7. Prihodchenko V.V., Prihodchenko O.V. Diagnostika zabolevanij molochnoy zhelezi s pomochyu zifrovogo kontaktnogo termomammografa [Diagnosis of breast disease with a digital contact thermograph]. Мед.-соц. проблемы семьи [Med.-social. family problems], 2005, vol.4, №10, pp. 61–65.

8. Rozenfeld L.G. Kolotilov N.N. Distanzionnaja infakrasnaja termografija v onkologii [Remote infrared thermography in oncology]. Onkologija [Oncology], 2001, vol. 3, № 2, pp 103–106.

9. Sverdan P.L. Vysshaja matematika. Analiz informazii v farmazii i medicine [Higher Mathematics. Analysis of information in the pharmacy and medicine]. Uchebnik [Textbook], Lvov, 1998, 332 p.

10. Andersson I., Janzon L. Reduced breast cancer mortality in women under age 50 : updater results from the Mammopgraphic Screening Program. J. Natl. Cancer Inst. Monogr., 1997, vol. 22, pp. 63–67.

11. Carbone A. Algorithm to estimate the Hurst exponent of high-dimensional fractals. Physical review, 2007, vol. E 76, 056703.

12. Carbone A., Stanley H.E. Scaling properties and entropy of long range correlated series. Physica A, 2007, vol. 384, pp. 21 – 24.

13. Curpen B. N., Sickles, E. A., Sollitto R. A. The comparative value of mammographic screening for women 40-49 years old versus women 50-59 years old. AJR, 1995, vol. 164, pp. 1099–1103.

14. Gautherie M. Thermopathalogy of breast cancer, measurement and analysis of in-vivo temperature and blood flow. Ann NY Acad Sci., 1980, pp. 365 – 383.

15. Gautterie M. Thermobiological assessment of being and malignant breast disease . Am J Obstet Gynecol., 1983, vol. 8, pp. 861–869.

16. Gorshkov O. Stabilogram diffusion analysis algorithm to estimate the Hurst exponent of high-dimensional fractals. J. Stat. Mech., 2012, vol. P04014, pp. 1-13.

17. Louis K., Walter J., Gautherie M. Long –Term Assessment of Breast Cancer Risk by Thermal Imaging. Alan R. Liss Inc., 1982, pp. 279-301.

18. Medical infrared imaging. Edited by Nicholas A. Diakides, Joseph D. Bronzino. CRC Press, Taylor & Francis Group. Boca Raton, U.S.A. 2008. 450 p.

19. Peng C.K., Hausdorff J.M, Goldberger A.L. Fractal mechanisms in neural control: Human heartbeat and gait dynamics in health and disease. Self-Organized Biological Dynamics and Nonlinear Control. Cambridge: Cambridge University Press, 2000.

20. Ring E., Ammer K. The technique of infrared imaging in medicine. Thermology International, 2002, vol.10, № 1, pp. 7–14.

21. Sterns E.E, Zee B, Sen Gupta J. Thermograpy: Its relation to pathologic characteristics, vascularity, proliferative rate and survival off patients with invasive ducatal canciroma of the breast. Cancer, 1996, vol. 77, pp. 124-128.

22. Stewart B, Kleihues PE. World cancer report. Lyon: IARCPress, 2003.

23. Tambasco M, Eliasziw M. Morphologic complexity of epithelial architecture for predicting invasive breast cancer survival. Journal of Translational Medicine, 2010, vol.8, pp.140.

24. Tavakol M., Lucas C., Saeed S. Analysis of Breast Thermography Using Fractal Dimension to Establish Possible Difference between Malignant and Benign Patterns. Journal of Healthcare Engineering, 2010, vol. 1, № 2, pp. 27–43.

УДК 004.89

ИДЕНТИФИКАЦИЯ ВЗАИМОСВЯЗЕЙ МЕЖДУ ТЕРМИНАМИ И ОБЪЕКТАМИ ЭКОНОМИЧЕСКОЙ ТЕМАТИКИ В ТЕКСТЕ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Статья поступила в редакцию 05.11.2015 г., в окончательном варианте 12.11.2015 г.

Дмитриев Александр Сергеевич, старший преподаватель, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. им. Ленина, 28, e-mail: golostos@yandex.ru

Соловьев Иван Сергеевич, разработчик, SurfStudio, 394018, Российская Федерация, г. Воронеж, ул. Среднемосковская, 1д, e-mail: issoloveyv@gmail.com

Орлова Юлия Александровна, кандидат технических наук, кандидат педагогических наук, доцент, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. им. Ленина, 28, e-mail: yulia.orlova@gmail.com

Розалиев Владимир Леонидович, кандидат технических наук, доцент, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. им. Ленина, 28, e-mail: vladimir.rozaliev@gmail.com

Константинов Василий Михайлович, аспирант, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. им. Ленина, 28, e-mail: konstantinovr1@gmail.com

В настоящее время происходит стремительное развитие систем обработки текста. Однако пока существует достаточно мало программных комплексов для эффективного поиска взаимосвязей внутри текстов. Это особенно ощущается при обработке значительного количества текстов экономической тематики, в т.ч. при извлечении из них важных (значимых) финансово-экономических терминов. Поэтому в рамках данной статьи описана созданная авторами программная система для идентификации объектов и терминов экономической тематики. Эта система позволяет также определить, с каким объектом связан тот или иной экономический термин. Был разработан шаблонный метод (на основе метода Snowball) для идентификации объектов и терминов в тексте. В статье описываются признаки терминов и объектов, особенности работы с контекстом для шаблонного метода. Для идентификации и уточнения отношений между объектами и терминами используется метод опорных векторов (SVM). Приводятся алгоритмы указанных методов и общая архитектура созданной программной системы.

Ключевые слова: экономические термины, идентификация объектов, текст, извлечение отношений, контекст предложения, программная система, интеллектуальный анализ данных, метод опорных векторов

RELATIONSHIP IDENTIFICATION BETWEEN THE ECONOMIC TERMS AND OBJECTS IN NATURAL LANGUAGE TEXT

Dmitriev Aleksandr S., senior lecturer, Volgograd State Technical University, 28 Lenin Ave., Volgograd, 400005, Russian Federation, e-mail: golostos@yandex.ru

Solovev Ivan S., software developer, SurfStudio, 1d Srednemoskovskaya St., Voronezh, 394018, Russian Federation, e-mail: issoloveyv@gmail.com

Orlova Yuliya A., Ph.D. (Engineering), Ph.D. (Pedagogics), Associate Professor, Volgograd State Technical University, 28 Lenin Ave., Volgograd, 400005, Russian Federation, e-mail: yulia.orlova@gmail.com

Rozaliev Vladimir L., Ph.D. (Engineering), Associate Professor, Volgograd State Technical University, 28 Lenin Ave., Volgograd, 400005, Russian Federation, e-mail: vladimir.rozaliev@gmail.com

Konstantinov Vasiliy M., post-graduate student, Volgograd State Technical University, 28 Lenin Ave., Volgograd, 400005, Russian Federation, e-mail: konstantinovr1@gmail.com

Despite the rapid development of systems for processing text, there are fairly small numbers of software systems for efficient search links within a text. It is particularly acute when processing a large number of economic texts and extraction of important financial and economic terms. As part of this work the software system was developed for the identification of objects and terms of economic subjects. The program also allows you to specify which object is associated one or another economic term. We have developed the template method on the basis of Snowball to identify objects and terms in the text. This paper de-

scribes the characteristics of terms and objects and features of the context for the template method. For identify and clarify the relationships between objects and terms a support vector machine (SVM) is used. Then shows the algorithms of described methods and shows the overall architecture of the software system.

Keywords: economic terms, object identification, text, relation extraction, context of the sentence, software system, data mining, support vector machines

Введение. Стремительный рост объемов информации в современном мире делает все более важным решение задач автоматизированной обработки текстов. Как следствие, возрастает количество соответствующих программных разработок, они приобретают дополнительные функциональные возможности, улучшается качество решения ими задач обработки текстов. Однако некоторые направления обработки текстов остаются исследованными не в полном объеме. Одним из таких направлений является автоматизированная обработка текстов экономической тематики. Поэтому целью данной статьи был комплексный анализ методов извлечения из текста экономических терминов и связанных с ними объектов.

Обоснование актуальности работы и основные задач. Современные классификаторы текстов не используют семантическую информацию для определения тематики текста. Это создает трудности для корректного разбиения больших предметных областей на меньшие множества текстов (соответствующие некоторым подобластям). Причина в том, что в данных текстах встречается большое количество одинаковых ключевых слов (например, для экономической тематики – это «долг», «кредит», «доход» и т.д.) и не учитывается контекстная информация [11].

Данная статья ограничивается экономической предметной областью. Связано это с тем, что в последние годы объем текстовой информации, содержащей различную экономическую информацию, стал чрезмерно большим для того, чтобы финансисты, экономисты и другие смежные аналитики успевали в достаточно сжатые сроки обрабатывать эту информацию. Особенно большую проблему представляет поиск актуальной финансовой информации и ее доставка аналитикам. Поэтому авторами было принято решение разработать систему извлечения экономических терминов и контекстной информации между ними из текстов на естественном языке (ЕЯ). Конечной целью являлось дальнейшее использование этой информации при кластеризации текстов по экономическим темам и создания подборок для аналитиков.

Объектами «экономической тематики» являются слова, связанные с экономическими терминами. В ходе выполнения работы были выполнены следующие задачи: проведен обзор аналогов систем установки взаимоотношений объектов в тексте на естественном языке (ЕЯ); исследовано современное состояние автоматизированного извлечения контекста из текста на ЕЯ; разработан алгоритм извлечения экономических терминов из текста и алгоритм установления (оценки) вероятности взаимосвязи пар; разработана программная система для поиска пар термин-объект и установления вероятности связи между ними.

Обзор аналогов. Были проанализированы существующие системы установки взаимоотношений объектов из неструктурированных текстов, такие как CoreNLP, Calais, NetOwl Extractor OntosMiner, Link Parser [4–8]. Большинство из них обеспечивает возможности построения семантических сетей; извлечения фактов, понятий; поиск по ключевым словам; создание таксономий и тезаурусов. Однако, ни одна из них не позволяет устанавливать взаимосвязи между объектами и терминами на русском языке – это и является главной особенностью работы, описываемой в настоящей статье. Сравнительные характеристики описанных систем приводятся в таблице 1.

Таблица 1

Сопоставление характеристик разных пакетов поиска отношений в тексте

	Поддержка русского языка	Шаблонное извлечение отношений	Поиск терминов	Вероятностная оценка связей «объект-термин»
CoreNLP	–	–	+	–
Calais	–	+	–	–
NetOwl Extractor	+	–	+	–
OntosMiner	+	–	–	–
Link Parser	–	+	–	–
Собственный алгоритм	+	+	+	+

Краткие характеристики рассмотренных программных комплексов.

Stanford CoreNLP (The Stanford Natural Language Processing Group) – это набор инструментов для анализа ЕЯ.

Calais (Thomson Reuters) – с использованием обработки ЕЯ (NLP) и машинного обучения, анализирует документ и выделяет объекты.

Netowl Extractor (SRA International, Inc.) предоставляет методы для извлечения имен личностей, отношений и событий; обеспечивает геотаггинг (расстановка геолокационных тегов для слов, описывающих географическое местоположение) и анализ тональности текста на нескольких языках.

Ontos AG (Ontos) является разработчиком и поставщиком семантических технологий – с акцентом на интеграцию и анализ информации.

Link Parser (Carnegie-Melon University), разработанный в Carnegie-Melon University - работает со словарем, включающем около 60000 словарных форм.

Исходя из того, что семантическая структура ЕЯ имеет предикатно-аргументное строение, использование систем разметки семантических ролей может улучшить результаты любого направления автоматической обработки текста на ЕЯ, например поиска информации.

Извлечение контекста из текста. Под контекстной информацией понимается совокупность объектов, терминов и связывающих их слов. Для генерации шаблонов поиска данной контекстной информации было принято решение использовать модифицированный метод Snowball [9]. Этот метод (рис. 1) основан на ключевых компонентах метода DIPRE [9]. Более конкретно, Snowball представляет новую технику для создания моделей и извлечения кортежей из текстовых документов. Кроме того, Snowball вводит стратегию для оценки качества моделей и наборов, которые создаются на каждой итерации процесса извлечения отношений. Только те кортежи, и модели, которые считаются «достаточно надежными», будут храниться в Snowball для следующих итераций системы. Эти новые стратегии для генерации и фильтрации образцов и наборов кортежей значительно улучшают качество извлеченных отношений.

Важнейшим шагом в процессе выделения связей является генерация шаблонов для нахождения новых кортежей в тексте. В идеале, шаблоны должны быть избирательными (так чтобы они не генерировали неправильные кортежи) и иметь высокий уровень охвата - для того чтобы находить разнообразные кортежи. Изначально для метода Snowball на вход подается несколько примеров кортежей. Для каждого такого кортежа $<o, l>$ метод находит сегменты текста в коллекции документов, где o и l обозначают две подстроки в тексте. Как правило, это объекты или именованные сущности, между которыми ищется (оценивается) взаимосвязь. Поскольку o и l обычно находятся (расположены) близко друг к другу в тексте, то метод анализирует слова, которые «связывают» o и l для генерации шаблонов [13].

Ключевое отличие Snowball от метода DIPRE заключается в том, что шаблоны Snowball включают в себя теги NLP. Примером такой модели является *<<МЕСТОПОЛОЖЕНИЕ> построено <ОРГАНИЗАЦИЯ>>*. Эта модель не будет выполняться для любой пары строк, связанных словом «построено». Вместо этого, *<МЕСТОПОЛОЖЕНИЕ>* будет соответ-

ствовать только строке, содержащей описание местонахождения. Кроме того, *<ОРГАНИЗАЦИЯ>* будет соответствовать только строке, содержащей описание типа организации.

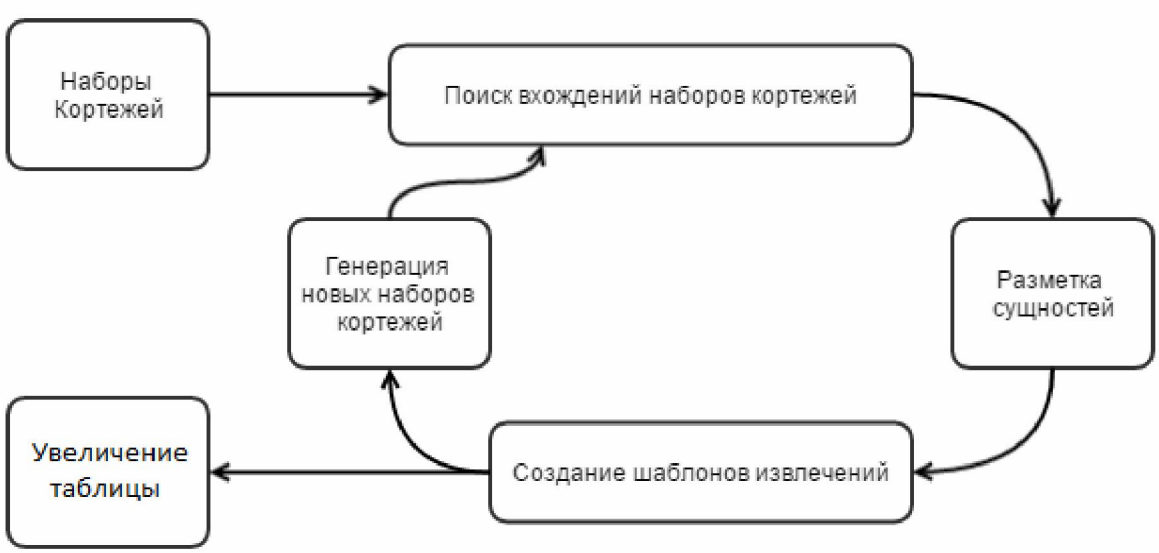

Рис. 1. Основные компоненты системы Snowball

Snowball получает контекст вокруг *<ОРГАНИЗАЦИЯ>* и *<МЕСТОПОЛОЖЕНИЕ>* в шаблонах в гибкой форме, с помощью которого создаются избирательные модели с высоким охватом. В результате незначительные изменения текста, такие как лишняя запятая или слово, нарушают согласование похожих контекстов. Более конкретно, Snowball представляет собой левый, средний и правый «контексты», связанные с шаблоном, как векторное пространство модели информационного поиска. Таким образом, левый, средний и правый контексты это три вектора, связывающие веса (т.е. числа от 0 до 1) с объектами (произвольный набор не пробельных символов). Эти веса указывают на важность каждого члена в соответствующем контексте [9].

После создания шаблонов, Snowball сканирует текст для того, чтобы найти новые кортежи. Во-первых, определяются предложения, которые включают в себя распознаваемую последовательность подстрок. Для данного сегмента текста, с соответствующими сущностями, Snowball генерирует кортеж $T = < l_c, t_1, m_c, t_2, r_c >$, где l_c, m_c, r_c – векторы, хранящие информацию о контексте, находящемся слева, между и справа от пары «термин-объект», а t_1 и t_2 – теги именованных сущностей (объект и термин). Каждый новый кортеж будет иметь ряд закономерностей, которые способствовали его генерации, каждая из них с соответствующими степенями совпадений. Метод использует эту информацию вместе с информацией о селективности шаблонов, чтобы решить, нужно ли добавлять новые кортежи в таблицу.

В качестве примера для генерации кортежа возьмем предложение «*В России с начала года ВВП упал на 3 %*». На основе этого предложения будет сгенерирован кортеж *< {<в, 0.2>}, <России, object>, {<с, 0.1>, <начала, 0.05>, <года, 0.07>}, <ВВП, term>, {<упал, 0.06>} >*.

Здесь, *<России, object>, <ВВП, term>* являются объектом и экономическим термином. *{<в, 0.2>}* представляет собой левый контекст пары «объект-термин», *{<с, 0.1>, <начала, 0.05>, <года, 0.07>}* – средний контекст и *{<упал, 0.06>}*, соответственно правый контекст. Веса при словах являются частотой появления этих слов в соответствующем контексте (в документе или совокупности документов). Соответственно, чем выше веса для отдельных слов и их совокупности в контексте, тем выше вероятность взаимосвязи между термином и объектом.

Выделение признаков терминов и объектов. Основными признаками терминов и объектов являются следующие: результат морфологического разбора слова; его местоположение

в тексте и предложении; сущности, стоящие между термином и объектом. Также, для улучшения работы алгоритма машинного обучения, используются признаки терминов, содержащиеся в их базе данных. К таким признакам можно, в частности, отнести ограничение на часть речи связываемого объекта и типа объекта (имя, название организации или страны и т.п.) [1].

Морфологический анализ изучает структуру слов и определяет морфемы языка (мельчайшие элементы, несущие значимую нагрузку). Любая форма слова может быть выражена в виде комбинации морфем. Слова известны в качестве основных единиц языка. Однако морфемы еще более маленькие синтаксической единицы, показывающие отношения между формами слов. В связи с этим морфологический анализ исследует структуру, формирование и функционирование слов; формулирует правила, которые моделируют язык.

Поиск терминов в тексте. В силу того, что исследуемая предметная область ограничена экономикой, поиск экономических терминов в тексте заметно упрощается по сравнению с абстрактным поиском абсолютно любого термина. Первоначально авторами была составлена структурированная база данных самых распространенных экономических терминов. Описание для удобной структуризации признаков проводилось в формате JSON. Структура описания признаков объектов для определенных экономических терминов представлена на рисунках 2 и 3.

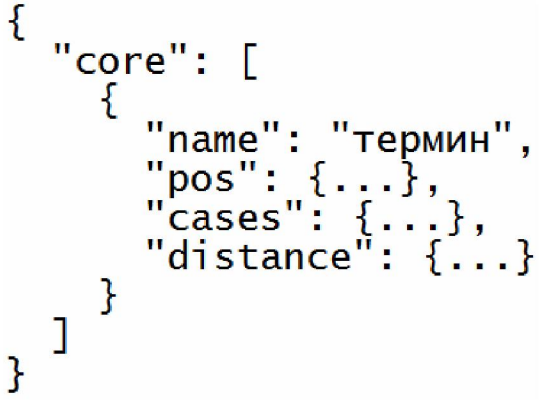

Рис. 2. Верхний уровень структуры описания терминов:
name – простая форма термина; **pos** – массив данных о признаке объекта «часть речи»;
cases – массив д

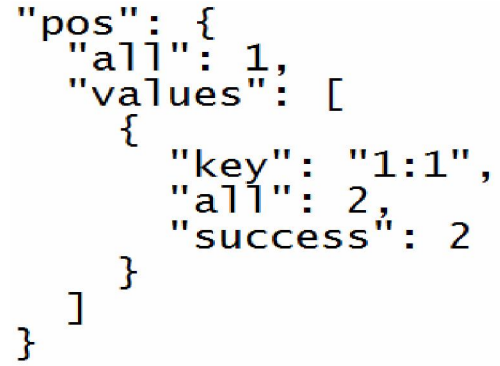

Рис. 3. Структура описания признака «Часть речи» для объектов:
all – количество примеров, содержащих данный признак; **values** – массив значений примеров данного признака; **key** – ключ, состоящий из нормализированного кода признака двух слов; **all** – количество примеров с данным ключом; **success** – количество положительных примеров с данным ключом

Как видно, термины представляют собой набор большого количества полей-признаков объектов для данного термина. В свою очередь все термины объединены в общий массив терминов.

В данной статье рассматриваются не все термины экономической тематики (на текущий момент – в основном термины из сферы макроэкономики). Поэтому для извлечения таких терминов будет применяться нормализация текста (т.е. перевод всех словоформ в некоторые стандартные значения) и поиск по уже известным терминам. Для обучения системы новому термину достаточно в обучающей выборке добавить тег термина у необходимого слова. Если системе этот термин ранее не был известен, то он автоматически будет добавлен в базу.

Для разметки обучающей выборки используется формат xml. Каждое слово обозначается тегом **<word></word>**, атрибутом которого является дополнительная информация. К таким атрибутам можно отнести принадлежность слова к терминам, позиция слова, зависимого от данного термина и т.п. Также в атрибутах записывается морфологический разбор этих слов – в связи с возможной неоднозначностью предложения (пример будет приведен далее).

Ручная разметка таким методом трудоемкая и неудобная. Поэтому в программной реализации алгоритма был создан подмодуль, который автоматически размечает входной текст, и пользователю остается только внести дополнительную информацию о терминах и их связях. Также, при необходимости, пользователь может исправить морфологический разбор слов [13].

```
<text>
  <word position="1" sentence_position="1" main_morph="S,муж,неод" sub_morph="им,ед"
  is_term="True">дефицит</word>
  <word position="2" sentence_position="1" main_morph="S,муж,неод" sub_morph="род,ед"
  is_term="True" to_term="1">бюджет</word>
  <word position="3" sentence_position="1" main_morph="S,гео,жен,неод" sub_morph="род,ед"
  is_term="False" to_term="2,5">россия</word>
  <word position="4" sentence_position="1" main_morph="V,пе" sub_morph="прош,ед,изъяв,муж,сов"
  is_term="False">составлять</word>
  <word position="5" sentence_position="1" main_morph="S,сокр" sub_morph="род,ед"
  is_term="True">ввп</word>
</text>
```

Рис. 4. Пример обучающей выборки терминов:
<word/> – тэг с информацией о слове; **main_morph** и **sub_morph** – морфологический разбор слова;
is_term – является ли слово экономическим термином; **to_term** – позиции слов-терминов, связанных с данным словом

На основе этих данных создается набор терминов, которым обучена система анализа текста. Следовательно, для увеличения охвата б*о*льшего множества терминов необходимо увеличивать размер обучающей выборки, содержащей разнообразные термины.

Пример обучающей выборки приведен на рисунке 4.

Разработка алгоритма для извлечения объектов экономической тематики из текста на русском языке. Чаще всего, такими объектами экономической тематики являются слова, находящиеся в тексте рядом с экономическими терминами. Например, словосочетания «Прибыль фирмы» или «Зарплата сотрудника» содержат в себе экономические термины и связанные с ними объекты экономической тематики. Чаще всего, в качестве этих объектов выступают существительные или прилагательные, согласованные по падежу и числу со связанными с ними экономическими терминами.

Как было указано выше, из числа существующих методов извлечения информации из текста на ЕЯ, для данной задачи хорошо подходит метод шаблонов с контекстной информацией. Поскольку для одного термина существует множество объектов экономической тематики, то для их выделения можно использовать шаблон «*<Термин> контекст <Объект (сущ. | прил. | местоимение)>*» с согласованием падежа и числа этих слов. Поскольку таких

наборов для термина может быть много, то каждое слово из последовательности имеет вес. Для его расчета используется частота появления слова во всех известных контекстах для данного термина, а также частота использования слова совместно со всеми остальными экономическими терминами.

Для выделения слов-кандидатов в отношении связи с термином, производится поиск контекстных слов для данного термина, и выбираются слова, подходящие под шаблон поиска [13].

При обучении системы, модуль поиска объектов экономической тематики подсчитывает веса для контекстных слов. В случае достаточного веса (это регулируется настройкой алгоритма) данное слово запоминается как контекст для рассматриваемого термина. Сохраняются также примеры согласований по падежу и числу между объектом и термином – для дальнейшего шаблонного поиска.

В основе разработанного авторами алгоритма лежит модифицированный метод Snowball [9], описанный выше. В этом модифицированном методе учитывается следующее: контекстная информация словосочетания; особенности каждого термина для коррекции оценки; порядок слов и расстояния между ними.

Алгоритм для извлечения объектов экономической тематики из текста на ЕЯ приведен на рисунке 5.

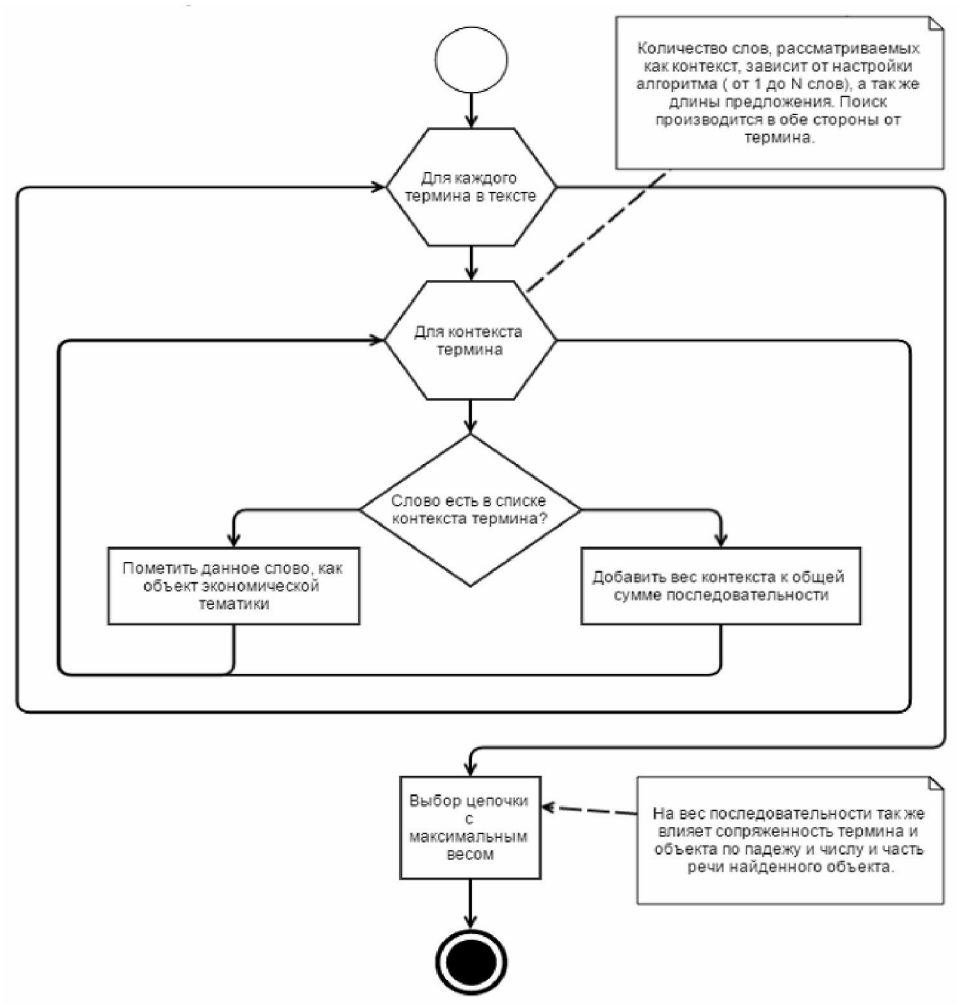

Рис. 5. Блок-схема алгоритма поиска объектов экономической тематики

Влияние признаков на взаимосвязи терминов и объектов. Связь между экономическим термином и объектом нельзя задать шаблонно, поэтому в алгоритме используется машинное обучение. Вне зависимости от выбранной модели машинного обучения, необходимо выделять признаки, которые влияют на принятие решения алгоритмом. В лингвистических задачах чаще всего такими признаками выступает морфологический разбор слов, их позиция в тексте и предложении, а также контекстная информация [15].

Морфологический разбор. Данный признак является номинальным. Он позволяет отсеивать варианты неправильных связей (объектами экономической тематики чаще всего являются существительные, прилагательные и местоимения).

Помимо части речи, морфологический разбор предоставляет информацию о падеже, роде и числе слова. Сопряжения по падежу и числу между термином и объектом, также позволяют сократить количество рассматриваемых пар «термин-объект».

Минусами таких признаков является неоднозначность значения слова. К примеру, слово «мыла» может быть морфологически разобрано как глагол или существительное. Поэтому необходимо рассматривать каждый вариант разбора как отдельно слово. Однако если использовать только результаты морфологического разбора как признак, то алгоритм может допускать погрешность в оценке взаимосвязей (в связи с неоднозначным разбором) [2].

Контекстная информация. Контекстом пары «термин-объект» называют слова, находящиеся слева или справа от пары, а также между ними. При обучении системы запоминаются слова, которые являются контекстом. Поскольку для одной пары может существовать множество контекстов, то их вес рассчитывается на основе частоты появления этих слов в различных контекстах. После этого важность слова в конкретном контексте вычисляется на основе его веса [14].

Контекстная информация позволяет выделять пары «термин-объект», основываясь на правильном контекстном смысле этих слов. Такой признак также номинальный – он представляет собой число (коэффициент контекста для выбранной пары).

Минусом использования такого признака является необходимость проверки слов, выбранных как контекст. Поэтому при обучении системы, в атрибутах слов, являющихся контекстом, можно указывать их веса – для выделения действительно важных контекстных слов. Веса таких контекстных слов не будут изменяться алгоритмом.

К контекстной информации можно также отнести положение термина и объекта в тексте и предложении. Поскольку в русском языке порядок слов в предложении может изменяться, то положение можно хранить как смещение относительно термина. Такой признак позволяет выбирать правильные пары из равных кандидатов, поскольку взаимное положение двух слов в тексте влияет на их смысл.

К примеру, в предложении «*На прошлой неделе дефицит Французского рынка достиг максимальной отметки*», для пары «*Дефицит-Рынок*», контекстной информацией являются такие слова: «*неделе*», «*Французского*» и «*достиг*». На основе информации, получаемой при обучении системы анализа текста, для каждого из слов рассчитывается их вес (как часто они появляются в контексте пар, являющихся положительными примерами отношений «термин-объект»).

Алгоритм установки вероятности взаимосвязи пар терминов и объектов в тексте на русском языке. В качестве метода машинного обучения был выбран метод опорных векторов (SVM) – рисунок 6. Как описано выше, все признаки номинальные, а результатом работы SVM является вектор классификации (+1|-1). Задачей классификации в данной работе можно считать определение принадлежности выбранной пары «термин-объект» к группе с достаточно высоким процентом вероятности взаимосвязи [10].

Для выбора наилучшей пары для термина алгоритм учитывает вес контекстной последовательности, а также дополнительную информацию о термине (признаки объектов для данного термина). Кроме того, алгоритм имеет настройку порогового значения вероятности взаимосвязи пар «термин-объект». Это пороговое значение позволяет сразу отбрасывать пары со слишком низкой вероятностью взаимосвязи для ускорения работы алгоритма [12].

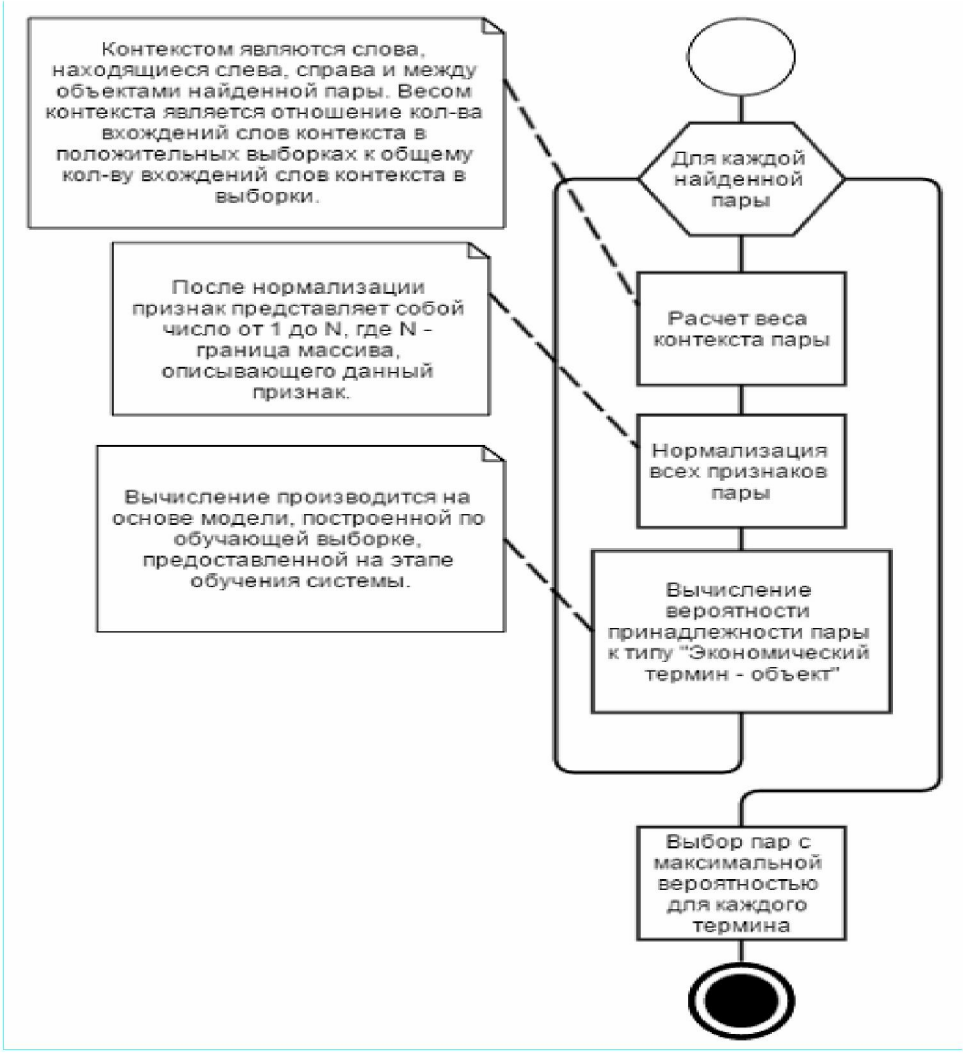

Рис. 6. Алгоритм установки вероятности взаимосвязи пар «термин-объект»

Программная система автоматизированной установки взаимоотношений объектов в тексте. Основной функцией разработанной программной системы (ПС) является поиск пар «экономический термин – объект» в русскоязычных текстах. При этом ставится задача минимизации ошибок 1-го (ошибочное выявление пары) и 2-го (пропуск пары) родов.

Подсистема извлечения терминов выполняет такие функции: нахождение известных системе терминов в тексте; обучение новым терминам на основе обучающей выборки.

Подсистема извлечения объектов экономической тематики выполняет следующие функции: поиск контекстных слов рядом с термином; поиск возможных объектов экономической тематики.

Графическое представление функциональной схемы программы для установления вероятности взаимосвязи «объект-термин» представлено на рисунке 7.

Описание входных и выходных данных. Входные параметры системы анализа текстов можно разделить на два типа: обучающая выборка и тексты для анализа.

При обучении системы на вход подается текст в формате xml. Телом тегов являются слова предложений, а их атрибутами – дополнительная информация об объектах. К таким

данным относятся связь между термином и объектом, вес контекстных слов, дополнительная информация о термине. Пример обучающей выборки был приведен ранее на рисунке 4.

Текст для анализа подается на вход системы анализа в виде текста на ЕЯ – без каких-либо модификаций.

Рис. 7. Функциональная схема программы

Выходные параметры являются также текстом в формате xml. Слова заключается в тег <**word**></**word**>, а для обозначения пары «объект-термин» используется тег <**pair**></**pair**>. Атрибутом тега <**pair**> является вероятность взаимосвязи термина и объекта. Для тега <**word**> в атрибутах указывается нормализированная форма слова, а также позиция в исходном тексте.

Общая архитектура системы. На рисунке 8 приведена общая архитектура разработанной системы анализа текстов, указаны функциональные и информационные взаимосвязи между объектами.

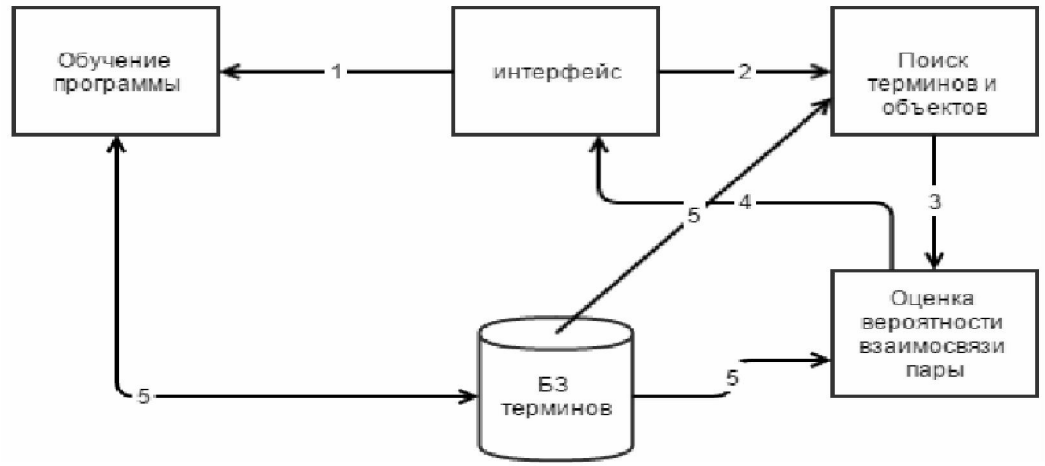

Рис. 8. Общая архитектура системы

Обозначения: 1 – обучающая выборка; 2 – текст на ЕЯ; 3 – набор пар термин-объект; 4 – набор пар термин-объект с вероятностью их взаимосвязи; 5 – информация о термине

Авторами ПС было написано на языке Python. Объем кода - около 1500 строк. Создана JSON база на 100 терминов и 2000 связанных с ними объектов.

Заключение. В процессе выполнения работы был выполнен аналитический обзор ПС-аналогов, предназначенных для установления взаимоотношений объектов в тексте на ЕЯ.

Исследованы также методы выделения терминов, объектов и связей между ними. Разработан алгоритм извлечения экономических терминов и объектов экономической тематики из текстов на русском языке.

Разработан алгоритм установления (определения) вероятности взаимосвязей пар «термин-объект» в тексте на русском языке.

Разработана и протестирована программная система для анализа текстов, реализующая описанные алгоритмы.

Список литературы

1. Ермаков, А. Е. Семантическая интерпретация в системах компьютерного анализа текста / А. Е. Ермаков, В. В. Плешко // Информационные технологии. – 2009. – N 6. – С. 2– 7.

2. Ермаков, А. Е. Автоматическое извлечение фактов из текстов досье: опыт установления анафорических связей / А. Е. Ермаков // Компьютерная лингвистика и интеллектуальные технологии : тр. междунар. конф. «Диалог'2007». – М., 2007. – С. 131–135.

3. Киселев, С. Л. Поиск фактов в тексте естественного языка на основе сетевых описаний / С. Л. Киселев, А. Е. Ермаков, В. В. Плешко // Компьютерная лингвистика и интеллектуальные технологии : тр. междунар. конф. «Диалог'2004». – М., 2004. – С. 72–75.

4. Официальный сайт компании «CoreNLP» [Электронный ресурс]. – 2015. – Режим доступа : http://nlp.stanford.edu/software/corenlp.shtml, свободный.

5. Официальный сайт компании «Calais» [Электронный ресурс]. – 2015. – Режим доступа : http://www.opencalais.com/about, свободный.

6. Официальный сайт компании «NetOwl Extractor» [Электронный ресурс]. – 2015. – Режим доступа : https://www.netowl.com/entity-extraction/, свободный.

7. Официальный сайт корпорации «Ontosminer» [Электронный ресурс]. – 2011. – Режим доступа: http://www.ontos.com/products/ontosminer/, свободный.

8. Официальный сайт компании «Link Parser» [Электронный ресурс]. – 2007. – Режим доступа: http://www.abisource.com/projects/link-grammar/, свободный.

9. Agichtein, E. Snowball: extracting relations from large plain-text collections / Eugene Agichtein, Luis Gravano // In Proceedings of the fifth ACM conference on Digital libraries. – 2000 – P. 85–94.

10. Minard, A.L. Multi-Class SVM for Relation Extraction from Clinical Reports / Anne-Lyse Minard, Anne-Laure Ligozat, Brigitte Grau // Proceedings of Recent Advances in Natural Language Processing. – 12-14 September 2011. – P. 604–609.

11. Dmitriev, A.S. Automatic identification of time and space categories in the natural language text / Dmitriev A.S., Zaboleeva-Zotova A.V., Orlova Y.A., Rozaliev V.L. // Applied Computing 2013 : proceedings of the IADIS International Conference (Fort Worth, Texas, USA, October 23-25, 2013) / IADIS (International Association for Development of the Information Society). – 2013. – P. 187-190.

12. Gildea, D. Automatic Labeling of Semantic Roles / D. Gildea, J. Daniel // Computational Linguistics. – 2002. – Vol. 28. – No 3. – P. 245–288.

13. Lao, N. Reading The Web with Learned Syntactic-Semantic Inference Rules / Ni Lao, Amarnag Subramanya, Fernando Pereira, William W. Cohen // Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. – 12-14 July 2012. – P. 1017–1026.

14. Stenchikova, S. QASR: Spoken Question Answering Using Semantic Role Labeling / S. Stenchikova, Dilek Hakkani-Tur, Gokhan Tur// State University of New York. – 2004. – No 3. – P. 11–17.

15. Wenlei, M. The phrase-based vector space model for automatic retrieval of free-text medical documents / M. Wenlei, W. Wesley // Data & Knowledge Engineering. – 2007. – P. 76–92.

References

1. Ermakov A.E., Pleshko V.V. *Semanticheskaia interpretatciia v sistemakh kompiuternogo analiza teksta* [*The semantic interpretation in the computer analysis of the text*]. Informatcionnye tekhnologii [Information technology], 2009, no. 6, pp. 2-7.
2. Ermakov A.E. *Avtomaticheskoe izvlechenie faktov iz tekstov dose: opyt ustanovleniia anaforicheskikh sviazei* [*Automatic extraction of facts from text files: the experience of the establishment of anaphoric relations*]. Kompiuternaia lingvistika i intellektualnye tekhnologii : tr. mezhdunar. konf. «Dialog'2007» [Computational Linguistics and Intellectual Technologies: Third Intern. Conf. «Dialog'2007»], 2007, pp. 131-135.
3. Kiselev S.L., Ermakov A.E., Pleshko V.V. *Poisk faktov v tekste estestvennogo iazyka na os-nove setevykh opisanii* [*Search the facts in the text of natural language based on network descriptions*]. Kompiuternaia lingvistika i intellektualnye tekhnologii : tr. mezhdunar. konf. «Dialog'2004» [Computational Linguistics and Intellectual Technologies: Third Intern. Conf. "Dialog'2004."], 2004, pp. 72-75.
4. CoreNLP official site. Available at: http://nlp.stanford.edu /software/corenlp.shtml (accessed 2015).
5. Calais official site. Available at: http://www.opencalais.com/about (accessed 2015).
6. NetOwl Extractor official site. Available at: https://www.netowl.com /entity-extraction/ (accessed 2015).
7. Ontosminer official site. Available at: http://www.ontos.com /products/ontosminer/ (accessed 2015).
8. Link Parser official site. Available at: http://www.abisource.com /projects/link-grammar/ (accessed 2015).
9. Agichtein Eugene, Gravano Luis. *Snowball: extracting relations from large plain-text collections.* In Proceedings of the fifth ACM conference on Digital libraries, 2000, pp. 85–94.
10. Minard Anne-Lyse, Ligozat Anne-Laure, Grau Brigitte. *Multi-Class SVM for Relation Extraction from Clinical Reports.* Proceedings of Recent Advances in Natural Language Processing, 12-14 September 2011, pp. 604-609.
11. Dmitriev A.S., Zaboleeva-Zotova A.V., Orlova Y.A., Rozaliev V.L. *Automatic identification of time and space categories in the natural language text.* Applied Computing 2013: proceedings of the IADIS International Conference (Fort Worth, Texas, USA, October 23-25, 2013), IADIS (International Association for Development of the Information Society), 2013, pp. 187-190.
12. Gildea D., Daniel J. *Automatic Labeling of Semantic Roles.* Computational Linguistics, 2002, Vol. 28, No 3, pp. 245-288.
13. Lao Ni, Subramanya Amarnag, Pereira Fernando, Cohen William W. *Reading The Web with Learned Syntactic-Semantic Inference Rules.* Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 12-14 July 2012, pp. 1017–1026.
14. Stenchikova S., Dilek Hakkani-Tur, Gokhan Tur. *QASR: Spoken Question Answering Using Semantic Role Labeling.* State University of New York, 2004, No 3, pp. 11–17.
15. Wenlei M., Wesley W. *The phrase-based vector space model for automatic retrieval of free-text medical documents.* Data & Knowledge Engineering, 2007, pp. 76–92.

УДК 004.912

МЕТОДЫ АДАПТАЦИИ ТЕКСТОВОЙ ИНФОРМАЦИИ ДЛЯ ЛИЦ С ОГРАНИЧЕННЫМИ ВОЗМОЖНОСТЯМИ ПО ЗРЕНИЮ[1]

Статья поступила в редакцию 04.11.2015 г., в окончательном варианте 15.11.2015 г.

Орлова Юлия Александровна, кандидат технических наук, кандидат педагогических наук, доцент, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. им. Ленина, 28, e-mail: yulia.orlova@gmail.com

[1] Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов № 13-07-00351, 14-07-97017, 15-07-07519, 15-07-05440.