The Attribute Selection for GRE Challenge: Overview and Evaluation Results

Anja Belz
Natural Language Technology Group
University of Brighton
A.S.Belz@brighton.ac.uk

Albert Gatt
Department of Computing Science
University of Aberdeen
agatt@csd.abdn.ac.uk

Abstract

The Attribute Selection for Generating Referring Expressions (ASGRE) Challenge was the first shared-task evaluation challenge in the field of Natural Language Generation. Six teams submitted a total of 22 systems. All submitted systems were tested automatically for minimality, uniqueness and 'humanlikeness'. In addition, the output of 15 systems was tested in a task-based experiment where subjects were asked to identify referents, and the speed and accuracy of identification was measured. This report describes the ASGRE task and the five evaluation methods, gives brief overviews of the participating systems, and presents the evaluation results.

1 Introduction

The Attribute Selection for Generating Referring Expressions (ASGRE) Challenge has come about as the result of a growing interest in comparative evaluation among NLG researchers over the past two years. The subfield of Generation of Referring Expressions (GRE) was an obvious choice for a first shared-task challenge, because it is one of the most lively and well-defined NLG subfields, with a substantial number of researchers working — unusually for NLG — on the same task, with very similar input and output specifications.

What made the ASGRE Challenge feasible, however, was the availability of the TUNA corpus, a collection of paired pictures of objects and human-produced references annotated with attribute sets (Gatt et al., 2007). We simplified and reduced the TUNA corpus somewhat for the purposes of the ASGRE Challenge, and divided it into training, development and test data. The training and development data was distributed to participants on 4 June, 2007. The test data became available on 21 June, and between then and 28 July, participants were able to request the test data at any time, but were required to return test data outputs within a week after receiving it, or by 28 July, whichever was earlier.

Participating teams were asked to submit a report describing their method before requesting test data. Teams were given a program for computing Dice coefficients, and were asked to compute Dice scores (see Section 4.1) on the development data set and include them in the report. The reports are included in this volume, and the evaluation results reported by participants are shown in Table 2.

Following the call for participation, 19 individual researchers registered their interest. Thirteen of these then formed six teams which submitted outputs of 22 systems by the deadline (see overview of teams and systems in Table 1). All submitted system outputs were tested automatically for minimality, uniqueness and humanlikeness (using Dice scores, see Section 4.1). In addition, 15 systems were tested in a task-based experiment where subjects were asked to identify referents, and the speed and accuracy of identification was measured.

This report presents the results of all evaluations (Section 5), along with an overview of the ASGRE Task (Section 2), brief descriptions of the participating systems (Section 3), and explanations of the evaluation methods (Section 4).

Team ID	Submitted Systems	Organisation
CAM	**CAM-B, CAM-BU, CAM-T, CAM-TU**	Computer Lab, Cambridge University, UK
DIT	**DIT-DS**, DIT-DI	Dublin Institute of Technology, Ireland
GRAPH	**GRAPH-SC, GRAPH-FP**	Universities of Twente and Tilburg, NL; Macquarie University, Australia
IS	**IS-FBN, IS-FBS, IS-IAC**	University of Stuttgart, Germany
NIL	**NIL**	Universidad Complutense de Madrid, Spain
TITCH	TITCH-BS-STAT, TITCH-BS-DYN	Tokyo Institute of Technology, Japan
	TITCH-AW-STAT, TITCH-AW-STAT-PLUS	
	TITCH-RW-STAT, TITCH-RW-STAT-PLUS	
	TITCH-AW-DYN, TITCH-AW-DYN-PLUS	
	TITCH-RW-DYN, TITCH-RW-DYN-PLUS	

Table 1: Overview of participating teams and systems. All systems were included in the automatic evaluations. The 15 systems included in the task-based evaluation are shown in bold.

Some of the terms we use in this report may benefit from up-front explanation: following DUC[1] terminology, a *peer system* is a system submitted to the shared-task challenge; *peer output* and *peer attribute set* refer to an output produced by a peer system; and a *reference attribute set* is an attribute set derived from the annotations of a human-produced referring expression in the TUNA corpus.

2 The Attribute Selection for GRE Task

The ASGRE Task has the same basic functionality as the majority of existing attribute selection for GRE methods: given a target referent and a set of distractors each with their own set of possible attributes, select a set of attributes for inclusion in a referring expression to be generated for the target referent.

However, we deliberately refrained from including in the task definition any aim that would imply assumptions about quality (such as producing minimal or uniquely distinguishing attribute sets) as is often the case in existing GRE task formulations. Instead, we simply told participants which evaluation criteria were going to be used: minimality, uniqueness, humanlikeness, and identification speed and accuracy (where identification means selecting among pictures of entities, see Section 4).

[1]The Document Understanding Conferences (DUC) are an established shared-task evaluation intiative in the field of document summarisation.

2.1 Data

We used all 780 singular items in the TUNA corpus in both the furniture and people domains. We simplified representations somewhat and removed the human-produced referring expressions, retaining only the attribute sets with which they had been annotated.

Each item in the ASGRE corpus consists of an input part, called a domain, and an output part, called a description. Each domain consists of seven domain entities: one target referent and six distractors. Each entity consists of a set of attribute-value pairs, as shown in Figure 1. Each output part, or description, consists of a subset of the attribute-value pairs of the target referent in the same format as shown in Figure 1.

We divided the data into 60% training data, 20% development data and 20% test data. Participants in the ASGRE Challenge were given both input and output parts in the training and development data, but just inputs in the test data. Participants were asked to submit the corresponding outputs for test data inputs.

3 Participating Methods and Systems

In this section, we give very brief descriptions of the participating systems. More details (and references) can be found in the individual participants' reports elsewhere in this volume.

```
<ENTITY ID="121" TYPE="target">
    <ATTRIBUTE NAME="colour" TYPE="literal" VALUE="blue" />
    <ATTRIBUTE NAME="orientation" TYPE="literal" VALUE="left" />
    <ATTRIBUTE NAME="type" TYPE="literal" VALUE="fan" />
    <ATTRIBUTE NAME="size" TYPE="literal" VALUE="small" />
    <ATTRIBUTE NAME="x-dimension" TYPE="gradable" VALUE="1" />
    <ATTRIBUTE NAME="y-dimension" TYPE="gradable" VALUE="3" />
</ENTITY>
```

Figure 1: Example of an entity representation from the furniture subdomain.

CAM-B, CAM-BU, CAM-T, CAM-TU

The CAM team submitted four adaptations of Siddharthan and Copestake's incremental algorithm for attribute selection in open domains. The CAM methods compute the discriminating quotient (DQ) of candidate attributes as the number of distractors which do not have the same attribute minus the number which do. CAM-B incorporates attributes in decreasing order of DQ. CAM-T additionally weights attribute values in terms of how discriminating an attribute is to humans. CAM-BU and CAM-TU are further variations in which DQ values are updated at each incremental step.

DIT-DS, DIT-DI

The DIT system is an incremental algorithm, where the order in which attributes are considered for selection is determined by the absolute frequency of attributes in the training corpus. The type attribute is always selected. Other attributes are selected if they exclude at least one distractor. In the DS version, frequencies are determined separately for the furniture and people domains; in the DI version, frequencies computed from both domains combined are used.

GRAPH-SC, GRAPH-FP

The GRAPH systems use Krahmer et al.'s graph-based framework with two different cost functions: one where the cost of selecting a property depends on its frequency in the corpus (GRAPH-SC), and a variation of this function where certain properties which are particularly salient to humans can be selected at zero cost (GRAPH-FP).

IS-FBS, IS-FBN, IS-IAC

IS-FBS is an extension of Dale's full brevity algorithm and selects the attribute set among the smallest candidate attribute sets that has the highest similarity with any one of the (human-produced) attribute sets found in the corpus for the same entity.[2] IS-FBN selects the attribute set of any length that has the highest similarity score in the same sense as IS-FBS. Similarity in both cases is computed using the Dice metric. IS-IAC is an incremental algorithm which uses a decision-tree built using Quinlan's C4.5. Given the set of remaining available attributes, and the set of attributes already selected, the decision tree returns the attribute to be selected next.

NIL

The NIL system is an adaptation of Reiter and Dale's fast efficient algorithm for referring expression generation, using relative groupings of attributes to determine the order in which they are considered. The relative groupings are obtained empirically from the training data.

TITCH-BS, TITCH-AW, TITCH-RW

TITCH-BS is an incremental algorithm which selects attribute-value pairs in order of their discrimination power (computed case by case, and defined as the number of distractors excluded) until the set uniquely identifies the target referent. In addition to discrimination power, TITCH-AW weights attributes according to their corpus frequency; and TITCH-RW weights attributes according to how frequently they are missing when compared to reference attribute sets. Discrimination power can be computed either dynamically (immediately before each new selection) or statically (once before any selections are made). These system variants are indi-

[2]The version of IS-FBS that was submitted to the ASGRE Challenge inadvertently did not always produce minimal attribute sets, which explains its score in Table 3. The last row in Table 2 (and Bohnet's report in this volume) shows self-reported scores for the corrected version of IS-FBS (marked with an asterisk). However, it was too late to update the other evaluation scores.

cated by -DYN and -STAT tags in the results tables. In addition, TITCH-AW and TITCH-RW optionally model the dependency between the HAIRCOLOUR and HASBEARD/HASHAIR attributes (indicated by the -PLUS tag in results tables).

4 Evaluation Methods

Experience in other shared-task evaluation challenges has shown that the use of a single method of evaluating participating systems can cause loss of trust and/or loss of interest, in particular if the method is seen as biased in favour of a particular type of system (e.g. BLEU in MT Eval) or if it severely restricts the definition of quality (e.g. the PARSEVAL metric in syntactic parsing).

We decided to use a range of different criteria of quality, including both automatically assessed and human-evaluated, both intrinsic and extrinsic methods. The five criteria we used were intended to address questions as follows:

1. *Uniqueness:* do peer attribute sets uniquely describe the target referent?
2. *Minimality:* are peer attribute sets of minimal size?
3. *Humanlikeness:* are peer attribute sets similar to reference attribute sets?
4. *Identification Accuracy:* do peer attribute sets enable people to identify the target referent accurately?
5. *Identification Speed:* do peer attribute sets enable people to identify a referent quickly?

4.1 Automatic Evaluation Methods

In the explanations of evaluation methods below, we refer to the following simplified example of a furniture domain:

```
<DOMAIN>
 <ENTITY ID="e1" TYPE="target">
  <ATTR NAME="colour" VALUE="red"/>
  <ATTR NAME="orientation" VALUE="right"/>
  <ATTR NAME="type" VALUE="chair"/>
  <ATTR NAME="size" VALUE="small"/>
 </ENTITY>
 <ENTITY ID="e2" TYPE="distractor">
  <ATTR NAME="colour" VALUE="blue"/>
  <ATTR NAME="orientation" VALUE="left"/>
  <ATTR NAME="type" VALUE="table"/>
  <ATTR NAME="size" VALUE="large"/>
 </ENTITY>
 <ENTITY ID="e3" TYPE="distractor">
  <ATTR NAME="colour" VALUE="green"/>
  <ATTR NAME="orientation" VALUE="right"/>
  <ATTR NAME="type" VALUE="chair"/>
  <ATTR NAME="size" VALUE="small"/>
 </ENTITY>
</DOMAIN>
```

Uniqueness:

Peer systems were tested to determine whether or not the attribute sets they produced uniquely distinguished the target referent. For example, the set {TYPE:chair, COLOUR:red} uniquely identifies the target referent e1 in the example above, since there is no other entity of which these attribute-value pairs are true. As an aggregate measure, we computed the proportion of outputs of peer systems which uniquely identify their target referents. Uniqueness is often seen as part of the standard problem definition for GRE, whereby an algorithm is successful if, and only if, the attribute set it returns uniquely identifies the referent (Bohnet and Dale, 2005).

Minimality:

A minimal attribute set is defined as an attribute set which uniquely identifies the referent such that there is no smaller attribute set which uniquely identifies the referent. For example, a minimal description of e1 in the above example is {COLOUR:red}, since it is the only red object. There may be more than one minimal attribute set. As an aggregate measure, we computed the proportion of minimal distinguishing outputs produced by peer systems. Minimality has frequently been cited as a desideratum for GRE algorithms (Dale, 1989; Gardent, 2002).

Humanlikeness:

We measured the similarity between the peer attribute sets and (human-produced) reference attribute sets, because (Grice's maxim of Clarity notwithstanding) humans choose to overspecify and underspecify for a variety of reasons; e.g. in the above example, humans are likely to use {TYPE:table, COLOUR:blue} in a description of e2 even though either {TYPE:table} or {COLOUR:blue} are distinguishing.

Similarity between the peer and reference attribute sets was calculated in terms of the Dice co-

efficient of similarity between pairs of peer and reference sets. In this Challenge, given two sets of attributes, A_1 (peer) and A_2 (reference), Dice was calculated as follows:

$$\frac{2 \times |A_1 \cap A_2|}{|A_1 \cup A_2|} \quad (1)$$

4.2 Design of Task-Based Evaluations

In order to evaluate peer systems in a task-based experiment, we created a simple realiser (see Section 4.3.1) to convert peer attribute sets into natural language descriptions that subjects in our experiments would be able to read. We then conducted an experiment in which the realisations of the peer attribute sets were shown to subjects along with pictures of the seven domain entities (referent and distractors). The pictures were the same as were used in the TUNA elicitation experiments. Subjects were given the task of finding the target referent (details described below).

4.3 Experimental set-up

The experimental design was Repeated Latin Squares in which each combination of peer system and test set item was allocated one trial. Because we had 148 items in the test set, but 15 peer systems, we randomly selected 2 test set items and duplicated them to give us a test set size of 150, and 10 Latin Squares.

We recruited 30 subjects from among the faculty and administrative staff at Brighton University. Just over a third of the subjects are computing science faculty members, but only one has an NLP background (not including NLG). The experiments were carried out on a single laptop[3], one subject at a time, in a quiet environment, under the supervision of the first author. Subjects were told only that the experiment formed part of an investigation into references to objects.

During the experiment, subjects were shown pictures of domain entities (referent and distractors) on the computer screen, along with a realised peer attribute set (description). The presentation of trials was randomised for each subject. Each subject was shown 75 trials, so the entire evaluation consisted of 2,250 individual trials. Subjects were told that the description was of one of the objects in the pictures, and were asked to mouse-click on the picture that was being described. Subjects initiated each trial, and each trial began with a bell sound and a small cross flashing on the screen, in order to focus subjects' attention and to direct their gaze to where the description would appear. Trials timed out after 15 seconds, but only 6 of the 2,250 trials reached time-out stage.[4] There were five practice trials at the beginning (the results of which were discarded), after which the real evaluation trials began.

4.3.1 Realiser

We used a very simple template-based realiser (written by Irene Langkilde-Geary, Brighton University, for this purpose) which always realises each attribute in the same way and in the same position regardless of context, except that it groups negated attributes contained in a list of premodifiers or post-modifiers together at the end of the list, in order to avoid ambiguity.[5] Some examples of realised peer attribute sets are shown in Figure 2. The shortest realisations of peer attribute sets were five words long (first example), the longest were 11 words long (two last examples).

4.3.2 Reaction time software

We used DMDX to display the identification experiment and to measure identification time and accuracy. DMDX was designed especially for language-processing experiments, to time the presentation of text, visual and audio material and to measure reaction times to such presentations with millisecond accuracy (Forster and Forster, 2003).

5 Evaluation Results

5.1 Self-reported humanlikeness scores

Table 2 shows the Dice scores for the furniture and people subdomains computed and reported by the participants themselves (using code provided by us). We computed the average of the scores for the furniture and people domains, weighted by the number

[3]The processor was a Pentium M 1.6 GHz with 560MHz Bus and 512MB RAM; 60GB Hard drive; running Windows XP. We used the maximum screen resolution of 1024×768.

[4]Time-outs were counted as missing trials because there were so few of them.

[5]For example, *the person with no beard and glasses* is ambiguous, whereas *the person with glasses and no beard* is not.

{Y-DIMENSION:1, TYPE:fan} the fan at the top
{TYPE:chair, COLOUR:grey, Y-DIMENSION:3, SIZE:large, ORIENTATION:back} the large grey chair facing back at the bottom
{Y-DIMENSION:2, HASSUIT:0, TYPE:person, HASGLASSES:1} the person wearing glasses and no suit in the middle row
{TYPE:chair, COLOUR:grey, Y-DIMENSION:1, SIZE:large, X-DIMENSION:3} the large grey chair in the center column at the top

Figure 2: Example realisations of peer attribute sets.

	w-Avg	Furniture	People	Trainable?	Dev seen?
IS-FBN	0.774	0.800	0.744	Y	N
CAM-TU	0.739	0.782	0.688	Y	Y
CAM-T	0.738	0.780	0.688	Y	Y
IS-IAC	0.726	0.752	0.696	Y	N
DIT-DS	0.726	0.752	0.695	Y	N
TITCH-RW-STAT-PLUS	0.694	–	0.678	Y	Y
GRAPH-FP	0.692	0.710	0.671	Y	N
TITCH-RW-DYN-PLUS	0.689	–	0.678	Y	Y
TITCH-AW-STAT-PLUS	0.684	–	0.683	Y	Y
TITCH-AW-DYN-PLUS	0.684	–	0.683	Y	Y
TITCH-RW-STAT	0.68	0.707	0.648	Y	Y
TITCH-RW-DYN	0.676	0.699	0.648	Y	Y
TITCH-AW-STAT	0.669	0.685	0.651	Y	Y
TITCH-AW-DYN	0.669	0.685	0.651	Y	Y
GRAPH-SC	0.659	0.661	0.656	Y	N
CAM-BU	0.632	0.585	0.688	N	–
CAM-B	0.62	0.563	0.688	N	–
NIL	0.612	0.752	0.448	Y	N
DIT-DI	0.607	?	?	Y	N
TITCH-BS-DYN	0.582	0.601	0.559	N	–
TITCH-BS-STAT	0.575	0.588	0.559	N	–
IS-FBS	0.505	0.56	0.44	N	–
IS-FBS*	0.357	0.39	0.32	N	–

Table 2: Self-reported mean Dice scores on development set. 'Trainable' indicates whether some aspect of the system was determined quantitatively from the corpus. 'Dev seen' indicates for trainable systems whether or not the development data set was used in training. To compute w-Avg for each TITCH-*-PLUS system we included the furniture score from the corresponding TITCH-* system (without the -PLUS extension). (For IS-FBS* see Footnote 2.)

of items in each domain (80 furniture items and 68 people items). In the table, systems are listed in order of this weighted average score (w-Avg).

5.2 Uniqueness

All systems except one uniquely described the target referent 100% of the time (as calculated on the test set). The exception was the TITCH-AW-DYNAMIC system, in whose output 23% of descriptions did not describe the target referent uniquely.

5.3 Minimality

Percentages of minimal peer outputs are shown for each system in Table 3. The negative correlation between minimality and Dice scores is considered in Section 5.6 below.

5.4 Humanlikeness

Table 4 displays the mean and standard deviation obtained by each system overall, as well as by domain (people or furniture). The systems are ordered by the overall mean score.

The differences apparent in the table were further confirmed via a 22 SYSTEM $\times$ 2 DOMAIN univariate Analysis of Variance (ANOVA) over the Dice scores. There were significant main effects of SYSTEM ($F(21, 2896) = 7.466$, $p < .001$) and DOMAIN ($F(1, 2896) = 79.73$, $p < .001$). The interaction was also significant ($F(17, 2896) = 5.413$,

	overall		furniture		people	
	Mean	SD	Mean	SD	Mean	SD
IS-FBN	0.7709	0.21	0.80	0.16	0.74	0.25
DIT-DS	0.7501	0.26	0.80	0.27	0.69	0.24
IS-IAC	0.7461	0.27	0.80	0.27	0.68	0.26
CAM-T	0.7249	0.27	0.79	0.24	0.65	0.28
CAM-TU	0.7214	0.27	0.78	0.25	0.65	0.28
GRAPH-FP	0.6898	0.25	0.71	0.26	0.67	0.24
GRAPH-SC	0.6715	0.26	0.71	0.26	0.63	0.25
TITCH-RW-STAT	0.6551	0.25	0.69	0.24	0.61	0.26
TITCH-RW-DYN	0.6551	0.25	0.69	0.24	0.61	0.26
TITCH-AW-STAT-PLUS	0.6532	0.28	–	–	0.65	0.28
TITCH-AW-DYN-PLUS	0.6532	0.28	–	–	0.65	0.28
TITCH-AW-STAT	0.6455	0.25	0.67	0.25	0.62	0.26
TITCH-AW-DYN	0.6411	0.25	0.66	0.25	0.62	0.26
TITCH-RW-STAT-PLUS	0.6400	0.28	–	–	0.64	0.28
TITCH-RW-DYN-PLUS	0.6400	0.28	–	–	0.64	0.28
CAM-BU	0.6300	0.27	0.61	0.26	0.65	0.28
NIL	0.6251	0.34	0.80	0.27	0.42	0.30
DIT-DI	0.6243	0.25	0.71	0.18	0.53	0.29
CAM-B	0.6203	0.27	0.59	0.25	0.65	0.28
TITCH-BS-DYN	0.5934	0.25	0.60	0.24	0.58	0.27
TITCH-BS-STAT	0.5928	0.26	0.60	0.26	0.58	0.27
IS-FBS	0.5276	0.28	0.62	0.23	0.42	0.29

Table 4: Mean Dice scores on test set and standard deviation (SD). Mean and SD are shown overall and separately for each subdomain. To compute overall means for each TITCH-*-PLUS system we included the furniture outputs from the corresponding TITCH-* system (without the -PLUS extension).

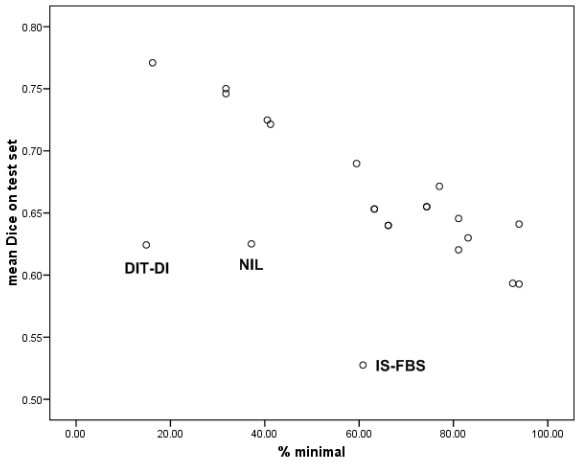

Figure 3: Minimality plotted against mean test set Dice coefficient (three outlier systems labelled).

$p < .001$). Table 5 displays the results of pairwise comparisons using Tukey's Honestly Significant Difference test. The table[6] shows systems and mean Dice scores, and indicates the homogeneous subsets in the data, so that systems which do not share a letter are significantly different at $p < .05$.

[6]Here and in Table 6 we use the same presentation format as in the DUC reports.

5.5 Results of task-based evaluation

In this section, we report on differences between systems on the two dependent variables in the identification experiment, identification time and identification accuracy. For each dependent variable, we report main effects and, when these are significant, further pairwise comparisons between systems. In all cases, our analysis uses items as the error term. We used SPSS 15.0 to perform all analyses of our experimental results.

5.5.1 Identification Accuracy

Because of the large number of zero values in error rate response proportions, and a high dependency of variance on the mean, we used a Kruskall-Wallis ranks test to compare identification accuracy rates across systems. This did not reveal any significant differences between systems at all ($\chi^2 = 8.971$, $= .8$).

5.5.2 Identification Speed

A univariate ANOVA was conducted over identification times using SYSTEM as the sole independent variable. The main effect was significant

System	minimal (%)
TITCH-AW-DYN	93.92
TITCH-BS-STAT	93.92
TITCH-BS-DYN	92.56
CAM-BU	83.10
CAM-B	81.08
TITCH-AW-STAT	81.08
GRAPH-SC	77.03
TITCH-RW-DYN	74.32
TITCH-RW-STAT	74.32
TITCH-RW-DYN-PLUS	66.18
TITCH-RW-STAT-PLUS	66.18
TITCH-AW-DYN-PLUS	63.24
TITCH-AW-STAT-PLUS	63.23
IS-FBS	60.08
GRAPH-FP	59.46
CAM-T	40.54
CAM-TU	41.21
NIL	37.16
DIT-DS	31.76
IS-IAC	31.75
IS-FBN	16.22
DIT-DI	14.86

Table 3: Proportion of minimal descriptions per system

System	Dice						
IS-FBN	0.7709	A					
DIT-DS	0.7501	A	B				
IS-IAC	0.7461	A	B	C			
CAM-T	0.7249	A	B	C	D		
CAM-TU	0.7214	A	B	C	D		
GRAPH-FP	0.6898	A	B	C	D	E	
GRAPH-SC	0.6715	A	B	C	D	E	
TITCH-RW-STAT	0.6551	A	B	C	D	E	
TITCH-RW-DYN	0.6551	A	B	C	D	E	
TITCH-AW-STAT+	0.6532	A	B	C	D	E	
TITCH-AW-DYN+	0.6532	A	B	C	D	E	
TITCH-AW-STAT	0.6455		B	C	D	E	F
TITCH-AW-DYN	0.6411		B	C	D	E	F
TITCH-RW-STAT+	0.6400		B	C	D	E	F
TITCH-RW-DYN+	0.6400		B	C	D	E	F
CAM-BU	0.6300			C	D	E	F
NIL	0.6251				D	E	F
DIT-DI	0.6243				D	E	F
CAM-B	0.6203				D	E	F
TITCH-BS-DYN	0.5934					E	F
TITCH-BS-STAT	0.5928					E	F
IS-FBS	0.5276						F

Table 5: Homogeneous subsets following post-hoc Tukey comparisons on mean Dice score. Systems which do not share a common letter are significantly different at $p < .05$.

($F(14, 449) = 6.401$, $p < .001$). Once again, post-hoc Tukey comparisons were used to compare the different systems. Homogeneous subsets of systems, together with mean identification times, are displayed in Table 6. Again, systems which do not share a letter differ significantly at $p < .05$.

5.6 Correlations between scores

There is a strong as well as highly significant positive correlation (Pearson's $r = .932$, $p < .001$) between the mean self-reported Dice scores for each system on the development set (shown in Table 2) and the mean Dice scores computed by us on the test data (shown in Table 4).

We also looked at the correlation between Dice scores and mean identification time, for those systems that were included in the task-based evaluation. The rationale was to obtain a rough indication of whether high agreement on attribute selection with the reference attribute sets would indicate faster identification. Therefore, the expected correlation is negative (i.e. a higher Dice score entails shorter identification times). Although the correlation is in the predicted direction, it is not very strong and fails to reach significance ($r = -.305$, $p > .2$).

We also looked at the relationship between humanlikeness and minimality. This is displayed in the scatter plot in Figure 3 which plots the mean Dice scores for each system against the proportion of minimal descriptions produced. With the exception of the three labelled outliers, there is a trend for the mean Dice score obtained by a system to decrease as the proportion of minimal descriptions increases.

6 Conclusions

Since the ASGRE Challenge was the first shared-task challenge in NLG, we regarded (and presented) it as a pilot event for a full-scale — and hopefully longer-term — NLG shared-task evaluation initiative. Our aim was to organise the Challenge in a relaxed, non-competetive and collaborative atmosphere, and initial feedback from participants indicates that we succeeded in this aim.

A crucial component of the ASGRE Challenge has been the use of five different criteria of quality, from the traditional criteria of uniqueness and minimality, and the more recent criterion of humanlikeness, to new extrinsic criteria of identification time and accuracy. Results showed the importance of using several evaluation criteria: some scores are negatively correlated (Dice and minimality), and can yield dra-

System	Mean IT			
TITCH-RW-STAT	2514.367	A		
CAM-TU	2572.821	A		
CAM-T	2626.022	A		
TITCH-AW-STAT-PLUS	2652.845	A		
CAM-BU	2659.369	A		
GRAPH-FP	2724.559	A		
TITCH-RW-STAT-PLUS	2759.758	A		
CAM-B	2784.804	A		
DIT-DS	2785.396	A		
GRAPH-SC	2811.091	A		
IS-IAC	2844.172	A	B	
TITCH-AW-STAT	2864.933	A	B	
NIL	2894.77	A	B	
IS-FBN	3570.904		B	C
IS-FBS	4008.985			C

Table 6: Homogeneous subsets following post-hoc Tukey comparisons on Identification Time. Systems which do not share a common letter are significantly different at $p < .05$.

matically different system rankings (e.g. Dice and identification time).

Another important aspect of the Challenge (one that is new in NLP shared-task challenges as far as we are aware) were the self-reported scores which gave participants a degree of control over the reported results.

The enthusiastic response from GRE researchers to the Challenge (and supportive comments from the wider NLG community) demonstrates that parts of the NLG field are willing and able to participate in comparative evaluation events, and we plan to organise similar events in the future.

As with all shared-task evaluations, the evaluation results of the ASGRE Challenge do not tell us what is in general terms the best way to do attribute selection for GRE. Rather, we have directly comparable results for 22 different systems and five quality criteria. This can help guide development and selection of attribute selection systems for similar domains in the future, in particular where such systems are required to maximise specific aspects of quality.

Acknowledgments

We gratefully acknowledge the contribution made to the evaluations by the faculty and staff at Brighton University who participated in the task-based evaluation. We would like to thank Irene Langkilde-Geary for providing the realiser for the task-based evaluation. Thanks are also due to Robert Dale, Kees van Deemter and Ielka van der Sluis for helpful commnents on the evaluations and this report. The biggest contribution was, of course, made by the participants who made the best of the short available time to create their systems.

References

B. Bohnet and R. Dale. 2005. Viewing referring expression generation as search. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence, IJCAI-05*.

R. Dale. 1989. Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, ACL-89.*

K. I. Forster and J. C. Forster. 2003. DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, 35(1):116–124.

C. Gardent. 2002. Generating minimal definite descriptions. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL-02*.

A. Gatt, I. van der Sluis, and K. van Deemter. 2007. Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proceedings of the 11th European Workshop on Natural Language Generation (ENLG'07)*, pages 49–56.