NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCES
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATION

PROGRAMME OF POSTGRADUATE STUDIES
LANGUAGE TECHNOLOGY

MASTER'S THESIS

Creation of a Dataset with utterances containing multiple intents including anaphora, cataphora & ellipsis

Vana I. Archonti

Supervisor: Themos Stafylakis, Associate Professor A.U.E.B (elected)

ATHENS
DECEMBER 2023

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΔΙΙΔΡΥΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΓΛΩΣΣΙΚΗ ΤΕΧΝΟΛΟΓΙΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Δημιουργία συνόλου δεδομένων με εκφωνήματα πολλαπλών προθέσεων που περιλαμβάνουν τα γλωσσολογικά φαινόμενα της αναφοράς, της καταφοράς και της έλλειψης

Βάνα Ι. Αρχοντή

Επιβλέπων: Θέμος Σταφυλάκης, Εκλεγμένος Αναπληρωτής Καθηγητής Ο.Π.Α

ΑΘΗΝΑ

ΔΕΚΕΜΒΡΙΟΣ 2023

MASTER’S THESIS

Creation of a Dataset with utterances containing multiple intents including the linguistic phenomena of anaphora, cataphora & ellipsis

Vana I. Archonti

A.M.: 7115182100002

SUPERVISOR: **Themos Stafylakis**, Associate Professor A.U.E.B (elected)

EXAMINATION COMMITTEE:

Themos Stafylakis, Associate Professor A.U.E.B (elected)

Stella Markantonatou, Research Director at ILSP/ ATHENA R.C

Athanasios Katsamanis, Principal Researcher at ILSP/ ATHENA R.C

December 2023

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Δημιουργία ενός συνόλου δεδομένων με εκφωνήματα πολλαπλών προθέσεων που περιλαμβάνουν τα γλωσσολογικά φαινόμενα της αναφοράς, της καταφοράς και της έλλειψης

Βάνα Ι. Αρχοντή

Α.Μ.: 7115182100002

ΕΠΙΒΛΕΠΩΝ: **Θέμος Σταφυλάκης**, Εκλεγμένος Αναπληρωτής Καθηγητής Ο.Π.Α

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Θέμος Σταφυλάκης,** Εκλεγμένος Αναπληρωτής Καθηγητής Ο.Π.Α

Στέλλα Μαρκαντωνάτου, Διευθύντρια Ερευνών ΙΕΛ / Ε.Κ ΑΘΗΝΑ

Αθανάσιος Κατσαμάνης, Κύριος Ερευνητής ΙΕΛ/ Ε.Κ ΑΘΗΝΑ

Δεκέμβριος 2023

ABSTRACT

Within the domain of TOD systems, intents are typically regarded as the fundamental units of recognition. In real-world applications, user utterances frequently include multiple intents, an aspect that is often ignored in most NLU datasets. Recent efforts to create such datasets, often contribute to the current trend of creating datasets containing single-intent utterances and tend to focus solely on the simple case of concatenating two single-intent utterances with conjunction. However, in real conversation scenarios, the two utterances may have the same referents or share common verbs and nouns, resulting in anaphoric, cataphoric, or elliptical constructions respectively. The primary objective of this thesis is to create a dataset consisting of multi-intent utterances that incorporate the linguistic phenomena of anaphora, cataphora and ellipsis. These utterances were created by deploying the pre-existing CLINC150 dataset. Regarding the construction of these anaphoric, cataphoric and elliptical structures, the English-GUM corpus was employed. However, the incorporation of these complex linguistic phenomena within the dataset necessitated the creation of the dataset through a manual process. The annotation process undertaken for this dataset was carried out by Canadian native speakers of English who volunteered their expertise as annotators during the dataset evaluation. Finally, two baseline experiments were carried out on the dataset: a multi-label learning technique treating double intents as an atomic label, and a threshold-based multi-label approach predicting single or double intents based only on single intents. The experimental results have indicated that the first approach exhibited positive outcomes compared to the threshold-based approach, which yielded less satisfactory results. Nevertheless, employing solely single intent labels for predicting both single and double intents could be a more effective strategy, especially considering its independence from double intents in the training set.

SUBJECT AREA: Natural Language Understanding (NLU)

KEYWORDS: multi-intent dataset, multi-intent classification, dialogue systems, cataphora, anaphora, ellipsis

ΠΕΡΙΛΗΨΗ

Στον τομέα των προσανατολισμένων διαλογικών συστημάτων, οι προθέσεις συνήθως αποτελούν τα κύρια συστατικά αναγνώρισης. Σε σενάρια του πραγματικού κόσμου, οι προτάσεις των χρηστών συχνά περιλαμβάνουν πολλαπλές προθέσεις, μια πτυχή που συχνά δεν λαμβάνεται υπόψη, με αποτέλεσμα να μην ενσωματώνεται στα περισσότερα σύνολα δεδομένων. Μάλιστα, πρόσφατες προσπάθειες κατασκευής τέτοιων συνόλων δεδομένων, ενισχύουν το κυρίαρχο σενάριο, δηλαδή αυτό της δημιουργίας συνόλου δεδομένων με προτάσεις ενός intent, είτε τείνουν να επικεντρώνονται αποκλειστικά στην απλή περίπτωση της παρατακτικής σύνδεσης δύο εκφωνημάτων μιας πρόθεσης με έναν σύνδεσμο. Ωστόσο, σε πραγματικά σενάρια συνομιλίας, τα δύο εκφωνήματα μπορεί να έχουν τα ίδια αντικείμενα αναφοράς, είτε να μοιράζονται κοινά ρήματα και ουσιαστικά, με αποτέλεσμα να δημιουργούνται αναφορικές, κατηφορικές ή και ελλειπτικές προτάσεις αντίστοιχα. Ο πρωταρχικός στόχος αυτής της διπλωματικής εργασίας είναι να δημιουργήσει ένα σύνολο δεδομένων με πολλαπλά intents που αποτελείται από προτάσεις που περιλαμβάνουν τα φαινόμενα της αναφοράς, της καταφοράς και της έλλειψης. Αυτές οι προτάσεις δημιουργήθηκαν αξιοποιώντας το ήδη υπάρχον σύνολο δεδομένων CLINC150. Επιπλέον, για την κατασκευή των αναφορικών, καταφορικών και ελλειπτικών εκφωνημάτων αξιοποιήθηκε το Σώμα Κειμένων English-Gum. Η ενσωμάτωση, ωστόσο, αυτών των σύνθετων γλωσσικά φαινομένων μέσα στο σύνολο δεδομένων κατέστησε αναγκαία τη δημιουργία του συνόλου δεδομένων χειροκίνητα. Για την αξιολόγηση, λοιπόν, του συνόλου δεδομένων ακολούθησε διαδικασία επισημείωσης, η οποία πραγματοποιήθηκε από Καναδούς φυσικούς ομιλητές της αγγλικής γλώσσας, οι οποίοι προσέφεραν εθελοντικά την γνώση τους ως φυσικοί ομιλητές στην αξιολόγηση μέρους των προτάσεων του συνόλου δεδομένων. Τέλος, πραγματοποιήθηκαν δύο πειράματα ακολουθώντας δύο βασικές προσεγγίσεις αντιμετώπισης πολλαπλών προθέσεων : μια τεχνική μάθησης πολλαπλών κατηγοριών που αντιμετώπιζε τις διπλές προθέσεις ως μια ενιαία οντότητα και μια μέθοδος ταξινόμησης πολλαπλών κατηγοριών βάση ενός κατωφλίου προβλέποντας μονές ή διπλές προθέσεις βασιζόμενη αποκλειστικά στις μονές προθέσεις. Αν και τα πειραματικά αποτελέσματα της πρώτης μεθόδου έδειξαν θετικά αποτελέσματα, συγκριτικά με την μέθοδο ταξινόμησης πολλαπλών κατηγοριών βάση ενός κατωφλιού, εντούτοις, η μεθοδολογία αξιοποίησης αποκλειστικά των μονών προθέσεων για την πρόβλεψη ταυτόχρονα μονών και διπλών προθέσεων μπορεί να αποβεί πιο αποτελεσματική, ειδικά δεδομένης της ανεξαρτησίας τους από τις διπλές προθέσεις στην χρήση τους στο σύνολο εκπαίδευσης.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Κατανόηση Φυσικής Γλώσσας

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ : Σύνολο δεδομένων πολλαπλών προθέσεων, Ταξινόμηση πολλαπλών προθέσεων, διαλογικά συστήματα, αναφορά, καταφορά, έλλειψη

ACKNOWLEDGMENTS

I would like to express my deep appreciation to Dr. Stafylakis, my supervisor, for his unwavering support and patient guidance throughout the entire process of developing my master's thesis. His mentorship and expertise have shaped my research skills and enabled me to complete my thesis. Additionally, I am sincerely thankful to Dr. Stella Markantonatou for her valuable insights and constructive suggestions that significantly enriched this work. Furthermore, I could not have undertaken this endeavor without my closest friends and colleagues, who have always been by my side in every way. Finally, I would also like to take this opportunity to express my sincere gratitude to the committed volunteers who selflessly devoted their time to our annotation process. Thank you for your support and dedication.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Without doubt, TOD systems have been widely deployed, as they enable users to interact with computer applications through conversations to achieve specific, well-defined tasks. ID recognizing the user's intent from their input, is a critical component of TOD systems. Thus, human-computer interaction has gained increasing attention due to its alluring commercial values and potential. In any dialogue system, the first step includes NLU. This component is responsible for parsing the user utterance into predefined semantic slots. Currently, state-of-the-art NLU systems excel in converting a user utterance into one single dialogue act.

However, in real-time conversations often a human-to-human conversation includes sentences that contain more than one intent, as humans possess the capability to comprehend and respond to such multi-intentional sentences, facilitating more fluid and natural dialogues, as opposed to the need to articulate each intention separately. In such cases, TOD systems must be able to handle such scenarios [1]. In conventional dialogue systems, it is usually assumed that each sentence contains only one intent. Users may need to wait for the system to finish processing one intent before moving on to the next. Such an assumption can restrict the flow of information and result in an unnatural conversational experience that hampers task completion applications [2]. Despite impressive performance and widespread use in commercial systems of intent detection models, current intent detection models fall short in capturing the complexity of human interactions, limiting their suitability for complex industry applications. In this way, current TOD systems are unable to capture real-time conversation features, which can include multiple intents within an utterance.

While the NLU methodology is advancing at a remarkable speed, it is important to note that the progress in constructing NLU datasets has not matched this rapid pace regarding modern production requirements [3]. This is partly due to constraints in existing training datasets primarily centered on single intent [4]. Such setups are not realistic focusing on simple single-label ID and lead to unnecessarily large intent sets. Although there has been some initial work on multi-label ID on MixATIS and MixSNIPS as multi-intent datasets; however, their multi-label examples are limited in the creation of multi-intent utterances through conjunction. Furthermore, there have been some efforts to create synthetic datasets through concatenation [5]. These datasets fall short of effectively capturing real-life situations where multiple intents can emerge within utterances, potentially leading to the recurrence or the absence of the same nouns or verbs, resulting in the creation of elliptical, anaphoric or cataphoric constructions.

As we move on this investigation into TOD systems, it is essential to recognize that the complexity of human interactions within such systems often involves linguistic phenomena that support modern needs. In the world of NLP, particularly within dialogues involving multiple intents, three fundamental linguistic constructs play a pivotal role: anaphora, cataphora, and ellipsis. These phenomena are essential components of natural language understanding and pose distinct challenges when dealing with multi-intent utterances.

Anaphora occurs in a sentence after the referent has been introduced, while in cases of backward anaphora or cataphora, a pronoun is used before the referent has been introduced. Although anaphora is a famous linguistic phenomenon in NLP, cataphora is a relatively rare phenomenon, as it has specific restrictions. Elliptical utterances are an integral part of information-seeking dialogue. Ellipsis occurs pervasively in natural language, especially in conversational settings and world languages use some or the other form of eliding redundant information, making this phenomenon universal and extremely important for linguistic research.

In the field of NLU, it becomes evident that contemporary research in the field confronts significant gaps and limitations. These gaps revolve around the inadequacies of existing NLU datasets to meet the evolving demands of industry and the ever-increasing complexity of human-computer interactions. Presently, NLU datasets primarily rely on crowd workers for data collection and annotation, leading to issues of limited lexical diversity and frequent annotation errors. Moreover, the conventional design of Intent Detection (ID) datasets assumes a single intent per sentence, a limitation that does not align with the intricacies of modern production requirements, where multi-intent utterances are commonplace [3].

The aim of the present master thesis is twofold: first, to bridge the existing gap in multi-intent datasets, and second, to incorporate three linguistic phenomena within multi-intent utterances. This overarching goal is to tackle the prevailing challenges presented by multi-intent utterances in NLU systems, ultimately driving forward the state-of-the-art in the field.

1.1 Thesis Structure

The research in this thesis is described according to the following chapters: Chapter 1 constitutes the introduction to the main topic and emphasizes the reasons for conducting this type of research. Chapter 2 mentions related work done by different researchers and the various definitions and approaches of the linguistic phenomena of anaphora, cataphora, and ellipsis, along with the first attempts at creating multi-intent datasets. Chapter 3 focuses on the description of the procedure that was followed to create the dataset. Chapter 4 describes the two experiments that have been carried out along with the experimental results. Finally, chapter 5 presents the experimental results and 6 includes the conclusion and the future work.

2. LITERATURE REVIEW

Dialogue-based systems have become increasingly widespread, expanding their conversational abilities from open-ended discussions to task-oriented settings. While open-ended dialogue systems aim to engage the user in a conversation, task-oriented dialogue systems prioritize accomplishing specific tasks as articulated by the user through written or spoken natural language utterances [6]. Natural Language Understanding (NLU), a pivotal component within TOD systems, has been a subject of scholarly exploration since the early 1990s. The inception of academic inquiries in this domain coincided with the onset of the ATIS project [7]. NLU comprises two fundamental elements: intent detection and slot-filling. Given the close relationship between intent detection and slot filling, recent research tends to approach these two tasks jointly [3], [8] – [11], taking into consideration the correlation between them. While recent years have witnessed a surge in datasets for developing and evaluating intent classification and slot-filling models in task-oriented dialogue systems, it's worth noting that surveys of dialogue systems tend to prioritize discussing models while giving less attention to datasets, as is observed in [2], [12], [13].

Furthermore, several studies in the research community have argued that contemporary NLU datasets primarily focus on single-intent utterances [3],[14],[15], noticing a lack of multi-intent datasets. [16],[17]. However, a comprehensive dataset should aim to faithfully replicate the complexities of human interactions, encompassing multiple intents within a single utterance. Traditional datasets do not necessarily align with current industry demands, as they fall short in addressing multi-intent detection. In recent research [18] it is underlined that modern NLU methods may also face challenges when dealing with utterances containing ellipsis and anaphora, despite these phenomena being common in natural conversations. This challenge arises from the frequent omission or reference to elements within ongoing dialogues, often traced back to the conversation history and pushes the need for specialized approaches to handle this problem.

Specifically, the issue of multi-utterances containing anaphoric elements has been indirectly referenced in [2]. This discussion refers to anaphoric sentences like "find avatar and play it," where processing these sentences separately could present obstacles to achieving a comprehensive understanding of their semantic meaning. Moreover, this issue is also addressed in NLU++ referring to intents with anaphoric elements as partial intents, giving the example “The savings one”. Regarding the incorporation of multiple intents within a single utterance, the emergence of multi-intent language understanding has evolved into a critical area of scholarly investigation within the dialogue systems domain. As a result, many researchers have shifted their focus to creating multi-intent datasets instead of solely presenting different approaches for multi-intent detection.

2.1 A REVIEW OF MULTI-INTENT DATASETS

A recent survey of intent classification methodologies [19] has documented four multi-intent corpora TOP, MixATIS, MixSNIPS and NLU++. However, a more careful look can show that the definition of multi-intent utterances is ubiquitous. The 'TOP' dataset, introduced in [20], has garnered attention due to its multi-slot annotation, where approximately 35% of its utterances exhibit multiple intents. Notably, TOP's multi-intent utterances follow a distinct structure where the nested intent is a direct child of a slot annotation, revealing intriguing patterns of intent co-occurrence. 'MixATIS' and 'MixSnips,' as presented in [16], provide valuable insights into multi-intent queries artificially constructed through conjunction words like “and", "," (comma), "and also", "and then". Additionally, 'NLU++,' introduced in [21], stands as a modular multi-intent dataset, likely affording researchers greater flexibility in modelling complex multi-intent utterances. In this context, it's important to refer to MULTI3NLU++ as well as the latest dataset including multi-intent utterances presented in [14], primarily known for its multi-intent nature and, more notably, its unparalleled emphasis on multilingualism.

While both MixSnips and MixATIS include multiple intents within datasets, their capabilities are limited by the source corpora, ATIS, and Snips, as well as the somewhat restricted use of conjunctions for connecting queries. Moreover, the random concatenation of two utterances from a dataset like Snips may not authentically respond to the way users typically interact with dialogue systems. When examining multi-intent utterances, their lack of naturalness can be attributed to the combination of intents from diverse domains, such as inquiring about the weather and rating a book, and the artificial way they are constructed. The concatenation of unrelated intents like "get_weather" and "rate_book" within a single utterance using random selection and linking words like "and then" or "and also" fails to capture the natural flow of conversational dialogues, presenting a limited perspective of reality. In [10] the issue of intent overlap is addressed, and tools for identifying intent categories with semantic overlaps are introduced in [22]. The NLU++ dataset is introduced as a multi-intent dataset, which includes utterances that have been annotated with multiple intent modules, thereby expanding the scope of intent classification. This research asserts that this ontology facilitates the representation of intricate combinations of intents. However, it's important to provide clarity when defining multi-intent utterances. In that study, some illustrative examples are as follows: For instance, an utterance might be "Make it higher," which includes both the "change" and "higher" intents. Similarly, another example could be "Cancel it," indicating the presence of only one intent, the "cancel" intent. Another research [15] supports that multi-intent detection mainly deals with short text without explaining it further. In the context of multi-intent detection, it is imperative to carefully define multi-intent utterances. This is crucial for effectively handling the task of precisely recognizing multiple intents present in users' utterances. As a result, there is a need for additional research within the field of linguistics. This is particularly important because, despite the presence of utterances with anaphoric elements in the examples provided in the aforementioned research, there is a lack of further analysis or discussion regarding the linguistic phenomenon of anaphora and its potential impact on the intent space.

2.2 Linguistic Phenomena on DIALOGUES

As a central problem, ellipsis and anaphora frequently occur in human conversations, creating additional challenges for dialogue understanding. During real-time conversations, people often use pronouns or demonstrative phrases to refer to previously mentioned or forthcoming subjects. Furthermore, humans also use elliptical structures for reasons of economy, style or even emphasis to express their intention in a dialogue. After examining existing multi-intent datasets in the preceding section, it is essential to define the linguistic phenomena of anaphora, cataphora and ellipsis that can occur in multi-intent utterances. Anaphora and ellipsis have undoubtedly gained attention due to the recognition of coreference and ellipsis resolution as crucial tasks of NLP, particularly because they frequently occur in dialogues. [23] has pointed out the need to annotate the linguistic phenomena of anaphora and ellipsis in dialogues, as popular datasets do not adequately reflect the actual performance of dialogue systems in real conversations. Many times, linguistic terms are used interchangeably or ambiguously for the sake of simplicity, especially in the field of NLP. In the present thesis, these terminologies are put into linguistic context.

2.2.1 Anaphora and antecedent

Defining anaphora is more challenging than one would expect. Typically, anaphora is defined as referring to items mentioned previously in discourse, as presented in [24]. Various types of anaphora are recognized, with pronominal anaphora being the most dominant in natural language. The definition proposed in [25], namely that “Anaphora is used most commonly in theoretical linguistics to denote any case where two nominal expressions are assigned the same referential value or range.” includes phenomena not classified as anaphora. Essentially anaphora is defined as the reference within the text itself and typically to entities already mentioned in the text. At the same time, coreference looks at references that connect entities, even if they have different linguistic forms, and often involves entities that exist outside of the text. Overall, it seems that coreference is a broader concept that goes beyond the immediate context of the text to consider how entities are connected in the real world or discourse. There are many different types of anaphora, but we will refer only to the types that we encounter in our datasets. These are pronominal anaphora and one-anaphora. Pronominal anaphora stands out as one of the most dominant types of anaphora and represents the most frequent type of anaphora in web dialogues. Pronominal anaphora is the most common phenomenon in which the pronouns are substituted with previously mentioned entities. This type of anaphora can be further divided into four subclasses, namely. Nominative: {he, she, it, they} Reflexive: {himself, herself, itself, themselves} Possessive: {his, her, its, their} Objective: {him, her, it, them}. One -anaphora, in terms of X-bar theory, the pro-form *one* is generally characterized as a substitute for an N/ constituent [26]. Although recent studies expand this theory and note that anaphoric "one" does not necessarily have to represent an N/constituent. Instead, it can also serve as a reference for various linguistic and non-linguistic elements, including a standalone noun, a multi-word noun phrase that is separate from its complement, a phrase with non-adjacent components, a component within a compound word, or even an entity within the broader non-linguistic context beyond just noun constituents [27]. Below are presented Anaphoric examples of the English-Gum Corpus and the new dataset presented below to explain the types of anaphora included in the dataset. In example (1) the anaphor ***he*** refers to the antecedent ***Daniel*** and the type of anaphora that is detected is called pronominal anaphora. Example (2) includes one-anaphora, where the anaphoric one substitutes the earlier mentioned noun ***song***.

(1) However, **Daniel** refused, because **he** wanted to study mathematics. (**pronominal anaphora**)

(2) What's the current **song** that we are listening to at this time, and could you also skip to the next **one**? (**one-anaphora**)

2.2.2 Cataphora and postcedent

Cataphora is used to describe the phenomenon where a pronoun is used before the referent has been introduced. Contrary to the phenomenon of anaphora, cataphora is a relatively rare phenomenon and there are conflicts in research when it is applied or not. In [28] it is supported that cataphora primarily occurs in subordinate clauses, but various approaches have been taken up to this point, leading to issues in characterizing the problem. The examples (4) and (5) represent two types of cataphora. In example (4) of English-Gum Corpus the postcedent "***its***" refers forward to the cataphoric "***the device***". Moreover, based on example (5) it appears that the English-Gum Corpus involves one more structure marked as cataphoric where the *"**dummy it"*** refers forward to the infinitive "***to define*"** and is annotated as a cataphoric structure. This structure is linked to inherent characteristics of the English language.

(3) By creating a Geofence with an erroneous location from **its** central location, **the device** will receive incorrect location notifications.

(4) For many scientific fields, however, there is no central listing of all tenure-track faculty, making **it** difficult **to define** a rigorous sample frame for analysis.

2.2.3 Ellipsis

Ellipsis along with its various types represents another widespread phenomenon that we encounter in dialogues [29]. Ellipsis is a linguistic phenomenon in which words or phrases of the sentence are omitted when they are redundant or have been previously referenced or mentioned. In all cases, context plays a vital role in the ellipsis resolution, as it seems that it depends on the contextual text for the recovery of the meaning [30]. It is classified into clausal, predicate, and nominal ellipsis, respectively [31]. Nominal ellipsis is about the deletion of the noun or the pronoun [32]. It is also called head noun ellipsis or Noun Phrase Ellipsis (NPE) [33]. A predicate ellipsis occurs when the main predicate of the clause is missing, often along with some of its internal arguments [34]. This type of ellipsis includes the pseudogapping, the Post Auxiliary Ellipsis (PAE) and the Verb Phrase Ellipsis (VPE). In this case, the ellipsis is permitted by a modal or auxiliary verb as in example (6) [35]. Finally, clausal ellipsis refers to the type of ellipsis where the whole clause is deleted as in example (8). In example (5) the phenomenon of VPE ellipsis is used as the main verb **move** is omitted to avoid the repetition in the second part of the sentence for the sake of economy. In example (6) the Post auxiliary Ellipsis (PAE) is illustrated.

(5) **Move** the finger backwards to about 4 inches (10 cm) away, then **[move]** back again. (**VPE ellipsis**)

(6) John will eat candy and Bill will do __, too. (**Modal Complement Ellipsis**)

(7) Matthew’s sweet tea is comparable to Granny’s **[sweet tea].** (**Nominal Ellipsis)**

(8) **We have a physics assignment due**, but I don’t remember where **(Clausal Ellipsis)**

2.3 Unresolved issues in NLU datasets

In recent years the above linguistic phenomena have gained attention in NLP, as it seems that both ellipsis and coreference could affect the performance of dialogue systems as addressed in [36], failing to understand such tricky utterances correctly [37]. It is well established that without resolving these phenomena, dialogue may fail to generate coherent responses [38]. At the same time, in the context of multi-intent detection, despite significant progress, certain gaps in the research still exist. Existing NLU datasets, despite the growing industry demand, have not kept pace with recent trends and continue to exhibit unresolved issues. These issues include the prevalence of domain-specific datasets, lack of lexical diversity due to the use of crowdsourcing workers which typically paraphrase, as well as synthetic datasets which do not illustrate natural dialogues. While multi-intent detection methodologies and ellipsis and coreference resolution have been addressed in recent years, there has been a lack of well-structured datasets that comprehensively handle these issues. Within the scope of multi-intent detection, three crucial areas require our attention. First, there is a need to refine the definition of multi-intent utterances. It is also essential to investigate strategies for studying multi-intent detection in conjunction with linguistic phenomena such as anaphora, cataphora, and ellipsis.

3. Dataset Creation

3.1 Data Collection

The present dataset was derived from the pre-existing CLINC150 dataset, which functioned as the primary source and it was initially introduced in [39]. This dataset serves as an evaluation benchmark for IC and out-of-scope prediction. The in-scope queries cover 150 distinct intent classes across 10 different domains: work, meta, banking, auto_and_commute, home, travel, utility, kitchen_and_dining, small_talk and credit_cards. The original dataset comprises 23,700 queries, of which 22,500 are within the scope of inquiry and 1,200 are out of scope. According to [2], in-scope queries are user-generated questions, and commands related to specific topic domains, collected through crowdsourcing, and used for training and evaluating task-driven dialogue systems In contrast, out-of-scope queries, neither align with any of the 150 intents nor can they be classified using scoping and scenario tasks rooted in topics from platforms such as Quora and Wikipedia. There are several variants of the CLINC150 dataset, including small, imbalanced, and OOS+.

For the development of our dataset, we selected the CLINC150 dataset because it is an IC dataset, offering a wide range of single-intent utterances of different domains and intents that closely resemble human-system interactions. We utilize the full version of the CLINC150 dataset and focus solely on in-scope examples, excluding the out-of-scope samples. Therefore, our experiments are conducted on the subset of the full dataset consisting of 22,500 in-scope utterances. Each intent category includes 100 training, 20 validation, and 30 testing in-scope queries. For the creation of our dataset, we utilize the training, validation and testing sets, which consist of 15,000, 4,500 queries and 3,000 queries respectively.

3.2 Dataset Expansion

The new dataset[1] is an expansion of the CLINC150 dataset, including single-intent and double-intent utterances. It comprises 85 single-intent categories and 65 double-intent categories. Single-intent categories determine the double-intent classes, as the 65 double-intent classes are essentially different combinations of the 85 single-intent utterances whenever feasible. Regarding the domains, it includes 18 domains, of which 10 are from the source dataset, and the remaining 8 are combined domains. We believe that combining intents for creating double intents should involve overlapping intents.

The intent ontology of the dataset will be described in more detail in the respective unit. Table 1 below illustrates the single intents available in our dataset along with examples.

[1] Our new dataset can be accessed on GitHub through the following link: https://github.com/vanaarxonti/Multi-intent-Dataset-including-Anaphora-Cataphora-and-Ellipsis

Table 1: Presentation of the SI that are available in our dataset, along with their respective domains and example utterances.

Domain	Intent	Utterances
auto_and_commute	tire_pressure	What is the air pressure in my tires?
credit_cards	report_lost_card	My visa card was stolen.
banking	rewards_balance	What is the reward balance on my Discover card?
utility	time	What time is it right now in Adelaide, Australia?
meta	sync_device	Could you connect with my phone?
travel	lost_luggage	I think my luggage got lost.
utility	measurement_conversion	Convert cm to inch.
credit_cards	credit_limit_change	Can I get my credit limit increased to $15,000?
home	what_song	Tell me the name of the song that is playing.
home	reminder	What did I put on my list of reminders?
small_talk	what_is_your_name	What's your full name?
auto_and_commute	current_location	Tell me how to find my current location.
auto_and_commute	gas	Is there enough fuel to make it to Walmart?
home	shopping_list	Display shopping list.
credit_cards	application_status	Can you check and see if my credit card application has been approved or not?
banking	bill_due	When do I have to pay my electric bill?
travel	car_rental	I need a rental car.
auto_and_commute	last_maintenance	Find out when my most recent oil change occurred.
small_talk	who_made_you	Who was your creator?
work	pto_used	Look up my total number of days off so far.
credit_cards	pin_change	Change my PIN for my checking account.
home	order	I need some more Lysol, could you order me some?
utility	find_phone	Is my phone in the house?
travel	travel_alert	What are the latest travel alerts for Dubai?
travel	travel_suggestion	Tell me some things to see in Tampa.
kitchen_and_dining	food_last	When will the milk expire?
credit_cards	damaged_card	Report the card has been damaged.
kitchen_and_dining	recipe	What do you need to do to make sushi?
small_talk	where_are_you_from	Where's home for you?

work	schedule_meeting	Schedule a meeting with Tom for 6 pm.
banking	report_fraud	Get rid of everything on my calendar for March 2nd.
home	calendar_update	Delete the hair appointment I had scheduled on May 1st, please.
home	next_song	Play the next musical number.
credit_cards	card_declined	I wish to know why my card was declined.
travel	exchange_rate	The exchange rate between Mexico and the U.S.
auto_and_commute	oil_change_when	How often should the oil be changed?
credit_cards	expiration_date	Do you know the expiration date for my visa card?
banking	freeze_account	Is it too much trouble to put a stop to my bank account?
banking	min_payment	I need to know my cable bill minimum payment.
credit_cards	pay_bill	I want to pay my tax bill.
meta	change_ai_name	Change the name of your system.
kitchen_and_dining	cook_time	How long should I cook the asparagus?
work	payday	What day do I get paid?
auto_and_commute	mpg	Do you know what this car's mpg is?
home	todo_list	Is 'cleaning the toilet' on my to-do list?
auto_and_commute	schedule_maintenance	What do I have to do tomorrow, according to my to-do list?
home	calendar	What's happening on May 3rd?
banking	balance	What is my bank balance for all accounts?
auto_and_commute	oil_change_how	How do you change the car oil?
work	income	How much do I make per day?
small_talk	tell_joke	What's the funniest thing you know about artificial intelligence?
home	todo_list_update	Put 'laundry' on my to-do list.
utility	spelling	How do I spell 'catheter'?
utility	text	Compose a text message.
small_talk	who_do_you_work_for	Do you know who you report to?
small_talk	how_old_are_you	What age is the AI?
small_talk	fun_fact	I want to learn a neat fact about black holes.
travel	book_flight	Look for a flight out of LA to Chicago on March 3rd for under $500.
credit_cards	international_fees	Do I incur extra fees if I use my card in London?

home	smart_home	Turn on the TV
banking	bill_balance	How do I look up my credit score?
credit_cards	credit_score	I wish to know my credit rating.
utility	date	Tell me what the date is today.
work	meeting_schedule	Is the gang getting together this afternoon?
auto_and_commute	jump_start	How do you jump-start a Subaru Forester?
utility	share_location	Can you forward my location to Tom?
travel	timezone	What timezone is Los Angeles in?
home	order_status	I need to track my package.
utility	make_call	Can I call a restaurant?
work	taxes	How much will I have to pay in Colorado taxes?
credit_cards	replacement_card_duration	What's the wait time for a replacement card?
credit_cards	improve_credit_score	How to protect my credit score?
meta	change_language	Adjust your language setting to English.
auto_and_commute	gas_type	What type of gas does my car use?
credit_cards	credit_limit	What is the credit limit for my Chase card?
credit_cards	apr	Hey Siri, tell me the APR on my Disney visa.
meta	change_user_name	People call me Gary.
work	next_holiday	I need to know when the next holiday is.
credit_cards	transactions	Help me get access to my recent transaction history.
utility	alarm	Wake me up at 6 am.
kitchen_and_dining	calories	What's the calorie count for tuna casserole?
kitchen_and_dining	meal_suggestion	Give me Italian meal ideas.
home	timer	Timer for 5 minutes.
meta	change_accent	Do a British male accent only.
travel	translate	Translate 'hello' in French.

The new dataset is an expansion, as the utterances from the original CLINC150 dataset were expanded with linguistic phenomena, occurring in double-intent utterances, which we call *utterances*, rather than *sentences* because we view them as acts of speech. The dataset including training, testing and validation set, comprises approximately 1.875 utterances, of which 910 exemplify the linguistic phenomena of anaphora, cataphora and ellipsis. Specifically, the training set consists of 975 double-intent utterances and 340 single-intent ones. Each of the test and validation sets includes 195 double-intent utterances and 85 single-intent utterances. Table 2 below illustrates the 65 intent classes along with examples of their respective utterances. The majority of the following utterances depict instances of double-intent uttrances with linguistic phenomena, annotated in this manner for illustrative purposes.

Table 2: Presentation of the 65 DI categories of our Dataset along with examples of their respective intents.

Domain	Double Intent Classes	Utterances
credit_cards	apr, credit_limit	I'd like to know the APR on my Visa card and its credit limit. [anaphora]
credit_cards	apr, credit_limit_change	Do you know my credit card's APR, as well as a way to get a higher limit on my Bank of America card? [ellipsis]
banking,credit_cards	balance, credit_score	Check my bank balance and my credit score. [ellipsis]
banking	bill_balance, min_payment	Please give me the outstanding balance on my water bill and the minimum amount I can pay on my phone bill [ellipsis]
banking	bill_due, pay_bill	When is it due? I want to pay my rent now [cataphora].
home	calendar, reminder	Please give me an overview of my calendar list and a rundown of my reminder list. [ellipsis]
travel	car_rental, book_flight	Book me a car rental in Dallas and I want to fly from it to Phoenix. [anaphora]
credit_cards	card_declined, international_fees	Let me know why my card was declined the other day and if is there a charge to use it in Japan. [anaphora]
credit_cards	card_declined, replacement_card_duration	Why was my card not accepted? How long will it take to replace it if it was stolen? [cataphora]
meta	change_language,change_accent	How can I adjust the spoken language, as well as the accent? [ellipsis]
meta	change_user_name,what_is_your_name	My name is Stu, not Sue! Tell me yours now [ellipsis].
kitchen_and_dining	cook_time, food_last	How long should I cook the frozen pizza? Also, I'd like to know how long I can keep it in the freezer before it goes bad. [anaphora]
credit_cards	credit_limit, credit_limit_change	Give me my credit limit, and then increase it to $1000. [anaphora]
banking,credit_cards	credit_limit_change, pin_change	Tell me the limit on my Amex card and set its PIN to 1234. [anaphora]

credit_cards	credit_score, improve_credit_score	Provide me with my credit score and with ways to build it. [ellipsis]
auto_and_commute,utility	current_location, share_location	Please tell me my location using GPS, then Darren and Stacey. [ellipsis]
credit_cards	damaged_card, replacement_card_duration	My card doesn't function anymore and when is the earliest date, I can get my new one? . [anaphora]
utility,auto_and_commute	date, current_location	I need to know today's date, as well as my location. [ellipsis]
credit_cards	expiration_date, application_status	On it, I mean my credit card, what's the expiration date? Plus, could you check and find out the current status of my credit card application? [cataphora]
credit_cards	expiration_date, replacement_card_duration	Can you check when my visa card expires, as well as when my replacement card will be mailed? [ellipsis]
utility,auto_and_commute	find_phone, current_location	I need to find my phone and my location. [ellipsis]
utility	find_phone, make_call	I lost it, and I need help finding my phone. Also, can you make a phone call to Dave? [cataphora]
utility, meta	find_phone, sync_device	Since I don't remember where it is, could you find my phone? Also, connect with it. [cataphora]
auto_and_commute	gas, gas_type	Please check the amount of gas I have, then the type of fuel to use with this car. [ellipsis]
small_talk	how_old_are_you, where_are_you_from	Tell me the amount of gas I have, and the type of gas this car uses. [ellipsis]
work	income, taxes	Please give me my salary figure, as well as the specifics of my federal taxes. [ellipsis]
auto_and_commute	last_maintenance, gas	Find the date of the last oil swap for my car and my current gas level. [ellipsis]
auto_and_commute	last_maintenance, gas_type	When was the last time I serviced my car and what type of fuel does it use? [anaphora]
auto_and_commute	last_maintenance, jump_start	When did I take my car to the mechanic, cause it's dead. Tell me how to jump-start it. [anaphora]
auto_and_commute	last_maintenance, mpg	Please tell me the last time I took my car to the shop. What's its mpg, too? [anaphora]
auto_and_commute	last_maintenance, schedule_maintenance	When did I last have it in the shop? My car needs fixing, so I want to schedule a car maintenance. [cataphora]

auto_and_commute	last_maintenance, tire_pressure	When was my car last looked at? Also, could you please check its tire pressure? [anaphora]
utility, travel	lost_luggage, find_phone	Help me find my luggage and my phone. [ellipsis]
travel, banking	lost_luggage,pin_change	I think my luggage has been lost, as well as my pin for my retirement account. [ellipsis]
utility	make_call,share_location	Please call Christie's number and inform the coach of my location.
utility, travel	measurement_conversion,exchange_rate	I want to convert feet to inches, as well as dollars to francs. [ellipsis]
auto_and_commute	mpg, oil_change_when	How much does it take to fuel this car and when should I change the oil in my car? [cataphora]
auto_and_commute	oil_change_how, oil_change_when	I need a manual on how to change the oil and how often I need to. [ellipsis]
home, utility	order_status, find_phone	Would you track my package and my phone? [ellipsis]
work	payday, pto_used	How many vacation days have I used up and how many until my next payday? [ellipsis]
work	pto_used, next_holiday	Access my job portal and inform me of the number of days I've taken off, as well as whether there is a holiday scheduled for next week. [ellipsis]
kitchen_and_dining	recipe, calories	Could you find me a recipe for sugar cookies, and also provide information about the number of calories in them? [anaphora]
kitchen_and_dining	recipe, meal_suggestion	I need you to find me a recipe for fried shrimp, and a meal suggestion for french dinner. [ellipsis]
banking,credit_cards	report_fraud, damaged_card	I'm reporting fraudulent activity on my card, and I also cracked it. Could you order a new one? [anaphora]
banking	report_fraud, freeze_account	There is a fraudulent charge for PayPal on it, so I would like to place a hold on my bank account immediately. [cataphora]
credit_cards	report_lost_card, replacement_card_duration	I need to report that my card is lost and inquire about the transit time to replace it. [anaphora]
banking,credit_cards	rewards_balance, transactions	How many reward points have stacked up for my Amex card. I would like to hear about its latest transactions, too. [anaphora]
work	schedule_meeting, meeting_schedule	Delete everything from the task list, and corn from my shopping list.

home	shopping_list, order	What are the contents of my shopping list? I want to order everything on it. [anaphora]
home, utility	smart_home, alarm	Can you see if I have my doors locked, as well as the alarm set? [ellipsis]
small_talk	tell_joke, fun_fact	Tell me something funny about cats and what would be a fun fact about them. [anaphora]
home, utility	text, calendar_update	Can you text him that I'm on my way before Christopher gets on his nerves? Also, please add the meeting with Carla to my schedule for July 4th. [anaphora]
utility	text,share_location	Could you send a text to Marty and say 'I am running behind' and let him know where I am located? [anaphora]
utility, travel	time, timezone	What time is it in Russia and what timezone is Boise in? [cataphora]
utility	timer, alarm	Set a 10-minute timer, then a 5 AM alarm, please. [ellipsis]
home	todo_list, todo_list_update	Read to me my to-do list and add the chore of 'vacuuming' to it. [anaphora]
home	todo_list_update, shopping_list	Add laundry to my to-do list, and cheerios to the grocery list. [ellipsis]
banking,credit_cards	transactions, apr	Show me my grocery transactions, and the APR on my MasterCard. [anaphora]
travel, meta	translate, change_language	How do you say "I need coffee" in Dutch, also switch your language setting to it. [anaphora]
utility, travel	translate, spelling	How do they say 'Where's the bathroom' in Spanish and find all the 'a's in it. [anaphora]
travel	travel_alert, timezone	Is it safe to travel to Norway, and what timezone is Detroit in? [cataphora]
travel	travel_suggestion, travel_alert	Give me some tourist suggestions for Jalisco, as well as the travel alerts. [ellipsis]
small_talk,meta	what_is_your_name, change_ai_name	Tell me your name, because from now on it will be 'Lord Vader'. [anaphora]
home	what_song, next_song	I need to know what this song is called. Could you give me the next one, too? [anaphora]
small_talk	who_made_you, who_do_you_work_for	It is needed, the name of your creator. Also, who do you work for? [cataphora]

Regarding the methodology employed for the dataset creation for this master's thesis project, the largest part of the dataset was manually created because requires the creation of double-intent utterances incorporating linguistic phenomena. On the other hand, the process of creating double intent utterances without these phenomena, as well as the extraction of simple intent utterances from the respective train, test, and validation sets of the CLINC150 dataset, was performed initially within Python programming language because the application of complex language patterns was not required. After the automatic concatenation of the two utterances, a manual process was applied to produce more natural utterances. In this way, other structures were employed too, “Not only- but also, besides, furthermore, moreover” etc.

3.3 Data Preparation

First, data were extracted with specific algorithms in the Python programming language; the extracted data were split into training, testing, and validation sets. As mentioned before, the dataset was derived from the pre-existing CLINC150 dataset, which served as a major source. The subsequent step involved categorizing the data into their respective sets. Regarding the type of our data, although the initial idea was to include out-of-scope utterances, only in-scope queries were employed in the final version of the dataset.

3.4 Creation of the dataset incorporating anaphora, cataphora and ellipsis

After categorizing our data, we studied them to determine the intent utterances that could be best combined and which is the more appropriate way to incorporate anaphora, cataphora, and ellipsis. However, before proceeding with the creation of double-intent utterances including these phenomena, we studied the aforementioned linguistic structures. Next, we looked for corpora containing annotations of these linguistic features. The search led to the study of CorefUD, a multilingual collection of corpora, particularly of the English version known as the “English-GUM Corpus” [40] from Georgetown University. This decision was primarily based on its accessibility, as it offers open-source access without the need for specific licenses. Additionally, it contains both cataphora and anaphora phenomena. Furthermore, it includes and distinguishes eight types of anaphoric links, encompassing pronominal anaphora, cataphora, lexical and predicative coreference, apposition, discourse deixis, split antecedents, and bridging. Observational scrutiny and meticulous selection were required to integrate pre-existing utterances. The challenges encountered in incorporating the phenomena of cataphora, anaphora, and ellipsis necessitated the modification of some of the initial utterances of the CLINC150 dataset, rather than adopting the straightforward conjunction of two utterances, we adhered to the rules governing anaphora, cataphora and ellipsis. This is one of the reasons why the implementation of these phenomena is a conscientious process that cannot always be applied automatically.

3.5 Combining intent-utterances from different domains

3.5.1 INTENT ONTOLOGY

The new dataset is an expansion of the CLINC150 dataset, encompassing both single-intent and double-intent utterances. It comprises 85 single-intent categories and 65 double-intent categories. Single-intent categories determine the double-intent classes, as the 65 double-intent classes are essentially different combinations of the 85 single-intent utterances whenever feasible. Regarding the domains, it includes 18 domains, of which 10 are from the source dataset, and the remaining 8 are combined domains. We believe that combining intents for creating double intents should involve overlapping intents.

The intent ontology of the dataset will be described in more detail in the respective unit. Table 3 below illustrates the single intents available in our dataset along with examples.

After the categorization of our data and the study of the linguistic structures, we had to select the intent utterances that could be combined based on their respective domains. The common approach is to combine intent utterances within the same domain. However, slots of different domains and intents are not entirely discrete and are often dependent on each other [41]. As a result, we resolved to combine in-domain and out-of-domain intent classes that exhibited semantic similarity [41]. This decision relied on the observation of overlapping intents across certain domains, as also indicated in [22]. A representative example of combining intents from different domains included in the new dataset is: “translate” and “change_language” as illustrated in Table 2. Each of these intents belongs to separate domains, with the former belonging to the *travel* domain and the latter to the *meta* domain. In this circumstance, the combination of '*translate*' and *'change_language'* allows the address of multilingual needs in a conversational context, as users often switch between languages and require translation within the same conversation. The combined domains are presented in Table 3. In this way, our new dataset ends up with 18 domains, as represented in Table 4, of which 8 are the result of pre-existing domains combined. Moreover, Table 5 illustrates all the overlapping intents that have been combined.

Table 3: Domains of the CLINC150 dataset that were selected to be combined to effectively combine intents across different domains based on semantic similarity.

DOMAINS COMBINED
banking,credit_cards
utility, travel
travel, meta
utility, meta
small_talk, meta
auto_and_commute, utility
travel, banking
home, utility

Table 4: Domains in our Dataset.

DOMAINS in our Dataset
work
utility
banking
Credit_cards
travel
meta
auto_and_commute
Kitchen_and_dining
Small_talk
home
banking,credit_cards
travel, meta
utility, travel
utility, meta
small_talk, meta
auto_and_commute, utility
travel, banking
home, utility

Table 5: Combination of Overlapping Intents.

Combined Domains	Combined Intents
banking, credit_cards	❖ balance, credit_score ❖ transactions, apr ❖ credit_limit_change, pin_change ❖ rewards_balance, transactions ❖ report_fraud, damaged_card
travel, meta	❖ translate, change_language
utility, meta	❖ find_phone, sync_device
small_talk, meta	❖ what_is_your_name, change_ai_name
auto_and_commute, utility	❖ current_location, share_location ❖ find_phone, current_location ❖ date, current_location
travel, banking	❖ lost_luggage, pin_change
home, utility	❖ order_status,find_phone ❖ text,calendar_update ❖ smart_home,alarm
utility, travel	❖ lost_luggage,find_phone ❖ translate, spelling ❖ time, timezone ❖ measurement_conversion, exchange_rate

3.6 Creation of Double intent utterances without linguistic phenomena and single intent utterances

The process of creating both double intent utterances without the linguistic phenomena of anaphora, cataphora and ellipsis, as well as the extraction of simple intent utterances from the respective train, test, and validation sets of the CLINC150 dataset, was performed within Python programming language. For the simple case of the combination of double-intent utterances without incorporating the linguistic phenomena, the following process was adopted. The open-source Python library of pandas was used for data manipulation along with the Python module of library ***random*** for the random selection of 10 double-intent utterances. To create double intent utterances, the unique intents are deployed to extract the associated utterances for each intent. These utterances are then combined applying the rule of the conjunction with the goal of the concatenation of the two utterances with the linking word 'and'. Regarding the extraction of single intent utterances, as in the case of the creation of double intents, the unique 85 intents are utilized to extract the utterances of each intent. Then, 4 utterances are randomly selected from each intent. The ***.sample () method*** of the random library was equally used in the creation of both simple intent utterances and double-intent utterances without phenomena ensuring a diverse array of utterances.

3.7 Data Understanding: A Statistical Perspective

The dataset contains utterances that include double intents, incorporating linguistic phenomena such as anaphora, cataphora, and ellipsis. Additionally, it includes utterances featuring double intents devoid of these linguistic phenomena, as well as those that are characterized by single intents. The dataset under investigation has been divided into three distinct subsets: a training set, a test set, and a validation set. It is essential to underline that this dataset is manually constructed, ensuring a high level of precision in its composition. This manual construction process, while labour-intensive, is essential to guarantee the quality of the dataset. As a result, due to the difficulties encountered during manual construction, the dataset's overall size may be considered relatively small. The dataset contains a total of 1,875 utterances. The training set comprises a total of 1,315 utterances, of which 975 represent double intents, with 650 instances exhibiting linguistic phenomena and 325 instances without these linguistic phenomena. Additionally, the training set includes 340 single-intent utterances. The test set consists of 280 utterances, with 195 representing double intents, comprising 130 examples with linguistic phenomena and 65 instances without, along with 85 single-intent utterances. The validation set mirrors the test set in terms of composition, with 280 utterances, 195 double-intent utterances (130 with linguistic phenomena and 65 without), and 85 single-intent utterances. Although the dataset is relatively small, it has been carefully constructed manually. In addition, a series of assessments by English native speakers from Canada were conducted.

Figure 1: Distribution of Intent Classes and Utterances count in Training Set

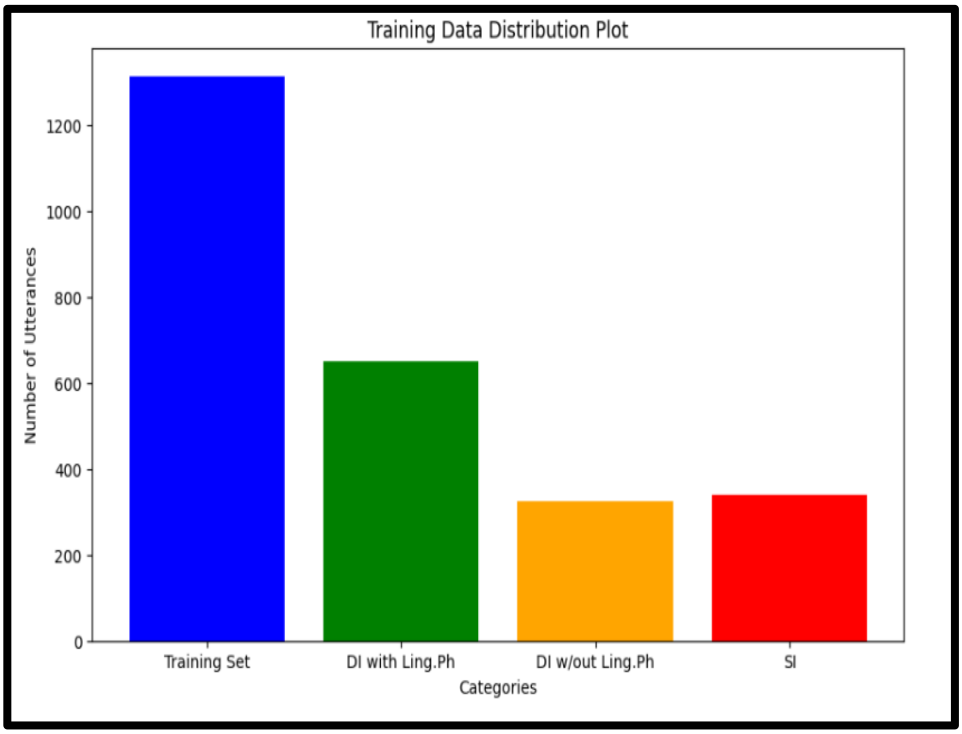


Figure 2: Distribution of Intent Types in Validation and Test Sets.

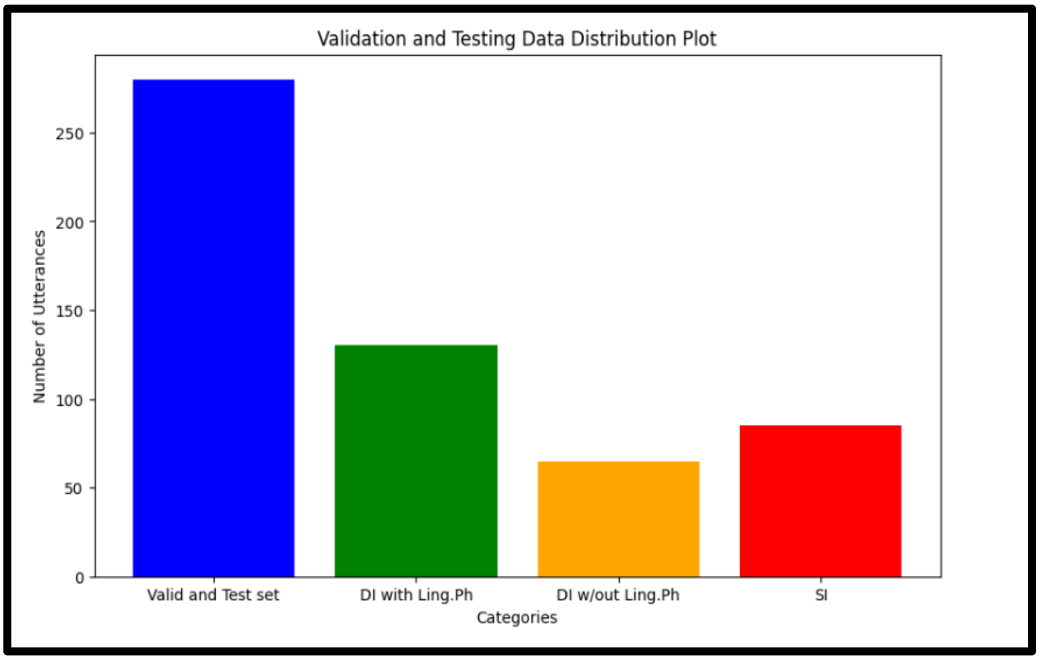


Table 6: Dataset Information.

Training Set Size	1.315
Development Set size	280
Testing Set size	280
Double Intents Unified Category	65
Single Intents	85
Total Classes	150

3.7.1 Anaphora, cataphora and ellipsis distribution

To improve our understanding of the dataset, distributions of different properties were calculated. Figure 6 illustrates the distribution of linguistic phenomena of interest across the training, the validation, and the test sets. The training set demonstrates a higher occurrence of such phenomena at around 49.47%, while the test and validation sets show a slightly lower presence, approximately 46.43% as depicted in Figure 3. That suggests a consistent trend in linguistic pattern distribution across the datasets, emphasizing the importance of diverse training data for robust natural language processing models.

Figure 3: Percentage Representation of Linguistic Phenomena Across Training, Test and Validation Sets.

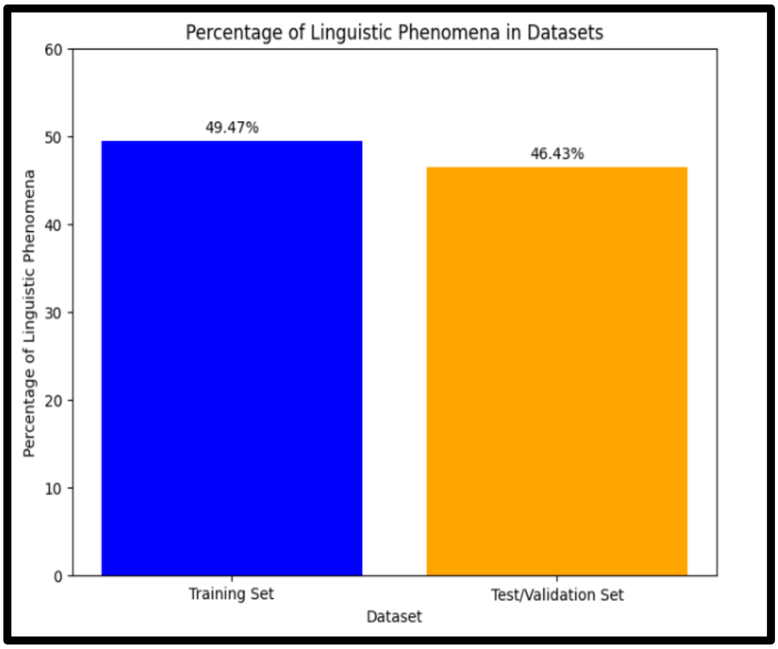


Figure 4: Distribution of Utterances with and without Linguistic Phenomena across Training Set.

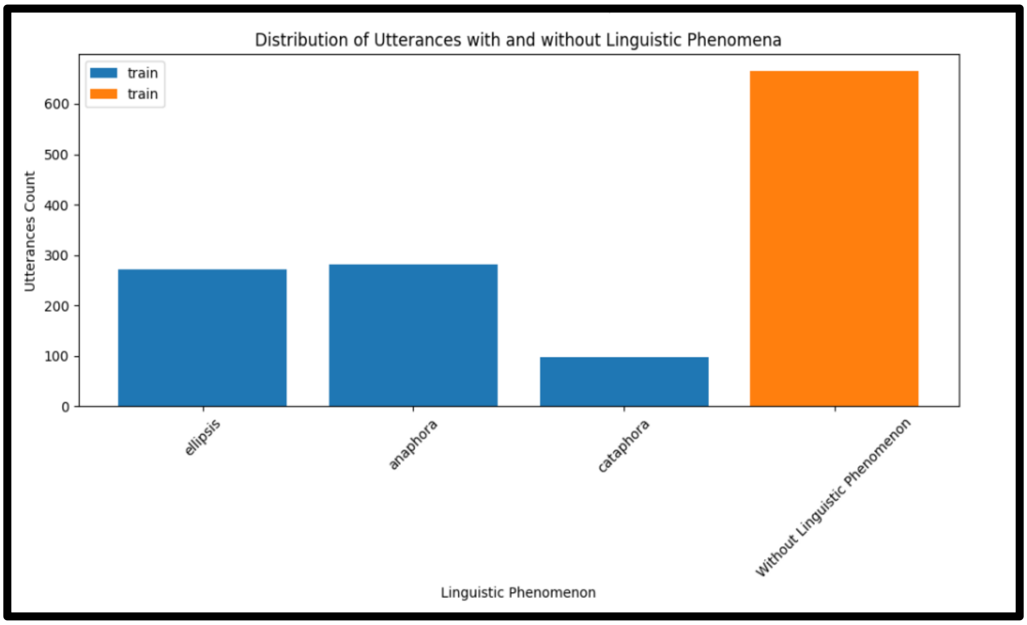


Figure 5: Distribution of Utterances with and without Linguistic Phenomena across Test Set.

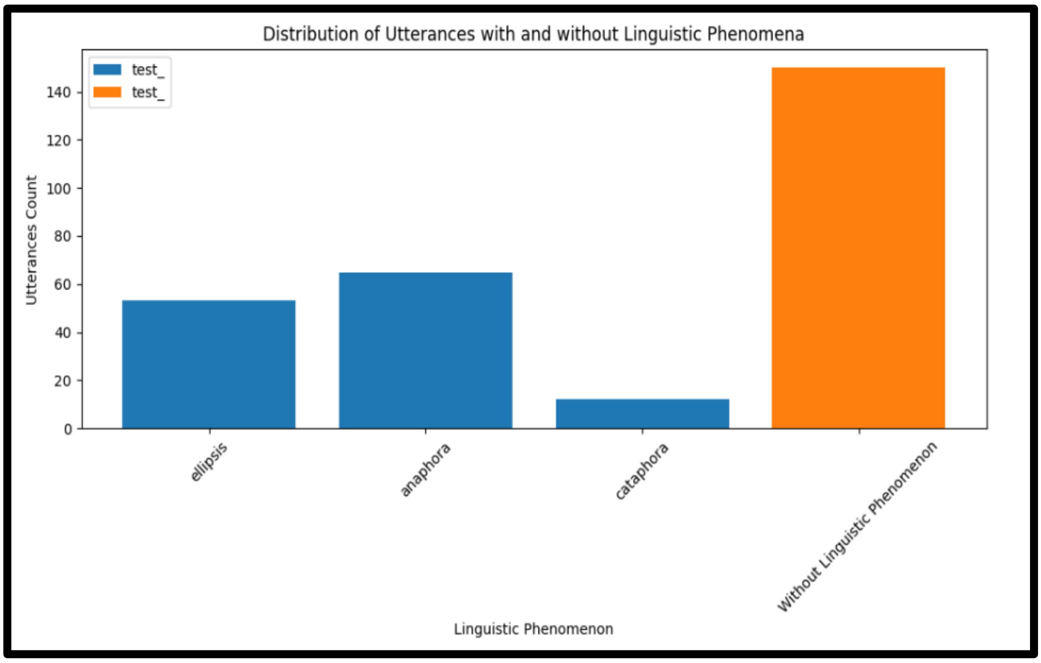
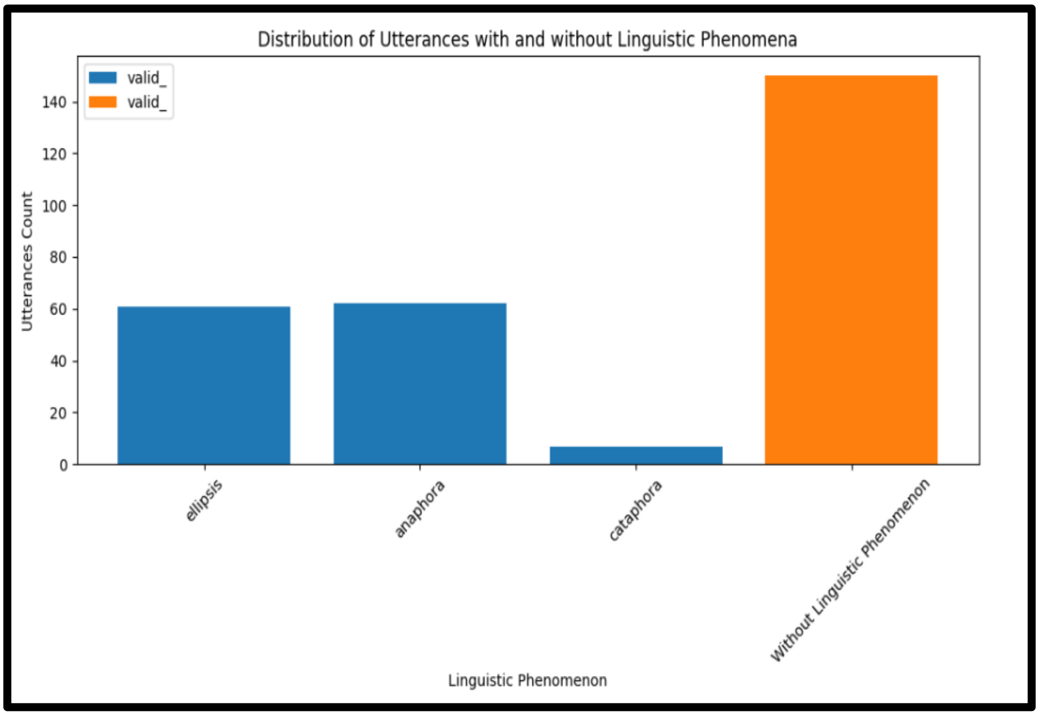


Figure 6: Distribution of Utterances with and without Linguistic Phenomena across Validation Set.



3.8 Rationale for Implementing an Annotation Process

The incorporation of anaphora, cataphora, and ellipsis phenomena raised questions regarding whether the creator and annotator identified the same referents of anaphora or cataphora, or whether they recognized the same elliptical elements. Furthermore, no native speakers were involved in the dataset development process. Due to these considerations, native speakers of English were recruited to assess part of the dataset. Out of the 1,895 utterances contained in the dataset, 360 were subjected to evaluation by native speakers (about 19 % of the dataset), primarily with a focus on phenomena such as anaphora, cataphora, and ellipsis.

3.8.1 Annotation Process

Annotators were recruited with an open call to social media for English native speakers. Applicants did not undergo a screening process that would include a review of their academic qualifications, prior experience in linguistic annotation or a linguistic proficiency test. Guidelines were also not considered necessary for this task, as their performance depended on their native speaker intuition. All the native speakers were Canadian. A task description was provided. Four annotators were involved in the annotation process. The annotation served two main purposes: first, to assess the grammatical correctness of the utterances, and second, to determine whether the same objects of referents, or elliptical elements were identified. As a response to both questions, annotators were expected to provide oral labels as either 'yes' or 'no' for grammaticality and 'yes' or 'no' regarding the identification of the objects of anaphora, cataphora, or elliptical elements. Annotators were encouraged not only to answer with 'yes' or 'no' labels but also to rectify potential grammatical errors and suggest changes.

3.8.2 Observations in Annotation

Throughout the annotation process, several observations have been made. Notably, some utterances where more than one linguistic phenomenon coexists have triggered a closer examination of the intricacies of language comprehension. Moreover, during the annotation process restrictions on the creation of anaphora were revealed due to semantic inconsistencies. Specifically, in the intents' **damaged_card** and ***replacement_card_duration'***, limitations have become apparent concerning the use of anaphora. This implies that certain linguistic or semantic conditions within these domains constrain the straightforward application of anaphoric references to previously mentioned elements. In the context of 'damaged_card' and 'replacement_card_duration', the lack of a coherent anaphoric link arises due to the nature of the referents. Specifically, the damaged card, which serves as the antecedent, cannot logically refer to the new replacement card. Anaphora relies on the continuity of reference between an earlier entity and a subsequent one, but in this case, the damaged card and the replacement card represent distinct entities, making it challenging to establish a meaningful connection between them. As a result, attempting to create an anaphoric link between these two distinct entities could potentially lead to confusion or misinterpretation within the discourse. Hence these restrictions have drawn attention to the need for precise application of linguistic mechanisms to ensure the coherence of communication. Furthermore, the thorough evaluation of the dataset has highlighted the need to address grammatical errors in several utterances that were not manually created, instead, they originated from the CLINC150 dataset. This finding confirms the assertion that the utilization of crowd workers in developing NLU datasets can potentially lead to errors [3].

3.9 Contributions and Limitations

The new dataset has been designed to overcome several significant limitations present in current Natural Language Understanding (NLU) datasets. Below, we outline its key accomplishments. First, it overcomes the existing limitations of current datasets which include only single-intent utterances and focus on one or two specific domains. It also overcomes the current trends through the incorporation of three linguistic phenomena within multi-intent datasets. We are not aware of any other dataset in this domain that includes these three phenomena. Moreover, since its source dataset includes intents from several domains, the new dataset encompasses intents of several domains too, helping to address the gap regarding domain-specific datasets. However, in this master thesis, we also attempted to combine intents that overlapped semantically across different domains.

While the present multi-intent dataset which was created in this master thesis, is valuable for its intended purpose, it exhibits several limitations. Firstly, its manual creation process proves to be time-intensive, consuming significant resources and effort. This time-expensive nature inherently restricts the dataset's scale, resulting in a relatively small-sized dataset. Additionally, the limited availability of annotators to annotate and evaluate the dataset poses a challenge, potentially impacting the dataset's depth and breadth of coverage. These limitations, while acknowledged, underline the necessity for further research and resource allocation to address them comprehensively.

4. EXPERIMENTS

Two baseline classification techniques were conducted in the context of this master thesis. The first experiment utilized a multi-label learning technique in which the combination of intent classes is treated as an atomic label. The goal of the first experiment was to predict the accurate intent label for each single or double intent utterance. On the other hand, the second experiment adopted a threshold-based multi-label approach, utilizing only the single labels to predict whether the utterances represent single or double intents, alongside their corresponding labels.

4.1 EVALUATION METRICS

Two different models are evaluated concerning their performance over the test datasets. Regarding the Intent Recognition task, accuracy is the evaluation metric of concern, following the related bibliography. Accuracy is defined as the ratio of correctly classified intents over the sum of the test examples. Nevertheless, we also recorded the F1 score in our dataset, as well as Precision and Recall. We report these metrics defined as follows:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall\ = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$F1\ Score\ = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$SI\ Custom\ Accuracy = \frac{1}{M}\sum_{i=1}^{M} 1\left(\hat{y}_i^{label} = y_i^{label}\right)$$

$$DI\ Custom\ Accuracy = \frac{1}{K}\sum_{i=1}^{K} 1\left(\hat{z}_i^{label} = z_i^{label}\right)$$

where the indicator function is defined as follows:

$$1\left(\hat{x}_i^{label}\right) := \begin{cases} 1, & if\ \hat{x}_i^{label} = x_i^{label} \\ 0, & if\ \hat{x}_i^{label} \neq x_i^{label} \end{cases}$$

4.2 Experiment 1

4.2.1 Approach

In this experiment, we adopted the baseline approach to address the scenario where double intents (DI) such as "*recipe, calories" are* present within a single utterance. This baseline approach treats double intents (separated by a comma in our case) as an atomic label made from the combination of two labels and is proven to be an effective technique. By treating double intents as an atomic class, the goal was to simplify the classification process while maintaining the integrity of the training process.

In the figure below the examples [2]-[7] where two intents are present will be treated as a single atomic class "balance,credit_score" for example, instead of two separate classes "balance" and "credit_score" respectively.

Table 7: SI and DI examples of our Dataset with and without Linguistic phenomena.

	UTTERANCES	INTENT
1	I would like to know the minimum payment for my credit card.	min_payment
2	Can you figure out how to find my credit score? Give me the recommendation to improve my credit score, too.	credit_score,improve_credit_score
3	Check my bank balance and my credit score.	balance,credit_score
4	While flying American Airlines I lost my luggage at O'hare and my phone, too	lost_luggage,find_phone
5	Please, tell me what my income is and how much tax there is on it.	income, taxes
6	How do I make the perfect omelette, and what is the calorie count for it?	recipe, calories
7	Put "laundry" on my to-do list.	todo_list_update
8	Text Sal and tell her hi.	text

4.2.2 Training details

For the first experiment in the row, we chose to utilize the BERT-base uncased as an Intent Classifier. The sequence length was restricted to a maximum of 44 tokens for optimal processing. We adopted the Adam optimizer with a learning rate of 1e-4 for effective model optimization. A dropout rate of 0.1 was integrated into the network to prevent overfitting. The model underwent training for 4 epochs. To control overfitting, we integrated an early stopping mechanism monitoring the validation loss, with a patience level of 5. A regularization technique, L2 regularization with a coefficient value of 0.00001, is also effectively applied to the final dense layer of the model, which also helps to prevent overfitting and improves the model's generalization. Finally, the batch size was set to 16.

4.3 Experiment 2

4.3.1 Approach

As opposed to the previous method that directly predicted the label combination, in this experiment, a threshold-based multi-label approach was implemented, treating the double intents labels as distinct entities during the training process. Therefore, only 85 single labels were used for predicting both double intents and single intents, given that these 85 single labels were used for the creation of combinations of the double intents.

In this experiment the goal was twofold; to predict the probability of an utterance belonging to one or two classes out of the 85 (if it is a single or double intent utterance) and to predict its equivalent labels.

The classification of this experiment is accomplished by comparing the predicted probability distribution and the true labels of test examples to belong to these 85 intent classes. Considering this aspect, Categorical Cross Entropy loss with logits = True is implemented. When using Categorical Cross Entropy loss with logits=True, the model is guided to understand how certain it is about its predictions. In this way, by comparing these probabilities, which are the transformed probabilities after the softmax function is applied, with the true labels, we help the model understand where it's making mistakes, allowing it to learn and improve its accuracy over time in classifying the multiple intent classes. Afterwards, the computed probabilities are sorted in descending order, assisting in identifying the most likely classes based on the highest predicted probabilities, providing crucial insights into the model's decision-making process, and helping in the accurate identification of the top predicted classes. To effectively classify if the test examples belong to single or double intent classes, we deploy only the first and second largest probability for the single and double intent classification as it is evident, that the predicted probabilities could vary regarding the difference between them. For that reason, we should find a way to compute the importance of this difference. To achieve this, we deploy the normalized score defined as follows:

$$normalized\ score = \frac{second\ largest\ probability}{first\ largest\ probability\ +\ second\ largest\ probability}$$

The values of the normalized score are in the range [0 – 0.5]. A value close to 0 means that we should have a single intent, while a value near 0.5 should be classified as a double intent. Consequently, the value of the threshold of this threshold-based multi-label approach should be in this range. Based on the value of the threshold, if the normalized score is under or equal to the threshold, then we consider the utterance to be a single-intent utterance. Otherwise, it is classified as a double-intent utterance. To determine the optimal threshold value, we experiment with different values to detect which value offers the best accuracy for both classifications respectively.

4.3.2 Training Details

During the training process, the data underwent an encoding process to encode double and single intents accordingly. The goal of this encoding was to assign 0.5 and 0.5 values to two out of 85 predicted classes 0 to the others, 1 to the predicted class and 0 to the others 84. For that reason, the dataset was initially preprocessed using the MultiLabelBinarizer technique to transform the list of 85 labels into a binary format where each label will be represented as a binary feature and then converted these labels to the values based on the sum of intents. In this experiment, the 'Bert-base-uncased' model variant was also deployed. The utterances were also tokenized and subsequently transformed into encodings with a maximum sequence length of 44. Regarding the batch size, in this experiment, it was set to 8. The model was also trained over 10 epochs, with an Adam optimizer utilizing a learning rate of 1e-5. The model's architecture encompassed a dropout layer with a rate of 0.2, followed by a dense layer, and was compiled using the Categorical Cross entropy with a logits loss function. Moreover, in the training process, the Early Stopping callback with patience of 3 was implemented.

5. Experimental Results

5.1 Results of Experiment 1.

Table 8 below depicts the performance of the first fine-tuned BERT model on the new dataset. The performance of the Bert model appears to be quite satisfactory, demonstrating its ability to correctly classify the intents of the utterances. Notably, the model's accuracy of 87.5 % on the test dataset highlights its ability to make almost accurate predictions. Additionally, the computed F1 score of 84.5 % suggests that there is a strong balance between precision and recall, implying that our model has a good level of accuracy in classifying both double and single intents with the approach of treating the double intents as a single atomic class, with a relatively low margin of error. That means there might still be some misclassifications, but the overall performance of our model is reliable. While this experiment appears to yield better results, it is accompanied by certain limitations. Training the model to predict double intents requires the inclusion of this predefined combined category within the training dataset. Moreover, in the context of this approach, training the model to predict double intents as an atomic class necessitates the availability of a sufficient number of instances within the training dataset where these combined classes are represented. This means that the dataset must include various examples that illustrate these combinations of predefined combined intents as unified categories. However, the challenge arises when attempting to ensure comprehensive coverage of all potential combinations in the utterances. When these combinations are underrepresented from the training data, the model might not adequately learn the distinct characteristics of these specific combined classes. Consequently, this can lead to a reduced capacity to accurately predict or handle previously unseen or less common combinations during the classification process. This limitation can potentially impact the model's overall effectiveness in handling complex real-world scenarios that involve a diverse array of intent combinations.

Table 8: Results of experiment 1.

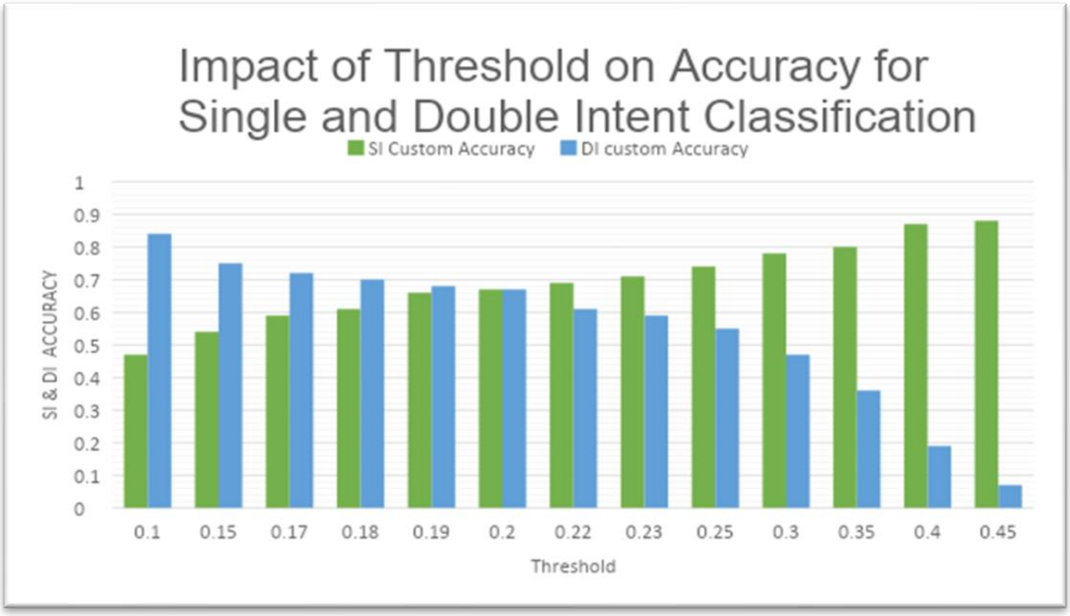
Experiment	Intent Accuracy	F1 Score	Precision	Recall
Bert-model (atomic model)	87.5 %	84.5 %	83.5 %	87.5%

5.2 Results of Experiment 2.

In the second experiment, we need to evaluate the performance of our model by creating two distinct custom accuracy metrics, one for single intent and one for double intent classification. For this purpose, the custom accuracies for Single Intents (SI) and Double Intents (DI) were created to evaluate the accuracy of the classification of Double Intents and Single Intents respectively. Thus, for single intents, the custom accuracy is calculated by accumulating the total number of single intents and identifying the instances where the predicted label matches the actual label. When and only this condition is met the accumulation occurs. The accumulation of the correctly classified intents, considering both the number of intents +++and their matching labels, provides a comprehensive view of the model's ability to correctly classify the test utterances. Similarly, for double intents, the process remains the same, with the focus on double intents. By summing up the intents that are at the same time both double and have correct labels as well, it is possible to effectively assess the model's accuracy in classifying double intents. Utilizing these custom accuracy measures allows for precise observation of the performance of our classification system, providing a detailed perspective that differs from its overall categorical accuracy for both single and double intents. Furthermore, regarding the correlation between the threshold and the custom accuracy results, the threshold is the value that affects the custom accuracy measures for both single and double intents. This value determines whether the utterance is a single or double intent. If the normalized score, computed based on the sorted test probabilities, is below or equal to the threshold, the utterance is classified as a single intent; otherwise, it is classified as a double intent.

As shown in Figure 7 below, the threshold value that provides the highest accuracy for both single and double intent classification is 0.2; hence, as the threshold deviates from the value of 0.2, the accuracy of intent classification is seen to depend on the variable threshold's decrease and increase for single intent classification and double intent classification, respectively. Consequently, at the threshold of 0.2, a balance is achieved where both Single Intent (SI) Custom Accuracy and Double Intent (DI) Custom Accuracy are relatively high, at around 70 %. This indicates a potential optimal threshold value.

Figure 7: Impact of Threshold on SI and DI Accuracy



Regarding the results of the second experiment, despite the slightly lower accuracy observed in the second experiment, this method only requires the inclusion of individual intents in the training set, without the requirement of including the double intent categories as unified categories in the training set, too. The incorporation of a normalized score further refines the model's classification precision by accounting for the variations between predicted probabilities. Through these improvements, the second approach achieves a more sophisticated and precise classification of single and double intents, leading to a more comprehensive understanding of the model's decision-making process. Furthermore, the second experiment demonstrates how the threshold value influences the custom accuracy measures for single and double intents. Notably, the identification of the optimal threshold value at 0.2 highlights a balance between high Single Intent (SI) Custom Accuracy and Double Intent (DI) Custom Accuracy, both approaching approximately 70%. This finding underscores the effectiveness of the threshold-based multi-label approach in accurately categorizing both single and double intents, confirming its efficacy in enhancing the overall classification accuracy.

However, supplementary metrics were necessary to gain a comprehensive understanding of our results. Consequently, we computed recall, precision, and F1-score for both single and double intents. Upon careful examination of the table, it becomes apparent that the F1-Score and Precision values for Single Intents are relatively low, at 58% and 50%, respectively. Although the recall rates are also the same as the accuracy rates, there is a noticeable difference in precision, with 50% for single intents and 84% for double intents. This suggests that the model is more precise in identifying double intents compared to single intents. Additionally, the F1 scores for single and double intents are 58% and 76%, respectively.

In consideration of precision's core emphasis on accurately predicting specific instances within a class, its susceptibility to the availability of representative samples becomes evident. When such samples are scarce for single intents, the precision score can decline, even if the overall accuracy remains high. It is crucial to note that precision is determined by the ratio of true positive results to the sum of true positives and false positives, reflecting the model's proficiency in correctly identifying relevant instances within a specific class. Notably, the significant impact of the 30% accuracy of double intents on single intents stands in contrast to the comparatively minor effect of the 30% accuracy of single intents.

Table 9: Results of the threshold-based model before the increase of SI samples.

Threshold-based model with the new dataset	**Single Intents**	**Double Intents**
Metrics		
Accuracy	69 %	70 %
Recall	69 %	70 %
Precision	**50 %**	84 %
F1-Score	**58 %**	76 %

Figure 8: Confusion Matrices before the addition of SI samples.

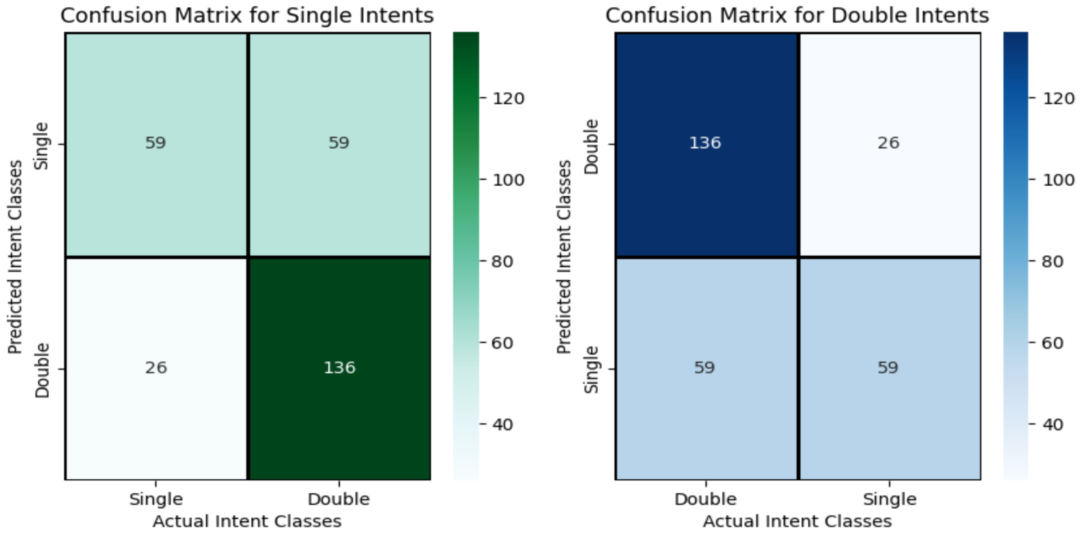


Figure 8 illustrates the Confusion Matrix for Single and Double Intents, before the addition of single-intent samples. The confusion matrices provided depict the performance metrics of our intent classification model for single and double intents. In the single-intent scenario, the model correctly identified 59 instances as single-intent, when they were, in fact, single (TP) from a total number of 85 single intents. However, it also misclassified 59 samples as double intents, when they were, in fact, single intents (FP). On the other hand, it erroneously predicted 26 samples as double intents, while they were single intents (FN). Nonetheless, the model accurately classified 26 samples that were not single and indeed they weren't (TN). As for the double-intent scenario, the model correctly identified 136 instances of the double intents (TP) from a total of 195 double intents, but erroneously labelled 26 instances as double, when they were not (FP). It also erroneously recognized 59 instances as single, while they were double intent (FN), but correctly detected that 59 instances weren’t double intents, and they were indeed single intents (TN).

6. Conclusion and Future Work

The major thrust of this master thesis has been the development of a multi-intent dataset comprising utterances exemplifying the linguistic phenomena of Anaphora, Cataphora, and Ellipsis, using the existing CLINC150 dataset. Two baseline classification methodologies were also employed for two supplemental experiments to examine our dataset: a multi-label learning technique treating the combination of intent classes as an atomic label and a threshold-based multi-label approach relying solely on the labels of single intents.

As a result, this research endeavours have primarily surpassed the limitations of existing NLU datasets that primarily feature single-intent utterances confined to a limited number of domains and being synthetically generated. Also, with the inclusion of three linguistic phenomena in a multi-intent setting, this study pioneers a unique dataset not previously observed. Additionally, the inclusion of diverse intents from multiple domains aids in bridging the gap associated with domain-specific datasets. Furthermore, the combination of semantically overlapping intents across different domains represents a novel attempt to enhance the sophistication of intent classification methodologies. Regarding the results of our two experiments, the concatenated version yielded an accuracy of 87.5%, while the threshold version experiment resulted in 70% accuracy, with the former being 17.5% higher. These results indicate that the first baseline approach is more effective at first glance.

Although the primary goal was the development of a large-scale database, the complexity of the linguistic phenomena included in the dataset resulted in a manual process to achieve more naturalness in the utterances of our dataset. While coreference and ellipsis have become crucial tasks of NLP, as they often occur in dialogues, the process of creating a dataset that includes multiple intents, coreferent and elliptical elements all automatically generated is a more complex task, because of the complexities associated with precisely capturing the linguistic phenomena of anaphora, cataphora and ellipsis concurrently. Because of the manual process adopted in creating this NLU dataset, the resulting dataset remains limited in size, characterized by its relatively small scale. However, this non-synthetic approach, which does not adopt the conventional method of creating double intents through simple conjunction words, as seen in benchmark multi-intent datasets such as 'MixAtis' and 'MixSnips' [42] effectively addresses the potential issue of random intent combinations. This approach enables a more meticulous process facilitated by human involvement. Moreover, to our knowledge, this is the first dataset that includes anaphora, cataphora, and ellipsis phenomena, along with multi-intents contributing to the creation of NLU datasets that are up to date with current industry requirements. By addressing these two core deficiencies in most NLU datasets, our dataset represents a small step, but significant one towards bridging some of the gaps mentioned in the literature.[3]

Despite the results favouring the first approach, our findings underscore the implications of this approach noted in another research, too. While this simplified approach demands the incorporation of these specific combinations of double intents within the training data, it also neglects the shared similarities and patterns that could be observed across various intent classes. As a result, this oversight creates a dearth of comprehensive data representation, leading to challenges in building robust models capable of accurate classification as also observed in [2]. Nevertheless, the threshold-based methodology still exhibits certain limitations. The results indicate a higher accuracy in predicting single intents compared to double intents. These findings confirm that the more commonly used threshold-based models perform better in predicting single intents within utterances but demonstrate decreased accuracy as the number of intents increases, while recent experiments with threshold-free methodology exhibit greater consistency across these two different scenarios [43].

While the current research primarily focuses on an intent-oriented approach, it could potentially benefit from transitioning to a joint-oriented methodology, which handles the two subtasks of intent detection and slot filing jointly. This methodology has attracted remarkable attention and success in recent years [1], [8], [16], [44]. By adopting this approach, a finer-grained analysis of the key elements within conversations becomes feasible, leading to an overall improvement in the precision and thoroughness of natural language processing tasks. Moreover, it facilitates the examination of how the creation of double intents via linguistic phenomena can affect the efficacy of our models. However, notably all these methods as our second experiment utilize a threshold-based approach for predicting multiple intents, where the standard practice involves estimating label-instance probabilities and selecting intent labels that surpass the predefined threshold value. Hence, it is essential to consider the adoption of threshold-free techniques in our future research endeavors to observe the outcomes. In addressing the challenges our models face when correctly classifying double intent classes in connection with linguistic phenomena, we have noted significant difficulties, particularly in instances involving ellipsis and cataphora. Nevertheless, we intend to incorporate these observations into future work.

Additionally, our research could be further expanded, integrating syntax annotation for the linguistic phenomena of anaphora, cataphora and ellipsis aligning to CONLL-U format. Although this approach is less conventional in this domain, it has the potential to contribute to the field of multi-intent detection. It enables us to assess whether it offers a deeper understanding of the fundamental linguistic structures and patterns associated with multiple intents. Through the incorporation of syntax annotation, we can capture the complex interconnections among words and phrases in utterances, facilitating a more advanced examination of how linguistic phenomena affect the recognition and the prediction of multiple intents within a single utterance. Finally, it would be crucial to integrate a comparative analysis of our models' outcomes in connection with other models to gain a deeper understanding of our dataset and its capabilities.

ACRONYMS

ATIS	Airline Travel Information System
DI	Double Intents
FN	False Negatives
FP	False Positives
IC	Intent Classification
ID	Intent Detection
NLP	Natural Language Processing
NLU	Natural Language Understanding
SI	Single Intents
TN	True Negatives
TP	True Positives

REFERENCES

[1] R. Gangadharaiah and B. Narayanaswamy, ‘Joint Multiple Intent Detection and Slot Labeling for Goal-Oriented Dialog’, in *Proceedings of the 2019 Conference of the North*, Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 564–569. doi: 10.18653/v1/N19-1055.

[2] P. Xu and R. Sarikaya, ‘Exploiting shared information for multi-intent natural language sentence classification’, in *Interspeech 2013*, ISCA, Aug. 2013, pp. 3785–3789. doi: 10.21437/Interspeech.2013-599.

[3] I. Casanueva, I. Vulić, G. Spithourakis, and P. Budzianowski, ‘NLU++: A Multi-Label, Slot-Rich, Generalisable Dataset for Natural Language Understanding in Task-Oriented Dialogue’, in *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, United States: Association for Computational Linguistics, Apr. 2022, pp. 1998–2013. doi: 10.18653/v1/2022.findings-naacl.154.

[4] S. Larson and K. Leach, ‘Redwood: Using Collision Detection to Grow a Large-Scale Intent Classification Dataset’. arXiv, Jul. 25, 2022. doi: 10.48550/arXiv.2204.05483.

[5] L. Qin, W. Che, Y. Li, H. Wen, and T. Liu, ‘A Stack-Propagation Framework with Token-Level Intent Detection for Spoken Language Understanding’, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, 2019, pp. 2078–2087. doi: 10.18653/v1/D19-1214.

[6] B. Liu and I. Lane, ‘End-to-End Learning of Task-Oriented Dialogs’, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, New Orleans, Louisiana, USA: Association for Computational Linguistics, Mar. 2018, pp. 67–73. doi: 10.18653/v1/N18-4010.

[7] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, ‘The ATIS Spoken Language Systems Pilot Corpus’, in *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*, 1990. Accessed: Sep. 29, 2023. [Online]. Available: https://aclanthology.org/H90-1021

[8] L. Chen, P. Zhou, and Y. Zou, ‘Joint Multiple Intent Detection and Slot Filling Via Self-Distillation’, in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Feb. 2022, pp. 7612–7616. doi: 10.1109/ICASSP43922.2022.9747843.

[9] L. Chen, N. Chen, Y. Zou, Y. Wang, and X. Sun, ‘A Transformer-based Threshold-Free Framework for Multi-Intent NLU’, in *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Jul. 2022, pp. 7187–7192. Accessed: Oct. 03, 2023. [Online]. Available: https://aclanthology.org/2022.coling-1.629

[10] S. Larson and K. Leach, ‘A Survey of Intent Classification and Slot-Filling Datasets for Task-Oriented Dialog’. arXiv, Jul. 26, 2022. Accessed: Oct. 02, 2023. [Online]. Available: http://arxiv.org/abs/2207.13211

[11] M. Song, B. Yu, L. Quangang, W. Yubin, T. Liu, and H. Xu, ‘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, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Sep. 2022, pp. 7967–7977. doi: 10.18653/v1/2022.emnlp-main.543.

[12] J. Liu, Y. Li, and M. Lin, 'Review of Intent Detection Methods in the Human-Machine Dialogue System', *J. Phys. Conf. Ser.*, vol. 1267, no. 1, p. 012059, Apr. 2019, doi: 10.1088/1742-6596/1267/1/012059.

[13] B. Kim, S. Ryu, and G. G. Lee, 'Two-stage multi-intent detection for spoken language understanding', *Multimed. Tools Appl.*, vol. 76, no. 9, pp. 11377–11390, Feb. 2017, doi: 10.1007/s11042-016-3724-4.

[14] N. Moghe, E. Razumovskaia, L. Guillou, I. Vulić, A. Korhonen, and A. Birch, 'MULTI3NLU++: A Multilingual, Multi-Intent, Multi-Domain Dataset for Natural Language Understanding in Task-Oriented Dialogue'. arXiv, Jun. 19, 2023. Accessed: Sep. 17, 2023. [Online]. Available: http://arxiv.org/abs/2212.10455

[15] J. Liu, Y. Li, and M. Lin, 'Review of Intent Detection Methods in the Human-Machine Dialogue System', *J. Phys. Conf. Ser.*, vol. 1267, no. 1, p. 012059, Apr. 2019, doi: 10.1088/1742-6596/1267/1/012059.

[16] L. Qin, X. Xu, W. Che, and T. Liu, 'AGIF: An Adaptive Graph-Interactive Framework for Joint Multiple Intent Detection and Slot Filling', in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online: Association for Computational Linguistics, Aug. 2020, pp. 1807–1816. doi: 10.18653/v1/2020.findings-emnlp.163.

[17] G. Bihani and J. T. Rayz, 'Fuzzy Classification of Multi-intent Utterances'. arXiv, Apr. 21, 2021. Accessed: Oct. 02, 2023. [Online]. Available: http://arxiv.org/abs/2104.10830

[18] H. Wu, K. Xu, L. Song, L. Jin, H. Zhang, and L. Song, 'Domain-Adaptive Pretraining Methods for Dialogue Understanding'. arXiv, May 28, 2021. doi: 10.48550/arXiv.2105.13665.

[19] S. Larson and K. Leach, *A Survey of Intent Classification and Slot-Filling Datasets for Task-Oriented Dialog*. 2022. doi: 10.48550/arXiv.2207.13211.

[20] S. Gupta, R. Shah, M. Mohit, A. Kumar, and M. Lewis, 'Semantic Parsing for Task Oriented Dialog using Hierarchical Representations', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 2787–2792. doi: 10.18653/v1/D18-1300.

[21] 'NLU++: A Multi-Label, Slot-Rich, Generalisable Dataset for Natural Language Understanding in Task-Oriented Dialogue - ACL Anthology'. Accessed: Sep. 26, 2023. [Online]. Available: https://aclanthology.org/2022.findings-naacl.154/

[22] S. Larson and K. Leach, 'Redwood: Using Collision Detection to Grow a Large-Scale Intent Classification Dataset', in *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Edinburgh, UK: Association for Computational Linguistics, Jun. 2022, pp. 468–477. Accessed: Oct. 02, 2023. [Online]. Available: https://aclanthology.org/2022.sigdial-1.45

[23] A. Gupta, T. Babtiwale, C. Jain, and K. Modi, 'HDIAL: Dataset for benchmarking dialogue systems on linguistic phenomena', Accessed: Oct. 03, 2023. [Online]. Available: https://tanaya-b.github.io/assets/slides/HDIAL_Granular_Benchmarking.pdf

[24] R. Mitkov, 'Robust Pronoun Resolution with Limited Knowledge', in *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational*

Linguistics, Volume 2, Montreal, Quebec, Canada: Association for Computational Linguistics, Dec. 1998, pp. 869–875. doi: 10.3115/980691.980712.

[25] 'varieties-of-anaphora.pdf'. Accessed: Oct. 03, 2023. [Online]. Available: https://pure.uvt.nl/ws/portalfiles/portal/757696/varieties-of-anaphora.pdf

[26] R. Dale, '`One'-anaphora and the case for discourse-driven referring expression generation', in *Proceedings of the Australasian Language Technology Workshop 2003*, Melbourne, Australia, Sep. 2003, pp. 16–21. Accessed: Oct. 12, 2023. [Online]. Available: https://aclanthology.org/U03-1002

[27] A. E. Goldberg and L. A. Michaelis, 'One Among Many: Anaphoric One and Its Relationship With Numeral One', *Cogn. Sci.*, vol. 41, no. S2, pp. 233–258, 2017, doi: 10.1111/cogs.12339.

[28] R. Trnavac and M. Taboada, 'Cataphora, backgrounding and accessibility in discourse', *J. Pragmat.*, vol. 93, pp. 68–84, Feb. 2016, doi: 10.1016/j.pragma.2015.12.008.

[29] J. G. Carbonell, 'Discourse Pragmatics and Ellipsis Resolution in Task-Oriented Natural Language Interfaces', in *21st Annual Meeting of the Association for Computational Linguistics*, Cambridge, Massachusetts, USA: Association for Computational Linguistics, Mar. 1983, pp. 164–168. doi: 10.3115/981311.981343.

[30] C. Phillips and D. Parker, 'The psycholinguistics of ellipsis', *Lingua*, vol. 151, pp. 78–95, Nov. 2014, doi: 10.1016/j.lingua.2013.10.003.

[31] J. Craenenbroeck and T. Temmerman, 'Ellipsis in natural language Theoretical and empirical perspectives', 2017. Accessed: Oct. 04, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Ellipsis-in-natural-language-Theoretical-and-Craenenbroeck-Temmerman/7959a28d1792ebba16423337f66c56241fe2ba67

[32] S. A. Bernhardt, 'Review of COHESION IN ENGLISH', *Style*, vol. 14, no. 1, pp. 47–50, 1980.

[33] N. Corver and M. van Koppen, 'NP-ellipsis with adjectival remnants: a micro-comparative perspective', *Nat. Lang. Linguist. Theory*, vol. 29, no. 2, pp. 371–421, May 2011, doi: 10.1007/s11049-011-9140-6.

[34] J. Van Craenenbroeck and J. Merchant, 'Ellipsis phenomena', in *The Cambridge Handbook of Generative Syntax*, 1st ed., M. Den Dikken, Ed., Cambridge University Press, 2013, pp. 701–745. doi: 10.1017/CBO9780511804571.025.

[35] I. A. Sag and J. Hankamer, 'Toward a theory of anaphoric processing', *Linguist. Philos.*, vol. 7, no. 3, pp. 325–345, Aug. 1984, doi: 10.1007/BF00627709.

[36] 'Ellipsis and Coreference Resolution in a Computerized Virtual Patient Dialogue System | SpringerLink'. Accessed: Oct. 03, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s10916-016-0562-x

[37] J. Quan, D. Xiong, B. Webber, and C. Hu, 'GECOR: An End-to-End Generative Ellipsis and Co-reference Resolution Model for Task-Oriented Dialogue', in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, Aug. 2019, pp. 4547–4557. doi: 10.18653/v1/D19-1462.

[38] B.-H. Tseng *et al.*, ‘CREAD: Combined Resolution of Ellipses and Anaphora in Dialogues’. arXiv, May 20, 2021. Accessed: Oct. 03, 2023. [Online]. Available: http://arxiv.org/abs/2105.09914

[39] S. Larson *et al.*, ‘An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction’. arXiv, Sep. 04, 2019. doi: 10.48550/arXiv.1909.02027.

[40] ‘(2) (PDF) The GUM corpus: creating multilayer resources in the classroom’. Accessed: Oct. 17, 2023. [Online]. Available: https://www.researchgate.net/publication/293195103_The_GUM_corpus_creating_multilayer_resources_in_the_classroom

[41] T. Saha, D. Gupta, S. Saha, and P. Bhattacharyya, ‘A hierarchical approach for efficient multi-intent dialogue policy learning’, *Multimed. Tools Appl.*, vol. 80, no. 28, pp. 35025–35050, Nov. 2021, doi: 10.1007/s11042-020-09070-7.

[42] H. Chen, X. Liu, D. Yin, and J. Tang, ‘A Survey on Dialogue Systems: Recent Advances and New Frontiers’, *ACM SIGKDD Explor. Newsl.*, vol. 19, no. 2, pp. 25–35, Nov. 2017, doi: 10.1145/3166054.3166058.

[43] L. Chen, N. Chen, Y. Zou, Y. Wang, and X. Sun, ‘A Transformer-based Threshold-Free Framework for Multi-Intent NLU’, in *Proceedings of the 29th International Conference on Computational Linguistics*, N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, Eds., Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Jul. 2022, pp. 7187–7192. Accessed: Nov. 02, 2023. [Online]. Available: https://aclanthology.org/2022.coling-1.629

[44] L. Qin, F. Wei, T. Xie, X. Xu, W. Che, and T. Liu, ‘GL-GIN: Fast and Accurate Non-Autoregressive Model for Joint Multiple Intent Detection and Slot Filling’, in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., Online: Association for Computational Linguistics, Dec. 2021, pp. 178–188. doi: 10.18653/v1/2021.acl-long.15.