НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2018 Том 18 № 5 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2018 Vol. 18 No 5 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 004.912:303.7

НЕЛОКАЛЬНЫЕ СЕМАНТИЧЕСКИЕ СВЯЗИ В РУССКОЯЗЫЧНЫХ ТЕКСТАХ

К.К. Боярский[a], Е.А. Каневский[b]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[b] Институт проблем региональной экономики РАН, Санкт-Петербург, 190013, Российская Федерация
Адрес для переписки: Boyarin9@yandex.ru

Информация о статье
Поступила в редакцию 21.06.18, принята к печати 25.07.18
doi: 10.17586/2226-1494-2018-18-5-863-869
Язык статьи – русский

Ссылка для цитирования: Боярский К.К., Каневский Е.А. Нелокальные семантические связи в русскоязычных текстах // Научно-технический вестник информационных технологий, механики и оптики. 2018. Т. 18. № 5. С. 863–869. doi: 10.17586/2226-1494-2018-18-5-863-869

Аннотация

Предмет исследования. Одним из способов автоматического анализа текстов является построение деревьев подчинения, в которых слова предложения связываются друг с другом семантико-синтаксическими связями. В работе выполнено исследование русскоязычных текстов, имеющих общеполитический, художественный и узкоспециальный характер. Особое внимание уделено случаям, когда связываются слова, удаленные друг от друга на значительное расстояние. **Метод**. С помощью семантико-синтаксического парсера строятся деревья подчинения, после чего производится подсчет распределения связей разных типов по длинам. Исследованы частоты появления нелокальных связей. **Основные результаты**. Показано, что доля нелокальных связей в зависимости от типа может доходить до десятков процентов. Особенно это существенно для связей, исходящих из предикатных вершин (субъектные, обстоятельственные и т.д.), а также для анафорических. Отмечено, что общедоступные семантические классификаторы и тезаурусы имеют ограниченную применимость для решения задачи правильного связывания удаленных слов в предложении. **Практическая значимость**. Показано, что при извлечении из текстов информации, носящей онтологический или сценарный характер, а также при решении задач кореференции нельзя пренебрегать длинными синтаксическими связями, образующими в результате нелокальный семантический контекст. Сделан вывод, что анализ только *n*-грамм недостаточен для адекватного выделения из текста информации, носящей онтологический или сценарный характер. В связи с этим возникает необходимость составления микрословарей, ориентированных на определенные синтаксические конструкции.

Ключевые слова
семантико-синтаксический анализ, синтаксические связи, дерево подчинения, *n*-граммы, кореференция

FEATURES OF NON-LOCAL SEMANTIC LINKS IN RUSSIAN TEXTS

K.K. Boyarsky[a], E.A. Kanevsky[b]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation
[b] Institute of Regional Economics Problems RAS, Saint Petersburg, 190013, Russian Federation
Corresponding author: Boyarin9@yandex.ru

Article info
Received 21.06.18, accepted 25.07.18
doi: 10.17586/2226-1494-2018-18-5-863-869
Article in Russian
For citation: Boyarsky K.K., Kanevsky E.A. Features of non-local semantic links in Russian texts. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2018, vol. 18, no. 5, pp. 863–869 (in Russian). doi: 10.17586/2226-1494-2018-18-5-863-869

Abstract

Subject of Research. One of the ways of automatic text analysis is the construction of subordination trees, in which the words of a sentence are connected with each other by semantic-syntactic links. The field of research is Russian-language texts, which have a general political, artistic and highly specialized character. Special attention is paid to the cases when the words are connected being far from each other at a considerable distance. **Method**. The subordination trees were built with the help of semantic-syntactical parser.Then the calculation of the distribution of links of different types by lengths was performed. The appearance frequencies of nonlocal links are studied. **Main Results**. It is shown that the fraction of non-local connections depending on the type can reach up to tens of percent. This is especially important for links coming from predicate nodes (subject, adverbial, etc.), as well as for anaphoric ones. It is noted that publicly available semantic classifiers

and thesaurus have limited applicability for solving the problem of correct linking of remoted words in a sentence. **Practical Relevance**. It is shown that when solving the problem of extracting information that is ontological or scenario-based, as well as coreference, the long syntactic links that form the non-local semantic context cannot be neglected. The conclusion is drawn that the analysis of *n*-grams only is insufficient for the adequate selection of information from the text that is ontological or scenario. In this regard, there is a need to compile micro-dictionaries, focused on certain syntactic structures.

Keywords
semantic-syntactical analysis, syntactical links, subordination tree, *n*-grams, coreference

Введение

При решении задачи автоматического извлечения информации из текстов на естественном языке обычно используется один из двух подходов. Самой распространенной является модель «мешка слов» – bag-of-words, в которой порядок слов в тексте не учитывается вообще [1]. Эту модель используют многие статистические методы: байесовская классификация, латентное размещение Дирихле и др. Модель bag-of-words удовлетворительно работает при обработке больших текстовых наборов, однако, естественно, она достаточно груба и не может учесть смысловых различий в текстах сходного лексического состава.

Более тонким методом является поиск *n*-грамм, т.е. определенных групп стоящих контактно слов. Для анализа отбираются *n*-граммы определенных типов, образующих шаблоны. Примерами лексических шаблонов могут являться сочетания «согласованные прилагательное + существительное», «существительное + существительное в родительном падеже» и более сложные [2]. В то же время даже устойчивый термин может быть разорван посторонними словами и, следовательно, не будет обнаружен при сравнении с шаблоном. Особенно это касается всевозможных глагольных, именных, предложных групп. Для повышения точности приходится существенно усложнять алгоритмы обработки текста [3, 4].

В то же время зачастую оказываются семантически связаны слова, достаточно удаленные друг от друга. В результате образуется нелокальный смысловой контекст. Обнаружение длинных семантических связей важно при выявлении свойств именованных сущностей, концептов онтологий, при построении сценариев, использующих предикатные элементы, при автоматическом реферировании и т.д. [5, 6]. Целью данной работы является анализ особенностей структур предложений, при которых метод *n*-грамм может быть не вполне релевантен задаче извлечения информации из текста.

Метод исследования

Для того чтобы понять, в каких случаях метод *n*-грамм адекватен, а в каких он может давать ошибки, был предпринят эксперимент по оценке встречаемости нелокальных синтаксических связей. К таким мы отнесли случаи, когда между связываемыми словами в предложении находится не менее четырех слов или знаков препинания (в дальнейшем – токенов). С помощью парсера SemSin [7] строились деревья подчинения, после чего производился подсчет распределения связей разных типов по длинам. Анализ производился на двух наборах текстов общеполитического характера из Интернета, объемом 34000 и 12500 слов (в дальнейшем результаты, полученные для этой группы текстов, обозначаются как «Полит»), а также на текстах узкой предметной области: главы объемом 2500 и 6000 слов, посвященные устройству парусного вооружения судов из книг [8, 9] («Парус»).

Из рассмотрения исключались связи, носящие служебный характер (например, связи, подключающие частицы), и сочинительные связи между частями сложносочиненных предложений. В последнем случае парсер связывает между собой удаленные предикатные вершины двух поддеревьев, но это объединение носит достаточно формальный характер.

В приводимых ниже примерах связываемые слова выделены полужирным шрифтом.

Типы нелокальных связей

Связи определительные и по родительному падежу. Определительные связи (обычно существительное – прилагательное, хотя возможны и другие варианты) встречаются наиболее часто. Как правило, в предложении определение стоит контактно с определяемым словом. То же относится и к связям по родительному падежу с самой разной семантикой. Среди этих типов связей только 1–3 процента нелокальных:
*А я не хочу быть **заложником** вот этих явно невыполнимых **обещаний.***

Причастные обороты. В предложении причастные обороты играют роль, близкую по смыслу к определениям, однако гораздо чаще далеко отстоят от определяемого слова.
*Сейчас возможности Искандера ограничиваются **Договором** о ракетах средней и меньшей дальности, **запрещающим** разработку, производство и развертывание ракет наземного базирования…*
*…стаксели могли присоединять к лееру и штагу еще специальным тросом — **слаблинем**, проходившим через люверсы косой шкаторины паруса и **огибавшим** леер или штаг…*
*Так, **решением** Сыктывкарского городского суда Республики Коми от 31 января 2001 года, **оставленным** без изменения постановлением Судебной коллегии по гражданским делам Верховного Суда Республики Коми, признан недействительным **договор** купли-продажи квартиры, **заключенный** гражданкой О. М. Мариничевой.*

Однородные члены. Совсем не обязательно это родственные слова, разделенные только запятыми или союзами. На большом расстоянии друг от друга могут находиться как сказуемые (т.е. предикатные вершины дерева подчинения), так и любые другие элементы.

*Для этого их равномерно **распределяли** по шкаторине паруса или концу, а в дальнейшем **соединяли** в один трос.*

*Украинская политика администрации Буша, в том числе **подталкивание** к вступлению в НАТО (несмотря на отсутствие в украинском обществе консенсуса по этому вопросу), а также финансовая **поддержка** местных неправительственных организаций усилила опасения Москвы, что США вновь взялись за политику стратегического сдерживания.*

Субъекты. Анализ показал, что нелокальные связи особенно характерны для предикатов. Особую важность для выявления смысла текста, построения сценариев по тексту и т.д. имеет связь предикат – субъект. В зависимости от типа текста нелокальные связи такого типа составляют от 10 до 20%.

*Сие **пространство**, которое занимает край полотнища, лежащий на другом, **называется** собственно швом паруса.*

Нелокальные связи предикат – субъект могут встретиться не только в теле текста, но даже в заголовке. Приведем пример из конкурса по выявлению парафраз [10, 11]. Имеются два заголовка новостных сообщений, сообщающих, очевидно (для человека), об одном и том же событии:

***Взрыв** на заправочной станции в Дамаске **унес** жизни 11 человек.*

*При взрыве на автозаправке в Дамаске **погибли 11** человек.*

Без нахождения удаленного субъекта в первом предложении невозможно построение правильного дерева разбора и дальнейшего сравнения смысла обоих заголовков.

ФИО. В общественно-политических текстах с частотой около 20% встречается удаленное расположение фамилии человека и его должности или специальности, а такие связи чрезвычайно важны для установления кореферентных отношений.

*Возбуждено уголовное дело, которое ведет старший **следователь** по особо важным делам Следственного комитета при прокуратуре РФ **Владимир Соловьев**.*

В некоторых случаях название должности включает в себя длинную цепочку слов, как правило, в родительном падеже, причем формально ФИО может быть соотнесено с несколькими маркерными словами – членами этой цепочки.

Рассмотрим сравнительно простой пример:

Президент Армении Роберт Кочарян принял сегодня заместителя помощника Государственного секретаря США по вопросам Европы и Евразии Линна Паско…

Очевидно, что первое ФИО (*Роберт Кочарян*) однозначно подключится к маркерному слову *президент*. Второе ФИО (*Линна Паско*) в стандартной ситуации будет подключено к первому же встретившемуся маркеру (*секретаря*). После этого предложная группа *по вопросам Европы и Евразии* сможет подключиться только к этому же слову (что, очевидно, ошибочно), поскольку любое подключение к слову, расположенному левее, будет невозможно из-за нарушения проективности [12].

Для разрешения этого противоречия принято правило, что если в предложении имеется последовательность маркерных слов, обозначающих должности типа заместителя, помощника и т.д., то ФИО подключается к самому левому, т.е. самому удаленному из них.

Обстоятельства. На большое расстояние от определяемого слова с частотой около 20% оказываются удалены различного типа обстоятельства.

*Эти паруса привязывают к рею длинной стороной; **в диаметральной плоскости** судна, по направлению к корме, их **растягивают** при помощи шкота* (место).

***В июне** на саммит Россия – ЕС мы опять их **зовем** в столицу энергоресурсов – в Ханты-Мансийск* (время, обстоятельство предшествует предикату).

*Головка самонаведения комплекса настолько чувствительна, что позволяет успешно **поразить** любые цели (подземные, малоразмерные, площадные, подвижные) даже **в безлунные ночи**, когда нет дополнительной природной подсветки* (время, обстоятельство после предиката).

Предложные группы. Большие сложности вызывает нахождение «хозяина» предложных групп, особенно при наличии омонимии. В качестве примера рассмотрим словосочетания *по вопросу* и *по вопросам*. Эти словосочетания могут являться предложным оборотом, а могут представлять сочетание предлога с существительным [13]. Для анализа этих связей был подобран специальный корпус размером более 35000 словоформ – около 1800 отдельных предложений из НКРЯ[1] и набора текстов объемом около 60 млн слов, составленного из ряда повестей, новостных и спортивных текстов. Среди этих предложений оказалось:

- с предложным оборотом *по вопросам* – 1050;
- с предложным оборотом *по вопросу* – 320;

[1] Национальный корпус русского языка // URL: http://www.ruscorpora.ru/

– с предлогом *по* и существительным *вопросам* – 135;
– с предлогом *по* и существительным *вопросу* – 290.

Таким образом, обнаруживается высокая степень синтаксической омонимичности, для словосочетания *по вопросу* предложный оборот и предлог + существительное встречаются с почти одинаковой частотой.

Особенностью данного предложного оборота является то, что он может подключаться к ограниченному набору существительных, которые можно разделить на несколько групп.

К первой группе относятся слова, обозначающие профессии или род занятий людей. Для выявления в тексте слов с определенным значением естественно было бы воспользоваться семантическим классификатором. Однако, как указывалось в [11], различные классификаторы очень сильно отличаются друг от друга и несут отпечаток субъективных предпочтений авторов. Сравним положение нескольких слов, способных подключать предложную группу *по вопросу* (или *по вопросам*) на онтологическом дереве классификаторов РуТез [14][1] и Тузова [15].

– *Заместитель* – по РуТез относится к ветви роль → должность → административная должность; по Тузову: человек → личность → профессия → глава → зам.
– *Эксперт* – по РуТез: человек → человек по способностям → знаток; по Тузову: человек → личность → профессия.
– *Представитель* – по РуТез: человек → человек по роли; по Тузову: человек → личность → профессия → глава → зам.

Казалось бы, указание класса человек → личность → профессия (с подклассами) по классификатору Тузова может решить вопрос о возможности подключения предложной группы *по вопросу*. Однако относящиеся к этим же или близким классам слова *гендиректор*, *губернатор*… никогда не подключают эту предложную группу.

Ко второй группе относятся слова, обозначающие некоторые учреждения, их части, а также временно созданные коллективы: *коллегия*, *департамент*, *делегация*, *совет* (но не *совет директоров*)… Большинство таких слов в классификаторе Тузова относятся к классу *поселения → учреждения → государственные*, по РуТез – к различным ветвям классов *субъект деятельности → организация* или *субъект деятельности → структурное подразделение*.

Третью группу слов можно условно назвать «информационные действия»: *агитация*, *конференция*, *дискуссия*… Слова с этим значением в классификаторе Тузова относятся к подклассам класса *сообщения → информация → речь*, расстояние от них до общей вершины на семантическом дереве 1–2 уровня. В то же время «семантическое расстояние» от слов *конференция* и *дискуссия* до общей вершины по РуТез составляет 4 уровня по одной ветви и 8 уровней по другой.

Четвертая группа слов, способных подключать предложную группу *по вопросу*, условно названа «решение проблем». К ней относятся слова *компромисс*, *конфликт*, *голосование* и т.д. Эти слова относятся к совершенно разным классам как по РуТез, так и по Тузову.

Таким образом, хотя оказывается, что классификатор Тузова лучше приспособлен для решения задачи построения синтаксического дерева, но и его недостаточно. К сожалению, приходится констатировать, что во многих случаях приходится составлять специализированные микрословари маркерных слов для каждой конкретной синтаксической конструкции.

Для предложной группы *по вопросам*/*по вопросу,* так же как и для ФИО, характерно, что она, как правило, присоединяется к максимально удаленному влево маркерному слову. Например, в предложении
Руководить подразделением будет Олег Плохой, до этого занимавший пост ***замначальника*** *управления президента* ***по вопросам*** *государственной службы и кадров*
подключение производится к слову *замначальника*, а не к более близкому слову *управления*. Такая же схема подключения к наиболее удаленному маркерному слову видна в следующей паре предложений:
В ответ государственный ***министр*** *Грузии* ***по вопросам*** *реинтеграции Темур Якобашвили заявил…*
Но:
В аппарате Пентагона введена должность ***помощника*** *министра* ***по вопросам*** *защиты национальной территории.*
Расстояние между маркерным словом и предложной группой может быть достаточно большим:
Мы провели ***переговоры*** *с коллегами из Великобритании, Германии и других стран* ***по вопросам*** *организации предварительной сортировки*

Если парсер не находит маркерное слово, то подключение предложной группы осуществляется к предикатной вершине дерева:
Правительство РФ ***по вопросу*** *разработки и принятия федеральной целевой программы модернизации жилищно-коммунального хозяйства не* ***определилось*** *пока*

[1] http://www.labinform.ru/pub/ruthes/index.htm

На рис. 1 показано распределение по длинам связей, подключающих предложную группу *по вопросу*/*по вопросам*. Более чем в 15% случаев мы имеем дело с нелокальными связями (длина больше 4), что указывает на недостаточность методов *n*-грамм при семантическом анализе.

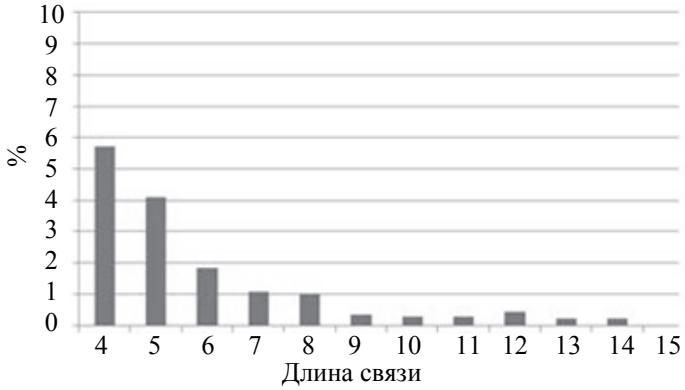

Рис. 1. Распределение по длинам связей для предложных групп *по вопросу*/*по вопросам*

Анафорические связи. Особый интерес представляют анафорические связи между местоимениями и их антецедентами. Среди таких связей нелокальных оказывается 40–50% (в зависимости от местоимения), причем их длина может быть очень большой. Распределение различных типов анафорических связей по длинам показано на рис. 2. Знак минус означает, что антецедент, как правило, предшествует анафору.

Приведем некоторые примеры.

***Запад** вынужден будет вступить в диалог, потому что **его** бьют на том поле, которое **он** всегда считал **своим**.*

*Каждый **парус** заимствует имя **свое** от мачты или рея, которым **он** принадлежит.*

Интересно, что местоимение *себя* в ряде случаев может стоять перед собственным антецедентом:

*29 октября снял **себя** с выборов N 2 по списку Пензенской области **Валерий Беспалов**.*

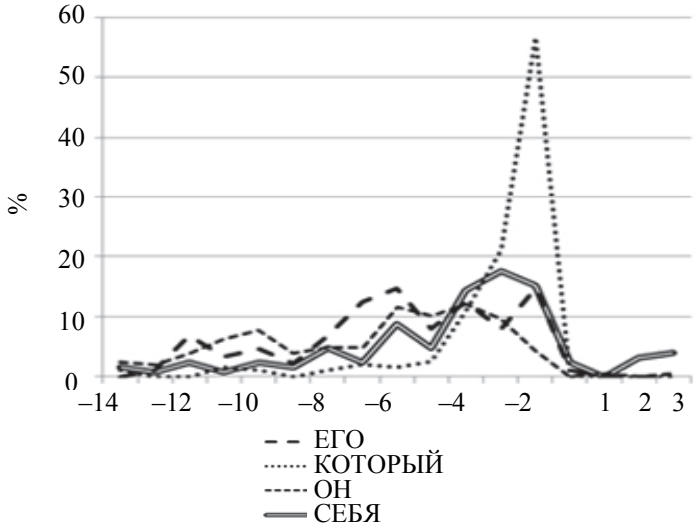

Рис. 2. Расстояние между антецедентом и анафором

Парсер SemSin позволяет выявлять анафорические связи не только в пределах предложения, но и в пределах целого абзаца. Естественно, в этом случае возрастает степень нелокальности:

*Тогда я понял: европейские **ценности** для Европы не абсолютны, а относительны и даже конъюнктурны. Парадокс, но мы не можем от **них** отказываться, именно потому, что сама Европа от **них** отказалась.* (28 токенов между антецедентом и местоимением).

Заключение

В настоящем исследовании показано, что если для целей классификации, выявления именованных сущностей достаточно анализа контактно стоящих групп слов в предложениях, то при извлечении из текстов информации, носящей онтологический или сценарный характер, а также при решении задач кореференции нельзя пренебрегать длинными синтаксическими связями, образующими в результате нелокальный семантический контекст.

В таблице приведен процент нелокальных связей различных типов.

Тип текста	Винительный падеж	Время	Место	Однородные	Причастный оборот	Субъект	ФИО	Анафора
Полит	5,53	14,59	14,63	18,20	13,76	9,45	18,18	46,0
Парус	7,85	22,22	20,63	27,27	10,53	19,64	–	

Таблица. Процент нелокальных связей

Для оценки важности специальной обработки нелокальных связей при построении дерева подчинения был проведен параллельный анализ с помощью общеизвестного парсера ЭТАП-3[1] 18 предложений, приведенных в качестве примеров в данной статье (за исключением анафорических связей).

Экспертный анализ построенных деревьев показал, что правильно подключенными оказалось только 55% нелокальных связей (10 предложений). Это связи определительные, между однородными членами, предикат–субъект. В то же время расположенные на большом удалении от хозяина причастные обороты, группы ФИО, обстоятельства, предложные группы *по вопросу* ЭТАП-3 неверно подключал более чем в половине случаев.

Это связано, в частности, с тем, что общедоступные классификаторы и тезаурусы не содержат всей полноты информации, необходимой для правильной установки нелокальных связей. Для решения возникающих проблем требуется составление микрословарей, т.е. списков слов, способных подключать определенные синтаксические конструкции. К сожалению, составление микрословарей не может быть полностью автоматизировано, а требует участия экспертов и большого объема ручного труда.

Для часто встречающихся нелокальных связей ФИО и предложных групп *по вопросу* (*вопросам*) разработаны правила и микрословари, позволяющие правильно определить хозяина этих групп.

Литература

1. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. 2-е изд. СПб.: БХВ-Петербург, 2007.
2. Большакова Е.И., Баева Н.В., Бордаченкова Е.А., Васильева Н.Э., Морозов С.С. Лексико-синтаксические шаблоны в задачах автоматической обработки текста // Труды международной конференции «Диалог 2007». Москва, 2007. С. 70–75.
3. Kormacheva D., Pivovarova L., Kopotev M. Automatic collocation extraction and classification of automatically obtained bigrams // Proc. Workshop on Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations. Tubingen, Germany, 2014. P. 27–33.
4. Enikeeva E.V., Mitrofanova O.A. Russian collocation extraction based on word embeddings // Proc. Int. Conf. Dialogue 2017. Moscow, 2017. P. 52–64.
5. Khomitsevich O., Boyarsky K., Kanevsky E., Bulusheva A., Mendelev V.S. Flexible context extraction for keywords in Russian automatic speech recognition results // Communications in Computer and Information Science. 2017. V. 661. P. 145–154. doi: 10.1007/978-3-319-52920-2_14
6. Дыбина А. Разработка текстовой базы на основе анализа структуры научного текста // International Journal Information Technologies & Knowledge. 2012. V. 6. N 1. P. 93–99.
7. Боярский К.К., Каневский Е.А. Семантико-синтаксический парсер SemSin // Научно-технический вестник информационных технологий, механики и оптики. 2015. Т. 15. № 5. С. 869–876. doi: 10.17586/2226-1494-2015-15-5-869-876
8. Курти О. Постройка моделей судов. Энциклопедия судомоделизма. Л.: Судостроение, 1977. 544 с.
9. Ромме М. L'Art de la Marine, ou Principes et Préceptes Generaux de l'Art de Construire, d'Armer, de Manœuvrer et de Conduire des Vasseaux. La Rochelle, 1787. 542 p.
10. Pivovarova L., Pronoza E., Yagunova E., Pronoza A. ParaPhraser: Russian paraphrase corpus and shared task // Communications in Computer and Information Science. 2017. V. 789. P. 211–225. doi: 10.1007/978-3-319-71746-3_18

References

1. Barsegyan A.A., Kupriyanov M.S., Stepanenko V.V., Kholod I.I. *Data Analysis Technologies: Data Mining, Visual Mining, Text Mining, OLAP*. 2nd ed. St. Petersburg, BKhV-Peterburg Publ., 2007. (in Russian)
2. Bol'shakova E.I., Baeva N.V., Bordachenkova E.A., Vasil'eva N.E., Morozov S.S. Lexicosyntactic patterns for automatic text processing. *Proc. Int. Conf. Dialogue 2007*. Moscow, 2007, pp. 70–75. (in Russian)
3. Kormacheva D., Pivovarova L., Kopotev M. Automatic collocation extraction and classification of automatically obtained bigrams. *Proc. Workshop on Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations*. Tubingen, Germany, 2014, pp. 27–33.
4. Enikeeva E.V., Mitrofanova O.A. Russian collocation extraction based on word embeddings. *Proc. Int. Conf. Dialogue 2017*. Moscow, 2017, pp. 52–64.
5. Khomitsevich O., Boyarsky K., Kanevsky E., Bulusheva A., Mendelev V.S. Flexible context extraction for keywords in Russian automatic speech recognition results. *Communications in Computer and Information Science*, 2017, vol. 661, pp. 145–154. doi: 10.1007/978-3-319-52920-2_14
6. Dybina A. Development of a textual base on the basis of the analysis of the structure of the scientific text. *International Journal Information Technologies & Knowledge*, 2012, vol. 6, no. 1, pp. 93–99. (in Russian)
7. Boyarsky K., Kanevsky E. SemSin semantic and syntactic parser. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2015, vol. 15, no. 5, pp. 869–876. (in Russian) doi: 10.17586/2226-1494-2015-15-5-869-876
8. Curti O. *Modelli Navali. Enciclopedia del Modellismo Navale*. Milano, 1980.
9. Ромме М. *L'Art de la Marine, ou Principes et Préceptes Generaux de l'Art de Construire, d'Armer, de Manœuvrer et de Conduire des Vasseaux*. La Rochelle, 1787, 542 p.
10. Pivovarova L., Pronoza E., Yagunova E., Pronoza A. ParaPhraser: Russian paraphrase corpus and shared task. *Communications in Computer and Information Science*, 2018, vol. 789, pp. 211–225. doi: 10.1007/978-3-319-71746-3_18
11. Boyarsky K., Kanevsky E. Effect of semantic parsing depth on

[1] http://proling.iitp.ru/ru/etap3

11. Boyarsky K., Kanevsky E. Effect of semantic parsing depth on the identification of paraphrases in Russian texts // Communications in Computer and Information Science. 2018. V. 789. P. 226–241. doi: 10.1007/978-3-319-71746-3_19
12. Кобзарева Т.Ю. Построение и использование проективных фрагментов именных и предложных групп // Труды международной конференции «Диалог 2007». Москва, 2007. С. 242–249.
13. Рогожникова Р.П. Толковый словарь сочетаний, эквивалентных слову. М.: Астрель, АСТ, 2003. 416 с.
14. Лукашевич Н.В. Тезаурус в задачах информационного поиска. М.: МГУ, 2011. 512 с.
15. Тузов В.А. Компьютерная семантика русского языка. СПб.: СПбГУ, 2004. 400 с.

Авторы

Боярский Кирилл Кириллович – кандидат физико-математических наук, доцент, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 56499298500, ORCID ID: 0000-0002-0306-8276, Boyarin9@yandex.ru

Каневский Евгений Александрович – кандидат технических наук, старший научный сотрудник, ведущий научный сотрудник, Институт проблем региональной экономики РАН, Санкт-Петербург, 190013, Российская Федерация, ORCID ID: 0000-0002-1498-4632, EAK300@mail.ru

the identification of paraphrases in Russian texts. *Communications in Computer and Information Science*, 2018, vol. 789, pp. 226–241. doi: 10.1007/978-3-319-71746-3_19
12. Kobzareva T.Yu. Building and use of projective fragments of attributive noun and prepositionalphrases. *Proc. Int. Conf. Dialogue 2007*. Moscow, 2007, pp. 242–249. (in Russian)
13. Rogozhnikova R.P. *Explanatory Dictionary of Combinations Equivalent to Word*. Moscow, Astrel'-AST Publ., 2003, 416 p. (in Russian)
14. Lukashevich N.V. *Thesaurus in Information Retrieval Problems*. Moscow, MSU Publ., 2011, 512 p. (in Russian)
15. Tuzov V.A. *Computer Semantics of Russian Language*. St. Petersburg, SPbSU Publ., 2004, 400 p. (in Russian)

Authors

Kirill K. Boyarsky – PhD, Associate Professor, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 56499298500, ORCID ID: 0000-0002-0306-8276, Boyarin9@yandex.ru

Eugeny A. Kanevsky – PhD, Senior researcher, Leading scientific researcher, Institute of Regional Economics Problems RAS, Saint Petersburg, 190013, Russian Federation, ORCID ID: 0000-0002-1498-4632, EAK300@mail.ru