The GREC Named Entity Generation Challenge 2009: 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

Abstract

The GREC-NEG Task at Generation Challenges 2009 required participating systems to select coreference chains for all people entities mentioned in short encyclopaedic texts about people collected from Wikipedia. Three teams submitted six systems in total, 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 an intrinsic evaluation involving human judges. This report describes the GREC-NEG Task and the evaluation methods applied, gives brief descriptions of the participating systems, and presents the evaluation results.

1 Introduction

The GREC-NEG task is about how to generate appropriate references to people entities in the context of a piece of discourse longer than a sentence. Rather than requiring participants to generate referring expressions (REs) from scratch, the GREC-NEG data provides sets of possible REs for selection. This was the first time we ran a shared task using this data. GREC-NEG is a step further from the related GREC-MSR Task in that it requires systems to generate plural as well as singular references, for all people entities mentioned in a text (GREC-MSR in contrast only had singular references to a single entity). Moreover in GREC-NEG, possible REs for each entity are provided as one set for each entity (rather than one set for each context), so the task of selecting an appropriate RE for a given context is harder than in GREC-MSR. The main aim for participating systems in GREC-NEG'09 was to select an appropriate *type* of RE (name, common noun, pronoun, or empty reference).

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 REs contained in them.

The *motivating theoretical interest* for the GREC Tasks is to discover what kind of information is useful in the input when making decisions about different properties of referring expressions when such expressions are being generated in context (this is in contrast to most traditional referring expression generation work in NLG which views the REG task as context-independent).

The GREC-NEG data is derived from the newly created GREC-People corpus which consists of 1,000 annotated introduction sections from Wikipedia articles in the category People.

Nine teams from seven countries registered for the GREC-NEG'09 Task, of which three teams ultimately submitted six systems in total (see Table 1). We also used the corpus texts themselves as 'system' outputs, and created four baseline systems. We evaluated the resulting 11 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-NEG data (Sections 2) and task (Section 3), the test sets and evaluation methods (Section 4), and the participating systems (Section 5).

Team	System name(s)
Univ. Delaware	UDel-NEG-1, UDel-NEG-2, UDel-NEG-3
ICSI, Berkeley	ICSI-CRF
Univ. Wolverhampton	WLV-STAND, WLV-BIAS

Table 1: GREC-NEG'09 teams and systems.

2 GREC-NEG Data

The GREC-NEG data is derived from the newly created GREC-People corpus which consists

Proceedings of the 2009 Workshop on Language Generation and Summarisation, ACL-IJCNLP 2009, pages 88–98,
Suntec, Singapore, 6 August 2009.

of 1,000 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, whichever comes sooner. Each text belongs to one of three subcategories: inventors, chefs and early music composers. For the purposes of the GREC-NEG'09 competition, the GREC-People corpus was divided into training, development and test data. The number of texts in the 3 data sets and 3 subdomains are as follows:

	All	Inventors	Chefs	Composers
Total	1,000	307	306	387
Training	809	249	248	312
Development	91	28	28	35
Test	100	31	30	39

In these texts we have annotated mentions of people by marking up the word strings that function as referential expressions (REs) and annotating them with coreference information as well as syntactic and semantic features. The subject of each text is a person, so there is at least one coreference chain in each text. The numbers of coreference chains (entities) in the 900 texts in the training/development sets are as shown in Table 2. The texts vary greatly in length, from 13 words to 935, with an average of 128.98 words.

2.1 Annotation of REs in GREC-People

This section describes the different types of referring expression (RE) that we annotated in the GREC-People corpus. These manual annotations were then automatically checked and converted to the XML format described in Section 2.2 (which encodes slightly less information, as explained below). In terminology and the treatment of syntax used in the annotation scheme and discussion of it in this report we rely heavily on *The Cambridge Grammar of the English Language* by Huddleston and Pullum which we will refer to as *CGEL* for short below (Huddleston and Pullum, 2002).

In the example sentences below, (unbroken) underlines are used for referential expressions (REs) that are an example of the specific type of RE they are intended to illustrate, whereas dashed underlines are used for other annotated REs. Coreference between REs is indicated by subscripts $i, j, ...$ immediately to the right of an underline (their scope is one example sentence, i.e. an i in one example sentence does not represent the same entity as an i in another example sentence). Square brackets indicate supplements. The syntactic component relativised by a relative pronoun is indicated by vertical bars. Supplements and their anchors (in the case of appositive supplements), and relative clauses and the component they relativise (in the case of relative-clause supplements) are coindexed by superscript $x, y, ...$ Dependents integrated in an RE are indicated by curly brackets. Supplements and dependents are highlighted in bold where they specifically are being discussed.

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 discussed in the following sections (e.g. syntactic category) are encoded as attributes on these tags (for details see Section 2.2). For GREC-NEG'09 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.

2.1.1 Syntactic Category and Function

This section describes the types of REs we annoted in the GREC-People Corpus.

I Subject NPs: referring subject NPs, including pronouns and special cases of VP coordination:

1. *He$_i$ was born in Ramsay township, near Almonte, Ontario, Canada, the eldest son of $|$Scottish immigrants, {John Naismith and Margaret Young}$|^x_{j,k}$ $[$who$_{j,k}$ had arrived in the area in 1851 and $\_\_{}_{j,k}$ worked in the mining industry$]^x$.*

2. *The Banū Mūsā brothers$_{i,j,k}$ were three 9th century Persian scholars, of Baghdad, active in the House of Wisdom.*

Ia Subjects of gerund-participials:

1. *His$_i$ research on hearing and speech eventually culminated in Bell$_i$ being awarded the first U.S. patent for the invention of the telephone in 1876.*

2. *Fessenden$_i$ used the alternator-transmitter to send out a short program from Brant Rock, which included his$_i$ playing the song O Holy Night on the violin and $\_\_{}_i$ reading a passage from the Bible.*

II Object NPs: referring NPs including pronouns that function as direct or indirect objects of VPs and prepositional phrases; e.g.:

1. *Many of the alpinists arrested with Vitaly Abalakov$_i$ were executed.*

2. *He$_i$ entrusted them$_{j,k,l}$ to Ishaq bin Ibrahim al-Mus'abi^x_m, $[$a former governor of Baghdad$]^x_m$.*

Entities	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Texts	437	192	80	63	38	31	16	18	4	7	9	1	1	0	0	0	0	0	0	1	1	0	1

Table 2: Numbers of person entities (hence coreference chains) in texts in the training/development data, e.g. there are 38 texts which mention exactly 5 person entities.

IIa Reflexive pronouns:

1. *Smith$_i$ called himself$_i$ the "Komikal Konjurer".*

III Subject-determiner genitives: genitive NPs (including genitive forms of pronouns) that function as subject-determiners, i.e. syntactic components that "combine the function of determiner, marking the NP as definite, with that of complement (more specifically subject)." (*CGEL*, p. 56):

1. *They$_{i,j,k}$ shared the 1956 Nobel Prize in Physics for their$_{i,j,k}$ invention.*
2. *On the eve of his$_i$ death in 1605, the Mughal empire spanned almost 500 million acres (doubling during Akbar's$_i$ reign).*

Note that this category excludes lexicalised cases, e.g. *the so-called "Newton's method"*.

IIIa REs in composite nominals: this is the only type of RE we have annotated that is not an NP, but a nominal. This type functions as integrated attributive complement, e.g.:

1. *The Eichengrün$_i$ version was ignored by historians ...*
2. *The new act was a great success, largely despite the various things Blackton$_i$ and Smith$_j$ were doing between the Edison$_k$ films.*

Note that this category too excludes lexicalised cases, e.g. *the Nobel Prizes; the Gatling gun*.

2.1.2 Annotation of supplements

We have annotated two kinds of supplements in the GREC-People corpus, **supplementary relative clauses** (*CGEL*, p. 1058), and **appositive supplements**. The former is not transferred to the XML annotation, for more information see (Belz, 2009).

The following examples illustrate annotation of appositive supplements (which are in bold):

1. *John W. Campbell, Jr.*^x_i [***the editor of Astounding magazine***$_i$]^x.
2. was *the eldest of the six children of Thomas Aspdin*^x_i, [***a bricklayer living in the Hunslet district of Leeds***$_i$]^x

In the XML version, anchor and supplement are simply annotated as two (or occasionally three) independent, usually adjacent REs (REFEXs); the syntactic function of the second (and third) RE is marked as appositive supplement (SYNFUNC="app-supp").

2.1.3 Further aspects of the annotation

As can be seen from some of the examples above, we annotated all **embedded references**. The maximum depth of embedding that occurs in the GREC-People corpus is 3.

We annotated all **plural REs** that refer to groups of people where the number of group members is known. For an explanation of our treatment of REs that are coordinations of NPs, see the GREC-NEG'09 documentation (Belz, 2009).

We have annotated all mentions of individual person entities even if they are not actually named anywhere in the text, and including cases of both definite and indefinite references, e.g.:

1. *The resolution's sponsor$_i$ described it as ...*
2. *... with the help of Robert Cailliau$_j$ and a {young} student staff {at CERN}$_k$.*

2.2 XML Annotation

Figure 1 shows one of the XML-annotated texts from the GREC-NEG data. Each such text consists of two initial lines of XML declarations followed by a GREC-ITEM. A GREC-ITEM consists of a TEXT element followed by an ALT-REFEX element. A TEXT has one attribute (an ID unique within the corpus), and is composed of one TITLE followed by any number of PARAGRAPHs. A TITLE is just a string of characters. A PARAGRAPH is any combination of character strings and REF elements.

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). A REF 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). For full details and ranges of values see (Belz, 2009). ENTITY and MENTION together constitute a unique identifier for a reference within a text; together

```
<?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.

with the `TEXT ID`, they constitute a unique identifier for a reference within the entire 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. See also next but one paragraph for embedding in `REFEX` elements that are contained in `ALT-REFEX` lists.

The second (and last) component of a `GREC-ITEM` is an `ALT-REFEX` element which is a list of `REFEX` elements. For the GREC-NEG'09 Task, these were obtained by collecting the set of all `REFEX`s 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. For example:

```
<ALT-REFEX>
  ...
  <REFEX ENTITY="2" REG08-TYPE="common" CASE="plain">
    a friend of <REF ENTITY="1" MENTION="?" SEMCAT=
    "person" SYNCAT="np" SYNFUNC="obj"></REF></REFEX>
  ...
</ALT-REFEX>
```

3 The GREC-NEG Task

The test data inputs were 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 em-

bedded 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 Evaluation Procedures

The GREC-NEG data set was divided into training, development and test data. We performed evaluations on the test data, using a range of different evaluation methods, including intrinsic and extrinsic, automatically assessed and human-evaluated, as described in the following sections.

Participants computed evaluation scores on the development set, using the `geval-2.0.pl` code provided by us which computes Word String Accuracy, REG'08-Type Recall and Precision, string-edit distance and BLEU.

4.1 Test sets

We created two versions of the test data for the GREC-NEG Task:

1. GREC-NEG Test Set 1a: randomly selected 10% subset (100 texts) of the GREC-People corpus (with the same proportion of texts in the 3 subdomains as in the training/development data).

2. GREC-NEG Test Set 1b: the same subset of texts as in (1a); 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 test set therefore contains three versions of each text where all the REFEXs in a given version were selected by one 'author'.

Test Set 1a has a single version of each text, and the scoring metrics below 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.

Test Set 1b, 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.

4.2 Automatic intrinsic evaluations

The chief humanlikeness measures we computed were REG08-Type Recall and Precision. REG08-Type Precision is defined as the proportion of REFEXs selected by a participating system which match the reference REFEXs (where match counts are obtained as explained in the preceding section). REG08-Type Recall is defined as the proportion of reference REFEXs for which a participating system has produced a match.

The reason why we use REG08-Type Recall and Precision for GREC-NEG rather than REG08-Type Accuracy as in GREC-MSR is that in GREC-NEG (unlike in GREC-MSR) there may be a different number of REFEXs in system outputs and the reference texts in the test set (because there are embedded references in GREC-People, and systems may select REFEXs with or without embedded references for any given REF).

We also computed String Accuracy, 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). For details, see GREC-MSR report in this volume.

4.3 Human-assessed intrinsic evaluations

Given that the motivating application context for the GREC-NEG Task is improving referential clarity and coherence in multiply edited texts, we designed the human-assessed intrinsic evaluation as a preference-judgment 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. The intrinsic human evaluation involved outputs for 30 randomly selected items from the test set from 5 of the 6 participating systems,[1] the four baselines and the original corpus texts (10 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 10x10 squares, and a total of 600 individual judgments in this evaluation (60 per system: 2 criteria x 3 articles x 10

[1] We left out UDel-NEG-1 given our limited resources and the fact that this is a kind of baseline system.

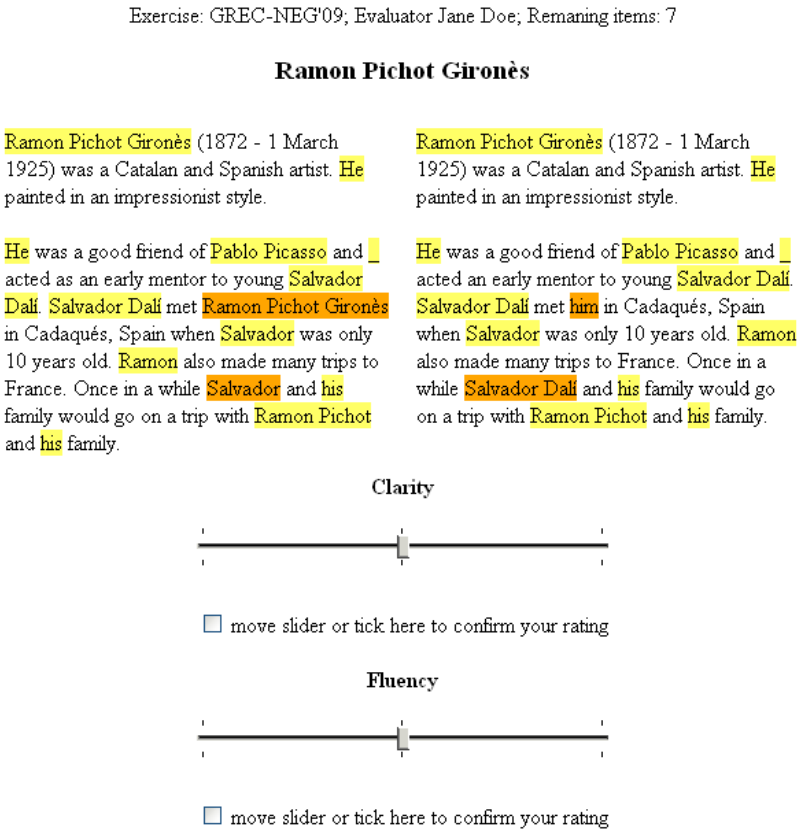

Figure 2: Example of text pair presented in human intrinsic evaluation of GREC-NEG systems.

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.

Following detailed instructions, subjects did two practice examples, followed by the 30 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. 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 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.

4.4 Extrinsic automatic evaluation

An evaluation we piloted in REG'08 was an automatic approach to extrinsic evaluation (for a more detailed description, see the GREC-MSR results report elsewhere in this volume). The basic premise is that poorly chosen reference chains seem likely to affect the reader's ability to resolve REs. In our automatic extrinsic method, the role of the reader is played by an automatic coreference resolution tool and the expectation is that the tool performs worse (is less able to identify coreference chains) with more poorly chosen referential expressions.

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[2] and OpenNLP (Morton, 2005)—and averaged results. For the same reason we used three different performance measures: MUC-6 (Vilain et al., 1995), CEAF (Luo, 2005), and B-CUBED (Bagga and Baldwin, 1998).

5 Systems

Base-rand, Base-freq, Base-1st, 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. *Base-freq* selects the first REFEX with a REG08-TYPE that is the overall most frequent (as determined from the training/development data) given the SYNCAT, SYNFUNC and SEMCAT of the reference. *Base-name* selects the shortest REFEX with attribute REG08-TYPE=name.

UDel: The UDel-NEG-1 system is identical to the UDel system that was submitted to the GREC-MSR Task (for a description of that system see GREC-MSR'09 results report in this volume), except that it was adapted to the different data format of GREC-NEG. UDel-NEG-2 is identical to UDel-NEG-1 except that it was retrained on GREC-NEG data and the feature set was extended by entity and mention IDs. UDel-NEG-3 additionally utilised improved identification of other entities.

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

WLV: The WLV systems start with sentence splitting and POS tagging. WLV-STAND then employs a J48 decision tree classifier to obtain a probability for each REF/REFEX pair that it is a good pair in the current context. The context is represented by the following set of features. Features of the REFEX word string: is it the longest of the possible REFEXs; number of words; all REFEX features supplied in GREC-NEG data. Features of the REF: is it part of the first chain in the text; is it the first mention of the entity; is it at the beginning of the sentence; all REF features supplied in GREC-NEG data. Other features: do the preceding words match ", but", "and then" and similar phrases; distance in sentences to last mention; REG08-Type selected for the two preceding REFs; POS tags of 4 words before and 3 words after; correlation between SYNFUNC and CASE values; size of the chain.

WLV-BIAS is the same except that it is retrained on reweighted training instances. The reweighting scheme assigns a cost of 3 to false negatives and 1 to false positives.

6 Results

This section presents the results of all the evaluation methods described in Section 4. We start with REG08-Type Precision and Recall, the intrinsic automatic metrics which participating teams were told was going to be the chief evaluation method, followed by Word String 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	REG08-Type		WS Acc.	Norm. SE
	Recall	Precision		
ICSI-CRF	83.05	83.05	0.786	0.197
WLV-BIAS	77.61	80.26	0.735	0.239
UDelNEG-3	75.27	75.27	0.333	0.636
UDelNEG-2	74.95	74.95	0.323	0.646
UDelNEG-1	68.87	68.87	0.315	0.658
WLV-STAND	66.20	68.46	0.626	0.351

Table 5: Self-reported evaluation scores for development set.

6.1 REG08-Type Precision and Recall

Participants computed scores for the development set (91 texts) themselves, using the geval evaluation tool provided by us. These scores are shown in Table 5, and are also included in the participants' reports elsewhere in this volume.[3]

REG08-Type Recall and Precision results for Test Set 1a are shown in column 2 of Table 3. As would be expected, results on the test data are

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

[3] ICSI-CRF scores obtained directly from ISCI team.

System	REG08-Type Precision and Recall Scores against Corpus (Test Set 1a)																	
	All													Chefs		Composers		Inventors	
	Precision							Recall						R	P	R	P	R	P
ICSI-CRF	79.12	A						76.92	A					70.01	73.54	78.11	80.18	80.05	81.86
WLV-BIAS	73.77		B					72.70	A					69.82	71.52	73.53	74.38	73.65	74.56
WLV-STAND	64.49			C				63.55		B				58.28	59.70	65.38	66.14	64.78	65.59
Base-freq	61.52			C				59.6		B				49.41	51.86	63.95	65.74	60.59	62.12
UDel-NEG-2	53.21				D			51.14			C			44.38	47.17	50.50	52.22	57.88	59.80
UDel-NEG-3	52.49				D			50.45			C			43.49	46.23	49.79	51.48	57.39	59.29
UDel-NEG-1	50.47				D			48.51			C			42.90	45.60	47.78	49.41	54.43	56.23
Base-rand	43.32					E		42.00				D		38.76	40.43	41.77	43.00	45.07	46.21
Base-name	40.60					E		39.09				D		44 97	47.80	39.06	40.32	34.24	35.28
Base-1st	10.99						F	10.81					E	12.43	12.73	9.30	9.43	12.07	12.22

Table 3: 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 1b)																		
	All													Chefs		Composers		Inventors	
	Precision							Recall							R	P	R	P	R	P
Corpus	82.67	A						84.01	A						84.24	82.25	84.47	83.26	83.04	82.02
ICSI-CRF	79.33	A	B					78.38		B					76.36	77.54	78.81	79.74	79.30	80.10
WLV-BIAS	77.78		B					77.78		B					77.58	77.58	77.86	77.86	77.81	77.81
WLV-STAND	67.51			C				67.51			C				65.76	65.76	68.60	68.60	67.08	67.08
Base-freq	65.38			C				64.37			C				58.48	59.94	68.07	68.97	62.84	63.64
UDel-NEG-2	57.39				D			56.06				D			55.15	57.23	54.86	55.92	58.85	60.05
UDel-NEG-3	57.25				D			55.92				D			55.76	57.86	54.57	55.62	58.35	59.54
Base-name	55.22				D			54.01				D			54.24	56.29	57.04	58.05	48.63	49.49
UDel-NEG-1	53.57				D			52.32				D	E		51.21	53.14	50.80	51.78	55.86	57.00
Base-rand	48.46					E		47.75					E		47.88	48.77	46.44	47.13	49.88	50.51
Base-1st	12.54						F	12.54						F	13.94	13.94	10.45	10.45	14.96	14.96

Table 4: 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.

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 given SEMCAT, SYNFUNC and SYNCAT (as done by the Base-freq system) provides a strong baseline for RE type selection.

The last 6 columns in Table 3 contain Recall (R) and Precision (P) results for the three subdomains. For most of the systems results are slightly better for Inventors than for Composers, and better for Composers than for Chefs. A contributing factor to this may be the fact that texts in Chefs tend to be much more colloquial. Base-1st has by far the worst results; this is because it selects the empty reference in almost all cases (because ALT-REFEX lists are sorted and if a list contains an empty reference it will end up at the beginning).

We carried out univariate ANOVAs with System as the fixed factor, and 'Number of REFEXs in a text' as a random factor, and REG08-Type Recall as the dependent variable in one ANOVA, and REG08-Type Precision in the other. The result for Recall was $F_{(10,704)} = 81.547, p < 0.001$.[4] The result for Precision was $F_{(10,722)} = 79.359, p < 0.001$. The columns containing capital letters in Table 3 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 4 shows analogous results computed against Test Set 1b (which has three versions of each text). These should be considered as the chief results of the GREC-NEG'09 Task evaluations, as stated in the participants' guidelines. Table 4 includes results for the corpus texts, computed (as are results for the system outputs in Table 4) against the three versions of each text in Test Set 1b. We performed univariate ANOVAs with System as the fixed factor, Number of REFEXs as a random factor, and Recall as the dependent variable in one, and Precision in the other. The result for Recall was $F_{(10,724)} = 72.528$, $p < .001$), and for Precision $F_{(10,722)} = 75.476$, $p < .001$. For both cases, we compared the mean scores with Tukey's HSD. As can be seen from the resulting homogeneous subsets (letter columns in Table 4), system ranks are the same for Precision and for Recall. In terms of Precision, the difference between the corpus texts and the ICSI-CRF system was not significant.

6.2 Other automatic intrinsic metrics

In addition to the chief evaluation measure reported on in the preceding section, we computed

[4] We included the corpus texts themselves in the analysis, hence 10 degrees of freedom (11 systems).

System	String similarity against Corpus (Test Set 1a)														
	Word String Accuracy											BLEU-3	NIST	SE	norm. SE
	All								Chefs	Composers	Inventors				
ICSI-CRF	74.84	A							68.24	76.63	77.10	0.75	5.78	0.70	0.23
WLV-BIAS	68.57		B						66.35	69.08	69.47	0.76	5.62	0.82	0.29
WLV-STAND	59.55			C					54.72	61.24	60.56	0.73	5.34	1.01	0.39
Base-name	28.48				D				35.53	27.51	24.43	0.5	4.09	1.80	0.67
UDel-NEG-1	16.58					E			20.13	15.09	16.28	0.43	2.47	2.1	0.82
UDel-NEG-2	16.44					E			19.81	14.79	16.54	0.45	2.37	2.08	0.83
UDel-NEG-3	16.37					E			19.18	15.09	16.28	0.45	2.41	2.08	0.83
Base-rand	8.22						F		8.49	7.10	9.92	0.17	0.9	2.43	0.89
Base-1st	7.28						F		7.23	6.36	8.91	0.16	0.98	2.54	0.90
Base-freq	2.52							G	4.40	2.37	1.27	0.31	1.91	2.34	0.90

Table 6: Word String Accuracy, BLEU, NIST, and string-edit scores, computed on Test Set 1a (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 1b)														
	Word String Accuracy											BLEU-3	NIST	SE	norm. SE
	All								Chefs	Composers	Inventors				
Corpus	81.90	A							83.33	82.25	80.15	0.95	7.15	0.71	0.25
ICSI-CRF	74.55		B						71.70	75.15	75.83	0.86	6.35	0.92	0.31
WLV-BIAS	69.07			C					69.50	68.49	69.72	0.88	6.17	1.03	0.36
WLV-STAND	59.70				D				58.18	60.36	59.80	0.84	5.81	1.21	0.45
Base-name	37.27					E			42.14	36.83	34.10	0.65	5.57	1.73	0.63
UDel-NEG-1	19.25						F		22.96	17.60	19.08	0.51	2.62	2.17	0.82
UDel-NEG-2	18.96						F		22.96	17.31	18.58	0.53	2.42	2.15	0.83
UDel-NEG-3	18.89						F		22.64	17.75	17.81	0.53	2.49	2.15	0.82
Base-rand	10.45							G	10.06	9.91	11.70	0.25	1.11	2.49	0.89
Base-1st	8.65							G	8.49	7.54	10.69	0.24	1.29	2.64	0.92
Base-freq	3.24								4.40	3.55	1.78	0.39	2.1	2.40	0.90

Table 7: Word String Accuracy, BLEU, NIST, and string-edit scores, computed on Test Set 1b (systems in order of Word String Accuracy); homogeneous subsets (Tukey HSD, alpha = .05) for String Accuracy.

Word String Accuracy and the other string similarity metrics described in Section 4.2. The resulting scores for Test Set 1a (the corpus texts) are shown in Table 6. Ranks for peer systems relative to each other are very similar to the results reported in the last section. However, the ranks of the baseline systems have changed substantially, both in relation to each other and to the peer systems. In particular, Base-freq has moved all the way down to the bottom of the table. The reason is that this method is geared towards selecting the correct type of RE, but pays no attention to whether it selects a syntactically appropriate RE for the given context, instead simply selecting the first RE from the ALT-REFEX list that has the selected type; in the GREC-NEG'09 Task (unlike the GRE-MSR task) this just happens to be an RE in the genitive case most of the time which is overall rarer than nominative/plain. It is likely that the Word String scores for the UDel-NEG systems are low for a similar reason.

We performed a univariate ANOVA with System as the fixed factor and Number of REFEXs as a random factor and Word String Accuracy as the dependent variable. The result for System was $F_{(10,726)} = 103.339$; the homogeneous subsets resulting from the Tukey HSD post-hoc analysis are shown in columns 3–9 of Table 6.

Table 7 shows analogous results for human topline Test Set 1b (which has three versions of each text). We carried out the same kind of ANOVA as for Test Set 1a; the result for System on Word String Accuracy was $F_{(10,726)} = 106.755, p < 0.001$. System rankings and homogeneous subsets are the same as for Test Set 1a; scores across the board are somewhat higher, because of the way scores are computed for Test Set 1b: it 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. Systems whose Word String Accuracy scores differ significantly are assigned the same ranks by NIST and the two string-edit distance metrics as by Word String Accuracy (except for Base-1st and Base-freq which swap ranks in some. BLEU-3 does the same and also flips ICSI-CRF and WLV-BIAS.

6.3 Human-assessed intrinsic measures

In the human intrinsic evaluation, evaluators rated system outputs in terms of whether they preferred them over the original Wikipedia texts. As a re-

Clarity											Fluency									
System	Mean							+	0	−	System	Mean						+	0	−
Corpus	0	A						0	30	0	Corpus	0	A					0	30	0
ICSI-CRF	-1.447	A	B					3	17	10	ICSI-CRF	-0.353	A					9	14	7
WLV-BIAS	-2.437	A	B	C				3	14	13	WLV-BIAS	-2.257	A	B				2	14	14
Base-name	-2.583		B	C				7	7	16	WLV-STAND	-5.823		B	C			1	3	26
WLV-STAND	-4.477			C	D			1	9	20	Base-name	-4.257			C	D		2	5	23
UDelNEG-3	-6.427				D	E		1	4	26	UDelNEG-3	-6.263			C	D	E	1	3	26
UDelNEG-2	-6.667				D	E		1	3	26	UDelNEG-2	-7.13				D	E	0	3	27
Base-rand	-8.183					E	F	0	1	29	Base-rand	-7.513				D	E	0	0	30
Base-freq	-8.26					E	F	0	0	30	Base-freq	-7.57				D	E	0	0	30
Base-1st	-9.357						F	0	0	30	Base-1st	-8.477					E	0	0	30

Table 8: 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.

sult 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 9-11 show corresponding counts of how many times each system was preferred (+), dispreferred (−), and neither (0), when compared to Wikipedia.

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 however a mild effect of Evaluator on both Fluency and Clarity. We ran the ANOVA again, this time with just System and Evaluator as fixed factors. The result for System on Fluency was $F_{(9,200)} = 37.925, p < .001$, and for System on Clarity it was $F_{(9,200)} = 35.439, p < .001$. 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 $r = 696, p < .01$, indicating that the two criteria covary to some extent.

Apart from Base-name and WLV-STAND 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. Perhaps the most striking result is that the ICSI-CRF system does succeed in improving Fluency compared to the original Wikipedia texts: it is preferred 9 times whereas the Wikipedia texts are preferred only 7 times.

System	(MUC+CEAF+B3)/3							M	C	B3
WLV-BIAS	62.64	A						57	62	69
ICSI-CRF	61.28	A	B					53	61	69
Base-name	61.11	A	B					55	61	68
Corpus	59.56	A	B	C				53	59	67
UDel-NEG-3	56.13		B	C	D			48	56	65
UDel-NEG-2	55.9		B	C	D			47	55	65
Base-freq	55.85		B	C	D			47	56	65
UDel-NEG-1	54.79			C	D			46	54	64
WLV-STAND	51.69				D			41	53	61
Base-rand	34.86					E		15	38	51
Base-1st	26.36						F	2	31	46

Table 9: MUC, CEAF and B-CUBED F-Scores for all systems; homogeneous subsets (Tukey HSD), alpha = .05, for mean of F-Scores.

6.4 Automatic extrinsic measures

We fed the outputs of all 11 systems through the two coreference resolvers, and computed mean MUC, CEAF and B-CUBED F-Scores as described in Section 4.4. The second column in Table 9 shows the mean of means of these three F-Scores, to give a single overall result for each of for this evaluation method. A univariate ANOVA with (text-level) mean F-Score as the dependent variable and System as the single fixed factor revealed a significant main effect of System on mean F-Score ($F_{(10,1089)} = 91.634, p < .001$). A post-hoc comparison of the means (Tukey HSD, alpha = .05) found the significant differences indicated by the homogeneous subsets in columns 3–8 (Table 9). The numbers in the last three columns are the separate MUC, CEAF and B-CUBED F-Scores

for each system, averaged over the two resolver tools (and rounded for reasons of space.

7 Concluding Remarks

This was the first time the GREC-NEG Task was run. It is a 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-NEG Task (as indicated by the nine teams that originally registered). However, only 3 teams were ultimately able to submit a system.

In particular because of the inclusion of plural references, multiple entities per text and embedded references, the GREC-NEG Task has a higher entrance level than the GREC-MSR Task. We are planning to run it again at Generation Challenges 2010 next year, and are considering the possibility of providing participants with a baseline system which would help e.g. with processing embedded references.

We are also planning to add a named entity recognition preprocessing task, so that this new task in combination with GREC-NEG can be used to perform end-to-end post-processing of extractive summaries (and other types of multiply edited texts) to improve the clarity and fluency of the referring expressions in them.

Acknowledgments

Many thanks to the members of the Corpora and SIGGEN mailing lists, and Brighton University colleagues who helped with the online MSRE selection experiments for GREC-NEG test set 1b. Thanks are also due to the Kings College London and University College London students who helped with 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. 2007. The GREC corpus: Main subject reference in context. Technical Report NLTG-07-01, University of Brighton.

A. Belz, 2009. *GREC Named Entity Generation Challenge 2009: Participants' Pack.*

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.

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.