Enhancing reference resolution in dialogue using participant feedback

Todd Shore, Gabriel Skantze

KTH Speech, Music and Hearing, Stockholm, Sweden

`tcshore@kth.se, gabriel@speech.kth.se`

Abstract

Expressions used to refer to entities in a common environment do not originate solely from one participant in a dialogue but are formed collaboratively. It is possible to train a model for resolving these referring expressions (REs) in a static manner using an appropriate corpus, but, due to the collaborative nature of their formation, REs are highly dependent not only on attributes of the referent in question (e.g. color, shape) but also on the dialogue participants themselves. As a proof of concept, we improved the accuracy of a words-as-classifiers logistic regression model by incorporating knowledge about accepting/rejecting REs proposed from other participants.

Index Terms: dialogue, situated, reference resolution, alignment, dialogue feedback

1. Introduction

Reference resolution in dialogue and the incorporation of multimodal knowledge into reference resolution is a popular research topic: dialogue is typically **situated** in a given context, whereby the dialogue participants share an environment. In the case where the participants are interacting in a common physical environment, features denoting physical properties of possible referents can be used for conditioning a model of reference resolution [1, 2, 3]. On the other hand, knowledge about the discourse itself can also be incorporated into training, such as encoding features which represent e.g. if the previous referring expression (RE) resolved to a given entity or the amount of time since a given entity was mentioned [4].

However, understanding a given utterance in a dialogue is not only dependent on the extra-linguistic context of the dialogue but also on the dialogue history itself: for example, REs are formed in a complex negotiation of **conceptual pacts** [5, 6], whereby the REs used by each dialogue participant eventually converge to a common pattern throughout the course of the dialogue. This phenomenon is called **alignment** in communication between multiple agents [7, 8]. The act of referring to an entity in a common environment is not relegated to individual utterances but spans multiple utterances from multiple users in a referential "sub-dialogue", collaboratively building appropriate REs to the point of **mutual acceptance** [5].

Table 1: *A dialogue participant describing a piece on a shared game board to another participant.*

Player	Time	Utterance
1	5:30	*the arrow I think you called it*
2	5:31	*the bottom left hand corner?*
1	5:32	*no in the middle*
2	5:34	*oh the middle*
1	5:36	*perfect*

The effects of these conceptual pacts in combination with dialogue grounding allow dialogue participants to refer to en-

tities for which they do not have any "canonical" term: Table 1 shows an example of a dyadic interaction in a collaborative online game. The participants discussed which piece should be moved on a board they both see but can only communicate via voice. The **initiated** reference *the arrow* is inadequate for resolution and is rejected by the other participant, proposing the **expansion** [*in*] *the bottom left hand corner*, which is subsequently **rejected** by the initiator and **replaced** with *in the middle* [5]. It can be said that the two participants iteratively propose potential referring traits represented by language (e.g. *arrow*, *bottom*, *middle*), which are then accepted or rejected as appropriate REs for the referent entity. What is notable, however, is that neither participant uses *all* possible referring language at once: In the dialogue in Table 1, only one semantically-rich token was uttered by both participants even though the sub-dialogue in question was ultimately successful.

It would be useful to model this mutual acceptance so that language used by both participants can be used for resolving REs. However, simply incorporating all utterances from all dialogue participants is a relatively naive approach, since not all referring language is (either explicitly or implicitly) accepted as valid.

In this paper, we show that this knowledge can be used to improve the training of a **words-as-classifiers (WaC)** model for classifying words as valid referring language for a particular set of physical features (e.g. a referent's color, shape, position) [1].

Previous work did not include knowledge of dialogic interaction between participants. We include the language from the participant initiating an RE in question as well as the language from other participants based on how strongly the terms observed in the others' language are accepted (or rejected) by the initiator.

2. Background

Previous studies used physical knowledge and dialogic information to improve reference resolution. Most relevantly in the physical domain. [1, 2, 3] used features of manipulable objects in combination with knowledge about the user's manipulations. They trained a WaC model representing the probability of a word (token) t being (part of) a valid RE for a referent r using a corpus of task-oriented dialogues [1, 9]:

$$p_t(r) \triangleq \sigma(w^T r + b) \qquad (1)$$

where w is a weight vector learned for the vector of physical features r representing a possible referent and σ is the logistic function.

Similarly, [10] classified sets of objects being referred to in a physical environment by training a joint model for both word meaning and visual perception. In the domain of dialogic knowledge, [4] defined a ranking SVM for classifying tangram pieces being manipulated by participants in a simulation. They utilized features representing participants' current

and past actions in the simulation as well as features which represent discourse-specific knowledge, such as if a particular referent r was referred to by the most recent RE and if a given referent r was referred to 10 seconds ago or less. Likewise, [11] incorporate anaphora and deixis into a unified model of reference resolution.

The studies involving physical knowledge and advanced dialogic knowledge were successful in improving reference resolution. However, in none of these works was the relationship between dialogue participants' dialogue acts encoded explicitly.

We create a similar experiment which incorporates dialogic knowledge to explore the feasibility of doing so in a well-established paradigm — namely, task-oriented dialogues wherein one participant has the role of **instructor** and one of **manipulator** [5, 7, 12, 13]. This "toy-domain" setup enables to control the environment, with less noise and fewer confounds.

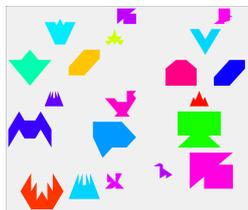## 3. Method

3.1. Data collection

Figure 1: *The game board as seen by both players during a game round.*

In order to test the feasibility of modeling RE collaboration, it is necessary to collect data of dialogues in which the participants create REs to refer to entities for which there are no fixed, "canonical" terms. The corpus design is inspired by the PentoRef corpus [9], but this corpus comprises dialogic instead of monologic language.

In order to observe the formation of these REs, participants played a collaborative online board game in pairs where, in each **round**, one player instructs the other to select a certain piece to move (see Figure 1). The board is the same on both players' screen with the exception that the instructor sees a piece randomly highlighted by the game, which they are then to describe to the manipulator sufficiently enough that they can select the relevant piece on their screen (which is not highlighted on their screen): If the piece is selected correctly, the players gain one point and proceed to the next round, where the roles are switched and the previously-selected piece moves to a random place on the board. However, if the wrong piece is selected, they lose two points and are required to try again. In each game, there were 20 pieces in total, all of which could be selected for movement at random by the program. In addition to the players' speech being recorded and transcribed, the state of the game at the time of each utterance was recorded — including features representing each piece (i.e. possible referent) on the board at any time. This is similar to work done by [1].

Each session lasts approximately fifteen minutes, which is usually the time it takes for players to establish REs and become more confident with their usage until there is nearly no "discourse" towards dialogue end: One player simply uses an RE and the other clicks on the correct piece with little communication. This setup encourages a freer dialogue between par-

ticipants than, e.g., the setup used by [1, 2, 3] which was largely one-sided, limited to the instructor referring to a particular piece and the selector saying little. However, our recorded language is free enough to include cases where a referring expression is negated (e.g. *no, not the blue one*). Therefore, for each game round, there are a number of individual **utterances** both by the instructor and manipulator. Recorded speech was transcribed and segmented manually into atomic units of dialogue-relevant meaning (e.g. [*it's the blue one?*] [*yeah good job*]), each of which being one utterance.

Pieces are procedurally generated for each game session, with piece color, size, shape and location chosen in a pseudo-random manner. The possible shapes were hand-chosen to be a roughly-even distribution of "familiar" shapes (recognizable as e.g. a bird or face) and abstract shapes. Procedural generation of a virtual environment allows extra-linguistic features to be controlled (see Table 2).

Table 2: *Referent features*

Set	Values
POSITIONX, POSITIONY	$\{x \in \mathbb{Q} \mid 0 \leq x \leq 1\}$
HUE	$\{x \in \mathbb{Q} \mid 0 \leq x \leq 1\}$
RED, GREEN, BLUE	$\{x \in \mathbb{N} \mid 0 \leq x \leq 255\}$
EDGECOUNT	$\{x \in \mathbb{N} \mid 6 \leq x \leq 16\}$, manually annotated
SIZE	$\{x \in \mathbb{N} \mid 0 \leq x \leq 2\}$
SHAPE	$\{x \in \mathbb{N} \mid 0 \leq x \leq 16\}$, one for each unique shape, e.g. 4 for the shape in Table 1

The sequence in which pieces are chosen to be moved are randomized across game sessions. All participants had English either as a native language or as a common language used in a professional context, and all had normal or corrected-to-normal vision.

3.2. Modeling

A WaC multi-class logistic regression model was trained similarly to that of [1, 2, 3], where, for each token in each utterance, the probability of the token being (part of) a valid RE is calculated. Using this model, the probability of a particular utterance comprising a sequence of tokens $u \triangleq \langle t_1 \dots t_n \rangle$ which refers to referent r can be computed as the normalized score of each word classifier $p_t(r)$:

$$p(u, r) \triangleq \frac{\sum_{t \in u} p_t(r)}{|u|} \qquad (2)$$

In order to account for unseen word classes, all word classes with fewer than 3 observations in the training data were merged into a single "out-of-vocabulary" class similarly to as done by [1]. Different preprocessing methods were tested, e.g. removing disfluencies, deduplicating tokens and then extracting NPs minus any child locational PPs[1] and tokenizing them (Adv)[2] as opposed to simply tokenizing all content (Basic) and removing stops (Stop) and fillers (Fill) (cf. Table 3).

Three different training methods were tested in total, all of which are different ways of selecting which individual utterances should be used as sources of language to use for training: During training, all tokens from the utterance(s) selected for usage during training are combined into a single bag of words

---

[1]Locational PPs to remove were identified by matching their head to a list of locational prepositions such as *above*, *beside*, etc.

[2]Parsed with Stanford CoreNLP's [14] greedy shift-reduce parser ("englishSR") [15] and bidirectional-dependency POS tagger version 3.7.0 ("english-bidirectional-distsim") [16].

$\{(t_1, w_1) \dots (t_n, w_n)\}$, each token weighted by its count observed in the set of utterances used. The logistic model for each token t in this bag is then provided a number of instances representing referent feature vectors r, the selection of which are explained below. First, two methods not using any dialogic knowledge were tested as a baseline:

- A "one positive, one random negative" (`OneNeg`) method as used by [1], where, for each frame in a dialogue (in our case, for each round in the game), one random negative example is trained from the feature vector representing one entity which is not selected in addition to training a positive example for the selected entity $\hat{r}$. All of the language used by the instructor is processed according to a combination of the aforementioned tokenization and filtering methods and every result token is then used for the one positive and one negative example generated. Manipulator language is disregarded both for training and testing.

- A "one positive, $n - 1$ negative" (`AllNeg`) method, where, for each round in the game, one negative example is trained from the feature vector representing *each* piece which is not selected $R \setminus \hat{r}$ in addition to the one positive example $\hat{r}$ (as stated above, $|R| = 20$ in our experiment). Just as with `OneNeg`, all of the instructor language is used for training with these 20 examples, and manipulator language is disregarded both for training and testing.

Lastly, we tested a "dialogically-aware" (`Dialogic`) method, where, for each utterance from the instructor, the `AllNeg` method is employed for training, and the utterances from the manipulator preceding each instructor utterance are evaluated as either being valid REs or not by computing an approximate "acceptance score" $a : U \mapsto \mathbb{R}$ for the given instructor utterance $u \in U$.

- If the instructor utterance is evaluated to have an acceptance value of $a(u) > 0$,[3] each preceding manipulator utterance is treated the same way that the instructor utterance is (training "one positive, $n - 1$ negative" examples). The rationale for this is, as explained in section 1, that a positive utterance like *yeah the blue bird that one* often serves as a confirmation of clarification requests from the manipulator, which often look like e.g. *do you mean the blue one*. Thus, the previous language most likely contained valid REs for the selected entity being mentioned by the instructor.

- If $a(u) < 0$, the preceding manipulator utterances are used as negative examples of the selected entity only $\hat{r}$, i.e. for each token in each utterance, one negative instance is created from the feature vector of $\hat{r}$. Similarly to above, a negative acceptance value likely means that the previous language from the manipulator is expressly *not* valid language for an RE referring to the selected entity.

The acceptance score was computed by matching the token sequence representing a given utterance to a curated set of patterns observed in the corpus to denote either acceptance or rejection, e.g. *yeah*, *that's it* and *that's right* or *no*, *that's not* and *oops*: The acceptance score is the sum of all patterns found in the utterance where acceptance patterns equal 1 and rejection equal -1.

---

[3]$a(u) > 0$ denotes the instructor's acceptance of referring language proposed by the manipulator and $a(u) < 0$ denotes rejection thereof.

3.3. Evaluation

13-fold cross-validation was run on 13 different game sessions, treating each session as its own set of dependent observations: As stated above, the REs used between two particular dialogue participants in a given dialogue are highly dependent on the participants themselves. For this reason, it is useful to test the effectiveness of partitioning the data into individual game sessions. The measurement is the rank of the gold-standard referent (in this case, the selected piece $\hat{r}$ in an n-best list of possible referents based on their classification score as defined in equation 2.

4. Results

The results of doing 13-fold cross validation using various methods of pre-processing are shown in Table 3 below — Since each game has the same amount of possible referents ($|R| = 20$), mean rank is an accessible metric for evaluating performance.

Table 3: *Comparison of results for cross-validation parameter combinations.*

Tokenization	Token filter	Training	M(rank)	SD(rank)
Basic	None	OneNeg	7.020	5.440
Basic	None	AllNeg	5.198	5.008
Basic	None	Dialogic	5.153	4.876
Basic	Stop, Fill	OneNeg	6.243	5.284
Adv	None	OneNeg	6.351	5.298
Adv	Stop, Fill	OneNeg	5.737	5.112
Adv	Stop, Fill	AllNeg	4.596	4.715
Adv	Stop, Fill	Dialogic	4.462	4.653

Using advanced tokenization `Adv` (described in section 3.2) as well as removing semantically-poor elements such as stopwords `Stop` and fillers `Fill` for processing utterances (see section 3.2) yields some improvement, but not as much as either the improved non-dialogic training method `AllNeg` or the dialogic method `Dialogic`.
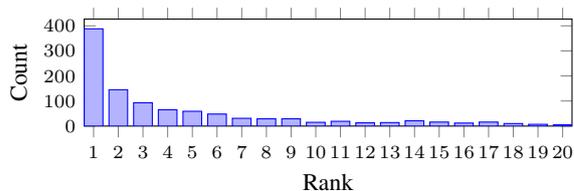
Figure 2: *Count of the referred entity at a given rank. Cross-validation parameters used: Extracting NPs without any child locational PPs (*`Adv`*), stopword and filler filters (*`Stop`, `Fill`*), dialogic training (*`Dialogic`*).*

Although the mean rank of 4.462 achieved for using the best pre-processing methods (Adv, `Stop`, `Fill`) in combination with dialogic knowledge-driven training (`Dialogic`) is relatively high as presented here (out of 20 ranked referents $|R| = 20$), the distribution does not follow a typical normal distribution: The rank with the highest rate of observance is in fact 1 (see Figure 2), occurring 388 times out of 1,035 tested sub-dialogues representing a game round, with a long tail of worse rankings.

The data was analyzed using a linear mixed model with TRAINING, TOKENCOUNT and SESSIONORDER as fixed effects and dyad ID (the pair of participants in a given session) as
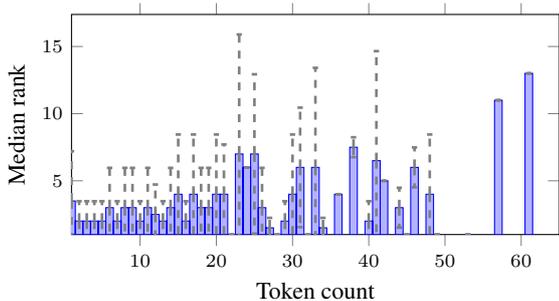
Figure 3: *Median ranks by the amount of tokens a round contains (limit 65). Error bars represent median absolute deviation (MAD).*

a random effect[4]: TRAINING represents one of the three training methods used as described in section 3.2, TOKENCOUNT represents the amount of spoken interaction (measured in tokens) overall in a given round regarding the referent which is to be manipulated in that round and SESSIONORDER represents the order of a given round in a single session, e.g. *2* denotes the second round in a particular game.

There is a significant interaction of TOKENCOUNT with training methods (TRAINING) in their effect on rank as shown by a likelihood-ratio test ($\chi^2 = 8.8$, $p < .05$ when comparing the additive and interaction model). This means that if the given round incorporates a large amount of language from the manipulator, which is only used by the training method Dialogic, rank reduces (i.e. improves) when using Dialogic compared to when using the better non-dialogic method AllNeg. Dialogic showed significantly lower ranks ($B = -.07$, $t = -2.3$, $p < .05$); The rank when using the naive "one positive, one negative" method OneNeg was significantly worse than when using AllNeg ($B = 1.23$, $t = 5.2$, $p < .001$).

In the data, a high token count often signifies a large amount of inter-participant "chatter" introducing a lot of noise into the model for a given referent, e.g. *uh somewhere in the middle you see a a red piece and to the bottom right of that piece*, meaning that for such utterances the rank is overall worse: ($B = 0.7$, $t = 2.7$, $p < .001$). However, when using dialogic knowledge in training, the token count in fact has a negative effect on rank, meaning that more spoken interaction improves modeling with Dialogic.

Interestingly, SESSIONORDER has a negative effect on rank, i.e. rank becomes better for a game as it goes on ($B = -.009$, $t = -3.9$, $p < .001$). This suggests that in fact the participants themselves converge to a similar model of referring language that the WaC classifier has learned, which shows a process of alignment between the two dialogue participants [8].

However, this model explains only 10% of the variance (conditional pseudo-R^2) and the effect sizes for the relevant effects are very small.

5. Discussion

The largest improvement in classifying REs in our corpus came from utilizing a very simple method for establishing a dialogic relation between an instructor's utterances and the manipulator's: If the instructor's utterances have a positive acceptance

---

score, it is likely that they are either explicitly or implicitly showing acceptance of the manipulator's choice of referring language. Likewise, a negative statement shows a degree of unacceptability of what the manipulator proposed.

This method, using simple pattern-matching, showed more improvement than an alternative, more sophisticated training method which nevertheless does not include dialogic knowledge (the "one positive, $n - 1$ negative model"). This experiment has shown that incorporating dialogic knowledge into static training is possible in a tractable manner, without utilizing any higher-order models.

5.1. Future work

In the near future, we hope not only to refine our methods for evaluating mutual acceptance between dialogue participants, but also to model the forming of conceptual pacts between them, as mentioned in section 1. but also to incorporate this knowledge to adapt RE classification models to a particular dialogue in a dynamic fashion, allowing this knowledge to improve classification for cases which have not yet been observed: For example, only one pair of participants out of the 13 recorded used the word *asteroid* to refer to a particular piece. With a statically trained model, the probability of resolving the RE language *asteroid* to r_{11} should be relatively low, as this usage was never seen in the data before that particular game session. However, during the course of the game dialogue, a conceptual pact was formed entailing that the probability of the RE *pink asteroid* referring to r_{11} is actually extremely high: For example, later in the dialogue, the other participant referred to the same entity again with the language *it's the pink asteroid*. At that time in the dialogue, the classifier should have been able to learn that such conceptual pact was formed and adjust the probability of $p_{asteroid}(r_{11})$ accordingly.

Finally, both modeling mutual acceptability in RE collaboration and conceptual pacts should also not only improve training but also to allow a pre-trained system to be adapted dynamically to as-yet-unseen data; In the future, we also intend to thus incorporate such knowledge not only for conditioning models but also for adapting existing ones and/or to do dynamic re-scoring based on this knowledge, similarly to as done by [20].

6. Acknowledgements

7. References

[1] C. Kennington, L. Dia, and D. Schlangen, "A discriminative model for perceptually-grounded incremental reference resolution," in *Proceedings of the 11th International Conference on Computational Semantics (IWCS) 2015*, 2015, pp. 195–205.

[2] C. Kennington and D. Schlangen, "Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution," in *Proceedings of the Conference for the Association for Computational Linguistics (ACL)*, 2015, pp. 292–301.

[3] ——, "A simple generative model of incremental reference resolution for situated dialogue," *Computer Speech & Language*, vol. 41, pp. 43–67, 2017.

[4] R. Iida, S. Kobayashi, and T. Tokunaga, "Incorporating extra-linguistic information into reference resolution in collaborative

task dialogue," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 1259–1267.

[5] H. H. Clark and D. Wilkes-Gibbs, "Referring as a collaborative process," *Cognition*, vol. 22, pp. 1–39, 1986.

[6] S. E. Brennan and H. H. Clark, "Conceptual pacts and lexical choice in conversation," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 22, pp. 1482–1493, 1996.

[7] D. J. Barr and B. Keysar, "Anchoring comprehension in linguistic precedents," *Journal of Memory and Language*, vol. 46, no. 2, pp. 391–418, 2002.

[8] M. J. Pickering and S. Garrod, "Alignment as the basis for successful communication," *Research on Language and Computation*, vol. 4, no. 2, pp. 203–228, 2006.

[9] S. Zarrieß, J. Hough, C. Kennington, R. Manuvinakurike, D. DeVault, R. Fernández, and D. Schlangen, "PentoRef: A corpus of spoken references in task-oriented dialogues," in *10th edition of the Language Resources and Evaluation Conference*, 2016.

[10] C. Matuszek, N. Fitzgerald, L. Zettlemoyer, L. Bo, and D. Fox, "A joint model of language and perception for grounded attribute learning," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 2012, pp. 1671–1678.

[11] K. Funakoshi, M. Nakano, T. Tokunaga, and R. Iida, "A unified probabilistic approach to referring expressions," in *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2012, pp. 237–246.

[12] M. E. Foster, T. By, M. Rickert, and A. Knoll, "Human-robot dialogue for joint construction tasks," in *Proceedings of the 8th International Conference on Multimodal Interfaces*, 2006, pp. 68–71.

[13] A. Ibarra and M. K. Tanenhaus, "The flexibility of conceptual pacts: Referring expressions dynamically shift to accommodate new conceptualizations," *Frontiers in Psychology*, vol. 7, pp. 561–574, 2016.

[14] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, pp. 55–60.

[15] M. Zhu, Y. Zhang, W. Chen, M. Zhang, and J. Zhu, "Fast and accurate shift-reduce constituent parsing," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2013, pp. 434–443.

[16] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, 2003, pp. 173–180.

[17] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: https://www.R-project.org/

[18] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[19] A. Kuznetsova, P. Bruun Brockhoff, and R. Haubo Bojesen Christensen, *lmerTest: Tests in Linear Mixed Effects Models*, 2016, R package version 2.0-33. [Online]. Available: https://CRAN.R-project.org/package=lmerTest

[20] T. Shore, F. Faubel, H. Helmke, and D. Klakow, "Knowledge-based word lattice rescoring in a dynamic context," in *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association*, 2012, pp. 1083–1086.