Evaluating Evaluation Methods for Generation in the Presence of Variation

Amanda Stent, Matthew Marge, and Mohit Singhai

Stony Brook University, Stony Brook, NY 11794, USA,
stent@cs.sunysb.edu, mmarge@ic.sunysb.edu, mohit@cs.sunysb.edu

Abstract. Recent years have seen increasing interest in automatic metrics for the evaluation of generation systems. When a system can generate syntactic variation, automatic evaluation becomes more difficult. In this paper, we compare the performance of several automatic evaluation metrics using a corpus of automatically generated paraphrases. We show that these evaluation metrics can at least partially measure adequacy (similarity in meaning), but are not good measures of fluency (syntactic correctness). We make several proposals for improving the evaluation of generation systems that produce variation.

1 Introduction

The task of surface realization is to select, inflect and order words to communicate the input meaning as completely, clearly and fluently as possible in context. Traditional grammar-based surface realizers, (*e.g.* [1]) focus on the production of at least one high quality output sentence for each input semantic form. By contrast, two-stage surface realizers (e.g. [2, 3]) produce many possible sentences for each input semantic form, but select only one for output. Comparatively little research has been performed on rule-based approaches to the generation of variation (but see [4, 5]). However, recently there has been increasing interest on corpus-based approaches to the generation of paraphrases, or text-to-text generation (e.g. [6–10]).

Variation in surface realization takes two basic forms: *word choice variation*, and *word order variation*. Example 1 shows both types of variation. Word order variation may entail word choice variation, as in example (1b).

Example 1
(a) I bought tickets for the show on Tuesday.
(b) It was the show on Tuesday for which I bought tickets.
(c) I got tickets for the show on Tuesday.
(d) I bought tickets for the Tuesday show.
(e) On Tuesday I bought tickets for the show.
(f) For the show on Tuesday tickets I bought.

Variation is widely used by humans both in text and dialog. However, not all variations are meaning-preserving. A variation may add meaning possibilities that were not there before, remove meaning possibilities (compare example (1a) with (1d) and