Generating Subsequent Reference in Shared Visual Scenes: Computation vs Re-Use

Jette Viethen
Tilburg University, The Netherlands
Macquarie University, Australia
jette.viethen@mq.edu.au

Robert Dale
Macquarie University, Australia
robert.dale@mq.edu.au

Markus Guhe
University of Edinburgh, Scotland
m.guhe@ed.ac.uk

Motivation

Content determination for referring expressions:

- Traditional algorithms for referring expression generation (REG) **compute** the attributes to best distinguish a target from a set of distractors in a deliberate way.
- Psycholinguistic accounts of reference in dialogue suggest that people **re-use** form and content of previous references from the same discourse.

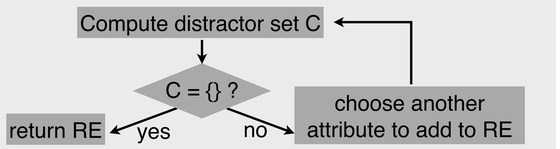
→ Which approach characterises human reference behaviour more accurately?

We use feature ablation in a machine learning approach in an attempt to answer this question.

Background

The TradREG approach: Computation

At the point where a reference is required:



(e.g.: Dale, 1989; Dale and Reiter, 1995)

Takes account of

- the visual distractors around the target
- other accessible discourse entities (Grosz and Sidner, 1986)

The Alignment approach: Re-Use

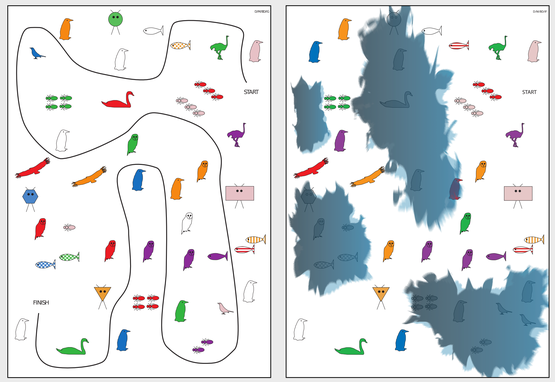
- Speakers (subconsciously) align the forms of reference they use to be similar to references that have been used before. (Pickering and Garrod, 2004)
- Once a form of reference to the intended referent has been established, they tend to re-use that form of reference, or perhaps an abbreviated version of it. (Clark and Wilkes-Gibbs, 1986)
- Speakers are more likely to use a dispreferred attribute if this attribute has recently been used by their conversational partner. (Goudbeek and Krahmer, 2010)

Takes account of

- form and content of previous references
- frequency and recency effects
- negotiation over reference

Reference in the iMap Corpus

- 32 participant-pairs:
  Instruction Giver (IG) and *Instruction Follower* (IF)
- IG describes a path, IF draws it on their map
- IG and IF maps show same landmarks with some discrepancies (ink, missing/different landmarks)



IG map

IF map

- 256 dialogues (8 per pair)
- 34,127 references to landmarks
- 16,358 referring expressions excluding plurals, pronouns and initial references

Landmark attributes:

type: bird, house, fish, car, alien, sign, bugs, trees
colour: green, blue, red, orange, purple, brown...
'other': kind (bird/house), pattern (fish/car), shape (alien/sign), number (bugs/trees)
relation: to other landmarks, the map, the path...

Semantic **Content Pattern**	count	%
⟨other⟩ – "The rectangle"	5893	36.0
⟨other, type⟩ – "the stripey fish"	3684	22.5
⟨other, colour⟩ – "the blue circle"	1630	10.0
⟨other, colour, type⟩ – "the single pink bug"	1021	6.2
⟨colour⟩ – "the red"	969	5.9
⟨relation⟩ – "the one below"	777	4.7
⟨other, relation⟩ – "the circle above"	587	3.6
⟨type⟩ – "the fish"	574	3.5
⟨colour, type⟩ – "the pink alien"	434	2.7
⟨other, relation, type⟩	312	1.9
⟨relation, type⟩	236	1.4
⟨colour, relation⟩	99	0.6
⟨other, colour, relation⟩	81	0.5
⟨other, colour, relation, type⟩	44	0.3
⟨colour, relation, type⟩	17	0.1
total	16358	

Modelling Referential Behaviour

The Task: predict the content pattern for each subsequent reference in the data set.

Each reference to a landmark can be characterised in terms of a large set of features:

- **TradREG** features
  - number of visual/discourse distractors
  - proportion of distractors with same type/colour/other
  - distance to the closest visual distractor
  - has the closest the same type/colour/other?
- **Align**ment features
  - was the speaker of the last mention the same?
  - how long ago was the last mention?
  - how long ago was type/colour/other/relation mentioned?
  - was type/colour/other/relation mentioned in the last RE?
  - how often as type/colour/other/relation been used?
  - quartile of the dialogue
  - dialogue number
  - mention number
- Theory-**Ind**ependent features
  - main map type
  - other attribute of main map type
  - other attribute of the target
  - was type/colour/other different on the IF map?
  - was the landmark missing on the speaker's map?
  - was the landmark inked-out on the IF map?
  - ID of the speaker
  - ID of the speaker pair
  - was the speaker IG or IF?

We used the C4.5 algorithm to build 4 decision trees that make binary decisions about the inclusion of the 4 attributes. The output of the 4 trees was then combined into a content pattern.

We built 7 models based on subsets of the features:

AllF: a model based on all features
Trad: using TradREG features only
Align: using Alignment features only
Ind: using Independent features only
Align+Ind: using all but the TradREG features
Trad+Ind: using all but the Alignment features
Trad+Align: using all but the Independent features

And 3 baseline models:

Head: only the attribute that is the most likely head noun
Repeat: repeat the last mention
Majority: the most common content pattern – ⟨other⟩

Evaluation Results

Acc (Accuracy): ratio of descriptions matching those in the corpus exactly in content.

Dice: coefficient of similarity between a candidate and a reference set of attributes. $DICE(A,B) = \frac{2\times|A\cap B|}{|A|+|B|}$

	col	other	type	rel	whole pattern	
	Acc	Acc	Acc	Acc	Acc	Dice
Head	–	–	–	–	23.1	0.49
Repeat	–	–	–	–	38.4	0.55
Majority	73.8	81.0	61.7	86.8	36.0	0.65
predicts	no	yes	no	no	⟨other⟩	
Trad	74.6	84.8	77.1	87.0	47.3	0.73
Align	83.6	84.1	80.7	87.5	54.6	0.78
Ind	81.9	82.8	81.4	88.0	52.7	0.78
Align+Ind	86.1	85.3	82.4	**88.7**	58.2	**0.81**
Trad+Ind	82.2	84.1	81.1	87.1	52.5	0.78
Trad+Align	84.1	84.0	80.1	86.8	53.9	0.78
AllF	**86.2**	**85.8**	**83.2**	88.5	**58.8**	**0.81**

Discussion and Conclusions

- All models significantly outperform all baselines.
- Align and Ind significantly outperform Trad.
- AllF performs best;
- but dropping TradREG does not hurt performance;
- while dropping Align or Ind *does* hurt performance.

Re-Use wins: The considerations underlying traditional computational approaches to REG do not seem to play as important a role in content selection for subsequent reference as the psycholinguistic considerations of alignment and conceptual pacts.

Error Analysis: The Trad model often choses too few attributes. This indicates that some of the over-specification often found in human reference behaviour might be due to alignment phenomena.

References

H. Clark and D. Wilkes-Gibbs (1986). Referring as a collaborative process. Cognition, 22(1):1– 39.
R. Dale (1989). Cooking up referring expressions. In Proceedings of the 27th ACL, Vancouver B.C., Canada.
R. Dale and E. Reiter (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. Cognitive Science, 19(2):233–263.
M Goudbeek and E Krahmer (2010). Preferences versus adaptation during referring expression generation. In Proc. of the 48th ACL, 55–59, Uppsala, Sweden.
B. Grosz and C Sidner (1986). Attention, intentions, and the structure of discourse. Computational Linguistics, 12(3):175–204.
M. Pickering and S. Garrod (2004). Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences, 27(2):169–226.