The GREC Main Subject Reference Generation Challenge 2009: Overview and Evaluation Results

Anja Belz **Eric Kow**
NLT Group
University of Brighton
Brighton BN2 4GJ, UK

Jette Viethen
Centre for LT
Macquarie University
Sydney NSW 2109

Albert Gatt
Computing Science
University of Aberdeen
Aberdeen AB24 3UE, UK

{asb,eykk10}@bton.ac.uk jviethen@ics.mq.edu.au a.gatt@abdn.ac.uk

Abstract

The GREC-MSR Task at Generation Challenges 2009 required participating systems to select coreference chains to the main subject of short encyclopaedic texts collected from Wikipedia. Three teams submitted one system each, and we additionally created four baseline systems. Systems were tested automatically using 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 an intrinsic evaluation involving human judges. This report describes the GREC-MSR Task and the evaluation methods applied, gives brief descriptions of the participating systems, and presents the evaluation results.

1 Introduction

The GREC-MSR 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-MSR data provides sets of possible referring expressions for selection. This was the second time we ran a shared task using the GREC-MSR data (following a first run in 2008). The task definition was again kept fairly simple, but in the 2009 round the main aim for participating systems was to select an appropriate word string to serve as a referring expression, whereas in 2008 it was to select an appropriate *type* of referring expression (name, common noun, pronoun, or empty reference).

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

The GREC-MSR data is an extension of the GREC 1.0 Corpus which had about 1,000 texts in the subdomains of cities, countries, rivers and people (Belz and Varges, 2007a). For the purpose of the GREC-MSR shared task, an additional 1,000 texts in the new subdomain of mountain texts were obtained and a new XML annotation scheme (Section 2.2) was developed.

Team	System Name
University of Delaware	UDel
ICSI, Berkeley	ICSI-CRF
Jadavpur University	JUNLG

Table 1: GREC-MSR'09 participating teams.

Nine teams from seven countries registered for GREC-MSR'09, of which three teams (Table 1) submitted one system each.[1] Participants had to submit their system reports before downloading test data inputs, and had to submit test data outputs within 48 hours of downloading the test data inputs. In addition to the participants' systems, we also used the corpus texts themselves as 'system' outputs, and created 4 baseline systems; we evaluated the resulting 8 systems using a range of intrinsic and extrinsic evaluation methods (for details see Sections 5 and 6). This report presents the results of all evaluations (Section 6), along with descriptions of GREC-MSR 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 introduc-

[1] One team submitted by the original deadline (Jan. 2009), one by the revised deadline (1 June 2009), one slightly later.

Proceedings of the 2009 Workshop on Language Generation and Summarisation, ACL-IJCNLP 2009, pages 79–87,
Suntec, Singapore, 6 August 2009.

tory sections in Wikipedia articles, in five different subdomains (cities, countries, rivers, people and mountains). In each text, three broad categories of Main Subject Reference (MSR)[2] have been annotated, resulting in a total of about 13,000 annotated REs. The GREC-MSR shared task version of 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.

2.1 Types of referential expression annotated

Three broad categories of main subject referring expressions (MSREs) are annotated in the GREC corpus[3] — subject NPs, object NPs, and genitive NPs and pronouns which function as subject-determiners within their matrix NP. These categories of referring expressions (RE) are relatively straightforward to identify and to 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 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 those 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.

[2]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 (and title) of one text is *London*.

[3]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-MSR training/development data set. 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 training/development data 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 possible alternative referential expressions (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 GREC-MSR corpus, the only type of RE that has `EMPHATIC=yes` is one which incorporates a reflexive pronoun used emphatically (e.g. *India itself*). The `REG08-TYPE` attribute indicates basic RE type. 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 use a pronoun (`pronoun`); or whether it can be left unrealised (`empty`).

2.3 The GREC-MSR 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 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

```
<?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 the GREC-MSR Training Data.

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 2009 GREC-MSR Task was to get the word strings contained in REFEXs right (whereas in REG'08 it was the REG08-TYPE attributes).

3 Test Data

1. 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. 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 them with 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 GREC-MSR Task make selections.[4] We obtained three versions of each text, where in each version all MSREs were selected by the same person. The motivation for this version of Test Set C was 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 that Wikipedia texts are edited by multiple authors which sometimes adversely affects MSR chains; we wanted to have additional reference texts where all references are selected by a single author.

3. Test Set L: 74 Wikipedia introductory texts from the subdomain of lakes (there were no lake texts in the training/development set).

4. 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 of different origin. We transcribed these texts from printed encyclopaedias published in the 1980s which are not available in electronic form. 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 than for the other test sets in both the 2008 and 2009 runs of the GREC-MSR task (for the latter, see Section 6).

Each test set was designed to test peer systems for generalisation to different kinds of unseen data. 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 for generalisation to a different corpus but the same subdomains.

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

4 Evaluation methods

4.1 Automatic intrinsic evaluations[5]

Accuracy of REFEX word strings: when computed against test sets (C-1, L and P), Word String Accuracy is simply the proportion of REFEX word strings selected by a participating system that are identical to the one in the corpus. When computed against test set C-2, which has three versions of each text, Word String Accuracy is computed as follows: first the number of correct REFEX word strings is computed at the text level for each of the three versions of a 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 Word String Accuracy score. The rationale behind computing the Word String 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 other MSREs in the same text.

Accuracy of REG08-Type: similarly to Word String Accuracy above, when computed against 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 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.

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 a length-normalised version of string-edit distance (denoted 'norm. SE' in results tables below). For test sets C-1, L and P, the global score is simply the mean of all RE-level scores. For Test Set C-2, the global score is the mean of the mean of the three text-level scores.

Other metrics: BLEU is a precision metric from machine translation that assesses peer translations in terms of the proportion of word n-grams ($n \leq 4$ is standard) they share 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. In both 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, and NIST are designed to work with one or multiple reference texts, so we did not need to use a different method for Test Set C-2.

4.2 Automatic extrinsic evaluation

As in GREC-MSR'08, we used an automatic extrinsic evaluation method based on coreference resolution performance.[6] The basic idea is that it seems likely that badly chosen reference chains affect the ability to resolve REs in automatic coreference resolution tools which will tend to perform worse with poorly selected MSR reference chains.

To counteract the possibility of results being a function of a specific coreference resolution algorithm or tool, we used two different resolvers—those included in LingPipe[7] and OpenNLP (Morton, 2005)—and averaged 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 mean.

4.3 Human intrinsic evaluation

The intrinsic human evaluation involved 24 randomly selected items from Test Set C and outputs for these produced by peer and basline systems as

[5] For GREC-MSR'09 we updated the tool that computes all automatic intrinsic scores and in the course of this eliminated a character encoding issue; as a result the results for baseline systems and corpus texts reported here are on the whole very slightly higher than those reported for GREC-MSR'08.

[6] However, for GREC'09 we overhauled the tool; the current version no longer uses JavaRAP, and uses the most recent versions of the other resolvers; the GREC-MSR'08 and GREC-MSR'09 results for this method are not entirely comparable for this reason.

[7] http://alias-i.com/lingpipe/

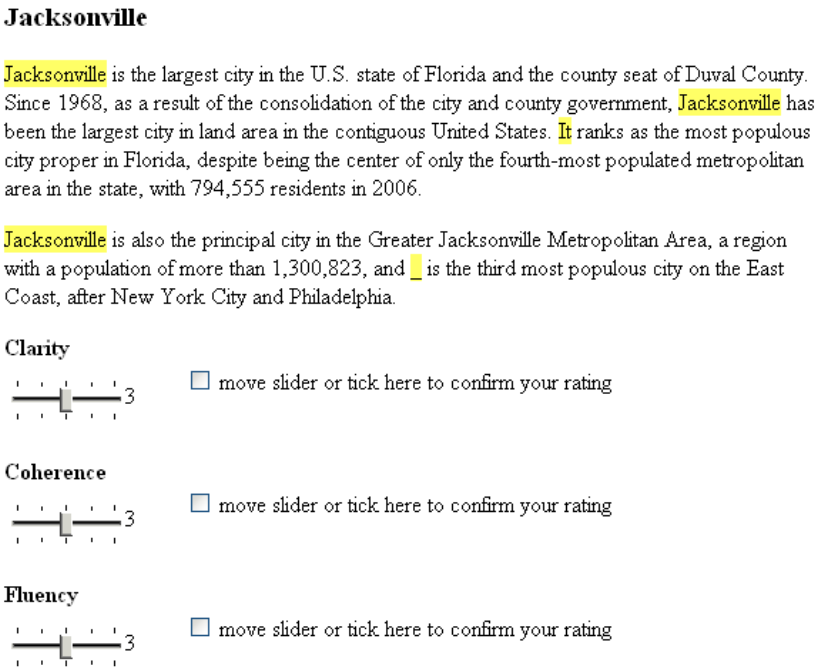

Figure 2: Example of text presented in human intrinsic evaluation of GREC-MSR systems.

well as those found in the original corpus texts (8 systems in total). We used a Repeated Latin Squares design which ensures that each subject sees the same number of outputs from each system and for each test set item. There were three 8x8 squares, and a total of 576 individual judgments in this evaluation (72 per system: 3 criteria x 3 articles x 8 evaluators).

We recruited 8 native speakers of English from among post-graduate students currently doing a linguistics-related degree at University College London (UCL) and University of Sussex.

Following detailed instructions, subjects did two practice examples, followed by the 24 texts to be evaluated, in random order. Subjects carried out the evaluation over the internet, at a time and place of their choosing. They were allowed to interrupt and resume the experiment (though discorouged from doing so). According to self-reported timings, subjects took between 25 and 45 minutes to complete the evaluation (not counting breaks).

Figure 2 shows what subjects saw during the evaluation of an individual text. All references to the MS are highlighted in yellow, and the task is to evaluate the quality of the REs in terms of three criteria which were explained in the introduction as follows (the wording of the explanations of Criteria 1 and 3 were taken from the DUC evaluations):

1. **Referential Clarity**: It should be easy to identify who or what the referring expressions in the text are referring to. If a person or other entity is mentioned, it should be clear what their role in the story is. So, a reference would be unclear if an entity is referenced, but their identity or relation to the story remains unclear.
2. **Fluency**: A referring expression should 'read well', i.e. it should be written in good, clear English, and the use of titles and names etc. should seem natural. Note that the Fluency criterion is independent of the Referential Clarity criterion: a reference can be perfectly clear, yet not be fluent.
3. **Structure and Coherence**: The text should be well structured and well organised. The text should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic. This criterion too is independent of the others.

Subjects selected evaluation scores by moving sliders (see Figure 2) along scales ranging from 1 to 5. Slider pointers started out in the middle of the scale (3). These were continuous scales and we recorded scores with one decimal place (e.g. 3.2). The meaning of the numbers was explained in terms of integer scores (1=very poor, 2=poor, 3=neither poor nor good, 4=good, 5=very good).

5 Systems

Base-rand, Base-freq, Base-1st, Base-name: Baseline system *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 `ALT-REFEX` list; and *Base-name* selects the shortest `REFEX` with attributes `REG08-TYPE=name`, `HEAD=nominal` and `EMPHATIC=no`.[8]

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

UDel: The UDel system consists of a preprocessing component performing sentence segmentation and identification of non-referring occurrences of main subject (MS) names, an RE type selection component (two C5.0 decision trees, one optimised for people and mountains, the other for the other subdomains), and a word string selection component. The RE type selection decision trees use the following features: is the MS the subject of the current, preceding and preceding but one sentence; was the last MSR in subject position; are there interfering references to other entities between the current and the previous MSR; distance to preceding non-referring occurrences of an MS name; sentence and reference IDs; other features indicating whether the reference occurred before and after certain words and punctuation marks. Given a selected RE type, the word-string selection component selects the longest non-emphatic name for the first named reference in an article, and the shortest for subsequent named references; for other types, the first matching word-string is used, backing off to pronoun or name.

ICSI-CRF: The ICSI-CRF system construes the GREC-MSR task as a sequence labelling task and determines the most likely current label given preceding labels using a Conditional Random Field model trained using the follow features for the current, preceding and preceding but one MSR: preceding and following word unigram and bigram; suffix of preceding and following word; preceding and following punctuation; reference ID; is this is the beginning of a paragraph. If more than one label remains, the last in the list of possible REs in the GREC-MSR data is selected.

JUNLG: The JUNLG system is based on co-occurrence statistics between REF feature sets and REFEX feature sets as found in the GREC-MSR data. REF feature sets were augmented by a paragraph counter and a within-paragraph REF counter. For each given set of REF features, the system selects the most frequent REFEX feature set (as determined from co-occurrence counts in the training data). If the current set of possible REFEXs does not include a REFEX with the selected feature set, then the second most likely feature set is selected. Several hand-coded default rules override the frequency-based selections, e.g. if the preceding word is a conjunction, and the current SYNCAT is np-subj, then the REG08-Type is empty.

6 Results

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

System	Word String Acc.	REG08-Type Acc.	Norm. Edit Dist.
ICSI-CRF	0.67	0.75	0.28
UDel	0.6357	0.7027	0.3383
JUNLG	0.532	0.62	0.421

Table 2: Self-reported evaluation scores for development set.

6.1 Word String Accuracy

Participants computed Word String Accuracy for the development set (97 texts) themselves, using an evaluation tool provided by us. These scores are shown in column 2 of Table 2, and are also included in the participants' reports in this volume. Corresponding results for test set C-1 are shown in column 2 of Table 3. Surprisingly, Word String Accuracy results on the test data are better (than on the development data) for the UDel and JUNLG systems. Also included in this table are results for the four baseline systems, and it is clear that selecting the most frequent word string given SEMCAT and SYNCAT (as done by the Base-freq system) provides a strong baseline.

The other two parts of Table 3 contain results for test sets L and P. 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 Word String Accuracy results for Test Set P are higher than for any other set, probably for the reasons discussed at the end of Section 3.

For each test set in Table 3 we carried out a univariate ANOVA with System as the fixed factor, 'Number of REFEXs in a text' as a random factor, and Word String Accuracy as the dependent variable. We found significant main effects of System on Word String Accuracy at $p < .001$ in the case of all three test sets (C-1: $F_{(7,1272)} = 90.058$; L: $F_{(7,440)} = 44.139$; P: $F_{(7,168)} = 21.991$).[9] The columns containing capital letters in Table 3

[9] We included the corpus texts themselves in the analysis, hence 7 degrees of freedom (8 systems).

Test Set C-1						Test Set L				Test Set P				
UDel	67.68	A				UDel	52.89	A		UDel	77.16	A		
ICSI-CRF	62.98	A				JUNLG	50.80	A		ICSI-CRF	72.22	A		
JUNLG	61.94	A				ICSI-CRF	49.20	A		JUNLG	71.60	A		
Base-freq	47.05		B			Base-name	21.06		B	Base-freq	53.09		B	
Base-name	28.74			C		Base-freq	20.74		B	Base-name	27.78			C
Base-1st	28.26			C		Base-1st	20.74		B	Base-1st	27.16			C
Base-rand	18.95				D	Base-rand	15.11		B	Base-rand	18.52			C

Table 3: Word String Accuracy scores against Test Sets C-1, L and P; homogeneous subsets (Tukey HSD, alpha = .05) for each test set (systems that do not share a letter are significantly different).

System	Word String Accuracy for multiple-RE Test Set C-2											
	All							Cities	Countries	Rivers	People	Mountains
Corpus	*71.58*	A						*65.25*	*69.11*	*76.47*	*80.40*	*66.87*
UDel	70.22	A	B					68.09	71.20	76.47	76.63	64.84
JUNLG	64.57		B	C				54.61	51.83	73.53	71.86	65.85
ICSI-CRF	63.69			C				58.87	56.54	64.71	72.11	60.98
Base-freq	57.01				D			51.06	57.07	58.82	63.82	53.05
Base-name	40.21					E		51.06	46.07	29.41	29.90	43.90
Base-1st	39.65					E		47.52	41.88	38.24	25.63	47.97
Base-rand	26.99						F	28.37	29.32	23.53	21.61	30.28

Table 4: Word String 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).

show the homogeneous subsets of systems as determined by post-hoc Tukey HSD comparisons of means. Systems whose Word String Accuracy scores are not significantly different (at the .05 level) share a letter.

The results for Word String Accuracy computed against Test Set C-2 are shown in Table 4. These should be considered the chief results of the GREC-MSR'09 Task evaluations, as stated in the participants' guidelines. Here too we performed a univariate ANOVA with System as the fixed factor, Number of REFEXs as the random factor and Word String Accuracy as the dependent variable. There was a significant main effect of System ($F_{(7,1272)} = 74.892$, $p < .001$). 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 the UDel system, but also there is no significant difference between the latter and the JUNLG system. In this analysis, all peer systems outperform all baselines; the Base-freq baseline outperforms all other baselines; and Base-name and Base-1st outperform the random baseline.

Overall, there is a marked improvement in Word String Accuracy compared to GREC-MSR'08 where peer systems' scores ranged from 50.72 to 65.61.

6.2 Other automatic intrinsic metrics

In addition to the chief evaluation measure reported on in the preceding section, we computed REG08-Type Accuracy and the string similarity metrics described in Section 4.1. The resulting scores for Test Set C-2 are shown in Table 5 (recall that in Test Set C-2 corpus texts are evaluated against 3 texts with human-selected alternative REs). The corpus texts again receive the best scores across the board. Ranks for peer systems are very similar to those reported in the last section.

We performed a univariate ANOVA with System as the fixed factor, Number of REFEXs as the random factor, and REG08-Type Accuracy as the dependent variable. The main effect of System was $F_{(7,1272)} = 75.040$, $p < .001$; the homogeneous subsets resulting from the Tukey HSD post-hoc analysis are shown in columns 3–5 of Table 5. The differences between the scores of the peer systems and the corpus texts were not found to be significant.

6.3 Human-assessed intrinsic measures

Table 6 shows the results of the human intrinsic evaluation. In each of the three parts of the table (showing the results for Fluency, Clarity and Coherence, respectively) systems are ordered in terms of their mean scores (shown in the second column of each part of the table). We first established that the main effect of Evaluator was weak (F between 2.1 and 2.6) on Fluency, Clarity and Coherence, and only of borderline significance (just below .05); and that the interaction between System and Evaluator was very weak and

System	Other similarity measures for Triple-RE Test Set C-2							
	REG08-Type				BLEU-3	NIST	SE	norm. SE
Corpus	*79.30*	*A*			*0.77*	*5.60*	*1.04*	*0.34*
UDel	77.71	A			0.74	5.32	1.11	0.37
JUNLG	75.40	A			0.53	4.69	1.34	0.40
ICSI-CRF	75.16	A			0.54	4.68	1.32	0.41
Base-freq	62.50		B		0.54	4.30	1.93	0.50
Base-name	51.04			C	0.46	4.76	1.80	0.63
Base-1st	50.32			C	0.39	4.42	1.93	0.63
Base-rand	48.09			C	0.26	3.02	2.30	0.72

Table 5: REG08-Type Accuracy, BLEU, NIST and string-edit scores, computed on test set C-2 (systems in order of REG08-Type Accuracy); homogeneous subsets (Tukey HSD, alpha = .05) for REG08-Type Accuracy only (systems that do not share a letter are significantly different).

Fluency						Clarity				Coherence			
Corpus	4.43	A				Base-name	4.62	A		*Corpus*	4.40	A	
UDel	4.27	A				*Corpus*	4.56	A		JUNLG	4.33	A	
JUNLG	4.26	A				JUNLG	4.50	A		UDel	4.27	A	B
ICSI-CRF	4.15	A	B			ICSI-CRF	4.45	A		ICSI-CRF	4.02	A	B
Base-freq	3.33		B	C		UDel	4.35	A		Base-freq	3.96	A	B
Base-name	2.84			C	D	Base-1st	4.27	A		Base-name	3.85	A	B
Base-1st	2.76			C	D	Base-freq	4.10	A		Base-1st	3.7	A	B
Base-rand	2.15				D	Base-rand	3.18		B	Base-rand	3.46		B

Table 6: Clarity, Fluency and Coherence scores (with homogeneous subsets) for all systems.

not significant in the case of Clarity and Coherence, and borderline significant in the case of Fluency. We then ran a (non-factorial) multivariate ANOVA, with Fluency, Coherence and Clarity as the dependent variables, and (just) System as the fixed factor. The main effect of System was as follows: Fluency: $F_{(7,128)} = 20.444, p < 0.001$; Clarity: $F_{(7,128)} = 5.248, p < 0.001$; Coherence: $F_{(7,128)} = 2.680, p < 0.012$. The homogeneous subsets resulting from a post-hoc Tukey analysis are shown in the letter columns in Table 6.

The effect of System was strongest on Fluency; here, the system ranks are also the same as for Word String Accuracy and REG08-Type Accuracy for Test Set C-2. This, together with the fair amount of significant differences found, indicates that the evaluators were able to make sense of the Fluency criterion and that there were interesting differences between systems under this criterion. However, differences between the three peer systems were not significant.

For Clarity, there were no significant differences among the peer systems and non-random baseline systems; all of these were significantly better than the random baseline. Base-name had the highest mean Clarity score, possibly because always chosing the name of an entity when referring to it ensures high referential clarity.

The Coherence results are perhaps the most difficult to interpret. Both the main effect of System on Coherence and its significance were weaker than for Fluency and Clarity. Only two significant pairwise differences were found: Corpus and JUNLG were better than the random baseline. The system ranks are roughly the same as for Fluency, but the mean scores cover a smaller range (from 3.46 to 4.4) than in the case of either of the other two criteria. Overall, the Coherence results probably indicate that the evaluators found it somewhat difficult to make sense of the Coherence criterion.

Computing Pearson's r for the three criteria on individual (text-level) scores showed that there were only moderate correlations between them (all around $r = 0.5$) which were all significant at $\alpha = 0.05$. This gives some indication that the evaluators were able to assess the three criteria independently from each other.

6.4 Automatic extrinsic measures

We fed the outputs of all eight systems through the two coreference resolvers, and computed mean MUC, CEAF and B-CUBED F-Scores as described in Section 4.2. The second column in Table 7 shows the mean of these three F-Scores, to give a single overall result for this evaluation method. A univariate ANOVA with mean F-Score as the dependent variable and System as the fixed factor revealed a significant main effect of System on mean F-Score ($F_{(7,1456)} = 73.061, p < .001$). A post-hoc comparison of the means (Tukey HSD, alpha = .05) found the significant differences indicated by the homogeneous subsets in columns 3–4 (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 two resolver tools. ANOVAs revealed the fol-

lowing effects of System on the separate scoring methods: on CEAF $F_{(7,1456)} = 43.471, p < .001$; on MUC: $F_{(7,1456)} =, p < .001$; on B-CUBED: $F_{(7,1456)} = 38.574, p < .001$. All three scoring methods separately and their mean yielded the same significant differences (as shown in columns 3–4 of Table 7).

The three F-Score measures (MUC, CEAF and B-CUBED) are all significantly correlated ($p < .001$, 2-tailed). However it is not a strong correlation, with Pearson's correlation coefficient around 0.5.

System	(MUC+CEAF+B3)/3			MUC	CEAF	B3
Base-name	65.19	A		62.35	63.14	70.06
Base-1st	63.77	A		59.95	62.08	69.28
Base-freq	63.14	A		59.08	62.04	68.3
UDel	46.19		B	34.85	46.86	56.86
ICSI-CRF	44.47		B	31.61	45.58	56.21
JUNLG	44.19		B	31.27	45.21	56.10
Base-rand	42.99		B	30.24	43.04	55.7
Corpus	*42.52*		*B*	*29.53*	*43.57*	*54.47*

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

6.5 Correlations

When assessed on the system-level scores and using Pearson's r, all evaluation methods above were strongly and significantly correlated with each other (at the 0.01 level, 2-tailed), with the following exceptions. Clarity was not significantly correlated with *any* of the other methods except NIST ($r = .902, p < .01$); apart from this, NIST was only correlated with Word String Accuracy on test set C-2, with non-normalised string-edit distance, Fluency and Coherence, moreover all at the weaker 0.05 level. Finally, the extrinsic method was not correlated with any of the intrinsic methods (and in fact showed signs of being negatively correlated with all of them except Clarity).

7 Concluding Remarks

The GREC-MSR Task is still a relatively new task not only for an NLG shared-task challenge, but also as a research task in general (post-processing extractive summaries in order to improve their quality seems to be just taking off as a research subfield). There was substantial interest in the GREC-MSR Task this year (as indicated by the nine teams that originally registered). However, only three teams were ultimately able to participate.

We continued the traditions of previous NLG shared tasks in that we used a wide range of evaluation metrics to obtain a well-rounded view of the quality of the participating systems. This included intrinsic human evaluations for the first time. However, we decided against an extrinsic human evaluation this year, given time constraints as well as the fact that this evaluation type yielded barely any significant results last year.

Overall, there was an improvement in system performance compared to last year, to the point where the performance of the top system was barely distinguishable from the human topline. We are not currently planning to run the GREC-MSR task again next year.

Acknowledgments

Many thanks to the UCL and Sussex students who participated in the intrinsic evaluation 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.

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.

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.