

# Frequency Based Incremental Attribute Selection for GRE

John D. Kelleher

School of Computing

Dublin Institute of Technology

[john.kelleher@comp.dit.ie](mailto:john.kelleher@comp.dit.ie)

# Approach

- The DIT system uses an incremental greedy search algorithm.
- It iterates over the list of attribute-values that are true of that target. This list is ordered by decreasing frequency of the attribute occurrences in the training corpus.
- An attribute-value pair is added to the description being generated if it is the type attribute or if it excludes at least 1 or the remaining distractors.

# Algorithm

## Initial conditions

- $T$  = target objects,  $DES = \{ \}$  (the set of attribute value pairs true of the target that have been selected for inclusion in the description),  $P_T$  = set of attribute value pairs true of  $T$ ,  $D$  = set of distractors that  $DES$  is true of

## Step 1. Check Success

- If  $D$  (the set of distractors that the set of attribute value pairs in  $DES$  are true of) is empty make sure the  $DES$  contains at least the type attribute and return  $DES$  as a distinguishing description.
- Otherwise, if  $P_T$  (the set of attributes true of  $T$ ) is empty make sure the  $DES$  contains at least the type attribute and return  $DES$  as a distinguishing description.

## Step 2. Choose next property

- Set  $p_I$  equal to the element of  $P_T$  whose attribute has the highest frequency of occurrence in the training corpus and remove  $p_I$  from  $P_T$

## Step 3. Extend descriptions

- If  $p_I$  is the type attribute or if its addition to  $DES$  would exclude at least 1 distractor add  $p_I$  to the  $DES$ . Let  $D$  = the set of distractors that the set of attribute values pairs in  $DES$  are true of. (Goto Step 1)

# Attribute Frequencies

Attribute	Furniture	People	Overall
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

# Results

Training Corpus	Test Corpus	DICE Score
Furniture Training	Furniture Development	<b>0.752</b>
People Training	People Development	<b>0.695</b>
Full Training	Full Development	<b>0.607</b>

The system's performance drops when it is trained and run on both domains at the same time. This is primarily due to the fact that some attributes occur in both domains and some do not. As a result, the frequency of attributes that occur in both domains are overestimated within each domain.

# Conclusions

- The system achieved reasonably results with a fairly straightforward algorithm.
- One area of future work would be to modify the algorithm to accommodate attribute co-occurrence.
  - For example, the x-dimension and y-dimension attributes often co-occur in the corpus. However, the system does not use this information and as a result often only includes one of these attributes in the generated description when two are required.