김일성종합대학학보
정보과학

주체110(2021)년
제67권 제3호

질문응답체계에서 통계적류사성과 품사적류사성, 의미적류사성을 고려한 질문－대답쌍사이의 류사성평가

리청한, 주혁위

최근시기 질문응답체계에서 질문－대답쌍자료기지를 리용한 질문응답방법[1]은 많은 사람들의 관심을 모으고있다.

질문－대답쌍자료기지를 리용한 질문응답방법에서 기본은 사용자의 질문과 질문－대답쌍자료기지에 있는 질문과의 류사성판정방법이다.

선행연구[2]에서는 격흐레임의 류사성에 의한 질문류사도계산방법과 의미속성의 류사성에 의한 문장류사도계산방법을 제기하였으나 정확도가 떨어지는 결함이 있다.

론문에서는 통계적류사성과 의미적류사성, 품사적류사성의 선형결합으로 사용자질문과 자료기지에 있는 질문사이 류사성을 계산하는 방법을 제안하였다.

1. 통계적류사성

통계적류사성은 자연언어질문과 질문－대답쌍자료기지의 문장사이에 같은 단어들이 어느 정도 존재하는가를 반영하는 류사성척도이다.

사용자질문과 질문－대답쌍자료기지안의 질문을 각각 Q_1, Q_2라고 할 때 단어모임

$$QS = Q_1 \cup Q_2$$

를 생각하면 이 단어모임은 두 질문에 들어있는 모든 단어이다.

이때 사용자의 질문과 자료기지의 질문은 벡토르 v로 표현할수 있으며 차원수는 단어모임 QS의 단어수와 같다.

이 벡토르의 매 성분은 단어모임 QS의 대응하는 단어로서 그 값은 다음과 같이 설정된다.

$$v_{ij} = \begin{cases} 0, & \text{단어가 질문문장에서 출현하지 않는 경우} \\ l, & \text{단어가 질문문장에서 출현하는 경우} \end{cases}$$

여기서 l은 질문문장에서 단어의 출현빈도수로서 정의 웅근수값이다.

실례로 다음의 2개 질문들에 대한 류사도계산은 다음과 같다.

Q_1: 장미의 색갈은 무엇인가?

Q_2: 황주사과의 색갈은 무엇인가?

두 질문문장에 대한 형태소해석을 진행한 이후의 단어모임은 다음과 같다.

$$QS = \{\text{장미, 색갈, 무엇, 황주, 사과}\}$$

즉 QS는 5차원벡토르이다.

이로부터 질문 Q_1의 벡토르 v_1과 질문 Q_2의 벡토르 v_2는 다음과 같다.

$$v_1 = \{1, 1, 1, 0, 0\},\ v_2 = \{0, 1, 1, 1, 1\}$$

이때 두 질문사이의 류사성은 다음과 같이 계산된다.

$$sim_{word} = \frac{v_1 \cdot v_2}{\|v_1\| \times \|v_2\|} = \frac{\sum_{i=1}^{k} v_{1i} \cdot v_{2i}}{\sqrt{\sum_{i=1}^{k} v_{1i}^2 \times \sum_{i=1}^{k} v_{2i}^2}} \quad (1)$$

여기서 k는 단어모임 QS의 단어수로서 벡토르의 차원수와 같다.

2. 의미적류사성

사용자질문과 자료기지질문들사이의 류사성을 정확히 계산하자면 질문을 이루는 단어들사이에 존재하는 의미적류사성을 계산하여야 한다.

실례로 질문《철이는 언제 학교로 갔습니까?》와《철이는 언제 대학으로 갔습니까?》에서 단어《학교》와《대학》은 서로 다르나 그것의 의미적정보는 류사하다.

론문에서는 단어사이의 의미적류사성계산을 위하여 의미지식기지인 조선어의미사전을 리용한다.

론문에서는 단어사이의 의미류사도를 다음과 같이 계산한다.

$$\begin{aligned} sim(W_1, W_2) = 2 \times DEPTH(LCS(W_1, W_2)) \times [Dis(W_1, LCS(W_1, W_2)) + \\ + Dis(W_2, LCS(W_1, W_2)) + 2 \times DEPTH(LCS(W_1, W_2))]^{-1} \end{aligned} \quad (2)$$

여기서 $LCS(W_1, W_2)$는 의미지식기지내에서 두 단어 W_1과 W_2의 최하위공통상위어이고 $DEPTH(LCS(W_1, W_2))$는 의미지식기지내에서 단어 W_1, W_2의 최하위공통상위어인 $LCS(W_1, W_2)$의 준위 즉 깊이, $Dis(W_1, LCS(W_1, W_2))$는 단어 W_1로부터 최하위공통상위어 $LCS(W_1, W_2)$까지의 거리, $Dis(W_2, LCS(W_1, W_2))$는 단어 W_2로부터 최하위공통상위어까지의 거리이다.

단어사이의 의미적류사성에 기초하여 두 질문문장들사이의 의미적류사성을 다음과 같이 정의한다.

$$sim_{semantic} = \frac{1}{2}\left(\frac{\sum_{a_i \in Q_1} \max\ ssim(a_i, Q_2)}{|Q_1|} + \frac{\sum_{b_j \in Q_2} \max\ ssim(b_j, Q_1)}{|Q_2|} \right) \quad (3)$$

여기서 $|Q_1|$과 $|Q_2|$는 두 질문 Q_1과 Q_2에서의 내용어들의 수를, a_i는 질문 Q_1의 i번째 단어를, b_j는 질문 Q_2의 j번째 단어를 의미한다.(단어는 명사를 의미) 그리고

$$\max\ ssim(a_i, Q_2) = \max(sim(a_i, b_1), sim(a_i, b_2), \cdots, sim(a_i, b_{|Q_2|}))$$

$$\max\ ssim(b_j, Q_1) = \max(sim(b_j, a_1), sim(b_j, a_2), \cdots, sim(b_j, a_{|Q_1|}))$$

이다.

의미적류사성의 값범위는 0～1이다.

두 문장사이의 의미적류사성을 다음의 실례를 통하여 보기로 한다.

질문 Q_1: 자료형 int는 표준자료형인가?

질문 Q_2: 클라스는 표준자료형인가?

질문 Q_1에 대한 단어는 자료형(a_1), int(a_2), 표준자료형(a_3)이다.

질문 Q_2에 대한 단어는 클라스(b_1), 표준자료형(b_2)이다.

표 1에 식 (2)에 의하여 계산된 두 단어사이의 의미적류사성을 보여주었다.

표 1. 두 단어사이의 의미적류사성

단어	자료형	Int	표준자료형
클라스	0.5	0.33	0.4
표준자료형	0.66	0.8	1

표 1로부터 $\max ssim(a_i, Q_2)$ 와 $\max ssim(b_j, Q_1)$ 을 계산하면 다음과 같다.

$$\max\ ssim(a_1,\ Q_2) = \max\{0.5,\ 0.66\} = 0.66$$
$$\max\ ssim(a_2,\ Q_2) = \max\{0.33,\ 0.8\} = 0.8$$
$$\max\ ssim(a_3,\ Q_2) = \max\{0.4,\ 1\} = 1$$
$$\max\ ssim(b_1,\ Q_1) = \max\{0.5,\ 0.33,\ 0.4\} = 0.5$$
$$\max\ ssim(b_2,\ Q_1) = \max\{0.66,\ 0.8,\ 1\} = 1$$

결국 두 질문문장사이의 류사도는 식 (3)에 의하여 0.785이다.

3. 품사적류사성

질문문장들에 대하여 형태소해석을 진행하여 품사정보를 얻으면 질문문장들을 품사들의 모임 즉 품사렬로 표시할수 있다.

대상으로 되는 두 질문문장들을 품사렬로 표시할 때 이 2개의 품사렬들사이에 존재하는 류사성을 품사적류사성이라고 부른다.

단어의 모호성은 형태론적으로 동음이의어적단어들에서 나타난다.

실례로 《높이》(명사), 《높이》(부사)는 음운구성은 동일하나 품사가 서로 다르면서 동음이의적관계를 가진다.

이 문제를 해결하기 위하여 론문에서는 형태부해석을 진행하여 단어에 대한 품사정보를 얻어 그것을 류사성평가에 리용한다.

품사적류사성은 다음과 같이 계산한다.

질문을 Q, i번째 형태부를 W_i 라고 하자.

그러면 이때 $Q = \{W_i \mid i = \overline{1,\ n}\}$ 이다.

이제 W_i 의 품사정보를 WF_i 라고 하면 품사적류사도는 다음과 같다.

$$sim_{pos} = \sum_{i=1}^{n} f_i(WF_1,\ WF_2) \tag{4}$$

여기서 $f_i(WF_1,\ WF_2)$ 는 품사의 류사도를 계산하는 값이다. 즉

$$f_i(WF_1,\ WF_2) = \begin{cases} 1, & \text{품사정보일치} \\ 0, & \text{기타} \end{cases}$$

이다.

품사적류사도는 질문문장의 문법적인 특성을 고려하자는데 있다.

론문에서는 두 질문사이의 전체적인 류사성을 통계적류사성과 의미적류사성, 품사적류사성사이의 선형결합으로 정의한다.

$$sim_{overall} = \alpha_1 sim_{word} + \alpha_2 sim_{semantic} + \alpha_3 sim_{pos}$$

여기서 α_1, α_2, α_3은 통계적류사성, 의미적류사성, 품사적류사성의 중요도를 나타내는 결수로서 이것들사이에는 $\alpha_1 \le \alpha_2 \le \alpha_3$이 존재한다.

EM알고리듬을 리용하여 계산한 결과 α_1, α_2, α_3값은

$$\alpha_1 = 0.23,\ \alpha_2 = 0.35,\ \alpha_3 = 0.42$$

일 때가 가장 좋은 결과를 주었다.

4. 실험 및 결과분석

성능비교를 위하여 선택한 방법들은 다음과 같다.

방법 1: 격흐레임의 류사성에 의한 질문류사도계산방법

방법 2: 의미속성의 류사성에 의한 문장류사도계산방법

표 2에서 보는바와 같이 론문에서 제안한 방법이 선행한 방법 1과 2보다 더 효과적이라는것을 알수 있다.

표 2. 질문류사도계산방법들의 성능비교

방 법	방법 1	방법 2	제안한 방법
정확도/%	78.3	80.6	88.3

맺 는 말

질문－대답쌍자료기지를 리용한 질문응답체계에서 사용자의 질문과 질문－대답쌍자료기지내의 질문과의 류사성을 통계적류사성과 의미적류사성, 품사적류사성사이의 선형결합으로 판정하는 방법을 새롭게 제안하고 론문에서 제안한 방법이 우월하다는것을 확증하였다.

참 고 문 헌

[1] M. Aqil et al.; Information Processing and Management, 54, 205, 2018.

[2] T. Mikolov et al.; Advances in Neural Information Processing Systems, 26, 1888, 2015.

주체110(2021)년 5월 5일 원고접수

Study on Evaluation of Similarity Between Query-Answer Considering the Statistical Similarity, POS Similarity and Semantic Similarity in QA System

Ri Chong Han, *Ju Hyok Wi*

In this paper, we newly propose the calculation of similarity between the user query and query in database by linear combination of the statistical similarity, semantic similarity and POS similarity.

Keywords: question answering system, statistical similarity, semantic similarity