김일성종합대학학보
정보과학
주체108(2019)년
제65권 제3호

# 사건기사모임에 대한 조선어다중문서요약실현의 한가지 방법

**정만흥, 김예화**

문서요약에는 문서 하나를 대상으로 하는 단일문서요약과 문서모임을 대상으로 하는 다중문서요약이 있으며 요약방법에 따라 추출요약과 추상요약으로 분류한다.

다중문서요약은 두가지 과제수행을 통하여 실현된다. 그중 하나의 과제는 다중문서모임에서 요약문장들을 추출하는 과제이며 다른 하나의 과제는 추출된 요약문장들을 순위화하는 과제이다.

요약문장들을 추출하는 과제에 대한 연구는 이미 많이 연구되였다.

선행연구[1, 4]에서는 문장 또는 문서들을 단어벡토르로 모형화하고 그것을 리용하여 요약문장들을 선택하며 다중문서요약의 중요한 과제의 하나인 문장순위화를 진행하였다.

최근에는 요약문장추출과 추출된 요약문장들을 순위화하는 과제가 동시에 연구되고 있다.

선행연구[2]에서는 혼합계층모형에 의한 다중문서요약을 고찰하였다. 여기서는 계층주제모형을 리용하여 잠재적인 특징에 기초한 문서무리짓기에서 문장들의 득점값을 계산하였다. 또한 선행연구[3]에서는 문장순위화의 전통적인 순위화기술에 무게를 고려하는 기술을 받아들여 시간사건무리짓기를 리용한 력사기사의 요약문제를 고찰하였다.

론문에서는 사건기사내용들에 대한 다중문서요약에서 고유실체의 중요성을 고려하여 고유실체에 대응하는 벡토르성분들을 반복시켜 만든 단어벡토르모형에 의해 다중문서요약을 실현하는 한가지 방법을 제안하였다.

## 1. 고유실체초점의 다중문서요약

우리는 다중문서요약을 다음의 절차에 따라 진행하였다.

첫째, 고유실체에 초점을 둔 문장에 대한 단어벡토르표현을 얻고 이 벡토르를 리용하여 문장들사이의 류사도를 계산하는것에 의해 문서모임의 문장들을 무리짓기한다.

둘째, 매개 무리들의 범위에서 문장들의 문장득점값을 계산하고 득점값의 크기순서로 일정한 개수의 요약문장들을 선택한다.

셋째, 무리준위의 대역적순위화를 진행한다.

넷째, 매개 무리들에 속하는 문장들에 대하여 문장준위의 국부적순위화를 진행한다.

다섯째, 무리준위순위화와 문장준위순위화결과를 종합하여 다중문서요약을 완료한다.

1) 문장의 벡토르표현과 무리짓기

단어사전 $D$는 앞부분과 뒤부분으로 나누어 설계한다.

사전의 앞부분에는 문서모임에 출현하는 명사(고유실체포함)들이 등록되며 뒤부분에는 고유실체단어들만이 재등록된다.

이때 단어사전 $D$를 리용하여 문장 $s$의 단어벡토르 $v_s$를 다음과 같이 정의한다.

$$v_s = (\delta_1, \delta_2, \cdots, \delta_n, \beta_1, \beta_2, \cdots, \beta_m)$$

여기서 $n$은 단어사전 $D$의 앞부분에 등록되여있는 단어들의 총수이며 $m$은 단어사전 $D$의 뒤부분에 등록되여있는 단어 즉 고유실체단어들의 개수이다. $\delta_i$와 $\beta_i$의 초기값들은 모두 0이다.

문장 $s$의 단어들을 차례로 조사하면서 조사되는 단어가 단어사전 $D$의 $i$째 단어와 일치한다면 $\delta_i = 1$로 재정의한다. 만일 조사되는 단어가 고유실체이면 마찬가지로 $\beta_i = 1$로 재정의한다.

문서모임에 속하는 문장들에 대한 무리짓기는 $k$-평균무리짓기의 2진분할법에 기초하여 진행한다.

무리짓기알고리듬은 다음과 같다.

걸음 1 무리짓기턱값 $\tau$를 입력하고 문서모임의 전체 문장을 포함하는 하나의 무리를 선택한다.

걸음 2 선택된 무리에서 문장의 류사도가 무리짓기턱값 $\tau$보다 작은 2개의 문장 $s_i$와 $s_j$를 임의로 선택한다. 즉

$$sim(v_{s_i}, v_{s_j}) < \tau$$

이다. 여기서 류사도함수 $sim$은 코시누스류사도이다.

걸음 3 문장 $s_i$와 $s_j$를 중심으로 하는 2개의 무리를 얻고 매 무리에 대한 오차를 계산한다. 무리에 속하는 매개 문장과 중심문장사이의 평균하밍거리로 매개 무리의 오차 $e$를 계산한다.

걸음 4 무리오차 $e$가 무리짓기턱값 $\tau$보다 큰 무리에 대하여 걸음 2로 이행하여 과정을 반복한다.

2) 무리에서의 요약

무리에 속하는 문장들가운데서 문장득점값이 높은 몇개의 문장들을 선택하여 요약문장모임을 얻는다.

문장 $s_i$의 득점값 $E(s_i)$를 다음과 같이 계산한다.

$$E(s_i) = \alpha \sum_j w_{ij} E(s_j) + \beta Q(s_i),\ 1 \le i \le N,\ \alpha,\ \beta \in [0, 1],\ \alpha + \beta = 1 \qquad (1)$$

여기서 $w_{ij}$는 문장들사이의 류사성을 특징짓는 류사도값이다.

$$w_{ij} = \begin{cases} sim(v_{s_i}, v_{s_j}), & s_i \ne s_j \\ 0, & s_i = s_j \end{cases}$$

그리고 $Q(s_i)$는 무리의 중심문장 $s_c$와 문장 $s_i$사이의 코시누스류사도이다.

$$Q(s_i) = sim(v_{s_c}, v_{s_i})$$

식 (1)을 행렬형태로 쓰면 다음과 같다.

$$(I-\alpha W)E=\beta Q \tag{2}$$

여기서 $I$는 $N$차단위행렬이며 $W$는 $w_{ij}$ 들을 성분으로 하는 $N$차행렬이다. 그리고 $E$와 $Q$는 각각 $E(s_i)$ 와 $Q(s_i)$ 들을 성분으로 하는 $N$차원벡토르이다. $N$은 문서모임의 문장의 총개수이다.

련립방정식 (2)의 결수행렬 $(I-\alpha W)$가 강한 대각선우세행렬로 되도록 파라메터 $\alpha$ 를 작게 선택한다. 이때 방정식 (2)를 가우스자이델법에 의해 풀수 있다.

다음 문장득점값이 크기순서로 무리의 전체 문장수에 비례하여 요약문장모임 $Cs_i(i=1, 2, \cdots, k)$ 를 얻는다. 여기서 $k$는 무리짓기의 결과로 얻어지는 무리의 개수이다.

3) 요약문장의 순위화

요약문장들에 대한 무리짓기를 수행한 후 문장들의 순위화를 다음의 두 단계를 거쳐 진행한다.

첫째 단계는 무리수준의 순위화이며 둘째 단계는 문장수준의 순위화이다. 즉 무리수준의 순위화는 무리들사이의 순위이며 문장수준의 순위화는 순위화된 무리들내에서의 문장들사이의 순위화이다.

(1) 무리수준의 순위화

무리수준의 순위화는 무리짓기의 결과로 얻어진 무리들사이의 순위화로서 대역적특징을 가진다.

그러므로 무리수준의 순위화를 대역적순위화라고 부른다.

알고리듬은 다음과 같다.

걸음 1 첫번째 순위무리 $G_1$의 선택

동일한 문서에 속하는 요약문장들에 문서자체에서의 문장순위에 따라 순위번호를 결정한다.

매개 무리에서 동일한 순위번호를 가진 문장들의 개수를 계수한다.

순위번호가 1인 문장개수가 제일 큰 클라스를 첫번째 순위무리로 결정한다. 개수가 같은 경우 순위번호 2를 론의한다.

걸음 2 $i$번째 순위의 무리 $G_i$의 선택

$i-1$개의 무리들이 $G_1, G_2, \cdots, G_{i-1}$과 같이 순위화되였다고 할 때 이미 순위화된 무리들과의 류사성이 최대로 되는 무리를 $i$번째 무리로 결정한다.

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

여기서 $G$는 순위화되지 않은 무리이다.

(2) 문장수준의 순위화

문장수준의 순위화 역시 무리수준의 순위화와 같은 원리에 따라 진행한다. 문장수준의 순서는 국부적특징을 반영하므로 문장수준의 순위화를 국부적순위화라고 부른다.

알고리듬은 다음과 같다.

$i=1, 2, \cdots, k$ 에 대하여 다음 걸음들을 수행한다. 여기서 $k$는 요약문장들을 무리짓기 하였을 때의 무리의 개수이다.

걸음 1 $i$째 무리 $G_i$에 속하는 문장들이 모두 동일한 문서내의 문장들이라면 해당 문서에서의 본문문장순위에 따라 무리 $G_i$안의 문장들을 순위화한다.

걸음 2 $i$째 무리 $G_i$에 속하는 문장들이 여러 문서들에 분산되여있으면 다음과 같이 순위화를 진행한다.

① $i=1$째 무리에서의 첫번째 순위문장 $s_{11}$의 선택

순위화의 두번째 무리 $G_2$에 속하는 모든 문장들과의 류사성이 최소로 되는 문장을 $s_{11}$로 결정한다.

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

② $i\neq 1$째 무리에서의 첫번째 순위문장 $s_{11}$의 선택

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

$$s_{i1}=\arg\min_{s\in G_i}\sum_{s'\in G_{i-1}} sim(s,\ s')$$

여기서 $sim(s,\ s')$는 문장 $s$와 $s'$사이의 코시누스류사도이다.

③ $p$번째 문장 $s_{ip}$의 선택

$p-1$개의 문장들이 $s_1, s_2, \cdots, s_{p-1}$과 같이 순위화되였다고 할 때 이미 순위화된 문장들과의 류사성이 최대로 되는 문장을 찾고 그 문장을 순위화의 $p$번째 문장으로 결정한다.

$$s_{i1}=\arg\max_{s}\sum_{j=1}^{p-1} sim(s_{ij},\ s'),\ \ p>1$$

여기서 $s$는 $i$번째 무리 $G_i$에 속하는 문장으로서 아직 순위화되지 않은 문장이다.

요약문장들의 순위화과정을 종합하면 다음과 같다.

첫째, 개선된 $k$-평균법에 의해 얻어진 요약문장들을 무리짓기하여 $k$개의 요약문장들의 무리를 얻는다.

둘째, 개선된 $k$-평균법에 의해 얻어진 문장들의 무리를 무리순위화알고리듬(대역적순위화알고리듬)을 리용하여 $k$개의 무리들을 순위화한다.

$$G_1, G_2, \cdots, G_k$$

셋째, 무리내에서의 문장들을 문장순위화알고리듬(국부적순위화알고리듬)을 리용하여 순위화한다. 이때 순위화된 문장렬은 다음과 같다.

$$s_{11}, s_{12}, \cdots, s_{1p1}, s_{21}, s_{22}, \cdots, s_{2p2}, \cdots, s_{k1}, s_{k2}, \cdots, s_{kpk}$$

여기서 $pi$는 $i$번째 무리에 속하는 문장의 개수이다.

## 2. 실험 및 결과분석

론문에서 우리는 실험자료로 조선어로 작성된 기사자료모임 KADS2000을 사용하였다. 자료모임의 기사의 개수는 127개이며 총문장의 개수는 1 583개이다.

자료모임에 들어있는 문장들에 대하여 이미 조선어형태부해석이 진행되여 명사단어와 고유실체표기가 되여있다.

문서모임의 무리짓기결과 무리의 개수가 75개정도가 되도록 무리짓기턱값 $\tau$를 설정

하였으며 매개 무리에 들어있는 문장의 17%에 해당한 문장들을 선택하여 요약문장들을 얻었다.

체계의 성능평가는 요약문서에 대한 성능평가와 요약문서에서 문장의 순위화정확도를 가지고 진행하였다.

성능비교실험은 OSVR방법과 NSVR방법을 리용하여 진행하였다. 여기서 OSVR방법은 문장의 단어벡토르를 다만 단어의 출현만을 고려한 $n$차원2진벡토르로 표현한 방법이며 NSVR방법은 론문에서 새롭게 제기한 고유실체에 초점을 둔 $(n+m)$차원2진벡토르로 문장의 단어벡토르를 표현하였을 때의 방법이다.

비교실험은 문장의 단어벡토르표현의 효과성을 검증하는데 기본을 두고 간단히 진행하였다.

다중문서요약결과에 대한 성능평가로는 정보검색체계에서 전통적으로 리용하는 척도로서 적중률, 완전률 그리고 $F$-척도들을 사용하였다.

적중률, 완전률 그리고 $F$-척도들은 다음과 같다.

① 적중률: $P=\dfrac{|S_s \cap S_h|}{|S_s|}$

② 완전률: $R=\dfrac{|S_s \cap S_h|}{|S_h|}$

③ $F$-척도: $F=\dfrac{(\beta^2+1)\times P\times R}{\beta^2\times P+R}$

여기서 $S_h$는 전문가에 의해 수동적으로 작성된 정답요약문장모임이며 $S_s$는 체계에 의해 추출된 요약문장모임이다.

간단히 하기 위해 우리는 체계에 의해 추출되는 요약문장의 개수와 전문가에 의해 수동으로 작성되는 요약문장의 개수를 일치시키였다. 이때 적중률과 완전률 그리고 $F$-척도는 모두 같으므로 $F$-척도에 대한 비교만을 진행한다.

요약문장들의 순위화평가를 위한 거리척도로는 $\tau$거리척도와 $AC$거리척도를 리용하였다.

$\tau$거리척도는 다음과 같이 계산된다.

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

여기서 웅근수 $N$은 요약문장의 개수이며 $m$은 순위화된 요약문장렬을 참고순위의 문장렬로 변환하기 위해 린접한 문장들끼리 순서를 바꾸는 총회수이다. 참고순위란 정답순위를 의미한다.

$\tau$의 값은 $-1$부터 1까지 변한다. $\tau=1$은 $m=0$인 경우로서 요약문장들의 순서와 참고문장들의 순서가 일치하는 경우이다. $-1$은 요약문장들의 순서와 참고문장들의 순서가 완전히 거꿀순서인 경우로서 최대로 나쁜 경우이다. 그러므로 우연적인 순서는 보통 평균값으로서 $\tau=0$인 경우이다.

$AC$거리척도는 평균련속성거리척도이다. 이 거리척도의 의미는 순위화의 정확도가 정확하게 순서화된 련속적인 문장들의 개수에 의해 평가된다는데 있다.

*AC*거리계산식은 다음과 같다.

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

여기서 $k$는 정확하게 순서화된 련속적인 문장들의 최대개수이며 $\alpha$는 $P_n=1$일 때 분모가 령이 되는 경우를 고려하여 첨부한 작은 상수값이다.(론문에서 $\alpha=0.01$)

$P_n$은 련속문장의 길이 $n$의 비률로서 다음과 같이 계산된다.

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

여기서 $m$은 순위화된 요약문장들과 참고문장들에서 길이가 $n$인 련속문장의 개수이며 $N$은 문장의 총개수이다.

표에서 보는것처럼 문장의 순위화와 요약문서추출알고리듬은 같다고 할지라도 알고리듬에서 리용되는 문장류사도평가기준이 다를 때 요약문서추출의 효률과 문장순위화의 정확도에서 차이난다는것을 보여준다. 이것은 문장류사도평가를 달리하면 문장무리짓기와 문장득점값계산, 문장의 전후관계에 따르는 순위가 달라진다는것을 의미한다.

표의 결과는 고유실체를 고려하여 단어벡토르를 리용하는 경우 고유실체를 리용하지 않는 경우에 비해 효과적이라는것을 보여준다.

표. 문서요약의 정확성과 순위화의 비교

| 방 법 | *F*-척도 | $\tau$거리 | *AC*거리 |
|---|---|---|---|
| OSVR | 0.610 9 | 0.315 2 | 0.120 7 |
| NSVR | 0.721 0 | 0.327 1 | 0.137 1 |

## 맺 는 말

우리는 고유실체에 초점을 둔 2진벡토르에 의해 문장의 단어표현을 얻고 그에 토대하여 요약문장의 선택과 순위화를 진행하는 방법으로 비용이 적게 드는 다중문서요약실현의 한가지 방법을 고찰하였다. 우리는 조선어로 작성된 기사자료모임 KADS2000에 대한 실험을 통해 론문에서 제기한 방법의 효과성을 확증하였다.

## 참 고 문 헌

[1] **김일성**종합대학학보(자연과학), 63, 4, 41, 주체106(2017).

[2] Asli Celikyilmaz, Dilek Hakkani-Tur; Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 815, 2010.

[3] James Gung, Jugal Kalita; Conference of the North American Chapter of the Association for Computational Linguistics, 631, 2012.

[4] Xiaojun Wan; Proceedings of the 23rd International Conference on Computational Linguistics, 1137, 2010.

주체108(2019)년 5월 5일 원고접수

# An Approach for Realizing Multi-Document Summarization for Event Article Collection

*Jong Man Hung, Kim Ye Hwa*

We suggested the way of getting term vector model of sentences by a binary vector focusing on named entities, and studied an approach of multi-document summarization that performed the selection and ranking of summary sentences based on it.

Key words: multi-document summarization, ranking of sentences