A Pointer Network-based Approach for Joint Extraction and Detection of Multi-Label Multi-Class Intents

[1]Ankan Mullick [1]Sombit Bose [1]Abhilash Nandy
[2]Gajula Sai Chaitanya and [1]Pawan Goyal
{ankanm, sbcs.sombit.24, nandyabhilash}@kgpian.iitkgp.ac.in
gsaichai@qti.qualcomm.com pawang@cse.iitkgp.ac.in
[1]Computer Science and Engineering Department, IIT Kharagpur, India [2]Qualcomm, India

Abstract

In task-oriented dialogue systems, intent detection is crucial for interpreting user queries and providing appropriate responses. Existing research primarily addresses simple queries with a single intent, lacking effective systems for handling complex queries with multiple intents and extracting different intent spans. Additionally, there is a notable absence of multilingual, multi-intent datasets. This study addresses three critical tasks: extracting multiple intent spans from queries, detecting multiple intents, and developing a multilingual multi-label intent dataset. We introduce a novel multi-label multi-class intent detection dataset (**MLMCID-dataset**) curated from existing benchmark datasets. We also propose a pointer network-based architecture (**MLMCID**) to extract intent spans and detect multiple intents with coarse and fine-grained labels in the form of sextuplets. Comprehensive analysis demonstrates the superiority of our pointer network based system over baseline approaches in terms of accuracy and F1-score across various datasets.

1 Introduction

Task-oriented dialogue systems have become a major field of study in recent years, significantly advancing the capabilities of Natural Language Understanding (NLU). These systems execute command-based tasks, demonstrating versatility in handling diverse user queries through a set of predefined skills, known as intents. Users interact with dialogue systems to fulfill their needs, and intent detection plays a pivotal role in comprehending user queries and generating appropriate responses in task-oriented conversations, thereby maintaining user engagement. The task of intent detection involves identifying the *intent(s)* within a given statement or query, which represents the underlying meaning conveyed by the user. For example, the query “How is the weather today?" would be associated with the $GetWeather$ intent. Dialogue systems rely on detecting these intents to understand user queries and provide suitable answers.

However, in real-world conversation, a query or a statement often contain multiple different intents. For instance, as shown in Fig. 1, for the query (from Facebook English dataset): “remind me to pick up contact lenses tomorrow, set the alarm for 5 mins and 30 seconds", contains two distinct intent categories with following spans: ‘remind me to pick up contact lenses tomorrow’ (‘set reminder’ intent) and ‘set the alarm for 5 mins and 30 seconds’ (‘set alarm’ intent). Both of these are fine intent categories. Multiple similar fine intents can be merged to create one coarse intent as explained in Table 1. Thus, the above query contains ‘reminder_service’ and ‘change_alarm_content’ coarse intents as shown in Fig. 1. In case of multiple intents in a sentence, one intent which is dominant and most important in that sentence can be termed as ‘Primary’ intent while the other intents can be considered ‘Non-Primary’. For example, in the query (From Mix-SNIPS dataset) “How is the weather today? It would be lovely to go for a movie" is a combination of two simple sentences ‘How is the weather today?’ and ‘It would be lovely to go for a movie’, whose intents are $GetWeather$ and $BookMovieTicket$ respectively. Out of the two possible intents, $BookMovieTicket$ is primary (primary and main focus of the sentence) and $GetWeather$ becomes non-primary. It would require an intent span extraction algorithm to extract multiple intent spans and a multi-label, multi-class classifier to detect different fine and coarse intents.

Over the past few years, researchers concentrate on intent identification across different domains. Flexible and adaptive intent class detection models have been developed for dynamic and evolving real-world applications. (Liao et al., 2023; Kuzborskij

Findings of the Association for Computational Linguistics: EACL 2024, pages 15664–15680
November 12-16, 2024 ©2024 Association for Computational Linguistics

Input Sentence	coarse_intent 1	coarse_intent 2	fine_Intent 1	fine_Intent 2
I tried withdrawing money in another country and the exchange rate was wrong. What should I do if my card is stolen? (BANKING)	exchange_rate_query	Card Problem	wrong_exchange_rate_for_cash_withdrawal	lost_or_stolen_card
remind me to pick up contact lenses tomorrow, set the alarm for 5 mins and 30 seconds (FACEBOOK)	reminder_service	change_alarm_content	set reminder	set alarm
Show me walking directions to MOMA and book a cab (SNIPS)	Location_Service	App_Service	GetDirections	RequestRide

Figure 1: Examples of multi-label multi intent datasets (SNIPS, Facebook and BANKING)

et al., 2013; Scheirer et al., 2012; Degirmenci and Karal, 2022) focus on streaming data to identify evolving new classes using incremental learning. SENNE Cai et al. (2019), IFSTC (Xia et al., 2021), SENC-MaS (Mu et al., 2017b), SENCForest (Mu et al., 2017a), ECSMiner (Masud et al., 2010) aim at SENC (streaming emerging new class) problem on intents on streams. (Sun et al., 2016) work on emergence and disappearance of intents. (Wang et al., 2020) uses high dimensional data for streaming classification. (Mullick et al., 2022d) identifies multiple novel intents using a clustering framework. (Na et al., 2018; Zhan et al., 2021; Larson et al., 2019; Yan et al., 2020; Zhou et al., 2022; Firdaus et al., 2023) detect new intents in the form of outlier detection. Unlike the previous single-intent detection models, which can easily utilize the utterance's sole intent to guide slot prediction, multi-intent SLU (Spoken Language Understanding) encounters the challenge of multiple intents, presenting a unique and worthwhile area of research. (Mullick et al., 2023, 2022b; Mullick, 2023b,a; Mullick et al., 2022a) explore intent detection in different directions. AGIF (Qin et al., 2020), GL-GIN (Qin et al., 2021), (Gangadharaiah, 2019), (Song et al., 2022) work on multiple intent identification problem but these approaches do not detect the sentence spans related to different intents and also do not distinguish the primary and non-primary intents. Based on Convert (Henderson et al., 2019) backed framework, (Coope et al., 2020) extract spans for different slots but does not extract and identify multiple intents. (Mullick et al., 2024; Guha et al., 2021; Mullick et al., 2022c) focus on entity extraction in different forms. Previous research also includes both pipeline-based approaches (Jiang et al., 2023) and end-to-end methods (Ma et al., 2021; Cui et al., 2019; Ma et al., 2022). However, our work is different from the fact that we identify multiple intent spans along with their corresponding fine and coarse labels.

Our work differs from the fact that, we extract multiple intent spans from a given sentence and detect its coarse and fine intent labels. In this paper, we seek to address the following research questions in the field of multi-label multi-class intent detection with span extraction:

1. We introduce a novel multi-label multi-class intent detection dataset (MLMCID-dataset) utilizing a diverse set of existing datasets with various intent sizes in multilingual settings (English and non-English languages), including coarse and fine-grained intent labeling along with primary and non-primary intent marking.

2. We thereafter, build a pointer network based encoder-decoder framework to extract multiple intent spans from the given query.

3. We propose a feed-forward network based intent detection module (MLMCID - **M**ulti-**L**abel **M**ulti-**C**lass **I**ntent **D**etection) to automatically detect multiple primary and non-primary intents for coarse and fine categories in a sextuplet form. We evaluate the performance of MLMCID for full and few shot-settings across several MLMCID datasets.

4. We experiment with different LLMs (Llama2, GPT) to assess their efficacy, comparing them with our approach, and providing a detailed qualitative analysis along with a specialized loss function for multi-label multi-class intent detection.

Empirical findings on various MLMCID datasets demonstrate that our pointer network based RoBERTa model surpasses other baselines methods including LLMs, achieving a higher accuracy with an improvement in macro-F1.

2 Dataset

We conduct different experiments to evaluate our framework on various datasets - all of which are benchmark datasets in NLU domain. We consider three different sizes of the datasets (as per intent class count - mentioned within bracket) -

(i) *Small*: a) SNIPS (10 intents) (Coucke et al., 2018), b) ATIS (21 intents) (Tur et al., 2010), c) Facebook Multi-lingual (12 intents) (Schuster et al., 2018) (consisting of the comparable corpus of English, Spanish and Thai data), abbreviated as Fb.

(ii) *Medium*: a) HWU (64 intents) (Liu et al., 2019a), b) BANKING (77 intents) (Casanueva et al., 2020).

(iii) *Large*: a) CLINC (150 intents) (Larson et al., 2019).

Intents of similar domains which convey a similar broader meaning and are manually grouped together to make coarse-grained labels from original fine-grained labels [1]. Table 1 shows an example of Facebook-English (Fb-en) combining multiple fine intents (like - 'cancel reminder', 'set reminder', 'show reminders') which are closely similar and convey similar broader meaning of 'reminder_service' so these are grouped together to form one single broad coarse grained intent label - 'reminder_service' and an example of SNIPS combining multiple fine intents (like - 'GetTrafficInformation', 'ShareETA') are merged into one single course intent class ('Traffic_update'). Finally, we end up with course intent class of 4 for SNIPS, 5 for Facebook, 18 for HWU, 12 for Banking and 120 for CLINC[2]. Due to space shortage, the details are in Appendix Table 12 and 13.

Fine Intents Combined	Coarse Intent
cancel reminder, set reminder, show reminders	reminder_service
GetTrafficInformation, ShareETA	Traffic_update

Table 1: Fine-Course Intent for Fb-en and SNIPS

All the above datasets are of single intent. In order to validate the broad applicability of the model, we follow the MixAtis and MixSNIPS data-generation guidelines (Qin et al., 2020) to prepare multi-intent datasets for Fb, HWU, BANKING and CLINC. We also use MixATIS and MixSNIPS datasets (Qin et al., 2020). All datasets are in English except for Facebook - which contains Spanish and Thai also along with English. Three annotators are selected after several discussions and conditions of fulfilling criteria like annotators should have domain knowledge expertise along with a good working proficiency in English. Each formed sentence instance is manually checked for correctness, coherence, grammatically meaningful and filter out many sentences which do not qualify. Annotators mark Multiple intents and their respective spans within the specified sentence. Annotators also point out which intent is *Primary*[3] and which one is *non-Primary*. If *Primary* and *non-Primary* intents can not be distinguished then both of the intents are considered as *Primary*.

Dataset	Train	Dev	Test
Mix-SNIPS	11000	2197	2198
Mix-ATIS	13161	600	829
FB-EN	800	100	100
FB-ES	800	100	100
FB-TH	800	100	100
HWU64	780	97	97
BANKING	1156	144	144
CLINC	1353	169	169
Yahoo	498	62	162
MPQA	284	36	136

Table 2: **MLMCID-dataset** statistics

To show the real world applicability of our framework, we also experiment on two different practical datasets: a) MPQA[4] (Multi Perspective Question Answering) (Mullick et al., 2016, 2017), b) Yahoo News article (Mullick et al., 2016, 2017). Intent can be broadly categorised as opinionated or factual. Each sentence from MPQA and Yahoo news articles is marked as opinion and fact. Further, opinions can be of four different subcategory (Asher et al., 2009) - 'Report', 'Judgment', 'Advise' and 'Sentiment' and facts can be subcategorised into five types (Soni et al., 2014) - 'Report', 'Knowledge', 'Belief', 'Doubt' and 'Perception'. So coarse intent can be sub-categorized in four opinionated fine-intents and five factual fine-intents. In MPQA and Yahoo news article, annotators are told to identify different clauses of compound and complex sentences and mark the fine label intent categories for opinion and fact. In all the annotation tasks - initial labeling is done by two annotators and any annotation discrepancy is checked and resolved by the third annotator after discussing with others. Overall inter-annotator agreement is 0.89 which is considered good as per (Landis and Koch, 1977). The detail statistics of train-dev-test divisions of different dataset intent dataset are shown in Table 2. We term this dataset as **MLMCID-dataset.**

We use the Facebook data from **MLMCID-dataset** comprising 1000 text instances and corresponding intent labels are annotated for its 3 vari-

[1]Course intent is a combination of multiple similar meaning or closely matching finer intents of higher hierarchy. One coarse-grained intent is a cluster of multiple closely matching fine-grained labels.

[2]For ATIS we keep fine intents as it is, without coarse intents due to high dis-similarity among intents

[3]Between two intents, we define one as primary which is more important than others and main focus of the sentence

[4]https://mpqa.cs.pitt.edu/

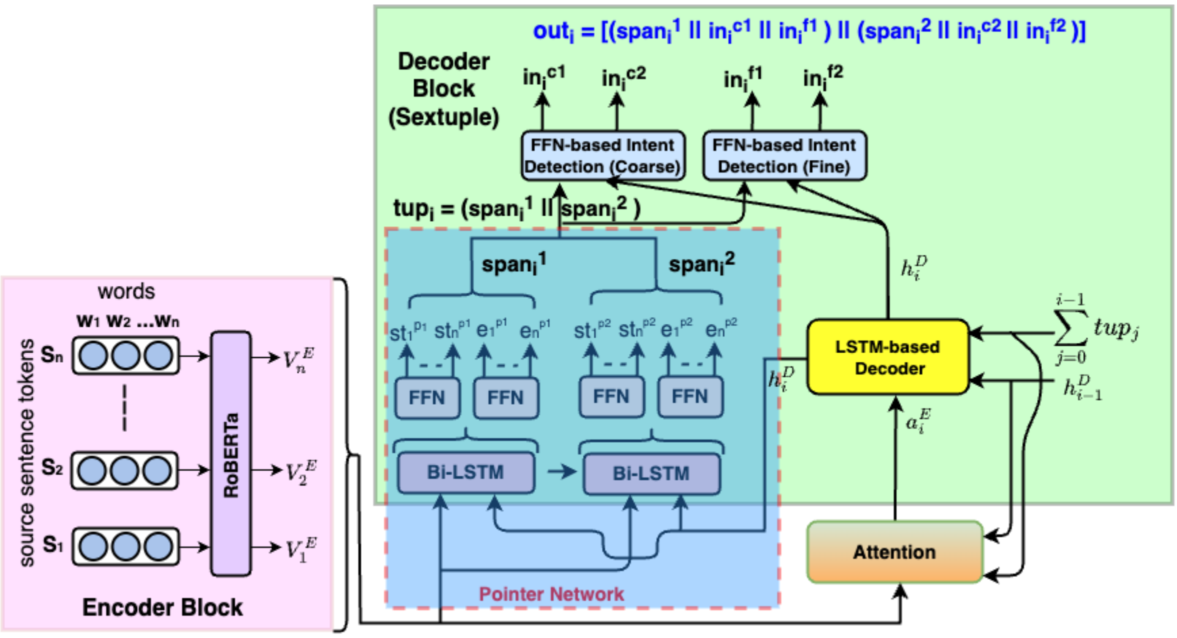


Figure 2: Pointer Network Based multi-label, multi-class intent detection (MLMCID) architecture

ations - English, Spanish and Thai. The text instances of English, Spanish and Thai languages are termed as Facebook (English), Facebook (Spanish) and Facebook (Thai) dataset respectively.

3 Problem Definition

To formally describe the multi-label, multi-class intent detection (MLMCID) problem setting, let there be an input sentence $S_i = \{w_1, w_2, ..., w_n\}$ contains n words. The model aims to extract multiple intent spans along with their coarse and fine classes in the form of a sextuple, $ST = \{out_i | out_i = [(st_i^{p_1}, e_i^{p_1}), in_i^{c_1}, in_i^{f_1}, (st_i^{p_2}, e_i^{p_2}), in_i^{c_2}, in_i^{f_2}]\}_{i=1}^{|ST|}$;

where t_i denotes the i^{th} triplet and $|ST|$ denotes the length of the sextuple set. $st_i^{p_1}$ and $st_i^{p_2}$ represents the beginning position of first intent span and second intent span respectively for the i^{th} sextuple. Similarly, $e_i^{p_1}$ and $e_i^{p_2}$ denotes the end position of first intent span and second intent span for the i^{th} sextuple. So ($st_i^{p_1}$ and $e_i^{p_1}$) mark the first intent span for the i^{th} sextuple. Similarly, ($st_i^{p_2}$ and $e_i^{p_2}$) mark the second intent span for the i^{th} sextuple. $in_i^{c_1}$ and $in_i^{f_1}$ represents the possible coarse and fine intent class of the first intent span. Similarly, $in_i^{c_2}$ and $in_i^{f_2}$ represents the possible coarse and fine intent class of the second intent span. p_1 and p_2 denote the two pointer network models. Pointer Network Model has the following advantages: it is a joint model for entity extraction and relation classification. Pointer network model can detect an intent in a sentence in a form of triplet (intent span, coarse intent label, fine intent label) even if there is an overlap with other intents. c_1 and c_2 mark the course labels. f_1 and f_2 indicates fine labels. out_i is the i^{th} output sextuple.

4 Solution Approach

For the task of multi-label, multi-class intent detection (MLMCID), our goal is to jointly extract the intent spans along with detecting multiple coarse and fine intents. Our MLMCID output representation is a sextuple format. We employ pointer network based architecture for joint extraction of the sextuple. Following are the different components of solution framework approach:

4.1 Encoder

We use four different embeddings in the encoder block (for English language datasets): a) BERT ('bert-base-uncased') (Devlin et al., 2019), b) RoBERTa ('roberta-base-uncased') (Liu et al., 2019b), c) DistilBERT (Sanh et al., 2019) and d) Electra (Clark et al., 2020). For non-English language datasets (Facebook Thai and Spanish), we utilise mBERT (multilingual BERT) (Pires et al., 2019), XLM-R (XLM-RoBERTa) (Conneau et al., 2020) and mDistilBERT (Sanh et al., 2019). mBERT architecture pre-trained on Wikipedia articles from 104 languages. XLM-RoBERTa is a large multi-lingual language model based on RoBERTa, trained on 2.5TB of filtered CommonCrawl data. mDistilBERT is a distilled version of mBERT containing 134 million parameters.

Let, S_i be the i^{th} sentence containing w_1, w_2 ... w_n words. After sentence encoding, the encoder generates a vector ($\mathbf{V}_i^E$) from the i^{th} sentence S_i. It is shown in the 'Encoder Block' in Fig 2.

4.2 Decoder

We apply a Pointer Network-based approach along with LSTM-based sequence generator, attention model and FFN (Feed-Forward Network) architecture (Similar to (Nayak and Ng, 2020)) to identify intent spans and predict the coarse and fine intent labels. Different blocks are as following:

LSTM-based Sequence Generator: The sequence generator structure is based on an LSTM layer with hidden dimension D_h to produce the sequence of two intent spans. Using the attention layer sentence encoding (a_i^E), pointer network based previous tuple ($\mathbf{tup}_i$) and hidden vectors (h_{i-1}^D) as input to generate the hidden representation of the current token (h_i^D). The tup_0 = ($\overrightarrow{0}$) denotes the dummy tuple. Following are LSTM outcomes:

$$\mathbf{tup}_i = \sum_{j=0}^{i-1} \mathbf{tup}_j \tag{1}$$

$$\mathbf{h}_i^D = \text{LSTM}(\mathbf{a}_i^E \| \mathbf{tup}_{i-1}, \mathbf{h}_{i-1}^D) \tag{2}$$

$$\hat{st}_i^1 = w_{st}^1 h_i^m + b_{st}^1, \quad \hat{e}_i^1 = w_e^1 h_i^m + b_e^1 \tag{3}$$

$$st_i^{p_1} = \text{softmax}(\hat{st}_i^1), \quad e_i^{p_1} = \text{softmax}(\hat{e}_i^1) \tag{4}$$

Attention Modeling: Utilizing Bahdanau et al. (2014) attention algorithm we use previous tuple (tup_{i-1}) and hidden vector (h_{i-1}^D) as input at timestamp t to produce the attention weighted context vector (a_i^E) for the current input sentence.

Pointer Network: A Bi-LSTM layer with hidden dimension $\mathbf{D}_H$, followed by two FFN (Feed Forward Networks), constitutes a pointer network. Here we use two-pointer networks for extracting two intent spans. We concatenate $\mathbf{h}_i^D$ and $\mathbf{V}_i^E$ (obtained from the encoding layer) to provide the input of a Bi-LSTM model (forward and backward LSTM), which provides a hidden representation to be fed to FFN models. Two FFNs with softmax provide scores between 0 and 1, the start (st) and end (e) index of one intent span.

where w_{st}^1 and w_e^1 are the weight parameters of FFN. b_{st}^1 and b_e^1 are the bias parameters of the feed-forward layers (FFN). $\hat{st}_i^1$ and $\hat{e}_i^1$ are normalized probabilities of the i^{th} source sentence. $st_i^{p_1}$ and $e_i^{p_1}$ denotes the begin and end token of the first intent span in the first pointer network model of the i^{th} source sentence. Then, the second pointer network model extracts the second entity. After concatenating the first Bi-LSTM output vector ($\mathbf{h}_i^m$) with decoder sequence generator output ($\mathbf{h}_i^D$) and sentence encoding ($\mathbf{V}_i^E$), we feed them to the second pointer network to obtain the position of the begin and end tokens of the second intent span. Together, these two pointer networks produce the feature vectors tup_i containing intent span 1 ($span_i^1$) and span 2 ($span_i^2$).

Intent Detector: We concatenate $\mathbf{tup}_i$ with $\mathbf{h}_i^D$ and pass it through a feed-forward network (FFN) with softmax to produce the normalized probabilities over intent sets and thereby predict the coarse ($in_i^{c_1}$, $in_i^{c_2}$) and fine ($in_i^{f_1}$, $in_i^{f_2}$) intent labels for first and second spans.

4.3 Baselines

We employ different open-source LLMs with prompt based fine-tuning on the training set to generate the two different intent spans and detect coarse and fine intents.

Llama2:[5] We apply Llama2-7b ((Touvron et al., 2023)) using Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023) (to optimize training efficiency) for supervised fine-tuning using MLMCID-Datasets.

GPT: We also use state-of-the-art large-size LLMs, developed by OpenAI: GPT-3.5 (gpt) [6] and GPT-4 (OpenAI, 2023)[7] with example based prompting to extract intent spans and identify coarse and fine intents (Computed on April, 2024).

5 Experiments

To validate our proposed framework, we compare the Pointer Network Model (PNM) of MLMCID while taking various embeddings as input: BERT, RoBERTa, DistilBERT, and Electra on all datasets. We also explore different large language models (Llama2-7b, GPT-3.5 and GPT-4) to check how effectively they can extract multiple intent spans and detect different intents. After that, we experiment with different variations of overall best performing RoBERTa model - varying the training data

[5]https://ai.meta.com/llama/

[6]https://chat.openai.com/

[7]https://openai.com/gpt-4

Dataset		BERT (p, av)	RoBERTa (p, av)	DistilBERT (p, av)	Electra (p, av)	Llama2 (p, av)	GPT-3.5 (p, av)	GPT-4 (p, av)
MIX_SNIPS	A	89.2,80.2	**90.0,81.9**	89.2,80.2	89.8,80.7	48.3,41.2	60.4,55.8	64.7,61.1
	F1	89.0,80.1	**89.7,82.1**	88.5,79.4	89.5,80.5	42.6,40.5	60.2,56.2	62.5,60.3
FACEBOOK (English)	A	98.0,80.8	**98.5,81.2**	97.2,80.2	97.4,80.5	21.0,19.2	70.7,62.1	75.6,76.5
	F1	**98.2,88.2**	92.8,82.8	92.8,82.2	92.8,83.1	20.6,19.6	65.3,60.8	72.6,70.5
MIX_ATIS	A	71.3,64.6	**70.2,63.5**	72.2,63.6	70.6,59.7	16.9,15.0	29.5,32.5	38.7,32.8
	F1	51.7,38.6	**53.4,38.8**	50.3,35.8	46.3,35.5	15.7,14.0	27.2,31.5	36.8,32.6
HWU64	A	83.5,68.0	**85.5,70.0**	82.5,66.2	83.0,66.2	35.8,38.1	56.0,52.3	59.1,53.1
	F1	81.9,65.9	**80.0,63.7**	79.9,64.1	79.4,62.5	32.9,30.5	50.6,51.2	57.3,56.4
BANKING	A	84.0,76.9	**85.4,78.5**	78.8,70.9	79.9,71.8	31.5,31.6	25.4,20.5	47.9,47.4
	F1	82.7,71.4	**85.2,75.2**	79.2,67.9	79.4,68.1	28.2,29.1	20.2,20.3	45.2,43.6
CLINC	A	86.3,72.7	**92.3,81.3**	79.8,68.0	88.7,71.7	57.5,55.9	58.7,57.2	64.3,56.6
	F1	77.1,64.1	**88.3,75.5**	71.7,60.0	81.3,63.0	51.2,50.3	56.3,55.3	63.7,54.3
Overall Average	A	84.1,75.7	**88.2,78.5**	82.2,73.2	85.7,72.2	34.1,37.0	49.2,38.1	60.6,53.3
	F1	80.8,73.9	**85.2,75.8**	81.4,70.6	80.9,71.3	30.5,32.8	44.9,41.4	58.7,53.6

Table 3: Overall Accuracy (A) and Macro F1-score (F1) in (%) of different models in **MLMCID** and LLMs for coarse labels (on English Datasets) - primary intent (p) and average(av). (The best outcomes are marked in **Bold**)

size to understand how much training data is required for decent performance. We also perform zero-shot and few-shot experiments to check the approach's usefulness in the presence of minimal data. Tables 3, 4 and 5 show the overall performances of different models for the English (Mix-SNIPS, Mix-ATIS, Facebook, HWU, BANKING and CLINC) and Non-English (Facebook Thai and Spanish) datasets. We use prediction accuracy and macro F1-score as evaluation metrics. Table 3 and 4 infer performances on primary and overall average of coarse and fine intent labels on English datasets. Following are the details of our findings:

Findings 1: For coarse label intent detection, as shown in Table 3, RoBERTa (with PNM) in MLMCID achieves superior performances in terms of accuracy and F1-score across all datasets of different intent sizes (Mix-SNIPS, Mix-ATIS, HWU, BANKING, CLINC) for both primary intent detection and overall average except for Facebook English where BERT is more effective in terms of F1-score for both primary and overall average.

Findings 2: Similar to coarse intent detection, for fine label intent detection, RoBERTa (with PNM) in MLMCID also produce better results than others in terms of accuracy and F1-score for most of the cases across all English datasets except for Facebook English dataset, where Electra provides better outcome in terms of accuracy and F1-score for both primary and overall intent detection. It is shown in Table 4.

Findings 3: For all English datasets, BERT, RoBERTa, DistilBERT and Electra performs almost similar with decent accuracy and F1-score which signifies the utility of pointer network model based **MLMCID** architecture.

Findings 4: We observe that the LLMs (Llama-2-7b, GPT-3.5, GPT-4) fall behind in performance from Pointer Network based approaches with different encoders, even though they are much larger than our proposed framework, thus strengthening the need for such a specialized **MLMCID** architecture. Llama2-7b performs poorly among three LLMs - this may be due to the fact of less contextual understanding in this specific task. More details in Appendix A.

Findings 5: RoBERTa with PNM in MLMCID performs better than any other models for overall average accuracy and F1-score across all English datasets for both primary and average course and

Dataset		BERT (p, av)	RoBERTa (p, av)	DistilBERT (p, av)	Electra (p, av)	Llama2 (p, av)	GPT-3.5 (p, av)	GPT-4 (p, av)
MIX_SNIPS	A	85.4,80.9	**89.6,85.0**	87.5,81.9	86.3,80.9	35.0,20.1	64.2,60.5	64.7,61.1
	F1	83.5,80.1	**89.0,85.9**	86.6,81.7	86.2,82.1	27.5,22.1	55.6,51.2	57.3,54.9
FACEBOOK (English)	A	96.5,81.3	97.5,80.7	96.5,79.7	**98.5,81.7**	11.1,12.1	44.4,46.4	73.4,77.6
	F1	87.5,79.5	94.5,82.0	78.4,73.1	**95.4,82.7**	9.2,9.7	40.2,41.3	69.5,69.8
MIX_ATIS	A	71.3,64.6	**70.2,63.5**	72.2,63.6	70.6,59.7	16.9,15.0	29.5,32.5	38.7,32.8
	F1	51.7,38.6	**53.4,38.8**	50.3,35.8	46.3,35.5	15.7,14.0	27.2,31.5	36.8,32.6
HWU64	A	74.1,57.2	**83.0,67.1**	75.1,57.7	70.1,53.9	29.8,20.3	41.8,33.2	52.5,48.2
	F1	57.9,43.6	**68.3,52.8**	61.0,44.6	54.5,41.6	25.6,19.6	31.6,30.5	48.9,46.3
BANKING	A	78.5,61.2	**82.3,71.2**	69.5,54.3	73.3,57.2	19.0,17.7	21.0,20.5	27.3,25.7
	F1	73.5,57.0	**80.0,68.4**	64.1,51.4	67.8,52.4	15.6,16.2	18.1,19.4	25.6,24.3
CLINC	A	88.1,73.9	**89.3,81.2**	81.6,68.1	84.9,70.8	43.0,37.8	47.0,40.9	55.7,48.1
	F1	81.7,66.9	**85.3,74.2**	75.2,60.8	79.4,63.4	39.6,35.7	45.4,39.5	51.2,45.3
Overall Average	A	82.3,69.9	**85.3,74.8**	80.4,67.5	80.6,67.4	25.8,20.5	41.3,39.0	52.1,48.9
	F1	72.7,60.9	**78.4,66.9**	69.3,57.9	71.6,59.7	22.2,19.6	36.4,35.6	48.2,45.5

Table 4: Overall Accuracy (A) and Macro F1-score (F1) in (%) of different models in **MLMCID** and LLMs for fine labels (on English Datasets) - primary intent (p) and average(av). (The best outcomes are marked in **Bold**)

Dataset			mBERT (p, av)	XLM-R (p, av)	mDistilBERT (p, av)	Llama-2 (p, av)	GPT-3.5 (p, av)	GPT-4 (p, av)
FACEBOOK (Spanish)	Coarse	A	98.0,80.7	**98.5,81.5**	98.0,80.2	51.2,39.9	64.6,61.6	70.7,75.6
		F1	91.3,82.2	**92.5,82.7**	91.1,82.9	47.2,39.6	62.6,61.3	69.4,69.3
	Fine	A	96.7,80.0	**97.5,81.0**	96.5,80.2	38.3,27.2	57.6,56.6	69.7,74.2
		F1	84.6,80.0	**86.0,81.7**	84.3,76.8	36.2,30.6	55.4,55.0	66.2,65.6
FACEBOOK (Thai)	Coarse	A	96.5,79.8	**97.0,80.0**	96.8,79.0	28.0,24.2	69.7,58.6	73.4,71.5
		F1	88.4,75.8	**96.6,78.8**	94.2,73.4	25.6,24.8	67.8,57.2	71.6,69.3
	Fine	A	96.0,79.5	**96.5,79.7**	95.5,77.2	16.3,15.2	18.2,18.7	68.7,64.9
		F1	84.1,74.2	**82.5,75.5**	68.8,62.7	15.7,14.9	17.9,16.8	59.2,58.7
Average	Coarse	A	97.2,80.3	**97.8,80.8**	97.4,79.6	39.6,32.1	67.2,60.1	72.1,73.6
		F1	89.8,79.0	**94.6,80.8**	92.6,78.2	36.4,32.2	65.2,59.3	70.5,69.3
	Fine	A	97.3,79.7	**97.0,80.8**	96.0,78.7	27.3,21.2	37.9,37.7	69.2,69.6
		F1	84.3,77.1	**84.3,78.6**	76.5,69.8	25.9,22.8	36.7,35.9	62.7,62.2

Table 5: Overall Accuracy (A) and Macro F1 (F1) in (%) of different models in **MLMCID** and LLMs for coarse and fine grained labels of Facebook Spanish and Thai datasets - primary intent (p) and overall average(av). (The best outcomes are marked in **Bold**)

fine intent detection after intent spans extraction.

Findings 6: For non-English languages like Spanish (Facebook) and Thai (Facebook) datasets , we observe that for both fine and coarse grained intent labels, XLM-R and mBERT both produce good results but XLM-R outperforms mBERT in all aspects across all datasets and overall for both primary intent detection and overall average intent detection with intent span extraction.

Findings 7: To check the effectivity of span extraction by pointer network, we vary the similarity (extracted intent span vs actual intent span) threshold utilise that extracted span to check the overall accuracy. We check for 50% - 90% similarity threshold range and overall framework (RoBERTA with PNM) accuracies (for both primary and average intent) across all datasets for coarse and fine intent labels are shown in Table 6 and 7. It is seen a good performance even with 50% similarity which shows the efficacy of the system.

Ablation Studies

1. K-shot setting: To evaluate the RoBERTa based PNM model of MLMCID architecture, we utilize K samples for all English datasets where K = 5 (5-shot) and 10 (10-shot) for coarse and fine intent labels. The accuracy and F1-score of primary and average intents are shown in Table 9. This shows even with very limited number of data-points (like in 5-shot), the system is able to achieve a decent performance across different datasets.

2. Practical Datasets: We test the trained RoBERTa models with PNM (using SNIPS, BANKING and CLINC dataset) in MLMCID to evaluate on external MPQA and Yahoo datasets. We also check LLMs - Llama2-7b (vanilla and fine-tuned), GPT-3.5 and GPT-4 on MPQA and Yahoo but RoBERTa based PNM in MLMCID outperfomrs LLMs in most of the cases and show decent performance as shown in Table 8. It is seen that, for Llama2-7b vanilla model performs poorly and

Th	Dataset (primary (p) and average (av) intent) in %							
	MIX_SNIPS	FB_en	FB_es	FB_th	MIX_ATIS	HWU64	BANKING	CLINC
50 %	89.2,80.9	96.0,78.5	94.5,77.4	89.9,82.4	95.1,90.2	85.5,70.0	81.8,74.7	90.1,79.2
60 %	87.7,78.9	95.0,77.9	86.5,71.2	77.4,70.3	91.9,90.2	85.5,68.9	79.4,72.0	88.4,77.5
70 %	79.4,70.8	91.0,74.6	75.6,63.1	75.2,67.7	85.1,89.2	84.6,68.1	75.9,68.3	84.0,73.0
80 %	70.4,63.5	83.0,68.8	72.6,59.4	71.4,62.9	83.8,88.2	81.9,66.6	69.9,62.8	79.1,67.6
90 %	59.2,54.2	75.0,63.2	61.6,50.3	69.4,59.6	80.9,86.2	77.5,62.6	63.4,56.0	67.5,58.2

Table 6: Overall Accuracy (A) in (%) of RoBERTa model in **MLMCID** for coarse grained labels (on English Datasets) - primary (p) and average (av) intents. ('Th' indicates threshold value)

Th	Dataset (primary (p) and average (av) intent) in %							
	MIX_SNIPS	FB_en	FB_es	FB_th	MIX_ATIS	HWU64	BANKING	CLINC
50 %	83.6,80.7	93.5,78.1	91.5,75.9	89.6,81.1	95.1,90.2	83.0,67.1	77.1,69.8	86.6,78.9
60 %	82.1,78.9	92.5,77.0	85.6,70.2	82.4,79.6	91.9,90.2	80.4,65.0	74.8,67.5	86.1,77.4
70 %	76.1,72.3	87.6,71.9	78.7,63.8	75.9,67.2	85.1,89.2	79.5,64.3	69.1,62.2	82.9,70.9
80 %	68.6,64.8	78.7,65.9	74.8,60.6	68.4,61.0	83.8,88.2	75.2,62.4	64.5,56.0	77.0,68.0
90 %	55.2,52.4	72.8,61.0	63.0,50.7	65.4,57.3	80.9,86.2	67.5,55.1	57.7,49.4	66.4,62.8

Table 7: Overall Accuracy (A) in (%) of RoBERTa model in **MLMCID** for fine grained labels (on English Datasets) - primary (p) and average (av) intents. ('Th' indicates threshold value)

Dataset		Llama2-7b Fine-tune (p,av)	Llama2-7b Vanilla (p, av)	GPT-3.5 (p, av)	GPT-4 (p, av)	RoBERTa-SNIPS(p, av)	RoBERTa-BANKING(p,av)	RoBERTa-CLINC(p, av)
MPQA	Fine	42.8,27.1	18.8,16.9	20.0,14.2	48.5,37.1	45.0,42.5	44.5,42.0	43.9,41.5
	Coarse	65.7,64.2	51.9,50.0	62.8,59.9	68.5,45.6	75.6,43.7	73.0,41.9	72.8,42.6
YAHOO	Fine	48.3,37.5	18.8,15.8	11.4,10.6	58.0,56.2	55.3,54.9	54.0,53.8	52.9,54.2
	Coarse	61.2,49.9	52.8,50.0	50.0,50.0	61.2,49.1	66.3,65.7	64.5,62.9	63.2,60.8

Table 8: Overall Accuracy (A) in (%) of RoBERTa model in **MLMCID** (trained on SNIPS, BANKING and CLINC) and LLMs for fine and course grained labels - primary (p) and average (av) intent.

fine-tune version perform better but does not outperform GPT and RoBERTa based models.

3. Intent Counts: All datasets have two intents (primary and non-primary) in one sentence except for Yahoo, 2.6% cases with more than 2 intents so we show all results considering the case of 2 intents in a sentence. Our system is also effective for more than two intents by utilizing more pointer network block in the decoder framework, as shown in Appendix A.2.

Dataset			Coarse (p, avg)	Fine (p, avg)
SNIPS	5-shot	A	61.0,49.2	70.9,53.3
		F1	58.1,46.4	67.9,51.7
	10-shot	A	61.4,52.1	75.9,63.1
		F1	60.7,47.4	75.1,61.0
FACEBOOK (English)	5-shot	A	83.5,62.0	76.0,58.3
		F1	58.0,42.8	26.7,20.4
	10-shot	A	87.5,67.8	83.5,64.3
		F1	59.5,45.9	34.3,25.2
HWU-64	5-shot	A	57.2,39.3	47.8,29.6
		F1	49.3,34.7	35.5,22.1
	10-shot	A	62.2,43.5	62.2,43.3
		F1	58.2,39.2	46.2,31.9
BANKING	5-shot	A	36.0,28.2	62.3,38.6
		F1	32.5,25.0	56.7,34.4
	10-shot	A	46.0,32.9	76.1,52.9
		F1	46.1,31.4	71.2,48.0
CLINC	5-shot	A	78.4,50.4	76.3,53.4
		F1	69.9,44.0	65.8,44.6
	10-shot	A	87.3,65.9	89.6,69.7
		F1	79.3,58.5	79.3,58.5

Table 9: Accuracy (A) and F1-Score for coarse and fine intents by RoBERTa(in %) for k-shot, k = {5, 10}

Experimental Settings: Our experiments are conducted on two Tesla P100 GPUs with 16 GB RAM, 6 Gbps clock cycle, GDDR5 memory and one 80GB A100 GPU, 210MHz clock cycle, 2*960 GB SSD with 5 epochs. We use Adam optimizer with learning rate: 10^{-5} with cross-entropy as the loss function, weight decay: 10^{-5} and a dropout rate of 0.5 is applied on the embeddings to avoid overfitting for all experiments (Details are in Appendix). All methods took less than 120 GPU minutes (except Llama2: ∼4-5 hrs) for fine tuning and ∼2 hrs for inference. All the hyperparameters are tuned on the dev set. We have used NLTK, Spacy, Scikit-learn, openai (version=0.28), huggingface_hub, torch and transformers python packages for all experiments and evaluation [8].

6 Loss Function

We calculate loss of different intent classes across all samples for primary, non-primary intents and their respective primary and non primary spans as shown in equation 5, 6 and 7 respectively. For training our model, we minimize the sum of negative log-likelihood loss for classifying the intent and the four pointer locations corresponding to the primary and non primary intent spans as shown in equation 8.

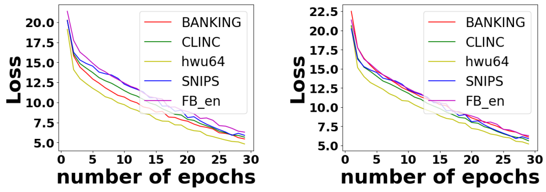
$$\mathcal{L}_p = -\frac{1}{N}\sum_{i=1}^{N}\Big[\sum_{j=1}^{C}(y_1)_{ij}log(p_{ij}) - \frac{1}{J}\sum_{j=1}^{J}\log((y_1)j^n)\Big] \quad (5)$$

$$\mathcal{L}_{np} = -\frac{1}{N}\sum_{i=1}^{N}\Big[\sum_{j=1}^{C}(y_2)_{ij}log(p_{ij}) - \frac{1}{J}\sum_{j=1}^{J}\log((y_2)j^n)\Big] \quad (6)$$

$$\mathcal{L}_{span} = -\frac{1}{N\times J}\sum_{n=1}^{N}\sum_{j=1}^{J}\Big[\log((st^{p_1})j^n\cdot (e^{p1})j^n) + \log((st^{p2})j^n\cdot(e^{p2})j^n)\Big] \quad (7)$$

Here, C is the number of intent classes and $(y_1) \in \{in^{c1}, in^{f1}\}$ and $(y_2) \in \{in^{c2}, in^{f2}\}$. $(y_1)ij$ and $(y_2)ij$ are the one-hot ground truth labels for sample i and class j for the primary and non-primary intents respectively, and p_{ij} is the predicted probability for sample i and class j. n represents the n^{th} training instance with N being the batch size, j represents the j^{th} decoding time step with J being the length of the longest target sequence among all instances in the current batch. $st^p, e^p;\ p \in \{p_1, p_2\}$ respectively represent the softmax scores corresponding to the true start and end positions of the primary and non primary spans. Fig 3 shows the

[8]All Code / Data details are in https://github.com/ankan2/multi-intent-pointer-network



(a) Combined loss - Coarse (b) Combined Loss - Fine

Figure 3: By RoBERTa based pointer network (PNM) model in ***MLMCID***

variation of the overall loss for course and fine intents with respect to the training progress (in terms of epochs) across different datasets. Loss decreases with larger epochs and after 10 epochs the loss decrement is significant to obtain decent outcome.

$$\mathcal{L} = \mathcal{L}_p + \mathcal{L}_{np} + \mathcal{L}_{span} \tag{8}$$

7 Conclusion

Intent detection is crucial in task-oriented conversation systems. Earlier works focus on scenarios with the presence of a single intent and do not extract intent spans. This work is one of the first to consider multiple intents in a single sentence within a conversation system, including primary and non-primary intents. First, we create novel datasets using state-of-the-art datasets with coarse and fine intent labels. Then, we develop a Pointer Network-based encoder-decoder framework (MLMCID - multi-label multi-class intent detection) using RoBERTa (for English data) and XLM-R (for non-English data) to jointly extract intent spans from sentences and detect corresponding coarse and fine intents. We show that the MLMCID model even outperforms various LLMs for these specific tasks across different datasets. The approach demonstrates efficacy even in few-shot scenarios. Qualitative analysis shows a reasonable grasp of primary and secondary intent concepts. Overall, this highlights the importance of multi-intent modeling for real-world conversational AI, with the datasets and models providing a strong foundation for future research.

Limitations and Discussion

Table 3, 4, 5 shows that even when our model fails to give the correct predictions exactly, it predicts the primary intent correctly most of the time. This is due to the fact we are using the top-2 intents to infer the primary and non-primary intents using the same classifier. Also, in some examples, the primary and non-primary intent Labels, when predicted wrongly, are swapped, suggesting that the model is still able to grasp the notion of intent. We shall work on these limitations in future.

Ethical Concerns

We use publicly available codes and datasets so there is no ethical concerns.

Acknowledgements

The work was supported in part by Prime Minister Research Fellowship (PMRF).

References

Gpt-3.5 turbo documentation.

Nicholas Asher, Farah Benamara, and Yvette Yannick Mathieu. 2009. Appraisal of opinion expressions in discourse. *Lingvisticæ Investigationes*, 32(2):279–292.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Xin-Qiang Cai, Peng Zhao, Kai-Ming Ting, Xin Mu, and Yuan Jiang. 2019. Nearest neighbor ensembles: An effective method for difficult problems in streaming classification with emerging new classes. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 970–975. IEEE.

Inigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. *arXiv preprint arXiv:2003.04807*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Sam Coope, Tyler Farghly, Daniela Gerz, Ivan Vulić, and Matthew Henderson. 2020. Span-convert: Few-shot span extraction for dialog with pretrained conversational representations. *arXiv preprint arXiv:2005.08866*.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Chen Cui, Wenjie Wang, Xuemeng Song, Minlie Huang, Xin-Shun Xu, and Liqiang Nie. 2019. User attention-guided multimodal dialog systems. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 445–454.

Ali Degirmenci and Omer Karal. 2022. Efficient density and cluster based incremental outlier detection in data streams. *Information Sciences*, 607:901–920.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Mauajama Firdaus, Asif Ekbal, and Erik Cambria. 2023. Multitask learning for multilingual intent detection and slot filling in dialogue systems. *Information Fusion*, 91:299–315.

Rashmi Gangadharaiah. 2019. Joint multiple intent detection and slot labeling for goal-oriented dialog.

Souradip Guha, Ankan Mullick, Jatin Agrawal, Swetarekha Ram, Samir Ghui, Seung-Cheol Lee, Satadeep Bhattacharjee, and Pawan Goyal. 2021. Matscie: An automated tool for the generation of databases of methods and parameters used in the computational materials science literature. *Computational Materials Science (Comput. Mater. Sci.)*, 192:110325.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2019. Convert: Efficient and accurate conversational representations from transformers. *arXiv preprint arXiv:1911.03688*.

Sheng Jiang, Su Zhu, Ruisheng Cao, Qingliang Miao, and Kai Yu. 2023. Spm: A split-parsing method for joint multi-intent detection and slot filling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 668–675.

Ilja Kuzborskij, Francesco Orabona, and Barbara Caputo. 2013. From n to n+ 1: Multiclass transfer incremental learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3358–3365.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. 2019. An evaluation dataset for intent classification and out-of-scope prediction. *arXiv preprint arXiv:1909.02027*.

Guobo Liao, Peng Zhang, Hongpeng Yin, Xuanhong Deng, Yanxia Li, Han Zhou, and Dandan Zhao. 2023. A novel semi-supervised classification approach for evolving data streams. *Expert Systems with Applications*, 215:119273.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019a. Benchmarking natural language understanding services for building conversational agents.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhiyuan Ma, Jianjun Li, Guohui Li, and Yongjing Cheng. 2022. Unitranser: A unified transformer semantic representation framework for multimodal task-oriented dialog system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 103–114.

Zhiyuan Ma, Jianjun Li, Zezheng Zhang, Guohui Li, and Yongjing Cheng. 2021. Intention reasoning network for multi-domain end-to-end task-oriented dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2273–2285.

Mohammad Masud, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani M Thuraisingham. 2010. Classification and novel class detection in concept-drifting data streams under time constraints. *IEEE Transactions on Knowledge and Data Engineering*, 23(6):859–874.

Xin Mu, Kai Ming Ting, and Zhi-Hua Zhou. 2017a. Classification under streaming emerging new classes: A solution using completely-random trees. *IEEE Transactions on Knowledge and Data Engineering*, 29(8):1605–1618.

Xin Mu, Feida Zhu, Juan Du, Ee-Peng Lim, and Zhi-Hua Zhou. 2017b. Streaming classification with emerging new class by class matrix sketching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Ankan Mullick. 2023a. Exploring multilingual intent dynamics and applications. *IJCAI Doctoral Consortium*.

Ankan Mullick. 2023b. Novel intent detection and active learning based classification (student abstract). *arXiv e-prints*, pages arXiv–2304.

Ankan Mullick, Akash Ghosh, G Sai Chaitanya, Samir Ghui, Tapas Nayak, Seung-Cheol Lee, Satadeep Bhattacharjee, and Pawan Goyal. 2024. Matscire: Leveraging pointer networks to automate entity and relation extraction for material science knowledge-base construction. *Computational Materials Science*, 233:112659.

Ankan Mullick, Pawan Goyal, and Niloy Ganguly. 2016. A graphical framework to detect and categorize diverse opinions from online news. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 40–49.

Ankan Mullick, Shivam Maheshwari, Pawan Goyal, and Niloy Ganguly. 2017. A generic opinion-fact classifier with application in understanding opinionatedness in various news section. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 827–828.

Ankan Mullick, Ishani Mondal, Sourjyadip Ray, R Raghav, G Chaitanya, and Pawan Goyal. 2023. Intent identification and entity extraction for healthcare queries in indic languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1825–1836.

Ankan Mullick, Abhilash Nandy, Manav Kapadnis, Sohan Patnaik, R Raghav, and Roshni Kar. 2022a. An evaluation framework for legal document summarization. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4747–4753.

Ankan Mullick, Abhilash Nandy, Manav Nitin Kapadnis, Sohan Patnaik, and R Raghav. 2022b. Fine-grained intent classification in the legal domain. *arXiv preprint arXiv:2205.03509*.

Ankan Mullick, Shubhraneel Pal, Tapas Nayak, Seung-Cheol Lee, Satadeep Bhattacharjee, and Pawan Goyal. 2022c. Using sentence-level classification helps entity extraction from material science literature. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4540–4545.

Ankan Mullick, Sukannya Purkayastha, Pawan Goyal, and Niloy Ganguly. 2022d. A framework to generate high-quality datapoints for multiple novel intent detection. *arXiv preprint arXiv:2205.02005*.

Gyoung S Na, Donghyun Kim, and Hwanjo Yu. 2018. Dilof: Effective and memory efficient local outlier detection in data streams. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1993–2002.

Tapas Nayak and Hwee Tou Ng. 2020. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8528–8535.

OpenAI. 2023. Gpt-4 technical report.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Libo Qin, Fuxuan Wei, Tianbao Xie, Xiao Xu, Wanxiang Che, and Ting Liu. 2021. Gl-gin: Fast and accurate non-autoregressive model for joint multiple intent detection and slot filling. *arXiv preprint arXiv:2106.01925*.

Libo Qin, Xiao Xu, Wanxiang Che, and Ting Liu. 2020. Agif: An adaptive graph-interactive framework for joint multiple intent detection and slot filling. *arXiv preprint arXiv:2004.10087*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. 2012. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2018. Cross-lingual transfer learning for multilingual task oriented dialog. *arXiv preprint arXiv:1810.13327*.

Mengxiao Song, Bowen Yu, Li Quangang, Wang Yubin, Tingwen Liu, and Hongbo Xu. 2022. Enhancing joint multiple intent detection and slot filling with global intent-slot co-occurrence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7967–7977.

Sandeep Soni, Tanushree Mitra, Eric Gilbert, and Jacob Eisenstein. 2014. Modeling factuality judgments in social media text. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 415–420.

Yu Sun, Ke Tang, Leandro L Minku, Shuo Wang, and Xin Yao. 2016. Online ensemble learning of data streams with gradually evolved classes. *IEEE Transactions on Knowledge and Data Engineering*, 28(6):1532–1545.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Gokhan Tur, Dilek Hakkani-Tür, and Larry Heck. 2010. What is left to be understood in atis? In *2010 IEEE Spoken Language Technology Workshop*, pages 19–24. IEEE.

Min Wang, Ke Fu, Fan Min, and Xiuyi Jia. 2020. Active learning through label error statistical methods. *Knowledge-Based Systems*, 189:105140.

Congying Xia, Wenpeng Yin, Yihao Feng, and Philip Yu. 2021. Incremental few-shot text classification with multi-round new classes: Formulation, dataset and system. *arXiv preprint arXiv:2104.11882*.

Guangfeng Yan, Lu Fan, Qimai Li, Han Liu, Xiaotong Zhang, Xiao-Ming Wu, and Albert YS Lam. 2020. Unknown intent detection using gaussian mixture model with an application to zero-shot intent classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1050–1060.

Li-Ming Zhan, Haowen Liang, Bo Liu, Lu Fan, Xiao-Ming Wu, and Albert Lam. 2021. Out-of-scope intent detection with self-supervision and discriminative training. *arXiv preprint arXiv:2106.08616*.

Yunhua Zhou, Peiju Liu, and Xipeng Qiu. 2022. Knn-contrastive learning for out-of-domain intent classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5129–5141.

A Experimental Findings

A.1 Why encoder decoder model performs well

Pointer Network model is a state-of-the-art approach which is ideal for extracting multiple spans from a sentence using the pointing mechanism to directly select positions in the input sequence, allowing for variable-length outputs and precise boundary identification. Their attention mechanism effectively handles context, enabling accurate span extraction in a computationally efficient manner. It is effective also because of -

- Dynamically predict entity spans within a sequence, enhancing adaptability across various NLP tasks
- capture the interdependence between spans and intents, crucial for tasks where one intent's prediction relies on another characteristics within the same context.
- Reduce the need for manual feature engineering, learning to predict spans directly from input data for more efficient models
- Finally, enable end-to-end learning by directly predicting entity span positions, facilitating seamless integration with other neural network components.

Dataset	Intent 1 (%)	Intent 2 (%)	Intent 3 (%)	Average (%)
MIX_SNIPS (fine)	81.2	73.8	60.3	71.7
MIX_SNIPS (coarse)	85.4	74.4	62.3	74.0
BANKING (fine)	79.3	60.0	56.3	65.2
BANKING (coarse)	83.3	68.9	59.6	70.6
CLINC (fine)	80.7	69.2	55.4	68.4
CLINC (coarse)	81.9	71.7	58.3	70.6

Table 10: 3-Intent Detection by Roberta based PNM

A.2 PNM for more than two intent cases

To evaluate the effectiveness of the Pointer Network framework for more than two intents, we experimented with a small sample from the MIX_SNiPS, BANKING, and CLINC datasets, incorporating three intents. For instance, the sentence "Will it snow this weekend? Please help me book a rental car for Nashville and play that song called 'Bring the Noise'" includes the intents: weather, car_rental, play_music. Table 10 presents the performance of RoBERTa on this annotated sample. The results demonstrate the effectiveness of our system in handling a larger number of intents, as reflected by the accuracy (in %).

A.3 Scalability

We experiment with datasets composed of two intents with the P100 server with 16GB GPU B where 6-9 GB GPU VRAM has been utilised. Further we experiment on the dataset with three intents in the same server which use 12-13 GB GPU VRAM so our approach is scalable and applicable in resource constrained environments. It is also seen that in case of larger numbers of intents with the introduction of additional pointer networks - the system is scalable and does not require large computational costs. So the framework can be useful in real time processing for large scale systems. Though it is also to be noted that most of the datasets are composed with two intents even in the real life sentences.

A.4 Single Intent Detection

We perform additional experiments on three datasets with various intent sizes - SNIPS (small), BANKING (medium) and CLINC (large) and detect the single-intent text using RoBERTa based pointer network architecture - which is shown in the following table (in %). It shows the effectiveness of our model for coarse (c) and fine (f).

B Experimental Settings

Our experiments are conducted on two Tesla P100 GPUs with 16 GB RAM, 6 Gbps clock cy-

Dataset	coarse (%)	fine (%)
SNIPS	90.0	85.9
BANKING	83.9	81.8
CLINC	80.0	75.3

Table 11: Single Intent Detection

cle, GDDR5 memory and one 80GB A100 GPU, 210MHz clock cycle, 2*960 GB SSD with 5 epochs. We use Adam optimizer with learning rate: 10^{-5} with cross-entropy as the loss function, weight decay: 10^{-5} and a dropout rate of 0.5 is applied on the embeddings to avoid overfitting for all experiments. All methods took less than 120 GPU minutes (except Llama2: $\sim$4-5 hrs) for fine tuning and $\sim$2 hrs for inference. All the hyperparameters are tuned on the dev set. We have used NLTK, Spacy, Scikit-learn, openai(version=0.28), huggingface_hub, torch and transformers python packages for all experiments and evaluation.

C Example

Figure 4 shows some examples from **MLMCID** dataset. Table 12 and 13 shows some examples of fine to coarse label conversion for **MLMCID** dataset. Table 14 shows some examples of the intent classes predicted with their respective confidence for PNM.

Sr. No.	Dataset	Text	Dominant Intent	non Dominant Intent
1	SNIPS	Book a table at Joan's on Third for my family reunion on Saturday, How much is a dinner there?	BookRestaurant	GetPlaceDetails
		Walking directions from home to my Halloween party stopping by a wine store, What will the weather be like from Home to there?	GetWeather	GetDirections
		Share my location with my Uber driver,Send a message to Michael with my ETA	Location Service	Traffic update
		Get me a table at Delmonico's next monday at 8pm,What's the weather forecast for the next week?	App Service	GetWeather
		Is it cold outside? How far am I from the Guggenheim museum?	GetWeather	Location Service
2	BANKING	I'd like to order another card, Why haven't I received my new card yet?	getting spare card	card arrival
		For the disposable cards, what are the restrictions? I made a card payment which is showing me pending.	card_service_enquiry	card problem
		The balance has not been updated. What does a pending cash withdrawal mean?	payment_inconsistency	General_Enquiry
		I was charged an additional amount of money for making a purchase with my card,is there something blocking me from making transfers	cancelled_transfer	card_payment_fee_charged
		I was made to pay an additional pound! When I got cash the exchange rate was wrong	wrong_exchange_rate	extra_charge
3	CLINC	just call me dennis, how many dollars can i exchange for 25 euros	exchange_rate	change_user_name
		i can't decide on dinner, what do you suggest,i need a really good recipe for making dinner	restaurant	recipe
		can you confirm i have a reservation at 5 pm on march 13th,can you tell me what reminders i've set	confirm_reservation	reminder
		how much money do i have coming in each month,search for travel alerts for kenya	income	travel_alert
		can you set the temp to 69 and check reservation availability for 2 at red lobster at 8pm	smart_home	restaurant_reservation
4	Facebook Multilingual Dialog Dataset	remind me to pick up contact lenses tomorrow, set the alarm for 5 mins and 30 seconds	reminder_service	change_alarm_content
		Recurdame que pare en la tienda cuando salga del trabajo, se supone que llover en Palm Springs la prxima semana?	reminder_service	get_weather
		ตั้งนาฬิกาปลุกสำหรับมื้อกลางวันพรุ่งนี้ อาทิตย์นี้จะมีพายุฝนฟ้าคะนองไหม	change_alarm_content	get_weather
		Necesito un informe meteorolgico para el viernes, recuerde que tengo una cita con el mdico el viernes a las 5:00 p. m.	reminder_service	get_weather
		พรุ่งนี้จะหนาวแค่ไหน เตือนหาคนเลี้ยงหมาพรุ่งนี้	reminder_service	get_weather
5	HWU64	what time have you set the alarm? do you think i should see a movie or go out to dinner.	general_query	alarm_query
		God! Can i get delivery from here	takeaway_query	general_query
		remind me at eight am tomorrow that i have a lunch meeting at noon in the conference room, are there any alarms	calender_set	alarm_query

Figure 4: Examples in ***MLMCID*** Dataset

Sr. No.	Dataset	Coarse Label	Fine Labels Combined
1.	SNIPS	Traffic_update	ComparePlaces, GetPlaceDetails, ShareCurrentLocation, SearchPlace, GetDirections
		App_Service	RequestRide, BookRestaurant
		Location_service	GetTrafficInformation, ShareETA
		GetWeather	GetWeather
2.	BANKING	Cancelled_ transfer	cancel_transfer, beneficiary_not_allowed
		Card_problem	card_arrival, card_linking, card_swallowed, activate_my_card, declined_card_payment, reverted_card_payment?, pending_card_payment, card_not_working, lost_or_stolen_card, pin_blocked, card_payment_fee_charged, card_payment_not_recognised, card_acceptance
		exchange_rate_query	exchange_rate, fiat_currency_support, card_payment_wrong_exchange_rate, wrong_exchange_rate_for_cash_withdrawal
		General_Enquiry	extra_charge_on_statement, card_delivery_estimate, pending_cash_withdrawal, automatic_top_up, verify_top_up, topping_up_by_card, exchange_via_app, atm_support, lost_or_stolen_phone, transfer_timing, transfer_fee_charged, receiving_money, top_up_by_cash_or_cheque, exchange_charge, cash_withdrawal_charge, apple_pay_or_google_pay
		Top_up	top_up_by_bank_transfer_charge, pending_top_up, top_up_limits, top_up_reverted, top_up_failed
		Account_opening	age_limit
		transaction_problem	contactless_not_working, wrong_amount_of_cash_received, transfer_not_received_by_recipient, balance_not_updated_after_cheque_or_cash_deposit, declined_cash_withdrawal, pending_transfer, transaction_charged_twice, declined_transfer, failed_transfer
		Card_service_enquiry	visa_or_mastercard, disposable_card_limits, getting_virtual_card, supported_cards_and_currencies, getting_spare_card, virtual_card_not_working, top_up_by_card_charge, card_about_to_expire, country_support
		Identity_verification	unable_to_verify_identity, why_verify_identity, verify_my_identity
		Service_request	order_physical_card, edit_personal_details, get_physical_card, passcode_forgotten, change_pin, terminate_account, request_refund, verify_source_of_funds, transfer_into_account, get_disposable_virtual_card
		Malpractice	compromised_card, cash_withdrawal_not_ recognised
		Payment_inconsistency	direct_debit_payment_not_recognised, Refund_not_showing_up, balance_not_updated_after_bank_transfer

Table 12: Fine to Coarse Labels Conversion Examples for SNIPS and BANKING Dataset

Sr. No.	Dataset	Coarse Label	Fine Labels Combined
3.	CLINC	health_suggestion	nutrition_info, oil_change_how, calories
		Restaurant	restaurant_reviews, accept_reservations, restaurant_reservation, meal_suggestion, restaurant_suggestion
		account	redeem_rewards, report_lost_card, balance, bill_balance, credit_limit, rewards_balance, bill_due, credit_score, transactions, spending_history, damaged_card, pin_change, replacement_card_duration, new_card, direct_deposit, credit_limit_change, payday, application_status, pto_request, pto_request_status, pto_balance, pto_used
		communication	make_call, text
		Reminder	remind_update, remind, reminder_update, reminder, meeting_schedule
		banking_enquiry	account_blocked, freeze_account, interest_rate
4.	Facebook Multilingual Dialog Dataset	change_alarm_content	cancel alarm, modify alarm, set alarm, snooze alarm
		reminder_service	cancel reminder, set reminder, show reminders
		sunset_sunrise	weather check sunrise, weather check sunset
		get_weather	weather find
		read alarm content	show alarm, time left on alarm
5.	HWU64	alarm	set, remove, query
		audio	audio_volume_mute, audio_volume_down, audio_volume_other, audio_volume_up
		iot	iot_hue_lightchange, iot_hue_lightoff, iot_hue_lighton, iot_hue_lightdim, iot_cleaning, iot_hue_lightup, iot_coffee, iot_wemo_on, iot_wemo_off
		calendar	calendar_query, calendar_set, calendar_remove
		play	play_music, play_radio, play_audiobook, play_podcasts, play_game
		general	general_query, general_greet, general_joke, general_negate, general_dontcare, general_repeat, general_affirm, general_commandstop, general_confirm, general_explain, general_praise
		datetime	datetime_query, datetime_convert
		takeaway	takeaway_query, takeaway_order
		news	news_query
		music	music_likeness, music_query, music_settings, music_dislikeness
		weather	weather_query
		qa	qa_stock, qa_factoid, qa_definition, qa_maths, qa_currency
		social	social_post, social_query
		recommendation	recommendation_locations, recommendation_events, recommendation_movies
		cooking	cooking_recipe, cooking_query
		email	email_sendemail, email_query, email_querycontact, email_addcontact
		transport	transport_query, transport_ticket, transport_traffic, transport_taxi
		lists	lists_query, lists_remove, lists_createoradd

Table 13: Fine to Coarse Labels Conversion Examples for Facebook and CLINC Dataset

Text	Predicted	True Label	Remarks about prediction
Find a store near Sia's place where I can buy champagne and find me a brunch spot in Lower Manhattan (SNIPS)	Location_Service (Dominant), App_Service (Non-dominant)	Location_Service, Location_Service	Non-Dominant Label predicted wrongly
Book a cab, is there traffic on the US 50 portion I'm going to take to go to my client meeting? (SNIPS)	App_Service (Dominant), Traffic_update (Non-Dominant)	Traffic_update, App_Service	Wrong Predictions - swapped ground-truth labels
What will the weather be like at my Airbnb this week end? Is there a parking at my hotel? (SNIPS)	GetWeather (Dominant), Location_Service (Non-Dominant)	GetWeather, Location_Service	Correct Predictions
Can you make a reservation at a lebanese restaurant nearby, for lunch, party of 5? How's the traffic from here? (SNIPS)	App_Service (Dominant), Traffic_update (Non-Dominant)	App_Service, Location_Service	Non-dominant label wrongly predicted
set alarm,remind me to pay electric monday (FACEBOOK)	set alarm (Dominant), set reminder (Non-Dominant)	set alarm, set reminder	Correct Predictions
is it going to snow in chicago tomorrow, any chance of rain today? (FACEBOOK)	weather find (Dominant), set reminder (Non-Dominant)	weather find, weather find	Non-dominant label wrongly predicted
how hot will it be, how long will it rain tomorrow (FACEBOOK)	weather find (Dominant), set reminder (Non-Dominant)	weather find, weather find	Non-dominant label wrongly predicted
what is the average wait for transfers, I'm still waiting on my identity verification.(BANKING)	General_Enquiry(Dominant), Identity_verification(Non-Dominant)	General_Enquiry, Identity_verification	Correct Predictions
My card is due to expire,Why can't I get cash out (BANKING)	card_about_to_expire(Dominant), declined_cash_withdrawal(Non-Dominant)	card_about_to_expire, declined_cash_withdrawal	Correct Predictions
I have a new email. I am in the EU. Can I get one of your cards? (BANKING)	Card_service_enquiry(Dominant), General_Enquiry(Non-Dominant)	Service_request, Card_service_enquiry	Incorrect Predictions; Predicted Dominant Intent is same as the Non-Dominant Ground Truth Label
Can other people top up my account? where did my funds come from? (BANKING)	verify_source_of_funds(Dominant), topping_up_by_card(Non-Dominant)	topping_up_by_card, verify_source_of_funds	Wrong Predictions - swapped ground-truth labels
Can you tell me my shopping list items, please? Is tomato on my shopping list? (CLINC)	shopping_list(Dominant), account(Non-Dominant)	shopping_list, shopping_list	Non-dominant label wrongly predicted
Change the name of your system. Your name from this point forward is george. (CLINC)	change_ai_name(Dominant), change_user_name(Non-Dominant)	change_ai_name, change_ai_name	Non-dominant label wrongly predicted
use my phone and connect please,tell me something that'll make me laugh(CLINC)	sync_device(Dominant), tell_joke(Non-Dominant)	sync_device, tell_joke	Correct Predictions
will there be traffic on the way to walmart,can you help me with a rental car(CLINC)	traffic(Dominant), car_rental(Non-Dominant)	traffic, car_rental	Correct Predictions

Table 14: Prediction of best-performing models and Respective Confidence