A Repository of Data and Evaluation Resources for Natural Language Generation

Anja Belz
University of Brighton
Lewes Road
Brighton BN2 4GJ, UK
a.s.belz@brighton.ac.uk

Albert Gatt
University of Malta
Tal-Qroqq
Msida MSD2080, Malta
albert.gatt@um.edu.mt

Abstract

Starting in 2007, the field of natural language generation (NLG) has organised shared-task evaluation events every year, under the Generation Challenges umbrella. In the course of these shared tasks, a wealth of data has been created, along with associated task definitions and evaluation regimes. In other contexts too, sharable NLG data is now being created. In this paper, we describe the online repository that we have created as a one-stop resource for obtaining NLG task materials, both from Generation Challenges tasks and from other sources, where the set of materials provided for each task consists of (i) task definition, (ii) input and output data, (iii) evaluation software, (iv) documentation, and (v) publications reporting previous results.

Keywords: Natural Language Generation, Evaluation Resources, Data Resources.

1. Introduction

Since 2007, the Natural Language Generation (NLG) community has organised a number of shared-task evaluation challenges under the Generation Challenges umbrella. These originated in various special sessions and workshops organised to address a growing interest in sharing data and evaluation methods for specific NLG sub-tasks. Such exercises are well established in other areas of NLP, including Summarisation,[1] Information Retrieval[2] and Machine Translation.[3] One of the benefits they bring to a research community is the creation of datasets for system development and evaluation, and the results of one or more evaluation measures for participating systems; in some cases, outputs from participating systems are also made available. In combination, these resources can serve as benchmarks for evaluating new approaches.

The present paper describes an effort to create a single repository of datasets and other resources related to different NLG tasks, most of which were organised as shared-task evaluations between 2007 and 2011. We describe each of the tasks, the evaluation metrics used and the data and software available. Since their inception in 2007, the range of different NLG shared tasks has grown substantially, with many new tasks being organised in the last three years. Hence, there is the potential for the repository to become a continuously developing resource.

2. The GenChal Repository

The GenChal Repository (https://sites.google.com/site/genchalrepository) is structured according to NLG subfield. Within each subfield, there are one or more datasets and other resources corresponding to a particular family of shared tasks, as shown in Figure 1. The subfields currently covered are the following:

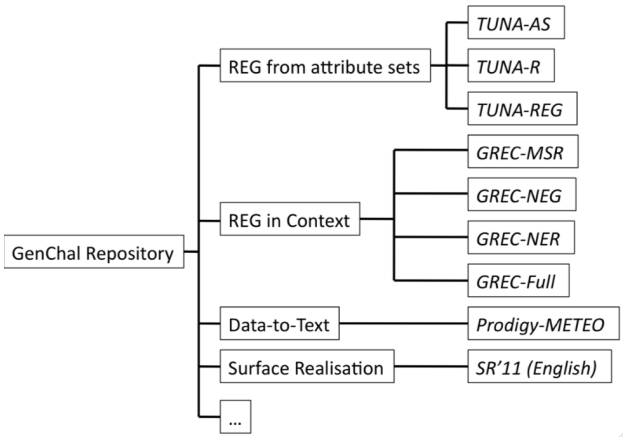

Figure 1: Structural overview of GenChal Repository.

1. *Referring expression generation* (REG): a sub-task required by many NLG systems, involving the selection of the semantic content and linguistic form of a noun phrase to identify an entity in the discourse context (Krahmer and van Deemter, 2012). Within this subfield, data is currently provided related to the TUNA tasks, which focus on the generation of one-off definite referring expressions (see Section 2.1.) and the GREC (Generation of Referring Expressions in Context) tasks which focus on the generation of chains of referring expressions in full discourse context (see Section 2.2.).

2. *Data-to-text generation*: a field of NLG in which text is generated from input data in raw (often numeric) form, rather than from inputs which are in some canonical semantic form (Reiter, 2007). Within this subfield, the repository currently provides the Prodigy-METEO Corpus, which focuses on the generation of weather reports from numerical data and provides input and output data as well as previous system outputs.

[1] See, for example, http://duc.nist.gov/pubs.html.
[2] See http://trec.nist.gov/.
[3] See http://www.itl.nist.gov/iad/mig/tests/mt.

3. *Surface realisation*: the task of rendering semantic inputs into a natural language string by applying morphosyntactic operations (Reiter and Dale, 2000). Here, the resources provided are related to the Surface Realisation shared task (SR'11) organised for the first time in 2011 and involving the conversion of input structures of varying degrees of complexity extracted from the Wall Street Journal corpus into English sentences (see 2.4.).

For each task, we provide the following resources in the repository:

1. *Task Definition*: A precise definition of the task, its inputs, outputs and the overall aim that peer systems are required to achieve.

2. *Input and Output Data*: A set of inputs and corresponding target outputs, subdivided into training, development and test data. We also provide outputs from existing systems (baseline and competitive), and in some cases additional human-authored outputs for the given training, development and test data.

3. *Evaluation Resources*:

   (a) A list of automatic evaluation tools, brief descriptions of the algorithms they implement and a link for downloading them; and

   (b) a list of human evaluation methods previously applied to the task, explained in sufficient detail to facilitate replication.

4. *Documentation*: In the case of tasks that were previously run as shared-task competitions, the original task documentation published for the task (if the task was run in multiple years, we provide the most recent documentation). In the case of other tasks, we provide another publication that describes the task in sufficient detail.

5. *List of Previous Publications*: A bibliography of existing publications that report results for the same task, using the same evaluation methods.

2.1. TUNA Resources

The TUNA shared tasks were run between 2007 and 2009 (Gatt and Belz, 2010) and focused on the NLG sub-task of referring expression generation (REG, for explanation see previous section).

The TUNA tasks were based on the TUNA Corpus (van Deemter et al., in press), and focused on the generation of full, identifying, definite noun phrases in domains where the entities are either furniture items or people. The corpus consists of descriptions elicited from human participants in an experiment in which they were asked to produce a description to identify an object in a visual domain where there were also a number of distractors. The corpus contains a representation of this visual domain (entities and their properties), together with the human-authored description.

```
<TRIAL ID="t101">
  <DOMAIN>

    <ENTITY ID="e1" TYPE="target">
      <ATTRIBUTE NAME="type" VALUE="sofa" />
      <ATTRIBUTE NAME="colour" VALUE="blue" />
      <ATTRIBUTE NAME="size" VALUE="large" />
      ...
    </ENTITY>

    <ENTITY ID="e2" TYPE="distractor">
      <ATTRIBUTE NAME="..." VALUE="..." />
      ...
    </ENTITY>
    ...
  </DOMAIN>

  <WORD-STRING>
    the blue sofa
  </WORD-STRING>

  <ANNOTATED-WORD-STRING>
    the
    <ATTRIBUTE NAME="colour"
               VALUE="blue">blue</ATTRIBUTE>
    <ATTRIBUTE NAME="type"
               VALUE="sofa">sofa</ATTRIBUTE>
  </ANNOTATED-WORD-STRING>

  <ATTRIBUTE-SET>
    <ATTRIBUTE NAME="colour" value="blue"/>
    <ATTRIBUTE NAME="type" value="sofa"/>
  </ATTRIBUTE-SET>
</TRIAL>
```

Figure 2: A TUNA Corpus instance.

The data used in the TUNA shared tasks comprises a subset of the TUNA Corpus, and the mark-up was converted to the following format: each corpus instance is an XML file pairing a DOMAIN node, which subsumes each domain ENTITY and its properties (represented as ATTRIBUTE nodes with values), with the human-written description identifying the *target* entity from the remaining entities (the *distractors*). There are three representations for the human description in each corpus instance: the WORD-STRING node reproduces the description as a string; the ANNOTATED-WORD-STRING node reproduces the string and identifies those substrings that correspond to ATTRIBUTEs in the DOMAIN, while the ATTRIBUTE-SET node lists only the domain attributes of the entity mentioned in the human description. An example is shown in Figure 2.

There were three different TUNA tasks. These are described below.

2.1.1. TUNA-AS

Task definition: TUNA-AS, organised in 2007 and 2008, focused on selecting the content for the description of a target entity; it required peers to develop a method that takes as input a DOMAIN, in which one ENTITY was the target, and return an ATTRIBUTE-SET consisting of a subset of

the target entity attributes which will help to distinguish it from its distractors.

Input and output data: The data from the TUNA Corpus, comprising only singular descriptions,[4] was randomly divided into 60% training, 20% development and 20% test datasets, each set containing exemplars from both the furniture and people domain.

For the 2008 edition, a new test set of 112 items was created by replicating the original TUNA elicitation experiment (this test set was also used in the 2008 and 2009 editions of the TUNA-REG Task, see below). The new experiment was run in order to obtain test sets in which each domain had two reference outputs.

Evaluation: For the TUNA-AS task, we developed a Java program that, given a corpus of TUNA instances, computes (i) two coefficients, Dice and MASI (Passonneau, 2006) that assessed the degree of overlap between ATTRIBUTE-SETs generated by a peer system with those produced by a human; (ii) *accuracy*, the proportion of peer outputs that were identical to human outputs; and (iii) *minimality*, the proportion of peer outputs which contain no more attributes than necessary to distinguish the target. The minimality criterion is theoretically motivated and based on the REG literature, where it has been proposed that descriptions should obey the Gricean maxim of quantity, which in this case implies including no more information in a description than required by the purpose of identification (Dale, 1989).

For the 2008 edition we added MASI as a metric and reimplemented some of the metrics.

Other metrics: In 2007, a laboratory experiment was run for this task in which subjects were shown referring expressions paired with their respective visual domains, and had to identify the target referent by clicking on it. The descriptions used in the experiment included those written by humans as well as those produced by peer systems. Since systems were only required to produce ATTRIBUTE-SETs, these were realised as strings using software purpose-built for this experiment.

The main measures in this experiment were (i) Identification Speed (the time it took human subjects to read the descriptions and identify the entity being described), and (ii) Identification Accuracy (the proportion of times that subjects identified the wrong entity).

2.1.2. TUNA-R

Task definition: This task, organised in 2008 and 2009, required participants to develop a method which, given an ATTRIBUTE-SET representing the semantic content of an identifying description for a target entity, outputs a WORD-STRING. Thus, the task focused on realisation in English of the semantic representation.

Input and output data: The training and development data consisted of the same TUNA corpus instances used for TUNA-AS. A new test set, consisting of 112 corpus instances and distinct from the test set used in the TUNA-AS and TUNA-REG tasks, was developed for this task using the same methodology as that used for the original TUNA corpus collection.

Evaluation software: The same software used for TUNA-AS was provided for this task (and extended). For WORD-STRINGs, the new version of the software compared human and peer outputs on the basis of (i) string edit (Levenshtein) distance; and (ii) *string accuracy*, that is, the proportion of peer output strings that were identical to human-produced descriptions.

Other metrics: In addition to edit distance and string accuracy, we also computed BLEU (Papineni et al., 2002) and NIST n-gram similarity scores comparing the peer and human outputs over the entire test sets.

2.1.3. TUNA-REG

Task definition: This task, also organised in 2008 and 2009, was a combination of TUNA-AS and TUNA-R, in that it required participants to develop a method which, given an input DOMAIN, outputs an identifying description for the target referent, that is, a WORD-STRING (thus, such systems would need to both select the ATTRIBUTEs and realise them, though peer systems were not required to do this in two separate steps).

Input and output data: The training and development data consisted of the same TUNA corpus instances used for TUNA-AS and TUNA-R. The test set was the same as the one newly created for the TUNA-AS'08 Task (as described above).

Evaluation software: The software program provided for this task was the same as for TUNA-R.

Other metrics: As in TUNA-R, we also computed BLEU and NIST scores to compare peer and human outputs. In 2008, as in TUNA-AS'07, we also ran a lab-based experiment with human participants, but this time Identification Speed was replaced by two separate measures, Reading Time and Identification Time, resulting in three measures: (i) Reading Time, i.e. the time it took to read the description; (ii) Identification Time, i.e. the time it took to identify the target referent given the description; (iii) the identification error rate.

In the 2009 edition of TUNA-REG, we ran a similar task-based experiment, but instead of reading descriptions, participants listened to a text-to-speech system reading out descriptions. The measures obtained were Identification Speed and Identification Accuracy (as in TUNA-AS). In 2009, we also ran an experiment in which system-generated and human-produced outputs were judged by linguists for their fluency and adequacy, as described in (Gatt and Belz, 2010).

2.2. GREC Resources

2.2.1. GREC-MSR

Input and output data: The GREC Corpus (version 2.0) consists of about 2,000 introductory sections in Wikipedia articles. In each text, three broad categories of Main Subject Reference (MSR) have been annotated (13,000 REs in total). The GREC-MSR shared task version of the corpus was randomly divided into 90% training data (of which

[4] The complete TUNA corpus also contains plurals, that is, references to two entities.

```
<?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.
...
</PARAGRAPH>
</TEXT>
```

Figure 3: Example text from the GREC-MSR Training Data.

10% were randomly selected as development data) and 10% test data.

Figure 3 is one of the texts in the GREC-MSR training/development data set. REFs indicate an instance of referring, REFEX is the selected RE and ALT-REFEX is a list of alternative REs for the referent. ALT-REFEX lists were generated for each text by an automatic method which collects all the (manually annotated) REs for the referent in the text and adds several defaults: pronouns and reflexive pronouns in all subdomains; and category nouns (e.g. *the river*), in all subdomains except people. Outputs generated by GREC-MSR systems are in the same format as the inputs, except that there are no ALT-REFEX lists and there is exactly one REFEX for each REF.

Task definition: The GREC-MSR Task is to develop a method for selecting one of the REFEXs in the ALT-REFEX list, for each REF in each TEXT in the test sets. The test data inputs are identical to the training/development data, except that REF elements contained only an ALT-REFEX list, not the preceding 'selected' REFEX. 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).

Evaluation software: For the GREC-MSR Shared Tasks we created an evaluation tool which computes the following metrics: (i) Accuracy of REFEX word strings, i.e. the proportion of REFEX word strings selected by a participating system that are identical to the one in the corpus; (ii) Accuracy of REG08-Type, i.e. the proportion of REFEXs selected by a participating system that have a REG08-TYPE value identical to the one in the corpus; (iii) String-edit distance; (iv) BLEU-3; and (v) NIST-5. In the case of the latter 3 string-comparison metrics, we assessed just the REs selected by peer systems (leaving out the surrounding text).

In the human evaluations, we assessed Fluency, Clarity and Coherence of REs within the textual context, as described in Belz et al. (2009).

2.2.2. GREC-NEG

Input and output 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, divided into training, development and test data.

We first manually annotated people mentions in the GREC-People texts by marking up the RE word strings and annotating them with coreference information, semantic category, syntactic category and function, and various supplements and dependents. Annotations included nested references, plurals and coordinated REs, certain unnamed references and indefinites. For full details see the GREC'10 documentation (Belz, 2010).

The manual annotations were then automatically checked and converted to XML format. The REF, REFEX and ALT-REFEX elements were the same as in the GREC-MSR annotations described above, except that here, all alternative REs are collected in a single list, appended at the end of the text, rather than to each reference. Also, here we allow arbitrary-depth embedding of references.

The training, development and test data for the GREC-NEG task is exactly as described above. The test data inputs are identical, except that REF elements in the test data do not contain a selected REFEX element.

Task definition: The GREC-NEG Task is to select one REFEX from the ALT-REFEX list for each REF in each TEXT in the test sets, including any embedded REFs. The aim is to select REs which make the text fluent, clear and coherent.

Evaluation software: We provide a software tool which computes the following metrics: (i) REG08-Type Precision is the proportion of REFEXs selected by a participating system which match the reference REFEXs; (ii) REG08-Type Recall is the proportion of target REFEXs for which a participating system has produced a match; (iii) String Accuracy is the proportion of word strings selected by a participating system that match those in the reference texts.

2.2.3. GREC-NER

Task definition: The GREC-NER task is a straightforward combined named-entity recognition and coreference resolution task, restricted to people entities. Systems insert REF and REFEX tags with coreference IDs around recognised mentions. The aim is to match the 'gold-standard' tags in the GREC-People data.

Input and output data: The GREC-NER training and development data come in two versions. The first is identical to the format described above (containing information about correct system outputs). The second is the test data input format, where texts do not have REFEXs, REFs, or ALT-REFEXs. Moreover, a proportion of REFEXs have been replaced with standardised named references.

System outputs have the same format as test data inputs, plus ALT-REFEX and REFEX tags inserted around recognised people references.

Evaluation software: To measure accuracy in the NER task, we provide a wrapper script which applies three commonly used performance measures for coreference resolution: MUC-6, CEAF, and B-CUBED.

2.2.4. GREC-Full

Task definition: The overall aim for GREC-Full Task is to improve the referential clarity and fluency of input texts. Systems should replace REs as and where necessary to produce as clear, fluent and coherent a text as possible. This task could be viewed as composed of three sub-tasks: (1) named entity recognition (as in GREC-NER); (2) a conversion tool to give lists of possible REs for each entity; and (3) named entity generation (as in GREC-NEG).

Input and output data: Inputs are as described for GREC-NER above, and outputs as for GREC-NEG above.

Evaluation software: We provide a tool which computes (i) BLEU-3; (ii) NIST; (iii) string-edit distance; and (iv) length-normalised string-edit distance. The human-assessed evaluation methods are preference-strength judgements using sliders for assessing Fluency and Referential Clarity.

2.3. Prodigy-METEO Resources

Input and output data The Prodigy-METEO inputs are data vectors of meteorological data about predicted wind characteristics; the outputs are corresponding 'wind statements' that form part of weather forecasts written by meteorologists for offshore oil platforms. The inputs and outputs were extracted from the SumTime-METEO corpus (Sripada et al., 2002). For example, the following is the target output for input 5Oct2000_03.num.1:

```
SSW 16-20 GRADUALLY BACKING SSE THEN FALLING VARIABLE
04-08 BY LATE EVENING
```

The wind data inputs are vectors of time stamps and wind characteristics (speed, direction, gusts etc.), e.g. the following is the input vector for output 5Oct2000_03.prn.1:

```
[[1,_SSW,16,20,-,-,0600],[2,_SSE,-,-,-,-,-1],[3,_VAR,04,
08,-,-,0000]]
```

In addition to the corpus-derived target outputs, the Prodigy-METEO data contains human-authored outputs for a subset of inputs, and outputs from 12 systems: a traditional deterministic rule-based generator, and 11 trainable generators (four probabilistic CFG-based systems, two probabilistic synchronous CFG systems and four systems based on phrase-based statistical machine translation).

Task definition: In order to be directly comparable with existing results using the data, systems must map from the inputs described above (which may be augmented by supplementary information not obtained by copying or converting other SumTime-METEO data) to wind statements. Trainable systems should either follow the 5-fold cross-validation regime facilitated by the data, or at least test on each of the five test data sets provided with the five folds, and average results. The aim for outputs is to be clear and fluent *as weather forecast text*, not as ordinary English text.

Evaluation software: Prodigy-METEO work has been evaluated by the BLEU metric (NIST scores are also sometimes reported), and by human intrinsic evaluation of Fluency, and Clarity using discrete rating scales and absolute quality judgements (rather than preference judgements). Parameter settings for BLEU and the experimental design for the human evaluations are provided in the repository.

2.4. Resources for Other Shared Tasks

The next set of resources that we plan to add to the GenChal Repository is from the SR'11 (English) Task in Surface Realisation. We also plan to add other resources from NLG Shared Tasks as and when the tasks are concluded.

3. Concluding Comments

In this paper we described the GenChal Repository of data and evaluation resources for Natural Language Generation which we have created in order to provide a free and easily accessible source of data to support research in NLG. Our core aims are (i) to lower the barrier to entry to the point where everything required for specific types of NLG work is provided in one place, so that the researcher can focus solely on building the generation system itself, and (ii) to facilitate work that can be directly compared to an existing body of work using an existing set of evaluation methods,

resulting in greater comparability of NLG research. The GenChal Repository should be particularly useful to students and other researchers new to NLG, but can also provide the more seasoned NLG researcher with the common ground necessary for comparing their work to related research. We plan to continue adding resources to the GenChal Repository and would welcome other researchers contributing their resources.

4. References

Anja Belz, Eric Kow, Jette Viethen, and Albert Gatt. 2009. The GREC main subject reference generation challenge 2009: Overview and evaluation results. In *Proceedings of the ACL-IJCNLP'09 Workshop on Language Generation and Summarisation (UCNLG+Sum)*, pages 79–87.

Anja 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. Dale and E. Reiter. 1995. Computational interpretation of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(8):233–263.

R. Dale. 1989. Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL'89)*.

A. Gatt and A. Belz. 2010. Introducing shared task evaluation to nlg: The TUNA shared task evaluation challenges. In E. Krahmer and M. Theune, editors, *Empirical Methods in Natural Language Generation*. Springer.

E. Krahmer and K. van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proc. ACL '02*, pages 311–318.

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

E. Reiter and R. Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, Cambridge, UK.

E. Reiter. 2007. An architecture for data-to-text systems. In *Proceedings of the 11th European Workshop on Natural Language Generation (ENLG'07)*, Schloss Dagstuhl, Germany.

S. Sripada, E. Reiter, J. Hunter, and J. Yu. 2002. SUMTIME-METEO: A parallel corpus of naturally occurring forecast texts and weather data. Technical Report AUCS/TR0201, Computing Science Department, University of Aberdeen.

K. van Deemter, A. Gatt, I. van der Sluis, and R. Power. in press. Generation of referring expressions: Assessing the incremental algorithm. *Cognitive Science*.