김일성종합대학학보
정보과학

주체109(2020)년
제66권 제1호

다중문서요약에서 클라스짓기에 의한 요약문장들의 순위화의 한가지 방법

정 만 흥

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《과학기술을 발전시키는것은 나라의 경제를 빨리 발전시키기 위한 중요한 담보입니다.》(《**김정일**선집》 증보판 제11권 133페지)

다중문서요약에서 어려운 과제의 하나는 요약문장선택후와 문장실현전의 정보순위화 과제이다.

단일문서요약에서의 순위화는 본문에서의 문장순서에 따라 진행할수 있으나 다중문서요약에서의 순위화는 본문에서의 문장순서에 따를수 없다.

최근의 선행연구[1, 2]들에서는 어휘적결합 혹은 실체중복과 변화를 고려한 국부결합관계에 초점을 두고있다. 그러나 대역적결합관계 즉 문장클라스들사이의 결합관계는 적게 고려되였다.

선행연구[1]에서는 요약문장들사이의 전후관계문장개념을 리용하여 요약문장들의 순위를 추리하기 위한 한가지 방법을 제기하였다. 여기서는 요약문장들의 관계그라프를 작성하고 그라프의 최량경로를 찾는 방법을 제기하였으나 이 문제가 전형적인 NP문제인것으로 하여 근사최량경로를 찾는 탐욕알고리듬을 리용하였다.

선행연구[2]에서는 요약문장들의 무리짓기에 토대한 클라스준위순위화와 클라스내에서의 문장준위순위화를 진행하는 방법을 제기하였다. 이 방법에서는 클라스와 클라스내에서의 문장순위를 결정하는데서 린접한 클라스들 또는 린접한 문장들사이의 련관성을 최대화하기 위한 탐욕알고리듬을 리용하였다.

론문에서는 무리짓기에 의한 순위화방법에서 첫순위클라스와 클라스내에서의 첫순위문장을 결정하는 새로운 방법에 토대한 요약문장순위화알고리듬을 제안하였다.

1. 순위화의 첫번째 클라스와 첫번째 문장의 선택

린접한 클라스들사이 또는 린접한 문장들사이의 련관성을 최대화하기 위한 선행한 탐욕알고리듬에서는 첫번째 순위의 클라스 또는 클라스내에서의 첫번째 순위의 문장을 선택하는데서 다른 모든 클라스 또는 다른 모든 문장들과의 류사성이 최대로 되는 클라스 또는 문장을 선택하였다.

① 첫번째 순위클라스의 선택

$$G_1 = \arg\max_{G} \sum_{G' \neq G} sim(G, G')$$

② 첫번째 순위문장의 선택

$$ss_1 = \arg\max \sum_{ss_i \neq ss_j} sim(ss_i, ss_j)$$

그러나 우와 같은 식에 의해 선택되는 클라스 또는 문장이 순위화에서 마직막클라스 또는 마지막문장이 될 가능성과 기타의 가능성도 동시에 주기때문에 이러한 방법으로 첫번째 클라스와 첫번째 문장의 순위를 결정하는 방법은 일반성을 가지지 못한다.

론문에서는 문서내에서 앞순위에 놓이는 문장들을 많이 포함하고있는 클라스가 순위화의 첫번째 클라스가 될 확률이 크며 또 주목하는 클라스에 속하는 문장들가운데서 앞순위를 가지는 클라스에 속하는 문장들과 류사도가 큰 문장이 고찰되는 클라스에서 첫번째 순위의 문장이 될 가능성이 크다는 가정하에서 첫번째 순위의 클라스와 클라스내에서의 첫번째 순위의 문장을 선택하는 방법을 제안하였다. 이리하여 요약문장들의 순위화에 대한 종전의 알고리듬을 개선하였다.

2. 개선된 요약문장순위화알고리듬

1) 클라스수준의 순위화

클라스수준의 순위화는 무리짓기의 결과로 얻어진 클라스들사이의 순위화로서 순위화의 대역적특징을 가진다.

알고리듬은 다음과 같다.

① 첫번째 순위클라스 G_1의 선택

첫번째 순위클라스는 원천문서들에서 앞선 순위에 위치하는 문장들을 가장 많이 포함하는 클라스일수록 첫번째 순위로 될 가능성이 크다는 가정에 기초한다.

클라스 G_i에 속하는 요약문장 s_j에 대하여 이 요약문장을 포함하는 원천문서 d_k에서의 문장순위 r_{ijk}를 계산하고

$$r_{ik} = \min\{r_{ijk} | \ j = 1, 2, \cdots, n_{ik}\}$$

로 놓는다. 여기서 n_{ik}는 클라스 G_i에 들어있는 요약문장의 개수이다.

클라스 G_i의 순위무게를 계산한다.

$$R_i = \sum_{s_j \in G_i} \frac{1}{r_{ik}}$$

첫번째 순위클라스 G_1을 결정한다.

$$G_1 = G_{i_0}, \ \ i_0 = \arg\max_i \{R_i\}$$

② i번째 순위의 클라스 G_i의 선택

$i-1$개의 클라스들이 $G_1, G_2, \cdots, G_{i-1}$과 같이 순위화되였다고 할 때 이미 순위화된 클라스들과의 류사성이 최대로 되는 클라스를 i번째 클라스로 결정한다.

$$G_i = \arg\max_G \sum_{j=1}^{i-1} sim(G_j, G), \ i > 1$$

여기서 G는 순위화되지 않은 클라스이다.

2) 문장수준의 순위화

문장수준의 순위화 역시 클라스수준의 순위화와 같은 원리에 따라 진행한다. 문장수준의 순위는 순위화의 국부적특징을 반영한다.

알고리듬은 다음과 같다.

$i = 1, 2, \cdots, K$(K는 클라스의 개수)

① i번째 클라스 G_i에 속하는 문장들이 모두 동일한 문서내의 문장들이라면 해당 문서에서의 본문문장순위에 따라 클라스 G_i안의 문장들을 순위화한다.

② i번째 클라스 G_i에 속하는 문장들이 여러 문서들에 분산되여 존재하면 다음과 같이 순위화를 진행한다.

ㄱ) 첫번째 클라스 G_1에서의 첫번째 문장 S_{11}의 선택

두번째 클라스 G_2에 속하는 모든 문장들과의 류사성이 최소로 되는 문장을 S_{11}로 결정한다.

$$S_{11} = \arg\min_{S \in G_1} \sum_{S' \in G_2} sim(S,\ S')$$

ㄴ) $i \neq 1$번째 클라스 G_i에서의 첫번째 문장 S_{i1}의 선택

클라스 G_{i-1}에 속하는 모든 문장들과의 류사성이 최대로 되는 문장을 S_{i1}로 결정한다.

$$S_{i1} = \arg\max_{S \in G_i} \sum_{S' \in G_{i-1}} sim(S,\ S')$$

여기서 $sim(S,\ S')$는 문장 S와 S'사이의 코시누스류사도이다.

ㄷ) p번째 문장 S_{ip}의 선택

i번째 클라스에서 이미 $p-1$개의 문장들이 $S_{i1}, S_{i2}, \cdots, S_{ip-1}$과 같이 순위화되였다고 할 때 이미 순위화된 문장들과의 류사성이 최대로 되는 문장을 찾고 그 문장을 p번째 문장으로 결정한다.

$$S_{ip} = \arg\max_{S} \sum_{j=1}^{p-1} sim(S_{ij},\ S),\ p > 1$$

여기서 S는 i번째 클라스 G_i에 속하는 문장으로서 아직 순위화되지 않은 문장이다.

3. 실험결과 및 분석

제기된 요약문장순위화알고리듬의 성능평가를 위해 성능평가척도로서 τ 거리척도와 AC거리척도[3]를 리용하였다.

τ 거리척도를 Kendall의 척도라고도 하는데 다음과 같이 계산된다.

$$\tau = 1 - \frac{4m}{N(N-1)}$$

여기서 N은 질문적합문장의 개수이며 m은 순위화된 질문적합문장렬을 참고순위의 문장렬로 변환하기 위해 린접한 문장들끼리 순서를 바꾸는 총회수이다.

τ의 값은 -1부터 1까지 변한다.

$\tau = 1$은 $m = 0$인 경우로서 질문적합문장들의 순서와 참고문장들의 순서가 일치하는 경우이다.

$\tau = -1$은 질문적합문장들의 순서와 참고문장들의 순서가 완전히 거꾸로 되는 경우로서 최대로 나쁜 경우이다.

그러므로 우연적인 순서는 보통 평균값으로서 $\tau = 0$인 경우이다.

AC(Average Continuity)거리척도를 평균련속성거리척도라고 부른다. 이 거리척도의 의미는 순위화의 정확도가 정확하게 순서화된 련속적인 문장들의 개수에 의해 평가된다는

데 있다.

AC거리의 계산식은 다음과 같다.

$$AC = \exp\left(\frac{1}{k-1}\sum_{n=2}^{k}\log(P_n+\alpha)\right)$$

여기서 k는 정확하게 순서화된 련속적인 문장들의 최대개수이며 α는

$$P_n = 0$$

일 때 로그함수가 값을 가지도록 정의되는 작은 상수값이다.

론문에서는 $\alpha = 0.01$로 하였다.

P_n은 련속문장의 길이 n의 비률로서 다음과 같이 계산된다.

$$P_n = \frac{m}{N-n+1}$$

여기서 m은 순위화된 질문적합문장들과 참고문장들에서 길이가 n인 련속문장의 개수이며 N은 문장의 총 개수이다.

순위화의 비교실험을 위해 요약문장의 개수 N의 각이한 값범위(5－12)에 따르는 평균값을 선택하였다.

비교실험의 결과를 표에 보여주었다.

표. 요약문장순위화의 비교실험

방법	τ거리	AC거리
Baseline	0.657 3	0.445 2
제안된 방법	0.728 6	0.568 8

표에서 보는것처럼 론문에서 제기한 무리짓기에 토대한 개선된 순서화알고리듬은 선행방법(Baseline)[2]에 비해 τ거리척도와 AC거리척도의 의미에서 효과적이라는것을 알수 있다.

맺 는 말

원천문서에서 앞선순위에 있는 문장들을 많이 포함하고있는 클라스가 첫번째 순위의 클라스로 될 확률이 크다는 가정과 앞선 클라스들에 속하는 문장들과의 류사도가 큰 문장일수록 해당한 클라스에서의 첫 순위의 문장이 될 가능성이 크다는 가정하에서 요약문장의 순위화를 진행하는 개선된 순위화알고리듬을 제기하고 그 효과성을 론증하였다.

참 고 문 헌

[1] D. Bollegala et al.; Information Processing & Management, 46, 89, 2010.

[2] A. S. Babar et al.; Procedia Comput. Sci., 46, 354, 2015.

[3] M. Lapata; Computational Linguistics, 32, 4, 1, 2006.

주체108(2019)년 11월 5일 원고접수

A Method of Summary Sentence Ordering by Clustering in Multi-document Summarization

Jong Man Hung

In this paper, we presented an improved method implementing summary sentence ordering through redetermining the first group and the first sentence of group-level ordering and sentence-level ordering in a greedy fashion of sentence ordering for summary generation.

Keywords: multi-document summarization, clustering, sentence ordering