Know what I'm talking about? Dual eye-tracking in multimodal reference resolution

Alan T. Clark
Dept. of Communication Studies
Northwestern University
alan-clark@northwestern.edu

Darren Gergle
Dept. of Communication Studies
Dept. of Electrical Engineering and
Computer Science
Northwestern University
dgergle@northwestern.edu

Copyright is held by the author/owner(s).

CSCW 2012, February 11-12, 2012, Seattle, Washington, USA.

Abstract

Multimodal reference resolution is an important component of intelligent user interfaces using natural language. We use multimodal context data – including dyadic gaze - to develop models to identify reference in natural conversation. Our approach uses a mixture of rule-based and automatically coded gaze-based, linguistic, spatial, and temporal features, and conversational conditions. We accurately identify the objects speakers were referring to at more than 60%, a rate substantially higher than random and majority baselines. We also explore the contribution to reference resolution of combinations of linguistics, spatial, and gaze features (both speaker and overlapping).

Keywords

Dual eye tracking, gaze, reference, multimodal, shared visual space

ACM Classification Keywords

H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces – collaborative computing, computer-supported cooperative work, computer-mediated communication

General Terms

Human Factors, Experimentation, Measurement, Design

Introduction

In this paper we consider how dual eye-tracking methods can be used to better understand the role gaze plays as a "conversational resource" [14] during reference – how people specify the person, object or entity that they are talking about [5]. Our work builds on a number of studies have begun to explore the relationship that exists between gaze and reference. For example, conversational partners' shared gaze toward referents is higher while they speak about those objects [3], addressees make use of the speaker's gaze as a cue for disambiguating references [11], and shared gaze toward local referents is reduced when speakers have an alternative indicator of attention, such as physical movement toward a referent [9].

Yet, while these studies demonstrate that reference is a multimodal process whereby objects are evoked through a conversational partner's actions, movement, gaze, etc., most computational accounts of reference focus primarily on spoken language. These models largely ignore the potential benefits of collecting and modeling features of non-verbal referential context that people often rely on in conversation [4,10].

Multimodal reference resolution is an increasingly important component of intelligent user interfaces using natural language. For example, in order for robots or virtual agents to interact with humans using speech in space, they will need to identify which objects people are talking about. In turn, they will need richer computational models that can understand the ambiguous speech and non-verbal contextual cues humans use in everyday conversation. The increasing prevalence of speech interfaces in consumer technology (e.g. Apple's "Siri"; Xbox 360 Kinect) further signal a trend toward reliance on intelligent user interfaces that will need to be able to understand naturalistic speech, including multimodal reference. Anticipating this, we sought to use multimodal input - with a particular focus on dyadic gaze data - to develop machine learning (ML) models of reference resolution. We generate and empirically evaluate a series of models that employ *gaze features (both individual and dyadic measures), spatial features, linguistic features*, *temporal features*, and *conversational conditions* to demonstrate their utility in identifying speakers' references.

Background

In earlier work [6,9], we identified several important aspects of communicative context that contribute to the multimodal reference process. Gaze is clearly an important conversational resource dyads use in multimodal reference. Briefly, conversation partners use each others' gaze to indicate attention to and understanding of references to objects in their environment [6,9,11]. Gaze overlap also interacts with other contextual factors. For example, gaze overlap on referents is generally above chance [3], but substantially lower than chance when mobile pairs use local deictic (pointing) references [9]. The language speakers use to refer is also highly sensitive to the visual information in the dyad's common ground [10,14]. What the dyad has already said, and have recently said, in the conversation affects the type of language speakers use. For example, if one referred with a highly specified expression like "the red box on the left", one can subsequently use a less specified reference like "it" and anticipate that the addressee will understand it's a reference to the same object. A dyad's spatial context, including the relative positions of the dyad and objects in shared visual space, affects how

Category	*Features*
Gaze Features	Speaker, addressee, and overlapping gaze on each object (12 variables); total gaze overlap
Spatial Features	Speaker & addressee position; Speaker & addressee time since last movement; Relative position; object(s) nearest speaker
Linguistic Features	Referential form; Plural; Local; Givenness; Givenness shift (from previous reference); Referent shift
General Context	Dyad, speaker, condition
Temporal Features	Onset time; time since last reference

Table 1. Features applied to reference resolution ML

and when speakers adapt their referring expressions to their addressee's perspective [2,15]. Temporal context also affects how dyads refer – for example, speakers might use more detailed language and rely less on gaze overlap early in a conversation. Finally, the conversational conditions of producing a particular reference can have idiosyncratic effects models need to account for. For example, the proportion of who speaks in a dyad varies widely as does the degree to which pairs talk about particular objects. We based our model's features on such findings.

Previous studies have used a variety of computational approaches to attempt to identify referential ambiguities in natural language [e.g., 8]. However, these studies have largely relied on linguistic data and do not take advantage of the potential benefits of using nonverbal conversational resources that people employ in referential communication [11]. Similar approaches that take nonverbal context into account exist. For example, [4] used visual context in a 3D virtual environment to aid in a reference resolution model.

Study

For this study we use the corpus collected in [9] drawn from a dyadic naturalistic referential elicitation task that has associated context data - linguistic features, proxemic context, timing, and gaze. Further details on the study and methods are found in [6,9].

Unlike existing NLP approaches to reference resolution, we focus on using easily acquired linguistic data that also takes a dyad's non-verbal coordination into account. We aim to use features which could provide a practical baseline for current technology, and we sought to test how well a semi-supervised approach could identify a speaker's intended referents using five types of features - gaze, spatial, linguistic, temporal, and general - that could be plausibly extracted from natural language without human annotation.

Model Implementation

We used the Orange data mining suite [8] to implement ML models using the feature sets described in Table 1. The goal of the learned models was to predict the speaker's intended referent for each of 1,546 referring expressions in our corpus. We include sample data from our corpus with a sample of coded features in Figure 1. The accuracy of these predictions was tested against a human-coded gold standard.

Gaze Features

Our models used thirteen gaze features: speaker's and addressee's proportion of gaze toward each of four referent objects; the dyad's rate of overlapping gaze on each of four objects; and the dyad's total rate of gaze

Time Elapsed	*Referring Expression*	*Speaker*	*Ref. Form*	*Givenness Shift*	*Gaze OverlapA*	*Speaker Position*	*True Referent*
6.15	**Th**is one?	1	this	--	.605	2	A
16.34	**It**'s a little Zen you know?	2	it	1	.895	4	A
21.24	I sort of see **it** as a face	1	it	0	.159	2	A

Figure 1. Sample data with sample subset of feature codes

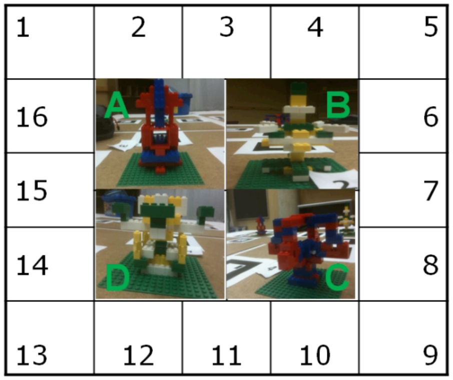

Figure 2. Grid for position coding

overlap across all four potential referent objects. In order to make our gaze data more resilient to saccades, we aggregated our gaze data into 3 'bins' per second containing points of gaze sampled at approximately 30hz. In these bins, we identified gaze targets only if they had a majority share of points of gaze within the bin. Each of these gaze statistics was calculated during a window of 9 bins (3 seconds) on either side of the onset of the referring term (e.g. "this"). So, for example, the Speaker Gaze toward Object A feature might have a value of .33 if the speaker was looking at object A in 6 of 18 bins. Overlap statistics used a speaker-based naive initiative calculation - so, for example, with a 1-bin offset, the gaze overlap was always calculated with the addressee's gaze compared with a point .33 seconds behind the speaker's. We used a 1-bin offset in all models discussed in the paper.

Spatial Features

We developed a spatial coding scheme that identified speaker and addressee position relative to each other and to referent objects, allowing us to incorporate proxemic data such as relative position of speaker and addressee [15] and distance from objects [2]. We used a top-down "grid" system (Figure 2) that, while not highly precise, could be quickly determined using head tracking or basic computer vision and an overhead camera. Using this coding scheme, we derived seven spatial features. We identified the speaker and addressee's current position at the time of the referring expression, including the addressee's position relative to the speaker (side-by-side, across, on the left, or on the right). We also noted the time that the speaker and addressee had last moved, which we anticipated as a potentially useful indicator of referents given the use of spatial evocation of referents noted in [9]. Finally, we included a feature indicating the object(s) nearest the speaker at the time of referring as a potential indicator of referent when local deixis was used.

Linguistic Features

We purposely avoided approaches such as deep parsing in order to make our six linguistic features easy to derive and feasible for end-to-end systems. Therefore, we employ simple linguistic features based on a keyword approach that captures core elements of referring expressions of theoretical interest. First, we used a simple code of the referential form of a given reference - for example, whether the reference was a definite description. We used this referential form as a keyword to derive other linguistic features using simple rules in order to roughly approximate the type of automated feature extraction that real interfaces would require. We used simple rules to indicate whether the reference was singular or plural and also whether it was proximal (local) or distal (remote). We also used a simple heuristic based on the "Accessibility of Referential Form Coding Scheme" in [3] to create a feature that indicated the “givenness” of a referent. Although this is a relatively simplistic notion of givenness [1], we wanted a feature to represent speakers' tendency to use less detail as they repeatedly referred to the same object. Along similar lines, we created a related ‘givenness shift’ feature indicating that dyads were going from a more specified to less specified reference, or vice versa. In turn, this allowed us to create a code for 'reference shifts' – in essence, capturing changes from discussion of one referent to another. When speakers shift from more specified (“this one”) to less specified (“it”) references, it often indicates that they are still referring to the same thing. Our rule for reference shifts combined the "time of last

	F-Measure		*% Refs. Correctly Resolved*
	Sing.	*Plur.*	
Naive Random Baseline	--	--	6.67%
Majority Baseline (A)	--	--	19.34%
Gaze-Only Baseline	.613	.093	52.2%
Gaze & Lang. Baseline	.659	.473	60.9%
Gaze & Spa. Baseline	.638	.279	54.4%
Lang. & Spa. Baseline	.256	.439	30.5%
Full Model	.648	.474	60.2%

Table 2. Machine learning true referent prediction performance using SVM with 1-bin gaze offset

reference" temporal feature and the "givenness shift" feature: if a reference occurred more than 10s after the last referring expression, or occurred within 10s and with a more specified form than the previous referring expression, we marked it as a reference shift. What such rule-based approaches lack in preciseness, they make up for with plausibility for real-world applications.

Temporal Features

We used two temporal features based on the place of each reference in the time course of the conversations. Onset time refers to the point in the discourse at which the referring expression was spoken (specifically, the onset of the referring term like "that"). Time since last reference indicates how much time had elapsed since the pair last referred to any of the objects.

Conversational Conditions

As control variables in our models we made use of three simple features that captured the speaker and experimental condition in which the reference was produced. We identified which dyad and which of speaker within that dyad produced each reference. We also included the experimental spatial condition that captured whether speakers were either seated across a table, seated side-by-side, or standing and mobile.

Machine Learning Approaches

We tested several ML approaches with a range of sophistication and computational intensity. Initial attempts revealed particularly poor performance for K-nearest-neighbor and CN2 approaches, so those were not tested further. We continued to test with three types of ML approaches - Naive Bayes classifiers (NB), Multiclass Support Vector Machines (SVM), and Classification Trees (CT). Describing the nuances of each approach is beyond the scope of this paper - we suggest reading [12,13] for further description. We initially tested models using all three approaches, but decided to use SVM in all the models reported here as it was generally the highest performing plausible approach - SVM was able to accurately identify referents at a higher rate (up to 10%) than NB or CT approaches. For each model, we used 5-fold cross validation[1] to test the validity of the predictions. This approach is commonly used way to provide validity, and is not particularly computationally expensive.

Results

As a first step, we tested these ML models against each other and against baselines with subsets of our features. In Table 2, we include in these baselines reference resolution using only gaze features (row 3), language and gaze (row 4), gaze and spatial features (row 5), and language and spatial data (row 6). All of our models included the conversational conditions and temporal features. For comparison, we also included a naive random baseline, which assumes that the model would pick from one of the 15 possible combinations of referents in our set. Similarly, we included a distribution-based majority baseline in which the model always guessed the referent (i.e., Object A) appearing most often in the corpus. To measure the effectiveness of the models, we used F-measure (a combination of precision and recall) for both singular and plural

[1] Although other approaches (e.g. 10-fold cross validation; leave-one-out) are more commonplace, we chose a method that would not be too computationally demanding. We also note that more intensive validation methods do not necessarily yield different results [12]. In our own case, a 10-fold validation produced nearly identical results for our full SVM model (differences of <.005 f-measure, <.5% accuracy).

	F-Measure		*% Refs. Correctly Resolved*
	Sing.	*Plur.*	
Gaze & Spa. (All)	.638	.279	54.4%
Gaze & Spa. (Standing)	.684	.458	52.4%
Gaze & Spa. (Seated)	.661	.319	58.1%

Table 3. ML true reference prediction performance with branching spatial

references, as well as the general prediction accuracy of the learner in picking the correct speaker's referent. We used a strict accuracy measure - if the speaker said "these two" talking about objects A and B, but the learner predicted objects A and C (or just object B), we didn't consider it a partial match. We discuss the performance of our models against reduced-input baselines, describe the comparative predictive power of speaker gaze vs. gaze overlap, and discuss how the model performs in reference resolution in our study's spatial conditions.

Baseline Performance

All of the reduced-input models listed in Table 2 perform well compared to a naive random baseline (row 1) and distribution-based majority baseline (row 2). Models with multimodal input (rows 3-7) proved useful for reference resolution, with accuracy substantially increased above these baselines. This was the case even in models without linguistic features included (row 6). This suggests value in our basic approach – even without deep parsing, our models were able to effectively identify referents more than 60% of the time. These reduced-input models highlight the relative strengths of features for reference resolution. Gaze features alone (row 3) were able to resolve references correctly in more than half of cases, although their accuracy was primarily driven by singular references - they performed very poorly in predicting plurals. Adding spatial features or linguistic features (rows 4 and 5) only marginally improved performance on singulars, but drastically improved prediction accuracy on plurals. Linguistic features, in particular, seemed to provide a major benefit to resolving plurals. In all cases, these feature baselines performed well above chance or distribution.

Interestingly, spatial features seemed to have some benefit for resolving plurals – adding spatial features to a gaze baseline (row 4) improved plural accuracy by nearly 20%, but netted only a small marginal gain in predicting singulars. Earlier findings indicate that speakers use movement toward objects to evoke local referents and reduce reliance on gaze coordination [9] and that speakers take addressee's relative position to speaker and object into account when formulating references [15]. We anticipated that this would have more of a benefit for singulars. One disadvantage of such semi-supervised models is that we don't pre-assign weights to particular features or situations - which may point to some performance benefits of using more highly supervised ML approaches or algorithms for multimodal reference resolution. However, performance benefits may be offset by reduced applicability across contexts.

We also note that the full model (row 7) performed slightly worse than the model which had language and gaze data but no spatial data (row 5). We suspect this may be because the models processed positional data from seated conditions and standing conditions without considering the relative role of spatial data in each context. In future work, we suspect that spatial information will be useful if applied more strategically. For example, performance might improve if better spatial measures (e.g., F-Formations) are used rather than the simple spatial features we used.

Spatial Branching

We also sought to initially inspect whether handling seated and standing dyads differently improved the utility of spatial data for reference resolution. The results of this are found in Table 3. We found that by

	F-Measure		*% Refs. Correctly Resolved*
	Sing.	*Plur.*	
Lang. & Spa. Baseline	.256	.439	30.5%
Baseline + Gaze (Speaker)	.606	.456	56.7%
Baseline + Gaze (Overlap)	.360	.430	37.9%

Table 4. ML true reference prediction performance using speaker vs. overlapping gaze

branching into a model for the standing condition (row 2) and a model for the seated conditions (row 3), we were able to slightly improve the combined classification accuracy from 56.3% from the 54.4% classification accuracy of an un-branched classifier (row 1). While we again ran into the problem of reducing the size of test and training sets (and possibly overfitting), we substantially improved performance metrics in resolving plurals and slightly improved performance in resolving singulars by branching the classifiers.

Speaker's Gaze vs. Gaze Overlap

Given previous findings that speaker's gaze is a strong cue that addressees use for disambiguating reference [11], we wanted to determine how much of the importance of gaze in the models was driven by the speaker's gaze as opposed to the joint measure of gaze overlap. We show these results in table 4. To do this, we added either speaker's gaze (row 2) or gaze overlap (row 3) to a baseline of language & spatial features (row 1). Speaker's gaze seemed to be the major driver of the gaze features' contribution, contributing a roughly 26% increase in classification accuracy when added to a baseline of the other feature sets. By comparison, gaze overlap data by itself contributed nearly 7.5% accuracy when added to the other features. While gaze overlap by itself may be insufficient as a predictor, it still proved useful. Finally, we note that if one is using speaker's gaze as input from a conversation, one is already gathering gaze data from both partners and might as well include gaze overlap features for their modest performance benefits.

Discussion

Given current trajectories of technology development, we anticipate that multimodal reference resolution will be an important component of the intelligent user interfaces. In this work, we took early steps toward using dual eye-tracking (and related non-verbal context data) as input for multimodal reference resolution. Using ML models, we demonstrated the utility of an approach that combines gaze-based (including both speaker gaze and overlapping gaze), keyword-based linguistic, and spatial features to resolve reference in naturalistic conversation. Compared to baselines of 7-19% accuracy, our models were able to predict referents with over 60% accuracy, including identification of plurals. We anticipate that more highly supervised approaches with refined features will improve performance even more going forward.

An additional benefit of our approach is that it relies on data that can plausibly be automatically gathered and processed without human annotation. For example, as described in [9], we've developed a system that uses computer vision to automatically identify gaze targets and post-process dyadic gaze statistics. Although we hand-coded some of our data (e.g. position), our models only included features that could plausibly be automated given current trajectories of speech-to-text, computational linguistics, computer vision, etc.

In early pilot testing of ML models using hand-coded linguistic features rather than rule-based features, our reference resolution accuracy was between 5-10% better. We note that the keyword-based rules used in this study to code plurals, local deixis, reference shifts, etc. may not be as accurate and introduce noise into the model. If reference resolution in natural language interfaces is to use similar ML approaches, it will be beneficial to employ more sophisticated NLP to automatically code linguistic features accurately.

We also note that we didn’t address gesture, a major component of other multimodal reference models, from our analysis. Gesture's contribution has already been established, so we opted to focus on the utility other forms of non-verbal coordination as a modular problem. However, if gesture were to be used in combination with the approaches outlined here, we would expect even better multimodal reference resolution performance, particularly in identifying plurals.

References

1. Ariel, M. Referring and accessibility. *Journal of linguistics 24*, 01 (1988), 65-87.

2. Bangerter, A. Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science 15*, 6 (2004), 415.

3. Bard, E.G., Hill, R., and Arai, M. Referring and gaze alignment: Accessibility is alive and well in situated dialogue. *31st Annual Conference of the Cognitive Science Society*, (2009), 1246-1251.

4. Byron, D.K., Mampilly, T., Sharma, V., and Xu, T. Utilizing visual attention for cross-modal coreference interpretation. *Lecture Notes in Computer Science: Modeling and Using Context*, (2005), 83-96.

5. Carlson, G.N. Reference. In L.R. Horn and G.L. Ward, eds., *The handbook of pragmatics*. Wiley-Blackwell, Oxford, UK, 2005, 74-96.

6. Clark, A.T. and Gergle, D. Mobile Dual Eye-Tracking Methods: Challenges and Opportunities. *DUET 2011: Dual Eye-Tracking Workshop at ECSCW*, (2011).

7. DeVault, D. and Stone, M. Learning to interpret utterances using dialogue history. *Proceedings of EACL 09*, ACL (2009), 184-192.

8. Demšar, J., Zupan, B., Leban, G., and Curk, T. Orange: From experimental machine learning to interactive data mining. *Knowledge discovery in databases: PKDD 2004*, (2004), 537–539.

9. Gergle, D. and Clark, A.T. See What I’m Saying? Using Dyadic Mobile Eye Tracking to Study Collaborative Reference. *CSCW ’11 Conference on Computer Supported Cooperative Work*, ACM (2011), 163-172.

10. Gergle, D., Rose, C.P., and Kraut, R.E. Modeling the impact of shared visual information on collaborative reference. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI ’07*, (2007), 1543.

11. Hanna, J. and Brennan, S. Speakers’ eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language 57*, 4 (2007), 596-615.

12. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *International joint Conference on artificial intelligence*, (1995), 1137–1145.

13. Kotsiantis, S. Supervised machine learning: A review of classification techniques. *Informatica 31*, (2007), 249-268.

14. Kraut, R.E., Fussell, S.R., and Siegel, J. Visual Information as a Conversational Resource in Collaborative Physical Tasks. *Human-Computer Interaction 18*, 1 (2003), 13-49.

15. Schober, M.F. How addressees affect spatial perspective choice in dialogue. *Representation and processing of spatial expressions*, (1998), 231-245.