김일성종합대학학보
(자연과학)
주체104(2015)년 제61권 제11호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 61 No. 11 JUCHE104(2015).

질문응답체계에서 정규식을 리용한 질문자동생성의 한가지 방법

김 예 화

질문응답체계를 실현하는데서 체계의 정확도를 높이기 위하여 자주 제기되는 질문(FAQ: Frequently Asked Question)에 대하여서는 질문과 응답쌍을 자료기지에 등록하고 입력된 질문이 자료기지에 있는가를 판정한 다음 있으면 질문에 따르는 응답을 출력하고 없는 경우에만 패쎄지검색을 통하여 응답추출을 진행하는 방법을 리용하고있다.[3]

그런데 FAQ에 대하여 질문과 응답쌍을 만드는것자체가 품이 많이 들기때문에 이 작업을 자동적으로 처리하는 문제가 중요하게 제기된다.

선행연구들[1－3]에서는 NER(Named Entity Recognition)기술을 도입한 조건에서 질문을 자동생성하는 방법을 제기하였지만 이것은 영어문장을 대상으로 NER기술을 적용한 결과를 전적으로 리용하기만 하며 날자외의 수자렬들에 대하여서는 론의하지 않고있다.

론문에서는 질문자동생성에서 제기되는 문제의 하나인 수자렬을 포함하고있는 문장들에 대한 질문자동생성의 한가지 방식을 제기한다.

1. 정규식에 의한 수자렬들의 표현

어떤 패턴에 대한 탐색에 정규식을 리용하면 프로그람적으로 실현하기 매우 쉬워진다. 그러므로 기계번역을 비롯한 자연언어처리응용에 정규식을 리용하고있다.[1]

다음과 같은 문장들을 실례로 보자.

《제13차세계청년학생축전은 1989년 7월에 평양에서 진행되였다.》

《주체사상탑의 높이는 총 170m이고 그중 봉화의 높이는 20m이다.》

《개선문은 1982년 4월 15일 혁명의 수도 평양에 건립되였다.》

《1inch는 2.54cm이다.》

실례에서 보는바와 같이 문장속에 있는 수자렬들은 시기와 그 어떤 대상의 순서나 수량을 나타내는데 이때에는 반드시 수자렬들에 cm, m 등과 같이 단위가 붙는다.

시기인 경우에도 어떤 시기는 년까지만, 어떤 시기는 년, 월, 일까지 다 표현하고 어떤 시기들은 년 또는 월뒤에 정확한 날자가 아니라 《어느 날》, 《상순》과 같이 대략적인 날자로 표현된다. 이 모든것들을 통속적으로 표현하고 이 패턴들을 정확히 찾아 대응하는 질문문장으로 변환한다면 프로그람작성이 매우 간단하고 쉬워질것이다.

수자렬의 몇가지 형태를 보자.

우선 첫번째 형태는 다음과 같은것들을 들수 있다.

① 주체 xx(xxxx)년 x월 x일

② 주체 xx(xxxx)년 x월 {어느 날|상순|중순|하순|말}
③ 주체 xx(xxxx)년 {정초|봄|여름|가을|겨울} {어느 날}～
④ xxxx년 x월 x일
⑤ xxxx년 x월 {어느 날|상순|중순|하순|말}
⑥ xxxx년 {정초|봄|여름|가을|겨울} {어느 날}～
⑦ xxxx년대 {초엽|중엽|말엽}
여기서 ～은 생략될수도 있다는 표식이다.)

다음 두번째 형태는 다음과 같다.

① x{개월|달|km|cm|mm|inch|km^2|cm^2|mm^2|…}
② {한|두|석|넉|다섯|여섯|일곱|여덟|아홉|열|열한}달

여기서 첫번째 형태는 날자에 대한 표기이고 두번째 형태는 수량을 나타내는 표기로서 우리는 편리상 첫번째 형태를 1형, 두번째 형태를 2형으로 약속한다.

이때 이것을 정규식으로 표현하면 다음과 같다.

① 1형에 대한 정규식

(주체 ?[0－9]{1,}\([0－9]{4}\)|[0－9]{4})년대? ?([0－9]{1,}월|정초|말|여름|겨울) ([0－9]{1, 2}일|어느 날|상순|중순|하순|말) (어느 날)?

② 2형에 대한 정규식

([0－9]{1,}[.,만～년]? ?([0－9]{1,})? ?만?|한|두|석|넉|다섯|여섯|일곱|여덟|아홉|열|열한)(개월|달|mm|cm|km|inch|t|메터|m|m^2|km^2|cm^2|mm^2)

단위들은 문서의 특성에 따라 달라지거나 보충될수 있다.

2. 정규식을 리용한 질문자동생성

이제 탐색대상문장을 s, 정합되는 문자렬의 개수를 n, 패턴탐색결과 정합되는 문자렬을 w1. w2, …, wn 그리고 패턴의 형은 T(wi), 2형인 경우 수자렬뒤에 있는 단위를 ei라고 하자.

그러면 수자렬에 대한 정규식을 리용한 질문자동생성알고리듬은 다음과 같다.

① 문장에 대한 정규식을 탐색한다.

정규식클라스를 리용하여 수자렬이 들어있는 문자렬에 대한 정규식객체들을 정의하고 문장을 단위로 처리를 진행한다. 즉 문장을 선택한 다음 문장에 대한 정규식을 탐색한다.

② 탐색이 실패이면 ⑦로 간다.
③ n=1인 경우
　　T(w1)=1이면 w1→언제
　　T(w1)=2이면 w1=n1+e1(단위분리), n1→몇, w1'=n1+e1
　　⑤로 이행
④ ∀i에 대하여
　　T(wi)=1이면 wi→언제
　　T(wi)=2이면

ㄱ) $e_1 = e_2 = \cdots = e_n$이면 $\forall i$, $w_i = n_i + e_i$, n_i →몇, $w_i' = n_i + e_i$

ㄴ) $\forall i,j$ $e_i \neq e_j$ $(i \neq j)$이면 i>=2에 대하여 $w_i = n_i + e_i$, n_i →몇 , $w_i' = n_i + e_i$

$\forall i, j$에 대하여 $T(wi) \neq T(wi)$이면

i>=2일 때 T(wi)=1이면 wi→언제

T(wi)=2이면 wi→몇

⑤ s의 술어를 물음문으로 대응표를 리용하여 치환한다.

실례로 였다→였는가, 이다→인가로 치환한다.

⑥ 질문과 응답쌍을 등록한다. 즉 변환된 s를 질문으로, w_i는 응답문장으로 하여 질문응답쌍자료기지에 추가한다. 이때 i>=2인 경우는 w_i 들을 차례대로 구분기호를 삽입하면서 렬거한다.

⑦ 끝

알고리듬을 리용하여 우의 실례문장들에 대하여 자동질문을 작성하면 다음과 같이 얻어진다.

제13차세계청년학생축전은 언제 평양에서 진행되였는가?

답: 1989년 7월

주체사상탑의 높이는 총 몇m이고 그 중 봉화의 높이는 몇m인가?

답: 170, 20

개선문은 언제 혁명의 수도 평양에 건립되였는가?

답: 1982년 4월 15일

1inch는 몇cm인가?

답: 2.54

3. 성 능 평 가

사실적질문응답자료가 들어있는 300개 문장을 대상으로 실험한 결과 다음과 같은 결과를 얻었다.

300개 문장속에 날자와 수량을 표현하는 단어들이 들어있는 문장이 211개로서 70.3%이며 그중 정확하게 작성된 자동질문은 207로서 98%, 전체 문서의 69%이다.

이것은 사실적질문응답자료의 대부분이 날자와 수량이라는것을 의미하며 정규식만 가지고서도 자동질문작성의 많은 량을 처리할수 있다는것을 보여준다.

참 고 문 헌

[1] 윤명환 등; 콤퓨터와 프로그람기술, 2, 27, 주체100(2011).

[2] 백승진 등; 정보과학과 기술, 5, 11, 주체101(2012).

[3] Daniel Jurafsky et al.; Speech and Language Processing, Springer, 21～50, 2009.

주체104(2015)년 7월 5일 원고접수

A Method of Query Auto-Generating using the Regular Expression in the Question-Answer System

Kim Ye Hwa

We use question-answer pair about frequently asked question to increase the accuracy of the system for implement of question answering system. It is important to automatically process this work. In this paper we propose a method of automatic generation of question about sentences including sequence of digits using regular expressions.

Key words: regular expression, question auto-generation