Murakami, N. (2009). Video switcher and video switching. Patent US 2009/0109334 A1.**8.** *Krivomaz, L. S.* (2010). Multi-camera live broadcasting as an effective training method for forming professional skills of cameramen and television directors. Innovative development of the society in the context of cross-cultural interactions. Abstracts from the reports from the 3rd International conference. Volume 2, 169 – 171.**9.** *Sokolov, A. G.* (2001). Montage: TV, movies, video. 625.**10.** *Fedorishin, V. I., Krivomaz, L. S., Pechenyuk, D. A.* (2009). Ukrainian music in the world culture, the TV version of the concert. National Pedagogical Dragomanov University, http://youtu.be/XDYGA3YhOXo. **11.** *Panchenko, B. E., Nikolenko, R. B., Pechenyuk, D. A.* (2014) Festival "Shid-Rock - 2014", the TV version of the concert. ODTRK Sumy, http://youtu.be/xFwSsbGSN7A.

Надійшла (received) 25.02.2015

УДК 004.89

Н. Ф. ХАЙРОВА, д-р техн. наук, проф., НТУ «ХПИ»;
АДЖИТ ПРАТАП СИНГХ ГАУТАМ, аспирант, НТУ «ХПИ»

ИНФОРМАЦИОННОЕ ПРОСТРАНСТВО ФАКТОВ ИНТЕГРИРОВАННОЙ КОРПОРАТИВНОЙ СИСТЕМЫ

В работе предлагается информационно-лингвистическая технология выделения фактов из слабоструктурированных и неструктурированных текстов. Технология основана на использовании специальных семантико-алгебраических (логических) методах, которые позволяют получать точность и полноту фактов, сравнимую с экспертными. Для извлечения и структурирования фактографической информации в тексте выделяются сущности, и используется структурированное представление семантики факта в терминах предикатных операций.

Ключевые слова: пространство фактов, автоматическая идентификация и экстракция, слабоструктурированный текст, предикатные операции.

Введение. Центральной составляющей современной интегрированной корпоративной системы является база знаний, которая должна включать в себя единое информационное пространство взаимосвязанных фактов или гипотез вне зависимости от типа источника получения информации. Сегодня система извлечения фактов является одним из наиболее эффективных инструментов выделения нужной для принятия решений информации и для проведения аналитической бизнес-разведки [1], практически заменяя обычный поиск информации. Факт о некоторой сущности представляет собой структурированную экстракцию из предложения текста документа в виде значения факта: его суть, время и место совершения, его участники [2].

Основной проблемой обработки фактографической информации является оценка достоверности автоматически определяемой фактографической информации [3], что особенно важно в связи с все более увеличивающейся плотностью потока текстовой информации в средствах масс-медиа и различного рода социальных сетях, форумах и блогах. Множественность значений факта обусловлена возможностью разной интерпретации одного и того же явления, а также противоречивостью, неточностью или нечеткостью поступающих из внешних источников сведений.

Цель работы. Целью работы является разработка подсистемы идентификации и экстракции фактов, позволяющая получить пространство фактов, динамически наполняемое из текстового-контента интегрированной корпоративной системы. На вход подсистемы поступают текстовые потоки разнородных источников информационной системы, на выходе формируется базовое пространство фактов интегрированной корпоративной системы.

Описание предметной области. Факт представляет собой знание в форме утверждения, достоверность которого строго установлена [1]. В сфере информационных технологий и теории обработки знаний под фактом, обычно, понимают зафиксированное и произошедшее событие, сопровождаемое временными и географическими метками, аргументирующей информацией, ссылками на источник и т. д.

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. При этом под объектом мы понимаем сущность, информация о которой накапливается в системе и может быть в ней само идентифицирована [2]. Для извлечения и структурирования фактографической информации в тексте выделяются сущности, и используется структурированное представление семантики факта в терминах предикатных операций. Факты выделяются из предложений, содержащих упоминание сущности или анафорические ссылки на нее. В свою очередь фактографическую информацию можно разделить на хорошо структурированную и плохо структурированную.

К хорошо структурированным сведениям (так называемая параметрическая информация) относятся, прежде всего, сведения количественного характера, а так же качественные сведения, имеющие хорошо регламентированную форму. К плохо структурированной фактографической информации относятся сведения, представленные различными нерегламентированными словесными конструкциями, представленными на естественном языке [3].

Алгоритмы фактографического анализа зависят, в свою очередь, от степени структурированности конкретного документа. По степени структурированности данные документа можно разделить, подобно общей классификации степени формализации информации, на табличные данные, отображенные в виде фактов; массивы однородных слабоструктурированных текстовых документов, обычно описывающие конкретную предметную область и документы произвольного слабоструктурированного типа [4].

Общее описание метода. Выделение фактов из слабоструктурированной текстовой информации включает следующие этапы [5]: извлечение слов или словосочетаний, важных для описания смысла текста; исследование связей между извлеченными понятиями; извлечение сущностей, распознавание фактов и действий.

Для реализации первого этапа выделения понятий используется стандартный лингвистический процессор [6], включающий графемную, морфологическую, синтаксическую и контекстную этапы обработки, с добавлением специализированных методов и алгоритмов обработки документов (рис. 1). Так как очень часто в задачах по извлечению фактографической информации нужно найти в тексте упоминания лиц, компаний, правительственных организаций и местоположений, и другие подобные типы сущностей, то для их выделения

используются специальные формализмы графемного анализа. На этапе морфологического анализа используется декларативный и алгоритмический методы. Каждый неправильный глагол английского языка представлен в базе данных во всех его формах, то есть глагол write иметь формы write-writes-wrote-written-writing, формы правильных глаголов определяются алгоритмически.

Факт представляет собой триплет: *Subject ->Predicate ->Object*, в котором предикат представляет собой отношение, а субъект и объект указывают на два предмета [7]. Практическое запись такого факта осуществляется строкой в таблице реляционной базы данных, поля которой представляют субъект и объект триплета, а имя таблицы соответствует отношению между предметами или предикатом триплета. Кроме того можно использовать представление в виде двухместного предиката в логике первого порядка. Следующим этапом после выделения слов или словосочетаний, представляющих узлы триплета факта, является выделение отношений, устанавливаемых данным фактом между словами.

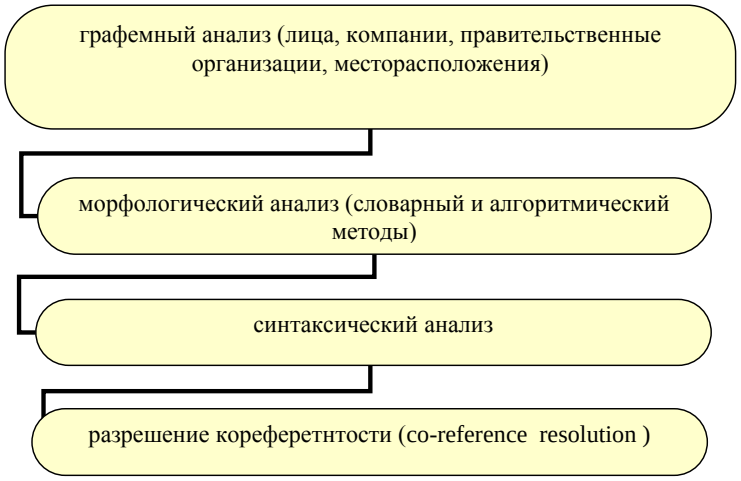

Рис. 1 – Базовые этапы используемого лингвистического процессора

Выделяем два типа 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*). Если второго участника связи в базе нет, то он создается автоматически.

Второй вид факта 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.

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

Такой подход рассматривается в рамках падежной грамматики и основывается на понятии глубинных падежей, введенных Ч. Филлмором, выделившим пропзицию, или основной смысл предложения, как предикат, выражаемый в поверхностной структуре чаще глаголом, связанным с помощью определенных глубинных падежей с участниками данной ситуации, или партиципантами [8]. Семантические падежи в различных естественных языках имеют разные формы

формального выражения, которые необходимо четко определить для автоматической идентификации и экстракции фактов из текстов. Например, в русском и украинском языках, семантическая информация партиципантов кодируется, в основном, грамматическими поверхностными падежами, тогда как в английском — она передается сочетанием с предлогом, порядком слов в предложении.

Описание математической модели. Введем на универсуме U, включающем все возможные понятия и объекты анализа сложной языковой системы [9], множество грамматических характеристик синтаксической сочетаемости слов английских предложений, влияющих на понятийные связи, $M = \{m_1, \dots, m_n\}$, где n – количество характеристик системы. Используя формальный аппарат алгебры конечных предикатов [10].

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 конъюнкция данных характеристик соответствует некоторой семантической функции или некоторому глубинному смысловому отношению между словами, грамматические характеристики которых выражаются $m_i, m_j, \dots, m_k$.

На множестве M введем систему предикатов S так, чтобы любой предикат $P(q_m) \in S$, обращался в 1 на множестве слов с грамматической информацией, соответствующей определенной семантической функции, и был равен 0 в противном случае. Таким образом, множество предикатов S можно сопоставить с множеством грамматических характеристик приписанных словам предложения, называющим сущности триплета факта.

Для формализации семантических функций предложений английского языка и их явного представления средствами поверхностной структуры были выделены и описаны следующие синтаксические и морфологические категории:

$$z^{to} \vee z^{by} \vee z^{with} \vee z^{about} \vee z^{of} \vee z^{on} \vee z^{at} \vee z^{in} \vee z^{out} = 1,$$
$$y^{ap} \vee y^{aps} \vee y^{out} = 1, \; x^{f} \vee x^{l} \vee x^{kos} = 1,$$
$$m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were} \vee m^{out} = 1,$$
$$p^{III} \vee p^{ed} \vee p^{I} \vee p^{ing} \vee p^{II} = 1,$$

где использованы предметные переменные, характеризующие следующие категории:

- наличие предлога *to*, *by*, *with*, *about*, *of* после предиката триплета или его отсутствие – $z^{to}, z^{by}, z^{with}, z^{about}, z^{of}, z^{at}, z^{on}; z^{in}, z^{out}$
- наличие или отсутствие апострофа в конце слова, определяющего притяжательный падеж у субъекта триплета – y^{ap}, y^{aps}, y^{out};
- расположение существительного, определяющего сущность, перед глаголом в личной форме, после глагола в личной форме или после косвенного дополнения – x^{f}, x^{l}, x^{kos};
- наличие или отсутствие любой формы глагол *to be* – $m^{is}, m^{are}, m^{havb}, m^{hasb}, m^{hadb}, m^{was}, m^{were}, m^{out}$;
- первая, вторая/третья и четвертая форма основного правильного глагола , и вторая, третья формы неправильного основного глагола – $p^{I}, p^{ed}, p^{ing}, p^{II}, {}^{III}$.

Семантические связи между извлеченными понятиями текста определяются через предикат P, связывающие категории наличия предлога после предиката,

существование апострофа, определяющего притяжательный падеж, расположения понятия, факт связи которого определяется, наличия глагола *to be* и формы основного глагола:

$$P(x, y, z, m, p) \rightarrow P(x) \wedge P(y) \wedge P(z) \wedge P(m) \wedge P(p). \quad (1)$$

Зададим на декартовом квадрате множества $S * S$ предикат $\gamma(x_n, y_n, z_n, m_n, p_n)$, принимающий значение 1, если комплекс выбранных категорий для фразы n формирует некоторые семантические связи понятий триплета, т.е. формирует некий факт, и значение 0 в противном случае. Таким образом, отношения грамматических элементов английского предложения, идентифицирующих некоторый факт, можно задать формулой:

$$P(x, y, z, m, p) = \gamma_k(x, y, z, m, p) \wedge P(x) \wedge P(y) \wedge P(z) \wedge P(m) \wedge P(p). \quad (2)$$

Практически никогда подмножество согласующихся категорий информации, выражающей факты, не совпадает с декартовым произведением на множестве грамматических признаков. Грамматические категории, которые в своей конъюнкции не формируют семантические связи и соответственно факты, исключаются из формулы (1) множителем $\gamma_k(x_n, y_n, z_n, m_n, p_n)$, $k \in [1;h]$, где h — количество, принятых к рассмотрению в системе типов фактов.

В процессе реализации модели был определен набор глаголов, соответствующих центральной части триплета идентифицируемых типов фактов. Одним из рассмотренных типов фактов является утверждение об обладании, приобретении (или наличии) у некоторой сущности субъекта некоторой сущности объекта. Такое утверждение в англоязычных текстах будет определяться предикатами (глаголами), заранее определенными в базе данных: *have, purchase, buy, acquire, get, gain, obtain*.

В соответствии с формулой (2) семантическая связь, выделяющая субъект триады данного утверждения будет определяться следующим предикатом:

$$\gamma_1(x_n, y_n, z_n, m_n, p_n) = z^{out} y^{out} x^{f} m^{out} p^{I} \vee z^{out} y^{out} x^{f} m^{out} p^{II} \vee z^{out} y^{out} x^{f} m^{out} p^{ed} \vee z^{by} y^{out} x^{l} p^{ed} (m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were}) \vee \vee z^{by} y^{out} x^{l} p^{III} (m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were}) \quad (3)$$

Объект данного факта будет явным образом выделен из предложения с помощью предиката, соответствующего конъюнкции предметных переменных грамматических категорий членов предложения:

$$\gamma_2(x_n, y_n, z_n, m_n, p_n) = z^{out} y^{out} x^{l} m^{out} p^{I} \vee z^{out} y^{out} x^{l} m^{out} p^{ed} \vee z^{out} y^{out} x^{l} m^{out} p^{II} \vee z^{out} y^{out} x^{f} p^{III} (m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were}) \vee z^{out} y^{out} x^{f} p^{ed} (m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were}) \quad (4)$$

Ко второму виду фактов, связанных с теми же глаголами можно отнести определение атрибутов времени, места, способа действия и т.д. Например, факт времени осуществленного действия выделяется из предложения с помощью предиката

$$\gamma_3(x_n, y_n, z_n, m_n, p_n) = z^{on} x^{kos} y^{out} m^{out} \vee z^{in} x^{kos} y^{out} m^{out} \vee z^{at} x^{kos} y^{out} m^{out} \quad (5)$$

Дополнительным лингвистическим условием выражения семантических

связей, определяющей атрибутивный факт места осуществления действия является представление объекта триплета факта в именем собственным (обычно графически выражаемым с большой буквы), так как в данном факте интерес представляет именно населенный пункт, а не местоположение, как например, *in mansion*.

Факт принадлежности или собственности объекта некоторому субъекту выделяется из предложений с вышеперечисленными глаголами, но определяется следующим предикатом

$$\gamma_3 (x_n, y_n, z_n, m_n, p_n) = z^{out} x^f(y^{ap} \vee y^{aps}) \quad (6)$$

Программная имплементация модели

Программная имплементация модели представляет собой веб-приложение, анализирующие текст или список анализируемых текстовых файлов. Извлеченная системой фактографическая информация представляется пользователю форме диалогового окна (рис. 2).

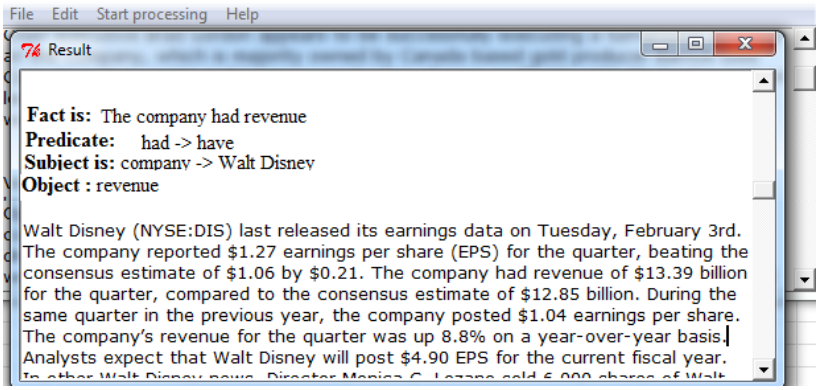

Рис. 2 – Фрагмент диалогового окна результата работы программы

Программа отображает изв-леченную фактографическую информацию в виде факта и первичных предложений, из которых данный факт был извлечен. В поле *"Fact is"* представлено извлеченное утверждение в виде триплета; в поле *"Predicate"* представлен извлеченный глагол и его каноническая форма; в полях *"Subject"* и *"Object"* соответственно извлеченные субъект и объект триплета. Кроме того, извлеченные названия сущностей анализируются по базе данных гипонемических отношений экономических терминов и подвергаются анализу, устанавливающему анафорические ссылки (рис. 1).

Идентифицированные факты записывается последовательно, перед абзацем текста, из которого он извлекается. Факты деятельности располагаются в порядке значимости, определенной системой**.**

Выводы. Результатом данного исследования является разработка логико-лингвистической модели извлечения фактов из слабоструктурированных текстов на английском языке. Используемая технология идентификации и экстракции фактов, основывающаяся на использовании специальных семантико-лингвистических методов, включающих специализированный лингвистический процессор, учитывающий как анафорические ссылки, так и словоизменительные формы, позволяют получить полноту и точность получаемого фактографического пространства, сравнимую с экспертными оценками.

Список литературы: **1.** *Киселев, С.* Модель информационной системы бизнес-разведи [Электронный ресурс] / *С. Киселев.* – Открытые системы #05-06/2005. Режим доступа: http://www.osp.ru/os/2005/05-06/185595/ **2.** *Andersen, P. M.* Knowledge engineering for the JASPER fact extraction system. [Text] / *Andersen, P. M., Huettner A. K.* // Intrgated Computer-Aided Enginrering. – 1 (6), 1994. – P. 473–493. **3.** *Ландэ, Д. В.* Интернетика: Навигация в сложных сетях: модели и алгоритмы [Текст] / *Д. В. Ландэ, А. А. Снарский , И. В. Безсуднов.* – М.: Либроком

(Editorial URSS), 2009. – 264 с. **4.** *Fader, A.* Identifying relations for open information extraction. [Text] / *Fader, S. Soderland, O. Etzioni.* // Conference on Empirical Methods in Natural Language Processing. – Edinburgh, Scotland, 2011. – P. 1535 – 1545. **5**. *Барахнин, В. Б.* Проблемы разработки технологии фактографического поиска [Текст] / *В. Б. Барахнин, А. М. Федотов*. – М.: Институт вычислительных технологий СО РАН, 1980. – 150 с. **6.** *Baeza-Yates, R.* Modern Information Retrieval. [Text] / *R. Baeza-Yates, B. Ribeiro-Neto* // Addison-Wesley, 1999. — 340 p. **7**. *Ritter, A.* Named entity recognition in tweets: an experimental study. [Text] / *A. Ritter, S. Clark, K. Mausam, O. Etzioni* // Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics. – Edinburgh/Scotland, 2011. – P. 1524–1534. **8**. *Филлмор, Ч.* Дело о падеже открывается вновь // Новое в зарубежной лингвистике [Текст] / *Ч. Филлмор.* – М.: Изд. иностр. лит., 1981, вып. 10. – С. 496-530. **9.** *Хайрова, Н. Ф.* Використання логіко-алгебраїчної моделі семантичних відмінків для семантичного аналізу речення [Текст] / *Н. Ф. Хайрова* // Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. – К.: ВІКНУ, 2012.- Вип. № 38. – С. 239 – 245. **10.** *Бондаренко, М. Ф.* Теория интеллекта [Текст] / *Бондаренко М. Ф., Шабанов-Кушнаренко Ю. П.* // Харьков: Комп. СМИТ, 2007. – 576 с.

Bibliography (transliterated): **1.** *Kiselev, S.* (2015). Model of information system business dissolve. Otkritie sistemi. #05-06/2005: http://www.osp.ru/os/2005/05-06/185595/ **2.** *Andersen, P. M., Huettner, A. K.* (1994). Knowledge engineering for the JASPER fact extraction system. Intrgated Computer-Aided Enginrering. – 1 (6), 473–493. **3.** *Lande, D. V., Snarskiy, A. A., Bezsudnov, I. V.* (2009). Internetika: Navigation in complex networks: models and algorithms. Moskow: Libkom (Editorial URSS). **4.** *Fader, S. Soderland, O. Etzioni, A.* (2011). Identifying relations for open information extraction. Conference on Empirical Methods in Natural Language Processings. Edinburgh, Scotland, 1535–1545. **5.** *Barahnin, V. B., Fedotov, A. M.* (1980). Problems of development of technology factual search. Moscow: Institute of Computational Technologies. **6.** *Baeza-Yates, R., Ribeiro-Neto, B.* (1999). Modern Information Retrieval. Addison-Wesley. **7.** *Ritter, A., Clark, S., Mausam, K., Etzioni, O.* (2011). Named entity recognition in tweets: an experimental study. Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics. – Edinburgh/Scotland, 1524–1534. **8.** *Charles, J. Fillmore* (1968). The Case for Case. In Bach and Harms (Ed.): Universals in Linguistic Theory. New York: Holt, Rinehart, and Winston, 1-88. **9.** *Khairova, N.* (2012). Using logic-algebraic model of semantic rules for semantic analysis of the sentence. Kiev: Zbirnik naukovich prach Viyskovogo institutu, Vol. 38, 239 – 245. **10.** *Bondarenko, M. F., Shabanov-Kushnarenko, U. P.* (2007). The theory of intelligence. Kharkov: SMIT Comp.

Надійшла (received) 21.02.2015

УДК 004.912

О. В. ЛОЗИНСЬКА, асистент, НУ «Львівська політехніка»;
М. В. ДАВИДОВ, канд. техн. наук, доц., НУ «Львівська політехніка»

МАТЕМАТИЧНА МОДЕЛЬ ГРАМАТИЧНО-ДОПОВНЕНОЇ ОНТОЛОГІЇ

Дослідження відомих методів вирішення проблеми багатозначності слів з використанням онтологій показало, що відомі методи обмежені лише контекстом слова і не надають додаткових переваг для граматичного і семантичного розбору речення. Для вирішення цієї проблеми розроблено математичну модель граматично-доповненої онтології. Ця модель використана для граматичного розбору речень української мови. Отримані результати показали адекватність розробленої моделі, але її використання вимагає наповнення словників нового типу.

Ключові слова: українська жестова мова, онтологія, граматично-доповнена онтологія, синсет.