김일성종합대학학보
(자연과학)
주체103(2014)년 제60권 제7호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 60 No. 7 JUCHE103(2014).

문서코퍼스로부터 자동생성된 질문문장에 의한 검색체계정확성개선의 한가지 방법

리남혁, 조성영

위대한 령도자 **김정일**동지께서는 다음과 같이 지적하시였다.

《우리는 정보기술, 나노기술, 생물공학을 발전시키는데 선차적으로 힘을 넣어야 하며 그 중에서도 정보기술 특히 프로그람기술을 빨리 발전시켜야 합니다.》(《**김정일**선집》 제22권 증보판 21페지)

현시기 과학기술의 급속한 발전으로 하여 정보량이 급격히 늘어나고있으며 따라서 대량의 정보검색체계에서 검색정확성을 개선하고 검색시간을 단축하는것이 중요한 문제로 나서고있다.

일반적으로 정보검색엔진은 사용자의 질문에 대하여 문서단위의 수많은 결과를 제시하므로 요구하는 정보를 찾기 위해서는 검색결과를 다시 검토해야 한다.

최근 이러한 문제점을 해결하기 위하여 정보검색기술과 자연언어로 된 질문문장에 대하여 문서보다 작은 단위로 정확한 대답을 제시하는 질문응답기술을 결합하기 위한 연구가 진행되고있다

질문응답체계에서 질문류형의 해석과 대답검색 등의 정확도를 높이기 위한 다양한 연구들이 진행[1－3]되였지만 아직까지 질문해석의 오유가능성을 완전히 없애지는 못하고있다.

이로부터 론문에서는 검색대상문서들의 모임(조선어문서코퍼스)을 지식기지로 간주하고 지식기지로부터 사용자가 질문할수 있는 질문과 정확한 대답을 자동적으로 미리 준비시키고 검색때에는 실마리어형식의 질문에 대하여 미리 대답을 가지고있는 질문을 검색하는 방식으로 문장과 문서수준에서 검색결과를 제시할수 있게 하는 한가지 방법을 제안한다.

1. 원천문장으로부터 질문문장의 생성방법

론문에서는 문서들의 모임으로부터 질문을 생성하기 위한 처리단위를 개별적인 문서들에 출현하는 문장으로 설정하고 이것을 원천문장이라고 부른다.

원천문장으로부터 질문문장을 자동적으로 생성하기 위하여서는 질문의 대상으로 될수 있는 요소들을 찾아야 하며 이를 위하여서는 자연언어처리수법을 리용하여 원천문장을 분석하여야 한다.

이때 생성된 질문문장의 정확성을 높이기 위하여 원천문장의 분석수준을 의미해석단계까지로 정하고 조선어원천문장의 분석을 격문법에 기초하여 다음과 같이 진행한다.

① 우선 형태부해석을 진행한다.

형태단어를 처리단위로 하여 형태부들로 분할하고 문법정보와 의미정보를 부여하며 문법적 및 의미적접속관계의 해석을 진행한다. 접속관계의 해석이 진행된 후 형태단어에 접속한 토를 제외한 나머지 형태부들을 결합하여 이것을 하나의 형태단어로 간주하며 그 단어의 문법 및 의미정보는 형태단어의 제일 마지막에 오는 자립적형태부(중심형태부)와 같은것으로 한다. 문법정보는 해당 형태부에 품사번호를 부여한다. 그리고 의미정보로서 체언형태부에는 의미분류에 기초한 의미표식이, 용언올림어에는 격틀이 부여된다.

실례로 원천문장이《1893년 12월 27일 평안북도 정주군 석산리에서 화전민의 아들로 출생한 계응상선생은 원산농업대학과 잠학연구소에서 유전학과 잠학을 연구하였다.》로 주어졌을 때 이것에 대한 형태부해석결과는 다음과 같다.

1893 년 12 월 27 일 평안북도 정주군 석산리 에서 화전민 의 아들 로 출생하 ㄴ 계응상 선생 은 원산 농업대학 과 잠학 연구소 에서 유전학 과 잠학 을 연구하 였 다.

② 다음 명사구해석을 진행한다.

형태부해석결과에 기초하여 명사구들을 갈라내고 병렬명사구해석, 명사구안에서 형태단어들사이의 수식관계에 대한 해석처리를 진행한다.

실례로 우의 실례문장에 대한 명사구해석결과는《1893년 12월 27일 평안북도 정주군 석산리에서 화전민의 아들로 출생한 계응상선생은 원산농업대학과 잠학연구소에서 유전학과 잠학을 연구하였다.》와 같다.

③ 원천문장이 확대문 또는 복합문인 경우 단일문들로 분할하고 수식관계(확대문의 경우)나 접속관계(복합문의 경우)를 해석한다.

실례로 우의 문장을 분할하면 다음과 같다.

계응상선생은 1893년 12월 27일 평안북도 정주군 석산리에서 화전민의 아들로 출생하였다.

계응상선생은 원산농업대학과 잠학연구소에서 유전학과 잠학을 연구하였다.

④ 단일문들에 대하여 격문법에 의한 의미해석을 진행한다.

실례로 주격(“계응상선생”, “사람/직업(111131)”), 시간격(“1893년 12월 27일”, “시간(1239)”), 장소격(“평안북도 정주군 석산리”, “지역(1122)”, 자격격(“화전민의 아들”,” 인간(11111)”), 시제격(“과거”,””)), 대격(“유전학과 잠학”, “학문(121111)”), 장소격(“원산농업대학과 잠학연구소”,” 시설(1121)”), 시제격(“과거”)으로 격을 가를수 있다.

의미해석결과에서 고유명사의 단어들은 밑줄을 그어 표시하였다.

문장해석이 끝나면 그 결과인 충족된 격틀에 기초하여 질문문장의 생성을 진행한다.

충족된 격틀에서 질문의 대상으로 될수 있는것은 격요소들이며 그것의 값(해당 격의 단어)의 의미표식은 질문의 구성요소로 될수 있는 대상의 류형을 판단하는데 리용하고 수식어는 대답에 대한 제약으로 리용한다. 우의 실례에서는 편리상 수식어를 그대로 두었다.

론문에서는 격요소로 될수 있는 대상(체언단어)들을 그것의 의미표식에 기초하여 6개의 대상류형 즉 PERSON, POSITION, LOCATION, TIME, ORGANIZATION, MISCEALLANY로 정하고 의미표식과 대상류형, 교체할 물음대명사들 사이의 대응표(표 1)를 작성하였다.

우의 실례에서 주격요소의 값《계응상선생》의 의미표식은《사람/직업(111131)》이므로 표 1로부터 대상류형은 PERSON으로 된다.

표 1. 의미표식과 대상류형, 물음대명사의 대응

의미표식(의미기호)	대상류형	물음대명사
인간(11111), 사람/직업(111131)	PERSON	누구
사람/지위(111132), 사람/역할(111133), …	POSITION	어디, 무엇
시설(1121), 지역(1122), 위치(12381), …	LOCATION	어디
시간(1239)	TIME	언제
기관(11121), 단체(11122), 국가(11125), …	ORGANIZATION	무엇
물건(113), 추상물(121), …	MISCELLANY	무엇

충족된 격틀과 표 1에 기초하여 원천문장으로부터 질문문장의 생성원칙은 다음과 같다.

① 원천문장에서 격요소의 값에 대응하는 형태단어(토는 제외)를 대답류형에 따르는 물음대명사로 교체한다.

② 형태단어의 수식어는 피수식어로 된 물음대명사에 대한 수식어로 한다.

③ 대격요소는 생성되는 질문문장에 반드시 포함시킨다.

이 원칙에 따라 생성된 질문문장이 특정한 대답을 얻을수 없거나 대답의 수가 너무 많으면 질문으로서의 의미가 없어진다.

질문문장은 또한 질문하는 대상을 명확하게 지시하여야 하며 그 대답의 수가 제한되여야 한다.

격요소의 값이 대명사이거나 보통명사이면 그 대상범위가 큰것으로 하여 질문에 대한 대답의 수가 너무 많아져 질문으로서의 효과성이 없어질수 있다.

한편 질문문장의 유효성은 격요소와 그것의 수식어부분에 존재하는 명사의 수에 따라서도 결정되며 명사의 수가 1개인 경우에는 의미없는 질문문장으로 될 가능성이 크다.

이로부터 론문에서는 다음의 조건(원천문장려과규칙)에 맞는 원천문장에 대해서만 질문문장을 생성한다.

① 격요소들에 대명사가 존재하지 말아야 한다.

② 교체할 격요소값(수식어도 포함)에는 고유명사가 1개이상 포함되여야 한다.

③ 격요소값과 그것의 수식어부분에 명사단어가 2개이상 존재하여야 한다.

려과규칙에 따라 지식기지로부터 선택된 원천문장으로부터 생성되는 질문문장은 주격형질문과 비주격형질문으로 나누어 생성한다.

주격형질문은 주격요소값으로 된 고유명사에 대해서 기타 격요소들에 포함된 고유명사를 각각 질문하는 질문문장을 의미한다. 이 방법으로는 질문문장을 최대로 주격을 제외한 격요소들에 존재하는 고유명사의 수만큼 생성할수 있다.

실례로《계응상선생은 언제 출생하였는가?》에는《1893년 12월 27일》,《계응상선생은 어디에서 출생하였는가?》에는《평안북도 정주군 석산리》,《계응상선생은 어디에서 유전학과 잠학을 연구하였는가?》에는《원산농업대학과 잠학연구소》라는 대답이 생성된다.

한편 비주격형질문은 비주격요소들을 그대로 놓고 주격요소의 고유명사를 질문하는 질문문장으로서 원천문장으로부터 1개만 생성된다.

실례로《누가 1893년 12월 27일 평안북도 정주군 석산리에서 화전민의 아들로 출생하였는가?》에도《계응상선생》,《누가 원산농업대학에서 유전학과 잠학을 연구하였는가?》에도《계응상선생》이라는 대답이 생성된다.

2. 질문생성에 기초한 검색체계의 구성

코퍼스로부터 질문문장의 생성 및 검색과정은 그림과 같다.

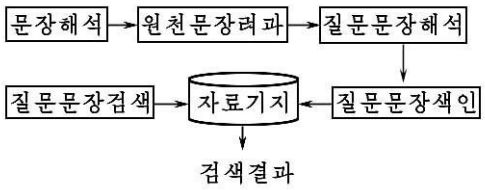

그림. 질문문장의 생성과 검색

① 문장해석

질문의 대상으로 되는 대상들을 찾기 위하여 격문법에 기초하여 해석을 진행한다.

② 원천문장려과

검색대상문서들의 모임으로부터 자동질문생성의 대상으로 되는 문장을 골라낸다.

③ 질문문장생성

주격형질문문장과 비주격형질문문장을 생성한다.

④ 질문문장색인

생성된 질문문장은 그것의 원천문장과 속한 문서정보와 함께 자료기지에 저장한다.

⑤ 질문문장검색

사용자가 입력한 열쇠어를 포함한 질문문장을 찾고 련관된 대답과 문서의 상세정보를 보여준다. 질문문장이 없는 경우 일반정보검색방법으로 문서들을 검색하여 보여준다.

맺 는 말

사용자가 질문할수 있는 질문과 대답을 자동적으로 생성하는 방법을 제기함으로써 검색체계의 정확성을 높이면서도 검색시간을 단축할수 있게 하였다.

참 고 문 헌

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

[2] 최정호; 언어정보처리, **김일성**종합대학출판사, 95～138, 주체96(2007).

[3] W. Salloum; Proceedings of the 2nd International Symposium on Knowledge Acquisition and Modeling, 383, 2009.

주체103(2014)년 3월 5일 원고접수

A Method for Improving the Precision of Retrieval System using Question Automatically Generated from Document Corpus

Ri Nam Hyok, Jo Song Yong

In general the information search engine retrieves many results in documents, so you have got to double-check the results to get your information.

In this paper we propose a method that generate the questions and correct answers from document knowledge automatically. Thus we improve the precision and speed of the retrieval system.

Key words: information retrieval, question answering