Paraphrase Generation as Monolingual Translation: Data and Evaluation

Sander Wubben, Antal van den Bosch, Emiel Krahmer
Tilburg centre for Cognition and Communication
Tilburg University
Tilburg, The Netherlands
{s.wubben,antal.vdnbosch,e.j.krahmer}@uvt.nl

Abstract

In this paper we investigate the automatic generation and evaluation of sentential paraphrases. We describe a method for generating sentential paraphrases by using a large aligned monolingual corpus of news headlines 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. Human judges prefer the PBMT paraphrasing system over the word substitution system. We demonstrate that BLEU correlates well with human judgements provided that the generated paraphrased sentence is sufficiently different from the source sentence.

1 Introduction

Text-to-text generation is an increasingly studied subfield in natural language processing. In contrast with the typical natural language generation paradigm of converting concepts to text, in text-to-text generation a source text is converted into a target text that approximates the meaning of the source text. Text-to-text generation extends to such varied tasks as summarization (Knight and Marcu, 2002), question-answering (Lin and Pantel, 2001), machine translation, and paraphrase generation.

Sentential paraphrase generation (SPG) is the process of transforming a source sentence into a target sentence in the same language which differs in form from the source sentence, but approximates its meaning. Paraphrasing is often used as a subtask in more complex NLP applications to allow for more variation in text strings presented as input, for example to generate paraphrases of questions that in their original form cannot be answered (Lin and Pantel, 2001; Riezler et al., 2007), or to generate paraphrases of sentences that failed to translate (Callison-Burch et al., 2006). Paraphrasing has also been used in the evaluation of machine translation system output (Russo-Lassner et al., 2006; Kauchak and Barzilay, 2006; Zhou et al., 2006). Adding certain constraints to paraphrasing allows for additional useful applications. When a constraint is specified that a paraphrase should be shorter than the input text, paraphrasing can be used for sentence compression (Knight and Marcu, 2002; Barzilay and Lee, 2003) as well as for text simplification for question answering or subtitle generation (Daelemans et al., 2004).

We regard SPG as a monolingual machine translation task, where the source and target languages are the same (Quirk et al., 2004). However, there are two problems that have to be dealt with to make this approach work, namely obtaining a sufficient amount of examples, and a proper evaluation methodology. As Callison-Burch et al. (2008) argue, automatic evaluation of paraphrasing is problematic. The essence of SPG is to generate a sentence that is structurally different from the source. Automatic evaluation metrics in related fields such as machine translation operate on a notion of similarity, while paraphrasing centers around achieving dissimilarity. Besides the evaluation issue, another problem is that for an data-driven MT account of paraphrasing to work, a large collection of data is required. In this case, this would have to be pairs of sentences that are paraphrases of each other. So far, paraphrasing data sets of sufficient size have been mostly lacking. We argue that the headlines aggregated by Google News offer an attractive avenue.

2 Data Collection

Currently not many resources are available for paraphrasing; one example is the Microsoft Paraphrase Corpus (MSR) (Dolan et al., 2004; Nelken and Shieber, 2006), which with its 139,000 aligned

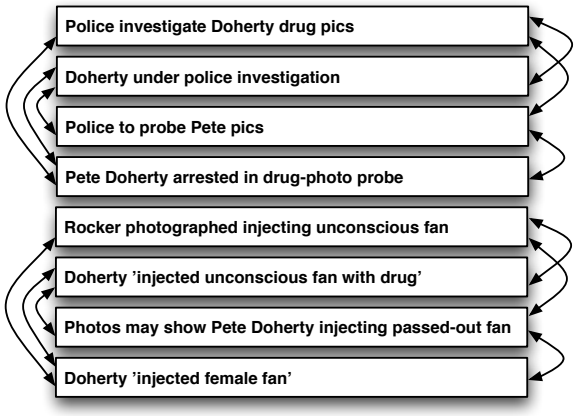

Figure 1: Part of a sample headline cluster, with aligned paraphrases

paraphrases can be considered relatively small. In this study we explore the use of a large, automatically acquired aligned paraphrase corpus. Our method consists of crawling the headlines aggregated and clustered by Google News and then aligning paraphrases within each of these clusters. An example of such a cluster is given in Figure 1. For each pair of headlines in a cluster, we calculate the Cosine similarity over the word vectors of the two headlines. If the similarity exceeds a defined upper threshold it is accepted; if it is below a defined lower threshold it is rejected. In the case that it lies between the thresholds, the process is repeated but then with word vectors taken from a snippet from the corresponding news article. This method, described in earlier work Wubben et al. (2009), was reported to yield a precision of 0.76 and a recall of 0.41 on clustering actual Dutch paraphrases in a headline corpus. We adapted this method to English. Our data consists of English headlines that appeared in Google News over the period of April to September 2006. Using this method we end up with a corpus of 7,400,144 pairwise alignments of 1,025,605 unique headlines[1].

3 Paraphrasing methods

In our approach we use the collection of automatically obtained aligned headlines to train a paraphrase generation model using a Phrase-Based MT framework. We compare this approach to a word substitution baseline. The generated paraphrases along with their source headlines are presented to human judges, whose ratings are compared to the BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and ROUGE (Lin, 2004) automatic evaluation metrics.

3.1 Phrase-Based MT

We use the MOSES package to train a Phrase-Based Machine Translation model (PBMT) (Koehn et al., 2007). Such a model normally finds a best translation $\tilde{e}$ of a text in language f to a text in language e by combining a translation model $p(f|e)$ with a language model $p(e)$:

$$\tilde{e} = arg \max_{e \in e^*} p(f|e)p(e)$$

GIZA++ is used to perform the word alignments (Och and Ney, 2003) which are then used in the Moses pipeline to generate phrase alignments in order to build the paraphrase model. We first tokenize our data before training a recaser. We then lowercase all data and use all unique headlines in the training data to train a language model with the SRILM toolkit (Stolcke, 2002). Then we invoke the GIZA++ aligner using the 7M training paraphrase pairs. We run GIZA++ with standard settings and we perform no optimization. Finally, we use the MOSES decoder to generate paraphrases for our test data.

Instead of assigning equal weights to language and translation model, we assign a larger weight of 0.7 to the language model to generate better formed (but more conservative) paraphrases. Because dissimilarity is a factor that is very important for paraphrasing but not implemented in a PBMT model, we perform post-hoc reranking of the different candidate outputs based on dissimilarity. For each headline in the testset we generate the ten best paraphrases as scored by the decoder and then rerank them according to dissimilarity to the source using the Levenshtein distance measure at the word level. The resulting headlines are recased using the previously trained recaser.

3.2 Word Substitution

We compare the PBMT results with a simple word substitution baseline. For each noun, adjective and verb in the sentence this model takes that word and its Part of Speech tag and retrieves from WordNet its most frequent synonym from the most frequent synset containing the input word. We use the Memory Based Tagger (Daelemans et al., 1996)

[1]This list of aligned pairs is available at http://ilk.uvt.nl/∼swubben/resources.html

System	Headline
Source	Florida executes notorious serial killer
PBMT	Serial killer executed in Florida
Word Sub.	Florida executes ill-famed series slayer
Source	Dublin evacuates airport due to bomb scare
PBMT	Dublin airport evacuated after bomb threat
Word Sub.	Dublin evacuates airdrome due to bomb panic
Source	N. Korea blasts nuclear sanctions
PBMT	N. Korea nuclear blast of sanctions
Word Sub.	N. Korea blasts atomic sanctions

Table 1: Examples of generated paraphrased headlines

trained on the Brown corpus to generate the POS-tags. The WordNet::QueryData[2] Perl module is used to query WordNet (Fellbaum, 1998). Generated headlines and their source for both systems are given in Table 1.

4 Evaluation

For the evaluation of the generated paraphrases we set up a human judgement study, and compare the human judges' ratings to automatic evaluation measures in order to gain more insight in the automatic evaluation of paraphrasing.

4.1 Method

We randomly select 160 headlines that meet the following criteria: the headline has to be comprehensible without reading the corresponding news article, both systems have to be able to produce a paraphrase for each headline, and there have to be a minimum of eight paraphrases for each headline. We need these paraphrases as multiple references for our automatic evaluation measures to account for the diversity in real-world paraphrases, as the aligned paraphrased headlines in Figure 1 witness.

The judges are presented with the 160 headlines, along with the paraphrases generated by both systems. The order of the headlines is randomized, and the order of the two paraphrases for each headline is also randomized to prevent a bias towards one of the paraphrases. The judges are asked to rate the paraphrases on a 1 to 7 scale, where 1 means that the paraphrase is very bad and 7 means that the paraphrase is very good. The judges were instructed to base their overall quality judgement on whether the meaning was retained, the paraphrase was grammatical and fluent, and whether the paraphrase was in fact different from the source sentence. Ten judges rated two paraphrases per headline, resulting in a total of 3,200 scores. All judges were blind to the purpose of the evaluation and had no background in paraphrasing research.

[2]http://search.cpan.org/dist/WordNet-QueryData/QueryData.pm

system	mean	stdev.
PBMT	4.60	0.44
Word Substitution	3.59	0.64

Table 2: Results of human judgements (N = 10)

4.2 Results

The average scores assigned by the human judges to the output of the two systems are displayed in Table 2. These results show that the judges rated the quality of the PBMT paraphrases significantly higher than those generated by the word substitution system ($t(18) = 4.11, p < .001$).

Results from the automatic measures as well as the Levenshtein distance are listed in Table 3. We use a Levenshtein distance over tokens. First, we observe that both systems perform roughly the same amount of edit operations on a sentence, resulting in a Levenshtein distance over words of 2.76 for the PBMT system and 2.67 for the Word Substitution system. BLEU, METEOR and three typical ROUGE metrics[3] all rate the PBMT system higher than the Word Substitution system. Notice also that the all metrics assign the highest scores to the original sentences, as is to be expected: because every operation we perform is in the same language, the source sentence is also a paraphrase of the reference sentences that we use for scoring our generated headline. If we pick a random sentence from the reference set and score it against the rest of the set, we obtain similar scores. This means that this score can be regarded as an upper bound score for paraphrasing: we can not expect our paraphrases to be better than those produced by humans. However, this also shows that these measures cannot be used directly as an automatic evaluation method of paraphrasing, as they assign the highest score to the "paraphrase" in which nothing has changed. The scores observed in Table 3 do indicate that the paraphrases gener-

[3]ROUGE-1, ROUGE-2 and ROUGE-SU4 are also adopted for the DUC 2007 evaluation campaign, http://www-nlpir.nist.gov/projects/duc/duc2007/tasks.html

System	BLEU	ROUGE-1	ROUGE-2	ROUGE-SU4	METEOR	Lev.dist.	Lev. stdev.
PBMT	50.88	0.76	0.36	0.42	0.71	2.76	1.35
Wordsub.	24.80	0.59	0.22	0.26	0.54	2.67	1.50
Source	60.58	0.80	0.45	0.47	0.77	0	0

Table 3: Automatic evaluation and sentence Levenshtein scores

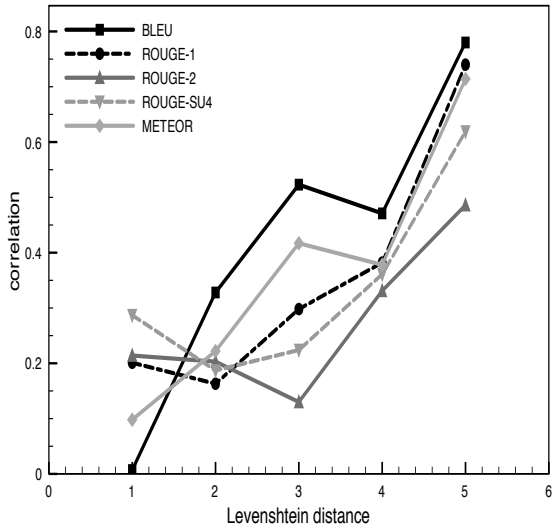

Figure 2: Correlations between human judgements and automatic evaluation metrics for various edit distances

ated by PBMT are less well formed than the original source sentence.

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, we observe that we generally get a higher correlation for all the tested metrics when the Levenshtein distance is higher, as visualized in Figure 2. 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.

5 Discussion

In this paper we have shown that with an automatically obtained parallel monolingual corpus with several millions of paired examples, it is possible to develop an SPG system based on a PBMT framework. Human judges preferred the output of our PBMT system over the output of a word substitution system. We have also addressed the problem of automatic paraphrase evaluation. We measured BLEU, METEOR and ROUGE scores, and observed that these automatic scores correlate with human judgements to some degree, but that the correlation is highly dependent on edit distance. 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 MT evaluation metrics might be best for SPG evaluation.

The data we used for paraphrasing consists of headlines. Paraphrase patterns we learn are those used in headlines and therefore different from standard language. The advantage of our approach is that it paraphrases those parts of sentences that it can paraphrase, and leaves the unknown parts intact. It is straightforward to train a language model on in-domain text and use the translation model acquired from the headlines to generate paraphrases for other domains. We are also interested in capturing paraphrase patterns from other domains, but acquiring parallel corpora for these domains is not trivial.

Instead of post-hoc dissimilarity reranking of the candidate paraphrase sentences we intend to develop a proper paraphrasing model that takes dissimilarity into account in the decoding process. In addition, we plan to investigate if our paraphrase generation approach is applicable to sentence compression and simplification. On the topic of automatic evaluation, we aim to define an automatic paraphrase generation assessment score. A paraphrase evaluation measure should be able to recognize that a good paraphrase is a well-formed sentence in the source language, yet it is clearly dissimilar to the source.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, June.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 16–23.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 17–24.

Chris Callison-Burch, Trevor Cohn, and Mirella Lapata. 2008. Parametric: an automatic evaluation metric for paraphrasing. In *COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics*, pages 97–104.

Walter Daelemans, Jakub Zavrel, Peter Berck, and Steven Gillis. 1996. Mbt: A memory-based part of speech tagger-generator. In *Proc. of Fourth Workshop on Very Large Corpora*, pages 14–27.

Walter Daelemans, Anja Hothker, and Erik Tjong Kim Sang. 2004. Automatic sentence simplification for subtitling in dutch and english. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1045–1048.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 350.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*, May.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 455–462, June.

Kevin Knight and Daniel Marcu. 2002. Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Artif. Intell.*, 139(1):91–107.

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. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL*.

Dekang Lin and Patrick Pantel. 2001. Dirt: Discovery of inference rules from text. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 323–328.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, page 10.

Rani Nelken and Stuart M. Shieber. 2006. Towards robust context-sensitive sentence alignment for monolingual corpora. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, 3–7 April.

Franz J. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, March.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 142–149, July.

Stefan Riezler, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu O. Mittal, and Yi Liu. 2007. Statistical machine translation for query expansion in answer retrieval. In *ACL*.

Grazia Russo-Lassner, Jimmy Lin, and Philip Resnik. 2006. A paraphrase-based approach to machine translation evaluation. Technical report, University of Maryland, College Park.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *In Proc. Int. Conf. on Spoken Language Processing*, pages 901–904.

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

Liang Zhou, Chin-Yew Lin, and Eduard Hovy. 2006. Re-evaluating machine translation results with paraphrase support. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 77–84, July.