УДК 004.93:159.95

О. В. Бисикало, д. т. н., проф.; И. А. Назаров

ОБЗОР МЕТОДОВ АВТОМАТИЧЕСКОГО АННОТИРОВАНИЯ ТЕКСТОВ

В статье рассмотрены методы автоматического аннотирования текстов. На основании проведенного обзора предложено использование модели распространения ограничений для усовершенствования метода карт текстовых отношений (TRM).

***Ключевые слова**: автоматическое аннотирование, модель распространения ограничений, метод TRM.*

Введение

Аннотация – это краткое изложение смысла текста, в котором перечислены основные изложенные вопросы. Аннотации классифицируют по содержанию и целевому назначению, по полноте охвата смысла и читательскому предназначению. По первому признаку выделяют справочные и рекомендательные, по второму – общие и специализированные, как отдельный вид существуют обзорные аннотации [1].

Впервые понятие аннотации появляется во второй половине I в. н. э., но функционально аннотацию использовали еще в каталогах Александрийской библиотеки (III в. до н. э.). Постоянное накапливание и увеличение объемов текстовой информации в условиях развития информационных технологий обусловливает актуальность задачи автоматического аннотирования естественно-языковых текстовых материалов. Данная задача является одним из главных направлений компьютерной лингвистики и тесно связана с автоматическим реферированием. С учетом отличий в сущности понятий аннотации и реферата подобные задачи решают при помощи похожих методов.

Автоматическая обработка естественно-языковых текстов предусматривает трудности в процессе формализации поставленных задач. С другой стороны, на способ формализации влияют наличие разнообразных видов аннотации (фактического результата работы системы) и подход к ее построению. В общем случае автоматическое аннотирование предусматривает для данного текста T формирование другого текста A (аннотации), который содержит короткое изложение основных, изложенных в T, вопросов:

$$T \to A. \qquad (1)$$

Благодаря их дискретной природе, тексты удобно рассматривать как конечные множества $T = \{t_1, t_2, ..., t_n\}$ и $A = \{a_1, a_2, ..., a_m\}$. Элементами множества T могут быть разные лексические единицы (предложения, абзацы, параграфы и т. д.) в зависимости от его текстового размера, а элементами множества A – только предложения (в силу ограничения его текстового размера). Если рассматривать как элементы обоих указанных множеств предложения, то из последнего утверждения следует необходимость выполнения условия

$$\{A\} << \{T\}. \qquad (2)$$

Основную сложность в задачах автоматического аннотирования представляет собой обеспечение совпадения основного смысла текста T с аннотацией A и, собственно, поиск такого смысла.

Постановка задачи

Задача исследования рассмотреть основные алгоритмы генерирования и вытягивания для решения задачи автоматического аннотирования естественно-языковых текстов; исходя из

приоритетности семантического анализа, среди алгоритмов генерирования выбрать метод для модификации его с использованием подхода на основе модели распространения ограничений.

Методы автоматического аннотирования

На сегодняшний день существует целый ряд подходов к решению задачи автоматического аннотирования. Их принято делить на две группы: методы составления выдержек (вытягивающие алгоритмы) и формирования короткого изложения (генерирующие алгоритмы). Вытягивающие алгоритмы формируют аннотацию, используя текстовые фрагменты исходного документа. Для этого выделяют блоки наибольшей лексической и статистической важности. В этом случае аннотация представляет собой объединение выбранных фрагментов. Генерирующие алгоритмы анализируют исходный документ для поиска информации, на основе которой формируют текст аннотации. Очевидно, что первый из названных подходов является простым в реализации и не требует огромных вычислительных ресурсов, но не обеспечивает достаточного качества составления аннотации в силу отсутствия семантического анализа текста. Второй подход предусматривает ряд преимуществ: отсутствие дублирования информации в основном тексте и в аннотации, полноту аннотации, учет семантических связей в тексте. Поэтому в данной работе признается приоритетность генерирующих алгоритмов как таковых, которые имеют перспективы применения для создания систем автоматического аннотирования высокого уровня.

На рис. 1 приведена схема существующих методов автоматического аннотирования с учетом приведенного выше классификационного признака.

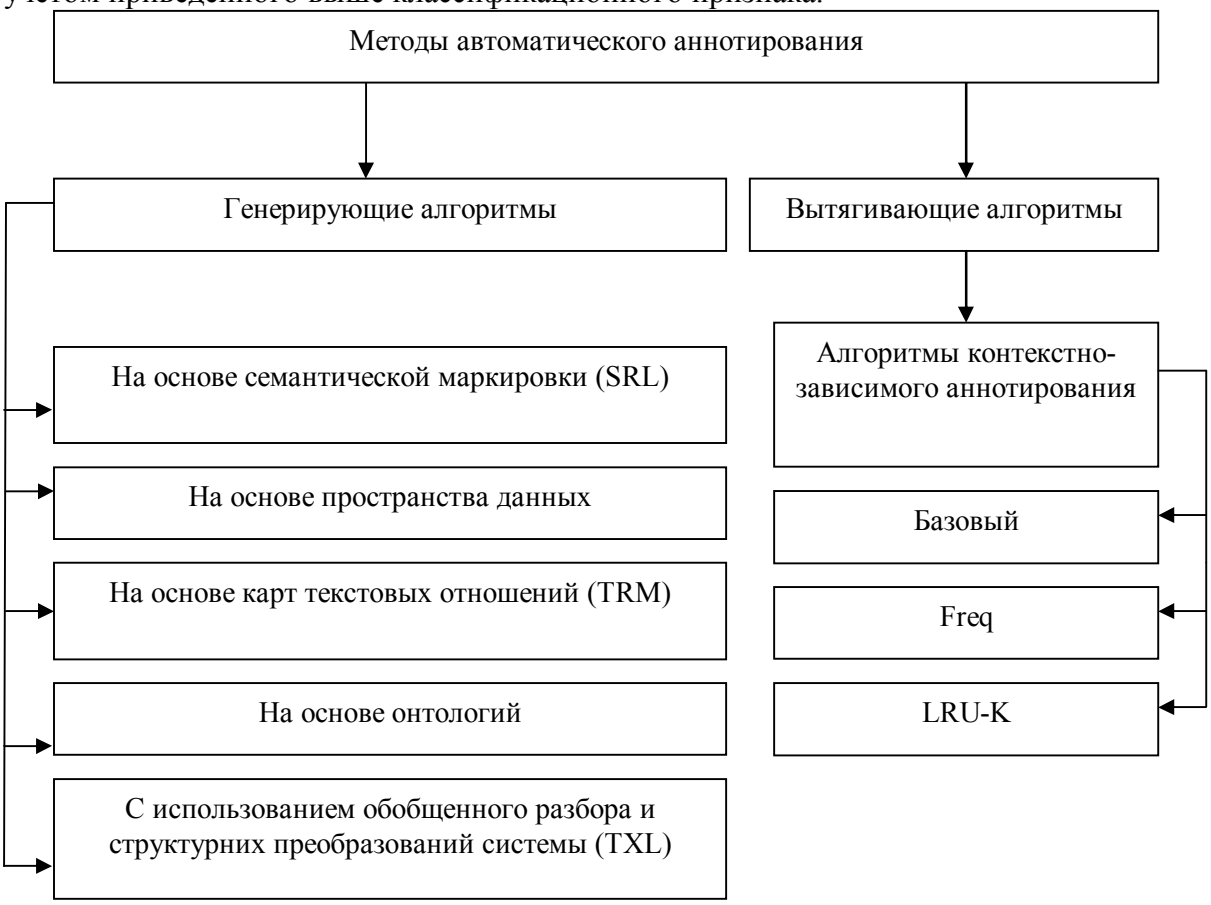

Рис. 1. Классификация методов автоматического аннотирования

Среди вытягивающих алгоритмов наиболее популярными являются алгоритмы

контекстнозависимого аннотирования HTML-документов. Аннотации, составленные подобными методами, используются поисковыми системами для описания результатов в виде коротких непрерывных фрагментов текста в соответствии с запросом пользователя. Выбор оптимального фрагмента текста осуществляют на основе расчета весов фрагментов.

Базовый алгоритм

Для расчета веса фрагмента в данном алгоритме используют формулу

$$W = \sum_{i=1}^{n} W_i + K\frac{n}{L}, \tag{3}$$

где W_i – вес i-го слова запроса, которое вошло в фрагмент; $K = const$; n – количество слов запроса, которые вошли в фрагмент; L – расстояние между первым и последним словами запроса.

Вес i-го слова W_i запроса вычисляют как

$$W_i = \frac{\log_2 N_i}{\log_2 N}, \tag{4}$$

где N_i – количество документов, в которых встретилось i-е слово; N – общее количество документов.

В список результатов поиска включают фрагмент текста с наибольшим весом. Если таких фрагментов более одного, алгоритм использует простое правило: в список включают наиболее близкий к началу текста фрагмент.

Проведенные эксперименты [2] свидетельствуют о том, что базовый алгоритм является наиболее эффективным по быстродействию, однако качество аннотирования (по оценкам экспертов) уступает другим алгоритмам.

Алгоритм Freq

Данный алгоритм являет собой усовершенствование предыдущего и, кроме количества слов поискового запроса, учитывает слова документа, которые наиболее часто повторяются. Вес фрагмента вычисляют по формуле

$$W = W_b + \sum_{i=1}^{n} \log_2 F_i, \tag{5}$$

где W_b – вес, вычисленный по базовому алгоритму; n – количество слов, которые наиболее часто повторяются; F_i – частота появления i-го слова.

Алгоритм Freq значительно уступает базовому по быстродействию, но обеспечивает более высокое качество аннотирования.

Алгоритм LRU-K

Данный алгоритм предложен в работе [2] и является вариантом алгоритма «последний, недавно использованный». Авторы применяют оценку локальной частоты появления слова при условии равномерного распределения слов. Экспериментальные исследования показали эффективность применения предложенного алгоритма для контекстнозависимых аннотаций: качество аннотирования несколько выше, нежели в алгоритме Freq при значительно большем быстродействии.

Алгоритм на основе семантической маркировки (SRL)

Основой алгоритма является блок анализа семантических структур аргумент-предикатов.

Суть алгоритма состоит в семантической маркировке связей в тексте. Для обеспечения связности предложений аннотации маркировку проверяют предикат-структурой. Аннотацию формируют трехуровневой обработкой:

1. Синтаксический анализ.
2. Построение дерева зависимостей.
3. Лексическая конструкция.

Основными преимуществами данного алгоритма является возможность формирования полной и завершенной аннотации, отсутствие повторов входного текста в аннотации. Определение связей между словами, их рода, падежа, числа позволяет замену, отбрасывание, сокращение слов. Недостатками алгоритма на основе семантической маркировки является сложность реализации, а также необходимость знания всех связей текста. Последний недостаток – отдельная сложная задача, без решения которой практическая реализация алгоритма невозможна.

Алгоритм на основе пространства данных

Автоматизированное аннотирование данных об определенном событии рассмотрено в работе [3]. Задачу генерации аннотации решают в два этапа:

1. Интеграция разрозненной информации и поиск информации о событии.
2. Аннотация события и вычисление коэффициентов подтверждения и опровержения информации.

Для решения первой из приведенных выше подзадач предлагаем подход на основе пространства данных. Пространство данных представляет собой структуру, которая состоит из данных (поданных в виде баз данных, хранилищ данных, статических веб-страниц), локальных хранилищ и индексов, средств поиска, обработки и интеграции информации.

При решении подзадачи вычисления коэффициентов подтверждения и опровержения информации используем построение адаптивной онтологии средств информации. Выведенные формулы позволяют количественно оценить данные коэффициенты. Найденные значения используют для построения аннотации, которая состоит из двух абзацев: первый из них подтверждает рассматриваемое событие, второй – опровергает. Соотношение между коэффициентами позволяет с некоторой вероятностью определить: произошло событие или нет.

Основным недостатком данного подхода следует признать отсутствие способа построения абзацев аннотирования. Эту задачу исследователи признают как достаточно сложную.

Алгоритм на основе карт текстовых отношений (TRM)

В основе алгоритма лежит использование карты текстовых отношений (Text Relationship Map - TRM) [4]. Идея состоит в формализации текста в виде графа

$$G = (P, V), \tag{6}$$

где $P = \{\overline{p_1}, \overline{p_2}, ..., \overline{p_n}\}$ – множество вершин графа; $E = \{e_1, e_2, ..., e_m\}$ – множество ребер между вершинами.

Каждая вершина такого графа представляет фрагмент входного текста и является взвешенным вектором, который включает весы отдельных слов фрагмента:

$$\overline{p_i} = (p_{i1}, p_{i2}, ..., p_{ik}). \tag{7}$$

Ребра соединяют вершины с большой мерой подобия, которая определяется как скалярное произведение векторов вершин:

$$m_{ij} = \overline{p_i}\,\overline{p_j}. \tag{8}$$

Наличие ребра между парой вершин свидетельствует о семантической близости данных фрагментов текста. Количество ребер, связанных с вершиной, определяет важность фрагмента текста, представленного этой вершиной. Построение аннотации позволяет определение наиболее важных фрагментов путем их сортировки по количеству связанных ребер.

Данный алгоритм гарантирует выполнение смыслового анализа текстов с целью их аннотирования. Кроме того, он может быть использован для поиска близких по смыслу документов, деления документов на группы по определённой тематике и т. д. Основную сложность в реализации алгоритма представляет построение карты текстовых отношений, которое предусматривает количественную оценку весов слов фрагментов текста и меры сходства фрагментов.

Алгоритм с использованием обобщенного парсинга и структурных преобразований системы TXL

Предложенный в работе [5] метод семантического аннотирования документов использует обобщенный парсинг и структурные преобразования системы TXL. TXL – это язык программирования, разработанный с целью поддержки анализа компьютерного программного обеспечения и задач преобразования документов. Процесс аннотирования в данном случае состоит из трёх этапов:

1. Представленный в TXL инструментарий парсинга используют для разбора входного текста, получают аппроксимированную (приблизительную) структуру фраз.
2. Перечисляют позитивные и негативные показатели семантических категорий для списка слов, получают первичную семантическую аннотацию документа.
3. Используют размеченный XML-текст для наполнения базы данных XML.

Данный алгоритм предназначен для полуавтоматического аннотирования естественноязыковых текстов. Полученную автоматически на втором этапе начальную аннотацию в ходе следующего этапа корректирует эксперт. Такое ограничение является существенным недостатком подобного подхода.

Выводы

В силу ограничения объема данной работы в ней изложены не все существующие на сегодняшний день подходы к решению задачи автоматического аннотирования естественноязыковых текстов. Существует большое количество методов полуавтоматического аннотирования, которые необходимо выделять в отдельную группу, поскольку они требуют участия эксперта в процессе составления аннотации. В ходе проведения исследования авторы отметили схожесть подходов к автоматическому аннотированию и реферированию, поэтому для составления аннотаций можно использовать модифицированные соответствующим образом методы реферирования.

Как следует из проведенного обзора, в генерирующих алгоритмах в той или иной степени используют семантический анализ входного текста. Основным недостатком большинства существующих подходов является несовершенство такого анализа и, как следствие, отсутствие заметных успехов в решении задачи. Поэтому целесообразным является использование метода определения смысла текстовой информации на основе модели распространения ограничений, предложенном в [6]. Эффективный семантический анализ может быть использован для улучшения рассмотренного выше алгоритма на основе карт текстовых отношений (TRM).

С целью адаптации модели текстовых отношений к предложенному подходу целесообразным является представление предложений в вершинах графа, в отличие от классического использования абзацев в качестве вершин. Это позволит усовершенствовать принятие решений в процессе аннотирования. В отличие от классического похода,

предлагаем возможность построения графа не для отдельного текста, а для коллекции документов. В таком случае граф строим следующим образом: каждое предложение текста представляем вершиной; ребра между вершинами определяют меру семантической связи предложений. Рациональным является представление текста в адаптированном виде, лишенном речевых единиц, которые не несут семантического значения. В качестве лексической меры сходства предложений возможно использование косинусной меры [7]. В качестве перспективного направления исследований необходимым считаем усовершенствование математического аппарата данного метода.

СПИСОК ЛИТЕРАТУРЫ

1. Ильичева Н. В. Аннотирование и реферирование / Н. В. Ильичева, А. В. Горелова, Н. Ю. Бочкарева. – Самара: Изд-во Самарского госуниверситета, 2003. – 100 с.

2. Губин М. В. Эффективный алгоритм формирования контекстно-зависимых аннотаций / М. В. Губин, А. И. Меркулов // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог-2005» (Звенигород, 1-6 июня 2005 г.). – 2005. – С. 116 – 120.

3. Про задачу автоматичного анотування події на основі простору даних [Електронний ресурс] / Шаховська Н. Б., Литвин В. В. // Науковий вісник Чернівецького національного університету ім. Юрія Федьковича. Збірник наук. праць. – Вип. 426: Фізика. Електроніка. – 2008. Режим доступу до журн.: http://www.nbuv.gov.ua/portal/natural/Nvchnu_ks/2008_426/426_09_Shakhovska.pdf.

4. Митрофанов М. С. Автоматическое аннотирование документов в многокомпонентной системе поиска и анализа естественно-языковой инфорации / М. С. Митрофанов, И. Е. Чижевский // Научная сессия МИФИ-2010. Ч. 1. XIV выставка-конференция. Телекоммуникации и новые информационные технологии в образовании. – С. 156 – 159.

5. Kiyavitskaya N. Text Mining through Semi Automatic Semantic Annotation / N. Kiyavitskaya, N. Zeni, L. Mich, J. Cordy, J. Mylopoulos // PAKM 2006. LNCS (LNAI). – vol. 4333. – 2006. – P. 143 – 154.

6. Квєтний Р. Н. Визначення сенсу текстової інформації на основі моделі розповсюдження обмежень / Р. Н. Квєтний, О. В. Бісікало, І. О. Назаров // Вимірювальна та обчислювальна техніка в технологічних процесах. – 2012. – № 1. – С. 93 – 96.

7. Salton G. Automatic text processing / G. Salton. – Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, – 1988. – 450 p.

Бисикало Олег Владимирович – д. т. н., профессор кафедры автоматики и информационно-измерительной техники.

Назаров Игорь Александрович – студент кафедры автоматики и информационно-измерительной техники.

Винницкий национальный технический университет.