Автоматическое извлечение ключевых слов и словосочетаний из русскоязычных текстов с помощью алгоритма KEA

Е.В. Соколова, О.А. Митрофанова

Санкт-Петербургский государственный университет

st049868@student.spbu.ru, o.mitrofanova@spbu.ru

Аннотация

В докладе представлены результаты работы по модификации алгоритма KEA (Keyphrase Extraction Algorithm), используемого для извлечения ключевых слов и словосочетаний. KEA широко известен своей эффективностью для извлечения ключевых слов и словосочетаний из англоязычных текстов. В статье представлены результаты применения данного алгоритма к текстам на русском языке. Для определения качества работы алгоритма с русскоязычными текстами были проведены эксперименты на материале представительных корпусов.

Ключевые слова: автоматическое выделение ключевых слов и словосочетаний, RAKE, русскоязычные корпусы текстов.

1. Введение

Наборы назначенных вручную или автоматически выделенных ключевых слов и словосочетаний текста используются для формирования у пользователя общего представления о содержании текста. Ключевые слова и словосочетания чаще всего понимают как структурные единицы текста, содержащие наиболее важную информацию о содержании текста. Необходимо различать два основных подхода к решению проблемы автоматизации выделения ключевых слов и словосочетаний: назначение ключевых слов и словосочетаний (keyphrase assignment) и их извлечение (keyphrase extraction) [1] [2]. Главное отличие заключается в том, что первый подход позволяет выделять только те ключевые слова и словосочетания, которые содержатся в некотором предусмотренном словаре, а второй подход предполагает выбор ключевой информации непосредственно из текста.

Мы опираемся на опыт работы с различными методами извлечения ключевых слов и словосочетаний [3] [4], среди которых хорошо известны статистические (например, метрика $TF \times IDF$), лингвистические (включающие семантический, синтаксический анализ и т.п.), методы машинного обучения

(наивный байесовский классификатор, метод опорных векторов и др.), а также гибридные подходы (KEA, TextRank и др.). Некоторые из них предполагают наличие словарей или фоновых корпусов, другие не требуют таковых. Стоит также отметить, что каждый отдельный алгоритм может выделять только ключевые слова, словосочетания или и те, и другие одновременно. Так как большая часть исследований в этой области проводится на материале английского языка, то в англоязычной литературе встречается несколько синонимичных терминов для обозначения ключевых слов и словосочетаний, например, «key terms», «keyphrases» или «keywords», но чаще всего используются последние два, причём как для униграмм, так и для *n*-грамм.

В качестве основного направления данной работы мы избрали исследование одного из алгоритмов автоматического извлечения слов и словосочетаний из текста, а именно KEA (Keyphrase Extraction Algorithm) [5] [6]. Данный алгоритм широко известен благодаря своим высоким результатам на материале английского языка [7] [8] [9], поэтому мы попытались применить его к русскоязычным текстам и найти оптимальный способ оценки его эффективности для русского языка.

Таким образом, основная цель нашего исследования — адаптация KEA для работы с русскоязычным материалом. Для её достижения необходимо дополнение алгоритма инструментами для работы с русским языком и подготовка данных для проведения экспериментов. Инструменты включают в себя модули графематического и морфологического анализа для русского языка, а также русскоязычный стоп-словарь. Материалом для проверки работы алгоритма и оценки его эффективности после адаптации служат четыре корпуса, содержащих тексты по ракетостроению и аэрокосмическим исследованиям.

2. Внутренняя организация KEA и этапы его работы

KEA был разработан Йеном Виттеном и его коллегами в Новой Зеландии в 1999 г. [5, 6]. Как было отмечено выше, KEA относится к классу гибридных алгоритмов и включает в себя два компонента:

- машинное обучение с учителем: на вход подаётся обучающая выборка с выделенными автором или экспертом ключевыми словами и словосочетаниями; в результате обучения строится модель для определения ключевых слов и словосочетаний;
- автоматическое извлечение ключевых слов и словосочетаний на основе построенной модели.

2. 1. Выделение кандидатов в ключевые слова и словосочетания

На каждом из этапов работы выбираются кандидаты в ключевые слова и словосочетания, для каждого из которых затем вычисляются значения определённых признаков. Выбор кандидатов, в свою очередь, предполагает три шага:

1) предварительная обработка подаваемых на вход документов:
   - токенизация;

- замена знаков пунктуации и цифр на символы, обозначающие границы словосочетаний;
- сегментация слов, предполагающих дефисное написание;
- удаление оставшихся небуквенных символов;

2) определение кандидатов с помощью следующего набора правил:

- ограничение максимальной длины словосочетания (как правило, три слова);
- отбрасывание кандидатов имён собственных;
- отбрасывание кандидатов, содержащих стоп-слово в начале или конце.

Все последовательности слов, полученных на шаге 1, рассматриваются с учётом правил, выделенных на шаге 2, в результате чего получается набор наиболее релевантных на данном этапе обработки кандидатов.

3) выравнивание регистра и стемминг: в оригинальном алгоритме для английского языка используется стеммер Джули Ловинс, созданный в 1968 г. Первоначальная форма и регистр слов сохраняются для представления пользователю в том случае, если кандидат действительно окажется ключевым словом или словосочетанием.

2.2. Вычисление значений основных признаков для каждого кандидата

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 нём.

2.3. Метрика TF×IDF

TF×IDF (term frequency — inverse document frequency) — это статистическая мера частоты встречаемости слова или словосочетания в конкретном документе, определяемая в сравнении с частотой его использования в других документах коллекции или в некотором фоновом корпусе.

Для этой цели KEA создаёт файл, в котором хранит информацию о частоте встречаемости слова или словосочетания в конкретном обрабатываемом документе и о количестве документов коллекции, содержащих данную структурную единицу.

Таким образом, для каждого слова или словосочетания P в документе D метрика *TF×IDF* рассчитывается по следующей формуле:

$$TF \times IDF = \frac{freq(P,D)}{size(D)} \times -log_2 \frac{df(P)}{N}, \text{где}$$

freq(*P*,*D*) — количество раз, которое данное слово или словосочетание встречается в *D*;

size(*D*) — количество слов в *D*;

df(*P*) — количество документов в некоторой коллекции или фоновом корпусе, содержащих *P*.

N — размер коллекции или фонового корпуса.

Второй множитель — это функция правдоподобия, отвечающая за вероятность появления данного слова или словосочетания в документе (представлена с отрицанием, так как вероятность меньше единицы). Если документ не является частью коллекции или фонового корпуса, перед вычислением функции параметры $df(P)$ и N увеличиваются на единицу, чтобы симулировать его появление.

2.4. Расстояние от начала документа до первого появления слова или словосочетания в нём

Расстояние от начала документа до первого появления слова или словосочетания в нём является отношением количества слов, предшествующих данному, и общего количества слов в документе. Результат представлен значениями в промежутке [0, 1] в зависимости от размера части документа до первого появления данного слова или словосочетания в нём.

2.5. Обучение алгоритма и построение модели прогнозирования кандидатов в ключевые слова и словосочетания

На данном этапе работы алгоритма используется обучающая выборка с выделенными автором или экспертом ключевыми словами и словосочетаниями. Во всех документах определяются кандидаты, для каждого из которых вычисляются значения описанных выше признаков. Чтобы уменьшить объём обрабатываемых данных, игнорируются слова с единичной частотой, после чего каждый кандидат помечается как «ключевой» или «неключевой». Это бинарное деление является классовым признаком, используемым наивным байесовским классификатором [10]. Наивный байесовский классификатор представляет собой классификатор, который определяет вероятность принадлежности рассматриваемого объекта к одному из заранее определённых классов. При этом процесс классификации строится на предположении о независимости классов друг от друга. Таким образом, данный классификатор относит объект X к классу C_i тогда и только тогда, когда выполняется условие $P(C_i|X) > P(C_j|X)$, где $P(C_i|X)$ – апостериорная вероятность принадлежности объекта X классу C_i, а $P(C_j|X)$ – апостериорная вероятность принадлежности объекта X классу C_j. Классификатор в KEA определяет веса, назначенные кандидату, и на их основе одну часть кандидатов помечает как «ключевые», а другую — как «неключевые». Далее строится модель, которая предсказывает, к какому из обозначенных классов относится то или иное слово или словосочетание в зависимости от значения вычисленных признаков.

2.6. Извлечение новых ключевых слов и словосочетаний

Чтобы выбрать ключевые слова и словосочетания из нового документа, KEA определяет кандидатов и вычисляет для каждого из них значения признаков, после чего применяет модель прогнозирования, построенную на предыдущем этапе. Она вычисляет полную вероятность того, что каждый кандидат является ключевым словом или словосочетанием, а затем, после обработки, выбирается лучший набор ключевых слов и словосочетаний.

Когда классификатор обрабатывает кандидата с признаками t ($TF{\times}IDF$) и d (distance), вычисляются две величины:

$$P[yes] = \frac{Y}{Y+N} P_{\text{TF}\times\text{IDF}}[t|yes] P_{distance}[d|yes]$$

и аналогичная для $P[no]$, где Y — количество положительных примеров в обучающей выборке, то есть слова и/или словосочетания, назначенные автором, а N — количество отрицательных примеров, то есть кандидаты, которые не являются ключевыми (чтобы избежать нулевой вероятности используется сглаживание Лапласа, которое Y и N заменяет на Y+1 и N+1).

Полная вероятность того, что кандидат является ключевым словом или словосочетанием, в свою очередь, вычисляется следующим образом:

$$p = P[yes]/(P[yes] + P[no])$$

Согласно значению этой величины, кандидаты ранжируются и осуществляются два следующих шага. Во-первых, значение $TF{\times}IDF$ используется в том случае, если вероятности двух кандидатов равны. Во-вторых, из списка удаляются все слова и словосочетания, которые содержатся в других выражениях, имеющих более высокий ранг. Из полученного ранжированного списка первые r предоставляются пользователю, где r — количество запрашиваемых ключевых слов и словосочетаний.

3. Адаптация KEA и оценка его работы на материале русского языка

3.1. Планирование эксперимента

KEA является универсальным лингвонезависимым алгоритмом, его программная реализация позволяет использовать его совместно с процессорами для любого естественного языка. Для проверки эффективности работы данного алгоритма на материале русскоязычных текстов мы осуществили адаптацию KEA, совместив его с модулями графематического и морфологического анализа для русского языка.

KEA реализован на языке программирования Java и поставляется разработчиками со всеми необходимыми инструментами для работы алгоритма на материале нескольких языков. Как было отмечено выше, на одном из этапов своей работы алгоритм подразумевает стемминг. Единственным доступным инструментом для работы с русскоязычными текстами на Java оказывается стеммер Портера [11]. Основываясь на особенностях языка, он отсекает лишь суффиксы и флексии, поэтому на данном этапе использовался другой инструмент, а именно морфологический анализатор pymorphy2 [12] для Python. Проводилась предварительная лемматизация текстов как из обучающей, так и из тестовой выборок, после чего к обработанным текстам применялся KEA. Следующим шагом, требующим адаптации, является удаление из документов стоп-слов. Для этой цели был использован стоп-словарь, который составлен на

основе данных из Национального корпуса русского языка (НКРЯ) [13], и включает в себя наиболее частотные предлоги, частицы, местоимения, междометия, некоторые вводные слова и конструкции, а также количественные и порядковые числительные, цифры и символы латиницы [14]. На основе уже имеющихся в пакете методов, предполагающих удаление стоп-слов из исходных текстов для других языков, был разработан отдельный метод для русскоязычного стоп-словаря.

3.2. Ход эксперимента, полученные данные

Эксперименты проводились на материале четырех корпусов русскоязычных текстов по ракетостроению и аэрокосмическим исследованиям, представляющих научный, публицистический, официально-деловой и художественный функциональные стили [15]. Объем каждого из корпусов составляет примерно 500 тыс. с/у, суммарный объем обработанных текстов, тем самым, оценивается в 2 млн. с/у.

Для корпусов с текстами научного и художественного стилей автоматически были получены списки ключевых выражений (по 40 наиболее частотных биграмм, ранжированных по значению *MI*, общее число — 80 биграмм). Корпус был поделен на обучающую и тестовую выборки. Для обучения использовались корпусы с текстами научного и художественного стилей, а для тестирования — публицистического и официально-делового. В обучающей выборке были размечены употребления выражений, совпадающих с биграммами из списка. В результате эксперимента для каждого документа из тестовой выборки было выделено 20 ключевых слов и словосочетаний.

Примеры ключевых слов и словосочетаний, выделенных KEA: *система координат, космический аппарат, система управления, источник энергии, анализировать причину, принимать решение, солнечная система, планета, химический состав, удельный импульс, сила тяги, решать проблему, процесс горения, стартовая масса, компонент топлива, слой атмосферы, космический корабль, космический аппарат*, и т.д.

3.3. Оценка качества автоматического выделения ключевых слов и словосочетаний

Существует два основных подхода к оценке качества ключевых слов и словосочетаний, выделенных автоматически, и они оба, так или иначе, предполагают участие авторов или информантов.

Первый подход основан на вычислении стандартных метрик из области информационного поиска — точности и полноты. Автоматически выделенная ключевая информация сравнивается с так называемым «золотым стандартом», представленным ключевыми словами и словосочетаниями, назначенными автором. Разумеется, у этого подхода есть свои недостатки. Во-первых, выделенные автором выражения не всегда наблюдаются в тексте. Во-вторых, их выбор порой преследует также несколько иные цели помимо краткого описания документа. В-третьих, далеко не каждый документ содержит ключевые слова, выделенные вручную. В-четвёртых, зачастую авторы выбирают лишь небольшое количество слов и словосочетаний.

Второй подход — оценивание автоматически выделенных ключевых выражений экспертами. Каждому информанту представляют документ и список ключевых слов и словосочетаний, выделенных для него автоматически, и предлагают тем или иным образом оценить релевантность каждого выражения по отношению к данному документу. Как и у предыдущего, у этого подхода тоже есть свои недостатки. Главным из них является, разумеется, субъективность оценки и её последующая объективизация. Второй существенный недостаток — очевидная трудность проведения эксперимента для объёмных документов.

Нами был выбран полностью автоматизированный способ оценки работы данного алгоритма. Во-первых, как отмечалось выше, 40 наиболее частотных биграмм для каждого документа из обучающей выборки были получены автоматически. Во-вторых, непосредственная оценка результатов применения алгоритма KEA при работе с экспериментальными корпусами осуществлялась с помощью построения тематических моделей для каждого из корпусов. Этот выбор обусловлен тем, что ключевые слова и словосочетания, составляющие семантическое ядро корпуса, находят соответствие в тематической модели корпуса (т.е. компоненты *n*-грамм должны быть представлены в составе кластеров, отражающих распределение слов по темам и тем по документам корпуса). При построении тематических моделей корпусов использовался алгоритм LDA (Latent Dirichlet Allocation) в пакете GenSim для Python [12]. В каждой тематической модели отбирались 200 статистически значимых лемм (по 10 из 20 тем), далее фиксировалось их наличие/отсутствие в списках ключевых выражений. За единичными исключениями все леммы в составе тем обнаружены в верхней трети списка ключевых выражений, которая оценивается как наиболее информативная.

4. Заключение

В ходе данного исследования была осуществлена адаптация KEA посредством дополнения алгоритма инструментами для работы с русским языком. Также мы осуществили тестирование работы KEA на материале русского языка и произвели оценку эффективности путем сравнения с тематическими моделями. Полученные данные дают основания считать результаты работы алгоритма KEA приемлемыми, а сам алгоритм в русскоязычной модификации пригодным для использования в лингвистических исследованиях.

Сравнив результаты работы KEA на материале русского языка с выдачей неоднократно протестированных и доказавших свою состоятельность тематических моделей, сгенерированных с помощью алгоритма LDA, мы приходим к выводу, что адаптированный нами алгоритм KEA является весьма полезным инструментом в автоматическом определении тематики текста. Он показывает удовлетворительные результаты, что подтверждается нахождением выделенных им ключевых слов и выражений в наиболее информативной части построенных тематических моделей, которые, как указывалось выше, призваны отражать статистическое распределение слов по темам и тем по документам корпуса.

Что касается перспектив дальнейшего исследования, то планируется сравнение KEA с другими алгоритмами и экспертиза результатов с участием информантов.

Исследование поддержано грантом РФФИ № 16–06–00529 «Разработка лингвистического комплекса для автоматического семантического анализа русскоязычных корпусов текстов с применением статистических методов» (2015–2018 гг.).

Литература

[1] Kaur J., Gupta V. Effective Approaches For Extraction Of Keywords // IJCSI International Journal of Computer Science Issues. Vol. 7. Issue 6. November 2010. http://www.ijcsi.org/papers/7-6-144-148.pdf.

[2] Beliga S. Keyword extraction a review of methods and approaches. 2014. http://langnet.uniri.hr/papers/beliga/Beliga_KeywordExtraction_a_review_of_methods_and_approaches.pdf.

[3] Красавина В.Д., Мирзагитова А.Р. Оптимизация поиска в системе Lead-Scanner с помощью автоматического выделения ключевых слов и словосочетаний // Труды международной конференции «Корпусная лингвистика–2015». СПб., 2015. С. 296−306.

[4] 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. 268−277.

[5] Keyphrase Extraction Algorithm. http://www.nzdl.org/Kea/index.html.

[6] Witten I.H., Paynter G.W., Frank E., Gutwin C., Nevill-Manning C.G. KEA: Practical Automatic Keyphrase Extraction. // Proceedings of the fourth ACM conference on Digital libraries. http://www.cs.waikato.ac.nz/~eibe/pubs/chap_Witten-et-al_Windows.pdf.

[7] Zesch T., Gurevych I. Approximate Matching for Evaluating Keyphrase Extraction. // International Conference RANLP 2009. Borovets, Bulgaria. Pp. 484−489. http://www.aclweb.org/anthology/R09-1086.

[8] Quarteroni S., Manandhar S. Adaptivity in Question Answering with User Modelling and a Dialogue Interface. // Proceedings of the Workshop on Cultural Heritage − 9th Conference of the Italian Association for Artificial Intelligence (AI*IA 2005). Milan, Italy, September 2005. http://www.aclweb.org/anthology/E06-2029 1086.

[9] Jones S., Paynter G.W. Human Evaluation of Kea, an Automatic Keyphrasing System // First ACM/IEEE-CS Joint Conference on Digital Libraries, Roanoke, Virginia, June 24−29, 2001. ACM Press. Pp. 148−156. http://researchcommons.waikato.ac.nz/bitstream/handle/10289/41/content.pdf;jsessionid=1D59C1B9A453FD1867F7ABED39121B06?sequence=1.

[10] Наивный байесовский классификатор.http://help.prognoz.com/ru/mergedProjects/Lib/06_datamining/lib_naivebayes.htm.

[11] Porter M. An algorithm for suffix stripping // Program. 14(3). 1980. Pp. 130−137. https://tartarus.org/martin/PorterStemmer/def.txt.

[12] Морфологический анализатор pymorphy2. http://pymorphy2.readthedocs.io/en/latest/.
[13] Национальный корпус русского языка. http://www.ruscorpora.ru/.
[14] Митрофанова О.А. Вероятностное моделирование тематики русскоязычных корпусов текстов с использованием компьютерного инструмента GenSim // Труды международной конференции «Корпусная лингвистика-2015». СПб., 2015. С. 332–343.
[15] Дубовик А.Р. Автоматическое определение стилистической принадлежности текстов по их статистическим параметрам // Сборник научных статей XIX Объединенной конференции «Интернет и современное общество» IMS-2017. Санкт-Петербург, 2 –23 июня 2017 г. СПб., 2017. [наст. изд.]

Automatic Keyphrase Extraction by applying KEA to Russian texts

E. Sokolova, O. Mitrofanova

Saint-Petersburg State University

KEA (Keyphrase Extraction Algorithm) is one of the well-known algorithms used by researchers working in the field of NLP (Natural Language Processing). KEA is aimed at effective automatic extraction of keywords and phrases. KEA shows good results in processing English texts and proves to be quite useful for linguistic purposes. This algorithm can be very helpful for the users who want to get a general idea of the material they are interested in. We present modification of KEA designed for processing Russian corpora. We linked KEA and necessary linguistic tools for processing Russian texts and then carried out experiments on representative Russian text corpora and examined reliability of our KEA modification. Therefore, in this paper we provide detailed description of KEA's internal organization, its functional peculiarities, experimental settings, results and evaluation procedure.

Keywords: automatic extraction of key words and phrases, KEA, Russian text corpora.