Paraphrase Generation as Monolingual Translation: Data and Evaluation[1]

Sander Wubben   Antal van den Bosch   Emiel Krahmer

Tilburg centre for Cognition and Communication,
Tilburg University, P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands

Abstract

We describe a sentential paraphrase generation (SPG) method that uses a large monolingual aligned corpus of paired news headlines belonging to the same news item, acquired automatically from Google News, and a standard Phrase-Based Machine Translation (PBMT) framework. The output of this system is compared to a word substitution baseline that replaces words by near-synonyms. Human judges prefer the PBMT paraphrasing system over the word substitution system. We demonstrate that the BLEU evaluation metric correlates well with human judgements, provided that the generated paraphrased sentence is sufficiently different from the source sentence.

1 Sentential Paraphrase Generation

SPG is a form of text-to-text generation. Given a source sentence, a new sentence is generated that differs in form but maintains the meaning of the original sentence. Currently, no sufficiently large paraphrase corpora exist. We collected a paraphrase corpus of automatically obtained aligned headlines crawled from Google News in order to train a paraphrase generation model using the the MOSES package, a phrase-based machine translation (PBMT) framework [1]. We use the method earlier described in [3] to align the paraphasing headlines in the Google News clusters, resulting in a corpus of 7,400,144 pairwise alignments of 1,025,605 unique headlines. We compare this MT approach to a word substitution baseline that makes use of WordNet to replace words with their (near-)synonyms. The generated paraphrases along with their source headlines are presented to human judges, whose ratings are compared to a set of automatic evaluation metrics, among which the BLEU metric [2].

2 Results

The average scores assigned by the human judges to the output of the two systems are displayed in Table 1. The judges rate the quality of the PBMT paraphrases significantly higher than those generated by the word substitution system ($t(18) = 4.11, p < .001$). The automatic measures also prefer the PBMT output over the baseline. There is an overall medium correlation between the BLEU measure and human judgements ($r = 0.41, p < 0.001$). We see a lower correlation between the various ROUGE scores and human judgements, with ROUGE-1 showing the highest correlation ($r = 0.29, p < 0.001$). Between the two lies the METEOR correlation ($r = 0.35, p < 0.001$). However, if we split the data according to Levenshtein distance between source and target over tokens, we observe that we generally get a higher correlation for all the tested metrics when the Levenshtein distance is higher, as visualized in Figure 1. A higher Levenshtein distance, or edit distance, means that the generated paraphrase differs more from the original. The distance expresses the number of insertions, deletions, or transpositions needed at the word level to change the source sentence to the paraphrase. At Levenshtein distance 5, the BLEU score achieves a correlation of 0.78 with human judgements, while ROUGE-1 manages to achieve a 0.74 correlation. Beyond edit distance 5, data sparsity occurs.

[1] This paper is accepted for publication in the Proceedings of the Sixth International Natural Language Generation Conference, INLG-2010, Dublin, Ireland

system	judges mean	BLEU	ROUGE-1	ROUGE-2	ROUGE-SU4	METEOR
PBMT	4.60	0.51	0.76	0.36	0.42	0.71
Word Substitution	3.59	0.25	0.59	0.22	0.26	0.54

Table 1: Results of human judgements (N = 10) and automatic measures

3 Discussion

With a parallel monolingual corpus with several millions of paired paraphrases as training data, it is possible to develop an SPG system by treating paraphrasing as an MT task. Human judges preferred the output of our PBMT system over the output of a word substitution baseline system. We have also addressed the problem of automatic paraphrase evaluation. We measured BLEU, METEOR and ROUGE scores, and observed that the outcomes of these automatic evaluation metrics correlate with human judgements to some degree, but that the correlation is highly dependent on the difference between the original and the paraphrase. At low edit distances, automatic metrics fail to properly assess the quality of paraphrases, whereas at edit distance 5 the correlation of BLEU with human judgements is 0.78, indicating that at higher edit distances these automatic measures can be utilized to rate the quality of the generated paraphrases. From edit distance 2, BLEU correlates best with human judgements, indicating that this MT evaluation metric might be best for SPG evaluation.

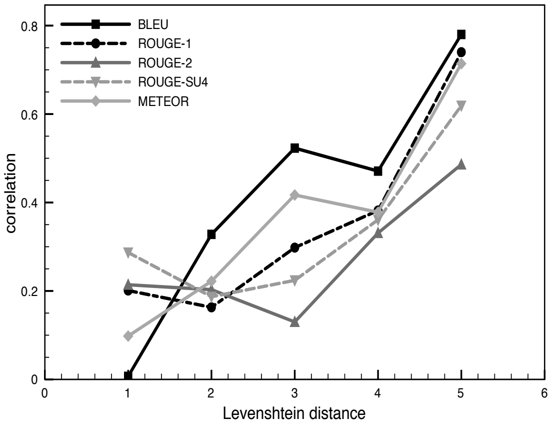

Figure 1: Correlations between human judgements and automatic evaluation metrics for various edit distances

References

[1] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris C. Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *ACL*. The Association for Computer Linguistics, 2007.

[2] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[3] Sander Wubben, Antal van den Bosch, Emiel Krahmer, and Erwin Marsi. Clustering and matching headlines for automatic paraphrase acquisition. In *ENLG '09: Proceedings of the 12th European Workshop on Natural Language Generation*, pages 122–125, Morristown, NJ, USA, 2009. Association for Computational Linguistics.