김일성종합대학학보
(자연과학)
주체105(2016)년 제62권 제3호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 62 No. 3 JUCHE105(2016).

문자렬패턴과 중요도에 의한 정의형질문응답 체계실현의 한가지 방법

김 동 수

정의형질문응답(Definitional QA)은 정의형질문《X는 무엇인가?》라는 정의형질문에 대한 대답으로서 사실적질문응답(Factoid QA)과 다른 방식이다.

사실적질문응답은 대답의 형이 명백하므로 사실적질문응답에 대한 대답을 찾는것은 어렵지 않다.

그러나 정의형질문응답은 기대되는 대답형을 명백히 암시하지 못하고 질문목표단어만을 제시하므로 정의형질문에 대한 정확한 대답을 찾는것은 어렵다.

선행한 연구들[2, 3]에서는 정의형질문에 의한 질문응답을 문장론적패턴들로 질문목표관련부분의 명사구와 동사구를 추출하여 대답후보로 하고 후보순위화평가에 의하여 우선순위가 높은 후보들로 대답을 결정하였다.

이 방법은 영어와 같은 언어에서 정의형문형이 명백한 패턴으로 주어지지 않는 언어에서는 적합하지만 조선어와 같은 언어에는 그대로 적용할수 없으므로 정의형질문에 대한 정확한 대답문장추출에는 적합하지 않다.

우리는 정의형질문문장의 질문목표단어로부터 질문확장을 실현하고 확장된 질문에 의하여 대답후보를 추출하며 정의형문자렬패턴에 정합되는 문장들로 대답을 추출하는 한가지 방법을 제안하였다.

1. 질문응답을 위한 질문확장

정의형질문응답에서 정의형질문은 기대되는 대답형을 명백하게 암시하지 못하지만 질문목표단어만은 명백히 암시하고있다. 그러므로 정의형질문응답에 의한 대답은 질문목표단어를 가지고 그에 대한 대답을 찾아야 한다.

질문목표단어는 질문문에서 고유실체인식(Named Entity Recognition)기술로 추출한다.

다음 질문목표단어를 가지고 정의형질문에 대한 대답문장들을 추출하기 위하여 질문목표단어와 관련한 단어들로 질문을 확장하여야 한다.

질문확장은 질문목표단어가 들어있는 문장에서 단어들의 호상정보량과 문장에서 형태부단어들사이 거리를 고려하여 진행한다.[1, 4]

질문목표단어를 q, 질문목표단어를 포함하고있는 문장을 S_q, $\forall w \in S_q$에 대하여 q와의 련관도를 $Rs(w)$라고 하면 련관도는 다음과 같이 표시된다.

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

여기서 $I(q, w)$는 q와 w사이의 호상정보량, $d(q, w)$는 q와 w사이의 형태단어거리, α는

거리의 영향을 조절하는 상수이다.

이때 $Rs(w)$에 의하여 결정되는 확장된 질문단어들의 모임 $EQ_1(q)$는 다음과 같이 결정된다.

$$EQ_1(q)=\{w\,|\,Rs(w)>T_1,\ \ w\in S_q\}$$

여기서 T_1은 문서의 종류에 따라 결정되는 상수이다.

다음 백과사전이나 코퍼스와 같은 외부자료들에서 동시출현빈도수에 의하여 질문목표단어와 련관성이 높은 단어들로 질문을 확장한다.

동시출현빈도수는 구문해석결과에 분리된 단순문안에서 질문목표단어와 단어들의 동시출현빈도수이다.

외부자료들에서 단어들의 모임을 EW, 단순문에서 q와 w의 동시출현빈도수를 $frq(q,\ w)$라고 하면 질문목표단어와의 련관도를 다음과 같이 계산한다.

$$Rt(w)=\log_2 frq(q,\ w)$$

이것은 문장의 중요도(질문목표단어와의 련관도)를 계산할 때 리용한다.

그러면 외부자료들을 리용한 질문확장은 다음과 같이 결정된다.

$$EQ_2(q)=\{w\,|\,frq(q,\ w)>T_2,\ \ w\in EW\}$$

여기서 T_2는 $\max\limits_{w} frq(q,\ w)-\gamma$로 결정되는 상수이다.

따라서 정의형질문의 질문목표단어에 의한 질문확장은 EQ_1과 EQ_2의 합모임으로 한다. 즉

$$EQ=EQ_1\cup EQ_2\,.$$

2. 정의형대답문장추출

질문목표단어에 의한 질문확장에 의하여 질문을 확장한 다음 확장된 단어들의 련관도에 의한 무게값에 의하여 정의형질문에 해당한 후보문장들을 추출한다.

정의형질문에 대한 대답후보문장들은 EQ에 속하는 단어들을 많이 포함하면서도 질문목표단어와 밀접한 련관관계를 가지는 문장이여야 한다.

이러한 문장을 추출하기 위하여 문서에 있는 매 문장의 중요도(질문과의 련관도)를 계산하고 중요도가 높은 문장들을 대답후보로 선정한다.

매 문장의 중요도는 문장속에 들어있는 EQ의 단어들이 질문목표단어와의 련관도값 $(Rs(w),\ Rt(w))$들을 합한값으로 설정한다.

그러면 문장의 중요도는 다음과 같이 계산한다.

$$\mathrm{Sscore}(S_i)=\lambda_1\sum_{w\in S_i\cap EQ_1}Rs(w)+\lambda_2\sum_{w\in S_i\cap EQ_2}Rt(w),\ \ i=1,\ 2,\ \cdots,\ n$$

여기서 λ_1과 λ_2는 문서와 외부원천과의 관계를 고려하여 설정한 상수이다.

다음 중요도에 따라 문장들을 순위화하고 중요도가 높은 문장들을 정의형질문대답문장들의 후보로 설정한다. 즉

$$\mathrm{CandS}(s)=\{S_i\,|\,\mathrm{Sscore}(S_i)>T_3,\ \ i=1,\ 2,\ \cdots,\ n\}\,.$$

얻어진 대답후보문장들가운데서 질문목표단어를 포함하면서 문형적으로 정의형문장으

로 되는 문장들과 함께 정의형문장은 아니지만 질문목표단어와 밀접히 련관되여있는 즉 추출된 정의형대답문장보다 중요도가 높은 문장들도 대답문장으로 추출하여야 한다. 그러면 정의형문장으로 될수 있는 문장의 언어적특징들이 추출되고 이 특징들을 리용하여 대답후보문장들가운데서 정의형질문에 대응한 정확한 대답문장을 선택하게 된다.

한편 조선어에서는 정의형문장들이 일정한 문자렬패턴들을 가지고있다.

실례로《과학기술출판사는 과학기술분야의 책들을 출판하는 기관이다.》라는 문장을 통하여 알수 있다. 이 실례에서 보는바와 같이 정의형질문의 대답을 찾는것은 매개 대답후보들이 질문목표단어에 관련되는 화제와 정의를 표현하면서 정의형문형으로 되는 문자렬패턴을 찾는 문제로 귀착된다.

조선어에서 정의형문장들의 문자렬패턴들은 다음과 같은것으로 분류할수 있다. 즉《질문목표단어》를 “KEYWORD”라고 하면 정의형문장의 문자렬패턴은

KEYWORD{는|은|란|이란}

~

{이다.|라고 한다.|이라고 한다.|라고 부른다.|이라고 부른다.}

로 표시된다.

선택된 대답후보들속에 우와 같은 문자렬패턴들을 포함하고있는 문장들은 정의형질문에 대한 대답문장으로 한다.

다음으로 대답후보문장들가운데서 추출된 정의형대답문장들보다 문장의 중요도가 더 높은 후보문장들도 정의형질문에 대한 대답문장으로 한다. 이것은 질문목표단어에 대한 내용을 충분히 설명하는 문장으로 되기때문이다.

이와 같이 정의형질문에 대한 질문응답체계에서는 질문목표단어와 밀접한 련관관계를 가지면서도 정의형문자렬패턴에 정합되는 문장들과 질문목표단어를 충분히 설명하는 문장들로 대답을 추출한다.

3. 실험 및 결과분석

제안된 방법과 선행한 방법과의 비교를 위해 완전률과 적중률에 대한 다음과 같은 종합적평가지표 F 를 리용하였다.

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

여기서 R는 완전률, P는 적중률, β는 적중률과 완전률의 중요도를 조절하는 상수이다.

한편 적중률과 완전률의 중요도를 균등화하면 F 를 다음과 같이 계산할수 있다.

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

비교실험은 선행한 방법과 론문에서 제안한 방법을 가지고 진행하였다. 비교실험결과는 표와 같다.

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

표. 실험결과 F 의 비교

질문	선행한 방법	제안된 방법
질문 1	0.780 3	0.825
질문 2	0.691 1	0.849
질문 3	0.732 0	0.811
질문 4	0.727 5	0.870

참 고 문 헌

[1] 김일성종합대학학보(자연과학), 59, 2, 33, 주체102(2013).

[2] Robert Navigli et al.; Proceeding of the 48th Annual Meeting of the Association for Computational Linguistics, 7, 1318, 2010.

[3] Hang Cui et al.; ACM Transactions on Information Systems, 25, 2, 8, 2007.

[4] M. Pasca; Computational Linguistics, 3, 1, 413, 2005.

주체104(2015)년 11월 5일 원고접수

A Method of Implementation of Definitional Question Answering by Character String Pattern and Importance Degree

Kim Tong Su

We study a method for definition question answering by importance degree of sentence and definitional string pattern related with the questioned words.

The paper describes about the method to realize the extension of question by the key word of definitional question sentence and extracts the candidate answer using extended question, and extracts the answer as the sentence which is adjusted to definitional string pattern.

Key words: definitional question answering, importance degree, string pattern