Проблематика оценивания алгоритмов автоматического извлечения ключевых слов

Храмцов Н.С., Академия ФСО России
nadya255@mail.ru

Аннотация

В данной работе рассматривается такая область автоматической обработки текстов, как извлечение ключевых слов (КС). На основе анализа 123 публикаций приведена описательная статистика ряда показателей перспективных из существующих алгоритмов извлечения КС. Обсуждается проблематика практического оценивания качества решений по извлечению КС с учетом специфики данной задачи.

1 Введение

Важную роль в поиске релевантной информации, индексации сайтов и автоматическом реферировании текстов выполняют ключевые слова (КС), которые либо указываются автором текста, в основном для научных работ, либо извлекаются автоматически с помощью программных средств. Постоянное увеличение информации, находящейся в Интернете, стало причиной создания алгоритмов, которые автоматически выделяют КС из текста. Ключевыми словами называют особо важные, общепонятные, ёмкие и показательные для отдельно взятой культуры слова в тексте, набор которых может дать высокоуровневое описание его содержания для читателя, обеспечив компактное представление и хранение его смысла в памяти [Ванюшкин, 2016]. Под данное определение попадают также фразы, состоящие из двух и более слов, называемые ключевыми фразами.

Важность тематики подтверждается существованием большого числа публикаций, в которых предлагаются подходы и конкретные алгоритмические решения по извлечению ключевых слов и фраз, описываются результаты применения известных решений к новым данным и языкам и т.д.

С целью систематизации данных о текущем состоянии предметной области и поиска алгоритма с лучшими показателями была составлена таблица, включающая данные из 123 просмотренных научных публикаций, описывающих разработки в области автоматического извлечения КС. В данной статье приводится описательная статистика по реализованной таблице, а также обсуждаются наблюдаемые результаты.

2 Состояние предметной области

Данные статей, посвященных автоматическому извлечению КС, были систематизированы в виде таблицы, состоящей из описания статей (ФИО автора, организация, которую они представляют, название статьи, цитируемость и год публикации) и характеристик алгоритмов (тип используемого в работе алгоритма математического аппарата, язык, для которого алгоритм применялся, значения метрик качества и наименование текстового корпуса, на котором они были получены).

Результаты анализа таблицы показали, что для англоязычных текстов существует множество алгоритмов, обладающих высокими значениями показателей качества и цитируемости. При этом объем работ, применимых к русскому языку, составил лишь 8% от общего числа рассмотренных статей. Отличительной особенностью алгоритмов для извлечения КС из русскоязычных текстов, является то обстоятельство, что за основу каждого из них брался какой-либо известный алгоритм для английского языка, а затем проводилась его адаптация к русскому языку. Типичным примером здесь может служить работа [Адаменко, 2017].

Стоит отметить, что значение F_1-меры – метрики, отражающей качество работы алгоритма – было указано авторами в 57% работ. В оставшейся части приведены либо только значения полноты или точности, либо полностью отсутствовали оценки эффективности. Одним из объяснений наблюдаемой картины является определенная сложность и вариативность расчета данных метрик применительно к задаче извлечения ключевых слов. Так, на рис. 1 приведены два из множества возможных вариантов сравнения ключевых фраз, приписанных тексту экспертом (эталонный список) и выделенных автоматически алгоритмом. В примере А (слева) выбор КС считается одинаковым, если существует пересечение хотя бы в одном слове. Т.е. перестановки одинакового

набора слов эквивалентны. Также схожими являются пересечения наборов слов в фразах. Напротив, вариант Б (справа) предполагает «жесткое» определение точности работы алгоритма через учет полных совпадений.

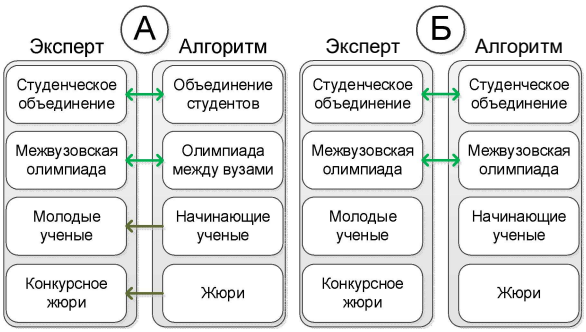

Рис. 1 Схема определения совпадений КС эксперта и алгоритма

В зависимости от того, каким из способов расчета пользовался автор алгоритма, значения метрик качества могут значительно отличаться. Только в четырех работах из рассмотренных был указан способ, по которому определялись совпадения КС, а как это производилось в остальных исследованиях, авторами статей не указано.

Помимо общности способов расчета показателей эффективности, для сравнения различных алгоритмов между собой необходимо добиться, чтобы они испытывались на одинаковых входных данных. В этом отношении заметим, что в рассмотренных статьях авторами было использовано 25 различных текстовых корпусов, на основе которых проводились вычисления. В зависимости от параметров включенных в корпус текстов (длина, стиль, сложность, наличие текстовых и нетекстовых объектов – таблиц, библиографических ссылок и т.д.), значения метрик будут отличаться. Поэтому сравнивать показатели алгоритмов, полученные на различных корпусах, нецелесообразно. Более того, далеко не каждый корпус обладает подходящими характеристиками с точки зрения извлечения КС [Ванюшкин, 2017; Vanyushkin, 2018].

Исходя из перечисленных аспектов, первым шагом для оценки качества работы алгоритма является определение того, одинаковый ли смысл вкладывают разработчики в это понятие (то есть, что именно понимается под ключевыми словами и сколько их должно быть). Второй шаг – выбор единого и адекватного задаче текстового корпуса, на основе которого будут проводиться вычисления. В завершение следует определить способ расчета метрик, используемых для оценки эффективности работы алгоритма.

3 Единство информационной базы

Обеспечение единства исходных данных – очевидное условие для сравнения между собой алгоритмов извлечения КС. Однако на практике достичь этого сложно.

Так, авторы только 56% рассмотренных публикаций указали источник, на основе которого они проводили вычисления показателей работы алгоритмов. Отсутствие в тексте статьи указания, на основе какого корпуса проводились расчеты, лишает нас возможности перепроверить представленные данные. На рис. 2 представлена цитируемость наиболее востребованных корпусов.

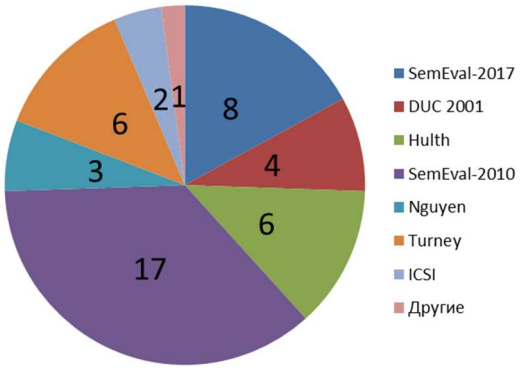

Рис. 2. Цитируемость текстовых корпусов, используемых в задачах извлечения ключевых слов

В категорию «другие» были отнесены корпуса, на которые ссылались один раз. Из гистограммы видно, что наибольшей цитируемостью обладают корпуса SemEval-2010[1] и SemEval-2017[2], используемые в качестве единого для соревнований, один из разделов которой был посвящен алгоритмам автоматического извлечения КС. Это было сделано для того, чтобы показатели алгоритмов были объективными и сравнимыми.

Авторы корпуса SemEval-2010 отмечают определенную проблему с формированием эталонных списков КС. Так, около 15% приписанных приглашенными экспертами и 19% приписанных авторами текстов КС отсутствовали в самих текстах. Поэтому, несмотря на

[1] http://semeval2.fbk.eu

[2] http://alt.qcri.org/semeval2017

высокую цитируемость, корпуса SemEval не в полной мере удовлетворяют требованиям эталонного набора данных. При этом, стоит отметить, что организация разметки текстового корпуса ключевыми словами далеко не тривиальна, и является самостоятельной научной задачей [Ванюшкин, 2018].

При этом, наиболее пригодным с точки зрения распределения длин текстов корпусом является корпус DUC-2001 [Ванюшкин, 2017], но он использовался исследователями лишь четыре раза.

Отсутствие единого и правильно размеченного тестового корпуса, на котором проводились бы испытания различных алгоритмов извлечения КС, является одной из причин того, что наблюдается широкий разброс приводимых исследователями значений F_1-меры, а также с течением времени не просматривается тенденции к улучшению этого показателя.

Перед тем, как проиллюстрировать это утверждение, отметим, что в зависимости от типа математического аппарата системы распознавания выделяют статистические, структурные и нейросетевые методы извлечения КС [Ванюшкин, 2016]. На рис. 3 представлен генезис приведенных авторами значений F_1-меры применительно к статистическим алгоритмам, а на рис. 4 – применительно к структурным алгоритмам.

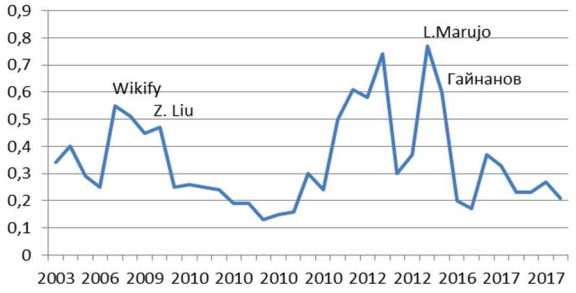

Рис. 3. Ретроспектива заявленных авторами значений F1-меры для статистических алгоритмов

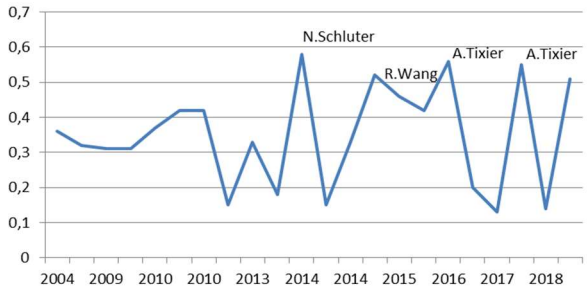

Рис. 4. Ретроспектива заявленных авторами значений F1-меры для структурных алгоритмов

На графике отмечены алгоритмы, которые представляют практический интерес ввиду того, что они имеют достаточно высокое значение F_1-меры, рассчитанное для размеченного корпуса. Для статистического типа наивысшими показателями обладают алгоритмы [Shulter, 2014; Wang, 2015; Tixier, 2016], а среди структурных алгоритмов выделяются [Mihalcea, 2007; Liu, 2009; Marujo, 2015].

Нейросетевое направление является сравнительно молодым и редко используемым. На долю таких разработок приходится 9% от общего числа рассмотренных статей. Нейросетевыми алгоритмами, имеющими значение F_1-меры 0.43 и 0.38 и протестированные на едином корпусе, являются работы [Ammar, 2017] и [Tsujimura, 2017] соответственно.

Показательно, что алгоритмы, оцениваемые на одном корпусе, демонстрируют, как правило, меньший разброс значений, рис. 5.

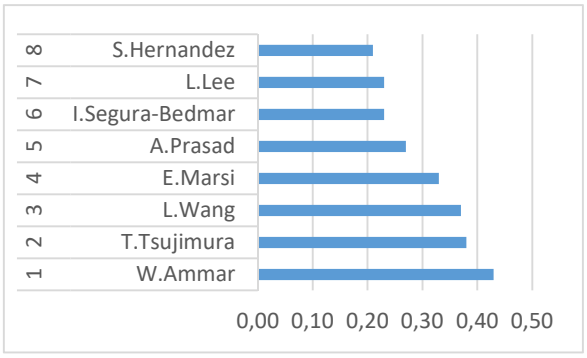

Рис. 5 Значения F_1-меры для алгоритмов, апробированных на корпусе SemEval-2017

Аналогично, на соревновании SemEval-2010 девятнадцать алгоритмов продемонстрировали разброс значений F_1-меры от 0.05 до 0.25 [Kim S.N. et al. 2010].

Это подтверждает тезис о том, что отсутствие единого корпуса может повлечь за собой неоправданно широкий диапазон значений метрик и как следствие, их необъективность. Таким образом, проведение расчетов метрик на едином текстов корпусе является обязательным условием при сравнении алгоритмов и определении среди них лучшего для заданных условий.

4 Проблематика оценивания качества извлечения ключевых слов

В данном параграфе будут рассмотрены особенности расчета показателей эффективности работы алгоритма. К таким показателям относят значения полноты (recall) и точности (precision). Точность – это доля выделенных алгоритмом ключевых слов, совпавших с выбором эксперта относительно эталонного набора ключевых слов (выделенного экспертами). Полнота – это доля выделенных алгоритмом ключевых слов, совпавших с выбором эксперта относительно набора ключевых слов, выделенных алгоритмом.

Эти значения можно рассчитать на основании таблицы, которая составляется для каждого класса отдельно.

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

TP – истинно-положительное решения;

TN – истинно-отрицательное решения;

FP – ложно-положительное решение;

FN – ложно-отрицательное решение На практике помимо полноты и точности используется значение F-меры (F-measure), которое вычисляется по формуле (1)

$$F = (1 + \beta^2)\frac{P \cdot R}{\beta^2 P + R}, \quad (1)$$

где P – точность, R – полнота [Hulth, 2003]. Зачастую в работах под F-мерой понимается значение F_1-меры, которая получается, принимая значение коэффициента $\beta = 1$. Стоит отметить, что выбирая значение $0 < \beta < 1$, при расчете F-меры, больший вес будет отдан значению точности, а при $\beta > 1$, приоритет будет отдан полноте.

Значение F_1-меры полностью зависит от величины полноты и точности работы алгоритма. В свою очередь значения полноты и точности зависят от способа определения совпадений ключевых слов. Например, в работе [Kim S.N. et al. 2010] по итогам соревнования SemEval-2010, одинаковыми признавались словосочетания типа «определение-определяемое» и «определяемое-определение», даже с учетом использования предлогов. того, совпали КС, выбранные экспертом и алгоритмом или нет. А в работе [Paukkeri, 2010] совпавшими считаются только те ключевые фразы, в которых и состав, и порядок слов одинаковый.

Также стоит обратить внимание на ряд особенностей, которые не учитываются большинством авторов. Из-за того, что количество авторских КС (эталонный список) может не всегда совпадать с количеством КС, выделенным алгоритмом, при тестировании системы ее следует настроить таким образом, чтобы количество КС для каждого текста строго совпадало. В противном случае эффективность испытываемого решения будет заведомо меньше, чем могла бы быть.

Еще одной особенностью оценивания качества алгоритмов извлечения КС является то, что эксперт в качестве КС может выбрать слова, отсутствующие в исходном тексте. В таком случае характеристики работы алгоритма будут определены не совсем верно. Поэтому с целью вычисления более качественной оценки не стоит при расчете показателей учитывать те КС, которые в тексте отсутствуют.

Альтернативный способ формирования эталонного списка КС и, соответственно, оценки качества работы алгоритма был предложен авторским коллективом во главе с М. Гринёвой. Эксперт должен был изучить ключевые термины, выделенные автоматически, и, по возможности, расширить собственный набор ключевых слов, то есть дополнить его теми терминами, которые, по его мнению, относятся к главным темам документа, но не были выделены на первом этапе [Grineva, 2009]. После переоценки значения полноты и точности изменились с 0.68 и 0.41 до 0.73 и 0.52 соответственно. Представленный способ также решает проблему различного количества КС, определенных экспертом и алгоритмом.

5 Заключение

Из результатов анализа статей, посвященных автоматическому выделению КС, следует, что вследствие отсутствия набора единых, правильно сформированных и размеченных тестовых корпусов, отсутствия единого способа расчета показателей эффективности и, в целом, понимания критериев отнесения слов к ключевым, затруднительно определить наилучший в настоящее время алгоритм извлечения КС. Сложно даже сказать, происходит ли поступательное улучшение характеристик систем, предлагаемых ежегодно.

Дальнейшим шагом в исследовании данной предметной области будет пересчет показателей алгоритмов, сведенных в описанную таблицу, на едином корпусе и последующее их сравнение с авторскими результатами. Для проведения исследования также потребуется привести к единому формату способ расчета. Выполнив указанные этапы, можно будет выделить те из них, которые наиболее целесообразно адаптировать к русскому языку.

Список литературы

Адаменко А.В. Адаптация алгоритма извлечения ключевых слов TextRank к русскому языку / А.В. Адаменко // Вестник науки. Сборник научных работ аспирантов, магистрантов и студентов физико-математического факультета/под общ. ред. Т.Н. Можаровой. – Выпуск 16. – Орел: ОГУ, 2017. – С. 6-12

Ванюшкин А.С. *Методы и алгоритмы извлечения ключевых слов* / А.С. Ванюшкин, Л.А. Гращенко // Новые информационные технологии в автоматизированных системах. – 2016. – №. 19 – С. 85-93.

Ванюшкин А.С. *Оценка алгоритмов извлечения ключевых слов: инструментарий и ресурсы* / А.С. Ванюшкин, Л.А. Гращенко // Новые информационные технологии в автоматизированных системах. – 2017. – №. 20 – С. 95-102.

Ванюшкин А.С. *О разметке корпусов текстов ключевыми словами* / А.С. Ванюшкин, Л.А. Гращенко // Новые информационные технологии в автоматизированных системах. – 2018. – №. 21 – С. 207-211.

Ammar W., Peters M., Bhagavatula C. and Power R. 2017. S*emisupervised end-to-end entity and relation extraction*. In SemEval-2017.

Grineva M., Grinev M., and Lizorkin D. 2009. *Extracting key terms from noisy and multitheme documents.* 18th International Conference on World Wide Web. ACM, New York, NY, USA.

Hong B. and Zhen D. 2012. *An Extended Keyword Extraction Method.* International Conference on Applied Physics and Industrial Engineering, Physics Procedia, Volume 24

Hulth A. 2003. *Improved automatic keyword extraction given more linguistic knowledge.* Conference on Empirical Methods in Natural Language Processing.

Kim S.N. et al. 2010. *SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles*. Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pp. 21-26.

Liu Z., Li P., Zheng Y., Sun M. 2009 *Clustering to find exemplar terms for keyphrase extraction.* Conference on Empirical Methods in Natural Language Processing, volume 1.

Marujo L. et al. 2015. *Automatic keyword extraction on twitter.* ACL.

Mihalcea R., Csomai A. 2007. *Wikify!: linking documents to encyclopedic knowledge.* ACM CIKM.

Paukkeri M., Honkela T. 2010. *Likey: Unsupervised language-independent keyphrase extraction.* 5th International Workshop on Semantic Evaluation. Uppsala, Sweden

Schluter N. 2014. *Centrality Measures for Non-Contextual Graph-Based Unsupervised Single Document Keyword Extraction.* TALN.

Tsujimura T., Miwa M., Sasaki Y. 2017. *Investigating Embeddings for End-to-End Relation Extraction from Scientific Papers* In SemEval-2017.

Wang R., Liu W., McDonald C. 2015. *Corpus-independent generic keyphrase extraction using word embedding vectors.* Software Engineering Research Conference.

Tixier A., Malliaros F., Vazirgiannis M. 2016. *A graph degeneracy-based approach to keyword extraction.* Conference on Empirical Methods in Natural Language Processing.

Vanyushkin A., Graschenko L. 2018. *An Overview of the Available Corpora for Evaluation of the Automatic Keyword Extraction Algorithms.* Proceedings of Computational Models in Language and Speech Workshop (CMLS 2018). Kazan, Russia. pp. 104-116.

Zhang C., Wang H., Liu Y., Wu D., Liao Y. and Wang B. 2008. *Automatic Keyword Extraction from Documents Using Conditional Random Fields.* Journal of Computational Information Systems.