김일성종합대학학보
정보과학

주체109(2020)년
제66권 제3호

질문응답에서 질문용어간 거리에 의한 근접함수결정의 한가지 방법

최명옥, 동승철

경애하는 최고령도자 **김정은**동지께서는 다음과 같이 말씀하시였다.

《첨단돌파전은 현대과학기술의 명맥을 확고히 틀어쥐고 과학기술의 모든 분야에서 세계를 앞서나가기 위한 사상전, 두뇌전입니다.》

질문응답(Question Answering)[1, 2]은 자연언어로 표현된 사용자질문에 대한 특정한 대답을 찾아내고 추출하여 현시하는 기술이다.

선행연구[1]에서는 용어근접이 용어들사이의 의존성에 강한 영향을 미친다고 지적하고 1에서 용어들사이의 거리를 더는 방법으로 근접도를 계산하여 검색정보를 순위화하는 방법을 제안하였다.

선행연구에서는 질문용어들의 발생빈도수와 발생위치, 발생순서를 고려하였으나 질문용어들의 근접정도에 대해서는 엄밀하게 론의하지 못한 결함이 있다.

론문에서는 질문응답에서 문장론적류사도의 한가지 방법으로 질문용어간 거리를 정의하고 그에 의한 근접함수를 결정하는 방법을 제안하였다.

1. 질문용어간 거리의 정의

류사도는 언어학, 인공지능학과 같은 여러 분야에서 리용된다. 류사도계산은 단어의 미애매성으로부터 시작하여 본문요약, 정보추출과 검색, 질문응답, 자동색인과 코드의 자동교정에 이르기까지 매우 중요한 문제로 되고있다.

선행연구들에서는 질문과 문서사이의 류사도계산을 위하여 질문용어들의 발생빈도수와 발생위치, 발생순서, 근접정도를 론의하였으나 근접정도에 대한 엄밀한 정의는 하지 못하였다.

론문에서는 문장에서 질문용어들의 근접정도의 측도로 근접점수를 받아들인다. 즉 근접점수는 실제대답에 대한 문장의 적합도를 나타낸다.

질문문장《고구려의 초대왕은 누구인가?》를 실례들어보자.

질문용어《고구려》,《초대왕》들을 서로 다른 단락들에 포함한 문서들보다 같은 문장에 포함한 문서에 더 높은 우선권이 할당되여야 한다.

문장의 질문용어들사이의 거리를 측정하는데《질문용어간 거리》라고 하는 측정단위를 정의한다.

정의 질문용어간 거리는 문장에서 정합되는 질문용어들을 모두 포함하는 최소길이의 련속적인 단어배렬안에 있는 비질문용어들의 개수이다.

질문용어간 거리의 실례를 표 1에 보여주었다.

표 1. 질문용어간 거리의 실례

문 장	질문용어간 거리
고구려를 건립한 초대왕은 성은 고씨요, 이름은 주몽이라 하였다.	1
고구려초대왕은 고주몽이다.	2
고려의 초대왕인 왕건은 천년강국인 고구려를 계승한다는 의미에서 국호를 고려라고 하였다.	0

질문용어간 거리를 표 1과 같이 정의함으로써 시간복잡도는 $O(n)$, 공간복잡도는 $O(m)$으로 계산할수 있다.

2. 질문용어간 거리에 의한 근접함수결정

문장의 정합된 용어수와 질문용어간 거리에 의하여 문장을 비교할수 있는 점수를 생성하여야 한다. 그리고 질문용어간 거리와 정합된 단어수를 취하여 점수를 귀환하는 함수를 결정한다.

질문용어간 거리를 x, 정합된 단어수를 n, 점수를 $y = f(x, n)$이라고 하자.

우선 문장의 정합된 질문용어개수와 +1사이에서 근접점수를 변화시킨다.

실례로 정합된 용어가 3개라면 3과 4사이의 점수를 얻는다.

다음의 문장들을 실례들어보기로 한다.

문장 1 고려의 초대왕인 왕건은 천년강국인 고구려를 계승한다는 의미에서 국호를 고려라고 하였다.

문장 2 고구려초대왕은 고주몽이다.

문장 1은 언급된 모든 단어를 고려하면 질문용어간 거리가 2이고 2개의 정합된 용어를 가진다.

문장 2는 2개의 정합된 단어를 가지지만 질문용어간 거리는 1이다. 명백히 문장 2는 대답을 포함할 가능성이 더 높지만 우에서의 방법을 적용하면 두 문장에 대하여 근접점수가 같게 된다.

결과 근접점수계산에 질문용어간 거리를 고려해야 한다는것을 알수 있다.

다음으로 문장을 일반화하여 문서에 대하여 론의한다. 어떤 문서에 대하여 턱값 th에 이를 때까지 $[n, n+1]$ 사이에서 점수를 주는 함수가 필요하다. 이것은 $n-1$개가 정합된 용어를 가진 문서보다 n개 정합된 용어를 가진 문서를 선택하도록 한다. 또한 턱값이 2인 용어에 비하여 10인 용어에 관하여 더 큰 질문용어간 거리를 유지할수 있게 정합된 용어수의 함수가 되는것이 적합하다. 따라서 th는 th(n)이 되여야 한다.

정합된 단어수가 2이며 질문용어간 거리가 4이상이라면 점수를 2이하로 준다.

$n = 2$에 대하여 $y = f(x, n)$의 선택비교를 그림 1에 보여주었다. 그림 1에서 보여준것처럼 부의 지수함수는 작은 x에 대해서는 적게 작용하지만 매우 빨리 하강한다. 부의 로그함수는 $x = 2n$까지는 작용하지 않지만 높은 질문용어간 거리에 많이 작용하지 않으며 부의 무한대로 매우 천천히 다가간다.

두 방식에 가장 좋게 접근하는것은 부의 선형함수이며 부의 무한대로 적당한 속도로

접근한다.

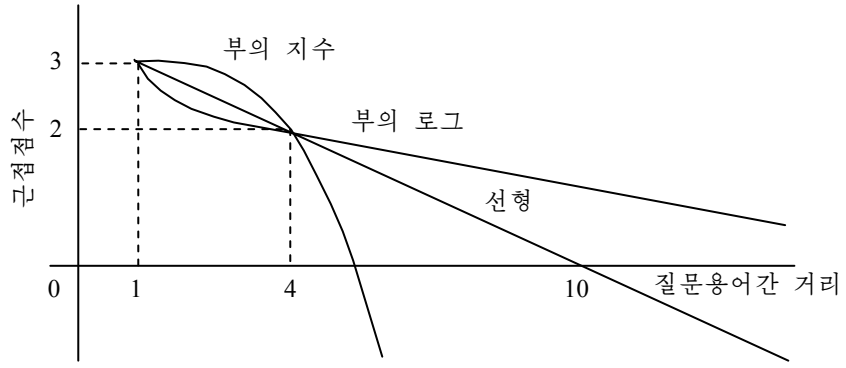

그림 1. $n=2$에 대하여 $y=f(x,n)$의 선택비교

실험을 통하여 $f(x,n)$을 선형, $\mathrm{th}(n)=2n$으로 선택하였다.

$$f(x,n)=\frac{-x}{2n-1}+\frac{2n^2+n}{2n-1}$$

식은 선택된 근접함수를 보여준다. 이것은 $f(x,n)$이 $n\in\{1,2,3,4,5,6\}$과 $x\in[1,15]$를 어떻게 찾는가를 보여준다.

선택된 근접함수를 그림 2에 보여주었다.

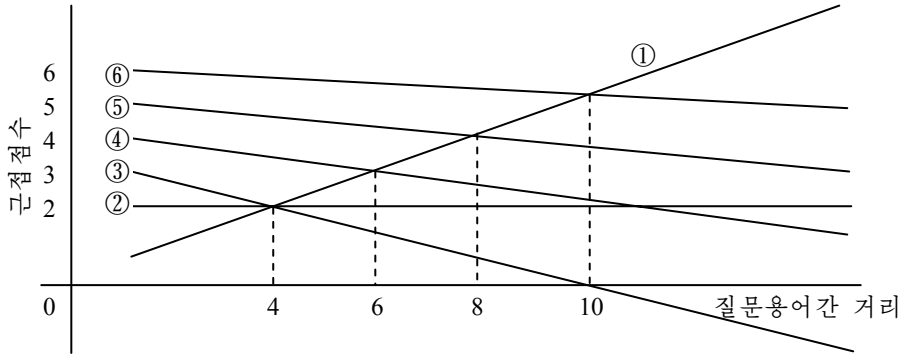

그림 2. 선택된 근접함수

그림 2에서 번호가 ①인 선은 점수가 n아래인 턱값함수를 나타낸다. 이로부터 경계내에서 점수를 주고 적당히 증감하는 $f(x,n)$을 얻을수 있다.

제안한 함수의 출력결과를 표 2에 보여주었다.

표 2. 제안한 함수의 출력결과

본문토막	근접점수
고구려의 초대왕은 고주몽이다.	3.8
발해는 고구려를 계승하였다.	2.4
발해의 초대왕	2
임의의 본문쪼각	0

3. 효과성평가

질문용어간 거리에 의한 근접함수적용의 효과성평가실험을 진행하였다. 용어근접정도를 고려한 방법을 비교하기 위하여 선행연구[1]에서 제안한 류사도계산방법과 론문에서 제안한 근접함수를 적용한 류사도계산방법을 비교하였다.

1 000개의 질문문장에 대한 대답추출의 효과성을 적중률과 완전률로 평가하였다.(표 2)

표 3. 실험결과비교

부류	선행방법	제안한 방법
적중률/%	75.8	85.2
완전률/%	65.4	78.4

표 3에서 보여준것처럼 제안한 방법이 선행방법에 비하여 적중률을 9.4%, 완전률을 13%만큼 올릴수 있다는것을 알수 있다.

맺 는 말

질문응답에서 질문용어간 거리를 정의한 다음 그것에 의한 근접함수를 결정하고 실험을 통하여 그 효과성을 검증하였다.

참 고 문 헌

[1] Shanthi Palaniappan et al.; International Journal of Pure and Applied Mathematics, 119, 12, 1969, 2018.

[2] Asad Abdi et al.; Soft Comput., 22, 213, 2018.

주체109(2020)년 5월 5일 원고접수

A Method of Deciding Proximity Function by Distance between Query Words in Question Answering

Choe Myong Ok, Tong Sung Chol

In this paper we defined the distance between query words in question answering, decided on proximity function by the distance and estimated the effectiveness.

Keywords: question answering, distance between query words, proximity function