Probabilistic Labeling for Efficient Referential Grounding based on Collaborative Discourse

Changsong Liu, Lanbo She, Rui Fang, Joyce Y. Chai
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
{cliu, shelanbo, fangrui, jchai}@cse.msu.edu

Abstract

When humans and artificial agents (e.g. robots) have mismatched perceptions of the shared environment, referential communication between them becomes difficult. To mediate perceptual differences, this paper presents a new approach using probabilistic labeling for referential grounding. This approach aims to integrate different types of evidence from the collaborative referential discourse into a unified scheme. Its probabilistic labeling procedure can generate multiple grounding hypotheses to facilitate follow-up dialogue. Our empirical results have shown the probabilistic labeling approach significantly outperforms a previous graph-matching approach for referential grounding.

1 Introduction

In situated human-robot dialogue, humans and robots have mismatched capabilities of perceiving the shared environment. Thus referential communication between them becomes extremely challenging. To address this problem, our previous work has conducted a simulation-based study to collect a set of human-human conversation data that explain how partners with mismatched perceptions strive to succeed in referential communication (Liu et al., 2012; Liu et al., 2013). Our data have shown that, when conversation partners have mismatched perceptions, they tend to make extra collaborative effort in referential communication. For example, the speaker often refers to the intended object iteratively: first issuing an initial *installment*, and then *refashioning* till the hearer identifies the referent correctly. The hearer, on the other hand, often provides useful feedback based on which further refashioning can be made. This data has demonstrated the importance of incorporating collaborative discourse for referential grounding.

Based on this data, as a first step we developed a graph-matching approach for referential grounding (Liu et al., 2012; Liu et al., 2013). This approach uses *Attributed Relational Graph* to capture collaborative discourse and employs a state-space search algorithm to find proper grounding results. Although it has made meaningful progress in addressing collaborative referential grounding under mismatched perceptions, the state-space search based approach has two major limitations. First, it is neither flexible to obtain multiple grounding hypotheses, nor flexible to incorporate different hypotheses incrementally for follow-up grounding. Second, the search algorithm tends to have a high time complexity for optimal solutions. Thus, the previous approach is not ideal for collaborative and incremental dialogue systems that interact with human users in real time.

To address these limitations, this paper describes a new approach to referential grounding based on probabilistic labeling. This approach aims to integrate different types of evidence from the collaborative referential discourse into a unified probabilistic scheme. It is formulated under the Bayesian reasoning framework to easily support generation and incorporation of multiple grounding hypotheses for follow-up processes. Our empirical results have shown that the probabilistic labeling approach significantly outperforms the state-space search approach in both grounding accuracy and efficiency. This new approach provides a good basis for processing collaborative discourse and enabling collaborative dialogue system in situated referential communication.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 13–18,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

2 Related Work

Previous works on situated referential grounding have mainly focused on computational models that connect linguistic referring expressions to the perceived environment (Gorniak and Roy, 2004; Gorniak and Roy, 2007; Siebert and Schlangen, 2008; Matuszek et al., 2012; Jayant and Thomas, 2013). These works have provided valuable insights on how to manually and/or automatically build key components (e.g., semantic parsing, grounding functions between visual features and words, mapping procedures) for a situated referential grounding system. However, most of these works only dealt with the interpretation of single referring expressions, rather than interrelated expressions in collaborative dialogue.

Some earlier work (Edmonds, 1994; Heeman and Hirst, 1995) proposed a symbolic reasoning (i.e. planning) based approach to incorporate collaborative dialogue. However, in situated settings pure symbolic approaches will not be sufficient and new approaches that are robust to uncertainties need to be pursued. DeVault and Stone (2009) proposed a hybrid approach which combined symbolic reasoning and machine learning for interpreting referential grounding dialogue. But their "environment" was a simplistic block world and the issue of mismatched perceptions was not addressed.

3 Data

Previously, we have collected a set of human-human dialogues on an object-naming task (Liu et al., 2012). To simulate mismatched perceptions between a human and an artificial agent, two participants were shown different versions of an image: the *director* was shown the original image containing some randomly placed objects (e.g., fruits), and the *matcher* was shown an impoverished version of the image generated by computer vision. They were instructed to communicate with each other to figure out the identities of some "named" objects (only known to the director), such that the matcher could also know which object has what name.

Here is an example excerpt from this dataset:

D[1]: there is basically a cluster of four objects in the upper left, do you see that (1)
M: yes (2)
D: ok, so the one in the corner is a blue cup (3)
M: I see there is a square, but fine, it is blue (4)
D: alright, I will just go with that, so and then right under that is a yellow pepper (5)
M: ok, I see apple but orangish yellow (6)
D: ok, so that yellow pepper is named Brittany (7)
M: uh, the bottom left of those four? Because I do see a yellow pepper in the upper right (8)
D: the upper right of the four of them? (9)
M: yes (10)
D: ok, so that is basically the one to the right of the blue cup (11)
M: yeah (12)
D: that is actually an apple (13)

As we can see from this example, both the director and the matcher make extra efforts to overcome the mismatched perceptions through collaborative dialogue. Our ultimate goal is to develop computational approaches that can ground interrelated referring expressions to the physical world, and enable collaborative actions of the dialogue agent (similar to the active role that the matcher played in the human-human dialogue). For the time being, we use this data to evaluate our computational approach for referential grounding, namely, replacing the matcher by our automatic system to ground the director's referring expressions.

4 Probabilistic Labeling for Reference Grounding

4.1 System Overview

Our system first processes the data using automatic semantic parsing and coreference resolution. For semantic parsing, we use a rule-based CCG parser (Bozsahin et al., 2005) to parse each utterance into a formal semantic representation. For example, the utterance "a pear is to the right of the apple" is parsed as

$$[a_1, a_2], [Pear(a_1), Apple(a_2), RightOf(a_1, a_2)]$$

which consists of a list of *discourse entities* (e.g., a_1 and a_2) and a list of first-order-logic predicates that specify the unary attributes of these entities and the binary relations between them.

We then perform pairwise coreference resolution on the discourse entities to find out the discourse relations between entities from different utterances. Formally, let a_i be a discourse entity extracted from the current utterance, and a_j a discourse entity from a previous utterance. We train a maximum entropy classifier[2] (Manning and Klein,

[1] *D* stands for the *director*; *M* stands for the *matcher*.

[2] The features we use for the classification include the distance between a_i and a_j, the determiners associated with them, the associated pronouns, the syntactic roles, the extracted unary properties, etc.

2003) to predict whether a_i and a_j should refer to the same object (i.e. *positive*) or to different objects (i.e. *negative*).

Based on the semantic parsing and pairwise coreference resolution results, our system further builds a graph representation to capture the collaborative discourse and formulate referential grounding as a probabilistic labeling problem, as described next.

4.2 Graph Representation

We use an *Attributed Relational Graph* (Tsai and Fu, 1979) to represent the referential grounding discourse (which we call the "*dialogue graph*"). It is constructed based on the semantic parsing and coreference resolution results. The dialogue graph contains a set A of N nodes:

$$A = \{a_1, a_2, \ldots, a_N\}$$

in which each node a_i represents a discourse entity from the parsing results. And for each pair of nodes a_i and a_j there can be an edge $a_i a_j$ that represents the physical or discourse relation (i.e. coreference) between the two nodes.

Furthermore, each node a_i can be assigned a set of "attributes":

$$\mathbf{x}_i = \left\{x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(K)}\right\}$$

which are used to specify information about the unary properties of the corresponding discourse entity. Similarly, each edge $a_i a_j$ can also be assigned a set of attributes $\mathbf{x}_{ij}$ to specify information about the binary relations between two discourse entities. The node attributes are from the semantic parsing results, i.e., the unary properties associated to a discourse entity. The edge attributes can be either from parsing results, such as a spatial relation between two entities (e.g., $RightOf(a_1, a_2)$); Or from pairwise coreference resolution results, i.e., two entities are coreferential ($coref = +$) or not ($coref = -$).

Besides the dialogue graph that represents the linguistic discourse, we build another graph to represent the perceived environment. This graph is called the "*vision graph*" (since this graph is built based on computer vision's outputs). It has a set Ω of M nodes:

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_M\}$$

in which each node ω_α represents a physical object in the scene. Similar to the dialogue graph, the vision graph also has edges (e.g., $\omega_\alpha \omega_\beta$), node attributes (e.g., $\breve{\mathbf{x}}_\alpha$) and edge attributes (e.g., $\breve{\mathbf{x}}_{\alpha\beta}$). Note that the attributes in the vision graph mostly have numeric values extracted by computer vision algorithms, whereas the attributes in the dialogue graph have symbolic values extracted from the linguistic discourse. A set of "symbol grounding functions" are used to bridge between the heterogeneous attributes (described later).

Given these two graph representations, referential grounding then can be formulated as a "*node labeling*" process, that is to assign a label θ_i to each node a_i. The value of θ_i can be any of the M node labels from the set Ω.

4.3 Probabilistic Labeling Algorithm

The probabilistic labeling algorithm (Christmas et al., 1995) is formulated in the Bayesian framework. It provides a unified evidence-combining scheme to integrate unary attributes, binary relations and prior knowledge for updating the labeling probabilities (i.e. $P(\theta_i = \omega_\alpha)$). The algorithm finds proper labelings in an iterative manner: it first initiates the labeling probabilities by considering only the unary attributes of each node, and then updates the labeling probability of each node based on the labeling of its neighbors and the relations with them.

Initialization:

Compute the initial labeling probabilities:

$$P^{(0)}(\theta_i = \omega_\alpha) = \frac{P(a_i \mid \theta_i = \omega_\alpha)\, \hat{P}(\theta_i = \omega_\alpha)}{\sum\limits_{\omega_\lambda \in \Omega} P(a_i \mid \theta_i = \omega_\lambda)\, \hat{P}(\theta_i = \omega_\lambda)}$$

in which $\hat{P}(\theta_i = \omega_\alpha)$ is the prior probability of labeling a_i with ω_α. The prior probability can be used to encode any prior knowledge about possible labelings. Especially in incremental processing of the dialogue, the prior can encode previous grounding hypotheses, and other information from the collaborative dialogue such as confirmation, rejection, or replacement.

$P(a_i \mid \theta_i = \omega_\alpha)$ is called the "compatibility coefficient" between a_i and ω_α, which is computed based on the attributes of a_i and ω_α:

$$\begin{aligned} P(a_i \mid \theta_i = \omega_\alpha) &= P(\mathbf{x}_i \mid \theta_i = \omega_\alpha) \\ &\approx \prod_k P\left(x_i^{(k)} \mid \theta_i = \omega_\alpha\right) \end{aligned}$$

and we further define

$$
\begin{aligned}
P\left(x_i^{(k)} \mid \theta_i = \omega_\alpha\right) &= p\left(x_i^{(k)} \mid \breve{x}_\alpha^{(k)}\right) \\
&= \frac{p\left(\breve{x}_\alpha^{(k)} | x_i^{(k)}\right) p\left(x_i^{(k)}\right)}{\sum\limits_{x_j^{(k)} \in L^{(k)}} p\left(\breve{x}_\alpha^{(k)} | x_j^{(k)}\right) p\left(x_j^{(k)}\right)}
\end{aligned}
$$

where $L^{(k)}$ is the "lexicon" for the k-th attribute of a dialogue graph node, e.g., for the *color* attribute:

$$L^{(k)} = \{red, green, blue, \ldots\}$$

and $p\left(\breve{x}_\alpha^{(k)} \mid x_i^{(k)}\right)$ is what we call a "symbol grounding function", i.e., the probability of observing $\breve{x}_\alpha^{(k)}$ given the word $x_i^{(k)}$. It judges the compatibilities between the symbolic attribute values from the dialogue graph and the numeric attribute values from the vision graph. These symbol grounding functions can be either manually defined or automatically learned. In our current work, we use a set of manually defined grounding functions motivated by previous work (Gorniak and Roy, 2004).

Iteration:

Once the initial probabilities are calculated, the labeling procedure iterates till all the labeling probabilities have converged or the number of iterations has reached a specified limit. At each iteration and for each possible labeling, it computes a "support function" as:

$$
\begin{aligned}
Q^{(n)}\left(\theta_i = \omega_\alpha\right) = & \prod_{j \in N_i} \sum_{\omega_\beta \in \Omega} P^{(n)}\left(\theta_j = \omega_\beta\right) \\
& P\left(a_i a_j \mid \theta_i = \omega_\alpha, \theta_j = \omega_\beta\right)
\end{aligned}
$$

and updates the probability of each possible labeling as:

$$P^{(n+1)}(\theta_i = \omega_\alpha) = \frac{P^{(n)}(\theta_i = \omega_\alpha) Q^{(n)}(\theta_i = \omega_\alpha)}{\sum\limits_{\omega_\lambda \in \Omega} P^{(n)}(\theta_i = \omega_\lambda) Q^{(n)}(\theta_i = \omega_\lambda)}$$

The support function $Q^{(n)}\left(\theta_i = \omega_\alpha\right)$ expresses how the labeling $\theta_i = \omega_\alpha$ at the n-th iteration is supported by the labeling of a_i's neighbors[3], taking into consideration the binary relations that exist between a_i and them. Similar to the node compatibility coefficient, the edge compatibility coefficient between $a_i a_j$ and $\omega_\alpha \omega_\beta$, namely the $P\left(a_i a_j \mid \theta_i = \omega_\alpha, \theta_j = \omega_\beta\right)$ for computing $Q^{(n)}\left(\theta_i = \omega_\alpha\right)$, is also based on the attributes of the two edges and their corresponding symbol grounding functions. So we also manually defined a set of grounding functions for edge attributes such as the spatial relation (e.g., $RightOf$, $Above$). If an edge is used to encode the discourse relation between two entities (i.e., the pairwise coreference results), the compatibility coefficient can be defined as (suppose edge $a_i a_j$ encodes a *positive* coreference relation between entities a_i and a_j):

$$
\begin{aligned}
& P\left(\overline{a_i a_j} = + \mid \theta_i = \omega_\alpha, \theta_j = \omega_\beta\right) \\
= & \frac{P\left(\theta_i = \omega_\alpha, \theta_j = \omega_\beta | \overline{a_i a_j} = +\right) P\left(\overline{a_i a_j} = +\right)}{P\left(\theta_i = \omega_\alpha, \theta_j = \omega_\beta\right)}
\end{aligned}
$$

which can be calculated based on the results from the coreference classifier (Section 4.1).

	Top-1	Top-2	Top-3
Random Guess[a]	7.7%	15.4%	23.1%
S.S.S.	19.1%	19.7%	21.3%
P.L.	24.9%	36.1%	45.0%
Gain[b]	5.8% ($p < 0.01$)	16.4% ($p < 0.001$)	23.7% ($p < 0.001$)
P.L. using annotated coreference	66.4%	74.8%	81.9%

[a] Each image contains an average of 13 objects.

[b] p-value is based on the Wilcoxon signed-rank test (Wilcoxon et al., 1970) on the 62 dialogues.

Table 1: Comparison of the reference grounding performances of a random guess baseline, Probabilistic Labeling (P.L.) and State-Space Search (S.S.S.), and P.L. using manually annotated coreference.

5 Evaluation and Discussion

Our dataset has 62 dialogues, each of which contains an average of 25 valid utterances from the director. We first applied the semantic parser and coreference classifier as described in Section 4.1 to process each dialogue, and then built a graph representation based on the automatic processing results at the end of the dialogue. On average, a dialogue graph consists of 33 discourse entities from the director's utterances that need to be grounded.

We then applied both the probabilistic labeling algorithm and the state-space search algorithm to ground each of the director's discourse entities onto an object perceived from the image. The averaged grounding accuracies of the two algorithms

[3] The set of indices N_i is defined as:

$$N_i = \{1, 2, \ldots, i-1, i+1, \ldots, N\}$$

are shown in the middle part of Table 1. The first column of Table 1 shows the grounding accuracies of the algorithm's top-1 grounding hypothesis (i.e., $\theta_i = \underset{\omega_\alpha}{\operatorname{argmax}} P\left(\theta_i = \omega_\alpha\right)$ for each i). The second and third column then show the "accuracies" of the top-2 and top-3 hypotheses[4], respectively.

As shown in Table 1, probabilistic labeling (i.e. P.L.) significantly outperforms state-space search (S.S.S.), especially with regard to producing meaningful multiple grounding hypotheses. The state-space search algorithm actually only results in multiple hypotheses for the overall matching, and it fails to produce multiple hypotheses for many individual discourse entities. Multiple grounding hypotheses can be very useful to generate responses such as clarification questions or nonverbal feedback (e.g. pointing, gazing). For example, if there are two competing hypotheses, the dialogue manager can utilize them to generate a response like "I see two objects there, are you talking about this one (pointing to) or that one (pointing to the other)?". Such proactive feedback is often an effective way in referential communication (Clark and Wilkes-Gibbs, 1986; Liu et al., 2013).

The probabilistic labeling algorithm not only produces better grounding results, it also runs much faster (with a running-time complexity of $O\left(MN^2\right)$,[5] comparing to $O\left(N^4\right)$ of the state-space search algorithm[6]). Figure 1 shows the averaged running time of the state-space search algorithm on a Intel Core i7 1.60GHz CPU with 16G RAM computer (the running time of the probabilistic labeling algorithm is not shown in Figure 1 since it always takes less than 1 second to run). As we can see, when the size of the dialogue graph becomes greater than 15, state-space search takes more than 1 minute to run. The efficiency of the probabilistic labeling algorithm thus makes it more appealing for real-time interaction applications.

Although probabilistic labeling significantly outperforms the state-space search, the grounding performance is still rather poor (less than 50%) even for the top-3 hypotheses. With no surprise, the coreference resolution performance plays an important role in the final grounding performance (see the grounding performance of using manually annotated coreference in the bottom part of Table 1). Due to the simplicity of our current coreference classifier and the flexibility of the human-human dialogue in the data, the pairwise coreference resolution only achieves 0.74 in precision and 0.43 in recall. The low recall of coreference resolution makes it difficult to link interrelated referring expressions and resolve them jointly. So it is important to develop more sophisticated coreference resolution and dialogue management components to reliably track the discourse relations and other dynamics in the dialogue to facilitate referential grounding.

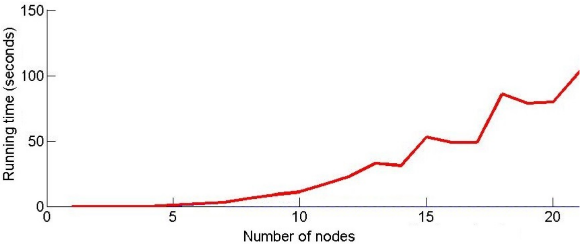

Figure 1: Average running time of the state-space search algorithm with respect to the number of nodes to be grounded in a dialogue graph.

6 Conclusion

In this paper, we have presented a probabilistic labeling based approach for referential grounding in situated dialogue. This approach provides a unified scheme for incorporating different sources of information. Its probabilistic scheme allows each information source to present multiple hypotheses to better handle uncertainties. Based on the integrated information, the labeling procedure then efficiently generates probabilistic grounding hypotheses, which can serve as important guidance for the dialogue manager's decision making. In future work, we will utilize probabilistic labeling to incorporate information from verbal and non-verbal communication incrementally as the dialogue unfolds, and to enable collaborative dialogue agents in the physical world.

Acknowledgments

This work was supported by N00014-11-1-0410 from the Office of Naval Research and IIS-1208390 from the National Science Foundation.

[4] The accuracy of the top-2/top-3 grounding hypotheses is measured by whether the ground-truth reference is included in the top-2/top-3 hypotheses.

[5] M is the number of nodes in the vision graph and N is the number of nodes in the dialogue graph.

[6] Beam search algorithm is applied to reduce the exponential $O\left(M^N\right)$ to $O\left(N^4\right)$.

References

Cem Bozsahin, Geert-Jan M Kruijff, and Michael White. 2005. Specifying grammars for openccg: A rough guide. *Included in the OpenCCG distribution*.

William J. Christmas, Josef Kittler, and Maria Petrou. 1995. Structural matching in computer vision using probabilistic relaxation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(8):749–764.

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

David DeVault and Matthew Stone. 2009. Learning to interpret utterances using dialogue history. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 184–192. Association for Computational Linguistics.

Philip G Edmonds. 1994. Collaboration on reference to objects that are not mutually known. In *Proceedings of the 15th conference on Computational linguistics-Volume 2*, pages 1118–1122. Association for Computational Linguistics.

Peter Gorniak and Deb Roy. 2004. Grounded semantic composition for visual scenes. *J. Artif. Intell. Res.(JAIR)*, 21:429–470.

Peter Gorniak and Deb Roy. 2007. Situated language understanding as filtering perceived affordances. *Cognitive Science*, 31(2):197–231.

Peter A Heeman and Graeme Hirst. 1995. Collaborating on referring expressions. *Computational Linguistics*, 21(3):351–382.

Krishnamurthy Jayant and Kollar Thomas. 2013. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association of Computational Linguistics*, 1:193–206.

Changsong Liu, Rui Fang, and Joyce Chai. 2012. Towards mediating shared perceptual basis in situated dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 140–149, Seoul, South Korea, July. Association for Computational Linguistics.

Changsong Liu, Rui Fang, Lanbo She, and Joyce Chai. 2013. Modeling collaborative referring for situated referential grounding. In *Proceedings of the SIGDIAL 2013 Conference*, pages 78–86, Metz, France, August. Association for Computational Linguistics.

Christopher Manning and Dan Klein. 2003. Optimization, maxent models, and conditional estimation without magic. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Tutorials - Volume 5*, NAACL-Tutorials '03, pages 8–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Alexander Siebert and David Schlangen. 2008. A simple method for resolution of definite reference in a shared visual context. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 84–87. Association for Computational Linguistics.

Wen-Hsiang Tsai and King-Sun Fu. 1979. Error-correcting isomorphisms of attributed relational graphs for pattern analysis. *Systems, Man and Cybernetics, IEEE Transactions on*, 9(12):757–768.

Frank Wilcoxon, SK Katti, and Roberta A Wilcox. 1970. Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test. *Selected tables in mathematical statistics*, 1:171–259.