



Реферирование и аннотирование

Ефремова Наталья Эрнестовна
Грацианова Татьяна Юрьевна
Большакова Елена Игоревна

Содержание



- Определение, назначение
- Виды реферирования
 - Экстрагирование
 - Абстрагирование
 - Стратегии отбора предложений
- Оценка качества реферата
- Задание



Определения



Российский ГОСТ

- Реферат — краткое точное изложение содержания документа, включающее основные фактические сведения и выводы, без дополнительной интерпретации или критических замечаний автора реферата;
- Сводный реферат — реферат, составленный на основе двух и более исходных документов;
- Аннотация — краткая характеристика документа с точки зрения его назначения, содержания, вида, формы и других особенностей

Abstract Summary Summarization

Аннотирование и реферирование документов



- Аннотирование и реферирование документов): сжатое изложение содержания
 - отдельного текста
 - совокупности близких по смыслу текстов – обзорное реферирование
 - ✓ обзорный реферат новостного кластера
 - ✓ историческая справка
 - ✓ аналитическая записка
- Назначение – быстрое ознакомление с содержанием документа/документов
- Степень сжатия м.б. задана:
 - ✓ относительная - в процентах (20%),
 - ✓ абсолютная - в словах (100, 200, 300, 400 слов)

Пример: сниппеты



Сниппет — это блок информации о найденном документе, который отображается в результатах поиска. Сниппет состоит из заголовка, описания или аннотации документа и пр.

Яндекс ✕ ⌵ Найти

[ПОИСК](#) [КАРТИНКИ](#) [ВИДЕО](#) [КАРТЫ](#) [МАРКЕТ](#) [НОВОСТИ](#) [ПЕРЕВОДЧИК](#) [ЕЩЁ](#)

S Аннотирование и реферирование
[StudFiles.ru > preview/1910691/](#) ▼
При **реферировании** сообщение освобождается от всего второстепенного, иллюстративного, поясняющего, сохраняется лишь сама суть содержания.

S реферирование и аннотирование текстов
[StudFiles.ru > preview/3579190/](#) ▼
Реферирование включает в себя некоторые сведения, касающиеся автора статьи, времени ее написания, указания на главную идею статьи...

Refeferирование — Педагогическое речеведение.
[ped_recheved.academic.ru > 191/Реферирование](#) ▼
Реферирование (от лат. *referre* — докладывать, сообщать) — один из видов речевой деятельности, занимающий промежуточное положение между продуктивными и репродуктивными видами речевой деятельности.

Области применения автоматического аннотирования



- Автоматизированный отбор документов, соответствующих тематике (пользователь делает это более релевантно, но быстрее по аннотации)
- Автоматизированная кластеризация научных статей: основные идеи и направления, развиваемые или критикуемые в статье
- Автоматическое аннотирование голосовых сообщений (установление приоритета звонка)
- Аннотирование форумных веток: установить значимость и интенсивность обсуждения темы)
- Аннотация совещаний и встреч

Виды реферирования. Классификация



- По технологии сжатия
- По количеству аннотируемых документов (один – много)
- По типу содержания
 - Индикативная (общее содержание)
 - Информативная (сохранение информативной ценности)
- По типу наполнения
 - Фрагмент текста (Common summarization)
 - Предложения (Headline summarization)
 - Ключевые слова (Keyword summarization)
- По языку генерируемой аннотации – совпадает или нет с языком исходного документа
- По потребности пользователя

Технология сжатия



Абстракты Vs. Экстракты

- Экстракты получаютс~~я~~ извлечением фрагментов исходных текстов (предложений) – основная применяемая технология с 70-80 гг.
Квазиреферирование
- Абстракты порождаются (генерируются)– экспериментальные методы
- Гибридные методы

Входной текст/тексты



- Аннотирование одного документа (Single-document summarization)
- Формирование обзорного реферата –аннотации набора документов (Multi document summarization).
Документы обычно связаны темой, сюжетной линией, и др. параметрами.
- Обработка потоков новостей. Задача решается в крупных коммерческих агрегаторах Rambler.News, Yandex .News, Google.News

Аннотация по потребности пользователя



- Общая аннотация (Generic)
- Аннотирование по запросу (*Query-based*) – *контекстная аннотация*,
тематически-ориентированная аннотация
- Обновленная аннотация (Update) – чувствительность ко времени, выбор новых фактов

Разные потребности пользователя – разные критерии адекватности

Экстрагирование: подходы



Выделение предложений может быть основано на:

- определении значимости (весе) предложения
- построении тематических моделей текста
- определении «центральности» (наибольшей значимости) предложения на основе графа предложений
- машинном обучении: использование различных признаков предложений и их комбинаций

Метод, основанный на весах предложений



- Частотные слова текста:
Вес предложения зависит от весов слов
- Заголовок (название) текста:
Вес предложения увеличивается, если в нем есть слова заголовка
- Присутствие особых слов и конструкций:
Подчеркнем, результатом будет ...
- Позиция предложений в тексте:
Предложения в начале более важны
- Длина предложения
- Именованные сущности
- Связность предложения с предыдущим

Расчет веса предложения



Из потенциального списка значимых слов исключаются

- Стоп-слова (предлоги, союзы, ...)
- Слова, широко употребляемые в данной предметной области («клетка» для биологии)
- Слова с низкой частотностью в рамках рассматриваемой текстовой коллекции

Модели назначения веса слова



1. Расчет вероятностей слов исходного кластера: наиболее частотные слова кластера с большей вероятностью должны оказаться в аннотации

$$p(w_i) = n / N$$

n – число вхождений слова w_i ; N – общее число слов-токенов

2. Метод *TF-IDF*

□ Частотное слово в исходном документе может оказаться недостаточно характерным для данного документа

□ Полезна информация из более широкой (*контрастной, фоновой*) текстовой коллекции

Вес некоторого слова пропорционален количеству употребления этого слова в документе, и обратно пропорционален частоте употребления слова в других документах коллекции.

$$\text{idf}_t = \log (F/\text{df}_t)$$

F – размер фоновой коллекции:

df – число документов фоновой коллекции, где встретилось слово t

Смысл: чем реже термин встречается во всех документах, тем он более информативен

Тематические модели для аннотирования



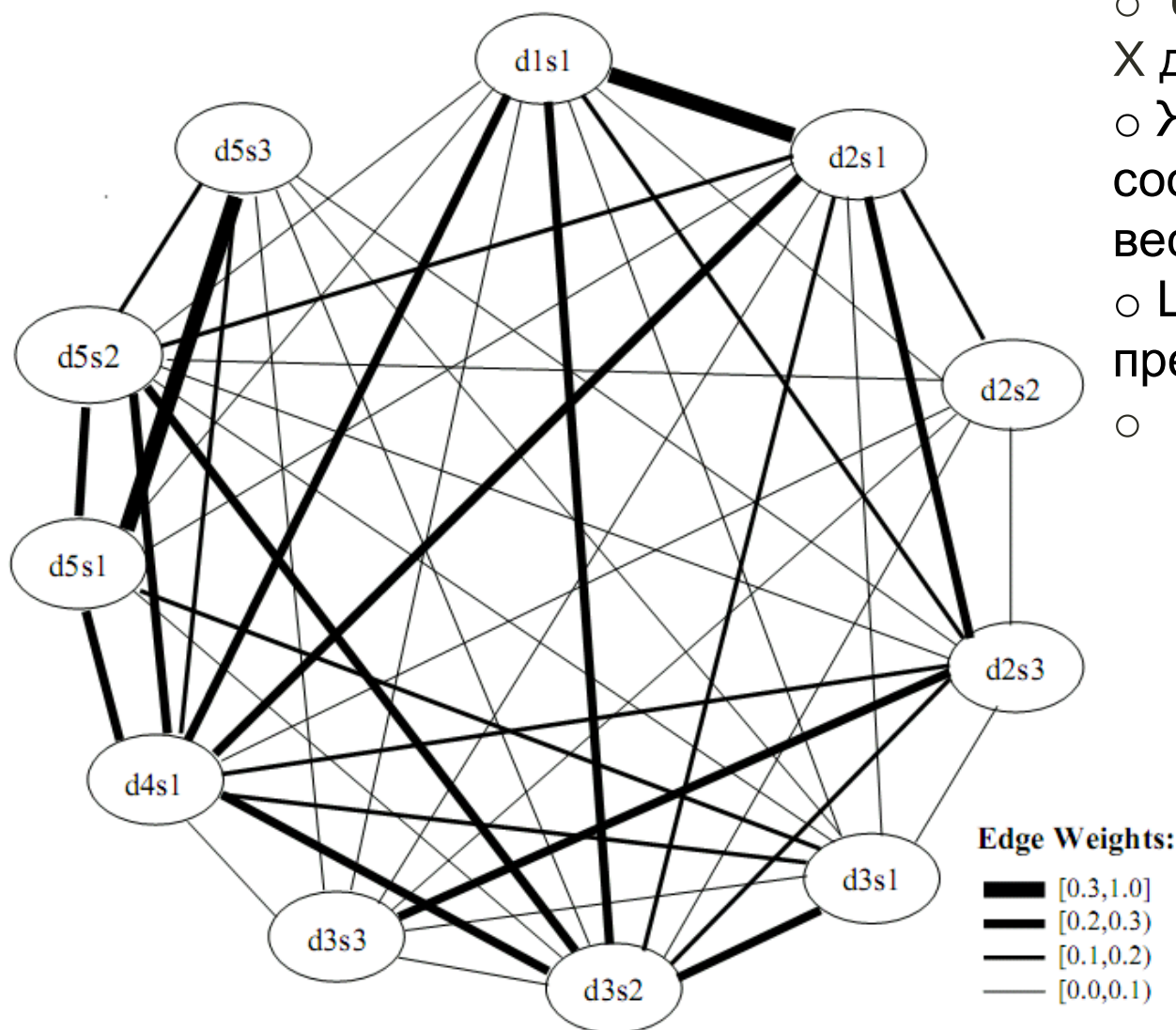
- Лексические цепочки (Lexical Chains): объекты могут описываться набором связанных слов и выражений
 - Для формирования цепочек используются тезаурусы
 - При каждой итерации отбираются
 - предложения содержащие узловой элемент наиболее значимой цепочки
 - Предложения, содержащие 2 или более значимые цепочки
- Латентный семантический анализ Latent Semantic Analysis, LSA – неявное представление семантических особенностей текста на основании статистики совместной встречаемости слов
- Байесовские тематические модели (Bayesian Topic Models) – исследуется вероятность появления пары (документ-слово)

Построение графа предложений



- Аннотирование наборов похожих документов: многие предложения схожи друг с другом
- Взвешенный граф сходства предложений:
 - Вершины графа – предложения
 - Дуги – сходство (*similarity*) между предложениями
 - сходство определяется косинусной мерой
- Важны значительные степени сходства, выше заданного порога: 0.1 / 0.2 / 0.3, связи ниже этого порога не рассматриваются.
- *Центральность предложения* – при заданном пороге оно связано с максимальным числом других предложений.
- Это предложение выбирается в аннотацию.

Граф предложений



- $dXsY$ – Y предложение в X документе
- Жирные линии соответствуют высокому весу
- Центральное предложение: $d4s1$
-

Машинное обучение



Задача бинарной классификации: разбиение всех предложений входной коллекции на принадлежащие и не принадлежащие итоговой аннотации

- Необходима коллекция документов и аннотаций-экстрактов
- При обучении подбирается комбинация из характеристик извлекаемого предложения:
 - ✓ длина предложения (больше/меньше порога)
 - ✓ позиция в тексте: номер абзаца, позиция внутри абзаца (в начале/середине/конце)
 - ✓ присутствие ключевых фраз (*результатом был...*)
 - ✓ наличие слов из заголовка документа
 - ✓ количество тематических слов (слов ПО)
 - ✓ присутствие имен собственных

Стратегии отбора предложений



□ Локальная оптимизация

Оптимизация выбора очередного предложения

Итеративный метод локальной оптимизации

Maximal Marginal Relevance – MMR (Carbonell, Goldstein, 1998)

□ Глобальная оптимизация

- Построении аннотации исходя из качества результирующей аннотации в целом
- Оптимизация некоторой функции, напр., присутствие наиболее частотных слов или выражений в аннотации; оценка информативности

Метод MMR



- На каждой итерации производится ранжирование предложений-кандидатов
- В итоговую аннотацию отбирается одно с самым высоким рангом:
 - Максимизируется сходство с исходным документом (набором документов)
 - Минимизируется сходство с уже отобранными в аннотацию предложениями
- При построении контекстно-зависимой аннотации максимизируется сходство с запросом

Метод MMR: выбор предложения



Контекстно-зависимая аннотация:

Пусть

Q – предложение-запрос к системе

S – множество предложений-кандидатов

s – рассматриваемое предложение-кандидат

E – множество уже выбранных предложений

Тогда:

$$\text{MMR} = \arg \max_{s \in S} \left[\lambda \cdot \text{Sim}_1(s, Q) - (1 - \lambda) \cdot \max_{s_j \in E} \text{Sim}_2(s, s_j) \right]$$

Используются различные меры схожести предложений Sim

Пример алгоритма.

Метод SumBasic



1. Расчет вероятностей слов исходного текста: наиболее частотные слова с большей вероятностью должны оказаться в аннотации

$$p(w_i) = n / N$$

n – число вхождений слова w_i ; N – общее число слов-токенов

2. Каждому предложению назначается вес, равный средней вероятности весов слов в данном предложении

$$weight(S_j) = \sum_{w_i \in S_j} \frac{p(w_i)}{|\{w_i \mid w_i \in S_j\}|}$$

3. Предложение с наибольшим весом отбирается в итоговую аннотацию
4. Пересчет вероятностей для слов из отобранного предложения: вес слов резко снижается.
5. Вычисляется общая длина получившейся аннотации, если она не превосходит порога, переход к шагу 2.

Постобработка аннотации-экстракта



После извлечения предложений производится улучшение связности и читаемости аннотации:

- Замена аббревиатур
- Приведение номеров и дат к стандартному виду
- Замена временных ссылок:
в конце следующего года → *в конце 2017 г.*
- Замена двусмысленностей и дискурсивных форм: *Но, это значит...* → *Это значит...*
- Конечная сортировка предложений

Проблемы аннотирования на основе извлечения



- Дублирование информации, лишние фрагменты (обороты) в предложении
- Отсутствие связности изложения (анафоры) и целостности
 - В парламентской гонке лидируют 3 партии: А, В и С. Партия С принимает участие в выборах впервые. Ее лидером является известный политик ...
- Возможно нарушение смысла

В ходе предвыборной гонки кандидат У набрал наибольшее количество голосов избирателей. Кандидат Х проиграл в теледебатах. На прессконференции он заявил, что принимает решение отказаться от участия в выборах.

Абстрагирование: идея



- Абстрагирование (*Abstracting*) – порождение нового текста
- В настоящее время возможно для текстов из узких предметных областей
- Общая схема построения аннотации:
 1. Извлечение определенной информации из текстов (обычно с использованием лингвистических шаблонов)
 2. Заполнение извлеченными данными заранее определенных шаблонов типичных фраз аннотаций

Абстракты: пример



Система SUMMONS (Radev D. R., McKeown K. R.)

- Основана на знаниях: новости о терроризме
- Шаблоны для извлечения информации используют специфику ПО
 - шаблон является подобием реляционной БД, содержит определенную информацию из новости, например, место, количество террористов, вид оружия и т.п.)
- Создается концептуальное представление текста – какие понятия из шаблонов включать в реферат
- Выражение концептуального представления текста на ЕЯ
 - Определяется структура предложения
 - Генерация предложений (используется системная грамматика английского языка для постройки синтаксического дерева, выбора класса слов и превращения дерева в предложение)

SUMMONS : пример шаблона



MESSAGE: ID	TST3-MUC4-0010
MESSAGE: TEMPLATE	2
INCIDENT: DATE	01 NOV 89
INCIDENT: LOCATION	EL SALVADOR
INCIDENT: TYPE	ATTACK
INCIDENT: STAGE OF EXECUTION	ACCOMPLISHED
INCIDENT: INSTRUMENT ID	-
INCIDENT: INSTRUMENT TYPE	-
PERP: INCIDENT CATEGORY	TERRORIST ACT
PERP: INDIVIDUAL ID	"TERRORIST"
PERP: ORGANIZATION ID	"THE FMLN"
PERP: ORG. CONFIDENCE	REPORTED: "THE FMLN"
PHYS TGT: ID	-
PHYS TGT: TYPE	-
PHYS TGT: NUMBER	-
PHYS TGT: FOREIGN NATION	-
PHYS TGT: EFFECT OF INCIDENT	-
PHYS TGT: TOTAL NUMBER	-
HUM TGT: NAME	-
HUM TGT: DESCRIPTION	"1 CIVILIAN"
HUM TGT: TYPE	CIVILIAN: "1 CIVILIAN"
HUM TGT: NUMBER	1: "1 CIVILIAN"
HUM TGT: FOREIGN NATION	-
HUM TGT: EFFECT OF INCIDENT	DEATH: "1 CIVILIAN"
HUM TGT: TOTAL NUMBER	-

SUMMONS: алгоритм



1. Ввод данных. Набор шаблонов.
2. Предварительная обработка
 1. Группировка шаблонов в хронологическом порядке
 2. Создание БД шаблонов
 3. Группировка шаблонов, относящихся к одному событию
3. Эвристическое объединение – выяснение отношений между шаблонами: генерализация, подтверждение информации другим источником, противоречия между источниками
4. Планирование структуры текста – представление информации внутри абзаца
5. Генерирование текста – лингвистическое резюмирование информации

Оценка качества аннотаций



Классификация

- ☐ Степень участия человека (автоматическое, ручное, полуавтоматическое)
- ☐ Внешняя и внутренняя
 - ☐ внешняя (extrinsic) - может ли аннотация заменить исходный текст
- ☐ Критерий оценки
 - ☐ Читабельность, «гладкость»
 - ☐ Соответствие по смыслу аннотируемому тексту
 - ☐ Соответствие запросу

Оценка качества аннотаций: мера ROUGE



Recall-Oriented Understudy for Gisting Evaluation

Сравнение сгенерированной автоматической аннотации с эталонной

- по количеству совпадающих N-грамм
- оценка длин максимальных общих подпоследовательностей по отношению к общей длине предложений
- Оценка с добавлением веса N-граммы

Бесплатные ресурсы



- visualworld.ru/ Лаборатория автоматизации научных исследований (Санкт-Петербург)
 - построение семантической сети (ассоциативного облака) по тексту
 - фильтрация семантической сети от малозначимых элементов
 - обратное преобразования семантической сети в текст
- support.office.com Microsoft – выбор предложений, имеющих «высший балл»



Соревнования



Российский семинар по Оценке Методов Информационного Поиска

<http://romip.ru/ru/2009/tracks/annotation.html>

<http://romip.ru/ru/2010/tracks/annotation.html>

Контекстно-зависимое аннотирование

Критерии оценки

- ☐ *Информативность: насколько аннотация понятна для принятия решения о полезности документа в контексте этого запроса.*
- ☐ *Читабельность: Понятен ли смысл?"*

47% документов признаны нерелевантными

Заключение. Перспективные направления аннотирования



- ❖ Новые направления
 - ❖ Построение заголовков к тексту
 - ❖ Гибридные источники, работа с мультимедиа
 - ❖ Новые форматы, использование дополнительной информации
- ❖ Новые методики
 - ❖ Порождение предложения из фрагментов разных предложений
 - ❖ Использование словарей синонимов, замена длинных цепочек короткими синонимичными

Президент Соединенных Штатов Америки Дональд Трамп

- ❖ Сжатие внутри предложения: удаление оборотов, перефразирование

Парламент на основе заслушанных докладов принял окончательное решение ...

Парламент решил

Задание



Вспомните свою выпускную работу (бывшую или будущую – у кого уже есть текст)

- Может ли аннотация к работе получиться методом экстракции предложений?
- Можно ли получить заглавие Вашей работы из аннотации методом экстракции словосочетаний?