The GREC Challenge: Overview and Evaluation Results

Anja Belz **Eric Kow**
NLT Group
University of Brighton
Brighton BN2 4GJ, UK
{asb,eykk10}@bton.ac.uk

Jette Viethen
Centre for LT
Macquarie University
Sydney NSW 2109
jviethen@ics.mq.edu.au

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

Abstract

The GREC Task at REG'08 required participating systems to select coreference chains to the main subject of short encyclopaedic texts collected from Wikipedia. Three teams submitted a total of 6 systems, and we additionally created four baseline systems. Systems were tested automatically using a range of existing intrinsic metrics. We also evaluated systems extrinsically by applying coreference resolution tools to the outputs and measuring the success of the tools. In addition, systems were tested in a reading/comprehension experiment involving human subjects. This report describes the GREC Task and the evaluation methods, gives brief descriptions of the participating systems, and presents the evaluation results.

1 Introduction

The GREC task is about how to generate appropriate references to an entity in the context of a piece of discourse longer than a sentence. Rather than requiring participants to generate referring expressions from scratch, the GREC data provides sets of possible referring expressions for selection. As this is a new referring expression generation (REG) task, the shared task definition was kept fairly simple and the aim for participating systems was to select the appropriate *type* of referring expression (more specifically, its `REG08-TYPE`, full details below).

The immediate *motivating application context* for the GREC Task is the improvement of referential clarity and coherence in extractive summarisation by regenerating referring expressions in summaries. There has recently been a small flurry of work in this area (Steinberger et al., 2007; Nenkova, 2008). In the longer term, the GREC Task is intended to be a step in the direction of the more general task of generating referential expressions in discourse context.

The GREC Task Corpus is an extension of GREC 1.0 which had about 1,000 texts in the subdomains of cities, countries, rivers and people (Belz and Varges, 2007a). for the purpose of the REG'08 GREC Task, we obtained an additional 1,000 texts in the new subdomain of mountain texts and developed a new XML annotation scheme (Section 2.2).

Five teams from four countries registered for the GREC Task, of which three teams eventually submitted 6 systems. We also used the corpus texts themselves as 'system' outputs, and created four baseline systems. We evaluated the resulting 10 systems using a range of intrinsic and extrinsic evaluation methods. This report presents the results of all evaluations (Section 6), along with descriptions of the GREC data and task (Section 2), test sets (Section 3), evaluation methods (Section 4), and participating systems (Section 5).

2 Data and Task

The GREC Corpus (version 2.0) consists of about 2,000 texts in total, all collected from introductory sections in Wikipedia articles, in five different domains (cities, countries, rivers, people and mountains). In each text, three broad categories of Main Subject Reference (MSR)[1] have been annotated, re-

[1]The main subject of a Wikipedia article is simply taken to be given by its title, e.g. in the cities domain the main subject

sulting in a total of about 13,000 annotated REs.

The corpus was randomly divided into 90% training data (of which 10% were randomly selected as development data) and 10% test data. Participants used the training data in developing their systems, and (as a minimum requirement) reported results on the development data. Participants had 48 hours to submit outputs for the (previously unseen) test data.

2.1 Types of referential expression annotated

Three broad categories of main subject referring expression (MSREs) are annotated in the GREC corpus[2] — subject NPs, object NPs, and genitive NPs and pronouns which function as subject-determiners within their matrix NP. These categories of referring expression (RE) are relatively straightforward to identify and achieve high inter-annotator agreement on (complete agreement among four annotators in 86% of MSRs), and account for most cases of overt main subject reference (MSR) in the GREC texts. The annotators were asked to identify subject, object and genitive subject-determiners and decide whether or not they refer to the main subject of the text. More detail is provided in Belz and Varges (2007b).

In addition to the above, relative pronouns in supplementary relative clauses (as opposed to integrated relative clauses, Huddleston and Pullum, 2002, p. 1058) were annotated, e.g.:

(1) *Stoichkov is a football manager and former striker who was a member of the Bulgaria national team that finished fourth at the 1994 FIFA World Cup.*

We also annotated 'non-realised' subject MSREs in a restricted set of cases of VP coordination where an MSRE is the subject of the coordinated VPs, e.g.:

(2) *He stated the first version of the Law of conservation of mass, __ introduced the Metric system, and __ helped to reform chemical nomenclature.*

The motivation for annotating the approximate place where the subject NP would be if it were realised (the gap-like underscores above) is that from a generation perspective there is a choice to be made about whether to realise the subject NP in the second and third coordinates or not.

(and title) of one text is *London*.

[2] In terminology and view of grammar the annotations rely heavily on Huddleston and Pullum (2002).

2.2 XML format

Figure 1 is one of the texts distributed in the GREC data sample for the REG Challenge. The `REF` element indicates a reference, in the sense of 'an instance of referring' (which could, in principle, be realised by gesture or graphically, as well as by a string of words, or a combination of these). `REF`s have three attributes: `ID`, a unique reference identifier; `SEMCAT`, the semantic category of the referent, ranging over `city`, `country`, `river`, `person`, `mountain`; and `SYNCAT`, the syntactic category required of referential expressions for the referent in this discourse context (`np-obj`, `np-subj`, `subj-det`). A `REF` is composed of one `REFEX` element (the 'selected' referential expression for the given reference; in the corpus texts it is simply the referential expression found in the corpus) and one `ALT-REFEX` element which in turn is a list of `REFEX`s which are alternative referential expressions obtained by other means (see following section).

`REFEX` elements have four attributes. The `HEAD` attribute has the possible values `nominal`, `pronoun`, and `rel-pron`; the `CASE` attribute has the possible values `nominative`, `accusative` and `genitive` for pronouns, and `plain` and `genitive` for nominals. The binary-valued `EMPHATIC` attribute indicates whether the RE is emphatic; in the present version of the GREC corpus, the only type of RE that has this attribute is one which incorporates a reflexive pronoun used emphatically (e.g. *India itself*). The `REG08-TYPE` attribute indicates basic RE type as required for the REG'08 GREC task definition. The choice of types is motivated by the hypothesis that one of the most basic decisions to be taken in RE selection for named entities is whether to use an RE that includes a name, such as *Modern India* (the corresponding `REG08-TYPE` value is `name`); whether to go for a common-noun RE, i.e. with a category noun like *country* as the head (`common`); whether to pronominalise the RE (`pronoun`); or whether it can be left unrealised (`empty`).

2.3 The REG'08 GREC Task

The task for participating systems was to develop a method for selecting one of the `REFEX`s in the `ALT-REFEX` list, for each `REF` in each `TEXT` in the test sets. The test data inputs were identical to the

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE TEXT SYSTEM "reg08-grec.dtd">
<TEXT ID="36">
<TITLE>Jean Baudrillard</TITLE>
<PARAGRAPH>
 <REF ID="36.1" SEMCAT="person" SYNCAT="np-subj">
  <REFEX REG08-TYPE="name" EMPHATIC="no" HEAD="nominal" CASE="plain">Jean Baudrillard</REFEX>
  <ALT-REFEX>
   <REFEX REG08-TYPE="name" EMPHATIC="no" HEAD="nominal" CASE="plain">Jean Baudrillard</REFEX>
   <REFEX REG08-TYPE="name" EMPHATIC="yes" HEAD="nominal" CASE="plain">Jean Baudrillard himself</REFEX>
   <REFEX REG08-TYPE="empty">_</REFEX>
   <REFEX REG08-TYPE="pronoun" EMPHATIC="no" HEAD="pronoun" CASE="nominative">he</REFEX>
   <REFEX REG08-TYPE="pronoun" EMPHATIC="yes" HEAD="pronoun" CASE="nominative">he himself</REFEX>
   <REFEX REG08-TYPE="pronoun" EMPHATIC="no" HEAD="rel-pron" CASE="nominative">who</REFEX>
   <REFEX REG08-TYPE="pronoun" EMPHATIC="yes" HEAD="rel-pron" CASE="nominative">who himself</REFEX>
  </ALT-REFEX>
 </REF>
 (born June 20, 1929) is a cultural theorist, philosopher, political commentator,
 sociologist, and photographer.
 <REF ID="36.2" SEMCAT="person" SYNCAT="subj-det">
  <REFEX REG08-TYPE="pronoun" EMPHATIC="no" HEAD="pronoun" CASE="genitive">His</REFEX>
  <ALT-REFEX>
   <REFEX REG08-TYPE="name" EMPHATIC="no" HEAD="nominal" CASE="genitive">Jean Baudrillard's</REFEX>
   <REFEX REG08-TYPE="pronoun" EMPHATIC="no" HEAD="pronoun" CASE="genitive">his</REFEX>
   <REFEX REG08-TYPE="pronoun" EMPHATIC="no" HEAD="rel-pron" CASE="genitive">whose</REFEX>
  </ALT-REFEX>
 </REF>
 work is frequently associated with postmodernism and post-structuralism.
</PARAGRAPH>
</TEXT>
```

Figure 1: Example text from REG'08 Training Data.

training/development data, except that REF elements contained only an ALT-REFEX list, not the preceding 'selected' REFEX. ALT-REFEX lists are generated for each text by an automatic method which collects all the (manually annotated) MSREs in a text including the title and adds several defaults: pronouns and reflexive pronouns in all subdomains; and category nouns (e.g. *the river*), in all subdomains except people. The main objective in the REG'08 GREC Task was to get the REG08-TYPE attribute of REFEXs right.

3 Test Data

1. GREC Test Set C-1: a randomly selected 10% subset (183 texts) of the GREC corpus (with the same proportions of texts in the 5 subdomains as in the training/testing data).

2. GREC Test Set C-2: the same subset of texts as in C-1; however, for C-2 we did not use the MSREs in the corpus, but replaced each of them with three human-selected alternatives. These were obtained in an online experiment as described in Belz & Varges (2007a) where subjects selected MSREs in a setting that duplicated the conditions in which the participating systems in the REG'08 GREC Task made selections.[3] We obtained three versions of each text, where in each version all MSREs were selected by the same person. The motivation for creating this version of Test Set C was firstly that having several human-produced chains of MSREs to compare the outputs of participating ('peer') systems against is more reliable than having one only; and secondly that Wikipedia texts are edited by multiple authors and so MSR chains may sometimes be adversely affected by this; we wanted to have additional reference texts without this characteristic.

3. GREC Test Set L: 74 Wikipedia introductory texts from the subdomain of lakes; participants did not know what this subdomain was until they received the test data (there were no lake texts in the training/development set).

4. GREC Test Set P: 31 short encyclopaedic texts in the same 5 subdomains as in the GREC corpus, in approximately the same proportions as in the training/testing data, but from a source other than

[3]The experiment can be tried out here: http://www.nltg.brighton.ac.uk/home/Anja.Belz/TESTDRIVE/

Wikipedia. We transcribed these texts from printed encyclopaedias published in the 1980s which are not available in electronic form, and this provenance was not revealed to participants. The texts in this set are much shorter and more homogeneous than the Wikipedia texts, and the sequences of MSRs follow very similar patterns. It seems likely that it is these properties that have resulted in better scores overall for Test Set P (see Section 6).

Each test set was designed to test peer systems for a different aspect of generalisation. Test Set C tests for generalisation to unseen material from the same corpus and the same subdomains as the training set; Test Set L tests for generalisation to unseen material from the same corpus but different subdomain; and Test Set P tests generalisation to a different corpus but same subdomains.

4 Evaluation methods

4.1 Automatic intrinsic evaluations

Accuracy of REG08-Type: when computed against the single-RE test sets (C-1, L and P), REG08-Type Accuracy is the proportion of REFEXs selected by a participating system that have a REG08-TYPE value identical to the one in the corpus.

When computed against the triple-RE test set (C-2), first the number of correct REG08-Types is computed at the text level for each of the three versions of a corpus text and the maximum of these is determined; then the maximum text-level numbers are summed and divided by the total number of REFs in all the texts, which gives the global REG08-Type Accuracy score. The rationale behind computing the REG08-Type Accuracy scores in this way for multiple-RE test sets (maximising scores on RE chains rather than individual REs) is that an RE is not good or bad in its own right, but depends on the other MSRs in the same text.[4]

String Accuracy: This is defined just like REG08-Type Accuracy, except here what is determined is identity between REFEX word strings (the MSREs themselves), not between REG08-Types.

String-edit distance metrics: String-edit distance (SE) is straightforward Levenshtein distance with a substitution cost of 2 and insertion/deletion cost of 1. We also used the version of string-edit distance described by Bangalore et al. (2000) which normalises for length. This version is denoted 'SEB' below. For the single-RE test sets, the global score is simply the average of all RE-level scores. For Test Set C-2, we used an approach analogous to that described above for REG08-Type Accuracy. We first computed the best string-edit distance at the text level (here, just the sum of RE-level distances) and then obtained the global distance by dividing the sum of best text-level distances by the number of REFs in all the texts.

Other metrics: BLEU is a precision metric from MT that assesses the quality of a peer translation in terms of the proportion of its word n-grams ($n \leq 4$ is standard) that it shares with several reference translations. We used BLEU-3 rather than the more standard BLEU-4 because most REs in the corpus are less than 4 tokens long. We also used the NIST version of BLEU which weights in favour of less frequent n-grams, as well as ROUGE-2 and ROUGE-SU4 (the two official automatic scores from the DUC summarisation competitions). In all cases, we assessed just the MSREs selected by peer systems (leaving out the surrounding text), and computed scores globally (rather than averaging over RE-level scores), as this is standard for these metrics.

BLEU, NIST and ROUGE are designed to work with either one or multiple reference texts, so we did not need to use a different method for Test Set C-2.

4.2 Human extrinsic evaluation

We designed a reading/comprehension experiment in which the task for subjects was to read texts one sentence at a time and then to answer three brief multiple-choice comprehension questions after reading each text. The basic idea was that it seemed likely that badly chosen MSR reference chains would adversely affect ease of comprehension, and that this might in turn affect reading speed and accuracy in answering comprehension questions.

We used a randomly selected subset of 21 texts from Test Set C, and recruited 21 subjects from among the staff, faculty and students of Brighton and Sussex universities. We used a Repeated Latin Squares design in which each combination of text and system was allocated three trials. During the experiment we recorded *SRTime*, the time subjects

[4] This definition is also slightly different from the one given in the Participants' Pack.

took to read sentences (from the point when the sentence appeared on the screen to the point at which the subject requested the next sentence).

We also recorded the speed and accuracy with which subjects answered the questions at the end (*Q-Time* and *Q-Acc*). The role of the comprehension questions was to encourage subjects to read the texts properly, rather than skimming through them, and we did not necessarily expect any significant results from the associated measures.

The questions were designed to be of varying degrees of difficulty and predictability. There was one set of three questions (each with five possible answers) associated with each text, and questions followed the same pattern across the texts: the first question was always about the subdomain of a text; the second about the location of the main subject; the third question was designed not to be predictable.

The order of the answers was randomised for each question and each subject. The order of texts (with associated questions) was randomised for each subject. We used the DMDX package for presentation of sentences and measuring reading times and question answering accuracy (Forster and Forster, 2003). Subjects did the experiment in a quiet room, under supervision.

4.3 Automatic extrinsic evaluation

As a new and highly experimental method, we tried out an automatic approach to extrinsic evaluation. The basic idea was similar to that in the human-based experiments described above: badly chosen reference chains seem likely to affect the reader's ability to resolve REs. In the automatic version, the role of the reader is played by an automatic coreference resolution tool and the expectation is that the tool performs worse (are less able to identify coreference chains correctly) with worse MSR reference chains.

To counteract the potential problem of results being a function of a specific coreference resolution algorithm or tool, we decided to use three different resolvers—those included in LingPipe,[5] JavaRap (Qiu et al., 2004) and OpenNLP (Morton, 2005)—and to average results.

There does not appear to be a single standard evaluation metric in the coreference resolution community, so we opted to use three: MUC-6 (Vilain et al., 1995), CEAF (Luo, 2005), and B-CUBED (Bagga and Baldwin, 1998), which seem to be the most widely accepted metrics.

All three metrics compute Recall, Precision and F-Scores on aligned gold-standard and resolver-tool coreference chains. They differ in how the alignment is obtained and what components of coreference chains are counted for calculating scores. Results for the automatic extrinsic evaluations are reported below in terms of the F-Scores from these three metrics, as well as in terms of their average.

5 Systems

Base-rand, Base-freq, Base-1st, Base-name: We created four baseline systems. *Base-rand* selects one of the `REFEX`s at random. *Base-freq* selects the `REFEX` that is the overall most frequent given the `SYNCAT` and `SEMCAT` of the reference. *Base-1st* always selects the `REFEX` which appears first in the list of `REFEX`s; and *Base-name* selects the shortest `REFEX` with attributes `REG08-TYPE=name`, `HEAD=nominal` and `EMPHATIC=no`.[6]

CNTS-Type-g, CNTS-Prop-s: The CNTS systems are trained using memory-based learning with automatic parameter optimisation. They use a set of 14 features obtained by various kinds of syntactic preprocessing and named-entity recognition as well as from the corpus annotations: `SEMCAT`, `SYNCAT`, position of RE in text, neighbouring words and POS-tags, distance to previous mention, `SYNCAT`s of three preceding `REFEX`s, binary feature indicating whether the most recent named entity was the main subject (MS), main verb of the sentence. For *Type-g*, a single classifier was trained to predict just the `REG08-TYPE` property of `REFEX`s. For *Prop-s*, four classifiers were trained, one for each subdomain, to predict all four properties of `REFEX`s (rather than just `REG08-TYPE`).

OSU-b-all, OSU-b-nonRE, OSU-n-nonRE: The OSU-2 systems are maximum-entropy classifiers trained on a range of features obtained by prepro-

[5] `http://alias-i.com/lingpipe/`

[6] Attributes are tried in this order. If for one attribute, the right value is not found, the process ignores that attribute and moves on the next one.

System	REG08-Type Accuracy for Development Set					
	All	Cities	Coun	Riv	Peop	Moun
CNTS-Type-g	76.52	64.65	75	65	85.37	75.42
CNTS-Prop-s	73.93	65.66	69.57	70	79.51	74.58
IS-G	66	54.5	64	80	66.8	65
OSU-n-nonRE	62.50	53.54	63.04	65	67.32	61.67
OSU-b-all	58.54	53.54	57.61	75	65.85	49.58
OSU-b-nonRE	51.07	51.52	53.26	40	57.07	45.83

Table 1: Self-reported REG08-Type Accuracy scores for development set.

cessing the text, as well as from the corpus annotations: SEMCAT, SYNCAT, position of RE in text, presence of contrasting discourse entity, distance between current and preceding reference to the MS, string similarity measures between REFEXs and title of text. *OSU-b-all* and *OSU-b-nonRE* are binary classifiers which give the likelihood of selecting a given REFEX vs. not selecting it, whereas *OSU-n-nonRE* is a 4-class classifier giving the likelihoods of selecting each of the four REG08-TYPEs. *OSU-b-all* also uses the REFEX attributes as features.

IS-G: The IS-G system is a multi-layer perceptron which uses four features obtained by preprocessing texts as well as from the corpus annotations: SYNCAT, distance between current and preceding reference to the MS, position of RE in text, REG08-TYPE of preceding reference to the MS, feature indicating whether the preceding MSR is in the same sentence.

6 Results

This section presents the results of all the evaluation methods described in Section 4. We start with REG08-Type Accuracy, an intrinsic automatic metric which participating teams were told was going to be the chief evaluation method, followed by other intrinsic automatic metrics (Section 6.2), the extrinsic human evaluation (Section 6.3) and the extrinsic automatic evaluation (Section 6.4).

6.1 REG08-Type Accuracy

Participants computed REG08-Type Accuracy for the development set (97 texts) themselves, using a tool provided by us. These scores are shown in Table 1, and are also included in the participants' reports elsewhere in this volume. Systems are ordered in terms of their overall REG08-Type Accuracy (column 1), and scores for each subdomain are also shown. Scores are highly consistent across the subdomains, except for the river subdomain which was the smallest set (containing only 4 texts), and results for it may be idiosyncratic for this reason.

Corresponding results for the (unseen) test set C-1 are shown in column 2 of Table 2. As would be expected, results are slightly worse than for the (seen) development set (although some systems managed to improve over their development set scores). Also included in this table are results for the four baseline systems, and it is clear that selecting the most frequent REG08-Type given SEMCAT and SYNCAT (as done by the Base-freq system) provides a strong baseline.

Other columns in Table 2 contain results for test sets L and P. Again as expected, results for Test Set L are lower than for Test Set C-1, because in addition to consisting of unseen texts (like C-1), Test Set L is also from an unseen subdomain (unlike C-1). The results for Test Set P are higher and on a par with those for the development set, probably for the reasons discussed at the end of Section 3.

For each test set in Table 2 we carried out a univariate ANOVA with System as the fixed factor. We found significant main effects at $p < .001$ in all three cases (C-1: $F = 95.426$; L: $F = 63.758$; P: $F = 21.188$). The columns containing capital letters in Table 2 show the homogeneous subsets of systems as determined by post-hoc Tukey HSD comparisons of means. Systems whose REG08-Type Accuracy scores are not significantly different (at the .05 level) share a letter.

The results for REG08-Type Accuracy computed against the triple-RE Test Set C-2 are shown in Table 3. These should be considered as the chief results of the GREC Task evaluations, as stated in the guidelines. Here too we performed a univariate ANOVA with System as the fixed factor and REG08-Type as the dependent variable. Having established by ANOVA that there was a significant main effect of System ($F = 86.946$, $p < .001$), we compared the mean scores with Tukey's HSD. As can be seen from the resulting homogeneous subsets, there is no significant difference between the corpus texts (C-1) and system CNTS-Type-g, but also there is no sig-

single-RE Test Set C-1							Test Set L						Test Set P						
CNTS-Type-g	68.15	A					CNTS-Type-g	62.06	A				CNTS-Type-g	75.31	A				
CNTS-Prop-s	67.04	A					CNTS-Prop-s	62.06	A				CNTS-Prop-s	72.84	A	B			
IS-G	66.48	A					IS-G	60.93	A				IS-G	67.90	A	B	C		
OSU-n-nonRE	63.69	A					OSU-n-nonRE	41.80		B			OSU-n-nonRE	66.67	A	B	C		
OSU-b-nonRE	53.11		B				OSU-b-nonRE	39.23		B			OSU-b-all	57.41		B	C	D	
OSU-b-all	52.39		B				OSU-b-all	37.62		B	C		OSU-b-nonRE	56.17			C	D	
Base-freq	43.47			C			Base-freq	35.53		B	C		Base-freq	44.44				D	F
Base-name	39.49			C			Base-rand	23.63			C	D	Base-rand	33.95					F
Base-1st	39.17			C			Base-name	23.63				D	Base-name	32.10					F
Base-rand	32.72				D		Base-1st	29.74				D	Base-rand	32.10					F

Table 2: REG08-Type Accuracy scores and homogeneous subsets (Tukey HSD, alpha = .05) for single-RE test sets. Systems that do not share a letter are significantly different.

System	REG08-Type Accuracy for multiple-RE Test Set C-2										
	All						Cities	Countries	Rivers	People	Mountains
Corpus	*78.58*	*A*					*70.92*	*77.49*	*85.29*	*84.67*	*75.81*
CNTS-Type-g	72.61	A	B				65.96	71.73	73.53	77.64	70.73
CNTS-Prop-s	71.34		B				64.54	67.02	70.59	75.38	71.75
IS-G	70.78		B				69.50	65.45	76.47	76.88	67.89
OSU-n-nonRE	69.82		B				65.25	64.92	79.41	78.14	65.65
OSU-b-nonRE	58.76			C			52.48	60.73	50.00	59.80	59.55
OSU-b-all	57.48			C			53.90	58.64	47.06	59.05	57.52
Base-name	50.00				D		53.19	54.45	35.29	43.22	53.86
Base-1st	49.28				D		53.19	49.21	38.24	43.22	53.86
Base-freq	48.17				D		43.97	42.41	55.88	56.78	44.11
Base-rand	41.24					E	41.84	36.13	32.35	44.47	41.06

Table 3: REG08-Type Accuracy scores against Test Set C-2 for complete set and for subdomains; homogeneous subsets (Tukey HSD, alpha = .05) for complete set only (systems that do not share a letter are significantly different).

nificant difference between the latter and systems CNTS-Prop-s, IS-G and OSU-n-nonRE. In this analysis, all systems outperform the random baseline; all peer systems outperform all of the baselines; and the four best peer systems outperform the remaining two.

6.2 Other automatic intrinsic metrics

In addition to the chief evaluation measure reported on in the preceding section, we computed the string similarity metrics described in Section 4.1 for all four test sets. Results were very similar to those for REG08-Type Accuracy, so we are reporting only scores for Test Set C-2 (Table 4). The corpus texts again receive the best scores across the board (SE is the odd one out, because here lower scores are better). Ranks for peer systems are very similar to the results reported in the last section.

We performed an ANOVA ($F = 138.159$, $p < .001$) and Tukey HSD post-hoc analysis for String Accuracy. The resulting homogeneous subsets (Table 4, columns 3–8) reveal significant differences similar to those for REG08-Type Accuracy. We also computed Pearson product-moment correlation coefficients between all automatic intrinsic evaluation measures we used. All pairwise correlations were significant at the .01 level (using a two-tailed test). One of the strongest correlations (.961) was between REG08-Type Accuracy and String Accuracy, implying that getting REG08-Type right gets you some way towards getting the actual RE right.

6.3 Human-based extrinsic measures

As a result of the experiment described in Section 4.2 we had SRTime measures (sentence reading times) for each sentence in each of the 21 texts that were included in the experiment. Table 5 shows the resulting SRTimes in milliseconds averaged per system. None of the differences were statistically significant. We also analysed SRTimes normalised by sentence length; SRTimes only from sentences that contained MRSs; and SRTimes normalised for subject reading speed. There were no significant differences under any of these analyses.

Much of the variance in SRTimes was due to subjects' very different average reading speeds: means of SRTime normalised for sentence length ranged from 188.45ms to 426.10ms for individual subjects.

System	Word string similarity for Triple-RE Test Set C-2												
	String Accuracy							BLEU-3	NIST	ROUGE-2	ROUGE-SU4	SE	SEB
Corpus	*71.18*	*A*						*0.7792*	*7.5080*	*0.66102*	*0.70991*	*0.7229*	*0.5136*
CNTS-Type-g	65.61	A	B					0.7377	6.1288	0.60280	0.64998	0.8838	0.3627
CNTS-Prop-s	65.29	A	B					0.6760	5.9338	0.60103	0.64963	0.9068	0.3835
OSU-n-nonRE	63.85		B	C				0.6715	5.7745	0.53395	0.57459	0.9666	0.0164
IS-G	58.20			C				0.5107	5.6102	0.50270	0.57052	1.1616	0.1818
OSU-b-nonRE	51.11				D			0.4964	5.5363	0.38255	0.42969	1.2834	0.0247
OSU-b-all	50.72				D			0.5050	5.6058	0.35133	0.39570	1.2994	0.3402
Base-freq	41.32					E		0.2684	3.0155	0.27727	0.33007	1.54299	-0.3250
Base-name	39.41					E		0.4641	5.9372	0.20730	0.25379	1.5175	-0.1912
Base-1st	39.09					E		0.3932	5.1597	0.21443	0.24037	1.6449	-0.0751
Base-rand	17.99						F	0.2182	2.9327	0.36056	0.41847	2.3217	-0.7937

Table 4: String Accuracy, BLEU, NIST, ROUGE and string-edit scores, computed on single-RE and triple-RE test sets (systems in order of String Accuracy); homogeneous subsets (Tukey HSD, alpha = .05) for String Accuracy only (systems that do not share a letter are significantly different).

	Mean SRTime (msecs)
CNTS-Prop-s	6305.8551
IS-G	6340.5131
OSU-n-nonRE	6422.5073
CNTS-Type-g	6435.6574
OSU-b-all	6451.7624
OSU-b-nonRE	6454.6749
Corpus	*6548.2734*

Table 5: Mean SRTimes for each system.

Question 1				Q2	Q3
Corpus	*1.00*	*A*		*.78*	*.63*
CNTS-Type-g	1.00	A		.83	.71
CNTS-Prop-s	.98	A	B	.86	.75
OSU-b-nonRE	.97	A	B	.83	.67
OSU-b-all	.95	A	B	.75	.62
IS-G	.95	A	B	.81	.63
OSU-n-nonRE	.90		B	.76	.76

Table 6: Question types 1–3, proportions correct; homogeneous subsets for Q1 (Tukey HSD, alpha = .05).

There was also variance from Text, i.e. some of the texts appear to be harder to read than others.

The other two measures from the task-performance experiment were Q-Acc (question answering accuracy) and Q-Time (question answering speed). ANOVAs revealed no significant main effect of System on Q-Time. For Q-Acc, we looked at each of the three question types Q1, Q2, Q3 (see Section 4.2) separately. ANOVAs showed no significant effect of System on Q-Acc for Q2 and Q3; there was a slight effect ($F = 2.193$, $p < .05$) of System on Q-Acc for Q1 (the easiest of the questions which simply asked for the subdomain of a text). Table 6 shows Q-Acc for Q1 and Q2, and the results of a post-hoc analysis (Tukey HSD) which revealed two homogeneous subsets with a lot of overlap (columns 2 and 3).

Table 6 shows the results of this analysis: there was

6.4 Automatic extrinsic measures

We used the same 21 texts as in the human extrinsic experiments, fed the outputs of each peer system as well as the corpus texts through the three coreference resolvers, and computed average MUC, CEAF and B-CUBED F-Scores as described in Section 4.3. The second column Table 7 shows the average of these three F-Scores, to give a single overall result for this evaluation method. A univariate ANOVA with the average F-Score (column 2) as the dependent variable and System as the single fixed factor revealed a significant main effect of System on average F-Score ($F = 5.051, p < .001$). A post-hoc comparison of the means (Tukey HSD, alpha = .05) found the significant differences indicated by the homogeneous subsets in columns 3–5 (Table 7). The numbers shown in the last three columns are the separate MUC, CEAF and B-CUBED F-Scores for each system, averaged over the three resolver tools. ANOVAs revealed the following effects of System: on CEAF $F = 9.984, p < .001$; on MUC: $F = 10.07, p < .001$; on B-CUBED: $F = 8.446, p < .001$.

The three F-Score measures (MUC, CEAF and B-CUBED) are all strongly and highly significantly cor-

related: Pearson's correlation coefficient is .947 for B-CUBED and CEAF, .917 for B-CUBED and MUC, and .951 for CEAF and MUC ($p < .01$, 2-tailed).

System	(MUC+CEAF+B3)/3				MUC	CEAF	B3
Base-1st	53.50	A			47.59	52.64	60.28
Base-name	52.84	A			45.99	51.73	60.81
OSU-n-nonRE	51.39	A			46.92	49.8	57.45
OSU-b-nonRE	51.27	A			47.68	48.62	57.50
OSU-b-all	50.87	A			47.06	48.40	57.14
CNTS-Type-g	48.64	A	B		43.77	46.32	55.82
IS-G	48.05	A	B		43.25	46.24	54.66
CNTS-Prop-s	46.35	A	B		42.82	43.36	52.88
Corpus	*43.32*	*A*	*B*		*37.89*	*41.6*	*50.47*
Base-freq	41.41		B	C	34.48	40.28	49.46
Base-rand	35.13			C	21.24	35.60	48.55

Table 7: MUC, CEAF and B-CUBED F-Scores for all systems; homogeneous subsets (Tukey HSD), alpha = .05, for average of F-Scores.

7 Concluding Remarks

The GREC Task is a new task not only for an NLG shared-task challenge, but also as a research task in general (improving referential clarity in extractive summaries seems to be just taking off as a research subfield). It was therefore not unexpected that only three teams were able to participate in this task.

We continued the traditions of the ASGRE'07 Challenge in that we used a wide range of evaluation metrics to obtain a well-rounded view of the quality of the participating systems. It had been our intention to use evaluation methods in all four possible extrinsic/intrinsic and automatic/human combinations. However, the combination intrinsic/human is missing from this report and will have to be left to future research.

There was no indication in the human task performance experiment that the different reference chains selected by different systems had any impact on subjects' reading speeds, and the evidence that there is an effect on comprehension was scant. This means that we will need to investigate alternative task-performance measures. Because of the lack of significant results from the human extrinsic experiment, we were also unable to validate the automatic extrinsic experiment against it, and so at this point we do not really know how useful it is (despite some correlation with intrinsic measures), something we will seek to establish in future research.

Acknowledgments

Many thanks to Jason Baldridge and Pascal Denis for help with selecting coreference resolution tools and metrics, and to the colleagues and students who helped with the task-performance experiment. Thanks are also due to the members of the Corpora and SIGGEN mailing lists, colleagues, friends and friends of friends who helped with the online MSRE selection experiment.

References

A. Bagga and B. Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the Linguistic Coreference Workshop at LREC'98*, pages 563–566.

S. Bangalore, O. Rambow, and S. Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of INLG'00*, pages 1–8.

A. Belz and S. Varges. 2007a. Generation of repeated references to discourse entities. In *Proceedings of ENLG'07*, pages 9–16.

A. Belz and S. Varges. 2007b. The GREC corpus: Main subject reference in context. Technical Report NLTG-07-01, University of Brighton.

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.

R. Huddleston and G. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.

X. Luo. 2005. On coreference resolution performance metrics. *Proc. of HLT-EMNLP*, pages 25–32.

T. Morton. 2005. *Using Semantic Relations to Improve Information Retrieval*. Ph.D. thesis, University of Pensylvania.

A. Nenkova. 2008. Entity-driven rewrite for multi-document summarization. In *Proceedings of IJCNLP'08*.

L. Qiu, M. Kan, and T.-S. Chua. 2004. A public reference implementation of the rap anaphora resolution algorithm. In *Proceedings of LREC'04*, pages 291–294.

J. Steinberger, M. Poesio, M. Kabadjov, and K. Jezek. 2007. Two uses of anaphora resolution in summarization. *Information Processing and Management: Special issue on Summarization*, 43(6):1663–1680.

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. 1995. A model-theoretic coreference scoring scheme. *Proceedings of MUC-6*, pages 45–52.