한국어 구어 음성 언어 이해 모델에 관한 연구

노용완, 홍광석
성균관대학교 정보통신공학부 휴먼컴퓨터연구실
Tel. 031-290-7196 , HP. 019-296-0594

A Study on Korean Spoken Language Understanding Model

Yong-Wan Roh, Kwang-Seok Hong
HCI Lab, School of Information and Communication Engineering, Sungkyunkwan University
elec1004@hotmail.com, kshong@skku.ac.kr

Abstract

In this paper, we propose a Korean speech understanding model using dictionary and thesaurus. The proposed model search the dictionary for the same word with in input text. If it is not in the dictionary, the proposed model search the high level words in the high level word dictionary based on the thesaurus. We compare the probability of sentence understanding model with threshold probability, and we'll get the speech understanding rate. We evaluated the performance of the sentence speech understanding system by applying twenty qusetions game. As the experiment results, we got sentence speech understanding accuracy of 79.8%. In this case probability of high level word is 0.9 and threshold probability is 0.38.

Ⅰ. 서론

인간의 지적 능력을 모델화하여 지적 시스템을 구축하고자 하는 인공지능의 연구가 활발해 짐에 따라 자연언어 정보처리 분야가 크게 부각되고 있다.

자연언어 이해 분야는 응용시스템에서 연속음성 인식이 가능하게 된 요즘에 그 비중을 더해가면서 지적인 언어 이해 시스템에 대한 연구가 활발해 지고 있다. 인식된 문장을 이해하여 이해된 결과에 의해 판단 및 반응을 하는 것은 초보적 단계에 있으나 많은 기업과 연구소에서 이를 위한 연구가 활발히 진행되고 있다.

최근에 사전을 기반으로 한 한국어 의미망 구축과 활용에 대해 사전이나 의미기반의 정보 검색에 대한 많은 연구들이 계속되고 있으며 이는 언어자원 구축의 방향을 제시 하고 있다. 객체 지향 시소러스에서 참조 질의 조건 완화 기법의 방식은 객체지향 시소러스의 구조적인 특징을 이용하여 질의 조건을 일반화 시키는 질의 처리 기법이다.[2] 과거에는 사전만을 사용하거나 시소러스 만을 사용하여 확률값을 얻거나 항상 같은 값의 확률값을 부여 하였다.

본 논문에서는 입력되는 문장마다 시소러스와 사전을 사용한 구어 음성 언어 이해 모델을 제안 하였다. 이해 모델은 사전의 내용과 시소러스의 정보를 사용하여 이해 모델의 확률 값을 구하며 실험을 통해 얻어진 임계값과 비교하고 관련이 있는지의 여부를 판단 한다. 사전과 시소러스를 관련 분야에 맞게 추가하면 다른 분야도 이해 시스템에 적용 가능하도록 설계되었다.

Ⅱ. 구어음성언어의 특징

2.1 묵음 구간

텍스트 문서와는 달리 한 발화 안에서 묵음 구간이 빈번하게 나타나고, 묵음 구간의 길이 또한 길어진다. 낭독체의 경우는 묵음 구간을 고려하지 않아도 상관없지만 대화체의 경우는 묵음 구간을 고려해야 하며 묵음 구간을 고려하지 않을 때 많은 삽입오류를 유발시킨다.

2.2 잡음

입술소리, 숨소리, 웃음소리, 기침소리 등의 잡음은 자연스러운 발화에서는 위치에 상관없이 나타나는 잡음 현상으로 음성인식 단계에서 많은 삽입 오류를 유발시키는 요소이다. 텍스트의 형태에서는 전혀 알 수 없는 잡음들은 거의 음성을 받을 때 내부, 또는 외부의 잡음을 첨가하게 된다. 이러한 잡음은 음향 모델의 질과 인식성능을 떨어뜨리는 요소로 작용하므로 견고한 음향 모델을 사용하는 것이 필요하다.

2.3 간투어

텍스트에 없는 말들이 대화체에서 빈번하게 나타나는 현상으로 음성 인식 성능을 떨어뜨리는 요소 중의 하나이다. 이전의 여러 연구에서 간투어는 입술소리, 숨소리

등과 같은 비언어적인 요소로 분류되기도 했으나, 언어적인 경계 정보를 가지고 있어 발화 위치에 따라 다음 단어에 대한 예측 기능을 가지고 있다. 간투어들은 문장의 시작 부분에서 가장 많이 타나나기 때문에 음성을 문장 단위로 할 때 간투어의 정보를 효과적으로 사용할 수 있다.

2.4 반복/수정발화

반복 및 수정 발화 같은 단어 또는 어절을 반복해서 말하거나 다른 단어나 어절로 수정해서 말하는 현상을 말한다. 한국어에서는 온전한 한 어절이 반복되는 현상은 매우 드물고, 주로 어절의 일부분만 발화 되거나 어미나 조사가 변형된 형태로 수정되는 것이 일반적이다. 이때 단어나 어절을 온전하게 다 발화하지 않고 발화도중에 중단하여 어절의 일부만 발화하는 현상을 단어의 조각화라고 한다. "있/있습니까", "우/우리나라에", 와 같은 예에서 "있", "우"가 조각난 단어들인데 이들을 어떻게 분류하여 처리할 것인가가 논의의 대상이 된다. 대체로 이들 조각난 단어들은 사전에 포함시키지 않는 것이 일반적이다.

2.5 발음 변이

텍스트나 또는 문어체와 다르게 발화하는 현상을 모두 발음 변이로 양성음의 음성음 발화, 음운축약 및 탈락, 패턴화된 발음변이, 발화 오류등이 있다. 여기서 많이 나타나는 것은 것입니까 -> 것임까. 합니까 -> 함까 형태로 발음이 된다.

Ⅲ. 이해 모델

이해 모델은 문장 text 또는 음성이 입력되면 입력된 문장을 이해 하는 것이다. 이해 모델은 시소러스와 영역별 사전, 상위어 사전을 사용하여 구현되며, 입력되는 문장에 대해서 사전을 검색하고 사전에 없는 경우에 시소러스를 이용하여 문장내의 단어에 대한 상위어를 추출하고, 상위어의 사전을 사용하여 검색한다. 문장음성 이해 시스템은 컴퓨터가 임의의 카테고리를 선택하여 선택된 카테고리에서 정보를 가져온다. 사용자는 임의의 문장 또는 텍스트를 사용하여 질문을 한다. 질문과 카테고리가 관계가 있으면 예, 관계가 없으면 아니오라고 판단한다. 사전과 상위어 정보를 가진 시소러스를 가지면 어느 분야이든 이해과정에 있어서 적용이 가능하다. 이해 모델에서는 사용자가 음성으로 카테고리 내용을 질문해서 컴퓨터가 생각한 카테고리를 맞추도록 구현하였다.

이해모델의 개념도를 그림3.1에 나타내었다.

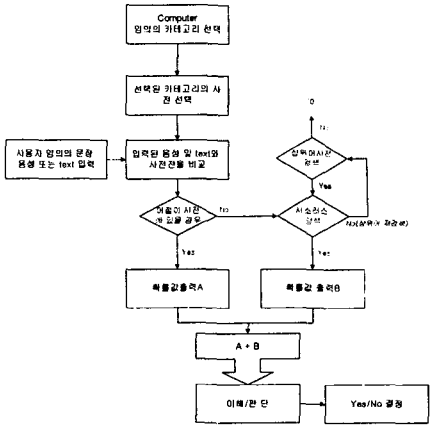

그림 3.1 이해 모델의 개념도

그림 3.1에서 컴퓨터가 임의의 카테고리를 선택하고 선택한 사전의 정보를 가져 온다. 사용자는 스무고개에 관련된 임의의 문장을 음성 또는 text로 질의 한다. 사용자의 질의와 컴퓨터의 사전을 비교하여 확률값 A를 출력하고 사전에 질의의 어절이 존재 하지 않을 경우 시소러스를 검색하여 상위어 정보를 가져오고, 상위어 사전을 비교하여 확률값 B를 출력한다.

출력된 두 개의 확률값을 더한 후 이해 및 판단을 한 다음 Yes, No를 결정한다.

3.1 사전

이해 모델에서 사전은 대용량 사전이나 영역별 사전, 상위어 하위어 사전등 여러 가지 사전을 이용하며 컴퓨터가 문장을 이해 하는데 기본적으로 필요한 도구이다. 사전의 사용 예로 스무고개 게임에서의 사전 구성을 설명한다. 본 논문에서는 일반 동물 사전과 상위어 사전으로 구성하였다. 일반 동물 사전은 Yahoo 동물 백과사전을 참조하여 구성하였으며 510개의 동물을 선정하였다. 동물 사전에는 510개의 동물들에 대한 설명이 나와있으며 그림 3.2와 같다.

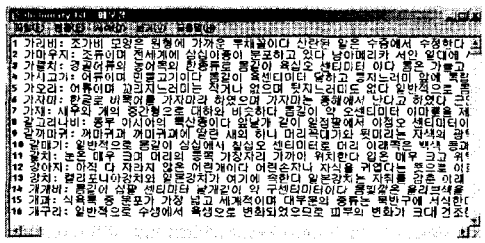

그림3.2 동물 사전

동물 사전은 맨 처음 동물 번호를 부여하였으며 가나다 순으로 정리 하였다. 510종류의 동물 선정은 동물

전공자가 아닌 일반인 5명이 한번이라도 들어본 적 있는 동물만 선택하여 구성하였다. 사전은 약 5만 어절정도로 구성되어 있으며 한 단어를 설명하는 평균어절이 100어절 정도이다.

상위어 사전은 동물 사전에 등록되어 있는 510개 동물 상위어에 대한 내용을 뜻 풀이 하였다. 시소러스에 동물의 상위어 정보들을 상위어 사전으로 따로 구성하여 뜻풀이를 하였다. 그 예로 시소러스에 호랑이, 척수동물, 포유류, 식육목, 고양이과에서 척수동물, 포유류, 식육목, 고양이과에 대한 사전은 상위어 사전에 나타내었으며 사전과 같은 뜻풀이문 형식으로 되어 있다. 상위어 사전은 일반사전에서 그에 관련된 정보가 없을 때 참조하게 된다. 상위어 사전은 그림 3.3과 같다.

3.2 시소러스

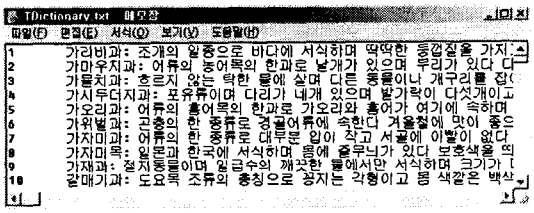

그림3.3 상위어 사전

시소러스는 사전에 없는 상하위관계를 나타내 주는 것으로 정보 검색에서 일반적으로 많이 사용한다.시소러스는 전체 시소러스와 영역별 시소러스로 나누어지며 전체 시소러스의 경우 모든 객체를 나타내주기 어렵기 때문에 영역별, 또는 부분 시소러스를 사용한다. 문장음성 이해 시스템에서는 사전과 시소러스를 사용하여 확률모델을 구성하고 확률값을 얻는다. 시소러스의 한 예로 동물 시소러스는 그림 3.4와 같이 구성한다.

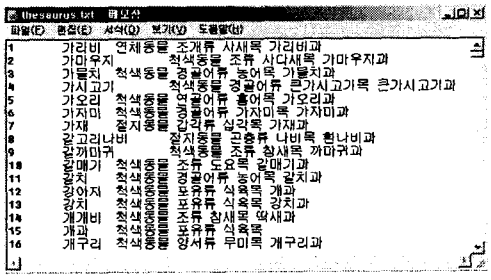

그림3.4 시소러스

동물 시소러스는 일반 동물 사전의 상위어 표시이며 사전의 상위어를 text로 나타내었다. 시소러스에서 첫 단어는 사전과 동일한 동물이 기재되어 있으며 선정한 동물 510개에 대한 상위어 정보를 기재하였다. 상위어는 척색동물, 포유류, 소목, 소과와 같은 형태 및 순서로 되어 있으며 최상위어부터 하위어 순으로 구성하였다. 한 동물의 상위어들의 최대 개수는 4개 이다.

동물 시소러스는 tree 형태로 구성되며 tree 형태를 바탕으로 상위어 정보가 기재된다.

3.3 확률모델

확률모델은 이해하는 과정의 일부분으로 판정 하기 이전의 단계에서 필요한 도구이다. 이 모델은 이해 판정하기 위해 필요한 과정이며 확률 임계값을 기준으로 문장의 이해도를 측정하게 된다.

$0<\alpha<1$, $0<\beta<1$ (α:상위어사전 검색시 확률값, β:threshold)

(1) $S=(W_1, W_2, \ldots W_n)$ (S:sentence, W:어절)

(2) $W\supset Po$이면 $w_i = W_i - Po$ 여기서,

$Po=(Po_1, Po_2 \ldots Po_n)$, ($P_o$:조사, w:단어)

$D\supset w_i$ 이면 w_{D_0}, 이것의 확률값 $\Pr(w_{D_0})$ 여기서 $w_i=(w_1, w_2, \ldots w_n)$, ($\Pr(w_{D_0})$:사전에 포함되는 단어의 확률값;포함된 단어 개수* $\frac{1}{n}$, D:사전)

(3) $TD_1\supset w_i$ 이고 $D\not\supset w_i$일때 w_{D_1}, 이것의 확률값 $\alpha\Pr(w_{D_1})$(TD_1:첫 번째 상위어 사전, $\Pr(w_{D_1})$: 첫 번째 상위어 사전에 포함되는 단어의 확률값)

(4) $TD_2\supset w_i$ 이고 $D\not\supset w_i$ 이고 $TD_1\not\supset w_i$일때 w_{D_2}, 이것의 확률값 $\alpha^2\Pr(w_{D_2})$ (TD_2:두번째 상위어 사전, $\Pr(w_{D_2})$: 두 번째 상위어 사전에 포함되는 단어의 확률값)

시소러스가 최상위어에 도달할때까지 계속,

(5) $TD_m\supset w_i$ 이고 $D\not\supset w_i$ 이고 $TD_{m-1}\not\supset w_i$ 일때 w_{D_m}, 이것의 확률값 $\alpha^m\Pr(w_{D_m})$ (m:상위어의 개수, $\Pr(w_{D_m})$: m번째 상위어 사전에 포함되는 단어의 확률값)

(6) $\Pr(S)=\sum_{j=0}^{m}\alpha^j\Pr(W_{D_j})$)(Pr(S) : 한 문장에 대한 확률값, j는 시소러스에서의 상위어 단계)

(7) 문장에 대한 확률값으로 판단

$\Pr(S)\geq\beta$ 이면 Yes, 아니면 No

3.4 판단

판단 및 결과 과정에서는 확률모델에 의해 나온 확률값

($Pr(S)$)를 사용하여 사용자 문장을 컴퓨터가 이해 판단한다. 판단의 임계값 β를 사용하여 예, 아니오를 판단한다. 본 알고리즘을 스무고개 게임에 적용시켜 β가 임계값 이상인 경우 예, 임계값 이하인 경우 아니오라고 판단한다. β를 결정하는 것은 실험 및 결과에서 초기치 확률을 정하는데 기반으로 하였다. α의 값을 얼마로 정하느냐에 따라 β의 값도 바뀌므로 어느 분야를 이해 하느냐에 따라 최적의 확률값을 구하기 위해 α와 β를 변화시켜서 가장 높은 이해도를 얻을수 있게 하는 α와 β를 실험을 통해 구한다.

Ⅳ. 실험 및 결과

본 논문에서의 실험은 두가지 형태로 이루어진다. 첫 번째는 문장음성 이해 확률 모델에서 사용할 최적의 α 값과 β값을 구하는 실험이다. 두 번째 실험은 두 번째 실험에서 구해진 최적의 확률값을 사용하여 스무고개 문장의 판단 정확도를 측정하였다.

4.1 확률모델 평가

확률모델을 평가하는 실험에서는 1000문장의 DB가 사용되었으며 판단의 기준이 되는 α와 β를 변화시키면서 최적의 확률값을 구하는 실험을 하였다. β를 0부터 0.5까지 0.02의 간격으로 하고 α을 0.5부터 1까지 0.05간격으로 변화 시켜 최적의 확률값을 구하여 α와 β 값을 찾는다.

실험 과정은 문장 10개를 추출하여 동물 510 종류에 대해 참값이 분포도와 거짓의 분포도를 구하였다. 총 5100개의 질문에 대해 참인 경우 1087개 거짓인 경우 4013개를 이용하였다. 상위어 값은 0.5일 때부터 1까지 0.05간격으로 11개, 판단의 기준은 0.02부터 0.52까지 0.02간격 나누어 실험하였다.

실험에 사용한 스무고개 문장이 일반적으로 /예/보다는 /아니오/를 답하는 문장이 많아서 /아니오/를 답하는 문장이 3.7배 많았다.

4.2 문장이해 스무고개 시스템

앞절에서 결정한 확률값을 사용하여 문장 이해 실험을 하였다. 많이 사용하는 200 종류 동물 단어에 대해 문장 이해 판단률을 구하였다.

입력되는 문장이 참인 문장일 때 예로 판단을 하는 경우와 거짓인 문장일 때 아니오라고 판단하는 경우가 올바른 판단이며, 반대의 경우는 거짓을 나타낸다. 즉 참->참, 거짓->거짓의 경우 올바른 판단, 참->거짓, 거짓->참이 잘못된 판단이다. 문장이해 판단 결과를 표 4.1에 나타내었다.

표4.1 문장이해 판단결과

	참->참	참->거짓	거짓->거짓	거짓->참
문장수	17359	7091	64066	13484

문장이해 판단 결과는 총 102000 종류 판단 결과 값을 나타낸 것이다. 정확도는 81425/102000로 79.8%의 판단 정확도를 나타내었다.

Ⅴ. 결론

본 논문에서는 문장음성 이해 분야에 대해 연구하였으며 이해 및 판단 모델을 확률적인 모델로 제안하여 구현하였다. 또한 제안한 확률적 모델의 성능평가를 위해 스무고개 게임을 구현하였다. 스무고개 게임은 문장음성 이해모델을 사용한 것으로 컴퓨터가 생각하고 있는 동물을 사람이 맞추어 나가면서 운영하도록 하였다.

입력된 문장을 이해하기 위하여 동물사전, 상위어 사전, 시소러스를 이용하였고 판단을 하기 위해 확률모델을 사용하였다. 최적의 확률모델을 얻기 위해서 α값과 β값을 변화 시켜 α값은 0.9, β가 0.38의 값을 최적의 값으로 얻었으며 판단률을 79.8%였다. 향후 판단의 정확도를 위해 사전의 확장이라든지 이해모델을 개선하여 더 높은 정확도를 얻을 수 있도록 개선하여야한다.

<감사의 글>
본 연구는 한국과학재단 목적기초연구(R05-2002-001007-0)지원으로 수행되었음.

참고문헌

[1] 이정민, "자연어 처리와 인지" 인지과학 Vol 3, No2, pp.161~176, 1992

[2] 김정애, 박종민, 김원중, 양재동, "객체지향 시소러스에서의 참조 질의 조건 완화 기법" 정보과학회 추계학술대회, pp.208~211, 2002

[3] 박계숙, "객체지향 기법을 이용한 시소러스 관리 시스템의 개발에 관한 연구" 한국정보처리 학회지 13권 2호 pp.5~18, 1996년

[4] 노용완, 윤재선, 홍광석 "스무고개 게임을 위한 음성인식", 전자공학회 하계종합 학술대회 논문집 제25권 1호 pp.203~206, 2002