2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Bridging the Gap to Natural Language-based Grasp Predictions through Semantic Information Extraction

Niko Kleer[1], Martin Feick[1], Amr Gomaa[1], Michael Feld[1], and Antonio Krüger[1]

Abstract— **Enabling multi-fingered robots to choose an appropriate grasp on an object from natural language instructions poses great difficulties for such systems. The diversity, imprecision, and limited information contained in the language make this task particularly challenging. However, speech serves humans as a natural communication interface that can aid robots in adapting to the environment more easily. Therefore, providing robots with relevant data about the objects they interact with is essential for them to understand how to carry out object manipulation tasks. By leveraging Named Entity Recognition (NER) to automatically extract semantic data, our work introduces a novel approach to text-based grasp predictions. Our methodology involves a multistage learning approach using a semantic information extractor that provides significant features to a grasp prediction model. To assess the effectiveness of our approach, we conducted experiments on an existing corpus and two corpora generated by ChatGPT. Our results demonstrate superior performance compared to similar grasp prediction models while overcoming limitations in the literature. Additionally, we open-source our training data for reproducibility and future research advancement.**

I. Introduction

Natural language in the form of speech or text has served a wide range of applications in Human-Robot Interaction (HRI) research. Its uses range from simple commands to control robot movement or end-effector [1], [2], [3], to the management of semantic knowledge about important concepts [4], [5], [6], [7], [8]. Within these applications, natural language is utilized primarily to aid robots in industrial environments and to engage in conversations with social robots. To establish natural HRI, systems often incorporate state-of-the-art Natural Language Processing (NLP) technology, such as Part-of-Speech (POS) tagging or Named Entity Recognition (NER). They present an efficient solution for parsing sentence structure or automatically extracting semantic data about concepts. In some cases, dialogue modeling is used to create an even more controlled environment for such robots [4], [7]. These applications aim to use natural language as an interface for facilitating interactions between humans and machines. Especially recent advances in the field of Large Language Models (LLMs) and the introduction of OpenAI's ChatGPT [9] may have implications for the future of natural language-based HRI research [10]. Due to their potential and versatility in many application domains, researchers are already exploring the use of ChatGPT in innovative chatbot systems [11], [12].

[1]DFKI, Saarland Informatics Campus, Saarland University, 66123 Saarbrücken, Germany {niko.kleer;martin.feick;amr.gomaa;michael.feld;antonio.krüger}@dfki.de

One of the domains that has resorted to the use of natural language is grasp (type) prediction for multi-fingered robotic grasping. Inspired by findings from the field of human grasp analysis, grasp predictors aim to simplify and improve multi-fingered robotic grasp planning by determining a canonic grasping pre-shape. The information about the grasp is then actively integrated into the robotic grasp planning procedure [13], [14], [15], [16], [17], [18], [19], [20]. Although several approaches specifically focus on the use of visual input as the only modality for grasp type identification [21], [22], grasp types have also been explored in the context of natural language [18], [23], [24], [25]. They involve incorporating information about the grasp into ontological frameworks or predicting appropriate grasping gestures from textual descriptions. Consider a lay human who wants to instruct a multi-fingered robot to grasp a generic bottle. Providing a statement such as "the bottle in front of you is a large container of cylindrical shape" presents an easy way for a human to describe the properties of the bottle during HRI without requiring knowledge about the robot. However, existing grasp predictors cannot deal with such instructions due to inflexible feature extraction that enforces strong assumptions on the format of the input data, or do not allow the extraction of significant information for future adaptation. An HRI system, on the other hand, should allow the use of natural communication and continuously adapt to unknown situations, necessitating deep knowledge about its environment. The approach we propose builds on prior research while addressing the above challenges, aiming to bridge the gap to natural language-based grasp prediction. To overcome existing limitations, this paper makes the following contributions:

- We introduce a novel approach to text-based grasp prediction by automatically extracting significant information about grasp affordances. We demonstrate its superior performance using three corpora and in direct comparison to existing models.
- We obtain two corpora by prompting ChatGPT with various queries and propose training a custom NER model for filtering semantic information that influences the choice of a grasp. By also predicting grasps based on these data, we demonstrate that results can strongly depend on prompt engineering.
- All data for training our models are made publicly available for reproducibility and future research[1].

[1]https://github.com/nikleer/EntityBasedGrasps

II. Related Work

A. Natural Language in HRI

Due to numerous technological advancements, the use of natural language in HRI (i.e., speech or text) has experienced notable developments. This field has suggested many approaches to tackle a multitude of different challenges, as subsequently described.

Robot Motion Control: Some works have investigated how to translate voice commands or a sequence of words into executable actions [1], [2]. These actions typically include directional movements or changing the state of the robotic gripper. Specifically, van Delden et al. [3] have investigated the problem of how to enable a non-expert to re-cast a generic pick-and-place task in a typical industrial environment. The authors include a vision component to capture what the user is pointing toward. Krupke et al. [26] have further explored a mixture of multimodal HRI methods. Their system combines voice commands with pointing gestures and head orientation in a mixed-reality environment. They effectively enable controlling a co-located robot for pick-and-place applications. However, robots often require a deeper understanding of the concepts they interact with.

Information Extraction: As robots are intended to perform increasingly complex tasks in everyday life (e.g., dexterous grasping), they may require a sophisticated understanding of their surroundings. To solve this problem in a direction understanding task, Kollar et al. [5] introduce the concept of a Spacial Description Clause (SDC). An SDC represents a sequence of words that describe a route instruction that includes a subject, an action, and a special relation to an object in the environment. Other approaches use well-established Natural Language Processing (NLP) methods to identify and filter significant features of textual data [4], [6], [7], [8]. Part-of-Speech (POS) tagging is commonly applied to identify the grammatical role of words in a sentence, such as nouns or verbs. Another technique often utilized is Named Entity Recognition (NER). The method allows extracting terms belonging to categories such as person names, organizations, and locations [27]. Paplu et al. [7] utilize these methods for including linguistic cues into the dialogue of their social robot to establish context-aware HRI. Some more recent methods have explored how to build interfaces on top of Large Language Models (LLMs) to exploit their pre-existing knowledge about objects and tasks.

ChatGPT in HRI: Artificial intelligence powered by LLMs has been gaining significant attention since the emergence of OpenAI's ChatGPT [9]. This is due to the potential and versatility displayed by models such as GPT-3 and the recently studied GPT-4 across various applications [28], [29], which have already led to the development of novel chatbot systems such as RoboGPT [12] and ROSGPT [11]. These systems capitalize on the capabilities of ChatGPT for HRI. Ye et al. [12] demonstrate that working with such systems reduces mental workload and fosters trust. However, the authors note that cases of miscommunication or inaccurate communication can be problematic, as ChatGPT applies its own understanding to instructions. Due to its recent introduction, how to effectively integrate the technology into HRI applications, such as instructing a robotic system on how to manipulate objects, remains an open question [10].

B. Grasp Prediction for Multi-fingered Robots

Grasp (type) prediction in multi-fingered robotics aims to predict a suitable grasping pre-shape for manipulating an object. The terminology originates from the field of human grasp analysis where grasp types have been studied since the 1950s [30]. Grasp types refer to the various ways in which humans hold and manipulate objects using their hands. Generic versions of these grasps can be categorized based on the position and configuration of the fingers and thumb during gripping [31], [32]. They are actively used to simplify multi-fingered robotic grasp planning [13], [14], [15], [16], [17], [18], [19], [20] and have been demonstrated to outperform similar approaches excluding this information [33]. While most existing approaches focus on the investigation of vision-based methods, grasp types have also been explored in natural language-related applications [18], [23], [24], [25]. For example, Varadarajan et al. [23] formulate an ontological framework that models functional and grasp affordances for task-based grasping. They utilize grasp types as a simple way to model the grasp affordances of objects. Li and Tian [24] similarly use grasp types for defining object manipulation constraints as a part of their ontological framework. To our knowledge, there currently exist two approaches for predicting grasp types from natural language. Rao et al. [18] use manually generated textual descriptions, in addition to precise quantitative data, that follow a specific format and only contain significant features influencing a grasp. They use a mixture of lemmatization, POS tagging, and regular expression-based sentence chunking. As the authors acknowledge, this restricts their extraction method to the specific scenario they have investigated. Our prior work partially solved this problem by retrieving unstructured textual descriptions from the Internet (e.g., Wikipedia) and predicting grasps based on the whole description [25]. While this approach enables processing arbitrary textual descriptions, extracting significant attributes and establishing a semantic representation of the objects is not possible.

Our proposed methodology enables processing arbitrarily formed textual descriptions while overcoming the challenge of feature extraction by automatically extracting semantic information about objects, therefore combining the strengths of previous text-based grasp predictors. Similar to other approaches in HRI, we leverage NER technology to extract this information. Since existing state-of-the-art NER models do not enable extracting information associated with grasp affordances (e.g., object geometry, material, hardness), we have trained a custom model specifically for this purpose. We obtained a large portion of our corpus by prompting ChatGPT with various queries and used this data to train our model. By applying this model, our results demonstrate superior performance in comparison to existing text-based grasp prediction models.

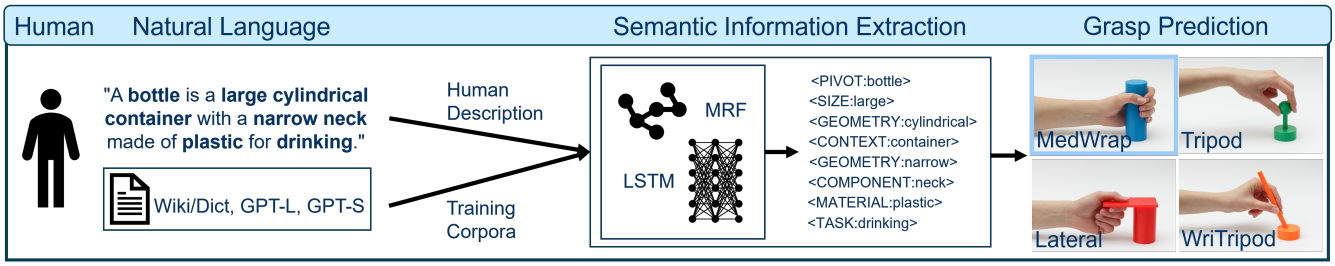

Fig. 1. Conceptually shows how to predict a grasp based on extracted semantic information from a description of a bottle.

III. Method

In this section, we elaborate on our underlying information extraction methodology that is used to model grasp affordances for a NER model. We subsequently describe the important attributes we want to extract. After that, we explain how to create a corpus from natural language that serves as a basis for (1) training the recognition model and (2) showing that natural language-based grasp prediction benefits from the extraction procedure. Figure 1 conceptually demonstrates this approach based on spoken language or text corpora.

A. Semantic Features for Automatic Information Extraction

Based on the literature [18], [31], [32], we have identified a set of significant attributes that may influence a suitable grasp and defined classes for our proposed NER model.

- **Pivot**: We consider the pivot to be the main point of reference in a sentence, and understanding it can help a learning model infer which poses are suitable for objects with similar semantic properties.
- **Geometry**: The geometry of an object has a strong influence on the type of grasp that's appropriate. For example, cylindrical objects like bottles require a completely different grasp than flat objects.
- **Material**: Objects made of fragile material that can more easily break may require a grasp suitable for careful manipulation instead of a strong grip.
- **Hardness**: Some objects are more easily deformable than others, which means they potentially require a grasp that provides more stability to manipulate an object effectively.
- **Texture**: In addition to the material and hardness of an object, its texture can also affect the stability of a grasp, especially if it's slippery.
- **Size**: When describing the size of an object, natural language often includes relative terms like "small" or "large" instead of providing exact numerical values. Although these words do not give an accurate measurement of an object's size, they provide a general idea about it, which affects the choice of a grasp.
- **Weight**: Similarly, the weight of an object, often described using words such as "lightweight" or "heavy," can affect the force required for a successful grasp.
- **Component**: One of the aspects sometimes considered is that some objects can be decomposed into multiple components (e.g., a mug having a body and a handle) [6], [24]. This is not only relevant for establishing semantic representations but also determines whether a grasp is applicable.
- **Context**: Understanding contextual information such as more abstract object categories (e.g., a bottle being a container) can further help to establish relationships and distinguish between different types of objects.
- **Task**: Tasks associated with an object can fundamentally change the grasp required for interacting with it. In particular, the applied grasp can impact how humans interpret the executed action [22], introducing another aspect to consider in natural HRI.
- **Color**: We only want to extract color as an auxiliary attribute to increase the recognizer's versatility in aiding vision-based applications in the future.

Taking the above-described entities into account, the natural language description "a bottle is a large cylindrical container made of plastic with a neck used for drinking", which contains a high density of relevant information, should result in the extraction of the following named entities:

$$\begin{aligned}\{&\text{bottle} \rightarrow \text{PIVOT}, \text{large} \rightarrow \text{SIZE},\\ &\text{cylindrical} \rightarrow \text{GEOMETRY}, \text{container} \rightarrow \text{CONTEXT},\\ &\text{plastic} \rightarrow \text{MATERIAL}, \text{neck} \rightarrow \text{COMPONENT},\\ &\text{drinking} \rightarrow \text{TASK}\}\end{aligned}$$

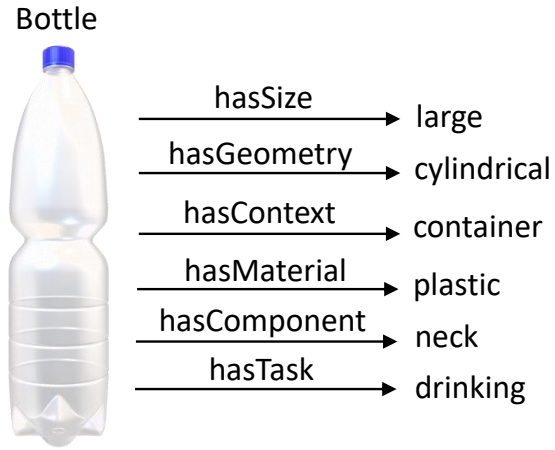

Fig. 2. Translation of entity types into semantic relationships commonly used in ontological frameworks.

TABLE I
WORD AND SENTENCE STATISTICS FOR EACH CORPUS.

Corpus	Number of Words	Number of Sentences
Wiki/Dict [25]	19,227	1,221
GPT-L	18,488	1,018
GPT-S	6,147	347
Total	43,862	2,586

We can translate this information into semantic relationships used in ontological frameworks, as visualized in Figure 2. The set of significant information we identified ultimately contains ten distinct entity types, excluding color. Distinguishing between these many classes naturally necessitates a corpus that allows extracting a rich vocabulary of terms.

B. Data Generation Procedure and Corpora Aggregation

Since state-of-the-art NER models, which are trained on standardized datasets, are not dedicated to extracting entities that influence the choice of a grasp, they are not useful for our investigations. Therefore, one of the main challenges of this research also lies in gathering data to establish a corpus. This corpus should contain a rich vocabulary used to describe the properties of objects. Our prior work dealt with the same challenge [25]. Here, we gathered textual data from online sources such as Wikipedia and online dictionaries for a set of 100 everyday objects. This dataset, which is fully labeled based on four grasp types for the task of holding an object, served as a starting point for our corpus. Additionally, we wanted to leverage crowdsourcing to gather a large corpus of object descriptions provided by humans. However, recent research shows that crowdsourcing study participants employ LLMs for automatizing text-generation tasks [34]. These findings are consistent with our pilot studies where we gathered a tremendous number of duplicate descriptions that were generated by an LLM instead of produced by a crowdsourcing worker. Therefore, similar to more recently developed systems that capitalize on the capabilities of ChatGPT for HRI [11], [12], we decided to leverage its potential to generate suitable data instead. This would not only serve the purpose of training a NER model but also allow us to explore whether data generated by ChatGPT can successfully be employed for predicting grasps. As noted in the literature, generating suitable data may require extensive prompt engineering [12], [28]. During our experiments, we also observed significant differences in how ChatGPT would describe the properties of an object. Including words such as "comprehensive", "thoroughly", "precise", or "short" strongly influenced the number of generated sentences as well as their detail. After thorough experimentation, we decided to generate two datasets that demonstrate vastly different behaviors. In variation one (hereafter referred to as **GPT-L**), we queried ChatGPT to provide as much information as possible by using the generic prompt:

> "What can you tell me about the physical properties of a/an [object]."

As a result of this query, the descriptions we retrieved contain large portions of useful data for training the NER model. This includes various geometric properties (e.g., cylindrical, spherical, or rectangular), a large number of materials (e.g., plastic, wood, or fabric), and details about the components of an object. On the other hand, they tend to be extremely extensive and generic, with an average length of over 184 words. In some cases, the number of words almost reaches 300. Therefore, in variation two (hereafter referred to as **GPT-S)**, we wanted to be more specific and queried ChatGPT to describe only the most significant properties using more precise language. To this end, we used the prompt:

> "Please concisely describe the most common physical attributes of a/an [object] and its typical uses."

In contrast to GPT-L, the information density of the generated descriptions is considerably higher and the language used is less generic. As a consequence, the average length of the retrieved sentences has decreased to 61 words. Table I provides a complete overview of the word and sentence statistics related to each corpus. We have also retrieved these statistics for the existing corpus [25]. Due to our previous extraction procedure, we refer to this data as **Wiki/Dict**. We observe that GPT-L generates nearly as much data as Wiki/Dict, which was gathered using nine different online sources. GPT-S, on the other hand, contains less than one-third in terms of the generated number of words and sentences. Overall, the corpora contain approximately 44,000 words and 2,600 sentences. To establish a training corpus for our custom NER model, we have manually labeled the data according to the semantic features described in the previous section. Sentences not containing the features we are interested in were ignored and not labeled. This led to reducing the number of sentences in our corpora to 677 for Wiki/Dict, 783 for GPT-L, and 257 for GPT-S. By aggregating this data, we obtained our final corpus for training the recognition model.

IV. EVALUATION

Our evaluation is divided into two parts. The first part involves assessing the performance of the proposed NER model. In the second part, we use the NER model to extract semantic information from each corpus introduced in the previous section. We then use the extracted entities to evaluate whether the automatic extraction of these features can benefit natural language-based grasp prediction. This further allows us to compare the performances we retrieve using each corpus.

A. Named Entity Recognizer

To train our proposed custom NER model, we use the state-of-the-art NLP framework FlairNLP [35]. It enables using a bi-directional LSTM in conjunction with a Conditional Random Field (CRF) for training an entity extraction model. Furthermore, it supports the incorporation and combination of many pre-trained word embeddings for a deep contextual understanding of the training data. Since there are no other

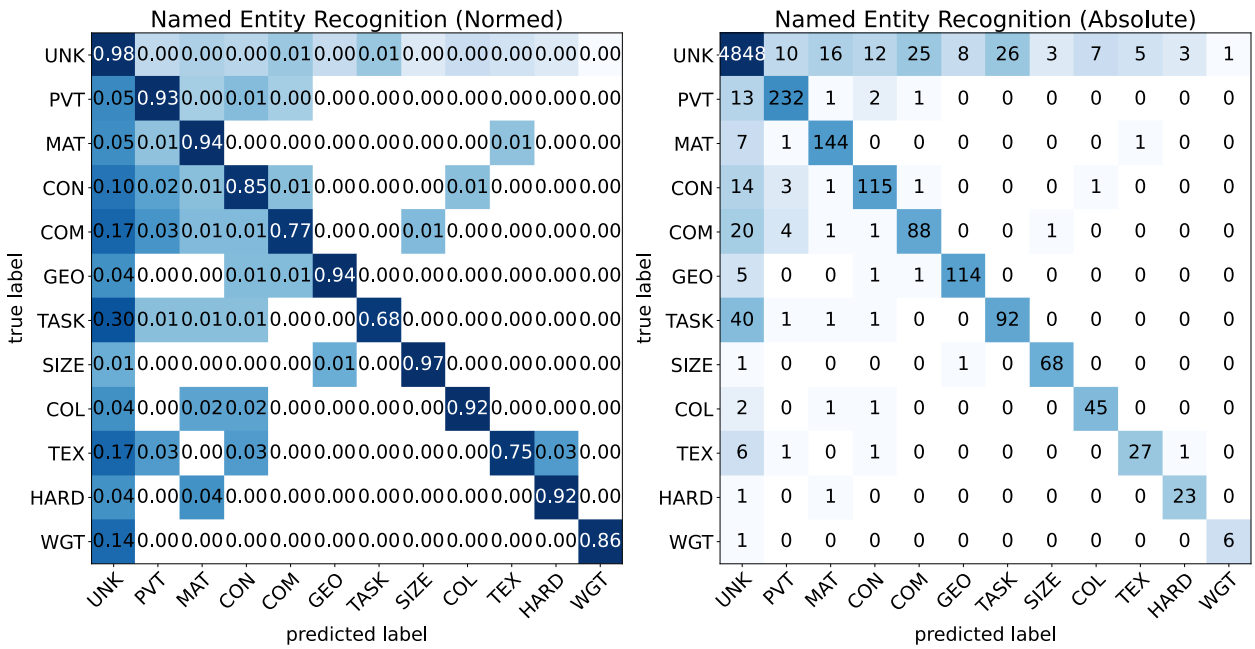

Fig. 3. Confusion matrices showing the normed (left) and absolute (right) value distribution of our NER prediction model averaged over all runs of the k-fold cross-validation. Average values in the interval $[0, 1]$ were rounded to 1. Entity type names are abbreviated for space reasons.

comparable NER models in the literature, we obtained all relevant evaluation metrics including Precision, Recall, F1-Score, Micro Average (Accuracy), and Macro Average. To effectively train and evaluate our custom NER model, we first pre-processed all labeled sentences, which served as the basis for using FlairNLP. Our training configuration includes pre-trained GloVe embeddings [36] and the bidirectional LSTM uses a hidden layer size of 512 cells. The initial learning rate is set to 0.1 and the model is trained using a batch size of 8 over 50 epochs. Table II shows the results based on a k-fold cross-validation using this configuration. The UNK entity type functions as a substitute for all words that did not receive a label. Support describes the total number of classified entities for each entity type. In terms of entity types we aim to extract, our trained model achieves an F1 Score above 0.9 for the PIVOT, MATERIAL, GEOMETRY, SIZE, and WEIGHT entity types. Especially tasks and components, both achieving an F1 Score close to 0.75, are much more challenging to extract. That's because these entity types are arguably the most complex to extract since the entities do not necessarily have a semantic relationship. For example, geometric features such as "spherical", "cylindrical", or "round" are semantically homogeneous whereas the terms "handle", "pages", and "nib" have no relationship to each other. The extraction of tasks deals with a different challenge as recognizing these entities is considerably more complex than tagging the verbs of a sentence. It is important to note that the NER model has to develop a deep understanding of the occurring words and sentence structure to identify what qualifies as a task related to an object. Because of these reasons, Precision and Recall are notably lower in comparison to other entity types. Our model's overall retrieved Micro F1 Score lies above 0.95 whereas the Macro F1 Score averages slightly above 0.88. However, due to a strongly unequal entity distribution in our data, the Micro F1 Score is not a suitable indicator for assessing the model's performance. As the UNK entity type, which represents all unlabeled words, has the highest support by a considerable margin and achieves excellent performance, the score is biased. Calculating this value without including the UNK entity type results in a Micro F1 Score of 0.8797, which is almost equivalent to our retrieved Macro F1 Score and a much better indicator. State-of-the-art models trained on standardized datasets reportedly achieve slightly higher Micro F1 Scores of 0.943 (4 classes) [37] and 0.913 (18 classes) [38] on well-established entity types (e.g., person names, organizations, and locations). We have also retrieved our entity extractor's confusion matrix to learn about the entity assignment (see Figure 3). Notably, the majority of False Positive and False Negative predictions are related to the UNK tag. As noted earlier, especially task-related information presents a challenge to our extractor, which leads to a high percentage of False Positive predictions. Only a small fraction of entities are mistakenly assigned a label of another significant entity type. The PIVOT entity type appears to be most prone to this misclassification. These results indicate that, while our NER model performs reasonably well, correctly tagging entities in unseen data can still be improved. We have used this entity extractor to predict grasp types, as described in the next section.

TABLE II

RESULTS OF A K-FOLD CROSS-VALIDATION ON AN ENTITY-LEVEL. THE AVERAGE ACCURACY (MICRO F1) AND MACRO AVERAGE INCLUDE ALL ENTITIES (MODEL-LEVEL EVALUATION).

Entity Type	Precision	Recall	F1 Score	Support
UNK (no tag)	0.9780	0.9758	0.9769	4966
PIVOT	0.9360	0.9444	0.9383	245
MATERIAL	0.8773	0.9475	0.9110	152
CONTEXT	0.8691	0.8553	0.8608	135
TASK	0.7778	0.6962	0.7394	133
GEOMETRY	0.9286	0.9409	0.9341	121
COMPONENT	0.7503	0.7775	0.7627	113
SIZE	0.9527	0.9804	0.9662	69
COLOR	0.8455	0.9299	0.8850	48
TEXTURE	0.8420	0.7625	0.7965	36
HARDNESS	0.8522	0.9252	0.8855	25
WEIGHT	0.9321	0.9042	0.9157	7
Micro Average			0.9585	6058
Macro Average	0.8782	0.8865	0.8803	6058

B. Grasp Prediction Model

To assess whether the extraction of semantic features benefits grasp prediction, we first extracted all entities for the objects in our corpora. To compare our results to the most recent natural language-based grasp method, we followed a similar evaluation methodology as in our prior work [25]. There, we found a Support Vector Machine (SVM) that uses a linear kernel and a Convolutional Neural Network (CNN) to be most effective. The SVM learns features from a tf-idf (term frequency-inverse document frequency) matrix and the CNN uses a pre-trained word embedding for its first layer. We demonstrate the performance of these models using three different input data configurations.

1) **Raw Data**: For the first configuration of our models, we use the raw textual data. This produces results equivalent to our prior work where we used complete textual descriptions [25]. As this does not involve using the NER model, it serves as a baseline.
2) **Entity and Entity Type**: This configuration leverages the extracted semantic entities in addition to their corresponding entity type (e.g., geometry cylindrical). Adding the information about the entity type presumably aids the CNN in establishing a better understanding of the extracted data and their relationships.
3) **Extracted Entities**: The final configuration we evaluate uses the smallest set of features by only learning grasps based on the extracted entities.

As before, we split the data according to a k-fold cross-validation. The accuracy and standard deviations of each configuration are summarized in Table III. We can derive several significant observations. First, our results show that the extraction of semantic information has improved grasp prediction for each corpus, regardless of the learning method used. With the Wiki/Dict pre-existing corpus, we achieved improvements ranging from 0.066 to 0.076 when compared to [25]. It further demonstrates an improvement of more than 0.05 compared to the work by Rao et al. [18], who used manually generated descriptions that follow a specific format. The extraction of entities led to an even more significant improvement for the GPT-S corpus, where the SVM gained 0.069 and CNN gained 0.088 in accuracy. For GPT-L, only the SVM's performance increased by nearly 0.05 in accuracy, while the CNN achieved a minor improvement. Furthermore, our hypothesis that the inclusion of the entity type may aid CNNs in establishing a better understanding of the relationships in the data is partially true using this data. While a direct comparison to predictions on raw data always leads to an improvement, the approach can be inferior (Wiki/Dict) or similarly effective (GPT-L) as predicting grasps using only the extracted entities. Moreover, a comparison between our retrieved results for the ChatGPT-based corpora shows significant differences in prediction quality. Using raw data, these differences can only be observed using the SVM classifier. However, by employing our NER model, both, SVM and CNN, strongly outperform predictions made on GPT-L by utilizing data extracted from the GPT-S corpus. Presumably, this could be caused by discrepancies in the extracted entities for these corpora. Overall, our evaluation demonstrates that the extraction of semantic features benefits grasp prediction across all models and corpora used in our experiments. Furthermore, a comparison of our retrieved corpora using ChatGPT shows that results can strongly differ based on the provided prompt.

V. DISCUSSION

When investigating the data qualitatively, an interesting observation that is not reflected in the quantitative results is that there is a small set of objects that is almost consistently misclassified across all datasets and model configurations. These objects include the "barrette" (i.e., a small hair clip), "bowl", "candle", "flashlight", "hammer", "lollipop", "pliers", "Rubik's cube", and "sponge". Although the data used in our evaluation originates from different sources, none of the extracted descriptions seem suitable for predicting an appropriate grasp using the applied models. For the task of securely holding an object, five of these objects are labeled using the same grasp (medium wrap). The objects "bowl", "lollipop", and "sponge" each require a different grasp. This issue potentially arises from non-optimal descriptions or the properties of an object being naturally more complex to describe. For example, a "lollipop" is sometimes described as round and hard. However, it is not adequately reflected in the data that these attributes refer only to the part of the object that is not usually grasped. Furthermore, even though our approach generally enables the extraction of semantic information, the NER model is currently unable to link components to their respective attributes. Consider the following short description of a hammer:

"A hammer is a heavy tool with a long handle".

When using the extracted information to predict appropriate grasps, the SVM cannot link the attribute "long" to the handle of the hammer by itself. Although the CNN may be able to learn about this relationship through the positioning

TABLE III

GRASP PREDICTION RESULTS FOR EACH CORPUS AND DIFFERING INPUT DATA.

Input	Model	Wiki/Dict	GPT-L	GPT-S
Raw Data (complete description)	SVM	0.79 [25]	0.690 ± 0.0253	0.785 ± 0.0108
Raw Data (complete description)	CNN	0.75 [25]	0.716 ± 0.0375	0.725 ± 0.0334
Extracted Entities and Types	SVM	0.806 ± 0.0126	0.690 ± 0.0141	0.806 ± 0.0117
Extracted Entities and Types	CNN	0.815 ± 0.0283	0.720 ± 0.0278	$\underline{0.813 \pm 0.0105}$
Extracted Entities	SVM	$\underline{0.856 \pm 0.0217}$	$\underline{0.737 \pm 0.0231}$	$\underline{0.854 \pm 0.0084}$
Extracted Entities	CNN	$\underline{0.826 \pm 0.0217}$	$\underline{0.720 \pm 0.0216}$	0.787 ± 0.0249

of such words, it would be desirable to model these relationships for managing knowledge even more effectively in the future. This representation of the extracted entities could be beneficial to our grasp prediction methods, especially when augmented with vision-based HRI applications for a robot that requires complete situational knowledge of its environment.

Since we only predict grasps in the context of securely holding an object, we similarly do not currently use the information about the extracted tasks. Instead, the information may serve our predictors in comparing objects used for the same tasks (e.g., all pens used for writing). Due to the flexibility of this method, it would be possible to simply provide instructions on the task during HRI. In contrast to the literature, which mainly exploits grasp types to plan multi-fingered robotic grasps, our work views them more as a potential interface for natural HRI. Since we have trained our model on textual data with a high information density, the recognizer can already extract attributes from natural language instructions such as:

"Please hold this large cylindrical object for me" $\rightarrow$

$\{\text{hold} \rightarrow \text{TASK}, \text{large} \rightarrow \text{SIZE}, \text{cylindrical} \rightarrow \text{GEOMETRY}\}$

or

"I need your help picking the rectangular parts" $\rightarrow$

$\{\text{picking} \rightarrow \text{TASK}, \text{rectangular} \rightarrow \text{SIZE}\}$

while only facing issues determining more generic pivots. Presumably, fine-tuning the recognition model based on the objects commonly manipulated in a specific environment or domain (e.g., ambient assisted living or industry 4.0) would enable the extraction of coherent information during HRI. Such instructions, or a single significant feature, could also be queried by a robot as part of a dialogue management system. Our work paves the way to solve these questions in future work to achieve more natural HRI using grasp types.

VI. CONCLUSION

This paper has presented a novel approach to natural language-based grasp prediction. We have proposed leveraging NER technology for the automatic extraction of semantic information influencing the choice of a grasp. To this end, we have identified relevant features associated with grasp affordances describing the properties of an object. To overcome the challenge of generating training data, we have retrieved two corpora by prompting ChatGPT with queries that demonstrate vastly different behaviors. An evaluation of our proposed NER model achieves a Macro Average F1 Score of 0.88 and a similar Micro F1 Score while only distinguishing between relevant entity types. We demonstrate that the automatic extraction of semantic features benefits grasp prediction using three corpora. We further show that results may depend on prompt engineering, necessitating the development of strategies for generating appropriate data using LLMs. Overall, our method demonstrates superior performance in comparison to similar grasp prediction models while overcoming currently existing limitations. By systematically post-processing the descriptions, we can more easily facilitate establishing semantic representations and potentially deal with instructions provided by a human.

Using our findings, we would like to explore the use of multimodal interaction techniques in the context of multi-fingered robotics for achieving natural HRI. Instead of treating grasp types purely as a learned prior for grasp planning, we find the idea of leveraging them as an interface to HRI, which has not been explored in the literature, very promising. Our research is intended to be a step in this direction using natural language and we look forward to seeing more research that considers this aspect in the context of HRI.

ACKNOWLEDGMENT

This work is supported by the German Federal Ministry of Education and Research (grant no. 01IW20008) as a part of CAMELOT - Continuous Adaptive Machine-Learning of Transfer of Control Situations.

REFERENCES

[1] J. Norberto Pires, "Robot-by-voice: experiments on commanding an industrial robot using the human voice," *Industrial Robot: An International Journal*, vol. 32, no. 6, pp. 505–511, Dec. 2005.

[2] W. S. Kit and C. Venkatratnam, "Pick and place mobile robot for the disabled through voice commands," in *2016 2nd IEEE International Symposium on Robotics and Manufacturing Automation (ROMA)*. Ipoh, Malaysia: IEEE, 2016, pp. 1–4.

[3] S. van Delden, M. Umrysh, C. Rosario, and G. Hess, "Pick-and-place application development using voice and visual commands," *Industrial Robot: An International Journal*, vol. 39, no. 6, pp. 592–600, 2012.

[4] L. Grassi, C. T. Recchiuto, and A. Sgorbissa, "Knowledge triggering, extraction and storage via human–robot verbal interaction," *Robotics and Autonomous Systems*, vol. 148, p. 103938, 2022.

[5] T. Kollar, S. Tellex, D. Roy, and N. Roy, "Toward understanding natural language directions," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2010, pp. 259–266.

[6] D. Nyga, M. Picklum, and M. Beetz, "What no robot has seen before—probabilistic interpretation of natural-language object descriptions," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 4278–4285.

[7] S. Paplu, H. Ahmed, A. Ashok, S. Akkus, and K. Berns, "Multimodal perceptual cues for context-aware human-robot interaction," in *IFToMM International Symposium on Science of Mechanisms and Machines (SYROM)*. Springer, 2022, pp. 283–294.

[8] O. Roesler, A. Aly, T. Taniguchi, and Y. Hayashi, "Evaluation of word representations in grounding natural language instructions through computational human-robot interaction," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 307–316.

[9] OpenAI, "Chatgpt," 2023. [Online]. Available: https://openai.com/blog/chatgpt/

[10] M. Aljanabi, "Chatgpt: Future directions and open possibilities," *Mesopotamian journal of Cybersecurity*, vol. 2023, pp. 16–17, 2023.

[11] A. Koubaa, "Rosgpt: Next-generation human-robot interaction with chatgpt and ros," 2023.

[12] Y. Ye, H. You, and J. Du, "Improved trust in human-robot collaboration with chatgpt," *IEEE Access*, 2023.

[13] Z. Deng, B. Fang, B. He, and J. Zhang, "An adaptive planning framework for dexterous robotic grasping with grasp type detection," *Robotics and Autonomous Systems*, vol. 140, p. 103727, 2021.

[14] Z. Deng, G. Gao, S. Frintrop, F. Sun, C. Zhang, and J. Zhang, "Attention based visual analysis for fast grasp planning with a multi-fingered robotic hand," *Frontiers in neurorobotics*, vol. 13, p. 60, 2019.

[15] D. Dimou, J. Santos-Victor, and P. Moreno, "Grasp pose sampling for precision grasp types with multi-fingered robotic hands," in *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*. IEEE, 2022, pp. 773–779.

[16] H. Liang, L. Cong, N. Hendrich, S. Li, F. Sun, and J. Zhang, "Multifingered grasping based on multimodal reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1174–1181, 2021.

[17] J. Lundell, E. Corona, T. N. Le, F. Verdoja, P. Weinzaepfel, G. Rogez, F. Moreno-Noguer, and V. Kyrki, "Multi-fingan: Generative coarse-to-fine sampling of multi-finger grasps," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 4495–4501.

[18] A. B. Rao, K. Krishnan, and H. He, "Learning robotic grasping strategy based on natural-language object descriptions," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 882–887.

[19] C. Wang, D. Freer, J. Liu, and G.-Z. Yang, "Vision-based automatic control of a 5-fingered assistive robotic manipulator for activities of daily living," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 627–633.

[20] Y. Zhang, J. Hang, T. Zhu, X. Lin, R. Wu, W. Peng, D. Tian, and Y. Sun, "Functionalgrasp: Learning functional grasp for robots via semantic hand-object representation," *IEEE Robotics and Automation Letters*, 2023.

[21] A. Das, A. Chattopadhyay, F. Alia, and J. Kumari, "Grasp-pose prediction for hand-held objects," in *Emerging Technology in Modelling and Graphics*. Springer, 2020, pp. 191–202.

[22] Y. Yang, C. Fermuller, Y. Li, and Y. Aloimonos, "Grasp type revisited: A modern perspective on a classical feature for vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 400–408.

[23] K. Varadarajan and M. Vincze, "Ontological knowledge management framework for grasping and manipulation," in *IROS Workshop: Knowledge Representation for Autonomous Robots*, 2011.

[24] C. Li and G. Tian, "Transferring the semantic constraints in human manipulation behaviors to robots," *Applied Intelligence*, vol. 50, no. 6, pp. 1711–1724, 2020.

[25] N. Kleer, M. Feick, and M. Feld, "Leveraging publicly available textual object descriptions for anthropomorphic robotic grasp predictions," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 7476–7483.

[26] D. Krupke, F. Steinicke, P. Lubos, Y. Jonetzko, M. Görner, and J. Zhang, "Comparison of multimodal heading and pointing gestures for co-located mixed reality human-robot interaction," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–9.

[27] R. Grishman and B. M. Sundheim, "Message understanding conference-6: A brief history," in *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996.

[28] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[29] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.

[30] J. R. Napier, "The prehensile movements of the human hand," *The Journal of bone and joint surgery. British volume*, vol. 38, no. 4, pp. 902–913, 1956.

[31] M. R. Cutkosky, "On grasp choice, grasp models, and the design of hands for manufacturing tasks," *IEEE Transactions on robotics and automation*, vol. 5, no. 3, pp. 269–279, 1989.

[32] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic, "The grasp taxonomy of human grasp types," *IEEE Transactions on human-machine systems*, vol. 46, no. 1, pp. 66–77, 2015.

[33] Q. Lu and T. Hermans, "Modeling grasp type improves learning-based grasp planning," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 784–791, 2019.

[34] V. Veselovsky, M. H. Ribeiro, and R. West, "Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks," *arXiv preprint arXiv:2306.07899*, 2023.

[35] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "FLAIR: An easy-to-use framework for state-of-the-art NLP," in *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 2019, pp. 54–59.

[36] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[37] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto, "LUKE: Deep contextualized entity representations with entity-aware self-attention," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020, pp. 6442–6454.

[38] J. Yu, B. Bohnet, and M. Poesio, "Named entity recognition as dependency parsing," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020, pp. 6470–6476.