On "redundancy" in selecting attributes for generating referring expressions

Philipp Spanger **Kurosawa Takehiro** **Tokunaga Takenobu**

Department of Computer Science
Tokyo Institute of Technology
Tokyo Meguro Ôokayama 2-12-1, 152-8550 Japan
{philipp, kurosawa, take}@cl.cs.titech.ac.jp

Abstract

We seek to develop an efficient algorithm selecting attributes that approximates human selection. In contrast to previous work we sought to combine the strengths of cognitive theories and simple learning algorithms. We then developed a new algorithm for attribute selection based on observations from a corpus, which outperformed a simple base algorithm by a significant margin. We then carried out a detailed comparison between our algorithm and Reiter & Dale's "Incremental Algorithm". In terms of achieving a *human-like* attribute selection, the overall performance of both algorithms is fundamentally equivalent, while differing in the handling of *redundancy* in selected attributes. We further investigated this phenomenon and draw some conclusions for further improvement of attribute-selection algorithms.

1 Introduction

Referring expressions are a key research area in human-agent communication. In the generation of referring expressions humans do not necessarily produce the most effective (i.e. minimal) expressions in a computational sense. Given evolutionary development of human linguistic capabilities, we can assume that human-produced expressions are generally optimal to identify a target for other human subjects. Thus the generation of human-like referring expressions is an important task as the generation of those expressions that are most easily understandable for humans.

The seminal work in this field is the "Incremental algorithm" (IA) (Dale and Reiter, 1995). Their work is based on an analysis of the overall cognitive tendencies of humans in the selection of attributes. In recent years, there have been a number of important extensions to this algorithm, dealing with very specific problems. This need for a systematic approach and unified evaluation of those vastly differing algorithms provided the motivation for the creation of the TUNA-corpus[1] that was developed at Aberdeen University as part of the TUNA project (van Deemter, 2007). Work has begun to use this corpus for evaluating different algorithms for attribute selection.

Our research is carried out within this general trend, seeking to take advantage of common resources (e.g. TUNA-corpus). A critical question is how to combine the generic human cognitive tendencies and the dependency of attribute selection on a specific distribution of attributes in a specific case. In this research we tackle this question in a corpus-based approach. Specifically, in a given environment, we seek to develop an efficient algorithm for selection of attributes that approximates human selection.

2 The corpus

We utilized a simplified version of the TUNA-corpus, which was also the basis for the GRE-challenge held as part of the UCNLG+MT workshop in 2007 (Belz and Gatt, 2007). The corpus consists of a collection of paired pictures of objects and human-produced referring expressions annotated with attribute sets. Figure 1 shows an image

[1] TUNA-corpus: www.csd.abdn.ac.uk/research/tuna

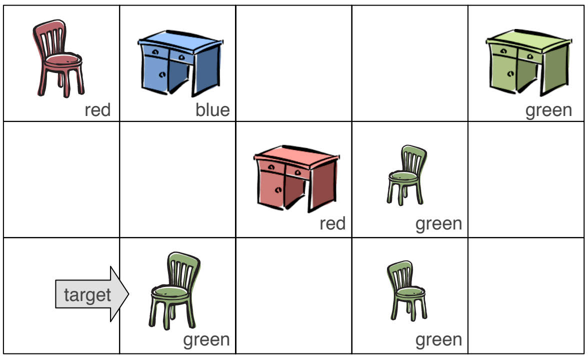

Figure 1: Image of a TUNA-corpus picture

of such a case[2]. This corpus provides information on the attribute-value pairs of the target and the distractors as well as of the referring expressions humans produced. Every item in our corpus consists of an input part ("case") and an output part ("description"). Each individual case consists of seven case entities: one target referent and six distractors. Every entity consists of a set of attribute-value pairs and all descriptions consist of a subset of the attribute-value pairs of the target referent in the same format as any entity. This corpus comprises two domains: a "Furniture" and a "Person" - domain. We note that within the corpus there were some cases that we judged as inappropriate for this study and thus excluded from the overall evaluation. This included cases where attribute-values were unspecified and/or inconsistent.

3 The base algorithm

We developed a base algorithm as a baseline for evaluation. We define "discriminative power" of a specific attribute as the number of entities in the case that have a different value from the target for this attribute.

We add attributes in descending order of discriminative power until the target can be identified uniquely. The generated attribute set is the output.

Every time an attribute is selected, we recalculate the discriminative power of the attributes of exclusively those distractors that could not be excluded by this stage.

4 Analysis of human-produced referring expressions

Our hypothesis is that in human generation of referring expressions, a combination of generic cognitive factors as well as case-dependent factors have to be dealt with. In order to account for the cognitive factor, we define a "selection probability" over a whole domain (i.e. independent from a specific case) and calculate the differences of this selection probability over the different attributes. We define the selection probability of a specific attribute a in a specific domain as equation (1).

$$SP(a) = \frac{C(a)}{\sum_{x \in X} C(x)} \tag{1}$$

where $C(x)$ denotes the number of occurrences of attribute x in the corpus.

We observe that in the Furniture-domain the attributes *colour* and *type* have extraordinarily high selection probabilities and in particular the attribute *type* is selected virtually unconditionally. We observe the same tendency of a very high selection probability for the attribute *type* in the Person-domain, even though all distractors as well as the target are of same type "person". Since the attribute *type* becomes the head of the noun phrase in the linguistic realisation of a referring expression, it is natural to mention the type. Overall, we can conclude that the different values for the selection probabilities reflect the cognitive load humans assign different attributes in a given domain.

4.1 Co-occurrence of attributes

We hypothesize that the selection of attributes is limited by co-occurence - dependencies between attributes.

In order to measure this degree of co-occurrence, we defined a "degree of dependency" between attributes as in equation (2). If the degree of dependency approaches 1, there is practically no dependency in the occurrence of attributes a and b. If this factor grows above 1, the two attributes easily occur jointly in the referring expression, on the other hand, the further it decreases below 1, the less likely are the two attributes to occur jointly. In the equation $\mathrm{P}(a,b)$ is the probability that the two attributes will be selected together, $\mathrm{P}(x)$ is the probability that the attribute x will be selected. $D(a,b)$ is the degree of dependency between attributes

$$D(a,b) = \frac{\mathrm{P}(a,b)}{\mathrm{P}(a) \times \mathrm{P}(b)} \tag{2}$$

We observed that in the Furniture-domain, size or orientation and dimension are less likely to oc-

[2] Actual pictures in the TUNA-corpus do neither show colour labels nor a target-marker.

cur together in a referring expression. Furthermore, in the Person - domain, *hairColour* and *hasHair* or *hasBeard* have a high degree of dependency, i.e. they likely occur together.

4.2 Redundancy of attributes

Even though in many referring expressions unique identification with few attributes is possible, humans show a tendency to add “redundant” attributes, i.e. that are in a strict sense not necessary for identification. By adding redundancy, humans add robustness to the expression as well as possibly reducing the cognitive load for humans in a specific context. Within the corpus, we counted the number of expressions containing redundancy. In the Furniture-domain there were 220 out of all 278 expressions and in the Person-domain there were 213 out of 230.

Table 1: Number of selected redundant attributes

Furniture (278 cases)		Person (230 cases)	
attribute	occurrences	attribute	occurrences
colour	110	*type*	201
orientation	15	*x-dimension*	4
size	10	*hasBeard*	42
type	210	*hasGlasses*	41
x-dimension	18	*hasHair*	32

This level of redundancy indicates that in order to produce human-like sets of attributes for the generation of referring expressions, it is not necessary to aim for a minimal set.

5 Our proposed algorithm for effective attribute selection

Based on our analysis of co-occurrence and redundancy of attributes, we centrally implemented the following improvements of the base algorithm.

Co-occurrence Based on the results from section 4.1, when a certain attribute is selected, we raise the selection probabilities of those attributes that have a tendency to co-occur with it, on the other hand we lower the selection probabilities of those attributes that have a tendency not to cooccur with this attribute.

Redundancy Based on the results in section 4.2, having selected the attributes to uniquely determine the target, we add the next candidate in the list of attributes as a redundant attribute .

Combination We combine both individual improvements. First of all, we add the type-attribute and then score the result based on the selection probability. With each selection of a specific attribute, we change the scores based on co-occurrence, and at the end we add a redundant attribute.

6 Evaluation of proposed algorithm

We measured the proximity of the sets of attributes by our system to the human-produced set of attributes. We utilize the Dice-coefficient (DC) – a measure of proximity for sets. For purposes of comparison, we implemented a version of the Incremental algorithm, where we calculated the order of selection of attributes according to the selection probabilities of attributes in the overall domain (Furniture or Person). It is of note that our algorithm (combination of all individual improvements) performs almost equivalent to the IA.

Table 2: Average DC for key improvements

	Furniture	Person
Base algorithm	0.305	0.314
Base+selection probability	0.784	0.669
Base+co-occurrence	0.254	0.314
Base+redundancy	0.401	0.341
Combination	0.811	0.703
Incremental algorithm	0.811	0.705

6.1 Comparison with Incremental Algorithm

We carried out a detailed analysis of the results of our algorithm and those of the IA. We found that the results of both algorithms in the Furniture - domain are exactly the same; however the results of the Person - domain show significant differences. Thus we concentrate on further analysis of the results in the Person - domain.

We divided all cases from the Person - domain into three sets; a set of cases where our algorithm performs better than the IA (sys-superior cases: 27 cases), a set of cases where the opposite is true (IA-superior cases: 24 cases) and a set of tie cases. We then compared the first two sets.

Investigating these sets, we observed that the key difference between these two algorithms lay in the treatment of redundancy. The IA often fails in the case where humans use fewer attributes and add only *type* as redundant attribute. On the other hand, our algorithm fails in the case where humans use more complex expressions, that is, more attributes including several redundant ones.

We investigated the redundant attributes which are selected by humans but not by the algorithms.

In the IA-superior cases, our system fails to select the *hasBeard* attribute compared with the IA in 20 out of 24 cases, while in the sys-superior cases both algorithms fail to select almost the same redundant attributes. We investigated for both algorithms, which attributes the algorithms wrongly select; i.e. which are not selected by humans. In the sys-superior cases, the IA wrongly selects attributes in all 27 cases, with 23 out of those including the wrongly-selected *hasBeard* attribute. In the IA-superior cases, the number of cases with wrongly selected attributes is much smaller (9 cases for each) and they are largely equiavalent.

Thus, our detailed analysis showed an overall opposite tendency in one attribute; *hasBeard*. While in sys-superior cases about 85% of the cases in which the IA output wrong attributes included *hasBeard*, in IA-superior cases our system failed to select exactly *hasBeard* at a largely equivalent rate (about 83%). At this moment, we do not have any reasonable explanation for this peculiarity of *hasBeard*, but suspect it might possibly be related to the characteristics of the corpus.

However, from the overall observation that our algorithm achieved an equivalent level of human-likeness to the IA while being weaker in cases of more complex redundancy, we conclude that further improvement in selecting redundant attributes is crucial to outperform the IA.

7 Concluding Remarks

Based on observations from the TUNA-corpus, we developed an algorithm for attribute-selection modeling human referring expressions. Our corpus-based algorithm sought to combine human generic tendencies of attribute selection in a certain domain with case-dependent variation of the salience of specific attributes. Our improved algorithm outperformed the base algorithm by a significant margin. However, we got qualitatively equivalent results to our implementation of the IA.

A detailed analysis of the characteristics of our algorithm in comparison to the IA pointed to the importance of the phenomenon of redundancy as possibly a central aspect that needs to be further investigated to achieve a qualitative improvement over the IA.

Our investigations into redundancy show that in those cases where our algorithm outperformed the IA, our algorithm almost exclusively added solely the *type*-attribute. In contrast in more complex cases of redundancy in referring expressions, the IA has shown to be superior. Since we achieved overall parity to the IA even though generally performing worse than the IA in cases of more complex redundancy, we can conclude that outside of this phenomenon our algorithm performs better than the IA in terms of human-likeness.

In previous research there has been some discussion on "redundancy" vs. "minimality" in referring expressions (e.g. (Viethen and Dale, 2006)). Through our research we have identified the phenomenon of redundancy as a critical topic for further research and for achieving further progress in the generation of human-like referring expressions.

Our algorithm includes some strong simplifications, e.g. our treatment of attributes did not take account of the fact that attribute-values are also of different type and did not explore what implications this has for the process of producing referring expressions; binary (*hasHair*), discrete (*hairColour*) or graded (*x-dim*). In future these factors should be integrated into attribute selection algorithms.

In future work, we will seek to provide a more detailed investigation of the phenomenon of redundancy, including its variation over different domains. Such an analysis should also contribute to further our understanding of the human cognitive process in the selection of attributes for the generation of referring expressions.

References

Belz, Anja and Albert Gatt. 2007. The attribute selection for GRE challenge: Overview and evaluation results. In *Proceedings of the MT Summit XI Workshop Using Corpora for Natural Language Generation: Language Generation and Machine Translation (UCNLG+MT)*, pages 75–83.

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

van Deemter, Kees. 2007. TUNA: Towards a unified algorithm for the generation of referring expressions - Final Report -. http://www.csd.abdn.ac.uk/research/tuna/pubs/TUNA-final-report.pdf.

Viethen, Jette and Robert Dale. 2006. Algorithms for generating referring expressions: Do they do what people do? In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 63–70.