김일성종합대학학보
(자연과학)
주체106(2017)년 제63권 제9호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 63 No. 9 JUCHE106(2017).

조선어질문응답체계에서 질문분류에 기초한 격관계나무구축의 한가지 방법

최 명 옥

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《현대과학기술의 빠른 발전은 기초과학의 성과에 토대하고있으며 과학기술분야에서의 자립성은 기초과학분야에서부터 시작됩니다.》(《**김정일**선집》 증보판 제10권 485페지)

질문응답(QA)체계[4]는 사용자의 질문에 대한 응답으로 문서목록대신 대답을 찾는 정보검색체계로서 자연언어처리와 의미웨브[1]와 같은 첨단과학기술을 리용하는 응용분야이며 이에 대한 연구가 세계적범위에서 활발히 벌어지고있다.

질문분류[2, 3]는 제시된 질문이 속하는 클라스를 식별함으로써 질문으로부터 유익한 정보를 뽑아내는데 리용되는 기술이다. 선행연구[2]에서 단어망과 Wikipedia에 토대한 분류는 그자체가 인증되지 않은 여분어휘자료기지를 리용하였으며 선행연구[3]에서 제안한 함수토대분류기는 질문을 사실, 목록, 정의, 추리, 풀이, 항행과 같은 6개 범주들의 모임으로 제한한 마르꼬브론리망을 리용한다. 그러나 사용자질문의 분류와 질문응답을 위한 페지들의 색인에서 부족점이 있다.

우리는 열린 령역에서 일반목적QA체계를 개발하기 위하여 질문분류를 진행하고 그것을 리용하여 자원의 문서들을 색인화함으로써 조선어질문응답체계를 구성하였다.

1. 질문분류에 기초한 색인화방법

백과사전자료기지에 있는 문서들을 분석하여 질문클라스에 따라 색인화한다.

질문분류모듈에서는 입력으로서 사용자의 질문을 취하여 질문클라스를 식별하며 다음과 같은 분류처리를 진행한다.

걸음 1 질문에서 질문클라스를 식별한다.

걸음 2 질문클라스를 제외하고 질문의 남아있는 부분이 질의로 변환된다.

질문클라스를《누구, 무엇, 어디, 언제, 어느, 왜, 어떻게, 관계, 구성》으로 제한한다.

질문분류에 기초한 색인을 리용하여 용어들이 나타나는 문장ID와 단어ID를 추출한다.

2. 질문형색인구조를 리용한 격관계나무구축방법

조선어와 일본어를 비롯한 교착어들에서는 문장을 이루는 문장성분들의 의미적역할이 그 성분들에 부여된 교착물(토 및 조사)들에 의하여 일정하게 주어지기때문에 이 언어들에 대한 해석에서는 격문법을 리용하여 문장의 의미구조를 해석하기 위한 연구가 진행되고 있다.

우리는 자료기지의 문장들에 출현하는 격들을 전면적으로 분석종합하여 질문응답용으로 25개의 격종류를 추출하였다.

문장분석의 목적은 대상지식으로 되는 문장이 어떠한 질문형에 대답해주는 문장인가를 구조적으로 표현하는것이다. 질문형과 관계되는 격들을 모두 찾아 분할하면 그 문장이 어떤 질문형에 대답하는 문장인가를 나타낼수 있다.

관계나무구조는 문장으로부터 단어들의 렬로 구성되며 꼬리표들의 모임과 2개의 함수, 매 단어 및 문자렬과 함께 있는 꼬리표들로 이루어져있다.

관계나무구조에서 입력문장 S가 단순문장일 때 S는 술어(P)와 띄여쓰기를 단위로 한 단어들의 모임 $W_i(i=\overline{1,\ n})$로 분할된다. 이때 뒤방향해석에 의하여 단어들의 순서는 술어로부터 가까운 단어들의 순서이다. 즉 $S=|W_n|W_{n-1}|W_{n-2}|\cdots|W_1|P|$.

또한 문장 S가 복합문장일 때 형태부해석에 의하여 단순문장 $S=|S_m|S_{m-1}|\cdots|S_2|S_1$들로 구성된다.

일반적인 문장은

$$S=\|W_n^m|W_{n-1}^m|\cdots|W_1^m|P^m\|,\|W_n^{m-1}|W_{n-1}^{m-1}|\cdots|W_1^{m-1}|P^{m-1}\|,\cdots,\|W_n^1|W_{n-1}^1|\cdots|W_1^1|P^1\|$$

로 표기된다. 여기서 m은 단순문의 수이며 n은 단순문안에서 단어의 개수이다.

이때 단순문의 술어들과 단어들사이에 격관계가 존재하는데 그때의 격을 $C_i(i=\overline{1,\ 25})$로 표시한다.

$OPEN$, $CLOSED\in V$에 대하여 다음의 알고리듬을 론의하자.

걸음 1 $OPEN$에는 문장 S를, $CLOSED$에는 빈모임 ϕ를 각각 대응시킨다. 즉

$$OPEN\leftarrow S,\ \ CLOSED\leftarrow\phi.$$

$OPEN$모임에서 소문장들의 개수 k를 구한다.

걸음 2 다음의 순환고리를 수행한다.

① $OPEN\neq\phi$이면 $P_k\in OPEN$을 임의로 취하고 $OPEN\leftarrow OPEN\{P_k\}$로, $OPEN=\phi$이면 귀환하고 $k=k+1$로 설정한다.

② $P_k\in CLOSED$이면 걸음 2에로 이행하고 $P_k\notin CLOSED$이면 $CLOSED\leftarrow CLOSED\cup\{P_k\}$로 한다.

③ P_k가 구분불가능하게 되면 걸음 2에로 이행한다.

P_k가 구분가능하게 되면 S_k의 어떤 구분에 대하여 P_k로부터 단어들의 개수 l을 구하고 $i=l$이면 걸음 2에로 이행한다.

$W_{ki}\rightarrow P_k$ $(i=\overline{1,\ l})$인 W_i에로의 호를 긋고 그것을 C_{lki} $(l=\overline{1,25})$로 표시한다. 그리고 $i=i+1$로 하고 다시 ③으로 이행한다.

우의 알고리듬을 적용하여 얻어지는 호들가운데서 서로 이웃하고있는 호들의 모임을 U라고 하고 기호들로 련결된 대상들의 모임을 V_{node}라고 하면 이것들로 구성되는 나무 $G=(V_{\text{node}}, U)$를 격관계나무라고 한다.

3. 효과성평가

제안한 방법을 Swing을 리용하여 Java로 실현하였으며 백과사전자료령역의 5 000개 문서를 대상으로 질문응답체계의 효과성실험을 진행하였다.

다음의 공식을 리용하여 매개 대답의 관련성점수(*ARS*, %)를 계산하였다.

$$ARS = \frac{RF}{TF} \cdot 100$$

여기서 *RF*는 체계에 의하여 귀환되는 관련인자들의 수, *TF*는 관련인자들의 전체 수이다.

실험에 의하여 평균대답관련성점수는 85.98%로부터 95.77%사이에서 얻어졌다.

실험결과들은 이 체계가 류사한 기존체계들보다 더 좋은 성능을 가진다는것을 보여준다.

맺 는 말

제안한 방법은 질문을 적당한 질문클라스로 구분하고 백과사전령역의 광범한 자료들을 리용하여 질문분류에 기초한 문서색인화를 진행하므로 전통적인 분류방법과 차이난다. 실험결과들은 체계에 의하여 귀환된 대답들이 더 높은 관련성점수를 얻는다는것을 보여주었다.

참 고 문 헌

[1] Mariana Damova et al.; Natural Language Interaction with Semantic Web Knowledge Bases and LOD, Chalmers University and GU, 87～91, 2014.

[2] Joseph Chang et al.; 21st Conference on Computational Linguistics and Speech Processing, 145, 2009.

[3] Fan Bu et al.; In Proceedings of the 2010 Conference on Empirical Methods on Natural Language Processing, 1119, 2010.

[4] Renu Mudgal et al.; International Journal of Computer Engineering & Applications, **2**, **2**, 27, 2013.

주체106(2017)년 5월 5일 원고접수

A Method building up the Case Relation Tree based on Question Classification in Korean Question Answering System

Choe Myong Ok

This paper presented a construction method of the case relation tree appending indices obtained by question classification in Korean question answering system and proved its effectiveness through the experiment.

Key words: question classification, indexing, case relation tree