Human Evaluation of a German Surface Realisation Ranker

Aoife Cahill\textsuperscript{1} and Martin Forst\textsuperscript{2}

\textsuperscript{1} Institut für Maschinelle Sprachverarbeitung (IMS)
University of Stuttgart
70174 Stuttgart, Germany
aoife.cahill@ims.uni-stuttgart.de

\textsuperscript{2} Powerset/Microsoft
475 Brannan Street
San Francisco, CA 94304, USA
martin.forst@microsoft.com

Abstract. In this chapter we present a human-based evaluation of surface realisation alternatives. We examine the relative rankings of naturally occurring corpus sentences and automatically generated strings chosen by statistical models (language model, log-linear model), as well as the naturalness of the strings chosen by the log-linear model. We also investigate to what extent preceding context has an effect on choice. We show that native speakers do accept quite some variation in word order, but that there are clearly also factors that make certain realisation alternatives more natural than others. We then examine correlations between native speaker judgements of automatically generated German text and automatic evaluation metrics. We look at a number of metrics from the MT and Summarisation communities and find that for a relative ranking task, most automatic metrics perform equally well and have fairly strong correlations to the human judgements. In contrast, on a naturalness judgement task, the correlation between the human judgements and the automatic metrics was quite weak, the General Text Matcher (GTM) tool providing the only metric that correlates with the human judgements at a statistically significant level.

Keywords: generation evaluation, surface realisation, human evaluation, German, human judgements, automatic metrics, correlation.

1 Introduction

An important component of research on surface realisation (the task of generating strings for a given abstract representation) is evaluation, especially if we want to be able to compare across systems. There is consensus that exact match with respect to an actually observed corpus sentence is too strict a metric and that BLEU score measured against corpus sentences can only give a rough impression of the quality of the system output. It is unclear, however, what kind of metric would be most suitable for the evaluation of string realisations, so that,
as a result, there have been a range of automatic metrics applied, including *inter alia* exact match, string edit distance, NIST SSA, BLEU, NIST, ROUGE, generation string accuracy, generation tree accuracy, word accuracy [13,16,26].

It is not always clear how appropriate these metrics are, especially at the level of individual sentences. Automatic evaluation metrics are an indispensable tool for rapid experiment turnover, but ideally, a metric for the evaluation of realisation rankers would rank alternative realisations in the same way as native speakers of the language for which the surface realisation system is developed, and not only globally, but also for individual sentences.

Another major consideration in evaluation is what to take as the gold standard. The easiest option is to take the original corpus string that was used to produce the abstract representation from which we generate. However, there may well be other realisations of the same input that are as suitable in the given context. Reiter and Sripada [21] argue that while we should take advantage of large corpora in NLG, we also need to take care not to introduce errors by learning from incorrect data present in corpora.

In order to better understand what makes good evaluation data and metrics, we designed and implemented an experiment in which human judges evaluated German string realisations. The main aims of this experiment were: (i) to establish how much variation in German word order is acceptable for human judges, (ii) to find an automatic evaluation metric that mirrors the findings of the human evaluation, (iii) to provide detailed feedback for the designers of the surface realisation ranking model and (iv) to establish what effect preceding context has on the choice of realisation. In this chapter, we concentrate on points (i), (ii) and (iv).

The remainder of the chapter is structured as follows: In Section 2 we outline the realisation ranking system that provided the data for the experiment. In Section 3 we outline the design of the experiment, and in Section 4 we present our findings. Section 5 outlines the correlation between the human judgements and automatic evaluation metrics. In Section 6 we relate this to other work, and finally we conclude in Section 7.

### 2 A Realisation Ranking System for German

We take the realisation ranking system for German described in Cahill et al. [6] and present the output to human judges. One goal of this series of experiments is to examine whether the results based on automatic evaluation metrics published in that paper are confirmed in an evaluation by humans. Another goal is to collect data that will allow us and other researchers [4] to explore more fine-grained and reliable automatic evaluation metrics for realisation ranking.

The system presented by Cahill et al. [6] ranks the strings generated by a hand-crafted broad-coverage Lexical Functional Grammar [5] for German [22] on the basis of a given input f-structure. Their system was trained and evaluated on data derived from the TIGER Corpus [4], a treebank comprising 50,472 sentences

1 The data is available for download from [http://www.ims.uni-stuttgart.de/projekte/pargram/geneval/data/](http://www.ims.uni-stuttgart.de/projekte/pargram/geneval/data/)