УДК 81'322.2

ЛИНГВИСТИЧЕСКИЙ СТАТУС ОПОРНЫХ СЛОВОФОРМ И ИХ РОЛЬ В ФОРМИРОВАНИИ РЕФЕРАТА

A REFLECTION OF THE LINGUISTIC STATUS OF KEY WORD FORMS AND THEIR ROLE IN ABSTRACT FORMATION

А.О. Варыгин, студент 5 курса
ФГБОУ ВО «Кубанский государственный университет»
г. Краснодар, Российская Федерация

Научный руководитель: Е.В. Князева, доцент, к.п.н.
ФГБОУ ВО «Кубанский государственный университет»
г. Краснодар, Российская Федерация

Аннотация. Любой человек, которому когда-либо приходилось быстро вникать в содержание научной статьи или книги, знает ценность хорошего резюме. В статье рассматривается один из подходов автоматического выделения наиболее важных частей текста: использование ключевых словоформ. Описан статистический метод определения опорных слов, основанный на коэффициенте важности, который учитывает не только частоту встречаемости слова, но и его распределение по абзацам. Этот метод был реализован на языке программирования VBA. Результаты проведенного эксперимента по обработке научно-технических текстов показали, что автоматически извлекаемые словоформы хорошо совпадают с теми, которые выбирают студенты-лингвисты, а в некоторых случаях даже более полно отражают содержание. Был сделан вывод, что ключевые словоформы (опорные словоформы) действительно служат семантической основой, на которой может быть построено высококачественное резюме.

Ключевые слова: ключевые словоформы, резюме, автоматическое обобщение, статистический анализ, коэффициент важности, семантическое сжатие, распределение слов в тексте.

Abstract. Anyone who has ever needed to quickly grasp the content of a scientific article or book knows the value of a good summary. But how can we automatically extract the most important parts of a text? This paper discusses one approach: the use of key word forms. We clarify what exactly can be considered a key word form and how it differs from a simple keyword. A statistical method based on an importance coefficient is described; it takes into account not only word frequency but also its distribution across paragraphs. This method was implemented in a VBA program, and we conducted an experiment

on a popular science text. The results showed that automatically extracted word forms coincide well with those chosen by linguistics students, and in some cases even reflect the content more fully. We conclude that key word forms indeed serve as a semantic framework upon which a high-quality summary can be built.

Keywords: key word forms, abstract, automatic summarization, statistical analysis, importance coefficient, semantic compression, word distribution in text.

Представим, что нужно просмотреть десяток статей на определенную тему за час. Прочитать каждую из них полностью невозможно – приходится полагаться на аннотации. Но кто и как их составляет? В идеале это делает автор статьи или редакторы. Однако в век цифровых технологий такой подход является слишком медленным и дорогостоящим. Именно поэтому уже несколько десятилетий разрабатываются методы автоматического обобщения, при которых компьютер сам извлекает основные моменты и составляет краткое резюме.

Часто возникает проблема: как объяснить машине, что в тексте важно, а что второстепенно? Наиболее распространенным подходом является поиск слов, несущих основную смысловую нагрузку. Они называются по-разному: ключевые слова, опорные слова, тематические слова, смысловые ориентиры. Мы будем использовать термин «ключевые словоформы». Почему именно «словоформы», а не просто «слова»? Потому что одна и та же лексема может фигурировать в разных грамматических формах, и все они должны рассматриваться как носители одного и того же значения.

Цель нашего исследования – продемонстрировать, что ключевые словоформы могут быть определены статистически и что на их основе можно построить абстракцию, близкую к тому, что создал бы человек. Для достижения этой цели уточним понятие ключевой словоформы. В лингвистической литературе нет единого мнения о том, что является ключевым элементом текста. Некоторые авторы, например, Гребенщикова А.В., приравнивают их к ключевым словам, в то время как другие (Зубов А.В. и Зубова И.И.) предлагают различать словоупотребление, словоформу и лексему. Мы придерживаемся мнения, что ключевая словоформа – это конкретная грамматическая форма слова, которая встречается в тексте. Но при анализе мы группируем их в лексемы, чтобы учесть синонимию и варианты. Если слово «компьютер» встречается 10 раз в единственном числе и 5 раз во множественном, то это явно один и тот же семантический узел. Если автор использует «компьютер» и «вычислительная машина» как синонимы, статистический метод без семантического анализа не позволит их объединить. На первом этапе анализа оставляем это ограничение в стороне, но в будущем его необходимо учитывать. Важно отметить, что ключевая словоформа выполняет несколько функций:

1) обозначает ключевую концепцию (репрезентативная функция);

2) связывает части текста, повторяясь в разных абзацах (текстообразующая функция);

3) позволяет читателю предвидеть последующее содержание (прогностическая функция).

Таким образом, ключевые словоформы – это не просто статистические выбросы, а реальные языковые единицы, которые организуют текст. Для автоматического извлечения ключевых словоформ существуют различные подходы: интуитивный (экспертная оценка), статистический, синтаксический и семантический. У каждого из них есть свои плюсы и минусы. Мы выбрали статистический подход, поскольку он легко алгоритмизируется и не требует сложных лингвистических ресурсов.

Основная идея заключается в том, что слово тем важнее, чем чаще оно встречается и чем более равномерно распределено по тексту. Простая повторяемость может ввести в заблуждение: например, в научной статье часто повторяются такие слова, как «исследование», «метод» и «результат», но они не всегда являются ключевыми для конкретной темы.

Равномерность распределения в тексте состоит в следующем: если термин встречается в каждом абзаце, это, скорее всего, отражает повторяющуюся тему. Для количественной оценки мы использовали коэффициент важности, предложенный А.В. Зубовым и И.И. Зубовой:

$$K_{важ} = \frac{F \times m}{N \times n},$$

где F – частота встречаемости словоформы в тексте; m – количество абзацев в тексте, в которых встречается эта словоформа; N – общее количество встречаемости слов в тексте; n – общее количество абзацев в тексте. В этом состоит вероятностный подход в автоматизации аннотирования и реферирования.

Чем больше m и F, тем выше коэффициент. Мы также ввели пороговые значения для классификации словоформ на главные опорные слова (ГОС) и второстепенные опорные слова (ВОС). Эти пороговые значения зависят от объема текста и определяются эмпирически. В нашем эксперименте для текста из 7 абзацев (395 слов) ГОС были те, у которых коэффициент $K \geq 0{,}0032$, а ВОС были те, у которых $0{,}0027 \leq K < 0{,}0032$.

Чтобы убедиться, что метод работает, мы выбрали научно-технический текст на английском языке об истории развития компьютеров. Он состоял из 7 абзацев и 395 слов. Текст был обработан программой, написанной на VBA в среде MS Word. Программа рассчитала частотность всех словоформ, вычислила коэффициенты и выделила ключевые слова. Основными ключевыми словоформами (лексемами) были: chip (0,0065), circuit (0,0052), computer (0,0380) и т.д. Второстепенным был microprocessor (0,0029). Эти слова действительно отражают основное

содержание текста – в нем обсуждались микросхемы, процессоры и память.

В рамках проведенного эксперимента студентами-лингвистами были отобраны в предложенном тексте и перечислены ключевые слова, которые, по их мнению, лучше всего передают его содержание. Затем мы сравнили их списки с ключевыми словоформами, полученными как результат работы программы. Все студенты перечислили компьютер, микросхему, микропроцессор, которые были включены в ГОС. Некоторые упомянули память, схему – они также были в числе ключевых форм. Однако были и несоответствия: например, один студент выделил слово «program», которое программа отнесла к общеупотребительной лексике и не включила в число ключевых словоформ. С другой стороны, программа обнаружила термин «integrated», который никто из студентов не упомянул, хотя в тексте ему был посвящен целый абзац. Студенты не включили слово в свой список, в то время как программа, следя за статистикой, посчитала его важным.

Таким образом, автоматический метод не уступает выводам обученных специалистов по точности, а в чем-то даже превосходит его – он не пропускает термины, которые равномерно распределены по всему тексту и встречаются чаще других.

Статистический метод прост в применении. Однако у него есть и слабые стороны.

Во-первых, в нем не проводится различия между омонимией (слова, которые совпадают по звучанию (произношению) или написанию, но различаются по значению) и многозначностью. Например, слово «память» может относиться к памяти человека или компьютера. В нашем тексте контекст был техническим, поэтому ошибки не возникло, но в общем случае это проблема.

Во-вторых, это зависит от стиля автора. Если автору нравится повторять одно и то же слово в каждом абзаце (например, «таким образом», «однако»), оно получит высокий балл, даже если не является существенным. Чтобы избежать этого, мы исключили из анализа так называемые стоп-слова (предлоги, союзы, местоимения, а также слишком общие слова, такие как «быть», «иметь»). Список стоп-слов необходимо тщательно подбирать для каждого языка и предметной области.

В-третьих, метод не учитывает синтаксические связи. Часто важны не отдельные слова, а фразы. Например, «искусственный интеллект» – это единый термин, а не два отдельных слова.

В-четвертых, не учитывается семантическая связь слов (синонимия). Если в тексте в качестве синонимов используются «компьютер», «вычислительная машина» и «автоматический компьютер», их следует сгруппировать вместе. Это требует использования тезаурусов (например, WordNet) и более сложных алгоритмов.

Для широкого круга практических задач, особенно в условиях обработки больших объемов однотипной документации, статистический метод остается вполне работоспособным инструментом. Он обеспечивает приемлемую точность, не требуя при этом сложной настройки или обучения, и без труда встраивается в действующие информационные системы.

Чтобы составить реферат на основе ключевых словоформ, проще всего выбрать предложения, содержащие много ключевых словоформ. Это называется экстрактивным обобщением. Этот подход был применен в программе. Полученный реферат сохранил связность, поскольку предложения следовали первоначальному порядку. Однако такой реферат может быть не совсем гладким – иногда в нем отсутствуют переходные фразы между предложениями.

Более продвинутым методом является абстрактное обобщение, при котором новые предложения генерируются на основе ключевых словоформ. Это сложная задача, требующая понимания грамматики и семантики. Здесь статистический подход должен быть дополнен лингвистическими правилами или моделями нейронных сетей. Однако для образовательных целей и для предварительной оценки содержания вполне подходит экстрактивный метод.

Список литературы

1) Батура Т.В. Математическая лингвистика и автоматическая обработка текстов. – Новосибирск: РИЦ НГУ, 2016. – 166 с.

2) Гребенщикова А.В. Основы квантитативной лингвистики и новых информационных технологий. – М.: ФЛИНТА, 2015. – 152 с.

3) Зубов А.В., Зубова И.И. Информационные технологии в лингвистике. – М.: Академия, 2004. – 208 с.

4) Рассел С., Норвиг П. Искусственный интеллект: современный подход. – М.: Вильямс, 2006. – 1408 с.

5) Щипицина Л.Ю. Информационные технологии в лингвистике. – М.: ФЛИНТА: Наука, 2013. – 128 с.