Generating Pseudo-Natural Language Explanations for Goal Selection*

Henrique M. R. Jasinski[0000−0003−3219−3073], Mariela Morveli-Espinoza[0000−0002−7376−2271], and Cesar A. Tacla[0000−0002−8244−8970]

Graduate Program in Electrical and Computer Engineering (CPGEI)
Federal University of Technology - Parana (UTFPR), Curitiba, Brazil
henriquejasinski@alunos.utfpr.edu.br, morveli.espinoza@gmail.com, tacla@utfpr.edu.br

Explainable Artificial Intelligence systems, including intelligent agents, must explain their internal decisions to other software agents or to human users. In the latter case, it is necessary that humans understand the outputs returned by the systems. Thus, like it was done in [2] and [3], we use a pseudo-natural language for improving the understanding of the explanations given by software agents.

ArgAgent[1] is a simulator for Belief-Based Goal Processing (BBGP) [1] which is an extension of the Belief-Desire-Intention (BDI) model [6]. Argumentation is used in the intention formation process, which is comprised of four sequential stages: **activation**, **evaluation**, **deliberation**, and **checking**. In this demo, we focus on the deliberation stage, whose purpose is to identify conflicts among goals and select which goals the agent will commit to. These conflicts can be: a) terminal (denoted by t), b) resource (denoted by r), or c) superfluity (denoted by s)[2]. Thus, the aim is to generate explanations for the question: why did the agent commit (or not) to a given goal? Besides, we enrich such explanations with information about the type of conflict that arose between goals.

The simulator input is an agent model, composed of initial beliefs, rules, preference values of goals, and plans. Explanatory arguments are constructed based on a set of six-domain-independent deliberation rules , whose premises are unified with beliefs that are generated from the Goal Argumentation Framework – which is constructed to determinate the set of chosen goals – and the set of chosen goals, which were obtained after applying a semantics based on conflict-free sets and a function that selects the extension that maximizes utility.

The explanations are generated from the set of explanatory arguments and the possible attacks between them, following the method presented in [5]. The pseudo-natural language explanations are built from: (i) the unified deliberation rules, which are used to construct explanatory arguments; and (ii) the respective schemes used to generate explanatory sentences from the explanatory arguments. Keywords, identified by '<' and '>' symbols, are replaced for an internal element (eg., predicate name, full predicate, or a term) of the argument itself.

* This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 - and the Centro Nacional de Pesquisa - Brasil (CNPq).

[1] Available at: https://github.com/henriquermonteiro/BBGP-Agent-Simulator/

[2] Further explanation of such conflicts is available in [4].

CMNA'20 - F. Grasso, N. Green, J. Schneider, S. Wells (Eds) - 8th September 2020

r_1	Rule	$incompatible(g, g', conflict), preferred(g, g') \rightarrow pursued(g)$
	Scheme	<goal_name_g> and <goal_name_g'> have the following conflicts: <predicate_1term_2>. Since <goal_name_g> is more preferable than <goal_name_g'>, I decided to commit to <goal_name_g>.
r_2	Rule	$maxUtility(g) \rightarrow pursued(g)$
	Scheme	Since <goal_name_g> belonged to the set of goals that maximize the utility, I decided to commit to <goal_name_g>.

Table 1: Examples of deliberation rules and explanations schemes

```
Cycle: 004 (Why mop(p1,p1))?
 > mop(p1,p1) and pickup(p5,p5) have the following conflicts: r.
     Since mop(p1,p1) is more preferable than pickup(p5,p5), I
     decided to commit to mop(p1,p1).
 > Since mop(p1,p1) belonged to the set of goals that maximize the
     utility, I decided to commit to mop(p1,p1).
```

Fig. 1: Pseudo-natural language explanation for ‘Why mop(p1,p1)?’

Table 1 shows the rules and explanation schemes used in the example shown in Fig. 1, where the explanation for ‘Why $mop(p1, p1)$?’ is given. Both rules are necessary to identify under what circumstances (conflicts and utility function) the goal was selected. In this example the agent seeks to clean its surroundings and at a given moment he has two options: to mop $at(p1, p1)$ or to pickup litter on $at(p5, p5)$. Both options can not be pursued at the same moment because the lack of resources available (conflict denoted by r). The agent then choose – based on argumentation process – what he believes is the best option ($mop(p1, p1)$).

In future work we seek to give a more complete explanation for the agents behaviour. Some work has already be done by extending the explanations for other stages of the intention formation process mentioned above. That includes allowing the agent designer to define explanation schemes for new rules.

References

1. Castelfranchi, C., Paglieri, F.: The role of beliefs in goal dynamics: prolegomena to a constructive theory of intentions. Synthese **155**(2), 237–263 (feb 2007)
2. Guerrero, E., Nieves, J.C., Lindgren, H.: An activity-centric argumentation framework for assistive technology aimed at improving health. Argument & Computation **7**(1), 5–33 (2016)
3. Koeman, V., Dennis, L.A., Webster, M., Fisher, M., Hindriks, K.: The “Why did you do that?” Button: Answering Why-questions for end users of Robotic Systems. Proc. of the 7th International Workshop on Engineering Multi-Agent Systems (EMAS) pp. 1–19 (2019)
4. Morveli-Espinoza, M.M., Nieves, J.C., Possebom, A.T., Puyol-Gruart, J., Tacla, C.A.: An argumentation-based approach for identifying and dealing with incompatibilities among procedural goals. International Journal of Approximate Reasoning **105**, 1–26 (2019)
5. Morveli-Espinoza, M., Tacla, C.A.: An argumentation-based approach for explaining goals selection in intelligent agents. In: 9th Brazilian Conference on Intelligent Systems (BRACIS) (2020)
6. Rao, A.S., Georgeff, M.P.: BDI agents: from theory to practice. In: ICMAS. vol. 95, pp. 312–319 (1995)