TITCH: Attribute selection based on discrimination power and frequency

Philipp Spanger philipp@cl.cs.titech.ac.jp
Kurosawa Takahiro kurosawa.t.aa@m.titech.ac.jp
Tokunaga Takenobu take@cl.cs.titech.ac.jp

Department of Computer Science
Tokyo Institute of Technology

Abstract

We provide a simple base algorithm for the given task of attribute selection as well as two improvements to this algorithm. We report on the results of their implementation and provide an error analysis. We then report some observations and conclusions about the human attribute selection process by comparing the output of our algorithms and the attributes selected by humans. Finally, we add some observations on the difficulty of the attribute selection task.
Note: We utilised both training and development data for the development and evaluation of our algorithm.

1 Base algorithm

We begin from the observation that humans tend to select the *type* attribute in virtually all cases in both domains, with the exception of circumstances where the type of any object that could be referred to is obvious and agreed upon by subjects. Hence, we always include the *type* attribute in attribute selection.

In selecting attributes other than *type*, at least two factors should be taken into account. The first factor is a human's general preference on object attributes, which would be related to cognitive load in recognising the attributes (Dale and Reiter, 1995). For instance, humans generally tend to refer to an object's colour rather than its orientation in 3-dimensional space.

However, it is obvious that the salience of a certain attribute depends on the case representing a particular situation in a domain. For instance, in a case where the target as well as almost all distractors have the same colour but largely different size, the attribute *size* becomes a critical attribute to be selected while colour might be much more salient in other cases. A critical question for research is how to combine these two factors: the case independent generic factor and the case dependent factor.

We first singled out the case dependence of the salience of a specific attribute and propose a simple method to calculate this salience. Our basic idea considers the discrimination power of an attribute-value pair as the case dependent salience. Given a certain case, attribute-value pairs of the target object are ranked according to their discrimination power, and they are selected one by one according to the ranking until the set of attribute-value pairs uniquely identifies the target object.

The discrimination power of an attribute-value pair is defined as the number of objects excluded by specifying that attribute-value pair; i.e, the fewer objects that share a certain attribute-value pair, the more discrimination power this attribute-value pair possesses.

In the ranking process, we have two options depending on if we regard discriminative power as static or dynamic. In the dynamic interpretation, each time an attribute-value pair is selected, we recalculate the number of objects so that the objects without the already selected attribute-value pairs are excluded from counting. In contrast, in the static interpretation, we calculate discrimination power once at the beginning.

2 Evaluation

We first evaluated the output of our algorithm by using the provided implementation of the Dice coefficient calculator. The result is shown in Table 1.

Table 1: Result of base algorithm (TITCH-BS)

Domain	Dice	
	static	dynamic
Furniture	0.588	0.601
People	0.559	0.559

As Table 1 indicates, the difference between static and dynamic variations of the discrimination power is very subtle; slightly increasing the Dice coefficient in the furniture domain, but without effect in the people domain. This tendency is observed throughout the rest of the experiments. Thus, in what follows, we concentrate on the results with the static discrimination power, although the figures of the dynamic version are also shown in the tables for reference.

As described in the previous section, our algorithm stops selecting attributes immediately after the selected attributes identify the target object uniquely. In this respect, our algorithm realises *full brevity* of referring expressions introduced by (Dale and Reiter, 1995).

Comparing the output of our algorithm and human selection, humans tend to select more attributes than our algorithm, i.e. humans produce redundant attribute sets. This corresponds to the observation by Dale and Reiter of the need to approximate full brevity. We observed that the redundant attributes by humans depend on the case. In the "Furniture" domain, our algorithm's output tends to lack the *colour* attribute in comparison with the human selection. Thus, we can presume that the attribute *colour* is particularly salient to human subjects in this domain. On the other hand, in the "People" domain, attributes specifying features of the human face are particularly salient (e.g. *hasGlasses* or *hasHair*). Table 2 shows frequently missing attributes in the output of our algorithm (static version) in comparison to human attribute selection. Their frequency is shown in percentage of the total number of missing attributes.

Table 2: Frequently missing attributes

Furniture		People	
Attribute	(%)	Attribute	(%)
colour	44.6	y-dimension	18.6
size	16.4	hasBeard	18.4
y-dimension	15.8	hasGlasses	16.8
orientation	15.4	hairColour	14.5
		x-dimension	11.0
		hasHair	9.9

The conclusion from the results is that besides an obvious case dependence of attribute selection, there is also an inherent preference in selecting attributes for referring expression generation. Namely, humans tend to use more attributes to produce referring expressions than the necessity minimum attribute set. We presume this difference comes from case independent nature of object attributes which is related to cognitive load to recognise the attributes. In the next section, we extend the base algorithm to include the cognitive aspect of attributes for humans, which would have a case independent nature.

3 Improvement of the base algorithm

Based on the above observations, we sought to test and compare different ways of implementing the case independent property of human preference in attribute selection. As an indirect indication of this property, we chose to use the frequency of the respective attributes in the data of human attribute selection. Of course, it is difficult to learn a human's perceptual tendency directly from the data. We presume that we can estimate it from the frequency of attributes mentioned by humans. An assumption behind this is that salient attributes tend to be mentioned frequently. That leads us to the idea to put weights on attributes based on the frequency of occurrence in the data. We then implemented two different ways of weighting attributes.

3.1 Absolute attribute weighting

The first improvement assigns weights to attributes in proportion to their frequency as selected by human subjects and integrates this weighting into the base algorithm. For instance, if the attribute a_i is used f_{a_i} times in the human data of a domain, the discrimination power of a_i with its value is multiplied by f_{a_i} in the ranking of attributes by the base algorithm. An implementation of this algorithm yielded the results as shown in Table 3.

Table 3: Results of absolute weighting (TITCH-AW)

Domain	Dice	
	static	dynamic
Furniture	0.685	0.685
People	0.651	0.651
People+	0.683	0.683

We note that in both domains, this weighting yielded an increase in the Dice coefficient by about 0.1, which confirms the conclusions drawn from the result of the base algorithm. Namely, both the case dependency and independency of attribute selection have to be accounted for.

In the "People" domain, there is a dependency between the attribute *hairColour* and both *hasHair* and *hasBeard*. That is, having *hairColour* logically entails having *hasHair* or *hasBeard*. In Table 3, the row "People+" indicates the results by a modified algorithm which takes this dependency into account. The modified algorithm adds one of *hasHair* and *hasBeard* when *hairColour* is selected according to their ranking. This modification improved the Dice coefficient slightly by 0.03. In a domain with a higher number of dependencies between attributes, we can expect more improvement.

3.2 Relative attribute weighting

As we mentioned in section 2, several attributes tended to be missing in the algorithm output. We calculated the difference set between the human selection and the output of our algorithm. The second improvement assigns weights to attributes in proportion to their frequency within this difference set. The idea behind this is that the difference of these two attribute sets should reflect the general cognitive factor in attribute preference. The result of implemented relative weighting on the base algorithm is shown in Table 4.

Table 4: Results of relative weighting (TITCH-RW)

Domain	Dice	
	static	dynamic
Furniture	0.707	0.699
People	0.648	0.648
People+	0.678	0.678

In comparison to absolute weighting, while the Dice coefficient in the "Furniture" domain slightly increased, that in the "People" domain slightly decreased.

3.3 Error analysis

We carried out a preliminary analysis of errors for the base algorithm as well as for both improvements by attribute weighting. We note that there are three qualitatively different cases (represented by "Correct", "Subset" and "Disjoint" in Table 5). The row "Correct" notes the absolute number of cases where our algorithm's output is the same as the human attribute selection respectively for the base algorithm, its improvement by absolute weighting (column "Abs.") and relative weighting (column "Rel."). "Subset" represents the case where our algorithm's output is a subset of the human attribute selection. "Disjoint" represents the case where our algorithm's output and the human attribute selection are disjoint.

Table 5: Distribution of error types (Furniture domain)

	Base	Abs.	Rel.
Correct	33	59	71
Subset	91	82	29
Disjoint	195	178	219

The number of "Correct" cases and "Subset" cases show an opposite tendency; while the number of "Correct" cases increases steadily from the base algorithm with both weighting improvements (more than double in total), the number of "Subset" cases steadily decreases (by almost 70 % of the initial number). In contrast, the number of "Disjoint" cases fluctuates at a high level and increases slightly overall.

This preliminary analysis allows the conclusion that introducing the different types of weightings into the base algorithm improved the cases where our base algorithm had already provided an approximation (a subset) of the human selection. In these cases, our proposed improvements yielded real gains. However, they do *not* seem at all to reduce the number of "Disjoint" cases. Overall the cases where our initial algorithm fails to produce at least a subset of the human selection, attribute weighting does not provide any gain. This means that in order to gain real improvements, we need to introduce some fundamentally different concept to the base algorithm. A first important step towards this would be a detailed analysis of the "Disjoint" cases, which we have not yet carried out.

Another point we need to mention is that such opposition of tendencies between "Correct" and "Subset" cases was not observed in the "People" domain. In order to elucidate this inconsistency, we need to further analyse the differences of the domains.

4 On the difficulty of the case

We propose to measure the difficulty of attribute-selection in a particular case by the number of possible different sets of attributes that uniquely identify the target object. Table 6 shows the correlation coefficient of the Dice coefficient and the difficulty of the case as defined above.

Table 6: Correlation of task difficulty and Dice coefficient

Static	Furniture	People
Base algorithm	-0.077	0.251
w/ abs. weighting	0.169	0.096
w/ rel. weighting	0.116	0.109

Dynamic	Furniture	People
Base algorithm	0.078	0.251
w/ abs. weighting	0.147	0.096
w/ rel. weighting	0.097	0.110

The correlation factor measures linear dependence, and the closer it is to 0, the less a linear dependence exists. The absolute values of all coefficients except one are less than 0.2. We can thus conclude that there is no linear correlation between the number of potential attribute-selections which uniquely identify the target object in a case, and the success of our algorithm to appropriately generate a set of attributes. In general, we can note this reflects the fact that humans select the attributes in their referring expressions from a very limited set of combinations, independent of the search space of attribute selection to identify the target object. For instance, in a case where a person could be uniquely identified by shoe colour, humans would rather use attributes referring to his facial features, even though it requires more attributes.

As we noted in the explanation of our initial base algorithm, it is not sufficient to simply produce a minimum set of attributes. In many cases, humans manifestly add other attributes to such a set, depending on the case and the target. Hence, as a measure of the difficulty of the task to specify a target object in a given case, the number of full-brevity descriptions is not sufficient; it is simply one factor. Further research into the process of human attribute-selection should provide more insight into the other factors and how they combine. This in turn will provide a better understanding of the difficulty of the task in a specific case.

References

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