Using Lexical Alignment and Referring Ability to Address Data Sparsity in Situated Dialog Reference Resolution

Todd Shore †
KTH Royal Institute of Technology
Speech, Music and Hearing
Stockholm, Sweden

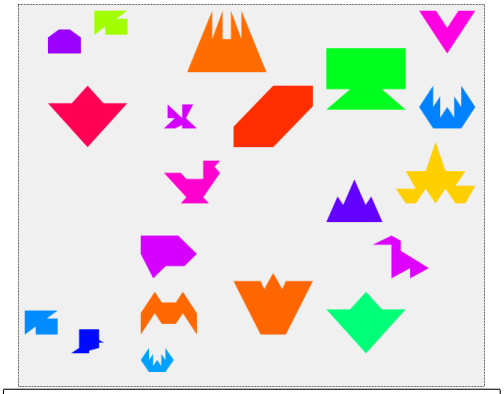
Gabriel Skantze
KTH Royal Institute of Technology
Speech, Music and Hearing
Stockholm, Sweden
skantze@kth.se

Abstract

Referring to entities in situated dialog is a collaborative process, whereby interlocutors often expand, repair and/or replace referring expressions in an iterative process, converging on conceptual pacts of referring language use in doing so. Nevertheless, much work on exophoric reference resolution (i.e. resolution of references to entities outside of a given text) follows a literary model, whereby individual referring expressions are interpreted as unique identifiers of their referents given the state of the dialog the referring expression is initiated. In this paper, we address this collaborative nature to improve dialogic reference resolution in two ways: First, we trained a words-as-classifiers logistic regression model of word semantics and incrementally adapt the model to idiosyncratic language between dyad partners during evaluation of the dialog. We then used these semantic models to learn the general referring ability of each word, which is independent of referent features. These methods facilitate accurate automatic reference resolution in situated dialog without annotation of referring expressions, even with little background data.

1 Introduction

A crucial part of dialog situated in a physical environment is **exophoric references**, i.e. language used by the participants to make entities in the shared environment salient to each other for the purposes of communication (Poesio and Vieira, 1998). Several studies in exophoric reference resolution have investigated how **referential semantics** can be learned automatically via the relationship of a referent's features to the language referring to it (cf. Kennington et al., 2015; Shore and Skantze, 2017) or the state of the interaction a dialog is situated in (cf. Prasov and Chai, 2008; Iida et al., 2010), inferring a relationship between e.g. the word *red* and the individual features it refers to, e.g. a particular range of hue values.



Speaker	Utterance
A	*it's the one to the left*
B	*the purple pinkish uh weird?*
A	*the pinkish one yeah it looks like an asteroid or something big*
B	*the very big?*
A	*yeah*

Figure 1: Collaborative reference in dialog situated in a reference communication task (cf. Krauss and Weinheimer, 1964).

Most works in exophoric reference resolution have assumed the identification of certain subsets of language known as **referring expressions** (REs) that have been either manually or automatically annotated (cf. Schutte et al., 2011; Meena et al., 2012; Zarrieß et al., 2016; Shore and Skantze, 2017). However, discerning REs from non-referring language in dialog is not trivial. For example, Figure 1 illustrates an interaction between two participants in a reference communication task like that of Krauss and Weinheimer (1964), whereby speaker **A** describes a particular

† Deceased 2 July 2018.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2288–2297
Brussels, Belgium, October 31 - November 4, 2018.

referent which must be resolved by speaker **B**. [1].

While REs are idealized as contiguous, single noun phrases (NPs) such as *the pinkish one*, reference in unrestricted, natural dialog is in fact a collaborative process to which both partners in a dyad contribute (Clark and Wilkes-Gibbs, 1986), and not all referring language (RL) is nominal, e.g. *big* in Figure 1. Both participants contribute RL in a cumulative fashion, but often no complete nominal RE is produced, e.g. *the big pinkish asteroid to the left*. Due to this, it is difficult to infer from syntax alone the **referring ability** (RA) of language, i.e. the overall ability of a subset of language to unambiguously refer to entities in discourse (Ariel, 1988; Reboul, 1997). In context (e.g. given the set of possible abstract shapes to choose from), it is easy to infer which words have the greatest RA, but without context this is more difficult. This makes recognizing "non-ideal" cases of RL difficult, as the boundary between RL and non-RL is often fuzzy.

Moreover, participants in dialog tend to develop so-called **conceptual pacts**, which means that they converge on commonly-used RL for unique referents in dialog (Brennan and Clark, 1996). As an example, they may repeatedly refer to a given entity as e.g. *the asteroid* even though *asteroid* may only rarely be used to refer to similar entities in the general population. Thus, RL varies less within a given dialog than across dialogs, and variation of RL has an inverse relationship with the length of the time two participants interact due to alignment of dialog participants' use of language (Clark and Wilkes-Gibbs, 1986; Garrod and Anderson, 1987; Brennan, 1996).

In this paper, we present two contributions to the automatic learning of referential semantics for reference resolution in situated dialog that address these problems: Firstly, we show the benefits of adapting models of RL semantics to a specific dialog as it progresses to accommodate the dyad's idiosyncratic use of RL. Secondly, we present a method for deriving a gradient (non-binary) measure of RA in situated dialog. Thus, instead of first identifying REs and then resolving which entity they refer to, we treat all language in the dialog as being more or less referential, and use this gradual measure together with the referential semantics to derive which entity is being talked about. Our assumption is that while the exact referential semantics of words vary greatly across dyads, the general ability of a given word to successfully refer to entities varies little across dyads. Thus, it should be possible to statistically measure the ability of a set of language to refer to entities in general, irrespective of the language's semantic content (e.g. the exact hue understood as *pink* by a dyad). This knowledge, combined with dialogic adaptation, facilitates accurate automatic reference resolution in situated dialog without annotation of REs, even with little background data.

[1]Examples are from the dataset of Shore et al. (2018)

2 Background

Both behavioral studies on reference resolution and RL and computational models thereof have illustrated the context-sensitive nature of reference resolution and RL and the gradient nature of RA.

2.1 Collaboration in Reference Resolution

Traditionally, reference resolution in dialog was analyzed using a literary model of reference, whereby individual REs are seen as unique identifiers of a referent as in written discourse, i.e. each RE is assumed to be "atomic" in its reference to a particular entity (Clark and Wilkes-Gibbs, 1986, 3). However, shortcomings in this approach have long since been identified (cf. Olson, 1970): REs often do not unambiguously identify their referent when initiated but rather comprise a larger process of collaborative reference resolution, whereby multiple dialog participants iteratively extend, repair and even replace REs initiated by themselves or others (Clark and Wilkes-Gibbs, 1986; Heeman and Hirst, 1995).

In Figure 1, speaker **A** initiates the RE *the one to the left* and immediately expands it in an episodic manner (Clark and Wilkes-Gibbs, 1986, 4, 17). Given a literary model of reference, they should have supplied exactly enough information to identify the referent and no more (*the big pinkish asteroid to the left*), adhering to Grice's (1975) maxim of quantity. However, RL can undergo not only expansion but also replacement: For color, speaker **B** proposes both *purple* and *pinkish*, but only *pinkish* is then accepted by **A**.

In contrast to a literary model of reference, a collaborative model represents reference resolution as a process of iteratively presenting RL to the other participant(s) in a dialog, which is then either accepted as being sufficient to identify a referent or rejected as insufficient (Clark and Wilkes-

Gibbs, 1986, 9). This more accurately models reference observed in spoken dialog.

2.2 Referring Language Syntax

Literary reference models fail to account not only for the collaborative nature of reference resolution but also for the syntactic structure of RL itself: Ideally, a reference is expressed linguistically as an NP, but this ideal does not hold in unrestricted dialog (Clark and Wilkes-Gibbs, 1986).

Since an RE cannot be defined as an atomic referring unit, a model of RL should ideally be able to measure the RA of any given set of language rather than simply classifying language as (part of) an RE in a binary decision, such as by using hand-crafted rules (cf. Shore and Skantze, 2017) or expert annotation (cf. Spanger et al., 2009; Kennington et al., 2015).

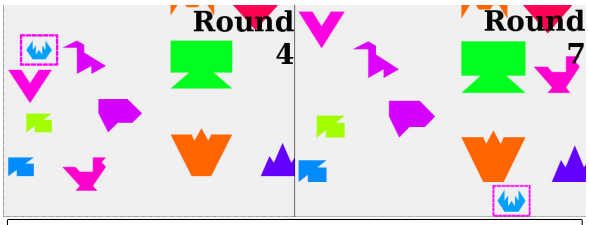
2.3 Modeling Referring Language

There have already been some efforts in automatic annotation of REs: For example, Schutte et al. (2011) algorithmically extracted RL as utterance(s) preceding a discrete event in a shared environment within a certain timeframe. However, one drawback to this method is that "the references must be contained in instructions that cause events involving the referents" and "it must be possible to automatically detect these events" (Schutte et al., 2011, 189). Thus, REs not referring to a detectable event cannot be detected in this manner. Moreover, not all language extracted is that of REs: For example, in the instruction *go through that door*, only half of the tokens constitute RL (*that door*). This means that this method must either be supplemented with additional methods to extract RL or tolerate a high noise-to-signal ratio.

Other approaches use language structure to infer RL, namely in parsing said language using a combination of statistical or rule-based methods. However, both entail that a solution be specialized for language specific to a given domain, such as for route-following instructions (Meena et al., 2012) or for a specific instructor-manipulator pair task (Shore and Skantze, 2017): Meena et al. (2012) used the highly-structured nature of route-following instructions to great effect, while Shore and Skantze (2017) used a phrase-structure parser pre-trained on out-of-domain data and supplemented it with hand-crafted rules to extract NPs according to the literary ideal of RL.

Finally, many works simply ignore the distinction between RL and non-RL and focus solely on learning reference resolution as a function of language and extra-linguistic knowledge such as entity features (cf. Kennington et al., 2015; Shore and Skantze, 2017), discourse and action history (cf. Iida et al., 2010), perception (cf. Matuszek et al., 2012) or gesture (cf. Matuszek et al., 2014). Although these methods improve the resolution of what RL *refers to*, they do not resolve what language *is RL*. Moreover, none of these works address the strong dyadic and dialogic entrainment effects on RL which include the formation of CPs, reinforcing the use of RL specific to a given dialog even if it diverges from population RL use.

3 Data Description



Speaker	Round	Utterance
B	4	*eh it looks like a blue crab sticking up his claws*
		...
A	7	*it's the same the the little crab again*

Figure 2: Example of repeated reference across rounds; the referent the participants have to collaboratively resolve is indicated here with a magenta square.

The data used is that of Shore et al. (2018), a set of $|D| = 42$ task-oriented dialogs (mean duration μ = 15:25 minutes, standard deviation SD = 1:13, total 647:35) in which one participant is an instructor referring to specific pieces on a shared game board which the other participant, the manipulator, must then attempt to resolve by selecting without the aid of extra-linguistic cues (see Figures 1–2): They sit at different locations and communicate solely through an audio channel. Upon successful selection, the piece moves to a random free place on the board and the participants alternate roles. This dataset is somewhat larger than that for similar tasks (cf. Iida et al., 2010; Matuszek et al., 2012; Malinowski and Fritz, 2014; Kennington et al., 2015). However, unlike in many other works, participants

were allowed to refer to pieces in any way they wish and both were allowed to speak freely.

Each dialog $d \in D$ has $|R| = 20$ randomly-generated game pieces and is divided into individual game **rounds** $d \triangleq \langle d'_1 \dots d'_n \rangle$, in each of which $d' \triangleq (R, \hat{r}, T)$ a single entity is pre-selected by the game as the entity $\hat{r} \in R$ which must be successfully resolved for that round. Each dialog presents the same 20 referents aside from their changing position, so participants must also refer to pieces which have already been referenced before: After 40 rounds, all pieces are guaranteed to have been referred to and so every reference thereafter is a coreference (see Figure 2).

Each entity $r \in R$ has features representing shape, size, color and position during the given round. A sequence of tokens T was transcribed from the speech of both participants using Penn Treebank tokenization rules (Marcus et al., 1993). See the Supplementary Material for further information on the dataset.

4 Baseline

The reference resolution method used as a baseline was a **words-as-classifiers** (WaC) regression model (cf. Kennington et al., 2015). In this framework, an individual logistic regression model $p_t(r) \triangleq \sigma(w_t^T r + b_t)$ is trained for each token type t, predicting the probability of a given entity r being the token's TRUE referent $\hat{r}$, given the feature vector r representing shape, size, color and position (see the Supplementary Material for details). For example, if trained successfully, the model for the token "red" should be sensitive to the entity's hue, but not to its size. Common non-descriptive words such as "the" should not be sensitive to any of the entity's properties, yielding an output of 0.5 for all entities. The score of a given entity r being the referent $r = \hat{r}$ of a set of RL tokens T is defined as the normalized linear combination of the tokens' corresponding classifiers $p_t(r)$:

$$p'(r = \hat{r}, T) \triangleq \frac{1}{n} \sum_{t \in T}^{n} p_t(r) \qquad (1)$$

For training, language in each round $(R, \hat{r}, T)$ is defined as a bag of words T referring to the referent $\hat{r}$. For each token $t \in T$, a training example is defined for the referent $\hat{r}$ (with a target score of 1) as well as for each non-referent entity $r \in R \setminus \hat{r}$ (with a target score of 0). To address model bias, the training example for $\hat{r}$ is weighted by its complement set size, $|R \setminus \hat{r}| = 19$.

Initial experiments showed that lemmatization did not affect the performance on our dataset. Thus, each inflected lexical form is considered a unique **word** (i.e., vocabulary item). Unlike Kennington et al. (2015), no smoothing was used, instead ignoring words of fewer than $\alpha \triangleq 3$ occurrences. The motivation for this is that a general out-of-vocabulary model is not expected to increase the performance, since it basically learns to ignore entity properties, similar to the models for common words such as "the". This was also confirmed in our initial experiments.

Note that all language from both the instructor and the manipulator in each round is used. This is unlike Kennington et al. (2015), who only used language from (manually annotated) REs. As argued above, REs cannot easily be identified in the type of dialog data we are addressing. This of course makes the task much more challenging, and the baseline performance can be expected to be lower than that reported in Kennington et al. (2015).

We did 42-fold cross-validation, in each fold using 40 dialogs for training as **background data**, one for testing and one for use as random data to compare the effects of dialog-specific data to (see Section 5 below). Each round in the test dialog is evaluated by the **reciprocal rank** (RR) of the referent $\hat{r}$ in the set of entities R ordered by their combined score for all word classifiers in the round $\sum_{t \in T} p_t(r)$, and its mean (MRR) is then calculated.

Statistic	Mean	SD	SEM
$Rank$	2.8060	3.2427	0.0566
RR	0.6892	0.3648	0.0064

Table 1: Baseline results for 42-fold cross-validation.

The cross-validation results for the baseline WaC model are shown in Table 1. As expected, this is indeed worse than e.g. Kennington et al. (2015)'s reported mean rank of 2.16 when only using speech features. The WaC model is nevertheless a simple and effective representation of referential semantics in domains where features for each individual referent can be easily represented (cf. Kennington and Schlangen, 2015). Still, it has two shortcomings: Firstly, it infers a static model of referential semantics which is good across di-

alogs but is suboptimal for language within dialogs due to effects of language alignment (Garrod and Anderson, 1987; Brennan, 1996; Brennan and Clark, 1996). Secondly, it encodes RA only indirectly: Given a large enough dataset, logistic regression for non-RL such as *okay now I'm ready* should have an even distribution between TRUE and FALSE classes, i.e. these classifiers should decide nothing. Conversely, strong RL such as *red* should entail strong relationships between certain features and decisions. However, due to the effects of idiosyncrasy and alignment on dialogic language, understanding low-frequency words is crucial despite that they cannot be conditioned for as well as can be done for high-frequency ones.

5 Dialogic Model Adaptation

We evaluated the benefits of adapting reference resolution parameters to the language of individual dialogs by initially conditioning WaC models on the training set as background data and then adapting the model during evaluation by retraining using data from previous states in the dialog being evaluated: The RR for the i^{th} round $(R, \hat{r}, T)_i$ is calculated using a model trained on both background data and **interaction data** defined as the rounds observed thus far in the given dialog $(R, \hat{r}, T)_{i'<i}$. The parameters for the logistic regression models representing individual words are optimized using quasi-Newton hybrid conjugate gradient descent from *Weka* v3.8.0 (Dai and Yuan, 2001; Frank et al., 2016). A ridge $\lambda = 100$ was used to avoid over-fitting of models for low-frequency words, tuned using cross-validation over the dataset (le Cessie and van Houwelingen, 1992). The same cross-validation method determined an optimal interaction data weight of 3 relative to background data, i.e. an observation in a given dialog is three times as relevant as one from the background data[2].

Figure 3 compares the improvement of RR from adapting model parameters using dialog interaction data (`Adt`) to the `Baseline` as well as effects of adding data from a randomly-chosen round from another unseen dialog (`RndAdt`): The condition `RndAdt` is used to rule out the possibility that model fit improves simply due to more training data in general. We fit a linear mixed model with conditions `Adt`, `RndAdt`, `Wgt` and scaled `Tokens` as linear fixed effects and game

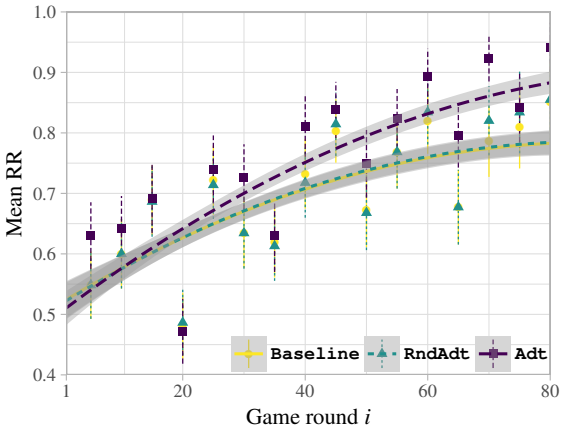


Figure 3: RR as a quadratic function of ROUND $i \cdot i^2$ for model adaptation using interaction data (CI 0.95; error bars are SEM; showing every 5th round).

round ordinality (ROUND) as a quadratic fixed effect: `Wgt` denotes weighting word classifiers by RA, which will be discussed in Section 6. `Tokens` denotes the number of word tokens produced by both speakers in the given round[3]. DYAD (the pair of participants in a given dialog) was included as a random intercept with a random slope for `Adt` and `Wgt`. We selected the best-fitting model using backwards selection with log-likelihood ratio tests: Starting from the maximally complex model (Barr et al., 2013), we first simplified the random structure and then removed fixed effects not contributing to fit. This showed that including `RndAdt` does not significantly improve fit ($\chi^2 = 0.00003, p = 0.99599$), meaning model fit improves from data specific to the given dialog and not merely from more training data.

We refit the final model using maximum-likelihood estimation with Satterthwaite approximation to degrees of freedom (see the Supplementary Material for details). Despite that RR correlates with ROUND i even for the baseline method due to dialogic lexical alignment (cf. Shore and Skantze, 2017; Shore et al., 2018), there is a significant improvement in RR from `Adt` ($B = 0.04882, t(40) = 7.65, p < 0.001$).

Since adding a small amount of data from a dialog significantly improves reference resolution for that dialog, dialogic reference resolution can be seen as a model adaptation problem, where in-domain data (that from the dialog being evaluated)

[2]Interaction data weight values tested were $1, 3, 5, 7, 10$.

[3]Adding the count of coreferences to a given referent as a fixed effect prevented model convergence when included with `Tokens`. Regardless, adding it in lieu of `Tokens` did not significantly improve fit ($\chi^2 = 3.7329,, p = 0.05335$).

is relatively sparse compared to out-of-domain data (that from other dialogs). This suggests that the effect of dyadic alignment on reference resolution is amplified: As ROUND i increases, not only is more data specific to the given dialog available, but the data observed becomes more homogeneous. Thus, the benefit of this method increases with time, as the ratio of interaction to background data increases.

6 Weighting by Referring Ability

While the method above facilitates the adaption of referential semantics models to dyad-specific language, not all language which is rare and/or observed in only one dyad has great RA: For example, in the dataset used, there were 19 observations of the word *awesome* but 15 of those were in a single dialog. Even when evaluating on that dialog, a classifier would be inferred from the $19 - 15 > \alpha$ remaining observations, and even adapting the word models with interaction data as done in Section 5 will only add noise since it only occurs as non-RL (e.g. *awesome good work*). Conversely, a word such as *piece* is semantically heavy in general English but is by itself a poor signifier of referents given the task at hand. So, we evaluated the benefit of weighting word classifiers by their RA in order to mitigate the effects of such spurious observations. To do this, we define the RA of a word t as the mean difference between the probability of the actual referent $\hat{r}$ being TRUE $p_t(\hat{r})$ and the mean probability for all other entities $R \setminus \hat{r}$ for every occurrence of the word in the training data:

$$w_t \triangleq \frac{1}{n} \sum_{\substack{d \in D \\ (R,\hat{r},T) \in d}}^{n} p''_t(R, \hat{r}, T)$$

$$p''_t(R, \hat{r}, T) \triangleq p'_t(\hat{r}, T) - \frac{1}{m} \sum_{r \in R \setminus \hat{r}}^{m} p'_t(r, T) \quad (2)$$

$$p'_t(r, T) \triangleq p_t(r) \sum_{i=1}^{|T|} [T_i = t]$$

One alternative to this metric that we considered was the area under the receiver operating characteristic (ROC) curve (AUC). However, the metric above is more conservative in cases of word models with few observations by penalizing their score due to the logistic ridge used, thus putting more "trust" in word models with more observations; The AUC does not account for this directly.

Although this metric is derived from referential semantics learned for a specific domain, the WaC logistic regression model(s) encoding referential semantics are simple and thus can easily be re-trained for other domains. It can also be derived from other models of referential semantics.

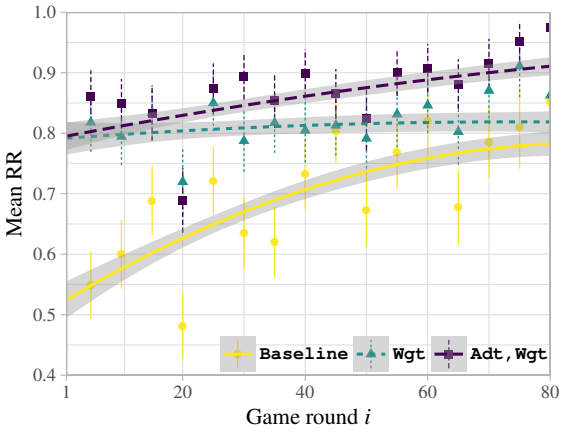


Figure 4: RR as a quadratic function of ROUND $i \cdot i^2$ for weighting model scores by RA $p_t(r) \cdot w_t$ (CI 0.95; error bars are SEM; showing every 5th round).

Figure 4 compares the improvement of RR from weighting each classifier $p_t(r)$ by its RA w_t (`Wgt`) to the `Baseline` and finally to that from combining adaptation from Section 5 with weighting (`Adt,Wgt`); Using the same linear mixed model described in Section 5, a significant improvement in RR was found for `Wgt` over the baseline ($B = 0.1314, t(39) = 11.79, p < 0.001$) although the effect weakens over time. However, this is likely not a weakness of the method but rather an effect of repeated reference on participants' RL use: With repeated reference, the length of RL reduces (Clark and Wilkes-Gibbs, 1986), meaning that the mean RA of each word increases due to fewer tokens of weak RA being uttered. Indeed, a significant interaction between `Round` and `Tokens` was found in their effects on RA (see Supplementary Material). Figure 5 shows a significant correlation of ROUND i and mean RA of all tokens for that round $\frac{1}{n}\sum_{t \in T_i}^{n} w_t$. Additionally, Figure 6 shows a significant inverse relationship with token count $|T_i|$. Since the referent $\hat{r}$ is chosen at random by the game, the amount of references to an entity increases with round ordinality, and so this corresponds with Clark and Wilkes-Gibbs (1986).

A qualitative assessment shows that vocabulary items with the great RA are typically nouns

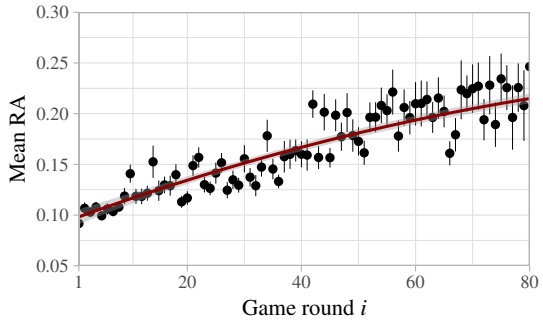


Figure 5: RA w_t for $t \in T_i$ as a quadratic function of ROUND $i \cdot i^2$ (CI 0.95; error bars are SEM).

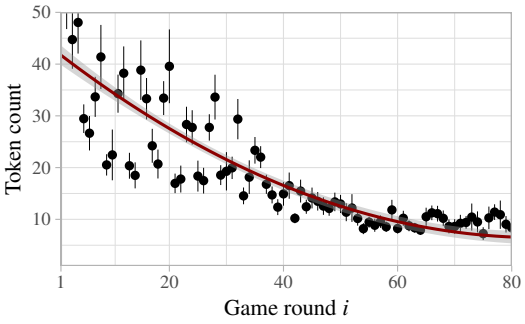


Figure 6: Token count $|T_i|$ as a quadratic function of ROUND $i \cdot i^2$ (CI 0.95; error bars are SEM).

strongly associated with the task at hand: The 31 words with greatest RA are all nouns referring to shapes. Despite this, however, great RA is not exclusive to nouns: In Table 2, *inside*, a preposition, is considered semantically lighter in general English than nouns are (Froud, 2001), but has RA greater than the mean ($\mu = 0.2424$, $SD = 0.1266$). On the other hand, the noun *color* has relatively little RA given the task at hand.

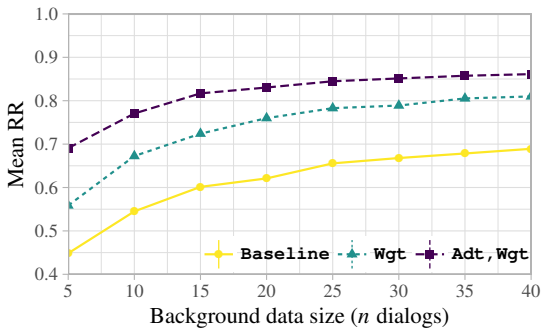


Figure 7: MRR per background training set size.

Moreover, including `Adt` and `Wgt` using dialog interaction data (`Adt,Wgt`) shows significant improvements over either alone: During model selection, including `Wgt` significantly improved fit

Word t	RA w_t	Count
K	0.7567	108
bat	0.6382	69
house	0.6374	153
chicken	0.6340	85
computer	0.6181	96
	...	
inside	0.3985	13
	...	
color	0.0291	195
	...	
's	0.0066	1593
it	0.0051	1731
okay	0.0048	1115
the	0.0040	2478

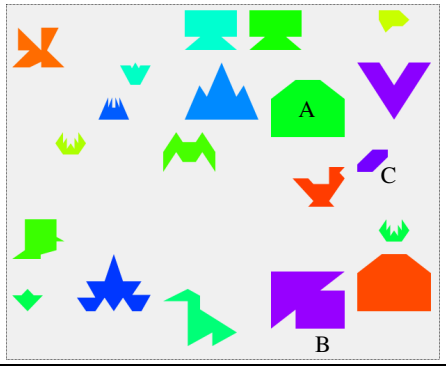
Table 2: Sample vocabulary items ordered by RA.

($\chi^2 = 61.425, p < 0.001$). This means that both methods can be used together and complement each other: Weighting is particularly beneficial for shorter interactions, where little in-domain interaction data is present, while adaptation provides greater benefit for longer interactions (cf. Figure 3). In fact, Figure 7 shows that `Wgt` has better MRR using only 12 randomly-chosen dialogs as background data than the `Baseline` does with 40, and adaptation and weighting together (`Adt,Wgt`) has better MRR with only 7. Figure 8 illustrates the effects of the two conditions on reference resolution in the task used for evaluation: The baseline classifier has a rank of 10 for the referent $\hat{r}$ out of $|R| = 20$ possible referents. In the baseline ***(A)***, the classifier for e.g. *color* has as much weight as e.g. *rectangle* although the former is not a useful signifier for the given task. When weighting by RA ***(B)***, however, the less-useful words contribute less to the total $\sum_{t \in T} (p_t(\hat{r}) \cdot w_t)$, improving rank to 5. Finally, when adapting the model with interaction data ***(C)***, models for semantically-heavy words like *violet* better fit the dyad's RL use, bringing rank to 1.

When both incrementally adapting semantic models with in-domain dialog data and weighting by RA, MRR for reference resolution was improved by 32.5% over the baseline (see Table 3).

7 Conclusion and Discussion

We have shown that it is possible to improve reference resolution for situated dialog by incrementally adapting word semantic model parameters to



(A) Baseline (Target rank 10)

	the	slanted	rectangle	with	two	triangles	violet	in	color
	the	slanted	rectangle	with	two	triangles	violet	in	color
Score	0.51	0.50	0.74	0.31	0.42	0.23	0.41	0.53	0.31

(B) Weighting by RA (Target rank 5)

	the	slanted	rectangle	with	two	triangles	violet	in	color
			rectangle	with	two	triangles	violet		color
Score	0.51	0.50	0.74	0.31	0.42	0.23	0.41	0.53	0.31
RA	0.00	0.00	0.32	0.15	0.19	0.31	0.34	0.02	0.05

(C) Adapting and Weighting by RA (Target rank 1)

	the	slanted	rectangle	with	two	triangles	violet	in	color
		slanted	rectangle	with	two	triangles	violet		color
Score	0.51	0.64	0.80	0.31	0.60	0.63	0.70	0.53	0.46
RA	0.00	0.56	0.34	0.15	0.19	0.30	0.34	0.02	0.04

Figure 8: Reference resolution with different conditions: The TRUE referent $\hat{r}$ is labeled *C*. Word hue denotes semantic score $p_t(\hat{r})$ (green = 1.0, yellow = 0.5, red = 0.0). Saturation denotes RA w_t.

a given dialog in order to accommodate idiosyncratic language use by dyad partners, and the effect of the partners' own alignment makes this method even more beneficial over time. Additionally, we have defined a metric of word referring ability which is derived from a word's referential semantics in situated dialog but holds across individual dialogs despite dyadic variation in RL use. We showed that this metric can be used to automatically determine the usefulness of a given word for reference resolution, meaning that RE annotation is not necessary. Both of these aspects are beneficial to natural language understanding (NLU) for situated dialog due to the difficulty of acquiring data domain-appropriate data.

Model adaption using dialogic knowledge can be effective for improving NLU (cf. Riccardi and Gorin, 2000) despite that little work has been done in this regard specifically for reference resolution. Our experiments with model adaptation in Section 5 suggest that it may be beneficial to treat reference resolution in situated dialog as a model adaptation task, where a given dialog being evaluated is considered "in-domain" data and all other dialogs considered "out-of-domain" data. Moreover, due to the fact that dialog participants' use of RL converges over time (Garrod and Anderson, 1987; Brennan, 1996; Brennan and Clark, 1996), the task should adapt a pre-trained reference resolution model not only for a given dialog but also to the given state of that dialog; On the other hand, Iida et al. (2010) incorporate intra-dialogic knowledge but do not adapt to inter-dialogic effects.

Condition	$Rank$	SD_{Rank}	SEM_{Rank}
`Baseline`	2.8060	3.2427	0.0566
`Adt`	2.4224	2.8614	0.0499
`Wgt`	1.9373	2.3288	0.0406
`Adt,Wgt`	1.5693	1.6685	0.0291
Condition	RR	SD_{RR}	SEM_{RR}
`Baseline`	0.6892	0.3648	0.0064
`Adt`	0.7372	0.3470	0.0061
`Wgt`	0.8099	0.3116	0.0054
`Adt,Wgt`	0.8613	0.2686	0.0047

Table 3: Overall results for conditions evaluated.

Lastly, weighting by RA w_t as derived from logistic word classifier scores $p_t(r)$ in Section 6 was shown to be effective and can be easily inferred from data. However, this inaccurately assumes inter-word independence, since it does not encode a word's context: For example, the RA of *not* was 0.0638, which is relatively low. While it is a poor signifier in itself, it reverses the polarity of the predicate it modifies. For example, in *it's the baby blue K the light one not the dark one*, the NP *the dark one* should in fact have negative RA: Entities with a low semantic score $\sum_{t \in \langle the, dark, one \rangle} p_t(r)$ should in fact be preferred over those those with a high score. This could be addressed via structural prediction (e.g. conditional random fields or neural networks) or even higher-order n-grams, but these methods cannot be easily utilized given the typically small size of situated dialog datasets.

Acknowledgments

This work is supported by the SSF (Swedish Foundation for Strategic Research) project COIN. Plots were made with *ggplot2* v2.2.1 (Wickham, 2009). The authors are grateful for Zofia Malisz's help with model selection.

References

Mira Ariel. 1988. Referring and accessibility. *Journal of Linguistics*, 24(1):65–87.

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Susan E. Brennan. 1996. Lexical entrainment in spontaneous dialog. In *Proceedings of the International Symposium on Spoken Dialogue (ISSD '96)*, pages 41–44, Philadelpha, PA, USA. Acoustical Society of Japan.

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):1482–1493.

S. le Cessie and J. C. van Houwelingen. 1992. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201.

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

Yu-Hong Dai and Ya-xiang Yuan. 2001. An efficient hybrid conjugate gradient method for unconstrained optimization. *Annals of Operations Research*, 103(1):33–47.

Eibe Frank, Mark A. Hall, and Ian H. Witten. 2016. The WEKA workbench. Online appendix for "Data mining: Practical machine learning tools and techniques", Morgan Kaufmann, fourth edition, 2016. Last accessed 21 February 2018.

Karen Froud. 2001. Prepositions and the lexical/functional divide: Aphasic evidence. *Lingua*, 111(1):1–28.

Simon Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218.

Herbert Paul Grice. 1975. Logic and conversation. In *Syntax and Semantics, Volume 3: Speech Acts*, pages 41–58. Academic Press, New York, NY, USA.

Peter A. Heeman and Graeme Hirst. 1995. Collaborating on referring expressions. *Computational Linguistics*, 21(3).

Ryu Iida, Syumpei Kobayashi, and Takenobu Tokunaga. 2010. Incorporating extra-linguistic information into reference resolution in collaborative task dialogue. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1259–1267, Uppsala, Sweden. Association for Computational Linguistics.

Casey Kennington, Livia Dia, and David Schlangen. 2015. A discriminative model for perceptually-grounded incremental reference resolution. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 195–205, London, England, UK. 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, Beijing, China. Association for Computational Linguistics.

Robert M. Krauss and Sidney Weinheimer. 1964. Changes in reference phrases as a function of frequency of usage in social interaction: a preliminary study. *Psychonomic Science*, 1(1):113–114.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1682–1690. Curran Associates, Inc.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2).

Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. 2014. Learning from unscripted deictic gesture and language for human-robot interactions. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 2556–2563, Palo Alto, CA, USA. AAAI Press.

Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012. A joint model of language and perception for grounded attribute learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, pages 1671–1678, New York, NY, USA. Omnipress.

Raveesh Meena, Gabriel Skantze, and Joakim Gustafson. 2012. A data-driven approach to understanding spoken route directions in human-robot dialogue. In *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association*, pages 226–229, Portland, OR, USA. International Speech Communication Association.

David Richard Olson. 1970. Language and thought: Aspects of a cognitive theory of semantics. *Psychological Review*, 77(4):257–273.

Massimo Poesio and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

Zahar Prasov and Joyce Y. Chai. 2008. What's in a gaze?: The role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*, IUI '08, pages 20–29, New York, NY, USA. ACM.

Anne Reboul. 1997. What (if anything) is accessibility? a relevance-oriented criticism of Ariel's accessibility theory of referring expressions. In John H. Connolly, Roel M. Vismans, Christopher S. Butler, and Richard A. Gatward, editors, *Discourse and Pragmatics in Functional Grammar*, pages 91–108. De Gruyter Mouton, Berlin, Germany and Boston, MA, USA.

Giuseppe Riccardi and Allen L. Gorin. 2000. Stochastic language adaptation over time and state in natural spoken dialog systems. *IEEE Transactions on Speech and Audio Processing*, 8(1):3–10.

Niels Schutte, John D. Kelleher, and Brian Mac Namee. 2011. Automatic annotation of referring expression in situated dialogues. *International Journal of Computational Linguistics and Applications*, 2(1-2):175–190.

Todd Shore, Theofronia Androulakaki, and Gabriel Skantze. 2018. KTH Tangrams: A dataset for research on alignment and conceptual pacts in task-oriented dialogue. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Todd Shore and Gabriel Skantze. 2017. Enhancing reference resolution in dialogue using participant feedback. In *Proceedings, GLU 2017 International Workshop on Grounding Language Understanding*, pages 78–82, Stockholm, Sweden.

Philipp Spanger, Yasuhara Masaaki, Iida Ryu, and Tokunaga Tokenobu. 2009. Using extra linguistic information for generating demonstrative pronouns in a situated collaboration task. In *Proceedings of the Workshop on the Production of Referring Expressions: Bridging the gap between computational and empirical approaches to reference*, Amsterdam, the Netherlands.

Hadley Wickham. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, New York, NY, USA.

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 *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).