Research Paper Summary on "Evaluation of Text Generation: A Survey"

This paper discusses natural language generation (NLG) evaluation methods, with an emphasis on neural NLG systems. The author begins by describing recent NLG models, which are generated by training deep neural network DNN models on a broad corpus of human-written texts. The Transformer architecture, which includes an encoder and a decoder that are both implemented using the self-attention mechanism, is being used by cutting-edge NLG systems. This research study divides NLG evaluation into three categories: (1) human-centric evaluation metrics, (2) automated metrics that do not require preparation, and (3) machine-learned metrics.

(1) Human-Centric Evaluation: Human evaluations are commonly regarded as the most important form of NLG system evaluation. Although human assessments provide the most information about how well a model performs in a role, they also present a number of challenges. Crowd-sourcing platforms like Amazon Mechanical Turk have allowed researchers to conduct larger-scale experiments at a lower cost. Other methods of using human subjects in the assessment process include training models on human decisions. In intrinsic evaluations, people are asked to evaluate the quality of generated text. Whereas, in extrinsic evaluation, people evaluate a system's performance on the task for which it was designed. Extrinsic are the most important since they demonstrate how a system performs in a downstream role. Intrinsic evaluations are more common than extrinsic evaluations because they are more expensive and harder to implement. Online evaluations of generated texts have grown in popularity. These platforms allow researchers to conduct large-scale evaluations in a timely manner. They also allow researchers to access a broader range of evaluators than they might otherwise hire in-person.

(2) Untrained Automatic Evaluation Metrics: Untrained automated NLG metrics are used to assess the efficacy of models that generate text, such as machine translation, image captioning, or query generation. These metrics create a score that indicates the degree of similarity (or dissimilarity) between an automatically generated text and a human-written comparison (gold standard) text. We categorize the untrained automated evaluation methods as follows:

1. **n-gram Overlap Metrics for Content Selection**: n-gram overlap metrics, which are commonly used for evaluating NLG systems, are used to assess the degree of "matching" between machine-generated and human-authored (ground-truth) texts.
2. **Distance-Based Evaluation Metrics for Content Selection**: In NLG applications, a distance-based metric employs a distance function to assess the similarity between two text units. The closer the gap between the two texts, the more similar they are. Embedding-based similarity measures are commonly used to measure similarity. Embeddings are vector representations of character or lexical units that are real-valued. They allow tokens with similar meanings to be represented in the same way.
3. **N-gram Based Diversity Metrics:** The lexical diversity score assesses the range and variety of words used in writing. N-gram based diversity reviews some of the metrics designed to measure the quality of the generated text in terms of Lexical diversity.
4. **Explicit Semantic Content Match Metrics**: Metrics for semantic content matching work at both the semantic and conceptual stages. They have been shown to be highly correlated with human decisions. The confidence scores obtained from semantic similarity methods have been used as an evaluation criterion in other text generation work in semantic similarity-based models metrics. Such models can evaluate a guide and a hypothesis document based on their task-level semantics.
5. **Syntactic Similarity-Based Metrics:** The similarity between a reference and a hypothesis text is captured by a syntactic similarity metric. POS tags are widely used in machine translation to assess translation accuracy. The study of the arrangement of words and phrases in well-formed sentences is known as syntactic analysis.

(3) Machine-Learned Evaluation Metrics: Untrained evaluation metrics assume that the generated text has significant word overlap with the ground-truth text. This assumption does not hold for many NLG tasks,

such as a social chatbot. In these cases, we can build machine-learned models (trained on human judgment data) to mimic human judges to measure quality metrics of output. Some of the machine learned evaluation model has been mentioned below:

1. **Sentence Semantic Similarity Based Evaluation:** Neural approaches to sentence representation look to capture semantic and syntactic meanings from various viewpoints Using DNN models, neural approaches attempt to map a sentence onto an embedding vector. The study suggests using SENTBERT, which is a fine-tuned BERT on a "common" mission.
2. **Evaluating Factual Correctness:** Zhang et al. suggest a way to address the issue of factual correctness in models. They introduce a reward-based generator optimization to ensure generators receive more rewards. Their evaluation model is based on the hypothesis that the consistency of a summary is correlated with the number of questions that can be answered by reading the summary.
3. **Regression-Based Evaluation:** Shimanaka et al. (2018) develop RUSE, a segment-level machine translation assessment metric. They approach the evaluation task as a regression problem with the goal of predicting a scalar value that indicates the quality of translating a machine-translated hypothesis (t) to a reference translation (r).
4. **Evaluation Models with Human Judgments:** Current evaluation methods are only useful to some degree as they have no knowledge about the dataset that the model is trained on. Hashimoto et al. (2019) propose a new evaluation metric named Human Unified with Statistical Evaluation (HUSE) which focuses on more creative and open-ended text generation tasks. On summarization and chit chat dialog tasks, HUSE was found to be successful.
5. **BERT-Based Evaluation:** Given the strong performance of BERT across many tasks, there has been work that uses BERT or similar pre-trained language models for evaluating NLG tasks. One of the BERT-based models for semantic evaluation is BERTSCORE (Zhang et al., 2020). Another model is fine-tuned on the Multi-Genre Natural Language Inference Corpus. Both model-based evaluators have been shown to be more robust and correlate better with human evaluation than automatic evaluation metrics such as BLEU and ROUGE.
6. **Composite Metric Scores:** The quality of many NLG models like machine translation and image captioning can be evaluated for multiple aspects, such as adequacy, fluency, and diversity. Sharif et al. (2018) presents a machine-learned composite metric for evaluating image captions.

Summary of Case Studies: As case studies, the researchers used two NLG tasks: automated document summarization and long-text generation. A text summarization method in Case Study #1 aims to extract valuable information from a reference document and produce a brief summary. There are various types of summarization methods, which can be classified according to their purposes. These methods may also be classified as extractive, abstractive, or query focused. In case study #2, a multi-sentence text generation method aims to create a single paragraph or a multi-paragraph document. This research area presents a particular challenge to state-of-the-art approaches that are based on statistical neural models. New criteria need to be implemented to measure the quality of long generated text.

Conclusions: This research study mentions recent advances in neural language models which have made significant progress in developing new NLG models and systems for evaluation. Research further indicates that the future NLG evaluation research should focus on developing easy-to-use, explainable evaluation tools. Clear reporting of human evaluations is very important, especially for replicability purposes. Research study also mentions that there is still a lack of systematic methods for evaluating how an NLG system can avoid generating improper language.

Questions: 1) Is there a percentage threshold on the results obtained by people in NLG evaluations, for example, if people evaluation is less than 25%, the text generation model is discarded or needs more improvement? 2)I'm curious about the differences in accuracy between human-centric and machine-learned evaluation models.