: предмет – атрибут – значение, где предмет – это объект, о котором фиксируется факт, атрибут – некоторое именованное, заранее определенное свойство, а значение представляет собой некоторое значение, область определения которого может быть в некоторых случаях известна. Например, это могут быть факты атрибутов места и времени осуществления некоторого действия.

Примеры идентификации фактов в английских текстах

Рассмотрим несколько примеров извлечения фактов из предложений английского языка. В предложении:

"The company bought back the business from OTIV",

с помощью глагола *bought* $\rightarrow$ *buy* выделяется факт, относящий к фактам наличия или приобретения некоторого объекта некоторым субъектом. Согласно уравнению (1) существительное "*company*" определяется как субъект данного факта. Грамматические и синтаксические характеристики данного существительного соответствуют конъюнкции

$$\gamma'_1 (x, y, z, m, p) = z^{out} y^{out} x^{f} m^{out} p^{II}.$$

Согласно уравнению (2) существительное "*business*" определяется как объект данного факта. Грамматические и синтаксические характеристики данного существительного соответствуют конъюнкции

$$\gamma'_2 (x, y, z, m, p) = z^{out} y^{out} x^{I} m^{out} p^{II}.$$

Второе рассматриваемое предложение:

"The companies' shares were sold by the investor on Tuesday ",

сообщает факт отсутствия или потери некоторого объекта. Факт определяется глаголом *sold* $\rightarrow$ *sell*. Согласно уравнению (1) существительное "*investor*" определяется как субъект данного факта. Грамматические и синтаксические характеристики данного существительного соответствуют конъюнкции

$$\gamma''_1 (x, y, z, m, p) = z^{by} y^{out} x^{I} p^{III} m^{were}.$$

Согласно уравнению (2) существительное " *shares* " ($\rightarrow$*share*) определяется как объект данного факта. Грамматические и синтаксические характеристики данного существительного соответствуют конъюнкции

$$\gamma''_2 (x, y, z, m, p) = z^{out} y^{out} x^{f} p^{III} m^{were}.$$

Слово "*Tuesday*" определяется как объект атрибута времени факта продажи. Слово выделяется с помощью конъюнкции предиката (3)

$$\gamma''_3 (x, y, z, m, p) = z^{on} x^{kos} y^{out} p^{III} m^{were}.$$

Слово " *companies'* " определяется как объект принадлежности субъекта факта продажи. Слово выделяется с помощью конъюнкции предиката (4)

$$\gamma''_4 (x, y, z, m, p) = z^{out} x^{f} y^{aps} p^{III} m^{were}.$$

Имплементация технологии в Web-приложение

Разработанная технология была имплементирована в Web-приложение, которое обрабатывает текст или тексты выбранных файлов. Извлеченная фактическая информация представляется в окне (рис. 2), отображающем: (1) исходное предложение, из которого экстрагируется факт; (2) предикат факта, в виде инфинитивной формы глагола; (3) субъект и объект факта, в

канонической форме существительного; (4) возможные атрибуты факта – время, место действия, принадлежность субъекта или объекта.

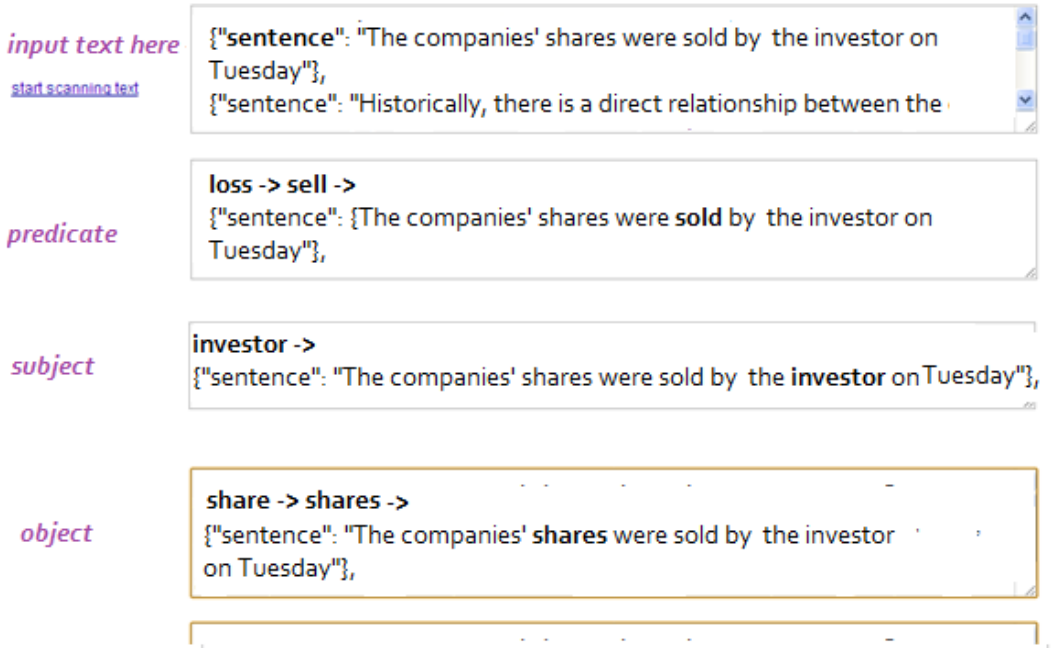

Рис. 2. Пример извлеченния фактической информации

В результате работы разработанного приложения отображаются: (1) исходное предложение, (2) предикат и тип предиката факта, (3) субъект и объект факта, (4) атрибуты времени, места действия и принадлежности субъекта или объекта факт.

Приложение определяет профильные типы фактов анализируемых объектов:

– факт наличия или приобретения некоторого объекта некоторым субъектом; в английских предложениях факты данного типа базируются на предикате, определяемом предварительно выделенным множеством глаголов $V_{possess}$ = {*possess*, *have*, *purchase*, *buy*, *acquire*, *get*, *gain*, *obtain*};

– факт отсутствия некоторого объекта у некоторого субъекта; в английских предложениях факты данного типа базируются на предикате, определяемом предварительно выделенным множеством глаголов $V_{lacking}$ = {*sell*, *market*, *realize*, *forfeit*, *lose*, … };

– факт перемещения некоторого объекта некоторым субъектом; в английских предложениях факты данного типа базируются на предикате, определяемом предварительно выделенным множеством глаголов $V_{transfer}$ = { *move*, *relocate*, *displace*, *transport*, *transfer*, …}

– факт атрибута времени, отображающий темпоральную характеристику события, позволяющую определить дату или некоторые временные характеристики (день недели, месяц, год) и их комбинации;

– факт атрибута места расположения (географическое расположение с указанием иерархии географического включения);

– атрибут принадлежности субъекта или объекта участников факта;

– факт атрибута, характеризующего деятельность организации (сферу и продукт ее деятельности).

Формальная оценка эффективности технологии экстракции фактов из слабоструктурированной текстовой информации

Для оценки эффективности технологии идентификации знаний из текстового контента необходимо определить метрики, представляющие объективно измеряемые показатели деятельности пользователей до и после внедрения предложенной технологии. На сегодняшний день не существует подобных стандартных метрик, позволяющих измерить качество и эффективность технологий идентификации и экстракции знаний, извлеченных из текстовых массивов [Хайрова, 2014]. В общем случае для оценки эффективности информационной обработки текстов, например классификации, кластеризации или информационного поиска, а также методов Text Mining, Opinion Mining, Web Mining, используется метод тестовых коллекций [Шабанов, 2003; Cormack, 1998]. При этом возникают дополнительные проблемы, связанные с субъективностью мнения эксперта, отсутствием общепринятого определения понятия “качества знания”, а также достоверностью полученного результата для всех рассматриваемых текстов.

Для определения объема экспериментально исследуемых текстов используется метод математической теории выборки, определяющий механизмы формирования репрезентативной выборки, изучение которой позволяет получить информацию о генеральной совокупности, из которой она была выбрана. Оценивая долю признака в генеральной совокупности, которая показывает отношение числа интересующих нас фактов к общему числу фактов в данной коллекции документов, по соответствующей доле признака в выборке, по соотношениям для выборок достаточно большого объема (> 20 элементов), задавшись допустимым для лингвистических исследований значением $Z_{0,05}$ = 1,96 и допустимой величиной предельной ошибки $e=0,05$, определяем искомый объем репрезентативной выборки N_{MAX} = 385 фактов, [Хайрова, 2014].

Определяя факт в виде триплета: *Subject* → *Predicate* → *Object*, центральной частью которого является действие, определяемое личным глаголом предложения, а сущности субъекта и объекта определяются существительными актантами данного действия, мы обозначаем возможность выражения факта одним простым предложением. Данное утверждение подтверждается лингвистами, утверждающими, что пропозициональная структура предложения служит способом кодирования языковыми средствами информации о конкретных положениях дел в действительности, т.е. о конкретных фактах и событиях. Этому служат конструктивные схемы с позициями для предиката и актантов [Сусов, 2006]. Так, например, согласно

С. Д. Кацнельсону [Кацнельсон, 1972]: “Непосредственной реальностью мира являются процессы, события, факты, данные в их пространственных и временных границах…; их отображает речь в своих предложениях”. Таким образом, каждое базовое предложение английского языка, представляющее некоторое высказывание, за исключением повелительных предложений, типа “Like sit!”, “And go!”, выражает тот или иной факт.

Обычно в английском языке используются достаточно короткие предложения. Martin Cutts в “Oxford Guide to Plain English” [Cutts, 2010] говорит о максимальной длине понимаемого предложения в 20 слов. В среднем длина предложения определяется как $10 \leq n \leq 20$, где n – количество слов в предложении. В научных текстах наблюдается и допускается большая длина предложений до 25-30 слов [Moore, 2011], но существуют исследования, показывающие уменьшение читабельности и уменьшения понимания таких текстов [Zinsser, 2016].

Кроме того, проведенный анализ показывает, что в деловых предложениях английского языка используются короткие слова. При максимальной длине предложения в 20 слов, средняя длина слова определяется как 5,67 символов [Ku, 2003], тогда средняя длина предложения определяется как 75-100 символов.

Используя знания о средней длине предложения, мы можем определить размер текстовой выборки, позволяющей осуществить достоверное оценивание результата эффективности разработанной информационной технологии:

$$\text{Length}_{\text{byte}} \approx 365^* 100+10^*100 = 37\,500$$

Таким образом, расчеты показывают, что превосходящий объем текста (ACCII кодировки), представляющего репрезентативную выборку предложений, обрабатываемых информационной системой, с учетом вводных и повелительных предложений должен быть приблизительно равен 38 кБайт. При этом, следует заметить, что объем выбран с большим превосходящим запасом и явным преувеличением количества повелительных предложений и междометий в деловых бизнес текстах.

В качестве метрики оценки эффективности разработанной технологии будем использовать интегральные показатели оценки качества излеченных из разнородных электронных источников знаний [Кураленок, 2002], основанные на показателях количественной оценки эффективности поиска, утвержденные межгосударственным стандартом по информации, библиотечному и издательскому делу. Такими показателями являются: коэффициент точности — *precision*, коэффициент полноты — *recall*.

Для вычисления данных коэффициентов, после проведения эксперимента любой определенный программным приложением факт, представляющий собой триплет (Subject $\rightarrow$ Predicate $\rightarrow$ Object), определим как принадлежаший к одному из классов:

1) правильно определенные приложением корректные факты;

2) неправильно определенные приложением некорректные факты;

3) корректные факты, содержащиеся в анализируемом тексте, который не определены приложением.

При этом, факт считается идентифицированным программным приложением корректно, если все элементы триплета (предикат, субъект и объект) определены правильно. Коэффициенты полноты и точности рассчитываются на базе значений следующих параметров:

– n_{yy} – число правильно идентифицированных программным приложением фактов;

– n_{yn} — число идентифицированных программным приложением некорректных фактов (фактов, которые определены экспертом как некорректные);

– n_{ny} — число фактов, которые остались не идентифицированными программным приложением.

Используя выше приведенные параметры, коэффициенты полноты и точности работы программы определяются по формулам:

$$precision = n_{yy} /(n_{yy} + n_{yn}),$$

$$recall = n_{yy} /(n_{yy} + n_{ny}).$$

Исследовалось около 400 предложений для каждого типа фактов: (1) факты наличия некоторого объекта у некоторого субъекта, (2) факты отсутствия некоторого объекта у некоторого субъекта, (3) факты перемещения некоторого объекта некоторым субъектом, (4) факты атрибута времени, (5) факты атрибута места расположения, (6) атрибуты принадлежности субъекта или объекта факта, (7) факты характеризующие деятельность корпорации или организации. Результаты проведенного эксперимента показаны в таблице 1

Таблица 1. Результаты расчетов коэффициентов точности и полноты

	Facts of the lack of…	Facts of possession of …	Facts of a displacement	Attributes of time	Attributes of location	Attributes of belonging	Attributes of industry
recall	0,92	0,91	0,91	0,97	0,96	0,95	0,94
precision	0,81	0,79	0,76	0,90	0,90	0,89	0,84

Средний коэффициент полноты, определяемый отношением числа правильно идентифицированных программным приложением фактов к общему числу корректных фактов данного типа, представленных в тексте, recall = 0,94

Средний коэффициент точности, определяемый отношением числа правильно идентифицированных программным приложением фактов к общему числу определенных системой фактов (как корректных так и не верных), precision = 0,84

Выводы

Разработанная логико-лингвистическая модель позволяет извлекать факты из слабоструктурированных текстов неограниченных предметных областей. В модели смысловые связи между сущностями выражаются через поверхностные характеристики партиципантов предложений английского языка. Рассматривались следующие грамматические характеристики: наличие после главного глагола конкретного предлога, наличие или отсутствие апострофа в конце партиципанта, позицию партиципанта по отношению к главному глаголу, наличие глагола “to be”, формы основного глагола. Были введены предикаты: субъекта и объекта трех типов фактов: факт наличия или приобретения некоторого объекта некоторым субъектом; факт отсутствия некоторого объекта у некоторого субъекта и факт перемещения некоторого объекта некоторым субъектом, а также предикаты атрибута времени, атрибута места расположения и атрибута принадлежности субъекта или объекта участников факта. Эффективность имплементированной в программное приложение модели оценивалась с помощью показателей полноты и точности. Для определения объема экспериментальной выборки исследуемых предложений использовался метод математической теории выборки, позволивший оценить объем текстов, экспериментальное исследование которых подтверждают достоверность построенной модели. Полученные средние коэффициенты полноты recall = 0,94 и точности precision = 0,84 выборки всех рассмотренных типов фактов выше, чем в аналогичных системах, работающих с неструктурированной текстовой информацией неограниченных предметных областей.

Литература:

[Andersen, 1994] Andersen, P. M., Huettner, A. K. Knowledge engineering for the JASPER fact extraction system. Integrated Computer-Aided Engineering. – 1 (6), 1994. – 473–493.

[Cormack, 1998] Cormack G.V. A Efficient construction of large test collections // G. V. Cormack , C. R. Palmer , C. L. Clarke // Proc. of the SIGIR'98 — P. 282—289.

[Cutts, 2010] Martin Cutts. Oxford guide to plain English. Oxford University Press, USA. – 2010.– 272 p.

[Ku, 2003] Anne Ku. The joys and pains of writing and editing [Электронный ресурс] // Le Bon Journal. 2003. Vol. 2. Issue Режим доступа: http://www.bonjournal.com/volume2/issue1writing.pdf

[Moore, 2011] Andrew Moore. The Long Sentence: A Disservice to Science in the Internet Age. – Bioessays, Vol. 33, No. 12. (2011), pp. 193-193

[Zinsser, 2016] William Zinsser. Writing Well / Harper Perennial. – 2016.– 336 p

[Витгенштейн, 2005] Людвиг Витгенштейн. Избранные работы / Пер. с нем. и англ. В. Руднева. М.: Издательский дом «Территория будущего», 2005. — с. 440

[Кацнельсон, 1972] Кацнельсон С.Д. Типология языка и речевое мышление. Ленинград: Наука, 1972. – 213 с.

[Кураленок, 2002] Кураленок И. Оценка систем текстового поиска/ И. Кураленок , И. Некрестьянов // Программирование. — 2002. — N 28(4). — С.226–242.

[Сусов, 2006] Сусов И. П. Введение в теоретическое языкознание М.: Восток - Запад, 2006. 382 с.

[Хайрова, 2014] Хайрова Н., Шаронова Н., Узлов Д. Решение проблемы формальной оценки эффективности технологий идентификации знаний в слабоструктурированной текстовой информации. // International Journal “Information Content & Processing” — Vol. 1, Number 3, 2014. — С. 239 —248.

[Хайрова, 2015] Хайрова Н.. Шаронова Н., Аджит Пратап Сингх Гаутам Логико-лингвистическая модель генерации фактов из текстовых потоков информационной корпоративной системы/ // International Journal Informatiom theories & application – 2015. vol. 22. № 2. – P 142-152.

[Шабанов, 2003] Шабанов В.И. Метод классификации текстовых документов, основанный на полнотекстовом поиске / В.И. Шабанов, А.М. Андреев // Труды РОМИП’2003. — СПб. : НИИ Химии СПб гос. ун-та, 2003. — С.52—71.

Authors' Information

Нина Хайрова *– профессор кафедры интеллектуальных компьютерных систем Национального технического университета «Харьковский политехнический институт», ул. Фрунзе, 21, Харьков, 61002, Украина e-mail: nina_khajrova@yahoo.com*

Научные интересы: искусственный интеллект, идентификация знаний из текстов, Text Mining, Opinion Mining, Web Mining, Natural language processing

Наталья Шаронова *– профессор, заведующий кафедрой интеллектуальных компьютерных систем Национального технического университета «Харьковский политехнический институт», ул. Фрунзе, 21, Харьков, 61002, Украина e-mail: nvsharonova@mail.ru*

Научные интересы: искусственный интеллект, математическое моделирование, автоматизированные библиотечные системы, прикладная лингвистика

Аджит Пратап Сингх Гаутам *– аспирант кафедры интеллектуальных компьютерных систем Национального технического университета «Харьковский политехнический институт», ул. Фрунзе, 21, Харьков, 61002, Украина. e-mail: apsgautam@gmail.com*

Научные интересы: интегрированные корпоративные системы, информационные технологии, модели представления знаний

Facts extraction from the semi-structured text information

Nina Khairova, Nataliya Sharonova, Ajit Pratap Singh Gautam

Abstract*: Fact extraction from the text is one of the most important areas of Natural Language Processing (NLP). Majority of existing approaches allows extracting facts from structured textual information of the specific subject areas. This paper proposes a logical-linguistic model extracting facts from semi-structured texts in English, which belong to unlimited subject areas. A fact is written in the form of a triplet: Subject - Predicate - Object, in which the Predicate defines the relations and Subject and Object define the subjects, objects or concepts. Our model defines meaning relations via grammatical and semantic features of the words in English sentences. In order to formalize and represent the participants of the fact triplet explicitly, we identify subject variables. The subject variables define a finite set of morphological and syntactic features of the words in sentences. The model was successfully implemented in the system of extraction and identification of a few types of the facts: the fact of lacking, the fact of ownership, the fact of transferring, and the fact of the presence of the attribute of time, location, and belonging for the first three fact actions. We estimated the effectiveness of our model via the coefficients of precision and recall. Results of the paper show that using of the model lets increase the numerical values of these coefficients.*

Keywords: facts extraction from the text, Natural Language Processing, semantic relations, the algebra of finite predicates, recall and precision.