Grounded language Understanding via Transformers in Italian for Human-Robot Interaction

Claudiu Daniel Hromei[1,*], Danilo Croce[1] and Roberto Basili[1]

[1]University of Roma Tor Vergata, Rome, Italy

Abstract

Situated natural language interactions between humans and robots are strictly necessary for complex applications: communication here implies the reference to the environment shared between a user and the robot. This paper proposes a transformer-based architecture that supports the integration of spatial information (as logical representation) about a semantic map of the environment and the input utterances. The generated interpretation is a logical form of the command that makes references to the state of the world through a single end-to-end process, stimulated at each interaction by an explicit linguistic description of the environment. In this specific work, the end-to-end capability of the targeted transformer is studied with respect to the processing of situated Italian Commands. The obtained experimental results confirm the applicability of transformers to grounded human-robotic interaction, with benefits in terms of both portability of the approach across domains and effectiveness in terms of reachable accuracy. Overall, the proposed architecture outperforms previous approaches and paves the way for sustainable multilingual architectures.

Keywords

Grounded Semantic Role Labeling, Human-Robot Interaction, End to End Sequence to Sequence Architectures, Robotics and Perception, Natural Language Understanding for Italian

1. Introduction

Ensuring that virtual assistants and robotic platforms understand human language is becoming increasingly important as these technologies become more prevalent in daily life. Virtual assistants are designed to fulfill various user needs, such as finding information or entertainment. Understanding commands and requests is thus crucial for satisfying these needs in a natural manner. This is especially important in critical scenarios, such as those involving robotic platforms performing sensitive or medical tasks that are typically controlled by speech [1, 2, 3]. The article [3] proposes the use of an innovative intelligent rehabilitation robot, *HeAL9000*. The robot is equipped with natural language interpretation capabilities to communicate with patients, manage dialogues, and coordinate physiotherapy sessions. Its primary objective is to assist patients in performing exercises aimed at restoring the use of an injured limb by providing instructions on the required movements. In the future, natural language could be a key factor in controlling these platforms as teaching them the movements or actions required for a task can be done vocally. Currently, domestic robots are used for tasks such as cleaning and cooking, but they face complex challenges, including self-localization, object and people recognition, physical object manipulation, and meaningful interaction with humans to fulfill their needs, as outlined in [4].

Home automation assistants need to be aware of their surroundings and the objects within them. To accurately interpret commands such as

$$\text{“}\textit{Take the volume on the table near the window}\text{”} \quad (1)$$

the assistant must have the capability to associate entities and retrieve the objects mentioned in the command (such as a *volume*, a *table*, and a *window*) while disambiguating between entities of the same type. For example, if there are multiple *tables*, the assistant should choose the one intended by the user, as it is *near the window*.

Several studies, including [5], propose specific methods for grounded language interpretation of robotic commands. In this paper, the approach presented in [6], known as *Grounded language Understanding via Transformers* (GrUT) is investigated. The article presented in [6] suggests using a Transformer-based architecture, such as BART [7], to produce linguistic interpretations of utterances. The architecture takes in the transcription of the input command prompted through the linguistic description of the surrounding map: this includes entities and their spatial and constitutive properties described through a natural language template. The output for the command (1) is the linguistic interpretation consistent with the Frame Semantics [8]:

$$\begin{aligned} &\textit{BRINGING}(\textit{THEME}(\text{“}\textit{the volume}\text{”}),\\ &\quad \textit{GOAL}(\text{“}\textit{on the table near the window}\text{”})) \end{aligned} \quad (2)$$

The meaning is that the robot is requested to take *the volume*, wherever it is and BRING it *on the table* that is

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy

*Corresponding author.

hromei@ing.uniroma2.it (C. D. Hromei); croce@info.uniroma2.it (D. Croce); basili@info.uniroma2.it (R. Basili)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

near the window. On the other hand, if *the volume* is already close to the window, the robot is requested to move there and just TAKE it. In essence, different dispositions of objects in the environment result in different interpretations. In this case, the output is:

$$TAKING(THEME(\text{“}\textit{the volume on the table near the window}\text{”})) \quad (3)$$

In general, as in [5], it is assumed that entities in an environment are denoted by one or more linguistic labels to enable grounded co-reference. Every entity in the Knowledge Base (KB) is identified through a unique identifier, e.g. b_1, and a conceptual category, e.g. BOOK, that describes its properties. For example, linguistic references, such as *volume* or *book*, correspond to a given object identified by b_1, that is in fact a book. GrUT[6] has been designed as an appealing approach, based on Transformers, for grounded language understanding as it reduces the need for task-specific model engineering and leverages the state-of-the-art Transformer paradigm, as in [9]. It only relies on a linguistically described map able to force the Transformer to generate grounded interpretations, consistent with both the input utterance and the KB. As a consequence, the same command can produce different linguistic interpretations when combined with different map descriptions.

In GrUT, the method adopted to inject knowledge about the environment (the map) is *textification*, i.e. the automatic compilation of a text describing the Semantic Map, hereafter referred to as *Textual Map Description*. The description is then added to input utterances in order to trigger the underlying Transformer. Objects of the map that are relevant to the input command are retrieved and described in natural language: this aims at leveraging the contextual awareness of the overall architecture. As a result, an end-to-end process able to directly compile the logic interpretation of commands is realized. Notice that the logic form achieved is consistent with both the utterance *and* the environment.

Moreover, the possibility of directly compiling grounded interpretations is explored. In this case, grounding means correctly associating each entity in the environment with the unique interpretation of arguments. As an example, the interpretations in Equations (2) and (3) refer to portions of text from the input command (e.g. *“the volume”*) rather than any entity. To do this, a way to refer to the entities is needed: in the Semantic Map, each entity is associated with a unique identifier. In a logic-oriented formalism for the output, the grounded interpretation could make use of identifiers as arguments of frame predicates, rather than spans of text: this would link the interpretation to the correct b_1 instance (i.e. the referred *volume*) and to t_1 for the corresponding *table*, as it is near the window. It is worth noting that the connection between words in a spoken command and the entities (here referred to as *linguistic grounding*) is not trivial. The result for the same command (1) is the following interpretation, where the spans of text are replaced with the entities identifiers:

$$BRINGING(THEME(b_1), GOAL(t_1)) \quad (4)$$

Frame Semantics is a language-independent approach, which means that it can be applied to various languages. In [10], GrUT was used to interpret commands in English. In this work, we apply GrUT to the interpretation of Italian commands since the grounded predicate for the Italian utterance should remain the same. Moreover, the application of this approach to Italian data could pave the way for an expansion of the *HeAL9000* experimentation [3]. The interpretation in *HeAL9000* is based on complex cascades of classifiers and handcrafted features, while this approach just needs a way to *textify* the information of the environment. Our experimental evaluation demonstrates that GrUT can achieve comparable quality when applied in either English or Italian without any specific adjustments to the model. In this work, we only assume that objects in the maps are referred to using lexical references in Italian.

In the following, section 2 presents the application of GrUT to grounded interpretation for Italian, section 3 reports the experimental evaluation, while section 4 derives some conclusions.

2. The GrUT approach

Transformer-based architectures like BART [7], T5 [11], their multi-lingual versions mBART [12], mT5 [13], and their Italian version BART-IT [14] and IT5 [15], have proven to be highly effective for the semantic interpretation of spoken commands and texts. Essentially, a Transformer takes natural language input and “translates” it into an artificial text that reflects the underlying linguistic predicate. The approach called *Grounded language Understanding via Transformers* (GrUT) [6] aims to extend the application of Transformers to Grounded SRL tasks by incorporating a natural language description of the map into the input sentence (*Textual Map Description*). To ensure that the interpretation is sensitive to entities’ properties, positions, and relational information (e.g., proximity or distance), a way to refer to them is necessary. This is achieved by identifying each entity through its noun, typically its most commonly used lexical reference (e.g., the word *“volume”*), and its conceptual type. The entity’s association with the environment (i.e., grounding) is accomplished through its identifier (Existence Constraint, *EC*), which is linked to the corresponding physical object’s position in the environment. For example, the map to be paired with the command in (1) must include at least

the following text connected with the corresponding EC: *"b_1, also known as volume or book, is an instance of the class BOOK, t_1, also known as table, is an instance of the class TABLE and w_1, also known as window, is an instance of the class WINDOW"*. All the entities described through the *EC* sentence are retrieved from the Semantic Map using a Retrieval Policy function. For more information about the functions used to select entities from a map based on a given command, please see [10]. Moreover, a Proximity Constraint (*PC*) acting over the selected entities is added in order to state which entities are close to each other in the environment. *PC: "b_1 is near w_1 and w_1 is near t_1"*. Finally, a Containability Constraint (*CC*), for each selected entity, to indicate whether it has the property of containing other objects is added. As an example, assume here the existence of a hypothetical cup, the CC would be: *"c_1 can contain other objects"*.

All the constraints derivable from a given map are appended, as a header, to the input command. In the resulting text, each constraint is separated from the next one by a "#" character as a delimiter. The resulting header serves as a micro-story that describes all the relevant properties of the underlying map in support of the SRL model, in order to distinguish between different situations in the environment. In this work, only these three constraints are defined and used, but in the future, a broader range of properties to enhance the analysis could be used.

In summarizing, when the necessary condition of the spatial constraint *PC* is true, the correct interpretation for the ambiguous situation, introduced in (1), corresponds to the following linguistic logical form:

$$TAKING\big(THEME(\text{``}\textit{the volume on the table near the window}\text{''})\big). \tag{5}$$

Since the book b_1, referenced through the noun *volume*, is close to t_1, it is interpreted thus as the THEME of the TAKING predicate. It is worth noticing that the linguistic description of the map enables the use of highly accurate transformers (such as BART [7] and T5 [11]). These are pre-trained on large natural language corpora and may take advantage of linguistic features and cross-dependencies to properly carry out SRL on the overall textual examples made by the informative pairs in GrUT.

In order to bring the interpretation from the linguistic level to the situated level, i.e. referring to the specific entities, it is necessary to link the identifiers (from the *EC* description) to the portions of the text in the command that refers to that specific entity. The transformer is trained to generate a predicate that conveys both the semantic information and the objects involved. The input for the command (1) to the SRL model remains consistent with that employed by GrUT in [6], but the output is:

$$TAKING\big(THEME(b_1)\big) \tag{6}$$

Notice that b_1 still refers to *the volume* defined in the *EC* previously. This schema results in a distinct input in scenarios where book b_1 is positioned **far** from table t_1. The map description of this scenario will include *PC: "b_1 is far from t_1 and t_1 is near w_1"*, while the output changes significantly:

$$BRINGING\big(THEME(b_1), GOAL(t_1)\big) \tag{7}$$

In this scenario, the robot is expected to take *the volume* b_1 from its position and BRING it to the GOAL, i.e. the position of t_1 *the table*.

In this paper, the Transformer is used to generate direct Grounded Interpretations of input commands for end-to-end processing. These interpretations are presented as logical forms consisting of Frames and Frame Elements [8], with the entity identifiers acting as fillers. In the following experimental evaluation, GrUT is applied in the interpretation of commands in Italian.

3. Experimental Evaluation

The evaluation is carried out in a home automation scenario, in which a robot receives spoken commands and interprets them to perform various actions, such as picking up a book, taking out the rubbish, or looking for the keys. The evaluation relies on the HuRIC[1] dataset, comprising 656 English and 241 Italian voice commands with corresponding interpretations in terms of predicates and arguments. The Grounding process described in the previous section is applied to these interpretations by linking them with entity identifiers in the surrounding environment. Predicates in HuRIC are defined based on a subset of the semantic frames in FrameNet [8], and their corresponding arguments are selected. Following the previous evaluation in [10], a 10-fold cross-validation scheme with an 80/10/10 data split between training, validation, and test sets is adopted to evaluate the same aspects of the previous versions of GrUT.

Given a command in natural (Italian) language, the GrUT approach was applied to fine-tune a wide range of models. The map description is generated using predefined templates, as in [10], in the Italian language, describing the entities evoked by the command, concatenated to the input as shown in section 2. To retrieve the entities from the robot's knowledge base, [10] proposes three Entity Retrieval Policies. In this paper, only the policy based on the Neural Semantic Similarity (NSS) is adopted, as it has been shown in [10] to be the best-performing policy in the end-to-end process.

To assess the quality of the overall interpretation process, the following tasks are evaluated and the results are reported in terms of F1 score: *i*) **Frame Prediction**

[1] https://github.com/crux82/huric

(FP), which measures the ability of the models to accurately generate the names of Frames evoked by the voice command; *ii*) **Argument Identification and Classification** as an **Exact Match** (AIC-ExM), where the system's ability to correctly generate the names of the Arguments evoked by the command and associate them with the entities that evoke that Argument is assessed; *iii*) **Argument Identification and Classification** as a **Head Match** (AIC-HeM) strategy, which is more relaxed than the ExM measure, and when it is performed the model is rewarded if the Argument contains at least the Semantic Head of the linguistic reference to the correct Entity. For example, let the SRL of the utterance "*Prendi il libro*" be

$$\text{TAKING}\big(\text{THEME}(\text{“}\textit{libro}\text{”})\big). \tag{8}$$

For the AIC-ExM test it has a score of 0, as "*libro*" instead of the entire correct span, i.e., "*il libro*", is returned as a filler of the THEME argument. However, "*libro*" is the semantic head[2] of the expression pointing to the entity b_1 in the class of BOOKs. For this reason, the AIC-HeM strategy assigns a score of 1 to the THEME argument of the system interpretation in Eq. (8).

3.1. Results

In this work, models such as mT5[13], IT5[15] and BART-IT[14] were trained with different learning rates (*lr*). The parameters used to fine-tune all the models here presented are the same as the ones in [10]. The Italian dataset used in this work is relatively small, consisting of only 241 utterances. The training dataset critically underrepresents some phenomena: Frames such as INSPECTING or GIVING account for only 2% (5/241 utterances) each of the total data samples. Due to the pre-training of the LLMs on a vast amount of data, fine-tuning them with such a small number of samples could make this task very challenging. In order to expand the Italian dataset, we translated the English version of HuRIC into Italian while preserving the original contexts in which they were uttered and paired them with the original interpretation. This takes advantage of the fact that the Frame Semantics theory produces language-independent interpretations, making it possible to reuse the logical forms. The translated dataset is used as additional training data (656 translated utterances), while the original Italian dataset is divided into training, evaluation, and testing with a ratio of 80/10/10 using a 10-fold cross-validation method.

[2]In this paper, the notion of semantic head has two different usages. First, it is the usual content word, such as "*libro*", the grammatical head in a phrase like "*il libro*": For the AIC-HeM strategy, it is admissible that the incomplete fillers are accepted if they include such heads. An alternative usage occurs when GrUT outputs grounded symbols as an argument's fillers, like b_1. As they are identifiers of entities then they are accepted by the AIC-HeM strategy if they correspond to the correct reference. Thus, the correct identifiers will get a score of 1 for the AIC-HeM measure.

DeepL[3] for automatic neural machine translation of English utterances has been used. To provide context for the translation, each lexical references for each entity ($LR(e_i)$) of the Semantic Map has been coupled with the corresponding English utterance. The Semantic Map defines a variety of entities in the environment in which the command was uttered. For example, the English noun *"glass"* can refer to either a brittle transparent solid or a glass container for holding liquids. Without context, the resulting translation in Italian could be ambiguous. In this work, no direct evaluation of the translation method was performed, the purpose is to extend the Italian dataset, and the quality of the translation is reflected in the models' performance.

Table 1 presents the results for the approach applied to Italian data. In the first row, LU4R [5] is reported as a soft baseline, which is the only model that performs SRL on the Italian dataset. However, its results are not directly comparable to the models evaluated in this work, as LU4R does not perform any Grounding step.

Table 1
Comparative Evaluation of GrUT in terms of F1. In **bold** the best performance for each task, in *italic* the performance of the LU4R model as it is not directly comparable.

Model	LR	FP	AIC-ExM	AIC-HeM
LU4R [5]	-	*95.32*	*77.67*	*86.35*
mT5	$1 \cdot 10^{-3}$	82.26	59.36	66.43
	$1 \cdot 10^{-4}$	90.61	73,21	82.89
IT5	$1 \cdot 10^{-4}$	72.00	64.44	65.97
	$5 \cdot 10^{-5}$	65.48	60.02	61.85
BART-IT	$1 \cdot 10^{-4}$	94.45	76.61	79.5
	$5 \cdot 10^{-5}$	**96.86**	**82.30**	**85.19**
	$2 \cdot 10^{-5}$	95.17	79.47	82.89

BART-IT with a *lr* of $5 \cdot 10^{-5}$ achieves state-of-the-art performance for both tasks: 96.86% F1 on the Frame Prediction task, and 82.10% as Exact Match and 85.19% as Head Match on the AIC task. It's important to note that LU4R solves the overall SRL task as a cascade of classification for each token in the utterance, while the Transformers evaluated here generate text for the interpretation task, and the FP and AIC subtasks are just a side effect of the entire process. The IT5 and mT5 models achieved a lower performance. The Italian dataset is relatively small and these LLMs are not able to effectively leverage the internal attention mechanism, as they contain more parameters. Some errors still persist, particularly for frames that are underrepresented in the training data, such as the GIVING frame. Let the command be *"Dammi le chiavi per favore"* and the description of the Semantic

[3]The translation was performed in February 2023 at https://deepl.com.

Map be composed of only one entity (*PC*: r_9 *conosciuto anche come chiavi è un'istanza della classe CHIAVI*). The Gold Standard interpretation is:

$$GIVING\big(RECIPIENT(\text{“}mi\text{”}), THEME(r_9)\big) \quad (9)$$

The prediction of the best-performing BART-IT model is similar to the Gold Standard. It differs only in frame written: BRINGING instead of GIVING. Nonetheless, the action implied by both interpretations (the predicted and the correct one) is the same: the robot is requested to navigate the environment, take the keys (r_9) wherever they are and give them to the speaker. Both GIVING and BRINGING should implement this behavior.

4. Conclusions

This work represents an extension to the GrUT approach from [10], initially developed using BART for the English language, to evaluate its applicability to the Italian language. In this study, various models were explored and evaluated for their effectiveness in handling Italian language commands. The experimental results suggest that state-of-the-art performance can be achieved in Italian by appropriately combining language-specific models that are competitive in terms of processing quality with respect to language-independent models.

There are several potential avenues for future research based on these findings. Primarily, these results suggest that an application of the GrUT approach to the *HeAL9000* experimentation could be possible in order to improve the interpretation performance. Moreover, portability to larger sets of natural languages should be tested by training, for example, mT5. Finally, the possibility of instructing robots to interact in a question-answering scenario about the Semantic Maps is appealing.

Acknowledgements The research for this paper was partly funded by Regione Lazio (prot. A0320-2019-28108).

References

[1] K. Zinchenko, C.-Y. Wu, K.-T. Song, A study on speech recognition control for a surgical robot, IEEE Transactions on Industrial Informatics 13 (2017) 607–615.

[2] U. Maniscalco, A. Messina, P. Storniolo, The human-robot interaction in robot-aided medical care, in: Proceedings of KES-HCIS 2020 Conference, Split, Croatia, June 17-19, 2020, volume 189, Springer, 2020, pp. 233–242.

[3] L. Cristofori, C. D. Hromei, F. S. di Luzio, C. Tamantini, F. Cordella, D. Croce, L. Zollo, R. Basili, Heal9000: an intelligent rehabilitation robot, in: Proceedings of the Workshop on Towards Smarter Health Care, Anywhere, November 29th, 2021, volume 3060, CEUR-WS.org, 2021, pp. 29–41.

[4] M. E. Foster, Natural language generation for social robotics: opportunities and challenges, Philosophical Transactions of the Royal Society B: Biological Sciences 374 (2019) 20180027.

[5] A. Vanzo, D. Croce, E. Bastianelli, R. Basili, D. Nardi, Grounded language interpretation of robotic commands through structured learning, Artif. Intell. 278 (2020).

[6] C. D. Hromei, L. Cristofori, D. Croce, R. Basili, Embedding contextual information in seq2seq models for grounded semantic role labeling, in: AIxIA 2022 – Advances in Artificial Intelligence, Springer International Publishing, Cham, 2023, pp. 472–485.

[7] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, ArXiv abs/1910.13461 (2020).

[8] C. J. Fillmore, Frames and the semantics of understanding, Quaderni di Semantica 6 (1985) 222–254.

[9] R. Blloshmi, S. Conia, R. Tripodi, R. Navigli, Generating senses and roles: An end-to-end model for dependency- and span-based semantic role labeling, in: Proceedings of IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021, ijcai.org, 2021, pp. 3786–3793.

[10] C. D. Hromei, D. Croce, R. Basili, Grounding end-to-end architectures for semantic role labeling in human robot interaction, in: Proceedings of NL4AI 2022, Udine, November 30th, 2022, volume 3287, 2022, pp. 24–38.

[11] C. Raffel, N. M. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, ArXiv abs/1910.10683 (2020).

[12] Y. Tang, C. Tran, X. Li, P. Chen, N. Goyal, V. Chaudhary, J. Gu, A. Fan, Multilingual translation with extensible multilingual pretraining and finetuning, CoRR abs/2008.00401 (2020).

[13] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, in: Proceedings of the 2021 Conference of NAACL-HLT 2021, Online, June 6-11, 2021, ACL, 2021, pp. 483–498.

[14] M. La Quatra, L. Cagliero, Bart-it: An efficient sequence-to-sequence model for italian text summarization, Future Internet 15 (2023).

[15] G. Sarti, M. Nissim, It5: Large-scale text-to-text pretraining for italian language understanding and generation, ArXiv preprint 2203.03759 (2022).