АВТОМАТИЧЕСКОЕ ОБНОВЛЕНИЕ АННОТАЦИИ НОВОСТНОГО КЛАСТЕРА

Автор: Алексеев Алексей

Определение новизны информации

- Определение новизны информации – важная и нерешённая задача.
- Проблема в общем виде:
  - поток информации и пользователь
  - в некоторый момент времени есть известная информация (известная пользователю)
  - Задача: извлечение новой информации из потока и предъявление пользователю

*

Конкретная задача

- Новостной кластер – набор документов по поводу некоторого события.
- Аннотация – краткое описание события, составленное из предложений документов кластера.
- В некоторый момент времени в кластер приходит ещё N документов.

Вопросы:

- ✔ **Что нового произошло?**
- ✔ **Как должна измениться аннотация?**
- **Как новое отобразить в аннотации?**
- **Какие предложения аннотации должны быть заменены?**

*

Конференция TAC

- Создана при поддержке и спонсируется Национальным Институтом Стандартов и Технологий (NIST) и Департаментом Защиты США.
- Проект был запущен в 2008 как продолжение конференции DUC.
- Участники – более 30 команд со всего мира.
- Назначение: поддержка исследований в области извлечения информации при помощи обеспечения инфраструктуры, необходимой для крупномасштабной оценки методов извлечения информации.

*

Постановка задачи «Обновление аннотации» в TAC - 1

- Данная задача впервые была поставлена в TAC в 2008 году и продолжает развиваться.
- **Постановка задачи:** Даны два упорядоченных и связанных множества документов (по 10 документов в каждом) и запрос пользователя.
- **Задача:** Сделать две аннотации, размером не более 100 слов, такие что:
  1. Первая аннотация покрывает первое множество документов.
  2. Вторая аннотация покрывает второе множество документов, при условии что пользователь уже ознакомлен с документами первого множества.

Постановка задачи «Обновление аннотации» в TAC - 2

- То есть по сути задача делилась на две основные и формально независимые подзадачи:
  1. Создание аннотации набора документов (Initial Summary)
  2. Создание обновлённой аннотации (Update Summary)

- Некоторые детали:
  1. Аннотации свыше 100 символов обрезались.
  2. Документы упорядочены по времени.
  3. Документы релевантные запросу пользователя.
  4. Независимая оценка аннотаций.

*

Входные данные для задачи «Обновление аннотации» в TAC - 1

- AQUAINT-2 collection
  1. New York Times
  2. Associated Press
  3. Los Angeles Times-Washington Post News Service
  4. Xinhua News Agency
  5. Agence France Presse
  6. Central News Agency (Taiwan)
  7. …
- 2.5 Гб текста – около 900.000 документов.
- Октябрь 2004 – Март 2006.
- Все документы на английском языке.
- Данная коллекция идеально подходит для поставленной задачи.

*

Входные данные для задачи «Обновление аннотации» в TAC - 2

- Специалисты NIST сделали 48 различных топиков.
- Каждому топику было отобрано по 20 релевантных документов.
- Документы были хронологически упорядочены и разделены на 2 множества, так что документы множества Б следовали за А хронологически.
- К каждому топику был составлен запрос, ответ на который содержался в предложенных документах.
- Запросы могли содержать вопросительные предложения и избыточную информацию *

Оценка результатов задачи «Обновление аннотации» в TAC

- Специалисты NIST сделали вручную по 4 «идеальных» аннотации к каждому топику.
- Применялось несколько различных и независимых способов оценки результатов:
  1. Автоматические ROUGE метрики.
  2. Оценка содержания аннотации методом «Пирамиды».
  3. Ручная оценка полноты, связности и читабельности.
- Все системы были независимо оценены

Автоматические ROUGE метрики - 1

- **ROUGE** или **Recall-Oriented Understudy for Gisting Evaluation** – набор метрик и комплекс программ для оценки автоматического аннотирования и машинного перевода текстов.
- Основная идея – сравнение генерированного текста с “эталонным”, сделанным человеком.
- Существуют различные формы метрики, сравнивающие:
  1. n-граммы (ROUGE-N)
  2. минимальные общие подстроки (ROUGE-L и ROUGE-W)
  3. монограммы и биграммы (ROUGE-1 and ROUGE-2) *

Автоматические ROUGE метрики - 2

- **Общая формула:**

$$ROUGE - N(A_i) = \frac{\sum_{M_{ij}} count(Ngram(A_i) \cap Ngram(M_{ij}))}{\sum_{M_{ij}} count(Ngram(M_{ij}))}$$

  - A_i – оцениваемая обзорная аннотация *i*-того кластера.
  - M_{ij} – ручные аннотации *i*-того кластера.
  - *Ngram(D)* – множество всех n-грамм из лемм соответствующего документа *D.*

- **Пример:**

1. Китай и Тайвань установили авиасообщение после 60-летнего перерыва.
2. После почти 60-летнего перерыва открылось регулярное авиасообщение между Тайванем и материковым Китаем.

$$\text{Rouge-1} = 7/12 = 0.58(3)$$

*

Метод «Пирамиды» - 1
(Pyramid Evaluation)

- Разработан в 2005 году Колумбийским университетом.
- Эксперты выделяют из «эталонных» аннотаций «информационные единицы» - Summary Content Units (SCUs).
- Каждый SCU получает вес, равный количеству «эталонных» аннотаций, где она встречалась.
- Оценка – суммарный вес входящих SCU.
- Неоднократное вхождение SCU в автоматическую аннотацию не

*

Метод «Пирамиды» - 2
(Pyramid Evaluation)

- **Итоговый результат:**

$$\frac{\text{[Суммарный вес найденных SCU]}}{\text{[Суммарный вес всех определённых SCU для данного топика]}}$$

- **Пример:**

SCU: Мини-субмарина попала в ловушку под водой.

1. мини-субмарина... была затоплена... на дне моря...
2. маленькая... субмарина... затоплена... на глубине 625 футов.
3. мини-субмарина попала в ловушку... ниже уровня моря.

*

Ручная оценка результатов на TAC

- Каждая автоматическая аннотация была прочитана несколькими экспертами NIST.
- Две оценки:
  - Содержание
  - Читабельность
- Пятибалльная система оценка – от 1 до 5.
- Результаты – заметный разрыв между автоматическими и «эталонными» аннотациями.
- Данная система оценки наиболее важна для нас, так как цель автоматического реферирования – человек, а не компьютер.*

Сравнение методов оценки

- **ROUGE:**

  + Малое участие человека, лёгкость применения

  - Отсутствие оценки читабельности, результат не всегда идеален с точки зрения человека

- **Метод «Пирамиды»:**

  + Наиболее объективная оценка содержания аннотации

  - Отсутствие оценки читабельности, большое участие человека

- **Ручная оценка:**

  + Оценка «пользователем», лучшая оценка читабельности

*

Результаты TAC 2008 – 1

- В целом не очень высокие результаты – заметный разрыв между «эталонными» и автоматическими аннотациями.
- Рассматриваем ручную оценку результатов.
- Лучший результат по содержанию:

  2.7917 - для 1-ой аннотации, **2.6042** – для второй.
- Лучший результат по читабельности:

  3.0000 – для 1-ой аннотации, **3.2083** – для второй.

(не учитывая «базовую» аннотацию NIST)

*

Результаты TAC 2008 – 2

-------Summaries A--------		----------Summaries B-----	
ID	Score	ID	Score
F	4.7917	D	4.875
D	4.7917	G	4.75
A	4.75	H	4.6667
G	4.6667	F	4.6667
B	4.5833	A	4.625
H	4.5	B	4.5833
C	4.4583	C	4.5417
E	4.4167	E	4.2917
50	2.7917	14	2.6042
26	2.7917	49	2.5833
12	2.7708	23	2.5833
49	2.75	11	2.5625
44	2.75	44	2.5208
42	2.75	24	2.5
23	2.75	50	2.4583
52	2.7083	41	2.4375

Результаты по содержанию аннотации

*

Результаты TAC 2008 – 3

Summaries A		Summaries B	
ID	Score	ID	Score
F	4.9167	D	4.9583
G	4.875	F	4.875
D	4.875	A	4.875
B	4.8333	G	4.8333
A	4.7917	B	4.7917
E	4.75	H	4.75
H	4.625	E	4.7083
C	4.625	C	4.5833
0	3.25	0	3.4167
50	3	49	3.2083
49	2.9375	23	3.1042
24	2.9375	52	2.9792
26	2.875	26	2.8958
51	2.8333	25	2.8958
52	2.8125	44	2.8542
23	2.8125	34	2.8542
1	2.75	46	2.8333

Результаты по читабельности аннотации

*

Анализ результатов TAC 2008

- Одна из лучших – система канадского университета Монтреаль для франкоговорящих. (Universit´e de Montreal)
- Стабильно высокие результаты для содержания аннотации и читабельности.
- Третье участие данной команды в DUC-TAC конференциях.
- Базовый алгоритм:

  «Максимальная граничная значимость»

  Maximal **M**arginal **R**elevance (**MMR**)

*

Maximal **M**arginal **R**elevance (**MMR**) - 1

- Итеративный метод.
- На каждой итерации производится ранжирование предложений-кандидатов.
- В итоговую аннотацию отбирается одно с самым высоким рангом.
- Давно используется для запрос - ориентированного аннотирования.
- Модификации алгоритма для «базовой» и «обновлённой» аннотаций.

*

Maximal Marginal Relevance (MMR) - 2

- **Для «базовой» аннотации:**

  Пусть:

  Q – запрос к системе.

  S – множество предложений кандидатов.

  s – рассматриваемое предложение кандидат.

  E – множество выбранных предложений.

  Тогда:

$$\mathrm{MMR} = \arg\max_{s \in S} \left[\lambda \cdot Sim_1(s, Q) - (1 - \lambda) \cdot \max_{s_j \in E} Sim_2(s, s_j) \right]$$

*

Maximal Marginal Relevance (MMR) - 3

- **Для «обновлённой» аннотации:**

  Пусть:

  Q – запрос к системе.

  s – рассматриваемое предложение кандидат.

  H – рассмотренные документы (история).

  f(H) –> 0 при увеличении H.

  Тогда:

$$\text{SMMR}(s) = Sim_1(s, Q) \cdot \left(1 - \max_{s_h \in H} Sim_2(s, s_h)\right)^{f(H)}$$

*

Maximal Marginal Relevance (**MMR**) - 4

- **Sim1(s,Q)** – стандартная косинусовая мера угла между векторами:

$$\cos\theta = \frac{\mathbf{v_1} \cdot \mathbf{v_2}}{\|\mathbf{v_1}\| \, \|\mathbf{v_2}\|}$$

  - **Sim2(s,s_h)** – максимальная общая подстрока (Longest Common Substring):

$$\text{Sim2}(s,s_h) = \frac{2 * \text{Length}(\text{ LCS}(s,s_h)\ )}{\text{Length}(s) + \text{Length}(s_h)}$$

*

Постпроцессинг (Post-processing)

- После отбора предложений производится улучшение связности и читаемости аннотации:

1. **Замена аббревиатур**
2. **Приведение номеров и дат к стандартному виду**
3. **Замена временных ссылок:**

   «в конце следующего года» □ «в конце 2010»
4. **Замена двусмысленностей и дискурсивных форм:**

   «Но, это значит…» □ «Это значит…»

*

Направление дальнейшей работы

- Поиск принципиально иных подходов к созданию «обновлённой» аннотации.
- Реализация существующих подходов с целью выявить их «слабые» места.
- Модификация существующих и создание новых (комбинированных?) методов.
- Поиск существующих и создание новых методов постпроцессинга (улучшение читабельности и связанности текста)
- Изучение связей документов, принадлежащих одному кластеру (ссылочная структура)

*

The End