김일성종합대학학보
(자연과학)
주체105(2016)년 제62권 제4호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 62 No. 4 JUCHE105(2016).

조선어질문응답체계에서 응답패턴과 의미정보의 결합에 의한 응답추출의 한가지 방법

김예화, 정만흥, 김순실

질문응답체계는 자연언어로 질문을 체계에 입력할 때 그에 대응하는 명백하고 정확한 응답을 결과로 출력하는 체계로서 자연언어처리분야에서 중요한 연구과제로 되고있다. 일반적으로 응답결과로는 추출된 응답후보문장들에 일정한 방법으로 득점값을 부여하고 그 득점값에 따라 순위화를 진행하여 득점값이 제일 높은것을 선택한다.

선행한 응답추출방법들[1－4]로는 단어의 출현빈도수에 기초한 방법, 질문용어들의 근접거리에 기초한 방법, 질문류형에 기초한 방법, 의존관계정보에 기초한 방법, 응답패턴에 기초한 방법 등을 들수 있다.

단어의 출현빈도수에 기초한 벡토르모형화방법과 거리에 기초한 방법은 속도가 빠르고 질문용어들과 응답후보문장들사이의 관계를 어느 정도 고려한것으로 하여 어순이 비교적 고정된 언어에서는 효과적이지만 조선어와 같이 어순이 다양한 언어들에서는 응답추출의 정확도가 크게 떨어지고있다.

의존관계정보에 기초한 방법[1]은 단어들사이의 의존관계를 고려한것으로 하여 응답추출의 정확도를 어느 정도 높이였지만 단어들사이의 의존관계를 서술하는데 많은 품을 요구한다.

응답패턴에 기초한 응답추출[5]에서는 언어의 구문적특징을 반영한 응답패턴을 리용하여 응답추출을 진행하지만 모든 구문을 반영한 패턴을 수동으로 작성하기 어렵고 학습을 통하여 패턴을 추출하는 경우에도 정확도가 떨어지는 결함을 가지고있다. 또한《누가, 언제, 어디》와 같이 사실적질문응답의 질문에 대하여 정확도는 비교적 높으나《무엇》형질문에 대하여서는 대답류형을 결정하기가 힘들기때문에 대답추출의 정확도가 떨어지는 결함이 있다.

론문에서는 지식망에 기초한 질문문장과 응답후보문장의 의미류사도계산방법과 응답패턴정합에 관한 류사도방법을 결합하여 응답추출을 진행하는 한가지 방법을 제안한다.

1. 지식망에 기초한 의미류사도계산

지식망(HowNet)은 언어지식기지로서 개념을 서술대상으로 하며 각종 개념과 개념사이의 관계, 개념이 가지고있는 속성과 속성사이의 관계를 기본내용으로 하는 지식기지이다.

개념은 사물현상의 일반적이며 본질적인 징표를 반영하는 사유형식으로서 여러개의 개별적사물현상들의 다양한 속성들가운데서 그 사물에만 있는 특수하거나 비본질적인 속성들을 버리고 일반적이며 본질적인 징표를 뽑아낸것이다. 실례로《아버지》라는 단어는《사

람, 남자, 어른》을, 《소년》이라는 단어는 《사람, 미성년》이라는 의미를 담고있다. 즉 개념은 단어의 의미를 구조적으로 서술한것으로서 지식표시언어를 리용하여 표현한다. 이 지식표시언어에서 리용하는 요소를 의미원이라고 부르는데 이러한 의미원은 개념의 최소의미단위이다.

1) 단어류사도계산

2개의 단어 w_1과 w_2가 각각 m개의 개념(의미원)과 n개의 개념을 가진다고 할 때 두 단어의 류사도는 매개 단어가 가지는 개념 혹은 의미원들사이의 류사도들의 최대값으로 한다. 즉 w_1과 w_2의 개념이 각각 s_{11}, s_{12}, …, s_{1m}, s_{21}, s_{22}, …, s_{2n}일 때 두 단어 w_1, w_2의 류사도는 다음과 같이 정의한다.

$$Sim(w_1,\ w_2) = \max_{i=1,\cdots,\ m,\ \ j=1,\cdots,\ n} Sim(s_{1j},\ s_{2j}) \tag{1}$$

결국 단어에 대한 류사도계산은 최종적으로 의미원의 류사도계산에 귀착된다.

한편 의미원모임은 의미원들사이의 상하위관계에 기초하여 나무형계층관계를 형성한다. 의미원나무에서 의미원 P_1과 P_2의 의미류사도는 다음과 같다.

$$Sim(P_1,\ P_2) = \frac{\alpha}{\alpha + d} \tag{2}$$

여기서 d는 의미원 P_1과 P_2의 의미원계층나무에서의 경로거리이고 α는 조절가능한 파라메터이다.

2) 문장의미류사도계산

언어학지식에서 알수 있는것처럼 임의의 문장은 여러개의 기본부분(주어, 술어, 보어)과 수식부분(규정어, 상황어)으로 나눌수 있다.

기본부분은 문장의 의미를 결정하는데서 기본역할을 하며 5종류의 품사(실사) 즉 명사, 동사, 형용사, 수사, 대명사로 구성된다. 그러므로 문장안의 실사들만을 추출하여 열쇠어로 하며 의미류사도계산에 참가시킨다.

문장 A와 B가 주어졌을 때 열쇠어배렬이 각각 $A = (A_1, A_2, \cdots, A_m)$, $B = (B_1, B_2, \cdots, B_n)$이라고 하자. 여기서 $A_i (1 \le i \le m)$와 $B_i (1 \le i \le n)$는 두 문장에 들어있는 열쇠어이다. 그러면 다음과 같은 행렬이 얻어진다.

$$M(A,\ B) = \begin{bmatrix} Sim(A_1, B_1), Sim(A_1, B_2), \cdots, Sim(A_1, B_n) \\ Sim(A_2, B_1), Sim(A_2, B_2), \cdots, Sim(A_2, B_{\mathrm{n}}) \\ \vdots \\ Sim(A_m, B_1), Sim(A_m, B_2), \cdots, Sim(A_m, B_{\mathrm{n}}) \end{bmatrix} \tag{3}$$

여기서 $Sim(A_i,\ B_i)$는 단어 A_i와 B_i의 류사도이다.

이때 문장 A와 B 사이의 의미류사도 $Sim(A,\ B)$는 다음과 같다.

$$Sim(A, B) = \sum_{i=1}^{m} \max(Sim(A_i, B_1), Sim(A_i, B_2), \cdots, Sim(A_i, B_n)) / m \tag{4}$$

마찬가지로 $Sim(B,\ A)$를 구할수 있다.

이렇게 얻은 두 문장의 평균의미류사도를 두 문장의 의미류사도로 정한다. 즉

$$Sim_1 = \frac{Sim(A,\ B) Sim(B,\ A)}{2}. \tag{5}$$

3) 응답패턴과 의미정보결합에 기초한 류사도계산

선행연구[2]에서 서술한바와 같이 응답패턴에 기초한 문장정합방법은 문법수준에서 응답후보와 리론적으로 정확한 응답사이의 류사도를 나타낸다.

열쇠어배렬에 기초한 문장의미류사도는 질문과 응답후보사이의 의미원들의 상관관계를 나타낸다. 그러므로 응답추출과정에서 문법정보와 의미정보의 두 측면의 정보를 고려하면 문법적 및 의미적정보를 고려한 응답후보를 추출할수 있다.

선형방법을 리용하여 이 두가지 정보를 결합하면 질문문장과 응답후보문장의 정합득점식은 다음과 같이 정의할수 있다.

$$Score(Q_i,\ A_j) = \lambda_s Sim_1 + \lambda_p Sim_2 \tag{6}$$

여기서 Sim_1 은 응답후보문장 A_j와 질문 Q_i의 의미류사도, Sim_2 는 응답후보문장 A_j와 질문 Q_i의 패턴정합류사도, λ_s, λ_p 는 각각 우의 두가지 특징의 무게이다.

2. 실험결과 및 성능평가

실험은 두 단계로 나누어 진행하였다.

먼저 선행연구[2]에서 제안한 응답패턴에 관한 문장정합방법을 리용하여 응답추출을 진행한 다음 여기에 의미정보를 결합하여 응답추출을 진행하였다.

응답추출부의 성능을 평가하기 위하여 창작가, 발명가, 지역, 날자에 대한 4가지 질문류형에 대하여 수동적으로 그 응답을 만들었다. 실험에서 선택한 질문은 1개의 질문요소만을 포함한다.

먼저 주어진 질문에 대하여 질문류형을 결정하고 응답패턴기지에서 질문류형에 대한 응답패턴을 선택하였다.

다음 패턴에서 열쇠어를 추출하여 검색엔진에 넣어 얻은 결과를 전처리한 다음 질문단어를 포함하는 모든 문장을 응답후보문장으로 보관하고 패턴정합정확도(정답확률)를 리용하여 점수가 높은 5개의 응답후보문장의 응답정합단어를 최종응답결과로 정하였다.

다음으로 식 (6)을 리용하여 응답후보문장들에 대한 득점식을 계산하고 점수가 제일 높은 5개의 문장을 취하였다.

실험은 응답순위거꿀평균(MRR: Mean Reciprocal Rank)을 응답추출의 성능평가를 위한 평가지표로 정하였는데 그 계산식은 다음과 같다.[2]

$$\mathrm{MRR} = \frac{1}{N}\left(\sum_{i=1}^{N}\frac{1}{r_i}\right) \tag{7}$$

여기서 N은 실험에 참가한 질문의 총개수이고 r_i는 i번째 질문에 대하여 얻어진 체계의 응답가운데서 정확한 응답이 놓이는 순위이다.

선행한 방법(응답패턴에 관한 문장정합방법[2])과 대비실험을 진행한 결과는 표와 같다.

표에서 P는 체계의 정확도(%)이며 다음의 식으로 정의한다.

$$P = \frac{n}{N}\times 100 \tag{8}$$

표. 대비실험결과

방법	MRR	$P(K=1)$/%	$P(K=5)$/%
선행한 방법	0.595	44.3	65.6
제안한 방법	0.6	44.9	66.0

여기서 N은 실험에 참가한 질문의 개수, n은 N개의 질문가운데서 정확한 응답이 얻어진 질문의 개수이다. 그리고 K는 응답후보의 개수이다.

표로부터 제안한 응답추출방법이 선행한 응답추출방법에 비하여 높은 성능개선을 가져왔다는것을 알수 있다.

참 고 문 헌

[1] 김일성종합대학학보(자연과학), 58, 10, 41, 주체101(2012).

[2] T. Mori; ACM Transactions on Asian Language Information Processing, 4, 3, 72, 2005.

[3] S. Sekine; ACM Transactions on Asian Language Information Processing, 4, 3, 35, 2005.

[4] C. Clarke; In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 43, 2003.

[5] 田卫东; 计算机工程与应用, 47, 13, 127, 2011.

주체104(2015)년 12월 5일 원고접수

A Method of Answer Extracting using the Combination of the Answer Pattern and Semantic Information in Korean Question-Answer System

Kim Ye Hwa, *Jong Man Hung* and *Kim Sun Sil*

Answer extraction is the key technology of the automatic question answering system. We proposed a method for extracting answer by combining sense similarity based on HowNet and similarity based on answer pattern matching.

The method of extracting answer proposed in this paper improved the performance than preceded method.

Key words: answer pattern, sentence matching, answer extraction, sense similarity