УДК 004.912

АВТОМАТИЧЕСКОЕ СВОДНОЕ РЕФЕРИРОВАНИЕ НОВОСТНЫХ СООБЩЕНИЙ

С.Д. Тарасов

Приводится обзор современных методов и алгоритмов автоматического сводного реферирования, обосновываются основные недостатки этих методов. Формируются требования к методу, который смог бы преодолеть эти недостатки. Рассматривается разработанный автором метод тематического связанного ранжирования для задач автоматического сводного реферирования. В обоснование эффективности разработанного метода приводятся результаты экспериментальной оценки качества полученных сводных рефератов.

Ключевые слова: сводное реферирование, автоматическое сводное реферирование, сводный реферат, методы и алгоритмы автоматического сводного реферирования.

Введение

Одной из самых серьезных проблем современного общества является лавинообразное увеличение объема информации, которую должен воспринимать, хранить и использовать человек в процессе своей трудовой деятельности. Согласно последним исследованиям, до 2020 г. количество информации и потребности в ней будут расти экспоненциально. В таких условиях особую важность приобретают методы автоматической классификации и реферирования информации, позволяющие знакомить специалистов и других заинтересованных людей с необходимыми им документами, представленными в сжатом виде, но с сохранением смысла.

Классическое реферирование – процесс сжатия текстового документа и получение реферата, в котором сохраняется смысл оригинала. Наибольший интерес представляют обзорные или сводные рефераты, составляемые на некоторое множество документов, содержащие основные положения из этих документов [1]. Использование сводных рефератов вместо первоисточников позволяет эффективнее работать с большими объемами информации. Автоматическое сводное реферирование может быть использовано как эффективный инструмент подготовки аналитических справок и информационных бюллетеней для информационной поддержки лиц, принимающих управленческие решения, а также как средство сжатия текстов, технических описаний, стандартов и регламентов, состоящих из множества документов, информация в которых частично дублируется.

Обзор существующих методов сводного реферирования

Все существующие методы реферирования, как классического (по одному документу), так и сводного (обзорного по набору документов), можно разделить на три направления:

- квазиреферирование (Sentence extraction);
- генерация реферата с порождением нового текста (Abstraction);
- методы, объединяющие предыдущие два подхода.

Квазиреферирование основано на экстрагировании фрагментов документов – выделении наиболее информативных предложений (иногда – фраз и словосочетаний) и формировании из них квазирефератов. Методы генерации реферата с порождением нового текста основываются на выделении из текстов с помощью методов искусственного интеллекта и специальных информационных языков наиболее важной информации и порождении новых текстов, содержательно обобщающих первичные документы. В силу ограниченности на практике методов понимания и синтеза текста на естественном языке и отсутствия необходимой базы семантических словарей достаточного объема и содержания данные методы на сегодняшний день не получили значительного распространения. Большинство современных методов реферирования, имеющих практическую реализацию, относятся к направлению квазиреферирования.

Задача получения сводных рефератов, в которых были бы представлены все основные вопросы, затрагиваемые в каждом документе, но в обобщенном виде без повторений информации, – намного более сложная задача, чем традиционное автоматическое реферирование одного документа, даже очень большого объема. Во-первых, это связано с неизбежной разнородностью формулировок тем документов, на которые, как правило, ориентированы методы автоматического сводного реферирования. Во-вторых, для сводного реферирования отдельной задачей является метод упорядочивания предложений, отобранных для включения в сводный реферат. Предложения могут выбраться из разных документов и в общем случае, как правило, не составляют связный текст.

За рубежом в рамках конференций по проблемам автоматического аннотирования DUC (Document Understanding Conference) и текстового реферирования TSC (Text Summarization Challenge) данному направлению исследований придается очень большое значение. Автоматическое сводное реферирование реализовано в таких системах, как:

- «NewsBlaster» (http://www.newsblaster.com/),
- «Ultimate Research Assistant» (http://ultimate-research-assistant.com),
- «iResearch Reporter» (http://iresearch-reporter.com/),
- новостных порталах «Google News» (http://news.google.com/), «Яндекс. Новости» (http://news.yandex.ru/), «Рамблер. Новости» (http://news.rambler.ru) и др.

На сегодняшний день предложено большое количество различных методов получения сводных рефератов. В традиционных методах реферирования чаще всего используются различные модификации подхода Г. Луна [2], известного с конца 50-х годов XX века, который заключается в отборе предложений с наибольшим весом для включения их в реферат. Вес предложения определяется как сумма частот входящих в него значимых слов (с учетом закона Ципфа). Предложены методы, в которых вместо слов используются словосочетания, концепты тезауруса [3, 4]. К наиболее перспективным можно отнести методы, описывающие связную модель текста документов с помощью формального математического аппарата. Данные методы, как правило, не привязаны к особенностям конкретного языка, не требуют большого количества лингвистических ресурсов.

В результате анализа были сформулированы критические недостатки существующих подходов, которые необходимо исправить для достижения требуемого качества реферирования, а также для расширения сферы применения метода.

- Большинство существующих методов требуют большого количества различных лингвистических ресурсов (толковые, лексические и частотные словари, грамматики, тезаурус). Большая сложность естественных языков не позволяет создать достаточно полные формализованные лингвистические ресурсы для всех языков, необходимые для работы алгоритмов автоматического реферирования.

- Большинство существующих методов ориентировано на особенности конкретного естественного языка.
- В существующих методах либо вообще не рассматривается вопрос о формировании связного текста итогового реферата, либо ему уделяется недостаточное внимание.
- Большинство подходов требуют ручной корректировки со стороны экспертов-лингвистов.
- Существующие средства синтеза текста на естественном языке, используемые рядом методов сводного реферирования, находятся на ранней стадии своего развития и не позволяют использовать данные методы в целях, отличных от научно-исследовательских.
- Ряд алгоритмов требует значительных вычислительных ресурсов, что нежелательно при их использовании в реальных условиях обработки больших объемов данных.

Необходимость учета вышеперечисленных недостатков, а также исследование качества автоматического и ручного сводного реферирования определяют требования к новым эффективным методам и алгоритмам:

– минимальная потребность в лингвистических ресурсах (словарях, грамматиках и т.д.);

– отсутствие привязки к особенностям конкретного естественного языка;

– не только сжатие информации и выделение из текста наиболее значимых предложений, но и формирование из этих предложений связного текста;

– полностью автоматическое порождение текста реферата без необходимости последующей корректировки со стороны эксперта;

– алгоритм должен быть прост с вычислительной точки зрения, чтобы его можно было использовать в реальных задачах автоматического реферирования больших объемов данных в условиях ограниченного времени.

Метод тематического связанного ранжирования

Суть разработанного автором метода тематического связанного ранжирования заключается в отборе предложений из исходных документов, наиболее полно отражающие темы этих документов. Метод имеет следующие особенности.

1. Для предварительного ранжирования предложений документов относительно тем может быть использован любой алгоритм, например, алгоритм Луна. Автором был использован алгоритм Manifold Ranking [5].
2. Темы, которые плохо отражают суть документа, исключаются. По результатам анализа такие темы имеют очень слабую связь с текстом документа.
3. Для обеспечения связности полученного реферата каждое последующее предложение реферата связано с предыдущим некоторой общей темой.
4. Для обеспечения уникальности каждого предложения каждое последующее предложение отражает основную тему предыдущего предложения, а также некоторую новую тему, отличную от предыдущей.
5. Для разрешения анафорических связей предложения, содержащие анафорическую связь, игнорируются, если предыдущее предложение уже не содержится в реферате.

Рассмотрим метод более подробно. Для набора документов $D=\{D_i\}$, где $\mathbf{T_i}$ – тема документа D_i вычисляется матрица $\Xi = \left\{\xi_{i,j}\right\}$, где столбцы этой матрицы соответствуют векторам ранга соответствующих предложений относительно заданных тем. Например, для кластера из двух документов по два предложения, первое и третье из которых используются как темы,

$$\Xi = \begin{matrix} \xi_{1,1} & \xi_{1,2} & \xi_{1,3} & \xi_{1,4} \\ \xi_{2,1} & \xi_{2,2} & \xi_{2,3} & \xi_{2,4} \\ \xi_{3,1} & \xi_{3,2} & \xi_{1,3} & \xi_{3,4} \\ \xi_{4,1} & \xi_{4,2} & \xi_{4,3} & \xi_{4,4} \end{matrix}, \Xi = \begin{matrix} \mathbf{0.9} & \mathbf{0} & \mathbf{0.12} & \mathbf{0} \\ \mathbf{0.7} & \mathbf{0} & \mathbf{0.33} & \mathbf{0} \\ \mathbf{0.3} & \mathbf{0} & \mathbf{0.7} & \mathbf{0} \\ \mathbf{0.21} & \mathbf{0} & \mathbf{0.5} & \mathbf{0} \end{matrix},$$

где $\xi_{\mathbf{j}} = \left\{\xi_{1,\mathbf{j}}, \xi_{2,\mathbf{j}}, \xi_{3,\mathbf{j}}, \xi_{4,\mathbf{j}}\right\}^{\mathrm{T}}$ – вектор ранжирования предложений кластера относительно предложения j (темы $\mathbf{T}_j$). Если для ранжирования используется метод Manifold Ranking, то ξ_j вычисляется итеративно:

$$\overline{\xi_j}(\mathbf{t}+\mathbf{1}) = \alpha \cdot S \cdot \overline{\xi_j}(\mathbf{t}) + (1-\alpha) \cdot \overline{\mathbf{y}_j},$$

где вектор $\overline{\mathbf{y}}_j = \left[\mathbf{y}_j^{\mathbf{0}}\mathbf{y}_j^{\mathbf{1}}, .. \mathbf{y}_j^n\right]^{\mathrm{T}}$, $y_j^j = 1$, и $y_j^i = 0, i \in (1,n), i \neq j$ для всех остальных предложений; α – коэффициент передачи ранга от источника, **S** – нормализованная матрица связей между предложениями. Традиционно в качестве матрицы связей использовалась матрица

$$\mathbf{W}_{i,j} = \mathbf{Sim}(\overline{\mathbf{x}_i}, \overline{\mathbf{x}_j}),$$

где $\mathbf{Sim}(\overline{\mathbf{x}_i}, \overline{\mathbf{x}_j}) = \dfrac{\overline{\mathbf{x}_i} \cdot \overline{\mathbf{x}_j}}{\|\mathbf{x}_i\| \cdot \|\mathbf{x}_j\|}$, $\mathbf{x}_i = [tf_0, tf_1, .. tf_n]^{\mathrm{T}}$, tf_k – стандартная TF-IDF мера относительной важности терма t_k. В [5] предложена модификация

$$\mathbf{W} = \lambda_1 \cdot \mathbf{W}_{\mathbf{inner}} + \lambda_2 \cdot \mathbf{W}_{\mathbf{intra}}$$

для учета различных весов связей предложений одного документа и разных документов и

$$S = D^{-1/2} \cdot W \cdot D^{-1/2}$$

для симметричной нормализации полученной матрицы. Автором была предложена и реализована следующая модификация матрицы **W**:

$$\mathbf{W} = \lambda_1 \cdot \mathbf{W}_{\mathbf{inner}} + \lambda_2 \cdot \mathbf{W}_{\mathbf{intra}} + \zeta \cdot \mathbf{W}_{\mathbf{path}},$$

где $\mathbf{W}_{\mathbf{path}}$ – матрица весов удаленности предложений друг от друга в тексте.

Далее матрица $\boldsymbol{\Xi}$ подвергается симметричной нормализации:

$$\tilde{\boldsymbol{\Xi}} = \mathbf{Z}^{-1/2} \cdot \boldsymbol{\Xi} \cdot \mathbf{Z}^{-1/2},$$

где **Z** – диагональная матрица, каждый элемент которой равен сумме элементов соответствующей строки исходной матрицы $\boldsymbol{\Xi}$. В результате этого строки матрицы $\boldsymbol{\Xi}$ содержат коэффициенты соответствия предложений кластера заданным темам документов:

$$\mathbf{T}(\mathbf{x}_i) = \sum_{j=1}^{n} \tilde{\boldsymbol{\Xi}}_{i,j} \cdot \mathbf{T}_j .$$

Например, для вышеприведенного кластера $\mathbf{T}(\mathbf{x_1}) = \mathbf{0,9} \cdot \mathbf{T_1} + \mathbf{0,1} \cdot \mathbf{T_3}$. Формально это означает, что предложение x_1 отражает 0,9 темы $\mathbf{T}_1$ (собственной темы) и 0,1 темы предложения x_3 $(\mathbf{T_3})$.

Алгоритм формирования связного текста итогового реферата

Для формирования связного текста итогового реферата используется следующий алгоритм.

1. На главной диагонали матрицы $\boldsymbol{\Xi}$ выделяется элемент, имеющий наименьшее значение. Это соответствует теме документа, наиболее сильно связанной с другими предложениями кластера. Эта тема используется как текущая основная тема $\mathbf{T}_{current}$ и выносится в заголовок итогового реферата.
2. Главная диагональ матрицы обнуляется.
3. В current-столбце матрицы $\boldsymbol{\Xi}$ определяется элемент $\xi_{i,current}$ с наибольшим значением. Это соответствует нахождению предложения x_i, наиболее близкого теме $\mathbf{T}_{current}$.
4. Предложение x_i помещается в итоговый реферат.
5. Для уменьшения ранга предложений, которые похожи на x_i, а также тем, которые уже нашли отражение в итоговом реферате, выполняется следующая процедура:

$$\xi_{i,j} = \xi_{i,j} - \boldsymbol{\omega} \cdot \overline{\mathbf{S}}_{i,current} \cdot \xi_{current}^{*},$$

где ω – коэффициент усечения похожих предложений (новизны), а $\xi_{current}^{*}$ – первоначальное значение вектора-столбца $\xi_{current}$.

6. В *i*-ой строке матрицы выполняется поиск элемента ξ_{next} с наибольшим значением. Исходя из

$$\mathbf{T}(\mathbf{x}_i) = \sum_{j=1}^{n} \tilde{\boldsymbol{\Xi}}_{i,j} \cdot \mathbf{T}_j ,$$

выполняется поиск темы $\mathbf{T}_{next}$, отраженной в предложении x_i и следующей по значимости после $\mathbf{T}_{\text{current}}$.

7. Процесс переходит на шаг 3 ($\mathbf{T}_{current} = \mathbf{T}_{next}$), пока объем итогового реферата не достигнет требуемой величины.

Предложения, содержащие анафорические связи, игнорируются в том случае, если предыдущее предложение документа не было включено в реферат на предыдущем шаге.

Реализация

Для научно-исследовательских целей автором был разработан программный комплекс для ручного и автоматического сводного реферирования на базе предложенного алгоритма тематического связанного ранжирования и оценки качества полученных сводных рефератов. Кроме данного алгоритма, в системе были реализованы такие алгоритмы, как BasicLine [6], Manifold Ranking [7], модифицированный алгоритм Manifold Ranking [7]. Созданный автором программный комплекс «MDS Evaluation» позволяет решать следующие задачи:

- автоматическое сводное реферирование в широком диапазоне различных параметров для различных нужд;
- ручное сводное реферирование в многопользовательском режиме;
- ручная и автоматическая оценка качества сводного реферирования;
- сравнение эффективности различных алгоритмов и методов.

Оценка

Традиционные методы оценки качества сводного реферирования включают в себя оценку сводного реферата специалистами-лингвистами по ряду критериев. К таким критериям относятся связность полученного текста, краткость (лаконичность), грамматическая правильность, сложность восприятия, содержание.

Однако даже простая ручная оценка качества сводного реферирования по нескольким критериям требует больших объемов человеческих ресурсов (согласно DUC, более 3000 часов работы лингвистов), что является очень дорогим. Одной из наиболее удачных реализаций систем для автоматической оценки качества сводного реферирования можно считать пакет ROUGE [8], используемый в DUC. Набор программ позволяет автоматически рассчитывать различные метрики ROUGE (Recall-Oriented Understudy for Gisting Evaluation): ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S, ROUGE-SU. Автором были реализованы алгоритмы оценки ручных и автоматических рефератов с помощью данных автоматических метрик для русского языка.

Для экспериментальной оценки качества работы предложенного автором метода реферирования была проведена ручная и автоматическая оценки получаемых различными методами рефератов, а также было выполнено построение ручных рефератов. В эксперименте приняло участие 13 человек (преподаватели и студенты 5 курса БГТУ «Военмех»). Эксперимент состоял из двух частей – построение ручных рефератов и их оценка. Исходными данными послужили 12 отобранных вручную новостных кластеров различной тематики («Россия», «Происшествия», «Наука и техника», «Спорт», «Культура» и др.) из системы «Google. News» за 2009 год. В рамках построения ручных рефератов участниками эксперимента было построено 156 ручных рефератов: каждый участник составил свой ручной сводный реферат для каждого кластера. В рамках оценки рефератов каждый участник оценил каждый сводный реферат (всего было получено 156 ручных и 2600 автоматических, порожденных различными методами с различными параметрами) по набору формальных критериев. Кроме того, была проведена автоматическая оценка всех сводных рефератов по метрикам ROUGE, для чего были использованы рефераты, построенные вручную. Результаты оценки приведены в таблице. Для вычислений использованы следующие обозначения и параметры:

- BL – усредненное значение для BasicLine (BL-1 – BL-7 – заведомо плохие рефераты [6]);
- МТСР – метод тематического связанного ранжирования, предложенный автором. Параметры: α=0,9, $\lambda1/\lambda2$=0,3, ω=50, ζ=0,1;
- Manifold – Метод Manifold Ranking [7]. Параметры: α=0,8, $\lambda1/\lambda2$=3, ω=50;
- Модифицированный Manifold – модифицированный метод Manifold Ranking [7]. Параметры: α=0,8, $\lambda1/\lambda2$=0,3, ω=50.

	BL	Manifold	Модифицированный Manifold	МТСР	Ручные
Связность	0,42	0,69	0,73	**0,81**	0,88
Содержание	0,45	0,78	0,78	**0,83**	0,84
Полнота	0,45	0,78	0,80	**0,82**	0,84
Общее впечатление	0,41	0,71	0,78	**0,85**	0,86
ROUGE-1	0,26	0,39	0,40	**0,41**	0,38
ROUGE-2	0,11	0,18	0,18	**0,19**	0,17
ROUGE-3	0,07	0,12	0,12	**0,12**	0,12
ROUGE-L	0,22	0,33	0,34	**0,36**	0,33

Таблица. Результаты оценки

Заключение

Ручное реферирование, несомненно, имеет ряд преимуществ перед автоматическим, однако, помимо крайне высокой стоимости построения рефератов, имеет и ряд недостатков. К ним относятся невозможность оперативного составления рефератов для очень большого количества исходных документов или документов большого объема; невозможность оперативного составления различных рефератов с заданными свойствами (например, объем реферата); элементы субъективности, так или иначе присутст-

вующие в конечном реферате (каждый эксперт выделяет те или иные значимые элементы и т.д.), и ряд других. Предложенный автором метод открывает возможность для построения сводных рефератов, представляющих связный текст в автоматическом режиме без использования сложных и труднодоступных лингвистических ресурсов и больших вычислительных мощностей. При этом эффективность метода и качество полученных сводных рефератов подтверждаются экспериментально.

Литература

1. ГОСТ 7.9–95. Система стандартов по информации, библиотечному и издательскому делу. Реферат и аннотация. Общие требования. – 2001 [Электронный ресурс]. – Режим доступа: http://www.standards.ru/document/4155011.aspx (дата обращения: 01.02.2010).
2. Luhn H.P. The Automatic Creation of Literature Abstracts // IBM Journal. – 1958, April. – P. 159–165.
3. Лукашевич Н.В., Добров Б.В. Автоматическое аннотирование новостных кластеров на основе тематического представления // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог». Периодическое издание. – 2009. – Выпуск 8 (15).
4. Абрамова Н.Н., Абрамов В.Е. Автоматическое составление обзорных рефератов новостных сюжетов. // Труды 9-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2007. – Переславль-Залесский, Россия, 2007.
5. Xiaojun Wan, Jianwu Yang, Jianguo Xiao. Manifold-Ranking Based Topic-Focused Multi-Document Summarization // DUC 2003 [Электронный ресурс]. – Режим доступа: http://www.ijcai.org/papers07/Papers/IJCAI07-467.pdf, своб.
6. Тарасов С.Д. Исследование и оптимизация параметров алгоритма Manifold Ranking на основе метрики автоматической оценки качества обзорного реферирования ROUGE-RUS // Труды XI Всероссийской научной конференции «Электронные библиотеки. Перспективные методы и технологии, электронные коллекции». – Петрозаводск, 2009. – С. 86–93.
7. Тарасов С.Д. Автоматическое составление обзорных рефератов новостных сюжетов // Вестник Балтийского государственного технического университета. – 2008. – № 3. – С. 61–67.
8. Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. Information Sciences Institute // University of Southern California. 2004.

Тарасов Сергей Дмитриевич – Балтийский государственный технический университет «Военмех» им. Д.Ф. Устинова, ассистент, tarasov_sd@mail.ru