GIVENNESS HIERARCHY THEORETIC NATURAL LANGUAGE GENERATION

by

Poulomi Pal

A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Master of Science (Computer Science).

Golden, Colorado

Date _____

Signed: _____
Poulomi Pal

Signed: _____
Dr. Tom Williams
Thesis Advisor

Golden, Colorado

Date _____

Signed: _____
Dr. Tracy Camp
Professor and Head
Department of Computer Science

ABSTRACT

Language-capable interactive robots participating in natural language dialogues with human interlocutors must be able to naturally and efficiently communicate about the objects, locations, and people found in human environments. A key aspect of natural language communication is the use of anaphoric language through *pronominal forms* such as *it*, *this*, and *that <NP>*. The linguistic theory of the *Givenness Hierarchy* (GH) suggests that humans use anaphora based on the *cognitive statuses* their referents have in the minds of their interlocutors. In previous work, researchers presented the first computational implementation of the full GH for the purpose of robot anaphora understanding, leveraging a set of rules informed by the GH literature. However, that approach was designed specifically for natural language understanding (NLU), oriented around GH-inspired memory structures used to assess the set of candidate referents with a given cognitive status. In contrast, natural language generation (NLG) requires a model in which cognitive status can be assessed for a given entity. In this work, we present a statistical model of cognitive status and demonstrate how this model can be used to facilitate robot anaphora generation. Specifically, we present an AI model that leverages the concept of cognitive status for the selection of pronominal forms for effective NLG.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Robots are becoming more common and an integral part of our lives nowadays. Initially, robots were mostly used in industries for navigation (e.g. moving from a start position to a destination) and manipulation (e.g. robot arms assembling products) tasks. But the application of robotics has widened since then where currently robots are not merely used as tools by humans to complete a task, but as partners/teammates to jointly achieve a particular task. When used as tools there is little to no interaction between a robot and its user (task-based interaction), whereas when working as partners the robot and the human are expected to interact with each other just as a human would interact with another human in a similar scenario (social interaction).

The area of social robotics deals with developing social robots, that is, autonomous agents that are able to communicate and interact with humans in their daily lives for jointly achieving particular tasks, following certain social cues and norms. The interaction capability is the most important aspect of a social robot because of human-human interaction being replaced by human-robot interaction in different environments such as schools, offices, public places, homes, etc.. Thus, it is crucial for a social robot to be able to understand, respond and behave appropriately with the person it is interacting with in different contexts.

The development and use of social robots has expanded from traditional industrial applications to commercial ones. The different types of social robots that have been developed and used over the years for different application scenarios are as described below:

1. Entertainment robots: Robots have been used for entertainment purposes such as toy robots for children and adolescents, robots as actors, robots used in exhibitions, etc. Sony's AIBO [1], a small dog-like pet-style robot was among the first toy robots developed for personal use and became highly popular. Similarly, robots looking like

animals (for example dinosaurs) are implemented in exhibitions and theme parks for entertainment of children and adults alike. In fact, as demonstrated in [2], robots can also collaborate with humans as actors in plays and thus facilitate creativity through interaction.

2. Robots used in healthcare and therapy: Social robots are used in the domain of healthcare and therapy, also known as socially assistive robotics (SAR) [3] that deals with eldercare, individuals in rehabilitation and patients with cognitive and social disabilities (such as autism, dementia, etc.). For example, the seal-like Paro robot (that responds to touch through wriggling and seal-like noises) has been shown to reduce loneliness and depression among people in eldercare homes by providing companionship and thus improving their psychological and mental health.

   Similarly, different types of social robots (animal-like and human-like) have been developed and used by therapists to assist their autistic patients augment their social and cognitive capabilities. An example is the NAO robot by Softbank Robotics.

3. Educational robots: Another use of social robots is in the field of learning, where the robot can be a tutor (supporting a teacher in their teaching), or a peer robot to a student/learner, helping each other in the process of learning or can be completely taught by a student. Research studies have shown that students are more excited and interested in learning with the social and physical presence of a robot. On the contrary, when a student teaches a robot it leads to better knowledge gain and confidence boost in the student [4].

4. Robots as Personal Assistants: Robots are being used as personal home assistants, which is kind of a successor technology to using smart-home assistants like Apple's Siri or Amazon's Alexa. But the advantage of having a robot personal assistant is that they have their social presence felt by engaging in simple conversations, tracking their user by turning, making expressions with virtual eyes, playing music, etc. An example

of such a social robot used as a personal assistant is JIBO by Jibo Inc.

5. Service Robots: The primary use of service robots is to help humans in doing their daily mundane repetitive tasks such as cleaning houses (for example, Roomba robot automatically cleans the house without the requirement of any human intervention), delivering objects from one place to another in warehouses or hotels, as security robots in public places such as shopping malls. Search and rescue robots are also service robots which can visit places that are otherwise harmful and dangerous for humans to visit.

   Apart from these, there exists service robots that directly interacts with people through spoken language to get their job done. One such example is that of a robot receptionist that offers information to its customers through conversations [5]. Another example is that of an autonomous robot tour guide in a museum setting where the robot moves around in the environment and when the visitors choose a guided tour, the robot leads them to different exhibits and provides explanations [6].

In this thesis work, we look into this aspect of human robot interaction (HRI) where robots working as teammates interact directly with humans using natural language and how it can better understand and generate spoken language using language models for effective and efficient collaboration with humans.

As human-robot interaction becomes increasingly common, language-capable robots need to be able to talk about the objects, locations, and people in their environments in the same way as humans do, to facilitate concise, easy, and unambiguous communication. The use of *pronouns* is an important aspect of natural language communication between humans. To reap these benefits, just like humans, robots must be able to understand and use pronouns like *it*, *this*, and *that*. For example, consider the following scenario where a person enters a building and enquires about the location of the reception to a fellow person.

Alice: Hi! Do you know where the Reception is?

Bob: Yes, it is on the second floor.

Alice: Thank you!

In this brief dialogue, Bob uses the pronominal form "it" to refer to the reception and the first person has absolutely no problem understanding that "it" refers to the "reception". Similarly, if there is a service robot in the building catering to the need of every person entering the building, then the robot must be equipped with understanding and generating pronominal forms such as *it, this, that* etc. to efficiently communicate with their human counterparts.

A common feature of natural language interaction between humans is the use of pronominal forms (as evident from the previous example scenario) and a robot working with a human will similarly have to be able to understand and generate pronominal forms for effective human-robot collaboration.

According to the linguistic theory of the *Givenness Hierarchy* (GH) [7], humans tend to use pronouns rather than longer referring expressions due to implicit assumptions about the *cognitive status* the referent has in the mind of their conversational partner. That is, the use of different referring forms is viewed as justified based on whether the referent is *In Focus*, *Activated*, *Familiar*, and so forth, within the current conversation. Thus, for robots to understand and generate human-like natural language they must be able to model this notion of cognitive status.

Previously, Williams and Scheutz [8] (see also [9, 10]) presented the first full computational implementation of the GH for the purpose of robotic natural language understanding (NLU), using a set of hand-crafted rules informed by the GH literature (no computational model of GH- theoretic language generation exists). However, that approach was designed specifically for robotic natural language understanding, oriented around GH-inspired memory structures used to assess what entities are candidate referents given a particular cognitive status. In contrast, natural language generation (NLG) requires a model in which cognitive status can be assessed for a given entity.

Such a model of cognitive status could either be developed as a rule-based model (not dissimilar from the rule-based approach to GH-theoretic language understanding taken by Williams and Scheutz [8]), or could instead be developed as a statistical model which would attempt to learn to predict an entity's cognitive status from data. While in practice both rule-based and data-driven empirical models are useful [11], data-driven models may be better able to handle unseen, uncertain situations [11, 12].

The research questions that we discuss in this thesis are as follows:

- How can we collect ground truth data on cognitive status assumptions?
- How can cognitive status be computationally modeled?
- How can we leverage the concept of cognitive status to directly select referring forms for natural language generation?

First, we develop and compare *two* models of cognitive status, both structured so as to be optimized for natural language generation rather than natural language understanding; a rule-based *Finite State Machine* model directly informed by the GH literature and a *Cognitive Status Filter* model designed to more flexibly handle uncertainty. The CSF is a statistical per-entity model of cognitive status in which a *Cognitive Status Engine* comprised of per-entity *Cognitive Status Filters* is used to maintain a distribution over cognitive statuses for each entity the robot believes to be at least familiar to all parties within the conversation. The models are trained and evaluated using a silver-standard [1] English subset of the OFAI Multimodal Task Description Corpus [13]. Specifically, the CSF seeks to predict the cognitive status for a given entity based on whether and how it has been referenced in natural language.

Such a computational model of cognitive status would give linguistic insights that can be used for the refinement of the GH itself, thus, being beneficial to different fields of study such as linguistics, cognitive psychology and designing of robots.

[1] This subset constitutes English transliteration of originally German dialogues.

Second, we demonstrate how the GH might need to be used quite differently to facilitate robot anaphora generation. Once we have the cognitive status information of a given entity, we can leverage this to select the appropriate *referring (pronominal)* form that can be used to refer to that entity. This is achievable because the GH indicates a strong connection between the assumed cognitive status and the associated referring form which has been validated through experimental studies as shown in Rosa and Arnold [14]. Specifically, we present a machine learning approach (using a decision tree classifier model) to select a referring form (*it*, *this*, *that*, $this\langle NP\rangle$, $that\langle NP\rangle$, $the\langle NP\rangle$) for a target referent given the information about that target entity as well as information about some set of distractors (other entities present in the same context and having at least the same cognitive status or higher as the target referent).

A GH-theoretic model of language generation would be an important contribution to the GH-theory itself. Thus, providing insights to linguists and cognitive psychologists. It will also help in advancing the state of the art of language-capable social robots, and across different fields of study such as robotics and natural language processing.

The remainder of this thesis work is organized as follows. In Chapter 2, we discuss related work. Next in Chapter 3, we present our first study, Givenness Hierarchy theoretic cognitive status filtering [15]. In Chapter 4, we present our second study, Givenness Hierarchy theoretic referring form selection. Finally, in Chapter 5, we discuss conclusion and possible future work.

CHAPTER 2

RELATED WORK

2.1 Theories of reference

An important aspect of linguistic behavior is the way in which humans communicate with each other referring to entities that are relevant in the current context. Almost every utterance that we generate involves referring to something, or identifying something. This is done through the generation and interpretation of referring expressions like proper nouns (like John, Peter), definite or indefinite descriptions (like the table, a box), demonstratives (like this, that, this building) or pronouns (like it, he, she).

Every time we speak we need to make a decision on how to refer to entities, that is, whether to use a full description or a shorter expression. If we decide to use a full description then additionally, we need to make a choice of what attributes and relations to use in that description. Thus, a challenge for language researchers is to understand and explain how speakers choose an appropriate referring form given the fact that there are so many options available.

The centering framework presented by Grosz et al. [16] uses the concept of linguistic focus and investigates the interactions between the focus of attention and the choice of referring expression and discourse coherence (where discourse coherence depends on centering, that is, a system of rules and constraints that define the relationship between discourse context and the choice of referring form).

Extending on the work by [16], Brennan et al. [17] presented a centering approach to the focus of attention in discourse and leveraged it to use as a foundation for their algorithm for tracking discourse context and binding pronouns.

These approaches used the concept of linguistic focus of attention and whether someone uses a full description or a pronoun like “it” was based on whether the target entity was in

the focus of the conversational context.

Ariel [18, 19] in contrast, used the concept of accessibility of information (discourse anaphora) being specified in the memory in varying degrees. The author argued that the speakers choice of referring expressions (such as definite descriptions, pronouns, demonstrative pronouns) marks different degrees of accessibility. Highly reduced forms like pronouns are reserved for the most accessible information and more complex referring expressions are associated with less accessible referents.

In a similar approach, Bard et al. [20] used the accessibility theory to investigate mentions of shapes in a collaborative task between dyads to analyse the effects of access on the listener's attention, player's actions and speaker's roles. Their experimental results demonstrated that speaker's actions affected the form of referring expression and only same role dyads correlated attention to accessibility theory.

These approaches used the concept of linguistic focus of attention and whether someone uses a full description or a pronoun like "it" was based on whether the target entity was somewhat salient within the discourse context.

The main approach to investigating theories of reference has been to recognize the underlying conditions or mechanisms that directs speakers to use different referring forms. As viewed by Levelt [21] several authors suggest that this linguistic choice of different forms is made based on the target entity's cognitive status (also known as information status or discourse status). These models share the idea that some information are more important and thus salient, or accessible or in in focus.

In our next section, we describe one such theory that takes into account the cognitive status of an object to determine how to refer to entities.

2.2 The Givenness Hierarchy

One of the most popular and established theories of reference is the Givenness Hierarchy, originally presented by Gundel et al. [7]. It consists of a nested hierarchy of six tiers of cognitive statuses: $\{$*in focus* $\subseteq$ *activated* $\subseteq$ *familiar* $\subseteq$ *uniquely identifiable* $\subseteq$ *referential* $\subseteq$

type identifiable}, each of which is associated with a set of referring (or pronominal) forms that can be used when referring to an entity with that status [22, 23] as shown in the following Table 2.1.

Table 2.1: Cognitive Statuses and the corresponding referring forms

Cognitive Status	Refering form
in focus	it
activated	this, that, thisNP
familiar	thatNP
uniquely identifiable	theNP
referential	indefinite thisNP
type identifiable	a NP

The "in focus" status at the top is the most restrictive search space for the expression type's referent, and "type identifiable" at the bottom, the least. The hierarchical nesting here means that an entity with one status can also be said to have all other statuses lower in the hierarchy. If a target referent is *in focus*, for example, it can also be inferred to be activated, familiar, and so forth. Accordingly, a speaker's selection of a pronominal form depends on their assumptions as to the cognitive status of their target referent in the mind of their conversational partner.

For example, if a speaker uses *it* to refer to an object, the listener can infer that the object being referenced must be one that is already *in focus*, whereas if a speaker uses $that\langle NP\rangle$, the speaker can only infer that the object is at least familiar (but may in fact be activated or even in focus), but a higher status cannot be inferred from a lower status. The GH model suggests that within a conversation, some information might be more salient than others, thus "in focus" of attention and occupying a prominent position within the mind and attentional state of the interlocutor. Moreover, it also represents the fact that the more precise/reduced referring forms are reserved for the highest status. But, due to the hierarchical nature of the GH, a speaker is also able to use more explicit descriptions for referring to an object that is "in focus".

The hierarchical structure of the GH is also important due to the way it parallels the hierarchical nesting of models of human memory, such as Cowan [24]'s, in which the focus of attention is a subset of short-term memory (or working memory), which is in turn a subset of long-term memory.

The GH *coding protocol*, presented by Gundel et al. [22], provides guidelines as to what features of linguistic and environmental context should dictate the cognitive status of a given entity. For example, this protocol suggests that an entity that is mentioned in a topic role in the preceding clause should be considered to be in focus, and that any entity that is mentioned at all should be considered to be at least activated [22, 23]. For example, if an object is the main topic of a sentence (one of the requirements of the protocol for an object to have the "in focus" status), then the linguists say it should be regarded as *in focus of attention* within the current conversation and will most likely be referred to by *it* in the following sentence.

Due to the GH's popularity within the research literature, and its validation across a wide variety of languages beyond English [25], many researchers have sought to computationally implement it in whole or in part, especially within the context of reference resolution algorithms. Kehler [26], for example, use the GH to justify an approach in which elements of an interface that are highlighted are considered to be "in focus", and referring expressions that use pronominal forms are automatically resolved to those highlighted referents. Building on this work, Chai et al. [27] proposed a probabilistic graph-matching algorithm for resolving referring expressions that are complex (involving multiple target referents) and ambiguous (involving gestures that could indicate multiple candidate referents) in multimodal user interfaces. Because this algorithm had high computational complexity, Chai et al. [28] demonstrated how the algorithm's performance could be improved using a greedy algorithm based on the theories of Conversational Implicature [29, 30] and the GH. Chai et al. combine these theories to create a reduced hierarchy: $Gesture \subseteq Focus \subseteq Visible \subseteq Others$, where Focus combines the "in focus" and "activated" tiers of the GH, and Visible combines

its "familiar" and "uniquely identifiable" tiers. When a referring expression is processed, the relationship between referring form and status is then used to help resolve that referring expression.

Finally, while the approaches above focused on modeling of reduced versions of the GH, Williams and Scheutz [8], Williams et al. [10] instead presented an implementation of the full GH, through a set of rules that associated different referring forms with different sequences of actions involving all six tiers of the GH. They demonstrated how this approach better enabled NLU in uncertain and open-world scenarios. This required, in part, four data structures corresponding to the top four tiers of cognitive statuses of the GH, while the last two tiers were instead associated with new "mnemonic actions" such as creating new mental representations [8].

2.3 Natural Language Generation

According to Reiter and Dale [31], Natural Language Generation (NLG) is a sub-area of artificial intelligence and computational linguistics that is concerned with the building of software systems and the automatic production of high-quality written or spoken content in either English or other human languages.

The general structure of a NLG system consists of different sub-processes. The typical stages of a NLG system are as follows [31]:

1. Content Determination - It is the task of deciding what information should be communicated in the output document.

2. Document Structuring - It is the task of deciding how chunks of content should be grouped in a document and how different chunks should be related in rhetorical terms.

3. Lexicalisation - It constitutes the task of deciding what specific words (or other linguistic resources, such as particular syntactic constructions) should be used to express the content selected by the content determination component.

4. Referring Expression Generation - It is concerned with deciding what expressions should be used to refer to entities.

5. Aggregation - It is the task of deciding how the structures created by document planning should be mapped onto linguistic structures such as sentences and paragraphs.

6. Linguistic Realisation - It is the task of converting abstract representations of sentences into the real text; it corresponds to the content aspect of surface realisation.

7. Surface Realisation - It is the task of converting abstract structures such as paragraphs and sections into the mark-up symbols understood by the document presentation component; this corresponds to the structural side of surface realisation.

The primary uses of an NLG system has been to output textual data, while the input can be of various forms. In applications such as automated machine translation and text summarization, the system generates text based on other, generally human-written text. Other applications, where the input to the NLG system is non-linguistic are weather or medical reports, data from environmental sensors, etc.

Applications in the area of NLG also includes automatically generating texts based on visual data such as static images and videos. In their work Kulkarni et al. [32] propose a language generation system that automatically generates natural language descriptions from images. Their system functions through two steps: first is content planning and second is surface realization and demonstrates that their system is more effective in generating relevant text for images than previous work.

Similarly, Khan and Gotoh [33], Kojima et al. [34] worked with the generation of natural language descriptions for human behavior, actions and relations with other objects as observed through video data. To achieve this, they used traditional image processing techniques to extract high level features (such as semantic features) from the videos and converting them into text using context free grammar. They evaluate their system using both quantitative and task based evaluations using human participants.

Recent work on NLG has also focused on replacing template and rule based techniques with data-driven approaches where systems are trained on large datasets as in [35–37] and the system automatically learns the rules from the data. In their work Krishnamoorthy et al. [38] presented a data-driven approach to automatically generate natural language descriptions from video data using the concept of text-mining. They demonstrated that their approach can be used to annotate even random videos without the requirement of previous training on similar video data.

2.3.1 NLG in social robotics

Another crucial application of NLG is to generate linguistic output that can be used in interactive systems, such as in a text-based chatbot, a spoken dialogue system or language capable interactive robots such as Nao, Pepper, etc. Currently, language interactions in realistic HRI scenarios involve either using "Wizard-of-Oz" techniques or template-based approaches or prerecorded speech.

Social robotics is an important application area (as described in the previous chapter) where humans and robots work as partners and share the same physical space while carrying out different tasks such as navigating an environment or referring to objects and locations.

Much of the work involving NLG and social robotics primarily focuses on referring expression generation (REG) as depicted through the GIVE challenges [39], where researchers competed to develop NLG systems that can generate instructions for helping human users navigate and interact with objects in a virtual world.

Though most of the work in NLG revolves around REG, it is worth noting that real world social interactions between humans and robots are not just restricted to referring to objects and "filling slots" for queries such as flight booking, but involves more human-like social interactions including linguistic output and non-linguistic factors such as behavioral aspects (gaze, gesture, prosody, etc.) [40].

For example, in their real-world project on HRI Foster et al. [41], had the goal of placing the language capable Pepper robot in a shopping mall where it is expected to socially

and intelligently engage with people and entertain them while carrying out various tasks. However, in their current work, Papaioannou and Lemon [42] emplys a task-based chatbot system that uses a template based approach for interactions.

Thus building efficient NLG systems for being used in the area of social robotics remains a challenge and as we will see in the next section real world interaction scenarios also involve humans using pronouns or shorter expressions as opposed to explicit descriptions in their daily conversations, which we also try to address through our work in this thesis.

2.4 Anaphora Generation

Anaphora essentially involves referring back to an entity previously mentioned in a discourse, a written text or within the context of a conversation using referring forms such as "he", "it", "that", etc. This referring back to an entity could be either within the same sentence (intrasentential anaphora) or within subsequent utterances (intersentential or discourse anaphora). For example, "I saw Paul. *He* was in a hurry." (intersentential) and "I kept my glasses on the table but cannot find *it* now." (intrasentential).

In their survey paper, Arnold and Zerkle [43] describe two broad categories of language production models, specifically pronoun generation, based on how the model treats the pronoun production process - as highly focused entities that the listener can easily retrieve from his/her memory (*pragmatic* models) or as producing easy short expressions rather than longer referring expressions to reduce the cost of production from the speaker's point of view (*rational* models). As discussed by Arnold et al. in their article [44], cognitive status information plays an important role in reference because people tend to use different referring forms depending on the current context of their conversation, their shared knowledge, background information, or whether introducing new entities or talking about previously mentioned entities.

Arnold and Nozari [45] investigate the relationship between the start/production of an utterance and the choice of referring form and accordingly devised a study to recognize the time elapsed between observing an action and verbally describing it which will provide

information on how much time is spent on pre-planning the utterance before speaking. They specifically worked on pronoun generation and did find a relationship between longer latencies and generation of pronouns.

In Fukumura's work [46], the author investigates the role of language production by speakers while taking into account the listener's perspective so that the listeners were able to easily recognize their intended target referents. The author conducted the study with both linguistic and visual context and observed that speakers tend to use more general referring forms for linguistic contexts than visual and specifically for pronoun generation speakers tend to use less pronouns when they had the knowledge of their listeners not being able to hear the linguistic antecedents.

In their work by Jaeger & Levy [47], the authors provides evidence on speakers choice of language production being dependent on optimizing the information communicated by speakers through syntactic reduction. Specifically, they demonstrate that there is an inverse relationship between redundant information and the choice of explicit words, including the use of "that" in "relative clauses". Depending on whether information is more or less predictable speakers tend to use less or more functional words respectively.

While a vast amount of research has been performed on referring expression generation, relatively little attention has been paid to anaphora generation [48]. Previous work on anaphora generation has explored factors such as discourse structure, coherence, and salience [49][50][51][52][53], and Ge et al. [52], for example, present a probabilistic model for generating the pronouns such as "it", "he", and "she" based on factors such as the distance, gender, and noun phrase repetition. However, these works directly associate these low-level features with pronoun usage, rather than using cognitive status as a mediating factor as suggested by the linguistics literature.

Moreover, these works primarily concentrate on either textual or simple visual domains rather than realistic human-robot interaction domains where both linguistic and non-linguistic factors such as gaze, gesture, and physical proximity, all of which might affect the cognitive

status of an entity. We thus believe that a GH-theoretic approach to anaphora generation would present a powerful new approach for robotic domains and enable valuable new language generation models for both psycholinguists and computational linguists.

The GH's claim that the choice of a particular referring form depends on the presumed cognitive status of an entity in the listener's mind [14] suggests that the GH should be a powerful tool for deciding when and how to use anaphoric expressions. Naively, one might assume that they could straightforwardly use the fact that each tier of the GH is associated with a different set of referring forms that can be used to refer to objects assumed to have that status (or higher).

However, because progressively lower tiers become associated with progressively greater numbers of possible referring forms, and because forms like "it", "this", and "that" can be ambiguous, these tier-form associations are not sufficient on their own for language generation. Indeed, although the knowledge of cognitive status affects language understanding and generation, no significant linguistic work exists that identifies the mechanisms by which it is actually *used* for selection of referring forms [54].

Accordingly, a contribution of this thesis work is to fill this gap in the literature by proposing a GH-theoretic model of natural language generation that leverages the GH-theoretic cognitive status modeling.

2.5 Anaphora Resolution

Anaphora resolution is the task of resolving potential antecedents for anaphoric expressions and it is well known that syntactic, semantic and pragmatic factors all play a key role in resolving the question of which of the many potential antecedents for a particular use of a pronoun is intended by the speaker as the antecedent.

One of the earliest work on pronominal anaphora resolution is by Lappin and Leass [55] where they present their algorithm RAP (Resolution of Anaphora Procedure) that resolves the pronoun interpretation of entities on textual domain. Specifically the algorithm maintains a list of all possible antecedents of a particular anaphora. Each antecedent was

associated with a salience value depending on different features such as head noun, subject, recency of mention etc. and the antecedent with the maximum salience value was selected as the preferred antecedent.

One of the classic theories of reference resolution is by Hobbs [56]. The author presents two approaches to solve pronoun references on the context of textual data. The first approach involves creating surface parse trees of each sentence in a text and then traversing the tree by using a simple algorithm that follows a predetermined order to search through the tree in order to resolve the likely referent of the pronoun. The second approach describes a system for "semantic analysis" that generates the referents of pronouns for "free" as a consequence.

Another classical approach to pronoun interpretation is the Centering Theory by Gordon et al. [57] which emphasizes on the fact that pronoun resolution depends solely on the grammatical structure of sentences and how information transitions from one sentence to another, that is, whether the focus entity of the current utterance is mentioned in the next utterance and what its grammatical role is in the current as well as next utterance.

Kehler and Rohde [58] attempted to merge classical pronoun resolution theories by Hobbs and Grosz by performing multiple psycholinguistic experiments. The experiments revealed that both coherence-based and centering-based approaches have their definitive roles to play in pronoun interpretation. Thus, the authors combined both approaches in a simple Bayesian probabilistic model whose conditional probability bias depends on both these factors.

Another rule-based algorithm for "automatic pronominal anaphora resolution" is that by Liang and Wu [59]. As mentioned by the authors the algorithm leverages the WordNet ontology and heuristic rules to perform anaphora/antecedent resolution by following the procedure of parsing, POS tagging, etc. for each sentence in an English text and storing these features alongwith other global features such as base nouns, gender, etc. This model also implemented a finite state machine for identifying noun phrases. Parsed sentences were checked for "anaphoric references" and everything else was considered as possible antecedents that were evaluated using a scoring function.

A recent work by Song and Kaiser [60] investigates into how to correctly interpret the antecedents of subject-position pronouns by performing two experiments and hypothesizes that to correctly interpret the referent of a pronoun we need to consider information that is available both before and after the pronoun is encountered. Specifically, in their two experimental studies, they worked with textual data consisting of particular verbs ("nonce" verbs) and how those verbs when used with one pronoun, two pronouns and one pronoun+name containing sentences caused the participants to resolve the correct referent of those pronouns. As expected, the authors did find that successful pronoun interpretation depends on both pre and post pronominal information [60], choice of verb and whether the pronoun is on subject position only or both subject and object positions.

Another recent work is by Nakos et al. [61] which is an extension of their previous work on reference resolution. Their work is based on the concept that humans often use a "two-stage process" to reference resolution first depending on its own knowledge and then rectifying errors through common ground consideration. The authors discuss their "Analogical Reference Resolution" model that takes into account the first part of the two-stage process relying heavily on the similarity between the target object and the hearer's own knowledge of such an entity. For the second part of the process they extend their model to account for the corrective measures depending on common ground between the speaker and the hearer.

While most of work on anaphora resolution algorithms has been done on either personal pronouns (he, she) or definite descriptions (the table, the cow), demonstrative pronouns (like this, that), play an important role in anaphoric references as well, since they appear frequently in text or within conversations and refer to integral content. Among other theories, as mentioned in the previous section, according to the GH, demonstrative expressions might indicate an entity to be "activated", which according to the coding protocol [22] informs the addressee that the entity has been recently mentioned or is immediately accessible outside the linguistic context.

In their work Brown-Schmidt et al. [62] investigated their hypothesis regarding "it" being used when referring to the most salient entity in a discourse and "that" referring to a conceptual composite. They performed three experiments, where in the first experiment participants followed spoken instructions while their eye gaze was monitored. In the first they found that people prefer to use "it" for single objects like a cup, and "that" when speaking about the cup and the saucer together. However, the main findings from all the three experiments demonstrated that entities without linguistic antecedents are sometimes preferred over entities with linguistic antecedents and a single factor like salience is not enough to account for the speaker's choice of differences referential forms [62].

In our next chapter we present a computational model of cognitive status for natural language generation as informed by the GH.

CHAPTER 3

GIVENNESS HIERARCHY THEORETIC COGNITIVE STATUS FILTERING

Reproduced with permissions from Cognitive Science Society [2]. Poulomi Pal [3,4], Lixiao Zhu[4], Andrea-Golden Lasher[4], Akshay Swaminathan[4], and Tom Williams[4]

3.1 Motivation

Robots capable of interacting with humans using natural language must be able to properly communicate about its surroundings for efficient collaboration. Since humans commonly use pronouns than longer referring expressions to refer to objects in their environment, robots must also be able to understand and generate pronouns for successful interaction. According to the linguistic theory of the Givenness Hierarchy (GH), humans use shorter forms due to the presumed cognitive status of an entity in the mind of their interlocutor.

In previous computational implementations, the GH is used to justify a set of data structures used to store representations for entities that could be referred to, and to justify which of these data structures should be considered (and how) when a given referring form is used. However, while this is sensible during natural language understanding, it may not be appropriate for the purposes of natural language generation.

During generation, the speaker already knows what object they wish to refer to, and do not need to search through these sorts of data structures. Instead, when a speaker decides what referring form to use to refer to a given object, we argue that they would instead start by determining the status of that object, and only then may they look through the data structure associated with that status, in order to determine what distractors must be ruled out. Also, using these data structures alone would require the speaker to search through

[2]Proceedings of the 42nd Annual Conference of the Cognitive Science Society (pp. 925-931),

[3]Corresponding author. Direct correspondence to poulomipal@mymail.mines.edu
[4]MIRRORLab, Colorado School of Mines, 1600 Illinois Street, Golden, CO 80401 USA

each data structure to identify whether their intended referent was a member of that set; a process requiring $O(\mid E \mid)$ worst-case time complexity for set of known objects E. Instead, what is needed during natural language generation (in *addition* to these data structures, is a means of quickly determining not what entities have a given cognitive status, but what cognitive status is most likely for a given entity.

Accordingly, in the next section 3.2 we propose an approach to this problem of quickly determining the cognitive status of an entity, which we term as *cognitive status modeling*.

3.2 Problem Formulation

We formulate cognitive status modeling as a Bayesian filtering problem. Let a dialogue consist of a set of utterances $U_0, ..., U_n$. For object o, let $S_o^t \in \{I, A, F\}$ denote the cognitive status of o at a particular timestep t after utterance U_t (either In Focus, Activated or Familiar), and let $L_o^t \in \{N, M, T\}$ denote the linguistic status of o in utterance U_t (e.g., either not mentioned in the utterance, mentioned in the utterance in a non-topic role, or mentioned in the utterance in a topic role). Using this formalism, our goal is to recursively estimate, for a given object, the probability distribution over cognitive statuses for object o at time t:

$$p(S_o^t) = p(S_o^{t-1})p(L_o^t)p(S_o^t \mid S_o^{t-1}, L_o^t) \tag{3.1}$$

We define a Bayesian filter of this form as a *Cognitive Status Filter* (CSF) for a given object o. Given a set of known objects, $O = \{o_1, ..., o_n\}$, our goal is then to estimate this distribution for each $o \in O$ at each time step. To do so, we use a Cognitive Status Modeling Engine C, consisting of a set of CSFs $\{c_0, ..., c_1\}$, one for each object believed to be of a status familiar or higher within the conversation. Here, we make the simplifying assumption that the same set of objects are known to both the robot and its conversational partner, meaning that the set of all objects with status *Uniquely Identifiable* or higher is simply the set of objects O. We assume that it is straightforward to determine whether one of these objects is or is not *Familiar* based on whether or not it has appeared in the current conversation.

This allows us to model whether or not an object is of status Familiar or higher based on whether or not a CSF $c \in C$ exists for that object, and to model *which* of those statuses the object likely has, using its associated CSF. Finally, we estimate whether each of the higher statuses (Activated, In Focus) hold for the object using the CSF c.

3.3 Data Collection

The core component of our CSF model that must be learned ahead of time is the conditional probability $p(S_o^t \mid S_o^{t-1}, L_o^t)$. To learn this, we trained our model using a silver-standard English translation of the German OFAI Multimodal Task Description corpus [13]. The corpus represents a collection of human-human and human-robot interactions where the human teacher shows and explains to a human or robot learner how to connect two separate parts of a tube and then how to mount the tube onto a box with holders, as shown in Figure 3.1 by actually moving around the objects and performing the task while explaining it to the learner.

The average length of a sentence that is used in this corpus has 8-9 words. As the name suggests, since the corpus is "multimodal", the corpus contains both verbal and non-verbal cues such as speech, gaze, and gestures. Realistic multimodal HRI scenarios require the use of such non-verbal cues; with our uncertainty sensitive model, however as our first step we begin in this work by looking only at our model's ability to handle the same kind of linguistic factors (even a simple subset of the linguistic factors) that are handled by the GH, leaving the ability to model other linguistic factors for future work.

While the OFAI MTD corpus contains data from four task scenarios, we only use the data from one particular task scenario (Task 3). The original dataset for this task consists of 16 monologues each having approximately 4 to 5 utterances. As a first step, in this work we begin by evaluating our model on a small subset of the original dataset, consisting of 4 of these monologues, each of which is comprised of just 4 utterances, to control for monologue length. As shown in Figure 3.1, this task context contains 8 objects, including the learner and teacher.

Task 3 was selected because it includes a larger number of objects than the other tasks in a dyadic instruction context, and contains data from both human-human and human-robot dyads. Specifically, Task 1 involved a human teacher explaining and performing a task in front of the camera without the presence of a learner in the scenario; Task 2 involved a human teacher and a human learner jointly performing the task of moving an object; and Task 4 is a pure "navigation task" involving both human-human and human-robot dyads [13].

3.3.1 Appearance Feature Annotation

To collect linguistic status information L, three annotators independently annotated the OFAI Multimodal Task Description Corpus [13] according to the following annotation procedure.

- Each annotator was provided a printed copy of all 16 monologues to annotate.
- For each sentence in each monologue, the annotator was instructed to underline any piece of the text that could refer to some object in the scene.
- For each of these underlined pieces of text, the annotator was instructed to indicate the correspondence between the underlined sentence fragment and the object in the scene it referred to.
- Finally, the annotator was required to circle the fragment-object mapping they believed to be the topic of the sentence.

There were a few cases in which annotators circled multiple objects as the topic of the sentence; in these cases, both objects were recorded as being equally probable topic referents[5].

[5]The inter-annotator agreement score as measured through Fleiss' Kappa was $\kappa_n = 0.37$, indicating fair agreement between annotators. It will be important in future work to adapt the annotation protocol to increase rate of agreement.

3.3.2 Cognitive Status Annotation

Ground-truth cognitive status information was then collected through a crowdsourced human-subject study. 160 US participants were recruited from Amazon Mechanical Turk. Two participants answered an attention check question incorrectly and were dropped from our analysis, leaving 158 participants (71 female, 85 male, 2 N/A). Participant ages ranged from 19 to 70 years (M = 35.03, SD = 11.36). Each participant was paid $0.25 for completing the study.

3.3.3 Procedure:

At the beginning of the study, each participant is shown the scene depicted in Figure 3.1, and is instructed to remember the objects and their labelings in order to performing their upcoming task. Participants were then shown the same scene without labels while listening to a portion of one of the study's four monologues, as read by the experimenter. Specifically, participants were randomly assigned to hear a random prefix of a randomly selected monologue (i.e., either only the first utterance of that monologue, the first two, the first three, or all four).

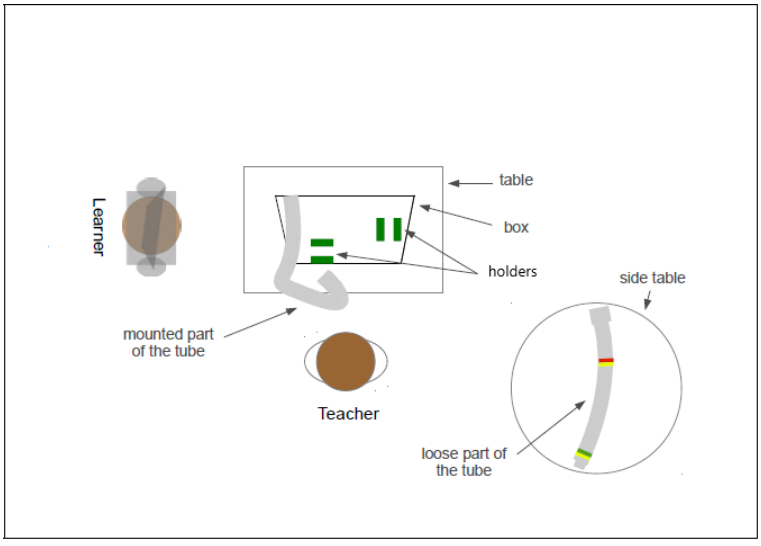

Figure 3.1: Scene (labeled)

At the end of this monologue excerpt, participants were asked to answer two questions, presented in a randomized order, with the second question becoming available after the first question was answered. The two questions are as follows:

- Q1: *Click on the object in the scene that you think the speaker would most likely be referring to if the speaker would have said "look at it" at the end of the monologue.*
- Q2: *Click on all the objects in the scene that you think the speaker would most likely be referring to if the speaker would have said "look at that" at the end of the monologue.*

Two of the monologues used in our study are as shown below.

- Monologue 1:

  U1: You must take the tube with your right hand.

  U2: And insert it in at the yellow-green connection here.

  U3: Put it on the tube.

  U4: Again, with your right hand insert it here in the holder.

- Monologue 2:

  U1: With the right hand stick the two tubes together.

  U2: You put that together here with the yellow-green mark.

  U3: It is okay that it is not holding firmly.

  U4: Now lead the one tube through here.

These questions allowed us to probe the user's implicit beliefs as to the cognitive status of the objects in the scene. From a GH-theoretic perspective, if a participant implicitly believed a given object to be in focus, they should click on that particular object for both Q1 and Q2, whereas if they believed the object to be activated, they should click on that object for Q2 but not for Q1. Because the context is narrowly defined and participants were

given time to examine each object in the scene, we assume that all objects in the scene are familiar or higher. Thus, if a participant believed the object to be familiar or lower, they should not click on the object at all. After completing the task, participants completed a check question (cf. Schreitter and Krenn [13]) requiring users to identify the scene they had viewed from among several distractors. This allowed us to ignore data from participants who did not pay sufficient attention while completing the task.

Using this coding procedure, we are thus able to determine the perceived cognitive status of each object in the scene for each participant after the completion of the monologue excerpt they were exposed to. When paired with the linguistic status annotations, this allowed us to train our CSF model, using the procedure described in the following section.

3.4 Training and Evaluation

3.4.1 Training

After collecting this dataset, our CSF was trained in the following way: First, we initialized a 9x3 matrix whose rows correspond to the nine cognitive/linguistic status pairs an object could have at time $t-1$ ((I_{t-1}, N_t), (I_{t-1}, M_t), (I_{t-1}, T_t), (A_{t-1}, N_t), (A_{t-1}, M_t), (A_{t-1}, T_t), (F_{t-1}, N_t), (F_{t-1}, M_t), (F_{t-1}, T_t)), and whose columns correspond to the three cognitive statuses that object could have at time t (I_t, A_t, F_t).

For each pair of adjacent utterances in each monologue (U_{t-1}, U_t), we consider the data from all participants (for all objects) who provided data immediately following utterance U_{t-1}, and from all participants who provided data immediately following utterance U_t. For each resulting pair of datapoints, we identify and increment the correct cell in this matrix.

For example, for the combination of a datapoint from a participant who heard some utterance and subsequently viewed that object as in focus, and a datapoint from a participant who heard the next utterance in the same monologue, containing object 1 in a non-topic role, and at that point viewed the object as being activated, we would increment the cell $((I_{t-1}, N_t), A_t)$. Once all data has been considered, we normalize each row of this table to produce a conditional probability table.

3.4.2 Evaluation

To evaluate our CSF model, we then considered each object o and each monologue M, and retrained our model using all data except that which was collected for object o or monologue M (for example, while testing for object o_1 in monologue M_1, we retrain our model with all the data except that concerned with M_1 and/or o_1), and used this model (along with a prior distribution over cognitive statuses for that object as described below) to simulate what status would be predicted for that object at each point in that monologue.

After each of these utterances, we evaluated the model's prediction by comparing it to the majority opinion from participants who *had* provided data for that object at that point in that monologue. Combining these prediction results for all eight objects in all four utterances in all four monologues produced a 128-element prediction vector for the model. Specifically, we computed these prediction vectors for each of two CSF models, each of which used a different prior distribution $p(S_o^{t-1})$ over cognitive statuses:

- U-Model: an *uninformed* prior in which each cognitive status was assigned a prior probability of 0.33.

- I-Model: a (weakly) *informed* prior, in which the three cognitive statuses were assigned prior probabilities I= 0.05, A= 0.1, F= 0.85. These probabilities reflect the fact that objects are a priori far more likely to be familiar than activated, and among the set of things that are currently activated it is more likely for a given object to be activated than in focus.

  While in theory this distribution could be learned from data, in a realistic environment it may be the case that hundreds or thousands of objects are familiar and only one is in focus, yielding an extremely unbalanced distribution. This weakly informed prior thus represents an *optimistic* belief state in which the prior probability of any given object being in focus is artificially boosted.

In addition to these two prediction vectors produced by different parameterizations of our CSF model, we also computed prediction vectors for two baseline models:

- Finite State Machine: First, we computed the decisions made by a rule-based finite state machine (FSM) model, which formalized a set of heuristics from the GH coding protocol (the same heuristics previously used in the work of [8]). In this FSM, the states correspond to cognitive statuses, and transitions are triggered based on linguistic statuses observed in incoming utterances. For example, for an FSM dedicated to some object, if that object is mentioned in a topic role, this will deterministically trigger a state transition to *in focus*.

- Random Baseline: Second, we computed the decisions made by a random baseline (RB) model, which predicted cognitive statuses at random.

3.5 Results

The overall accuracy of each model (i.e., the proportion of correct entries in each model's prediction vector) is shown in Table 3.1. This demonstrates that our U-model had the highest accuracy, and that our I-model and the theoretical FSM model had the same accuracy, slightly less than the U-model.

The accuracy measure of the FSM model suggests that the heuristics encoded in the GH coding protocol are a good representation of the patterns that can be learned from the data we collected, given our choice of data annotations. The similarity of the CSF model's accuracy to that of the FSM similarly demonstrates that the CSF did a good job of automatically learning these patterns from our data.

The slightly higher accuracy of the U-model over the I-model suggests that the uniformly distributed prior probabilities may have been more helpful than the weakly informed prior distribution.

Finally, the performance advantage of all of these models over the RB model provides a good baseline measurement of success.

Table 3.1: Accuracy measure of each model

model	accuracy
U-model	82.03
I-model	81.25
FSM	81.25
RB	32.81

To validate these intuitive assessments, we formally compared our four models using six pairwise McNemar's Tests [63, 64], whose results are shown in Table 3.2 and Table 3.4.

Table 3.2: Contingency Table entries for model pairs

$model_1$	$model_2$	N_{ss}	N_{sf}	N_{fs}	N_{ff}
U-model	I-model	104	1	0	23
U-model	FSM	89	16	15	8
U-model	RB	34	71	8	15
I-model	FSM	89	15	15	9
I-model	RB	33	71	9	15
FSM	RB	34	70	8	16

Table 3.2 (see also Figure 3.2) shows the contingency table values used by McNemar's test for each pairwise comparison, where the four N counts refer to the contingency table cells shown in Table 3.3. That table layout simply depicts a general 2x2 contingency table [65, 66] comparing the performance of two models A and B. Here, N_{ff} and N_{ss} respectively denote the number of instances where both models failed and succeeded. N_{fs} and N_{sf} respectively denote the instances where one model failed and the other succeeded.

Table 3.3: A 2X2 Contingency Table

	model A success	model A fail
model B success	N_{ss}	N_{sf}
model B fail	N_{fs}	N_{ff}

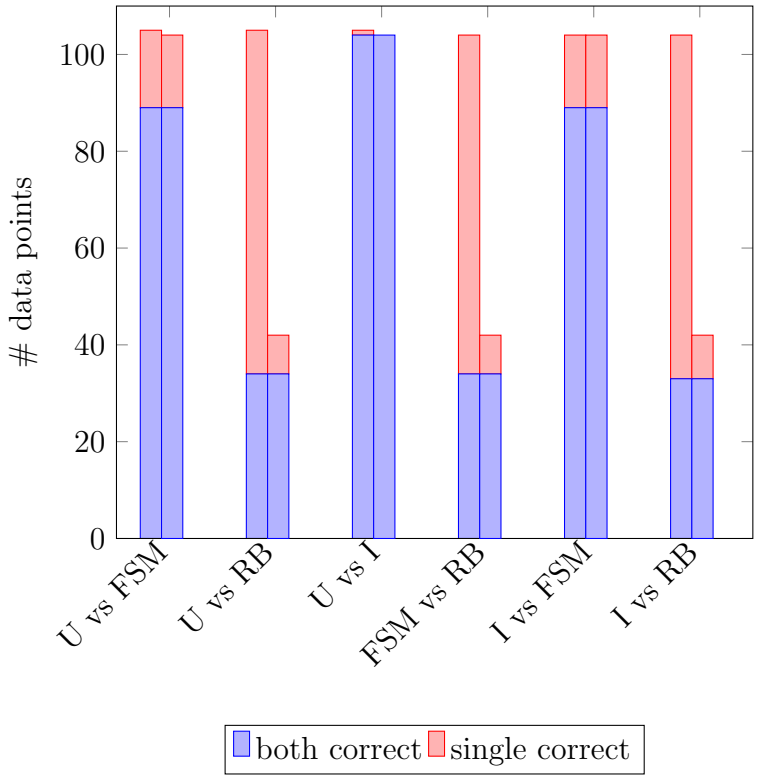

Figure 3.2: Comparison between models

The McNemar's Test statistics χ^2 (with 1 degree of freedom) and p-values [64, 65, 67] are calculated for each pair of models as shown in Table 3.4. By looking at the McNemar's Test results the following deductions can be formally made:

1. The U-model and I-model show similar performance ($\chi^2 \approx 0$ and p-value = 1).

2. The U-model and FSM also show similar performance ($\chi^2 \approx 0$ and p-value = 1).

3. The FSM and RB models show significant difference in their performance ($\chi^2 = 47.705$ and p-value = 0.0001).

4. The CSF model and RB model differ significantly in performance regardless of model parametrization.

5. The performance difference between the CSF model and the FSM model is not statistically significant.

Table 3.4: McNemar's Test statistic (χ^2) and p-values

	χ^2	p-value
U-model, I-model	0.000	1.000
U-model, FSM	0.000	1.000
U-model, RB	48.658	<0.0001
I-model, FSM	0.033	0.8551
I-model, RB	46.513	<0.0001
FSM, RB	47.705	<0.0001

3.6 Conclusion

This chapter concludes our first study where we present the notion of a *Cognitive Status Filter*: a statistical model for estimating the cognitive status of some entity that may be referenced in conversation. We then described a Mechanical Turk experiment used to gather ground truth data to train this model, and demonstrate how the accuracy of this model compares to a rule-based FSM model and a random baseline.

The overall accuracy of our CSF model in predicting the cognitive status of an object was slightly better than that achieved by a FSM. This simultaneously speaks in favor of the heuristics encoded in the GH coding protocol, while also demonstrating that those heuristics can be learnt from data. However, there are a number of directions for future work that may significantly improve the potential performance of the statistical CSF model over the rule-based FSM model.

3.7 Future Work

First, this experiment used a relatively small corpus collected in a single task; given the fact that our model works on this small dataset one follow up step would be to collect a larger dataset from a broader set of HRI scenarios (preferably a gold-standard English corpus), as that could yield a model with better generalizability.

Second, our CSF model currently only uses linguistic status information that is already explicitly called for by the subset of the GH coding protocol used to design the FSM model.

However, the CSF model could straightforwardly be extended to include additional non-linguistic cues like gaze and gesture which are critical in both human-human and human-robot communication (e.g., for establishing joint attention [68, 69]), which although not well described in the GH coding protocol would clearly play a role in informing notions of cognitive status.

Similarly, we considered only three simple linguistic features (topic mentioned, mentioned, and not mentioned) given by the GH coding protocol, whereas more complex and varied linguistic features could improve performance. Finally, one of the theoretical advantages of the CSF model is its ability to handle uncertainty. This will be critical for integrating gaze and gesture, which are inherently ambiguous and uncertain cues.

Finally, in future work, we intend to leverage our CSF model to implement a GH-theoretic anaphora generation model that uses an object's cognitive status when selecting a referring form during natural language generation. We further plan to integrate this model into the DIARC cognitive robotic architecture [70] and demonstrate its use in realistic HRI scenarios.

3.7.1 Limitation

In addition, one limitation of our experimental paradigm is that users may have been coerced into selecting an object in the scene as a candidate referent for "it" (question Q1, i.e., as opposed to selecting nothing at all) even when they believed that no felicitous referent existed. This could be addressed in future work by modifying the question asked to participants in order to allow them to not select any present object if they did not believe them to be sufficiently likely candidates.

CHAPTER 4

GIVENNESS HIERARCHY THEORETIC REFERRING FORM SELECTION

Language-capable interactive robots participating in dialogues with human interlocutors must be able to naturally and efficiently communicate about the entities in their environment. A key aspect of such communication is the use of anaphoric language. The linguistic theory of the *Givenness Hierarchy* (GH) suggests that humans use anaphora due to (subconscious) assumptions about the *cognitive statuses* their referents have in the minds of their interlocutors. In Chapter 3 (see also [15]), we presented a statistical per-entity model of cognitive status. In this chapter, we will now describe how this cognitive status model can be used in conjunction with data structure oriented methods to facilitate efficient and effective algorithms for deciding when and how to use anaphora.

4.1 Motivation

For referring to an entity, a speaker needs to make a decision regarding how to refer to that entity, whether to use a pronoun like *it*, demonstrative adjectives like *this, that*, a definite description like *the*<*NP*> or some other referring form. As mentioned earlier, the GH suggests a strong connection between the presumed cognitive status and the associated referring forms that can be used to refer to objects having a particular status.

According to Arnold [54], though a lot of work has been done on how the GH affects linguistic forms, none of the previous linguistic or psycholinguistic work on cognitive status has been able to identify the underlying mechanisms that involve cognitive status and the selection of referring forms associated with it.

From the computational linguistic perspective, different issues appear relating to cognitive status and the choice of referring forms. As discussed in their survey, Gatt and Krahmer [48] provide several reasons motivating our work on GH-theoretic referring form selection. First, previous work shows that in spite of referring form selection being a research topic

for long, traditionally more importance has been given to other natural language generation tasks. Second, while some of this related work has drawn on concepts of discourse focus/salience [49–51, 53], none of them leveraged the concept of cognitive status. Third, most of the previous work also has been either on just using pronouns (not considering other referring forms given by the GH) or focused on simple visual/textual domains limiting their use on realistic human-robot interaction scenarios, where several non-linguistic factors might play a role in informing the cognitive status of an entity.

Our ultimate goal is to decide how to refer to a target O based on its cognitive status. Our first task is thus to identify the cognitive status of O. We do so using the concept of a *Cognitive Status Engine*. As described in Chapter 3, we leverage the concept of *Cognitive Status Filter* (CSF) for a given object o. We maintain distributions over cognitive statuses for each object presumed to be Familiar or higher through the CSE comprised of CSFs for each such object (if an entity is not tracked by a CSF, it is assumed to be at most Uniquely Identifiable (UID)).

Once we have determined the cognitive status of O, we use it to decide how to refer to O. Each tier of cognitive status is associated with a set of referring forms. One might thus naively assume that to decide how to refer to O, one could simply use a referring form associated with O's presumed cognitive status. While we may do so when O is at most Familiar or Uniquely Identifiable, when O is In Focus or Activated, this is infeasible for the following two reasons.

- First, there may be multiple candidate referents that could plausibly be referred to using that referring form (e.g., when a form like "this" is used to cue an activated object, and more than one activated object exists). Accordingly, when this type of ambiguity is identified (i.e., when the set of status-sensitive distractors for an object are non-empty, it may be avoided by selecting a referring form *that involves a definite description* at either the same level or a lower level of cognitive status, e.g. selecting "this NP" rather than "it".

To construct a status-sensitive distractor set, we must identify all objects with at least the same status as our target. The cognitive status of the target can be quickly retrieved from the CSE. The distractors to be ruled out are then the union of the data structures associated with the target's current most probable cognitive status and higher.

- Second, some referring forms (e.g., "this" and "that") are only appropriate for objects that are physically (or temporally) close. An entity may be "temporally close" depending on when and how often it is mentioned within an utterance or utterances. When this type of conflict is identified, it may be avoided by selecting a referring form that *does not violate such constraints* at either the same level or a lower level of cognitive status, e.g., selecting "that N" rather than "this N".

4.2 Problem formulation

We formulate our problem using a machine learning approach. Since, we are trying to choose a possible referring form (like"it"/"this"/"that"/"thisNP"/"thatNP"/"theNP") for an entity depending on its cognitive status and other extralinguistic factors (such as ambiguity and proximity), we consider this as a classification problem, because we have discrete input features and discrete output labels. Mathematically, classification can be represented as

$$f : X \rightarrow Y \tag{4.1}$$

that is, learning a function or a mapping (f) from input domain (features X) to output (Y label). Since we have a multiclass classification here, it implies that we have non-linear data and so we need a non-linear classification model, like, a decision tree to solve our problem.

4.2.1 A Decision Tree classifier model

A Decision tree represents a function that takes as input a vector of feature values and returns a *decision*, that is a single output value. A decision tree is also often visualized as a tree like structure where the main components are nodes, branches/edges and leaves and

the steps on which the model is built are splitting, stopping and pruning.

- Nodes: The root node (also known as decision node) represents a choice that divides the entire dataset into two or more mutually exclusive subsets of data. The root node represents the entire dataset and does not have any parent node. The internal nodes represent one of the possible choices available at that point in the tree. The leaf nodes represent the final outcome as a result of all the previous decisions.

- Branches: The branches depict outcomes that emerge from the root node and the internal nodes. They can be thought of as "if-else" rules. Thus each path from the root to a leaf node represents a classification decision rule in the tree.

- Splitting criterion: The decision tree makes decisions by splitting its nodes. Thus, splitting is the process of dividing a node into multiple sub-nodes (child nodes) to create relatively "pure" nodes. Depending on the type of the target variable (continuous or discrete), node splitting is done in multiple ways:

  - reduction in variance (for continuous target variable),
  - gini impurity or information gain or chi-square (for categorical target variable)

- Stopping criterion: Stopping rules are applied to a decision tree to avoid making it highly complex. This means, that if the tree continues to grow maximally until each leaf node corresponds to the "purest", then the model suffers from overfitting of data. Thus, the tree will suffer from poor generalizability of data and lack robustness.

  Similarly, if we stop the splitting too early, the model suffers from underfitting, which implies that error on training data is not sufficiently high and the model will perform poorly due to bias.

- Pruning: Underfitting and overfitting of a decision tree model can be prevented by either setting constraints on the size of the tree or by pruning. Pruning is the process

of reducing the size of a decision tree by removing sections of the tree. This tends to reduce the complexity of the final classifier model and thus improves the predictive accuracy by minimizing overfitting.

Advantages of a Decision tree classifier: The Decision tree classifier is one of the most popular models used in the fields of machine learning, data mining, etc. Decision trees are easy to visualize, interpret and understand (intuitive). They can handle both categorical and numerical data and does not require normalization of numerical data. They are able to perform both binary and multi class classification.

4.2.2 Comparison between different non-linear classifier models

A brief comparison between the different non-linear classifiers, that is, kernel Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naive Bayes and Decision tree is as mentioned below.

- SVM: Achieves high performance on non-linear classification problems but it is complex and can be computationally expensive.
- KNN: Classifies an example input by identifying its k nearest neighbors' class. Simple to understand, is fast and efficient but we need to manually choose the number of neighbors (k-value).
- Naive Bayes: It is a probabilistic classifier model based on the Bayes' theorem. It assumes that all features are independent of each other and contribute equally to the outcome, which is not always true in real-life scenarios.

Thus, we chose to solve our GH-theoretic referring form selection problem using a decision tree classifier because of the earlier mentioned advantages over the other classifier models.

4.3 Dataset

The dataset that we used to implement the decision tree modeling approach to our problem is the data collected by Bennett et al. [71] for their human-subject experiment in

which each participant collaborated with a human learner and a robot learner to complete a task together. The initial setup of the task setting is as shown in Figure 4.1.

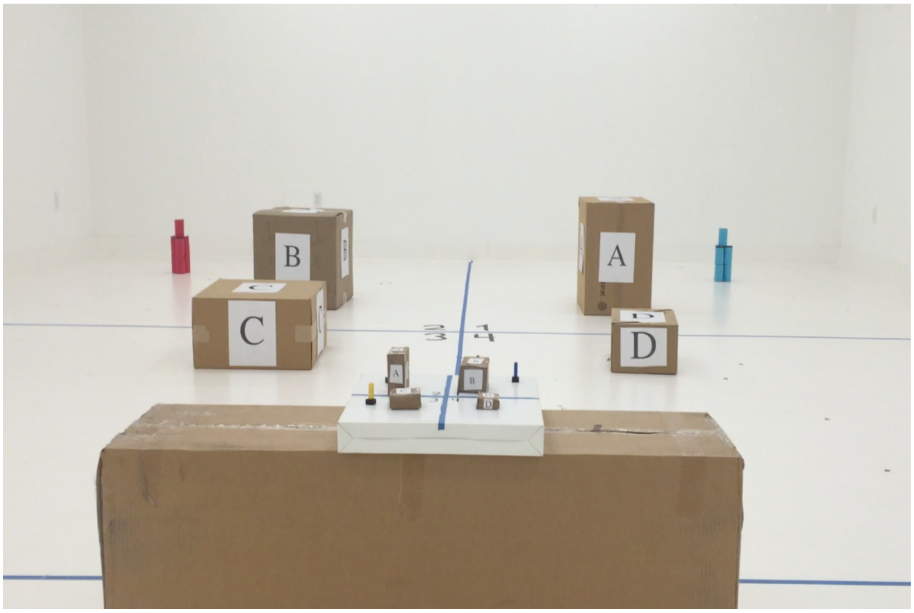

Figure 4.1: Initial whole setup [71]

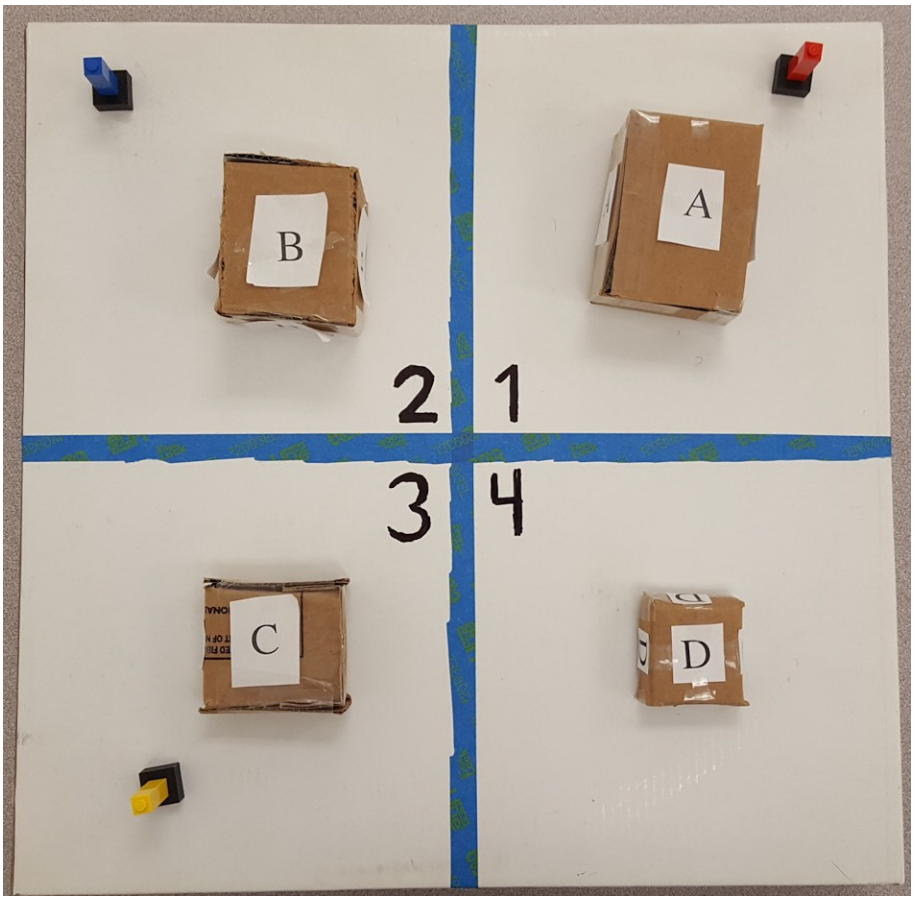

Figure 4.2: Initial setup for participant[71]

The experimental environment consisted of 7 objects - four different boxes labeled A, B, C and D, three colored towers (yellow, red and blue). The experiment room is divided into four quadrants with the boxes and towers are placed as shown in the Figure 4.2 (also same as the initial setting). Each participant would sit in front of the small demo setup and change the initial setting by moving the boxes and knocking down the towers and achieve a goal setting as shown in Figure 4.3. After the human/robot learner entered, the participant gave instructions to the learner to achieve the same goal setting on the actual setup in the room.

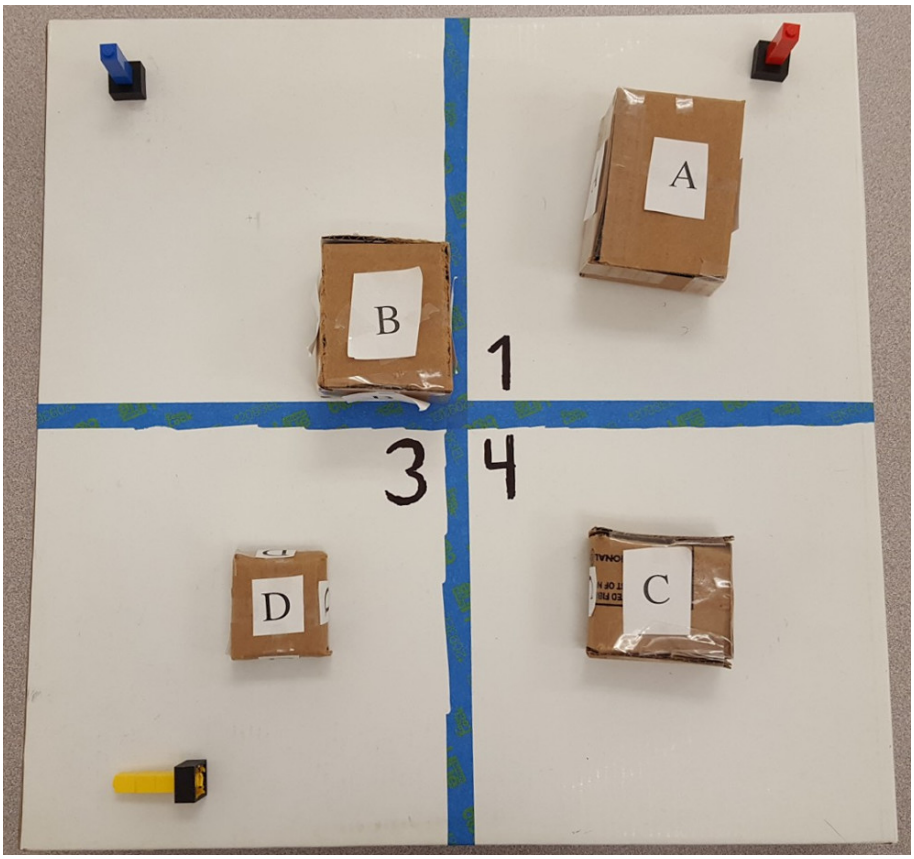

Figure 4.3: A goal setup [71]

The dataset consists of both textual data (instructions given by the participant) and video data where we can watch each interaction between the participant and the human/robot learner. The dataset consists of 66 monologues and 485 utterances in total (after removing the utterances that did not mention any object). Each monologue contains a different number of utterances, ranging from 5 to 24. An example monologue from the dataset is as shown below:

- Monologue:

  U1: Um, first, we're gonna go push block A to the back-center of box 1.

  U2: Um, and knock over the blue tower.

  U3: Then we're gonna take box B and move it to the center of the 3rd quadrant.

U4: Um, box C is going to go right on the number 4.

U5: Um, and then box D is going to go to the corner on the line with box 1 on it, up against the wall.

U6: Um, box C needs to be on the number 4.

U7: Um, Box D needs to be closer to the line.

Each row in the final dataset represents a reference to an object, that is, how someone in the video referred to that object. the final dataset has 603 rows of data. The features used in our dataset are *cognitive status*, *number of distractors*, *temporal distance*, *physical distance* and the target label is the *referring form* used. While "cognitive status", "physical distance" and "referring form" correspond to *categorical* data, the "number of distractors" and "temporal distance" correspond to *numerical* data.

The features are described as below:

1. The Cognitive Status information (in focus/activated/familiar/uniquely identifiable) for each object in a monologue is obtained by using our CSF model [15]. Thus, four possible values for this feature.

2. The number of distractors for each object in the monologue is then obtained by calculating the number of objects having the same status or higher (as mentioned previously).

3. The physical distance of an object from the speaker is obtained through watching the videos. For each object mentioned in an utterance by the speaker, if the object was placed in either quadrants 1 or 2 then we considered it as "far", if the object was placed in quadrants 3 or 4 then we considered it as "close", and if the object was placed just on the borders separating the quadrants then we considered it as neither close nor far, so "none". Thus three possible values for this feature.

4. The temporal distance is calculated from the text data (utterances) that we have where we can see when exactly and how many times was an object mentioned (thus calculated

on the basis of reference). The temporal distance of an object in an utterance is defined as the inverse of its total count of the number of times it was mentioned previously.

For example, if an object is most recently mentioned then its count is 1 and its temporal distance is 1; if the object was the second most recently mentioned then its count is 2 and temporal distance is 0.5 and so forth. This means that if an object is never mentioned in an utterance then its temporal distance is 0.

5. The referring form used is again obtained from the text and video data. The six possible values for the target feature are *it*, *this*, *that*, *thisNP*, *thatNP*, *theNP*.

4.4 Training

We used the publicly available open source software WEKA (Waikato Environment for Knowledge Analysis). It is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand [72].

REPTree (Reduced Error Pruning Tree) algorithm of WEKA software is a popular machine learning algorithm based on Quinlan [73] C4.5 algorithm. It is a fast decision tree learner and builds a decision or regression tree model using information gain or variance as the splitting criterion and prunes it using reduced-error pruning (with backfitting). It only sorts values for numeric attributes once. The missing values are dealt with by splitting the corresponding instances into pieces (i.e. as in C4.5) [73–75].

WEKA performs stratified 10-fold cross-validation by default. This means that the dataset is split into 10 subsets, and one subset in turn is the test data with the remaining 9 subsets as the training data. We experimented with the different values of cross-validation folds (5, 10, 15), to obtain the model with the best accuracy and finally used 15 (best accuracy obtained for the main model).

4.5 Evaluation

Our main dataset, with all the previously mentioned features, is named as D1. D1 is further modified to obtain three more datasets:

- dataset D2 (without temporal distance feature)
- dataset D3 (without physical distance feature)
- dataset D4 (without distance sensitivity)

This is done to evaluate our main model obtained through D1, with the baseline models created by D2, D3 and D4 respectively, so that we can compare the performance of our main model with these baseline models. In our case, it also helps in providing insights regarding how the distance feature affects the performance of the models.

Our chosen evaluation metrics as obtained through WEKA are: *accuracy* of prediction of model, *root mean squared error* (RMSE), *precision*, *recall* and *F1 score*.

In the next section, we discuss the results and visualize the decision tree models as obtained through WEKA.

4.6 Results

After performing a 15 fold-cross validation on each of the datasets, the final results are as shown in Table 4.1 below. As evident from the accuracy values, the model generated for D2 has the highest accuracy (86.07%) followed by the main model with D1 (84.74%). The models generated with D3 and D4 have similar accuracy. The model without distance sensitivity has the worst accuracy. Thus, it is evident that the distance parameter is an important factor while selecting the appropriate referring form for an entity. The model with D2 (with only physical distance feature) has the highest accuracy which might indicate that only physical distance information is more important than temporal distance alone or when both the distances are considered.

The model with D2 also has the least root mean square error, highest precision and F1 score. With regard to all the metrics, our main model (with D1) performs fairly good than the models generated with D3 and D4 respectively.

Table 4.1: Evaluation Metrics

	model(D1)	model(D2)	model(D3)	model(D4)
Accuracy	84.74	86.07	83.58	83.57
RMSE	0.1971	0.1949	0.2077	0.208
Precision	0.858	0.882	0.840	0.841
Recall	0.847	0.861	0.836	0.836
F1 measure	0.843	0.858	0.838	0.839

The decision tree model as generated for the main dataset D1 is as in Figure 4.4 below. As evident from the tree, instances of all the six referring forms are generated here.

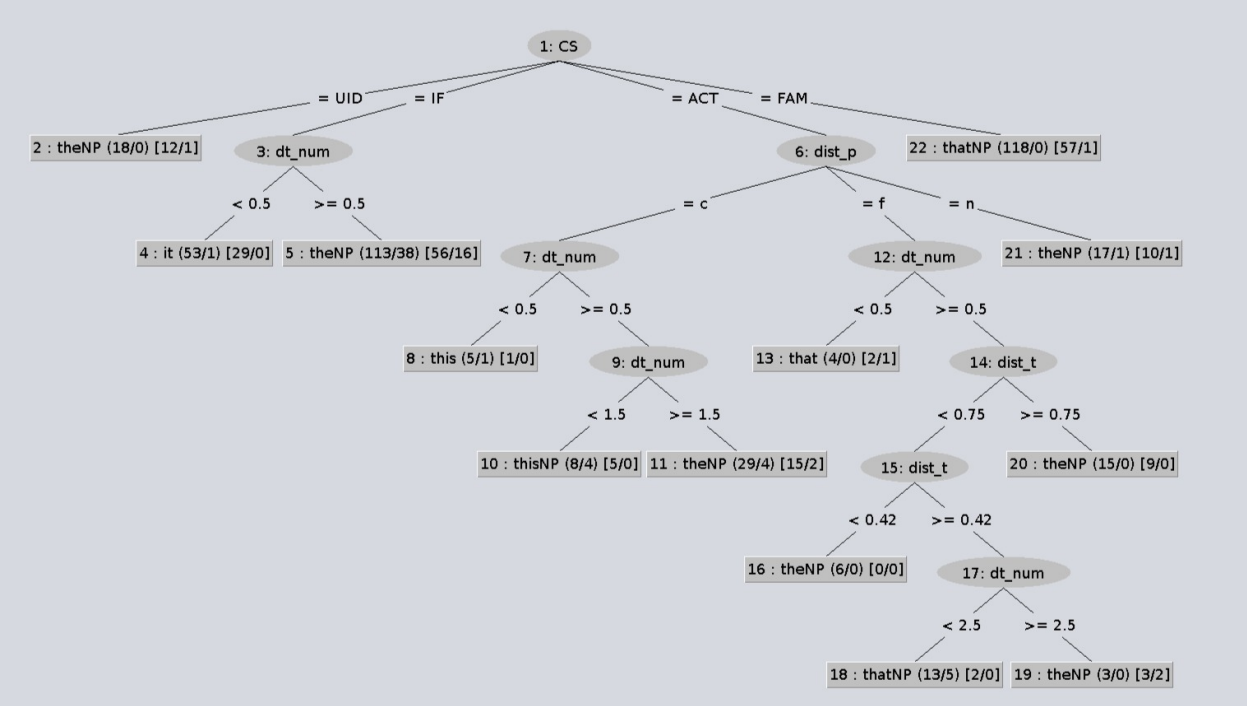

Figure 4.4: Decision Tree model for D1

The interpretation of the decision tree in Figure 4.4 is as explained below:

The root node represents the entire dataset samples, and the main split happens on the *cognitive status* (CS) feature. The root node splits into four possible cognitive statuses. There are no further splits on the familiar (FAM) and uniquely identifiable (UID) statuses. Thus, the leaf nodes with the referring forms "thatNP" and "theNP" represents the target labels for the FAM and UID statuses respectively.

For the in focus (IF) status, the only split that happens is on the *number of distractors* feature (dt_num), and as evident, when there are no distractors the tree decides on the referring form "it" and if there is at least one distractor then the tree decides on the referring form "theNP".

For the activated (ACT) status multiple splits happens. The main split happens on the *physical distance* feature (dist_p). Thus, when the distance is *none*, the referring form is "theNP". When the distance is *close*, another split happens with respect to dt_num. Thus, if there are no distractors then the referring form is "this", else, for at most one distractor the referring form is "thisNP", and for at least 2 distractors the referring form is "theNP". When the distance is *far*, another split happens on the dt_num feature. Thus, if there are no distractors then the referring form is "that", else, another split happens on the *temporal distance* (dist_t) feature. Thus, if most recently mentioned object then referring form is "theNP", otherwise, another split happens on dist_t. Thus, if at most 3rd most recently mentioned object then the referring form is "theNP", if at least 2nd most recently mentioned object then the referring form is "theNP", else, final split happens on dt_num. Thus, when there are at most 2 distractors the referring form is "thatNP" and when there are at least 3 distractors the referring form is "theNP".

The decision tree model as generated for the dataset D2 is as in Figure 4.5 below and the interpretation of this decision tree is given similar to the previous one.

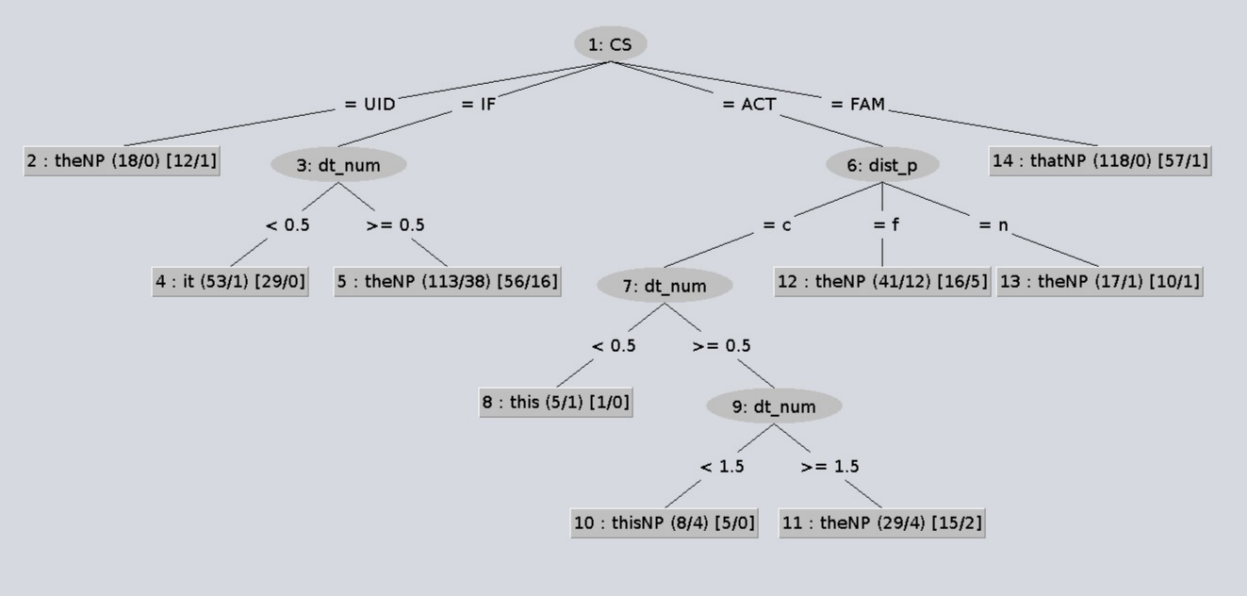

Figure 4.5: Decision Tree model for D2

The root node represents the entire dataset samples, and the main split happens on the *cognitive status* (CS) feature. The root node splits into four possible cognitive statuses. There are no further splits on the familiar (FAM) and uniquely identifiable (UID) statuses. Thus, the leaf nodes with the referring forms "thatNP" and "theNP" represents the target labels for the FAM and UID statuses respectively.

For the in focus (IF) status, the only split that happens is on the *number of distractors* feature (dt_num). Then, when there are no distractors the tree decides on the referring form "it" and if there is at least one distractor then the tree decides on the referring form "theNP".

For the activated (ACT) status multiple splits happens. The main split happens on the *physical distance* feature (dist_p). Thus, when the distance is *none*, the referring form is "theNP", when the distance is *far*, the referring form is "theNP" and when the distance is *close*, another split happens with respect to dt_num. Then, if there are no distractors, the referring form is "this", else, another split happens on dt_num. Then, for at most one distractor the referring form is "thisNP", and for at least 2 distractors the referring form is

"theNP".

The decision tree model as generated for D3 is as in Figure 4.6 below. Though this model is not able to generate instances of all the referring forms, it is worth noting that this model is much simpler than the other two models while having similar accuracy to other models. The model generated with D4 has the same tree structure as in Figure 4.6 for the same 15-fold cross validation used.

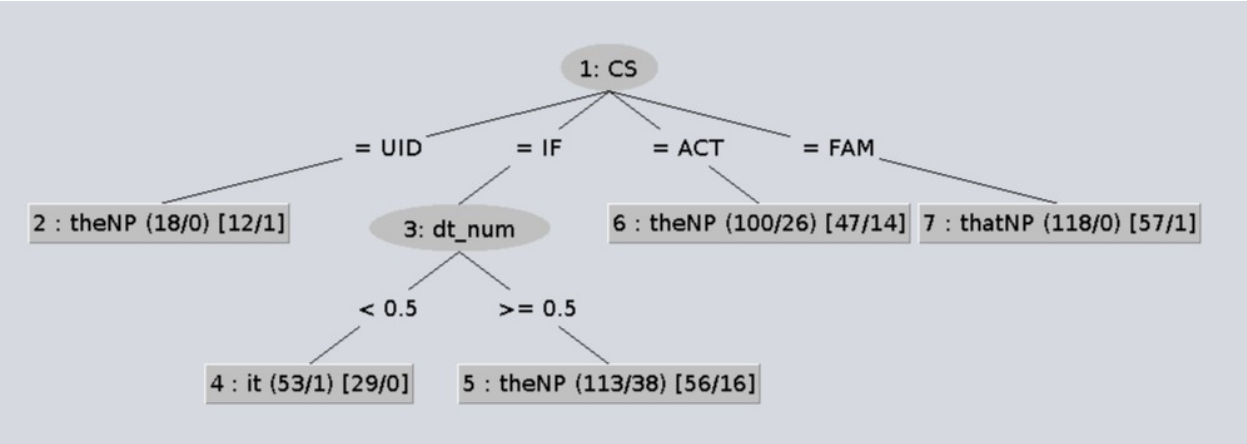

Figure 4.6: Decision Tree model for D3

4.6.1 Conclusion

We presented our second study, a machine learning approach to the problem of GH-theoretic referring form selection using a decision tree classifier model and demonstrated that the model automatically learns the decisions from the data. We trained the model as well as several baseline models on a corpus of human-human and human-robot interactions.

We compared the performance of our trained model with the baseline models and demonstrated that the physical distance feature is more informative than the temporal distance feature.

We demonstrated that using non-linguistic features (such as ambiguity and proximity) along with the cognitive status of an entity might inform the choice of a referring form.

Since the accuracy measures of all the models are close to each other, we cannot prove that one model is better than the other (without doing hypothesis testing). This is because of two reasons:

- First, our CSF model's accuracy is around 82%, which means that the cognitive status information (which is one of the features) might not be correct all of the times, which in turn might affect the accuracy of our decision tree model.

- Second, the referring forms chosen by people really are context specific and depends on their own preferences of referring to an object. For example, while referring to a box, they can say "the box", "this box", "that box", or something else. Every time our models generated a wrong prediction for an object, it meant that the referring form predicted by the model did not match the exact referring form used by a particular participant for that particular object.

4.6.2 Future work

One aspect of this particular dataset [71] is that we did not have many instances for the referring forms "this" and "that", so in future we would like to collect a bigger and more balanced dataset that has relatively similar number of occurrences of all the referring forms.

In future work, we would also like to include all the 6 cognitive statuses and all the referring forms as encoded in the Givenness Hierarchy to generate a full anaphora generation model for referring form selection. We will extend our anaphora generation model for selection of referring content.

Finally, we intend incorporate the GH-theoretic anaphora generation model into the DIARC [70] robotic architecture to demonstrate its use in real-world HRI contexts.

CHAPTER 5

CONCLUSION AND FUTURE WORK

In our first study, we have presented the notion of a *Cognitive Status Filter*: a statistical model for estimating the cognitive status of some entity that may be referenced in conversation. We then described a Mechanical Turk experiment used to gather ground truth data to train this model, and demonstrate how the accuracy of this model compares to a rule-based FSM model and a random baseline model. The overall accuracy of our CSF model in predicting the cognitive status of an object is similar to that achieved by a FSM. This simultaneously speaks in favor of the heuristics encoded in the GH coding protocol, while also demonstrating that those heuristics are learnable from data.

In our second study, we used our CSF model to inform GH-theoretic natural language generation. Specifically, For achieving our goal of leveraging the concept of cognitive status to directly select referring forms for effective NLG and depending on our intuitions as to what extralinguistic factors (such as ambiguity, proximity, etc.) interact with the cognitive status to determine the referring form we should use to describe the target entity, we further presented a technique for GH-theoretic NLG.

We adopted a machine learning technique, that is, we used a decision tree classifier model to automatically generate the decisions for selecting referring forms given the input parameters. We depicted that this decision tree modeling approach has an accuracy of 84% on the dataset used. We also demonstrated that the physical distance parameter of an object from the speaker is an important non-linguistic factor for the choice of referring forms (possibly more important than temporal distance). Though we could not strongly suggest on how good or bad our model is, because of the inherent issues such as moderate accuracy of our CSF model in predicting the status of an object and the referring forms being person-specific.

Through this study we demonstrated that cognitive status is indeed an important factor that needs to be considered for effective language generation. We also demonstrated that cognitive status information augmented with non-linguistic factors affect language generation.

5.1 Future work

We believe that a computational model of cognitive status that takes into account extralinguistic factors (gaze, gesture) and may be other informative sources (like human behavior recognition) and thus accounts for uncertainty (which is expected for real life scenarios) would not only generate a cognitively informed approach to reference generation, but would provide a more robust and successful model of reference understanding as well. Moreover, such a model would provide us with linguistic insights that would allow for refinement of the Givenness Hierarchy itself, thus providing benefits not only for robot designers, but for linguists and cognitive psychologists as well.

We expect that a situated, GH-theoretic approach to referring form selection will be beneficial to both computational linguistics and psycholinguistics communities, giving us a more robust and generalized model that accounts for a wider range of referring forms and cognitive statuses and greater applicability to situated domains.

From the perspective of applications in real world HRI scenarios where social robots are expected to behave and converse like humans, we believe that a GH-theoretic model of language generation can be used to take a step forward in the development of "Theory of Mind" capabilities among language capable robots.

Theory of Mind (ToM) is defined as the attribution of mental states such as beliefs, goals, intentions and desires to other agents or humans [76–79]. Accordingly, researchers in robotics field working in ToM, believe that in order to build human-like robots with social skills as humans, robots need to be able to learn about the properties of both animate and inanimate objects present in the world and have thus examined what other agents believe about particular entities. Research by Scassellati [80] deals with inferring intentions of other

agents, and research by Hiatt et al. [81] deals with deducing the beliefs of other agents.

In contrast, the Givenness Hierarchy accounts for a different aspect of ToM, concerned not with what propositional content others associate with an entity, but whether an entity is known and what cognitive status it holds in their mind. As demonstrated by Gundel et al. [82], children (by age of 3 or earlier) are able to use the full range of referring forms given by the GH, specifically, being implicitly able to judge "non-propositional" cognitive statuses and "focus of attention" with respect to the target entity and being able to explicitly judge "knowledge and belief" [83]. When compared with adults it has been shown that children acquire this implicit form of ToM before learning to follow the Gricean Maxims [29] (that affects the GH).

Thus, our approach to developing a GH-theoretic model of language generation could provide the foundation of a robust Theory of Mind for robots and can improve on the design of language capable collaborative robots. This in turn will also help advance future work on different application areas such as social robotics and natural language processing (NLP), while providing valuable insights to the areas of linguistics and cognitive psychology.

REFERENCES CITED

[1] Masahiro Fujita. Aibo: Toward the era of digital creatures. *The International Journal of Robotics Research*, 20(10):781–794, 2001.

[2] Peter H Kahn Jr, Takayuki Kanda, Hiroshi Ishiguro, Solace Shen, Heather E Gary, and Jolina H Ruckert. Creative collaboration with a social robot. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 99–103, 2014.

[3] Adriana Tapus, Maja J Mataric, and Brian Scassellati. Socially assistive robotics [grand challenges of robotics]. *IEEE Robotics & Automation Magazine*, 14(1):35–42, 2007.

[4] Fumihide Tanaka and Takeshi Kimura. Care-receiving robot as a tool of teachers in child education. *Interaction Studies*, 11(2):263, 2010.

[5] Rachel Gockley, Allison Bruce, Jodi Forlizzi, Marek Michalowski, Anne Mundell, Stephanie Rosenthal, Brennan Sellner, Reid Simmons, Kevin Snipes, Alan C Schultz, et al. Designing robots for long-term social interaction. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1338–1343. IEEE, 2005.

[6] Wolfram Burgard, Armin B Cremers, Dieter Fox, Dirk Hähnel, Gerhard Lakemeyer, Dirk Schulz, Walter Steiner, and Sebastian Thrun. The interactive museum tour-guide robot. In *Aaai/iaai*, pages 11–18, 1998.

[7] Jeanette K Gundel, Nancy Hedberg, and Ron Zacharski. Cognitive status and the form of referring expressions in discourse. *Language*, pages 274–307, 1993.

[8] Tom Williams and Matthias Scheutz. Reference in robotics: A givenness hierarchy theoretic approach. *The Oxford Handbook of Reference*, 2019.

[9] Tom Williams, Evan Krause, Bradley Oosterveld, and Matthias Scheutz. Towards givenness and relevance-theoretic open world reference resolution. In *RSS Workshop on Models and Representations for Natural Human-Robot Communication*, 2018.

[10] Tom Williams, Saurav Acharya, Stephanie Schreitter, and Matthias Scheutz. Situated open world reference resolution for human-robot dialogue. In *Proceedings of the International Conference on Human-Robot Interaction (HRI)*, pages 311–318. IEEE, 2016.

[11] Srinivas Bangalore and Owen Christopher Rambow. Probabilistic model for natural language generation, 2005. US Patent 6,947,885.

[12] Srinivas Bangalore and Michael Johnston. Balancing data-driven and rule-based approaches in the context of a multimodal conversational system. In *Proceedings of the 2003 IEEE Workshop on Automatic Speech Recognition and Understanding.* IEEE, 2003.

[13] Stephanie Schreitter and Brigitte Krenn. The OFAI multi-modal task description corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 1408–1414, 2016.

[14] Elise C Rosa and Jennifer E Arnold. The role of attention in choice of referring expressions. *Proceedings of PRE-Cogsci: Bridging the gap b/w computational, emp. and theoretical app. to reference*, 2011.

[15] Poulomi Pal, Lixiao Zhu, Andrea Golden-Lasher, Akshay Swaminathan, and Tom Williams. Givenness hierarchy theoretic cognitive status filtering. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society, https://cogsci.mindmodeling.org/2020/papers/0164/0164.pdf*, pages 925–931, 2020.

[16] Barbara J Grosz, Aravind K Joshi, and Scott Weinstein. Centering: A framework for modelling the local coherence of discourse. 1995.

[17] Susan E Brennan, Marilyn W Friedman, and Carl Pollard. A centering approach to pronouns. In *25th Annual Meeting of the Association for Computational Linguistics*, pages 155–162, 1987.

[18] Mira Ariel. The function of accessibility in a theory of grammar. *Journal of pragmatics*, 16(5):443–463, 1991.

[19] Mira Ariel. Accessibility theory: An overview. *Text representation: Linguistic and psycholinguistic aspects*, 8:29–87, 2001.

[20] Ellen Gurman Bard, Robin L Hill, Mary Ellen Foster, and Manabu Arai. Tuning accessibility of referring expressions in situated dialogue. *Language, Cognition and Neuroscience*, 29(8):928–949, 2014.

[21] Willem JM Levelt. *Speaking: From intention to articulation*, volume 1. MIT press, 1993.

[22] Jeanette K Gundel, Nancy Hedberg, Ron Zacharski, Ann Mulkern, Tonya Custis, Bonnie Swierzbin, Amel Khalfoui, Linda Humnick, Bryan Gordon, and Mamadou Bassene. Coding protocol for statuses on the givenness hierarchy. *Unpublished manuscript (1993/2006). http://www. sfu. ca/hedberg/Coding_ for_Cognitive_Status. pdf*, 2006.

[23] Nancy Hedberg. Applying the givenness hierarchy framework: Methodological issues. *International workshop on information structure of Austronesian languages*, 2013.

[24] Nelson Cowan. *Attention and memory: An integrated framework*, volume 26. Oxford University Press, 1998.

[25] Jeanette K Gundel, Mamadou Bassene, Bryan Gordon, Linda Humnick, and Amel Khalfaoui. Testing predictions of the givenness hierarchy framework: A crosslinguistic investigation. *Journal of Pragmatics*, 42(7):1770–1785, 2010.

[26] Andrew Kehler. Cognitive status and form of reference in multimodal human-computer interaction. In *Proceedings of AAAI/IAAI*, 2000.

[27] Joyce Y Chai, Pengyu Hong, and Michelle X Zhou. A probabilistic approach to reference resolution in multimodal user interfaces. In *Intelligent User Interfaces (IUI)*, 2004.

[28] Joyce Yue Chai, Zahar Prasov, and Shaolin Qu. Cognitive principles in robust multimodal interpretation. *Journal of Artificial Intelligence Research*, 27:55–83, 2006.

[29] Herbert P Grice. Logic and conversation. In *Syntax and Semantics 3: Speech acts*, pages 41–58. 1975.

[30] Robert Dale and Ehud Reiter. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263, 1995.

[31] Ehud Reiter and Robert Dale. *Building natural language generation systems*. Cambridge university press, 2000.

[32] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013.

[33] Muhammad Usman Ghani Khan and Yoshihiko Gotoh. Describing video contents in natural language. In *Proceedings of the workshop on innovative hybrid approaches to the processing of textual data*, pages 27–35, 2012.

[34] Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 50(2):171–184, 2002.

[35] Ondřej Dušek and Filip Jurčíček. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. *arXiv preprint arXiv:1606.05491*, 2016.

[36] Xiao Li, Kees van Deemter, and Chenghua Lin. Statistics-based lexical choice for nlg from quantitative information. In *Proceedings of the 9th International Natural Language Generation conference*, pages 104–108, 2016.

[37] Verena Rieser, Oliver Lemon, and Simon Keizer. Natural language generation as incremental planning under uncertainty: Adaptive information presentation for statistical dialogue systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(5):979–994, 2014.

[38] Niveda Krishnamoorthy, Girish Malkarnenkar, Raymond Mooney, Kate Saenko, and Sergio Guadarrama. Generating natural-language video descriptions using text-mined knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, 2013.

[39] Kristina Striegnitz, Alexandre Denis, Andrew Gargett, Konstantina Garoufi, Alexander Koller, and Mariët Theune. Report on the second second challenge on generating instructions in virtual environments (give-2.5). 2011.

[40] Ioannis Papaioannou, Christian Dondrup, and Oliver Lemon. Human-robot interaction requires more than slot filling-multi-threaded dialogue for collaborative tasks and social conversation. In *FAIM/ISCA Workshop on Artificial Intelligence for Multimodal Human Robot Interaction*, pages 61–64, 2018.

[41] Mary Ellen Foster, Rachid Alami, Olli Gestranius, Oliver Lemon, Marketta Niemelä, Jean-Marc Odobez, and Amit Kumar Pandey. The mummer project: Engaging human-robot interaction in real-world public spaces. In *International Conference on Social Robotics*, pages 753–763. Springer, 2016.

[42] Ioannis Papaioannou and Oliver Lemon. Combining chat and task-based multimodal dialogue for more engaging hri: A scalable method using reinforcement learning. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 365–366, 2017.

[43] Jennifer E Arnold and Sandra A Zerkle. Why do people produce pronouns? pragmatic selection vs. rational models. *Language, Cognition and Neuroscience*, 34(9):1152–1175, 2019.

[44] Jennifer E Arnold, Elsi Kaiser, Jason M Kahn, and Lucy K Kim. Information structure: linguistic, cognitive, and processing approaches. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(4):403–413, 2013.

[45] Jennifer E Arnold and Nazbanou Nozari. The effects of utterance timing and stimulation of left prefrontal cortex on the production of referential expressions. *Cognition*, 160:127–144, 2017.

[46] Kumiko Fukumura. Interface of linguistic and visual information during audience design. *Cognitive science*, 39(6):1419–1433, 2015.

[47] T Jaeger and Roger Levy. Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, 19:849–856, 2006.

[48] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018.

[49] Massimo Poesio, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. Centering: A parametric theory and its instantiations. *Computational linguistics*, 30(3): 309–363, 2004.

[50] Kathleen F McCoy and Michael Strube. Generating anaphoric expressions: pronoun or definite description? In *The Relation of Discourse/Dialogue Structure and Reference*, 1999.

[51] Rodger Kibble and Richard Power. Optimizing referential coherence in text generation. *Comp. Ling.*, 2004.

[52] Niyu Ge, John Hale, and Eugene Charniak. A statistical approach to anaphora resolution. In *Sixth Workshop on Very Large Corpora*, 1998.

[53] Charles B Callaway and James C Lester. Pronominalization in generated discourse and dialogue. In *Proc. ACL*, pages 88–95, 2002.

[54] Jennifer E Arnold. Explicit and emergent mechanisms of information status. *Topics in Cog. Sci.*, 2016.

[55] Shalom Lappin and Herbert J Leass. An algorithm for pronominal anaphora resolution. *Computational linguistics*, 20(4):535–561, 1994.

[56] Jerry R Hobbs. Resolving pronoun references. *Lingua*, 44(4):311–338, 1978.

[57] Peter C Gordon, Barbara J Grosz, and Laura A Gilliom. Pronouns, names, and the centering of attention in discourse. *Cognitive science*, 17(3):311–347, 1993.

[58] Andrew Kehler and Hannah Rohde. A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, 39(1-2): 1–37, 2013.

[59] Tyne Liang and Dian-Song Wu. Automatic pronominal anaphora resolution in english texts. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 9, Number 1, February 2004: Special Issue on Selected Papers from ROCLING XV*, pages 21–40, 2004.

[60] Jina Song and Elsi Kaiser. Forward-looking effects in subject pronoun interpretation: What comes next matters.

[61] Constantine Nakos, Irina Rabkina, Samuel Hill, and Kenneth D Forbus. Corrective processes in modeling reference resolution.

[62] Sarah Brown-Schmidt, Donna K Byron, and Michael K Tanenhaus. Beyond salience: Interpretation of personal and demonstrative pronouns. *Journal of Memory and Language*, 53(2):292–313, 2005.

[63] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

[64] Betul Bostanci and Erkan Bostanci. An evaluation of classification algorithms using McNemar's test. In *Bio-Inspired Computing: Theories and App. (BIC-TA)*. Springer, 2013.

[65] Douglas Liddell. Practical tests of 2 $\times$ 2 contingency tables. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 25(4):295–304, 1976.

[66] Adrian F Clark and Christine Clark. Performance characterization in computer vision: a tutorial. 1999.

[67] Michael P Fay. Exact McNemar's test and matching confidence intervals, 2011.

[68] Chris Moore, Philip J Dunham, and Phil Dunham. *Joint attention: Its origins and role in development*. Psychology Press, 2014.

[69] DGT Peeters, Z Azar, and A Özyürek. The interplay between joint attention, physical proximity, and pointing gesture in demonstrative choice. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, pages 1144–1149. Austin, Tx: Cognitive Science Society, 2014.

[70] Matthias Scheutz, Thomas Williams, Evan Krause, Bradley Oosterveld, Vasanth Sarathy, and Tyler Frasca. An overview of the distributed integrated cognition affect and reflection DIARC architecture. In *Cognitive Architectures*, pages 165–193. Springer, 2019.

[71] Maxwell Bennett, Tom Williams, Daria Thames, and Matthias Scheutz. Differences in interaction patterns and perception for teleoperated and autonomous humanoid robots. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6589–6594. IEEE, 2017.

[72] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[73] JR Quinlan. Program for machine learning. *C4. 5*, 1993.

[74] Ian H Witten and Eibe Frank. Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, 31(1):76–77, 2002.

[75] Dr B Srinivasan and P Mekala. Mining social networking data for classification using reptree. *International Journal of Advance Research in Computer Science and Management Studies*, 2(10), 2014.

[76] Wilfrid Sellars et al. Empiricism and the philosophy of mind. *Minnesota studies in the philosophy of science*, 1(19):253–329, 1956.

[77] Bertram F Malle, Louis J Moses, and Dare A Baldwin. *Intentions and intentionality: Foundations of social cognition*. MIT press, 2001.

[78] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

[79] Simon Baron-Cohen, Alan M Leslie, Uta Frith, et al. Does the autistic child have a "theory of mind". *Cognition*, 21(1):37–46, 1985.

[80] Brian Scassellati. Theory of mind for a humanoid robot. *Autonomous Robots*, 12(1): 13–24, 2002.

[81] Laura M Hiatt, Anthony M Harrison, and J Gregory Trafton. Accommodating human variability in human-robot teams through theory of mind. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[82] Jeanette K Gundel, Dimitris Ntelitheos, and Melinda Kowalsky. Children's use of referring expressions: some implications for theory of mind. *ZAS Papers in Linguistics*, 48:1–21, 2007.

[83] Jeanette K Gundel. Children's use of referring expressions: What can it tell us about theory of mind? *Children*, 97(98):99, 2009.

APPENDIX

PERMISSIONS

The permissions from co-authors for Chapter 3 is included in the Supplemental file named "copyright_permission_co-authors.pdf"