Автоматическое составление обзорного реферата на основе кластеризации предложений

Гнездилов Дмитрий, гр. 524

Научный руководитель
к.ф.-м.н., с.н.с. НИВЦ МГУ Лукашевич Н.В.

Автоматическое составление обзорного реферата

- Одна из важнейших практических задач автоматической обработки текста
- **Обзорный реферат** это совокупность предложений, позволяющих пользователю за небольшое время ознакомиться с основным содержанием тематически связанного набора документов
- К обзорному реферату предъявляются **требования**
  - По содержанию
  - По читабельности

Постановка задачи

- Разработка модели кластеризации предложений с учетом информации об отношениях между словами, описанной в тезаурусе РуТез

Тысячи людей в Мексике, Панаме, Колумбии с раннего утра занимали очередь в банк, чтобы как можно быстрее снять деньги со своих счетов.

Напуганные вкладчики из стран Латинской Америки выстаивают громадные очереди, чтобы снять свои вклады из банка.

- Проведение оценки качества кластеризации предложений
- Создание системы автоматического составления обзорного реферата на основе метода кластеризации предложений для обеспечения:
  - полноты покрытия содержания набора документов
  - снижения повторяемости информации в реферате.

Исходные данные

- Набор новостных статей
  - Статьи с единой темой
- Результаты графематического анализа
  - Размеченные предложения
- Результаты морфологического анализа
  - Слова, приведенные к начальной форме
- Выделенные концепты набора статей
  - Концепт – слово, определение которого найдено в тезаурусе
- Связи концептов в тезаурусе
  - *Мексика – Государство*
  - *Колумбия – страна Латинской Америки*

Описание метода кластеризации предложений - 1

- Определение меры близости для каждой пары предложений
  - Векторное представление предложения

$$sent = (tfids_{w_1}, tfids_{w_2} \dots tfids_{w_n})$$

$$tfids = \frac{n_i}{\sum_k n_k} \times \log \frac{|S|}{|S_w|}$$

n_i – число употреблений слова w в предложении

$\sum_k n_k$ – общее число слов в предложении

$|S|$ – общее число предложений

$|S_w|$ – предложения, в которых встречается слово w

  - Мера близости пары предложений

$$sim = \lambda \times sim_w + (1 - \lambda) \times (sim_c + sim_{rel})$$

$$sim_{w,c} = \frac{(sent1, sent2)}{|sent1|, |sent2|}$$

$\lambda \in [0,1]$ – параметр кластеризации

sim_{rel} – мера близости по отношениям концептов

Описание метода кластеризации предложений - 2

- Мера близости по отношениям концептов

Тысячи людей в Мексике, Панаме, Колумбии с раннего утра занимали очередь в банк, чтобы как можно быстрее снять деньги со своих счетов.

Напуганные вкладчики из стран Латинской Америки выстаивают громадные очереди, чтобы снять свои вклады из банка.

106106	ЛАТИНСКАЯ АМЕРИКА	104676	МЕКСИКА	1
566	ДЕНЬГИ	115362	ВКЛАД (ВЛОЖЕННЫЕ ДЕНЬГИ,ЦЕННОС	1
106106	ЛАТИНСКАЯ АМЕРИКА	105034	ЭКВАДОР	2
106106	ЛАТИНСКАЯ АМЕРИКА	104551	ПЕРУ	2
106106	ЛАТИНСКАЯ АМЕРИКА	104542	ПАНАМА	2
106106	ЛАТИНСКАЯ АМЕРИКА	103620	ВЕНЕСУЭЛА	2

$$sim_{rel} = penalty \times \frac{(sent1(c_1^1 \cdots c_n^1), sent2(c_1^2 \cdots c_n^2))}{|sent1| \times |sent2|}$$

$c_i^1 \in (sent1 \backslash sent2); c_i^1 \neq c_j^1, i \neq j$

$c_i^2 \in (sent2 \backslash sent1); c_i^2 \neq c_j^2, i \neq j$

$penalty$ – штраф, параметр кластеризации

$distance(c_i^1, c_i^2) \leq \max\_dist$

$\max\_dist$ – параметр кластеризации, максимальное расстояние между концептами

Описание метода кластеризации предложений - 3

- Описание алгоритма агломеративной кластеризации

  Каждое предложение – отдельный кластер,

1. Определение R_{max}

$$R_{max} = \max\left(sim\left(clust_i, clust_j\right)\right) = sim(U,V)$$

2. $threshold <= R_{max}$ - объединение U и V в один кластер N,
   иначе остановка кластеризации

$$threshold \in [0.1,1] - \text{порог кластеризации}$$

3. Пересчет расстояний от нового кластера до остальных кластеров

$$sim(N,M) = \frac{k \times sim(U,M) + l \times sim(V,M)}{k+l}$$

M – кластер, $M \neq U, M \neq V$

k – количество предложений в кластере U

l – количество предложений в кластере V

4. Переход на шаг 1

Оценка кластеризации

- Ручная кластеризация
- Попарное сравнение

Предложение	*Ручная кластеризация*	*Автоматическая кластеризация*
Тысячи людей в Мексике, Панаме, Колумбии с раннего утра занимали очередь в банк, чтобы как можно быстрее снять деньги со своих счетов	+	+
Напуганные вкладчики из стран Латинской Америки выстаивают громадные очереди, чтобы снять свои вклады из банка.		

- Вычисление F-меры

$$F_{measure} = \frac{2 \times R \times P}{(R + P)}$$

$$P = \frac{\{\text{количество правильно определенных пар}\}}{\{\text{общее количество определелленных пар}\}} - \text{точность}$$

$$R = \frac{\{\text{количество правильно определенных пар}\}}{\{\text{общее количество правильных пар}\}} - \text{полнота}$$

Составление аннотации

- Определение наиболее важных кластеров
  - Выбор кластеров с наибольшим количеством предложений
- Определение и извлечение центра кластера

$$sent^{i}_{cent} = \max_{sent \in clust_i} (\lambda \times |sent_w| + (1-\lambda) \times |sent_c|) - \text{центр } i-\text{го кластера}$$

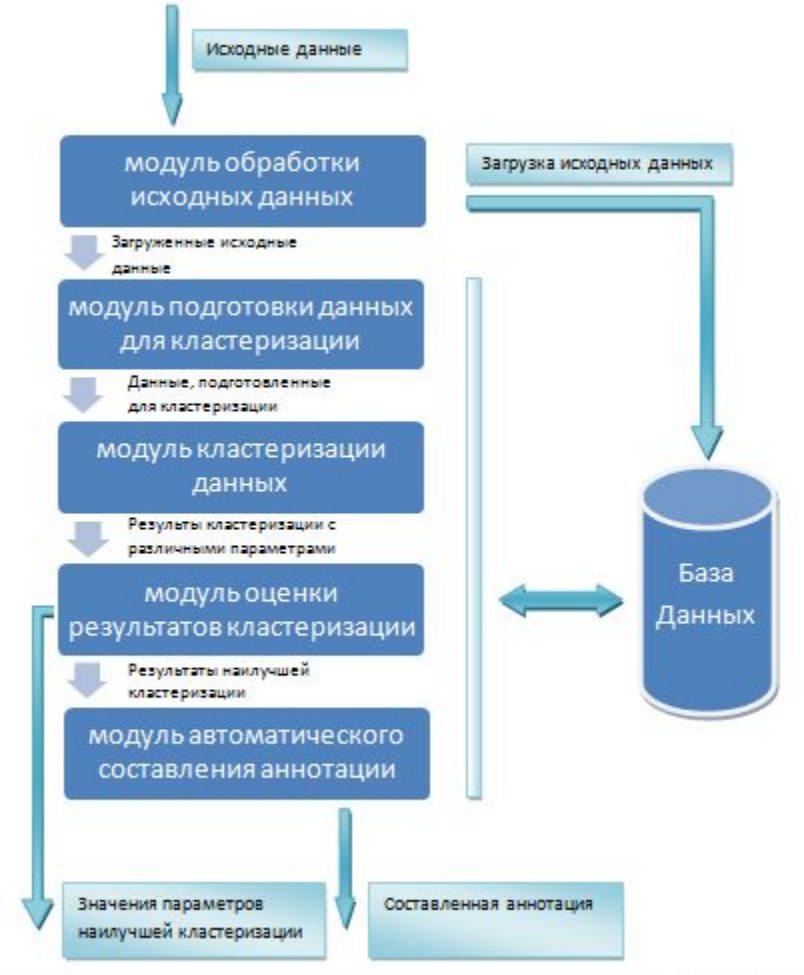
- Определение порядка выбранных предложений

$$timestamp_{sent^{i}_{cent}} = \min_{sent \in clust_i} (timestamp_{sent}) - \text{временная метка центра } i-\text{го кластера}$$

Программная реализация

Используемые инструментальные средства:

- СУБД
  - Microsoft Access
- Язык программирования
  - Microsoft Visual Basic



Описание эксперимента

- Дано 10 наборов по 30 новостных статей в каждом
- Необходимо вычислить значения параметров наилучшей кластеризации

λ – относительный вес слов и концептов
$threshold \in [0.1,1]$ – порог кластеризации
$penalty$ – штраф за сопоставление по отношениям
max_dist – параметр кластеризации, максимальное расстояние между концептами

- Составить аннотацию на основе полученных значений параметров

Результаты эксперимента

- Улучшение $F_{measure}$ *на 7% при точности* $P > 0.7$

max_dist	penalty	lambda	threshold	$F_{measure}$
5	1	0.8	0.3	0.297
0	0	1	0.3	0.277

- Пример составленной аннотации

Стэнфорд попытался арендовать частный самолет, однако из-за того, что его счета заморозили, компания-авиаперевозчик не приняла к оплате его кредитную карту.

Властям США неизвестно место нахождения миллиардера Аллена Стэнфорда, которого обвиняют в мошенничестве в крупных размерах.

Ассоциация крикета Англии и Уэльса отказалась от спонсорских отношений со Стэнфордом до окончания расследования.

В США техасский миллиардер обвиняется в мошенничестве на сумму около 8 млрд долл. По данным Комиссии по ценным бумагам и биржам США, в течение последних 15 лет принадлежащая миллиардеру компания Stanford Financial Group реализовывала мошенническую схему продажи ценных бумаг, суливших инвесторам получение высоких доходов.

Тем временем латиноамериканские издания отмечают, что паника началась в Мексике, Панаме, Колумбии, Эквадоре, Перу и некоторые филиалы (Эквадор и Перу) были вынуждены на неопределенное время приостановить свою работу.

Заключение

- В ходе выполнения дипломной работы:
  - Предложена модель кластеризации предложений с учетом тезаурусной информации
  - Реализована программная система, производящая кластеризацию предложений и составляющая обзорный реферат
  - Произведено тестирование созданной программной системы на различных наборах новостных статей
  - В ходе эксперимента были проанализированы и выбраны оптимальные параметры метода
  - Показано улучшение кластеризации предложений за счет тезаурусных знаний на 7%