Автоматическое выделение словосочетаний из текстов славянского происхождения: сравнение подходов[1]

В. Б. Барахнин, О. Ю. Кожемякина,
Е. В. Рычкова, Ю. С. Борзилова

Институт вычислительных технологий
Сибирского отделения Российской академии наук (ИВТ СО РАН)
630090, Новосибирск, пр. Академика Лаврентьева, 6

e-mail: bar@ict.nsc.ru

Аннотация. В рамках постановки и решения задачи аналитического обзора рассмотрен ряд исследований, проводимых с целью извлечения словосочетаний (коллокаций) из текстов на русском, украинском и польском языках. Описан ход развития текущих исследований для автоматизации анализа поэтических текстов, проводимых на базе ИВТ СО РАН.
Ключевые слова: словосочетания, славянские языки, автоматическая обработка текста, коллокации, работа с корпусами текстов, средства автоматизации лингвистических исследований.

1. Введение

Комплексные лингвистические исследования, в частности, анализ русских поэтических текстов, включают в себя комплекс подзадач. Например, составление словарей языка поэтов, который может включать не только слова, но и словосочетания, которые характеризуют речь данного автора. Автоматизация названной подзадачи значительно облегчит автоматизацию анализа поэтических текстов.

В рамках постановки и решения задачи аналитического обзора рассмотрен ряд исследований, проводимых с целью извлечения словосочетаний (коллокаций) из текстов. Общим в анализируемых работах является объект исследования, а именно: работа с корпусами текстов, написанных на славянских языках восточного (русский и украинский) и западного (польский) происхождения. Рассматривались достаточно современные исследования, поскольку мы не ставим целью углубляться в исторический аспект проблемы.

Затрагивая проблему поиска информации в современном информационном пространстве, нельзя не отметить, что современные информационные технологии предоставляют исследователю мощный аппарат для «манипулирования данными».

[1] Работа выполнена в рамках темы государственного задания № АААА-А17-117120670141-7 (№ 0316-2018-0009) при частичной поддержке гранта Российского фонда фундаментальных исследований № 18-07-01457 и гранта Министерства образования и науки Республики Казахстан № BR05236839.

В электронной форме данные приобретают новую форму, обеспечивая более широкое распространение и эффективное использование. Может сложиться впечатление, что развитие информационных технологий (ИТ) само по себе должно вывести работу с информацией на качественно новый уровень, но это не всегда так, поскольку ИТ в полной мере не могут предоставить адекватный аппарат для оперирования с информацией и информационными ресурсами [1].

В связи с проблемой поиска и формирования информации важно подчеркнуть, что анализ уже существующей информации в общем был и остается актуальным для научных работников.

Современный подход к исследованию текстовых сообщений предполагает использование многоуровневой модели информации, изложенной, например, в работе германского исследователя В. Гитта [2]. Анализируя эту модель, можно увидеть, что ее нижний уровень соответствует шенноновскому значению термина «информация», три последующих — семиотической триаде (синтактика — семантика — прагматика), а верхний уровень носит, скорее, философский характер. При этом наличие в некотором сообщении информации высокого уровня влечет за собой наличие информации всех низших высоких уровней, но, разумеется, не наоборот [3].

Модель В. Гитта является важным этапом в понимании обобщенной структуры информации и зависимости ее, в том числе, от отправителя. Однако модель не учитывает уровни лингвистического представления фраз и их частей в тексте, что важно при непосредственной работе с корпусами текстов. В таком ракурсе наиболее значима модель И. А. Мельчука «Текст-Смысл», которая описывает уровни лингвистического представления фраз и их частей согласно компонентам: фонология, морфология, синтаксис, семантика. Значимость модели в данном исследовании обусловлена ее точностью и прикладным значением — описать структуру языка на синтаксическом уровне для понимания предметного поля исследования.

И. А. Мельчук в своей работе [4] приводит описание уровней лингвистического представления фраз и их частей с сопоставлением модулей модели «Текст-Смысл» (рис. 1).

В контексте данной статьи будет уместно остановиться на описании поверхностно-синтаксической структуры, являющейся компонентом поверхностно-синтаксического представления (ПСинтП), поскольку наш анализ посвящен обзору исследований, имеющих дело со структурными единицами в виде предложений и словосочетаний. ПСинтП — это упорядоченная четверка вида:

$$\text{ПСинтП} = \langle \text{ПСинтС}; \text{ПСинт-КоммС}; \text{ПСинт-ПросС}; \text{ПСинт-АнафС} \rangle,$$

где ПСинтС — поверхностно-синтаксическая структура; ядро ПСинтП выражает организацию фразы в терминах входящих лексических единиц и связывающих их между собой;

ПСинт-КоммС — поверхностно-синтаксическая-коммуникативная структура фразы; задает разделение ПСинтС на коммуникативные зоны;

ПСинт-ПросС — поверхностно-синтаксическая-просодическая структура фразы; описывает семантически нагруженные просодии (ударения, например, декларативное, экспрессивное, вопросительное);

ПСинт-АнафС — поверхностно-синтаксическая-анафорическая структура фразы; указывает кореферентность лексических единиц, составляющих ПСинтС.

Рисунок 1. Уровни лингвистического представления и модули модели «Текст-Смысл». Обозначения (снизу вверх): СемП — семантическое представление; ГсинтП — глубинно-синтаксическое представление; ПСинтП — поверхностно-синтаксическое представление; ГморфП — глубинно-морфологическое представление; ПморфП — поверхностно-морфологическое представление; ГфонП — глубинно-фоническое представление; ФонолП — фонологическое представление; ПфонП — поверхностно-фоническое представление; ФонетП — фонетическое представление

Поверхностно-синтаксическая структура (ПСинтС) некоторой фразы представляет собой линейное неупорядоченное дерево зависимостей [4]. Вершины дерева помечены именами всех реальных лексем, снабженными семантически полными граммемами (грамматическими значениями). Дуги помечены именами поверхностно-синтаксических отношений (ПСинтО), которые обозначают конкретные синтаксические конструкции данного языка. ПСинтО специфичны для каждого

языка (как и фонемы, морфемы, лексемы). ПСинтС является линейно неупорядоченным деревом зависимостей.

Описанные схемы Гитта и Мельчука дают теоретическое основание для проводимого ниже обзора.

2. Обзор экспериментальных исследований по тематике

На текущий момент существующие алгоритмы извлечения словосочетаний можно разделить на категории по нескольким признакам.

1. *По обучению*. Можно выделить необучаемые, самообучаемые и обучаемые.
   - Необучаемые методы подразумевают контекстно-независимое выделение ключевых слов на основе заранее составленных моделей и правил.
   - Обучаемые методы при принятии решений о выделение ключевого словосочетания из текста используют различные лингвистические ресурсы.
   - Самообучаемые алгоритмы — это те алгоритмы, обучение которых происходит без учителя или с подкреплением.

2. *По математическому аппарату распознавания*. Наиболее важными алгоритмами являются статистические, структурные и нейросетевые.
   - Статистические методы учитывают частоту встречаемости морфологических, лексических и синтаксических единиц.
   - Второй тип — структурный. В основе его лежит представление текста как о системе взаимосвязанных элементов — слов. Выделяют графовые и шаблонные (синтаксические) алгоритмы.
   - Нейросетевые алгоритмы используют свойство нейронных сетей к обобщению и выделяют скрытые взаимосвязи.

3. *По лингвистическим ресурсам*. Алгоритмы извлечения словосочетаний из текста могут не использовать какие-либо лингвистические ресурсы, использовать или использовать разного рода словари, онтологии и тезаурусы, а также корпуса текстов (без разметки или с разметкой).

Что касается автоматического извлечения ключевых слов, динамика исследований в этой области подробно описывается в работе [5].

2.1. Польша

Проект SyntLex [6] (2007 г.) посвящен извлечению вербо-номинальных словосочетаний из корпуса текстов на польском языке. Эксперименты проводились в институте IPI PAN [7]. Словарь получил название BR (Basic Resource). Он содержит в себе описание предикативных существительных; 7 500 существительных из

40 000 записей бумажного словаря были помечены как предикативные; также был сформирован словарь из 2 826 существительных, обозначающих абстрактные виды деятельности. Данные существительные использовались как шаблоны фильтров для проведения эксперимента. В ходе исследования [6] выполнялись следующие шаги:

Шаг 1. Применение шаблонов фильтров; выходной файл A имел размерность 235 742 слова.

Шаг 2. Нормализация глаголов, удаление повторений; выходной файл B имел размерность 82 464 слова.

Шаг 3. Удаление всех слов, кроме существительного и глагола, приведение существительного в начальную форму; выходной файл C имел размерность 12 485 слов.

Шаг 4. Ручная обработка файла C для определения возможных словосочетаний.

Шаг 5. Восстановление контекста для отобранных словосочетаний (из файла B); выходной файл B' имел размерность 12 783 слова.

Исследователи предлагают данный алгоритм как методологическую парадигму при автоматическом выделении сочетаний (на основе словаря).

Другим исследованием для польских корпусов текстов является работа [8] 2008 г. Авторами подчеркиваются недостатки статистической меры вероятности совпадения (co-occurrence). Такой подход хорошо работает с английским языком, но имеет недостатки при работе с флективными языками, как, например, польский. Во-первых, фиксированный порядок слов в предложении не работает для польского синтаксиса; во-вторых, польские лексемы выражаются большим количеством словоформ. Для исследования использовалась авторская система Kolokacje. Процесс распознавания словосочетаний был разделен на 3 этапа:

- Приведение словоформ к леммам.
- Статистическое распознавание — частые последовательности лемм отмечаются как возможные словосочетания.
- Статистическая синтаксическая фильтрация — данные о частотах возможных словосочетаниях, удовлетворяющих заданному синтаксическому ограничению, собираются из корпуса, проверяются на статистическую значимость и создается список размеченных словосочетаний.

Результаты практического эксперимента сравнивались со словарями, вручную обработанными лингвистами (совпадение около 25%). Авторы делают акцент на дальнейшем использовании системы в полуавтоматическом режиме.

Важным элементом рассматриваемых исследований является то, что авторы отделяют методики работы с польским текстом, от методик работы с английским текстом.

2.2. Украина

В статье [9] предложена общая модель перевода с одного языка на другой, используя «знание-ориентированный подход». Реализация модели предполагает построение базы данных, которая включает словарь лексико-семантических валентностей глаголов и словарь семантических интерпретаций. Основа распознанных синтаксических правил определяет устойчивые словосочетания и понятия в заданной предметной области. Авторами высказаны рекомендации по использованию модели в системах машинного перевода (скорее, только для украинского языка, так как приведенные в статье эксперименты касались исключительно украинских текстов).

В исследовании [10] используется модель, основанная на логико-алгебраических уравнениях конечной алгебры предикатов. Проводится сравнение между украинским и английским языками в плане принципа построения словосочетаний (коллокаций) «глагол-существительное» и «существительное-прилагательное». Авторы условно разделили методы автоматического извлечения коллокаций на 2 группы:

- статистические методы (co-occurrence frequencies): оконные методы (window-based methods), потенциальные меры взаимной информации (Pointwise mutual information (PMI) measures), измерение T-оценки (T-score measure), распределение хи-квадратов (Chi-squared distribution);
- аналитические методы, основанные на синтаксической структуре коллокаций.

Предложенная логико-лингвистическая модель формализует семантическую эквивалентность словосочетаний с помощью их семантических и грамматических характеристик. Основная идея этого подхода заключается в том, что между словосочетаниями, имеющими семантические корреляции, существует общий контент (значение). Для формального выражения сходства сочетаний авторы используют логико-алгебраические уравнения конечной алгебры предикатов. В статье высказано предположение, что реализация модели позволит автоматически распознавать семантически эквивалентные сочетания.

Для некоторых украинских работ характерно, в отличие от работ польских исследователей, совпадение методики выделения словосочетаний (коллокаций) без привязки к особенностям конкретного языка.

2.3. Россия

В российском научном сообществе задачи по проблематике выделения словосочетаний (в частности, ключевых), также активно обсуждались. Нельзя не отметить вклад исследовательского отдела компании Яндекс: в 2010 г. был разработан свободно распространяемый продукт MyStem [11]. Программа работает на основе словаря и способна формировать морфологические гипотезы о незнакомых словах. Свободное распространение дало толчок для развития дальнейших исследований в отечественной компьютерной лингвистике. Развивая идею MyStem, в 2013 г. был разработан морфологический анализатор pymorphy2 [12], заявленный для использования применительно к русскому и украинскому языкам. Среди полезных особенностей программы можно отметить следующие возможности [13]:

- приводить слово к начальной форме;
- ставить слово в нужную форму (например, менять падеж);
- получать грамматическую информацию о слове (число, род, падеж, часть речи).

В работе программы используется словарь OpenCorpora [14]; для незнакомых слов строятся гипотезы. Проект также является свободно распространяемым, исходный код доступен на GitHub [15].

В работе [16] описывается ряд экспериментов, проведенных на базе Санкт-Петербургского государственного университета в 2014 г., с целью сравнить методы автоматического выделения устойчивых глагольно-именных словосочетаний. В качестве меры ассоциации и ранжирования при выделении словосочетаний была выбрана мера Mutual Information (*MI)* — коэффициент силы синтагматической связанности:

$$MI = \log_2 (f(n,c) \cdot N)/(f(n) \cdot f(c)),$$

где n — ключевое слово (node); c — коллокат (collocate); $f(n,c)$ — частота встречаемости ключевого слова n в паре с коллокатом c; $f(n)$; $f(c)$ — абсолютные (независимые) частоты ключевого слова n и слова c в корпусе (тексте); N — общее число словоформ в корпусе (тексте).

Инструментом для выделения глагольно-именных словосочетаний послужило программное средство IntelliText [17], разработанное Центром переводческих исследований в университете г. Лидса (Великобритания). Эксперименты по выделению глагольно-именных словосочетаний проводились для глаголов *выполнять, нарушать, принимать*. На основе проведенных экспериментов авторы сделали следующие выводы:

1. Качество выделения устойчивых глагольно-именных словосочетаний с учетом части речи коллоката выше, чем без учета его частеречной принадлежности.

2. Увеличение ширины контекста при учете частеречной принадлежности коллоката в основном повышает полноту, но понижает точность и F-меру выделяемых словосочетаний.

3. Увеличение ширины контекста при учете части речи коллоката имеет неоднозначный характер.

4. Результаты данных экспериментов обнаруживают, что статистическая мера *MI* является эффективной мерой ассоциации и ранжирования при выделении глагольно-именных словосочетаний.

Проблема дальнейшего использования вышеописанного исследования затрудняется тем, что авторы использовали инструмент только для выделения устойчивых глагольно-именных словосочетаний (для 3 глаголов). Возможно, проведение экспериментов с другими словосочетаниями по каким-то причинам было невозможно.

Также представляют интерес одни из последних исследований О. А. Митрофановой: в коллективе были проведены эксперименты по применению алгоритмов RAKE [18] и KEA [19] к русскоязычным корпусам текстов. Рассмотрим каждое из исследований подробнее.

Алгоритм RAKE основан на предположении о том, что ключевые выражения часто представляют собой не только отдельные слова, но и фразы. На первом этапе обработки текста необходимо выделить фразы-кандидаты в ключевые словосочетания. Для этого текст разбивается на отрывки по знакам препинания и словарю, содержащему так называемые стоп-слова (артикли, местоимения, вводные слова, и т. д.). Полученные цепочки слов являются кандидатами на роль ключевых словосочетаний. Для каждого слова на основе общей частоты слова и средней длины фразы, в которую оно входит, рассчитывается вес, в то время как вес фразы-кандидата, в свою очередь, рассчитывается как сумма весов, входящих в нее слов. В исходном виде алгоритм пригоден для автоматической обработки англоязычных корпусов текстов. Авторами были предприняты шаги, направленные на модификацию RAKE для работы с русскоязычным материалом. Предобработка текста включает разбиение текста на условные слова (по пробелам) и проставление границ условных синтаксических групп (используется знак-разделитель “|”). На этапе морфологического анализа входного текста используется морфоанализатор pymorphy2 [12].

Выделялись группы:

- прилагательное + существительное;
- существительное + прилагательное + существительное (при наличии согласования);
- одиночные леммы.

Правила составлялись на основе грамматики синтаксического парсера NLTK4RUSSIAN [20]. В некоторых случаях возможным недостатком является выделение слишком длинных конструкций, которые формально имеют ту же синтаксическую структуру, что и словосочетания, состоящие из двух-трех слов. Полученные данные дают авторам основания считать результаты работы алгоритма RAKE приемлемыми, а сам алгоритм RAKE в русскоязычной модификации пригодным для использования в лингвистических исследованиях [18].

Другой выбранный алгоритм — KEA (Keyphrase Extraction Algorithm) также исправно работает на англоязычных текстах [19]. В своей работе авторы предприняли попытку по адаптации алгоритма к русскоязычным текстам. На каждом из этапов работы выбираются кандидаты в ключевые слова и словосочетания, для каждого из которых затем вычисляются значения определенных признаков. Выбор кандидатов, в свою очередь, предполагает три шага:

1. Предварительная обработка подаваемых на вход документов (токенизация, удаление небуквенных символов и др.); использовался морфологический анализатор pymorphy2.

2. Определение кандидатов с помощью следующего набора правил:

- ограничение максимальной длины словосочетания (как правило, три слова);
- отбрасывание кандидатов имен собственных.

3. Выравнивание регистра и стемминг.

Для каждого кандидата вычисляются значения двух основных признаков, используемых в дальнейшем как в обучающей выборке, так и для тестового набора документов: метрика TF-IDF и расстояние от начала документа до первого появления рассматриваемого слова или словосочетания в нем (использование машинного обучения). В качестве оценки качества работы алгоритма авторы использовали автоматизированный способ оценки, который использовал тематических модели, основанные на алгоритме LDA.

Рассматривая работы по анализу текста, нельзя не упомянуть о международной конференции по компьютерной лингвистике ДИАЛОГ [21]. Среди последних работ, посвященных выделению словосочетаний, можно выделить [22–24]. Следует подчеркнуть, что эти работы основаны на методах машинного обучения.

Работы, не применяемые отдельно для задачи выделения словосочетаний (или выполняемые как промежуточные), также имели место в научной среде. Так, исследование [25] 2010 г. посвящено кластеризации текстовых документов на основе составных ключевых термов. Для автоматического выделения ключевых слов авторы использовали ранее упоминаемый продукт MyStem. Ключевые словосочетания отбирались по морфологическим шаблонам с учетом словоформ языка. Авторы

применяли данный продукт для достижения промежуточной цели — выделения ключевых слов из текста.

Этим же коллективом авторов в 2015 г. [26] решалась задача извлечения ключевых слов (словосочетаний) из корпуса текстов однородной тематики с целью дальнейшего использования извлеченных ключевых слов в качестве возможных значений атрибутов сущностей, описываемых в создаваемой онтологии предметной области, предназначенной для организации фактографического поиска в расширенном корпусе текстов соответствующей тематики. Предлагаемая технология основана на применении метода опорных векторов для разметки в текстах частей речи с последующим использованием метода случайных блужданий для извлечения семантически связанных ключевых слов (словосочетаний). К набору этих словосочетаний с целью отнесения конкретного словосочетания к определенному атрибуту описываемой в тексте сущности применяется обученная нейронная сеть со скрытым слоем.

В настоящее время на базе Института вычислительных технологий СО РАН (г. Новосибирск) продолжаются работы по выделению всех словосочетаний из поэтических корпусов. Работа выполняется в рамках общего проекта по автоматизации работы с проведением анализа текстов поэтического характера (рис. 2).

В общем случае работа с поэтическими текстами проходит следующие этапы:

Этап 1. Пакетная выгрузка поэтических текстов из базы данных. Анализ с помощью системы [27]. Полученная метроритмическая статистика выгружается обратно в базу данных.

Этап 2. Автоматическое извлечение словосочетаний с учетом поэтической синонимии. Выгрузка полученных результатов в базу данных.

Этап 3. Определение жанрово-стилевых характеристик с применением методов машинного обучения. Выгрузка результатов в базу данных.

Этап 4. Организация работы с полученными характеристиками с возможностью их сравнения.

В качестве основного средства морфологического анализа применяется технология компании Яндекс (свободно распространяемый парсер).

Семантический анализ (рис. 2) включает извлечение словосочетаний и характеристик слов, связанных с синонимией: эти процессы тесно связаны между собой и взаимозависимы.

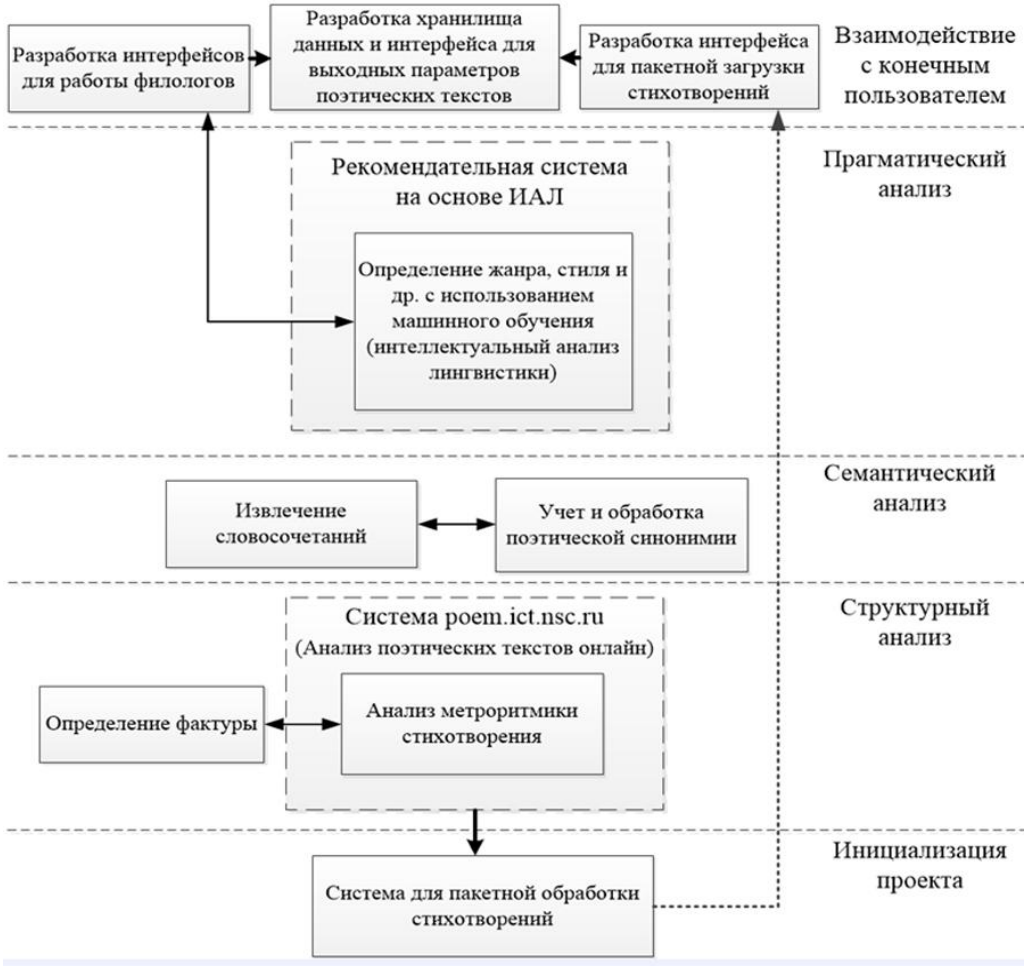

Рисунок 2. Общая схема проекта ИВТ СО РАН для автоматизации работы по проведению анализа поэтических текстов

3. Заключение

Подводя итог, можно сделать несколько выводов. С точки зрения задействованного в анализируемых работах теоретического аппарата использовались методы:

- статистические (объединение модели совместной встречаемости и критерия χ^2);
- лингвистические (основанные на синтаксических алгоритмах и словарных данных);
- гибридные (KEA, RAKE).

Выбор алгоритма определялся особенностями решаемой задачи. В вышеописанных работах применялись все из этих методов (в разных их сочетаниях). С точ-

ки зрения нашего аналитического обзора мы выделяем следующие принципиальные положения из рассмотренных работ:

- в большинстве работ процедура выделения словосочетаний ограничивалась одним их типом (глагольно-именные и пр.);
- поставленные в работах задачи ограничивались поиском только ключевых словосочетаний, т. е. часть сочетаний исключалась из анализа;
- как для славянских, так и для английского языков применялись однотипные алгоритмы (такое обобщение целесообразно только для некоторых видов словосочетаний);
- исследования (с использованием алгоритмов KEA и RAKE) требуют высокой технической подготовки (высокий уровень вхождения в область);
- не во всех работах обозначены используемые корпусы текстов.

Таким образом, для задач автоматизации комплексного анализа русских поэтических текстов, предусматривающих в частности, словари языка поэта, которые в идеале должны включать не только слова, но и словосочетания, характеризующие поэтическую речь данного автора, следует разработать алгоритм, позволяющий выделять из текста все словосочетания, а также выявлять среди них эквивалентные, отличающиеся только грамматической формой. В настоящее время коллективом авторов данной статьи ведутся работы по созданию указанных алгоритмов.

Литература

[1] *Шокин Ю. И., Федотов А. М., Барахнин В. Б.* Проблемы поиска информации. — Новосибирск : Наука, 2010.

[2] *Gitt W.* Ordnung und Information in Technik und Natur // In: Gitt W. (Hrsg.): Am Anfang war die Information. Graefeling: Resch KG, 1982. P. 171–211.

[3] *Барахнин В. Б., Кожемякина О. Ю.* Об автоматизации комплексного анализа русского поэтического текста // CEUR Workshop Proceedings. 2012. Т. 934. С. 167–171. http://ceur-ws.org/Vol-934/paper27.pdf

[4] *Мельчук И. А.* Язык: от смысла к тексту. — М. : Языки славянских культур, 2012. http://biblioclub.ru/index.php?page=book&id=219899

[5] *Ванюшкин А. С., Гращенко Л. А.* Методы и алгоритмы извлечения ключевых слов // *Новые информационные технологии в автоматизированных системах.* 2016. № 19. С. 85–93. https://cyberleninka.ru/article/v/metody-i-algoritmy-izvlecheniya-klyuchevyh-slov

[6] *Vetulani Z., Obrębski T., Vetulani G.* Towards a Lexicon-Grammar of Polish: Extraction of Verbo-Nominal Collocations from Corpora // Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference. 2007. P. 267–268. https://pdfs.semanticscholar.org/6a85/761345a366948cf9a30bffb808ffeaea67d5.pdf

[7] Instytut Podstaw Informatyki PAN. https://ipipan.waw.pl/

[8] *Broda B., Derwojedowa M., Piasecki M.* Recognition of structured colloca-tions in an inflective language // *Systems Science*. 2008. Vol. 34(4). https://www.researchgate.net/publication/266593107_Recognition_of_structured_collocations_in_an_inflective_language

[9] *Толубко В. Б., Литвиненко Л. О.* Розробка моделі автоматичного синтаксичного аналізу і синтезу тексту в системі машинного перекладу // *Вісник Київського національного університету імені Тараса Шевченка.* 2013. Вып. 2(31). С. 57–59. http://www.library.univ.kiev.ua/ukr/host/10.23.10.100/db/ftp/visnyk/viyskovi_31_2013.pdf

[10] *Khairova N., Petrasova S., Gautam A. P. S.* The logic and linguistic mod-el for automatic extraction of collocation similarity // *Econtechmod*. 2015. Vol. 4. No. 4. P. 43–48. http://journals.pan.pl/dlibra/publication/99282/edition/85586/content

[11] Программа MyStem морфологического анализа текста на русском языке. https://tech.yandex.ru/mystem/

[12] Морфологический анализатор pymorphy2. http://pymorphy2.readthedocs.io/en/latest/

[13] *Korobov M.* Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts. AIST 2015. Communications in Computer and Information Science. 2015. Vol. 542. P. 320–332. http://dx.doi.org/10.1007/978-3-319-26123-2_31

[14] OpenCorpora: открытый корпус русского языка. http://opencorpora.org/

[15] Морфологический анализатор pymorphy2 на сайте GitHub. https://github.com/kmike/pymorphy2

[16] *Кощеева С. С.* Сравнение методов автоматического выделения глагольно-именных словосочетаний // Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Труды XVII Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2014). — СПб. : Университет ИТМО, 2014. С. 298–303. http://ojs.ifmo.ru/index.php/IMS/article/view/270/266

[17] Программное средство IntelliText. http://corpus.leeds.ac.uk/it/

[18] *Москвина А. Д., Митрофанова О. А., Ерофеева А. Р., Харабет Я. К.* Автоматическое выделение ключевых слов и словосочетаний из русскоязычных корпусов текстов с помощью алгоритма RAKE // Труды международной конференции «Корпусная лингвистика–2017». — СПб, 2017. С. 268–274.

[19] *Соколова Е. В., Митрофанова О. А.* Автоматическое извлечение ключевых слов и словосочетаний из русскоязычных текстов с помощью алгоритма KEA // Компьютерная лингвистика и вычислительные онтологии. Вып. 1. Труды XX Международной объединенной научной конференции «Интернет и современное общество» (IMS-2017). — СПб. : Уни-т ИТМО, 2017. С. 157–165. http://openbooks.ifmo.ru/ru/file/6522/6522.pdf

[20] *Москвина А. Д., Орлова Д., Паничева П. В., Митрофанова О. А.* Разработка ядра синтаксического анализатора для русского языка на основе библиотек NLTK // Компьютерная лингвистика и вычислительные онтологии. Труды XIX Международной объединенной

научной конференции «Интернет и современное общество» (IMS-2016). — СПб. : Университет ИТМО, 2016. С. 44–45. http://openbooks.ifmo.ru/ru/file/4103/4103.pdf

[21] Международная конференция по компьютерной лингвистике ДИАЛОГ. http://www.dialog-21.ru

[22] *Enikeeva E. V., Mitrofanova O. A.* Russian Collocation Extraction Based on Word Embeddings // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2017”. http://www.dialog-21.ru/media/3908/enikeevaev mitrofanovaoa.pdf

[23] *Kazennikov A. O.* Part-of-Speech Tagging: The Power of the Linear SVM-based Filtration Method for Russian Language // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2017”. http://www.dialog-21.ru/media/3916/kazennikovao.pdf

[24] *Stepanova M. E., Budnikov E. A., Chelombeeva A. N., Matavina P. V., Skorinkin D. A.* Information Extraction Based on Deep Syntactic-Semantic Analysis // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016”. http://www.dialog-21.ru/media/3431/stepanovameetal.pdf

[25] *Барахнин В. Б., Ткачев Д. А.* Кластеризация текстовых документов на основе составных ключевых термов // *Вестник Новосибирского государственного университета. Серия: Информационные технологии.* 2010. Т. 8. № 2. С. 5–14. https://nsu.ru/xmlui/bitstream/handle/nsu/284/01.pdf

[26] *Барахнин В. Б., Пастушков И. С.* Технология автоматизированного наполнения онтологии фактографической поисковой системы // *Вестник Новосибирского государственного университета. Серия: Информационные технологии*. 2015. Т. 13. № 4. С. 5–13. https://nsu.ru/xmlui/bitstream/handle/nsu/10148/2015_13_4_01.pdf

[27] Анализ поэтических текстов онлайн. http://poem.ict.nsc.ru/

Авторы:

Владимир Борисович Барахнин — доктор технических наук, доцент, ведущий научный сотрудник лаборатории информационных ресурсов, Институт вычислительных технологий СО РАН; профессор кафедры общей информатики, Новосибирский национальный исследовательский государственный университет

Ольга Юрьевна Кожемякина — кандидат филологических наук, старший научный сотрудник лаборатории информационных ресурсов, Институт вычислительных технологий СО РАН

Елена Владимировна Рычкова — кандидат физико-математических наук, доцент, научный сотрудник лаборатории информационных ресурсов, Институт вычислительных технологий СО РАН; доцент Гуманитарного института, Новосибирский национальный исследовательский государственный университет

Юлия Сергеевна Борзилова — аспирант, Институт вычислительных технологий СО РАН

The Automatic Extraction of the Collocations from the Texts of Slavic Origin: The Comparison of Approaches

V. B. Barakhnin, O. Yu. Kozhemyakina, E. V. Rychkova, Yu. S. Borzilova

Institute of Computational Technologies of the Siberian Branch of the Russian Academy of Sciences, 6, Ac. Lavrentieva ave., Novosibirsk, Russia, 630090
e-mail: bar@ict.nsc.ru

Abstract. As part of the statement and the solution of the task of analytical review, a number of studies, which were conducted in the purpose of the extraction of the phrases (collocations) from the texts in Russian, Ukrainian and Polish, are considered. The development of current researches for the automation of the analysis of poetic texts, conducted on the basis of ICT SB RAS, is described.
Keywords: R-FMEA, FMEA, specific customer requirements, risk, quality.

References

[1] *Shokin Yu. I., Fedotov A. M., Barakhnin V. B.* (2010) Problemy poiska informatsii. Novosibirsk. [In Rus]
[2] *Gitt W.* (1982) Ordnung und Information in Technik und Natur. In: Gitt W. (Hrsg.): Am Anfang war die Information. Graefeling: Resch KG. P. 171–211.
[3] *Barakhnin V., Kozhemyakina O.* (2012) Ob avtomatizatsii kompleksnogo analiza russkogo poeticheskogo teksta. CEUR Workshop Proceedings. 934:167–171. http://ceur-ws.org/Vol-934/paper27.pdf [In Rus]
[4] *Melchuk I. A.* (2012) Yazyk: ot smysla k tekstu. Moscow, Yazyki slavyanskikh kultur. [In Rus]
[5] *Vanyushkin A. S., Grashchenko L. A.* (2016) *Novyye informatsionnyye tekhnologii v avtomatizirovannykh sistemakh*. 19:85–93. [In Rus]
[6] *Vetulani Z., Obrębski T., Vetulani G.* (2007) Towards a Lexicon-Grammar of Polish: Extraction of Verbo-Nominal Collocations from Corpora. In Proc. of the Twentieth International Florida Artificial Intelligence Research Society Conference. P. 267–268. https://pdfs.semanticscholar.org/6a85/761345a366948cf9a30bffb808ffeaea67d5.pdf
[7] https://ipipan.waw.pl/
[8] *Broda B., Derwojedowa M., Piasecki M.* (2008). *Systems Science*. **34**(4).
[9] *Tolubko V. B., Lytvynenko L. O.* (2013) *Herald of Taras Shevchenko National University of Kyiv*. **2**(31):57–59. [In Ukr]
[10] *Khairova N., Petrasova S., Gautam A. P. S.* (2015) *Econtechmod*. **4**(4):43–48. http://journals.pan.pl/dlibra/publication/99282/edition/85586/content
[11] https://tech.yandex.ru/mystem/
[12] http://pymorphy2.readthedocs.io/en/latest/
[13] *Korobov M.* (2015) Morphological Analyzer and Generator for Russian and Ukrainian Languages. In Proc. Analysis of Images, Social Networks and Texts. AIST 2015. Communications in Computer and Information Science. 542:320–332. http://dx.doi.org/10.1007/978-3-319-26123-2_31
[14] http://opencorpora.org/
[15] https://github.com/kmike/pymorphy2
[16] *Koshcheeva S.* (2014) Sravneniye metodov avtomaticheskogo vydeleniya glagol'no-imennykh slovosochetaniy. In Proc. of the Tekhnologii informatsionnogo obshchestva v nauke, obrazovanii i kul'ture. Trudy XVII Vserossiyskoy ob"yedinennoy konferentsii «Internet i sovremennoye obshchestvo» (IMS-2014). P. 298–303. http://ojs.ifmo.ru/index.php/IMS/article/view/270/266 [In Rus]
[17] http://corpus.leeds.ac.uk/it/
[18] *Moskvina A. D., Mitrofanova O. A., Erofeeva A. R., Charabet Ja. K.* (2017) Avtomaticheskoye vydeleniye klyuchevykh slov i slovosochetaniy iz russko-yazychnykh korpusov tekstov s pomoshch'yu algoritma RAKE. In Proc. of the International conference "Corpora linguistics-2017". P. 268–274. [In Rus]

[19] *Sokolova E., Mitrofanova O.* (2017) Avtomaticheskoye izvlecheniye klyuchevykh slov i slovosochetaniy iz russko-yazychnykh tekstov s pomoshch'yu algoritma KEA. In Proc. of the Komp'yuternaya lingvistika i vychislitel'nyye ontologii. Vol. 1. Trudy XX Mezhdunarodnoy ob"yedinennoy nauchnoy konferentsii «Internet i sovremennoye obshchestvo» (IMS-2017). P. 298–303. [In Rus]

[20] *Moskvina A., Orlova D., Panicheva P., Mitrofanova O.* (2016) Razrabotka yadra sintaksicheskogo analizatora dlya russkogo yazyka na osnove bibliotek NLTK. In Proc. of the Komp'yuternaya lingvistika i vychislitel'nyye ontologii. Trudy XIX Mezhdunarodnoy ob"yedinennoy nauchnoy konferentsii «Internet i sovremennoye obshchestvo» (IMS-2016). P. 44–45. [In Rus]

[21] http://www.dialog-21.ru

[22] *Enikeeva E. V., Mitrofanova O. A.* (2017) Russian Collocation Extraction Based on Word Embeddings. In Proc. of the International Conference "Dialogue 2017". http://www.dialog-21.ru/media/3908/enikeevaevmitrofanovaoa.pdf

[23] *Kazennikov A. O.* (2017) Part-of-Speech Tagging: The Power of the Line-ar SVM-based Filtration Method for Russian Language. In Proc. of the International Conference "Dialogue 2017". http://www.dialog-21.ru/media/3916/kazennikovao.pdf

[24] *Stepanova M. E., Budnikov E. A., Chelombeeva A. N., Matavina P. V., Skorinkin D. A.* (2016) Information Extraction Based on Deep Syntactic-Semantic Analysis. In Proc. of the International Conference "Dialogue 2016". http://www.dialog-21.ru/media/3431/stepanovameetal.pdf

[25] *Barakhnin V. B., Tkachev D. A.* (2010) *Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Informatsionnyye tekhnologii*. **8**(2):5–14. [In Rus]

[26] *Barakhnin V. B., Pastushkov I. S.* (2015) *Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Informatsionnyye tekhnologii*. **13**(4):5–13. [In Rus]

[27] http://poem.ict.nsc.ru/