Исследование и оптимизация параметров алгоритма Manifold Ranking на основе метрики автоматической оценки качества обзорного реферирования ROUGE-RUS

© С.Д. Тарасов

Балтийский Государственный Технический Университет им. Д.Ф.Устинова «ВОЕНМЕХ»

Аннотация

В статье рассматривается используемая в DUC метрика *ROUGE*, а также выполненная авторами ее модификация для русского языка *ROUGE-RUS*. Разработана система для автоматической оценки качества обзорного реферирования на основе *ROUGE-RUS* с Web-интерфейсом. Проведен эксперимент по составлению ручных аннотаций новостных кластеров и автоматической оценке качества обзорного реферирования по метрике *ROUGE-RUS*. Введено понятие базовой величины *ROUGE-RUS* на кластере. Рассмотрен алгоритм обзорного реферирования *Manifold Ranking*. Проведен анализ степени влияния выбора базовых параметров и начальной темы на результат работы алгоритма. На основе результата исследований влияния выбора темы на работу алгоритма разработан модифицированный алгоритм.

Введение

Задача автоматического построения обзорных рефератов на сегодняшний день является очень актуальной. Это вызвано, в первую очередь, необходимостью в условиях постоянного роста информации знакомить специалистов и других заинтересованных людей с необходимыми им документами, представленными в сжатом виде, но с сохранением смысла. В обзорной статье [1] описывается современное состояние в области автоматического реферирования, а также основные направления и пути развития.

На текущий момент существует огромное количество различных методов получения обзорных рефератов. В традиционных методах реферирования чаще всего используются различные модификации подхода Г. Луна [3], 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. В работе [7] описан метод, в котором в качестве значимых элементов выбираются не слова, а словосочетания. В работе [8] представлены методы обзорного реферирования с использованием концептов тезауруса. 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. К таким методам относятся метод регрессии опорных векторов, метод ранжирования связных структур [5, 9], а также ряд других.

Не менее актуальной является и задача оценки полученных автоматических рефератов. Несомненно, наиболее правдоподобные оценки качества можно проводить в ручном режиме с привлечением большого числа экспертов. Однако такие ручные оценки являются чрезвычайно дорогими. Методики автоматической оценки качества реферирования не только делают этот процесс более доступным, но и позволяют в реальном времени производить настройку параметров работы определенного алгоритма, производить их оптимизацию.

1 Оценка качества обзорного реферирования

На сегодняшний день предложено огромное количество методов обзорного реферирования. Работа каждого метода определяется некоторым набором внешних условий; кроме того, каждый метод содержит набор параметров, подбор которых изменяет качество реферирования при заданных условиях в широком диапазоне. Эти параметры, как правило, определяются для заданных внешних условий эмпирическим путем. В связи с этим, одним из немаловажных вопросов является оценка качества обзорного реферирования. Это позволяет в реальном времени производить настройку параметров работы конкретного алгоритма, производить оптимизацию этих параметров, сравнивать эффективность разных алгоритмов, а также делать окончательный вывод о возможности

Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.

практического применения данного алгоритма автоматического реферирования.

Традиционные методы оценки качества обзорного реферирования включают в себя оценку обзорного реферата по ряду критериев специалистами-лингвистами. К основным критериям относятся связность, краткость (лаконичность), грамматическая правильность, сложность восприятия, содержание. Однако даже простая ручная оценка качества обзорного реферирования по нескольким критериям требует больших объемов человеческих ресурсов (согласно DUC, более 3000 часов работы лингвистов), что является очень дорогим. Кроме того, нет возможности проводить оценку качества в «реальном времени», например, при оптимизации работы некоторого метода реферирования. В связи с этим, вопрос о возможности автоматизации оценки качества является очень актуальным. За последние несколько лет было предложено несколько методик автоматической оценки качества обзорного реферирования. Все они основаны на автоматическом сравнении реферата, полученного с помощью метода автоматического реферирования, с одним или несколькими обзорными рефератами, составленными экспертами. В этом случае, так или иначе, критерием качества можно считать «схожесть» автоматического реферата с ручным. В качестве меры «сходства» в DUC были предложены «cosine similarity», «unit overlap(unigram or bigram)», «longest common subsequence».

2 ROUGE: метрика для автоматической оценки качества обзорного реферирования

Одной из наиболее удачных реализаций систем для автоматической оценки качества обзорного реферирования можно считать пакет *ROUGE* [6], 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*. К наиболее часто используемым в DUC относятся *ROUGE-1* и *ROUGE-2*.

2.1 ROUGE-N

Метрика *ROUGE-N* представляет собой обобщенную статистическую меру, выражающую какой процент лексических единиц (*N-gram*,-последовательностей из *N* лексем), входящих в состав ручного, построенного независимым экспертом, реферата, повторяется в автоматическом реферате:

$$ROUGE-N=\frac{\sum_{S\in RefSum}\sum_{n-gram\in S}Count_{match}(n-gramm)}{\sum_{S\in RefSum}\sum_{n-gram\in S}Count(n-gramm)}.$$

В случае использования нескольких ручных рефератов для оценки автоматического, в [2] предлагается сравнивать автоматический реферат с каждым ручным по метрике *ROUGE-N*, а затем выбирать максимальное значение.

$$ROUGE-N_{multi}=\arg\max_i \mathrm{ROUGE\text{-}N}(r_i, s),$$

где r_i – *i*-й ручной реферат, *s* – оцениваемый автоматический реферат. Такая же процедура используется и для *ROUGE-L*.

Важно отметить, что *ROUGE* является метрикой полноты, и соответственно несимметричной относительно сочетаний ручной - автоматический реферат.

2.2 ROUGE-L

Для оценки степени совпадения автоматического реферата с ручным, также используется метод «наибольшей совпадающей подпоследовательности»[2]. Величина *LCS* (*Longest Common Subsequence*) представляет собой длину наибольшей подпоследовательности между двумя предложениями *X* и *Y*. В качестве элементов последовательностей выбираются лексемы. При вычислении величины *ROUGE-L* для автоматического реферата, содержащего *v* предложений (всего *n* слов) и ручного реферата, содержащего *u* предложений (всего *m* слов), производится вычисление объединенной *LCS* между каждым предложением ручного реферата r_i и всеми предложениями автоматического c_j.

$$R_{lcs}=\frac{\sum_{i=1}^{u}LCS_{\cup}(r_i,C)}{m},\ P_{lcs}=\frac{\sum_{i=1}^{u}LCS_{\cup}(r_i,C)}{n}$$

$$ROUGE-L=F_{lcs}=\frac{(1+\beta^2)R_{lcs}\cdot P_{lcs}}{R_{lcs}+\beta^2\cdot P_{lcs}},$$

где $LCS_{\cup}(r_i,C)$ - длина наибольшей подпоследовательности между предложением ручного реферата r_i и всеми предложениями автоматического реферата *C*[2]. В DUC полагается $\beta\to\infty$, таким образом, учитывается только R_{lcs}-составляющая.

В [2] на основе кластеров из DUC 2001, 2002, 2003, 2004 метрики *ROUGE* показали высокую степень корреляции (Коэффициент Пирсона до 0.99) с ручными оценками. Это позволяет использовать метрики *ROUGE* для автоматической оценки качества обзорного реферирования, для сравнительной оценки различных методов, а также для оптимизации работы отдельно взятого метода.

Пакет программ, реализующий автоматическую оценку *ROUGE*, является свободно распространяемым набором *Perl*-скриптов, однако ориентирован только на использование кластеров в формате DUC, а также на документы на английском языке. Это создает определенные трудности при попытке оценки по метрике *ROUGE* русскоязычных обзорных рефератов. Кроме того, «отличные» результаты, полученные в DUC для англоязычных

кластеров, вовсе не свидетельствуют о таких же результатах для кластеров на русском языке.

3 ROUGE-RUS

С учетом недостатков существующего пакета *ROUGE* была разработана модифицированная метрика *ROUGE-RUS*, обладающая следующими отличительными особенностями:

- Русская морфология, список стоп-слов для русского языка;
- Возможность учитывать синонимы (с использованием концептов тезауруса);
- Усреднение (а не максимум) значения при наличии нескольких ручных аннотаций. По полученным результатам это позволяет сделать метрику более устойчивой и использовать меньшее количество ручных рефератов.

Для расчета метрики *ROUGE-RUS* была разработана система с Web-интерфейсом, являющаяся частью системы "MDS Evaluation Framework"[9].

4 Эксперимент по оценке метрики ROUGE-RUS

Для изучения свойств метрики *ROUGE-RUS*, а также для исследования возможности использования ее для автоматической оценки качества обзорного реферирования был проведен следующий эксперимент.

4.1 Исходные данные

В качестве исходных данных были взяты новостные кластеры различной тематики («Россия», «Происшествия», «Спорт», «Культура» и др.) из системы «Google.News» за конец ноября – начало декабря 2008 года. Всего было обработано 67 кластеров из 613 документов, полученных из 21 источника («РБК», «РИА Новости», «Российская Газета» и т.д.).

4.2 Построение ручных рефератов

К построению ручных рефератов были привлечены студенты 5 курса БГТУ «ВОЕНМЕХ», обучающиеся по специальности «Автоматизированные и управляющие системы». Всего в исследовании приняло участие 67 человек. Каждому участнику было предложено составить одну аннотацию для каждого из случайно выбранных 50 кластеров. При составлении ручного реферата для каждого кластера участник должен был выбрать 4 различных предложения из всех документов кластера. В результате было получено 2385 ручных аннотаций.

4.3 Исследование метрики ROUGE-RUS на наборе ручных рефератов

Далее было отобрано $N=50$ кластеров, для которых имелось по $M>40$ ручных аннотаций, порожденных разными пользователями.

$$A_i \in A, 1 \le i \le N,$$

$$A_i^j \in A_i, 1 \le j \le M.$$

Для исследования метрики *ROUGE-RUS*, были произведены вычисления величин *ROUGE-1*, *ROUGE-2*, *ROUGE-3*, *ROUGE-4* и *ROUGE-L* для каждой пары аннотаций из множества A_i по всем кластерам:

$$RR_i^{l,m} = ROUGE - RUS(A_i^l, A_i^m), 1 \le l, m \le M, 1 \le i \le N,$$

где A_i^l выступает в роли ручной, а A_i^m в роли автоматической аннотации, оцениваемой по метрике. Цель такого исследования – выявить особенности метрики, оценить распределение ее величины на множестве «заведомо хороших» ручных аннотаций, определить минимальное необходимое количество ручных аннотаций для стабильной оценки одной автоматической. Для произвольно взятой аннотации $A_i^j \in A_i$ распределение величины *ROUGE-RUS*(A_i^l , A_i^j) имеет следующий вид (см. Рис. 1). Для всех кластеров распределение имеет примерно такой же вид. Из этого следует, что:

[1] Ручные рефераты, порожденные разными пользователями, слабо согласуются друг с другом.

[2] Использование одного ручного реферата для оценки недостаточно.

[3] В ручных рефератах, порожденных разными пользователями, практически отсутствует кластеризация. Если таковая и имеет место быть, то, как правило, не в области максимума, а где-то «посередине».

[4] Использование морфологии, списка стоп-слов и словаря синонимов положительно сказывается на пологости кривой, что обеспечивает меньший разброс величины для разных ручных рефератов в пределах одного кластера.

Отсутствие кластеризации, особенно в области максимального значения, говорит о том, что использование метода максимума, предложенного в [6], является неоправданным. В этом случае во внимание, фактически, принимается только одна ручная аннотация (наиболее близкая к автоматической), что при таком разбросе значения величины не является допустимым. Исходя из этих соображений, способ усреднения должен рассматриваться как основной для учета нескольких ручных рефератов при оценке одного автоматического.

Далее были сформированы выборки величины *ROUGE-RUS* для $K=1..10$, где K – количество ручных аннотаций, принимаемых в расчет при оценке одной автоматической. В этом случае из

множества $A_i^j \in A_i$ выбиралась *1* аннотация, как автоматическая, подлежащая оценке, и *K* из оставшихся, как ручные. Были использованы следующие вычисления:

$$RR_i^{l,m} = ROUGE - *(A_i^l, A_i^m), 1 \le l, m \le 40, 1 \le i \le 50,$$

а затем вычислялся максимум и среднее для данного значения *K*. После этого были сформированы выборки из величин *ROUGE-RUS* для каждого *K*, и было произведено усреднение по всему множеству кластеров. Таким образом, были оценены такие параметры, как: зависимость среднего, минимального и максимального значения, дисперсии, ср. кв. откл. величины *ROUGE-RUS* от *K*. Результаты представлены в Таблице 1. Следует отметить, что дисперсия величины не рассматривалась как критерий качества самой метрики, однако ее минимализация необходима для того, чтобы метрика была более стабильной.

Таким образом, можно считать, что использование усреднения с учетом слабого согласия ручных аннотаций друг с другом, дает более стабильный результат. В этом случае для оценки одной автоматической аннотации достаточно 4-5 ручных.

5 Исследование и оптимизация параметров алгоритма обзорного реферирования на основе метрики ROUGE-RUS

Метрики автоматической оценки, к которым относится *ROUGE*, могут быть использованы, в первую очередь, для исследования влияния различных параметров на качество аннотаций, порождаемых алгоритмами обзорного реферирования, а также для оптимизации этих параметров. Авторами была произведена подробная оценка работы алгоритма *Manifold Ranking* для русского языка [9], а также сравнение результатов работы с *«Basic Lines»* по метрике *ROUGE-RUS* и исследование влияния различных параметров на работу алгоритма.

5.1 Алгоритм ранжирования связных структур для задачи обзорного реферирования

Алгоритм *Manifold Ranking*[5] позволяет описать связную структуру текста при помощи матриц. Изначально алгоритм предполагает выделение элементов (предложений) наиболее близких заданному (теме). Такая интерпретация характерна задаче информационного поиска. Для автоматического реферирования также выделяется набор предложений, наиболее близких заданной теме кластера, однако обязательным является применение алгоритма отсечения «похожих» предложений, что особенно актуально для многодокументного аннотирования.

Автоматическое реферирование набора документов с использованием алгоритма ранжирования связных структур состоит из двух этапов:

[1] Вычисление ранга каждого предложения. Этим решается задача ранжирования всех предложений в соответствии с их «близостью» заданной теме кластера.

[2] Применение алгоритма отсечения предложений, наиболее похожих на те, что уже попали в обзорный реферат. Этим решается задача исключения из обзорного реферата одинаковых или близких предложений.

Основной особенностью алгоритма ранжирования связных структур является учет внутренней связной структуры объектов, составляющих текст. Объекты должны быть представлены векторами в Евклидовом пространстве. В этом случае полагается, что «близость» двух объектов представленных векторами может быть вычислена, как Евклидова мера или скалярное произведение векторов. Целью алгоритма является упорядочить объекты, с учетом их внутренних связей между собой.

Формально, связная структура объектов представляется как некий взвешенный граф, вершинами которого являются сами объекты, а в качестве весов дуг задаются евклидовы расстояния между ними. Алгоритм ранжирования заключается в постепенном распространении объектами своего ранга на смежные объекты-вершины. Таким образом, ранг f_i каждого предложения x_i вычисляется не только с учетом «близости» его к эталонному объекту (теме кластера *T*), но и с учетом связной структуры текста, т.е. ранг «распространяется» по графу с учетом весов связей структур. В результате некоторое количество предложений с наибольшим рангом выбирается для результирующего реферата.

5.2 Сравнение результатов MR с результатами «ручная-ручная». Базовая величина метрики ROUGE-RUS на кластере

Для оценки результатов работы алгоритма по метрике *ROUGE-RUS* было введено понятие «Базовой величины метрики *ROUGE-RUS* на кластере». В качестве этой величины было использовано усредненное и максимальное значение *ROUGE-RU*S для данного кластера. Для оценки каждой автоматической аннотации, порождаемой алгоритмом было выбрано 10 ручных, построенных разными пользователями. Таким образом, результат работы алгоритма для заданного кластера оценивался относительно этих двух базовых величин (среднее и максимум).

При значении параметров из [5, 9] базовый алгоритм *Manifold Ranking* показал, в среднем, результаты, несколько худшие, чем базовые оценки ручных с ручными.

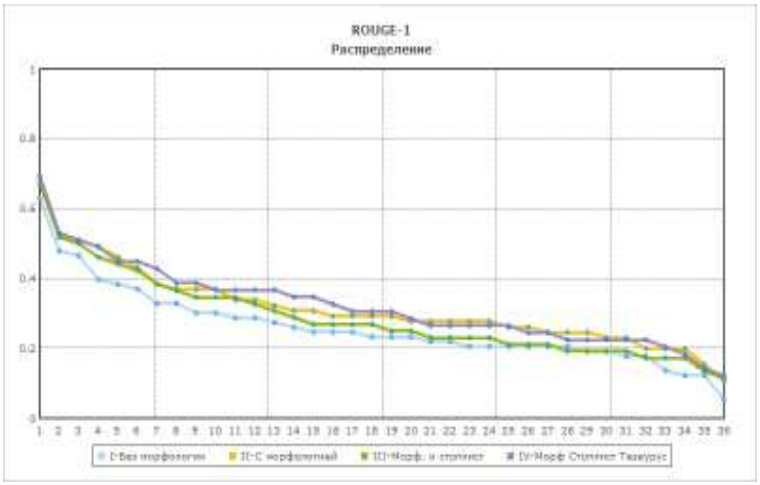

Рисунок 1 - Распределение величины *ROUGE-1* при сравнении одной ручной аннотации со всеми остальными для произвольного кластера. Значения отсортированы по убыванию.

Таблица 1 - Относительный разброс величин *ROUGE-RUS* при различных значениях *K*

K	δ, Метод Максимума, %					δ Метод усреднения, %				
	ROUGE-1	***ROUGE-2***	***ROUGE-3***	***ROUGE-4***	***ROUGE-L***	***ROUGE-1***	***ROUGE-2***	***ROUGE-3***	***ROUGE-4***	***ROUGE-L***
1	42.18	79.31	108.69	127.03	45.90	42.18	79.31	108.69	127.03	45.90
5	28.52	41.47	49.00	56.47	30.25	25.10	42.57	57.64	67.85	26.59
10	24.58	33.24	37.91	43.38	26.19	22.05	35.39	47.55	56.23	23.07

5.3 Сравнение результатов MR с «Basic Lines»

Также были проведены оценки с псевдо-автоматическими аннотациями *«Basic Lines»*:

[1] **BL1** – 4 первых предложения 1-го документа,
[2] **BL2** – первые предложения 4-х первых документов,
[3] **BL3** – последние предложения 4-х первых документов,
[4] **BL4** – заголовки 4-х первых документов,
[5] **BL5** – 4 первых предложения последнего документата,
[6] **BL6** – последние предложения 4-х первых документов,
[7] **BL7** – последние предложения 4-х последних документов,
[8] **BL8** – заголовки 4-х последних документов.

5.4 Оптимизация общих параметров

К общим параметрам алгоритма можно отнести: α (определяет относительный вклад близких предложений в ранк текущего и начальный ранк каждого предложения), λ_1 (коэффициент учета веса связности предложений из одного документа), λ_2 (коэффициент учета веса связности предложений из разных документов), ω (коэффициент усечения сходных предложений). В среднем, по кластерам, были получены следующие значения оптимальных параметров: α=0.9, λ_1=0.3, λ_2=0.8, ω=10. Полученные значения несколько отличаются от использованных в DUC [5]. Это связано в первую очередь со спецификой взятых новостных кластеров, а также с особенностями русского языка. Кроме того, нет сведений о том, проводили ли авторы алгоритма подбор этих параметров или значения были взяты исходя из эмпирических соображений.

На имеющихся в наличии кластерах была выявлена сильная устойчивость алгоритма к значениям базовых параметров. С одной стороны, это хорошо, т.к. значительно упрощается задача подбора оптимальных значений этих параметров, с другой, нет возможности управлять работой алгоритма, изменяя в широком диапазоне значения параметров.

5.5 Ограничение длины документов

В отобранных новостных кластерах существует большое количество «очень длинных» документов, содержащих более 30 предложений. Учитывая новостной принцип «перевернутой пирамиды», предложения, настолько удаленные от начала документа, как правило, не несут в себе большой смысловой нагрузки и, как кандидаты для включения в обзорный реферат, не представляют большого интереса. Было проведено исследование влияния данного параметра на работу алгоритма. При этом документы усекались до 200, 50, 20, 15, 10, 7, 5, 4 предложений. В среднем, лучшие результаты были получены при укорачивании документов до 10 предложений (см. Таблицу 2).

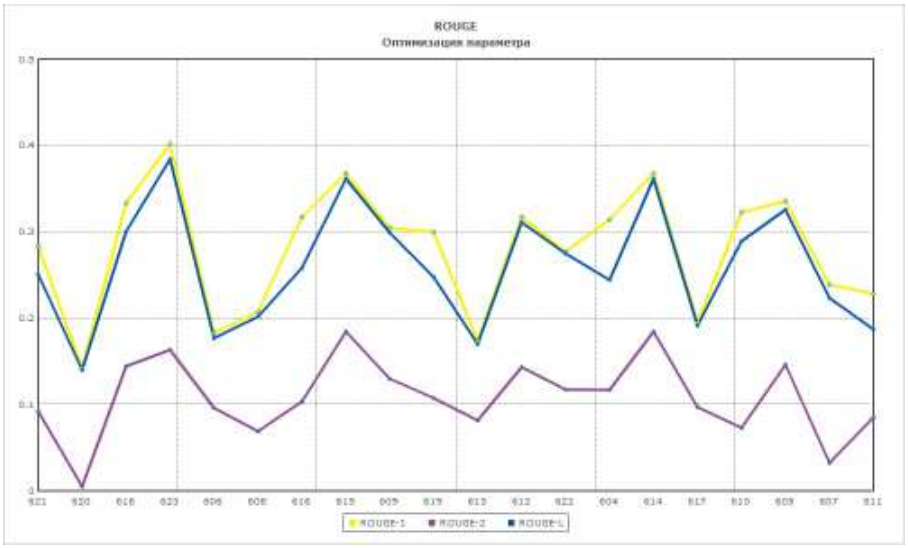

Рисунок 2 - Алгоритм *Manifold Ranking* является очень чувствительным к выбору темы. По горизонтальной оси отложены номера документов, заголовок которых был использован в качестве темы

Таблица 2 – Укорачивание длины документов

Кол-во предложений	***ROUGE-1***	***ROUGE-2***	***ROUGE-L***
200	0.32	0.11	0.27
50	0.32	0.11	0.27
20	0.36	0.12	0.30
15	0.36	0.12	0.30
10	0.42	0.18	0.39
7	0.40	0.17	0.37
5	0.38	0.16	0.36
4	0.38	0.16	0.36

Таблица 3 – Модификация алгоритма: выбор темы

Выбор темы	***ROUGE-1***	***ROUGE-2***	***ROUGE-L***
Заголовок одного документа	0.04-0.48	0.00-0.22	0.04-0.41
Заголовки всех документов	0.37	0.09	0.29
Заголовки из первых двух документов	0.31	0.07	0.23
Заголовки из первых четырех документов	0.18	0.06	0.17
Заголовки из последних двух документов	0.40	0.06	0.27
Заголовки из последних четырех документов	0.38	0.08	0.29

Таблица 4 - Результаты

	ROUGE-1	***ROUGE-2***	***ROUGE-L***
Базовое значение метрики (среднее)	0.28	0.11	0.25
Базовое значение метрики (максимум)	0.50	0.30	0.47
Базовый алгоритм *MR*	0.18	0.10	0.18
BL-1	0.33	0.16	0.32
BL-2	0.34	0.16	0.33
BL-3	0.28	0.08	0.24
BL-4	0.20	0.08	0.20
BL-5	0.01	0.00	0.01
BL-6	0.50	0.28	0.47
BL-7	0.30	0.15	0.25
BL-8	0.33	0.13	0.30
Модифицированный алгоритм	**0.46**	**0.19**	**0.42**

5.6 Исследование влияния выбора темы на работу алгоритма обзорного реферирования Manifold Ranking

Общеизвестным фактом является сильная чувствительность практически любого алгоритма обзорного реферирования к выбору начальной темы кластера. Так как алгоритм *Manifold Ranking* по определению является *«topic focused»*, то был получен действительно довольно большой разброс значений *ROUGE-RUS* для разных тем (Рис. 2).

Т.к. базовый алгоритм предполагает использование одной темы, то авторы попытались выявить зависимость величины *ROUGE-RUS* от выбора темы по следующим критериям:

[1] Дата публикации документа, откуда выбирается тема
[2] Кол-во слов в предложении темы
[3] Кол-во существительных в предложении темы

На имеющихся кластерах нам не удалось выявить различимой закономерности между выбором темы по вышеуказанным критериям и значениями величины *ROUGE-RUS*. Исходя из этого, проанализировав базовый алгоритм ранжирования абстрактных связных структур [6], авторы нашли возможность использовать несколько тем (предложений) как элементов, являющихся источником ранка.

$$y = [y_0, y_1 .. y_n]^T,$$
$$y_i = 1, i \in (0, n),$$

если x_i – предложение, отмеченное как тема, и

$$y_i = 0, i \in (0, n),$$

для всех остальных предложений. Были рассмотрены несколько вариантов модифицированного алгоритма в отношении использования нескольких тем:

[1] Заголовки всех документов
[2] Заголовки из первых двух документов
[3] Заголовки из первых четырех документов
[4] Заголовки из последних двух документов
[5] Заголовки из последних четырех документов.

На имеющихся у авторов кластерах, в среднем, было получено, что наилучшую оценку дает использование всех тем и тем из последних документов; наихудшую – использование одной темы (нестабильно) и тем из первых документов (см. Таблицу 3). Кроме того, была предпринята попытка использовать в качестве тем не заголовки документа, а первое и второе предложения, т.к. в новостях заголовок зачастую призван привлечь внимание, и не всегда достоверно отражает суть вопроса. Однако никакие изменения в сторону повышения качества при использовании с учетом игнорирования тем при использовании заголовков получены не были.

5.7 Результаты подбора параметров

В результате подбора параметров и модификации алгоритма для использования нескольких тем авторам удалось получить результаты оценки мо метрике *ROUGE-RUS,* превосходящие базовую величину метрики на кластере. Результаты оптимизации параметров показаны в Таблице 4.

Заключение

Задача автоматического построения обзорных рефератов на сегодняшний день является очень актуальной. Не менее актуальной является и задача оценки полученных автоматических рефератов. Несомненно, наиболее правдоподобные оценки качества можно проводить в ручном режиме с привлечением большого числа экспертов. Однако такие ручные оценки являются чрезвычайно дорогими.

Методики автоматической оценки качества реферирования не только делают этот процесс более доступным, но и позволяют в реальном времени производить настройку параметров работы определенного алгоритма, производить их оптимизацию.

Авторами была рассмотрена используемая в DUC автоматически вычисляемая метрика *ROUGE*. Адаптировав ее под русский язык и, немного видоизменив, авторы ввели метрику *ROUGE-RUS* и разработали систему для автоматической оценки качества обзорного реферирования на основе этой метрики с Web-интерфейсом.

Исследования метрики *ROUGE-RUS* показали возможность ее применения для оценки качества обзорного реферирования. Для получения стабильных результатов оценки необходимо использовать как минимум 4-5 ручных аннотаций, составленных различными пользователями.

Используя новую метрику *ROUGE-RUS*, была исследована степень влияния выбора базовых параметров и начальной темы на результат работы алгоритма. Алгоритм показал высокую степень устойчивости относительно выбора базовых параметров. На основе результатов исследования влияния выбора темы на работу алгоритма был разработан модифицированный алгоритм. Учет тем всех или нескольких документов позволил повысить качество работы алгоритма.

В результате подбора параметров и модификации алгоритма для использования нескольких тем были получены результаты оценки мо метрике *ROUGE-RUS* не хуже, чем результаты сравнения ручных аннотаций друг с другом. На ряде кластеров были получены результаты, превосходящие базовую величину *ROUGE-RUS* на кластере. Кроме того, удалось улучшить показания метода по сравнению с псевдо-автоматическими аннотациями *«Basic Lines»*.

Таким образом, проведенный эксперимент по составлению ручных аннотаций новостных кластеров различными людьми позволил на основе полученного материала, исследовать как саму метрику *ROUGE-RUS*, так и алгоритм обзорного реферирования *Manifold Ranking*, а также построить

модифицированный алгоритм, показавший значительно лучше результаты по сравнению с базовым.

Литература

[1] H.T. Dang. Overview of DUC 2006. http://duc.nist.gov/pubs/2006papers/duc2006.pdf National Institute of Standards and Technology (NIST)

[2] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. Information Sciences Institute. University of Southern California 2004

[3] Luhn The Automatic Creation of Literature Abstracts (context) http://citeseer.ist.psu.edu/context/74679/0 1958

[4] MDS Evaluation Framework http://mdsevaluation.ru/

[5] Xiaojun Wan, Jianwu Yang and Jianguo Xiao. Manifold-Ranking Based Topic-Focused Multi-Document Summarization http://www.aaai.org/Papers/IJCAI/2007/IJCAI07-467.pdf. DUC 2003. Institute of Computer Science and Technology Peking University, Beijing 100871, China

[6] Zhou et al., 2003b D. Zhou, J. Weston, A. Gretton, O. Bousquet and B. Schlkopf. Ranking on data manifolds. In Proceedings of NIPS'2003.

[7] Белоногов Г.Г., Калинин Ю.П., Хорошилов А.А. Компьютерная лингвистика и перспективные информационные технологии.

[8] 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. Выпуск 8 (15), 2009.

[9] С.Д. Тарасов. Автоматическое составление рефератов новостных сюжетов. Труды 10-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2008, Дубна, Россия, 2008.

The research and parameter's optimization of Manifold Ranking Algorithm based on automatically summarization evaluation metric by ROUGE-RUS

S.D. Tarasov

In this article we review the ROUGE metric adopted by DUC for automatic summarization evaluation and also it modification for Russian language ROUGE-RUS implemented by author.

Developed prototype with WEB interface for automated evaluation of multi-document summarization performance based on ROUGE-RUS

Concluded the experiment with objective to compose the handmade annotations for news clusters and automated evaluation of multi-document summarization.

Defined the base variable ROUGE-RUS on cluster. Reviewed the algorithm of Multi-document summarization "Manifold Ranking".

Researched the impact of diversity penalty (base parameters and initial input) on algorithm final result. The algorithm was modified to compensate above diversity penalty.