Аннотирование по запросу: связность или информативность?

© Кондратьев Михаил

Санкт-Петербургский Государственный Университет
mikhail@oasis.apmath.spbu.ru

Аннотация

Статья описывает результаты экспериментальной оценки алгоритмов аннотирования по запросу, выполненного в рамках РОМИП'2005. Целью этого исследования являлось экспериментальное сравнение нескольких подходов к построению аннотаций для выявления зависимости качества получаемых аннотаций от фрагментированности и информационного наполнения.

1. Введение

Аннотация является эффективным способом представления информации, позволяющим значительно ускорить работу пользователя с коллекцией документов. Одним из важных применений аннотаций является представление результатов работы поисковой системы в виде краткой информации о найденных документах [4].

Двумя основными характеристиками аннотации являются связность и информативность. Идеальная аннотация по запросу представляет собой фрагмент связного текста, описывающего информацию исходного документа, соответствующую запросу пользователя.

Исследования в области аннотирования текстовых документов продолжаются уже более 40 лет и за это время было разработано множество различных способов аннотирования. По методу построения все множество предлагаемых алгоритмов можно поделить на две группы: генерирующие и извлекающие [7].

Генерирующие алгоритмы анализируют исходный документ или документы, с ним связанные, для поиска информации, на основе которой генерируется текст аннотации. При использовании таких подходов текст аннотации строится алгоритмом, основываясь на

правилах естественного языка и специфике информационного домена. В отличие от генерирующих, извлекающие алгоритмы аннотирования формируют аннотацию используя тестовые фрагменты документа или его контекста. Под контекстом документа обычно понимается все множество документов, так или иначе ссылающихся на исходный. Отметим, что многие подходы представляют собой смешанные алгоритмы, использующие различные источники информации и методы ее обработки для построения аннотации.

Сравнение и оценка методов контекстного аннотирования являются трудными задачами вследствие следующих причин [6]:

- автоматическая оценка аннотаций автоматически невозможна, а ручная обработка имеет высокую трудоемкость;
- переиспользование результатов оценки крайне затруднительно.

Интуитивно ясно, что соответствующая запросу пользователя информация расположена во фрагментах текста, находящихся недалеко от термов запроса или содержащих их. На этом предположении основывается большинство методов построения аннотаций по запросу. Однако, термы могут встречаться во многих фрагментах документа и возникает вопрос о том, какие из них должны попасть в аннотацию.

Выбрав аннотацию в виде цельного фрагмента документа мы гарантируем высокую связность аннотации, но рискуем потерять в информационной насыщенности. Такая аннотация с большей вероятностью будет включать в себя небольшую часть термов запроса и может отражать содержимое только какой-то одной части исходного документа. Формируя аннотацию из малых по размеру фрагментов текста позволяет увеличить информационную насыщенность аннотации за счет включения большего количества текстовых элементов, содержащих термы запроса пользователя, но в то же время с уменьшением размера использованных тестовых фрагментов разрушается связность аннотации.

Целью этого исследования являлось экспериментальное сравнение нескольких подходов к построению аннотаций для выявления зависимости качества получаемых аннотаций от фрагментированности и информационного наполнения.

2. Методы построения аннотаций

Для решения задачи аннотирования по запросу необходимо ответить на следующие взаимосвязанные вопросы:

- какого вида фрагменты должны входить в аннотацию?
- сколько фрагментов должно входить в аннотацию?
- какой размер должны иметь текстовые фрагменты?

Нами были рассмотрены следующие три метода построения аннотаций, каждый из которых дает свой ответ на поставленную задачу

2.1 Выбор наилучшего фрагмента

В основе этого алгоритма лежит предположение, что оптимально выбранный в соответствии с запросом единый фрагмент связного текста позволит наилучшим образом отразить ту часть содержания документа, которая интересует пользователя.

В качестве элемента построения аннотации использовался неразрывный фрагмент исходного текста, содержащий максимальное количество термов запроса. При этом к фрагменту выдвигалось следующее требование: два соседних предложения фрагмента, содержащих термы запроса, могут разделяться не более чем одним предложением, не содержащим ни одного терма запроса.

Построение аннотации осуществлялось по следующему алгоритму:

1. Идентификация фрагментов текста, удовлетворяющих вышеупомянутому условию
2. Выбор фрагментов текста, содержащих наибольшее количество различных термов запроса.
3. В случае существования фрагментов, одержащих равное число уникальных термов запросов происходила оценка фрагмента на основе частоты вхождения термов фрагмента в исходный документ. Вес фрагмента при этом вычислялся по следующей формуле:

$$W_F = \sum_{t \in T} \ln(F_t)$$

где T – множество уникальных термов исходного документа, а F_t – частота терма t в документе.

Размер отобранного фрагмента сравнивается с максимально допустимым размером аннотации. В случае, если размер фрагмента меньше размера аннотации, аннотация дополняется первым заголовком, предшествующим выбранному фрагменту текста, который помещается в начало аннотации. Если размер полученной аннотации все еще меньше максимально допустимого, то в аннотацию включаются предложения, следующие в исходном документе после вы-

бранного фрагмента, но не выходящие за границы параграфа, которому принадлежит выделенный фрагмент.

2.2 Выбор комбинации предложений

По сравнению с предыдущим методом, использующим единый связный фрагмент, данный подход позволяет включить в аннотацию большее количество предложений, содержащих термы запроса, однако делает это в ущерб связности результирующего текста.

В качестве единицы построения аннотации использовались предложения исходного документа. Аннотации формировались как комбинации C предложений документа, определяемые следующим образом:

пусть Q – множество термов запроса,
Si – i предложение документа, тогда

$$C = (S_i, S_j, ...), \forall i, j (i < j)$$

$$\forall \text{i}(\text{Si} \cap \text{Q} \neq \text{O}) \text{ и } |\text{C}| < MAX$$

Каждой аннотации присваивался вес, определяемый как число термов запроса, принадлежащих предложениям комбинации:

$$W = \left| (\sum_i S_i) \cap Q \right|, S_i \in C$$

В случае, если находилось более одной комбинации с равными весами W, для каждой из них вычислялся дополнительный вес (аналогично фрагментам текста), позволяющий принять окончательное решение:

$$W_F = \sum_{t \in T} \ln(F_t)$$

Данный алгоритм основывается на предположении, что наиболее информативным представлением документа является множество предложений, содержащее наибольшее количество термов запроса и при этом сохраняющее последовательность изложения информации исходного документа. Построении комбинаций указанным выше образом позволяет гарантировать некоторую степень связности аннотации, так как сохраняет упорядоченность и целостность.

2.3 Выбор комбинации фрагментов предложений

Важной характеристикой информативности аннотации можно считать количество содержащихся в ней термов запроса. В условиях жестких ограничений на размер аннотации для достижения наибольшей концентрации термов запроса необходимо оперировать

единицами текста, меньшими, чем предложение. Для выполнения этой задачи было принято решение использовать модифицированный алгоритм построения комбинаций фрагментов документа, который вместо полных предложений будет использовать только их часть, обрамляющую термы запроса.

При реализации данного алгоритма как элементы комбинации использовались части предложения, имеющие следующий вид: начало фрагмента совпадает с началом предложения, а последний терм фрагмента имеет индекс не больший, чем индекс последнего терма запроса в предложении плюс четыре. Радиус окрестности равный 4 был выбран эмпирически после проведения нескольких экспериментов.

3. Участие в РОМИП'2005

В этом году мы предоставили три прогона, реализующие описанные выше алгоритмы.

3.1 Технические особенности

Для сбора информации об исходном документе и построения аннотаций был разработан набор средств на языке Java. Как основа обработчика HTML документов была использована библиотека tagsoup [14], позволяющая работать с 'грязным' HTML как с обыкновенным XML документом. Во всех случаях, когда построение аннотаций требовало работы на уровне термов, использовался простой стеммер Портера для русского языка [13].

Для всех вычислений использовался копьютер с процессором Intel Pentium 4 2,4 Ghz и объемом оперативной памяти 2 Gb. Построение аннотаций для каждого из трех прогонов не потребовало значительных усилий и было выполнено менее чем за час.

3.2 Результаты

Результаты оценки для уровня relevant minus приведены в следующей таблице.

В случае дорожки аннотирования все системы показали достаточно близкие результаты. Наш исследовательский прототип продемонстрировал средние значения на большинстве из оценок. В то же время, анализ полученных данных дает возможность сделать некоторые выводы, которые могут быть использованы для дальнейшего исследования предложенных алгоритмов аннотирования.

	Accuracy	Error	Precision (macro)
	our / best	our / best	our / best
Метод выбора фрагмента			
and_and	0.73 / 0.79	0.36 / 0.32	0.35 / 0.43
and_or	0.91 / 0.95	0.62 / 0.59	0.35 / 0.43
or_and	0.67 / 0.68	0.27 / 0.18	0.62 / 0.70
or_or	**0.89 / 0.89**	0.51 / 0.41	0.62 / 0.70
Метод комбинации предложений			
and_and	0.75 / 0.79	**0.32 / 0.32**	0.41 / 0.43
and_or	0.93 / 0.95	**0.59 / 0.59**	0.41 / 0.43
or_and	0.66 / 0.68	0.19 / 0.18	0.68 / 0.70
or_or	0.87 / 0.89	0.45 / 0.41	0.68 / 0.70
Метод комбинации фрагментов предложений			
and_and	0.72 / 0.79	**0.32 / 0.32**	**0.43 / 0.43**
and_or	0.91 / 0.95	**0.59 / 0.59**	**0.43 / 0.43**
or_and	0.65 / 0.68	**0.18 / 0.18**	**0.70 / 0.70**
or_or	0.87 / 0.89	**0.41 / 0.41**	**0.70 / 0.70**

4. Анализ результатов

В связи с временными ограничениями мы успели провести лишь ограниченный анализ результатов и здесь приводим самые очевидные выводы из полученных оценок.

4.1 Расхождение в оценке

Сравнение результатов оценки аннотаций асессорами по методу strong и weak позволяет сделать вывод о большом расхождении во мнении асессоров при выставлении оценки на уровне vital. Разница в абсолютных значениях для прогонов, оцениваемых по методу strong (то есть аннотация признается релевантной, если релевантной ее признал каждый из асессоров) и методу weak (достаточно признания аннотации релевантной хотя бы одним из асессоров) достигает 700%. Так, например, для одного из прогонов 'Золушки' соответст-

вующие значения по метрике PrecisionAnnotation равны 0,04 и 0,28 соответственно. На уровне relevant minus расхождение значительно меньше и колеблется от 30 до 100 процентов.

Полученные значения могут свидетельствовать, что сгенерированные аннотации не позволяли асессорам составить точное мнение о содержимом документа и, соответственно, принять решение о его полной релевантности. Такое предположение подтверждается еще и тем фактом, что точность на множестве релевантных аннотаций иногда более чем в 10 раз выше при оценке relevant minus по сравнению с vital.

Причиной столь сильного расхождения мнений асессоров, скорее всего, послужило жесткое ограничение размера аннотации, вследствие чего степень сжатия достигала 10 000 раз. Несмотря на то, что это ограничение оправданно, зачастую указанные рамки представляются слижком узкими для описания содержимого большого по объему документа. Особенно явно это проявляется на запросах, целью которых не являлся поиск конкретного факта в документе.

4.2 Сравнение прогонов

Из общих тенденций следует отметить, что прогоны, выполненные нашим исследовательским прототипом в большинстве случаев показали наибольшее значение по метрике PrecisionAnnotations среди всех участников. При этом прогон, использующий части предложений для построения аннотаций всегда показывал наилучшее значение по этой метрике. Если говорить о всех трех алгоритмах в целом, возможной причиной такого поведения является применение стратегии максимального использования доступного объема аннотации и термов запроса в документе, однако выяснение точной причины требует дополнительного исследования. При условии, что значения метрик AnnotationAccuracy и AnnotationError приблизительно равны результатам других участников, большое количество аннотаций, признанных асессорами релевантными, позволяет говорить об относительной эффектиности предложенных алгоритмов.

Приведенная ниже таблица показывает, какой алгоритм и при каком методе оценки показывал наилучшие и наихудшие результаты. Введено следующее обозначение прогонов: L – алгоритм выбора фрагмента, S – алгоритм выбора комбинации предложений, SF – алгоритм выбора комбинации фрагментов предложений. Разница менее 1 процента считалась пренебрежимо малой.

	Худшая точность	Лучшая точность	Наибольшая ошибка	Наименьшая ошибка
Vital (and_and)	L	SF	SF, S, L	SF, S, L
Vital (and_or)	L, S	SF	SF, S, L	SF, S, L
Vital (or_and)	SF	S	SF, S, L	SF, S, L
Vital (or_or)	SF	S, L	SF, S, L	SF, S, L
Rel + (and_and)	L	SF, S	SF, L	S
Rel + (and_or)	S	L, S	SF, S, L	SF, S, L
Rel + (or_and)	SF	L	L	SF, S
Rel + (or_or)	SF, S	L	S, L	SF
Rel– (and_and)	S	SF	L	SF, S
Rel– (and_or)	S	SF, L	L	SF, S
Rel– (or_and)	SF	S, L	L	SF
Rel– (or_or)	L	SF, S	L	SF

Как уже отмечалось, все системы, принимавшие участие в дорожке аннотирования, показали достаточно близкие результаты. Анализ результатов прогонов, выполненных нашим исследовательским прототипом не позволил выявить явного фаворита среди трех предложенных алгоритмов.

Несмотря на это, интересно вкратце рассмотреть зависимости, проявляющиеся в оценках каждого из прогонов.

4.3 Метод выбора наилучшего фрагмента

Несмотря на опасения, что данный алгоритм окажется неприменим к столь малым по объему аннотациям, результаты прогона на некоторых уровнях релевантности оказались даже лучше, чем у других подходов. Так, например, на уровне релевантности relevant plus прогон показал наилучший результат при всех методах оцеки за исключением and_and.

В то же время, значения метрики AnnotationError для этого алгоритма были всегда максимальны. Особенно ярко это проявилось на уровне relevant minus, где алгоритм показал наихудшее значение ошибки при всех методах оценки, причем в случае оценки по методу weak отличие от лучшего результата составляло более 10%. Возможной причиной относительно высокого значения метрики Annota-

tionError является тот факт, что алгоритм дает представление лишь об одном аспекте, а не о содержании документа в целом.

Для данного алгоритма интересно опробовать его применение на аннотациях с меньшей степенью сжатия документа. В случае, когда на аннотацию не накладывается жесткое ограничение по размеру, может иметь смысл использование комбинации наилучших фрагментов текста для представления содержимого исходного документа.

4.4 Алгоритмы выбора комбинации предложений и частей предложений

Данные алгоритмы показали средние результаты на фоне других систем по значениям метрики AnnotationAccuracy, лишь в 1 случае сравнявшись с наилучшим результатом, но никогда и не показывая наихудший результат среди представленных. Интересно, что алгоритмы показали хорошие значения по метрике AnnotationError, чье значение часто оказывалось наилучшим среди результатов, показанных всеми участниками. Особенно явно это прослеживается на значениях метрик, полченных при оценке аннотаций на уровне релевантности relevant minus для алгоритма использующего фрагменты предложений для формирования аннотаций. Оценки, полученные на этом уровне оказались наилучшими среди всех участников как для мягкого, так и для строгого типа оценок.

Полученные значения метрик свидетельствуют, что аннотация, построенная как комбинация фрагментов предложений, содержит большое количество информации релевантной запросу и будет скорее отнесена к релевантным, чем нет. Это хорошо соотносится с использованным алгоритмом конструирования аннотации и тем фактом, что на всех уровнях релевантности и при всех методах оценки предложенный алгоритм продемонстрировал значения метрики PrecisionAnnotation наибольшие среди участников.

Большой интерес представляет оценка алгоритмов, использующих различные по размеру фрагменты предложений документа для построения аннотации. Построение зависимости между размером фрагмента и качеством аннотирования позволило бы найти оптимальную стратегию для краткого представления информации.

5. Заключение

В данной работе рассматривались алгоритмы извлекающего аннотирования по запросу на основе текста исходного документа. Такой подход обусловлен происхождением используемой коллекции документов, которая составлялась как смешение web страниц, принадлежащих домену narod.ru, и коллекции юридических документов. Большая часть страниц из вышеупомянутых подколлекций имеет минимальный контекст, что позволяет сделать предположение о низкой эффективности алгоритмов, использующих контекст документа для формирования текста аннотации. В то же время, при разработке алгоритмов учитывалась гипертекстовая разметка исходных документов, которая позволяет извлечь дополнительную информацию и использовать ее в процессе аннотирования.

В рамках семинара РОМИП была выполнена основная задача – проведен сравнительный анализ алгоритмов с целью выявить оптимальное соотношение между связностью и информационной насыщенностью текста аннотации в завивисимости от размера используемых текстовых фрагментов. Несмотря на то, что результаты оценки не позволили выявить бесспорного лидера, полученные данные показали важные зависимости, проявляющиеся при использовании предложенных алгоритмов и дали информацию для дальнейших исследований в этой области.

Литература

[1] М. Губин, А. Меркулов. Эффективный Алгоритм Формирования Контекстно-Зависимых аннотаций, Диалог 2005

[2] Труды РОМИП'2004, http://romip.narod.ru/romip2004/index.html

[3] A. Berger, V. O. Mittal. Query-Relevant Summarization using FAQs, in proc. of the 38th ACL

[4] A. Tombros, M. Sanderson. Advantages of Query Biased Summaries, in proc. of the ACM SIGIR, 1998

[5] I. Mani, T. Firmin, B. Sundheim, The TIPSTER SUMMAC Text Summarization Evaluation, in proc. of the EACL, 99

[6] I. Mani, Summarization Evaluation: An Overview, in proc. of the NTCIR Workshop, 2001

[7] I. Mani. Automatic Summarization, JohnBenjamin's Publishing Company, 2001.

[8] M. Sanderson, Accurate user directed summarization from existing tools, in proc. of the CIKM, 1998.

[9] M. Wu, R. Wilkinson, C. Paris. Evaluation of a Query-biased Document Summarisation Approach for the Question Ansnwering Task, in proc. of the Australasian Language Technology Workshop 2004
[10] R. White, J. M. Jose, I. Ruthven. Query-Biased Web Page Summarisation: A Task-Oriented Evaluation, in proc. of the SIGIR, 2001
[11] Y. Chali, S. Matwin, S. Szpakowicz. Query-Biased Text Summarization as a Question-Answering Technique, in proc. of the AAAI Fall Symposium Series, 1999
[12] Snowball Web site, 2005. http://snowball.tartarus.org
[13] Tagsoup library web site, 2005 http://mercury.ccil.org/~cowan/XML/tagsoup/

Query-based summarization: readability or details

Mikhail Kondratyev

The article describes results of the experimental evaluation of different query summarization approaches in the frame of RIRES'2005 summarization track. The goal of experiments was to evaluate the impact of summary readability and information value on the quality of summarization.