Sentence-level Natural Language Generation Evaluation with Semantic Parsing

Motivation. Generalized neural sequence-to-sequence models [seq2seq; 7] provide new data-driven approaches in tasks which require capability in natural language generation (NLG), such as machine translation [MT; 8], dialogue, and summarization. Seq2seq models consist of an encoder and decoder, and the decoder of seq2seq models is usually parameterized by gated variants of recurrent neural networks [LSTM; 5]. LSTMs have the flexibility to output an arbitrary sequence of tokens but are not explicitly designed to generate with respect to syntax and semantics. To determine the quality of NLG systems, evaluation has become a central methodological concern. In MT, intrinsic evaluations, where humans rate machine produced translations, are the main indicator of progress in the field. However, human based evaluations can be time consuming and costly, so an evaluation based on automatic metrics computed from reference corpora can be useful during development cycles.

Automatic metrics. The goodness of an automatic metric is determined by the metric's correlation with human judgment in evaluating output language. At least two of the challenges in automatic metrics are outlined here. First, the metric must measure many dimensions of the output that are relevant. Second, in evaluations of many NLG tasks, several outputs may be acceptable. In translation, we may be concerned with a translation's fluency (how natural does it sound?) and adequacy (is the meaning correct?), and several outputs may be valid translations. The most common automatic metric used in MT is BLEU [6]. The BLEU scores of an output translation is calculated based on the n-gram overlap of the output to the reference translation. The authors claims that overlap of lower order n-grams capture adequacy, while higher order n-grams capture fluency. While using n-grams alleviate the problem of evaluating against multiple reference translations, they are shallow linguistic features. As a result, it is a coarse indicator of quality of individual translations, and only a strong indicator of system-level quality when aggregating over a test corpus. Much recent work in NLG adopts metrics from the MT community for evaluation.

Parsing. Parsing identifies the underlying linguistic representation for a given sentence, from a formalism of interest. In syntactic parsing, the linguistic constructions and sentence structure is derived; in semantic parsing, formal representations of the sentence meaning are derived. The semantic representations often abstract away from specific syntactic constructions and preserve sentence meaning (e.g. "She pursues him" will have a semantic representation that is similar or equivalent to "He was pursued by her"). In Abstract Meaning Representation [AMR; 1], a sentence's meaning is represented as a graph with nodes as entities or concepts, and edges as relations. By parsing natural language into these semantic representations, performing certain computations on these representation data structure can solve different natural language understanding (NLU) task. Reframing the evaluation of NLG systems as an NLU task, scoring an output to a reference text can be viewed as comparing representations, if our parsers are capable. Since parsing has been crucial to natural language processing, the community has developed robust parsers and linguistic formalisms.

Intellectual merit. My graduate research will explore the use of deep linguistic features from semantic parsers to evaluate NLG. By leveraging the recent advancements in semantic parsing, a metric that achieves high sentence-level correlation with human judgment of NLG output may be possible. Metrics in evaluations of language generation have previously used features from shallow semantic annotations and dependency parses , but the use of general

Source text	Reference text	MT-system output
团队赛始终是最棒的。	The team event is always the best.	The team is always the best.

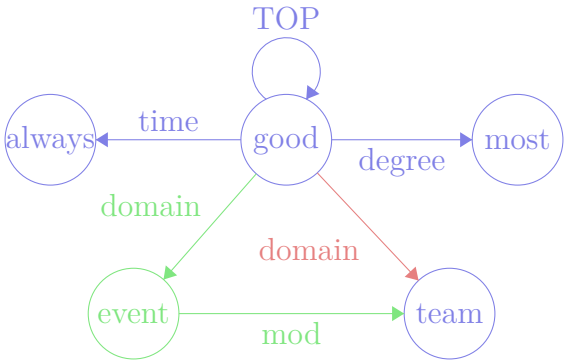

Figure 1: An example from the dataset collected in the WMT'17 metrics shared task. The human rating score assigned to the MT translation is -0.33 (low). Both the AMR representation of the reference and MT translations are aligned and overlaid. AMR structures are simplified. Blue represents overlapping graph structures, green for structures unique to the reference representation, and red for those unique in the MT-system representation.

purpose, graph based semantics such as AMR is underexplored. Refer to Figure 1. My aim is to research a metric that accounts for how the differences in the semantic graphs account for the difference in human ratings. Besides the advantage of using deeper linguistic features, a metric with this approach can also provide error diagnostics by summarizing the deviations of the semantic graphs from MT to reference translations.

Resources. My research will mainly be developed under the setting of the metrics track of the Conference of Machine Translation [WMT'17; 2]. This annual shared task collects human ratings for MT translations (as in Figure 1). In other NLG domains, several similar datasets such as ratings for image captioning are also available. AMR has an active research community, and recent work such as Flanigan et al. [4] provide robust parsers to build upon. To align our semantic graph representations (as aligned in Figure 1), we will use Smatch, which is predominantly used to calculate the accuracy of semantic parsers. [3]

Broader impact. Automatic metrics have the potential to advance the pace of development for an entire field. In the speech recognition community, the word error rate (WER) facilitated rapid development and comparisons of new systems, as did BLEU for MT. Successful research on NLG metrics will greatly accelerate development of NLG systems and ensure the reliability of results from automatic evaluations. Goals in NLG are broad and far reaching, with application to many fields beneficial to society. Encompassing tasks such as dialogue and summarization, these systems readily apply to educational chatbots and healthcare literature, respectively.

[1] Banarescu et al. "Abstract Meaning Representations for Sembanking" (2013). [2] Bojar et al. "Findings of the 2017 Conference on Machine Translation (WMT17)" (2017). [3] Cai et al. "Smatch: an Evaluation Metric for Semantic Feature Structures" (2013). [4] Flanigan et al. "A discriminative graph-based parser for the abstract meaning representation" (2014). [5] Hochreiter et al. "Long Short-Term Memory" (1997). [6] Papineni et al. "Bleu: a Method for Automatic Evaluation of Machine Translation" (2002). [7] Sustkever et al. "Sequence to Sequence Learning with Neural Networks" (2014). [8] Wu et al. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation" (2016).