서술형질문응답을 위한 문장득점값계산의 한가지 방법

정만흥, 김예화

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《프로그람을 개발하는데서 기본은 우리 식의 프로그람을 개발하는것입니다. 우리는 우리 식의 프로그람을 개발하는 방향으로 나가야 합니다.》(《**김정일**선집》 증보판 제21권 42페지)

자연언어형식으로 서술한 질문을 받아들여 서로 다른 구조의 많은 자료로부터 대답을 찾아내는 정보검색체계를 질문응답체계라고 부른다. 다시말하여 질문응답체계는 지식표현과 정보검색, 자연언어처리와 추론 등의 기술이 하나로 일체화된 정보검색체계[4]라고도 할수 있다.

질문응답체계에는 《백두산의 높이는 얼마인가?》의 질문에 2 750m와 같이 구체적인 사실자료에 대답을 주는 사실형(Factoid)질문응답체계, 《사과의 종류들을 렬거하시오.》와 같이 하나이상의 대답을 요구하는 목록형(List)질문응답체계, 《프락탈이란 무엇인가?》와 같이 어떤 개념의 정의에 대한 대답을 주는 정의형(Definition)질문응답체계 그리고 《참매는 어떤 새인가?》와 같은 보다 일반적인 질문에 대답을 주는 서술형(Description)질문응답체계 등이 있다. 사실형 및 목록형질문응답체계 그리고 정의형질문응답체계는 그 실현이 서술형질문응답체계에 비해 비교적 쉽다.

서술형질문응답은 질문중심문서요약[1－4]이라고 할수 있다.

론문에서는 질문응답후보문장들사이의 류사도에 따르는 수학적모형에 기초하여 서술형질문응답체계 즉 질문중심문서요약체계를 실현하는 방법에 대하여 고찰하였다.

1. 질문응답후보문장선택의 수학적모형

단순한 문서요약은 문서의 정보적내용을 가장 많이 담고있는 문장들을 추출하는데 있다. 질문에 대한 대답정보를 많이 포함할 가능성이 큰 문장을 질문응답후보문장이라고 부른다.

질문응답후보문장의 중요도평가를 위한 문장득점값계산의 수학적모형을 보기로 하자.

이를 위해 다음과 같은 가정을 한다.

높은 문장득점값을 가지는 문장들과 밀접한 류사성관계를 많이 가지는 문장일수록 그 문장이 주어진 문서에서 질문응답후보문장이 될 가능성이 크다.

문장들의 모임을 $S=\{s_i \mid 1\le i\le n\}$과 같이 표시하자.

문장 s_i의 득점값을 $E(s_i)$라고 할 때 우의 가정으로부터 문장득점값을 계산하는 다음과 같은 모형식을 고찰한다.

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

여기서 w_{ij}는 문장들사이의 류사성을 특징짓는 류사도값이며 $Q(s_i)$는 질문과 관계되는

문장 s_i의 특징값이다.

1) 문장들사이의 류사도값 w_{ij}의 계산

$$w_{ij}=\begin{cases} sim(s_i,\ s_j),\ i\neq j \text{ 인 경우} \\ 0, \quad \text{기타} \end{cases}$$

여기서 $sim(s_i,\ s_j)$는 문장 s_i와 s_j사이의 코시누스류사도이다.

이때 코시누스류사도 $sim(s_i,\ s_j)$의 계산에 참가하는 문장 s_i와 s_j에 대응하는 특징벡토르들은 각각 해당 문장들에서의 용어출현빈도수벡토르이다.

2) 특징값 $Q(s_i)$의 계산

문서속에 들어있는 단어들사이의 린접정도와 호상정보량을 리용하여 단어들사이의 련관도

$$A(w_1,\ w_2)=I(w_1,\ w_2)\cdot\exp(-\delta\rho(w_1,\ w_2))$$

를 계산한다. 여기서 δ는 구간 [0, 1]의 파라메터이고 $\rho(w_1,w_2)$는 동일한 문장속에 들어있는 두 단어 w_1과 w_2사이의 린접성정보를 나타내는 량으로서 두 단어사이에 놓이는 단어개수를 num이라고 할 때 다음과 같이 계산된다.

$$\rho(w_1,w_2)=\min\left\{num_k\mid k:\text{문장번호}\right\}+1$$

$I(w_1,\ w_2)$는 단어 w_1과 w_2에 대한 호상정보량이다.

$$I(w_1,\ w_2)=\log\frac{p(w_1,\ w_2)}{p(w_1)p(w_2)}$$

여기서 $p(w_1,\ w_2)$는 단어 w_1과 w_2가 동시출현하는 문장준위출현확률이며 $p(w_1)$과 $p(w_2)$는 각각 단어 w_1과 w_2의 출현확률이다.

질문용어 q가 들어있는 문장속의 단어 w들에 대하여 $A(q,w)$의 값이 작아지는 순서로 순위화하고 t개의 단어 w들을 앞순위의 순서로 선택한다. 이때 선택된 t개의 단어 w들을 질문용어 q의 1차련관단어라고 부른다. 마찬가지로 1차련관단어들의 모임 $\{w_1,w_2,\cdots,w_t\}$에 속하는 매개 단어들에 대한 1차련관단어들을 정의하며 이 단어들을 모두 질문용어 q의 2차련관단어라고 부른다.

질문용어 q의 1차련관단어 및 2차련관단어들로 이루어진 단어들의 모임을 질문련관단어모임으로 정의하고 질문련관단어모임에 의해 질문과 관계되는 문장의 특징값 $Q(s_i)$를 다음과 같이 계산한다.

첫째로, 질문련관단어모임에 속하는 련관단어 w_i와 w_j들의 쌍 $<w_i,w_j>$의 무게를 다음과 같이 결정한다.

$$<w_i,\ w_j>\text{의 무게}=\begin{cases} p_{ij},\ w_i\text{와 } w_j\text{가 } q\text{의 1차련관단어인 경우} \\ 0.7p_{ij},\ w_i\text{와 } w_j\text{중의 하나만이 } q\text{의 1차련관단어인 경우} \\ 0.3p_{ij},\ w_i\text{와 } w_j\text{가 } q\text{의 2차련관단어인 경우} \end{cases}$$

여기서 파라메터 p_{ij}는 단어 w_i와 w_j들이 동시에 출현하는 문장빈도수이다.

둘째로, 무게가 큰 련관단어쌍들을 많이 포함하는 문장일수록 중요한 문장이라고 보고 매 문장에 포함되여있는 가능한 련관단어쌍 $<w_1,w_2>$들의 무게합을 그 문장의 특징

값으로 설정한다.

$$Q(s_i) = \sum_{w_p, w_q \in s_i} < w_p, w_q > \text{의 무게}$$

셋째로, 문장특징값 $Q(s_i)$를 다음과 같이 정규화한다.

$$Q(s_i) = \frac{Q(s_i)}{\sum_{i=1}^{n} Q(s_i)}$$

2. 문장득점값 $E(s_i)$의 계산

문장득점값 $E(s_i)$를 계산하자면 식 (1)로 주어지는 모형식을 풀어야 한다. 이를 위해 우선 다음의 행렬 W, 벡토르 $\boldsymbol{u}$와 $\boldsymbol{p}$를 정의한다.

$$W = \alpha \begin{bmatrix} w_{11} w_{12} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} w_{n2} & \cdots & w_{nn} \end{bmatrix}, \quad \boldsymbol{u} = \begin{bmatrix} E(s_1) \\ \vdots \\ E(s_n) \end{bmatrix}, \quad \boldsymbol{p} = \beta \begin{bmatrix} Q(s_1) \\ \vdots \\ Q(s_n) \end{bmatrix}$$

이때 모형식 (1)을 벡토르－행렬형식으로 쓰면 다음과 같다.

$$(I - W)\boldsymbol{u} = \boldsymbol{p} \tag{2}$$

여기서 I는 단위행렬이다.

이제 련립방정식 (2)의 풀이가 유일존재하도록 하기 위해 행렬 W의 매 원소들에 감쇠인자 $\theta (0 < \theta < 1)$를 곱한다. 이와 같이 행렬 W를 변경시켜도 행렬이 담고있는 문장의 중요특징은 변화되지 않는다는것을 알수 있다.

최종적으로 다음의 련립방정식을 얻는다.

$$(I - \theta W)\boldsymbol{u} = \boldsymbol{p} \tag{3}$$

련립방정식 (3)의 결수행렬 $(I - \theta W)$가 강한 대각선우세행렬로 되도록 감쇠인자 θ를 선택한다. 감쇠인자 θ로는 실험적으로 구간 [0.4, 0.8]의 값을 취할 때 좋다는것을 확인하였다. 론문에서는 0.6으로 설정하였다.

가우스－자이델법을 리용하여 련립방정식 (3)을 푼다.

문장득점값 $E(s_i)$를 계산한 다음 문서를 구성하고있는 문장들가운데서 전체 문장의 20%정도 되는 문장들을 $E(s_i)$값이 큰 순서로 선택하여 요약문장으로 한다.

3. 실 험 분 석

사용자중심비부값행렬분해에 기초한 일반자동문서요약방법[2]과 제안방법을 적중률과 완전률, F-척도 등의 지표로써 비교하였다.

적중률, 완전률, F-척도는 다음과 같이 계산된다.

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

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

여기서 S_h는 전문가에 의해 작성된 요약문장모임, S_s는 체계가 출력시킨 요약문장모임, $|\cdot|$는 모임의 농도이다.

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

여기서 β는 적중률과 완전률의 중요도를 조절하는 파라메터상수이다. 론문에서는 β=3으로 설정하였다.

적중률 P, 완전률 R, F-척도의 실험값은 표와 같다.

표. 적중률 P, 완전률 R, F-척도의 실험값

방법	적중률 P	완전률 R	F-척도
선행방법[2]	0.57	0.62	0.593
제안방법	0.56	0.76	0.734

표에서 보는바와 같이 선행방법과 거의 같은 적중률을 보장하면서도 완전률을 1.23배 높였다. 이것은 론문에서 제안한 방법이 효과적이라는것을 말해준다.

맺 는 말

론문에서 제안한 방법을 사용자중심비부값행렬분해에 기초한 일반자동문서요약방법과 비교하여 효과성을 검증하였다.

참 고 문 헌

[1] 정만흥, 박련금; 정보기술통보, 1, 12, 주체103(2014).

[2] S. Park; International Conference on Computer Engineering and Applications IPCSIT, 2, 2011, 2009.

[3] H. Morita et al.; Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 223, 2011.

[4] M. Surdeanu et al.; Computational Linguistics, 37, 352, 2011.

주체107(2018)년 11월 5일 원고접수

A Method of the Sentence Score Value Calculation for the Descriptive Question and Answer

Jong Man Hung, Kim Ye Hwa

In this paper, we considered a method for calculating the sentence score value representing the importance of QA candidate sentence using the similarity characteristics between the sentences and the similarity characteristics between the question sentence and the sentences.

Key words: descriptive question and answer, document summary, sentence similarity