DIT - Frequency Based Incremental Attribute Selection for GRE.

J.D. Kelleher
School of Computing
Dublin Institute of Technology
john.kelleher@comp.dit.ie

1 System Description

The DIT system uses an incremental greedy search to generate descriptions, (similar to the incremental algorithm described in (Dale and Reiter, 1995)) incremental algorithm). The selection of the next attribute to be tested for inclusion in the description is ordered by the absolute frequency of each attribute in the training corpus. Attributes are selected in descending order of frequency (i.e. the attribute that occurred most frequently in the training corpus is selected first). The *type* attribute is always included in the description. Other attributes are included in the description if they excludes at least 1 distractor from the set of distractors that fulfil the description generated prior the the attribute's selection.The algorithm terminates when a distinguishing description has been generated or all the targets attributes have been tested for inclusion in the description. To generate a description the system does the following:

Initial conditions: Let T = target object, P_T = the attributes true of T, $DES = \{\}$, D = the set of distractors elements that DES is true of.

Step 1. Check success: if $|P_T| = 0$ **then** return DES as non-distinguishing description **else if** $|D| = 0$ **then** return DES as distinguishing description. **goto Step 2**

Step 2. Choose next property: select the $p_i \in P_T$ that has the highest frequency of occurrence in the training corpus. Let $P_T \leftarrow P_T - p_i$. **goto Step 3**

Step 3. Extend description: if $p_i = type$ **then** include p_i in DES **else if** the inclusion of p_i in DES excludes distractors from D **then** $DES \leftarrow DES \cup p_t$. **goto Step 1**.

Table 1 lists the frequencies of each attribute in the corpus. Column 1 lists the attribute name, Column 2 the frequency of that attribute in the furniture domain, Column 3 the frequency of the attribute in the people domain, Column 4 the overall frequency of the attribute in the training corpus. A dash (-) indicates that the attribute does not occur in that domain.

Attribute	Furniture	People	Both
type	233	185	418
colour	210	2	212
orientation	84	4	88
size	86	-	86
y-dimension	62	63	125
x-dimension	49	50	99
other	5	10	15
hasGlasses	-	90	90
hasBeard	-	88	88
hairColour	-	62	62
hasHair	-	33	33
age	-	15	15
hasSuit	-	3	3
hasShirt	-	2	2
hasTie	-	1	1

Table 1: Attribute frequencies in the training corpus.

Results: When the system was trained on the furniture training corpus and run on the furniture development corpus it achieved a DICE score of 0.752. When it was trained on the people training corpus and run on the people development corpus it achieved a DICE score of 0.695. Finally, when it was trained on the full training corpus and run on the full development corpus it achieved a DICE score of 0.607.

As is evident from the results the system's performance drops when it is trained and run on both domains at the same time. This is largely due to the fact that some of the attributes span both domains while others do not. Consequently, when the system is trained on both domains the frequency of attributes that span both domains are overestimated within each domain. For example, the *orientation*, *y-dimension* and *x-dimension* attributes will be selected before the *hasBeard* for trials in people domain even thought the *hasBeard* attribute occurs more frequently in than these other attributes in that domain.

References

R. Dale and E. Reiter. 1995. Computatinal interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 18:233–263.