Correlating Human and Automatic Evaluation of a German Surface Realiser

Aoife Cahill
Institut für Maschinelle Sprachverarbeitung (IMS)
University of Stuttgart
70174 Stuttgart, Germany
aoife.cahill@ims.uni-stuttgart.de

Abstract

We examine correlations between native speaker judgements on automatically generated German text against 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 General Text Matcher (GTM) tool correlates best overall, although in general, correlation between the human judgements and the automatic metrics was quite weak.

1 Introduction

During the development of a surface realisation system, it is important to be able to quickly and automatically evaluate its performance. The evaluation of a string realisation system usually involves string comparisons between the output of the system and some gold standard set of strings. Typically automatic metrics from the fields of Machine Translation (e.g. BLEU) or Summarisation (e.g. ROUGE) are used, but it is not clear how successful or even appropriate these are. Belz and Reiter (2006) and Reiter and Belz (2009) describe comparison experiments between the automatic evaluation of system output and human (expert and non-expert) evaluation of the same data (English weather forecasts). Their findings show that the NIST metric correlates best with the human judgements, and all automatic metrics favour systems that generate based on frequency. They conclude that automatic evaluations should be accompanied by human evaluations where possible. Stent et al. (2005) investigate a number of automatic evaluation methods for generation in terms of adequacy and fluency on automatically generated English paraphrases. They find that the automatic metrics are reasonably good at measuring adequacy, but not good measures of fluency, i.e. syntactic correctness.

In this paper, we carry out experiments to correlate automatic evaluation of the output of a surface realisation ranking system for German against human judgements. We particularly look at correlations at the individual sentence level.

2 Human Evaluation Experiments

The data used in our experiments is the output of the Cahill et al. (2007) German realisation ranking system. That system is couched within the Lexical Functional Grammar (LFG) grammatical framework. LFG has two levels of representation, C(onstituent)-Structure which is a context-free tree representation and F(unctional)-Structure which is a recursive attribute-value matrix capturing basic predicate-argument-adjunct relations.

Cahill et al. (2007) use a large-scale handcrafted grammar (Rohrer and Forst, 2006) to generate a number of (almost always) grammatical sentences given an input F-Structure. They show that a linguistically-inspired log-linear ranking model outperforms a simple baseline tri-gram language model trained on the Huge German Corpus (HGC), a corpus of 200 million words of newspaper and other text.

Cahill and Forst (2009) describe a number of experiments where they collect judgements from native speakers about the three systems compared in Cahill et al. (2007): (i) the original corpus string, (ii) the string chosen by the language model, and (iii) the string chosen by the linguistically-inspired log-linear model.[1] We only take the data from 2 of those experiments since the remaining experiments would not provide any

[1] In all cases, the three strings were different.

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 97–100,
Suntec, Singapore, 4 August 2009.

informative correlations. In the first experiment that we consider (A), subjects are asked to rank on a scale from 1–3 (1 being the best, 3 being the worst) the output of the three systems (joint rankings were not permitted). In the second experiment (B), subjects were asked to rank on a scale from 1–5 (1 being the worst, 5 being the best) how natural sounding the string chosen by the log-linear model was. The goal of experiment B was to determine whether the log-linear model was choosing good or bad alternatives to the original string. Judgements on the data were collected from 24 native German speakers. There were 44 items in Experiment A with an average sentence length of 14.4, and there were 52 items in Experiment B with an average sentence length of 12.1. Each item was judged by each native speaker at least once.

3 Correlation with Automatic Metrics

We examine the correlation between the human judgements and a number of automatic metrics:

BLEU (Papineni et al., 2001) calculates the number of *n*-grams a solution shares with a reference, adjusted by a brevity penalty. Usually the geometric mean for scores up to 4-gram are reported.

ROUGE (Lin, 2004) is an evaluation metric designed to evaluate automatically generated summaries. It comprises a number of string comparison methods including *n*-gram matching and skip-*n*grams. We use the default ROUGE-L longest common subsequence f-score measure.[2]

GTM General Text Matching (Melamed et al., 2003) calculates word overlap between a reference and a solution, without double counting duplicate words. It places less importance on word order than BLEU.

SED Levenshtein (String Edit) distance

WER Word Error Rate

TER Translation Error Rate (Snover et al., 2006) computes the number of insertions, deletions, substitutions and shifts needed to match a solution to a reference.

Most of these metrics come from the Machine Translation field, where the task is arguably significantly different. In the evaluation of a surface realisation system (as opposed to a complete generation system), typically the choice of vocabulary is limited and often the task is closer to word reordering. Many of the MT metrics have methods for attempting to account for different but equivalent translations of a given source word, typically by integrating a lexical resource such as WordNet. Also, these metrics were mostly designed to evaluate English output, so it is not clear that they will be equally appropriate for other languages, especially freer word order ones, such as German.

[2]Preliminary experiments with the skip *n*-grams performed worse than the default parameters.

	Experiment A			Experiment B
	GOLD	LM	LL	LL
human A (rank 1–3)	1.4	2.55	2.05	
human B (scale 1–5)				3.92
BLEU	1.0	0.67	0.72	0.79
ROUGE-L	1.0	0.85	0.78	0.85
GTM	1.0	0.55	0.60	0.74
SED	1.0	0.54	0.61	0.71
WER	0.0	48.04	39.88	28.83
TER	0.0	0.16	0.14	0.11
DEP	100	82.60	87.50	93.11
WDEP	1.0	0.70	0.82	0.90

Table 1: Average scores of each metric for Experiment A data

	Sentence		Corpus	
	corr	p-value	corr	p-value
BLEU	-0.615	<0.001	-1	0.3333
ROUGE-L	-0.644	<0.001	-0.5	1
GTM	-0.643	<0.001	-1	0.3333
SED	-0.628	<0.001	-1	0.3333
WER	0.623	<0.001	1	0.3333
TER	0.608	<0.001	1	0.3333

Table 2: Correlation between human judgements for experiment A (rank 1–3) and automatic metrics

The scores given by each metric for the data used in both experiments are presented in Table 1. For the Experiment A data, we use the Spearman rank correlation coefficient to measure the correlation between the human judgements and the automatic scorers. The results are presented in Table 2 for both the sentence and the corpus level correlations, we also present p-values for statistical significance. Since we only have judgements on three systems, the corpus correlation is not that informative. Interestingly, the ROUGE-L metric is the only one that does not rank the output of the three systems in the same order as the judges. It ranks the strings chosen by the language model higher than the strings chosen by the log-linear model. However, at the level of the individual sentence, the ROUGE-L metric correlates best with the human judgements. The GTM metric correlates at about the same level, but in general there seems to be little difference between the metrics.

For the Experiment B data we use the Pearson correlation coefficient to measure the correlation between the human judgements and the automatic

	Sentence Correlation	P-Value
BLEU	0.095	0.5048
ROUGE-L	0.207	0.1417
GTM	0.424	0.0017
SED	0.168	0.2344
WER	-0.188	0.1817
TER	-0.024	0.8646

Table 3: Correlation between human judgements for experiment B (naturalness scale 1–5) and automatic metrics

metrics. The results are given in Table 3. Here we only look at the correlation at the individual sentence level, since we are looking at data from only one system. For this data, the GTM metric clearly correlates most closely with the human judgements, and it is the only metric that has a statistically significant correlation. BLEU and TER correlate particularly poorly, with correlation coefficients very close to zero.

3.1 Syntactic Metrics

Recently, there has been a move towards more syntactic, rather than purely string based, evaluation of MT output and summarisation (Hovy et al., 2005; Owczarzak et al., 2008). The idea is to go beyond simple string comparisons and evaluate at a deeper linguistic level. Since most of the work in this direction has only been carried out for English so far, we apply the idea rather than a specific tool to the data. We parse the data from both experiments with a German dependency parser (Hall and Nivre, 2008) trained on the TIGER Treebank (with sentences 8000-10000 heldout for testing). This parser achieves 91.23% labelled accuracy on the 2000-sentence test set.

To calculate the correlation between the human judgements and the dependency parser, we parse the original strings as well as the strings chosen by the log-linear and language models. The standard evaluation procedure relies on both strings being identical to calculate (un-)labelled dependency accuracy, and so we map the dependencies produced by the parser into sets of triples as used in the evaluation software of Crouch et al. (2002) where each dependency is represented as `deprel(head,word)` and each word is indexed with its position in the original string.[3] We compare the parses for both experiments against the parses of the original strings. We calculate both a weighted and unweighted dependency f-score, as given in Table 1. The unweighted f-score is calculated by taking the average of the scores for each dependency type, while the weighted f-score weighs each average score by its frequency in the test corpus. We calculate the Spearman and Pearson correlation coefficients as before; the results are given in Table 4. The results show that the unweighted dependencies correlate more closely (and statistically significantly) with the human judgements than the weighted ones. This suggests that the frequency of a dependency type does not matter as much as its overall correctness.

[3]This is a 1-1 mapping, and the indexing ensures that duplicate words in a sentence are not confused.

	Experiment A		Experiment B	
	corr	p-value	corr	p-value
Dependencies	-0.640	<0.001	0.186	0.1860
Unweighted Deps	-0.657	<0.001	0.290	0.03686

Table 4: Correlation between dependency-based evaluation and human judgements

4 Discussion

The large discrepancy between the absolute correlation coefficients for Experiment A and B can be explained by the fact that they are different tasks. Experiment A ranks 3 strings relative to one another, while Experiment B measures the naturalness of the string. We would expect automatic metrics to be better at the first task than the second, as it is easier to rank systems relative to each other than to give a system an absolute score.

Disappointingly, the correlation between the dependency parsing metric and the human judgements was no higher than the simple GTM string-based metric (although it did outperform all other automatic metrics). This does not correspond to related work on English Summarisation evaluation (Owczarzak, 2009) which shows that a metric based on an automatically induced LFG parser for English achieves comparable or higher correlation with human judgements than ROUGE and Basic Elements (BE).[4] Parsers of German typically do not achieve as high performance as their English counterparts, and further experiments including alternative parsers are needed to see if we can improve performance of this metric.

The data used in our experiments was almost always grammatically correct. Therefore the task

[4]The GTM metric was not compared in that paper

of an evaluation system is to score more natural sounding strings higher than marked or unnatural ones. In this respect, our findings mirror those of Stent et al. (2005) for English data, that the automatic metrics do not correlate well with human judges on syntactic correctness.

5 Conclusions

We presented data that examined the correlation between native speaker judgements and automatic evaluation metrics on automatically generated German text. We found that for our first experiment, all metrics were correlated to roughly the same degree (with ROUGE-L achieving the highest correlation at an individual sentence level and the GTM tool not far behind). At a corpus level all except ROUGE were in agreement with the human judgements. In the second experiment, the General Text Matcher Tool had the strongest correlation. We carried out an experiment to test whether a more sophisticated syntax-based evaluation metric performed better than the more simple string-based ones. We found that while the unweighted dependency evaluation metric correlated with the human judgements more strongly than almost all metrics, it did not outperform the GTM tool. The correlation between the human judgements and the automatic evaluation metrics was much higher for the relative ranking task than for the naturalness task.

Acknowledgments

This work was funded by the Collaborative Research Centre (SFB 732) at the University of Stuttgart. We would like to thank Martin Forst, Alex Fraser and the anonymous reviewers for their helpful feedback. Furthermore, we would like to thank Johan Hall, Joakim Nivre and Yannick Versely for their help in retraining the MALT dependency parser with our data set.

References

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *Proceedings of EACL 2006*, pages 313–320, Trento, Italy.

Aoife Cahill and Martin Forst. 2009. Human Evaluation of a German Surface Realisation Ranker. In *Proceedings of EACL 2009*, pages 112–120, Athens, Greece, March.

Aoife Cahill, Martin Forst, and Christian Rohrer. 2007. Stochastic Realisation Ranking for a Free Word Order Language. In *Proceedings of ENLG-07*, pages 17–24, Saarbrücken, Germany, June.

Richard Crouch, Ron Kaplan, Tracy Holloway King, and Stefan Riezler. 2002. A comparison of evaluation metrics for a broad coverage parser. In *Proceedings of the LREC Workshop: Beyond PARSEVAL*, pages 67–74, Las Palmas, Spain.

Johan Hall and Joakim Nivre. 2008. A dependency-driven parser for German dependency and constituency representations. In *Proceedings of the Workshop on Parsing German*, pages 47–54, Columbus, Ohio, June.

Eduard Hovy, Chin yew Lin, and Liang Zhou. 2005. Evaluating duc 2005 using basic elements. In *Proceedings of DUC-2005*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and recall of machine translation. In *Proceedings of NAACL-03*, pages 61–63, NJ, USA.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2008. Evaluating machine translation with LFG dependencies. *Machine Translation*, 21:95–119.

Karolina Owczarzak. 2009. DEPEVAL(summ): Dependency-based Evaluation for Automatic Summaries. In *Proceedings of ACL-IJCNLP 2009*, Singapore.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL-02*, pages 311–318, NJ, USA.

Ehud Reiter and Anja Belz. 2009. An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems. *Computational Linguistics*, 35.

Christian Rohrer and Martin Forst. 2006. Improving Coverage and Parsing Quality of a Large-Scale LFG for German. In *Proceedings of LREC 2006*, Genoa, Italy.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and Ralph Weischedel. 2006. A study of translation error rate with targeted human annotation. In *Proceedings of AMTA 2006*, pages 223–231.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presense of variation. In *Proceedings of CICLING*, pages 341–351.