The GREC Challenges 2010: Overview and Evaluation Results

Anja Belz **Eric Kow**
Natural Language Technology Group
School of Computing, Mathematical and Information Sciences
University of Brighton
Brighton BN2 4GJ, UK
{asb,eykk10}@bton.ac.uk

Abstract

There were three GREC Tasks at Generation Challenges 2010: GREC-NER required participating systems to identify all people references in texts; for GREC-NEG, systems selected coreference chains for all people entities in texts; and GREC-Full combined the NER and NEG tasks, i.e. systems identified and, if appropriate, replaced references to people in texts. Five teams submitted 10 systems in total, and we additionally created baseline systems for each task. Systems were evaluated automatically using a range of intrinsic metrics. In addition, systems were assessed by human judges using preference strength judgements. This report presents the evaluation results, along with descriptions of the three GREC tasks, the evaluation methods, and the participating systems.

1 Introduction

Until recently, referring expression generation (REG) research focused on the task of selecting the semantic content of one-off mentions of listener-familiar discourse entities. In the GREC research programme we have been interested in REG as (i) grounded within discourse context, (ii) embedded within an application context, and (iii) informed by naturally occurring data.

In general terms, the GREC tasks are about how to select appropriate references to an entity in the context of a piece of discourse longer than a sentence. In GREC'10, there were three subtasks: identification of references to people in free text (GREC-NER); selection of references to people in text (GREC-NEG); and regeneration of references to people in text (GREC-Full) which can be thought of as combining the NER and NEG tasks.

The immediate motivating application context for the GREC Tasks is the improvement of referential clarity and coherence in extractive summaries and multiply edited texts (such as Wikipedia articles) by regenerating referring expressions contained in them. The motivating theoretical interest for the GREC Tasks is to discover what kind of information is useful for making choices between different kinds of referring expressions in context.

The GREC'10 tasks used the GREC-People corpus which consists of 1,100 Wikipedia texts about people within which we have annotated all references to people.

Five teams participated in the GREC'10 tasks (see Table 1), submitting 10 systems in total. Two of these were created by combining the NER system of one of the teams with the NEG systems of two different teams, producing two 'combined' systems for the Full Task. We also used the corpus texts themselves as 'system' outputs, and created baseline systems for all three tasks. We evaluated systems using a range of intrinsic automatically computed and human-assessed evaluation methods. This report describes the data (Section 2) and evaluation methods (Section 3) used in the three GREC'10 tasks, and then presents task definition, participating systems, evaluation methods, and evaluation results for each of the three tasks separately (Sections 4– 6).

2 GREC'10 Data

The GREC'10 data is derived from the GREC-People corpus which (in its 2010 version) consists of 1,100 annotated introduction sections from Wikipedia articles in the category People. An introduction section was defined as the textual content of a Wikipedia article from the title up to (and excluding) the first section heading, the table of contents or the end of the text, which ever comes first. Each text belongs to one of six subcategories: inventors, chefs, early music composers, explorers, kickboxers and romantic composers. For the

Team	Affiliation	NEG systems	NER systems	Full systems
UDel^x	University of Delaware	UDel-NEG	UDel-NER	UDel-Full
UMUS	Université du Maine Universität Stuttgart	UMUS	–	–
JU^x	Jadavpur University	JU	–	–
Poly-co	École Polytechnique de Montréal	–	Poly-co	–
XRCE^y	Xerox Research Centre Europe	XRCE	–	–
UDel/UMUS	(see above)	–	–	UDel-UMUS-Full
UDel/XRCE	(see above)	–	–	UDel-XRCE-Full

Table 1: GREC-NEG’09 teams and systems (combined teams in last two rows). ^{x} = resubmitted after fixing character encoding problems and/or software bugs; ^{y} = late submission.

	All	Inventors	Chefs	Early Composers	Explorers	Kickboxers	Romantic Composers
Training	809	249	248	312	–	–	–
Development	91	28	28	35	–	–	–
Test (NEG)	100	31	30	39	–	–	–
Test (NER/Full)	100	–	–	–	33	34	33
Total	1,100	307	306	387	33	34	33

Table 2: Overview of GREC’10 data sets.

purposes of the GREC task, the GREC-People corpus was divided into training, development and test data. The number of texts in the subsets are as shown in Table 2.

In the GREC-People annotation scheme, a distinction is made between *reference* and *referential expression*. A reference is ‘an instance of referring’ which is unique, whereas a referential expression is a word string and each reference can be realised by many different referential expressions. In the GREC corpora, each time an entity is referred to, there is a single reference, but there may be one or several referring expressions provided with it: in the training/development data, there is a single RE for each reference (the one found in the corpus); in the test set, there are four REs for each reference (the one from the corpus and three additional ones selected by subjects in a manual selection experiment).

We first manually annotated people mentions in the GREC-People texts by marking up the word strings that function as referential expressions (REs) and annotating them with coreference information as well as semantic category, syntactic category and function, and various supplements and dependents. Annotations included nested references, plurals and coordinated REs, certain unnamed references and indefinites. In terminology and the treatment of syntax used in the annotation scheme we relied heavily on *The Cambridge Grammar of the English Language* by Huddleston and Pullum (2002). For full details of the manual annotation please refer to the GREC’10 documentation (Belz, 2010).

The manual annotations were then automatically checked and converted to XML format. In the XML format of the annotations, the beginning and end of a reference is indicated by <REF><REFEX>... </REFEX></REF> tags, and other properties mentioned above (e.g. syntactic category) are encoded as attributes on these tags. For the GREC tasks we decided not to transfer the annotations of integrated dependents and relative clauses to the XML format. Such dependents are included within <REFEX>...</REFEX> annotations where appropriate, but without being marked up as separate constituents.

Figure 1 shows one of the XML-annotated texts from the GREC data. For full details of the manual annotations and the XML version, please refer to the GREC’10 documentation (Belz, 2010). Here we provide a brief summary.

The REF element indicates a reference, and is composed of one REFEX element (the ‘selected’ referential expression for the given reference; in the corpus texts it is the referential expression found in the corpus). The attributes of the REF element are ENTITY (entity identifier), MENTION (mention identifier), SEMCAT (semantic category), SYNCAT (syntactic category), and SYNFUNC (syntactic function). ENTITY and MENTION together constitute a unique identifier for a reference within a text; together with the TEXT ID, they constitute a unique identifier for a reference within the entire

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE GREC-ITEM SYSTEM "genchal09-grec.dtd">
<GREC-ITEM>
<TEXT ID="15">
<TITLE>Alexander Fleming</TITLE>

<PARAGRAPH> <REF ENTITY="0" MENTION="1" SEMCAT="person" SYNCAT="np" SYNFUNC="subj">
  <REFEX ENTITY="0" REG08-TYPE="name" CASE="plain">Sir Alexander Fleming</REFEX>
</REF> (6 August 1881 - 11 March 1955) was a Scottish biologist and pharmacologist.
<REF ENTITY="0" MENTION="2" SEMCAT="person" SYNCAT="np" SYNFUNC="subj">
  <REFEX ENTITY="0" REG08-TYPE="name" CASE="plain">Fleming</REFEX>
</REF> published many articles on bacteriology, immunology, and chemotherapy.
<REF ENTITY="0" MENTION="3" SEMCAT="person" SYNCAT="np" SYNFUNC="subj-det">
  <REFEX ENTITY="0" REG08-TYPE="pronoun" CASE="genitive">his</REFEX>
</REF> best-known achievements are the discovery of the enzyme lysozyme in 1922 and the discovery
of the antibiotic substance penicillin from the fungus Penicillium notatum in 1928, for which
<REF ENTITY="0" MENTION="4" SEMCAT="person" SYNCAT="np" SYNFUNC="subj">
  <REFEX ENTITY="0" REG08-TYPE="pronoun" CASE="nominative">he</REFEX>
</REF> shared the Nobel Prize in Physiology or Medicine in 1945 with
<REF ENTITY="1" MENTION="1" SEMCAT="person" SYNCAT="np" SYNFUNC="obj">
  <REFEX ENTITY="1" REG08-TYPE="name" CASE="plain">Florey</REFEX>
</REF> and
<REF ENTITY="2" MENTION="1" SEMCAT="person" SYNCAT="np" SYNFUNC="obj">
  <REFEX ENTITY="2" REG08-TYPE="name" CASE="plain">Chain</REFEX>
</REF>.</PARAGRAPH>
</TEXT>

<ALT-REFEX>
  <REFEX ENTITY="0" REG08-TYPE="empty" CASE="no_case">_</REFEX>
  <REFEX ENTITY="0" REG08-TYPE="name" CASE="genitive">Fleming’s</REFEX>
  <REFEX ENTITY="0" REG08-TYPE="name" CASE="genitive">Sir Alexander Fleming’s</REFEX>
  <REFEX ENTITY="0" REG08-TYPE="name" CASE="plain">Fleming</REFEX>
  <REFEX ENTITY="0" REG08-TYPE="name" CASE="plain">Sir Alexander Fleming</REFEX>
  <REFEX ENTITY="0" REG08-TYPE="pronoun" CASE="accusative">him</REFEX>
  <REFEX ENTITY="0" REG08-TYPE="pronoun" CASE="genitive">his</REFEX>
  <REFEX ENTITY="0" REG08-TYPE="pronoun" CASE="nominative">he</REFEX>
  <REFEX ENTITY="0" REG08-TYPE="pronoun" CASE="nominative">who</REFEX>
  <REFEX ENTITY="1" REG08-TYPE="empty" CASE="no_case">_</REFEX>
  <REFEX ENTITY="1" REG08-TYPE="name" CASE="genitive">Florey’s</REFEX>
  <REFEX ENTITY="1" REG08-TYPE="name" CASE="plain">Florey</REFEX>
  <REFEX ENTITY="1" REG08-TYPE="pronoun" CASE="accusative">him</REFEX>
  <REFEX ENTITY="1" REG08-TYPE="pronoun" CASE="genitive">his</REFEX>
  <REFEX ENTITY="1" REG08-TYPE="pronoun" CASE="nominative">he</REFEX>
  <REFEX ENTITY="1" REG08-TYPE="pronoun" CASE="nominative">who</REFEX>
  <REFEX ENTITY="2" REG08-TYPE="empty" CASE="no_case">_</REFEX>
  <REFEX ENTITY="2" REG08-TYPE="name" CASE="genitive">Chain’s</REFEX>
  <REFEX ENTITY="2" REG08-TYPE="name" CASE="plain">Chain</REFEX>
  <REFEX ENTITY="2" REG08-TYPE="pronoun" CASE="accusative">him</REFEX>
  <REFEX ENTITY="2" REG08-TYPE="pronoun" CASE="genitive">his</REFEX>
  <REFEX ENTITY="2" REG08-TYPE="pronoun" CASE="nominative">he</REFEX>
  <REFEX ENTITY="2" REG08-TYPE="pronoun" CASE="nominative">who</REFEX>
</ALT-REFEX>
</GREC-ITEM>
```

Figure 1: Example XML-annotated text from the GREC-NEG’09 data.

corpus.

A REFEX element indicates a referential expression (a word string that can be used to refer to an entity). The attributes of the REFEX element are REG08-TYPE (name, common, pronoun, empty), and CASE (nominative, accusative, etc.).

We allow arbitrary-depth embedding of references. This means that a REFEX element may have REF element(s) embedded in it.

The second (and last) component of a GREC-ITEM is an ALT-REFEX element which is a list of REFEX elements. For the GREC tasks, these were obtained by collecting the set of all REFEXs that are in the text, and adding several defaults including pronouns and other cases (e.g. genitive) of REs already in the list.

REF elements that are embedded in REFEX elements contained in an ALT-REFEX list have an unspecified MENTION id (the ‘?’ value). Furthermore, such REF elements have had their enclosed REFEX removed.

The two test data sets exist in two versions:

1. Version a: each text has a single human-selected referring expression for each reference (i.e. the one found in the original Wikipedia article).

2. Version b: the same subset of texts as in (a); for this set we did not use the REs in the corpus, but replaced each of them with human-selected alternatives obtained in an online experiment as described in (Belz and Varges, 2007); this version of the test set therefore contains three versions of each text where all the REFEXs in a given version were selected by one ‘author’.

The training, development and test data for the GREC-NEG task is exactly as described above. The training and development data for the GREC-NER/Full tasks comes in two versions. The first is identical to the standard XML-annotated version of the GREC-People corpus as described above (Section 2). The second is in the test data input format.

In this format, texts have no REFEX and REF tags, and no ALT-REFEX element. A further difference is that in the test data format, a proportion of REFEX word strings have been replaced with standardised named references. All empty references have been replaced in this way, whereas (non-relative) pronouns, and previously seen named references that are not identical to the standardised named reference, are replaced with a likelihood of 0.5.

The reason for this replacement is to make both tasks easier (as we are running them for the first time) as well as more realistic (in an extractive summary, reference chains are unlikely to be as good as in the Wikipedia texts).

3 Evaluation Procedures

Table 3 is an overview of the evaluation measures we applied to the three tasks in GREC'10. Version a of the test sets has a single version of each text, and the scoring metrics that are based on counting matches (Word String Accuracy counts matching word strings, REG08-Type Recall/Precision count matching REG08-Type attribute values) simply count the number of matches a system achieves against that single text. Version b, however, has three versions of each text, so the match-based metrics first calculate the number of matches for each of the three versions and then use (just) the highest number of matches.

3.1 Automatic Evaluations

REG08-Type Precision is defined as the proportion of REFEXs selected by a participating system which match the reference REFEXs. REG08-Type Recall is defined as the proportion of reference REFEXs for which a participating system has produced a match.

String Accuracy is defined as the proportion of word strings selected by a participating system that match those in the reference texts. This was computed on complete, 'flattened' word strings contained in the outermost REFEX i.e. embedded REFEX word strings were not considered separately.

We also computed BLEU-3, NIST, string-edit distance and length-normalised string-edit distance, all on word strings defined as for String Accuracy. BLEU and NIST are designed for multiple output versions, and for the string-edit metrics we computed the mean of means over the three text-level scores (computed against the three versions of a text).

To measure accuracy in the NER task, we applied three commonly used performance measures for coreference resolution: MUC-6 (Vilain et al., 1995), CEAF (Luo, 2005), and B-CUBED (Bagga and Baldwin, 1998).

3.2 Human-assessed evaluations

We designed the human-assessed intrinsic evaluation as a preference-judgement test where subjects expressed their preference, in terms of two criteria, for either the original Wikipedia text or the version of it with system-generated referring expressions in it. For the GREC-NEG systems, the intrinsic human evaluation involved system outputs for 30 randomly selected items from the test set. 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 10 × 10 squares, and a total of 600 individual judgements in this evaluation (60 per system: 2 criteria × 3 articles × 10 evaluators). We recruited 10 native speakers of English from among students currently completing a linguistics-related degree at Kings College London and University College London.

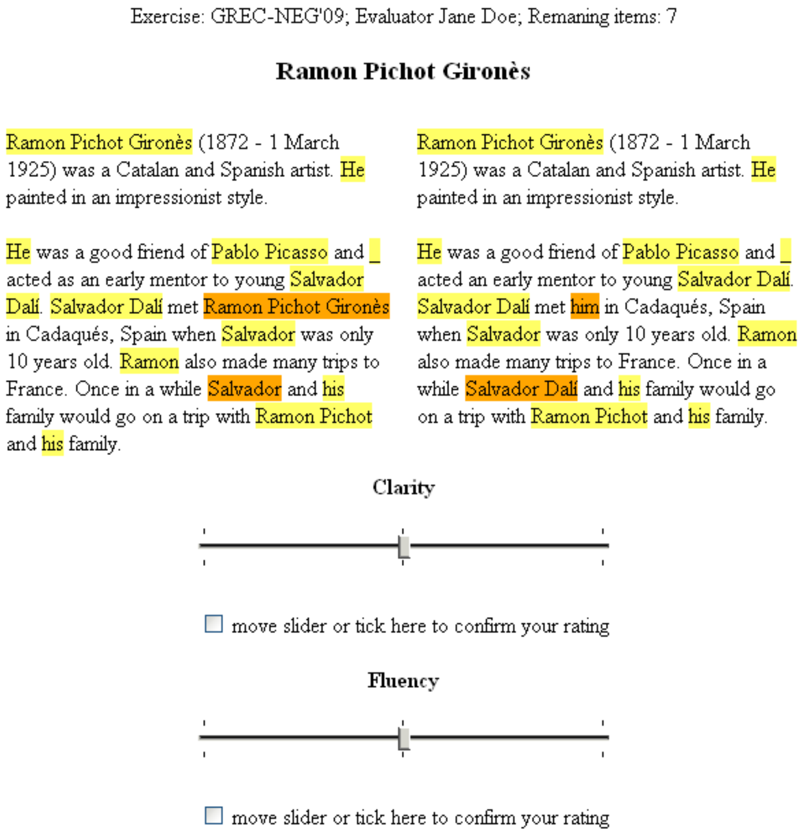
For the GREC-Full systems, we used 21 randomly selected test set items, a design analogous to that for the GREC-NEG experiment, and 7 evaluators from the same cohort. This experiment had three 7 × 7 squares, and 294 individual judgements.

Following detailed instructions, subjects did two practice examples, followed by the 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 discouraged from doing so).

Figure 2 shows what subjects saw during the evaluation of an individual text pair. The place (left/right) of the original Wikipedia article was randomly determined for each individual evaluation of a text pair. People references are highlighted in yellow/orange, those that are identical in both texts are yellow, those that are different are orange (in the GREC-Full version, there were only yellow highlights). The evaluator's task is to express their preference in terms of each quality criterion by moving the slider pointers. Moving the slider to the left means expressing a preference for

Quality criterion:	Type of evaluation:	Task:	Evaluation Method(s):
Humanlikeness	Intrinsic/automatic	NEG	1. REG'08-Type Recall and Precision 2. String Accuracy 3. String-edit distance
		NEG, Full	1. BLEU 2. NIST version of BLEU
		NER	CEAF, MUC-6, B-CUBED
Fluency	Intrinsic/human	NEG, Full	Human preference-strength judgements
Referential Clarity	Intrinsic/human	NEG, Full	Human preference-strength judgements

Table 3: Overview of GREC'10 evaluation procedures.



Exercise: GREC-NEG'09; Evaluator Jane Doe; Remaning items: 7

Ramon Pichot Gironès

Ramon Pichot Gironès (1872 - 1 March 1925) was a Catalan and Spanish artist. He painted in an impressionist style.	Ramon Pichot Gironès (1872 - 1 March 1925) was a Catalan and Spanish artist. He painted in an impressionist style.
He was a good friend of Pablo Picasso and _ acted as an early mentor to young Salvador Dalí. Salvador Dalí met Ramon Pichot Gironès in Cadaqués, Spain when Salvador was only 10 years old. Ramon also made many trips to France. Once in a while Salvador and his family would go on a trip with Ramon Pichot and his family.	He was a good friend of Pablo Picasso and _ acted an early mentor to young Salvador Dalí. Salvador Dalí met him in Cadaqués, Spain when Salvador was only 10 years old. Ramon also made many trips to France. Once in a while Salvador Dalí and his family would go on a trip with Ramon Pichot and his family.

Clarity

□ move slider or tick here to confirm your rating

Fluency

□ move slider or tick here to confirm your rating

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

the text on the left, moving it to the right means preferring the text on the right; the further to the left/right the slider is moved, the stronger the preference. The two criteria were explained in the introduction as follows (the wording of the first is from DUC):

1. **Referential Clarity**: It should be easy to identify who the referring expressions are referring to. If a person is mentioned, it should be clear what their role in the story is. So, a reference would be unclear if a person 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 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.

It was not evident to the evaluators that sliders were associated with numerical values. Slider pointers started out in the middle of the scale (no preference). The values associated with the points on the slider ranged from -10.0 to +10.0.

4 GREC-NEG

4.1 Task

The GREC-NEG test data inputs are identical to the training/development data (Figure 1), except that REF elements in the test data do not contain a REFEX element, i.e. they are 'empty'. The task for participating systems is to select one REFEX from the ALT-REFEX list for each REF in each TEXT in the test sets. If the selected REFEX contains an embedded REF then participating systems also need to select a REFEX for this embedded REF and to set the value of its MENTION attribute. The same applies to all further embedded REFEXs, at any depth of embedding.

4.2 Systems

NEG-Base-rand, NEG-Base-freq, NEG-Base-1st, NEG-Base-name: We created four baseline systems each with a different way of selecting a REFEX from those REFEXs in the ALT-REFEX list that have matching entity IDs. *Base-rand* selects a REFEX at random. *Base-1st* selects the first REFEX (unless the first is the empty reference in which case it selects the second).[1] *Base-freq* selects the first REFEX with a REG08-TYPE and CASE combination that is the overall most frequent (as determined from the training/development data) given the SYNCAT, SYNFUNC and SEMCAT of the reference.[1] *Base-name* selects the shortest REFEX with attribute REG08-TYPE=name.

UMUS: The UMUS system maps REFEXs to class labels encoding REG08-TYPE, CASE, pronoun type, reflexiveness and recursiveness. References are represented by a set of features encoding the attributes given in the corpus, information about intervening references to other entities, preceding punctuation, sentence and paragraph boundaries, surrounding word and POS n-grams, etc. A Conditional Random Fields method is then used to map features to class labels. The problem is construed as predicting a sequence of class labels for each entity, to avoid repetition. If there is more than one REFEX available with the predicted label then the longest one is chosen the first time, and selection iterates through the list subsequently.

UDel: The UDel system is a set of decision-tree classifiers (separate ones for the main subject and other person entities) using psycholinguistically inspired features that predict the REG08-TYPE and CASE of the REFEX to select. Then the system applies rules governing the length of first and subsequent mentions. There are back-off rules for when the predicted type/case is not available. An ambiguity checker avoids the use of a pronoun if there has been an intervening reference to a person of the same gender.

JU: The JU baseline system is similar to our NEG-Base-freq system described above. The submitted JU system adds features to the set of REF attributes available from the corpus, including indices for paragraph, sentence and word. It also adds features to the REFEX attributes available from the corpus, in order to distinguish between several REFEXs that match the predicted REG08-TYPE and CASE combination.

XRCE: The XRCE system uses a conditional random field model in combination with the SampleRank algorithm for learning model parameters. The feature functions used include unary ones (>100 features encoding the attributes provided in the corpus as well as position within sentence, adjacent POS tags, etc.) and binary ones (distance to previous mention, distribution of type and case). Some binary feature functions are activated only if the previous mention was a name and control overuse of pronouns.

4.3 Evaluation results

Participants computed evaluation scores on the development set, using the geval code provided by us which computes Word String Accuracy, REG'08-Type Recall and Precision, string-edit distance and BLEU. The following is a summary of teams' self-reported scores:

	Recall	Precision	WSA
UMUS	0.816	0.829	0.813
UMUS'09	0.830	0.830	0.786
XRCE	0.771	0.771	0.702
UDel	0.758	0.758	0.650
JU	0.66	0.63	0.54

REG08-Type Recall and Precision results for Test Set NEG-a (version a of the test set with just one REFEX for each REF) are shown in Table 4. As would be expected, results on the test data are somewhat worse than on the development data. Also included in this table are results for the 4 baseline systems, and it is clear that selecting the most frequent RE type and case combination given SEMCAT, SYNFUNC and SYNCAT (as done by the Base-freq system) provides a strong baseline, although it is a much better predictor for Composer and Inventor texts than Chef texts.

The last 6 columns in Table 4 contain Recall (R) and Precision (P) results for the three subdomains. For most of the systems results are slightly better for Composers than for Chefs. A contributing factor to this may be the fact that Chef texts tend to be much more colloquial. A striking detail is the collapse in scores in the Inventors subdomain for

[1] Note that this is a change from GREC'09.

System	REG08-Type Precision and Recall Scores against Corpus (Test Set NEG-a)													
	All								Chefs		Composers		Inventors	
	Precision				Recall				P	R	P	R	P	R
UMUS	80.71	A			78.31	A			79.19	75.44	80.88	78.68	81.66	80.05
UMUS'09	80.17	A			77.06	A			75.16	70.71	82.25	79.54	80.66	78.08
XRCE	74.26	A			71.38	A			68.55	64.50	75.44	72.96	76.84	74.38
JU	66.98	A	B		64.38	A	B		79.56	74.85	84.32	81.55	26.97	26.11
Base-freq	61.52	A	B	C	59.60	A	B	C	51.86	49.41	65.74	63.95	62.12	60.59
UDel-NEG	60.92	A	B	C	58.56	A	B	C	55.35	52.07	62.43	60.37	62.85	60.84
Base-rand	43.32		B	C	42.00		B	C	40.43	38.76	43.00	41.77	46.21	45.07
Base-name	40.60			C	39.09			C	47.80	44.97	40.32	39.06	35.28	34.24
Base-1st	40.25			C	39.64			C	47.88	46.75	39.71	39.20	34.91	34.48

Table 4: REG08-Type Precision and Recall scores against corpus version of Test Set for complete set and for subdomains; homogeneous subsets (Tukey HSD, alpha = .05) for complete set only.

System	REG08-Type Precision and Recall Scores against human topline (Test Set NEG-b)														
	All									Chefs		Composers		Inventors	
	Precision				Recall					P	R	P	R	P	R
Corpus	82.67	A			84.01	A				82.25	84.24	83.26	84.47	82.02	83.04
UMUS	81.64	A			80.49	A				82.92	80.91	80.59	79.54	82.41	81.80
UMUS'09	80.46	A			78.59	A	B			80.50	77.58	80.62	79.10	80.15	78.55
XRCE	73.76	A	B		72.04	A	B	C		73.58	70.91	74.11	72.71	73.28	71.82
UDel-NEG	65.54	A	B	C	64.01	A	B	C	D	66.04	63.64	66.12	64.88	64.12	62.84
Base-freq	65.38	A	B	C	64.37	A	B	C	D	59.94	58.48	68.97	68.07	63.64	62.84
JU	63.73	A	B	C	62.25	A	B	C	D	76.42	73.64	76.04	74.60	32.32	31.67
Base-name	55.22		B	C	54.01		B	C	D	56.29	54.24	58.05	57.04	49.49	48.63
Base-1st	54.68		B	C	54.68			C	D	55.45	55.45	57.68	57.68	48.88	48.88
Base-rand	48.46			C	47.75				D	48.77	47.88	47.13	46.44	50.51	49.88

Table 5: REG08-Type Recall and Precision scores against human topline version of Test Set for complete set and for subdomains; homogeneous subsets (Tukey HSD, alpha = .05) for complete set only.

the JU system. As a side effect, the resulting variation led to fewer significant differences between systems being found in the results than would have been the case otherwise.

We carried out univariate ANOVAs with System as the fixed factor, and REG08-Type Recall as the dependent variable in one ANOVA, and REG08-Type Precision in the other. The F-ratio for Recall was $F_{(9,990)} = 13.253, p < 0.001$.[2] The F-ratio for Precision was $F_{(9,990)} = 12.670, p < 0.001$. The columns containing single capital letters in Table 4 show the homogeneous subsets of systems as determined by a post-hoc Tukey HSD analysis. Systems whose scores are not significantly different (at the .05 level) share a letter.

Table 5 shows analogous results computed against Test Set NEG-b (which has three versions of each text). Table 5 includes results for the corpus texts, also computed against the three versions of each text in test set GREC-NEG-b. We performed univariate ANOVAs with System as the fixed factor, and Recall as the dependent variable in one, and Precision in the other. The result for Recall was $F_{(9,990)} = 5.248$, $p < .001$), and for Precision $F_{(9,990)} = 5.038$, $p < .001$. We again compared the mean scores with Tukey's HSD.

One would generally expect results on test set NEG-b to be better than on NEG-a. This is the case for all baseline systems and some of the participating systems, but not all. The JU system in particular drops in score (and rank).

We also computed Word String Accuracy and the other string similarity metrics described in Section 3 for the GREC-NEG Task. The resulting scores for Test Set NEG-a are shown in Table 6. Ranks for peer systems relative to each other are very similar to the results for REG08-Type reported above.

We performed a univariate ANOVA with System as the fixed factor, and Word String Accuracy as the dependent variable. The F-ratio for System was $F_{(9,990)} = 41.308, p < 0.001$; the homogeneous subsets resulting from the Tukey HSD post-hoc analysis are shown in columns 3–7 of Table 6.

Table 7 shows analogous results for human topline Test Set NEG-b (which has three versions of each text). We carried out the same kind of ANOVA as for Test Set NEG-a; the result for System on Word String Accuracy was $F_{(9,990)} = 35.123, p < 0.001$. System rankings are the same as for Test Set NEG-a (the differences between JU and Base-freq, which swap ranks, are not significant); scores across the board (again, except for the JU system) are somewhat higher, because of the way scores are computed for version b test

[2] We included the corpus texts themselves in the analysis, hence 9 degrees of freedom (10 systems).

System	String similarity against Corpus (Test Set NEG-a)											
	Word String Accuracy									BLEU-3	NIST	SE	norm. SE
	All						Chefs	Composers	Inventors				
UMUS	78.51	A					76.42	79.29	78.88	0.7968	7.4986	0.6063	0.2019
UMUS'09	75.05	A					69.18	77.66	75.32	0.7615	6.9865	0.6806	0.2233
XRCE	65.25	A					61.01	66.12	67.18	0.7031	6.0264	0.8969	0.3131
JU	60.71	A					72.96	76.63	23.41	0.5720	5.7264	1.1810	0.3671
Base-freq	57.10	A	B				50.31	60.65	56.49	0.5913	4.9860	1.2249	0.4191
UDel-NEG	38.21		B	C			37.42	39.20	37.15	0.5498	5.0211	1.6222	0.5869
Base-name	28.48			C	D		35.53	27.51	24.43	0.4966	4.9355	1.8017	0.6662
Base-rand	8.22				D	E	8.49	7.10	9.92	0.1728	1.2501	2.4290	0.8928
Base-1st	4.69					E	3.46	5.47	4.33	0.1990	2.4018	2.9906	0.8152

Table 6: Word String Accuracy, BLEU, NIST, and string-edit scores, computed on Test Set NEG-a (systems in order of Word String Accuracy); homogeneous subsets (Tukey HSD, alpha = .05) for String Accuracy only.

System	String similarity against human topline (Test Set NEG-b)												
	Word String Accuracy										BLEU-3	NIST	SE	norm. SE
	All							Chefs	Composers	Inventors				
Corpus	81.90	A						83.33	82.25	80.15	0.9499	9.1087	0.7082	0.2517
UMUS	77.29	A	B					79.25	76.48	77.10	0.9296	8.1746	0.8383	0.2906
UMUS'09	74.84	A	B	C				73.58	75.59	74.55	0.8968	7.5005	0.9096	0.3083
XRCE	63.95	A	B	C				66.35	63.02	63.61	0.7960	6.0780	1.1577	0.4060
Base-freq	59.84		B	C	D			55.97	62.72	58.02	0.7393	5.4920	1.3949	0.4717
JU	56.31			C	D	E		68.87	66.86	27.99	0.5765	5.8764	1.5114	0.4720
UDel-NEG	41.60				D	E		44.34	40.38	41.48	0.6503	5.9571	1.7138	0.6057
Base-name	37.27					E		42.14	36.83	34.10	0.6480	6.6551	1.7299	0.6287
Base-rand	10.45						F	10.06	9.91	11.70	0.2468	1.4828	2.4869	0.8884
Base-1st	8.58						F	5.66	10.95	6.87	0.2824	3.5790	2.9226	0.7868

Table 7: Word String Accuracy, BLEU, NIST, and string-edit scores, computed on Test Set NEG-b (systems in order of Word String Accuracy); homogeneous subsets (Tukey HSD, alpha = .05) for String Accuracy.

sets: a score is the highest score a system achieves (at text-level) against any of the three versions of a test set text that is taken into account.

Results for BLEU-3, NIST and the two string-edit distance metrics are shown in the rightmost 4 columns of Tables 6 and 7. With the exception of Base-freq/Basename on Test Set NEG-b, systems whose Word String Accuracy scores differ significantly are assigned the same relative ranks by all other string-similarity metrics as by Word String Accuracy.

In the human intrinsic evaluation, evaluators rated system outputs in terms of whether they preferred them over the original Wikipedia texts. As a result of the experiment we had (for each system and each evaluation criterion) a set of scores ranging from -10.0 to +10.0, where 0 meant no preference, negative scores meant a preference for the Wikipedia text, and positive scores a preference for the system-produced text.

The second column of the left half of Table 8 summarises the Clarity scores for each system in terms of their mean; if the mean is negative the evaluators overall preferred the Wikipedia texts, if it is positive evaluators overall preferred the system. The more negative the score, the more strongly evaluators preferred the Wikipedia texts. Columns 8–10 show corresponding counts of how many times each system was preferred (+), dispreferred (−), and neither (0).

The other half of Table 8 shows corresponding results for Fluency.

We ran a factorial multivariate ANOVA with Fluency and Clarity as the dependent variables. In the first version of the ANOVA, the fixed factors were System, Evaluator and Wikipedia_Side (indicating whether the Wikipedia text was shown on the left or right during evaluation). This showed no significant effect of Wikipedia_Side on either Fluency or Clarity, and no significant interaction between any of the factors. There was also no significant effect of Evaluator on Fluency, and only a weakly significant effect of Evaluator on Clarity. We ran the ANOVA again, this time with just System as the fixed factor. The F-ratio for System on Fluency was $F_{(9,290)} = 22.911, p < .001$, and for System on Clarity it was $F_{(9,290)} = 13.051, p < .001$. Post-hoc Tukey's HSD tests revealed the significant pairwise differences indicated by the letter columns in Table 8.

Correlation between individual Clarity and Fluency ratings as estimated with Pearson's coefficient was $r = 0.66, p < 0.01$, indicating that the two criteria covary to some extent.

Clarity										Fluency									
System	Mean					+	0	−		System	Mean						+	0	−
Corpus	0.000	A					1	28	1	Corpus	0.133	A					1	29	0
UMUS	-2.023	A	B				1	13	16	UMUS	-1.640	A	B				4	12	14
UMUS'09	-2.527	A	B	C			0	15	15	UMUS'09	-2.130	A	B				3	11	16
Base-name	-2.900		B	C			1	7	22	XRCE	-3.587		B	C			2	8	20
Base-1st	-3.160		B	C			4	3	23	JU	-4.057		B	C	D		0	10	20
XRCE	-3.500		B	C	D		1	9	20	Base-freq	-4.990			C	D		1	3	26
JU	-3.577		B	C	D		0	10	20	Base-name	-6.620				D	E	0	1	29
UDel-NEG	-5.137			C	D	E	0	1	29	Base-1st	-7.823					E	1	0	29
Base-freq	-6.190				D	E	0	2	28	Base-rand	-7.950					E	1	0	29
Base-rand	-7.663					E	1	0	29	UDel-NEG	-7.970					E	0	1	29

Table 8: GREC-NEG: Results for Clarity and Fluency preference judgement experiment. Mean = mean of individual scores (where scores ranged from -10.0 to + 10.0); + = number of times system was preferred; − = number of times corpus text (Wikipedia) was preferred; 0 = number of times neither was preferred.

The relative ranks of the peer systems are the same in terms of both Fluency and Clarity. However, there are interesting differences in the ranks of the baseline systems. For Clarity, Base-name and Base-1st are scored fairly highly (presumably because both tend to pick named references which are clear if not always fluent), but both go back to not being significantly better than Base-rand in the Fluency rankings. Base-freq does badly in the Clarity scores, but is significantly better than the bottom three systems in terms of Fluency.

5 GREC-NER

5.1 Task

The GREC-NER task is a straightforward combined named-entity recognition and coreference resolution task, restricted to people entities. The aim for participating systems is to identify all those types of mentions of people that we have annotated in the GREC-People corpus, and to insert REF and REFEX tags with coreference IDs into the texts.

5.2 Systems

Baselines: We used the coreference resolvers included in the LingPipe[3] and OpenNLP Tools[4] packages as baseline systems.

Poly-co: The Poly-co system starts by applying a POS tagger to the input text. A Conditional Random Fields classifier (trained on an automatically annotated Wikipedia corpus) is then used to detect named mentions, using word and POS based features. Logical rules then detect pronoun mentions, using named-entity, word and POS features. Coreference of named mentions is determined by clustering with a similarity measure based on words, POS tags and sentence position, applied to mentions in order of their appearance. Coreference of pronouns is determined with the Hobbs algorithm for anaphora resolution.

UDel-NER: The UDel-NER system starts by (1) parsing the input text with the Stanford Parser, from which it extracts syntactic functions of words and relationships between them; and (2) separately applying the Stanford Named Entity Recognizer. Pronoun and common noun mentions are identified using lists of all English pronouns and of common nouns which could conceivably be used to refer to people (occupations like 'painter', family relations like 'grandmother', etc.). Values for all REF and REFEX attributes except coreference ID are obtained. Finally, the system applies a coreference resolution tool which compares each reference to all previous references in reverse order, on the basis of case, gender, number, syntactic function, and REG'08-Type.

5.3 Results

The coreference resolution accuracy scores for the GREC-NER systems are shown in Table 9. The two participating systems are both significantly better than the two baslines in terms of their mean coreference resolution accuracy scores.

6 GREC-Full

6.1 Task

The aim for GREC-Full systems was to improve the referential clarity and fluency of input texts. Participants were free to do this in whichever way they chose. Participants were encouraged, though not required, to create systems which replace referring expressions as and where necessary to produce as clear and fluent a text as possible. This task could be viewed as composed of three subtasks: (1) named entity recognition (as in GREC-

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

[4] http://opennlp.sf.net

	Test set					
	Mean			B-3	CEAF	MUC
UDel-NER	72.71	A		80.51	77.53	60.09
Poly-co	66.99	A		76.92	70.29	53.77
LingPipe	58.23		B	71.19	61.58	41.92
OpenNLP	54.03		B	67.61	59.17	35.32

Table 9: MUC-6, CEAF and B-3 scores for GREC-NER systems. Systems shown in order of average scores.

NER); (2) a conversion tool to give lists of possible referring expressions for each entity; and (3) named entity generation (as in GREC-NEG).

6.2 Systems

All GREC-Full systems in our evaluations are composed of a GREC-NER and a GREC-NEG system. We created three baseline systems. Two of these we created by combining the two GREC-NER baseline systems with the random GREC-NEG baseline system (Base-rand). For this purpose we created a simple conversion utility which adds default REFEXs. The third baseline system combines the UDel-NER system with Base-rand.

The only team that submitted both a GREC-NER and a GREC-NEG system was UDel. All other GREC-Full systems therefore combine the efforts of two teams (for overview of system combinations, please refer to Table 1). The two system combinations involving the UDel-NER system did not require a conversion utility, because UDel-NER already outputs full GREC-People format.

6.3 Results

NIST and BLEU scores computed against the Wikipedia texts for the GREC-Full systems are shown in Table 10. Note that these have been computed on the complete texts, not just the referential expressions (which explains the high BLEU scores). The scores in the second row (Corpus, test set vers.) are obtained by comparing the test set versions of the corpus texts (in which some of the references have been replaced with standardised named references, as explained in Section 2) against the Wikipedia texts. The two halves of the table show scores computed against version a of the test set (the original Wikipedia texts) on the left, and against version b of the test set (which has three versions of each text with human-selected REs) on the right.

In the human intrinsic evaluation of GREC-Full systems, evaluators again rated system outputs in terms of whether they preferred them over the original Wikipedia texts. Table 11 shows the results in the same format as in Table 8 for the GREC-NEG systems.

We ran the same two factorial multivariate ANOVAs with Fluency and Clarity as the dependent variables. In the first version of the ANOVA, there were no effects of Evaluator (apart from a mild one on Clarity) and Wikipedia_Side and no significant interaction between any of the factors. There was no effect of Evaluator on Fluency and only a mild effect of Evaluator on Clarity. The second ANOVA just had System as the fixed factor. The F-ratio for Fluency was $F_{(6,140)} = 13.054, p < .001$, and for System on Clarity it was $F_{(6,140)} = 14.07, p < .001$. Post-hoc Tukey's HSD tests revealed the significant pairwise differences indicated by the letter columns in Table 11.

Correlation between individual Clarity and Fluency ratings as estimated with Pearson's coefficient was $r = 0.696, p < .01$, indicating that the two criteria covary to some extent.

Apart from UDel-Full and OpenNLP/Base-rand switching places, system ranks are the same for Fluency and Clarity. Moreover, system ranks are very similar to those produced by the string-similarity scores above. UDel-Full is a much harder task than GREC-NEG and it is a very good result indeed for a system to be preferred over Wikipedia once or twice and to be rated equally good as Wikipedia 4–7 times.

7 Concluding Remarks

GREC'10 has, for the first time, produced systems which can do end-to-end named-entity generation, moreover most of which can do it well enough for human judges do rate them as good as Wikipedia or better around one third of the time.

This was the second time the GREC-NEG Task was run, and the first time GREC-NER and GREC-Full were run. As in 2009, many more teams registered than were able to submit a system by the deadline, but we hope that the GREC data (which is now freely available) will lead to many more re-

Test Set NEG-Full-a									Test Set NEG-Full-b								
System	Mean text-level BLEU-4						BLEU-4	NIST	System	Mean text-level BLEU-4						BLEU-4	NIST
Corpus	1.00	A					1.000	13.71	Corpus	.991	A					0.985	13.74
Corpus (test set vers.)	.941		B				0.923	12.92	Corpus (test set vers.)	.946		B				0.929	13.20
UDel/UMUS	.934		B	C			0.925	13.13	UDel/UMUS	.939		B	C			0.928	13.29
UDel/XRCE	.921		B	C			0.898	12.98	UDel/XRCE	.928		B	C			0.907	13.15
UDel-Full	.905			C			0.870	12.59	UDel-Full	.912			C			0.882	12.82
UDel/Base-rand	.812				D		0.809	12.17	UDel/Base-rand	.823				D		0.821	12.43
OpenNLP/Base-rand	.809				D		0.775	11.49	OpenNLP/Base-rand	.817				D		0.785	11.72
LingPipe/Base-rand	.752					E	0.753	11.48	LingPipe/Base-rand	.763					E	0.764	11.70

Table 10: GREC-FULL: Mean text-level BLEU-4 scores, system-level BLEU-4 and NIST scores.

Clarity								Fluency								
System	Mean				+	0	−	System	Mean					+	0	−
Corpus	-0.033	A			1	20	0	Corpus	0	A				0	30	0
UDel/XRCE	-2.209	A	B		0	6	15	UDel/XRCE	-3.424		B			1	4	16
UDel/UMUS	-2.638	A	B		1	6	14	UDel/UMUS	-4.057		B	C		2	5	14
UDel-Full	-2.833		B		0	7	14	OpenNLP/Base-rand	-4.671		B	C		2	4	15
OpenNLP/Base-rand	-3.486		B		1	7	13	UDel-Full	-4.967		B	C		0	4	16
UDel/Base-rand	-4.667		B		0	5	16	UDel/Base-rand	-6.800			C	D	0	2	19
LingPipe/Base-rand	-7.829			C	0	0	21	LingPipe/Base-rand	-8.405				D	0	0	21

Table 11: GREC-FULL: Results for Clarity and Fluency preference judgement experiment. Mean = mean of individual scores (where scores ranged from -10.0 to + 10.0); + = number of times system was preferred; − = number of times corpus text (Wikipedia) was preferred; 0 = number of times neither was preferred.

sults being produced and reported over time.

Acknowledgments

Many thanks to the members of the Corpora and SIGGEN mailing lists, and Brighton University colleagues who helped with the online RE selection experiments for the b-versions of the test sets. Thanks are also due to the Oxford, Kings College London and University College London students who helped with the intrinsic evaluation experiments.

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. 2007. Generation of repeated references to discourse entities. In *Proceedings of ENLG'07*, pages 9–16.

A. Belz. 2010. GREC named entity recognition and GREC named entity regeneration challenges 2010: Participants' Pack. Technical Report NLTG-10-01, Natural Language Technology Group, 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.

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.