OSTIS-2014
(Open Semantic Technologies for Intelligent Systems)

УДК 004.82

ЛОГИКО-ЛИНГВИСТИЧЕСКАЯ МОДЕЛЬ ИДЕНТИФИКАЦИИ СЕМАНТИЧЕСКИХ ОТНОШЕНИЙ СУЩНОСТЕЙ СРЕДСТВАМИ АЛГЕБРЫ КОНЕЧНЫХ ПРЕДИКАТОВ

Хайрова Н.Ф., Узлов Д.Ю., Шаронова Н.В.

Национальный технический университет "Харьковский политехнический институт", г. Харьков, Украина

nina_khajrova@yahoo.com

poputcik@mail.ru

nvsharonova@mail.ru

В работе предлагается логико-лингвистическая модель извлечения слабоструктурированных фактов из естественно языковых текстов. Для идентификации факта в тексте определяются некоторые сущности, выраженные лексическими единицами, и семантические связи между ними. Семантические связи определяются семантическими функциями партиципантов предложения, которые описаны предикатами алгебры конечных предикатов. Модель применяется на семантическом этапе лингвистического процессора информационной подсистемы идентификации криминалистически значимых фактов в слабоструктурированных текстах.

Ключевые слова: слабоструктурированные факты; семантические функции; алгебра конечных предикатов; лингвистический процессор.

Введение

Знания о некоторой предметной области (ПО) представляют собой совокупность сведений об объектах/субъектах данной ПО, их существенных свойствах и связывающих отношениях, а также о фактах, семантически объединяющих партиципанты ПО и их отношения в триаду субъект – атрибут – значение (или субъект – отношение – объект).

Для извлечения фактов, представленных в хорошо структурированных текстовых документах, существуют достаточно надежные алгоритмы [Baeza-Yates, 1999].

К плохо структурированной фактографической информации относятся сведения, представленные различными нерегламентированными словесными конструкциями на естественном языке. Задача извлечения плохо структурированных фактов из произвольных текстов до сих пор не имеет сколько-нибудь общего решения [Ландэ, 2009].

Для идентификации некоторого знания, представленного в форме слабоструктурированного факта, необходимо извлечь из текстовой информации некоторые объекты/субъекты, выраженные лексическими единицами, и определить семантические отношения между ними.

Так как факт выражается на естественном языке в форме законченного высказывания, то необходимо построить некий шаблон, отображающий семантические (или понятийные) связи партиципантов предложения (участников действия, выраженных существительными).

Для задания таких смысловых связей предлагается использовать семантические функции, выражаемые через отношения морфологических и семантических категорий партиципантов предложения средствами алгебры конечных предикатов (АКП).

1. Описание используемой модели.

Введем конечное множество грамматических и семантических характеристик партиципантов предложения $M = \{m_1, \dots, m_n\}$, где n количество указанных характеристик. Отношения между характеристиками можно представить в виде $m_i * m_j * \dots * m_k$, где $m_i, m_j, \dots, m_k \in M$, а знак * – обозначает, что данные характеристики соответствуют существительному, выполняющему некоторую семантическую функцию.

На множестве M введем систему предикатов S так, чтобы любой предикат $P(q_m) \in S$, обращался в 1 на множестве существительных с грамматико-семантической информацией, соответствующей определенной семантической функции, и был равен 0 в противном случае. Таким образом, множество предикатов S можно сопоставить с множеством семантико-грамматических характеристик приписанных партиципанту предложения.

Для формализации семантических функций партиципантов предложения русского языка и их явного представления средствами поверхностной структуры были выделены и описаны предметными переменными морфологические (грамматический падеж) и семантические категории существительных [Бондаренко, 2007]. Рассматривались семантические категории: живой / неживой, инструмент, часть тела, объемное пространство, пункт назначения, место отправления, плоскость/точка, механизм, определенный час, период, месяц/сезон.

Область изменения введенных переменных формально задается следующим образом:

$$x^{о} \vee x^{н} = 1,$$
$$z^{и} \vee z^{р} \vee z^{д} \vee z^{в} \vee z^{т} \vee z^{п} = 1,$$
$$y^{м} \vee y^{с} \vee y^{и} \vee y^{ч} \vee y^{т} \vee y^{о} \vee y^{в} \vee y^{п} \vee y^{ц} = 1,$$

где x^i, z^j, y^k – предметные переменные, характеризующие: $x^{о}$– категорию живого, $x^{н}$ – категорию неживого; $y^{м}$ – наличие признака «механизм», $y^{и}$ – наличие семантического признака «инструмент», $y^{ч}$ – наличие семантического признака «часть тела», $y^{т}$ – наличие семантического признака «плоскость/точка», $y^{о}$ – семантического признака «объемное пространство», $y^{в}$ – семантического признака «определенное время», $y^{п}$ – наличие семантического признака «период», $y^{ц}$ – наличие семантического признака «пункт назначения»; $z^{и}$, $z^{р}$, $z^{д}$, $z^{в}$, $z^{т}$, $z^{п}$ — грамматическая категория падежа.

Семантическая функция существительного — партиципанта предложения описывается предикатом P(x, y, z) = 1, связывающим элементы семантического значения существительного x и y с его грамматическими значениями z [Хайрова, 2012]. Тогда, используя конъюнкцию предикатов, можно записать:

$$P(x, y, z) \rightarrow P(x) \bullet P(y) \bullet P(z), \qquad (1)$$

где • — операция конъюнкции.

Таким образом, отношения между морфологическими и семантическими признаками существительных предложения, выражающие семантические функции, можно записать логическим произведением:

$$P(x_n) * P(y_n) * P(z_n) = \gamma_k(x_n, y_n, z_n) \bullet P(x_n) \bullet P(y_n) \bullet P(z_n), \qquad (2)$$

Логическое произведение предикатов $P(x_n)$, $P(y_n)$ и $P(z_n)$, описывает всевозможные отношения морфологических и семантических характеристик, а предикат $\gamma_k(x_n, y_n, z_n)$ исключает часть связей, которые не реализуются в данной семантической функции. Предикат γ_k принимает значение 1, если морфосемантическая информация словоформы n формирует некоторую семантическую функцию лексемы, и значение 0 в противном случае.

Множество рассмотренных в системе семантических категорий значения существительного задается предикатом:

$$P(y_n) = y_n^{м} \vee y_n^{с} \vee y_n^{и} \vee y_n^{ч} \vee y_n^{т} \vee y_n^{о} \vee y_n^{в} \vee y_n^{п} \vee y_n^{ц}. \qquad (3)$$

Множество значений грамматических категорий, определяющих грамматические падежи существительного, выражается предикатом:

$$P(z_n) = z^{и}_n \vee z_n^{р} \vee z_n^{д} \vee z_n^{в} \vee z_n^{т} \vee z_n^{п}. \qquad (4)$$

Признак одушевленности выражается предикатом:

$$P(x_n) = x_n^{о} \vee x_n^{н}. \qquad (5)$$

Семантическая функция агенс, представляющая субъект действия, обычно выступающего инициатором действия, лицо или предмет, имеющее потенцию на осуществление действий, выражается предикатом:

$$\gamma_A(x_n, y_n, z_n) = x_n^{о} z_n^{и} \vee z^{и} x_n^{н} y_n^{м} \vee z^{и} x_n^{о} y_n^{с}. \qquad (6)$$

Семантическая функция инструменталис, определяющая непосредственную причину действия, играющую определенную роль в совершении процесса, выражается предикатом:

$$\gamma_И(x_n, y_n, z_n) = z_n^{т} x_n^{н} y_n^{и} \vee z_n^{т} x_n^{н} y_n^{ч}. \qquad (7)$$

Семантическая функцию локатив, выражающая характеристики месторасположения, пространственной ориентации действия или состояния, выражается предикатом:

$$\gamma_Л(x_n, y_n, z_n) = z^{п} x_n^{н} y_n^{т} \vee z^{п} x_n^{н} y_n^{м} \vee z^{п} x_n^{н} y_n^{ц} \vee z^{п} x_n^{н} y_n^{о}. \qquad (8)$$

Семантическая функция объектив, определяющая объект, над которым непосредственно осуществляется действие, выражается предикатом:

$$\gamma_O(x_n, y_n, z_n) = z^{в} x_n^{н} \vee z^{в} x_n^{о}. \qquad (9)$$

Семантическая функция темпоралис, выражающая временную характеристику действия, выражается предикатом:

$$\gamma_Т(x_n, y_n, z_n) = z^{в} x_n^{н} y_n^{в} \vee z^{п} x_n^{н} y_n^{п}. \qquad (10)$$

Семантические функции в различных естественных языках имеют разные формы формального выражения и соответственно определяются предикатами АПК различного вида. Например, в русском, украинском и белорусском языках они выражаются грамматической формой и семантическими категориями существительного,

стоящего после или перед определенным глаголом, как представлено в нашей модели.

Тогда как в английском языке семантические функции во многом определяются синтаксическими категориями, в частности отношением с предлогом, стоящим после глагола. Такие категории могут быть как уникальными для определенных глаголов, так и общими, как, например, признак направления движения, определяемый предлогом (в случае его наличия) после глаголов go, run, drive, ride, transport, ship и т.д.

2. Идентификация криминалистически значимых фактов из текстов

Рассмотрим применение данной модели для идентификации криминалистически значимых фактов в источниках полнотекстовой информации. Из огромных информационных потоков текстовой информации, обрабатываемой в процессе оперативно-служебной деятельности (сводки, объяснительные/служебные записки, отчеты, газетные и интернет публикации, словесные портреты фигурантов и т. п.), и сопровождающего ее шума следователь (или иное процессуально-должностное лицо) должен извлечь факты конкретного уголовного или иного дела. В подавляющем числе случаев к таким фактам относятся: сведения о фигурантах, сведения об объекте посягательства, сведения о механизме и способе совершения преступления.

Для извлечения из неструктурированной текстовой информации фактов о дате и месте рождения некоторых персоналий (фигурантов дела), с целью исключения или привлечения потенциального кандидата, а так же о дате, месте, субъекте и объекте некоторого противоправного действия использовались семантические функции, выражающие информацию, соответствующую требованиям:

- темпоралис – временная характеристика события, позволяющая определить дату (в нашем случае: дату рождения человека или некоторого противоправного деяния);
- локатив – функция, характеризующая местонахождение, положение или состояние объекта, определяя место (в нашем примере, место рождения человека или некоторого преступного события);
- объектив – функция, определяющая сферы и продукты деятельности человека (в данном случае: сведения об объекте посягательства);
- агенс – семантическая функция, представляющая субъект действия, обычно выступающего инициатором действия (в нашем случае: лицо/субъект противоправного действия).

Были выбраны наиболее распространенные глаголы, соответствующие фактам преступного действия и идентифицирующие личность, а также определены семантические функции, являющиеся центральной частью триплета субъект – атрибут – значение и предикаты (формулы 6-9), описывающие отношения морфологических и семантических категорий существительных, соответствующих внешним элементам данного триплета (табл. 1).

Таблица 1 – Формализм модели идентификации семантических отношений

Глагол	Падеж	Предикат семантической функции
родиться, похищать, похитить, убивать, убить, красть, выкрасть, украсть, обмануть, обманывать, грабить, ограбить и др.	Темпоралис whenacted	10
	Локатив whereacted	8
	Объектив toactsmth	9
	Агенс tobeactedbysmth	6

3. Этапы работы лингвистического процессора рассмотренной задачи

Предложенная модель используется на семантическом этапе лингвистического процессора, включающего также графемный, морфологический, контекстный и семантический этапы анализа (рис. 1).

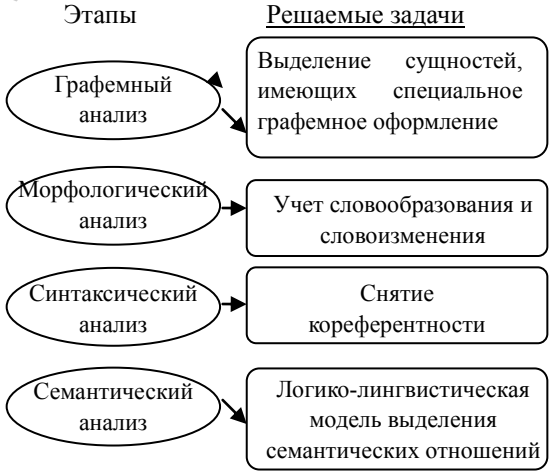

Рис. 1. Структурная схема лингвистического процессора подсистемы идентификации криминалистически значимых фактов в слабоструктурированных текстах.

Подавляющую часть субъектов криминалистически значимых фактов, составляют персоналии, компании и организации, кроме того, часто необходимо определять факт местоположения того или иного субъекта. Специальное графическое оформление таких сущностей позволяет для их выделения использовать формализмы графемного анализа. Графемный этап используется для выявления антропонимов, топонимов (включающих названия населенных мест, внутригородских объектов и др.), названий организаций, предприятий

и учреждений и базируется на базе имен собственных и специальных алгоритмах.

На этапе морфологического анализа осуществляется обработка словоизменительных и словообразовательных форм, позволяющая, например, отнести глаголы ―красть" и ―украсть" к одному и тому же факту.

Качественное построение шаблона факта предполагает определение возможных имен сущностей, которые могут встречаться в текстах под разными знаками. Для повышения точности определения факта на этапе синтаксического анализа необходимо решить проблему кореферентности, в частности, ассоциировать местоимения со своими антецедентами, соотнося их с именуемой сущностью.

Задача построения связей шаблона факта представляет собой одну из центральных задач извлечения знаний из текстов, она направлена на распознавание в тексте набора возможных отношений между элементами шаблона, разработанными на предыдущем этапе. Для решения данной задачи на этапе семантического анализа лингвистического процессора используется разработанная логико-лингвистическая модель идентификации отношений между сущностями.

4. Практическая реализация разработанного метода

Задача извлечения криминалистически значимых фактов из полнотекстовой информации, обрабатываемой в процессе оперативно-служебной деятельности, была реализована для извлечения даты, места рождения персоналии, а также сведений о дате, месте, субъекте и объекте некоторого противоправного действия.

Программа представляет собой веб-приложение, анализирующее текст или список рассматриваемых текстовых файлов. Определенная фактографическая информация представляется пользователю в форме диалогового окна, в котором отображается извлеченный факт и предложение или несколько предложений, представляющие описание данного факта.

В разработанном приложении определяются семантические функции, выражающие информацию, соответствующую поставленной задаче: темпоралис - временная характеристика события, локатив - характеристика местонахождения, положения или состояния; агенс и объектив – участники заданного действия. В процессе реализации модели был определен набор глаголов, требующих выполнения данных семантических функций партиципантов предложения (табл. 1).

Тестовая проверка разработанного приложения показала приемлемость использования предложенной модели.

Заключение

Таким образом, разработанная логико-лингвистическая модель идентификации семантических отношений сущностей позволила извлекать факты из слабоструктурированной текстовой информации виде триплета субъект – отношение – объект. Отношения между сущностями в данном триплете описывается предикатом взаимосвязи морфологических и семантических категорий партиципанта того или иного действия. Модель применяется на семантическом этапе лингвистического процессора информационной подсистемы идентификации криминалистически значимых фактов в слабоструктурированных текстах для чего к настоящему моменту описаны функции: темпоралис, локатив, объектив и агенс.

Библиографический список

[Baeza-Yates, 1999] Baeza-Yates R., Ribeiro-Neto B. Modern Information Retrieval. — Addison-Wesley, 1999. — 340 p.

[Ландэ, 2009] Ландэ Д. В. Интернетика: Навигация в сложных сетях: модели и алгоритмы: моногр. / Д. В. Ландэ, А. А. Снарский , И. В. Безсуднов — М.: Либроком (Editorial URSS), 2009. 264 с.

[Бондаренко, 2007] Бондаренко М. Ф. Теория интеллекта: учебник/ Бондаренко М. Ф., Шабанов-Кушнаренко Ю. П. Харьков: Комп. СМИТ, 2007. — 576 с.

[Хайрова, 2012] Хайрова Н. Ф. Використання логіко-алгебраїчної моделі семантичних відмінків для семантичного аналізу речення/ Н. Ф. Хайрова // Зб. наук. пр. Військового ін-ту Київ. нац. ун-ту. — К.: ВІКНУ, 2012. — Вип. № 38. — С. 239—245.

LOGICAL-LINGUISTIC MODEL FOR IDENTIFICATION OF SEMANTIC RELATIONSHIPS BETWEEN ENTITIES ON BASE OF ALGEBRA OF FINITE PREDICATES

Khairova N., Uzlov D., Sharonova N.

National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine

nina_khajrova@yahoo.com

poputcik@mail.ru

nvsharonova@mail.ru

This paper proposes a logical-linguistic model extracting semi-structured facts of natural language texts. To identify the fact some entities expressed by lexical units as well as semantic relations between them are defined in the text. The semantic relations are expressed by semantic functions of sentence participants. The functions are described by predicates of algebra of finite predicates. The model is applied to the semantic stage of linguistic processor of information subsystem for facts identification, which are essential for criminalistics, in the framework a semi-structured texts.

Key words: semi-structured facts; semantic functions; algebra of finite predicates; linguistic processor.