Towards a Types-As-Classifiers Approach to Dialogue Processing in Human-Robot Interaction

Julian Hough[1,*], **Lorenzo Jamone**[1,†]
David Schlangen[2,‡], **Guillaume Walck**[2,§] and **Robert Haschke**[2,§]
[1] * Cognitive Science Group, †Centre for Advanced Robotics,
School of Electronic Engineering and Computer Science,
Queen Mary University of London, UK
[2] ‡Dialogue Systems Group, §Neuroinformatics Group,
CITEC, Bielefeld University, Germany
j.hough@qmul.ac.uk

Abstract

We propose a novel Types-As-Classifiers approach to dialogue processing for robots using probabilistic type judgments. In our proposal, incoming sensory data is converted to a world belief record in real time, and then derived beliefs such as intention attribution to a user, or the prediction of affordances of visible objects, are made as record type judgements of that record. The record can be updated dynamically like a dialogue state, allowing information of different perceptual sources to be easily combined in real time.

1 Introduction

The combination of computer vision and natural language processing is now incredibly popular. Thanks to increased computing power and the development of new deep learning techniques, huge strides forward have been made in several tasks, including: automatic image retrieval from key words, reference resolution of objects in photographs from text (Kennington and Schlangen, 2015), generating referring expressions to objects given probabilistic estimation of object properties (Mast et al., 2016), caption generation and visual question answering (Antol et al., 2015).

A more challenging task, beyond the use of single sentence texts with images, is the creation of dialogue systems designed for real-world human-robot interaction (HRI) which combines probabilistic information encoding visual and physical properties of objects and information about the interaction more commonly encoded in a dialogue state. This uniform approach not only requires the use of complex visual information and semantic parsing, but needs to permit fluid interaction with a collaborative robot to help a user complete a manual task. This requires an incrementally and dynamically evolving dialogue state which encodes the robot's own action state as well as its estimation of the user's intentions in real time.

In this paper we address this challenge by formulating a simple interaction state for a robot using concepts from Type Theory with Records (TTR) (Cooper, 2005). We characterize the robot's world belief as a constantly updating record, and use type classifiers of different kinds which operate on the state record to make type judgements on the world belief. Once a judgement is made and used (committed), this can be added to the world belief for further classification and update. For the classification we use a combination of lattice theory and probabilistic TTR (Cooper et al., 2014). Inspired by the recent work using TTR for perceptual classification (Dobnik et al., 2012; Yu et al., 2016) and the simple Words-As-Classifiers (WAC) model (Kennington and Schlangen, 2015) to reference resolution of objects in real-world scenes, here we propose a general Types-As-Classifiers (TAC) approach.

2 Types-As-Classifiers for human-robot interaction

Typical raw perceptual information for a collaborative pick-and-place robot may be as in Fig. 1. The left side shows a camera feed, and computer vision based segmentation and tracking of objects as described in (Ückermann et al., 2014a,b), and perceptual classifiers, such as that for 'yellow', which classify the degree to which an object has that perceptual property. The current words recognized by the robot's speech recognizer (ASR) are also added to the state as they arrive. On the right side, the diagram shows how the robot tracks its own current task state and action state of its arm through a Hierarchical State Machine (HSM).

SCENE:

OBJECTS (segmentation and visual classifiers):
object_0:
yellow = 0.69
blue = 0.38
..
object_1:
yellow = 0.10
blue = 0.86
...

USER SPEECH (current user utterance):
'put the left green apple in the basket'

ROBOT ACTION AND TASK STATE:

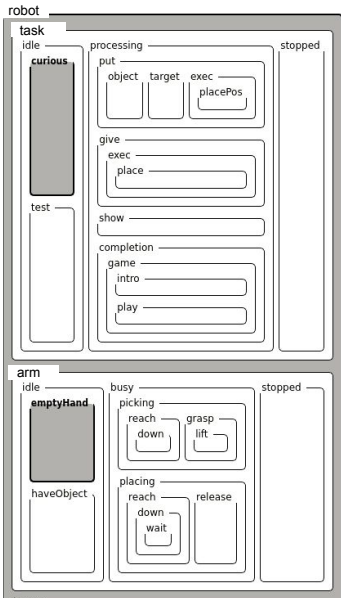


Figure 1: A typical state according to the robot. Objects are segmented and properties can be obtained for each object. The robot's internal action state is controlled by a Hierarchical State Machine (HSM)

2.1 Encoding the robot's sensory state as an updating TTR record

In this paper we use TTR *record types*, and the inhabitants of record types, *records*, as our primary formal apparatus – see Cooper (2005) for details. We characterize the state as a *world belief record*- for an in-robot control system for our purposes it will be of the format in (1).[1]

$$\left[\begin{array}{lll} objects & = & \left[\begin{array}{lll} obj\_0 & = & [\ ...\ =\ ...\ ] \\ obj\_1 & = & [\ ...\ =\ ...\ ] \\ ... & = & ... \\ obj\_N & = & [\ ...\ =\ ...\ ] \end{array}\right] \\ robot & = & \left[\begin{array}{lll} arm & = & [\ ...\ =\ ...\ ] \\ task & = & [\ ...\ =\ ...\ ] \\ intention & = & [\ ...\ =\ ...\ ] \end{array}\right] \\ human & = & \left[\begin{array}{lll} c{-}utt & = & \left[\begin{array}{lll} parse & = & ... \\ words & = & ... \end{array}\right] \\ status & = & ... \\ intention & = & [\ ...\ =\ ...\ ] \end{array}\right] \end{array}\right] \quad (1)$$

For HSMs as in Fig. 1, we can formulate the state at a given time as a record via the use of recursive structure. The record gets constructed from the highest level down, whereby each parallel/concurrent state, such as the *task* and *arm* substates of *robot* in Fig. 1, are encoded as separate fields in the record. If the current state is an embedded substate, for example the *emptyHand* and *holdsObject* substates within the *idle* substate of the *arm* state in Fig. 1, that will be encoded in the record structure as an embedded record (a record within a record). When a state is atomic, that will be encoded as a single value in the record.

Given this recursive formulation, the robot's current action and task state as shown by the darkened areas in Fig. 1 can be formulated as in (2). This is an efficient way of encoding the state, as not all the inactive substates need be encoded.

$$\left[\ robot\ =\ \left[\begin{array}{lll} task & = & [\ idle\ =\ curious\ ] \\ arm & = & [\ idle\ =\ emptyHand\ ] \end{array}\right]\ \right] \quad (2)$$

3 Record Type classifiers applied to the world belief for higher-level perception

The driving incremental interpretation process of the system is a probabilistic classification of the current world belief record wb (with the structure in (1)) as being of a given situation record type i within a set of possible record types I, conditioned by current evidence record type e.

In the following sub-sections we outline different perceptual classifiers which operate on wb to get the probability judgement that wb is of a given type. This can be done recursively, as once a type judgement is made (for a given purpose), this can be added to wb, and then further judgements of its

[1] This is an example record where many of the labels and values are just represented by '...' to indicate at least one such field would be present in the full representation.

$$i = \left[\begin{array}{l} human : \left[\begin{array}{l} intention : \left[\begin{array}{ll} goal & : \left[\begin{array}{ll} landmark & : obj\_2 \\ rel\_location & : INTO \end{array} \right] \\ objects & : \{obj\_1\} \\ action & : PUT \end{array} \right] \end{array} \right] \end{array} \right]$$

Figure 2: A user intention record type to effect the movement of an object.

type can be made and added to it. While we suggest a pipeline here by presentation order, we are not committed to a specific classification ordering or algorithm for inter-leaving these processes, and leave investigation into this for future work. However, we are committed to the distribution over possible record type judgements being stored in a record type lattice– see (Hough and Purver, 2017).

3.1 Perceptual classification 1: predicting object affordances

The robot's perception of object properties is vital for complex interaction with the human user. Specifically, the perception of object *affordances* (Gibson, 1979), i.e. the possible actions associated to the objects (e.g. *graspable*), is crucial for the robot to be able to manipulate them (Jamone et al., 2016). Recently, probabilistic computational models of affordance perception have been proposed, using Bayesian Networks (Gonçalves et al., 2014) and variational auto-encoders (Dehban et al., 2016)- these can be used obtain the probability of an object having different affordances from visual and linguistic features. In our model, affordance prediction is part of the probabilistic type judgement of wb, such that the probabilities of each object having each affordance property are part of the available type judgements. In future work, we will investigate how affordance prediction can best integrate with natural language processing decisions – e.g. (Salvi et al., 2012).

3.2 Perceptual classification 2: parsing

The next higher-level perception classification is the incremental semantic parsing of the recognized words from the ASR. For this we use the Dylan ('DYnamics of LANguage') parser (Purver et al., 2011).[2] The parser fulfills the criteria for incremental semantic construction outlined in (Hough et al., 2015): it consumes words one-by-one and outputs a maximal semantic record type (RT) based on a pre-defined Dynamic Syntax-TTR (DS-TTR) grammar– see (Eshghi et al., 2011) for full details. A typical parse for 'put the red apple in the big basket' is as in (3):

$$\left[\begin{array}{ll} r1 & : \left[\begin{array}{ll} x & : e \\ p_{=basket(x)} & : t \\ p1_{=big(x)} & : t \end{array} \right] \\ x2_{=\iota(r.x)} & : e \\ r & : \left[\begin{array}{ll} x & : e \\ p_{=apple(x)} & : t \\ p1_{=red(x)} & : t \end{array} \right] \\ x1_{=\iota(r.x)} & : e \\ e2_{=INTO} & : es \\ x_{=addressee} & : e \\ e_{=PUT} & : es \\ p3_{=obj(e2,x2)} & : t \\ p2_{=indObj(e,e2)} & : t \\ p1_{=obj(ev,x1)} & : t \\ p_{=subj(ev,x)} & : t \end{array} \right] \quad (3)$$

The best parse is added to the $human.c{-}utt.parse$ field. Now other inference can be done using this information, primarily recognizing the user's intention word-by-word.

3.3 Perceptual classification 3: user intention recognition

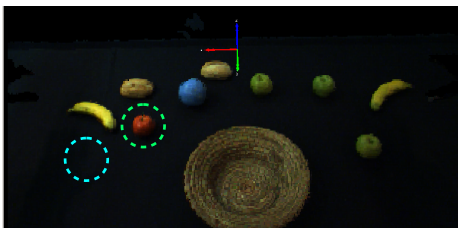
As DyLan's DS-TTR parser provides RTs word-by-word incrementally, the user's intention can also be estimated word-by-word as wb is updated. Given a set of possible user intention record types I, where a typical intention may look like i in Fig. 2, and the conditioning evidence e, a record type representing a sub-part of wb, we characterize a standard Maximum Likelihood multi-class probabilistic classifier to estimate the best prediction for the $human.intention$ field and its probability (or *confidence*) in its prediction $Ev(human.intention)$ by the standard $arg\ max$ and max functions in (4) and (5), respectively.

$$human.intention = \arg\max_{i \in I} p(wb : i | wb : e) \quad (4)$$

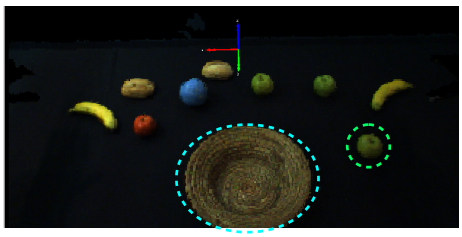
$$Ev(human.intention) = \max_{i \in I} p(wb : i | wb : e) \quad (5)$$

In our current implementation, e simply consists in judgements on the $human.c{-}utt.parse$

[2] Available open-source at https://bitbucket.org/dylandialoguesystem/dsttr.



put the apple in front of the banana



... in the basket

Figure 3: Syntactic ambiguity causing the system changing its top hypothesis about the user's intention.

and $objects$ fields of wb, but it can be more than these, and in future, we plan to learn which parts are relevant for estimating user intentions.

In our current implementation, to calculate the conditional likelihood $p(wb : i \,|\, wb : e)$ for two given RTs i and e, we create a directed graph of the current parse RT based on its field dependencies, beginning from the head event field $e_{=PUT}$ (which determines the action), and recursively traverse all fields which depend on it, applying the relevant type classifiers. We match the field values in the embedded entity restrictor RTs such as $red(x)$ to the low-level classifier results in $objects$. If the relevant type judgement (e.g. $red(x)$) appears in the parse, the corresponding low-level classification strength for each object (e.g. $obj\_1.red = 0.8$) will be used, using the product rule to multiply the probability of the relevant fields for a given object. The overall likelihood of $wb : i$ is calculated recursively, beginning with the likelihood of the embedded RTs such as $intention.goal$ and the target objects $intention.objects$. The likelihood of the judgements of each of the embedded fields is multiplied together to get the overall probability of the intention.

3.4 Perceptual classification 4: estimating legibility of robot intentions

Dual to confidence about the user's intention, we can also estimate the *legibility* of the robot's intention (Dragan et al., 2013), which is similar in structure to the human intention in Fig. 2. Legibility is important for estimating when the robot's intention has become distinct enough from other possible intentions, and consequently what can be considered grounded with the user through the robot's action so far (Hough and Schlangen, 2017). We estimate the strength-of-evidence function $Ev(robot.intention)$ as in (6) where e is taken to be all of wb excluding $robot.intention$:

$$Ev(robot.intention) = p(wb : robot.intention | wb : e) \quad (6)$$

(6) is the likelihood that the robot's current intention will be recognized by the user as such. In practice, this legibility measure can be estimated via a number of physics-based methods such as the proximity of the arm to the target object compared to the other objects, or through using movement trajectories– see (Dragan et al., 2013).

4 Conclusion

We have given an overview of a Types-As-Classifiers (TAC) approach to dialogue processing in human-robot interaction. We believe our approach is complementary to the Words-As-Classifiers (WAC) approach to reference resolution (Kennington and Schlangen, 2015), and we believe it brings several advantages. Firstly, it is not constrained by individual word classifiers alone, but can use the structure from a parser to compute likelihood of complex intentions, all the while maintaining word-by-word incrementality. Secondly, it gives a uniform way to process different multimodal information such as robotic task and action states and visual and physical properties of objects within a dialogue state. In future, we intend to show how it allows the different processes to help each other- e.g. the online resolution of parsing ambiguity such as that in Fig. 3, where the first 'in' is taken not to modify 'the apple', but this decision is changed once the user continues talking. We are also planning to test our current implementation with users.

Acknowledgments

We thank the reviewers for their useful comments. This work was supported by the DFG Center of Excellence EXC 277, the DFG Transregional Research Centre CML, TRR-169.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Robin Cooper. 2005. Records and record types in semantic theory. *Journal of Logic and Computation*, 15(2).

Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. 2014. A probabilistic rich type theory for semantic interpretation. In *Proceedings of the EACL Workshop on Type Theory and Natural Language Semantics (TTNLS)*, Gothenburg, Sweden. ACL.

Atabak Dehban, Lorenzo Jamone, Adam R Kampff, and José Santos-Victor. 2016. Denoising auto-encoders for learning of objects and tools affordances in continuous space. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 4866–4871. IEEE.

Simon Dobnik, Robin Cooper, and Staffan Larsson. 2012. Modelling language, action, and perception in type theory with records. In *International Workshop on Constraint Solving and Language Processing*, pages 70–91. Springer.

Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. 2013. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 301–308. IEEE.

Arash Eshghi, Matthew Purver, and Julian Hough. 2011. DyLan: Parser for Dynamic Syntax. Technical Report EECSRR-11-05, School of Electronic Engineering and Computer Science, Queen Mary University of London. ISSN 2043-0167. Available from http://sf.net/projects/dylan/files/dylan/DSImp_TechReport.pdf.

James J Gibson. 1979. The theory of affordances. *The people, place, and space reader*, pages 56–60.

Afonso Gonçalves, João Abrantes, Giovanni Saponaro, Lorenzo Jamone, and Alexandre Bernardino. 2014. Learning intermediate object affordances: Towards the development of a tool concept. In *Joint IEEE International Conferences on Development and Learning and Epigenetic Robotics (ICDL-Epirob), 2014*, pages 482–488. IEEE.

Julian Hough, Casey Kennington, David Schlangen, and Jonathan Ginzburg. 2015. Incremental semantics for dialogue processing: Requirements, and a comparison of two approaches. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 206–216.

Julian Hough and Matthew Purver. 2017. Probabilistic record type lattices for incremental reference processing. In *Modern perspectives in type-theoretical semantics*, pages 189–222. Springer.

Julian Hough and David Schlangen. 2017. It's Not What You Do, It's How You Do It: Grounding Uncertainty for a Simple Robot. In *Proceedings of the 2017 Conference on Human-Robot Interaction (HRI2017)*.

Lorenzo Jamone, Emre Ugur, Angelo Cangelosi, Luciano Fadiga, Alexandre Bernardino, Justus Piater, and José Santos-Victor. 2016. Affordances in psychology, neuroscience and robotics: a survey. *IEEE Transactions on Cognitive and Developmental Systems*.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. Proceedings of the Conference for the Association for Computational Linguistics (ACL). ACL.

Vivien Mast, Zoe Falomir, and Diedrich Wolter. 2016. Probabilistic reference and grounding with pragr for dialogues with robots. *Journal of Experimental & Theoretical Artificial Intelligence*, 28(5):889–911.

Matthew Purver, Arash Eshghi, and Julian Hough. 2011. Incremental semantic construction in a dialogue system. In *Proceedings of the 9th IWCS*, Oxford, UK.

G. Salvi, L. Montesano, A. Bernardino, and J. Santos-Victor. 2012. Language bootstrapping: Learning word meanings from perception-action association. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(3):660–671.

Andre Ückermann, Christof Eibrechter, Robert Haschke, and Helge Ritter. 2014a. Real-time hierarchical scene segmentation and classification. In *Humanoid Robots (Humanoids), 2014 14th IEEE-RAS International Conference on*, pages 225–231. IEEE.

Andr Ückermann, Christof Elbrechter, Robert Haschke, and Helge Ritter. 2014b. Hierarchical Scene Segmentation and Classification.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2016. Training an adaptive dialogue policy for interactive learning of visually grounded word meanings. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 339.