Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 9, S. 78-84, 4, 2025

Araştırma Makalesi

IJANSER
https://as-proceeding.com/index.php/ijanser
ISSN:2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 9, pp. 78-84, 4, 2025

Research Article

Intent Discovery Pipeline using Z-Bert-A

Abdullah Aijaz[1], Edlira Vakaj[2] and Eda Tabaku [3*]

[1] *Faculty of Computing, Engineering and the Built Environment, Birmingham City, abdullah.aijaz@mail.bcu.ac.uk*
[2] *Faculty of Computing, Engineering and the Built Environment, Birmingham City edlira.vakaj@bcu.ac.uk*
[3]*Faculty of Information Technology, Aleksandër Moisiu University Of Durrës edatabaku@uamd.edu.al*

[*]*(edatabaku@uamd.edu.al) Email of the corresponding author*

(Received: 11 April 2025, Accepted: 23 April 2025)

(5th International Conference on Engineering, Natural and Social Sciences ICENSOS 2025, April 15-16, 2025)

ATIF/REFERENCE: Aijaz, A., Vakaj, E. & Tabaku, E. (2025). Intent Discovery Pipeline using Z-Bert-A. *International Journal of Advanced Natural Sciences and Engineering Researches*, 9(4), 78-84.

Abstract – This paper introduces a novel approach to handling unknown intents in dialogue systems by proposing a custom intent discovery pipeline using Z-BERT-A. Developed in Python, this pipeline is specifically designed to address intents that are not predefined within the system. The development of this solution is guided by a comprehensive literature review, which examines existing models and techniques, including rule-based, machine learning, and deep learning approaches. Experimental results on the SNLI and banking datasets demonstrate that Z-BERT-A outperforms other models in managing unknown intents. The proposed pipeline integrates Z-BERT-A and customizes its source code, creating a flexible and generalized solution for intent discovery. Capable of handling unseen intents, this pipeline is crucial for modern dialogue systems used in triage scenarios. Additionally, it is resource-efficient, easily adaptable to various domains, and integrates seamlessly into existing systems. The pipeline also incorporates preprocessing and postprocessing steps to ensure accuracy, efficiency, and scalability. The paper concludes by evaluating the pipeline's performance through multiple metrics, comparing it to other state-of-the-art models.

Keywords – Unknown Intents, Z-Bert-A, Intent Discovery Pipeline, Dialogue Systems, Deep Learning.

I. INTRODUCTION

In natural language processing (NLP) and conversational AI, intent discovery refers to the task of automatically identifying the user's goals or needs when interacting with a system [1]. This process aims to assign meaningful intent labels to a set of unlabeled utterances within a specific domain, without relying on pre-labeled examples. The primary objective of intent discovery is to detect the most common intents in a given domain and generate a set of accurate, relevant intent labels[2]. This task plays a critical role in conversational systems such as chatbots, virtual assistants, and customer service applications, but it also extends to other NLP tasks like text classification and information retrieval [3]. As conversational systems continue to grow in popularity, the importance of intent discovery increases, enabling companies to swiftly develop tailored systems without requiring vast amounts of labeled data or domain-specific expertise [4].

This paper explores the significance of intent discovery in NLP and conversational AI, highlighting its potential applications . The main objective of this paper is to develop a more accurate model for discovering unseen intent categories along with their corresponding labels, and to integrate this model into a fully

functional NLP pipeline using Python. Previous approaches to this challenge have employed unsupervised methods, including Semantic Clustering, Dependency Parsing, TEXTOIR, Z-BERT-A (a Python package), Deep Aligned Clustering, Deep Embedding Clustering, and KMeans Clustering across various datasets, which will be further examined in this study.

II. BACKGROUND

Intent discovery is a critical task in natural language processing (NLP) that aims to automatically identify the user's goal or purpose during an interaction with a computer system [5]. In dialog systems, this process, often referred to as "intent understanding," is essential for the system to generate appropriate responses. The primary objective of intent discovery is to detect the most common intents within a specific domain and generate accurate, relevant intent labels [6]. This task is widely used in conversational systems such as chatbots, virtual assistants, and customer service applications [7]. Moreover, it extends to other NLP tasks like text classification and information retrieval, where the goal is to categorize or extract information from text based on its underlying intent.

With the increasing adoption of conversational systems, such as chatbots and virtual assistants, understanding user intent has become more important [8]. These systems are now commonplace in customer service, offering users quick solutions without human intervention[9]. However, to be effective, these systems must accurately interpret the user's intent to provide relevant responses. Intent discovery plays a central role in this capability [10].

Typically, intent discovery is applied in domain-specific conversational systems, such as those tailored for a particular company's customer service. The aim is to automatically identify common intents—such as "cancel an order" or "track a package"—and generate relevant intent labels [11]. This process eliminates the need for manually labeled examples, which can be time-consuming and require expert knowledge [12].

The process begins with a large set of unlabeled utterances, which are analyzed to detect patterns and themes using techniques such as clustering, NLP, and machine learning algorithms [13]. The result is a set of intent labels that can guide the training of the conversational system.

Intent discovery also has applications beyond conversational systems, including text classification and information retrieval [14]. In text classification, the goal is to categorize documents, such as news or sports articles, while in information retrieval, the goal is to extract relevant information from a large text corpus [15]. In both scenarios, intent discovery can be used to identify the purpose of the text, thereby improving system performance.

Traditionally, intent understanding is treated as a supervised learning problem, requiring labeled examples for training [16]. However, labeling data for new domains can be resource-intensive and necessitate domain expertise. Unsupervised methods, like intent discovery, offer a solution by automatically identifying meaningful intent labels from unlabeled utterances [17]. This method is increasingly important for industrial applications, such as chatbots and virtual assistants, as it enables the rapid creation of conversational systems without relying on large amounts of labeled data or domain-specific knowledge [18].

In summary, intent discovery is a foundational task in NLP and conversational AI that helps determine a user's goal during an interaction. By automatically identifying intent labels from unlabeled data, this task supports the development of more accurate and efficient conversational systems [19]. As these systems become more prevalent across industries like e-commerce, healthcare, and finance, the ability to understand user intent is crucial for providing appropriate and timely responses [20]. This is especially important for learning environments that require continuous access to critical systems such as data servers, online learning platforms, and student resources [21].

III. LITERATURE REVIEW

This section reviews key literature in the areas of deep learning-based unsupervised clustering and intent detection methods, highlighting their contributions to the development of advanced models for clustering and recognizing user intents. Several significant advancements in these areas have been proposed, each with its distinct approach and contributions[22].

[23] introduced a novel method called Deep Embedded Clustering (DEC), which aims to cluster data without relying on labeled information. DEC leverages a deep neural network architecture consisting of an encoder and a decoder network. The encoder maps the input data to a low-dimensional space, creating an embedding that facilitates clustering. The decoder reconstructs the data back to its original space, minimizing the reconstruction error. DEC also integrates a clustering objective that encourages the model to group similar data points together while separating dissimilar ones by minimizing the KL divergence between the low-dimensional representations and their respective cluster assignments. This approach is highly efficient, outperforming traditional clustering methods like k-means, PCA, and Autoencoders, and it does not require data preprocessing or fine-tuning. DEC was shown to perform well on benchmark datasets such as MNIST and Reuters, demonstrating both higher clustering accuracy and computational efficiency compared to other methods [24].

[25] proposed the Simultaneous Deep Learning and Clustering (SDLAC) network, which addresses the limitations of traditional clustering methods like k-means when dealing with high-dimensional data. The SDLAC network transforms the data into a "K-means-friendly" space using a deep neural network, thus improving the separability of clusters. The network comprises an encoder (either a CNN or Autoencoder) and a clustering layer that directly associates the encoder's output with cluster assignments. The model uses a combination of supervised pre-training and unsupervised fine-tuning via a k-means clustering loss function. This method improves clustering performance, particularly when data is non-linearly separable, and it has been shown to outperform the traditional k-means algorithm on various benchmark datasets. SDLAC also demonstrated better robustness to k-means initialization, further enhancing its utility in real-world applications.

[26]proposed a method for discovering user intents from unstructured text without the need for labeled data. Their approach combines unsupervised semantic clustering and dependency parsing to identify underlying patterns in text that correspond to different user intents. The first step involves clustering similar words and phrases from input text, followed by dependency parsing to examine the grammatical relationships between words. This method was tested on customer service conversations, achieving high accuracy and identifying previously unknown intents. Notably, the authors demonstrated that their approach could generalize to unseen data, making it a cost-effective and scalable solution for natural language understanding tasks in various domains.

[27] introduced TEXTOIR, a comprehensive platform for intent recognition that integrates both machine learning and rule-based approaches. The platform includes several modules: a text pre-processing module, an intent recognition module using machine learning models (e.g., SVMs and deep neural networks), and a rule-based system for handling complex or domain-specific intents. TEXTOIR also provides a visual interface for users to interact with and understand intent classification results. In a user study, the platform demonstrated high accuracy in recognizing intents from customer service queries, and participants reported that the interface was easy to use and informative. TEXTOIR thus represents a powerful tool for combining traditional machine learning methods with practical user interaction in intent recognition tasks.

In another study, [28]presented a method for discovering new intents in conversational AI systems using deep aligned clustering. This method utilizes a deep neural network to learn representations of conversational data, which are then clustered using aligned clustering. Aligned clustering minimizes the distance between the learned representations and the true intents of the data. The approach was tested on customer service conversation datasets, where it outperformed traditional clustering methods. By aligning the learned representations with actual intents, this technique enables the discovery of new intents, which is especially useful in evolving conversational AI systems.

[29] introduced Z-BERT-A, a zero-shot approach for detecting unknown intents using the BERT transformer architecture. The Z-BERT-A pipeline includes two stages: first, a fine-tuned BERT model

classifies known intents, and second, an adversarial detector identifies unknown intents. This zero-shot capability allows the model to recognize intents it has not encountered during training. Z-BERT-A outperformed baseline models in recognizing previously unknown intents and organizing known ones. The method's ability to detect unknown intents makes it highly suitable for real-world applications where new intents may emerge over time. The model was evaluated on several datasets, including SNLI and Banking77-OOS, and demonstrated excellent zero-shot performance, further validating its practical applicability.

The reviewed studies provide valuable insights into various methods for clustering and intent recognition. Key developments such as DEC, SDLAC, and deep aligned clustering have contributed significantly to improving clustering techniques by combining deep learning with unsupervised learning. Approaches like Liu et al.'s unsupervised semantic clustering and Zhang et al.'s TEXTOIR platform offer practical solutions for intent detection in real-world applications, especially in natural language understanding tasks. Moreover, Z-BERT-A's zero-shot capability opens new avenues for intent recognition in dynamic environments where new intents may emerge. Collectively, these methods demonstrate the power of combining deep learning with clustering techniques to tackle complex problems in clustering and intent recognition, advancing the state-of-the-art in AI systems [30].

IV. METHODOLOGY

This study aims to develop an intent discovery pipeline capable of handling unknown intents, which are intents not predefined within the system. To achieve this, a comprehensive literature review of existing intent discovery models was conducted, providing insights into the problem and existing approaches. Various techniques, including rule-based, machine learning-based, and deep learning-based methods, were evaluated. Based on these evaluations, Z-BERT-A was selected for the unknown intent discovery task due to its superior performance in experiments conducted on both the SNLI and a banking dataset. Z-BERT-A demonstrated the ability to generate intents that were semantically similar to their ground-truth counterparts.

The next step involved customizing the Z-BERT-A source code to create a generalized pipeline. The pipeline was broken down into modular components—data ingestion, training, and prediction modules—to facilitate reuse across different domains. A data ingestion module was developed to handle training and testing datasets in CSV format. A flexible training module was also created to allow the use of different training algorithms and fine-tuning based on the characteristics of the dataset. The prediction module was designed for efficient processing of new inputs, capable of large-scale predictions and easy integration of new features. This module generates a results.csv file, containing the newly generated intent labels along with the original dialogue or sentence.

To ensure that the intent labels were generalized, singularization techniques were applied to group similar intent labels, and similarity metrics and text embedding techniques were used to cluster context-based intents that were similar or identical. These techniques leveraged the Spacy module, which uses optimized English NLP pipelines, improving the generalization of the model.

For training and testing, a banking dataset was chosen, [31]as it provided a diverse set of data covering different intent types and variations. The dataset, in CSV format, was divided into train.csv and test.csv files, used respectively for model training and performance evaluation on unseen data. The dataset underwent preprocessing steps such as tokenization, data cleaning, and transformation before being used for training. The use of CSV format offers flexibility, as future datasets with similar structures can easily replace the current ones, allowing for the retraining of the model with updated data. This feature makes the pipeline adaptable and suitable for various domains and use cases.

V. RESULTS

For model training, the train.csv and test.csv files were placed in the data directory of the pipeline, where they were used during the training process in conjunction with the SNLI Corpus. The train.csv and test.csv files contained text and corresponding text category columns. To facilitate model training, a train.py file was created and placed in the root directory of the pipeline. This file also incorporated other Python scripts responsible for preprocessing tasks such as tokenization and data cleaning before the model was trained. Additionally, the model's parameters were fine-tuned to optimize its performance and accuracy on the test dataset.

After executing the train.py file, the trained model, named "model.pt," was saved in the train/model directory for future use. This allowed the model to be loaded and utilized for predictions without needing to retrain it each time. The model could also be adapted to different domains or use cases simply by modifying the train.csv and test.csv files.

For intent generation, the prediction module, implemented in a script called predict.py, was used. This script, located in the root directory of the pipeline, processed the text in the test.csv file to run predictions using the trained model. Upon execution, the generated intent labels were saved in a new file named results.tsv, along with the original text inputs. This results.tsv file contained the predicted intent labels for each text input, facilitating the evaluation and analysis of the model's performance.

VI. CONCUSIONS

The proposed pipeline leverages Z-BERT-A, a BERT-based model fine-tuned with Adapters on Natural Language Inference (NLI), to perform zero-shot predictions for known intents. This approach allows the pipeline to handle unseen intents, a critical challenge in intent discovery. The primary strength of this pipeline lies in its ability to make zero-shot predictions, thereby eliminating the need for extensive labeled data for every new intent. However, one limitation identified in Z-BERT-A is the reliance on high-quality dependency parsing for the new intent generation stage, which may affect the overall accuracy.

Having worked extensively with the Z-BERT-A source code, there is significant potential for future research and improvement in the intent generation pipeline. Future contributions could focus on enhancing the accuracy of the intent generation process by exploring advanced techniques in dependency parsing or investigating alternative methods for generating new intents. Furthermore, the data ingestion module currently uses the CSV format, which limits its flexibility. A potential area of improvement involves developing a module capable of handling multiple data formats such as TSV, JSON, and various database types, thus increasing the pipeline's adaptability to different sources and use cases.

VII. FUTURE WORK

Future work on this pipeline can explore several key avenues for improvement. First, integrating additional zero-shot learning techniques could further enhance the model's ability to generalize across unseen domains, improving its performance and applicability in a broader range of contexts. Another promising direction is investigating the application of the pipeline to multilingual datasets, which would expand its usability to diverse linguistic environments and enhance its versatility in global applications. Additionally, addressing the pipeline's scalability to efficiently handle larger datasets would be crucial for ensuring its effectiveness in real-world deployment, particularly in high-volume environments. These improvements, coupled with enhancements in the intent generation and data ingestion modules, would contribute to making the pipeline a more robust and adaptable tool for various conversational AI applications.

ACKNOWLEDGMENT

The heading of the Acknowledgment section and the References section must not be numbered.

REFERENCES

[1] C. B. Chandrakala, R. Bhardwaj, and C. Pujari, "An intent recognition pipeline for conversational AI," *International Journal of Information Technology*, vol. 16, no. 2, pp. 731–743, Feb. 2024, doi: 10.1007/s41870-023-01642-8.

[2] D. Comi, D. Christofidellis, P. F. Piazza, and M. Manica, "Z-BERT-A: a zero-shot Pipeline for Unknown Intent detection," Aug. 2022.

[3] D. Comi, D. Christofidellis, P. Piazza, and M. Manica, "Zero-Shot-BERT-Adapters: a Zero-Shot Pipeline for Unknown Intent Detection," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 650–663. doi: 10.18653/v1/2023.findings-emnlp.47.

[4] A. S. Adekotujo, T. Enikuomehin, B. Aribisala, M. Mazzara, and A. F. Zubair, "Computational treatment of natural language text for intent detection," *Computer Research and Modeling*, vol. 16, no. 7, pp. 1539–1554, Dec. 2024, doi: 10.20537/2076-7633-2024-16-7-1539-1554.

[5] H. Zhang *et al.*, "Generic Intent Representation in Web Search," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: ACM, Jul. 2019, pp. 65–74. doi: 10.1145/3331184.3331198.

[6] J. Atuhurra, H. Kamigaito, T. Watanabe, and E. Nichols, "Domain Adaptation in Intent Classification Systems: A Review," Mar. 2024.

[7] R. V. V. Vishruth and S. G. Mohan, "Online video conference analytics: A systematic review," in *Applied Data Science and Smart Systems*, London: CRC Press, 2024, pp. 435–447. doi: 10.1201/9781003471059-57.

[8] J. Singh, S. Goyal, R. Kumar Kaushal, N. Kumar, and S. Singh Sehra, *Applied Data Science and Smart Systems*. London: CRC Press, 2024. doi: 10.1201/9781003471059.

[9] E. Tabaku, E. Duçi, and A. Lazaj, "From Physical Stores to Virtual Marketplaces: The Evolution of Shopping," *Interdisciplinary Journal of Research and Development*, vol. 11, no. 3, p. 175, Dec. 2024, doi: 10.56345/ijrdv11n324.

[10] R. Mohammad, O. S. Alkhnbashi, and M. Hammoudeh, "Optimizing Large Language Models for Arabic Healthcare Communication: A Focus on Patient-Centered NLP Applications," *Big Data and Cognitive Computing*, vol. 8, no. 11, p. 157, Nov. 2024, doi: 10.3390/bdcc8110157.

[11] A. Hrytsyna and R. Alves, "From Representation to Response: Assessing the Alignment of Large Language Models with Human Judgment Patterns," *ACM Trans Intell Syst Technol*, Dec. 2024, doi: 10.1145/3709148.

[12] H. Hettiarachchi *et al.*, "CODE-ACCORD: A Corpus of building regulatory data for rule generation towards automatic compliance checking," *Sci Data*, vol. 12, no. 1, p. 170, Jan. 2025, doi: 10.1038/s41597-024-04320-x.

[13] R. Mohammad, O. Favell, S. Shah, E. Cooper, and E. Vakaj, "Utilisation of open intent recognition models for customer support intent detection," Jul. 2023.

[14] D. Al-Turki *et al.*, "Human-in-the-Loop Learning with LLMs for Efficient RASE Tagging in Building Compliance Regulations," *IEEE Access*, pp. 1–1, 2024, doi: 10.1109/ACCESS.2024.3512434.

[15] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Mar. 2016, pp. 31–35. doi: 10.1109/ICASSP.2016.7471631.

[16] K. Berahmand, S. Bahadori, M. N. Abadeh, Y. Li, and Y. Xu, "SDAC-DA: Semi-Supervised Deep Attributed Clustering Using Dual Autoencoder," *IEEE Trans Knowl Data Eng*, vol. 36, no. 11, pp. 6989–7002, Nov. 2024, doi: 10.1109/TKDE.2024.3389049.

[17] P. Liu, Y. Ning, K. K. Wu, K. Li, and H. Meng, "Open Intent Discovery through Unsupervised Semantic Clustering and Dependency Parsing," Apr. 2021.

[18] L. W. Y. Yang *et al.*, "Development and testing of a multi-lingual Natural Language Processing-based deep learning system in 10 languages for COVID-19 pandemic crisis: A multi-center study," *Front Public Health*, vol. 11, 2023, doi: 10.3389/fpubh.2023.1063466.

[19] A. B. Saka, L. O. Oyedele, L. A. Akanbi, S. A. Ganiyu, D. W. M. Chan, and S. A. Bello, "Conversational artificial intelligence in the AEC industry: A review of present status, challenges and opportunities," 2023. doi: 10.1016/j.aei.2022.101869.

[20] E. Tabaku, E. Vyshka, R. Kapçiu, A. Shehi, and E. Smajli, "UTILIZING ARTIFICIAL INTELLIGENCE IN ENERGY MANAGEMENT SYSTEMS TO IMPROVE CARBON EMISSION REDUCTION AND SUSTAINABILITY," *Jurnal Ilmiah Ilmu Terapan Universitas Jambi*, vol. 9, no. 1, pp. 393–405, Feb. 2025, doi: 10.22437/jiituj.v9i1.38665.

[21] E. Tabaku and E. Duçi, "Optimizing High Availability in Educational Systems Using Xen Paravirtualization," *Journal of Educational and Social Research*, vol. 15, no. 2, p. 205, Mar. 2025, doi: 10.36941/jesr-2025-0054.

[22] S. C. Ivan, R. Ş. Győrödi, and C. A. Győrödi, "Sentiment Analysis Using Amazon Web Services and Microsoft Azure," *Big Data and Cognitive Computing*, vol. 8, no. 12, p. 166, Nov. 2024, doi: 10.3390/bdcc8120166.

[23] Y. Ren, K. Hu, X. Dai, L. Pan, S. C. H. Hoi, and Z. Xu, "Semi-supervised deep embedded clustering," *Neurocomputing*, vol. 325, pp. 121–130, Jan. 2019, doi: 10.1016/j.neucom.2018.10.016.

[24] E. Tabaku, "Improving High Availability Services Using KVM Full Virtualization," *European Journal of Computer Science and Information Technology*, vol. 13, no. 1, pp. 1–15, Jan. 2025, doi: 10.37745/ejcsit.2013/vol13n1115.

[25] A. Y. Yousif and B. Al Sarray, "Convex Optimization Techniques for High-Dimensional Data Clustering Analysis: A Review," *Iraqi Journal for Computer Science and Mathematics*, vol. 5, no. 3, Jan. 2024, doi: 10.52866/ijcsm.2024.05.03.022.

[26] P. Liu, Y. Ning, K. K. Wu, K. Li, and H. Meng, "Open Intent Discovery through Unsupervised Semantic Clustering and Dependency Parsing," Apr. 2021.

[27] S. Moradizeyveh, "Intent Recognition in Conversational Recommender Systems," Dec. 2022.

[28] A. Chatterjee and S. Sengupta, "Intent Mining from past conversations for conversational agent," May 2020.

[29] D. Comi, D. Christofidellis, P. F. Piazza, and M. Manica, "Z-BERT-A: a zero-shot Pipeline for Unknown Intent detection," Aug. 2022.

[30] E. Tabaku, "The Evolution of Technology in Accounting and Corporate Finance: Implications for Business Adaptation and Competitiveness," *International Research Journal of Modernization in Engineering Technology and Science*, Jan. 2025, doi: 10.56726/IRJMETS66318.

[31] S. Kumar, S. Deep, and P. Kalra, "Enhancing Customer Service in Banking with AI: Intent Classification Using Distilbert," *International Journal of Current Science Research and Review*, vol. 07, no. 05, May 2024, doi: 10.47191/ijcsrr/V7-i5-32.