ARL-TR-9324 • SEP 2021

DEVCOM ARMY RESEARCH LABORATORY

Extending Generation and Evaluation and Metrics (GEM) to Grounded Natural Language Generation (NLG) Systems and Evaluating their Descriptive Texts Derived from Image Sequences

by Stephanie M Lukin and Clare R Voss

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-9324 ● SEP 2021

Extending Generation and Evaluation and Metrics (GEM) to Grounded Natural Language Generation (NLG) Systems and Evaluating their Descriptive Texts Derived from Image Sequences

by Stephanie M Lukin and Clare R Voss
Computational and Information Sciences Directorate, DEVCOM Army Research Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
September 2021	Technical Report	April 1–30, 2021

4. TITLE AND SUBTITLE
Extending Generation and Evaluation and Metrics (GEM) to Grounded Natural Language Generation (NLG) Systems and Evaluating their Descriptive Texts Derived from Image Sequences

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Stephanie M Lukin and Clare R Voss

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
DEVCOM Army Research Laboratory
ATTN: FCDD-RLC-IT
2800 Powder Mill Rd, Adelphi, MD 20783

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-9324

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Contact author email: <stephanie.m.lukin.civ@army.mil>

14. ABSTRACT
We present here, for consideration in a future Generation and Evaluation and Metrics (GEM) challenge, a graduated, task-based approach to evaluating *grounded natural language generation (NLG) systems* that generate descriptive texts derived from sequences of input images. We start by characterizing *grounded NLG tasks* that generate descriptive texts at increasing levels of complexity, then step through examples of these levels with image sequences and facet targets (input) and their derivative descriptive texts (output) from our human-authored data set. For evaluating whether a grounded NLG system is "good enough" for users' needs, we first ask if the user can recover the images the system used to derive descriptive texts at the relevant, graduated level of complexity. The texts judged as adequate in this *image-selection task* are then analyzed for their semantic facet units (SFUs), which form the basis for scoring descriptive texts generated by other grounded NLG systems. The image-selection and SFU scoring together constitute the evaluation we are piloting for grounded, data-to-text NLG systems.

15. SUBJECT TERMS
natural language generation, evaluation and metrics, images and text, narrative facets

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Stephanie M Lukin
Unclassified	Unclassified	Unclassified	UU	17	19b. TELEPHONE NUMBER (Include area code) 310-448-5396

Contents

List of Figures

List of Tables

1. Introduction

Evaluating the output of natural language generation (NLG) systems has long been a topic of study*, yet a gold standard is often elusive due to the subjective and varying nature of the task at hand and the humans who interact with the system. The GEM benchmark (natural language Generation, its Evaluation, and Metrics) presents a "living benchmark", focusing on feasible tasks with pragmatic evaluation metrics incorporating many topics in the research community, including summarization, dialogue, structure-to-text, and simplification.[2] The inputs in the GEM data sets are primarily text or structured text, although they may be a derivative of another form, for example, a concept from CommonGEN that is grounded and situated.[3]

This report proposes extending GEM to include *grounded NLG systems*, systems that generate descriptive texts derived from another type of input, images. The primary motivation for designing such a benchmark challenge is our need for an NLG engine running on board a robot so that it can provide, to one or more human operators at another location, a detailed description of what only it can see as it explores an environment. The robot's generated text should be descriptive enough for the operators to perform particular tasks or make informed decisions, such as in a disaster, determining whether conditions in a street may not be safe with power lines and trees down.

For grounded NLG systems, we describe the input and their generated texts by **levels** that are defined in terms of the following:

- An image source: a single image, a sequence of two images, or a sequence of three images.
- A facet: identifying entities, describing a scene, or crafting a narrative.

Section 2 introduces the data set of human-authored derivative descriptions we are collecting, which consists of different combinations of image sources and target facets.

*Refer to Howcroft et al.,[1] for a 20-year review in NLG evaluation.

Given that specific user-robot tasks can vary by scenario, we propose an extrinsic evaluation, a more general task to assess the adequacy and utility of the derivative descriptions, by asking if the description is "good enough" for a user to select the image source(s) from which the text was derived (Section 3).

Once this assessment in determining if a user can successfully perform an *image-selection task* from derivative descriptions is complete, we next propose to analyze each description for the facets they convey. We propose a variant of the Pyramid method for summary evaluation[4] applied to our derivative descriptions for Semantic Facet Units (SFUs). This would spell out guidelines for human annotators providing ground truth in assessing a grounded NLG system. It additionally supplies automated metrics that can be used to score the adequacy of the system output (Section 3).

We expect there will be other situations of interest besides our robot scenario to researchers outside the NLG community, in particular, the multimedia and video understanding communities who would be interested in NLG for their input data sets. As an example, NIST has been running TRECVID with a Video-to-Text (VTT) description task in which they evaluate systems that generate natural language sentence descriptions of short video inputs (between 3 and 10 s long).[5] We conclude in Section 4 with a discussion of future work in these communities.

2. Input Run Levels and their Derivative Descriptions

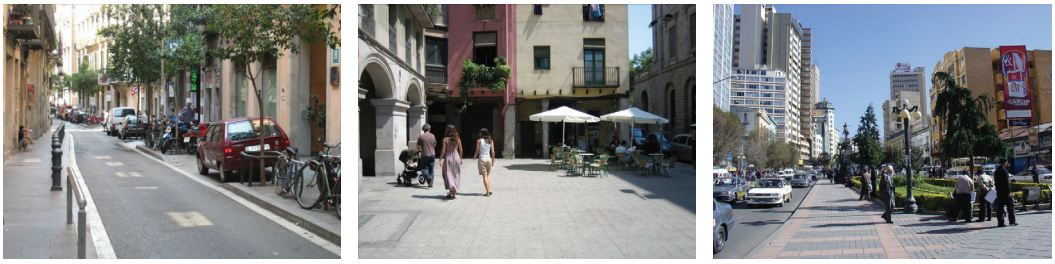
We define various **input run levels***, where a level is characterized by 1) an input run type, and 2) a facet. Table 1 shows the combination of input run types and facets we explore in this report to create four distinct input run levels. Each level builds off the previous level and its facets.

Table 1. Definition of input run levels

Input run level	Input run type	Facet
Level-I	one image	Entities
Level-II	one image	Scene
Level-III	two images	Narrative
Level-IV	three images	Narrative

*The phrase "input run level" inspired by Awad et al.,[5] use of the phrase "run type" to refer to the cross product of a training data type and a training feature type.

The *input run type* dictates the input visual medium. It ranges from a video clip, to a sequence of still-frame images extracted from a video or selected from a photo album, down to a single image from either of these sources. In this report, we explore three-image, two-image, and one-image input run types that are temporally ordered, where applicable. Figure 1 shows three images collected from the Visual Genome data set.[6] Taken together, these images ($image_1$ – $image_3$) constitute a three-image sequence where the images are arranged in the temporal order in which they were stored in the Visual Genome data set*. A two-image sequence is created by isolating two of the three images in temporal order.



(a) image$_1$ **(b) image$_2$** **(c) image$_3$**

Fig. 1. A three-image sequence

The *facets* we explore are a) entities, b) scenes, and c) narratives. These facets are refined from existing writing tasks, such as those proposed by Huang et al.,[7] and Lukin et al.[8] The *entity facet* isolates and enumerates visible entities in each input run type, and may include particular attributes or relations of the entities, including spatial (e.g., co-location, orientation), observational (e.g., color, size), and confidence (e.g., if the entity is unexpected, or if something about it is unclear).

The *scene facet* is a snapshot of events, located in a specific point in time and space. It supplies answers to the questions "what is happening", "where is this", and "when is this." This facet is akin to a literal explanation of what appears in the input run type, and is dependent upon the entity facet for helping to identity the "who" or "what" is in the input run type.

Finally, the *narrative facet* weaves together a span of events within the input run type, evoking a temporal arrangement of scenes and supplying subjective evaluations and orientation.[9] This facet involves creative storytelling, and explores a tem-

*Images in the Visual Genome are organized by a unique ID, and photo albums were scraped in sequential order

poral leap into future possibilities beyond the visually observable in the entity or scene facets observable in the input run type.

We elicit human annotators to curate a data set of these texts with instructions about how each facet is defined. These texts, called *derivative descriptions*, are derived from the facets of an input run type. Table 2 shows four derivative descriptions for each level, referencing image_1, image_2, and image_3. In level-I, we see that the annotator has listed the key visible entities in image_1, including the street and cars, as well as important aspects, such as the "narrow, one way" nature of the street, and the European look of the cars. In Level-II, the annotator has isolated the snapshot of image_1 by calling it a "busy spot", and again referencing the European style.

Table 2. Human-authored derivative descriptions for each level, by annotator B. Punctuation, capitalization, and misspellings are exactly as the annotator typed.

Level-I image_1; entity facet	street: narrow, one way; cars: look European by license plates; bikes: parked outside of shops; motorbikes: parked outside of shops; people: talking and hanging out; signs: outside of businesses
Level-II image_1; scene facet	This looks like a busy spot for people to shop. It is down a one way street and is quite cramped. This is very characteristic of European cities.
Level-III image_1–image_2; narrative facet	I have always wanted to go to Europe. The lovely architecture, food, and culture are all so picturesque. I went to Germany on my honeymoon. We stayed in a small city in an Air BnB that overlooked a courtyard and a small cafe at the end of a street. We rented motor bikes for the week since they seem much more convenient to get around on in these small streets.
Level-IV image_1–image_3; narrative facet	One day we go into the modern areas of the town. It reminds me of home, because of the constant traffic. A lot of people in suits are walking around with brief cases. We park nearby and decide to walk and find land marks. There is a huge statue in the middle of it all. It looks like a man on a horse. This place is very pedestrian friendly with plenty of places to stop and sit. There are street lights around various gardens. I mention to my family that we should come back at night to see everything lit up.

In level-III, we see the annotator crafting a narrative that centers around setting the stage for the narrative based on the imagery in image_1 and image_2. They include a few events, such as renting motor bikes, and conduct creative extrapolation, such as getting the bikes because they would be more convenient than walking or driving

on the small streets. The level-IV derivative description extends level-III by adding image_3 into consideration. Here, we see a hypothetical scenario unfold given the level-III derivative description as the background.

To date, we have curated 300 three-image input run types. We have collected 730 level-III and level-IV derivative descriptions from annotators on Mechanical Turk (146 three-image input run types annotated by 5 unique annotators), and 2,190 level-I and level-II derivative descriptions (438 individual images across the 146 three-image sequences, annotated by 5 unique annotators). Data collection is ongoing, and the input run types and their derivative descriptions will be released in a future publication.

3. Proposed Evaluation Methodology

We propose two connected evaluations that measure the adequacy of the derivative descriptions and analyze the faceted components of the derivative description. Both evaluations can be conducted on the human-authored texts in our data set, as well as the NLG output of an automated system.

3.1 Evaluating the Utility of Descriptive Texts

Given our practical need for a grounded NLG system (a robot that reports back what it sees), we pursue an extrinsic evaluation to assess whether users can find what they need as a result of reading system output. Empirically, our approach to assessing the adequacy of derived descriptions is to ask whether participants in a study, given a collection of several images, can identify the image or images from which that text was derived. We propose to vary the images and derivative descriptions presented to participants along the four levels defined previously. Although the texts and their facets and characteristics will vary, this image selection task itself is only concerned with success, and can thus be measured for precision, recall, and f-measure. For example, if a user's goal is to determine if an area has a paved road, they would pay particular attention to a level-I description with entities relevant to roads in order to make the correct image determination.

Determining the "best" level of derivative description is a non-trivial task. If a user's goal is to assess if a location is crowded, then a level-II description providing a snapshot of the scene in a single image might be adequate. However, if the user goal is to determine whether the classroom in a school building is readily accessible to

navigating with a wheelchair, then the entity enumeration that level-I provides is inadequate; the user would instead want a level-III or level-IV derived description of a sequence of images that could describe a visibly clear path from outside to the inside of the building. This utility-driven evaluation is independent of task, and can be used to inform rules for NLG engines based on different user goals.

3.2 Evaluating the Facets of Descriptive Texts

If a descriptive text achieves high performance in the image-selection evaluation, the text's component semantics can be extracted. We propose a variation of the pyramid method for summarization,[4] which uses Summarization Content Units (SCUs), annotations of clauses within a corpus of summaries. Our evaluation proposes to annotate derivative descriptions for SFUs. This evaluation does not rely on string matching, and is measured by a majority of human annotators.

Consider the derivative descriptions written by annotator-B in Table 2. For the *entity facet*, the SFUs isolate salient entities and their attributes, for example, the "narrow, one way" aspect of the street and the European license plates on the cars. Extracting SFUs from other annotators' responses for this particular input run type shows that the vehicles are also highlighted as being important*:

A1: Vehicles-red, silver car and white van
B2: Cars: Look European by license plates
C2: parked cars
D2: Multiple parked cars
E2: Vehicles

Attributes can additionally be categorized into SFUs and counted for overlap.

The SFUs for the *scene facet* isolate the function of the space with respect to the activities and events that take place, for example, annotator-B's statement "this looks like a busy place for people to shop." The other annotators extracted similar semantic units revolving around shopping:

A2: People come here either to live or shop.
B1: This looks like a busy spot for people to shop.

*The letter indicates the annotator, and the number indicates the position of the entity in their response. Punctuation, capitalization, and misspellings are exactly as the annotators typed.

C2: People go inside and shop the adjacent buildings.
D2: There are storefronts to an assortment of locally owned shops.
E3: The ground level on the buildings seems to be comprised mostly of businesses.

The SFUs for the *narrative facet* turn to the highest level of abstraction. We propose to follow narrative clauses for orientation, action, and evaluation to characterize the component parts,[9] as well as plot units to characterize the narrative arc and compare each annotator's construction.[10]

This semantic-driven evaluation is also independent of task, and can be used to inform rules for NLG engines based on different user goals.

4. Conclusion and Future Work

This report proposes a benchmark challenge for a future GEM 2.0, in which a faceted, textual description is derived from an image or a sequence of images. We propose two methods to evaluate the texts, an image-selection task that measures the text's utility, and a semantic facet analysis that serves to illuminate the critical aspects of the derived description.

Future work can explore expanding the input run type to include video clips, similar to the TRECVID challenge, as well as perform an automated down-selection from videos into the one-, two-, and three-image sequences, rather than assembling them from preconstructed photo albums.

There are many open questions to consider as we move forward in developing an NLG engine that can emulate human-authored descriptive texts for this benchmark challenge. In which scenarios are level-I descriptions enough for the user to perform the task, and when are the higher level descriptions a requirement? Computationally, how can a computer vision algorithm detect concepts like "cramped" and what does it take to produce derivative descriptions with temporal connectivity? As we answer these questions, we are guided by our evaluations, which measure the adequacy and utility of the human-authored texts from our ongoing data collection and construct a benchmark that can serve as a human topline in development.

5. References

1. Howcroft DM, Belz A, Clinciu MA, Gkatzia D, Hasan SA, Mahamood S, Mille S, van Miltenburg E, Santhanam S, Rieser V. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In: Proceedings of the 13th International Conference on Natural Language Generation; p. 169–182.

2. Gehrmann S et al. The GEM Benchmark: Natural language generation, its evaluation and metrics. 2021. arXiv preprint arXiv:2102.01672.

3. Lin BY, Zhou W, Shen M, Zhou P, Bhagavatula C, Choi Y, Ren X. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In: Findings of the Association for Computational Linguistics: EMNLP 2020; Online: Association for Computational Linguistics; 2020. p. 1823–1840.

4. Nenkova A, Passonneau RJ. Evaluating content selection in summarization: The pyramid method. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004; p. 145–152.

5. Awad G, Butt AA, Curtis K, Lee Y, Fiscus J, Godil A, Delgado A, Zhang J, Godard E, Diduch L, Smeaton AF, Graham Y, Kraaij W, Quenot G. TRECVID 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval. 2020.

6. Krishna R et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision. 2017;123(1):32–73.

7. Huang TH et al. Visual storytelling. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; p. 1233–1239.

8. Lukin S, Hobbs R, Voss C. A pipeline for creative visual storytelling. In: Proceedings of the First Workshop on Storytelling; p. 20–32.

9. Labov W, Waletzky J. Narrative analysis: Oral versions of personal experience. 1997;7:3–38.

10. Lehnert WG. Plot units and narrative summarization. Cognitive science. 1981;5(4):293–331.

List of Symbols, Abbreviations, and Acronyms

GEM – Generation, its Evaluation, and Metrics

NIST – National Institute of Standards and Technology

NLG – Natural Language Generation

SCU – Summarization Content Unit

SFU – Semantic Facet Unit

TRECVID – Text Retrieval Conference Video Retrieval Evaluation

VTT – Video-to-Text

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	DEVCOM ARL FCDD RLD DCI TECH LIB
2 (PDF)	DEVCOM ARL FCDD RLD IT C R VOSS S M LUKIN