Journal of The Korea Society of Computer and Information
Vol. 23 No. 1, pp. 17-24, January 2018
www.ksci.re.kr
https://doi.org/10.9708/jksci.2018.23.01.017

Out-Of-Domain Detection Using Hierarchical Dirichlet Process

Young-Seob Jeong*

Abstract

With improvement of speech recognition and natural language processing, dialog systems are recently adapted to various service domains. It became possible to get desirable services by conversation through the dialog system, but it is still necessary to improve separate modules, such as domain detection, intention detection, named entity recognition, and out-of-domain detection, in order to achieve stable service offer. When it misclassifies an in-domain sentence of conversation as out-of-domain, it will result in poor customer satisfaction and finally lost business. As there have been relatively small number of studies related to the out-of-domain detection, in this paper, we introduce a new method using a hierarchical Dirichlet process and demonstrate the effectiveness of it by experimental results on Korean dataset.

▸Keyword: Out-of-domain detection, Dialog system, Topic modeling, Hierarchical Dirichlet process

I. Introduction

음성인식 기술과 자연어 처리 기술의 발달은 대화시스템이 우리 삶의 다양한 곳에 적용될 수 있는 기회를 제공하고 있다. 특히 아마존 Echo [1], SK 텔레콤 Nugu [2], Naver WAVE [3] 등과 같은 인공지능형 대화시스템에 기반한 가정용 스마트 스피커는 일상대화를 통해 각종 서비스를 제공받을 수 있게 해주었으며, 많은 기업 및 기관에서 대화시스템을 적용하기 위해 시도하고 있다 [4, 5].

사람이 컴퓨터 시스템과 소통하여 원하는 정보를 찾거나 서비스를 제공받는 방법은 두 단계의 변화를 거쳐왔다고 볼 수 있다. 웹 기반 검색 서비스의 등장은 사람들이 원하는 정보 및 서비스를 찾는 과정을 단순화하고 시간을 크게 절약하도록 도왔으며, 심지어 사람이 관심을 가질만한 정보를 선별하여 추천해주는 서비스까지 지원하게 되었다 [6, 7]. 컴퓨터 앞에 앉아서 검색 서비스를 사용하던 시대는 스마트폰의 대량 보급과 함께 새로운 국면을 맞이하였는데, 스마트폰 앱을 설치하여 이를 통해 서비스를 제공받게 된 것이다. 사람들은 휴대 가능한 스마트폰을 통해 언제 어디서든 정보를 검색하고 앱을 설치하여 서비스를 제공받을 수 있게 되었으며, 이는 스마트폰 앱 시장이라는 새로운 시장을 여는 계기가 되었다. 이렇게 두 단계의 변화를 거쳐왔던 컴퓨터 시스템과의 소통 방법은 대화시스템의 등장과 함께 세 번째 변화를 겪고 있다. 대화시스템은 검색어 입력을 통해 정보 또는 서비스를 찾던 방식을 탈피하여 '자연어 문장'으로 소통할 수 있도록 하였으며, 음성 인식 기술을 접목하면 사람과 대화하듯이 음성을 통해 컴퓨터 시스템과 소통할 수 있게 되었다. 대화시스템의 등장은 컴퓨터 시스템과의 새로운 소통 방법을 제공함으로써 다양한 기관, 서비스 등에 접목이 가능하므로, 이와 관련된 시장이 점차 커지고 있다.

대화시스템의 등장이 삶의 질 향상에 큰 변화를 가져올 것은 의심할 여지가 없지만, 아직 초창기인만큼 안정적인 서비스 제공을 위해서 가야 할 길이 많이 남아있다. 예를 들어, 현재 시중에 판매되는 인공지능 스피커들은 시끄러운 상황에서의 음성 인식 성공률이 매우 낮아지게 되는데, 사람이 시끄러운 상황에서 다른 사람의 음성을 정확히 인지하는 것처럼 되기까지 음성 인식 기술 성능 개선이 필요하다. 자연어 처리에서는 새롭게 등

• First Author: Young-Seob Jeong, Corresponding Author: Young-Seob Jeong
*Young-Seob Jeong (bytecell@sch.ac.kr), Dept. of Big Data Engineering, Soonchunhyang University
• Received: 2017. 11. 13, Revised: 2017. 12. 15, Accepted: 2018. 01. 10.
• This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP: Ministry of Science, ICT & Future Planning) (No. 2017017836). This work was supported by the Soonchunhyang University Research Fund.

장한 지명 혹은 인물에 대한 개체명 인식률 개선을 위한 연구가 지속되어야 할 것이다. 특히, 대화시스템에서 제공하지 않는 서비스 도메인, 즉 미지원 도메인인지 여부를 검출하는 기능의 성능 개선이 매우 필요하다.

본 연구는 대화시스템의 자연어 이해 모듈의 도메인 추출 단계에서 수행되는 미지원 도메인 검출에 대한 새로운 방법을 제시한다. 문서에 내재된 토픽을 자동으로 검출하는 토픽 모델링 기술의 일종인 계층적 디리클레 프로세스 (Hierarchical Dirichlet Process)는 문서 특성을 고려하여 토픽 개수를 자동으로 결정한다. 미지원 도메인 데이터가 없거나 부족한 상황에서, 지원 도메인 데이터들에 내재된 임의의 패턴 혹은 토픽을 검출하는 HDP 모델에서 생성되는 결과물을 대화시스템의 미지원 도메인 검출에 자질로써 효과적으로 적용되는지를 실험을 통해 입증한다.

본 논문의 2장에서는 본 연구와 관련된 기반 지식과 관련 연구들을 소개하고, 3장에서는 새롭게 제시하는 방법에 대하여 설명하며, 4장에서는 실험 결과를 통해 제시된 방법의 효과를 입증하고 5장에서 결론을 맺는다.

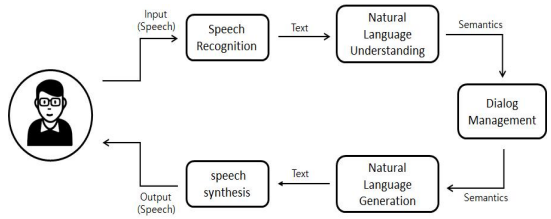

Fig. 1. Flow of dialog system

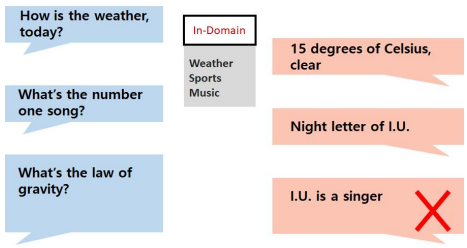

Fig. 2. Failure example of OOD detection

II. Background

1. Research Necessity

대화시스템은 보통 내부적으로 Fig. 1과 같이 여러 단계를 거쳐 동작을 수행한다. 음성 인식 단계는 음성을 입력받아 N개의 가장 유망한 텍스트를 출력으로 제공한다. 텍스트가 아닌 다른 형태, 이를테면 Lattice 형태의 결과물을 제공할 수도 있지만 [8], 자연어 이해 단계에서 처리하기에 텍스트가 용이하므로 흔히 텍스트를 음성 인식 단계의 결과물로 제공하게 된다. 자연어 이해 단계에서는 입력 문장(텍스트)에 대하여 내부적으로 여러 소단계를 거치게 되는데, 그 중 하나가 '도메인 추출'이다. 대화시스템이 제공하는 서비스 목록 S = {S1, ..., Sk}에 속한 각 Si를 i번째 도메인이라고 지칭할 때, 대화시스템에 입력되는 임의의 문장 텍스트가 k개의 도메인들 중 어디에도 속하지 않는 '미지원 도메인 문장'인지를 체크하는 과정이 필요하다.

미지원 도메인 문장인지 체크하는 기능이 없거나 이 기능을 수행하는 모델의 성능이 매우 낮은 경우, 이는 사용자의 서비스 만족도를 급감시키는 원인이 될 수 있다. Fig. 2에서 미지원 도메인 검출 실패로 인한 결과 예시를 보여주고 있다. 지원 도메인이 {날씨, 스포츠, 음악}인 경우, '만유 인력의 법칙이 뭐야?'라는 문장이 미지원 도메인임을 검출하지 못하고, '음악' 도메인으로 잘못 인식한 경우에는 Fig. 2과 같이 엉뚱한 결과를 제공하게 되며, 사용자 만족도를 급감시킬 수 있다. 미지원 도메인임을 성공적으로 검출한다면, 이러한 문장에 대하여 '지원하지 않는 기능입니다'라는 결과를 제공함으로써 사용자 만족도 감소의 폭을 줄일 수 있게 된다.

2. Related Work

미지원 도메인 검출을 위해 수행된 기존 연구는 크게 3가지 관점으로 볼 수 있다. 첫 번째 관점은 '데이터 확보'에 대한 관점, 두 번째는 분류 모델에 대한 관점, 세 번째는 자질 정의에 대한 관점이다.

2.1 Data

대화시스템을 개발할 때, 제공하고자 하는 서비스 목록 S를 우선적으로 결정하게 된다. 서비스 목록 S에 속한 각 서비스 또는 각 도메인에 해당하는 데이터를 수집 및 태깅하여, 대화시스템을 구성하는 각 단계를 데이터 기반 기계학습 모델들을 사용하게 학습하고 평가할 수 있도록 작업하게 된다. 즉, S에 k개의 도메인이 속해있는 경우, 각 도메인 S_i에 해당하는 데이터를 m개 수집한 데이터셋을 D_i라고 지칭한다면, 각 데이터 $d_j \in D_i = \{d_1, ..., d_m\}$는 해당 데이터가 속한 도메인이 S_i인 것이 태깅되어 있어야 기계학습 모델 학습에 사용될 수 있을 것이다. 대화시스템 개발을 위해 k개의 서비스 도메인에 해당하는 데이터셋들을 수집 및 태깅하는 것은 당연히 해야할 일이지만, 미지원 도메인에 대한 데이터는 대부분 수집하지 않는다. 무한한 개수만큼 존재하는 미지원 도메인에 해당하는 데이터를 수집하는 것은 매우 어려울뿐더러, 어떤 방법과 정책 하에 수집해야 하는지 결정하는 것도 어렵기 때문이다. 이는 미지원 데이터 연구를 위한 데이터셋 수집을 어렵게 만들었고, 미지원 도메인 관련연구가 다른 분야에 비해 미비하게 수행되게 만든 가장 큰 이유가 되었다. 이 문제점을 어느정도 피할 수 있는 방법으로 Deleted Interpolation 기법 [9]이 소개되었는데, 이는 각 서비스 도메인 S_i를 한 번씩 미지원 도메인으로 가정하고 실험을 수행하는 방법이다.

2.2 Classifier

미지원 도메인 판별 모델로서 여러 가지가 적용되었는데, Linear Discriminant Model (LDM) [16], Latent Semantic

Analysis (LSA) [17], Support Vector Machines (SVM) [18], Logistic regression [19], Maximum Entropy Model (MEM) [21], IB1 등이 주로 사용되었으며, SVM과 LDM을 활용한 연구들이 전반적으로 좋은 성능을 보여왔다 [9, 12, 13, 20, 29]. 최근에는 인공신경망을 활용한 연구가 등장하였으며 [22], 워드 임베딩 벡터와 지원 도메인 데이터에 Long Short-Term Memory (LSTM) [31] 를 적용하여 문장 임베딩 벡터를 생성한 후, 다양한 분류기에 자질로써 활용하여 비교 실험하였고, Auto-Encoder (AE)를 활용하였을 때 Equal Error Rate (EER) 값이 7.02%로 가장 좋은 성능을 보였다. EER은 False Acceptance Rate (FAR)과 False Rejection Rate (FRR)이 같을 때의 error rate를 의미하는 값으로서, 모델 간의 성능 비교를 위해 활용된다. SVM과 LDM을 주로 활용하였던 과거 연구들의 EER이 전반적으로 15%부터 19%였던 것을 감안하면, 대단히 높은 성능 향상을 이룬 것이라고 볼 수 있다.

2.3 Feature

자질의 관점에서 보면 기존 연구들은 크게 3가지로 분류할 수 있다. 첫 째는 사용자 발화 텍스트 자체로부터 추출하는 자질이다. 사용자 발화 텍스트에 대한 N-gram 기반의 lexical 자질이나 syntactic 자질 (예: POS 태그, syntactic parse), semantic 자질 (예: semantic role label) 등을 그 예로 들 수 있다. 둘 째, 외부 자원 기반의 자질이다. Wordnet [14] 기반으로 자질을 정의하거나 자체 제작한 사전 기반의 자질을 사용할 수 있다 [10]. 이 자질을 사용하면 외부 자원의 양적, 질적 수준에 크게 의존한다는 약점을 가지게 된다. 셋 째, 대화시스템 내부의 로그 데이터 기반 자질이다. 이 자질은 대화시스템에서 미지원 도메인 여부를 판별하는 시점과 매우 밀접한데, 크게 두 가지 시점이 대표적으로 고려된다. 첫 번째는 자연어 이해 단계의 시작 시점에 사용자 발화의 도메인을 추출함과 동시에 미지원 도메인 여부를 판별하는 것이고, 두 번째는 자연어 이해가 모두 수행된 후, 자연어 이해 최종 결과물을 종합하여 미지원 도메인 여부를 판별하는 것이다. 물론, 위 두 시점에 미지원 도메인 검출 수행을 병행하는 것도 가능한데, 이 두 가지 시점의 장단점을 고려하여 수용하는 것이 좋을 것이다.

자연어 이해 단계의 시작 시점에 미지원 도메인 여부를 판별하는 경우, 자연어 이해 단계가 마치는 시점에 판별할 때보다 상대적으로 적은 정보를 가지고 미지원 도메인 여부를 판별해야 하는 단점이 있는 반면, 미지원 도메인 여부를 일찍 검출하는 것이 가능한만큼 미지원 도메인에 대한 대화시스템 응답 속도 향상에 기여할 수 있다는 장점도 가지고 있다. 자연어 이해 단계를 모두 마친 후 미지원 도메인 판별을 하게 된다면, 자연어 이해 결과물, 이를테면 개체명 인식 결과, Semantic Role Labels, 사용자 의도 파악 결과, 슬롯 추출 결과 등을 미지원 도메인 판별을 위한 자질 정의에 사용할 수 있게 되므로, 높은 판별 성능 달성을 위해 비교적 유리하다고 볼 수 있을 것이다.

기존 연구에서 주로 이용되었던 자질들은 사용자 발화 자체로부터 추출 가능한 자질로서 N-gram 기반의 자질들이 사용되었으며 [9, 20, 29], 이전 발화와 현재 발화에 대한 대화시스템의 자연어 이해 결과물을 모두 활용한 자질들이 사용되기도 하였고 [11, 12], 외부 언어 리소스를 활용한 자질들이 적용되기도 하였다 [13]. 특히, 사용자 발화 자체에서 추출 가능한 자질에 대한 다양한 시도가 부족하여, N-gram 기반 자질 외에 다른 시도가 연구된 바가 거의 없었으며, 최근에 이르러서야 문장의 단어 위치에 기반한 자질 추출 연구가 수행되었다 [15].

본 연구에서는 토픽 모델 기술로써 많은 분야에 널리 활용되었던 Hierarchical Dirichlet Process (HDP) [23]로 생성되는 토픽 분포를 미지원 도메인 검출에 자질로써 활용하는 기법을 제시한다. 토픽 모델 기술을 활용하여 자동으로 추출되는 '토픽'은 주어진 문서에 내재된 주제를 의미하므로, 토픽 모델의 '토픽'을 대화시스템에서 제공하는 '도메인'으로 간주할 수 있다. 임의의 문장에 대하여 HDP를 통해 생성된 토픽들에 대한 분포가 미지원 도메인 검출을 위한 자질로써 유용하다는 것을 실험을 통해 증명한다.

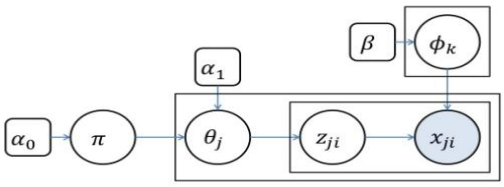

Fig. 3. Graphical representation of hierarchical Dirichlet process

3. Topic Modeling

본 연구에서 자질 생성을 위해 활용하는 토픽 모델링 기술은 문서집합 D에 내재된 T개의 토픽을 추출한다. 가령, 문서집합 D가 영화에 배경음악으로 사용되었던 곡들에 대한 감상평들이고 이로부터 2개의 토픽을 검출하도록 할 경우, {음악, 영화}라는 2개의 토픽이 검출될 수 있을 것이다. 토픽 모델의 구조에 따라, 문서 저자의 관점에서 토픽들이 추출되거나 [25], 문서 관점에서 추출되거나 [24], 각 개체들의 관점에서 추출되는 등 [26], 서로 다른 종류의 토픽들이 추출될 수 있다. 토픽 개수 T는 흔히 사람에 의해 주어지며, 토픽 모델은 그 구조에 따라 임의의 T개의 토픽을 추출하게 된다. 특히, 토픽의 개수도 모델에 의해 자동으로 결정되는 기법들이 소개되었는데 HDP 모델이 가장 대표적인 모델이다.

Fig. 3에서 HDP 모델의 구조를 보여주고 있다. 노드 Θ_j는 j번째 문서에 대한 토픽 분포, Φ_k는 k번째 토픽, π는 전체 토픽 분포라고 볼 수 있다. HDP 모델을 흔히 Chinese Restaurant Franchise (CRF)에 빗대어 표현하기도 하는데, 본 논문에서도 설명의 용이함을 위해 CRF를 빗대어 설명해보면 다음과 같다. 각 문서는 restaurant, 각 단어는 손님으로 빗댈 수 있으며, restaurant 안의 각 테이블은 토픽으로 볼 수 있다. 문서 안에 N개의 단어가 있다면 restaurant 안에 N명의 손님이 들어오는 것을 빗댈 수 있고, 각 손님은 특정 기준에 의거하여 테이블을

선택하여 앉게 된다. 손님이 이미 앉아있는 테이블은 다른 손님이 합석할 확률이 높아지고, 아무도 앉지 않은 테이블에 앉을 확률은 α_1에 비례하게 된다. 여기까지의 비유는 Chinese Restaurant Process (CRP)이며, 여러 restaurant에 대한 비유로 확장하면 CRF가 된다.

모든 restaurant를 관리하는 본사가 있다고 할 경우, 이 본사에서는 모든 restaurant에서 서비스하는 토픽들의 분포를 관리하는데, 마치 각 restaurant의 테이블이 손님인 것처럼 취급함으로써 전체 토픽 분포를 관리한다. 가령, 모든 지점 restaurant에 존재하는 테이블의 총합이 M개라면, 본사에 M명의 손님이 들어와 각 토픽 테이블에 앉는 것으로써 모든 토픽에 대한 분포를 관리하게 된다. 이 때, 다른 손님이 이미 앉아있는 테이블이 아닌 새로운 테이블에 앉는 행위는 “새로운 토픽”을 추가하는 효과를 낳게 되며, 이를 통해 토픽 개수가 자동으로 증가할 수 있게 된다. 그래서 일반적으로 HDP 모델에서는 토픽 개수를 1로 초기화하여 학습을 시작한다.

토픽 모델 학습은 Markov Chain Monte Carlo의 일종인 collapsed Gibbs sampling [27], variational inference [24] 등의 알고리즘을 사용하여 반복적으로 파라메터 값들을 추산한다. 특히 collapsed Gibbs sampling 알고리즘은 구현하기 쉬우면서도 학습된 모델의 성능이 뛰어난 것으로 알려져있어 비교적 널리 사용되고 있다. HDP 모델도 collapsed Gibbs sampling 알고리즘으로 학습이 가능하며, 각 학습 스텝마다 모든 단어들에 대한 새로운 토픽이 샘플링된다. 각 토픽은 restaurant의 테이블로 빗댈 수 있으므로, 단어들에 원래 할당되어 있던 토픽을 지우고 새로운 토픽을 할당하는 작업은, 마치 각 손님이 자리에서 일어나 새로운 테이블로 이동하는 것을 묘사한다고 볼 수 있다. 이 때, 원래 존재하던 테이블에 있던 손님들이 모두 일어나 다른 테이블로 이동하게 되면, 해당 테이블은 사라지게 될 것이다. 이러한 현상이 모든 restaurant에서 전반적으로 발생하여 특정 토픽에 해당하는 테이블에 앉은 손님이 모두 사라지게 되는 경우, 그 토픽은 다음 학습 스텝에서는 사라지는 것으로 간주된다. 따라서, 토픽 개수의 초기값을 1로 설정하여 반복적인 학습 알고리즘을 통해 토픽 개수가 증가했다가 감소하는 현상이 반복되며, 충분한 반복 횟수가 경과하는 동안 안정적인 토픽 개수를 자동으로 찾아가게 되는 것이다.

본 연구에서는 HDP 모델이 문서에 내재된 임의의 토픽들의 개수를 자동으로 결정하는 특징을 이용하여, HDP 모델을 통해 생성된 토픽 분포에 대한 값을 미지원 도메인 검출을 위한 자질로 활용한다.

III. Out-Of-Domain Detection Using Hierarchical Dirichlet Process

대화시스템에서 제공하는 k개의 서비스 도메인들을 $S = \{S_1, ..., S_k\}$ 라고 할 때, 토픽 모델에서 추출되는 k개의 토픽 $T = \{T_1, ..., T_k\}$이 이상적으로 추출된다면 S와 T 사이에 일대일 대응이 만족하게 될 것이다. 즉, k개의 서비스 도메인에 대한 데이터셋 $D = \{D_1, ..., D_k\}$에 대하여 토픽 모델이 k개의 토픽을 추출하도록 할 경우, $1 \le i \le k$인 i에 대하여 $S_i = T_i$를 만족하고, D_i는 T_i에 대한 내용을 담고 있을 것이다. 하지만, 도메인들 간에 사람이 인지하기 어려운 임의의 패턴들이 존재할 수 있으며, 이 패턴들 또한 별도의 도메인 혹은 토픽으로 취급할 수 있다. 가령, 영화 배경음악에 대한 문서들로부터 {음악, 영화}라는 2개의 토픽이 추출되는 것을 일반적으로 기대할 수 있다면, {발라드, 락, 힙합, 액션, 로맨스, 광고, ...}와 같이 좀 더 세분화된 토픽이나 새로운 관점의 토픽이 추출될 수도 있다. 즉, 관점 혹은 토픽 개수에 따라 추출되는 토픽들이 달라질 수 있으며, 이는 결국 미지원 도메인 검출을 위한 자질로서의 품질도 달라질 수 있음을 의미한다.

본 연구는 k개의 서비스 도메인에 대한 문서로부터 $m \neq k$ $(m>0)$인 m개의 임의의 토픽들을 토픽 모델을 통해 추출하여, 이를 자질로서 활용하는 기법에 대한 연구이다. 이 토픽들은 서비스 도메인과 정확히 일치하지 않을 가능성이 높으며, 사람이 인지하기 어려운 임의의 패턴을 기반으로 추출된다. 특히, 본 연구에서 토픽 추출을 위해 사용하는 HDP 모델은 토픽에 대한 분포를 디리클레 프로세스를 통해 모델링함으로써, 토픽 개수가 증가할 수 있도록 허용하는 모델이다. HDP 모델은 학습 알고리즘을 통해 파라메터들을 학습하는 과정에서 기존에 존재하던 토픽이 없어지는 현상이 발생하기도 하며, 결과적으로 HDP 모델은 학습 데이터에 내재된 패턴에 기반하여 토픽 개수를 자동으로 증감해가면서 최적의 토픽 개수를 찾아가게 된다.

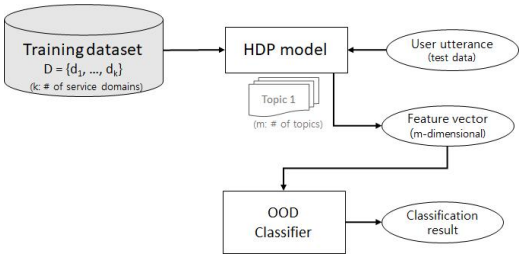

Fig. 4. Process of training HDP model and OOD detection

HDP 모델이 추출한 m개의 미지의 $T = \{T_1, ..., T_m\}$에 대한 토픽 분포를 미지원 도메인 검출을 위한 자질로 활용하는 과정을 Fig. 4에서 보여주고 있다. 학습 데이터의 서비스 도메인은 k개인데 HDP 모델이 추출한 토픽 개수는 $m \neq k$ $(m>0)$인 m이라는 점을 주목한다. 학습에 사용되지 않은 테스트용 사용자 발화 문장 $S_{test} = \{w_i \mid 1 \le i \le |S_{test}|\}$가 $|S_{test}|$개의 단어로 이루어져 있다고 할 때, 학습이 완료된 HDP 모델의 각 단어 w_i에 대한 토픽 분포 값 Φ_{kwi} 값으로 이루어진 m 차원의 벡터 $F_{Stest} = (\Phi_{1wi}, ..., \Phi_{2wi}, ..., \Phi_{mwi})$를 자질 벡터로 생성한다. 테스트 문장$S_{test}$ 이 미지원 도메인인지 여부를 나타내는 class label C_{Stest}와 자질 벡터 F_{Stest}를 사용하여 미지원 도메인 분류기가 cross validation 등의 기법을 사용하여 학습 및 테스트된다. 미지원 도메인 분류기는 SVM, Decision tree,

Random forest [28] 등의 다양한 분류기가 사용될 수 있다.

자질 생성을 위한 토픽 모델은 HDP 모델이 아닌 다른 토픽 모델들, 이를테면 latent Dirichlet allocation (LDA) 모델 [24] 등을 사용할 수도 있다. LDA 모델 및 이에 기반한 기존 토픽 모델들은 정해진 토픽 개수, 즉 서비스 도메인 개수만큼 토픽들을 생성해내는 반면, HDP 모델은 문서에 내재된 패턴을 기반으로 토픽 개수를 자동으로 인식한다는 차이가 있는데, 본 연구에서는 이러한 HDP 모델의 특징이 OOD 검출에 더 유리하다고 보고, HDP 모델에서 생성되는 토픽에 대한 분포를 자질로써 활용하는 것이다. 미지원 도메인은 무수히 많은 개수의 도메인이라고 생각할 수 있으며, 사람이 인지하지 못하는 많은 패턴을 가지고 있을 것이라고 짐작할 수 있다. 불행히도 활용 가능한 데이터는 서비스 도메인 데이터밖에 없는 경우가 많기 때문에, 이 데이터로부터 HDP 모델을 통해 가능한 한 많은 개수의 미지의 패턴을 추출하여 미지원 도메인 검출을 위한 자질로써 활용하고자 하는 것이다. 다음 장에서는 HDP 모델에서 생성되는 토픽이 기존 토픽 모델(예: LDA 모델)에서 생성되는 토픽에 비해 미지원 도메인 검출에 더 유용하다는 것을 실험을 통해 검증한다.

Table 1. Statistics of dataset

Domain	Number of sentences
alarm	2,729
calendar	2,460
campus	127
control	4,388
emergency	166
sports	3,057
weather	4,792

Table 2. Sample sentences of each domain

Domain	Sample sentences
alarm	- 1시 알람 풀자 (Cancel alarm 1 PM) - 오후 3시 알람 제거 (Remove alarm 3 PM)
calendar	- 할 일이 뭐예요? (Anything to do?) - 오늘의 일정 (Today schedule)
campus	- 내 학점 몇이야? (How is my grade?) - 제 학번은 20171234입니다 (My student ID is 20171234)
control	- 이거 꺼 (Turn this off) - 취소하자 (Cancel it)
emergency	- 불불불 (Fire, fire, fire) - 도둑이야 (Thief)
sports	- 오늘 넥센 경기하니? (Any game of Nexen, today?) - 오늘 경기 하려나 (Any game, today?)
weather	- 빨리 날씨를 알고 싶어 (Wanna know today weather ASAP) - 오늘의 기온 (Today temperature)

IV. Experiments

본 연구에서 실험에 사용한 데이터는 7개의 대화시스템 도메인에 대하여 구축한 한국어 데이터셋을 사용하였다. 각 도메인 별 문장 개수는 Table 1과 같으며, 도메인 별 예시 문장은 Table 2과 같다.

각 도메인을 한 번씩 미지원 도메인으로 가정하여 실험하는 deleted interpolation [9] 기법을 사용하여 실험하였으며, HDP 모델의 토픽이 OOD 검출을 위한 자질로서 더욱 적합한지 확인하기 위해 가장 대표적인 토픽 모델인 LDA 모델에서 생성되는 자질과 비교하였다. LDA 모델이 아닌 다른 토픽 모델에서 생성되는 토픽들과 추가비교하면 더 좋겠지만, 미지원 도메인 검출을 위해 설계되었던 토픽 모델들이 없었던만큼, 가장 대표적인 토픽 모델인 LDA 모델과 비교하는 것이 현 시점에서는 타당하다고 볼 수 있다. 또한, 본 실험은 토픽 개수를 자동으로 찾아내는 HDP 모델의 토픽이 미지원 도메인 검출에 더 적합하다는 것을 증명하는 것이 목적이므로, 기존의 대표적인 토픽 모델인 LDA 모델과 비교하도록 한다. LDA 모델 학습을 위해 토픽 개수는 서비스 도메인 개수와 일치하도록 7개로 하였고, HDP 모델은 1개부터 시작하여 자동으로 토픽 개수가 결정되도록 하였다. LDA 모델의 학습을 위한 파라메터 α는 0.01, β는 0.0001로 하였으며, HDP 모델 파라메터 α_0는 0.01, α_1은 0.01, β는 0.0001로 설정하였다.

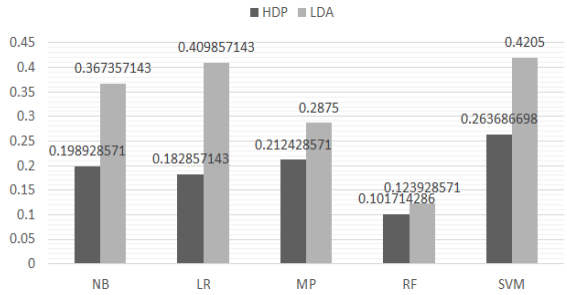

Fig. 5. Equal error rate (EER) comparison between different classifiers with HDP-generated or LDA-generated features, where vertical axis represents EER and horizontal axis indicates classifiers

HDP 모델과 LDA 모델 학습에 사용하였던 데이터는 미지원 도메인 분류기 성능 검증에 사용하지 않았다. 다시 말해서, 서비스 도메인에 해당하는 데이터를 9:1의 비율로 나누고, 9에 해당하는 데이터는 HDP 모델 학습과 미지원 도메인 분류기 검증을 위해 사용하고 나머지는 분류기 검증에 사용하였다. 만약 alarm 도메인을 미지원 도메인으로 가정할 경우, 나머지 6개의 서비스 도메인에 해당하는 데이터는 9:1의 비율로 나누어서 HDP 모델 학습과 미지원 도메인 분류기 검증에 사용하고, 미지원 도메인인 alarm 도메인의 데이터는 미지원 도메인 분류기 검증에만 사용하였다. 이처럼 토픽 모델 학습에 사용되는 데이터를 미지원 도메인 분류기에 사용하지 않는 이유는, 미지원 도메인 분류기를 대화시스템 등에 적용할 경우 미지원 도메인이 대한 데이터 없이 토픽 모델 학습이 이루어지는 것을 감안한 것이다. 즉, 실제 대화시스템 제품 개발에 적용되는 상황을 가정하여 실험하였다고 볼 수 있다.

HDP 모델과 LDA 모델에서 생성된 토픽들에 대해 다양한 기계학습 모델들을 미지원 도메인 분류기로서 실험을 수행하였으며 결과는 Fig. 5와 같다. Deleted interpolation 기법을 통해 7개 도메인을 한 번씩 미지원 도메인으로 가정하였으므로 7번 실험하여 얻은 수치들의 평균을 사용하였다. 성능 비교를 위해 적용한 지표는 equal error rate (EER)인데, 이 값은 false acceptance rate (FAR)과 false rejection rate (FRR) 값이 같을 때의 값을 의미하며, 이 값은 서로 다른 미지원 도메인 분류기 간의 성능 비교를 위해 주로 사용된다. EER 값이 낮을수록 더 좋은 분류기라고 볼 수 있다. 미지원 도메인 분류기로 사용한 각 모델에 대한 설명 및 파라메터 설정은 Table 3과 같다. 특히, MP의 자질 개수 설정 시, 자질 개수에 따라 적절한 hidden layer 개수를 설정하기 위해 hidden layer 개수는 (자질 개수 + 클래스 개수)/2로 설정하였다. 즉, LDA 모델에 대해서는, 토픽 분포로부터 생성되는 자질은 7개이고 클래스는 2개이므로 hidden layer는 4개가 된다.

Fig. 5에서처럼, HDP 모델에서 생성된 토픽 분포를 자질로 사용한 경우가 LDA 모델의 토픽 분포를 사용한 경우보다 모든 모델에서 전반적으로 성능이 향상된 것을 알 수 있다. 이는 HDP 모델에서 추출된 불특정한 개수의 토픽들이 미지원 도메인 검출을 위한 임의의 패턴을 잡아냈고, 이 토픽 분포를 활용하였기 때문에 성능이 향상된 것으로 해석할 수 있다.

가장 좋은 성능을 보인 분류기 모델은 약 0.1017의 EER 수치를 보인 random forest (RF)이다. 사용한 데이터가 다르기는 하지만 기존 관련연구들에서 EER 수치가 0.196 [9], 0.153 [20], 0.149 [29], 약 0.11 [13] 이었고, 최근에 이르러서야 딥러닝 기술을 적용하여 0.0702 [22]를 기록했었다. 특히, 딥러닝 기술을 적용한 연구 [22]에서는 사전에 학습된 워드 임베딩 벡터를 사용하였다는 점을 감안할 때, 순수 텍스트에 토픽 모델을 적용하여 얻은 토픽 분포를 자질로 활용한 random forest 분류기의 성능은 기존 연구들과 견주어 손색이 없다고 볼 수 있다.

LDA 모델은 토픽 개수를 7개로 정하여 학습한 반면, HDP 모델은 토픽 개수가 자동으로 설정되게 하였기 때문에 토픽 개수가 다르다는 점에서 공정한 비교가 아니었다고 볼 수도 있다. 가령, Fig. 5 실험결과에서 HDP 모델을 학습하였을 때 자동으로 설정된 토픽 개수는 도메인 별로 작게는 68개부터 크게는 82개였는데, LDA 모델은 7개의 토픽만을 추출하였기 때문에 공정하지 않다고 볼 수도 있다. 하지만, 본 연구는 HDP 모델이 미지의 패턴을 자동으로 감지하여 토픽 개수를 설정한다는 점에 착안하여 미지원 도메인 검출에 적용한 것이므로, 공정한 비교를 위해 HDP 모델이 7개의 토픽만을 추출하도록 억지로 조정하는 것은 바람직하지 않다. 반대로, HDP 모델에서 자동으로 설정된 토픽 개수만큼 LDA 모델에서 추출하도록 하는 것은 HDP 모델과의 성능 비교는 가능하겠지만, 미지원 도메인 인식 기능이 대화시스템 등에 실제 적용될 때 LDA 모델에 설정할 적절한 토픽 개수를 알 수 없기 때문에 실용적인 관점에서 의미 없는 실험이라고 볼 수 있다. 정리하자면, LDA 모델과 HDP 모델은 그 구조와 기능이 다르며, HDP 모델이 자동으로 적절한 토픽 개수를 찾아낸다는 특징을 미지원 도메인 검출 문제에 적용하는 것이므로, 동일한 토픽 개수를 설정하여 두 모델을 비교하는 것은 목적에 어긋나므로 바람직하지 않다.

Table 3. Classifiers and settings

Classifier	Definition and setting
NB (naive bayes)	- Probabilistic classification model that is based on bayes' theorem which assums the independence between features
LR (logistic regression)	- Probabilistic model utilizing a linear combination of independent features - Trained via ridge estimator
MP (multilayer perceptron)	- Feed-forward neural networks having at least three perceptron layers - Number of hidden layer = (#features + #classes)/2 - Mini-batch size=100, learning rate=0.3, momentum=0.2, #epochs=500 - Trained via back-propagation algorithm
RF (random forest)	- Kind of ensemble model that generates final result by incorporating results of multiple decision-trees - #trees=100 - #features=log(#trees)+1 - Each tree has no depth-limitation
SVM (support vector machine)	- Non-probabilistic binary classification model that finds a decision boundary with a maximum distance between two classes - Kernel: Poly - Exponent=1.0, Complexity c=1.0 - Trained via sequential minimal optimazation algorithm [30]

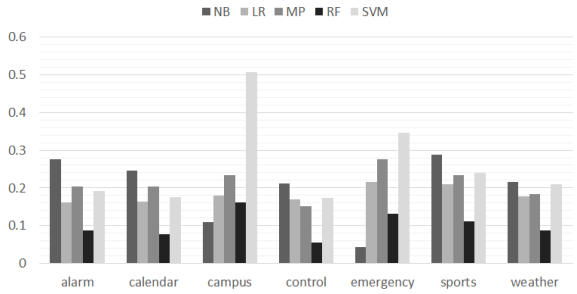

Fig. 6. Equal error rate (EER) of the classifiers between different targeted out-of-domains with HDP-generated features, where vertical axis represents EER and horizontal axis indicates domains

Fig. 5는 각 도메인을 미지원 도메인으로 가정한 실험들의 평균을 표시함으로써 HDP 모델에서 생성된 토픽 분포가 자질로서 유용함을 증명했다면, Fig. 6은 각 도메인 별로 HDP 모델의 자질을 사용한 분류기 모델의 성능을 표시하였다. 전반적으로 random forest 분류기가 가장 낮은 EER 수치를 보였지만, campus 도메인과 emergency 도메인에서는 naive bayes 분류기가 random forest보다 더 낮은 EER 수치를 기록한 것을 볼 수 있다. 이는 campus 도메인과 emergency 도메인의 문장 개수가 타 도메인의 문장 개수보다 적기 때문에 타 도메인에 비해 비

교적 적은 패턴들이 존재하고, 이로 인해 모델 복잡도가 상대적으로 작은 naive bayes가 이 도메인들에 대하여 더 좋은 성능을 보였다고 해석할 수 있다. 만약 campus와 emergency 도메인의 문장 개수가 충분히 많아지게 되면 이 도메인들에서도 random forest 분류기가 가장 좋은 성능을 보이게 될 것이다.

V. Conclusions

본 연구에서는 대화시스템 등에 필요한 기능인 미지원 도메인 인식 기능을 위한 새로운 시도를 소개하였다. 토픽 모델에서 생성되는 토픽들에 대한 분포를 자질로 활용하여 미지원 도메인 분류기 학습에 이용하였는데, 특히 자동으로 토픽 개수를 찾아내는 HDP 모델을 사용함으로써 서비스 도메인에 해당하는 문서들로부터 임의의 토픽들을 추가적으로 감지하게 하였다. 7개 도메인 데이터에 대한 실험을 통해, 기존의 대표적인 토픽 모델인 LDA 모델보다 HDP 모델에서 생성된 토픽 분포가 더 자질로서 미지원 도메인 검출에 유용하다는 점을 증명하였다. 인공지능형 대화시스템 개발의 난제로 꼽히는 미지원 도메인 검출 성공률을 높임으로써, 대화시스템과 관련된 다양한 산업 분야에 활용될 것으로 기대된다. 향후 연구의 방향은, 대화시스템 데이터에 특화된 응용 모델을 HDP 기반으로 설계함으로써 검출 성능 향상을 꾀할 수 있을 것이다. 또한 나아가, 딥러닝 계열의 최신 기술 기반의 분류기 모델에 HDP 모델의 자질을 활용하는 연구도 진행할 수 있을 것이다.

REFERENCES

[1] Amazon Echo, https://www.amazon.com

[2] SK Telecom Nugu, http://www.nugu.co.kr/main

[3] Naver WAVE, https://clova.ai/ko

[4] Naver AWAY IVI platform, http://www.naverlabs.com/newsroom/IVI.html

[5] Hyundai-card Buddy, www.hyundaicard.com/

[6] Naver search, https://www.naver.com

[7] Google search, https://www.google.com

[8] Lidia Mangu and Eric Brill, “Lattice Compression in the Consensual Post-Processing Framework,” in Proceedings of the Third World Multiconference on Systemics, Cybernetics and Informatics, pp. 246-252, 2000.

[9] Ian R. Lane, Tatsuya Kawahara, Tomoko Matsui, and Satoshi Nakamura, “Out-Of-Domain Detection Based on Confidence Measures From Multiple Topic Classification,” in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, pp. 757-760, 2004.

[10] Seonghan Ryu and Donghyeon Lee, Gary Geunbae Lee, Kyungduk Kim, and Hyungjong Noh, “Exploiting Out-Of-Vocabulary Words For Out-Of-Domain Detection in Dialog Systems,” in Proceedings of the International Conference on Big Data and Smart Computing, pp. 165-168, 2014.

[11] Deirdre Hogan, Johannes Leveling, Hongyi Wang, Paul Ferguson, and Cathal Gurrin, “SMS Normalisation, Retrieval and Out-of-Domain Detection Approaches for SMS-Based FAQ Retrieval,” Multilingual Information Access in South Asian Languages, pp. 184-196, 2013.

[12] Mikio Nakano, Shun Sato, Kazunori Komatani, Kyoko Matsuyama, Kotaro Funakoshi, and Hiroshi G. Okuno, “A Two-Stage Domain Selection Framework for Extensible Multi-Domain Spoken Dialogue Systems,” in Proceedings of the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Portland, Oregon, USA, pp. 18–29, June 17-18, 2011.

[13] Yoko Fujita, Shota Takeuchi, Hiromichi Kawanami, Tomoko Matsui, Hiroshi Saruwatari, and Kiyohiro Shikano, “Out-of-Task Utterance Detection Based on Bag-of-Words Using Automatic Speech Recognition Results,” in Proceedings of the third Annual Summit and Conference of Asia-Pacific Signal and Information Processing Association, Xi’an, China, October 18-21, 2011.

[14] Wordnet, https://wordnet.princeton.edu/

[15] Young-Seob Jeong, “Experimental Analysis for Out-Of-Domain Detection Using Features of Word Positions in Sentences,” in Proceedings of the Spring Conference of Korean Society for Internet Information, April 21-22, 2017.

[16] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, “The elements of statistical learning,” Springer, New York, 2003.

[17] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman, “Indexing by Latent Semantic Analysis,” Journal of the American Society for Information Science (JASIS), Vol. 41, No. 6, pp. 391-407, 1990.

[18] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik, “A Training Algorithm For Optimal Margin Classifiers,” in Proceedings of the fifth Annual Workshop on Computational Learning Theory, Pittsburgh, PA, USA, pp. 144-152, 1992.

[19] David A. Freedman, “Statistical Models: Theory and Practice,” Cambridge University Press, 2009.

[20] Ian R. Lane, Tatsuya Kawahara, and Shinichi Ueno,

"Example-Based Training of Dialogue Planning Incorporating User and Situation Models," in Proceedings of the 8th International Conference on Spoken Language Processing, Jeju, Korea, October 4-8, 2004.

[21] Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra, "A Maximum Entropy Approach to Natural Language Processing," Computational Linguistics, Vol. 22, No. 1, pp. 39-71, 1996.

[22] Seonghan Ryu, Seokhwan Kim, Junhwi Choi, Hwanjo Yu, and Gary Geunbae Lee, "Neural Sentence Embedding Using Only In-Domain Sentences For Out-Of-Domain Sentence Detection in Dialog Systems," Pattern Recognition Letters, Vol. 88, pp. 26-32, 2017.

[23] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei, "Hierarchical Dirichlet Processes," Journal of the American Statistical Association, Vol. 101, No. 476, pp. 1566-1581, 2006.

[24] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, 3, pp. 993-1022, January, 2003.

[25] Machal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers, "Learning Author-Topic Models," ACM Transactions on Information Systems, Vol. 28, No. 1, pp. 1-38, 2010.

[26] Young-Seob Jeong and Ho-Jin Choi, "Sequential Entity Group Topic Model for Getting Topic Flows of Entity Groups Within One Document," in Proceedings of the 16th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Kuala Lumpur, Malaysia, pp. 366-378, 29 May ~ 1 June, 2012.

[27] Thomas L. Griffiths and Mark Steyvers, "Finding Scientific Topics," in Proceedings of the National Academy of Sciences of the United States of America, pp. 5228-5235, 2004.

[28] Leo Breiman, "Random Forests," Machine Learning, Vol. 45, No. 1, pp. 5-32, 2001.

[29] Ian R. Lane and Tatsuya Kawahara, "Incorporating Dialogue Context and Topic Clustering in Out-of-Domain Detection," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, Pennsylvania, USA, pp. 1045-1048, March 18-23, 2005.

[30] John C. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines," Advances in Kernel Methods - Support Vector Learning, MIT Press, 1998.

[31] Sepp Hochreiter and Jurgen Schmidhuber, "Long Short-term Memory," Neural Computation, Vol. 9, No. 8, pp. 1735-1780, 1997.

Authors

Young-Seob Jeong received a BS in Computer Science from Hanyang University, Korea, in 2012, an MSc in Computer Science from KAIST, Korea, and in 2016, a PhD in School of Computing from KAIST, Korea. He joined the faculty of the Department of Big Data Engineering at Soonchunhyang University, Asan city, Korea, in 2017. His current research topics are text mining, information extraction, action recognition, and dialog systems, where his favorite techniques are topic modeling and deep learning.