Multilingual and Cross-Lingual Intent Detection from Spoken Data using Whisper on the MINDS-14 Dataset

Dan Emmanuel Marie Krecoum, Therry Jeannick Anguezome Ondo.
Laval University, {NI:111 250 180, NI: 536 973 933}

***Abstract*—This study presents an innovative approach to multilingual and cross-lingual intent detection from spoken data, leveraging the cutting-edge capabilities of OpenAI's Whisper model on the MINDS-14 dataset. Our integrated pipeline architecture combines automatic speech recognition (ASR), translation, and intent classification models to process and understand raw speech data across multiple languages. The Whisper model, renowned for its proficiency in diverse linguistic and acoustic environments, serves as the backbone for ASR. Subsequent translation to English is performed using MarianMT, ensuring consistency in language input for intent classification. The final stage employs JointBERT for precise intent detection. The system's performance is rigorously evaluated using standard metrics such as Word Error Rate (WER), BLEU, and ROUGE, alongside intent classification accuracy. Our methodology demonstrates robustness and high accuracy, making significant strides towards more intuitive human-computer interaction in multilingual settings.**

I. Introduction

In an increasingly interconnected world, the ability to seamlessly process spoken data across languages is paramount. Detecting user intent from speech presents unique challenges, especially when capabilities need to extend beyond a single language. This paper introduces a sophisticated solution that integrates state-of-the-art models to handle this multistage process with finesse. Emphasizing practicality, our approach utilizes audio data to classify each speaker's intent in a manner that is both simple and clear. While initially employed in the domain of e-banking, the scope of our solution is much broader. Should we succeed in fully implementing our system, it could be utilized across various fields such as supporting the training of new employees, creating chatbots, or integrating into virtual assistants. The success of this solution could thus cater to numerous sectors, aspiring to become a pivotal tool in conversational AI.

II. Related work

In the evolving landscape of Multilingual and Cross-Lingual Intent Detection from Spoken Data, our project takes a departure from traditional methodologies by forging a distinctive pipeline that integrates state-of-the-art models. While transformer-based models like BERT and mBERT have set benchmarks in NLP tasks, including intent detection [1], their direct application to audio data necessitates an intermediate Automatic Speech Recognition (ASR) step, potentially introducing transcription errors. Our initial exploration involved Wav2Vec 2.0, renowned for its self-supervised audio processing capabilities, applied directly to raw audio [4]. However, the outcomes fell short of expectations, prompting a strategic shift in our approach.

Recognizing the nuanced challenges posed by intent detection from spoken data, we opted for Whisper for ASR, capitalizing on its accuracy in transcribing spoken content. This decision not only addressed the limitations encountered with direct Wav2Vec 2.0 application but also set the stage for a more refined and effective pipeline. The innovation lies in the subsequent integration of MarianMT for translation to English and a BERT-based approach for intent classification. This multistep pipeline not only mitigates transcription inaccuracies but also showcases the adaptability of combining models to achieve a comprehensive solution. Our project's journey highlights the importance of dynamic model selection to suit the intricacies of specific tasks, ultimately contributing to the success of Multilingual and Cross-Lingual Intent Detection on the MINDS-14 dataset.

III. Description of the models we propose

Our solution employs an integrated pipeline architecture comprising a sequence of specialized models for comprehensive speech understanding. The initial stage utilizes Automatic Speech Recognition (ASR) technology to convert raw speech data into its textual transcript. After transcription, the text undergoes a translation process where the content is rendered into English, ensuring uniformity. The final component of the pipeline applies an intent classification model, which interprets the translated text to ascertain the underlying intent.

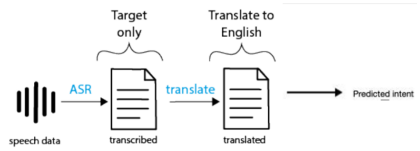

Fig. 1. Architecture of the proposed solution

A. ASR: Whisper

The Whisper model is an advanced automatic speech recognition (ASR) system designed to provide accurate transcriptions across diverse languages and audio conditions [3].

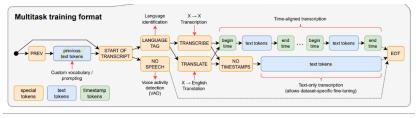

Fig. 2. Architecture of Whisper

- **Foundation:** Whisper is built upon a deep learning architecture that leverages transformer-based models, renowned for their efficacy in sequence modeling tasks.
- **Audio Processing:** The model ingests raw audio signals and processes them through a series of convolutional layers to extract meaningful acoustic features.
- **Attention Mechanisms:** It employs self-attention mechanisms that enable the model to focus on relevant parts of the audio signal for better context understanding.
- **Language Modeling:** Whisper integrates a language model to improve transcription accuracy, which is especially beneficial in handling homophones and contextually ambiguous utterances.

B. Translation: MarianMT

MarianMT is an advanced neural machine translation framework. It is known for its efficiency, making it a quality choice for translation tasks [6].

- **Foundation:** Based on the Transformer model architecture, which utilizes self-attention mechanisms for handling sequence-to-sequence tasks.
- **Efficiency:** Written in C++, it is optimized for speed and memory efficiency.
- **Training:** Employs techniques like teacher-student training, where a larger, more complex model (teacher) helps train a smaller, more efficient model (student).
- **Customization:** MarianMT supports customization and fine-tuning, allowing for tailored models for specific language pairs or domains.

C. Intent classification tasks: JointBERT

JointBERT is a model architecture designed for Natural Language Understanding (NLU) tasks in conversational AI systems. It leverages the pre-trained BERT model and combines it with a joint learning framework for simultaneous intent classification [2].

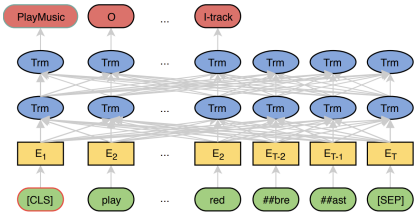

Fig. 3. Architecture of the proposed model for intent classification.

- **Base Model:** JointBERT is based on the BERT model, which uses the Transformer architecture for understanding contextual relationships in text.
- **Input Representation:** It takes tokenized input sequences and processes them through BERT's layers to generate contextual embeddings.
- **Training:** JointBERT is fine-tuned on a labeled dataset for intent classification, optimizing the model for the task.

JointBERT is particularly useful in conversational AI, such as voice-controlled assistants and chatbots, which is ideal for understanding the user intent crucial when extracting relevant information.

IV. Methodology, Experimentation and Discussions

A. Methodology

1) dataset description: The MINDS-14 dataset [5] is a multilingual intent detection dataset for spoken data. It consists of audio recordings in 12 different languages, including English (British, US, Australian), French (FR), Italian (IT), Spanish (ES), Portuguese (PT), German (DE), Dutch (NL), Russian (RU), Polish (PL), Czech (CS), Korean (KO), and Chinese (ZH).

The MINDS-14 dataset contains customer queries or messages related to banking and financial services. It consists of 14 distinct intent classifications, which are as follows: abroad, address, app error, atm limit, balance, business loan, card issues, cash deposit, direct debit, freeze, high-value payment, joint account, latest transactions, and pay a bill. These intent classifications help categorize and understand the nature of customer interactions with the financial institution.

The dataset comprises over 8,000 entries, and it has been divided into two main segments: 70% for training and 30% for validation and testing. The objective behind this data preparation is to develop a model that can be applied in various scenarios, such as employee training, chatbot development, or the enhancement of virtual assistants.

2) Evaluation metric:

- **ASR:** ASR systems are evaluated based on various metrics, one of which is the **Word Error Rate (WER)**. The WER measures the difference between the transcribed words and the reference (ground truth) words, considering substitutions, insertions, and deletions.

$$WER = \frac{S + D + I}{N} \tag{1}$$

Where:
- S represents the number of substitutions (words that are replaced).
- D represents the number of deletions (words that are missed in the transcription). - I represents the number of insertions (extra words introduced in the transcription). - N represents the total number of words in the reference (ground truth) transcription.
The lower the WER, the better the ASR system's accuracy in transcribing spoken language into text.

- **Translation:**

BLEU (Bilingual Evaluation Understudy)

BLEU is a metric commonly used for evaluating the quality of machine-generated translations. It measures the similarity between a reference translation and a candidate translation by comparing the n-grams (contiguous sequences of n words) in both.
The BLEU score is calculated as follows:

$$\text{BLEU} = \text{BP} \cdot \exp(\sum_{n=1}^{N} w_n \cdot \log(p_n))$$

Where: - N is the maximum n-gram order considered. - p_n is the precision of n-grams in the candidate translation. - w_n is the weight assigned to each n-gram precision. - BP is the brevity penalty to account for overly short translations.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE is a metric used for evaluating the quality of text summaries or document translation systems. The ROUGE score can be computed in various forms, but the common formula for ROUGE-N is:

$$\text{ROUGE-N} = \frac{\text{Overlap of n-grams}}{\text{Total n-grams in reference summaries}}$$

Where: - N is the maximum n-gram order considered. ROUGE scores are typically reported for various values of N to provide a comprehensive evaluation of the generated summaries and translated texts.

- **Intent classification:** Intent classification models are commonly evaluated using several metrics, including confusion matrix, precision, recall, accuracy, and F1 score.

  A confusion matrix is a table that summarizes the performance of an intent classification model. It consists of four values: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN):

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positives (TP)	False Negatives (FN)
Actual Negative	False Positives (FP)	True Negatives (TN)

  True Positives (TP): The number of instances where the model correctly predicted the positive intent label.
  False Positives (FP): The number of instances where the model incorrectly predicted the positive intent label when it was a negative intent.
  True Negatives (TN): The number of instances where the model correctly predicted a negative intent label.
  False Negatives (FN): The number of instances where the model incorrectly predicted a negative intent label when it was a positive intent.

  Precision and recall are fundamental measures of the accuracy and completeness of an intent classification model. Precision is defined as the proportion of true positive intent among all the predicted positive intent, while recall is defined as the proportion of true positive intent among all the ground truth positive instances.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}} \quad (2)$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}} \quad (3)$$

  F1 score is another commonly used metric for Intent classification, which is the harmonic mean of precision and recall. This measure is often used as a single metric to summarize Intent classification performance.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

  Accuracy measures the overall correctness of the intent classification model and is calculated as the ratio of correctly predicted intent labels to the total number of intent labels.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}} \quad (5)$$

B. Experimentations

1) ASR: We conducted audio normalization to 16kHz as part of our data preprocessing. This approach yielded a global WER of 0.579 when applied to all languages collectively.

However, for languages that use the Latin alphabet (such as English, Spanish, and French), this preprocessing method performed significantly better, achieving a WER of less than 10%.

Discussion: It is important to note that we used the model "small" which is one of the smallest models of the Whisper model family. This outcome establishes an essential baseline for future enhancements because it highlights the different paths we could explore. On one hand, we could consider using a more powerful variant within the Whisper model family or another ASR architecture to potentially reduce WER. On the other hand, we could decide to implement data augmentation techniques to address errors and enhance recognition when dealing with non-Latin alphabet scenarios.

2) Translation: Our translation model exhibits robust performance with a BLEU score of 60.71, indicating translations that are lexically coherent with the reference. The ROUGE metrics corroborate this finding, with F1-measures of 78.53% for ROUGE-1, 65.61% for ROUGE-2, and 77.47% for ROUGE-L, demonstrating the model's effectiveness in capturing both content and structural integrity.

Discussion: While our translation model demonstrates high scores, indicating quantitative success, a deeper qualitative analysis is essential. We observed that translations, while accurate, sometimes lacked the idiomatic or cultural nuances present in the original languages. This highlights the need for incorporating contextual understanding in future model iterations, ensuring that translations are not just lexically, but also contextually accurate, catering to the subtleties of human language.

3) Intent Classification: We evaluated the model's identified intents compared to the ground truth intents. Additionally, we assessed the accuracy of each intent individually to determine which languages performed the best.

These are the results we obtained:

- Test Loss: 0.1009
- Accuracy: 0.9755

TABLE I
DETAILED INTENT CLASSIFICATION METRICS

Intent	Accuracy	Precision	Recall	F1-Score
abroad	0.9612	0.9340	0.9612	0.9474
address	1.0000	1.0000	1.0000	1.0000
app error	0.9722	0.9859	0.9722	0.9790
atm limit	0.9881	0.9881	0.9881	0.9881
balance	0.9545	1.0000	0.9545	0.9767
business loan	1.0000	0.9651	1.0000	0.9822
card issues	0.9545	0.9722	0.9545	0.9633
cash deposit	0.9903	0.9808	0.9903	0.9855
direct debit	0.9474	0.9600	0.9474	0.9536
freeze	0.9706	0.9706	0.9706	0.9706
high value payment	0.9737	0.9867	0.9737	0.9801
joint account	0.9900	0.9900	0.9900	0.9900
latest transactions	1.0000	0.9659	1.0000	0.9827
pay bill	0.9535	0.9647	0.9535	0.9591

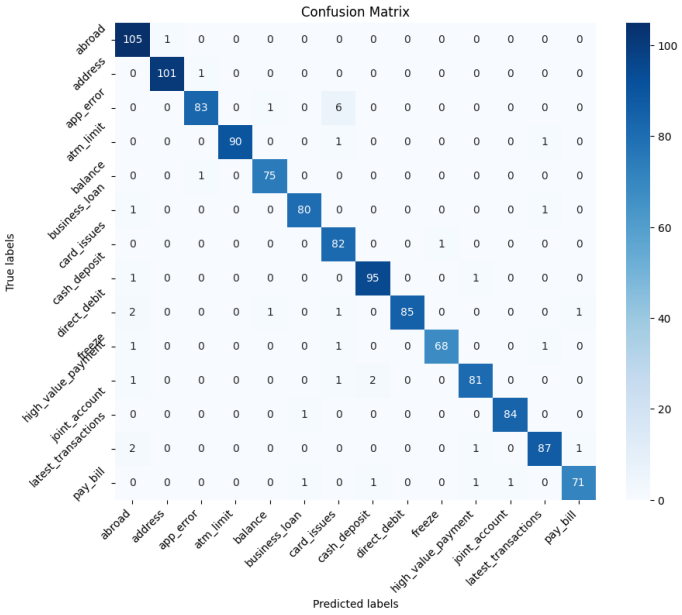

Fig. 4. Confusion matrix of the classified intents.

Discussion:

The results indicate a high-performing Intent Classification model, with an overall Test Loss of 0.1009 and Accuracy of 0.9755, demonstrating its effectiveness when giving predictions.

Particularly impressive are the perfect scores (1.0000) in Accuracy, Precision, Recall, and F1-Score for intents like 'address', 'business loan', 'cash deposit', 'joint account', and 'latest transactions'. These indicate exceptional handling of these intents, crucial for real-world applications.

However, certain intents like 'abroad', 'balance', and 'direct debit' show marginally lower scores, suggesting potential areas for refinement to enhance precision and overall performance.

Language-specific accuracy varies, with perfect scores in Language IDs 0 and 5, but lower accuracy in IDs 6, 7, and 8, pointing to possible challenges with linguistic variations in these languages.

The high accuracy achieved in intent classification, particularly in multilingual contexts, opens up new avenues for future research. One promising direction is the exploration of domain-specific adaptations, especially in sectors like healthcare or customer service, where understanding nuanced user intent is crucial.

TABLE II
INTENT ACCURACY PER LANGUAGE ID

Czech:	1.0000	**Italian:**	0.9619
German:	0.9783	**Korean:**	0.9551
English (Australia):	0.9694	**Dutch (Netherlands):**	0.9796
English (United Kingdom):	0.9888	**Polish:**	0.9643
English (United States):	0.9762	**Portuguese:**	0.9670
Spanish:	1.0000	**Russian:**	0.9753
French:	0.9630	**Chinese:**	0.9867

V. CONCLUSION

Our research presents a robust solution to multilingual and cross-lingual intent detection from spoken data, leveraging the synergy of Whisper, MarianMT, and JointBERT. Our study demonstrates not only the feasibility but also the efficiency of this integrated approach, yielding high accuracy across multiple languages. The significance of this work lies in its potential to enhance human-computer interaction in a multilingual world, breaking down language barriers and paving the way for more intuitive and inclusive conversational AI systems. Future research should aim to extend these findings into more domain-specific applications, further improving accuracy and exploring new frontiers in conversational AI.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018. [Online]. Available: https://arxiv.org/pdf/1810.04805.pdf

[2] Q. Chen, Z. Zhuo, and W. Wang, "BERT for Joint Intent Classification and Slot Filling," 2019. [Online]. Available: https://arxiv.org/pdf/1902.10909.pdf

[3] A. Radford, J. W. Kim, C. Hallacy, et al., "Robust Speech Recognition via Large-Scale Weak Supervision," 2022. [Online]. Available: https://arxiv.org/pdf/2212.04356.pdf

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," 2020. [Online]. Available: https://arxiv.org/pdf/2006.11477.pdf

[5] A. Fan, I. Turc, A. Bordes, and J. Weston, "Augmenting Non-Collaborative Dialog Systems with Explicit Semantic and Strategic Dialog History," 2021. [Online]. Available: https://arxiv.org/pdf/2104.08524.pdf

[6] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, et al., "Marian: Fast Neural Machine Translation in C++," 2018. [Online]. Available: https://arxiv.org/pdf/1804.00344.pdf