Conceptual Pact Models of Reference in Building Games with Dialogue: Scaling up from Pentomino Puzzles to the challenge of Minecraft

Julian Hough[1], Chris Madge[2],
Matthew Purver,[2], Massimo Poesio[2]

[1]School of Mathematics and Computer Science, Swansea University
[2]School of Electronic Engineering and Computer Science, Queen Mary University of London
Correspondence: julian.hough@swansea.ac.uk

Abstract

A version of Brennan and Clark's conceptual pact model can be applied to referring expressions with some success to the Pentomino puzzle building domain where an instructor gives instructions to build puzzles from a simple set of 12 puzzle pieces. We discuss how such a model could be scaled up to a much more complex domain in the game of Minecraft, outlining the key differences between the two domains and a plan for scaling up the models with language models.

1 Introduction

Following work in embodied reference resolution (Kennington and Schlangen, 2015; Yu et al., 2016; Suglia et al., 2022), conversational grounding (Poesio and Rieser, 2011; Ginzburg, 2012) and language acquisition in the spirit of Steels and Vogt (1997), we explore how a computational model of Brennan and Clark (1996)'s psycholinguistic theory of conceptual pacts in dialogue has had some success in a simple reference domain, and how the challenge of a more complex domain could be met.

2 Modelling Conceptual Pact Building in Dialogue with Language Models

We follow the description of conceptual pacts using language models described by Hough et al. (2024). We capture two ways conceptual pacts can work in conversation: Firstly, different dialogue pairs can develop different pacts for naming different objects which have quite different lexical content, but remain consistent throughout their interaction. Secondly, the convention of naming a object can stabilize over time in the interaction.

To capture the contribution of local conceptual pacts, we use local updating language models for each object r, p_r^{pact}, e.g. in a Pentomino puzzle domain, for the X piece $p_X^{pact}(w_0..w_n)$ gives the probability value that a referring expression $w_0..w_n$ will be used for X based on the previous references to the piece seen so far. For our simulated interactive learning element, we make the simplifying assumption that after trying to resolve $w_0..w_n$, our agent receives a signal of the correct piece then adds $w_0..w_n$ to the training data for the relevant p_r^{pact} model. We allow the possibility of incorporating prior experience from observing other interactions, with language models $p_r^{ex}(w_0..w_n)$. The experience models return the probability of the words being generated to refer to piece r based on prior conversations they have observed and do not update during the current interaction, much like standard static machine learning models. We assume that an effective model will make use of both sources of knowledge, optimally using the locally built language model in combination with the experience model with some weight λ in reference resolution, for example in a simple Bayesian model as in (1).

$$\arg\max_{r \in refs} p_r^{ex}(w_0..w_n) + \lambda p_r^{pact}(w_0..w_n) \cdot p(r) \quad (1)$$

Results on the Pento-CV corpus Hough et al. (2024) show that in a simple reference resolution system trained and tested on the PentoRef Pento-CV corpus (Zarrieß et al., 2016), using the probabilities from these combined pact models as features improved accuracy compared to an equivalent static system. As can be seen in Figure 1, some pieces, like the red X piece (left graph) have very distinct separation in their models' probabilities being applied to their references compared to those of the other pieces, while some, like the N piece (right graph) take longer to separate out from some competitor piece models. When trained on 7 dialogues and using the updating LM probabilities in its feature set during the 8th test dialogue, there was significant improvement (88% vs. 83% accuracy) and when limiting training to just a single prior dialogue the dynamic system is substantially better than the static one (81% vs 59%).

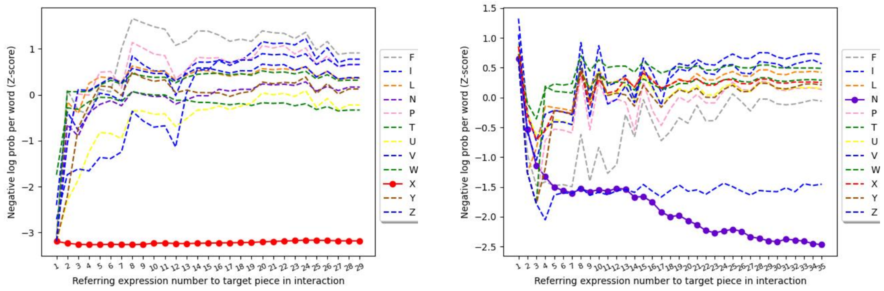

Figure 1: Plots of the moving average of the per-word cross-entropy (per-word negative log probability) of 10 different Pentomino pieces being referred by one conversational pair according to the model for that piece (solid line with solid circular markers), vs that assigned by models for other pieces (dashed lines).

PENTO-CV	**MDC-R**
8 dialogues (8 different pairs - switching roles)	101 dialogues (49 different pairs in fixed roles)
11,000 words per session (mean)	300 words per session (mean)
1,899 referring expressions (only pieces)	7,600 referring expressions (exhaustive)
Reference chain length for pieces: median=19	Reference chain length for block sets: median=2

Table 1: Comparison of size and format of datasets.

3 Minecraft Dialogue Corpus with Reference and comparison to Pento-CV

The Minecraft Dialogue Corpus with Reference (MDC-R) corpus (Madge et al., 2025) annotates the original MDC with reference annotations, as part of the ARCIDUCA project (Poesio et al., 2022).[1] The MDC-R uses a 11×9×11 Cartesian coordinate based Minecraft world, with blocks of 6 different colours (maximum 20 blocks of each colour). Much like Pento-CV, the Architect instructs the Builder to lay blocks into positions, though into a virtual 3D grid world. There are several differences to Pento-CV which we will briefly layout here.

Number of possible referents While in Pento-CV the number of individual piece referents was only 12, and in theory 2^{12} possible subsets of pieces, in MDC-R, the number is far higher: while there are a maximum of 120 coloured blocks that could be used in the game (and 2^{120} possible subsets thereof), Architects also refer to blank spaces, so the referent set could be one of 1089 places (or an enormous 2^{1089} subsets thereof).

Dialogue length and pact length The potential pact length for objects in the two corpora is as shown in Table 1. While the MDC-R has many more references annotated, the length of reference chains is significantly shorter (median=2) as blocks are introduced and used within a single game.

Reference annotations and types While in PENTO-CV referring expression annotations are only made for pieces present in the building area, MDC-R has all references annotated, not only for the blocks present, but for all referents to whole structures, which may not yet have been created in the playing area, with "bridging" references.

4 Conclusion

While there are differences between the two corpora described, the model used for the superficially simpler reference situation in Pento-CV could be adapted for MDC-R. One of the main problems is the massive potential set of referents. The possible referent set could be reduced by filtering on the possible subsets at a given point in the dialogue. Some solutions could involve:

- exploiting the difference between blocks still in storage and those in the game space.
- allowing co-reference to block (set) types rather than precise tokens in fixed positions - e.g. a pact for a line of 8 green blocks.
- using part-whole relations, where the pact involves a hierarchical map from concepts to words ("[the back of [the chair]]"), not just a flat language model, where the volume hierarchies of structures could also be exploited.

While challenging, we remain optimistic that conceptual pact models are useful for complex reference domains using some of the above adaptations.

[1] https://www.arciduca.org/

Acknowledgments

All authors were supported by EPSRC grant EP/W001632/1 ‘ARCIDUCA’. Hough’s work is supported by the EPSRC grant EP/X009343/1 ‘FLUIDITY’.

References

Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, 22(6):1482.

Jonathan Ginzburg. 2012. *The interactive stance: Meaning for conversation*. Oxford University Press.

Julian Hough, Sina Zarrieß, Casey Kennington, David Schlangen, and Massimo Poesio. 2024. Conceptual pacts for reference resolution using small, dynamically constructed language models: A study in puzzle building dialogues. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3689–3699, Torino, Italia. ELRA and ICCL.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 292–301.

Chris Madge, Maris Camilleri, Paloma Carretero Garcia, Mladen Karan, Juexi Shao, Prashant Jayannavar, Julian Hough, Benjamin Roth, and Massimo Poesio. 2025. Mdc-r: The minecraft dialogue corpus with reference. *Preprint*, arXiv:2506.22062.

Massimo Poesio, Richard Bartle, Jon Chamberlain, Julian Hough, Chris Madge, Diego Perez-Llebana, Matthew Purver, and Juntao Yu. 2022. Arciduca: Annotating reference and coreference in dialogue using conversational agents in games. In *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue - Poster Abstracts*, Dublin, Ireland. SEMDIAL.

Massimo Poesio and Hannes Rieser. 2011. An incremental model of anaphora and reference resolution based on resource situations. *Dialogue Discourse*, 2:235–277.

Luc Steels and Paul Vogt. 1997. Grounding adaptive language games in robotic agents. In *Proceedings of the fourth european conference on artificial life*, volume 97. Citeseer.

Alessandro Suglia, Bhathiya Hemanthage, Malvina Nikandrou, George Pantazopoulos, Amit Parekh, Arash Eshghi, Claudio Greco, Ioannis Konstas, Oliver Lemon, and Verena Rieser. 2022. Demonstrating EMMA: Embodied MultiModal agent for language-guided action execution in 3D simulated environments. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 649–653, Edinburgh, UK. Association for Computational Linguistics.

Yanchao Yu, Oliver Lemon, and Arash Eshghi. 2016. Comparing dialogue strategies for learning grounded language from human tutors. In *Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, New Brunswick, NJ. SEMDIAL.

Sina Zarrieß, Julian Hough, Casey Kennington, Ramesh Manuvinakurike, David DeVault, Raquel Fernández, and David Schlangen. 2016. PentoRef: A Corpus of Spoken References in Task-oriented Dialogues. In *10th edition of the Language Resources and Evaluation Conference*, Portorož (Slovenia).