УДК 004.89
DOI: 10.17586/0021-3454-2019-62-11-976-981

АЛГОРИТМ АВТОМАТИЧЕСКОГО ВЫДЕЛЕНИЯ КОЛЛОКАЦИЙ ИЗ ТЕКСТА

К. В. Ненаусников, С. В. Кулешов

Санкт-Петербургский институт информатики и автоматизации РАН,
199178, Санкт-Петербург, Россия
E-mail: nenausnikov@iias.spb.su

Для повышения точности системы ассоциативного поиска предложен алгоритм автоматического выделения коллокаций из корпуса текстов на естественном языке. Разработанный алгоритм предназначен для аддитивной оценки биграмм (пар элементов) текста на основе статистического подхода и выделения наиболее релевантных биграмм с использованием распределения Ципфа. Выполнен анализ методов выделения коллокаций из случайного корпуса текстов, размещенных в сети Интернет, на основе таких ассоциативных мер, как частота вхождения биграмм в текст, *t*-тест, MI и χ^2, с использованием грамматического фильтра, с удалением стоп-слов и последующей оценкой указанных мер. Применение метода аддитивного оценивания при построении распределения Ципфа позволяет определить область корректных коллокаций, что приводит к уменьшению количества ошибок в полученных списках коллокаций.

Ключевые слова: *семантический анализ, понятие, коллокация, словарь, ассоциативная мера, лингвистический шаблон, MI, t-тест, χ^2, ассоциативный поиск, распределение Ципфа*

Введение. Автоматическая обработка естественного языка является важным направлением, которое нашло применение в таких областях, как поисковые системы, технологическая документация, медицинские базы данных. Одна из базовых функций автоматической обработки — выделение атомарных элементов. В зависимости от задачи такими элементами могут быть символы, токены, понятия, высказывания и т.д.

Структура любого естественного языка не является семантически однородной, поэтому информация, содержащаяся в тексте, не может передаваться по принципу суперпозиции отдельных слов. Информация содержится в различных комбинациях слов, причем одно и то же значение может быть представлено разными комбинациями.

В настоящей статье для повышения точности системы ассоциативного поиска предложен алгоритм автоматического извлечения двусложных коллокаций из корпуса текстов на естественном языке. Списки выделенных коллокаций используются в синтаксическом анализе для повышения точности построения структуры предложения [1, 2], а в семантическом анализе — для выделения понятий, информационного поиска, построения онтологий и т.д. [3—6].

Методы выделения коллокаций. Коллокация — это устойчивое словосочетание, содержащее синтаксические или семантические связи.

При автоматическом выделении коллокаций из текста рассматриваются два подхода: на основе частоты вхождения слов и словосочетаний в рассматриваемый текст (частотный подход) и на основе структуры предложения [7, 8]. Наилучший результат достигается при использовании комбинированного подхода, включающего оба вышеуказанные.

Известными методами при анализе текста на основе частотного подхода являются следующие ассоциативные меры: частота вхождения; MI (Mutual Information — совместная информация); χ^2, или критерий Пирсона; *t*-тест, или критерий Стьюдента, и log-likelihood (вероятностная мера). Для упорядочения оценки предложенных мер выделяют четыре класса коллокаций: терминологические, традиционные, экспрессивные и этнокультурные [9]. При из-

влечении терминологических и этнокультурных коллокаций наиболее эффективны методы на основе мер MI и χ^2, а при извлечении традиционных и экспрессивных коллокаций — частота вхождения и *t*-тест [10—12].

Для выделения коллокаций на основе структуры предложения выполняется синтаксический анализ, результатом которого является иерархическая структура предложения, — такой подход наиболее востребован для языков со свободным порядком слов в предложении. Также для повышения точности системы автоматического выделения коллокаций применяется синтаксический шаблон [13].

Рассматриваемая в настоящей статье задача заключается в составлении списка коллокаций для улучшения работы системы ассоциативного поиска [14]. Для решения поставленной задачи разработана общая схема выделения коллокаций из случайного корпуса текстов, размещенных в сети Интернет, которая представлена на рис. 1.

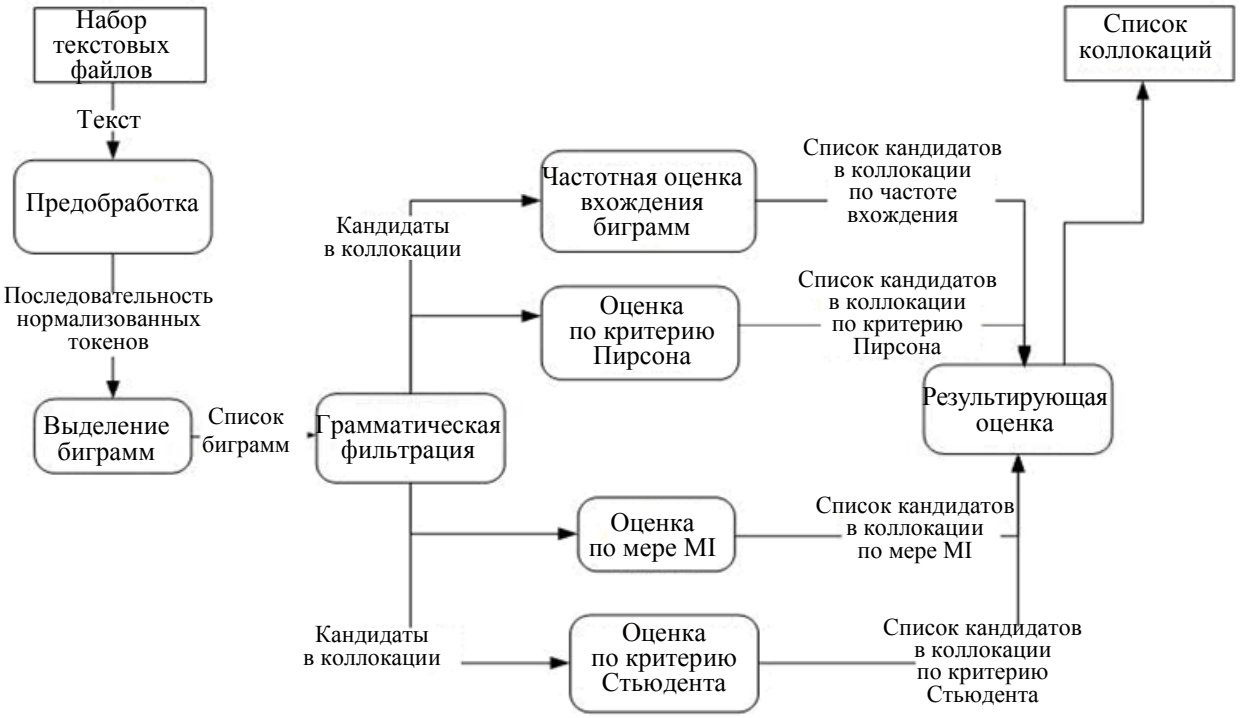

Рис. 1

На этапе предобработки текст разбивается на отдельные элементы — токены (от англ. „token“ — знак, символ), выделение которых осуществляется с использованием правил синтаксиса русского языка. Токеном считается словоформа или знак препинания. Для каждого из выделенных токенов определяются его нормальная грамматическая форма и часть речи. Все знаки препинания заменяются на условный символ — „разделитель“, из текста удаляются стоп-слова (например, междометия, предлоги).

На следующем этапе из текста выделяются биграммы, т.е. пары слов, последовательно расположенные в предложении, причем биграммой не может считаться пара слов, между которыми стоит „разделитель“. Каждая биграмма сопоставляется с заданным грамматическим шаблоном [14]. Русский язык характеризуется следующим набором шаблонов:

1) глагол с существительным (например, составить кроссворд, любоваться городом);
2) глагол с инфинитивом (например, помочь сделать, любить читать);
3) глагол с наречием (например, сказать сгоряча, вернуться поздно);
4) существительное с существительным (например, колесо фортуны, делу время);
5) существительное с наречием (например, взгляд исподлобья, борщ по-украински);
6) существительное с инфинитивом (например, готовность помочь, повод поговорить);
7) наречие с прилагательным (например, насыщенно красный, тяжело больной);

8) наречие с наречием (например, крайне осторожно, невыносимо больно);

9) прилагательное с существительным (например, певучая речь, длинная дорога).

Биграммы, выделенные из текста и соответствующие одному из предложенных шаблонов, будем считать кандидатами в коллокации.

Для каждого кандидата вычисляется аддитивная (результирующая) оценка — средневзвешенное значение оценок на основе ассоциативных мер: частотной меры, *t*-теста, MI, и χ^2. На рис. 2 приведены графики оценок кандидатов в коллокации, представленных в алфавитном порядке (здесь σ — оценка биграмм по данной мере, *N* — номер биграммы в алфавитном списке).

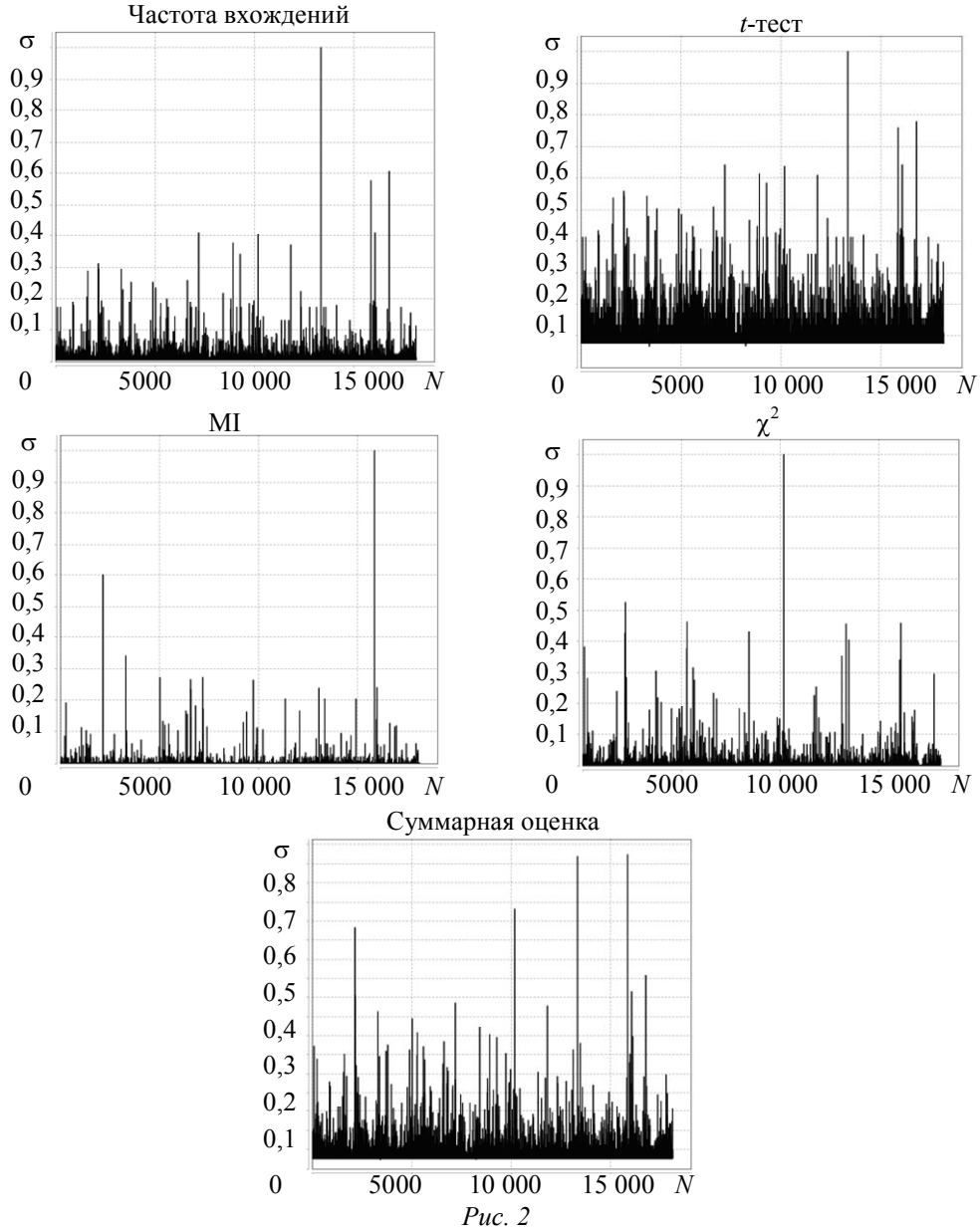

Рис. 2

Анализ графиков показывает, что суммарная оценка сглаживает разницу между кандидатами в коллокации: уменьшает значение „выбросов“, т.е. высоких ошибочных значений для каждого метода.

Распределение коллокаций в естественном языке осуществляется по закону Ципфа. Распределения для каждой из оценок представлены на рис. 3 (по оси ординат — относительная частота вхождения — ω, по оси абсцисс — положение слова в частотном словаре (ранг) — *x*).

На данном графике можно выделить 3 участка: I участок — небольшое число кандидатов, которые с высокой вероятностью являются коллокациями; II участок — кандидаты, которые потенциально могут считаться коллокациями, т.е. содержат и корректно выделенные коллокации, и ошибки; III участок — кандидаты, имеющие оценки, близкие к нулю, — это биграммы, элементы которых случайно оказались в тексте рядом. В зависимости от поставленной задачи, например высоких требований к точности, выбираются элементы только I участка, а при требованиях к большому объему — элементы I и II участков.

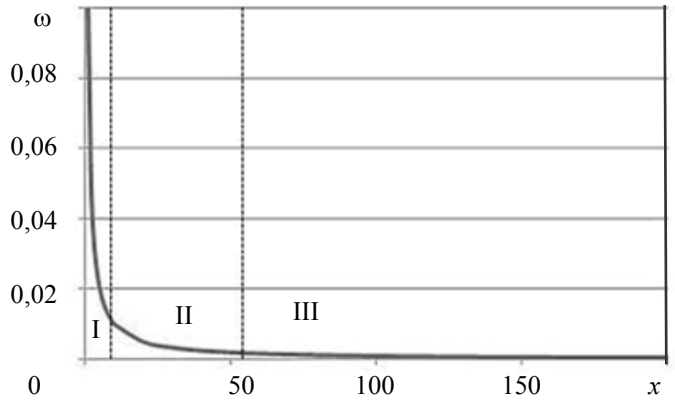

Рис. 3

Определение границ участков является нетривиальной задачей, так как для распределений, описывающих разные тексты, необходимы различные границы, что требует участия эксперта. Для решения этой задачи предлагается следующий алгоритм. Список кандидатов в коллокации, отсортированный в порядке уменьшения значения результирующей оценки, обходится окном заданного размера — на основе экспериментов размер окна был выбран равным пяти элементам — до тех пор пока вектор, координаты которого задаются двумя точками (границами окна), не выполнит поворот в 30° относительно своего начального положения. Значение 30° определено экспериментально, так как оно ближе остальных к границам, выбранным экспертным путем для набора текстов.

Эксперимент. Для эксперимента выбран набор из 24 361 случайных текстовых документов из сети Интернет. В результате автоматической обработки выделено около 500 000 биграмм, 35 000 из которых определены как коллокации. Итоговый список, сформированный на основании аддитивной оценки (см. рис. 2), содержит коллокации всех типов. Значительная часть словосочетаний, содержащих ошибки, получила меньшую оценку, чем при использовании одной (любой) из рассмотренных мер.

Заключение. Выполнен анализ методов извлечения двусложных коллокаций из текстов на основе ассоциативных мер с использованием грамматического фильтра и удалениея стоп-слов. Использовались оценки на основе таких мер, как частота вхождения биграмм в текст, *t*-тест, MI и χ^2. Каждый из полученных списков кандидатов в коллокации может содержать ошибки, количество которых удается сократить посредством аддитивной оценки полученных кандидатов. Для определения конечного списка коллокаций применен подход на основе распределения Ципфа. Предложенный подход может быть использован для выделения коллокаций без разделения их на классы, а в дальнейшем — как модуль для работы системы ассоциативного поиска. Направлением развития разработанного подхода может быть извлечение из текста многословных коллокаций.

Работа выполнена в рамках реализации Государственного задания на 2019 г., № 0073-2019-0005.

СПИСОК ЛИТЕРАТУРЫ

1. *Chen W. T., Bonial C., Palmer M.* English light verb construction identification using lexical knowledge // Proc. of the 29th AAAI Conf. on Artificial Intelligence, Austin, TX, USA. 2015. P. 2368—2374.
2. *Kolesnikova O., Gelbukh A.* Binary and Multi-class classification of lexical function in Spanish verb-noun collocations // Lecture Notes in Computer Science Ed.: *M. Gonzalez-Mendoza, F. Castro, S. Miranda-Jimenez.* Mexico, 2018. P. 3—14. DOI:10.13140/RG.2.1.2610.0242.
3. *Bobkova A.* The use of collocations for English nouns disambiguation // Thought Elaboration: Linguistics, Literature, Media Expression / Ed.: *D. Satkauskaite*. Vilnius: Vilnius Univ., 2017. P. 64—78.
4. *Granger S.* Formulaic sequences in learner corpora: Collocations and lexical bundles // Understanding Formulaic Language: A Second Language Acquisition Perspective / Ed.: *A. Siyanova-Chanturia, A. Pellicer-Sanchez.* N. Y.: Routledge, 2018. P. 228—247. DOI:10.4324/9781315206615.
5. *Gyllstad H., Wolter B.* Collocational processing in light of the phraseological continuum model: does semantic transparency matter? // Language Learning. 2017. Vol. 45, iss. 3. P. 296—323. DOI:10.1111/lang.12143.
6. *Лескина С. В., Шаранова В. Б.* Структурная и семантическая соотнесенность коллокаций и фразеологических единиц в русском и в английском языках // Вестн. Южно-Уральского гос. ун-та. Сер. Лингвистика. 2014. № 1. С. 22—28.
7. *Verma R., Vuppuluri V., Nguyen A., Mukherjee A., Mammar G., Baki S., Armstrong R.* Mining the Web for collocations: IR models of term associations // Lecture Notes in Computer Science. Ed.: *A. Gelbukh*. Springer Verlag, 2018. P. 177—194. DOI:10.1007/978-3-319-75477-2_11.
8. *Влавацкая М. В.* Комбинаторная лексикология: функционально-семантическая классификация коллокаций // Филологические науки. Вопросы теории и практики. 2015. № 11, ч. 1. С. 56—60.
9. *Ягунова Е. В., Пивоварова Л. М.* Природа коллокаций в русском языке. Опыт автоматического извлечения и классификации на материале новостных текстов // Науч.-техн. информация. Сер. 2. Информационные процессы и системы. 2010. № 6. С. 30—40.
10. *Захаров В. П., Хохлова М. В.* Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии. 2010. № 9 (16). С. 137—143.
11. *Liu X., Huang D., Yin Z., Ren F.* Recognition of collocation frames from sentences // IEICE Transact. on Information and Systems. 2019. P. 620-627. DOI: 10.1587/transinf.2018EDP7255.
12. *Петров А. С., Шульга Т. Э.* Математическая модель русскоязычного текстового документа для решения задачи автоматического извлечения терминов из текста // Вестн. Воронеж. гос. ун-та. Сер. Системный анализ и информационные технологии. 2017. № 3. С. 195—203.
13. *Кулешов С. В., Зайцева А. А., Марков В. С.* Ассоциативно-онтологический подход к обработке текстов на естественном языке // Интеллектуальные технологии на транспорте. 2015. № 4. С. 40—45.
14. *Найханова Л. В.* Технология создания методов автоматического построения онтологии с применением генетического и автоматного программирования: Монография. Улан-Удэ: Изд-во БНЦ СО РАН, 2008. 244 с.

Сведения об авторах

Константин Вячеславович Ненаусников — СПИИРАН, лаборатория автоматизации научных исследований; мл. науч. сотрудник; E-mail: nenausnikov@iias.spb.su

Сергей Викторович Кулешов — д-р техн. наук; СПИИРАН, лаборатория автоматизации научных исследований; гл. науч. сотрудник; E-mail: kuleshov@iias.spb.su

Поступила в редакцию
12.08.19 г.

Ссылка для цитирования: *Ненаусников К. В., Кулешов С. В.* Алгоритм автоматического выделения коллокаций из текста // Изв. вузов. Приборостроение. 2019. Т. 62, № 11. С. 976—981.

ALGORITHM OF AUTOMATIC SELECTION OF COLLOCATIONS FROM THE TEXT

K. V. Nenausnikov, S. V. Kuleshov

St. Petersburg Institute for Informatics and Automation of the RAS,
199178, St. Petersburg, Russia
E-mail: nenausnikov@iias.spb.su

To improve the accuracy of the associative search system, an algorithm for automatic selection of collocations from the corpus of natural language texts is proposed. The developed algorithm is intended for additive estimation of bigrams (pairs of elements) of the text on the basis of statistical approach and selection of the most relevant bigrams with the use of Zipf distribution. Methods of extracting collocations are analyzed on the example of a random corpus of texts obtained from the Internet on the base of such associative measures as the frequency of occurrence of bigrams in the text - t-test, MI and χ^2, using a grammatical filter, with removal of stop words and subsequent evaluation of these measures. The application of the additive estimation method in the construction of Zipf distribution makes it possible to determine the area of correct collocations, which leads to a decrease in the number of errors in the obtained collocation lists.

Keywords: semantic analysis, entity, collocation, dictionary, associative measure, linguistic pattern, MI, *t*-test, χ^2, associative search, Zipf distribution

REFERENCES

1. Chen W.T., Bonial C., Palmer M. *Proc. of the 29th AAAI Conf. on Artificial Intelligence*, Austin, TX, USA, 2015, pp. 2368–2374.
2. Kolesnikova O., Gelbukh A. *Lecture Notes in Computer Science,* Gonzalez-Mendoza M., Castro F., Miranda-Jimenez S., ed., Mexico, 2018, pp. 3–14. DOI:10.13140/RG.2.1.2610.0242.
3. Bobkova A. Thought Elaboration: Linguistics, Literature, Media Expression, Satkauskaite D., ed., Vilnius, Vilnius Univ., 2017, pp. 64–78.
4. Granger S. *Understanding Formulaic Language: A Second Language Acquisition Perspective*, Siyanova-Chanturia A., Pellicer-Sanchez A., ed., NY, Routledge, 2018, pp. 228–247. DOI:10.4324/9781315206615.
5. Gyllstad H., Wolter B. *Language Learning*, 2017, no. 3(45), pp. 296–323. DOI:10.1111/lang.12143.
6. Leskina S.V., Sharanova V.B. *South Ural State University Bulletin. Linguistics*, 2014, no. 1, pp. 22–28. (in Russ.)
7. Verma R., Vuppuluri V., Nguyen A., Mukherjee A., Mammar G., Baki S., Armstrong R. *Lecture Notes in Computer Science*, Springer Verlag, 2018, pp. 177–194. DOI:10.1007/978-3-319-75477-2_11.
8. Vlavatskaya M.V. *Philological Sciences. Issues of Theory and Practice*, 2015, no. 11, pt. 1, pp. 56–60. (in Russ.)
9. Yagunova E.V., Pivovarova L.M. *Automatic Documentation and Mathematical Linguistics*, 2010, no. 6, pp. 30–40. (in Russ.)
10. Zakharov V.P., Khokhlova M.V. *Computational Linguistics and Intelligent Technologies*, 2010, no. 9(16), pp. 137–143. (in Russ.)
11. Liu X., Huang D., Yin Z., Ren F. *IEICE Transact. on Information and Systems*, 2019, pp. 620–627. DOI: 10.1587/transinf.2018EDP7255.
12. Petrov A.S., Shul'ga T.E. *Proc. of Voronezh State University. Series: Systems analysis and information technologies*, 2017, no. 3, pp. 195–203. (in Russ.)
13. Kuleshov S.V., Zaytseva A.A., Markov V.S. *Intellectual Technologies on Transport*, 2015, no. 4, pp. 40–45. (in Russ.)
14. Naykhanova L.V. *Tekhnologiya sozdaniya metodov avtomaticheskogo postroyeniya ontologii s primeneniyem geneticheskogo i avtomatnogo programmirovaniya* (The Technology of Creating Methods for Automatically Constructing Ontologies Using Genetic and Automatic Programming), Ulan-Ude, 2008, 244 p. (in Russ.)

Data on authors

Konstantin V. Nenausnikov — St. Petersburg Institute for Informatics and Automation of the RAS, Laboratory of Automation of Scientific Research; Junior Researcher E-mail: nenausnikov@iias.spb.su

Sergey V. Kuleshov — Dr. Sci.; St. Petersburg Institute for Informatics and Automation of the RAS, Laboratory of Automation of Scientific Research; Principal Researcher; E-mail: kuleshov@iias.spb.su

For citation: Nenausnikov K. V., Kuleshov S. V. Algorithm of automatic selection of collocations from the text. *Journal of Instrument Engineering*. 2019. Vol. 62, N 11. P. 976—981 (in Russian).

DOI: 10.17586/0021-3454-2019-62-11-976-981