Метод контекстно-зависимого аннотирования документов на основе спектральных оценок лексем

© Зябрев И.Н., Пожарков О.В.

AlterTrader Research Ltd.
info@altertrader.com

Аннотация

В статье описан метод контекстно-зависимого аннотирования, основанный на спектральных оценках лексем документов, учитывающих их внутренние частоты. Представлены алгоритм и результаты его обучения на основе экспертных оценок.

1. Введение

Контекстно-зависимое аннотирование документов является важной задачей лингвистического моделирования, и для ее решения существует множество методов. Основу большинства из них составляет отбор фрагментов исходного документа, максимально удовлетворяющих введенному критерию качества построения аннотации относительно запроса. Задачу контекстно-зависимого аннотирования можно представить в следующем виде.

Исходные данные:

- Запрос Q – множество лексем запроса q_i
- Документ D – множество лексем документа d_i
- Аннотация A – множество лексем аннотации a_i

Ограничения:

В данном случае вводятся ограничения на длину аннотации в 300 символов:

$$\sum_{A} length(a_i) \le 300 \quad (1),$$

где length(a_i) – длина лексемы или разделителя аннотации.

Требуется сформировать множество A, удовлетворяющее введенным ограничениям и отражающее содержание документа D относительно запроса Q.

2. Описание алгоритма

Нами задача получения множества A решалась в следующем виде:

Множество лексем документа разбивается на фрагменты F. Для этого используются следующие способы:

1. Метод скользящего окна с длиной, заданной в количестве предложений. $F_i=\{P_k\}$, где P - множество предложений документа D, k=i,i+1,..i+L-1, L - заданная длина скользящего окна. Т.е. каждое предложение является фрагментом, за исключением случаев, когда предложение не укладывается в ограничения (1). Подмножества F не пересекаются.
2. Метод скользящего окна с длиной, заданной в количестве слов. $F_i=\{d_k\}$, где k=i,i+1,..i+L-1, где L - заданная длина скользящего окна. Данный метод применяется в том случае, если предложение не укладывается в ограничение (1). Подмножества F в данном случае пересекаются.

На основе анализа критериев качества аннотирования были формализованы частотно-зависимые метрики ранжирования фрагментов, для вычисления которых используются характеристики, описанные в [1]:

Внутренняя частота леммы $IF(L,d)$ - число вхождений леммы L в документ d.

Условная частота леммы - число документов, удовлетворяющих заданным условиям. В контексте данной работы используются следующие условия:

1. Лемма L имеет встречаемость IF(L,d)=v:

   $CLF(L,v)=card(d|IF(L,d)=v)$, где card(d|A) - число документов коллекции, удовлетворяющих условию A, v – целое число.
2. Лемма L имеет встречаемость IF(L,d)≥v

   $CLF2(L, v)=card(d|IF(L,d)\ge v)$

   Абсолютная частота слова AF(L) – число вхождений леммы некоторого слова во все документы коллекции.

$AF(L) = \sum_{d=1}^{N} IF(L,d)$, где N – мощность множества документов коллекции.

Относительная условная частота леммы RCLF – отношение условной частоты леммы некоторого слова к его абсолютной частоте:

$$RCLF(L,v) = \frac{CLF(L,v)}{AF(L)}$$

Абсолютная документальная частота слова DF(L) – число документов, в которые лемма L входит не менее 1 раза

DF(L)=CLF2(L,1)

Обратная условная частота леммы:

$$ICLF(L,v) = \frac{DF(L)}{CLF(L,v)} \quad (2)$$

На основе (2) были сформулированы следующие оценки для ранжирования фрагментов:

1. Мера вхождения запроса во фрагмент

$$IFQ_i = \sum_{Q, q_j \notin F_i} ICLF(q_j, IF(q_j, D)) \quad (3),$$

где q_j- лемма j-го слова запроса Q.

2. Мера вхождения фрагмента в запрос

$IQF_i = \sum_{F_i, f_{i,j} \notin Q} ICLF(f_{i,j}, IF(f_{i,j}, D))$ (4), где $f_{i,j}$- лемма j-го слова фрагмента F_i

Так как сформировано две метрики фрагментов, то для критериального оценивания и последующего ранжирования необходимо ввести свертку. Будем использовать свертку вида:

$$FM_i = b_1 IFQ_i^{a_1} + b_2 IQF_i^{a_2} \quad (5),$$

где a_1, a_2, b_1, b_2 – коэффициенты, которые в дальнейшем определяются на основе экспертных оценок по обучающей выборке.

Согласно вычисленным оценкам производится ранжирование множества фрагментов по возрастанию FM_i, в результате чего получается множество RF. Из полученного списка путем последовательного выбора формируется аннотация:

$$A = \{RF_k \mid \sum_{k=1}^{K} length(RF_k) \le 300\} \quad (6)$$

Из коллекций документов KM.ru 2007 и BY.web 2007 была сформирована обучающая выборка, на основе которой в

дальнейшем для различных наборов коэффициентов формулы свертки (5) были построены аннотации. С целью сокращения объема обрабатываемых данных было сформировано ограниченное множество значений коэффициентов, каждое из подмножеств которого обладает определенными особенностями. Далее экспертами была оценена каждая из аннотаций, полученная на всем множестве параметров формулы свертки. Результаты оценивания сведены в таблицу (Таблица 1).

Номер набора	a1	b1	a2	b2	Средняя оценка экспертов			
					1	2	3	4
1	1	1	1	1	0,3	0,3	0,5	0,3
2	1	1	1	A	2,8	2,5	2,3	3,3
3	1	1	-1	1	0,2	0,2	0,3	0,2
4	1	1	-1	B	2,3	2,2	2,2	1,7
5	1	A	1	1	0,4	0,3	0,3	0,2
6	1	A	1	A	4,9	4,7	2,9	5,2
7	1	A	-1	1	0,2	0,3	0,3	0,2
8	1	A	-1	B	5	3,5	5,9	5,1
9	-1	1	1	1	5,4	3,8	6	6,7
10	**-1**	**1**	**1**	**A**	**9,1**	**7,1**	**7,2**	**8,1**
11	-1	1	-1	1	5,3	3,8	4,5	5,5
12	**-1**	**1**	**-1**	**B**	**9,7**	**9,2**	**8,4**	**9,8**
13	-1	B	1	1	4,3	3,6	2,9	2,8
14	-1	B	1	1	4,6	4,2	5,6	5,2
15	-1	B	-1	1	5	4,5	4,3	4,6
16	-1	B	-1	B	6,1	3,2	3,8	5,3

Коэффициент A=10^{-18}, коэффициент B=10^{-6}, коэффициенты A и B применяются для того, чтобы нормализовать значение соответствующего слагаемого в тех случаях, когда необходимо уменьшить его влияние на конечный результат.

Коэффициент согласия экспертов 0,92, т.е. довольно высокий. Причем максимальные оценки у наборов 10 и 12, каждый из которых имеет свои особенности ранжирования фрагментов документа. Так набор 10 отдает предпочтение фрагментам текста, максимально соответствующим запросу и при этом максимально развернутым. Набор 12 также дает наивысшие оценки фрагментам текста, максимально соответствующим запросу, но, в отличие от предыдущего набора, наиболее лаконичным. Причем на некоторых документах оба метода дают один и тот же результат аннотирования. Т.к. нельзя однозначно отдать предпочтение одному из наборов, было принято решение использовать оба по следующему алгоритму. Фрагменты документа для аннотации выбираются поочередно из ранжированных списков, полученных по каждому из наборов, начиная с 12 до тех пор, пока выполняются ограничения по длине аннотации. Если соответствующие элементы списков совпадают, то фрагмент выбирается однократно, чтобы не было дублирующихся предложений.

Таким образом, предлагаемый алгоритм контекстно-зависимого аннотирования можно разбить на следующие шаги:

1. Фрагментация исходного документа $D \longrightarrow F$.
2. Оценивание фрагментов документа по формулам $FM1_i = \frac{1}{IFQ_i} + \frac{10^{-6}}{IQF_i}$, $FM2_i = \frac{1}{IFQ_i} + 10^{-18} IQF_i$
3. Построение ранжированных списков по возрастанию критериев FM1 и FM2. $F \xrightarrow{FM1} RF1, F \xrightarrow{FM2} RF2$
4. Построение аннотации путем последовательного выбора элементов списков RF1,RF2 до тех пор, пока выполняется ограничение (1), исключая повторное включение фрагментов текста.

3. Заключение

Особенностями предложенного алгоритма являются использование частотно-зависимых оценок лексем [1] и использование двух методов ранжирования фрагментов документа, имеющих принципиальное отличие друг от друга, что дает возможность получить более содержательную аннотацию по сравнению с алгоритмом, использующим один метод. Стоит отметить, что параметры алгоритма ранжирования получены на небольшой обучающей выборке и основаны на мнении четверых экспертов. При формализации критериев качества построения аннотации появится возможность машинного обучения алгоритма на большой выборке документов и множестве значений параметров**,** не ограниченном представленными выше наборами. В частности для решения этой задачи можно использовать генетические оптимизационные алгоритмы.

Литература

[1] Зябрев И.Н, Пожарков О.В. Спектральное оценивание лексических единиц в задачах лингвистического моделирования. http://www.altertrader.com/publications16.html

Context-depended annotation method based on spectral estimations of lexemes

Zyabrev I.N., Pozharkov O.V.

The paper describes context-depended annotation method based on spectral estimations of documents lexemes. Also presents algorithm and results of its learning on basis expert estimation.