Conceptual Pacts for Reference Resolution using Small, Dynamically Constructed Language Models: A Study in Puzzle Building Dialogues

Julian Hough[1], **Sina Zarrieß**[2], **Casey Kennington**[3]
David Schlangen[4,5] **and Massimo Poesio**[6]
[1]School of Mathematics and Computer Science, Swansea University
[2]Faculty of Linguistics and Literature, Bielefeld University
[3]Department of Computer Science, Boise State University
[4]Department of Linguistics, University of Potsdam
[5]German Research Center for Artificial Intelligence (DFKI), Berlin
[6]School of Electronic Engineering and Computer Science, Queen Mary University of London
[1]julian.hough@swansea.ac.uk

Abstract

Using Brennan and Clark's theory of a Conceptual Pact, that when interlocutors agree on a name for an object, they are forming a temporary agreement on how to conceptualize that object, we present an extension to a simple reference resolver which simulates this process over time with different conversation pairs. In a puzzle construction domain, we model pacts with small language models for each referent which update during the interaction. When features from these pact models are incorporated into a simple bag-of-words reference resolver, the accuracy increases compared to using a standard pre-trained model. The model performs equally to a competitor using the same data but with exhaustive re-training after each prediction, while also being more transparent, faster and less resource-intensive. We also experiment with reducing the number of training interactions, and can still achieve reference resolution accuracies of over 80% in testing from observing a single previous interaction, over 20% higher than a pre-trained baseline. While this is a limited domain, we argue the model could be applicable to larger real-world applications in human and human-robot interaction and is an interpretable and transparent model.

Keywords: reference resolution, small language models, situated dialogue

1. Introduction

The rise of Large Language Models (LLMs) in automatic dialogue processing like the Generative Pre-Trained Transformer (GPT) models (Radford et al., 2018) offer the promise of bigger and better data-driven language models that can be fine-tuned for different computational linguistics tasks with high degrees of success. While these models perform well across a range of tasks, they are large and resource-intensive, and the contributions of different interactions on their models cannot easily be decomposed. Furthermore, while they can be fine-tuned offline, doing so in a time-linear manner during interaction remains challenging. In this paper, we put several language models of a much smaller size and more traditional kind to use in situated reference resolution in a small data domain, to maximise speed, modularity, and human-interpretability.

This paper draws on recent work in situated reference resolution using classical and neural models (Kennington and Schlangen, 2015; Yu et al., 2016; Jayannavar et al., 2020; Suglia et al., 2022; Loáiciga et al., 2022; Poesio et al., 2022; Kiseleva et al., 2022), while trying to overcome the possible shortcomings of these models for interaction, particularly those driven by LLMs, in so far as they rely on training on (standardly) large amounts of static data with no model of dynamic update in reference models *during* interactions.

In human interaction, we observe that in dialogue within populations of speakers of the same official language, different pairs or groups of participants can use very different referring expressions to other groups. Furthermore, these conventions stabilize over time within the group, consistent with experimental evidence that there are conventions which emerge for collaboratively referring to objects (Clark and Wilkes-Gibbs, 1986), and that dynamically constructed sub-languages emerge over interaction time (Healey, 2008; Mills, 2011) when seeing language as constantly evolving and dynamically changing through interaction (Gregoromichelaki et al., 2022). This causes problems for machine learning and statistical models which cannot adapt during the current interaction.

To move beyond static models, here we assume a dynamic, interactive learning scenario where an agent such as a robot receives live feedback from an interlocutor as it guesses which object is being referred to, and can update its own reference model during the conversation, assuming an interactive machine learning set-up (Kulesza et al., 2015). To

Figure 1: The video feed which the Instruction Giver sees during the data collection of PENTO-CV. The Pentomino piece names from left to right are, top row: F, X, Z, P, V; middle row: Y, I, U, T; bottom row: N, L, W

build a cognitively realistic and interpretable model, we use the insight from Brennan and Clark (1996)'s argument that when interlocutors agree on a name for an object, they are forming a temporary agreement on how to conceptualize that object, a *Conceptual Pact*. We model a Conceptual Pact process as it occurs over time between different conversation pairs, using small language models assigned to each referent which are trained from scratch by the agent from the beginning of each interaction with a new participant. We simulate a teaching process where a positive example for a referent becomes available immediately after a reference has been made to it. We use these models to extend a simulation of an interactive reference resolution setting by using existing human-human data.

Our aim here is to provide a method applicable to any reference situation with interactive feedback from a human user, and are guided by several motivating factors for the choice of method. Firstly, we purposefully make use of efficient models which do not require resource-intensive re-training regimes: where two models are equal in terms of accuracy, we assume the less resource-intensive one is preferable, not only for reasons of speed, but for environmental sustainability (Scuri et al., 2022). We also want to use the approach to experiment with reduced data situations, inluding situations where the robot may have only had a single interaction to learn from. Finally, we also prefer human-interpretable models over black box models. In the remainder of this paper we present our domain and data in §2, our Conceptual Pact model for reference in §3, a first experiment on optimizing our model for reference resolution in §4, a second experiment on reducing the amount of training data available in §5, and a final discussion in §6.

Speaker	End time (s)	Referring expression	English translation
B	167.5	das hellblaue L	the light blue L
B	407.2	das zweite L	the second L
B	454.0	das oben rechts liegende L	the L at the top right
B	519.6	das große L	the big L
B	785.7	das L	the L
B	1083.5	das element	the part
B	1101.9	dem blauen	the blue one
B	1233.6	das blaue andere L	the other blue L
B	1283.8	das blaue element	the blue part
A	1545.5	das blaue element	the blue part
A	1626.0	das blaue L	the blue L
A	1635.4	das blaue L	the blue L
A	1646.4	das blaue L	the blue L
A	1661.9	das blaue L	the blue L
A	1853.9	das blaue L	the blue L
A	1922.8	das blaue L	the blue L
A	1970.9	der blaue	the blue one
A	2114.2	das blaue L	the blue L
A	2296.3	das L	the L
A	2297.3	das blaue	the blue one
A	2298.1	das blaue L	the blue L
A	2546.6	das blaue L	the blue L
A	2645.0	das blaue L	the blue L
A	2806.5	das blaue L	the blue L
A	2834.6	das blaue L	the blue L

Figure 2: A pair's evolving referring expressions for the V pentomino piece over time, eventually stabilizing.

(a)

Spkr	End time (s)	Referring expression	English translation
B	326.1	das plus	the plus
B	547.3	das rote plus	the red plus
B	700.6	das plus	the plus
B	725.8	plus	plus
B	743.3	plus	plus
B	1060.7	das plus	the plus
B	1070.0	plus	plus
A	1246.4	das plus	the plus
A	1249.6	das plus	the plus
A	1429.0	das plus	the plus
A	1447.9	das plus	the plus
A	1666.1	das plus	the plus
A	1695.4	das plus	the plus
A	1808.5	plus	plus
A	1811.3	das plus	the plus

(b)

Spkr	End time (s)	Referring expression	English translation
B	141.7	das rote kreuz	the red cross
B	311.5	das rote kreuz	the red cross
B	325.1	das kreuz	the cross
B	338.2	das kreuz	the cross
B	629.3	das kreuz	the cross
B	645.9	das kreuz	the cross
B	768.7	das kreuz	the cross
B	825.7	das kreuz	the cross
B	862.6	dem roten kreuz	the red cross
B	862.6	dem kreuz	the cross
B	1265.3	das kreuz	the cross
A	1488.4	das X	the X
A	1490.0	das kreuz	the cross
A	1539.9	das X	the X
A	1553.4	das kreuz	the cross
A	1565.3	dem X	the X
A	1815.6	das kreuz	the cross
A	1839.2	dem X	the X
A	1945.0	das kreuz	the cross
A	2068.4	das kreuz	the cross
A	2260.6	das kreuz	the cross
A	2271.0	kreuz	cross
A	2283.4	das kreuz	the cross
A	2294.2	das X	the X
A	2385.6	das kreuz	the cross
A	2416.2	dem kreuz	the cross

Figure 3: Two different pacts established for the X piece by two different conversation pairs.

2. Domain and Data

We propose a component of a dynamic Spoken Language Understanding (SLU) model of human participants building puzzles with pieces of different shapes and colours, specifically in the domain of Pentomino puzzles (Golomb, 1996). This domain requires that speakers use ways to refer to puzzle pieces that they may not have a conventionalized name for (Götze et al., 2022). Participants tend to align and form a conceptual pact for each piece for efficient interaction.

Specifically, we use the PENTO-CV corpus data from the PentoRef data collection (Zarrieß et al.,

2016) (Zarrieß et al., 2016) from Github.[1] PENTO-CV is a corpus of situated interactions between 16 German speakers wherein 8 pairs of participants instruct one another via video and audio feed to manually complete a Pentomino puzzle , using the 12 puzzle pieces as shown in Fig. 1. This data is motivated by a long history of work on collaborative referring in interactive settings - see e.g. (Clark and Wilkes-Gibbs, 1986) for Tangram experiments for forcing pact forming to refer to unfamiliar objects and (Anderson et al., 1991) for the interactive Edinburgh map task where participants have to align to each other's reference strategies to complete a task. Furthermore, PENTO-CV has high enough audio, video and transcription and annotation quality for developing automatic methods.

In PENTO-CV both participants have a turn at playing one of two roles: the *instruction giver* (IG) is given a photograph of the final goal configuration of the puzzle pieces and can see the puzzle being constructed by the *instruction follower* (IF). Audio access is full-duplex and bidirectional while only the IG has a video feed showing an overhead view of the area of play including the puzzle pieces and IF's hands as per Fig. 1. Each game is further subdivided into an initial 'selection' phase and a following 'building' phase. In the selection phase, the IF chooses some pieces and presents them to the IG, often using speech and an indicative gesture. The IG enters these objects into a user interface that retrieves an image of a complete puzzle configuration from a database containing the selected pieces in addition to others. After that, in the building phase, the IG directs the IF in creating that target configuration. In this paper, we only use the building phases of the participants' interactions due to inconsistent levels of transcription and annotation in the selection phases. We use transcribed speech data from both dialogue partners which is annotated for references to the piece name according to the Pentomino letter conventions as described in the caption in Fig. 1.

In this paper we only deal with referring expressions to individual pieces, rather than multiple ones. We further filter the data by removing anaphoric references.[2] We preprocess the referring expressions by removal of reparandum words (Shriberg, 1994) in repair disfluencies and of any filled pauses marked up in the DUEL corpus style annotation (Hough et al., 2016).[3]

Once this cleaning process is complete, there are a total of 1899 referring expressions to the 12 pentomino pieces across the 8 pairs of speakers. The median number of referring expressions each pair makes to a piece is 19, with a lower bound of 2 mentions for a piece and the highest at 41. For an example of the referring expressions used for the V piece in the order they occur for an interaction pair, see Fig. 2, and for two different pairs' references to the X piece over time, see Fig. 3.

3. Conceptual Pact Models for Reference Resolution using Small Dynamically Constructed Language Models

We intend to capture two ways conceptual pacts can work in conversation. Firstly, as shown in Fig. 3, different dialogue partners can develop different pacts for naming different objects which have quite different lexical content, but remain consistent throughout their interaction - in this example pair (a) use 'plus' while (b) alternate between 'kreuz' and 'X' for the shape description of the X and they consistently use these conventions. Secondly, as Fig. 2 shows, the convention of naming a piece can stabilize over time in the interaction after initial variation, in this case alighting on 'das blaue L' for the V piece. These observations suggest a model which does not dynamically update with feedback from the current interaction and only draws on previous pacts it has observed in its training data may perform poorly when dealing with a novel pair, particularly in small data situations.

To capture the contribution of local conceptual pacts, we use local updating language models for each piece r, p_r^{pact}, e.g. 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.

While our focus is on the locally constructed models for capturing pacts, the practical challenge is that initially they will be maximally uncertain when encountering the first instance of a referent as they have had no training data for it and may continue to be uncertain in the early mentions of the piece. 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 con-

[1] https://github.com/clp-research/pentoref

[2] This was done procedurally by removing the words in {"es", "das", "er", "da", "der", "ihn", "den", "sie", "die", "damit", "daran", "dem" }

[3] The reparandum is taken to be any material before the repair point + in the in-line annotation and filled pauses annotated such as the 'uh' here are removed: (the + the) {F uh} red cross $\rightarrow$ the red cross

versations 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).

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

The final hyper-parameter of the model we posit is motivated by the fact that while the prior experience model should be drawn on initially, as the agent interacts more it should become more confident in its local pact model in accordance with our observations of stabilizing pacts over time such as in Fig. 2. Technically, this parameter determines what the λ weight should be dynamic, making λ a function of how many times a certain referent has been seen so far, based on a linear degradation of the influence of the prior experience model p_r^{ex}. This weighting begins with complete use of p_r^{ex} and λ=0 for the local model before the first mention of a piece due to it having no training data, then increases it linearly with each mention until the final λ is used after a certain number of occurrences. We define the number of mentions until the stabilization of the pact as an integer variable $stable$ which can be optimized. The λ weighting after a given number of occurrences o_r of reference r is below, replacing the λ in (1), is therefore:

$$\lambda_r(o_r) = \lambda - (\frac{1}{stable} \cdot max(stable - o_r, 0) \cdot \lambda) \quad (2)$$

The final model for each referent r is a joint score given to the words in a referring expression $w_0..w_n$, computed by the weighted interpolation of the relevant language models' probability estimations of $w_0..w_n$: one based on the experience of object r from previous interactions p_r^{ex} and one dynamically built during the current interaction p_r^{pact}. The interpolated weight depends on the λ hyper-parameter and the number of times the referent r has been encountered so far using (2) to give:

$$(1 - \lambda_r(o_r))p_r^{ex}(w_0..w_n) + \lambda_r(o_r)p_r^{pact}(w_0..w_n) \quad (3)$$

In application, after such a score has been obtained for a referring expression $w_0..w_n$ and ground-truth piece g, a simulated teaching episode is carried out using a positive example of g, by incrementing o_g by 1 and updating the local pact model p_g^{pact} with $w_0..w_n$ as a training example.

4. Experiment 1: Conceptual Pacts as Local Language Models blended with Prior Experience Language Models

To test the efficacy of the models we set up a simple reference resolution task with baselines and competitor models, implementing the experiments in Python.[4] We use a simple classifier model throughout for all experiments, the Linear Support Vector Classifier (LSVC) with square hinge loss, and investigate the benefit of including scores from our model described above in the instance data. The primary baseline we compare against is a standard classification set-up which is trained on all available other data, using lexical features without our model's features, with no dynamic updating with the current interaction. We also compare our model against a competitor which exhaustively re-trains the LSCV after each new instance is encountered, again just using the lexical data but without information from our models.

4.1. Obtaining model features

For each candidate referent r, we build two simple n-gram language models (Shannon, 1948) with Lidstone (add-k) smoothing, one trained using the referring expressions to r from other pairs p_r^{ex} and one dynamically built during the interaction for that piece p_r^{pact} which is initialized with empty counts.

For each referring expression $w_0..w_n$ encountered in time-linear order, we compute the per-word negative log probability (*cross-entropy*) of (3) for each possible referent r then compute a relative score for those values with Z-score normalization. After the scores have been obtained, the simulated teaching episode for ground-truth referent g is carried out by incrementing observation count o_g and updating the local pact model p_g^{pact} by adding $w_0..w_n$ into its n-gram counts for subsequent probability estimations. This update is a simple incremental modification to count dictionaries and is extremely computationally efficient.

Fig. 4 shows the moving average of the Z-score normalized cross-entropy values for referring expressions generated by each model for 10 of the 12 pieces for one interaction, though note these plots are shown for a different ground truth referent in each graph. As can be seen, the model for a given ground-truth piece (solid line with solid markers) successfully separates out its cross-entropy values from the other models (dashed lines) over time as its local pact model is built dynamically, becoming more certain about the expressions referring to its piece over time as the others become less certain.

[4]The code for all experiments is available at https://github.com/julianhough/conceptualpacts.

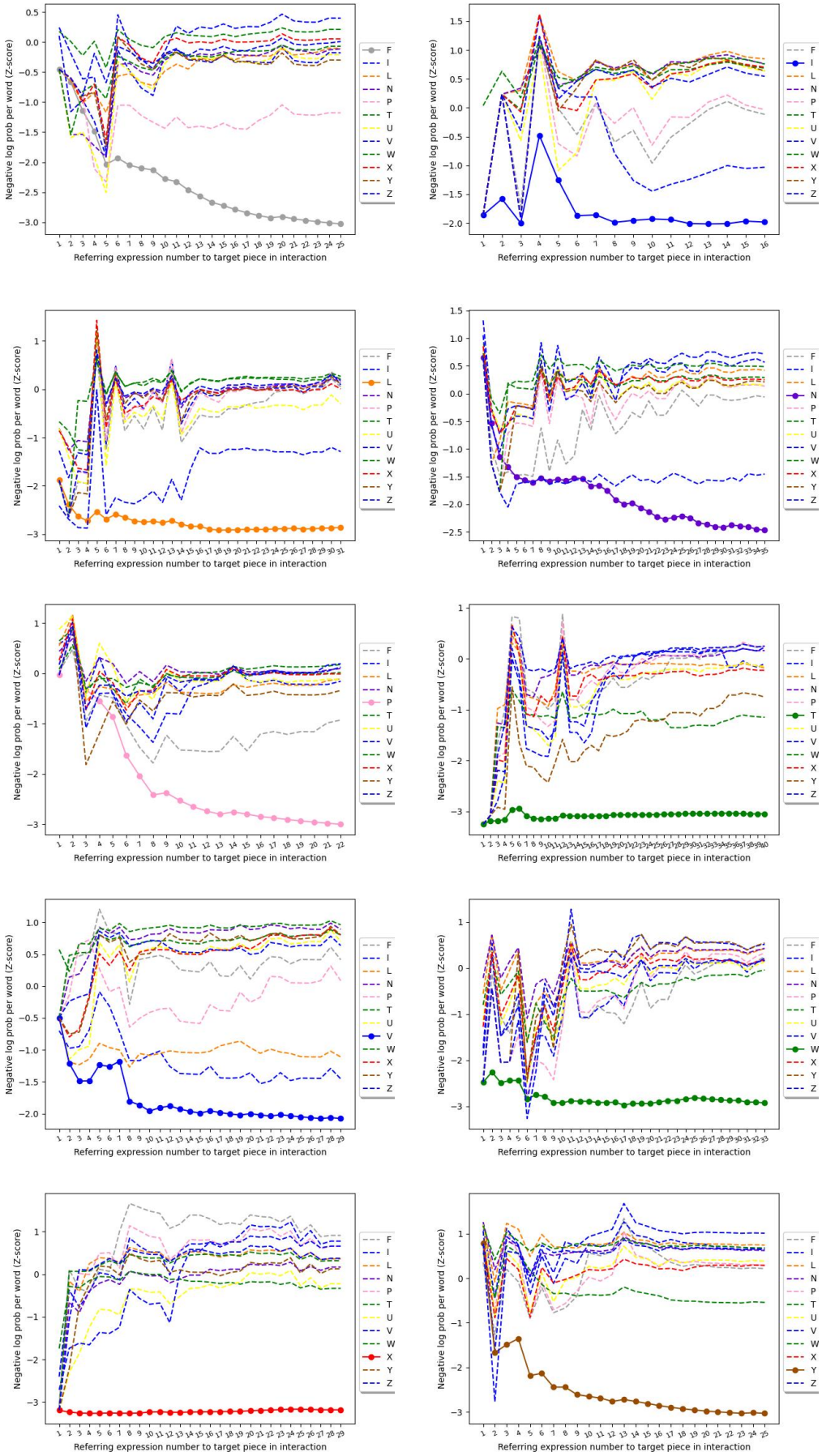

Figure 4: Plots of the moving average of the per-word cross-entropy (per-word negative log probability) of 10 different Pentomino pieces being referred by one conversational pair according to the model for that piece (solid line with solid circular markers), vs that assigned by models for other pieces (dashed lines).

System	X-val Accuracy	Test Accuracy	X-val Time (s)
Pact-ex+Lex: Pact+Experience Model + Lexical	**86.5**	**88.0**	10.73
Static baseline: Lexical Only no update	82.5***	83.3***	10.12
Retrain: Exhaustive re-training Lexical Only	**86.4**	**89.9**	31.47

Table 1: Experiment 1 results. MacNemar test of difference to top row model ***p<=0.001

Some objects have distinct referring expressions early on such as the X (bottom row, left), while others take longer to separate out from the others, such as the N (second row, right).

While in some cases p_r^{ex} models do not always help to provide useful initial values due to the whole dataset being quite small and the pacts on pieces being diverse, it is hoped that overall, the Z-scored cross-entropies for the ground-truth piece's model versus the rest will be sufficiently lower than that produced by other models to provide useful information for the reference resolution classifier.

4.2. Experimental set-up and Evaluation

We evaluate in two settings as follows: 1) *cross-validation*: after selecting a test conversation pair that is closest in size to the mean number of overall referring expressions, we exclude that from cross-validation experiments to give 7 interactions. We train on 6 folds and evaluate on the 7th iteratively and obtain the mean accuracy of the predictions of all folds. To obtain the p_r^{ex} values for the training data for each fold iteration, we train p_r^{ex} models on 5 folds and apply them on the 6th, again in a cross-fold way, to ensure we are simulating the test scenario where the test data will never have been used before, even for language modeling. For each example we combine the Z-score normalized scores of (3) for each candidate referent with bag-of-words length-normalized counts for the lexical features. 2) *testing*: we use the test pair removed from the cross-validation experiments to test on, using all the data from the 7 other pairs for training. While in theory, the training data could use p_r^{ex} models with 6 of the pairs, this is kept to 5 to be consistent with the size of the cross-validation p_r^{ex} models so as not to alter the range of expected probability values the models assign during testing.

We employ a standard $accuracy$ metric for reference resolution in all experiments and significance test between the models' predictions using MacNemar's test on their predictions and the ground-truth labels as applied to machine learning models by Dietterich (1998). In the cross-validation case to compute this we concatenate the predictions and labels across all folds into single lists.

4.3. Optimization

We use the cross-validation setting to exhaustively optimize the 6 hyper-parameters for our model: n_{ex} and n_{pact}, the order of n-grams for the experience and local pact language models respectively; k_{ex} and k_{pact}, their two add-k smoothing parameters; λ for the pact model weighting, and the count $stable$ at which point λ is set to its final value as per (2). The optimal parameter values for our 7-fold cross-validation set-up were found to be the following:

n_{ex}	k_{ex}	n_{pact}	k_{pact}	λ	$stable$
1	0.1	1	0.9	0.1	6

The optimization shows that the model needs to rely on the experience model more than the final 1-λ weight up until the sixth mention of a piece, where we could see this as the typical stabilisation point of the pacts for each piece. The fact the λ weighting on the local p_r^{pact} models is only 0.1 is more due to their size being roughly a fifth of the p_r^{ex} models rather than them not being very useful - with less prior experience data and smaller p_r^{ex} models we expect this to rise towards or above 0.5, as will be shown in Experiment 2.

4.4. Results

When using the optimized hyper-parameters, the results of the cross-validation and testing evaluations are as per Table 1. Our model achieves reference resolution accuracies of 86.5% and 88% in the two settings. The application of the MacNemar test on the error distributions shows the optimized model significantly outperforms the static baseline in both cross-validation and in final testing. The competitor model which exhaustively re-trains after each reference with lexical data does not perform statistically significantly differently to our model. However, the exhaustive retraining model has a far increased training time and resource use - Table 1 shows the full running time of our model training and testing on the 7-fold cross-validation, running on an Apple M2 Pro chip with a 16GB memory, being around 3 times faster than the exhaustive retraining model, and with very little extra time needed over that of the static baseline. The very low computational resource intensity of our model is due to the fact that dynamically adding counts to existing n-grams or adding novel n-grams to the object language models consists of simply updating the relevant Python dictionaries, with average case operation times of constant time *O(1)* and worst case times of linear time *O(n)*, where *n* is the number of n-gram types in the model.

5. Experiment 2: The effect of limiting prior experience

While in Experiment 1 in cross-validation and testing, 6 and 7 interactions were available for training, we imagine a scenario where a robot has little to no previous experience. To simulate this, we look at the effect of reducing the training data by one more interaction iteratively, all the way down to allowing observation of a single interaction as training data.

5.1. Optimization

Using cross-validation, we optimize the parameters separately for each number of interactions allowed for training (*#experience*). For each setting, using a Python random seed, we pseudo-randomly choose the data of the appropriate #experience size. The exceptional case is when #experience=1, where the standard way of applying language models does not hold as the experience models p_r^{ex} have no data to build from to obtain the training data: to alleviate this we simulate language model training data by using the same interaction to train p_r^{ex} as for p_r^{pact}, which is a set-up not available at test time as it would require predicting the whole interaction before it happened. While this accommodation is not ideal, it yielded quite promising results in testing. The optimal parameters found in cross-validation were as below:

	#experience				
	1	2	3	4	5
n_{ex}	1	1	1	1	1
k_{ex}	0.1	1.0	0.95	0.75	0.7
n_{pact}	1	1	1	1	1
k_{pact}	0.1	0.1	0.1	0.1	0.1
λ	0.1	0.8	0.65	0.4	0.3
$stable$	7	6	8	5	7

As can be seen, with the exception of #experience=1, the λ weighting starts high and gradually diminishes with more training data pairs, as the models trust the p_r^{ex}s from other pairs more as more data is made available. The p_r^{ex} is still used more than the final weighting in all models in the opening mentions of the pieces, with $stable$ being between 5-8 mentions.

5.2. Results

The reference resolution results from Experiment 2 are shown in Table 2, against the same competitor models as in Experiment 1 with their training sets reduced to the same data available to our models and the comparison of our model's performance against the static baseline on the test set is shown in Figure 5. In cross-validation, the model outperforms the static baseline for all #experience, reaching an accuracy of 78.2% with only 1 pair to train on,

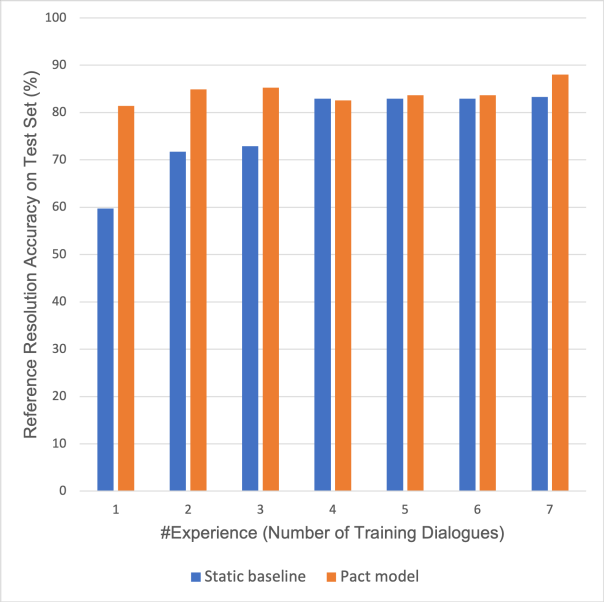

Figure 5: Reference resolution accuracy showing faster convergence of the Pact Model with limited experience compared to the static baseline.

exhibiting a steady improvement as #experience increases, with the exception of 5. The model performs the same as the exhaustive re-training model for #experience={3-6}, however it is outperformed when training on just 1 and 2 pairs: this may be due to their uniquely small p_r^{ex} models and the unique set-up for #experience=1.

In testing, the model is outperformed by the exhaustive re-training model in all set-ups. As shown in Figure 5, our model outperforms the static baseline for #experience={1-3} but not for #experience={4-6}, however it does again for the full version in Experiment 1 (i.e. #experience=7). We discovered in development that the size and selection of the language model training interactions can greatly affect results, so particular fold pairings can make a difference and further experimentation is needed in future. It is still promising that in testing with one single dialogue to learn from, the model can exceed the baseline static lexical model by a large margin (81.4% vs 59.7%), making good use of the dynamically available data despite the challenges of the set-up, and as per the run times in Table 1 at a very marginal extra computational cost.

6. Discussion

The two experiments we present using features generated by our conceptual pact model show promise for their use in reference resolution in an interactive learning set-up. While in Experiment 1, the method using the full dataset was shown to be effective compared to a non-interactive lexical

System	X-val Accuracy (#experience)					Test Accuracy (#experience)					
	1	2	3	4	5	1	2	3	4	5	6
Pact-ex+Lex	78.2	82.5	**83.6**	**85.2**	**84.5**	81.4	84.9	85.3	82.6	83.7	83.7
Static baseline	66.0***	72.0***	72.9***	79.6***	79.2***	59.7***	71.7***	72.9***	82.9	82.9	82.9
Retrain	**81.4*** **	**83.8** **	**83.7**	**85.1**	**85.4**	**85.7***	**88.4***	**88.8***	**89.9*** **	**90.3*** **	**90.3*** **

Table 2: Experiment 2 results. MacNemar test of difference to top row model ***p<=0.001,**p<=0.01,*p<=0.05

feature baseline, in Experiment 2 where training data was reduced, the picture is more mixed, yet in cross-validation it still outperforms the baseline for all amounts of reduced training data tested. While it does not outperform the exhaustive re-training competitor, as shown by the run times, it is considerably less resource-intensive to update counts of simple n-gram models than retraining classifiers from scratch. Furthermore, complex discriminative classifiers like the LSVC using lexical and other features do not allow a clear decomposition of the contributions of the different parts of their training data, so our model also brings the benefits of modularity and human interpretability for analysis.

Our work follows work in embodied reference resolution (Kennington and Schlangen, 2015; Yu et al., 2016; Suglia et al., 2022) and live grounded language acquisition in the spirit of (Steels and Vogt, 1997). We maintain principles of small data use for a more realistic language acquisition models than is currently employed for large language model-based systems. In future, we plan for a more realistic evaluation in an in-robot system adopting the instruction follower role, such as PentoRob (Hough and Schlangen, 2016), where the system would consume raw audio data and use automatic speech recognition hypotheses and use visual features similar to PENTO-CV's provided features or those extracted by another system.

While modelling a zero, or near-to-zero, initial linguistic knowledge approach for the pact reference models, technically, neural models of any kind could substitute the n-gram models for the prior experience language models here, including LLMs like the GPT models or LLaMA (Touvron et al., 2023). As described here, they could be used initially, likely with appropriate prompt engineering for reference resolution in this small domain, with their influence attenuated over time as the local pact is formed. However, our intention was to show the strength of the framework for language models in general, with promising results for classical, efficient, and interpretable n-gram models being a strong starting point with scant use of compute power and maintaining complete transparency of the training data. We propose that the components of the model, such as the weight of the local pact model's contribution λ and stabilization point $stable$ can be experimented with in a range of interactive reference scenarios, using different language models, to test the difference between different domains and participant pairs.

In terms of how well this model could generalize beyond the micro-domain of Pentomino puzzle building, we claim this is a proof-of-concept for a much more general model. We argue for a multi-domain, rather than domain-general, approach to dialogue systems, where individual pacts can be established for different situations either starting with existing knowledge or starting anew for new sound-meaning pairs, conversational genres (Ginzburg, 2012) or, in Wittgenstinian terms, different language games (Wittgenstein, 1953). We try to model language as a completely dynamic process in a state of constant change, in the spirit articulated by interactivism (Gregoromichelaki et al., 2022). If models only rely on static knowledge, even if using very large amounts of it, they will ultimately be limited.

7. Conclusion

We have presented a model of Brennan and Clark's conceptual pacts in spoken interaction situations where agents refer to objects in a puzzle-building task. We simulate the pact-building process with small local language models for each referent which are updated through a simulation of interactive learning. The models capture how uncertainty in determining which referent is being referred to decreases over time as a pact stabilises and also how different interaction pairs build different models between them. We show how our model can combine the local language models from the current interaction with those from prior experience in an optimal way and their features can help a reference resolution classifier outperform a lexical model without them. When reducing the amount of prior training data, the model is still robust, performing well after just a single interaction has been observed. While it does not outperform an exhaustive retraining baseline, it exhibits parity to it in a number of situations, and its interpretability, modularity, and far lower resource intensity make this an interesting model to explore in more complex domains and for fully interactive systems.

Limitations and Ethical Considerations

The study has several limitations: Firstly, our experiments were limited to German speakers and in a fairly narrow reference domain, so cross-lingual and cross-domain claims must be limited from this study alone. Secondly, we did not carry out full, end-to-end reference resolution from the transcripts as we only used the ground-truth annotated referring expressions, simulating perfect mention detection. Thirdly, the fact certain kinds of anaphoric reference and referring expressions for multiple objects were filtered out means not all references to objects were used from the transcription data due to inconsistent annotation protocols, where these could be the most challenging cases. Fourth, while it is a simple domain, for a more solid empirical basis and comparison to human-human agreement, a sample of the reference annotations should be checked using inter-annotator agreement measures as described by e.g. (Artstein and Poesio, 2008).

In terms of ethical considerations, all participants from the PENTO-CV data gave explicit permission for their data to be used for research purposes, and were paid at above the German minimum wage rate for their participation. The publicly downloadable transcribed data does not include audio or video data which could identify the participants. We consider our use of relatively low resource models compared to modern large language models to be part of a more sustainable approach to this kind of theoretically informed computational linguistics.

Acknowledgements

Hough is supported by the UKRI Engineering and Physical Sciences Research Council (EPSRC) grant EP/X009343/1 'FLUIDITY' and Poesio and Hough are supported by EPSRC grant EP/W001632/1 'ARCIDUCA'. We thank the three LREC-COLING reviewers for their thorough reviews and for discussions with Arash Eshghi on the content.

Bibliographical References

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

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.

Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.

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

Solomon W Golomb. 1996. *Polyominoes: puzzles, patterns, problems, and packings*, volume 16. Princeton University Press.

Jana Götze, Karla Friedrichs, and David Schlangen. 2022. Interactive and cooperative delivery of referring expressions: A comparison of three algorithms. In *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Dublin, Ireland. SEMDIAL.

Eleni Gregoromichelaki, Arash Eshghi, Christine Howes, Gregory J Mills, Ruth Kempson, Julian Hough, PG Healey, Matthew Purver, et al. 2022. Language and cognition as distributed process interactions. In *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue*, pages 160–171.

Patrick Healey. 2008. Interactive misalignment: The role of repair in the development of group sub-languages. *Language in Flux. College Publications*, 212.

Julian Hough and David Schlangen. 2016. Investigating fluidity for human-robot interaction with real-time, real-world grounding strategies. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 288–298.

Julian Hough, Ye Tian, Laura de Ruiter, Simon Betz, Spyros Kousidis, David Schlangen, and Jonathan Ginzburg. 2016. DUEL: A Multi-lingual Multimodal Dialogue Corpus for Disfluency, Exclamations and Laughter. In *10th edition of the Language Resources and Evaluation Conference*, Portorož (Slovenia).

Prashant Jayannavar, Anjali Narayan-Chen, and Julia Hockenmaier. 2020. Learning to execute

instructions in a Minecraft dialogue. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2589–2602, Online. Association for Computational Linguistics.

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.

Julia Kiseleva, Ziming Li, Mohammad Aliannejadi, Shrestha Mohanty, Maartje ter Hoeve, Mikhail Burtsev, Alexey Skrynnik, Artem Zholus, Aleksandr Panov, Kavya Srinet, et al. 2022. Interactive grounded language understanding in a collaborative environment: Iglu 2021. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 146–161. PMLR.

Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 126–137.

Sharid Loáiciga, Anne Beyer, and David Schlangen. 2022. New or old? exploring how pre-trained language models represent discourse entities. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 875–886, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Gregory Mills. 2011. The emergence of procedural conventions in dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Massimo Poesio, Richard Bartle, Jon Chamberlain, Julian Hough, Chris Madge, Diego Perez-Llebana, Matt 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.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Sabrina Scuri, Marta Ferreira, Nuno Jardim Nunes, Valentina Nisi, and Cathy Mulligan. 2022. Hitting the triple bottom line: Widening the hci approach to sustainability. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Elizabeth Shriberg. 1994. Preliminaries to a theory of speech disfluencies. *Doctoral dissertation, University of California at Berkeley*.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ludwig Wittgenstein. 1953. Philosophical investigations, mcmillan. *New York*.

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).

Language Resource References

Zarrieß, Sina and Hough, Julian and Kennington, Casey and Manuvinakurike, Ramesh and

DeVault, David and Fernández, Raquel and Schlangen, David. 2016. *PentoRef: A Corpus of Spoken References in Task-oriented Dialogues*. ELRA, 1.0. PID https://github.com/clp-research/pentoref.