공학석사학위논문

과제 지향형 대화에서의 노이즈에 강인한 대화 의도 인식

Noise Robust Dialogue Act Recognition for Task-oriented Dialogues

2015 년 8 월

서울대학교 대학원

전기컴퓨터공학부

김 태 연

초록

대화 시스템과 이메일, 게시글 요약 시스템 구축에 있어 대화 의도 분류는 중요한 역할을 한다. 이는 각각의 시스템들이 발화, 메일, 게시글 형태의 데이터에 대하여 대화 의도를 분류하고 이 정보를 하위 작업의 입력으로 사용하기 때문이다. 그래서 대화 의도 분류 성능이 해당 시스템의 전체 성능에 크게 영향을 주기 때문에 성능 향상 측면에 있어 중요하다. 대화 의도 분류는 대화 내 발화에 대화 의도를 할당하는 문제이다.

특히 대화 시스템에서는 음성 인식 에러가 존재하기 때문에 에러에 강인한 대화 의도 분류 모델이 필요하다. 따라서 본 논문에서는 두 명의 사람이 특정 목적을 가지고 진행하는 과제 지향형 대화라는 상황에서 발화, 화자, 대화 의도를 고려하여 대화 구조를 모사하는 생성모델을 만들어 노이즈 데이터에 대응하였다. 이 모델의 기반이 되는 가정은 화자는 어떠한 행위를 수행하고자 하는 목적을 가지고, 그 목적에 맞는 적절한 어휘 집합을 사용하여 상대방에게 말을 한다는 것이다. 즉 제안한 모델은 이러한 가정을 고려하여 마르코프 모델을 개선하였다.

과제 지향형 데이터인 HCRC map task, live chat, SACTI-1 말뭉치를 이용한 실험을 통해 제안한 모델이 기존 마르코프 모델에 비하여 더 나은 성능을 보이고, 현재까지도 대화 의도 분류 성능이 높은 SVM-HMM와 경쟁력 있는 결과를 보이는 것을 확인 하였다. 특히 대화 시스템의 음성 인식 모듈의 에러를 모방한 SACTI-1 말뭉치에 대하여 제안한 모델이 SVM-HMM에 비하여 노이즈에 강인함을 보였다.

주요어 : 대화 의도, 대화 의도 분류, 마르코프 모델, SVM-HMM, 대화 시스템

학 번 : 2013-23110

목차

그림 목차

표 목차

제 1 장 서론

1.1 연구의 배경

최근 스마트 기기의 보급으로 Apple의 Siri, Google의 Google Now, Microsoft의 Cortana, Samsung의 S voice, Amazon의 Echo와 같은 대화 시스템(spoken dialog system)들을 쉽게 접할 수 있게 되었다. 특히 이들은 단순한 질문에 대답만 하는 대화 시스템을 넘어 조금 더 복잡하고 구체적인 상황에 대해 인간과 대화를 할 수 있는 대화 시스템을 만들고자 한다. 이 뿐만 아니라 자연어처리 기술과 기계학습 기술의 발전에 따라 IBM의 지능형 질문 응답 시스템인 Watson이 그 가능성을 보였다. 그리고 교육, 교통, 의료, 금융 등 사회의 여러 응용 분야에서 대화시스템을 구축 및 서비스하는 시도들이 많아지고 있다.

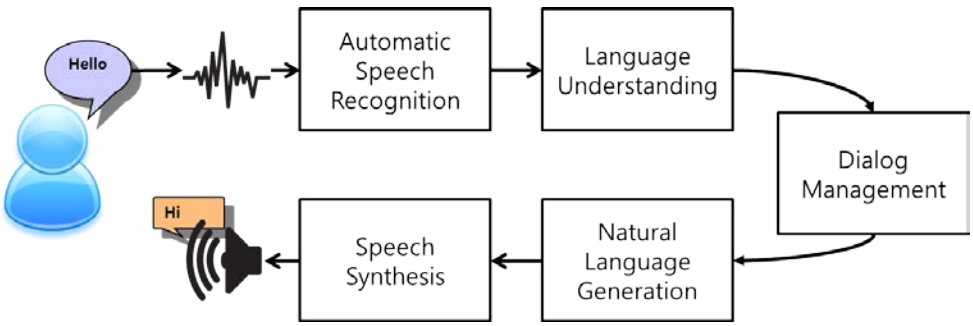

그림 1 대화시스템의 일반적인 아키텍처

이러한 대화시스템은 목적 및 기능에 따라 다양한 형태의 아키텍처를 가질 수 있지만 일반적으로 5가지의 모듈로 설명할 수 있다[10]. 이러한

모듈에는 사람의 발화(utterance)인 음성신호를 평문(plain text)으로 바꾸어 주는 자동 음성 인식(Automatic Speech Recognition, ASR)모듈, 평문으로 변환된 발화에 대해 대화시스템에서 필요로 하는 자질(feature)들을 추출하는 대화문 이해(Spoken Language Understanding, SLU)모듈, 이러한 자질들을 사용하여 대화시스템 목적에 따라 정의된 행위(action)를 수행하는 대화 관리(Dialog Management, DM)모듈, 결과로 나오는 행위를 자연어 형태로 만들어주는 자연어 생성(Natural Language Generation, NLG)모듈, 이 자연어를 사람이 인지할 수 있는 음성신호로 바꾸어주는 음성 합성(Speech Synthesis)모듈로 이루어져 있다. 특히 대화문 이해 모듈인 SLU에서는 발화의 여러 자질인 화자(speaker)의 식별 정보, 개체명(named entity), 품사(part-of-speech), 의미역(semantic role), 주제(topic), 대화 의도(dialogue act) 등을 추출한다. 여러 중요한 자질들 중에서 발화의 대화 의도를 파악하는 것이 중요한데 이는 다음 단계인 DM 모듈에서 이를 사용하여 대화시스템 목적에 맞는 시스템 행위를 취하기 때문이다[12, 4]. 대화 의도를 파악하는 SLU 모듈에서의 성능이 대화시스템의 전체적인 성능에 영향을 주기 때문에 각 기능들의 성능을 향상시키는 연구들이 자연어 처리 분야 및 기계학습 분야에서 진행되고 있다.

대화시스템에서 발화의 대화 의도 추출의 중요성뿐만 아니라 이메일, 글타래(thread) 요약 시스템(summary system)에서도 그 중요성을 보여준다[3, 13]. 이러한 시스템은 방대한 온라인 포럼의 게시글(post)나 이메일 데이터들에 대하여 요약을 해주는 것을 목표로 하는데, 기존의 문서요약 방법과 다르게 대화 의도를 사용한다. 특히 게시글과 이메일은 시간 순서로 순차적으로 기록되는 특징과 여러

사람이 참여하여 이야기가 진행되는 면에서 대화문과 그 형태가 비슷하다. 그래서 이러한 특성을 반영하여 게시글이나 이메일에 대하여 대화 의도를 분류하는 여러 연구들이 진행되고 있다.

지금까지 설명한 이러한 분야가 대화 의도를 이용하는 대표적인 응용 시스템이다. 그리고 대화 의도 추출의 정확성이 이러한 시스템의 전체적인 영향을 주기 때문에 이들의 추출 성능을 향상시키고자 여러 가지의 기계학습 접근 방법들이 제시되었다. 접근 방법들은 보통 대화문에서 추출할 수 있는 문법적(grammatical), 구조적(structural), 음향적(acoustic) 형태의 자질들을 학습 데이터로 만들어 분류기(classifier)를 학습시키거나, 대화문의 구조적인 정보인 순차성과 여러 자질들을 분류기 모델에 반영하여 모델을 정의한 후 학습시키는 연구들이 최근까지 활발하게 진행되고 있으며, 이러한 연구는 대화 의도 분류 정확도(accuracy)를 크게 향상시키는 데 기여하고 있다. 본 논문에서는 대화문의 구조적인 정보와 자질들을 사용하여 정의한 분류 모델을 이용하여 대화 의도 분류 문제를 다룬다. 이 접근방법에서는 대화의 특징 정보를 모델로 잘 모사하는 것에 따라 대화 의도 분류의 정확도를 크게 좌우하기 때문에, 대화 의도를 분류하는데 있어 대화문을 모델링을 하는 것이 상당히 중요한 작업이라 할 수 있다.

1.2 연구의 내용 및 범위

본 연구는 대화문의 특성인 발화 간에 의존 관계와 화자의 정보를 고려하여 대화 의도 분류 문제를 해결하기 위하여 입출력 은닉 마르코프 모델(input-output hidden Markov model)의 아키텍처를 변형한 모델을 제안한다. 그리고 모델의 상태(state)를 표현하는데 있어 마르코프 모델

처럼 상태는 명시적으로 표현한다. 이 모델에서는 시간에 따라 변하는 대화의 흐름 속에 그 흐름을 변화시키는 것과 제일 맞닿아 있는 대화 의도 간 의존 관계를 모델에 반영하도록 노력하였다. 또한 각 발화를 이루는 단어들의 분포형태는 현재의 대화 상황에 의해서도 다르고, 화자가 맡고 있는 역할에 따라서도 다르다. 결국 이러한 어휘집합 분포는 각 대화 의도와 화자마다 다르게 형성된다는 가정하에 이러한 특성을 모델에 반영하도록 연구하였다.

기존 연구들[15, 7, 11]도 대화의 의존 관계 및 화자의 정보를 사용하였지만, 이들 연구들은 제안하는 모델과 아키텍처가 다를 뿐만 아니라 은닉 마르코프 모델(Hidden Markov Model, HMM)을 사용한 비지도 학습(unsupervised learning) 방법이다. 그렇기 때문에 대화 의도를 모델에 상태(state)로 명확히 반영하지 못 한다. 그래서 HMM의 은닉상태(hidden state)를 역으로 대화 의도로 변환해주어야 하는 작업이 필요하고 그 성능 또한 지도 학습(supervised learning) 방법에 비하여 좋지 못하다. 또한 지도학습 방법으로 위의 대화문의 특성을 고려하여 모델을 제안한 연구들[23, 25, 22]도 있다. 이 연구에서는 SVM-HMM, CRF, Log-linear 모델 등을 이용하여 순차 데이터에 관하여 분류문제를 해결하지만, 분류기에 사용한 자질 등에서 문제가 있다. 바로 대화시스템에서 실시간 처리할 때 얻을 수 없는 자질인 대화 내 발화의 상대위치에 관한 자질과 일련의 발화에 동일하게 할당되는 대화의 세부주제(subtopic)[1]자질이다. 상대위치는 대화의 시작과 끝을 알아야지 구할 수 있는 자질인데 실제 상황에서는 대화의 끝을 알기 어렵기 때문에 이 자질은 이용할 수 없다. 또한 대화의 세부 주제 역시 얻기 어려운 정보이다. 왜냐하면 세부주제는 연속된 발화 내용들의 주제를 대표하는 것으로 그 시작과 끝을

[1] 어느 문헌에서는 게임(game)이라고 언급을 하기도 한다

알기 어렵기 때문에 실시간으로 이 정보를 얻기가 어렵다. 물론 상대위치는 통계적인 분포를 사용하여 추정된 값으로 대체할 수 있고, 세부주제 역시 주제모델(topic model) 혹은 세부주제로 이루어진 학습 집합을 이용해서 분류기로 만들어 실시간에 세부주제를 발화에 할당할 수 있다. 하지만 위의 논문들은 이러한 상황을 언급은 하였지만 실험 부분에서 전혀 고려하지 않고 정답 데이터에 있는 자질들을 그대로 사용하였다. 본 논문에서는 이러한 자질들을 전혀 고려하지 않고 오로지 대화 데이터에서만 얻을 수 있는 자질들만 이용하였다. 실험에서도 기존 연구들과 성능평가를 수행할 때 위에서 문제를 언급한 자질들을 전혀 고려하지 않고, 발화의 텍스트 정보와 의존관계 및 화자 정보만을 사용하여 제안한 모델이 최신연구와 견주어 분류 정확도를 높이는 것을 목표로 한다.

대부분의 연구들은 대화의 노이즈(noise)가 없는 말뭉치(corpus)에 대하여 실험을 진행하였는데, 이 부분에 있어 부족한 점이 있다. 실제로 대화 시스템에서 대화 의도 분류는 자동 음성 인식 ASR 모듈이 음성을 텍스트로 변환한 평문에 대하여 수행이 된다. 하지만 현재까지 ASR 모듈이 음성을 평문으로 완벽히 변환하지 못하기 때문에, 변환된 평문에는 많은 노이즈가 포함이 되어 있다. 그렇기 때문에 노이즈 데이터에 대하여 대화 의도 분류 성능을 높이는 것도 중요한 문제이다. 현재까지 이 문제를 크게 다루지 않았고, 최근에 들어서 이러한 노이즈 데이터를 사용한 연구[5, 15]가 나오고 있다. 엄밀하게 말하여 [15]는 ASR 노이즈와 다르게 트위터 데이터에 존재하는 오타, 약어, 은어들이 노이즈가 된다. 반면 ASR 노이즈는 잡음, 강세, 연음에 의한 음성학적 노이즈에 해당한다. 한 예로 ASR이 “we can ~”을 “week an ~”으로 잘 못 변환하여 노이즈를 만들 수 있다. 즉 이러한 노이즈는 단순 단어의 치환이나 알파벳을 삽입, 삭제, 교환을 통해 만드는 노이즈와 다르다. 그래서 본 논문에서는

ASR 노이즈에 대해서도 제안한 모델이 대화 의도 분류 정확도의 큰 저하 없이 잘 동작하는 것을 목표로 한다.

1.3 논문의 구성

본 논문의 구성은 다음과 같다. 2장에서는 대화문의 구성요소를 살펴보고 용어 및 대화 의도 분류 문제에 대한 정의를 한다. 그리고 3장에서는 대화 의도 분류 문제에 대하여 지도 학습 기반의 모델들에 대하여 개괄적인 설명과 성능에 대해서 이야기를 한다. 또한 대화 의도 분류 문제에 있어서 기존 연구들의 한계점들을 알아본다. 다음으로 4장에서는 모델을 제시하기 전에 설명에 필요한 배경지식을 언급하고, 대화 의도 분류를 위한 대화문의 구조적인 정보와 화자정보를 고려하여 입출력 마르코프 모델을 변형한 모델을 제안한다. 5장에서는 과제 지향형(task-oriented) 대화 데이터인 map task 말뭉치, live chat 말뭉치, SACTI-1 말뭉치를 사용하여 본 논문에서 제안한 모델과 문헌에서 성능이 제일 좋은 SVM-HMM모델을 다각도로 비교 분석을 한다. 특히 분류 정확도와 확장성(scalability), 강인성(robustness)면에서 실험적인 결과를 보인다. 마지막으로 6장에서는 본 논문의 결론을 내리고 향후 연구 방향에 대하여 논의한다.

제 2 장 문제 정의

이 장에서는 본 연구에서 다루는 대화문의 구성요소와 그 예를 살펴보고 사용하는 용어의 정의와 대화문에서 발화의 대화 의도를 인지하는 문제에 대하여 수학적으로 정의한다.

2.1 대화문의 구성요소

대화문을 분석하고 설명하기 위하여 대화문을 구성하고 있는 여러 가지 요소들에 대하여 형식적으로 정의한다.

- 대화 (Dialogue)

하나의 대화는 여러 개의 발화로 구성이 된다. 대화라는 것은 분류의 단위가 되는 발화를 묶어주는 상위 단위로 사용된다. 하나의 대화 내에서는 공통된 주제 혹은 여러 주제들에 대해서 화자 간에 대화가 이루어 질 수 있다. 본 논문에서 대화는 하나의 공통된 주제를 가지고 두 명의 사람이 진행하는 일련의 발화의 연속으로 정의한다.

$$d = u_1 u_2 \dots u_{T-1} u_T,$$

d는 대화, u는 발화, T는 대화의 길이, d에 속하는 모든 발화 u에 대하여 $da(u) \in DAs$ (단, $da(u)$는 발화 u가 속하는 대화 의도)

- 대화 말뭉치 (Dialogue Corpus)

말뭉치는 언어 연구를 위해 컴퓨터가 읽을 수 있는 형태의 텍스트들을

모아 놓은 데이터 집합 혹은 데이터 셋을 지칭하는 용어로 대화 말뭉치는 여러 개의 대화들을 모아 놓은 대화 집합을 의미한다. 이를 형식적으로 나타내면 다음과 같다.

$$DC = \{d_1, d_2, \ldots, d_{|DC|}\},$$

DC는 대화 말뭉치 혹은 대화 집합, d는 한 대화

또한 각 대화들은 말뭉치 내에 다른 대화들과 서로 식별할 수 있는 번호로 구별이 되는데, 이러한 번호를 대화 번호라고 정의한다.

- 화자 (Speaker)

화자는 대화를 이끌어 나가는 중심 인물이고, 발화를 언급하는 주체이다. 하나의 대화에는 여러 화자들이 등장할 수 있지만, 본 논문에서는 과제 지향형(task-oriented) 말뭉치를 다루기 때문에 대화는 두 명의 화자로 이루어져 있다고 본다. 주로 정보를 제공하는 화자가 있고, 정보를 제공받는 화자가 있다. 예를 들어 본 실험 부분에서 사용하는 HCRC map task 말뭉치는 길 안내 정보를 제공하는 화자 Giver(G)가 있고, 정보를 제공받는 화자 Follower(F)로 이루어져 있다. 마찬가지로 live chat 말뭉치는 상품 상담을 하는 화자 Agent(A)가 있고, 질문이나 불만을 토로하는 화자 Client(C)가 있고, SACTI-1 말뭉치는 여행 정보를 제공하는 대화 시스템인 Wizard(W)와 정보를 제공 받는 화자 User(U)가 있다.

$$S = \{A, B\},$$

S는 화자 집합, 정보를 제공하는 화자 A, 정보를 제공받는 화자 B

● 발화 (Utterance)

발화는 대화 의도를 분류하는데 있어 가장 기본적인 단위로 사용된다. 발화는 화자가 소리를 내어 말을 하여 산출된 일정한 음의 연쇄체를 뜻하며, 텍스트들로 이루어진 대화 말뭉치 범위에서 발화는 화자가 전달하고자 하는 음의 연쇄체를 텍스트 형태인 단어의 연쇄체로 표현한 것이다. 특히 한번의 발화로 여러 문장을 말을 할 수 있는데, 본 논문에서는 발화의 단위는 문장으로 본다. 즉 여러 문장으로 말을 할 경우에는 여러 발화로 분할되어 있다.

$$u = w_1 w_2 \dots w_{|u|},$$

u는 발화, w는 단어, u에 속하는 모든 단어 w_i에 대하여, $w_i \in V$

● 턴 (Turn)

앞에서 발화는 하나의 문장으로 이루어져 있고, 여러 문장으로 말을 할 경우 여러 발화로 걸쳐서 이야기 한다고 하였다. 즉 동일한 화자가 문장 단위의 발화를 이어나갈 경우 이를 턴이라고 정의한다. 턴은 하나의 발화일 경우에도 적용된다.

$$turn = u_i u_{i+1} \dots u_{i+|turn|-1},$$

$turn$은 턴, i는 턴의 시작 번호, 모든 발화 u의 화자 $s \in S$는 동일

● 문장 (Sentence)

생각이나 감정을 말과 글로 표현할 때 완결된 내용을 나타내는 최소의 단위로 주어와 서술어를 갖추고 있는 것이 원칙이지만 대화 상황에서는 생략되는 상황이 빈번하기 때문에 본 논문에서는 문장의 끝에 해당하는 구두점 ‘.’, ‘?’, ‘!’ 등을 기준으로 문장을 정의한다. 그리고 이러한 문장은

발화의 정의에 의하여 발화와 일대일로 대응이 된다. 또한 대화 말뭉치 내에서 구두점이 없이 발화가 끝이 나는 경우가 있어도 이를 하나의 문장으로 본다.

- 단어 (Word)

하나의 발화는 여러 개의 단어로 구성된다. 이러한 단어들은 대화 의도 분류기에서 발화를 분류하기 위한 자질로 사용이 된다. 기본적으로 발화를 공백문자 기준으로 토큰(token)으로 나누고, 토큰을 접미사를 자르는 어근추출(stemming) 혹은 사전에 있는 단어로 치환하는 표제어복원(lemmatization)을 수행해야지 토큰을 단어(word)[2]로 볼 수 있지만, 본 연구에서는 토큰 그 자체를 단어로 부른다. 즉 오타와 사전에 없는 은어들도 단어로 보는 것이다. 또한 채팅에서 많이 사용하는 특수문자도 대화 의도에 영향을 줄 것이라는 판단 하에 넓은 의미로 특수문자도 단어로 취급을 한다. 추가적으로 ASR를 통하여 만들어진 노이즈들도 단어로 본다. 그리고 이러한 단어들을 모아놓은 집합을 어휘 집합(Vocabulary, V)라 부른다.

- 대화 의도 (Dialogue Act)[3]

Austin과 Searle에 의하여 체계화된 화행이론(speech act theory)[1, 17]에서 나오는 개념으로 사람은 발화 속에 있는 글자 그대의 의미를 뜻하며 말하는 것이 아닌 어떠한 행위가 수행되기를 바라면서 말을 한다는 이론을 가지고 화행(speech act)를 정의하고 설명하고 있다. 즉 발화가 뜻할 수 있는 행위는 일반적으로 세 가지가 있으며 글자 그대로의 의미를 나타내는 언표적 행위(locutionary act)와 말에 숨어 있는 피상적 의미를 나타내는 언표내적 행위(illocutionary act)와 언표내적 행위로 청

[2] 엄밀하게는 텀(term)이지만 본 논문에서는 단어로 부른다.
[3] 어느 문헌에서는 화행과 구분 없이 부르기도 하고, 대화 이동(dialogue move)으로 부르기도 한다.

자에게서 얻고 싶은 효과를 나타내는 언향적 행위(perlocutionary act)가 있다. 그림 2를 설명하는 예시는 다음과 같다.

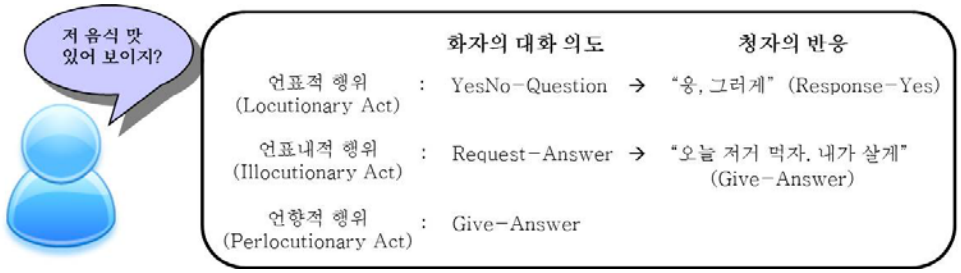

그림 2 화행이론에 대한 예시

예를 들어 "저 음식 맛있어 보이지?" 라고 물어볼 경우 이는 언표적 행위 YN-Question(예/아니요 대답을 요구하는 화행)를 뜻하고 이에 대한 응답으로 Response-Yes를 뜻하는 "네/아니요"라고 대답을 할 수 있지만, 그렇게 되면 이 대화는 자연스럽게 흘러가기가 어려울 것이다. 물론 화자의 의도가 단순한 사실에 대한 응답 물어보는 경우에는 "네/아니요"가 올바른 대답일 수 있다. 하지만 일반적인 의미에서는 '저 음식을 먹고 싶다'라는 의도를 갖는다. 이런 경우에는 청자에게 일반적인 대답으로 응답을 요구하는 언표내적 행위인 Request-Answer(일반적인 대답을 요구하는 화행)를 행하였음을 알 수 있다. 이에 대하여 화자가 청자에게 기대하고 하는 반응은 대답을 원하는 것이므로 언향적 행위로 Give-Answer(일반적인 대답을 제공하는 화행)를 행하였다고 볼 수 있다. 이 예제에서 언급하고 있는 화행은 Austin과 Searle의 논문에서 언급하고 있는 화행의 종류와 다르다. 본 저자가 예시 목적으로 임의의 화행을 표기한 것이다.

특히 대화 의도는 언표내적 행위로 발화의 의미를 전달하는 측면에서 화행과 유사한 개념이다[19]. 차이점으로는 대화시스템의 목적에 따라

조금 더 세분화 되어 화행이 정의되는 점이 다르다. 예를 들어 '질문하기' 화행이 있으면 '(대화시스템에서)무엇에 대한 질문하기'가 대화 의도가 되는 것이다. 그리고 이러한 대화 의도는 대화시스템 목적에 따라 정의되는 개수가 다양하다. 실험에서 사용하는 HCRC map task 말뭉치는 13개의 대화 의도로 이루어져 있는 반면에 Switchboard 말뭉치는 220개의 대화 의도로 이루어져 있다. 그리고 이러한 대화 의도는 화행이론을 잘 알고 있는 전문가에 의해서 대화 의도 코딩(coding) 정책에 맞게 태깅(tagging)이 된다.

- 전문가 (Expert)

전문가는 발화의 실제 대화 의도를 판단할 수 있는 사람이라고 정의한다. 즉, 전문가는 대화 의도 코딩 규칙에 따라 발화 u의 대화 의도 $da(u)$를 판단한다. 전문가는 학습 데이터 구축을 담당한다.

2.2 대화 의도 분류 문제 정의

대화문은 일련의 발화들로 구성되어 있으며, 각 발화의 대화 의도를 찾는 것이 대화 의도 분류 문제의 목표이다. 이를 수학적으로 정의를 하면 다음과 같다.

문제 정의. *대화 말뭉치 (dialogue corpus)에 있는 각 대화(dialogue)* d *가 발화(utterance)* u *의 일련의 시퀀스* $d = (u_1, u_2, \dots, u_T)$ *(*T*는 대화의 길이)로 주어지고, 대화 의도(dialogue act)* da*가* $da = \{a_1, a_2, \dots, a_m\}$ *(*m*은 대화 의도 수)가 주어질 때, 각 발화* u_i*에 해당하는 대화 의도* a_i*를 찾는 문제.*

2.3 대화문의 특징 및 문제 해결의 어려운 점

지금까지 대화문의 구성요소와 대화 의도 분류 문제를 살펴보았다. 실제 데이터 내용을 다음 표 1에 제시를 하였다. 이 데이터는 HCRC map task 말뭉치의 q1ec3 대화 번호를 갖는 대화에서 일부를 발췌한 것이다. 발화 번호 36과 38을 보면 단어의 형태는 'right'로 같지만 대화 의도는 서로 다른 'acknowledge'와 'ready' 임을 알 수 있다. 이처럼 대화 의도 분류 문제에서는 동일한 언표적 행위를 하였어도 언표내적 행위가 다를 수 있기 때문에 단어가 같아도 다른 대화 의도를 갖는 경우가 매우 빈번하다. 이러한 문제가 대화 의도 분류에 있어 제일 어려운 점이다. 즉 단순히 단어만 사용해서는 쉽게 분류하기가 어렵다.

또한 뉴스와 같은 단어의 개수가 많은 문서에 비하여 발화는 상대적으로 적은 단어로 이루어진 특징이 있다. 그렇기 때문에 단어를 bag-of-words 방식으로 분류 문제를 다루는 분류 모델 같은 것을 적용하여 큰 성능을 기대하기 어렵다. 그리고 대화는 화자간에 발화가 엇갈린 형태로 비순차적으로 진행될 수 있다. 그렇기 때문에 발화 간에 의존 정보를 모델링 하기도 어렵다. 마지막으로 실제 대화 시스템의 ASR 모듈이 평문 형태로 만들어 내는 발화에 대해서는 노이즈가 들어 있기 때문에 이러한 문제들을 해결할 필요성이 있다. 이에 대한 기존의 연구들은 3장에서 살펴보고, 4장에서는 본 논문에서 제안하는 모델을 설명할 것이다.

표 1 map task 말뭉치의 대화 번호 q1ec3의 대화문의 예시

대화 번호	턴 번호	발화 번호	화자 유형	발화	대화 의도
		…	…	…	…
	19	28	G	mmhmm	acknowledge
		29	G	um	uncodable
	20	30	F	so you have	check
	21	31	G	well	ready
		32	G	do you have an apache camp	query_yn
	22	33	F	yeah got an apache camp	reply_y
	23	34	G	okay	acknowledge
	24	35	F	but that's like at forty-five degrees south east	explain
	25	36	G	right	acknowledge
		37	G	okay then	ready
q1ec3		38	G	right	ready
		39	G	you're going to go below the diamond mine	instruct
	26	40	F	right	acknowledge
	27	41	G	do you have a graveyard	query_yn
	28	42	F	no no graveyard	reply_n
	29	43	G	don't have a graveyard	check
	30	44	F	no	reply_n
	31	45	G	do you have a desert	query_yn
	32	46	F	yes got a desert	reply_y
		47	F	which if i continue straight down past the diamond mine	explain
		…	…	…	…

제 3 장 관련 연구

대부분의 연구들이 대화 의도 분류 문제를 풀기 위하여 여러 기계학습 알고리즘을 사용하여 접근한다. 초기에는 지도 학습(supervised learning)기반의 모델이 주로 사용이 되어왔다면 최근에 들어서는 비지도 학습(unsupervised learning) 및 준지도 학습(semi-supuervised learning)기반의 모델들이 등장하고 있다. 하지만 대화 의도가 태깅 되어 있는 대화 말뭉치가 필요하지 않는 장점을 지닌 비지도 학습 기반의 방법은 아직도 지도 학습 기반 방법에 비하여 성능이 좋지 못한 문제가 있다. 본 연구에서는 비지도 학습, 준지도 학습 기법에 대하여 다루지 않기 때문에 이 부분에 관한 관련 연구는 생략한다. 그러므로 이 장에서는 분류 성능이 높은 지도 학습 기반의 알고리즘들을 설명하고 한계점들을 언급한다.

3.1 지도 학습 기반의 대화 의도 분류 연구

초기의 연구들은 룰 기반의 접근 방법[16]이었다. 이러한 방법들은 발화 속에는 대화 의도를 분별해주는 핵심 단어(cue word)와 핵심 절(cue phrase)이 있다고 보고 이들과 대화에서 추출할 수 있는 자질들을 조합하여 룰로 만든다. 그리고 변환기반학습(Transformation-Based Learning, TBL) 알고리즘을 사용하여 룰들 간에 최고의 분류 성능을 만들어 내는 최적의 룰 적용 순서를 찾아내어 이를 이용하여 대화 의도 분류를 시도한다. 특히 핵심 단어와 절을 가지고 대화 의도를 분류하는 점

에서 상당히 성공적인 연구였다. VERBMOBIL 말뭉치에 대하여 75.12% 분류 정확도를 달성하였다. 본 논문에서는 룰 기반의 방법은 다루지 않는다.

다음으로 LSA[18]를 사용하여 대화 의도를 분류하려는 시도가 있었다. 이 연구에서는 단순히 LSA에다가 대화 말뭉치에서 얻을 수 있는 여러 형태의 자질들을 행렬의 차원으로 추가하여 분류 성능을 만들어낸다. 특히 대화의 소주제(subtopic)에 해당하는 대화의 게임(game)정보를 자질로 추가 사용해서 HCRC map task 말뭉치에 73.91%라는 성능을 만들어 낸다. 하지만 게임 정보를 사용하지 않았을 때의 실험 결과는 47.09%를 보여준다. 또한 [2]에서도 게임 정보를 사용하여 대화의 의존 관계를 모델링한 모델로 같은 말뭉치에 대하여 74.9%라는 성능을 올렸다. 하지만 이 둘의 연구는 실제 대화 시스템을 서비스하는 상황에서 얻을 수 없는 소주제를 사용하기 때문에 정확한 대화 의도 분류 방법이라 할 수 없다. 소주제 정보를 사용하지 않고 map task 말뭉치에 대하여 대화 의도 분류를 실험한 연구들[9, 27]이 있다. 이들은 신경망 네트워크(Neural Network), 은닉 마르코프 모델(Hidden Markov Model, HMM), TBL을 사용하여 62~64%의 분류 성능을 얻었다. [8]저자는 대화 내에서 추출할 수 있는 자질들의 모든 조합에 대하여 NB, SVM-HMM, CRF에 학습데이터로 넣어주고 분류기의 성능 평가를 하였다. 특히 대화의 자질들 중에 발화의 상대적 위치와 이전 대화 의도, 화자의 정보가 중요하다는 것을 실험적으로 보여주었다. 하지만 여기서도 발화의 상대적 위치 자질은 실제 상황에서는 알 수 없는 정보이기 때문에 이용할 수 없다.

3.2 대화 의도의 의존 관계를 모델링 한 연구

지금까지는 발화의 의존관계 정보를 이용하지 않거나 이 정보를 학습 데이터로 구축해서 모델을 학습 시키는 연구들뿐 이였다. 이러한 정보를 직접 모델에 반영한 모델들[22, 23, 25, 19]이 연구되어왔다. [19]는 대화를 그래프 모델로 모델링을 한 첫 연구이다. 이 모델에서의 대화를 모사하는 가정은 대화는 대화 의도의 흐름이고, 발화는 대화 의도에 의하여 제시 된다는 것이다. 이 가정을 담고 있는 마르코프 모델(Markov Model, MM)은 그림 3의 아키텍처와 같다. 흰색으로 칠해진 영역은 관측 할 수 없는 값을 뜻하고, 노란색으로 칠해진 영역은 관측 할 수 있는 값을 뜻한다. 그리고 사각형 박스는 오른쪽 아래 첨자의 수만큼 해당 영역이 여러 번 반복 됨을 의미한다.

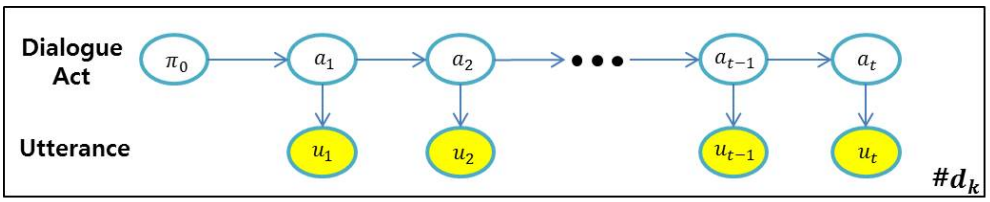

그림 3 마르코프 모델의 아키텍처

이 아키텍처에 따르면 대화 의도 간에는 전이확률(transition probability) $P(a_t|a_{t-1})$ 에 의하여 대화 의도가 변화하고 방출확률(emission probability) $P(u_t|a_t)$에 의하여 발화가 생성이 된다. 그리고 대화 의도를 추론할 때는 알고리즘 1의 Viterbi decoding 알고리즘을 사용하여 발화에 대응되는 최적의 대화 의도를 구한다.

[25]는 대화를 모델링 하기 위하여 주제(topic)[4]과 대화 의도 정보를

[4] 여기서 언급하는 주제는 앞에서 언급한 소주제, 게임과 같은 개념이다. 즉 일련의 연속된 발화에서 다루어지는 주제이다. 그렇기 때문에 LDA와 같은 주제 모델(topic model)을 통하여 구한 주제와 다르다.

사용한다. 그리고 발화의 생성확률을 만들어 내기 위하여 로그선형 모델(Log-linear model)을 사용한다. 이 모델에서는 대화의 흐름은 주제와 대화 의도에 의해서 변화가 되고, 발화는 주제와 대화 의도에 따라 제시가 된다는 가정을 가진다.

Algorithm 1 Viterbi Decoding for MM/ HMM

Input : Transition matrix $P(a_t|a_{t-1})$, emission matrix $P(u_t|a_t)$, initial state matrix $P(a_1|\pi_0)$, dialogue $d = u_1u_2 \dots u_{T-1}u_T$,

Output : The optimal path of dialogue act states $\widehat{a_1}, \widehat{a_2}, \dots, \widehat{a_T}$

initialize $\delta_{1..m,1..T}$, $\psi_{1..m,1..T}$
for $(i = 1\ to\ m)$ **do**
  $\delta_{i,1} = P(u_1|a_1 = i) * P(a_1 = i|\pi_0)$
end for
for $(t = 2\ to\ T)$ **do**
  for $(i = 1\ to\ m)$ **do**
    $\delta_{i,t} = P(u_t|a_t = i) * max_{1\le j\le m}(P(a_t = i|a_{t-1} = j) * \delta_{j,t-1})$
    $\psi_{i,t} = argmax_{1\le j\le m}(P(a_t = i|a_{t-1} = j) * \delta_{j,t-1})$
  end for
end for
$\widehat{a_T} = argmax_{1\le i\le m}\delta_{i,T}$
for$(t = T - 1\ to\ 1)$ **do**
  $\widehat{a_t} = \psi_{\widehat{a_{t+1}},t+1}$
end for
return $\widehat{a_1}, \widehat{a_2}, \dots, \widehat{a_T}$

이를 나타낸 아키텍처는 그림 4에 표현되어 있다. 이 아키텍처에서는 주

제는 이전 주제와 이전 대화 의도에 의해 전이가 일어나고, 대화 의도 역시 이전 주제와 이전 대화 의도에 의하여 전이가 일어난다. 그리고 발화는 대화 의도뿐만 아니라 주제에 의해서도 영향을 받는다는 것을 보여준다.

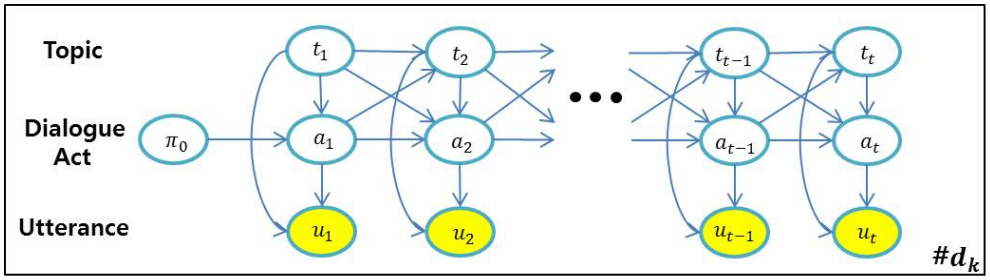

그림 4 주제 및 대화 의도를 고려한 로그선형모델의 아키텍처

이 모델 역시 대화 의도를 추론하기 위하여 Viterbi decoding 알고리즘을 사용하여 최적의 대화 의도를 구한다. 한편 이 모델의 단점으로는 실제 대화 시스템 상에서 발화에서 주제를 이용할 수 없다는 것이다. 그렇기 때문에 이 연구는 이 주제 정보를 사용할 수 있다는 가정하에 모델을 제시하였다. 본 연구에서는 이 주제 정보를 실제 상황에서 이용할 수 없기 때문에 대화 말뭉치에 따로 주제 정보를 태깅을 하여 구축하지 않았다.

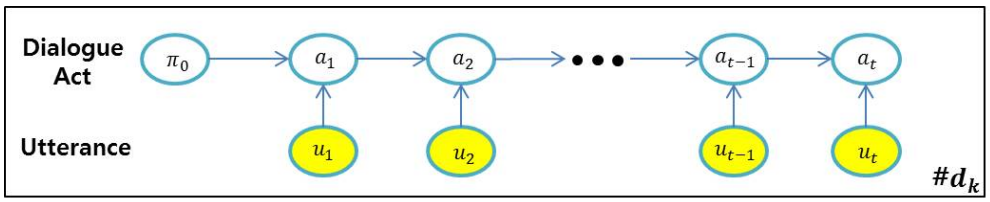

그림 5 SVM-HMM의 아키텍처

[22, 23]에서는 은닉 마르코프 모델을 변형하여 지지벡터기계(Support

Vector Machine, SVM) SVM과 결합한 SVM-HMM을 제시한다. 이 모델의 가정은 발화가 대화 의도를 결정한다는 것이다. 그림 5는 SVM-HMM의 아키텍처로 그림 3과 그 형태가 유사하다. 하지만 차이점으로는 [19]의 마르코프 모델에서는 대화 의도가 발화의 생성에 영향을 미쳤다면 이 모델에서는 가정에 따라 발화가 대화 의도에 영향을 주는 것이다. 그렇기 때문에 발화에서 대화 의도로 영향을 주는 화살표가 존재한다. 즉 이 차이로 인한 결과는 비지도 학습 방법인 은닉 마르코프 모델의 Viterbi decoding 알고리즘 결과처럼 대화 의도와 대응이 안 되는 은닉 상태의 리스트가 나오는 것이 아니라, 대화 의도와 일대일로 대응이 되는 상태로 결과가 나온다는 것이다. SVM-HMM에서 최적의 대화 의도 리스트를 구하는 알고리즘은 알고리즘 2에 제시되어 있다. 알고리즘 1과의 차이점은 각 상태 별 초기 확률 분포가 없다는 것이고, 방출확률이 없는 대신에 학습 데이터에서 학습한 SVM 모델의 결과값을 사용하는 점이 다르다. 알고리즘 2의 2~4번째 줄에서 초기 확률 분포를 사용하지 않고, 첫 번째 발화가 만들어 내는 대화 의도의 SVM 출력 값으로 초기화한다. 5~10번째 줄에서는 t를 늘려 나가면서 각 상태가 가지는 값 δ를 전이 확률 $P(a_t|a_{t-1})$과 SVM 출력 값 $P(a_t|u_t)$을 사용하여 갱신하여 나간다. 즉 이 부분에서 알고리즘 1과 차이가 있다. 상태에서 관측 값을 생성해 내는 방출확률 대신에 역 방향으로 SVM 출력 값을 사용한다는 점과 상태 전이를 통해 δ값 전이에 있어 최대(max)함수를 사용하지 않고, 합(sum)함수를 사용하는 점이다. 첫 번째 변형 부분에서는 확률적으로 정확하지는 않지만 이러한 변형이 성능을 좋게 만들었다. 그리고 알고리즘의 나머지 부분에서는 알고리즘 1과 동일하고 최적의 대화 의도 경로가 출력 값으로 만들어진다. 이 모델의 단점으로는 대화에 존재하는 여러 자질들을 SVM 분류기의 학습 데이터로만 구축해서 사용한다는 것이다. 즉 모델에서 고려하면 성능 개선의 여지가 있는 자질들을 단순히 학습 데이

터의 한 차원으로만 고려하는 한계가 있다. 또한 대화 말뭉치 및 대화 의도 태깅 정보가 변경될 때 마다 SVM의 최적 파라미터를 구해야 하는 노력이 매번 필요하다.

Algorithm 2 Modified Viterbi Decoding for SVM-HMM

Input : Transition matrix $P(a_t|a_{t-1})$, SVM output matrix $P(a_t|u_t)$, dialogue $d = u_1u_2 \dots u_{T-1}u_T$,

Output : The optimal path of dialogue act states $\widehat{a_1}, \widehat{a_2}, \dots, \widehat{a_T}$

initialize $\delta_{1..m,1..T}$, $\psi_{1..m,1..T}$
for $(i = 1\ to\ m)$ **do**
  $\delta_{n,1} = P(a_1 = i|u_1)$
end for
for $(t = 2\ to\ T)$ **do**
  for $(i = 1\ to\ m)$ **do**
    $\delta_{i,t} = P(a_t = i|u_t) * \sum_{1\le j\le m}(P(a_t = i|a_{t-1} = j) * \delta_{j,t-1})$
    $\psi_{i,t} = argmax_{1\le j\le m}(P(a_t = i|a_{t-1} = j) * \delta_{j,t-1})$
  end for
end for
$\widehat{a_T} = argmax_{1\le i\le m}\delta_{i,T}$
for$(t = T - 1\ to\ 1)$ **do**
  $\widehat{a_t} = \psi_{\widehat{a_{t+1}},t+1}$
end for
return $\widehat{a_1}, \widehat{a_2}, \dots, \widehat{a_T}$

3.3 기존 연구의 한계점

대화 의도 분류 문제에 대한 주된 연구 방향은 대화의 자질들을 어떻게 학습데이터로 구축할 것 인지와 발화 간에 의존 관계를 모델로 어떻게 표현할 수 있는 지를 연구 하는 쪽으로 진행이 되어 왔다. 특히 많은 연구들이 실제 상황에서 얻을 수 없는 소주제 정보를 이용해서 분류의 성능을 높이는 연구를 진행하였는데 이는 얻기가 어려운 자질이고, 이러한 자질을 실행시간에 분류 모델을 통하여 예측한다고 해도 분류 성능에 한계가 있기 때문에 기존 연구들이 정답 데이터의 소주제 자질을 사용해서 실험한 성능보다는 낮게 나올 것이다. 그렇기 때문에 이러한 자질을 사용하지 않는 방향으로 연구가 진행이 되어야 한다.

또한 발화의 의존 관계를 고려한 모델만 사용할 뿐 모델에 대화의 자질들을 고려하는 연구가 확인 되지 않는다. 물론 비지도 학습 기반의 방법의 은닉 마르코프 모델에 대해서는 대화의 여러 자질들을 고려하여 모델을 새로 정의하고 더 나아가 주제 모델도 함께 고려하는 모델도 연구가 되고 있지만, 지도 학습 방법에서는 그러한 모델이 없고 시도한 연구가 보이지 않는다. 따라서 이러한 부분에서 개선의 여지가 있다.

한편 실제 대화 시스템에서는 ASR 모듈의 음성인식 에러가 시스템에 크게 영향을 주는데, 이를 해결하고자 하는 연구가 현저히 부족하다. 대부분 연구의 실험에서 사용하는 말뭉치는 사람 간에 대화로 이루어진 것이고, 이들의 대화에는 텍스트로 옮겨 적을 때 생기는 노이즈가 드물게 존재하고 ASR이 만들어 내는 노이즈는 전혀 고려가 되지 않고 있다. 즉, 대부분 연구들이 ASR이 완벽하게 음성을 인식을 했다는 가정하에 실험을 수행하고 있다. 이러한 부분에서 추가적으로 ASR 노이즈에 대한 실

험을 수행해 보고, 노이즈 상황에서도 모델의 적합한지를 보여야 한다.

제 4 장 마르코프 모델 기반 대화 의도 분류

제안한 모델을 설명하기 위하여 필요한 배경지식을 먼저 설명하고 제안하는 모델에 대해서 이야기를 한다.

4.1 배경지식

4.1.1 언어모델

언어모델(language model)은 언어가 생성되는 현상을 확률적으로 모델링 한 것을 말하며, 정보검색 및 기계번역 분야에서 많이 사용이 된다. 특히 많이 사용하는 언어모델은 유니그램(unigram)과 바이그램(bigram), 트라이그램(trigram)등이 있다. 유니그램 언어모델은 단어간 의존관계 없이 단어 자체가 가지는 비율에 따라 단어가 선택되어 이루는 집합이 언어를 만든다고 보는 모델이다. 바이그램, 트라이그램 언어모델은 유니그램 모델에서 단어 간에 의존 관계가 더 추가 되어 이전 단어에 의해 단어가 선택 될 확률이 달라지고, 이 분포에 따라 단어가 선택되어 이루는 체인이 언어를 만든다고 보는 모델이다. 특히 학습 데이터에 따라 등장하는 단어의 빈도가 다르므로 언어모델을 다르게 만들 수 있다. 보통 정보검색 분야에서 쓰이는 기술로 쿼리 우도 언어모델(query likelihood language model)[14]이 있는데, 이는 검색이 되는 문서마다 고유한 언어모델 LM_d을 가지고 있어 사용자가 입력하는 쿼리 q를 생성하는 확률 $P(q|LM_d)$이 높은 문서가 검색이 되게끔 하는 모델이다. 즉 이러한 개념

을 대화 의도 분류문제에서도 생각해 볼 수 있다. 즉 어떤 화자가 특정 대화 의도를 가지고 말을 할 경우, 화자에 의해 말하여지는 발화들은 특정 언어모델에 의하여 만들어진다고 생각할 수 있다. 만약 화자와 대화 의도마다 생성되는 언어모델이 다르다면 우리는 이것을 사용하여 대화 의도 분류 문제를 해결할 수 있다. 실제로 YN-Question 대화 의도에서는 'hello'와 같은 단어가 Opening 대화 의도에서 보다 상대적으로 적게 나올 것이기 때문에 두 대화의도의 언어모델은 차이가 있다. 그리고 이는 ASR 노이즈가 있어도 대응할 수 있다. 왜냐하면 특정 대화의도마다 주로 생성되는 노이즈의 형태는 제한되어 있을 것이기 때문에, 노이즈로 인한 기존 단어의 형태는 달라지더라도 언어모델을 이루는 구성요소들의 분포는 크게 변하지 않을 것이다.

4.1.2 마르코프 모델과 은닉 마르코프 모델

순차 데이터를 처리하기 위해서 마르코프 모델과 은닉 마르코프 모델은 자주 사용된다. 모델의 복잡성 문제 때문에 보통은 1차 마르코프 가정에 따라 현재의 상태는 이전의 상태에만 영향을 받는 모델이 많이 사용이 된다. 이들 모델은 공통적으로 상태(state)의 전이로 모델이 동작을 한다. 그리고 각 상태에 따라 관측 값이 만들어 진다고 생각을 한다. 여기서 마르코프 모델과 은닉 마르코프 모델의 차이는 상태의 정의에서 생긴다. 마르코프 모델은 상태가 관측할 수 있는 관측 값의 수가 제한되어 있다. 다르게 해석을 하면 하나의 상태에서 모든 관측 값을 관측할 수 없다. 반면 은닉 마르코프 모델에서는 모든 관측 값은 모든 상태에서 관측이 될 수 있다는 것이 차이점이다. 하지만 이러한 차이로 인하여 은닉 마르코프 모델은 1차 마르코프 모델에 비하여 강력한 추론을 할 수 있다. 추론에 사용하는 방법은 앞에서 언급했듯이 관측 값을 가지고 최적의 상

태 열을 찾아주는 Viterbi decoding 알고리즘을 사용한다. 그리고 이를 대화 의도 분류문제에 적용할 수 있다고 언급하였다. 관측 값은 발화이고 상태를 대화 의도로 보고 알고리즘을 수행하여 발화의 적합한 대화 의도를 찾을 수 있다. 하지만 [19]에서 이야기 했듯이 관측 값들을 높은 확률로 만들어낼 수 있는 최적의 상태 열이 발화와 실제 대화 의도와 일치가 잘 되지 않는다. 오히려 이전 몇 개의 대화 의도만 고려해도 높은 분류 성능을 낼 수 있다고 말한다. 즉 과거의 모든 대화 의도를 다 고려해서 모델로 만들 필요 없이 단순히 이전 대화 의도만 고려를 해도 충분하다.

4.2 입출력 마르코프 모델을 변형한 대화 의도 분류 모델

배경지식에서도 이야기 했듯이 화자와 대화 의도에 따라 언어모델이 다를 것이고, 대화 의도는 이전에 모든 대화 의도에 영향을 받는 것이 아니라 단지 몇 개의 대화 의도에만 영향을 받는다는 것이다. 이를 고려하여 마르코프 모델의 아키텍처를 개선시킬 수 있다. 다음 그림 6을 보자. 그림 3에서 발화 u_t를 단어 레벨로 w_i까지 표현한 것이다.

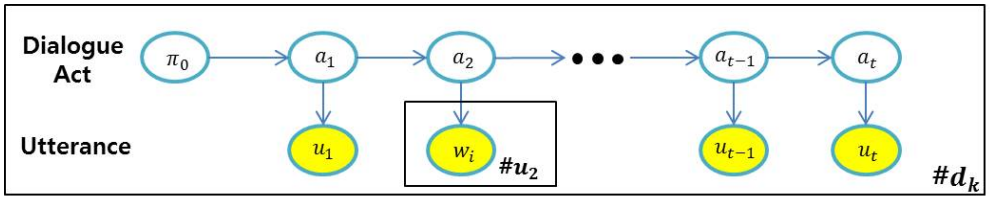

그림 6 대화 의도 분류를 위한 간단한 마르코프 모델

이는 매우 간단한 형태의 마르코프 모델이다. 전체 대화의 수만큼 반복이 되고, 각 상태 노드(node)는 대화 의도고 관측 값 노드는 발화를 뜻한다. 그리고 발화는 단어의 bag-of-words로 표현되어 있다. 이 부분에서 은

닉 마르코프 모델과 다른 점은 상태의 전이확률을 학습데이터에 있는 대화 의도 정보를 사용해서 직접 학습시킬 수 있는 것이다.

보통 마르코프 모델에서도 은닉 마르코프 모델에서처럼 관측 값의 대화 의도를 추정할 때 Viterbi decoding 알고리즘을 사용한다. 하지만 실험적으로 성능이 좋지 않기 때문에 본 논문에서는 대화 의도간에 일차적인 전이 관계만을 사용하여 대화 의도를 추정한다. 이는 알고리즘 3에 제시하였다. 이 모델에서는 t 시점의 a_t를 추정하기 위해서 a_{t-1}에서 a_t로 전이 되는 전이확률을 곱하고 a_t에서 발화 u_t를 생성하는 방출확률을 곱한 후 이 값을 크게 하는 a_t가 t 시점 대화 의도가 된다. 여기서 전이확률과 방출확률은 학습데이터에서 최대 우도 추정법(the method of maximum likelihood estimation)을 사용하여 확률 값들을 추정할 수 있다. 최대 우도 추정법은 데이터에서 관측되는 값들을 단순히 셈을 통하여 구할 수 있다.

Algorithm 3 Simple Dialogue Act Prediction

Input :	Transition matrix $P(a_t\|a_{t-1})$, language model-based emission matrix $P(u_t\|a_t)$, current utterance u_t, previous dialogue act a_{t-1}
Output :	The predicted dialogue act
1	**initialize** da
2	$da = argmax_{1\le i\le m} P(a_t = i\|a_{t-1}) * P(u_t\|a_t = i)$
3	**return** da

그림 7은 기존 마르코프 모델에서 화자 정보를 입력으로 넣어 모델의 아키텍처를 입출력 마르코프 모델과 비슷한 형태로 개선시킨 것이다. 즉

화자는 대화 의도에도 영향을 주고 발화의 생성에도 영향을 준다. 그리고 다음 사람의 대화 의도에 영향을 주게 된다.

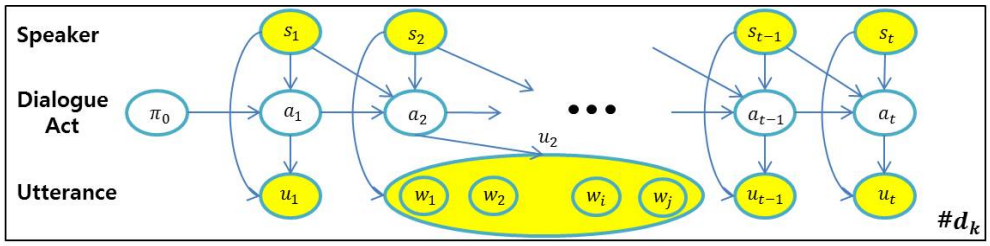

그림 7 화자정보를 반영한 변형된 입출력 마르코프 모델

이 모델을 사용하여 대화 의도를 구하기 위해서는 다음 수식 1의 결합 확률(joint probability)을 구해야 한다

$$\begin{aligned} & P(a_t, a_{t-1}, u_t, s_{t-1}, s_t) \\ &= P(u_t|a_t, s_t) * P(a_t|a_{t-1}, s_{t-1}, s_t) \\ & * P(a_{t-1}) * P(s_{t-1}) * P(s_t) \end{aligned} \qquad \text{수식 1}$$

이는 t 시점의 상태 a_t 와 연결 되어있는 베이지안 네트워크(Bayesian network)상에 노드들의 결합확률을 구하는 것과 같다. 그리고 이식에서 우리가 궁금한 대화 의도 a_t를 찾는 것이므로 결합확률을 최대화하는 a_t 찾는 것으로 볼 수 있다.

$$\begin{aligned} & argmax_{a_t} P(a_t, a_{t-1}, u_t, s_{t-1}, s_t) \\ &= argmax_{a_t} P(u_t|a_t, s_t) * P(a_t|a_{t-1}, s_{t-1}, s_t) \\ & * P(a_{t-1}) * P(s_{t-1}) * P(s_t) \end{aligned} \qquad \text{수식 2}$$

$$= argmax_{a_t} P(u_t|a_t, s_t) * P(a_t|a_{t-1}, s_{t-1}, s_t) \qquad \text{수식 3}$$

즉 수식 2처럼 유도할 수 있으며, a_t에 관하여 $P(a_{t-1}) * P(s_{t-1}) * P(s_t)$는 상수 이므로 제거한다. 그러면 최종 수식 3을 얻을 수 있다. 이것의 의미를 살펴보면 $P(u_t|a_t, s_t)$는 화자가 있고 화자가 말하고자 하는 대화 의도가 있을 때, 발화를 생성하는 언어모델로 볼 수 있다. 즉 앞에서 언급한

이야기를 모델에 반영한 것이다. 반면 $P(a_t|a_{t-1}, s_{t-1}, s_t)$는 상태 간에 전이확률을 뜻하는 것으로 대화 의도의 전이는 이전 대화 의도와 이전 화자와 현재 화자에 따라 전이되는 것으로 모델링을 하였다.

이제는 각 확률들의 값을 추정해야 한다. 전이확률 $P(a_t|a_{t-1}, s_{t-1}, s_t)$는 발화에 대화 의도가 태깅 되어있는 학습 데이터를 이용하여 최대 우도 추정법을 사용하여 추정한다.

$$P(u_t|a_t, s_t) = \prod_{w \in u_t} P(w|a_t, s_t) = \prod_{w \in u_t} \frac{\#(w, a_t, s_t) + 1}{\#(a_t, s_t) + |V|} \quad \text{수식 4}$$

그리고 발화의 생성확률인 $P(u_t|a_t, s_t)$를 구해야 한다. 이 역시 학습 데이터를 사용하여 최대 우도 추정 법을 사용한다. 특히 a_t와 s_t를 만족시키면서 관측되는 값들을 셈하여 한다. 그리고 셈이 되지 않는 단어들에 대해서는 보간 방법 중에 하나인 add-one smoothing 기법을 사용하여 값을 보간 한다.

Algorithm 4 Proposed Dialogue Act Prediction

Input : Transition matrix $P(a_t|a_{t-1})$, language model-based emission matrix $P(u_t|a_t)$, current utterance u_t, current speaker s_t, previous speaker s_{t-1}, previous dialogue act a_{t-1},

Output : The predicted dialogue act

1 initialize da

2 $da = argmax_{1 \le i \le m} P(a_t = i|a_{t-1}, s_t, s_{t-1}) * P(u_t|a_t = i, s_t)$

3 **return** da

수식 3으로 돌아와서 앞에서 최대 우도 추정법으로 구한 확률 값들을 곱한 것들 중에서 제일 확률 값들을 크게 만드는 a_t를 찾는다. 즉 이것이 발화 u_t의 대화 의도가 된다. 이는 알고리즘 4에 나타내었다.

제 5 장 성능 평가

5.1 대화 말뭉치

실험에서는 길 안내 목적으로 만들어진 HCRC map task 말뭉치[5]와 상품의 트러블슈팅에 관한 대화로부터 수집한 live chat 말뭉치[6], 대화 시스템 구축을 위하여 만들어진 SACTI-1(Simulated ASR-Channel: Tourist Information, part 1) 말뭉치[7]를 사용한다. Map task 말뭉치, live chat 말뭉치, SACTI-1 말뭉치의 통계 정보는 표 2에 정리되어 있다. 표 2에서 알 수 있듯이 map task 말뭉치가 다른 두 말뭉치에 비하여 발화 개수의 크기가 더 크다. 대화당 평균 발화 수를 보면 대화가 7배 정도로 더 길게 진행 됨을 알 수 있다. 그리고 어휘 집합의 크기를 살펴보면 두 말뭉치에 비하여 live chat의 어휘의 수가 더 많음을 알 수 있다. 이는 live chat이 채팅 기반으로 기록된 말뭉치이기 때문에 오타와 같은 단어의 노이즈 뿐만 아니라 동일한 상품 개체 명에 대해 여러 가지의 약어가 많고, 명사나 동사, 특수문자 등을 채팅 은어로써 표현하기 때문에 어휘 집합의 수가 더 큼을 알 수 있다. 전체 어휘 개수를 살펴보면 map task의 발화 수가 많기 때문에 다른 말뭉치에 비하여 많은 양의 단어로 이루어져 있음을 알 수 있다. 그리고 이 말뭉치들은 특정 문제를 가지고 대화가 행하여지는 특징이 있다. Map task는 한 사람이 길을 안내하고 다른 사람은 안내에 따라 목적지까지 도착할 때까지 이루어지는 대화를 기록해 놓은 것이고, live chat은 문제가 있는 상품의 해결책 정보를 얻기

[5] http://groups.inf.ed.ac.uk/maptask/
[6] 산학협력 연구과제에서 제공 받은 데이터임
[7] http://mi.eng.cam.ac.uk/projects/sacti/corpora/SACTI-1/

위하여 유저가 온라인에서 에이전트와 대화한 채팅 내용을 기록해 놓은 것이고, SACTI-1은 여행 가이드 주제에 대한 두 사람의 대화를 WOZ(Wizard-Of-Oz)방식을 통하여 기록해 놓은 것이다.

표 2 map task 말뭉치와 live chat 말뭉치 비교

	HCRC map task	live chat	SACTI-1
대화 개수	128	152	144
발화 개수	27,083	5,148	5,195
대화당 평균 발화 수	211.6	33.9	36.1
어휘 집합의 크기 (전처리 수행함)	2,137	3,367	1,534
전체 어휘 개수 (전처리 수행함)	160,092	62,022	65,507
대화 유형	과제 지향형 (task-oriented)	과제 지향형 (task-oriented)	과제 지향형 (task-oriented)
화자 수	2 (giver, follower)	2 (agent, client)	2 (wizard, user)
대화 의도 개수	13	18	14
ASR 노이즈 유무	없음	없음	있음
대화 기록 형태	말을 글로 옮김 (transcript)	채팅 (chatting)	말을 글로 옮김 (transcript)
kappa 상관계수	0.83	0.51	-

이 WOZ 방식은 실제 대화 시스템 역할인 위자드(wizard)를 사람이 맡게 하고, 유저(user)가 대화 시스템을 사람이 아니라 기계로 인식하게 하여 둘 간에 대화를 진행시키는 방법이다. 이러한 말뭉치들은 모두 특정 목적을 이루고자 하는 과제 지향형 특징을 갖고 있고, 두 사람이 역할을 가지고 대화하는 상황에서 제안한 모델의 성능 평가를 함에 있어 적합한 말뭉치들 이다. 특히 SACTI-1은 다른 두 말뭉치와 다르게 대화 시스템을 모사하여 만들어진 말뭉치이기 때문에, 대화 시스템의 ASR 모듈에서

만들어지는 노이즈도 포함한다. 이 노이즈는 특정 확률 분포를 따라 단순히 단어를 치환해서 만들어 낼 수 없다. 왜냐하면 말의 연음과 같은 발음 문제로 ASR이 단어를 추가로 생성하거나 삭제할 수 있기 때문에 단어 치환 방법과 다른 노이즈 생성 방법이 요구된다. 그래서 실제 노이즈와 비슷하게 만들어 내기 위하여 음성 혼동 모델(phonetic confusion model)과 언어모델을 사용하고[26, 20] 이 모듈을 노이즈 생성 모델 혹은 노이즈 채널이라 명명한다. 그리고 노이즈 정도는 수식 6의 단어 에러율(Word Error Rate, WER)을 사용하여 나타낸다. 이는 올바른 문장과 노이즈 문장 사이의 단어 단위인 Levenshtein 거리로 생각할 수 있다. 삽입(insertion), 삭제(deletion), 치환(substitution) 연산 사용한 회수를 전체 단어의 수로 나눈 비율로 계산된다.

$$WER = \frac{\#\ of\ word\ insertions,\ deletions,\ and\ substitutions}{\#\ of\ words} \qquad \text{수식 } 5$$

그리고 이 비율에 따라 SACTI-1 말뭉치는 4개의 말뭉치로 나눌 수 있는데. WER이 0%인 none 그룹, WER이 32%인 low 그룹, WER이 46%인 Med 그룹, WER이 63%인 Hi 그룹으로 구별이 된다. 이 에러율 수치는 말뭉치 제작자가 정한 수치이다.

이러한 노이즈를 바탕으로 시스템 역할을 하는 위자드와 유저는 대화를 진행하게 되는 과정을 설명하면 다음과 같다. 먼저 유저가 말을 하면 이를 타자수가 텍스트로 타이핑을 한다. 그리고 이 텍스트를 ASR 노이즈 생성 모델에 넣어 노이즈를 유발시킨다. 이 노이즈가 생성된 텍스트를 위자드가 보고 유저와 대화를 진행하게 된다. 그리고 유저는 다시 말을 하고 이 과정이 반복이 된다. 항상 우선순위는 유저의 말이기 때문에 위자드 말 중간에 유저가 말을 하면 잠시 보류가 된다. 그리고 대화 의도

분류 모델 학습에서는 타자수가 타이핑한 텍스트와 이를 ASR 노이즈 생성 모델에 넣어 노이즈가 유발된 텍스트 2개를 사용하여 학습하고 이 둘의 결과를 비교해본다.

단어의 각 말뭉치는 발화에 대화 의도가 태깅이 되어 있는데, map task는 13개의 대화 의도를 갖고, live chat은 18개의 대화 의도, SACTI-1은 14개의 대화 의도를 갖는다. SACTI-1은 Traum이 제안한 conversation act[24]의 speech act와 grounding act를 참고하여 태깅이 되었다. 그리고 한 발화에 대하여 여러 개의 대화 의도를 가질 수 있는데, 본 실험에서는 첫 번째로 태깅된 대화 의도를 사용한다. 다음 표 3, 4, 5는 각 말뭉치에서 대화 의도 비율과 해당 말뭉치에서 정의된 대화 의도의 설명을 기록해 놓았다. 이 말뭉치들의 대화 의도 태깅의 신뢰도는 Cohen의 kappa 상관계수로 나타낼 수 있는데, map task는 0.83이고, live chat은 0.51이다. 이 수치로부터 live chat이 map task에 비하여 태깅 일치율이 많이 낮음을 알 수 있다. 그리고 SACTI-1은 kappa 상관계수가 문헌에 나와 있지 않아 표기하지 못 하였다.

표 3 map task 말뭉치의 대화 의도에 관한 설명 및 통계정보

Dialogue Act	% of Corpus	Description
Acknowledge	20.7%	대화를 경청하고 있음을 보여주는 행위를 뜻함
Reply_y	11.9%	예와 같은 긍정적인 대답을 제시하는 행위를 뜻함
Check	7.9%	불확실한 정보에 대하여 파트너에게 맞는지 물어보는 행위를 뜻함
Align	6.6%	파트너의 반응 및 동의를 확인 하는 행위를 뜻함
Clarify	4.4%	이미 언급한 정보를 다시 알려주는 행위를 뜻함
Reply_n	3.3%	아니오 같은 부정적인 대답을 제시하는 행위를 뜻함
Uncodable	0.1%	12개 대화 의도에 속하지 않는 것
Instruct	15.8%	파트너에게 어떠한 행동을 지시하는 행위를 뜻함
Explain	8.0%	파트너가 이해하지 못한 것에 대하여 설명해주는 행위를 뜻함
Ready	7.6%	새로운 주제로 대화의 시작을 파트너에게 알리는 행위를 뜻함
Query_yn	6.5%	예/아니요 질문을 하는 행위를 뜻함
Reply_w	3.4%	임의의 대답을 하는 행위를 뜻함
Query_w	2.9%	임의의 질문을 하는 행위를 뜻함

표 4 live chat 말뭉치의 대화 의도에 관한 설명 및 통계정보

Dialogue Act	% of Corpus	Description
PresentSolution	16.5%	문제의 해답을 제시하는 행위를 뜻함
PresentProblem	10.8%	문제를 제시하는 행위를 뜻함
OtherAnswer	8.9%	임의의 짧은 답을 하는 행위를 뜻함
YNQuestion	7.3%	예/아니요 질문을 하는 행위를 뜻함
Opening	4.5%	대화를 시작하는 행위를 뜻함
Closing	2.1%	대화를 끝내는 행위를 뜻함
ConnectionCheck	1.7%	대화가 지속되고 있는지를 확인하는 행위를 뜻함
Wait	1.7%	대화를 잠시 멈추고자 하는 행위를 뜻함
ProblemSolved	0.6%	문제가 해결 되었음을 알리는 행위를 뜻함
Acknowledgement	13.8%	대화를 경청하고 있음을 보여주는 행위를 뜻함
OtherStatement	10.7%	임의의 긴 답을 하는 행위를 뜻함
YesAnswer	7.5%	예와 같은 긍정적인 대답을 제시하는 행위를 뜻함
OpenQuestion	6.8%	임의의 질문을 하는 행위를 뜻함
ProblemNotSolved	2.3%	문제가 해결되지 않았음을 알리는 행위를 뜻함
ConfirmProblem	1.9%	문제를 파악하는 행위를 뜻함
NoAnswer	1.7%	아니오 같은 부정적인 대답을 제시하는 행위를 뜻함
Transfer	0.6%	다른 담당자로 교환을 하는 행위를 뜻함
StateCause	0.5%	문제의 증상을 언급하는 행위를 뜻함

표 5 SACTI-1 말뭉치의 대화 의도에 관한 설명 및 통계정보

Dialogue Act (+Grounding Act)	% of Corpus	Description
Request	28.6%	질문 혹은 요청을 하는 행위를 뜻함
Inform	25.0%	진술, 사실의 반복, 욕구, 예/아니요 질문에 대한 응답을 알리고자 하는 행위를 뜻함
ExplAck	18.4%	대화를 경청하고 이해하고 있음을 알리는 행위를 뜻함
StateInterp	8.7%	상대방의 가설, 생각, 의도를 언급하여 묻는 행위를 뜻함
GreetingFarewell	6.4%	대화의 시작과 끝을 알리는 행위 혹은 잡담을 하는 행위를 뜻함
DisAck	4.3%	이해하고 있지 못하는 현재 상황을 알리는 행위를 뜻함
RejectOther	2.4%	ReqAck가 선행하지 않고, 상대방의 의도, 해석을 부정하는 행위를 뜻함
ReqRepeat	1.6%	반복을 요구하는 행위를 뜻함
RespondNegate	1.4%	ReqAck에 대한 부정적인 응답
RespondAffirm	1.3%	ReqAck에 대한 긍정적인 응답
IncompleteUnknown	0.9%	노이즈 혹은 의미가 없는 발화를 뜻함
UnsolicitedAffirm	0.6%	상대방이 잘 이해하고 있음을 알리는 행위를 뜻함
HoldFloor	0.2%	대화를 잠시 멈추거나, 발언을 유지하는 행위를 뜻함
ReqAck	0.2%	상대방이 대화를 따라오고 있는지 확인하는 행위를 뜻함

5.2 비교모델 및 개발환경

본 논문에서 제안하는 대화 의도 분류기의 비교 모델로 다음의 모델들을 사용한다. 말뭉치에 제일 비중이 높은 대화 의도를 발화에 할당하는 모델인 Zero-R을 기본적인 베이스라인(baseline)으로 사용을 한다. 그리고 문서 분류에 있어 성능이 좋고 제일 많이 사용하는 Naïve Bayes(NB)와 SVM을 비교 모델로써 이용한다. 텍스트 자질에 대해서는 [8]의 연구 결과의 것을 사용한다. 즉 발화에 있는 단어들의 유니그램(unigram)과 이에 대응하는 빈도수 대신에 존재 유무만 고려하는 불린(Boolean) 값으로 바꾸어 모델을 학습시켰다. 이 부분에서 대화 내 상대적인 위치는 자질로 고려하지 않았다. 또한 앞부분의 모델인 NB와 SVM에 추가적으로 화자 정보와 대화문의 발화간 의존 정보라고 할 수 있는 이전에 대화 의도 정보를 자질로 추가하여 모델을 학습시켰다. 이때 대화 의도 정보는 과거 이력을 고려하여 여러 개를 넣어줄 수 있지만 본 실험에서는 이전 시점의 대화 의도만 넣어주었다. 이를 NB(+S+DA)와 SVM(+S+DA)으로 부르겠다. 그리고 이들의 구현은 Weka 라이브러리[8]를 사용하여 구현을 하였다. 그리고 CRF, LSA 모델 등은 이전 실험에서 SVM-HMM 모델 보다 성능이 안 좋게 나와 고려하지 않았고, SVM-HMM은 현재까지 성능이 제일 좋은 모델이기 때문에 이 모델을 이기는 것을 목표로 한다. 이 모델의 구현은 Cornell 대학에서 제공하는 라이브러리[9]를 사용한다. 이때 사용하는 자질은 앞에서와 같이 화자 정보와 단어의 유니그램이다. 그리고 이를 SVM-HMM(+S)으로 명명한다. 그리고 SVM-HMM의 최적의 파라미터 c와 e값을 구하기 위하여 학습 데이터의 일정 부분을 사용하여 파라미터 값을 변화시키면서 정확도가 제일 높게 나오는 c와 e를 택하였다. 본 실험에서 사용하는 데이터에서는

[8] http://www.cs.waikato.ac.nz/ml/weka/
[9] https://www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html

map task 말뭉치에서는 c가 34.5, e는 0.5이고, live chat 말뭉치에서는 c는 7, e는 0.5일 때 성능이 제일 높았다. SACTI-1 말뭉치에서는 e는 0.5이고, c는 단어 에러율에 따라 최적의 c값 1, 3.5, 2.5, 2를 선택하였다. 마지막으로 본 논문에서 제안하는 모델의 명칭은 MM(+S)으로 한다. 이는 화자 정보와 단어의 유니그램 모델을 사용하기 때문에 명명하였다. 그리고 기존 마르코프 모델(Markov Model, MM)보다 성능이 높음을 보이기 위하여 마르코프 모델도 비교모델로 사용하고, 이를 MM으로 명명한다. 이때 상태 추론 기법으로는 Viterbi decoding 알고리즘이 쓰일 수 있지만, 이전 대화 의도만 가지고 현재 대화 의도를 추론한 것이 성능이 더 높아 이 방법을 사용한다.

5.3 성능 평가 측정치

분류 문제에 대하여 성능 평가 측정치로 많이 사용하는 것이 precision과 recall이 있다. 이는 다음 표 6의 혼동 행렬(confusion matrix)을 통하여 구할 수 있다.

표 6 분류 결과의 혼동 행렬

		분류기 예측 결과	
		Positive	Negative
실제 정답	Positive	True Positive(TP)	False Negative(FN)
	Negative	False Positive(FP)	True Negative(TN)

$$precision = \frac{TP}{TP + FP} \quad \text{수식 6}$$

$$recall = \frac{TP}{TP + FN} \quad \text{수식 7}$$

Precision은 분류기가 정답이라고 말한 것들(TP+FP) 중에 실제 정답

의 개수(TP)의 비율로 분류기의 예측 정확성을 알려주는 척도로 수식 7과 같다. Recall은 실제 정답인 것들(TP+FN) 중에 분류기가 예측한 결과의 실제 정답 개수(TP)의 비율로 분류기가 얼만큼 많은 정답을 예측하였는지 알려주는 척도로 수식 8과 같다. 이 둘의 관계는 보통 반비례하며 분류 문제에서는 이 둘의 성능을 최대화 하는 쪽으로 연구가 진행이 된다. 이 두 측정치는 분류기를 평가하는데 있어 중요하기 때문에, 이 둘의 중요성을 모두 반영한 측정치인 F1 측정치가 수식 9에 제시 되어 있다. 그리고 전체 분류의 성능을 나타내는 accuracy 측정치가 수식 10와 같다. 이는 다음과 같이 구할 수 있다.

$$F1 = \frac{2 * precision * recall}{precision + recall} \quad \text{수식 8}$$

$$accuracy = \frac{correctly\ classified\ instances}{total\ number\ of\ instances} \quad \text{수식 9}$$

위에서 설명한 recall과 precision, F1 측정치는 각 분류 카테고리의 성능을 개별적으로 평가할 때 사용하는 것이고, accuracy는 전체 카테고리에 대한 성능을 평가할 때 사용하는 것이다. 본 실험에서는 k-fold cross validation을 수행한 후 분류기의 성능을 평가해야 하기 때문에, 각 fold에서 얻어지는 recall, precision, F1의 평균치와 accuracy를 전체 k개의 fold 개수만큼을 평균을 내는 macro-averaging 방법을 사용하여 구한다. 그리고 이 측정치들을 이용하여 전체적인 대화 의도 분류기의 성능을 평가한다.

5.4 실험 결과 및 분석

실험은 각 말뭉치 별로 5-fold cross validation을 수행하였다. 보통은

각 fold가 균등한 데이터 크기를 가져야 하지만 말뭉치의 대화문들은 발화의 개수가 서로 다르기 때문에 균등한 크기로 나누기가 어렵다. 그렇기 때문에 발화의 개수가 균등하지는 않지만, 동일한 대화 개수를 가지도록 대화들을 나누고 모델 평가에 관한 실험을 진행하였다.

5.4.1 분류 성능

그림 8, 9, 10은 map task 말뭉치와 live chat 말뭉치에 대하여 베이스라인 모델들과 본 논문에서 제안하는 모델의 성능을 보여주는 실험 결과이다. X축에는 평가 측정치인 precision, recall, F1, accuracy를 나열하였고, Y축은 각 평가 측정치의 0~1 사이의 값을 보여준다. 따라서 막대 그래프가 높이 있을수록 각 측정치 관점에서 모델의 성능이 높다고 말할 수 있다. 특히 4가지의 평가 측정치 중에서 accuracy를 중요하게 살펴볼 것이다. 여러 문헌에서는 분류기의 성능을 비교하기 위하여 F1보다 accuracy를 많이 사용한다[18, 8, 2, 23]. 이는 F1값이 높게 나와도 전체 분류되는 개수의 차이가 많이 생기기 때문에 accuracy를 더 중요한 측정치로 여긴다. 비슷한 이유로 분류기가 빈도가 적은 대화 의도를 하나도 예측하지 못 하여 precision, recall, F1의 값이 0이 나오는 경우가 있다. 이런 경우 모든 대화 의도를 macro-averaging 하게 되면 측정치의 평균 값이 낮아지는 문제도 있다.

이제 각 모델 별 실험 결과를 분석해보자. 먼저 두 말뭉치에서 Zero-R을 제외하고 NB의 accuracy가 다른 모델들에 비하여 제일 낮다. 그 뒤로는 SACTI-1 말뭉치를 제외하고 SVM이 성능이 낮았다. 즉 단순히 발화의 단어들만 보고 대화 의도를 분류하기 때문에 일련의 발화에 따라 대화 의도가 변하는 특성을 해결하지 못 하는 점을 알 수 있다. 그래서

화자 정보와 이전 대화 의도를 추가하여 만든 학습 데이터를 학습한 NB(+S+DA)와 SVM(+S+DA)의 결과를 보면 두 말뭉치에서 성능 향상이 되었음을 알 수 있다. SACTI-1 말뭉치에서 SVM(+S+DA)는 SVM과 비교하여 성능 향상이 보이지 않는데, 이는 이전 대화 의도를 학습 데이터의 자질로 사용하는 것이 이 모델에서 적합하지 않음을 보여주는 결과로 해석된다. 또한 이 모델 모두 MM보다 성능이 낮은데, 이 결과 역시 대화 의도를 학습 데이터의 자질로 사용하는 것이 부적합한 것으로 보여준다.

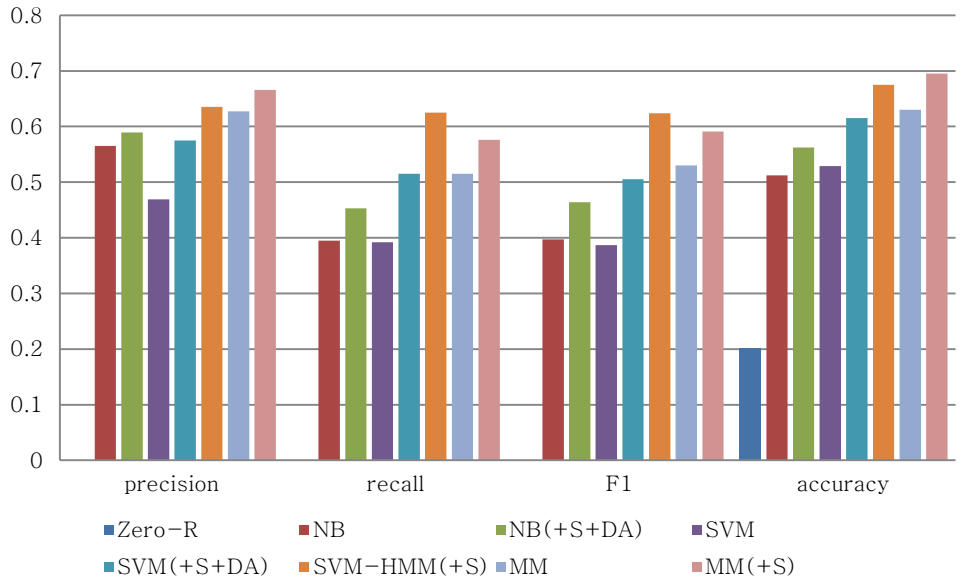

그림 8 map task 말뭉치에서 각 측정치 별 모델들의 성능

또한 대화 의도의 의존 관계를 고려한 제일 기본적인 마르코프 모델 MM에서 Viterbi decoding 알고리즘을 사용한 결과는 그래프에는 표기하지 않았지만 이전 대화 의도만 가지고 추론하는 방법보다 좋지 못하였다. 즉 대화 의도가 서로 의존 관계가 있어도 모든 대화 의도간에 의존 관계를 고려하여 최적의 상태 열을 찾아내는 Viterbi decoding 알고리즘은 이전 대화 의도만 고려하는 것보다 성능이 좋지 못하였다. 결국 대화

의도라는 것은 이전에 모든 대화 의도에 영향을 받는 것이 아니라 이전 몇 개의 대화 의도에 의하여 영향을 받는다는 것을 알 수 있다.

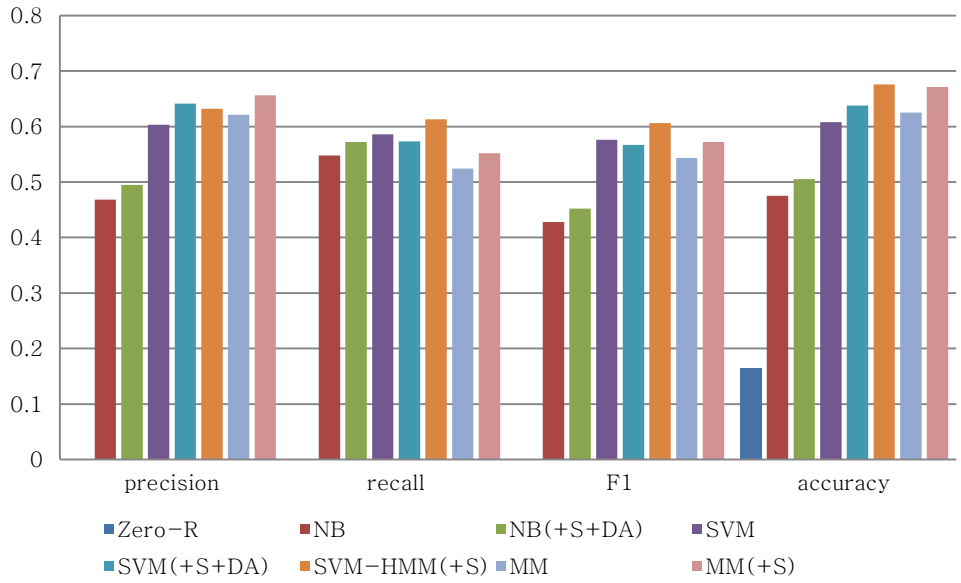

그림 9 live chat 말뭉치에서 각 측정치 별 모델들의 성능

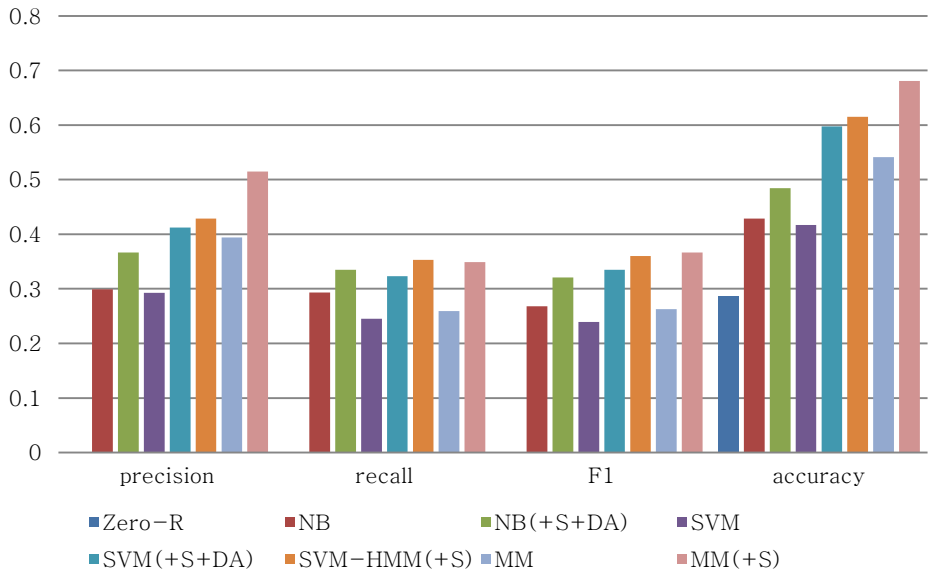

그림 10 SACTI-1 말뭉치에서 각 측정치 별 모델들의 성능

그리고 제안한 모델인 MM(+S)은 SVM-HMM(+S)와 비교하여 map task 말뭉치에서는 accuracy가 2.5%이상 높았으며, live chat 말뭉치에서는 SVM-HMM(+S)와 비교하여 0.4%가 낮았다. 이 실험결과에 대하여 대응표본 t검증(paired t-test)를 수행하여 얻은 결과는 map task 말뭉치에 대해서는 p값이 0.01으로 유의수준 0.05보다 낮아 두 실험 성능은 차이가 없다라는 귀무가설이 기각되어 MM(+S)과 SVM-HMM(+S)의 2.5% 실험결과 차이는 통계적으로 의미가 있다. 반면 live chat 말뭉치에 대해서는 p값이 2.61로 유의수준 0.05보다 높아 귀무가설이 채택되어 이 두 모델의 성능은 통계적으로 유사하다고 말을 할 수 있다. 이는 live chat 데이터가 채팅으로 기록 된 것이기 때문에 데이터의 노이즈가 많고, 어휘 집합의 크기가 크면서 태깅된 데이터의 수가 작고 kappa 상관계수가 낮은 여러 요인으로 언어모델 학습이 잘 안되어 성능이 낮은 것으로 생각된다. 또한 SACTI-1 말뭉치에 대해서는 6.5%이상의 accuracy가 차이가 났고 p값이 0.0으로 유의수준 0.05보다 낮아 실험결과 차이가 통계적으로 의미가 있다. 이는 노이즈가 있는 데이터에 대하여 제안한 모델이 기존 모델들 보다 높은 성능으로 잘 동작함을 보여준다. 이는 언어모델을 사용하기 때문에 노이즈가 발생하여도 대화 의도 별로 고유한 단어들의 분포가 형성되어 노이즈에 강함으로 분석이 된다. 이에 대한 분석으로 다음 절에서 알아보겠다.

Accuracy를 제외한 다른 측정치에 대해서도 accuracy의 결과와 비슷한 모양을 이룬다. 다만 제안한 모델은 SVM-HMM(+S)와 비교했을 때, precision 측정치에 대해서 우위에 있지만 recall 측정치에 대해서는 그렇지 못하였다. 이는 언어모델을 사용하기 때문에 생기는 결과로 예상이 된다. 몇몇 단어가 특정 대화 의도에서 주로 나오는 경우에는 그 단어들

이 대화 의도를 잘 분별해주어 precision이 높은 반면에, 분별력 있는 단어들의 수가 적은 특정 대화 의도들이 다른 대화 의도로 예측이 되어 recall이 낮게 나온 것으로 해석이 된다. 결국 precision과 recall은 반비례 관계이기 때문에 어느 한 쪽이 좋다고 모델을 평가할 수 없지만, 제안한 모델은 높은 정확도로 발화의 대화 의도를 맞추어야 하는 상황에서 높은 성능을 낼 것으로 기대된다. 이러한 상황으로는 대화에서 높은 품질로 필요한 정보만 추출하고자 할 때를 말하며, 주로 대화 데이터 기반의 지식 베이스(knowledge base)를 구축하거나 게시글 요약 시스템 등을 구축할 때 높은 정확도가 요구 된다.

5.4.2 ASR 노이즈에 대한 강인성

ASR 노이즈가 존재하는 SACTI-1 말뭉치에 대하여 제안한 모델 MM(+S)의 성능이 제일 높았다. 그렇다면 모델을 학습할 때 유저의 발화를 텍스트로 타이핑한 것과 이를 노이즈 생성 모델을 통하여 노이즈가 생성된 텍스트를 사용하여 SVM-HMM(+S)와 비교해보면 제안한 모델의 강인성(robustness)을 확인할 수 있다. 이에 대한 정당성으로 시스템 역할을 하는 위자드는 노이즈가 있는 텍스트를 보고 머릿속으로 노이즈를 제거하거나 추론하는 과정을 거쳐서 실제로 타이핑되어 입력으로 들어온 텍스트의 의미를 추론하고, 이를 바탕으로 유저와 대화를 진행한다. 그렇기 때문에 이 두 가지 타입의 텍스트를 각각 학습하고 비교 분석하는 것이 유의미하다. 앞서 노이즈 정도를 나타내는 WER의 각 레벨에 따라 말뭉치를 none, low, med, hi 그룹으로 나누었다고 언급하였고 편의상 모델의 학습 데이터 타이핑 한 텍스트를 사용한 그룹은 A, B, C, D라 명명하고, 노이즈 생성 모델을 통하여 이 텍스트에 노이즈가 추가된 텍스트를 노이즈 그룹별로 A, B', C', D'라 명명한다. 이때 A는 none 그룹에

속하므로 동일하다. 또한 각 그룹은 서로 다른 대화이고 대화의도 분포 역시 다르다. 하지만 본 실험의 목적은 노이즈에 따른 성능 감소를 확인하여 제안한 모델이 SVM-HMM(+S)보다 강인성이 좋을지 확인하는 것이기 때문에 각 그룹별 말뭉치가 달라도 문제가 없다.

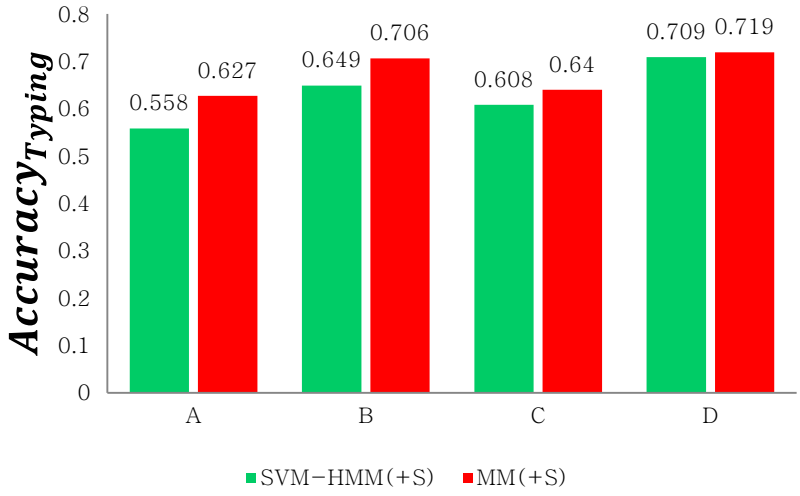

그림 11 SACTI-1 말뭉치에서 노이즈 별 분류 정확도 (타이핑한 텍스트로 학습)

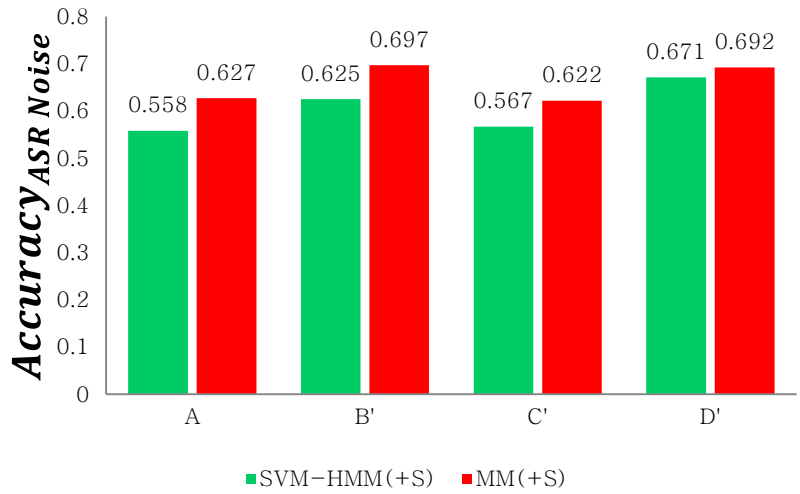

그림 12 SACTI-1 말뭉치에서 노이즈 별 분류 정확도 (노이즈 텍스트로 학습)

그림 11와 12은 두 가지 타입의 텍스트를 각각 학습하여 SVM-HMM(+S)와 MM(+S)의 분류 정확도를 비교한 것이다. 그림 11에서 타이핑한 텍스트로 각 모델을 학습한 경우, 노이즈 그룹별로 제안한 모델인 MM(+S)가 SVM-HMM(+S)보다 성능이 높음을 알 수 있다. 마찬가지로 그림 12에서 노이즈 텍스트로 각 모델을 학습한 경우에도 제안한 모델이 성능이 좋음을 알 수 있다. 즉 노이즈 정도에 상관없이 제안한 모델이 최신연구보다 성능이 좋으므로 노이즈에 적합한 모델임을 알 수 있다. 그리고 노이즈에 따른 분류 성능 감소는 그림 13에서 보여준다.

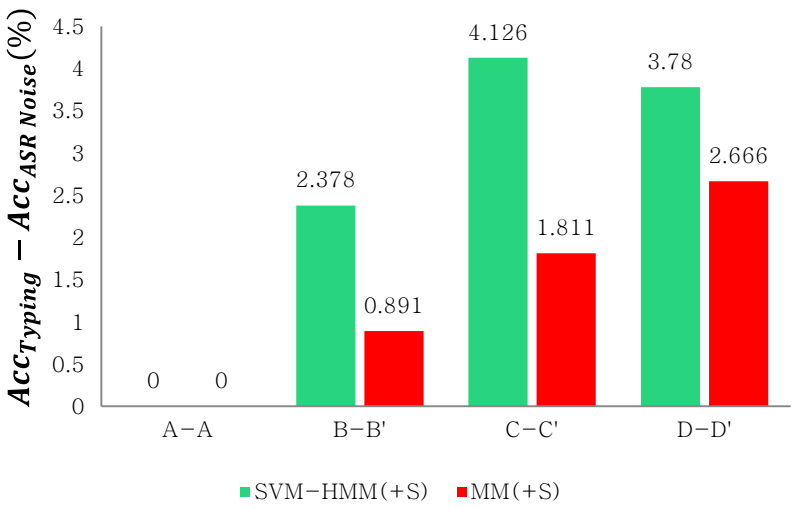

그림 13 노이즈 유무에 따른 분류 정확도 저하 정도

위 그래프는 타이핑한 텍스트로 학습했을 때 분류 성능에서 노이즈 텍스트로 학습했을 때 분류 성능을 각 노이즈 그룹별로 차이를 나타낸 것이다. None 그룹에 대해서는 동일한 말뭉치이기 때문에 성능차이가 없고, low 그룹에 대해서는 두 모델 모두 성능 하락을 보이지만 그 하락의 폭은 SVM-HMM(+S)가 더 크다. Med 그룹에서는 그 하락의 폭이 더 커

진 것 알 수 있고 두 모델의 성능 하락 차이도 커졌다. Hi 그룹에서는 두 모델의 성능 하락 차이는 줄어 들었지만 여전히 SVM-HMM(+S)의 성능 하락이 더 크다. 결과적으로 각 노이즈 그룹별로 제안한 모델이 절대값 면에서 SVM-HMM(+S)보다 성능 하락이 낮아 더 강인하다고 말할 수 있다.

5.4.3 확장성

마지막으로 모델의 확장성(scalability)을 비교해본다. Live chat 말뭉치와 SACTI-1 말뭉치는 발화의 수가 적어 확장성을 실험하기에는 부적합하므로 map task 말뭉치를 사용하여 확장성을 비교해본다. 그리고 이 크기 또한 작아 map task 말뭉치를 복제하여 10배 크기로 확장하였다. 왜냐하면 실험에 사용한 말뭉치의 태깅된 개수가 적더라도 실제 응용 측면에서는 더 큰 말뭉치가 사용될 수도 있고, 지도 학습, 비지도 학습, 준지도 학습 모델들이 예측한 대화 의도들을 다시 학습 데이터로 이용할 경우 그 크기는 몇 배가 될 수 있기 때문에 이러한 상황이 충분히 있을 수 있어 말뭉치를 복제를 해도 크게 무리가 없다. 결과적으로 복제를 통해 총 1000개의 대화가 있고, map task에서 하나의 대화는 약 211개의 발화로 구성이 되므로 총 211,000개의 발화에 대하여 확장성 실험을 진행하였다. 그림 15에서 X축은 대화 말뭉치에서 학습데이터로 사용되는 대화의 수를 나타내었다. Y축은 모델의 학습시간을 초 단위로 나타낸 것이다.

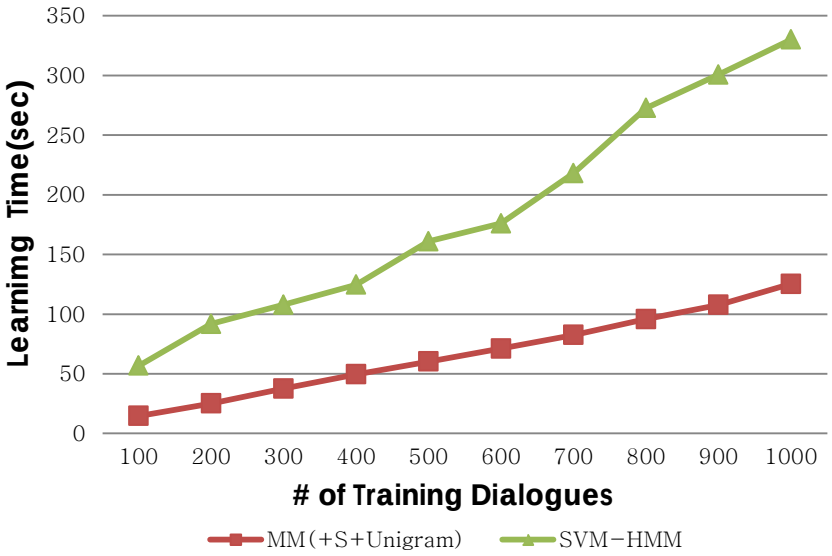

그림 14 학습데이터 증가에 따른 모델의 학습시간 비교

실험결과에서 알 수 있듯이 본 논문에서 제안한 모델이 SVM-HMM(+S)와 비교하여 모델의 학습시간이 상수 배 정도로 작음을 알 수 있다. SVM-HMM(+S)같은 경우는 데이터의 크기가 늘어날수록 SVM과 HMM에서 수행되는 모델의 학습 시간이 크게 늘어남을 알 수 있다. 반면에 제안한 모델은 확률 테이블을 구축하기 위하여 관측 값인 학습 데이터에서 단어의 빈도수만 셈하면 되기 때문에 학습 시간이 크게 오래 걸리지 않는다. 즉 확장성 면에서 제안한 모델이 SVM-HMM(+S)보다 더 좋음을 알 수 있다.

제 6 장 결론 및 향후 연구

6.1 결론

본 논문에서는 대화를 구성하는 발화에 대화 의도를 분류하는 문제를 다루었다. 이를 해결하기 위해서 기존 연구들은 발화의 의존관계 및 대화에서 추출할 수 있는 여러 자질들을 어떻게 모델링을 할지를 고민하였으며 단순히 학습데이터에 의존관계 정보와 자질들을 넣어주거나 순차 데이터 처리 모델들인 HMM, CRF, SVM-HMM 등을 이용하기에 머물러 있었다. 본 저자는 대화 의도는 말하는 사람마다 그 의도를 표현하는 방법이 다를 것이고, 이전 대화 의도에도 크게 영향을 받는 다는 생각을 시작으로 대화 의도를 마르코프 모델의 상태로 표현을 하고 대화 의도에 영향을 주는 화자, 단어 정보들을 마르코프 모델의 입출력 형태로 표현을 하였다. 특히 대화 의도와 화자에 따라 생성되는 단어들이 달라지는 것을 표현하기 위하여 언어모델을 사용하였다. 실험에서 제안한 모델이 최신연구인 SVM-HMM와 비교하여 경쟁력 있는 분류 성능을 얻었으며, ASR 노이즈에 대해서는 SVM-HMM보다 분류 성능이 좋았고, 성능 저하면에서 강인성이 있음을 확인하였다. 또한 모델의 간결함으로 학습시간 부분에서 SVM-HMM보다 확장성이 높음을 보였다. 결과적으로 제안하는 가정이 대화를 표현하는데 있어 적합하고, 실제 대화 시스템의 ASR 모듈이 만드는 노이즈에 대해서 기존 모델들에 비하여 대응 가능한 수준임을 확인하였다.

6.2 향후 연구

향후 연구로는 다음과 같은 점을 개선하려고 한다. 첫 번째로 제안한 모델에서는 유니그램 기반의 언어모델을 사용한다. 실험부분에서 표기하지는 않았지만 바이그램(bigram), 트라이그램(trigram)으로도 실험을 해보았다. 물론 결과가 그램 수가 늘어날수록 나빠졌다. 즉 이 부분에서 아직 개선의 여지가 남아 있다. 본 연구에서는 모든 단어들에 대하여 동일한 가중치를 가지고 단어의 생성확률을 사용하였다. 하지만 몇몇 발화에서는 대화 의도를 결정짓는 핵심 단어(cue word) 혹은 핵심 절(cue phrase)들이 존재한다. 예를 들어 'can you~'와 'do you have~' 등이 있으면 YN-Question 대화 의도를 가질 확률이 높은 것이 그 이유이다. 하지만 제안한 모델에는 이 부분이 명확히 모델링이 되어있지 않다. 물론 언어모델을 화자와 대화 의도에 따라 나뉘어진 학습 데이터해서 사용하기 때문에 이러한 문제를 간접적으로 해결을 하고 있지만 여전히 발화의 모든 단어들의 생성확률을 계산하여 고려해야 하는 문제가 있다. 두 번째로 이 모델은 과제 지향형 대화에 대해서 동작하는 모델이다. Switch Board 말뭉치와 같이 개방형 주제를 다루는 대화에서는 화자간에 역할이 존재하지 않기 때문에 적절한 언어모델을 생성하기가 어렵다. 그렇기 때문에 개방형 대화에서도 제안한 모델과 유사한 접근 방법으로 새로운 모델을 연구해 보고자 한다.

다른 측면에서 준지도 학습 방법으로 제안한 모델을 확장해 보려고 한다. 지도학습의 성능이 비지도 학습보다 성능이 좋을지라도 한계점이 있는데, 매번 구축하려는 대화시스템의 도메인에 따라 학습에 사용되는 대화 말뭉치에 시스템 목적에 맞는 대화 의도 및 여러 가지의 자질들을 태깅 해주어야 한다는 것이다. 하지만 이는 오랜 시간이 걸리기 때문에 많

은 양의 태깅 된 데이터를 얻기가 어렵다는 문제가 있다. 또한 전문적인 지식을 갖고 있지 않는 사람이 태깅을 할 경우 대화 의도의 일치율이 떨어져 신뢰할 수 없는 데이터를 만드는 문제가 있다. 즉 지도학습의 성능은 태깅 데이터의 크기와 품질에 영향을 받기 때문에 태깅 된 대화 말뭉치의 부족은 성능에 영향을 주는 매우 중요한 문제이다.

이러한 이유로 지도학습과 더불어 태깅 데이터를 사용하지 않고 비슷한 대화 의도를 갖는 발화들을 클러스터링하는 비지도 학습의 연구들이 꾸준히 나오고 있지만, 클러스터링 성능이 지도학습에 비하여 좋지 못하고 클러스터링 결과로 나오는 클러스터들을 다시 대화 의도에 매칭해야 하는 문제가 있다. 물론 클러스터 식별 아이디를 대화 의도로써 사용하여 대화시스템에서 이용할 수 있지만, 대화시스템의 시스템 행위를 정의하고 설계하는데 있어 역으로 클러스터 아이디에 해당하는 대화 의도를 지니는 발화들을 분석하는 작업이 필요로 하여 어려움이 있다.

이러한 문제들을 해결하고자 하는 시도로 여러 연구들에서 준지도 학습 방법이 제안되고 있다[6, 21, 28]. 공통적으로 제안된 방법들이 태깅되어 있는 데이터를 이용하여 태깅 되어 있지 않은 데이터로 대화 의도의 태그를 확장시키는 쪽으로 진행이 된다. 이러한 부분에서 지도학습을 통하여 만들어진 분류기를 사용하거나 검색분야에서 쓰이는 유사한 문서를 검색하는 기술을 이용할 수 있다[29]. 어떠한 방법을 사용하든지 높은 품질의 태깅 데이터를 얻을 수 있다면, 여러 도메인에서 대화시스템을 구축하는데 있어 지도학습 방법을 통하여 만들어진 모델이 대화 의도 분류문제를 푸는데 있어 높은 성능을 낼 것으로 기대된다.

참고문헌

[1] J. L. AUSTIN, *How to do things with words*, Clarendon Press, Oxford, 1962.

[2] S. BANGALORE, G. DI FABBRIZIO and A. STENT, *Learning the structure of task-driven human-human dialogs*, Audio, Speech, and Language Processing, IEEE Transactions on, 16 (2008), pp. 1249-1259.

[3] S. BHATIA, P. BIYANI and P. MITRA, *Summarizing Online Forum Discussions-Can Dialog Acts of Individual Messages Help?*, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014.

[4] C.-P. CHEN, C.-H. WU and W.-B. LIANG, *Robust dialogue act detection based on partial sentence tree, derivation rule, and spectral clustering algorithm*, EURASIP Journal on Audio, Speech, and Music Processing, 2012 (2012), pp. 1-9.

[5] H. CUAY HUITL, N. DETHLEFS, H. HASTIE and O. LEMON, *Impact of ASR N-Best Information on Bayesian Dialogue Act Recognition*, SIGDIAL, 2013.

[6] M. JEONG, C.-Y. LIN and G. G. LEE, *Semi-supervised speech act recognition in emails and forums*, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, Association for Computational Linguistics, 2009, pp. 1250-1259.

[7] S. JOTY, G. CARENINI and C.-Y. LIN, *Unsupervised modeling of dialog acts in asynchronous conversations*, *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, 2011, pp. 1807.

[8] S. N. KIM, L. CAVEDON and T. BALDWIN, *Classifying dialogue acts in one-on-one live chats*, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2010, pp. 862-871.

[9] T. LAGER and N. ZINOVJEVA, *Training a dialogue act tagger with the μ-TBL system*, *The Third Swedish Symposium on Multimodal Communication*, 1999.

[10] C. LEE, S. JUNG, K. KIM, D. LEE and G. G. LEE, *Recent Approaches to Dialog Management for Spoken Dialog Systems*, JCSE, 4 (2010), pp. 1-22.

[11] D. LEE, M. JEONG, K. KIM and G. G. LEE, *Unsupervised modeling of user actions in a dialog corpus*, *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, IEEE, 2012, pp. 5061-5064.

[12] W.-B. LIANG, C.-H. WU and C.-P. CHEN, *Semantic information and derivation rules for robust dialogue act detection in a spoken dialogue system*, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, Association for Computational Linguistics, 2011, pp. 603-608.

[13] T. OYA and G. CARENINI, *Extractive Summarization and Dialogue Act Modeling on Email Threads: An Integrated Probabilistic Approach*, *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2014, pp. 133.

[14] J. M. PONTE and W. B. CROFT, *A language modeling approach to information retrieval*, *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, 1998, pp. 275-281.

[15] A. RITTER, C. CHERRY and B. DOLAN, *Unsupervised modeling of twitter conversations*, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, 2010, pp. 172-180.

[16] K. SAMUEL, S. CARBERRY and K. VIJAY-SHANKER, *Dialogue act tagging with transformation-based learning*, *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, Association for Computational Linguistics, 1998, pp. 1150-1156.

[17] J. R. SEARLE, *Speech acts: An essay in the philosophy of language*, Cambridge university press, 1969.

[18] R. SERAFIN and B. DI EUGENIO, *FLSA: Extending latent semantic analysis with features for dialogue act classification*, *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 2004, pp. 692.

[19] A. STOLCKE, K. RIES, N. COCCARO, E. SHRIBERG, R. BATES, D. JURAFSKY, P. TAYLOR, R. MARTIN, C. VAN ESS-DYKEMA and M. METEER, *Dialogue act modeling for automatic tagging and recognition of conversational speech*, Computational linguistics, 26 (2000), pp. 339-373.

[20] M. N. STUTTLE, J. D. WILLIAMS and S. YOUNG, *A framework for dialogue data collection with a simulated ASR channel*, *INTERSPEECH*, 2004.

[21] A. SUBRAMANYA and J. BILMES, *Semi-supervised learning with measure propagation*, The Journal of Machine Learning Research, 12 (2011), pp. 3311-3370.

[22] D. SURENDRAN and G.-A. LEVOW, *Dialog act tagging with support vector machines and hidden Markov models*, *INTERSPEECH*, 2006.

[23] M. TAVAFI, Y. MEHDAD, S. JOTY, G. CARENINI and R. NG,

Dialogue act recognition in synchronous and asynchronous conversations, *Proceedings of the SIGDIAL 2013 Conference*, Citeseer, 2013, pp. 117-121.

[24] D. R. TRAUM, *A Computational Theory of Grounding in Natural Language Conversation*, *Computer Science Dept.*, Rochester Univ., 1994.

[25] B. C. WALLACE, T. A. TRIKALINOS, M. B. LAWS, I. B. WILSON and E. CHARNIAK, *A Generative Joint, Additive, Sequential Model of Topics and Speech Acts in Patient-Doctor Communication*, *EMNLP*, 2013, pp. 1765-1775.

[26] J. D. WILLIAMS and S. YOUNG, *Characterizing task-oriented dialog using a simulated ASR chanel*, *INTERSPEECH*, Citeseer, 2004.

[27] H. WRIGHT, *Automatic utterance type detection using suprasegmental features*, (1998).

[28] X. YANG, J. LIU, Z. CHEN and W. WU, *Semi-supervised learning of dialogue acts using sentence similarity based on word embeddings*, *Audio, Language and Image Processing (ICALIP), 2014 International Conference on*, IEEE, 2014, pp. 882-886.

[29] Q. ZHANG, J. KANG, J. QIAN and X. HUANG, *Continuous word embeddings for detecting local text reuses at the semantic level*, *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, ACM, 2014, pp. 797-806.

Abstract

Noise Robust Dialogue Act Recognition for Task-oriented Dialogues

Taeyeon Kim
School of Computer Science and Engineering
The Graduate School
Seoul National University

In spoken dialog system, e-mail summary system and thread summary system development, dialogue act classifier plays an important role because the systems depend on the performance of classifying dialogue acts of utterances, e-mails and posts to improve completeness of the system. The dialogue act classification problem is a well-known problem to assign the dialogue acts to utterances in a conversation.

One of the main challenges in the development of robust dialog systems is especially to deal with noisy input due to imperfect results from Automatic Speech Recognition (ASR) module. The challenge in dialogue act recognition is the mapping from noisy user utterances to dialogue acts. In this paper, to cope with noisy utterances, we describe a noise robust generative model of task-oriented conversation that captures both the speaker information and the dialogue act associated with each utterance under the assumption that a speaker says about something by using appropriate vocabulary with the aim of getting someone to do somethings. The proposed model is based on Markov model and is modified to reflect the assumption.

In the experiments, we evaluate the classification results by comparing them to the simple Markov model and state-of-the-art SVM-HMM results. The proposed model is a better conversation model than the simple Markov model and shows the

competitive classification results in comparison with SVM-HMM in the task-oriented HCRC map task corpus, live-chat corpus and SACTI-1 corpus. Results based on SACTI-1 corpus which simulates ASR errors particularly show that the proposed model is robust against noisy user utterances.

Keywords : Dialogue Act, Dialogue Act Classifier, Markov Model, SVM-HMM, Dialog System

Student Number : 2013-23110