Hidden Markov Model을 이용한 대화 의도 모델링

이승원[a], 김은솔[b], 장병탁[b]
서울대학교 전기정보공학부[a], 서울대학교 컴퓨터공학부[b]
seungwon91@gmail.com, eskim@bi.snu.ac.kr, btzhang@snu.ac.kr

Modeling of Speech Intention using the Hidden Markov Model

요 약

음성 인식 시스템은 사람의 음성 신호를 문자로 변환하는 역할을 하며, 여러 장점이 있지만 잡음과 개개인의 독특한 음성 신호 때문에 다양한 용도에 사용될 정도로 충분한 인식 성능을 얻지 못하고 있다. 사람은 말하고자 하는 의도에 따라 말할 문장을 생성하며, 이 의도는 대화상 이전에 주고받은 말의 영향을 받기 때문에 Hidden Markov Model과 같은 temporal model을 사용하면 문장의 의도를 좀 더 정확하게 판단할 수 있으며, 문장의 의도를 반영하여 음성 인식을 한다면 음성 인식 시스템의 성능을 향상시킬 수도 있다. 이 논문에서는 Hidden Markov Model을 이용한 문장 의도 분류기가 이전의 연구에서 사용한 Decision Tree와 같이 이전의 대화 내용을 고려하지 않는 분류기보다 성능이 향상될 수 있음을 제시한다.

1. 서 론

음성 인식 시스템은 사람이 말하는 음성 신호를 문자 데이터로 변환하는 시스템으로, 구글의 음성 검색 시스템이나 애플의 음성 인식 앱인 시리(Siri)가 대표적인 음성 인식 시스템이다. 음성은 사람이 별도의 기술을 학습할 필요없이 쉽게 사용할 수 있는 정보 전달 방법으로, 문자 입력 방식에 비해 짧은 시간에 많은 정보를 전달할 수 있고 손이나 발이 자유롭지 못한 상황에서도 사용할 수 있는 장점이 있다. 하지만 주변의 소음에 영향을 많이 받고 사람마다 다른 음성 신호를 갖고 있는 특징으로 인해 음성 인식 시스템의 성능이 이전에 비해 많이 발전했음에도 불구하고 아직 시스템의 신뢰도가 낮다.

사람들이 대화를 하는 상황을 고려해보면, 사람이 매 순간 임의의 문장을 말하는 것이 아니라 이전의 대화 내용과 발화자의 의도에 따라 새로운 문장을 말한다는 것을 알 수 있다. 따라서 사람들의 대화 속에서 의도를 분류하고 사람이 다음에 말할 것으로 예상되는 의도를 예측할 수 있다면, 잡음으로 인해 음성 신호가 인식에 사용되기에 좋지 못한 상황이더라도 좀 더 실제와 근접한 결과를 얻을 수 있을 것이다. 그러므로 본 연구에서는 카페의 종업원과 손님의 대화라는 제한적인 상황에 대해 temporal model 중 하나인 Hidden Markov Model(HMM)을 이용하여 각 문장들의 의도를 파악하는 분류기를 제안하려 한다.

이전의 유사한 연구로는 같은 데이터를 바탕으로 Decision Tree를 이용한 분류기를 만든 연구가 있는데, 이 연구에서는 분류할 대상인 문장 이전에 있었던 대화들을 함께 고려한다는 점에서 차별성이 있다고 할 수 있다. 또한 본 연구는 사람의 청각을 나타내는 복잡한 model과 음성 신호 특성을 이용하여 감정 및 화자의 의도를 파악하기보다는 이전 상태와의 연관성을 바탕으로 의도를 파악하고 이를 이용하여 음성 인식의 성능도 향상시킬 수 있다는 점에서 최근 연구가 활발한 speech emotion recognition과 다르다고 할 수 있다.

2. 데이터 수집

본 연구에서는 커피숍에서 대화하는 상황에서 손님과 점원 발화의 의도를 파악하는 것을 목표로 한다. 이를 위하여 실제 카페에서 일어나는 대화를 녹음기로 수집하였다. 실제로 수집한 커피숍 대화 데이터는 약 20시간 분량이며, 손님과 점원이 연속해서 발화를 주고받은 것을 기준으로 대화를 정의했을 때 130건의 대화를 수집하였다.

분류 문제로 정의하기 위하여 모든 발화 문장에 대하여 의도를 레이블로 추가하였다. 커피숍 상황에서 일어나는 대화를 구분할 수 있는 23개의 의도를 정의하였고, 직접 문장마다 의도를 태깅하였다. 수집한 실제 데이터에 대한 예시 문장들과 의도를 표 1에 정리하였다.

표 1. 실제 수집한 데이터. 본 연구에서 정의한 의도 23개와 각각에 대한 예시 문장을 정리하였다.

문장 의도	예시 문장
대화 시작/종료	대화 시작과 종료를 구분하기 위한 dummy state
인사	"안녕하세요."
주문 권유	"뭐 드시겠어요?"
주문 요청	"저기요, 주문할게요."
주문 계속	"스나이더 하나랑요"
주문 끝	"시리얼 두유 팥빙수 하나랑 카모마일 하나요."
주문 확인	"밀크 코코아랑 핫 초코요"
주문 질문	"팥빙수는 어떤 게 있나요?"
주문 변경 및 취소	"그럼 그냥 보이차로 주세요."
선택사항 확인	"아이스로 드릴까요?"
선택사항 요청	"아니요. 따뜻한 거로 주세요."
결제금액 안내	"3800원입니다."
결제 질문	"지금 내야 되나요?"
결제금액 재확인	손님이 금액을 듣고 따라하는 말. "8900원"
포장 확인	"테이크 아웃인가요?"
메뉴 안내	"견과류 과일 토핑으로 홍차, 녹차, 초코, 팥빙수가 있어요."
결제	"여기요.(카드 제시)"
카드 서명 요청	"싸인 부탁드려요"
간단한 대답(네/아니오)	손님의 질문에 대한 '네'
맞장구형 대답(네 등)	손님의 주문 중에 점원이 하는 '네'
감사 인사	"감사합니다."
기타 질문	"노트북이랑 아이패드는 여기 있는 거 쓰면 돼요?"
기타	"영수증 버려주시겠어요?"

3. 분류기 모델

이 연구에서 사람이 대화를 할 때 이전에 나온 문장의 의도를 바탕으로 새로운 의도의 문장을 말한다는 가정 하에 HMM을 적용하여 문장 의도 분류기를 구성하였다.

의도를 분류하려는 문장은 한 글자만 바뀌어도 다른 문자열이 되기 때문에 무수히 많은 종류의 값이 가능한 변수이다. 따라서 state와 evidence 모두 유한하고 이산적인 HMM으로 나타내기 위해 문장을 몇 가지 간단한 규칙에 따라 구분하여 HMM의 evidence variable로, 문장의 의도를 HMM의 state로 구성하였다. 이러한 모델 구성은 그림 1에 나타나 있다.

문장 의도는 '인사', '주문 권유', '결제 금액 안내' 등 23가지로 나누었으며, 문장은 화자 정보(종업원, 손님)와 메뉴의 포함 유무, 금액 포함 유무, 선택사항 포함 유무와 같은 기준을 종합하여 몇 개의 분류 집합을 구

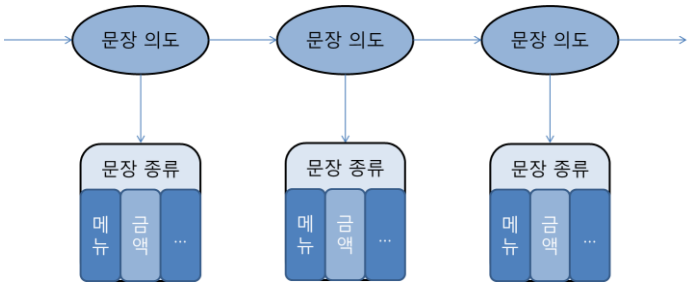

그림 1. 문장 의도 분류기 모델

표 2. 문장 분류 기준들

분류 집합	분류에 사용한 기준
분류 1	메뉴, 금액, 기타
분류 2	메뉴(완결/ 미완결), 금액, 기타
분류 3	분류 2의 기준, '선택사항'

성하였다. 문장 분류 집합에 사용한 분류 조건들은 표 2에 정리하였으며, 각 문장 분류 기준에 속하는 예시 문장들을 표 3에 나타내었다. 표 2의 문장 분류 기준에서 「분류1」은 문장에 메뉴와 금액을 포함하고 있는지의 여부에 따라 문장을 나누는 것을 의미한다. 「분류2」는 「분류1」의 메뉴 포함 기준을 세분화하여 메뉴를 포함한 문장이 완결된 문장인가의 여부도 분류 기준으로 포함하고 있으며, 「분류3」은 「분류2」에서 메뉴로 함께 고려하던 선택사항(커피의 원 샷이나 아이스 등)을 다른 분류 기준으로 나눈 것이다.

여기서 제시한 문장 분류 기준의 종류와 개수에 따라 분류기의 정확도에 차이가 발생하는데, 이는 구현 결과에서 자세히 다룰 것이다.

표 3. 분류 기준에 속하는 예시 문장들

문장 분류	예시 문장
손님 – 메뉴 포함 완결형 문장	"두유 라떼 아이스로 하나 주세요."
손님 – 메뉴 포함 비완결형 문장	"생망고 생딸기 요구르트하구요."
손님 – 선택사항 포함 문장	"아니요. 투샷으로 주세요."
손님 – 금액 포함 문장	"5800원이요?"
손님 – 기타 문장	"노트북이랑 아이패드는 여기 있는 거 쓰면 돼요?"
점원 – 메뉴 포함 완결형 문장	"두유 라떼 아이스"
점원 – 메뉴 포함 비완결형 문장	"네 매실 토닉워터랑요"
점원 – 선택사항 포함 문장	"카페모카 아이스요?"
점원 – 금액 포함 문장	"15000원입니다."
점원 – 기타 문장	"싸인해주시구요."
NULL	'대화 시작/종료'를 나타내는 state의 evidence를 표현하기 위한 dummy

4. 분류기 구현 결과

분류기의 state간 transition probability와 state-to-evidence probability를 구하고, Cross-Validation으로 성능을 test하기 위해 앞에서 설명한 카페에서 수집한 데이터를 이용하였다.

문장을 분류하는 기준에 따라 검증 결과가 달라지는데, 이 논문에서는 표 2에서 설명한 분류 집합들에 대해 비교하였으며 결과는 그림 2와 같다. 그림 2는 10-fold cross-validation을 이용하여 이전 연구와 본 연구에서의 결과를 비교한 그래프로, DT는 Decision Tree를 이용한 결과를, HMM은 이 연구에서 얻은 결과를 의미한다. HMM 뒤에 붙은 Crit1과 Crit2, Crit3는 표 1에서 설명한 분류 집합을 나타낸다. 분류 기준을 적게 사용한 HMM-Crit1의 경우 이전 연구보다

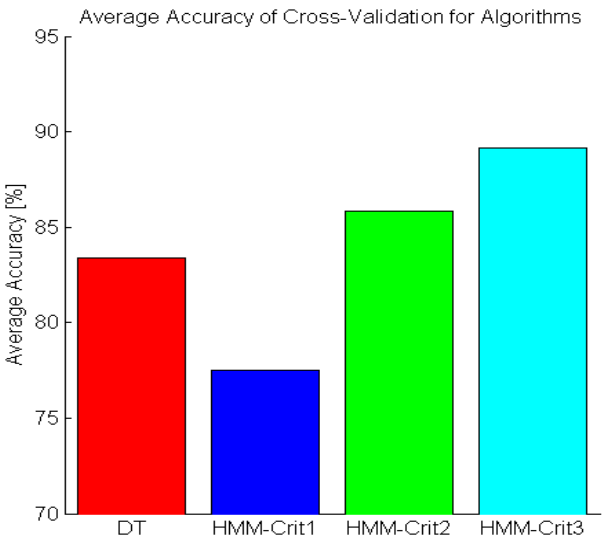

그림 2. 알고리즘별 평균 정확도(%)

성능이 안 좋지만 문장을 분류하는 세부적인 기준을 추가함에 따라 성능이 향상됨을 볼 수 있다. 이는 문장 의도가 23가지나 되고 메뉴와 관련된 의도만 10개, 금액이 포함된 것이 3개, 나머지 의도가 10개나 되기 때문에 메뉴에 대한 세부 조건이 추가될수록 성능이 향상되는 것이 당연한 결과이다.

Cross-Validation을 통한 성능 비교 외에도 HMM transition model의 conditional probability로부터 다른 정보를 찾아볼 수 있다. Transitional probability가 높은 경우는 이전 대화 의도와 다음 대화 의도 사이에 강한 인과관계가 있다는 것을 의미하기 때문이다. 그림 3은 transitional probability를 크기에 따라 색이 밝아지도록 나타낸 도식이며, 각 행은 이전 대화 의도가 주어졌을 때 다음 대화 의도의 conditional probability를 의미한다.

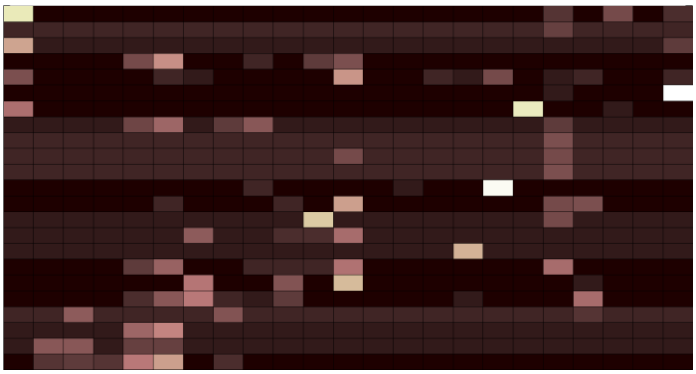

그림 3. Transitional Probability의 크기에 따른 도식. 세로축이 이전 문장의 의도, 가로축이 새 문장의 의도를 나타내며, 밝을수록 높은 확률을 의미한다. 왼쪽 아래가 '시작/종료'이며, 각 의도는 표 1의 순서를 따른다.

확률이 가장 높은 경우부터 4가지를 살펴보면, 점원이 서명을 요청한 후 손님이 특별한 의도로 분류되지 않는 말이나 행동을 하는 경우가 0.79의 확률로 가장 높았다. 이는 표 1에 나온 '기타'에 해당하는 문장 의도에 '서명하기'의 동작이 포함되어 있고, 대체로 서명이 끝난 후 다양한 종류의 대화가 이루어지기 때문이다. 그 다음으로 높은 것은 점원이 결제 금액을 안내한 후 손님이 결제하는 경우로, 확률이 0.75에 해당하며 당연한 결과이다. 손님이 결제한 후 점원이 서명을 요청하는 경우와 분류 되지 않은 의도에서 대화가 종료되는 경우가 0.61의 확률로 그 다음을 따르고 있으며, 카드 결제 후 서명을 요청하는 점과 주문부터 결제가 모두 끝난 후에 다양한 종류의 대화가 많이 이루어지는 점을 생각하면 두 결과 모두 쉽게 이해할 수 있는 것들이다.

5. 결론 및 향후 연구 과제

본 논문에서는 문장에 대한 화자의 의도를 분류하기 위해 문장을 몇 가지 기준에 따른 소규모 집합으로 분류하고 Hidden Markov Model을 이용하였다. Decision Tree를 이용한 이전의 연구와 비교할 때 보다 간단한 구조의 모델로 더 높은 성능을 얻을 수 있었다.

문장의 형태나 구성에 대한 언어학적 지식을 활용하여 성능을 향상시킬 수 있는 문장 분류 기준을 찾는 것도 가능한 연구 과제이며, 이 연구에서 구현한 문장 의도 분류기를 이용하여 문장 생성기를 구현하는 것도 확장 가능한 연구 주제이다.

참 고 문 헌

[1] Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

[2] Russell, Stuart and Peter Norvig. (2009). *Artificial Intelligence: A Modern Approach* (3rd ed.). Prentice-Hall.

[3] Yuncu, Enes, Huseyin Hacihabiboglu and Cem Bozsahin. (2014). "Automatic Speech Emotion Recognition using Auditory Models with Binary Decision Tree and SVM". Retrieved from http://www.metu.edu.tr/~bozsahin/icpr2014-emotion.pdf