김일성종합대학학보
(자연과학)
주체103(2014)년 제60권 제9호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 60 No. 9 JUCHE103(2014).

의미범주와 문장정합을 리용한 응답추출의 한가지 방법

김예화, 리명일

질문응답체계에서 응답추출단계는 검색된 패쎄지속에서 이름을 가진 실체(Named Entity:NE)들을 식별하는 NE인식단계와 검출된 여러개의 NE들가운데서 정확한 응답을 얻어내는 응답평가단계로 구성된다.[2, 3] NE인식단계에서는 여러개의 응답후보(NE)들이 얻어지는데 그속에는 질문과 맞지 않는 응답후보들이 들어있을수 있다.

응답평가단계에서는 응답후보들속에서 정확한 응답을 찾기 위해 응답후보들에 대하여 일정한 방법으로 득점을 부여하고 그 득점에 따라 순위화를 진행하여 득점이 제일 높은 응답후보를 선택한다.

선행한 응답추출방법으로는 열쇠단어의 수에 기초한 방법, 거리에 기초한 방법, 류형에 기초한 방법, 의존관계정보에 기초한 방법 등을 들수 있다.[1－4] 선행연구[2]에서는 검색된 패쎄지속의 NE의 류형과 질문류형분류에서 얻어진 질문류형사이의 일치성에 기초하여 응답을 찾고있다. 이 방법은 속도가 빠르지만 검색된 문장들속에 질문류형과 일치하는 응답후보가 하나인 경우 정확한 응답을 얻을수 없다.

론문에서는 의미범주들사이의 류사도평가를 위하여 의미범주체계를 구축하고 의미범주에 기초한 문장정합득점방법을 제기한다.

1. 의미범주와 문장정합을 리용한 응답추출

1) 의미범주사전의 구성

질문응답에서는 보통 응답후보를 정확히 선택하기 위하여 매 단어에 의미범주정보를 부여하고 질문류형과 응답후보에 대한 류사성을 판정하기 위하여 의미범주들로 구성된 의미범주사전(혹은 개념계층사전) 즉 씨소라스(thesaurus)를 리용한다. 여기서 의미범주는 명사의 품사를 가지는 매 단어의 일반적특징을 반영하는 개념이다. 실례로《김철호》의 의미범주는 사람이다.

이러한 의미범주들에는 질문응답체계가 대상으로 하는 령역에 따라 일부 차이날수 있으나 일반적으로 사람이름(인명), 날자, 장소이름(지명), 조직이름(조직명), 수량 등이 속할수 있다.

론문에서는 질문에 대한 응답으로 될수 있는 NE들의 의미범주들을 질문응답의 특성에 맞게 크게 2개의 계층으로 구성하였다.

표 1에 의미범주의 일부를 보여주고있다. 두번째 층의 의미범주들은 첫번째 층의 의

미범주들의 하위범주들이다.

표 1. 의미범주사전의 일부

첫번째 층	두번째 층
동물(100)	새(101), 물고기(102), 포유동물(103), 파충류(105), 곤충(106) 등
식물(200)	풀(201), 나무(202) 등
장소(400)	주소(401), 건물(402), 도시(403) 등
날자(900)	일(901), 월(902), 년(903), 요일(904) 등
시간(1000)	시(1001), 분(1002), 초(1003) 등
수량(1100)	나이(1101), 길이(1102), 온도(1103), 무게(1104), 돈(1105) 등
조직(1200)	회사(1201), 학교(1202) 등

표 1에서 보는바와 같이 매 의미범주에는 류사성평가를 위하여 100이상의 자연수를 대응시킨다.

2) 의미범주에 기초한 질문류형의 일치성

질문류형은 사용자의 질문의도를 반영한 정보이며 질문해석단계에서 질문속의 의문사와 열쇠단어들에 의하여 결정된다.[3]

검색된 문장속의 응답후보가 질문류형과 같은 의미범주를 가진다면 그 응답후보는 정확한 응답이 될 가능성이 큰것으로 된다.

질문류형의 일치성은 응답후보의 의미범주와 의문사에 의하여 표현된 질문류형사이의 일치성을 나타낸다.

이와 같은 질문류형에 관한 정합득점 $St(AC,\ L_j,\ L_q)$는 다음과 같이 계산한다.

$$St(AC,\ L_j,\ L_q) = C_t \cdot \sum_{k_i \in SKW(AC,\ L_j,\ L_q)} w(k_i) \cdot sim(t_{AC},\ t_Q) \qquad (1)$$

여기서 C_t는 결수, t_{AC}는 응답후보 AC의 의미범주, t_Q는 질문류형이다. 그리고 $w(k)$는 열쇠단어 k에 대한 무게이고 $sim(t_{AC},\ t_Q)$는 t_{AC}와 t_Q의 류사도평가함수로서 0과 1사이의 값을 가진다. 즉

$$sim(t_{AC},\ t_Q) = \begin{cases} 1 & t_{AC} = t_Q \\ a & t_{AC}\text{와 } t_Q\text{의 상주범위가 같은 경우, } 0 < a < 1. \\ 0 & \text{기타 경우} \end{cases}$$

이때 다음의 식 (2)가 만족될 때 t_{AC}와 t_Q의 상위범주가 같다고 평가한다.

$$(t_{AC}\ \text{div}\ 100) = (t_Q\ \text{div}\ 100) \qquad (2)$$

여기서 연산자 div는 나누기연산의 상을 돌려준다. 실례로 $560\ \text{div} 100 = 5$이다.

한편 식 (1)에서 $SKW(AC,\ L_i,\ L_q)$는 다음과 같다.

$$SKW(AC,\ L_i,\ L_q) = (KW(L_i) \setminus \{AC\}) \cap KW(L_q)$$

여기서 $KW(L)$은 문장 L에 들어있는 열쇠단어들의 모임이다.

질문류형에 관한 정합득점계산실례는 그림과 같다.

그림에서 보는바와 같이 응답후보 2013년 6월 12일이 다른 응답후보들에 비하여 높은 정합득점을 가진다.

이것은 응답후보 2013년 6월 12일의 의미범주가 날자(900)로서 질문류형과 일치하기때문이다.

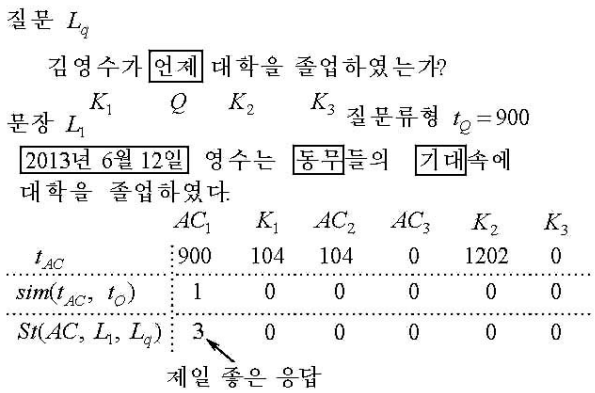

그림. 질문류형에 관한 정합득점계산실례

3) 문장정합에 의한 응답후보득점방법

질문류형의 일치성만으로는 정확한 응답을 기대할수 없으며 여러가지 방법들을 적절히 결합하여야 체계의 성능을 효률적으로 높일수 있다는 관점에서 응답후보 AC에 대한 득점부여식을 다음과 같이 준다.

$$S(AC,\ L_i,\ L_q)=Sk(AC,\ L_i,\ L_q)+Sd(AC,\ L_i,\ L_q)+St(AC,\ L_i,\ L_q) \qquad (3)$$

여기서 AC는 응답후보, L_q는 질문문장, L_i는 i번째로 검색된 문장, $S(AC,\ L_q,\ L_i)$는 L_q에 대한 검색문장 L_i속의 응답후보 AC의 정합득점이다.

식 (3)은 의문사 Q를 가지는 L_q에 대한 L_i속의 응답후보 AC에 대한 부분정합득점들의 선형결합이다.

$Sk(AC,\ L_i,\ L_q)$와 $Sd(AC,\ L_i,\ L_q)$는 L_q속의 의문사 Q의 문맥과 검색된 문장 L_i속의 응답후보 AC의 문맥사이의 류사성을 평가하기 위한 득점들이다.[1, 2]

따라서 $S(AC,\ L_i,\ L_q)$는 응답후보 AC를 의문사 Q라고 볼 때 질문문장 L_q와 L_i가 얼마나 잘 정합되는가를 나타낸다.

결국 정합득점 $S(AC,\ L_i,\ L_q)$가 클수록 응답후보 AC를 정확한 응답이라고 볼수 있다.

2. 실험 및 결과

응답추출부의 성능을 평가하기 위하여 도서《조선전사》1～20권을 준비하였다. 다음 우의 도서에 대하여 날자형질문 100개, 사람형질문 20개, 장소형질문 60개, 기타 여러가지 형의 질문 20개 총 200개의 질문과 그에 대한 정확한 응답을 준비하였다.

선행한 방법[1]과 제안한 방법에 대한 대비실험결과는 표 2와 같다.

표 2. 실험결과

구분	MRR	$P(K=1)$	$P(K=5)$
선행한 방법	0.516	43.5%	64.6%
제안한 방법	0.59	44.2%	65.5%

표에서 MRR는 가장 정확한 응답의 거꿀순위평균이다.[2]

그리고 P는 체계의 정확도(%)이며 다음의 식으로 정의한다.

$$P = \frac{n}{N} \times 100 \tag{4}$$

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

표로부터 새롭게 제안한 응답추출방법이 선행한 응답추출방법에 비하여 성능이 개선되였다는것을 알수 있다.

참 고 문 헌

[1] **김일성**종합대학학보(자연과학), 58, 10, 41, 주체101(2012).

[2] T. Mori; ACM Transactions on Asian Language 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.

주체103(2014)년 5월 5일 원고접수

A Method of Answer Extracting using the Semantic Category and Sentence Matching

Kim Ye Hwa, Ri Myong Il

We build a semantic category system to evaluate similarity between semantic categories and propose a method for scoring the sentence matching based on semantic category.

Through the experiment, we confirmed that our method is superior to pervious one.

Key words: semantic category, sentence matching, answer extracting