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

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

Jette Viethen[1,2]
jette.viethen@mq.edu.au
[1]TiCC
University of Tilburg
Tilburg, The Netherlands

Robert Dale[2]
robert.dale@mq.edu.au
[2]Centre for Language Technology
Macquarie University
Sydney, Australia

Markus Guhe[3]
m.guhe@ed.ac.uk
[3]School of Informatics
University of Edinburgh
Edinburgh, UK

Abstract

Traditional computational approaches to referring expression generation operate in a deliberate manner, choosing the attributes to be included on the basis of their ability to distinguish the intended referent from its distractors. However, work in psycholinguistics suggests that speakers align their referring expressions with those used previously in the discourse, implying less deliberate choice and more subconscious reuse. This raises the question as to which is a more accurate characterisation of what people do. Using a corpus of dialogues containing 16,358 referring expressions, we explore this question via the generation of subsequent references in shared visual scenes. We use a machine learning approach to referring expression generation and demonstrate that incorporating features that correspond to the computational tradition does not match human referring behaviour as well as using features corresponding to the process of alignment. The results support the view that the traditional model of referring expression generation that is widely assumed in work on natural language generation may not in fact be correct; our analysis may also help explain the oft-observed redundancy found in human-produced referring expressions.

1 Introduction

Computational work on referring expression generation (REG) has an extensive history, and a wide variety of algorithms have been proposed, dealing with various facets of what is recognised to be a complex problem. Almost all of this work sees the task as being concerned with choosing those attributes of an intended referent that distinguish it from the other entities with which it might be confused (see, for example, Dale (1989), Dale and Reiter (1995), Krahmer et al. (2003), van Deemter and Krahmer (2007), Gardent and Striegnitz (2007)). Independently, an alternative way of thinking about reference has arisen within the psycholinguistics community: there is now a long tradition of work that explores how a dialogue participant's forms of reference are influenced by those previously used for a given entity. Most recently, this line of work has been discussed in terms of the notions of *alignment* (Pickering and Garrod, 2004) and *conceptual pacts* (Clark and Wilkes-Gibbs, 1986; Brennan and Clark, 1996).

We suspect that neither approach tells the full story, and so we are interested in exploring whether the two perspectives should be integrated. Using a large corpus of referring expressions in task-oriented dialogues, this paper presents a machine learning approach that allows us to combine features corresponding to the two perspectives. Our results show that models based on the alignment perspective outperform models based on traditional REG considerations, as well as a number of simpler baselines.

The paper is structured as follows. In Section 2, we outline the two perspectives on subsequent reference, and summarise related work. In Section 3, we describe the iMAP Corpus and the referring expressions it contains. In Section 4, we describe the approach we take to learning models of referential behaviour using this data, and in Section 5 we discuss the results of a number of experiments based

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 1158–1167,
Edinburgh, Scotland, UK, July 27–31, 2011.

on this approach, followed by an error analysis in Section 6. Section 7 draws some conclusions and discusses future work.

2 Related Work

2.1 The Algorithmic Approach

We use the term *algorithmic approach* here to refer to the perspective that is common to the considerable body of work within computational linguistics on the problem of referring expression generation developed over the last 20 years. Much of this work takes as its starting point the characterisation of the problem expressed in (Dale, 1989). This work has focused on the design of algorithms which take into account the context of reference in order to decide what properties of an entity should be mentioned in order to distinguish that entity from others with which it might be confused. Early work was concerned with *subsequent* reference in *discourse*, inspired by Grosz and Sidner's (1986) observations on how the attentional structure of a discourse made particular referents accessible at any given point. More recently, attention has shifted to *initial* reference in *visual* domains, driven in large part by the availability of the TUNA dataset and the shared tasks that make use of it (Gatt et al., 2008). The construction of *distinguishing descriptions* has consistently been a key consideration in this body of work.

Scenarios that require the generation of references in multi-turn dialogues that concern visual scenes are likely to be among the first where we can expect computational approaches to referring expression generation to be practically useful. Surprisingly, however, the more recent work on initial reference in visual domains and the earlier work on subsequent reference in discourse remain somewhat distinct and separate from each other, despite much the same algorithms having been used in both. There is very little work that brings these two strands together by looking at both initial and subsequent references in dialogues that concern visual scenes. An exception here is the machine learning approach developed by Stoia et al. (2006), who aimed at building a dialogue system for a situated agent giving instructions in a virtual 3D world. However, their approach was concerned with choosing the *type* of reference to use (definite or indefinite, pronominal, bare or modified head noun), and not with the *content* of the reference; and their data set consisted of only 1242 referring expressions.

2.2 The Alignment Approach

Meanwhile, starting with the early work of Carroll (1980), a quite distinct strand of research in psycholinguistics has explored how a speaker's form of reference to an entity is impacted by the way that entity has been previously referred to in the discourse or dialogue. The general idea behind what we will call the *alignment approach* is that a conversational participant will often adopt the same semantic, syntactic and lexical alternatives as the other party in a dialogue. This perspective is most strongly associated with the work of Pickering and Garrod (2004). With respect to reference in particular, speakers are said to form *conceptual pacts* in their use of language (Clark and Wilkes-Gibbs, 1986; Brennan and Clark, 1996). Although there is disagreement about the exact mechanisms that enable alignment and conceptual pacts, the implication of much of this work is that one speaker introduces an entity by means of some description, and then (perhaps after some negotiation) both conversational participants share this form of reference, or a form of reference derived from it, when they subsequently refer to that entity.

Recent work by Goudbeek and Krahmer (2010) supports the view that subconscious alignment does indeed take place at the level of content selection for referring expressions. The participants in their study were more likely to use a dispreferred attribute to describe a target referent if this attribute had recently been used in a description by a confederate.

There is some work within natural language generation that attempts to model the process of alignment (Buschmeier et al., 2009; Janarthanam and Lemon, 2009), but this is predominantly concerned with what we might think of as the 'lexical perspective', focussing on lexical choice rather than the selection of appropriate semantic content for distinguishing descriptions.

2.3 Combined Models

This paper is not the first to look at how the algorithmic approach and the alignment approach might be integrated in REG. An early machine learning ap-

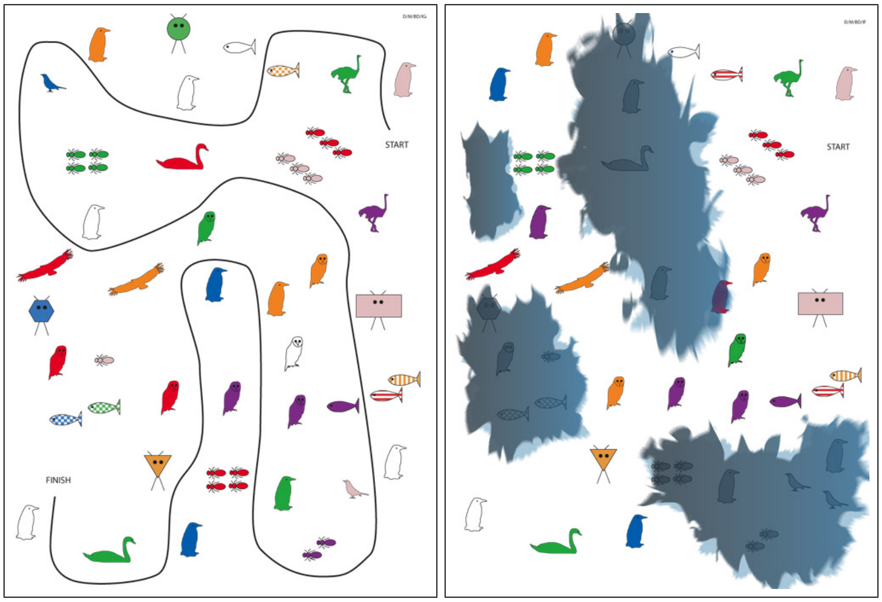

Figure 1: An example pair of maps.

proach to content selection was presented by Jordan and Walker (2000; 2005); they were also interested in an exploration of the validity of different psycholinguistic models of reference production, including Grosz and Sidner's (1986) model of discourse structure, the conceptual pacts model of Clark and colleagues, and the intentional influences model developed by Jordan (2000). However, their data set consists of only 393 referring expressions, compared to our 16,358, and these expressions had functions other than identification; most importantly, the entities referred to were not part of a shared visual scene as is the case in our data.

Gupta and Stent (2005) instantiated Dale and Reiter's (1995) Incremental Algorithm with a preference ordering that favours the attributes that were used in the previous mention of the same referent. In a second variant, they even require these attributes to be included in a subsequent reference. Differently from most other work on REG, they extended the task to include ordering of the attributes in the surface form. They therefore create a special evaluation metric that takes ordering into account, which makes it hard to compare the performance they report to that of any system that is not concerned with attribute ordering, such as ours. Their evaluation set was also considerably smaller than ours: they used 1294 and 471 referring expressions from two different corpora, compared to our test set of 4947 referring expressions.

More recently in (Viethen et al., 2010), we presented a rule-based system that addressed a specific instance of the problem we consider here, using the same corpus as we do: we singled out 2579 first references to landmarks by the second speaker ('second speaker initial references') and attempted to reproduce these using a system based on Dale and Reiter's (1995) Incremental Algorithm. Although the data set was a subset of the one used here, the system did not reach the same performance (see Section 5).

3 Referring Expressions in the iMAP Corpus

The iMAP Corpus (Louwerse et al., 2007) is a collection of 256 dialogues between 32 participant-pairs who contributed 8 dialogues each. Both participants had a map of the same environment, but one participant's map showed a route winding its way between the landmarks on the map; see Figure 1. The task was for this participant (the instruction giver, IG) to describe this route in such a way that their partner (the instruction follower, IF) could draw it onto their map; this was complicated by some discrepancies between the two maps, such as missing landmarks, the unavailability of colour in some regions due to ink stains, and small differences between some landmarks.

The landmarks differ from each other in type, colour, and one other attribute, which is different for each type of landmark. For example, there are different *kinds* of birds (eagle, ostrich, penguin . . .); fish differ by their *patterns* (dotted, checkered, plain . . .), aliens have different *shapes* (circular, hexagonal . . .), and bugs appear in small clusters of differing *numbers*. In addition to these inherent attributes of the landmarks, participants used spatial relations to other items on the map. Each referring expression in the corpus is annotated with a unique identifier corresponding to the landmark that it describes and the semantic values of the attributes that it contains. This collection of annotations forms the basic data we use in our experiments.

For each landmark R referred to in a dialogue, we view the sequence of references to this landmark as a *coreference chain*, notated $\langle R_1, R_2, \ldots, R_n \rangle$. By convention, R_1 is termed the *initial reference*, and all other references in the chain are *subsequent references*. From the corpus as a whole we extracted 34,127 referring expressions in 9558 chains. The average length of a chain is 4.74; and the longest coreference chain contains 43 references. References may be contributed to a chain by either speaker, and can be arbitrarily far apart: in the data, 4201 references are in the utterance immediately following the preceding reference in the chain, but the distance between references in a chain can be as high as 423 utterances.

We removed from the data any annotation that was not concerned with the four landmark attributes, type, colour, relation, or the landmark's other distinguishing attribute. For example, 'semantically empty' head nouns such as *thing* or *set*. Ordinal numbers that were annotated as the use of the number attribute were re-tagged as spatial relations, as these usually described the position of the target within a line of landmarks.

As a result of the removal of annotations not pertaining to the use of the four landmark attributes, 2785 referring expressions had no annotation left; we removed these instances from the final data set. We also do not attempt to replicate the remaining 5552 plural referring expressions or the 3062 pronouns found in the corpus.[1] However, we do include all of these instances in the feature extraction step, on the assumption that they might impact on the content of subsequent references. Similarly, we filter out 6369 initial references after we have extracted features from them, since we focus here on the generation of subsequent reference only. The remaining 16,358 referring expressions form the data which we use in our experiments.

Content Pattern	Count	Proportion
⟨other⟩	5893	36.0%
⟨other, type⟩	3684	22.5%
⟨other, colour⟩	1630	10.0%
⟨other, colour, type⟩	1021	6.2%
⟨colour⟩	969	5.9%
⟨relation⟩	777	4.7%
⟨other, relation⟩	587	3.6%
⟨type⟩	574	3.5%
⟨colour, type⟩	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	16,358	

Table 1: The 15 content patterns by frequency.

Contrary to findings from other corpora, in which colour was used much more frequently (Gatt, 2007; Viethen and Dale, 2008), the colour attribute was used in only 26.3% of the referring expressions in our data set. This is probably due to the often low reliability of colour in this task caused by the ink stains. The proportion of referring expressions mentioning the target's type might, at 38.7%, also seem low. This can be explained by the fact that one quarter of the landmarks, namely birds and buildings, are more likely to be described in terms of their specific kind than in terms of their generic type. This also helps explain why the overall use of the other attribute, which for some landmarks was their kind, was used in 81.0% of all instances. Spatial relations were used in 13.16% of the referring expressions, comparable to other corpora in the literature.

[1] The additional issues that arise in generating plural references and deciding when to use pronouns considerably complicate the problem; see (Gatt, 2007).

We can think of each referring expression as being a linguistic realisation of a *content pattern*: this is the collection of attributes that are used in that instance. The attributes can be derived from the property-level annotation given in the corpus. So, for example, if a particular reference appears as the noun phrase *the blue penguin*, annotated semantically as $\langle \textsf{blue}, \textsf{penguin} \rangle$, then the corresponding content pattern is $\langle \textsf{colour}, \textsf{kind} \rangle$. Our aim is to replicate the content pattern of each referring expression in the corpus. Table 1 lists the 15 content patterns that occur in our data set in order of frequency.

4 Modelling Referential Behaviour

4.1 The Two Perspectives

Our task is defined simply as follows: for each subsequent reference R in the corpus, can we predict the content pattern that will be used in that reference? As we noted at the outset of the paper, the literature would appear to suggest two distinct approaches to this problem. What we have characterised as the algorithmic approach can be summarised thus:

> At the point where a reference is required, a speaker determines the relevant features of other entities in the context, then computes the content of a referring expression which distinguishes the intended referent from the other entities.

The alignment approach, on the other hand, can be summarised thus:

> Speakers align the forms of reference they use to be similar or identical to references that have been used before. In particular, 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.

The alignment approach would appear to be preferable on the grounds of computational cost: we would expect that retrieving a previously-used referring expression, or parts thereof, generally requires less computation than building a new referring expression from scratch.

Map Features	
Main_Map_type	most frequent type of LM on this map
Main_Map_other	other attribute if the most frequent type of LM
Mixedness	are other LM types present on this map?
Ink_Orderliness	shape of the ink blot(s) on the IF's map
Lmprop Features	
other_Att	type of the other attribute of the target
*[att]*_Value	value for each *att* of target
*[att]*_Difference	was *att* of target different between the two maps?
Missing	was target missing one of the maps?
Inked_Out	was target inked]_out on the IG's map?
Speaker Features	
Dyad_ID	ID of the pair of participant-pair
Speaker_ID	ID of the person who uttered this RE
Speaker_Role	was the speaker the IG or the IF?

Table 2: The Ind feature set.

On the other hand, if the context has changed in any way, then a previously-used form of reference may no longer be effective in identifying the intended referent, and recomputation may be required.[2] This is precisely the consideration on which the initial work on referring expression generation was based, inspired by Grosz and Sidner's (1986) observations about how the changing attentional structure of a discourse moves different entities in and out of focus. However, a straightforward recomputation of reference based on the current context carries the risk that the most effective set of properties to use may change quite radically; if no account is taken of the history of previous references to the entity, it's conceivable that one could produce a description that is so different from the previous description that they are virtually unrecogisable as descriptions of the same entity. Ideally, what we want to do is *modify* a previous description to do the job.

These observations suggest that, in order to choose the most appropriate form of reference for an entity, we need to simultaneously take account of:

- the other entities from which it must be distinguished, both in the visual context and in the preceding discourse (in other words, exactly the information that traditional algorithmic approaches consider);
- how this entity, and perhaps other entities, have been referred to in the past (precisely the information that the alignment approach considers).

[2] Unfortunately, determining what counts as a change of context, especially in visual scenes, is fraught with difficulty.

TradREG Features (Visual)	
Count_Vis_Distractors	number of visual distractors
Prop_Vis_Same_*[att]*	proportion of visual distractors with same *att*
Dist_Closest	distance to the closest visual distractor
Closest_Same_*[att]*	has the closest distractor the same *att*?
Dist_Closest_Same_*[att]*	distance to the closest distractor of same *att* as target
Cl_Same_type_Same_*[att]*	has the closest distractor of the same type also the same *att*?
TradREG Features (Discourse)	
Count_Intervening_LMs	number of other LMs mentioned since the last mention of the target
Prop_Intervening_*[att]*	proportion of intervening LMs for which *att* was used AND which have the same *att* as target

Table 3: The TradREG feature set.

Alignment Features (Recency)	
Last_Men_Speaker_Same	who made the last mention of target?
Last_Mention_*[att]*	was *att* used in the last mention of target?
Dist_Last_Mention_Utts	distance to the last mention of target in utterances
Dist_Last_Mention_REs	distance to the last mention of target in REs
Dist_Last_*[att]*_LM_Utts	distance in utterances to last use of *att* for target
Dist_Last_*[att]*_LM_REs	distance in REs to last use of *att* for target
Dist_Last_*[att]*_Dial_Utts	distance in utterances to last use of *att*
Dist_Last_*[att]*_Dial_REs	distance in REs to last use of *att*
Dist_Last_RE_Utts	distance to last RE in utterances
Last_RE_*[att]*	was *att* mentioned in the last RE?
Alignment Features (Frequency)	
Count_*[att]*_Dial	how often has *att* been used in the dialogue?
Count_*[att]*_LM	how often has *att* been used for target?
Quartile	quartile of the dialogue the RE was uttered in
Dial_No	number of dialogues already completed +1
Mention_No	number of previous mentions of target +1

Table 4: The Alignment feature set.

The set of features we describe next attempts to capture these two aspects of the problem.

4.2 Features

The number of factors that can be hypothesised as having an impact on the form of a referring expression in a dialogic setting associated with a visual domain is very large. Attempting to incorporate all of these factors into parameters for rule-based systems, and then experimenting with different settings for these parameters, is prohibitively complex. Instead, we here capture a wide range of factors as features that can be used by a machine learning algorithm to automatically induce from the data a classifier that predicts for a given set of features the attributes that should be used in a referring expression.

The features we extracted from the data set are listed in Tables 2–4.[3] They fall into five subsets. **Map** Features capture design characteristics of the maps the current dialogue is about; **Speaker** Features capture the identity and role of the participants; and **LMprop** Features capture the inherent visual properties of the target referent. For our experiments, we group the Map, LMprop and Speaker feature sets into one theory-independent set (**Ind**). Most importantly for our present considerations, **TradREG** Features capture factors that the traditional computational approaches to referring expression generation take account of, in particular properties of the discourse and visual distractors; and **Alignment** Features capture factors that we would expect to play a role in the psycholinguistic models of alignment and conceptual pacts.

4.3 The Models

For the experiments described here, we used a 70–30 split to divide the data into a training set (11,411 instances) and a test set (4,947 instances). In addition to the main prediction class *content pattern*, the split was stratified for Speaker_ID and Quartile to ensure that training and test set contained the same proportion of descriptions from each speaker and each quartile of the dialogues. We used the J48 algorithm implemented in the Weka toolkit (Witten and Frank, 2005) to train decision trees with the task of judging, based on the given features, which content pattern should be used.

First, we have three separate baseline models:

HeadNounOnly generates only the property that is the most likely head noun for the target, which is kind for birds and buildings and type for all

[3] In these tables, *att* is an abbreviatory variable that is instantiated once for each of the four attributes type, colour, relation, and the other distinguishing attribute of the landmark. The abbreviation LM stands for landmark

other landmarks. This is a form of 'reduced reference' strategy.

RepeatLast represents a very simplistic alignment approach. It generates the same content pattern that was used in the previous mention of the target referent.

MajorityClass generates the content pattern most commonly used in the training set.

We then have a number of models that use subsets of the features described above:

AllFeatures is a decision tree trained on all features;

TradREG is a decision tree trained on the TradREG features only;

Alignment is a decision tree trained on the Alignment features only;

Ind is a decision tree trained on the Ind features only;

Alignment+Ind is a decision tree trained on all but the TradREG features;

TradREG+Ind is a decision tree trained on all but the Alignment features; and

TradREG+Alignment is a decision tree trained on all but the Ind features.

5 Results

In this section we report how the models described in the previous section performed on the held-out test set in comparison to each other and to the three baselines.

We use Accuracy and average DICE score for our comparisons; these are the most commonly used measures in the REG literature (see, for example, Gatt et al., 2008). Given two sets of attributes, A and B, DICE is computed as

$$\text{DICE} = \frac{2 \times |A \cap B|}{|A| + |B|}. \tag{1}$$

This gives some measure of the overlap between two referring expressions, assigning a partial score if the two sets share attributes but are not identical. The Accuracy of a system is the proportion of test instances for which it achieves a DICE score of 1, signifying a perfect match.

	col	other	type	rel	Comb. Pattern	
	Acc	Acc	Acc	Acc	Acc	DICE
HeadOnly	n/a	n/a	n/a	n/a	23.1	0.49
RepLast	n/a	n/a	n/a	n/a	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
AllFeatures	**86.2**	**85.8**	**83.2**	88.5	**58.8**	**0.81**

Table 5: Performance of our models compared to the baselines. Model names are abbreviated for space reasons. The Accuracy (given in %) of all models is significantly better than that of the highest performing baseline at p<.01 according to the χ^2 statistic.

We tested two different ways of generating content patterns based on the different feature sets described above: **PatternAtOnce** builds a decision tree that chooses one of the 15 content patterns that occur in our data set; whereas **CombinedPattern** builds attribute-specific decision trees (one for each of the four attributes that occur in the data: colour, other, type, and relation), and then combines their predictions into a complete content pattern. We found that CombinedPattern slightly outperformed PatternAtOnce, although the difference is not statistically significant for all feature sets. For space reasons, we report in what follows only on the slightly better-performing CombinedPattern model.

Table 5 compares the performances of the three baselines and the decision trees based on the five feature subsets for each of the individual attributes and for the combined content pattern; note that the HeadNounOnly and RepeatLast baselines do not make attribute-specific predictions. The table shows that the learned systems outperform all three baselines for the individual attributes as well as for the combined content pattern.

A comparison of the Alignment feature set and the TradREG feature set shows that the former outperforms the latter for the attribute-specific trees which predict the use of the colour attribute and the use of relation, and that the combined patterns resulting from the Alignment trees are a better match of the human-produced patterns both in terms of Accuracy (p<.01 for all three categories, using χ^2) and DICE. Interestingly, even the theory-independent Ind features outperform the TradREG features.

The comparison between TradREG+Ind and Alignment+Ind again shows a clear advantage for the Alignment features: dropping them from the complete feature set significantly hurts performance compared to AllFeatures (χ^2=80.5, p<.01), while dropping the TradREG features has no significant impact. Also consistent with the results of the three individual feature sets, dropping the Ind features hurts performance more than dropping the TradREG features, but less than dropping the Alignment features. Training on the complete feature set (AllFeatures) achieves the highest performance, which is significantly better than that of all other features sets (p<.01 using χ^2) except Alignment+Ind.

These results suggest that considerations at the heart of traditional REG approaches do not play as important a role as those postulated by alignment-based models for the selection of semantic content for subsequent referring expressions.

We also note that the Accuracy scores achieved by our learned systems are similar to the best numbers previously reported in the REG literature. While Jordan and Walker's (2005) data set is not directly comparable, they achieved a maximum of 59.9% Accuracy, against our 58.8%. Stoia et al.'s (2006) best Accuracy was 31.2%, albeit on a slightly different task. Even in the arguably much simpler non-dialogic domains of the REG competitions concerned with pure content selection, the best performing system achieved only 53% Accuracy (see Gatt et al., 2008). The most comparable approach, the rule-based system we presented in (Viethen et al., 2010) for a subset of the data used here, was not able to outperform a RepeatLast baseline at 40.2% Accuracy and an average DICE score of 0.67.

6 Error Analysis

An important question to ask is how wrong the models really are when they do not succeed in perfectly matching a human-produced reference in our test set. It might be that they choose a completely different set of attributes from those included by the human speaker; however, the Accuracy score also counts as incorrect any set that only partly overlaps with the reference found in the test set.

	Acc	Dice	Super	Sub	Inter	Noover
Trad	47.3	0.75	14.4	22.2	5.5	10.5
Align	54.6	0.78	16.0	16.1	3.9	9.4
Ind	52.7	0.78	17.1	17.2	3.9	9.0
Align+Ind	58.2	0.81	16.0	14.8	3.1	7.9
Trad+Ind	52.5	0.78	17.4	17.5	3.8	8.8
Trad+Align	53.9	0.78	17.1	15.6	4.3	9.0
AllFeature	58.8	0.81	16.5	14.5	3.1	7.2

Table 6: The proportions of test instances for which each model produced a subset, a superset, some other form of intersection or no-overlap to the human reference.

The DICE score gives us a partial answer to this question, as it assigns a score that is based on the size of the overlap between the attribute set chosen by the model and that included by the human speaker. A DICE score that is equal to the Accuracy score would mean that each referring expression was either reproduced perfectly, or that a set of attributes was chosen that did not overlap with the original one at all. The fact that all our models achieved DICE scores much higher than their Accuracy scores shows that they only rarely got it completely wrong.

Table 6 gives a more fine-grained picture by listing, for each model, what percentage of the referring expressions it produced contained a subset of the attributes included in the human reference, what percentage were a superset, what percentage had another form of partial intersection, and what percentage had no commonality with the human reference. Interestingly, a large number of the referring expressions produced by the model trained only on TradREG features are subsets of the human reference. This indicates that human speakers tend to include more attributes than are strictly speaking necessary to distinguish the landmark.[4] The Alignment model does not as often produce a subset of the gold standard content pattern, suggesting that it might be alignment considerations that account for some of

[4] That humans often produce 'redundant' descriptions, in opposition to the target behaviour of some of the early REG algorithms, is of course an oft-observed fact.

	both	both	1st	2nd	either	pot.
	corr.	wrong	corr.	corr.	corr.	Acc
Trad vs Ind	1797	1794	545	811	3153	63.7
Trad vs Align	1742	1647	600	958	3300	66.7
Trad vs Align+Ind	1849	1574	493	1031	3373	68.2
Align vs Trad+Ind	1908	1557	792	690	3390	68.5
Align vs Ind	1872	1511	828	736	3436	69.5
Ind vs Trad+Align	1840	1511	768	828	3436	69.5

Table 7: Comparison of the predictions for the combined content pattern between the models trained on mutually exclusive feature sets.

the apparent redundancy that human-produced referring expressions contain.

A second important question is whether the different feature sets are doing the same work, or whether they complement each other. Table 7 lists for those pairings of our learned models which were based on mutually exclusive feature sets how many referring expressions both models predicted correctly, how many both failed to predict, and how many were predicted correctly by either of the two models.

Note the high numbers in the columns listing the counts of instances which both models got either correct or wrong: these show that there is considerable overlap between all pairings. The smallest agreement lies at 3424 instances (68.2%) between TradREG (the least successful model) and Alignment+Ind (the most successful model). However, they also each predict correct solutions that the other misses: 493 (10.0%) for TradREG and 1031 (20.8%) for Alignment+Ind.

The last two columns of Table 7 show the number of instances that at least one of the two models in each pairing got correct and the proportion out of all test instances that this number represents. This proportion is the maximum Accuracy that could be achieved by a model that combines the two models in a pairing and then correctly chooses which one to use in each instance. The maximum Accuracies that could be achieved in this way on our data set lie just below 70%, significantly higher than any numbers reported in the literature on the task of generating subsequent reference.

7 Conclusions

Using the largest corpus of referring expressions to date, we have shown how both the traditional computational view of REG and the alternative psycholinguistic alignment approach can be captured via a large set of features for machine learning. Additionally, we defined a number of theory independent features. Using this approach we have presented three main findings.

First, we have demonstrated that a model using all these features to predict content patterns in subsequent references in shared visual scenes delivers an Accuracy of 58.8% and a DICE score of 0.81, outperforming models based only on features inspired by one of the two approaches. However, we found that the features based on traditional REG considerations do not contribute as much to this score as those based on the alignment approach, and that dropping the traditional REG features does not significantly hurt the performance of a model based on alignment and theory-independent features.

Second, our error analysis showed that the main reason for the low performance of a model based on traditional algorithmic features is that it often chooses too few attributes. The fact that the model based on the alignment features does not make this mistake so frequently suggests that it may be the psycholinguistic considerations incorporated in our alignment features that lead people to add those additional attributes.

Finally, while the different models make the same correct predictions about the content of referring expressions in many cases, there are also a considerable number of cases where the models based on either the traditional algorithmic features (10.0%) or the alignment and independent features (20.8%) alone make correct predictions that the other gets wrong; this suggests that a system with the ability to choose the correct model in each of those cases (perhaps based on a hypothesis as to whether or not the relevant context has changed) could reach an accuracy of almost 70% on our data set. In future work we plan to identify further features that will allow us to inform this choice so that we can move towards this level of performance.

References

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1482–1493.

Hendrik Buschmeier, Kirsten Bergmann, and Stefan Kopp. 2009. An alignment-capable microplanner for natural language generation. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 82–89, Athens, Greece.

John M. Carroll. 1980. Naming and describing in social communication. *Language and Speech*, 23:309–322.

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

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Robert Dale. 1989. Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver B.C., Canada.

Claire Gardent and Kristina Striegnitz. 2007. Generating bridging definite descriptions. In Harry C. Bunt and Reinhard Muskens, editors, *Computing Meaning*, volume 3, pages 369–396. Kluwer, Dordrecht, The Netherlands.

Albert Gatt, Anja Belz, and Eric Kow. 2008. The TUNA Challenge 2008: Overview and evaluation results. In *Proceedings of the 5th International Conference on Natural Language Generation*, pages 198–206, Salt Fork OH, USA.

Albert Gatt. 2007. *Generating Coherent Reference to Multiple Entities*. Ph.D. thesis, University of Aberdeen, UK.

Martijn Goudbeek and Emiel Krahmer. 2010. Preferences versus adaptation during referring expression generation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 55–59, Uppsala, Sweden.

Barbara J. Grosz and Candance L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Surabhi Gupta and Amanda Stent. 2005. Automatic evaluation of referring expression generation using corpora. In *Proceedings of the Workshop on Using Corpora for Natural Language Generation*, pages 1–6, Brighton, UK.

Srinivasan Janarthanam and Oliver Lemon. 2009. Learning lexical alignment policies for generating referring expressions for spoken dialogue systems. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 74–81, Athens, Greece, March. Association for Computational Linguistics.

Pamela W. Jordan and Marilyn Walker. 2000. Learning attribute selections for non-pronominal expressions. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, Hong Kong, China.

Pamela W. Jordan and Marilyn Walker. 2005. Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, 24:157–194.

Pamela W. Jordan. 2000. *Intentional Influences on Object Redescriptions in Dialogue: Evidence from an Empirical Study*. Ph.D. thesis, University of Pittsburgh, Pittsburgh PA, USA.

Emiel Krahmer, Sebastiaan van Erk, and André Verleg. 2003. Graph-based generation of referring expressions. *Computational Lingustics*, 29(1):53–72.

Max M. Louwerse, Nick Benesh, Mohammed E. Hoque, Patrick Jeuniaux, Gwyneth Lewis, Jie Wu, and Megan Zirnstein. 2007. Multimodal communication in face-to-face computer-mediated conversations. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 1235–1240.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–226.

Laura Stoia, Darla M. Shockley, Donna K. Byron, and Eric Fosler-Lussier. 2006. Noun phrase generation for situated dialogs. In *Proceedings of the 4th International Conference on Natural Language Generation*, pages 81–88, Sydney, Australia, July.

Kees van Deemter and Emiel Krahmer. 2007. Graphs and Booleans: On the generation of referring expressions. In Harry C. Bunt and Reinhard Muskens, editors, *Computing Meaning*, volume 3, pages 397–422. Kluwer, Dordrecht, The Netherlands.

Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring expression generation. In *Proceedings of the 5th International Conference on Natural Language Generation*, pages 59–67, Salt Fork OH, USA.

Jette Viethen, Simon Zwarts, Robert Dale, and Markus Guhe. 2010. Dialogue reference in a visual domain. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, Valetta, Malta.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco CA, USA.