УДК 81’322

МЕТОДЫ И МОДЕЛИ АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ КЛЮЧЕВЫХ СЛОВ

С.О. Шереметьева, П.Г. Осминин
Южно-Уральский государственный университет, г. Челябинск

Дается обзор и классификация основных методов автоматического извлечения ключевых слов из текстовых документов, среди которых выделяются статистические и гибридные с использованием корпуса текстов или на основе отдельного документа. Анализируются преимущества и недостатки каждого из подходов. Отмечается проблематичность применения статистических методик для флективных языков, таких как русский. Формулируются требования к эффективной модели извлечения ключевых слов из текстов на русском языке и даются конкретные рекомендации для их достижения. Подчеркивается, что для создания эффективных экстракторов ключевых слов следует учитывать лингвистические типы естественных языков (аналитический, флективный, агглютинативный, изолирующий), предметную область (подъязык) и наличие необходимых лингвистических и программных ресурсов. Подход иллюстрируется на примере автоматического экстрактора ключевых слов Lana-Key-RU из русскоязычных статей по математическому моделированию.

Ключевые слова: автоматическое извлечение, ключевые слова, русский язык.

Введение

Ключевые слова – это одно- и многокомпонентные лексические группы, отражающие содержание документа [9]. Автоматическое извлечение ключевых слов представляет собой необходимый этап обработки текста в таких важных приложениях как системы автоматического информационного поиска, аннотирования, реферирования и т. д. Однако, несмотря на достаточно большое количество исследований, автоматическое извлечение ключевых слов представляет собой проблему, которая до сих пор не решена [5, 7, 27, 28]. Проблематичным является автоматическое извлечение многокомпонентных ключевых слов, особенно, если делается попытка автоматически извлечь определенные типы лексических групп, например, именные группы. При всех методиках алгоритм верхнего уровня извлечения ключевых слов универсален и включает этапы: а) формирования множества «кандидатов» в ключевые слова и б) фильтрации этого множества для получения результирующего списка ключевых слов. Достаточно часто до извлечения ключевых слов из текста удаляются стоп-слова. Стоп-слова – это слова, которые не несут никакой смысловой нагрузки (артикли, предлоги, союзы, частицы, местоимения, вводные слова, междометия и т. д.). Различие методик определяется процедурами обработки текста на каждом из этапов и количеством необходимых для этих процедур лингвистических знаний. Основные типы методов и моделей автоматического извлечения ключевых слов можно разделить на чисто статистические и гибридные. В рамках указанных подходов можно выделить методы, требующие наличия корпуса текстов одной тематики и методы, не требующие такого корпуса текстов.

1. Статистические модели автоматического извлечения ключевых слов

Наиболее простой статистический метод извлечения ключевых слов предполагает построение множества кандидатов ключевых слов путем ранжирования всех словоформ или лексем документа по частоте. Фильтрация заключается в отборе в качестве ключевых определенного количества наиболее частотных лексем. Этот метод является первым методом автоматического извлечения ключевых слов. Он разрабатывался, например, в работах Г.П. Луна [23], Р.Г. Пиотровского [3] и широко используется до сих пор. Распространенность метода отбора ключевых слов исключительно на основе частот лексем объясняется его простотой.

При использовании частоты слова в документе в качестве единственного параметра для автоматического извлечения ключевых слов подсчет общей частоты словоформ из парадигмы одной лексемы чаще всего осуществляется следующим образом: общая частота ключевых слов подсчитывается путем сравнения словоформ, нормализованных к одной форме, как правило, к основе или лемме. Автоматическая нормализация словоформы по сути дела представляет собой задачу морфологического анализа и достаточно проблематична сама по себе.

При статистических подходах к извлечению ключевых слов используются простые эвристические алгоритмы, чаще всего нормализующие словоформу к ее квази-основе, отсекая от словоформы определенное количество букв. Такие алгоритмы называют стемминг-алгоритмами, наиболее известным из которых является стемминг-алгоритм Портера [30]. Нормализованные словоформы ранжируются по частоте и те из них, чья частота выше заданного порога, считаются ключевыми. Ключе-

вые слова, как правило, выдаются в усеченном виде квази-основ. Статистические методы извлечения многокомпонентных ключевых слов в качестве необходимого этапа построения множества кандидатов включают вычисление n-грам [16, 34].

С одной стороны, частота употребления слова несомненно характеризует важность слова для данного документа, но, с другой стороны, ключевые слова, как подчеркивали исследователи группы «Статистика речи» Р.Г. Пиотровского, и другие, не всегда являются самыми частотными [2, 32]. Часто именно уникальные термины более точно сигнализируют о теме документа, например, о новизне изобретения в патентных документах.

Для учета параметров частотности и уникальности лексем текста, для вычисления релевантности ключевых слов документа широко используется метод TF-IDF [17, 31] с применением корпуса одинаковых по тематике документов. Релевантность ключевых слов в данном случае определяется как произведение двух мер: частоты слова в документе (TF = Term Frequency) и обратной частоты слова в коллекции документов (IDF = Inverse Document Frequency). Последнее означает количество документов в корпусе, где термин употреблен по крайней мере один раз.

Использование корпуса текстов для повышения корректности извлечения ключевых слов получило достаточно широкое распространение, однако отсутствие таких корпусов для каждой конкретной предметной области в реальной жизни делает применение таких корпусных моделей и методов весьма проблематичным.

В стремлении более точно отразить содержание документа разрабатываются методики, использующие в качестве меры релевантности вес лексемы, складывающийся из некоторой комбинации значений различных параметров лексем, например, частоты лексемы в документе, расположения в определенной части текста (например, в заголовке или начале параграфа), статистики совместной встречаемости слов в документе/корпусе и их дисперсии [24, 36] или отношения логарифмического правдоподобия [10].

Преимуществами чисто статистического подхода являются универсальность алгоритмов извлечения ключевых слов и отсутствие необходимости в трудоемких и времязатратных процедурах построения лингвистических баз знаний. Несмотря на указанные преимущества статистических методов извлечения ключевых слов, чисто статистические методы часто не обеспечивают удовлетворительного качества результатов. При этом область их применения ограничена языками с бедной морфологией, такими как английский, где частотность словоформ одной лексемы велика. Чисто статистические модели извлечения ключевых слов, удовлетворительно работающие, например, на материале английского языка, не пригодны для естественных языков с богатой морфологией, в частности, для русского языка, где каждая лексема характеризуется большим количеством словоформ с низкой частотностью в каждом конкретном тексте.

2. Гибридные модели автоматического извлечения ключевых слов

Для повышения корректности автоматического извлечения ключевых слов используются гибридные методики, в которых статистические методы обработки документов дополняются одной или несколькими лингвистическими процедурами (морфологическим, синтаксическим, и семантическим анализами) и лингвистическими базами знаний различной глубины (словарями, онтологиями, грамматиками, лингвистическими правилами и т. д.).

Гибридные методы извлечения ключевых слов из документа, также как и статистические, могут требовать или не требовать корпуса текстов. Не требующие корпуса гибридные методы извлечения ключевых именных групп описаны, например, в работах [6, 14, 20, 35]. Метод Кена Баркера и др., представленный в [6], включает поиск в тексте документа базовых именных групп (БИГ) с использованием морфо-синтаксического анализа на основе словарей и вычисление релевантности БИГ. Ключевыми считаются именные группы с показателем релевантности выше заданного порога.

Гибридный метод извлечения ключевых именных фраз, разработанный С.О. Шереметьевой [35] для английского языка, не требует наличия корпуса текстов, предусматривает построение множества кандидатов посредством вычисления всех n-грам документа и фильтрацию этого множества с помощью правил удаления n-грам, не являющихся именными фразами, и вычисления релевантности «уцелевших» n-грам-именных групп (см. более подробное описание в следующем разделе).

В гибридных методах извлечения ключевых слов на основе графов [11, 21, 25, 29], а также в работах Р. Михальцеа [26], Д. Усталова [4] основной процедурой является построение взвешенного графа, в вершинах которого стоят лексемы-кандидаты в ключевые слова, а дуги взвешены в соответствии со степенью близости кандидатов-вершин. Ключевые слова отбираются в процессе обработки графа алгоритмами из теории графов. Различие между этими методами заключается в способах отбора множества кандидатов-вершин и определения близости отдельных кандидатов, которые, наряду со статистическими параметрами отбора (например, близостью расположения в тексте, вычисляемой по количеству слов между двумя терминами), основаны на морфологическом, синтаксическом, а иногда и семантическом анализе, например, с помощью статей Википедии (чему посвящены, например, работы Гриневой [1, 12].

К числу гибридных методов извлечения ключевых слов можно отнести методы на основе машинного обучения, где задача извлечения ключе-

вых слов рассматривается как задача классификации. Методы на основе машинного обучения для создания обучающей выборки и построения модели-классификатора, как правило, требуют корпуса документов с размеченными ключевыми словами. Помеченные ключевые слова считаются положительным примером, остальные слова – отрицательным примером. Далее высчитывается релевантность каждого слова тренировочного текста путем сопоставления ему вектора значений различных параметров, например, меры TF-IDF, длины слова, части речи, положения слова в заголовке, положения слова в первом абзаце, последнем абзаце, в списках литературы и т. д. Фиксируются отличие значений векторов этих параметров для ключевых слов и не ключевых. Далее вычисляется вероятность отнесения каждого слова к группе ключевых и задается ее порог, т. е. модель обучается. Извлечение ключевых слов из нового документа происходит путем вычисления релевантности слов и их вероятности отнесения к ключевым в соответствии с построенной моделью.

Среди методов на основе машинного обучения можно отметить:

- байесовские методы [8, 18, 41, 38];
- метод опорных векторов [13, 15, 19];
- деревья решений [37];
- использование нейронных сетей [22, 33, 40].

Анализ существующих методов автоматического извлечения ключевых слов показывает, что для создания эффективных экстракторов ключевых слов следует учитывать, лингвистические типы естественных языков (аналитический, флективный, агглютинативный, изолирующий), предметную область (подъязык) и наличие необходимых лингвистических и программных ресурсов.

3. Экстрактор ключевых слов для русскоязычных текстов LanAKey_Ru

При разработке экстрактора ключевых слов LanaKey_Ru для текстов на русском языке нашей целью было разработать модель с последующей программной реализацией, которая была бы достаточно универсальной и позволяла настройку на извлечение различных лексических групп и тексты различных предметных областей. При этом модель должна отвечать следующим требованиям:

- обеспечить лингвистически корректные результаты; извлеченные ключевые слова должны быть грамматически правильными лексическими группами (именными, глагольными и т. д.), а не усеченными цепочками квази-основ;
- не требовать заранее построенного корпуса предметной области, поскольку в реальной жизни таких корпусов для каждой конкретной предметной области, как правило нет и их создание – далеко не тривиальная задача;
- обладать вычислительно привлекательными свойствами и обеспечивать высокую скорость обработки текстов;
- обеспечить корректное извлечение не только высокочастотных, но и низкочастотных лексических единиц, что позволило бы извлекать все лексические единицы определенного типа, а не только ключевые;
- обеспечить достаточно быстрое создание программного инструмента посредством повторного использования и адаптации уже существующих методик и программного обеспечения.

При разработке экстрактора ***LanAKey_Ru*** мы повторно использовали программную оболочку и методику извлечения ключевых слов, созданную С.О. Шереметьевой для извлечения номинативных многокомпонентных ключевых слов из патентов на английском языке [35]. Преимуществами этой модели является то, что она

- отвечает сформулированным выше требованиям;
- основана на универсальном (не зависящем от конкретного языка) алгоритме извлечения лексических единиц определенных типов;
- дает возможность ранжирования извлеченных лексических на основе различных векторов релевантности;
- требует неглубокой, зависимой от конкретного языка базы знаний, которая состоит:
  - из набора стоп-листов иконически перечисленных словоформ лексем, запрещенных для использования в лексических группах определенного типа,
  - правил исключения из списка кандидатов групп, содержащих словоформы из стоп-листов;
  - правил нормализации кандидатов-словоформ к одной лексеме;
- имеет программную оболочку, которая позволяет менять (обновлять) базу знаний без вмешательства программиста.

Процедура извлечения ключевых слов состоит из следующих этапов:

1. Вычисление n-грам (n=1,2,3,4) из исходного документа (статьи).

2. Удаление n-грам, которые не могут быть лексическими группами требуемого типа, с использованием стоп-лексиконов и правил исключения и получение списка словоформ кандидатов-именных групп.

3. Нормализация словоформ к одной лексеме

4. Вычисление релевантности каждой извлеченной лексической группы.

5. Выдача списка ключевых слов, удовлетворяющего определенному порогу релевантности или всех лексических (например, именных) групп.

Первая экстраполяция английского экстрактора на русский язык сделана для извлечения номинативной терминологии, т. е. именных групп (ИГ), в том числе и ключевых, для подъязыка

математического моделирования. Именные группы считаются наиболее релевантными ключевыми словами, поскольку они наиболее близко отражают содержание документа [39]. Процедура экстраполяции заключалась в замене английских стоп-листов русскими с учетом грамматики русской именной группы и выборе вектора релевантности для отбора из всех извлеченных именных групп ключевых.

База знаний (стоп-листы) этого экстрактора построена на основе статистического анализа корпуса научных статей по математическому моделированию объемом 140 000 словоупотреблений. Для нормализации словоформ-кандидатов ИГ разработан упрощенный алгоритм на основе совпадения определенного количества знаков слов. Вектор релевантности ключевых именных групп $R = (F, N, n, T, M, U)$ вычисляется по эмпирически определенной формуле:

$$R = (F/N)\cdot 30 + n\cdot 30 + U + T + M/n,$$

где R – релевантность ИГ, F – частота ИГ, N – средняя частота ИГ определенной длины, n – длина ИГ (1–4 компонента), T – количество самых частотных слов, содержащихся в ИГ; наиболее частотными считаются 30 % 1-грам (слов) с наиболее высокой частотой; M – сумма частот слов, входящих в ИГ, U – уникальность; этот параметр показывает, что этот кандидат функционирует индивидуально, а не в составе более длинной именной группы. Уникальность вычисляется как разность между частотой данной именной группы и суммой частот более длинных именных групп, содержащих данную.

В качестве ключевых извлекаются первые десять наиболее релевантных именных групп. Экстрактор допускает извлечение именных групп в текстовой форме и в основной форме с кумулятивной релевантностью.

Описанный экстрактор экстраполирован на другие предметные области русского языка и адаптирован для извлечения других лексических групп.

Заключение

Основные типы методов и моделей автоматического извлечения ключевых слов делятся на чисто статистические и гибридные. Преимуществами чисто статистического подхода являются универсальность алгоритмов извлечения ключевых слов и отсутствие необходимости в трудоемких и времязатратных процедурах построения лингвистических баз знаний. Однако статистические методы часто не обеспечивают удовлетворительного качества результатов. При этом область эффективного применения статистических моделей ограничена языками с бедной морфологией, они, как правило, не пригодны для естественных языков с богатой морфологией, в частности, для русского языка.

Большим потенциалом обладают гибридные методики, в которых статистические методы обработки документов дополняются одной или несколькими лингвистическими процедурами и лингвистическими базами знаний различной глубины. Не всегда возможным оказывается использование методик с использованием корпусов текстов в связи с отсутствием таковых для каждой конкретной предметной области.

Анализ существующих методов автоматического извлечения ключевых слов показывает, что для создания эффективных экстракторов ключевых слов следует учитывать лингвистические типы естественных языков (аналитический, флективный, агглютинативный, изолирующий), предметную область (подъязык) и наличие необходимых лингвистических и программных ресурсов, что иллюстрируется на примере разработки экстрактора ***LanAKey_Ru.***

Литература/References

1. Гринева М., Гринев М. Анализ текстовых документов для извлечения тематически сгруппированных ключевых терминов. Труды ИСП РАН. 2009. Т. 16. С. 155–165. [Grineva M. Analiz tekstovykh dokumentov dlya izvlecheniya tematicheski sgruppirovannykh klyuchevykh terminov(Analysis of Text Documents for Extraction of the Thematically Grouped Keyterms). *Trudy ISP RAN* (Proceeding of ISP RAS). 2009, vol. 16, pp. 155–165.]

2. Алексеев П.М., Герман-Прозорова Л.П., Пиотровский Р.Г., Шепетова О.П. Основы статистической оптимизации преподавания иностранных языков. Статистика речи и автоматический анализ текста. Л., 1974. С. 195–234. [Alekseev P.M., German-Prozorova L.P., Piotrovskii R.G., Shepetova O.P. Osnovy statisticheskoy optimizatsii prepodavaniya inostrannykh yazykov (Basics of the Statistical Optimization of Foreign Languages Teaching). *Statistika rechi i avtomaticheskiy analiz teksta* (Statistics of Speech and Automatic Analysis of the Text). Leningrad, 1974, pp. 195–234.]

3. Пиотровский Р.Г., Бектаев К.Б., Пиотровская А.А. Математическая лингвистика: учеб. пособие для пед. институтов. М.: Высшая школа, 1977. 383 с. [Piotrovskiy R.G.., Bektaev K.B., Piotrovskaya A.A. *Matematicheskaya lingvistika.* (Mathematical Linguistics). Moscow, Vysshaya shkola, 1977, 383 p.]

4. Усталов Д. Извлечение терминов из русскоязычных текстов при помощи графовых моделей. http://koost.eveel.ru/science/ CSEDays2012. pdf (дата обращения: 30.11.2014). [Ustalov D. *Izvlechenie terminov iz russkoyazychnykh tekstov pri pomoshchi grafovykh modeley* (Term Extraction by Means of Graph Model from Russian texts). Available at: http://koost.eveel.ru/science/CSEDays2012.pdf (accessed: 30.11.2014)]

5. Liu Z., Huang W., Zheng Y., Sun M. Automatic keyphrase extraction via topic decomposition. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, Massachusetts, 2010, pp. 366–376.

6. Barker K. Cornacchia N. Using Noun Phrase Heads to Extract Document Keyphrases. *Advances in Artificial Intelligence*. 2000, vol. 1822, pp. 40–52.

7. Piao S.S., Rayson P., Archer D., McEnery T. Comparing and Combining a Semantic Tagger and a Statistical Tool for MWE Extraction. *Computer Speech & Language*. 2005, vol. 19, no. 4, pp. 378–397.

8. Frank E., Paynter G.W., Witten I.H., Gutwin C, Nevill-Manning C.G.. Domain-Specific Keyphrase Extraction. *Proceeding of 16th International Joint Conference on Artificial Intelligence*. Stockholm, Sweden, 1999, pp. 668–673.

9. Dostal M. Automatic Keyphrase Extraction Based on NLP and Statistical Methods. *Proceedings of the Dateso 2011: Annual International Workshop on Databases, Texts, Specifications and Objects.* Pisek, Czech Republic, 2011, pp. 140–145.

10. Dunning T. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics – Special Issue on Using Large Corpora*. 1993, vol. 19, no. 1, pp. 61–74.

11. Girish K.P. Keyword Extraction from a Single Document Using Centrality Measures. *Pattern Recognition and Machine Intelligence.* Springer Berlin Heidelberg, 2007, pp. 503–510.

12. Grineva M. Effective Extraction of Thematically Grouped Key Terms From Text. Available at: http://www.aaai.org/Papers/Symposia/Spring/2009/SS-09-08/SS09-08-010.pdf (accessed 30.11.2014)

13. Herbrich R. *Large Margin Rank Boundaries for Ordinal Regression. Advances in Large Margin Classifiers*. MIT Press, 2000, pp. 115–132.

14. Hulth A. Improved Automatic Keyword Extraction Given More Linguistic Knowledge. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*. Sapporo, July, 2003, pp. 216–223.

15. Jiang X. A Ranking Approach to Keyphrase Extraction. *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Boston, MA, USA, 2009, pp. 756–757.

16. Jiao H. Chinese Keyword Extraction Based on N-Gram and Word Co-occurrence. Proceeding CISW '07 Proceedings of the 2007 International Conference on Computational Intelligence and Security Workshops. Harbin, 2007. pp. 152–155.

17. Jones K.S. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation.* 2004, vol. 60, no. 5, pp. 493–502.

18. KEA: Practical Automatic Keyphrase Extraction. I.H. Witten, G. W. Paynter, G. W. Paynter, E. Frank, C. Gutwin, C. G. Nevill-Manning. *DL '99 Proceedings of the Fourth ACM Conference on Digital Libraries*. Berkeley, CA, USA, 1999, pp. 254–255.

19. Keyword Extraction Using Support Vector Machine. K. Zhang, H. Xu, J. Tang, J. Li. Advances in Web-Age Information Management. *Springer Berlin Heidelberg*, 2006, pp. 85–96.

20. Krulwich B. Learning User Information Interests through Extraction of Semantically Significant Phrases. Available at: http://www.aaai.org/ Papers/Symposia/Spring/1996/SS-96-05/SS96-05-018. pdf (accessed: 30.11.2014).

21. Litvak M. Graph-based Keyword Extraction for Single-Document Summarization. *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization*. Manchester, United Kingdom, 2008, pp. 17–24.

22. Lopez P. HUMB: Automatic Key Term Extraction from Scientific Articles in GROBID. *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden, 2010, pp. 248–251.

23. Luhn H.P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*. 1957, vol. 1, no. 4, pp. 309–317.

24. Matsuo Y. Keyword Extraction from a Single Document Using Word co-occurrence Statistical Information. *International Journal on Artificial Intelligence Tools*. 2004. V. 13, no. 1, pp. 157–169.

25. Matsuo Y. KeyWorld: Extracting Keywords from Document s Small World. *Discovery Science*. Springer Berlin Heidelberg, 2001, pp. 271–281.

26. Mihalcea R. TextRank: Bringing Order into Texts. *Proceedings of EMNLP 2004. Barcelona*, Spain, 2004, pp. 404–411.

27. Multiword Expressions: A Pain in the Neck for NLP. I.A. Sag, T. Baldwin, F. Bond, A. Copestake, D. Flickinger. *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing CICLing '02*. London, UK. 2002, pp. 1–15.

28. Nallapati R. Extraction of Key Words from News Stories. Available at: https://sites.google.com/site/nmramesh77/research-papers/2002_synthesis_report.pdf?attredirects=0 (accessed: 30.11.2014).

29. Ohsawa Y. KeyGraph: Automatic Indexing by co-occurrence Graph Based on Building Construction Metaphor. *ADL '98 Proceedings of the Advances in Digital Libraries Conference*. Santa Barbara, CA, USA, 1998, pp. 12–18.

30. Porter M.F. An Algorithm for Suffix Stripping. Readings in Information Retrieval. Morgan Kaufmann Publishers Inc., 1997, pp. 313–316.

31. Salton G.A Vector Space Model for Automatic Indexing. *Communications of the ACM*. 1975, vol. 18, no. 11, pp. 613–620.

32. Salton G. On the Specification of Term Values in Automatic Indexing. Journal of Documentation. 1973, vol. 29, no. 4, pp. 351–372.

33. Sarkar K., Nasipuri M., Ghose S. A New Approach to Keyphrase Extraction Using Neural Networks. International Journal of Computer Science Issues. 2010, vol. 7, no. 2, pp. 16–25.

34. Sarkar, K. An N-Gram Based Method for Bengali Keyphrase Extraction / K. Sarkar // Information Systems for Indian Languages. Springer Berlin Heidelberg, 2011, pp. 36–41.

35. Sheremetyeva S. An efficient patent keyword extractor as translation resource. MT Summit XII: Third Workshop on Patent Translation. Ottawa, 2009. Pp. 25–32.

36. Smadja F. Retrieving collocations from text: Xtract. Computational Linguistics – Special issue on using large corpora: I. 1993, vol. 19, no. 1, pp. 143–177.

37. Turney P.D. Learning Algorithms for Keyphrase Extraction. Information Retrieval. 2000, vol. 2, no. 4, pp. 303–336.

38. Uzun Y. Keyword Extraction Using Naive Bayes. Available at: http://www.cs.bilkent.edu.tr/~guvenir/courses/CS550/Workshop/Yasin_Uzun.pdf (accessed: 30.11.2014).

39. Vadas D. Statistical Parsing of Noun Phrase Structure. Available at: http://sydney.edu.au/ engineering/it/~dvadas1/papers/vadas09_phd_thesis. pdf (accessed 30.11.2014).

40. Wang J., Peng H., Hu J.-S. Automatic Keyphrases Extraction from Document Using Neural Network. Advances in Machine Learning and Cybernetics. Springer Berlin Heidelberg, 2006, pp. 633–641.

41. Wasserman S., Faust K.. Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press, 1995. 857 p.

Шереметьева Светлана Олеговна, доктор филологических наук, профессор кафедры лингвистики и межкультурной коммуникации, Южно-Уральский государственный университет (Челябинск), linklana@yahoo.com

Осминин Павел Григорьевич, преподаватель кафедры лингвистики и межкультурной коммуникации, Южно-Уральский государственный университет (Челябинск), osperevod@gmail.com

Поступила в редакцию 28 ноября 2014 г.

ON METHODS AND MODELS OF KEYWORD AUTOMATIC EXTRACTION

***S.O. Sheremetyeva**, South Ural State University, Chelyabinsk, Russian Federation, linklana@yahoo.com*
***P.G. Osminin,** South Ural State University, Chelyabinsk, Russian Federation, osperevod@gmail.com*

The paper presents an overview and classification of major approaches to the automatic extraction of keywords from text documents. The approaches can be divided into statistical and hybrid approaches. Both of these types can be further classified into corpora-based and document-based. Advantages and shortcomings of particular approaches are analyzed. It is claimed that the use of statistical keyword extraction methods for inflecting languages, such as Russian, is problematic. Requirements to the efficient model of automatic keyword extraction from texts in Russian are formulated and particular recommendations to meet these requirements are given. It is emphasized that to create effective keyword extractors one should take into consideration the linguistic types of natural languages (analytical, inflecting, agglutinative, isolating), the domain (sublanguage) and the availability of linguistic and programming resources. The approach is illustrated by a case study of a keyword extractor for Russian texts on mathematical modeling.

Keywords: automatic extraction, keywords, Russian.

Svetlana O. Sheremetyeva, PhD (Habilitation), professor of the Linguistics and Intercultural Communication department, South Ural State University (Chelyabinsk), linklana@yahoo.com

Pavel G. Osminin, assistant professor of the Linguistics and Intercultural Communication department, South Ural State University (Chelyabinsk), osperevod@gmail.com

Received 28 November 2014

БИБЛИОГРАФИЧЕСКОЕ ОПИСАНИЕ СТАТЬИ

Шереметьева, С.О. Методы и модели автоматического извлечения ключевых слов / С.О. Шереметьева, П.Г. Осминин // Вестник ЮУрГУ. Серия «Лингвистика». – 2015. – Т. 12, № 1. – С. 76–81.

REFERENCE TO ARTICLE

Sheremetyeva S.O., Osminin P.G. Kokhanova L.A. On Methods and Models of Keyword Automatic Extraction. *Bulletin of the South Ural State University. Ser. Linguistics.* 2015, vol. 12, no. 1, pp. 76–81. (in Russ.)