The TUNA Challenge 2008: Overview and Evaluation Results

Albert Gatt
Department of Computing Science
University of Aberdeen
Aberdeen AB24 3UE, UK
a.gatt@abdn.ac.uk

Anja Belz **Eric Kow**
Natural Language Technology Group
University of Brighton
Brighton BN2 4GJ, UK
{asb, eykk10}@brighton.ac.uk

Abstract

The TUNA Challenge was a set of three shared tasks at REG'08, all of which used data from the TUNA Corpus. The three tasks covered attribute selection for referring expressions (TUNA-AS), realisation (TUNA-R) and end-to-end referring expression generation (TUNA-REG). 8 teams submitted a total of 33 systems to the three tasks, with an additional submission to the Open Track. The evaluation used a range of automatically computed measures. In addition, an evaluation experiment was carried out using the peer outputs for the TUNA-REG task. This report describes each task and the evaluation methods used, and presents the evaluation results.

1 Introduction

The TUNA Challenge 2008 built on the foundations laid in the ASGRE 2007 Challenge (Belz and Gatt, 2007), which consisted of a single shared task, based on a subset of the TUNA Corpus (Gatt et al., 2007). The TUNA Corpus is a collection of human-authored descriptions of a referent, paired with a representation of the *domain* in which that description was elicited.

The 2008 Challenge expanded the scope of the previous edition in a variety of ways. This year, there were three shared tasks. TUNA-AS is the Attribute Selection task piloted in the 2007 ASGRE Challenge, which involves the selection of a set of attributes which are true of a target referent, and help to distinguish it from its distractors in a domain. TUNA-R is a realisation task, involving the mapping from attribute sets to linguistic descriptions. TUNA-REG is an 'end to end' referring expression generation task, involving a mapping from an input domain to a linguistic description of a target referent. In addition, there was an Open Submission Track, where participants were invited to submit a report on any interesting research that involved the shared task data, and an Evaluation Track, for which submissions were invited on proposals for evaluation methods. This year's TUNA Challenge also expanded considerably on the evaluation methods used in the various tasks. The measures can be divided into *intrinsic*, automatically computed methods, and *extrinsic* measures obtained through a task-oriented experiment involving human participants.

The training and development data for the Challenge included the full dataset used in the ASGRE Challenge, that is, all of the 2007 training, development and test data. For the 2008 edition, two new test sets were constructed. Test Set 1 was used for TUNA-R, Test Set 2 was used for both TUNA-AS and TUNA-REG.

1.1 Overview of submissions

Overall, 8 research groups submitted 33 systems by the deadline. Table 1 provides a summary of the submissions. The extrinsic evaluation experiment was carried out on peer outputs in the TUNA-REG task only, using outputs from at most 4 systems per participating group. The 10 systems included are indicated in boldface in the table. An additional submission was made by the USP team to the Open Track. No submissions were made to the Evaluation Track. Given the number of submissions, space restrictions do not permit us to give an overview of the characteristics of the various systems; these can be found in the reports authored by each participating group, which are included in this volume.

Group	Organisation	TUNA-AS	TUNA-R	TUNA-REG
ATT	AT&T Labs Research Inc.	ATT-DR-b ATT-DR-sf ATT-FB-f ATT-FB-m ATT-FB-sf ATT-FB-sr	ATT-R	ATT-TemplateS-ws **ATT-TemplateS-drws** ATT-Template-ws ATT-Template-drws **ATT-PermuteRank-ws** **ATT-PermuteRank-drws** **ATT-Dependency-drws** ATT-Dependency-ws
DIT	Dublin Institute of Technology	DIT-FBI DIT-TVAS	DIT-CBSR DIT-RBR	**DIT-FBI-CBSR** **DIT-TVAS-RBR**
GRAPH	University of Tilbug etc	GRAPH-FP		**GRAPH-4+B***
IS	University of Stuttgart	IS-FP	IS-GT	**IS-FP-GT**
JUCSENLP	Jadavpur University	JU-PTBSGRE		
NIL-UCM	Universidad Complutense de Madrid	NIL-UCM-MFVF	NIL-UCM-BSC	**NIL-UCM-FVBS**
OSU	Ohio State University	OSU-GP		**OSU-GP***
USP	University of Sao Paolo	USP-EACH-FREQ		

Table 1: Overview of participating teams and systems, by task. TUNA-REG peer systems whose outputs were included in the extrinsic, task-based evaluation are shown in boldface. Systems marked * were submissions to TUNA-AS which made use of the off-the-shelf ASGRE realiser for their entries to TUNA-REG.

Participants in TUNA-AS and TUNA-R were also given the opportunity to submit peer outputs for TUNA-REG, and having them included in the extrinsic evaluation, by making the use of off-the-shelf modules. For systems in TUNA-AS, we made available a template-based realiser, written by Irene Langkilde-Geary at the University of Brighton. Originally used in the 2007 ASGRE Challenge, this was re-used by some TUNA-AS participants to realise their outputs. Systems which made use of this facility are marked by a (*) in Table 1.

In the rest of this report, we first give an overview of the tasks and the data used for the Challenge (Section 2), followed by a description of the evaluation methods (Section 3). Section 4 gives the comparative evaluation results for each task, followed by a few concluding remarks in Section 5. In what follows, we will use the following terminology, in keeping with their usage in Belz and Gatt (2007): a *peer system* is a system submitted to the shared-task challenge, while *peer output* is an attribute set or a description (in the form of a word string) produced by a peer system. We will refer to a description in the TUNA corpus as a *reference output*.

2 Data and task overview

2.1 The TUNA Data

The TUNA corpus was constructed via an elicitation experiment as part of the TUNA project[1]. Each file in the data consists of a single pairing of a domain (representation of entities and their attributes) and a human-authored description (reference output)

```
<TRIAL CONDITION="+/-LOC" ID="...">

    <DOMAIN>
        <ENTITY ID="..." TYPE="target" IMAGE="...">
            <ATTRIBUTE NAME="..." VALUE="..." />
            ...
        </ENTITY>
        <ENTITY ID="..." TYPE="distractor" IMAGE="...">
            <ATTRIBUTE NAME="..." VALUE="..." />
            ...
        </ENTITY>
        ...
    </DOMAIN>

    <WORD-STRING>
        the string describing the target referent
    </WORD-STRING>

    <ANNOTATED-WORD-STRING>
        the string in WORD-STRING annotated
        with attributes in ATTRIBUTE-SET
    </ANNOTATED-WORD-STRING>

    <ATTRIBUTE-SET>
        the set of domain attributes in the description
    </ATTRIBUTE-SET>

</TRIAL>
```

Figure 1: Format of corpus items

which is intended to describe the target referent in the domain. Only the singular descriptions in the corpus were used for the TUNA Challenge.

The descriptions in the corpus are subdivided by *entity type*: there are references to people, and references to furniture items. In addition, the elicitation experiment manipulated a single condition, ±LOC. In the +LOC condition, experimental participants were told that they could refer to entities using any of their properties, including their location. In the −LOC condition, they were *discouraged* from doing so, though not prevented.

Figure 1 is an outline of the XML format used in the Challenge. Each file has a root TRIAL node with a unique ID and an indication of the experimental condition. The DOMAIN node subsumes 7

[1] http://www.csd.abdn.ac.uk/research/tuna/

ENTITY nodes, which themselves subsume a number of ATTRIBUTE nodes defining the properties of an entity in attribute-value notation. The attributes include properties such as an object's colour or a person's clothing, and the location of the image in the visual display which the DOMAIN represents. Each ENTITY node indicates whether it is the target referent or one of the six distractors, and also has a pointer to the image that it represents. Images were made available to the TUNA Challenge participants.

The WORD-STRING is the actual description typed by a human author, and the ATTRIBUTE-SET is the set of attributes belonging to the referent that the description includes. The ANNOTATED-WORD-STRING node was only provided in the training and development data, to display how substrings of a human-authored description were mapped to attributes to determine the ATTRIBUTE-SET.

Training and development data: For the TUNA Challenge, the 780 singular corpus instances were divided into 80% training data and 20% development data. This data consists of all the training, development and test data used in the 2007 ASGRE Challenge.

Test data: Two new test sets were constructed by replicating the original TUNA elicitation experiment. The new experiment was designed to ensure that each DOMAIN in the new test sets had two reference outputs. Thus, this year's corpus-based evaluations are conducted against multiple instances of each input DOMAIN. Both sets consisted of 112 items, divided equally into furniture and people descriptions, sampled from both experimental conditions (±LOC). Test Set 1 was used for the TUNA-R Task. Participants in this task received a version of the test set whose items consisted of a DOMAIN node and an ATTRIBUTE-SET node. There were 56 unique DOMAINs, each represented twice in the test set, with two attribute sets from two different human authors. Because each DOMAIN and ATTRIBUTE-SET combination in this test set is unique, the results for this task are reported below over the whole of Test Set 1. Test Set 2 was used for the TUNA-AS and TUNA-REG Tasks. For these tasks, the test items given to participants consisted of a DOMAIN node only. There were 112 unique DOMAINs; the evaluations on these tasks were conducted by comparing each peer output to two different reference outputs for each of these domains. Therefore, in the TUNA-AS and TUNA-REG tasks, the data presented here averages over the two outputs per DOMAIN.

2.2 The tasks

Task 1: Attribute Selection (TUNA-AS): The TUNA-AS task focused on *content determination* for referring expressions, and follows the basic problem definition used in much previous work in the area: given a domain and a target referent, select a subset of the attributes of that referent which will help to distinguish it from its distractors. The inputs for this task consisted of a TRIAL node enclosing a DOMAIN node (a representation of entities and properties). A peer output was a TRIAL node enclosing just an ATTRIBUTE-SET node whose children were the attributes selected by a peer system for the target entity.

Task 2: Realisation (TUNA-R): The TUNA-R task focussed on *realisation*. The aim was to map an ATTRIBUTE-SET node to a word string which describes the ENTITY that is marked as the *target* such that the entity can be identified in the domain. The inputs for this task consisted of a TRIAL node enclosing a DOMAIN and an ATTRIBUTE-SET node. A peer output for this task consisted of a TRIAL node enclosing just a WORD-STRING node.

Task 3: 'End-to-end' Referring Expression Generation (TUNA-REG): For the TUNA-REG task, the input consisted of a DOMAIN, and a peer output was a word string which described the entity marked as the *target* such that the entity could be identified in the domain. The input for this task was identical to that for TUNA-AS, i.e. a TRIAL node enclosing just a DOMAIN node. A peer output for this task was identical in format to that for the TUNA-R task, i.e. a TRIAL enclosing just a WORD-STRING node.

3 Evaluation methods

The evaluation methods used in each task, and the quality criteria that they assess, are summarised in Table 2. Peer outputs from all tasks were evaluated using intrinsic methods. All of these were automatically computed, and are subdivided into (a)

Task	Criterion	Type	Methods
TUNA-AS	Humanlikeness	Intrinsic	Accuracy, Dice, MASI
	Minimality	Intrinsic	Proportion of minimal outputs
	Uniqueness	Intrinsic	Proportion of unique outputs
TUNA-R	Humanlikeness	Intrinsic	Accuracy, BLEU, NIST, string-edit distance
TUNA-REG	Humanlikeness	Intrinsic	Accuracy, BLEU, NIST string-edit distance
	Ease of comprehension	Extrinsic	Self-paced reading in identification experiment
	Referential Clarity	Extrinsic	Speed and accuracy in identification experiment

Table 2: Evaluation methods used per task

those measures that assess humanlikeness, i.e. the degree of similarity between a peer output and a reference output; and (b) measures that assess intrinsic properties of peer outputs. Peer outputs from the TUNA-REG task were also included in a human, task-oriented evaluation, which is extrinsic insofar as it measures the adequacy of a peer output in terms of its utility in an externally defined task. In the remainder of this section, we summarise the properties of the intrinsic methods. Section 3.1 describes the experiment conducted for the extrinsic evaluation.

Dice coefficient (TUNA-AS)**:** This is a set-comparison metric, ranging between 0 and 1, where 1 indicates a perfect match between sets. For two attribute sets A and B, Dice is computed as follows:

$$Dice(A,B) = \frac{2 \times |A \cap B|}{|A| + |B|} \quad (1)$$

MASI (TUNA-AS)**:** The MASI score (Passonneau, 2006) is an adaptation of the Jaccard coefficient which biases it in favour of similarity where one set is a subset of the other. Like Dice, it ranges between 0 and 1, where 1 indicates a perfect match. It is computed as follows:

$$MASI(A,B) = \delta \times \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

where δ is a *monotonicity coefficient* defined as follows:

$$\delta = \begin{cases} 0 & \text{if } A \cap B = \emptyset \\ 1 & \text{if } A = B \\ \frac{2}{3} & \text{if } A \subset B \text{ or } B \subset A \\ \frac{1}{3} & \text{otherwise} \end{cases} \quad (3)$$

Accuracy (all tasks): This is computed as the proportion of the peer outputs of a system which have an exact match to a reference output. In TUNA-AS, Accuracy was computed as the proportion of times a system returned an ATTRIBUTE-SET identical to the reference ATTRIBUTE-SET produced by a human author for the same DOMAIN. In TUNA-R and TUNA-REG, Accuracy was computed as the proportion of times a peer WORD-STRING was identical to the reference WORD-STRING produced by an author for the same DOMAIN.

String-edit distance (TUNA-R**,** TUNA-REG)**:** This is the classic Levenshtein distance measure, used to compare the difference between a peer output and a reference output in the corpus, as the minimal number of insertions, deletions and/or substitutions of words required to transform one string into another. The cost for insertions and deletions was set to 1, that for substitutions to 2. Edit distance is an integer bounded by the length of the longest description in the pair being compared.

BLEU (TUNA-R**,** TUNA-REG)**:** This is an n-gram based string comparison measure, originally proposed by Papineni et al. (2002) for evaluation of Machine Translation systems. It evaluates a system based on the proportion of word n-grams (considering all n-grams of length $n \leq 4$ is standard) that it shares with several reference translations. Unlike Dice, MASI and String-edit, BLEU is by definition an aggregate measure (i.e. a single BLEU score is obtained for a system based on the entire set of items to be compared, and this is generally not equal to the average of BLEU scores for individual items). BLEU ranges between 0 and 1.

NIST (TUNA-R**,** TUNA-REG): This is a version of BLEU, which gives more importance to less frequent (hence more informative) n-grams. The range of NIST scores depends on the size of the test set. Like BLEU, this is an aggregate measure.

Uniqueness (TUNA-AS): This measure was included for backwards comparability with the ASGRE Challenge 2007. It is defined as the proportion of peer ATTRIBUTE-SETs which identify the target referent uniquely, i.e. whose (logical conjunction of)

attributes are true of the target, and of no other entity in the DOMAIN.

Minimality (TUNA-AS**)**: This measure was defined as the proportion of peer ATTRIBUTE-SETs which are *minimal*, where 'minimal' means that there is no attribute-set which uniquely identifies the target referent in the domain which is smaller. Note that this definition includes Uniqueness as a prerequisite, since the description must identify the target entity uniquely in order to qualify for Minimality.

All intrinsic evaluation methods except for BLEU and NIST were computed (a) overall, using the entire test data set (i.e. Test Set 1 or 2 as appropriate); and (b) by object type, that is, computing separate values for outputs referring to targets of type *furniture* and *people*.

3.1 Extrinsic evaluation in TUNA-REG

The experiment for the extrinsic evaluation of TUNA-REG peer outputs combined a self-paced reading and identification paradigm, comparing the peer outputs from 10 of the TUNA-REG systems shown in Table 1, as well as the two sets of human-authored reference outputs for Test Set 2. We refer to the latter as HUMAN-1 and HUMAN-2 in what follows[2].

In the task given to experimental subjects, a trial consisted of a description paired with a visual domain representation corresponding to an item in Test Set 2. Each trial was split into two phases: (a) in an initial reading phase, subjects were presented with the description only. This phase was terminated by subjects once they had read the description. (b) In the second, identification phase, subjects saw the visual domain in which the description had been produced, consisting of images of the domain entities in the same spatial configuration as that in the test set DOMAIN. They clicked on the object that they thought was the intended referent of the description they had read.

The experiment yielded three dependent measures: (a) reading time (RT), measured from the point at which the description was presented, to the point at which a participant called up the next screen via mouse click; (b) identification time (IT), measured from the point at which pictures (the visual domain) were presented on the screen to the point where a participant identified a referent by clicking on it; (c) error rate (ER), the proportion of times the wrong referent was identified.

This design differs from that used in the 2007 ASGRE Challenge, in which descriptions and visual domains were presented in a single phase (on the same screen), so that RT and IT were conflated. The new experiment replicates the methodology reported in Gatt and Belz (2008), in a follow-up study on the ASGRE 2007 data. Another difference between the two experiments is that the current one is based on peer outputs which are themselves realisations, whereas the ASGRE experiment involved attribute sets which had to be realised before they could be used.

Design: We used a Repeated Latin Squares design, in which each combination of SYSTEM[3] and test set item is allocated one trial. Since there were 12 levels of SYSTEM, but 112 test set items, 8 randomly selected items (4 furniture and 4 people) were duplicated, yielding 120 items and 10 12×12 latin squares. The items were divided into two sets of 60. Half of the participants did the first 60 items (the first 5 latin squares), and the other half the second 60.

Participants and procedure: The experiment was carried out by 24 participants recruited from among the faculty and administrative staff of the University of Brighton, as well as from among the authors' acquaintances. Participants carried out the experiment under supervision in a quiet room on a laptop. Stimulus presentation was carried out using DMDX, a Win-32 software package for psycholinguistic experiments involving time measurements (Forster and Forster, 2003). Participants initiated each trial, which consisted of an initial warning bell and a fixation point flashed on the screen for 1000ms. They then read the description and called up the visual domain to identify the referent. Trials timed out after 15000ms.

Treatment of outliers and timeouts: Trials which

[2]Note that HUMAN-1 and HUMAN-2 were both sets of descriptions randomly sampled from the data collected in the experiment. Each set of human descriptions contains output from different human authors.

[3]The SYSTEM independent variable in this experiment includes HUMAN-1 and HUMAN-2.

timed out with no response were discounted from the analysis. Out of a total of (24×60 =) 1440 trials, there were 4 reading timeouts (0.3%) and 7 identification timeouts (0.5%). Outliers for RT and IT were defined as those exceeding a threshold of mean $\pm 2SD$. There were 64 outliers on RT (4.4%) and 191 on IT (13.3%). Outliers were replaced by the overall mean for RT and IT (see Ratliff (1993) for discussion of this method).

4 Evaluation results

This section presents results for each of the tasks. For all measures, except BLEU and NIST, we present separate descriptive statistics by entity type (people vs. furniture subsets of the relevant test set), and overall.

4.1 Results for TUNA-AS

Descriptive statistics are displayed for all systems in Table 3. This includes the Accuracy and Minimality scores (proportions), and mean MASI and Dice scores. Values are displayed by entity type and overall. The standard deviation for Dice and MASI is displayed overall. Scores average over both sets of reference outputs in Test Set 2. All systems scored 100% on Uniqueness, and either 0 or 100% on Minimality. These measures are therefore not included in the significance testing, though Minimality is included in the correlations reported below.

Two 15 (SYSTEM) $\times$ 2 (ENTITY TYPE) univariate ANOVAs were conducted on the Dice and MASI scores. We report significant effects at $p \leq .001$. There were main effects of SYSTEM (Dice: $F(13, 1540) = 193.08$; MASI: $F(13, 1540) = 93.45$) and ENTITY TYPE (Dice: $F(1, 1540) = 91.75$; MASI: $F(1, 1540) = 168.12$), as well as a significant interaction between the two (Dice: $F(13, 1540) = 7.45$, MASI: $F(13, 1540) = 7.35$). Post-hoc Tukey's comparisons on both Dice and MASI yielded the homogeneous subsets displayed in Table 4.

Differences among systems on Accuracy were analysed by coding this as an indicator variable: for each peer output, the variable indicated whether it achieved perfect match with *at least one* of the two reference outputs on the same DOMAIN. A Kruskall-Wallis test showed that the difference between systems was significant ($\chi^2 = 275.01, p < .001$).

	Minimality	**Accuracy**	**Dice**	MASI
Minimality		-0.877	-0.959	-0.901
Accuracy	-0.877		0.973	0.998
Dice	-0.959	0.973		0.985
MASI	-0.901	0.998	0.985	

Table 5: Correlations for TUNA-AS; all values are significant at $p \leq .05$

Pairwise correlations using Pearson's r are shown in Table 5, for all measures except Uniqueness. All correlations are positive and significant, with the exception of those involving Minimality, which correlates negatively with all other measures (i.e. the higher the proportion of minimal descriptions of a system, the lower its score on humanlikeness, as measured by Dice, MASI and Accuracy). This result corroborates a similar finding in the 2007 ASGRE Challenge.

4.2 Results for TUNA-R

Table 6 shows descriptives for the 5 participating systems in TUNA-R. Once again, mean Edit scores and Accuracy proportions are shown both overall and by entity type, while BLEU and NIST are overall aggregate scores.

A 15 (SYSTEM) $\times$ 2 (ENTITY TYPE) univariate ANOVA was conducted on the Edit Distance scores. There was no main effect of SYSTEM, and no interaction, but ENTITY TYPE exhibited a main effect ($F(1, 550) = 19.99, p < .001$). Given the lack of a main effect, no post-hoc comparisons between systems were conducted. A Kruskall-Wallis test also showed no difference between systems on Accuracy. Pairwise correlations between all measures are shown in Table 7; this time, the only significant correlation is between NIST and BLEU.

	Edit	**Accuracy**	NIST	BLEU
Edit		0.195	-0.095	0.099
Accuracy	0.195		0.837	0.701
NIST	-0.095	0.837		0.900*
BLEU	0.099	0.701	0.900*	

Table 7: Correlations for the TUNA-R task (* indicates $p \leq .05$).

4.3 Results for TUNA-REG

4.3.1 Tests on the intrinsic measures

Results for the intrinsic measures on the TUNA-REG task are shown in Table 8. As in Section 4.1,

	Dice				MASI				Accuracy			Minimality
	furniture	*people*	*both*	SD	*furniture*	*people*	*both*	SD	*furniture*	*people*	*both*	*both*
GRAPH	0.858	0.729	0.794	0.160	0.705	0.465	0.585	0.272	0.53	0.56	0.40	0.00
JU-PTBSGRE	0.858	0.762	0.810	0.152	0.705	0.501	0.603	0.251	0.55	0.58	0.41	0.00
ATT-DR-b	0.852	0.722	0.787	0.154	0.663	0.441	0.552	0.283	0.52	0.54	0.36	0.00
ATT-DR-sf	0.852	0.722	0.787	0.154	0.663	0.441	0.552	0.283	0.50	0.52	0.36	0.00
DIT-FBI	0.850	0.731	0.791	0.153	0.661	0.451	0.556	0.280	0.50	0.53	0.36	0.00
IS-FP	0.828	0.723	0.776	0.165	0.641	0.475	0.558	0.278	0.52	0.54	0.37	0.00
NIL-UCM-MFVF	0.821	0.684	0.753	0.169	0.601	0.383	0.492	0.290	0.44	0.46	0.31	0.00
USP-EACH-FREQ	0.820	0.663	0.742	0.176	0.616	0.404	0.510	0.291	0.46	0.48	0.33	0.00
DIT-TVAS	0.814	0.684	0.749	0.166	0.580	0.383	0.482	0.285	0.43	0.46	0.29	0.00
OSU-GP	0.640	0.443	0.541	0.226	0.352	0.114	0.233	0.227	0.17	0.20	0.06	0.00
ATT-FB-m	0.357	0.263	0.310	0.245	0.164	0.119	0.141	0.125	0.13	0.14	0.00	1.00
ATT-FB-f	0.231	0.307	0.269	0.215	0.093	0.138	0.116	0.104	0.13	0.12	0.00	1.00
ATT-FB-sf	0.231	0.307	0.269	0.215	0.093	0.138	0.116	0.104	0.13	0.12	0.00	1.00
ATT-FB-sr	0.231	0.307	0.269	0.215	0.093	0.138	0.116	0.104	0.13	0.12	0.00	1.00

Table 3: Descriptives for the TUNA-AS task. All means are shown by entity type; standard deviations are displayed overall.

Dice				MASI					
ATT-FB-f	A			ATT-FB-f	A				
ATT-FB-sf	A			ATT-FB-sf	A				
ATT-FB-sr	A			ATT-FB-sr	A				
ATT-FB-m	A			ATT-FB-m	A	B			
OSU-GP		B		OSU-GP		B			
USP-EACH-FREQ			C	DIT-TVAS			C		
DIT-TVAS			C	NIL-UCM-MFVF			C	D	
NIL-UCM-MFVF			C	USP-EACH-FREQ			C	D	E
IS-FP			C	ATT-DR-b			C	D	E
ATT-DR-b			C	ATT-DR-sf			C	D	E
ATT-DR-sf			C	DIT-FBI			C	D	E
DIT-FBI			C	IS-FP			C	D	E
GRAPH			C	GRAPH				D	E
JU-PTBSGRE			C	JU-PTBSGRE					E

Table 4: Homogeneous subsets for systems in TUNA-AS. Systems which do not share a common letter are significantly different at $p \leq .05$

	Edit				Accuracy			NIST	BLEU
	furniture	*people*	*both*	SD	*furniture*	*people*	*both*	*both*	*both*
IS-GT	7.750	9.768	8.759	6.319	0.02	0.00	0.01	0.4526	0.0415
NIL-UCM-BSC	7.411	9.143	8.277	6.276	0.05	0.04	0.04	1.7034	0.0784
ATT-1-R	7.143	9.268	8.205	6.140	0.02	0.00	0.01	0.1249	0
DIT-CBSR	7.054	10.286	8.670	6.873	0.09	0.02	0.05	1.1623	0.0686
DIT-RBR	6.929	9.857	8.393	6.668	0.04	0.00	0.02	0.9151	0.0694

Table 6: Descriptives for the TUNA-R task.

	Edit				Accuracy			BLEU	NIST
	furniture	*people*	*both*	SD	*furniture*	*people*	*both*	*both*	*both*
ATT-PermuteRank-ws	8.339	8.304	8.321	3.283	0.00	0	0	0.007	0.0288
ATT-Template-ws	8.304	8.161	8.232	3.030	0.00	0	0	0	0.0059
ATT-Dependency-ws	8.232	8.000	8.116	3.023	0.00	0	0	0.0001	0.0139
ATT-TemplateS-ws	8.214	8.161	8.188	3.063	0.00	0	0	0	0.0057
OSU-GP	7.964	13.232	10.598	4.223	0.00	0	0	1.976	0.0236
ATT-PermuteRank-drws	7.464	8.411	7.938	3.431	0.02	0.04	0.03	0.603	0.0571
DIT-TVAS-RBR	6.893	8.161	7.527	3.358	0.05	0	0.03	1.0233	0.0659
ATT-TemplateS-drws	6.786	7.679	7.232	3.745	0.07	0.02	0.04	0.6786	0.0958
ATT-Template-drws	6.768	7.696	7.232	3.757	0.07	0.02	0.04	0.6083	0.0929
NIL-UCM-FVBS	6.643	8.411	7.527	3.618	0.07	0.04	0.05	1.8277	0.0684
IS-FP-GT	6.607	7.304	6.955	3.225	0.05	0.02	0.04	0.8708	0.1086
DIT-FBI-CBSR	6.536	7.643	7.089	3.889	0.16	0.05	0.11	0.8804	0.1259
ATT-Dependency-drws	6.482	7.446	6.964	3.349	0.07	0	0.04	0.3427	0.0477
GRAPH	5.946	9.018	7.482	3.541	0.18	0	0.09	1.141	0.0696

Table 8: Descriptives for TUNA-REG on the intrinsic measures.

means for the intrinsic measures average over both sets of reference outputs in Test Set 2.

A 15 (SYSTEM) $\times 2$ (ENTITY TYPE) univariate ANOVA was conducted on the Edit Distance scores. There were significant main effects of SYSTEM ($F(13, 1540) = 8.6, p < .001$) and ENTITY TYPE ($F(1, 1540) = 47.5, p < .001$), as well as a significant interaction ($F(13, 1540) = 5.77, p < .001$). A Kruskall-Wallis test on Accuracy, coded as an indicator variable (see Section 4.2), showed that systems differed significantly on this measure as well ($\chi^2 = 26.27, p < .05$).

Post-hoc Tukey's comparisons were conducted on Edit Distance; the homogeneous subsets are shown in Table 9. The table suggests that the main effect of Edit Distance may largely have been due to the

difference between OSU-GP and all other systems.

Correlations between these measures are shown in Table 10. Contrary to the results in Section 4.2, the correlation between BLEU and NIST does not reach significance. The negative correlations between Edit distance and Accuracy, and between Edit and BLEU are as expected, since higher Edit cost implies greater distance from a reference output.

IS-FP-GT	A	
ATT-Dependency-drws	A	
DIT-FBI-CBSR	A	
ATT-Template-drws	A	
ATT-TemplateS-drws	A	
GRAPH-4+B	A	
DIT-TVAS-RBR	A	
NIL-UCM-FVBS	A	
ATT-PermuteRank-drws	A	
ATT-Dependency-ws	A	
ATT-TemplateS-ws	A	
ATT-Template-ws	A	
ATT-PermuteRank-ws	A	
OSU-GP		B

Table 9: Homogeneous subsets for systems in TUNA-REG, Edit Distance measure. Systems which do not share a common letter are significantly different at $p \leq .05$

	Edit	**Accuracy**	NIST	BLEU
Edit		-0.584*	0.250	-0.636*
Accuracy	-0.584*		0.383	0.807**
NIST	0.250	0.383		0.371
BLEU	-0.636*	0.807**	0.371	

Table 10: Correlations for TUNA-REG (* indicates $p \leq .05$; ** indicates $p \leq .01$).

4.3.2 Tests on the extrinsic measures

Table 11 displays the results for the extrinsic measures. Reading time (RT), identification time (IT) and error rate (ER), are displayed only for the systems that participated in the evaluation experiment, as well as for the two sets of reference outputs HUMAN-1 and HUMAN-2.

Separate univariate ANOVAs were conducted testing the effect of SYSTEM and ENTITY TYPE on IT and RT. For IT, there was a significant main effect of SYSTEM ($F(11, 1409) = 5.66, p < .001$) and ENTITY TYPE ($F(1, 1409) = 23.507, p < .001$), as well as a significant interaction ($F(11, 1409) = 2.378, p < .05$). The same pattern held for RT, with a main effect of SYSTEM ($F(11, 1412) = 9.95, p < .001$) and ENTITY TYPE ($F(1, 1412) = 9.74, p < .05$) and a significant interaction ($F(11, 1412) = 2.064, p < .05$). A Kruskall-Wallis test conducted on ER showed a significant impact of SYSTEM on the extent to which experimental participants identified the wrong referents ($\chi^2 = 35.45, p < .001$). The homogeneous subsets yielded by post-hoc Tukey's comparisons among systems, on both RT and IT, are displayed in Table 12.

Finally, pairwise correlations were estimated between all three extrinsic measures. The only significant correlation was between RT and IT ($r = .784, p < .05$), suggesting that the longer experimental subjects took to read a description, the longer they also took to identify the target referent.

5 Conclusion

The first ASGRE Challenge, held in 2007, was regarded and presented as a pilot event, for a research community in which there was growing interest in comparative evaluation on shared datasets. Referring Expression Generation was an ideal starting point, because of its relatively long history within the NLG community, and the widespread agreement on inputs, outputs and task definitions.

The tasks described and evaluated in this report constitute a broadening of scope over the 2007 Challenge. Like the previous Challenge, the 2008 edition emphasised diversity in terms of the measures of quality used. This year, there was also an increased emphasis on broadening the range of tasks, with the inclusion of realisation and end-to-end referring expressions generation. This extends the scope of the REG problem, which has traditionally been focussed on content determination (attribute selection) for the most part. As for evaluation, the diversity of measures can shed light on different aspects of quality in these tasks. The fact that the correlation among measures based on different quality criteria is not straightforward is in itself an argument for maintaining this diversity, particularly as comparative evaluation exercises such as this one provide the opportunity for further investigation of the nature of these relationships.

Another indicator of the growing diversity in this year's Challenge is the range of algorithmic solutions in the three tasks, ranging from new models based on classical algorithms, to data-driven methods, evolutionary algorithms, and graph- and tree-based frameworks. The body of work represented by submissions to the TUNA-R and TUNA-REG tasks is also interesting for its exploration of how to apply

	RT				IT				ER		
	furniture	*people*	*both*	SD	*furniture*	*people*	*both*	SD	*furniture*	*people*	*both*
HUMAN-1	2155.376	2187.737	2171.693	2036.462	1973.369	1911.742	1942.297	809.5139	11.864	6.780	9.322
OSU-GP	2080.532	3204.198	2637.644	1555.003	2063.441	2274.690	2167.275	682.8325	6.667	18.966	12.712
HUMAN-2	1823.553	2298.467	2061.010	1475.005	1873.621	1945.880	1909.750	761.3386	16.667	5.000	10.833
ATT-PremuteRank-drws	1664.911	1420.087	1543.528	1392.729	1765.731	1719.456	1742.788	675.3462	10.000	8.475	9.244
DIT-FBI-CBSR	1581.535	1521.799	1551.667	1170.031	1528.119	1932.806	1732.163	694.9878	10.169	10.000	10.084
NIL-UCM-FVBS	1561.291	1933.833	1747.562	1428.490	1531.378	1723.148	1627.263	672.9894	6.667	3.333	5.000
GRAPH	1499.582	1516.804	1508.193	952.158	1706.153	2026.268	1866.211	704.0210	5.000	5.000	5.000
DIT-TVAS-RBR	1485.149	1442.573	1463.861	998.332	1559.953	1734.853	1647.403	588.4615	8.333	13.333	10.833
ATT-Dependency-drws	1460.152	1583.887	1522.019	1177.817	1505.059	2078.336	1791.697	725.9459	1.667	18.333	10.000
ATT-TemplateS-drws	1341.245	1641.539	1490.130	1098.304	1656.401	1720.365	1687.841	650.8357	3.333	10.345	6.780
IS-FT-GT	1292.754	1614.712	1453.733	1374.652	1616.855	1884.557	1750.706	732.4362	6.667	1.667	4.167
ATT-PermuteRank-ws	1218.136	1450.603	1334.369	1203.975	1876.680	1831.485	1854.082	688.3493	31.667	13.333	22.500

Table 11: Descriptives for the extrinsic measures in TUNA-REG.

IT				RT				
NIL-UCM-FVBS	A			ATT-PermuteRank-ws	A			
DIT-TVAS-RBR	A			IS-FT-GT	A			
ATT-TemplateS-drws	A	B		DIT-TVAS-RBR	A			
DIT-FBI-CBSR	A	B		ATT-TemplateS-drws	A			
ATT-PremuteRank-drws	A	B		GRAPH-4+B	A	B		
IS-FT-GT	A	B		ATT-Dependency-drws	A	B		
ATT-Dependency-drws	A	B		ATT-PremuteRank-drws	A	B		
ATT-PermuteRank-ws	A	B		DIT-FBI-CBSR	A	B		
GRAPH-4+B	A	B		NIL-UCM-FVBS	A	B	C	
HUMAN-2	A	B	C	HUMAN-2		B	C	
HUMAN-1		B	C	HUMAN-1			C	D
OSU-GP			C	OSU-GP				D

Table 12: Homogeneous subsets for systems in TUNA-REG, extrinsic time measures. Systems which do not share a common letter are significantly different at $p \leq .05$

realisation techniques to the specific problem posed by referring expressions.

The outcomes of this evaluation exercise are obviously not intended to be a 'final word' on the right way to carry out evaluation in referring expressions generation. Rather, comparative results open up the possibility of improvement and change. Another important aspect of a shared task of this nature is that it results in an archive of data that can be further exploited, either through follow-up studies, or for the provision of baselines against which to compare novel approaches. We have already used the data from ASGRE 2007 for further investigation, particularly in the area of extrinsic evaluation. We plan to carry out more such studies in the future.

Acknowledgements

Our heartfelt thanks to the participants who helped to make this event a success. Thanks to Advaith Siddharthan, who proposed MASI for TUNA-AS.

References

A. Belz and A. Gatt. 2007. The attribute selection for gre challenge: Overview and evaluation results. In *Proc. UCNLG+MT: Language Generation and Machine Translation*.

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.

A. Gatt and A. Belz. 2008. Attribute selection for referring expression generation: New algorithms and evaluation methods. In *Proceedings of the 5th International Conference on Natural Language Generation (INLG-08)*.

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

S. Papineni, T. Roukos, W. Ward, and W. Zhu. 2002. Bleu: a. method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 311–318.

R. Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proc. 5th International Conference on Language Resources and Evaluation (LREC-06)*.

R. Ratliff. 1993. Methods for dealing with reaction time outliers. *Psychological Bulletin*, 114(3):510–532.