Journal of Convergence for Information Technology
Vol. 9. No. 9, pp. 1-12, 2019
e-ISSN 2586-4440
DOI : https://doi.org/10.22156/CS4SMB.2019.9.9.001

Survey on Out-Of-Domain Detection for Dialog Systems

Young-Seob Jeong*, Young-Min Kim
Professor, Department of Big Data Engineering, Soonchunhyang University

대화시스템 미지원 도메인 검출에 관한 조사

정영섭*, 김영민
순천향대학교 빅데이터공학과 교수

Abstract A dialog system becomes a new way of communication between human and computer. The dialog system takes human voice as an input, and gives a proper response in voice or perform an action. Although there are several well-known products of dialog system (e.g., Amazon Echo, Naver Wave), they commonly suffer from a problem of out-of-domain utterances. If it poorly detects out-of-domain utterances, then it will significantly harm the user satisfactory. There have been some studies aimed at solving this problem, but it is still necessary to study about this intensively. In this paper, we give an overview of the previous studies of out-of-domain detection in terms of three point of view: dataset, feature, and method. As there were relatively smaller studies of this topic due to the lack of datasets, we believe that the most important next research step is to construct and share a large dataset for dialog system, and thereafter try state-of-the-art techniques upon the dataset.

Key Words : Dialog system, User utterance, Out-of-domain detection, Natural language understanding, Text classification

요 약 대화시스템은 인간과 컴퓨터 사이의 새로운 의사소통 수단으로 떠오르고 있다. 대화시스템은 인간의 음성을 입력으로 취하여, 적절한 음성 답변 또는 서비스를 제공하게 된다. 아마존 에코, 네이버 웨이브 등과 같은 대화시스템 제품들이 등장하고 있음에도 불구하고, 이 대화시스템들은 공통적으로 미지원 도메인을 제대로 처리하지 못한다는 문제점을 안고 있다. 이와 관련한 몇몇 연구들이 있었지만, 이 문제를 풀기 위한 더욱 많은 연구가 진행될 필요가 있다. 이 논문에서는, 미지원 도메인 검출과 관련한 기존 연구들에 대하여 3가지 관점, 즉 데이터, 자질, 방법에 대한 관점으로 요약한 정보를 제공한다. 데이터셋이 부족하다는 점으로 인해 타 연구분야에 비해 적은 연구가 수행되어왔으므로, 앞으로 가장 시급한 연구 주제는 대화시스템의 미지원 도메인 검출을 위한 공개용 데이터셋을 구축하고 배포하는 것이다.

주제어 : 대화시스템, 사용자 발화, 미지원도메인 검출, 자연어 이해, 텍스트 분류

*This work was supported by the Soonchunhyang University Research Fund. 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. 2019021348).

*Corresponding Author : Young-Seob Jeong(bytecell@sch.ac.kr)

Received July 2, 2019
Revised August 19, 2019
Accepted September 20, 2019
Published September 28, 2019

1. Introduction

Since the World Wide Web (WWW) has appeared, it has become possible to find the designated information by navigating the web pages. We navigate the web by just giving queries to search engines such as Google [1] or Yahoo [2], and we are able to get desired information by just clicking the one among listed items. With the appearance of smartphones later, arbitrary applications providing desirable services can be found by giving the queries to app markets (e.g., Android market). This gave a huge change to human life, as the smartphone users could be provided with any kinds of services by just touching their smartphones anywhere. We are now facing another huge change provided by dramatically improved techniques of speech recognition and natural language understanding. The speech recognition (SR) converts a given human-voice signal into a list of promising texts or a lattice; thereby allowing of development of various services (e.g., question-answering system, dialog system) using the techniques of natural language understanding. The natural language understanding (NLU) converts the natural language texts into a formatted information within computer's comprehension. The advance of speech recognition techniques makes it possible to accurately recognize spoken utterances, while the improvement of natural language understanding opens the conversational way of interaction between computers and human. This will allow smartphone users to get desirable services by just saying their needs to intelligent conversational agent, which is also typically called as chatbot or dialog system.

The standard pipeline architecture of the dialog system is described in Fig. 1. The first step of the architecture is Automatic Speech Recognition (ASR), which takes the speech of user utterance as an input and gives one or more hypotheses of the utterance as an output. The output of ASR step might be a list of top promising text of the user utterance [3], or may be a lattice of weighted sequence of promising words [4]. The second step, namely Natural Language Understanding (NLU), usually takes the top N hypotheses of the utterance as an input and generates structured information as an output. The structured information may contain the promising domain or intention of the given utterance hypothesis, or subject/object of the intention. The NLU part usually works without considering contextual information, such as current date (or time), previous utterance, and previous system response. In the third step, Dialogue Manager (DM) is supposed to deal with the contextual information, so that the system reaction can be consistent with the conversational context. Natural Language Generation (NLG), the fourth step, takes the generated system reaction as an input, and generates a system response in a form of natural language. The final step, Text-To-Speech (TTS), generates speech of system response that will be given to the user. There are some studies that do not follow this pipeline. For example, end-to-end systems [5] generate system responses directly from the user speech, and scheduler-based approaches [6] add an abstract layer, namely a scheduler, that manages the context of user utterance in real-time. The scheduler does not always follow the pipeline as it manages the context without the DM.

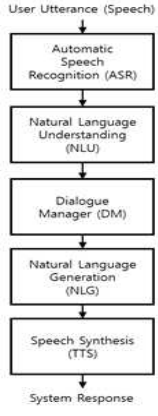

Fig. 1. Pipeline architecture of dialog system

Natural Language Processing (NLP) framework usually takes word tokenization, sentence splitter, morphological analysis, Part-Of-Speech (POS) tagging, shallow parsing (i.e., chunking), Named Entity Recognition (NER), syntactic parsing, and semantic role labeling [7]. The NLU process of dialog system has similar structure to the NLP framework, except that the NLU process usually has domain detection and intention prediction [8]. For the user utterance 'How is the weather today?', then its domain and intention might be 'weather' and 'inform', respectively. As the domain detection is usually performed as a first step of the NLU process, the result of it has a huge impact on the following other steps (e.g., slot filling, NER).

There have been many studies related to the domain detection [9-12], but relatively smaller studies aimed at out-of-domain (OOD) detection problem. The purpose of OOD detection is to predict whether the given sentence belongs to the service domains or not. Given service domains {whether, sports}, the sentence 'Who is Lady Gaga?' must be classified as OOD, whereas the sentence 'How is the weather today?' must be classified as in-domain (ID). If the first sentence above is misclassified as ID, then the system will probably give a poor answer to the user. For example, if the first sentence is misclassified as 'sports', then the system may give the answer 'There is no soccer player named Lady Gaga'. There have been only few studies related to the OOD detection, and there were no survey or review papers focused only on this topic, as far as we know. In this paper, we give a summary of the previous studies related to the OOD detection, and discuss about future research direction.

We examine the previous studies about the OOD detection in terms of three points of view: data, feature, and method. We showed that support vector machines (SVM) with n-gram features was the best in earlier studies [13], and deep learning technique proved its potential recently. We suggest that the most important step of future research direction must be to construct a public dataset.

The remainder of this paper is organized as follows. Section 2 describes definition of OOD detection. Section 3 gives a summary of previous studies, and Section 4 discusses about the future research direction and concludes.

2. Out-of-domain Detection

Domain of a user utterance is a field at a certain level that the utterance belongs to. For example, the utterance 'Who is the singer Michael Jackson?' might belong to 'music' domain or 'person' domain. Another utterance 'Play any pop music' may belong to 'pop-music' domain at a level of music genre or 'music' domain at a greater level. If there is a set S of service domains, then the number of domains |S| and variety of the domains are determined by human (e.g., a project leader). It should be noted that the domain is different from 'intention'. The intention means a purpose of the utterance while the domain is the field of the utterance. For the sentence 'Who is the MVP of tonight game?', the domain and intention can be 'sports' and 'inform', respectively. Both of domain and intention are strongly related to the services or functions provided by the product.

During communication between a user and a dialog system, the user is supposed to give utterances of service domains. If the user says an utterance that does not belong to any service domain, then the dialog system should give a response that the utterance is not comprehensible. While previous studies denote such utterance with different names such as out-of-domain utterance, out-of-task utterance or orphan utterance [14], we call it as out-of-domain (OOD) utterance in this paper. More formally, we define in-domain (ID) utterance as the one belonging to any service domain and corresponding service is provided, and

define out-of-domain (OOD) utterance as the one that does not belong to any service domain or corresponding service is not available. The corresponding service is closely related to the intention. Even if an utterance belongs to a certain service domain, it will be OOD when its corresponding function or service is not provided by the system. For example, if there is a service domain 'music' with only one function 'play music', then the intention of the utterance 'who is this singer?' will not be recognizable because there is only one function 'play music'. According to the definition of OOD in this paper, all of these out-of-intention utterances are also OOD.

The task of OOD detection is to recognize whether a given utterance is OOD utterance or ID utterance. There is no intersection between the OOD and the ID, so the OOD detection is essentially a binary classification problem. One may argue that the OOD detection is similar to addressee detection [15] or dialog act tagging [16, 17]. The addressee detection is to detect who the speaker is talking to, and it might be related to the OOD detection problem if we assume that the dialog system is one of the addressees. The biggest difference between OOD detection and the addressee detection is that the OOD detection works in an assumption that the addressee of user utterances is the system. That is, the dialog systems usually employ particular names (e.g., Alexa, OK Google, Sally); thereby allowing the users to call the system whenever they want. Thus, the OOD detection is a problem at a different level from the addressee detection. The dialog act tagging is to predict the type of an utterance (e.g., question, order), and it might be related to the OOD detection problem if we assume that the dialog system offers only a certain type of utterance. For example, if a dialog system provides only question-answer services, then only question-style utterances can be ID utterances. However, the dialog system probably has a finite set of target domains for question-answer service, and some question-style utterances that do not belong to the target domains might be accepted. For example, if the dialog system offers question-answer service on a limited set of domains {music, movie}, the question-style utterance 'Miller is the MVP of tonight game?' might be accepted because its dialog act is 'yes-no-question'. Thus, the OOD detection problem is different from the dialog act tagging.

The OOD detection is usually performed either at the beginning part or at the ending part of the NLU process. Assume that we have two service domains {weather, transport}, and we are given a sentence 'Who is the Michael Jackson?.' If we perform the OOD detection at the beginning part, then we will have to consider two tasks concurrently: domain classification and OOD detection. The domain classification will results in a promising domain with its confidence (e.g., weather (30%)), while the OOD detection will give the probability that the sentence is OOD (e.g., 80%). The final decision can be made by incorporating the two results in particular ways (e.g., thresholds). On the other hand, if we perform the OOD detection at the ending part, then there must be much more information about the given sentence (e.g., extracted slots, intention, dependency parsing tree, semantic relations). It is obvious that it is time-consuming that we perform the whole NLU process for the potential OOD sentence; we may achieve better decision about the OOD detection with a loss of efficiency. Anyhow, there is no doubt that the dialog system will fail without carefully designed OOD detection.

To summarize, the OOD detection is to predict whether a given utterance is OOD or ID, where the OOD utterance does not belong to any service domain or corresponding service is not available. In the following section, related studies are introduced.

3. Previous Studies

Since there have been many studies about domain detection [9-12], relatively smaller studies about OOD detection have appeared. The reason is that there is no sufficient dataset because the number of OOD can be infinite while the number of service domains is finite. Thus, in many cases, only a dataset of in-domain is provided. Including this data-related issue, we provide three points of view to discuss the previous studies: datasets, features, and methods. In the following subsections, detail of each of the view-points is described.

3.1 Datasets

Although there are several datasets that can be used for domain detection, most of these datasets are useless for OOD detection because all documents of the datasets are in-domains. Moreover, it is not trivial to construct a dataset for OOD detection, because the number of OOD can be infinite. Deleted interpolation detours this problem by treating each service domain as OOD [18-20]. If we have three service domains {music, movie, sports}, then the 'music' domain is assumed to be OOD, while the other two domains are regarded as in-domain. To make it fair, of course, each of the other two domains {movie, sports} is also assumed to be OOD.

As the deleted interpolation makes it possible to make use of the datasets of in-domains for OOD detection, we specify some available datasets for domain-detection in Table 1. The first four datasets can be found in UCI machine learning repository, while the last one can be found in [21]. Spam collections can be used to OOD detection when the ham messages and spam messages are assumed to be ID and OOD, respectively. Scientific articles have research domains and review documents are collected from multiple product domains, so that these documents can be used for OOD detection. Unfortunately, all of these datasets have a common limitation that their target application is not a dialog system. Most of previous studies assumes that it works as a part of a dialog system, because the OOD detection is necessary only when the system input can be any domain. The dialog system is the one of such systems and receives huge public attention recently. Thus, it is necessary to construct and share a large amount of utterances for the dialog system, and greater number of domains will be more helpful.

Table 1. Datasets for domain detection, which can also be available to OOD detection

Name	Explanation
Sentence Classification Dataset	Scientific articles annotated with Argumentative Zones annotation scheme, where the articles come from three domains {PLoS Computational Biology, The machine learning repository on arXiv, The psychology journal Judgment and Decision Making}
SMS Spam Collection Dataset	A collection of SMS spam messages and randomly chosen ham messages
YouTube Spam Collection Dataset	A collection of spam/ham messages extracted from five YouTube videos
OpinRank Review Dataset	Car reviews and hotel reviews collected from Tripadvisor and Edmunds
Yahoo Answers	Documents of 10 topics: Society & culture, Science & Mathematics, Health, Education & Reference, Computers & Internet, Sports, Business & Finance, Entertainment & Music, Family & Relationships, Politics & Government

3.2 Features

The features used in OOD detection fall into three categories: text-based features, resource-based features, and system-based features. The summary is shown in Table 2. The text-based feature is extracted from the text itself. The early studies focused on the n-gram features [18-20], where the n-gram features were usually constructed using a word-base form, a word-surface form, and part-of-speech (POS) tags. For a verb 'gave', its word-base form and word-surface form are 'give' and 'gave', respectively.

The intuition behind of the n-gram features is that the more familiar (i.e., already seen) words in a sentence, the more likely the sentence is in-domain. It is possible, of course, that we take the opposite intuition; the more unseen words in a sentence, the more likely the sentence is OOD. In [22], indeed the number of words and out-of-vocabulary (OOV) frequency are adopted as features, where the OOV frequency means the number of words that have not appeared in the in-domain documents. The n-gram feature basically assumes that the words are independent to each other, so it does not incorporate a sequence or a position of the words. In [23], it shows that the word positions can help to achieve better accuracy of the OOD detection.

Other than the n-gram features and word positions, there are some other text-based features: semantic role (SR) labels, named-entity (NE) labels, and head/tail relations of syntactic parse trees. There was very few studies that utilized such features [9], because these features are usually determined for a specific domain. For the sentence 'Michael Jackson has died', the term 'Michael Jackson' can be annotated with NE label 'singer' of music domain, 'scientist' of science domain, or 'soldier' of military domain. Note that, for the same entity 'Michael Jackson', there are many possible NE labels of different domains. Thus, if one may utilize the NE labels, then NE labels from all possible domains should be obtained before the OOD detection. Obviously this will degrade the overall efficiency (i.e., response time) of the dialog system.

The resource-based feature is extracted from other arbitrary linguistic resources (e.g., Wordnet [24]) or methods. For example, scores computed using a database of question-answer pairs are adopted [22]. Although the database was not built for the OOD detection, using the database contributed to performance improvement. In [25], results from search engines are employed. It firstly normalizes the given text using a set of rules, then extracts features by applying the query to the search engine. The features include performance estimates of the search engine for the given query, the size of result, and a document score computed by the search engine. These features are given to the OOD classifier. In [26], based on a hand-crafted dictionary and results from previously constructed domain model, unigram features are defined. Note that the resource-based feature strongly depends on the resources or methods. Moreover, if one needs to achieve better efficiency (i.e., response time), then it will be better to minimize dependency to the resource-based features because of time and expense.

The dialog system is supposed to take an action and give a response to a given user utterance. When all of the system actions and user utterances are recorded, they might be used to recognize user intention more accurately. For example, given a previous utterance 'How old is Obama?', it is natural to infer the intention of the current utterance 'How about Trump?' is to ask the age of Trump. The system-based feature is extracted from the recorded history of system actions and utterances. It also includes the context of the other modules in the system. In [27], a type of previous utterance and a previous response of the system are adopted. It also uses the number of results from the automatic speech recognition (ASR) module as a feature, and takes the confidence scores for all domains except the previous domain as a clue for final decision. These features, essentially, are related to the structure of the dialog system, because they are obtainable only when the dialog system computes confidence values of multiple domains. That is, the decision for OOD detection is made when the dialog system partially completes the analysis on the utterance, so these features will not be available to a particular dialog system that performs the OOD detection as a first step. If one may use the system-based features, then the

structure of dialog system must be carefully designed in order to keep extensibility and robustness [27].

The text-based feature is extracted from an utterance, so it is relatively faster than the other types of features. That is, the resource-based feature uses the results of some other resources or methods, so it will take more time to generate features. The system-based feature also needs more time, because it is often to consider all of the previous system actions, utterances, and system response of every possible domain. Therefore, if one wants to achieve better efficiency, then it will be better to take only the text-based feature.

Table 2. Three categories of features for OOD detection

Category	Example
Text-based (extracted from the raw text)	- Lexical features (e.g., n-grams), - Syntactic parse, - Semantic role labels, - Named entity labels
Resource-based (extracted from other arbitrary linguistic resource or method)	- Synonyms - Word category - Result of search engines
System-based (extracted from the system itself)	- Previous system responses - Previous user utterances - Current system actions - Context of the other modules

3.3 Methods

Previous studies adopted various methods such as linear discriminant model (LDM) [28], latent semantic analysis (LSA) [29], support vector machines (SVM) [13], logistic regression [30], maximum entropy model (MEM) [31], IB1 algorithm, and neural networks. To measure the performance of the methods, equal error rate (EER) is widely adopted. The EER value is obtained when false rejection rate (FRR) and false acceptance rate (FAR) are equal to each other. The FRR is the measure of the likelihood that the utterance of ID is rejected, while the FAR is the measure of the likelihood that the utterance of OOD is accepted. Smaller values of FRR and FAR must be better, but it is difficult to make both of them to be small at the same time. If a method achieves very small FRR value, then it usually loses FAR value, and vice-versa. When the FAR and the FRR intersect at a certain point, it is called the EER and it is used as a measurement to compare the performance of different methods. In this paper, we assume that the EER value ranges from 0 to 1, where smaller EER value is better.

In early studies, the SVM with n-gram features was the best. The SVM with trigram features achieved EER 0.196 in [18] and 0.153 [19]. In [20], the SVM is used to get distances between domains, and a hierarchical clustering algorithm is applied to the distances. The clusters of domains are used to compute confidence scores, and LDM with a threshold value as a classifier achieved EER 0.149. The datasets used by these studies are different to each other, so it is not fair if we simply compare the EER values of them. Nevertheless, from these EER values, we can get an insight about which model will be better and how much accurate it will be.

In addition to the n-gram features, some studies employed resource-based features or system-based features. In [21], given a question-answer database, the SVM with features of similarity scores between utterance words and the database recorded about EER 0.11. When the SVM is applied to 10-best ASR results, then it gave about EER 0.13. The SVM is also used with features of previous system responses and previous utterances [27]. It basically consists of two steps: (1) promising domain prediction using logistic regression, and (2) decision whether keep the previous domain or not. The second step uses the SVM, and achieved precision 0.829, recall 0.824, and F1 score 0.826. This study did not provide an EER value. In [25], a search engine is used to generate features such as performance estimates of the search engine, result set size, and document score. It has three steps: (1)

a rule-based text normalization, (2) getting results from a search engine, and (3) a decision by IB1 algorithm. It recorded an accuracy 0.848, and it also did not provide an EER value. In [26], boosted by results from previously constructed domain model, it utilized MEM and achieved EER 0.043. This study strongly depends on a hand-crafted list that used to generate features, which implies that it is hard to extend if there are some additional service domains to incorporate. Syntactic parse, Semantic parse, and n-grams are used as features for the SVM in [9], and it gave precision 0.18 and recall 1.0.

Recently, deep learning technique attracts attention, and it is successfully applied to various areas such as object recognition of images, speech recognition, speech synthesis, and text analysis. In [32], based on pre-trained word embedding vectors, sentence embedding vectors are computed using long short-term memory (LSTM) [33], and the sentence embedding vectors are fed to an auto-encoder as a decision maker for OOD detection. The intuition behind this study is that the clue for OOD detection should be not only in a sentence level, but also in a word level. It firstly constructs the word embedding vectors, and uses them to generate the sentence embedding vectors for the in-domain sentences. The auto-encoder works as a one-class classifier for the sentence embedding vectors. It determines whether the given sentence is OOD or not by a particular threshold value, and achieved EER 0.0702. In [34], a neural joint model for domain classification and OOD detection was proposed. It aims at satisfying a given false acceptance rate (FAR) while maximizing the domain classification accuracy. It achieved 0.05 FAR with domain classification accuracy 0.9038 on the dataset of 21 domains. This was the first approach to jointly address the OOD detection and the domain classification.

Note that all the above existing studies do not use the same dataset, so it is not fair to simply compare the EER values of them. Nonetheless, the EER values of them can give an insight that the deep learning technique is now promising, so it should be encouraged to conduct many studies using the deep learning technique for this task.

4. Future Research Direction

Summary of previous studies is described in Table 3. Widely used features were n-grams and dictionary-based tags, where some studies tried dialog context information (e.g., previous utterances, system response) or resources given by other systems (e.g., search engines). Such features might contribute somehow, but there was no intensive investigation to various features. It is necessary to try other types of features such as features based on Out-Of-Vocabulary (OOV) or domain-dependent natural language understanding. OOV words may be an evidence for OOD detection, but some OOV words belong to service domains. To address such case, the difference of character distributions between in-domain words and OOV words may be examined. Natural language understanding (NLU) module of dialog system extracts information or clues from a given textual utterance, in order to convert the unstructured user utterance into a structured form (i.e., frame). For a given utterance 'Play a music of Michael Jackson', named-entity (NE) extractor of the NLU module will extract a 'person' NE tag for 'Michael Jackson'. The NE tags are strongly related to its corresponding domain, so NE tags of different domains may be used as features for OOD detection.

For the beginning phase (i.e., 1990s ~ 2000s), the most dominant method was Support Vector Machines (SVM) that achieved around EER 11%. This implies that it is about 89% likely to correctly predict whether a given utterance is OOD or not. Deep-learning approach was recently adopted and it achieved about EER 7%. Although this may seem

a great advancement by deep-learning technique, we insist that we need to keep investigating other methods as well, because of the special characteristic of the task of OOD detection. Most previous problems are essentially similar to each other, as they are commonly 'finding particular patterns' over data. However, the OOD detection is not finding some desirable patterns, but 'finding prohibited patterns'. This is different from 'anomaly detection' that is to find unfamiliar observations. The biggest difference is that the OOD detection is to find prohibited patterns which may be familiar. In other words, the OOD detection requires to find some OOD utterances that even seem very similar to in-domain utterances. For example, given a set of service domains {weather, music}, the utterance 'Today is rainy, will the baseball game be cancelled?' must be OOD, even though its words are familiar to the 'weather' domain. Moreover, the datasets used by the previous studies are different, which means that it is not fair to compare the methods according to EER values.

To make it fair comparison, we firstly need to construct a publically available dataset for OOD task. Several previous studies [9,25,27] did not present EER values, which makes it hard to compare them with other studies. As it is supposed to use the EER value for comparison between different methods, it will be better to adopt the EER value as a measurement.

Table 3. Summary of previous studies

Authors	Features and method	Performance
Ian R. Lane et al. [18]	* Features: n-grams of word base form, surface form, and word+POS * Method: Support Vector Machines (SVM)	EER 19.6~23.3% on text data
Ian R. Lane et al. [19]	* Features: n-grams of word base form, surface form, and word+POS * Method: SVM of 1-versus-all approach	EER 15.3% on text data
Ian R. Lane and Tatsuya Kawahara [20]	* Features - n-grams of word base form, surface form, and word+POS - dialog context information: occurred word vector, topic distribution * Method: Linear Discriminant Model (LDM)	EER 14.9% on text data
Yoko Fujita et al. [22]	* Features - n-grams - OOV frequency - Similarity scores between words * Method: SVM	- EER 11~16% on text data - EER 13~26% on speech data
Mikio Nakano et al. [27]	* Features - dialog context information: previous utterance type, system response, # of words of previous utterance * Method - Logistic regression (LR) for choosing promising domain - SVM for decision making about state transition	F1 0.826 (precision 0.829, recall 0.824) on text data
Deirdre Hogan et al. [25]	* Features: performance estimates of search engine, # of documents and document scores given by search engine * Method: IB1 algorithm	Accuracy 84.8% on text data
Seonghan Ryu etl al. [26]	* Features: n-grams, dictionary-based OOV tags (OOV-LSP) * Method: Maximum Entropy Model (MEM)	EER 4.3% (FAR 6.6, FRR 3.6) on text data
Gokhan Tur et al. [9]	* Features: n-grams, syntactic/semantic parse tree * Method: SVM	Precision 0.18, recall 1.0 on text data
Seonghan Ryu et al. [32]	* Features: word-embeddings, sentence embeddings * Method: Auto-Encoder (AE)	EER 7.02% on text data
Young-Seob Jeong [35]	* Features: Topic distributions * Method: Hierarchical Dirichlet Process (HDP)	EER 6~20% on text data
Joo-Kyung Kim and Young-Bum Kim [34]	* Features: unigram * Method: a neural joint model (Bi-LSTM with embedding)	FAR 5% (with about 90% accuracy of domain classification)

5. Conclusion

Since the dialog system attracts much attention in both of industry and academic areas, it becomes important to develop a system to predict a given utterance is in-domain or out-of-domain. Compared to other research fields, there were relatively smaller studies of OOD detection due to the lack of sufficient datasets. In this paper, we

reviewed the previous studies about the OOD detection in terms of three points: dataset, feature, and method. In terms of the dataset, when we adopt the deleted interpolation, the existing datasets for domain detection becomes available to the study of OOD detection. These datasets have a common limitation that they are not for development of the dialog system, so it is necessary to construct and share a large dataset for the dialog system. The most widely used features were n-grams, and several studies employed other resources or methods to generate additional features. Some studies made use of the result of dialog system (e.g., previous utterance, previous system response), and defined features based on the result. The SVM with n-gram features was the best in earlier studies, and deep learning technique proved its potential recently. We believe that the most important step of future research direction is to construct a public dataset for the dialog system, and this will eventually accelerate the advance of OOD detection techniques, so that it contributes to improvement of the quality of industrial products. Furthermore, it is also important to take a consistent measure (e.g., EER, FAR) for a fair comparison.

REFERENCES

[1] Google. http://www.google.com

[2] Yahoo. http://www.yahoo.com

[3] M. S. Seigel. (2013). *Confidence Estimation for Automatic Speech Recognition Hypotheses*. Doctoral dissertation, St Edmund's College.

[4] J. C. Chappelier, M. Rajman, R. Aragües. & A. Rozenknop. (1999). Lattice Parsing for Speech Recognition. *Traitement Automatique du Langage Naturel*, 95-104.

[5] A. Graves. & N. Jaitly. (2014). Towards End-to-End Speech Recognition with Recurrent Neural Networks. *Proceedings of the 31th International Conference on Machine Learning*, 1764-1772.

[6] H. Khouzaimi, R. Laroche & F. Lefevre. (2014). An Easy Method to Make Dialogue Systems Incremental. *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 98-107.

[7] C. Shi, M. Verhagen & J. Pustejovsky. (2014). A Conceptual Framework of Online Natural Language Processing Pipeline Application. *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, 53-59.

[8] X. Liu., R. Sarikaya., L. Zhao., Y. Ni. & Y. C. Pan. (2016). Personalized Natural Language Understanding. *Proceedings of the 17th Annual Conference of the International Speech Communication Association*, 1146-1150.

[9] D. Wang., D. H. Tur & G. Tur. (2013). Understanding Computer-Directed Utterances in Multi-User Dialog Systems. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 8377-8381.

[10] P. Xu. & R. Sarikaya. (2014). Contextual Domain Classification in Spoken Language Understanding Systems Using Recurrent Neural Network. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 136-140.

[11] C. Lee., S. Jung., S. Kim & G. G. Lee. (2009). Example-Based Dialog Modeling for Practical Multi-Domain Dialog System. *Speech Communication, 51*, 466-484.

[12] R. Meena. (2016). *Data-Driven Methods for Spoken Dialogue Systems*. Doctoral dissertation, KTH Royal Institute of Technology.

[13] B. E. Boser, I. M. Guyon & V. N. Vapnik. (1992). A Training Algorithm For Optimal Margin Classifiers. *Proceedings of the fifth Annual Workshop on Computational Learning Theory*, 144-152.

[14] G. Tur, A. Deoras & D. Hakkani-Tur. (2014). Detecting Out-Of-Domain Utterances Addressed to A Virtual Personal Assistant. *Proceedings of the 15th Annual Conference of the International Speech Communication Association*, 283-287.

[15] E. Shriberg, A. Stolcke, D. Hakkani-Tur & L. Heck. (2012). Learning When to Listen: Detecting System-Addressed Speech in Human-Human-Computer Dialog. *Proceedings of the 13th Annual Conference of the International Speech Communication Association*, 334-337.

[16] A. Stolcke et al. (2000). Dialogue Act Modeling For Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics,*

26(3), 339-373.

[17] M. Core. & J. Allen. (1997). Coding Dialogs With the DAMSL Annotation Scheme. *Proceedings of the Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*.

[18] I. R. Lane., T. Kawahara., T. Matsui. & S. Nakamura. (2004). Out-Of-Domain Detection Based On Confidence Measures From Multiple Topic Classification. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 757-760.

[19] I. R. Lane., T. Kawahara., T. Matsui. & S. Nakamura. (2004). Topic Classification and Verification Modeling For Out-Of-Domain Utterance Detection. *Proceedings of the 8th International Conference on Spoken Language Processing*.

[20] I. R. Lane. & T. Kawahara. (2005). Incorporating Dialogue Context and Topic Clustering in Out-of-Domain Detection. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1045-1048.

[21] X. Zhang., J. Zhao. & Y. Lecun. (2015). Character-Level Convolutional Networks for Text Classification. *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 649-657.

[22] Y. Fujita. S. Takeuchi, H. Kawanami, T. Matsui, H. Saruwatari & K. Shikano. (2011). Out-of-Task Utterance Detection Based on Bag-of-Words Using Automatic Speech Recognition Results. *Proceedings of the third Annual Summit and Conference of Asia-Pacific Signal and Information Processing Association*.

[23] Y. S. Jeong. (2017). Experimental Analysis for Out-Of-Domain Detection Using Features of Word Positions in Sentence. *Proceedings of the Spring Conference of Korean Society for Internet Information*, *18(1).*

[24] Wordnet. https://wordnet.princeton.edu/

[25] D. Hogan, J. Leveling, H. Wang, P. Ferguson & C. Gurrin. (2013). SMS Normalisation, Retrieval and Out-of-Domain Detection Approaches for SMS-Based FAQ Retrieval. *Multilingual Information Access in South Asian Languages*, 184-196.

[26] S. Ryu, D. Lee, G. G. Lee, K. Kim & H. Noh. (2014). Exploiting Out-Of-Vocabulary Words For Out-Of-Domain Detection in Dialog Systems. *Proceedings of the International Conference on Big Data and Smart Computing*, 165-168.

[27] M. Nakano, S. Sato, K. Komatani, K. Matsuyama, K. Funakoshi & H. G. Okuno. (2011). A Two-Stage Domain Selection Framework for Extensible Multi-Domain Spoken Dialogue Systems. *Proceedings of the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 18-29.

[28] Springer. (2003). *The Elements of Statistical Learning*. Berlin: T. Hastie., R. Tibshirani. & J. Friedman.

[29] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer & R. Harshman. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science, 41(6),* 391-407.

[30] D. A. Freedman. (2009). *Statistical Models: Theory and Practice. Cambridge University Press*.

[31] A. L. Berger, S. A. D. Pietra & V. J. D. Pietra. (1996). A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, *22(1)*, 39-71.

[32] S. Ryu, S. Kim, J. Choi, H. Yu & G. G. Lee. (2017). Neural Sentence Embedding Using Only In-Domain Sentences for Out-Of-Domain Sentence Detection in Dialog Systems. *Pattern Recognition Letter, 88,* 26-32.

[33] S. Hochreiter. & J. Schmidhuber. (1997). Long Short-Term Memory. *Neural Computation, 9(8),* 1735-1780.

[34] J. K. Kim. & Y. B. Kim. (2018). Joint Learning of Domain Classification and Out-of-Domain Detection with Dynamic Class Weighting for Satisficing False Acceptance Rates. *Proceedings of 19th Annual Conference of the International Speech Communication Association*, 556-560.

[35] Y. S. Jeong. (2018). Out-Of-Domain Detection Using Hierarchical Dirichlet Process. *Journal of The Korea Society of Computer and Information, 23(1),* 17-24.

정 영 섭(Young-Seob Jeong) [정회원]

· 2016년 2월 : 한국과학기술원 전산학과 (공학박사)
· 2016년 2월 ~ 2016년 12월 : Naver
· 2017년 1월 ~ 현재 : 순천향대학교 빅데이터공학과 교수
· 관심분야 : 인공지능, 자연어처리
· E-Mail : bytecell@sch.ac.kr

김 영 민(Young-Min Kim) [정회원]

· 2009년 8월 : 국민대학교 비즈니스IT 학과(경영학사)
· 2015년 8월 : 연세대학교 산업공학과 (공학박사)
· 2018년 3월 ~ 현재 : 순천향대학교 빅데이터공학과 교수
· 관심분야 : 금융빅데이터분석, 지능형정보시스템
· E-Mail : kimym38@sch.ac.kr