김일성종합대학학보
(자연과학)
주체103(2014)년 제60권 제12호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 60 No. 12 JUCHE103(2014).

질문중심요약에서 질문과의 련관도를 리용한 중요문장추출의 한가지 방법

김동수, 리명일

선행한 질문응답[1, 3]에서는 질문단어에 대하여 호상정보량과 단어사이의 거리에 의하여 그것과 련관된 단어들로 질문을 확장한 다음 확장된 질문단어들에 의하여 문장의 중요도를 계산하고 중요도에 따라 대답문장을 추출하고있다. 이러한 방법은 단어의 련관관계와 질문의 의미를 잘 반영하지 못하므로 질문중심요약의 완전률과 적중률을 높이지 못하고있다.

론문에서는 질문에 대한 열쇠단어들에 대하여 문서에서의 련관관계를 가지는 단어들로 질문을 확장한 다음 이 확장된 질문단어들과 온톨로지에서 련관관계를 가지는 용어들의 련관도[2]를 계산하여 련관도가 높은 용어들로 질문을 확장하고 확장된 질문단어들로 문장의 중요도를 계산하여 중요도에 따라 대답문장을 추출하는 방법을 제안하였다.

1. 질문확장방법

1) 문서에서 련관단어추출방법

조선어에서는 문장안의 단어들이 그 문장의 의미를 결정하는데 기여하므로 같은 문장에 나타나는 단어들은 어느 정도로 의미적류사성을 가지고있다. 그러므로 사용자가 입력하는 질문과 직접 련관이 있는 단어들은 질문단어가 있는 문장에서 추출하여야 한다.

질문단어와 단어들과의 련관은 문장안에서 형태부들사이의 의미적류사도로 평가하는데 이러한 의미적류사도는 련관성척도[1]로 표현할수 있다. 이 련관성척도는 질문단어와 문서에서의 단어들의 련관도이다. 즉

$$Lscore(w) = I(q,\ w) \times e^{-\alpha(d(q,\ w)-1)}$$

여기서 q는 질문단어, $I(q,\ w)$는 q와 w의 호상정보량, $d(q,\ w)$는 q와 w의 거리(형태단어개수), α는 거리의 영향을 조절하는 상수이다. 그리고 e는 거리가 먼 단어일수록 련관이 없다는것을 반영한다.

그러므로 질문관련단어모임은 다음과 같이 결정한다.

$$LW = \{w \mid Lscore(w) > T_1,\quad w \in W_T\}$$

여기서 T_1은 문서의 종류와 사용자의 의도에 따라 결정되는 상수이다.

2) 온톨로지에서 련관단어추출방법

일반적으로 온톨로지는 정보검색, 질문응답, 요약에서 질문을 확장하는데 리용된다.

온톨로지를 리용하여 질문확장을 진행할 때 질문의 열쇠단어들에 대하여 온톨로지에서 열쇠단어들의 상하위개념들과 측면관계를 가지는 개념들에 대한 련관도를 해석하고 련관도가 높은 개념들로 질문을 확장한다.

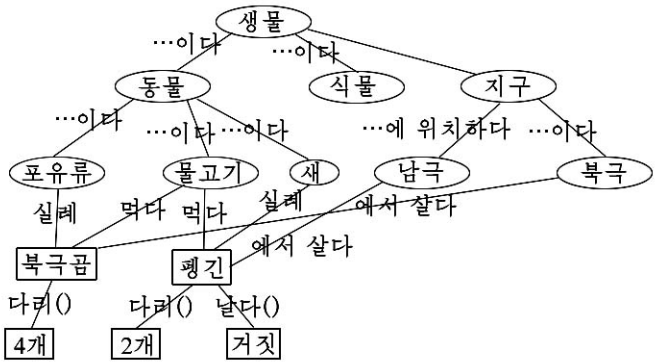

그림. 령역온톨로지의 실례

실례로 령역온톨로지에서 열쇠단어《북극곰》과의 련관도를 고려할 개념들은《포유류》,《물고기》,《북극》이다.(그림)

온톨로지에서 설정된 용어들사이의 련관도[2]는 기초련관도와 관계련관도를 고려하여 설정한것으로서 개념들의 의미적련관성을 잘 반영하고있다.

이제 용어 c_1, c_2사이의 련관도를 $S(c_1,\ c_2)$라고 하면 다음과 같다.

$$S(c_1,\ c_2)=\alpha S_I(c_1,\ c_2)+(1-\alpha)S_R(c_1,\ c_2)\quad \alpha\in[0,\ 1]$$

여기서 $S_I(c_1,\ c_2)$는 기초련관도, $S_R(c_1,\ c_2)$는 관계련관도이다.

한편 질문열쇠단어들의 모임을 Q, Q에 속하는 열쇠단어들을 $w_i(\overline{1,\ m})$이라고 하자. 그리고 온톨로지에서 질문열쇠단어들과 련관된 용어들을 t_{ij} $(i=\overline{1,\ m},\ j=\overline{1,\ n})$, 질문열쇠단어들과의 련관도를 S_{ij}라고 하면

$$S_{ij}=S(w_i,\ t_{ij})$$

이다.

여기로부터 질문련관단어들의 모임을 RW라고 하면 다음과 같이 결정된다.

$$RW=\left\{t_{ij}\mid S_{ij}>T_1,\ i=\overline{1,\ m},\ j=\overline{1,\ k}\right\}$$

여기서 T_1은 련관에 따라 설정되는 상수이다.

따라서 질문확장은 질문열쇠단어들과 온톨로지에서 련관도가 높은 단어들로 구성된다.

3) 질문확장

질문확장은 먼저 질문단어들과 문서에서의 련관단어들로 1차질문확장을 진행하고 다음 이 확장된 단어들과 온톨로지에서 련관도가 높은 단어들로 최종질문확장을 진행한다.

문서에서 련관단어모임 LW와 온톨로지에서 련관단어모임 RW로부터 질문확장단어모임 EW를 다음과 같이 결정한다.

문서에서 련관단어모임은 LW로 선택하고 이 LW와 질문단어모임 Q의 합모임으로 온톨로지입력단어모임 QLW를 결정한다. 즉

$$QLW=Q\bigcup LW=\{w_i\mid w_i\in Q(i=\overline{1,\ m}),\ w_i\in LW(i=\overline{m+1,\ l})\},\quad l=m+k$$

다음 QLW의 단어들과 온톨로지에서의 련관용어들을 t_{ij}라고 하면 EW는 다음과 같이 결정된다.

$$EW=\{t_{ij}\mid S_{ij}>T_1,\ i=\overline{1,\ l},\ j=\overline{1,\ n}\}\bigcup\{w_i\mid i=\overline{1,\ l}\}$$

2. 중요문장추출

질문에 대한 요약문장은 EW에 속하는 단어들을 많이 포함하면서도 질문과 밀접한 련관관계를 가지는 문장으로서 이것은 문장의 중요도(질문과의 련관도)로 평가할수 있다.

문장의 중요도계산을 위하여 EW에 속하는 단어들이 문장에서 차지하는 무게를 설정하고 문장에 포함된 EW의 단어들의 무게를 합하여 문장의 중요도를 결정한다.

단어들이 문장에서의 무게는 온톨로지에서 용어들과 질문열쇠단어들과의 련관도에 의하여 다음과 같이 결정한다.

질문열쇠단어들 $w_i(i=\overline{1,\ l})$의 무게를 Sw_i, 용어 $t_{ij}(i=\overline{1,\ l},\ j=\overline{1,\ n})$의 무게를 St_{ij}라고 하면 그것들은 다음과 같다.

$$Sw_i=\begin{cases}1 & i=\overline{1,\ m}\\ Lscore(w_i) & i=\overline{m+1,\ l}\end{cases}$$

$$St_{ij}=Sw_i\times S_{ij}$$

그러면 문장의 중요도는 다음과 같이 계산할수 있다.

$$Sscore(S_t)=r_1\sum_{w_i\in S_t\cap EW}Sw_i+r_2\sum_{t_{ij}\in S_t\cap EW}St_{ij},\quad t=1,\ 2,\ \cdots,\ s_n$$

여기서 r_1, r_2는 문헌과 온톨로지에서의 류사도를 고려하여 설정한 상수이다.

다음 턱값 T_2를 설정하여 질문에 대한 요약문장을 결정한다. 즉

$$AT=\left[S_t\mid Sscore(S_t)>T_2,\quad t=1,\ 2,\ \cdots,\ s_n\right]$$

3. 평 가 실 험

제안된 방법과 선행한 방법을 완전률(R)과 적중률(P)을 리용하여 평가할수 있는데 이것들을 리용한 종합적평가지표(F)는 다음과 같다.

$$F_\beta(R,\ P)=\frac{(\beta^2+1)PR}{\beta^2P+R}$$

여기서 β는 적중률과 완전률의 중요도를 조절하는 상수이다.

한편 적중률과 완전률의 중요도를 균등화하여 F를 다음과 같이 계산한다.

$$F(R,\ P)=\frac{2PR}{P+R},\quad(\beta=1)$$

이 값이 큰 체계가 더 좋은 체계로 평가된다.

비교실험은 호상정보량과 단어사이의 거리에 의한 중요문장추출방법과 론문에서 제안한 방법을 가지고 진행하였다. 비교실험결과는 표와 같다.

표에서 알수 있는바와 같이 제안된 방법은 질문에 대한 요약문장추출에서 종전의 방법에 비하여 1.2～1.5배의 개선을 가져왔다.

표. 실험결과 F의 비교

질문	선행한 방법	제안된 방법
1	0.780 3	0.835
2	0.691 1	0.859
3	0.732	0.841
4	0.727 5	0.889

참 고 문 헌

[1] 김일성종합대학학보(자연과학), 59, 2, 33, 주체102(2013).
[2] 김철범 등; 정보과학과 기술, 2, 11, 주체100(2011).
[3] M. Pasca; Computational Linguistics, 3, 1, 413, 2005.

주체103(2014)년 8월 5일 원고접수

A Method for Important Sentences Extraction using the Relation with Question in Query-Biased Summarization

Kim Tong Su, Ri Myong Il

The question is extended with the words, from the document containing the questioned words, related to the questioned key words. This extended question is once more extended with the words, from ontology, related to extended question words. And then metrics of importance is calculated with these extended question words and based on this metrics summarization is done.

Key words: summary, ontology, related word