김일성종합대학학보
(자연과학)
주체105(2016)년 제62권 제3호

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

조선어질문응답체계에서 응답패턴을 리용한 응답추출의 한가지 방법

김예화, 정만흥

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

선행한 응답추출방법들[1－4]에는 단어의 출현빈도수에 기초한 방법, 질문용어들의 근접거리에 기초한 방법, 질문류형에 기초한 방법, 의존관계정보에 기초한 방법 등이 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 방법은 단어들사이의 의존관계를 고려한것으로 하여 응답추출의 정확도를 어느 정도 높이였[1]지만 단어들사이의 의존관계를 서술하는데는 많은 품을 요구한다.

문장은 일정한 문장론적규칙에 부합되는 단어들의 순서렬이다.[5] 이러한 문장의 구성상특징을 고려하여 론문에서는 질문용어와 응답후보문장들사이의 관계 즉 조선어문장의 구문적특징을 반영한 응답패턴들을 작성하고 그에 기초하여 응답추출을 진행하는 한가지 방법을 제안한다.

1. 응답패턴정합에 기초한 응답추출

《누가 시 〈백두산〉을 창작하였는가?》라는 질문에 대하여 만일 질문용어 《시, 백두산, 창작》으로 구성되는 벡토르모형을 리용할 때 다음과 같은 응답후보문장이 얻어졌다고 하자.

《시창작활동을 갓 시작한 영남이는 자기의 첫 사업으로 장편서사시 〈백두산〉에 대한 연구에 달라붙었다.》

이러한 경우 질문에 대한 응답의 결과로 《영남》을 출력할수 있다.

정확한 응답이 들어있는 응답후보문장 실례로 《위대한 수령님의 영광찬란한 혁명력사를 깊이 연구학습하는 과정에 그것을 예술적화폭으로 재현할것을 열렬히 소망하여온 조기천은 1947년 2월에 장편서사시 〈백두산〉을 창작하였다.》를 분석하면 이 문장은 질문용어들과 응답항목사이에 다음과 같은 문장구조를 가진다는것을 알수 있다.

…[조기천]은… 〈《백두산》〉을 창작하였다.

이 사실은 질문용어와 응답항목사이에 일정한 문장구조패턴(응답패턴)을 고려할 때 응답결과가 들어있는 정확한 후보문장을 찾을수 있다는것을 알수 있다.

1) 패턴학습과 평가

응답패턴은 응답후보가 들어있는 문장구성에 대한 형식화로서 다음과 같이 표시한다.

[A]+ [To]+<Qt>: P

여기서 [A]는 응답항목, [To]는 토, <Qt>는 질문용어, P는 패턴정합에 의해 후보로 선택된 응답의 정답확률이다.

매 질문류형에 대한 응답패턴과 정답확률을 얻기 위하여 질문류형의 훈련자료((Qt, A))에 대하여 다음과 같이 학습을 진행한다.

① 주어진 질문류형에 대하여 훈련자료기지로부터 1개의 쌍 (Qt, A)를 선택하여 탐색엔진에 넘겨 질문단어(Qt)와 응답항(A)을 포함하는 문장을 추출한다. 이때 귀환된 문장구조 접미사나무에 대하여 매 문장의 접미사렬 및 접미사렬의 길이를 얻을수 있다.

접미사나무의 렬을 려과하여 질문단어와 응답항을 포함하는 렬만을 남긴다.

② 질문단어와 응답항을 포함하는 접미사렬에 대하여 질문단어와 응답항목을 <Qt>와 [A]로 교체한다.

한편 동등한 질문류형의 매개 훈련견본에 대하여 우의 과정을 반복한다.

응답패턴학습에 의하여 추출된 응답패턴에 대하여 그 정확성을 완전히 믿을수 없으므로 이 패턴들에 대하여 정확성을 평가해야 한다.

응답패턴에 대한 평가는 다음의 방법으로 진행한다.

① 훈련자료중의 질문단어를 탐색엔진에 입력하여 질문단어를 포함하는 문장을 추출하여 응답후보모임 S를 구성한다.

② 질문류형에 따르는 응답패턴학습에서 얻은 응답패턴에 대하여 후보응답모임 S에서 정합을 진행한다.

③ 패턴정합에 참가한 총회수 *n*과 정합이 되는 정합추출회수 m을 구하여 정답확률 P=m/n을 얻는다.

2) 응답추출

우의 응답패턴학습과 응답패턴평가방법을 통하여 질문류형에 따르는 응답패턴 및 그것의 정답확률이 얻어지는데 이 패턴을 리용하여 구체적인 질문에 대한 응답을 추출한다.

그 과정을 보면 다음과 같다.

① 주어진 질문에 대하여 질문류형을 해석하고 질문단어를 식별한다.

실례로 질문《누가 시 〈백두산〉을 창작하였는가?》의 질문류형은《창작가(사람)》이다.

② 질문단어를 탐색엔진 혹은 응답본문자료기지에 입력한다.

③ 귀환된 페지의 본문을 추출하고 형태소해석, 품사결정, 구문해석을 비롯한 전처리를 진행하고 질문단어를 포함하는 모든 문장을 응답후보문장으로 선택하여 보관한다.

④ 후보문장의 질문단어를 <Qt>로 교체한다.

⑤ 질문류형에 대한 응답패턴을 리용하여 응답후보문장과의 정합을 진행한다. 그중 정합이 잘 진행된 후보문장에 대한 응답패턴의 정답확률을 리용하여 후보문장에 대한 평가분석을 진행한다.

⑥ 후보문장의 평가분석에 근거하여 순서화를 진행하고 앞의 5개 문장에서 응답패턴중의 [A]와 정합이 되는 응답항을 추출한다.

실례로 <창작가>형질문의 응답패턴과 그것의 응답패턴의 정합과정은 다음과 같다.

우선 응답패턴은 다음과 같다.

<Question Type=창작가(CREATOR)>

<Answer Pattern>

1.0: [A]{은, 는, 가, 이}<Qt>{을, 를}창작하다.

0.9 :<Qt>{은, 는, 가, 이}[A]{에 의하여, 에 의해}창작되다.

0.85: [A]{은, 는}<Qt>{이, 가}창작하다.

0.83: <Qt>의 창작은 [A]

0.75: <Qt>은 창작가[A]

0.5 : [A](<Qt>)

……

</Answer pattern>

우의 패턴을 리용하였을 때 정합결과는 다음과 같다.

질문: 장편서사시《백두산》을 누가 창작하였는가?

질문류형: 창작가

질문단어:《백두산》

귀환된 본문:

《위대한 수령님에 대한 열렬한 흠모의 정을 안고 수령님의 영광찬란한 혁명력사를 깊이 연구학습하는 과정에 그것을 예술적화폭으로 재현할것을 열렬히 소망하여온 조기천은 1947년 2월에 장편서사시 <백두산>을 창작하였다.》

정합된 패턴: 1.0: [A]{은, 는, 가, 이}<Qt>{을, 를}창작하다.

정합된 응답: 조기천

2. 실험결과 및 성능평가

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

여기에 기초하여 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 하였다.

마지막으로 패턴정합정확도(정답확률)를 리용하여 점수가 높은 5개의 응답후보문장의 응답정합단어를 최종응답결과로 정하였다.

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

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

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

선행한 방법(의존구조에 관한 문장정합방법)과 제안한 방법과의 대비실험결과는 표와 같다.

표. 대비실험결과

방법	MRR	$P(K=1)$/%	$P(K=5)$/%
선행한 방법	0.516	43.5	64.6
제안한 방법	0.595	44.3	65.6

표에서 P는 체계의 정확도(%)로서 다음의 식으로 정의한다.

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

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

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

참 고 문 헌

[1] **김일성**종합대학학보(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 11월 5일 원고접수

A Method of Answer Extraction using the Answer Pattern in Korean Question-Answer System

Kim Ye Hwa, Jong Man Hung

The answer extraction is the key technology of the automatic question answering system. In this paper we created the answer pattern for Korean question-answering system and proposed the method for extracting answer based on it.

The method for extracting answer based on answer pattern proposed in this paper improved the performance than previous method.

Key words: answer pattern, sentence matching, answer extraction