Dublin Institute of Technology
ARROW@DIT

Conference papers

School of Computing

2007-01-01

DIT frequency based incremental attribute selection for GRE.

John D. Kelleher
Dublin Institute of Technology, john.d.kelleher@dit.ie

Follow this and additional works at: http://arrow.dit.ie/scschcomcon

Part of the Computer Engineering Commons

Recommended Citation

Kelleher, John: DIT frequency based incremental attribute selection for GRE. Proceedings of the MT Summit XI Workshop Using Corpora for Natural Language Generation: Language Generation and Machine Translation (UNLG+MT), Blez and Varges (eds), 2007

This Article is brought to you for free and open access by the School of Computing at ARROW@DIT. It has been accepted for inclusion in Conference papers by an authorized administrator of ARROW@DIT. For more information, please contact yvonne.desmond@dit.ie, arrow.admin@dit.ie.

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). 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). Where two or more attributes have the same frequency of occurrence the first attribute found with that frequency is selected. The *type* attribute is always included in the description. Other attributes are included in the description if they exclude at least 1 distractor from the set of distractors that fulfil the description generated prior that attribute's selection.The algorithm terminates when a distinguishing description has been generated (i.e., all the distractors have been excluded) or when all the target's attributes have been tested for inclusion in the description. To generate a description the system does the following:

Initial conditions:

$T =$ target object, $DES = \{\}$,
$P_T =$ the set of attributes true of T,
$D =$ the set of distractor objects

Step 1. Check success:

if $|D| = 0$ **then**
return DES as distinguishing description
else if $|P_T| = 0$ **then**
return DES as non-distinguishing description
end if
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:

Let $D_i \leftarrow \{x \in D : p_i(x)\}$
if $p_i =$ *type* **then**
include p_i in DES
else if $|D_i| < |D|$**then**
$DES \leftarrow DES \cup p_i$
end if
Let $D \leftarrow D_i$
goto Step 1

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

2 Results

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

As is evident from the results the system's performance drops when it is trained and run on both

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

Table 1: Attribute frequencies in the training corpus.

domains at the same time. This is primarily due to the fact that some of attributes occur in both domains while others do not. Consequently, when the system is trained on both domains the frequency of attributes that occur in both domains are overestimated within each domain. For example, for trials in the people domain the *y-dimension* and *x-dimension* attributes will be selected before the *hasBeard* attribute even though the *hasBeard* attribute occurs more frequently than these 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.