Тинькова Лариса Игоревна
магистрант
Каменский Михаил Васильевич
доцент, преподаватель
Гуманитарный институт
ФГАОУ ВО «Северо-Кавказский федеральный университет»
г. Ставрополь, Ставропольский край

ПОДХОДЫ К АЛГОРИТМИЗАЦИИ ПРОЦЕССА ИССЛЕДОВАТЕЛЬСКОГО ПОИСКА ЛЕКСИЧЕСКИХ И СИНТАКСИЧЕСКИХ КОНСТРУКЦИЙ

Аннотация*: в данной статье авторами рассмотрены различные подходы к составлению алгоритмов для идентификации лексических и синтаксических конструкций в тексте, не зависимо от его тематики и стиля.*

Ключевые слова*: алгоритм, лексические конструкции, синтаксические конструкции, лексема, фраза.*

Языковая культура практически всех языков мира обладает большим количеством языковых конструкций, однако сфера их применения одна – передача письменной и устной речи. При этом языковые конструкции могут быть как простыми, так и сложными. Основная цель любой языковой конструкции (независимо от уровня её сложности) – правильность передачи мысли или смысла сказанного.

С целью оптимизации исследовательского поиск синтаксических языковых конструкций эффективным является использование алгоритмов.

Как известно, алгоритм – это:

1. Определенная последовательность операций или вычислений (в математике).

2. Программа для электронной вычислительной машины, позволяющая от исходных данных прийти к искомому результату.

3. Алгоритм означает точное описание некоторого процесса, инструкцию по его выполнению» [1].

Разработка алгоритма является сложным и трудоемким процессом. Алгоритмизация – это техника разработки (составления) алгоритма для решения задач на ЭВМ.

Нами рассмотрены основные алгоритмы извлечения различных лексических и синтаксических средств в тексте. Так, подходы к автоматическому извлечению лексем менялись по мере развития моделей лексических средств и теории распознавания образов. Языконезависимый графориентированный алгоритм TextRank был предложен в 2004 году на основе известного алгоритма ранжирования веб-страниц PageRank [6]. Значимость вершины в графе рассчитывается через значимости смежных вершин. Ребром графа здесь может служить любое отношение между лексическими единицами. Для задачи выделения необходимых лексем и синтаксических это отношение смежности или совместного появления, задаваемое расстоянием между словами. Две вершины смежны, если соответствующие им лексические единицы появляются внутри окна ширины $2 \leq N \leq 10$. Перед добавлением вершин в графы может использоваться фильтрация лексики, например по принадлежности к частям речи. После построения графа выполняется подсчет значимости узлов, их ранжирование, и первые 5 – 20 сохраняются для дальнейшей обработки.

В алгоритме Rake сначала формируется список потенциальных модальных средств с помощью заданного словаря разделителей фраз, а затем строится граф, вершины которого – отдельные слова. Особенность такого графа состоит в том, что вершины графа могут быть представлены одинаковыми словами [7].

Значимость для слова определяется набором показателей: частота появления вершины, степень вершины, отношение степени к частоте. Значимость потенциальных модальных слов и выражений рассчитывается как сумма значимостей каждого входящего в него слова. В качестве модального компонента для данного текста отбирается первая треть упорядоченного по убыванию значимости списка вершин. В алгоритме DegExt сперва удаляются стоп-слова, а затем

строится граф, в котором дуги между вершинами проводятся только для соседствующих в любом предложении слов, не разделенных знаками пунктуации [5].

Вершины с наибольшими степенями соответствуют кандидатам в модальное пространство текста. Для извлечения необходимых лексем или синтаксических конструкций необходимо выделить из списка последовательности смежных слов (заданной длины), встречающиеся в одном предложении.

Для каждой фразы или лексемы вычисляется значимость как средняя степень составляющих ее букв. По сравнению с TextRank, данный алгоритм вычислительно менее сложен. Авторы рекомендуют использовать его для выделения большого числа лексических и синтаксических конструкций (около15). Показано, что лучшие результаты извлечения необходимых выражений и слов были достигнуты при использовании нормированной по длине центральности по посредничеству (Length-scaled Betweenness Centrality). Другие графовые алгоритмы отличаются от вышеописанных использованием дополнительной информации о позициях слов, их длине, разметке и форматировании текста форматирование текста и т.д. [2].

В других гибридных алгоритмах на основе обучения применяются байесовская классификация метод условных случайных полей (CRF) и метод опорных векторов [3]. В общем, несмотря на продолжающееся совершенствование классических статистических и структурных алгоритмов извлечения лексических и синтаксических средств, акцент разработчиков сместился в область гибридных решений на основе текстовых корпусов [4].

Анализ существующих методов и алгоритмов показывает, что на фоне роста доступных вычислительных ресурсов в настоящее время усилия исследователей направлены на развитие гибридных технологий. При этом вычислительно более простые графориентированные алгоритмы обладают рядом дополнительных преимуществ, таких как независимость от языка и размеченных корпусов (онтологий).

Список литературы

1. Баканова Н.Б. Обзор программных средств автоматизированного поиска и анализа ключевых слов документов // Проблемы современной науки. – 2013. – №7–3. – С. 40–45.

2. Воронина И.Е. Функциональный подход к выделению ключевых слов: методика и реализация // Вестник Воронежского государственного университета. – 2009. – №1. – С. 68–72.

3. Matsuo Y. Extracting Keywords from Documents Small World. Discov. Sci. Springer Berlin Heidelb. – 2001. – P. 271–281.

4. Hulth A. Improved Automatic Keyword Extraction Given More Linguistic Knowledge. EMNLP'03 Proc. 2003 Conf. Empir. Methods Nat. Lang. Process. – 2003. – №2000. – P. 216–223.

5. Litvak M. DegExt: A language-independent keyphrase extractor. J. Ambient Intell. Humaniz. Comput. / M. Litvak, M. Last, A. Kandel. – 2013. – Vol. 4. – P. 377–387.

6. Mihalcea R. TextRank: Bringing order into texts. Proc. EMNLP. / R. Mihalcea, P. Tarau. – 2004. – Vol. 4. – P. 404–411.

7. Rose S. Automatic Keyword Extraction from Individual Documents. Text Min. Appl. Theory / S. Rose, D. Engel, N. Cramer, W. Cowley. – 2010. – P. 1–20.

8. Ванюшкин А.С., Гращенко Л.А. Методы и алгоритмы извлечения ключевых слов: Текст научной статьи по специальности «Языкознание» [Электронный ресурс]. – Режим доступа: http://cyberleninka.ru/article/n/metody-i-algoritmy-izvlecheniya-klyuchevyh-slov (дата обращения: 27.03.2017).