KIPS Trans. Softw. and Data Eng.
Vol.8, No.5 pp.213~222 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2019.8.5.213

A Semi-Automatic Semantic Mark Tagging System for Building Dialogue Corpus

Junhyeok Park† · Songwook Lee†† · Yoonseob Lim††† · Jongsuk Choi††††

ABSTRACT

Determining the meaning of a keyword in a speech dialogue system is an important technology for the future implementation of an intelligent speech dialogue interface. After extracting keywords to grasp intention from user's utterance, the intention of utterance is determined by using the semantic mark of keyword. One keyword can have several semantic marks, and we regard the task of attaching the correct semantic mark to the user's intentions on these keyword as a problem of word sense disambiguation. In this study, about 23% of all keywords in the corpus is manually tagged to build a semantic mark dictionary, a synonym dictionary, and a context vector dictionary, and then the remaining 77% of all keywords is automatically tagged. The semantic mark of a keyword is determined by calculating the context vector similarity from the context vector dictionary. For an unregistered keyword, the semantic mark of the most similar keyword is attached using a synonym dictionary. We compare the performance of the system with manually constructed training set and semi-automatically expanded training set by selecting 3 high-frequency keywords and 3 low-frequency keywords in the corpus. In experiments, we obtained accuracy of 54.4% with manually constructed training set and 50.0% with semi-automatically expanded training set.

Keywords : Dialogue Corpus, Semantic Mark Tagging, Context Vector Similarity

대화 말뭉치 구축을 위한 반자동 의미표지 태깅 시스템

박 준 혁† · 이 성 욱†† · 임 윤 섭††† · 최 종 석††††

요 약

지능형 음성 대화 인터페이스 구현에 있어 핵심어의 의미표지는 사용자 의도 파악을 위한 중요한 요소이다. 대화시스템은 사용자 발화의 의도를 파악하기 위해 핵심어와 그 의미표지를 이용하여 발화의 의도를 결정한다. 하나의 핵심어는 여러 개의 의미표지를 가질 수 있는 중의성을 지닌다. 이러한 중의성을 지닌 핵심어를 사용자의 의도와 일치하는 의미표지로 결정하는 것은 단어 의미 분별 문제와 유사하다. 우리는 전사된 대화 말뭉치의 약 23%를 수동으로 의미를 부착하여 핵심어에 대한 의미표지 사전, 유의어 사전, 문맥벡터 사전을 먼저 구축한 후, 나머지 77% 대화 말뭉치에 존재하는 핵심어의 의미를 자동으로 부착한다. 중의성을 가진 핵심어는 문맥벡터 사전으로부터 문맥 벡터 유사도를 계산하여 의미를 결정한다. 핵심어가 미등록어인 경우에는 유의어 사전을 이용하여 가장 유사한 핵심어를 찾아 그 핵심어의 의미를 부착한다. 중의성을 가진 고빈도 핵심어 3개와 저빈도 핵심어 3개를 말뭉치에서 선정하여 제안 시스템의 성능을 평가하였다. 실험결과, 수동으로 구축한 말뭉치를 사용하였을 때 약 54.4%의 정확도를 얻었고, 반자동으로 확장한 말뭉치를 사용하였을 때 약 50.0%의 정확도를 얻었다.

키워드 : 대화 말뭉치, 의미 표지 태깅, 문맥 벡터 유사도

1. 서 론

음성 대화 시스템은 사용자의 발화를 분석하여 의도를 파악한 후, 최선의 대화 전략을 결정하는 시스템이다. 본 연구의 목적은 전사된 긴급 및 응급전화 음성 대화 말뭉치에 존재하는 사용자 발화에서 추출된 핵심어에 대하여 올바른 의미표지를 결정하는 것이다. 이 의미표지는 대화 관리자가 사용자의 의도를 파악할 때 사용되는 필수적인 요소 중 하나이다.

다음 Fig. 1은 주어진 대화 말뭉치 중 일부 대화를 나타낸 예제이다.

Receiver: "네 어디가 아프신데요?"
("Where are you sick?")
Caller: "힘이 없으시고 일어날 힘이 없으셔서 토할 거 같고 어지럽고"
("He has no power to get up and seems to vomit and dizzy")
Keywords: '없다', '일어나다', '없다', '구토', '어지럼증'
('have nothing', 'get up', 'have nothing'
'vomit', 'dizziness')

Fig.1. An Example of the Dialogue Corpus

※ 이 논문은 한국연구재단(NRF-2017R1D1A1A02019411)의 지원을 받았음.
† 비 회 원 : 한국교통대학교 컴퓨터정보공학과 석사과정
†† 정 회 원 : 한국교통대학교 컴퓨터정보공학전공 교수
††† 비 회 원 : 한국과학기술연구원 지능로봇연구단/치매DTC융합연구단 선임연구원
†††† 비 회 원 : 과학기술연합대학원(UST) HCI 및 로봇 세부전공 주임교수
Manuscript Received : December 28, 2018
First Revision : March 11, 2019
Accepted : March 21, 2019
* Corresponding Author : Songwook Lee(leesw@ut..ac.kr)

본 연구의 목적은 Fig. 1과 같이 신고자 발화와 그 핵심어가 주어졌을 때 핵심어의 의미표지를 다음과 같이 태깅하는 것이다.

의미표지: '없다/Others_Predicate',
'일어나다/FirstAid_Predicate',
'없다/Others_Predicate',
'구토/FirstAid_Subject',
'어지럼증/FirstAid_Subject'

차후 대화시스템은 이와 같이 올바른 의미표지가 부착된 핵심어를 토대로 신고자의 대화 주제가 '응급구조(FirstAid)'라는 것을 파악하여 응급구조 도메인에 맞는 대화 전략을 수립할 수 있게 된다.

핵심어는 문맥에 따라 여러 가지 의미로 사용되며, 하나의 핵심어가 여러 의미표지로 사용될 수 있다. 예를 들어 "불이 나다"의 '나다'는 'Fire_Predicate'이며, "사고가 나다"의 '나다'는 'Rescue_Predicate'이다. 우리는 핵심어의 의미를 신고자의 의도에 맞게 결정하는 문제를 단어 의미 분별(Word Sense Disambiguation) 문제로 간주하였다.

우리는 전사된 대화 말뭉치의 약 23%를 먼저 수동으로 의미표지를 부착한 후, 구축된 말뭉치를 이용하여 나머지 77%의 말뭉치에 존재하는 핵심어의 의미표지를 자동으로 부착하였다.

우리는 핵심어의 의미표지 부착 문제를 크게 핵심어가 핵심어-의미표지 사전에 존재하는 등록어인 경우와 존재하지 않는 미등록어인 경우로 나눠 해결한다. 등록어인 경우에는 대화 말뭉치의 문맥정보를 Word2Vec[1]를 통해 벡터로 변환한 후 이를 사용하여 의미표지를 결정한다. 미등록어인 경우에는 유의어 사전을 이용하여 미등록어와 가장 유사한 등록어를 찾은 후, 이를 핵심어로 간주하여 의미표지를 결정한다.

본 논문의 구성은 다음과 같다. 2장에서 핵심어의 의미결정 문제와 유사한 문제인 단어 의미 분별과 관련된 연구들을 살펴보고, 3장에서 본 시스템의 의미 결정 방법을 설명한다. 4장에서 본 시스템에 사용된 말뭉치와 이를 이용한 실험 결과를 보이며, 5장에서 결론을 맺는다.

2. 관련 연구

본 연구는 주어진 핵심어에 대한 올바른 의미표지를 결정한다는 점에서 의미역 결정(Semantic Role Labeling) 연구와 단어 의미 분별 연구와 유사하다. 의미역 결정 연구들은 크게 문장에서 나타나는 의미를 표현하기 위해 정의된 격틀(frame) 사전을 이용한 방법과 말뭉치를 이용한 방법으로 나뉜다. 격틀 사전 기반 의미역 결정 방법은 사전에 기술된 격틀에 따라 서술어-논항 관계와 격틀 사이의 유사도를 계산하여 의미역을 결정함으로서 높은 정확률을 보이지만 격틀 사전 구축이 어렵고, 격틀에 미등록된 문장 형태에는 적용하지 못하는 문제가 발생한다. 의미역 결정을 위해 [2]는 표준국어대사전을 기반으로 구축한 격틀 사전 정보와 한국어 어휘망 UPropBank에 있는 하위 범주 정보를 결합한 자질을 CRF(Conditional Random Field) 모델에 적용한 시스템을 제안하였다.

말뭉치 기반 의미역 결정 방법은 의미역이 부착된 말뭉치를 지지벡터기계, 신경망 등을 이용한 기계 학습기법으로 의미역을 결정하기 때문에 격틀 사전기반 방법보다 적용률이 높은 장점이 있으나, 양질의 의미역 부착 말뭉치 확보에 어려움이 있다. [3]은 의미역 부착 말뭉치 Korean Propbank[4]를 Neural Network Language Model(NNLM)을 통해 학습하여 축소된 차원의 단어 벡터를 얻은 후, 의미역을 결정하고자 하는 단어와 그 주변 단어, 단어로부터 추출된 자질을 Feed-Forward Neural Network(FFNN)의 입력으로 받아 최적화된 자질의 설계 및 조합을 제시하였다.

단어 의미 분별 연구들은 활용 자원에 따라 크게 지식 기반 방법과 말뭉치 기반 방법, 이를 혼용한 방법으로 분류할 수 있다. 지식 기반 방법은 사전(Machine Readable Dictionary)이나 시소러스 등의 자원들을 바탕으로 단어의 의미를 추론하고 접근하는 방법이다. 사전에는 여러 의미들을 지닌 단어의 뜻을 풀이한 문장들이 정의되어 있으며, 이를 문맥 정보로 추출하여 주어진 단어의 의미를 결정하는 방법으로 구현이 간단하지만 사전의 기재된 문장의 양과 질에 따라 자료 부족 문제가 발생한다[5].

자료 부족 문제는 WordNet[6]과 같은 시소러스를 이용하는 연구들이 진행되어 왔다. [7]은 세종 의미 부착 말뭉치에서 빈번히 사용된 용언 4개를 선정한 후, 주어진 용언이 문장에서 등장한 경우 공기한 명사들을 세종 전자사전의 용언 하위범주화 정보를 이용하여 규칙을 설정하였다. 이때 한국어 어휘의미망-KorLex를 이용하여 공기한 명사들을 확장하여 자료 부족 문제를 해소하였다. 주어진 용언이 사용된 문맥을 추출한 후 특정 규칙에 적용되는 의미를 결정하지만 규칙이 없거나 문맥이 KorLex에 없는 경우 문제를 해결하지 못하는 단점이 있다.

말뭉치 기반 의미 분별 방법은 의미태그가 부착되지 않은 원시말뭉치나 부착된 말뭉치를 나이브 베이지언 분류기, 지지벡터기계, 결정트리, 신경망 등을 이용한 기계학습 기법으로 단어의 의미를 결정한다. 기계학습 방법에 따라 지도학습(supervised learning)과 비지도학습(unsupervised learning)으로 나뉘며, 대부분의 경우 의미태그가 부착된 말뭉치를 이용한 지도학습 기반 시스템이 비지도학습보다 성능이 높다. 하지만 의미태그를 부착하는데 많은 시간과 자원이 필요하다.

[8]은 1,000만 어절의 세종말뭉치를 학습하면서 단어의 의미를 분별하는 은닉 마르코프 모델(Hidden Markov Model: HMM) 기반의 전이모델을 제안한다. 전이모델은 인접한 두 어절의 빈도를 이용하며, 추가로 한국어가 가진 교착어의 특성을 고려하여 앞 어절, 앞 어절의 마지막 문법 형태소와 뒤 어절의 첫 번째 실질형태소간의 관계(AF, EF)를 이용한다. 자료 부족에 따라 단계별로 전이모델을 제안한 결과 품사와 동형이의어를 동시에 태깅하면서 높은 성능을 보인다.

대규모의 말뭉치를 효율적으로 처리하는 벡터 공간 모델

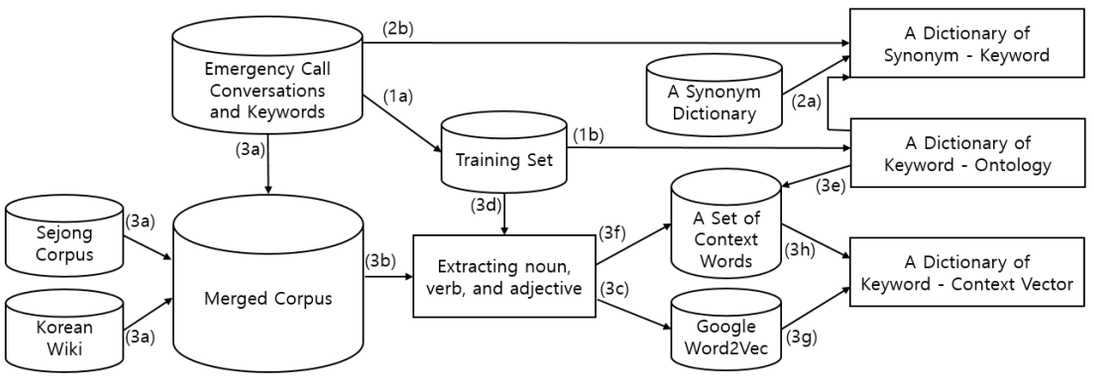

Fig. 2. The Process of Building Dictionaries for the Proposed System

[9]를 이용한 연구들은 간단하면서도 효과적이다. [10]은 의미태그가 부착된 세종 의미 부착 말뭉치를 학습하면서 중의성을 지닌 단어를 주어진 윈도우 크기 안에 있는 공기한 단어들의 빈도와 거리로 벡터 공간 모델을 만든 후, 주어진 단어 벡터와의 코사인 유사도를 계산하여 의미를 결정한다.

[11]은 자료 부족 문제를 해결하기 위해 세종 의미 부착 말뭉치에서 의미태그가 부착된 명사들을 추출한 후, 표준국어대사전에 대응하여 각 명사에 대하여 의미별로 정의된 속담과 용례 문장들을 말뭉치에 결합한다. 확장된 말뭉치를 학습하면서 [10]과 같은 벡터 공간 모델을 만들어 나이브 베이지언 분류기로 의미를 결정한다.

말뭉치를 구성하고 있는 각각의 단어들을 서로 다른 벡터로 표현하는 인공신경망 Word2Vec를 이용하여 단어들의 유사도를 측정하거나 중의성을 해소하는 연구들이 진행되어 왔다. [12]은 뉴스데이터를 수집하여 Word2Vec 모델에 학습하여 벡터값을 추출 한 후, 존재하는 모든 명사간의 코사인 유사도를 계산하여 상위 10개를 대체어로 추출하는 시스템을 제안하였다. [13]는 세종 의미 부착 말뭉치에서 4가지 중의성 단어를 선별한 후, 이 단어가 사용된 문맥을 추출한다. 이를 Word2Vec 모델에 학습하여 벡터 공간 모델로 생성한 후, 지지벡터기계(Support Vector Machine: SVM)를 이용하여 의미를 분별하였다.

3. 문맥 유사도를 이용한 의미표지 결정

우리는 먼저 전사된 긴급 및 응급전화 음성 대화 말뭉치의 약 23%를 학습 데이터로 사용한다. 이를 위해 학습집합의 발화와 이에 대응하는 핵심어에 대하여 수동으로 의미표지를 부착하여 핵심어-의미표지 사전을 구축한다. 이후 나머지 77%의 말뭉치의 핵심어는 핵심어-의미표지 사전을 기반으로 자동으로 의미표지를 부착한다. 핵심어를 자동으로 부착할 때, 핵심어가 구축된 사전의 의미표지와 1:1로 대응하는 경우, 1:N으로 대응하는 경우, 그리고 사전에 존재하지 않는 경우로 총 세 가지로 나뉜다.

우리는 핵심어가 사전에 1:N으로 대응되거나 존재하지 않는 경우의 문제를 해결하기 위해 Word2Vec 모델을 이용하여 핵심어의 문맥을 벡터로 표현한 후 단어 간 유사도 계산에 사용함으로써 문제를 해결한다.

본 시스템은 핵심어의 의미표지 결정을 위해 유의어-핵심어, 핵심어-의미표지, 핵심어-문맥벡터 총 세 개의 사전을 구축하며, Fig. 2는 사전들을 구축하는 과정을 나타낸다.

3.1 핵심어-의미표지 사전 구축

핵심어-의미표지 사전 구축 과정은 Fig.2의 다음 단계와 같다.

(1a) 전사된 긴급 및 응급전화 음성 대화 말뭉치의 약 23%를 학습집합으로 사용한다.

(1b) 대화 말뭉치는 신고자의 발화와 이에 대응하는 핵심어 k로 구성되며, 핵심어 k에 대하여 의미표지 t를 수동으로 부착하여 핵심어-의미표지 사전을 Equation (1)과 같이 구축한다. 우리는 중의성을 가지는 핵심어 k를 위해 n개의 의미표지 t를 부착하였다. 따라서 핵심어 k와 의미표지 t와의 관계는 1:1 또는 1:N이다.

$$dic_{keyword}(k) = < t_1, t_2, \dots, t_n > \quad (1)$$

Fig. 3은 중의성을 지닌 핵심어가 있는 예문이다.

Input 1) Fire situation
Caller: “천장이 불에 타다”
(“The ceiling is on fire”)
Keywords: ‘천장’, ‘불’, ‘타다’
(‘ceiling’, ‘fire’, ‘burning’)
Semantic mark: ‘천장/Others_Subject’(ceiling),
‘불/Fire_Subject’(fire),
‘타다/Fire_Predicate’(burning)

Input 2) General situation
Caller: “자전거를 타다가”
(“riding a bicycle”)
Keywords: ‘자전거’, ‘타다’
(‘bicycle’, ‘ride’)
Semantic mark: ‘자전거/Others_Subject’(bicycle),
‘타다/Others_Predicate’(ride)

Fig. 3. Examples of an Ambiguous Keyword

위와 같이 핵심어 '천장'과 '자전거'는 1개의 의미표지 <'Others_Subject'>가 부착되지만 핵심어 '타다'는 상황에 따라 다른 의미를 갖기 때문에 총 2개의 의미표지 <'Fire_Predicate', 'Others_Predicate'>가 부착된다.

3.2 유의어-핵심어 사전 구축

주어진 핵심어가 핵심어-의미표지 사전에 존재하지 않는 미등록어인 경우 말뭉치의 존재하는 핵심어 중 가장 유사한 핵심어를 찾아 문제를 해결한다. 이를 위해 유의어-핵심어 사전을 구축하게 되는데 그 과정은 Fig. 2의 다음 단계와 같다.

(2a) 유의어 사전과 핵심어-의미표지 사전을 로드한다.

(2b) 유의어 사전에서 모든 핵심어 k에 대응하는 유사한 단어 s를 추출하여 유의어-핵심어 사전을 Equation (2)와 같이 구축한다. 하나의 유의어 s는 m개의 핵심어를 가질 수 있으며, 핵심어 k는 핵심어-의미표지 사전에 따라 1개 또는 n개의 의미표지를 가질 수 있다. 따라서 유의어 s에 대응하는 핵심어 k와 의미표지 t와의 관계는 M:N이다.

$$dic_{synonym}(s) = < k_1, k_2, \dots, k_m > \tag{2}$$

3.3 핵심어-문맥벡터 사전 구축

핵심어가 핵심어-의미표지 사전에 1:N으로 대응하거나, 미등록어이지만 유의어-핵심어 사전에 존재하여 핵심어와 의미표지의 관계가 M:N으로 대응하는 경우 문맥벡터를 이용하여 문제를 해결한다. 다음은 핵심어-문맥벡터 사전 구축을 위한 Fig.2의 과정이다.

(3a) 대화 말뭉치, 세종 말뭉치, 한국어 위키 백과를 결합한 말뭉치를 이용한다.

(3b) 결합된 말뭉치의 모든 문장은 명사, 동사, 형용사의 품사로 구성하며, 형태소 분석에는 CNU 형태소 분석기[14]를 이용한다.

(3c) Word2Vec의 CBOW 모델을 이용하여 결함된 말뭉치의 모든 문맥을 학습한다. 이때 윈도우 크기는 5, 예측되는 단어의 고유벡터의 차원은 50으로 설정하였다.

(3d) 대화 말뭉치는 수보자와 신고자의 발화, 신고자의 발화를 분석하여 추출한 핵심어로 이루어져 있다. 우리는 신고자의 발화에 대응하는 수보자의 발화 또한 문맥을 결정하는 중요한 자질을 지닌 것으로 판단하여 하나의 문맥벡터를 생성하는데 한 쌍의 발화를 사용한다. 모든 발화는 명사, 동사, 형용사의 품사로 구성한다.

(3e) 3.1절에서 구축한 핵심어-의미표지 사전을 로드한다.

(3f) 단어의 위치정보는 중의성을 지닌 단어의 의미를 분별하는 데 있어 중요한 역할을 한다. 하지만 본 연구에서 사용한 대화 말뭉치에는 핵심어의 위치정보가 존재하지 않기 때문에 핵심어 k의 위치와 형태소 분석기를 이용하여 추출된 문맥 단어 c의 위치와의 연관성을 찾을 수 없기 때문에 윈도우 크기로만 문맥을 설정하였다. 따라서 핵심어 k가 의미표지 t를 가질 때 문맥을 구성하는 모든 문맥 단어 c를 사용하여 크기가 l인 집합 C_{k_t}을 Equation (3)과 같이 생성한다.

$$C_{k_t} = \{c_1, c_2, ..., c_l\} \tag{3}$$

Receiver: "네 어디가 아프신데요?"
("Where are you sick?")
Caller: "힘이 없으시고 일어날 힘이 없으셔서 토할 거 같고 어지럽고"
("He has no power to get up and seems to vomit and dizzy")
Semantic mark: '없다/Others_Predicate'(have nothing),
'일어나다/FirstAid_Predicate'(get up),
'없다/Others_Predicate'(have nothing),
'구토/FirstAid_Subject'(vomit),
'어지럼증/FirstAid_Subject'(dizziness)

C_{k_t} = { 아프(apeu)/VV:1, 힘(power)/NNG:2, 없(eobs)/VA:2, 일어나(il-eona)/VV:1, 토하(toha)/VV:1, 거(geo)/NNB:1, 같(gat)/VA:1, 어지럽(eojileob)/VA:1 }

Fig. 4. An Example of Building the Context Words Set

Fig. 4는 학습 말뭉치에서 문맥 단어 집합을 만드는 예제이다. 총 4개의 핵심어에 대한 의미표지('없다/Others_Predicate', '일어나다/FirstAid_Predicate', '구토/FirstAid_Subject', '어지럼증/FirstAid_Subject')는 모두 동일한 크기 l의 문맥 단어 집합 C_{k_t}을 가진다. 차후 다른 문맥들이 학습되면 중의성을 지닌 핵심어들이 학습되면서 의미표지 t에 따라 다른 크기 l을 가진 문맥 단어 집합 C_{k_t}가 구축된다. 따라서 n개의 문맥 단어 집합 C_{k_t}의 집합 U_k을 Equation (4)와 같이 생성한다.

$$U_k = \left\{C_{k_{t_1}}, C_{k_{t_2}}, ..., C_{k_{t_n}}\right\} \tag{4}$$

(3g) Word2Vec 모델을 로드한다.

(3h) 문맥 단어 집합 C_{k_t}에서 i번째에 위치하는 각각의 문맥 단어 c_i에 빈도수에 대한 가중치 w_i를 Equation (5)와 같이 계산한다. 이후 문맥 단어 집합 C_{k_t}에 있는 모든 문맥 단어 c_i들을 Word2Vec를 이용하여 문맥 단어 벡터 c_i^{w2v}로 치환하여 각각의 가중치 w_i를 Equation (6)와 같이 적용하여 문맥벡터 v_{k_t}를 얻는다. 우리는 Word2Vec 모델의 고유벡터 크기 d를 50차원으로 설정하였으며, Equation (6)은 각각의 가중치 w_i가 적용된 문맥 단어 벡터를 더하여 최종적으로 크기 d가 50인 문맥벡터 v_{k_t}를 얻는다. 이와 같은 과정을 거쳐 n개의 의미표지 t를 가진 핵심어-문맥벡터 사전을 Equation (7)과 같이 구축한다.

$$w_i = \frac{-1}{\log\left(\dfrac{freq(c_i)}{\sum_{c_i \in U_k}^{n} freq(c_i)}\right)} \tag{5}$$

$$v_{k_t} = \frac{1}{l}\sum_{i}^{l} w_i \cdot c_i^{w2v}, \qquad c_i^{w2v}\begin{cases} w2v(c_i), & c_i \in w2v \\ 0, & c_i \notin w2v \end{cases} \tag{6}$$

$$dic_{context}(k) = < v_{k_{t_1}}, v_{k_{t_2}}, ..., v_{k_{t_n}} > \tag{7}$$

3.4 의미표지 결정

Fig. 5는 구축된 사전들을 이용하여 의미표지를 결정하는 과정을 나타낸다.

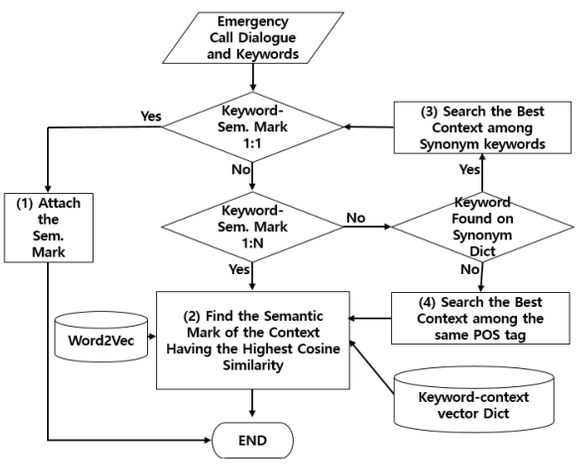

Fig. 5. The Process of Determining Semantic Mark Using the Dictionaries

Fig. 5는 의미표지를 자동으로 부착하는 과정을 나타낸 그림이다. 먼저 입력 핵심어가 핵심어-의미표지 사전에 1:1로 대응하는 경우, Fig. 5의 (1)번과 같이 해당하는 의미표지를 부착한다. 핵심어가 핵심어-의미표지 사전에 1:N으로 대응하는 경우, Fig. 5의 (2)번과 같이 가장 코사인 유사도가 높은 문맥을 가진 의미표지를 찾는다. 이 때 입력 문맥의 벡터 v_s를 Equation (6)을 이용하여 구한 후, Equation (8)을 이용하여 코사인 유사도(cosine similarity)를 계산한다. 이 때, 입력 문맥과 핵심어의 의미표지 n개와의 문맥유사도를 각각 계산하여 Equation (9)와 같이 유사도가 가장 큰 의미표지를 결정한다.

$$\cos(v_s, v_{k_t}) = \frac{\sum_{i}^{d}(v_{si} \times v_{k_t i})}{\sqrt{\sum_{i}^{d}(v_{si} \times v_{si})} \times \sqrt{\sum_{i}^{d}(v_{k_t i} \times v_{k_t i})}} \tag{8}$$

$$SemMark(k) = \arg max_{k \in dic_{context}}(\cos(v_s, v_{k_t})) \tag{9}$$

핵심어가 핵심어-의미표지 사전에 존재하지 않는 미등록어인 경우, Fig. 5의 (3)번과 같이 유의어-핵심어 사전을 이용하여 의미가 유사한 핵심어들을 구한다. 이 때, 유의어-핵심어 사전에 존재하는 경우 유사한 핵심어들의 개수에 따라 발생하는 경우의 수는 총 2가지이다.

만약 유사한 핵심어의 개수가 1개이면 그 핵심어의 의미표지를 결정하는 문제로 간주하여 핵심어가 등록어일 때와 동일하게 의미표지를 결정하면 된다.

만약 유사한 핵심어의 개수가 m개이면 미등록어의 의미표지는 1:M×N으로 대응하는 경우로 볼 수 있으며 주어진 핵심어의 문맥과 m개의 유사 핵심어들에 대한 모든 문맥들과의 유사도를 측정한 후 가장 높은 유사도를 갖는 문맥의 의미표지를 부착한다. 이 경우도 유사도 비교 대상 문맥의 개수만 많아졌을 뿐, 등록 핵심어의 1:N 대응을 처리하는 과정으로 처리할 수 있다.

Fig. 5의 (4)번은 핵심어가 유의어-핵심어 사전에도 존재하지 않는 경우이다. 이때는 동일 품사를 가진 모든 핵심어와의 문맥 유사도를 측정하여 그 중 가장 높은 유사도를 갖는 의미표지를 부착한다.

Fig. 5는 대화 말뭉치에 있는 핵심어에 대한 의미표지를 결정하기 위해 문맥 벡터를 계산하는 과정을 나타낸 예이다.

Receiver: "예"
("Yes")
Caller: "산에서 연기가 나요. 담뱃불을 던졌는지."
("Smoke from the mountain. I guess someone throw a cigarette")
Keywords: '산', '연기', '나다', '담뱃불'
('mountain', 'smoke', 'nada', 'cigarette')

$v_s =$
< foreach i
$w_i \ c_i^{w2v}$,
expect 산(mountain)/NNG:1, 연기(smoke)/NNG:1,
나(na)/VV:1, 담뱃불(cigarette)/NNG:1, 던지(deonji)/VV:1
>

Fig. 6. An Example of Producing a Context Vector

Fig. 6에서 주어진 발화의 문맥 단어들은 {산/NNG:1, 연기/NNG:1, 나/VV:1, 담뱃불/NNG:1, 던지/VV:1}로 표현되며 Equation (5)와 Equation (6)에 의해 문맥 벡터 v_s로 변환된다.

Fig. 6의 핵심어 '산'과 '연기'는 의미표지와의 관계가 1:1 대응이기 때문에 사전에 있는 의미표지 'Mountain'과 'Smoke'를 각각 부착한다.

Fig. 6의 핵심어 '나다'는 의미표지와의 관계가 1:N이기 때문에 Fig. 5의 (2)번, (3)번 과정을 거쳐 아래와 같이 의미표지를 부착한다.

$$SemMark(\text{나다}) = \arg max_{\text{나다} \in dic_{context}}(\cos(v_s, v_{k_t}))$$

Fig. 6의 핵심어 "담뱃불"은 미등록어이며 Fig. 5의 (4)번 과정을 거쳐 유의어-핵심어 사전에서 2개의 유사어{'불', '산불'}를 가진다. Fig. 5의 (2)번, (3)번 과정을 거쳐 '불'과 '산불'의 모든 문맥에 대하여 다음과 같이 유사도를 계산하여 의미표지 'Fire_Subject'로 결정한다.

$$SemMark(\text{담뱃불}) = \arg max_{\text{불,산불} \in dic_{context}}(\cos(v_s, v_{k_t}))$$

최종적으로 대화 말뭉치에 다음과 같은 의미표지가 추가된다.

의미표지: '산/Mountain',
'연기/Smoke',
'나다/Fire_Predicate',
'담뱃불/Fire_Subject'

4. 실험 및 결과

우리는 Word2Vec 모델을 학습하기 위해 대화 말뭉치, 세종 말뭉치, 한국어 위키 백과를 결합한 말뭉치를 사용한다. 대화 말뭉치는 신고자의 발화가 수보자의 질의에 대한 간단히 대답하는 발화가 많아 형태소 분석 결과 어절이 5개 이상인 20,726개의 문장들만 사용하였다. 어휘 사전의 크기는 70,278개이며, 학습된 말뭉치의 용량은 약 184MB이다.

Table 1은 Word2Vec 모델을 위해 말뭉치별로 학습에 사용된 문장과 어절의 수를 나타낸다.

Table 1. Dataset for Word2Vec Modeling

Corpus	# of sentences	# of words
Emergency Call Conversations	20,726	172,264
Korean Wiki	612,030	9,755,245
Sejong Corpus	746,873	7,639,770
Total	1,379,629	17,567,279

우리는 대화 말뭉치의 약 23%를 학습집합 A로 선정하여 핵심어에 대한 의미표지를 수동으로 부착하였다. 그 후 제안 시스템으로 나머지 약 77%의 대화들의 핵심어에 대한 의미표지를 자동으로 부착하였다. 수동으로 의미표지를 부착한 학습집합 A와 자동으로 의미표지를 부착한 집합을 통합하여 학습집합 B를 구축하였다. 따라서 B 집합은 대화 말뭉치 전체이며 반자동으로 의미표지가 부착된 말뭉치이다.

Table 2는 학습집합 A와 확장된 학습집합 B의 구성을 나타낸다.

Table 2. The Manual Tagged Set and the Expanded Set

Set	# of conversations	# of sentences	# of keywords
A	1,878	8,779	14,467
B	8,296	36,882	59,994

A 집합은 8,779개 문장, 14,467개 핵심어로 구성되며, 우리는 제안 시스템을 이용하여 28,103개 문장, 45,527개 핵심어에 대하여 자동으로 의미표지를 부착하였다. 본 시스템의 주요 의미표지는 크게 Subject와 Predicate로 나뉘며 신고 상황에 따라 Fire, FirstAid, Rescue, Others 등으로 나뉜다.

Table 3은 B 집합의 상위 20개의 고빈도 의미표지를 나타낸다. 긴급 및 응급전화의 신고 내용은 일반적으로 화재상황보다 응급상황과 구조상황이 많은 것을 알 수 있다.

Table 3. A Evaluation Result of New Keyword

Semantic Mark	Sense	Frequency
Others_Subject	subject	4,211
Others_Predicate	predicate	1,880
FirstAid_Predicate	firstaid predicate	928
Rescue_Subject	rescue subject	917
FirstAid_Subject	firstaid subject	701
Hospital	hospital	599
SingleHouse	house	502
Apartment	apartment	459
GeneralBuilding	school, post office	437
Child	child	435
Family_Member	mother, father	359
Rescue_Predicate	rescue predicate	358
Telephone	telephone	317
FirstAider	911	287
Fire_Subject	fire subject	264
Location	location, place	259
Road	road	196
Fire_Predicate	fire predicate	154
Market	market, mart	121
FirstAid_Case	emergency center	110

Table 4는 A 집합을 이용하여 대화 말뭉치에 의미표지를 자동으로 부착할 때 등록어와 미등록어의 빈도를 나타낸 표이다. 대부분의 사전에 등록된 핵심어는 의미표지와 1:1로 대응하며, 미등록어의 경우 유의어 사전을 이용하여 처리할 수 있는 비율이 적은 것을 알 수 있다.

Table 4. Number of Cases by Automatic Tagging

Case			Frequency
Registered keywords	# of 1:1 mapping		41,743
	# of 1:N mapping		2,944
Non-Registered keywords	Using synonym dictionary	# of 1:1 mapping	75
		# of 1:N mapping	0
		# of M:N mapping	31
	# of noun mapping		637
	# of verb mapping		97
# of keywords in set B			45,527

Table 5는 A 집합과 확장된 B 집합에 대하여 핵심어-의미표지 사전의 크기를 나타낸 표이다.

Table 5. Entries of Keyword-Sematic Mark Dictionary

Set	# of 1:1 mapping	# of 1:N mapping	# of total
A	842	17	859
B	1,135	39	1,174

확장된 학습집합 B에서 추가로 중의성을 지닌 핵심어는 22개이다. 이는 미등록어 처리 과정에서 발생되는 결과로, A 집합의 사전에 존재하지 않는 단어가 문맥 벡터간의 유사도를 이용하여 의미를 부착할 때 의미가 두 개 이상으로 나뉘는 경우 발생한다.

우리는 A 집합과 B 집합을 이용하여 구축한 시스템의 성능을 평가하기 위해 핵심어와 의미표지가 1:N으로 대응하는 중의성 핵심어 중 고빈도 핵심어 3개-(나다, 타다, 먹다)와 저빈도 핵심어 3개-(일어나다, 걷다, 맞다)를 선정하였다. A 집합을 이용하여 자동으로 B 집합을 구축하는 과정에서 선정된 중의성이 있는 핵심어를 포함하는 문장의 약 10%를 각각 랜덤하게 추출하여 평가집합을 구성하였다.

Table 6. Frequency of Semantic Marks and the Result of the System

Ambiguous word (# of senses)	Semantic mark	# in A set	# in B set	# in test	acc with A set	acc with B set
나다 (nada) (4)	Fire_Pred.	89	153	24	29.2	33.3
	FirstAid_Pred.	85	368	24	87.5	91.7
	Rescue_Pred.	87	518	20	85.0	40.0
	Others_Pred.	30	74	16	6.3	0
	Overall	291	1,113	84	54.8	45.2
타다 (tada) (2)	Fire_Pred.	15	133	6	100.0	83.3
	Others_Pred.	79	214	21	52.4	66.7
	Overall	94	347	27	63.0	70.4
먹다 (meogda) (2)	Rescue_Pred.	31	204	7	100.0	100.0
	Others_Pred.	62	153	20	30.0	30.0
	Overall	93	357	27	48.1	48.1
일어나다 (il-eonada) (3)	Fire_Pred.	1	12	0	-	-
	FirstAid_Pred.	18	57	5	60.0	40.0
	Others_Pred.	7	21	2	0	0
	Overall	26	90	7	42.9	28.6
걷다 (geodda) (2)	FirstAid_Pred.	8	58	2	100.0	100.0
	Others_Pred.	9	22	5	40.0	60.0
	Overall	17	80	7	57.1	71.4
맞다 (majda) (2)	Rescue_Pred.	11	38	1	100.0	100.0
	Others_Pred.	10	40	5	40.0	20.0
	Overall	21	78	6	50.0	33.3
Overall	-	-	-	-	54.4	50.0

Table 6은 선정된 핵심어가 A 집합, B 집합, 평가집합에 등장하는 빈도와 두 집합을 이용하여 구축한 시스템의 성능을 평가집합의 정확도로 나타낸 표이다. 학습집합 A를 이용한 시스템의 성능은 약 54.4%이었고 확장 학습집합 B를 이용한 시스템의 성능은 50.0%이었다. 이는 수동 구축 말뭉치를 사용했을 때의 성능보다 반자동 구축 말뭉치를 사용했을 때의 성능이 약 4.4% 떨어진다고 볼 수 있다. 수동 부착 비용을 고려하면 큰 성능 저하는 아니라고 볼 수 있다.

가장 많이 발생하면서 가장 많은 수의 의미표지(4개)를 갖는 핵심어 '나다'의 의미 표지 부착 성능은 대체로 'FirstAid'와 'Rescue' 상황일 때 높게 나타났으며, 이는 Table 6에서와 같이 위 의미들을 지닌 대화가 비교적 자주 발생하면서 문맥 단어들이 두 상황을 구분하는 데 중요한 역할을 한다고 해석된다. 실제로 'FirstAid' 상황인 경우에는 문맥 단어가 '출혈', '구토' 등과 같은 단어들이 빈번하게 발생하고, 'Rescue' 상황인 경우에는 '차', '사고' 등의 단어들이 빈번하게 발생한다. 반면 'Others' 상황인 경우 특정 상황이 아닌 나머지 일반적인 상황으로서 다양한 단어들이 주변에 등장하여 문맥 단어로 그 상황을 구분하기 힘든 경우가 많다. 마찬가지로 나머지 핵심어들 또한 'Others' 상황일 때 성능이 다른 상황보다 비교적 낮게 나타났다.

핵심어 '나다'의 'Rescue' 상황에서 성능은 학습집합 A를 사용했을 때보다 반자동으로 구축한 학습집합 B를 이용할 때 성능이 다른 의미표지에 비해 월등히 저하되었다. 이는 두 집합의 의미표지의 분포가 서로 다른 데서 그 원인을 찾을 수 있다. Table 6에서 핵심어 '나다'의 의미표지 'Fire', 'FirstAid', 'Rescue'의 빈도는 학습집합 A에서 각각 89회, 85회, 87회로 의미 표지별로 거의 비슷한 분포를 가진다. 그러나 학습집합 B에서의 빈도는 각각 153회, 368회, 518회로 특히 'Rescue'에 크게 편중된 결과를 보이며 결과적으로 자동으로 의미 표지를 부착하는 과정에서 Rescue로 잘못 부착하는 오류를 유발한다. 그 외에도 두 학습집합의 의미 표지별 문맥 단어의 분포의 차이에서도 성능 저하의 원인을 찾을 수 있다.

Table 7은 '나다'의 의미 표지별로 빈번하게 등장하는 상위 4개의 문맥 어휘와 각 의미 표지별 문맥 어휘의 총 빈도수를 나타낸 표이다. '나다'의 의미 표지 'Fire', 'FirstAid', 'Rescue'의 문맥 단어의 총 빈도수는 학습집합 A에서 각각 659회, 650회, 625회이지만 학습집합 B에서 각각 1,171회, 2,729회, 3,920회로 의미 표지별 문맥 단어의 분포도 두 집합이 서로 다르다. 반자동으로 말뭉치를 구축하는 과정에서 'Rescue'로 의미표지가 부착되는 오류가 다른 의미표지보다 많이 발생하였고, 결과적으로 오류 문맥들이 많이 섞여 학습집합 B를 사용한 시스템의 정확도가 저하된 것으로 분석된다.

제안 시스템의 성능이 비교적 낮은 이유는 크게 세 가지를 들 수 있다. 첫 번째로 현재 대화 말뭉치에 구축된 핵심어의 위치 정보가 없어 각 핵심어의 정확한 문맥 단어들을 추출할 수 없어 하나의 대화에 대해 모든 핵심어들이 동일한 문맥을 가지도록 설계되었다.

Table 7. Context Words Frequency of '나다(nada)'

Semantic mark	in A set		in B set	
	Context	Freq	Context	Freq
Fire_ Predicate	불(fire)/NNG	46	불(fire)/NNG	105
	연기(smoke) /NNG	35	연기(smoke) /NNG	65
	지금(now) /NNG	21	지금(now) /NNG	27
	냄새(smell) /NNG	17	냄새(smell) /NNG	24
	Total	659	Total	1,171
FirstAid_ Predicate	피(blood)/ NNG	56	피(blood) /NNG	143
	하(ha)/VV	18	지금(now) /NNG	70
	열(fever)/NNG	16	있(iss)/VV	54
	지금(now) /NNG	16	머리(head) /NNG	50
	Total	620	Total	2,729
Rescue_ Predicate	사고(accident) /NNG	73	사고(accident) /NNG	221
	교통(traffic) /NNG	33	지금(now) /NNG	113
	지금(now) /NNG	21	교통(traffic) /NNG	94
	일일구(911) /NNP	15	일일구(911) /NNP	71
	Total	625	Total	3,920
Others_ Predicate	소리(sound) /NNG	10	하(ha)/VV	36
	지금(now) /NNG	9	있(issda)/VV	34
	거(geo)/NNB	6	지금(now) /NNG	27
	하(ha)/VV	6	소리(sound) /NNG	21
	Total	238	Total	843

Table 7에서 신고 전화에서 자주 나오는 단어 중 하나인 '지금/NNG'은 모든 의미표지의 문맥에 등장한다. 이는 핵심어-문맥 벡터 사전의 변별력을 떨어뜨리는 결과를 가져오게 된다. 두 번째는 의미표지를 수동으로 부착할 때 모호한 신고 상황에서 일관성있는 부착이 어렵고 신고자의 의도도 불명확한 경우도 많다. 예를 들어, 주로 'Fire' 상황에 관찰되는 '불', '화재' 등의 핵심어가 'FirstAid'와 'Rescue' 상황에도 많이 관찰되므로 세 가지 상황 중 특정한 상황으로 결정하기 어려운 문제점이 있다. 마지막으로 실험에 사용한 대화 말뭉치의 크기가 약 2만 문장에 불과해 자료부족 문제가 발생한다. 다양한 상황에 구사되는 발화를 많이 수집한다면 양질의 문맥 벡터를 구축하여 제안 시스템의 성능을 향상할 수 있을 것이다. 그 외 품사 태깅 오류로 인해 좋은 자질이 될 수 있는 단어가 문맥 단어에 누락되어 발생하는 오류도 있다.

5. 결론 및 향후 과제

의미표지는 대화 관리자가 사용자의 의도를 파악할 때 사용되는 필수적인 요소 중 하나이다. 우리는 음성 대화시스템에서 구축된 대화 말뭉치의 약 23%의 말뭉치에 수동으로 핵심어의 의미표지를 부착하였다. 부착된 말뭉치를 이용하여 핵심어 의미표지를 결정하는 시스템을 구축하였으며 이를 이용하여 나머지 77%의 말뭉치에 자동으로 의미표지를 부착하였다. 핵심어의 의미표지별 문맥벡터 사전을 Word2Vec로 먼저 구축하여 주어진 핵심어의 문맥과 의미표지별 문맥을 비교하여 가장 유사한 문맥을 가진 의미표지를 그 핵심어의 의미표지로 결정하였다. 중의성을 가진 핵심어 6개를 말뭉치에서 선정하여 제안 시스템의 성능을 평가하였다. 실험결과, 수동으로 구축한 말뭉치를 사용하였을 때 약 54.4%의 정확도를 얻었고, 반자동으로 확장한 말뭉치를 사용하였을 때 약 50.0%의 정확도를 얻었다. 대화 말뭉치에 존재하는 핵심어 정보의 위치정보가 없어 올바른 문맥을 추출할 수 없어 반자동 구축 말뭉치를 이용한 시스템은 비교적 낮은 성능을 보였다.

본 핵심어 의미 표지 부착 시스템의 성능을 향상시킬 수 있도록 문맥의 유사도를 계산할 때, [10]에서와 같이 사전(prior) 확률을 이용해 볼 필요가 있다. 말뭉치를 반자동으로 구축하는 과정에서 발생하는 오류를 필터링 할 수 있는 연구를 진행하여 확장된 말뭉치를 이용한 시스템의 성능을 향상시켜야 한다. 발화에서 적절하고 올바른 문맥을 추출할 수 있도록 구문 분석기 등을 이용한 방법을 연구하여야 하며, 구축된 의미표지 부착 말뭉치를 활용한 대화 관리자 및 응답 발화 생성에 대한 연구가 필요하다.

References

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv:1301.3781*, 2013.

[2] Wansu Kim and Cheolyoung Ock, "Korean Semantic Role Labeling Using Case Frame Dictionary and Subcategorization," *Journal of KIISE*, Vol.43, No.12, pp.1376-1384, 2016.

[3] Jangseong Bae, Changki Lee, and Soojong Lim, "Korean Semantic Role Labeling using Deep Learning," *Proc. of the KIISE Korea Computer Congress 2015*, pp.690-692, 2015.

[4] Martha Palmer, Shijong Ryu, Jinyoung Choi, Sinwon Yoon, and Yeongmi Jeon, Korean Propbank, [Online]. Available: http://catalog.ldc.upenn.edu/LDC2006T03.

[5] Michael Lesk, "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone," in *Proceedings of the 5th Annual International Conference on Systems Documentation*, 1986.

[6] G. A. Miller, "WordNet : An On-Line Lexical Database," *International Journal of Lexicography*, Jan. 1990.

[7] Sangwook Kang, Minho Kim, Hyukchul Kwon, Sungkyu

Jeon, and Juhyun Oh, “Word Sense Disambiguation of Predicate using Sejong Electronic Dictionary and KorLex,” *KIISE Transactions on Computing Practices*, Vol.21, No.7, pp.500-505, 2015.

[8] Joonchoul Shin and Cheolyoung Ock, “A Stage Transition Model for Korean Part-of-Speech and Homograph Tagging,” *Journal of KIISE*,, Vol.39, No.11, pp.889-901, 2012.

[9] H. Schutze, “Automatic Word Sense Discrimination,” *Computational Linguistics*, Vol.24, No.1, 1998.

[10] Yongmin Park and Jaesung Lee, “Word Sense Disambiguation using Korean Word Space Model,” *Journal of The Korea Contents Association*, Vol.12, No.6, pp.41-47, 2012.

[11] Hanjo Jeong and Byeonghwa Park, “Korean Word Sense Disambiguation using Dictionary and Corpus,” *Journal of Intelligence and Information Systems*, Vol.21, pp.1-13, 2015.

[12] Sangyun Kim and Soowon Lee, “Automatic Extraction of Alternative Word Candidates using the Word2vec model,” *Korean Institute of Information Scientists and Engineers*, Vol.2015, No.12, pp.769-771, 2015.

[13] Junhyeok Park, and Songwook Lee, “Word Sense Classification Using Support Vector Machines,” *KIPS Tr.*, Vol.5, No.11, pp.563-568, 2016.

[14] Kongjoo Lee and Songwook Lee, “Error-driven Noun-Connection Rule Extraction for Morphological Analysis”, *Journal of the Korean society of Marine Engineering*, Vol. 36, No.8, pp.1123-1128, 2012.

박 준 혁

https://orcid.org/0000-0002-3653-9425
e-mail : jhpark@a.ut.ac.kr
2017년 한국교통대학교 컴퓨터정보공학과 (학사)
2018년~현 재 한국교통대학교 컴퓨터정보공학과 석사과정
관심분야 : 자연언어처리, 기계학습, 의미분별, 인공지능

이 성 욱

https://orcid.org/0000-0002-6224-4241
e-mail : leesw@ut.ac.kr
1996년 서강대학교 전자계산학과(학사)
1998년 서강대학교 컴퓨터학과(석사)
2003년 서강대학교 컴퓨터학과(Ph.D.)
2003년~2004년 서강대학교 산업기술연구소 연구원
2003년~2005년 서강대학교 정보통신대학원 대우교수
2004년~2005년 LG전자 기술원 선임연구원
2005년~2007년 동서대학교 컴퓨터공학과 전임강사
2007년~현 재 한국교통대학교 컴퓨터정보공학전공 교수
관심분야 : 자연언어처리, 대화인터페이스, 기계학습, 인공지능

임 윤 섭

https://orcid.org/0000-0003-4754-6038
e-mail : yslim@kist.re.kr
2002년 서울대학교 기계항공공학부(학사)
2004년 서울대학교 전기공학부(석사)
2014년 Boston University, Computational Neuroscience (Ph.D.)
2014년~현 재 한국과학기술연구원 지능로봇연구단/치매DTC융합연구단 선임연구원
관심분야 : 인공지능, 청각인지, 뇌-컴퓨터 인터페이스, 인간-로봇 상호작용

최 종 석

https://orcid.org/0000-0002-0399-4675
e-mail : cjs@kist.re.kr
1994년 KAIST 전기및전자공학과(학사)
1996년 KAIST 전기및전자공학과(석사)
2001년 KAIST 전기및전자공학과(Ph.D.)
2001년~현 재 KIST 지능로봇연구단 연구원/선임연구원/책임연구원
2017년~현 재 KIST 지능로봇연구단 단장
2017년~현 재 과학기술연합대학원(UST) HCI 및 로봇 세부전공 주임교수
관심분야 : 로봇청각, 인간-로봇 사회적 상호작용, 소셜로봇

부 록

Semantic Marks and Corresponding Examples

Semantic mark	Examples
Others_Subject	남자, 여자, 아저씨, 아가씨, 머리, 목, 가슴, 다리, 손. 발, 치아, 허리, 등반, 등산, 발언불가능, 사고, 장마, 폭우 등
Others_Predicate	있다, 없다, 잃다, 일어나다, 차다, 터지다, 걷다, 나다, 먹다, 맞다 등
Fire_Subject	불, 불길, 불꽃, 불똥, 산불, 소화기, 소화전, 연기, 연막탄, 화재 등
Fire_Predicate	끄다, 번지다, 붙다, 빨갛다, 솟다, 타다 등
FirstAid_Subject	경기, 경련, 고열, 골절, 구역질, 구토, 근육통, 기상불가, 기절, 탈진, 피, 환자 등
FirstAid_Predicate	가렵다, 걷다, 일어나다, 나다, 어지럽다, 삐끗하다, 쓰러지다, 열나다, 체하다 등
Rescue_Subject	교통사고, 구조, 말벌, 벌집, 독극물, 살인, 약, 자살, 자살시도, 전복, 접촉사고 등
Rescue_Predicate	갇히다, 나다, 먹다, 맞다, 구르다, 깔리다, 싸우다, 사고나다, 잠기다, 잠그다 등
Apartment	아파트, 주공
Building	건물
Child	아이
Disability	장애
Dormitory	고시원, 교도소, 기숙사
F_Car	버스, 승용차, 트럭
Factory	공장
Familiy_Member	아버지, 어머니, 남편, 아내, 동생 등
Farm	과수원, 농가, 농장, 밭, 비닐하우스
FirstAid_Case	구급, 응급, 응급실
Flame	스파크
Forest	숲
Friend	친구
GasStation	주유소
GeneralBuiliding	가게, 고등학교, 식당, 편의점 등
Hospital	내과, 병원, 장례식장 등
In_Person	나, 저
KoreanSauna	목욕탕, 찜질방
LargeWareHouse	물류센터
Location	위치, 현장, 언덕, 회룡역 등
Market	마트, 시장
Mountain	산
MultiplexTheater	영화관
MultistoryBuilding	백화점
NursingHome	노인정, 요양원
Park	공원, 유원지
Passer_By	손님, 옆집, 지나가는사람
Rescue_Case	구조
Resort	리조트, 모텔, 캠핑장, 펜션
RicePaddy	논
Road	가로수, 도로, 전봇대, 정류장 등
SingleHouse	주택, 가정집, 빌라 등
Smell	냄새, 탄냄새
Telephone	전화
Tunnel	터널
Valley	계곡