An Empirical Evaluation of Data-Driven Paraphrase Generation Techniques

Donald Metzler
Information Sciences Institute
Univ. of Southern California
Marina del Rey, CA, USA
metzler@isi.edu

Eduard Hovy
Information Sciences Institute
Univ. of Southern California
Marina del Rey, CA, USA
hovy@isi.edu

Chunliang Zhang
Information Sciences Institute
Univ. of Southern California
Marina del Rey, CA, USA
czheng@isi.edu

Abstract

Paraphrase generation is an important task that has received a great deal of interest recently. Proposed data-driven solutions to the problem have ranged from simple approaches that make minimal use of NLP tools to more complex approaches that rely on numerous language-dependent resources. Despite all of the attention, there have been very few direct empirical evaluations comparing the merits of the different approaches. This paper empirically examines the tradeoffs between simple and sophisticated paraphrase harvesting approaches to help shed light on their strengths and weaknesses. Our evaluation reveals that very simple approaches fare surprisingly well and have a number of distinct advantages, including strong precision, good coverage, and low redundancy.

1 Introduction

A popular idiom states that "variety is the spice of life". As with life, variety also adds spice and appeal to language. Paraphrases make it possible to express the same meaning in an almost unbounded number of ways. While variety prevents language from being overly rigid and boring, it also makes it difficult to algorithmically determine if two phrases or sentences express the same meaning. In an attempt to address this problem, a great deal of recent research has focused on identifying, generating, and harvesting phrase- and sentence-level paraphrases (Barzilay and McKeown, 2001; Bhagat and Ravichandran, 2008; Barzilay and Lee, 2003; Bannard and Callison-Burch, 2005; Callison-Burch, 2008; Lin and Pantel, 2001; Pang et al., 2003; Pasca and Dienes, 2005)

Many data-driven approaches to the paraphrase problem have been proposed. The approaches vastly differ in their complexity and the amount of NLP resources that they rely on. At one end of the spectrum are approaches that generate paraphrases from a large monolingual corpus and minimally rely on NLP tools. Such approaches typically make use of statistical co-occurrences, which act as a rather crude proxy for semantics. At the other end of the spectrum are more complex approaches that require access to bilingual parallel corpora and may also rely on part-of-speech (POS) taggers, chunkers, parsers, and statistical machine translation tools. Constructing large comparable and bilingual corpora is expensive and, in some cases, impossible.

Despite all of the previous research, there have not been any evaluations comparing the quality of simple and sophisticated data-driven approaches for generating paraphrases. Evaluation is not only important from a practical perspective, but also from a methodological standpoint, as well, since it is often more fruitful to devote attention to building upon the current state-of-the-art as opposed to improving upon less effective approaches. Although the more sophisticated approaches have garnered considerably more attention from researchers, from a practical perspective, simplicity, quality, and flexibility are the most important properties. But are simple methods adequate enough for the task?

The primary goal of this paper is to take a small step towards addressing the lack of comparative evaluations. To achieve this goal, we empirically

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 546–551,
Portland, Oregon, June 19-24, 2011.

evaluate three previously proposed paraphrase generation techniques, which range from very simple approaches that make use of little-to-no NLP or language-dependent resources to more sophisticated ones that heavily rely on such resources. Our evaluation helps develop a better understanding of the strengths and weaknesses of each type of approach. The evaluation also brings to light additional properties, including the number of redundant paraphrases generated, that future approaches and evaluations may want to consider more carefully.

2 Related Work

Instead of exhaustively covering the entire spectrum of previously proposed paraphrasing techniques, our evaluation focuses on two families of data-driven approaches that are widely studied and used. More comprehensive surveys of data-driven paraphrasing techniques can be found in Androutsopoulos and Malakasiotis (2010) and Madnani and Dorr (2010).

The first family of approaches that we consider harvests paraphrases from monolingual corpora using distributional similarity. The DIRT algorithm, proposed by Lin and Pantel (2001), uses parse tree paths as contexts for computing distributional similarity. In this way, two phrases were considered similar if they occurred in similar contexts within many sentences. Although parse tree paths serve as rich representations, they are costly to construct and yield sparse representations. The approach proposed by Pasca and Dienes (2005) avoided the costs associated with parsing by using n-gram contexts. Given the simplicity of the approach, the authors were able to harvest paraphrases from a very large collection of news articles. Bhagat and Ravichandran (2008) proposed a similar approach that used noun phrase chunks as contexts and locality sensitive hashing to reduce the dimensionality of the context vectors. Despite their simplicity, such techniques are susceptible to a number of issues stemming from the distributional assumption. For example, such approaches have a propensity to assign large scores to antonyms and other semantically irrelevant phrases.

The second line of research uses comparable or bilingual corpora as the 'pivot' that binds paraphrases together (Barzilay and McKeown, 2001; Barzilay and Lee, 2003; Bannard and Callison-Burch, 2005; Callison-Burch, 2008; Pang et al., 2003). Amongst the most effective recent work, Bannard and Callison-Burch (2005) show how different English translations of the same entry in a statistically-derived translation table can be viewed as paraphrases. The recent work by Zhao et al. (Zhao et al., 2009) uses a generalization of DIRT-style patterns to generate paraphrases from a bilingual parallel corpus. The primary drawback of these type of approaches is that they require a considerable amount of resource engineering that may not be available for all languages, domains, or applications.

3 Experimental Evaluation

The goal of our experimental evaluation is to analyze the effectiveness of a variety of paraphrase generation techniques, ranging from simple to sophisticated. Our evaluation focuses on generating paraphrases for verb phrases, which tend to exhibit more variation than other types of phrases. Furthermore, our interest in paraphrase generation was initially inspired by challenges encountered during research related to machine reading (Barker et al., 2007). Information extraction systems, which are key component of machine reading systems, can use paraphrase technology to automatically expand seed sets of relation triggers, which are commonly verb phrases.

3.1 Systems

Our evaluation compares the effectiveness of the following paraphrase harvesting approaches:

PD: The basic distributional similarity-inspired approach proposed by Pasca and Dienes (2005) that uses variable-length n-gram contexts and overlap-based scoring. The context of a phrase is defined as the concatenation of the n-grams immediately to the left and right of the phrase. We set the minimum length of an n-gram context to be 2 and the maximum length to be 3. The maximum length of a phrase is set to 5.

BR: The distributional similarity approach proposed by Bhagat and Ravichandran (2008) that uses noun phrase chunks as contexts and locality sensitive hashing to reduce the dimensionality of the contextual vectors.

BCB-S: An extension of the Bannard Callison-Burch (Bannard and Callison-Burch, 2005) approach that constrains the paraphrases to have the same syntactic type as the original phrase (Callison-Burch, 2008). We constrained all paraphrases to be verb phrases.

We chose these three particular systems because they span the spectrum of paraphrase approaches, in that the PD approach is simple and does not rely on any NLP resources while the BCB-S approach is sophisticated and makes heavy use of NLP resources.

For the two distributional similarity approaches (PD and BR), paraphrases were harvested from the English Gigaword Fourth Edition corpus and scored using the cosine similarity between PMI weighted contextual vectors. For the BCB-S approach, we made use of a publicly available implementation[1].

3.2 Evaluation Methodology

We randomly sampled 50 verb phrases from 1000 news articles about terrorism and another 50 verb phrases from 500 news articles about American football. Individual occurrences of verb phrases were sampled, which means that more common verb phrases were more likely to be selected and that a given phrase could be selected multiple times. This sampling strategy was used to evaluate the systems across a realistic sample of phrases. To obtain a richer class of phrases beyond basic verb groups, we defined verb phrases to be contiguous sequences of tokens that matched the following POS tag pattern: (TO | IN | RB | MD | VB)+.

Following the methodology used in previous paraphrase evaluations (Bannard and Callison-Burch, 2005; Callison-Burch, 2008; Kok and Brockett, 2010), we presented annotators with two sentences. The first sentence was randomly selected from amongst all of the sentences in the evaluation corpus that contain the original phrase. The second sentence was the same as the first, except the original phrase is replaced with the system generated paraphrase. Annotators were given the following options, which were adopted from those described by Kok and Brockett (2010), for each sentence pair: 0) Different meaning; 1) Same meaning; revised is grammatically incorrect; and 2) Same meaning; revised is grammatically correct. Table 1 shows three example sentence pairs and their corresponding annotations according to the guidelines just described.

Amazon's Mechanical Turk service was used to collect crowdsourced annotations. For each paraphrase system, we retrieve (up to) 10 paraphrases for each phrase in the evaluation set. This yields a total of 6,465 unique (phrase, paraphrase) pairs after pooling results from all systems. Each Mechanical Turk HIT consisted of 12 sentence pairs. To ensure high quality annotations and help identify spammers, 2 of the 12 sentence pairs per HIT were actually "hidden tests" for which the correct answer was known by us. We automatically rejected any HITs where the worker failed either of these hidden tests. We also rejected all work from annotators who failed at least 25% of their hidden tests. We collected a total of 51,680 annotations. We rejected 65% of the annotations based on the hidden test filtering just described, leaving 18,150 annotations for our evaluation. Each sentence pair received a minimum of 1, a median of 3, and maximum of 6 annotations. The raw agreement of the annotators (after filtering) was 77% and the Fleiss' Kappa was 0.43, which signifies moderate agreement (Fleiss, 1971; Landis and Koch, 1977).

The systems were evaluated in terms of coverage and expected precision at k. *Coverage* is defined as the percentage of phrases for which the system returned at least one paraphrase. *Expected precision at k* is the expected number of correct paraphrases amongst the top k returned, and is computed as:

$$\mathrm{E}[p@k] = \frac{1}{k}\sum_{i=1}^{k} p_i$$

where p_i is the proportion of positive annotations for item i. When computing the mean expected precision over a set of input phrases, only those phrases that generate one or more paraphrases is considered in the mean. Hence, if precision were to be averaged over all 100 phrases, then systems with poor coverage would perform significantly worse. Thus, one should take a holistic view of the results, rather than focus on coverage or precision in isolation, but consider them, and their respective tradeoffs, together.

[1] Available at `http://www.cs.jhu.edu/~ccb/`.

Sentence Pair	Annotation
A five-man presidential council for the independent state newly proclaimed in south Yemen was named overnight Saturday, it **was officially announced** in Aden. A five-man presidential council for the independent state newly proclaimed in south Yemen was named overnight Saturday, it **was cancelled** in Aden.	0
Dozens of Palestinian youths held rally in the Abu Dis Arab village in East Jerusalem **to protest** against the killing of Sharif. Dozens of Palestinian youths held rally in the Abu Dis Arab village in East Jerusalem **in protest of** against the killing of Sharif.	1
It says that foreign companies have no greater right to compensation – **establishing** debts at a 1/1 ratio of the dollar to the peso – than Argentine citizens do. It says that foreign companies have no greater right to compensation – **setting** debts at a 1/1 ratio of the dollar to the peso – than Argentine citizens do.	2

Table 1: Example annotated sentence pairs. In each pair, the first sentence is the original and the second has a system-generated paraphrase filled in (denoted by the bold text).

Method	C	Lenient			Strict		
		P1	P5	P10	P1	P5	P10
PD	**86**	.48	.42	.36	**.25**	**.22**	**.19**
BR	84	**.83**	**.65**	**.52**	.16	.17	.15
BCB-S	62	.63	.45	.34	.22	.17	.13

Table 2: Coverage (C) and expected precision at k (Pk) under lenient and strict evaluation criteria.

Method	Lenient			Strict		
	P1	P5	P10	P1	P5	P10
PD	**.26**	.22	**.20**	**.19**	**.16**	**.15**
BR	.05	.10	.11	.04	.05	.05
BCB-S	.24	**.25**	**.20**	.17	.14	.10

Table 3: Expected precision at k (Pk) when considering redundancy under lenient and strict evaluation criteria.

Two binarized evaluation criteria are reported. The *lenient* criterion allows for grammatical errors in the paraphrased sentence, while the *strict* criterion does not.

3.3 Basic Results

Table 2 summarizes the results of our evaluation. For this evaluation, all 100 verb phrases were run through each system. The paraphrases returned by the systems were then ranked (ordered) in descending order of their score, thus placing the highest scoring item at rank 1. Bolded values represent the best result for a given metric.

As expected, the results show that the systems perform significantly worse under the strict evaluation criteria, which requires the paraphrased sentences to be grammatically correct. None of the approaches tested used any information from the evaluation sentences (other than the fact a verb phrase was to be filled in). Recent work showed that using language models and/or syntactic clues from the evaluation sentence can improve the grammaticality of the paraphrased sentences (Callison-Burch, 2008). Such approaches could likely be used to improve the quality of all of the approaches under the strict evaluation criteria.

In terms of coverage, the distributional similarity approaches performed the best. In another set of experiments, we used the PD method to harvest paraphrases from a large Web corpus, and found that the coverage was 98%. Achieving similar coverage with resource-dependent approaches would likely require more human and machine effort.

3.4 Redundancy

After manually inspecting the results returned by the various paraphrase systems, we noticed that some approaches returned highly redundant paraphrases that were of limited practical use. For example, for the phrase "were losing", the BR system returned "are losing", "have been losing", "have lost", "lose", "might lose", "had lost", "stand to lose", "who have lost" and "would lose" within the top 10 paraphrases. All of these are simple variants that contain different forms of the verb "lose". Under the lenient evaluation criterion almost all of these paraphrases would be marked as correct, since the

same verb is being returned with some grammatical modifications. While highly redundant output of this form may be useful for some tasks, for others (such as information extraction) is it more useful to identify paraphrases that contain a diverse, non-redundant set of verbs.

Therefore, we carried out another evaluation aimed at penalizing highly redundant outputs. For each approach, we manually identified all of the paraphrases that contained the same verb as the main verb in the original phrase. During evaluation, these "redundant" paraphrases were regarded as non-related.

The results from this experiment are provided in Table 3. The results are dramatically different compared to those in Table 2, suggesting that evaluations that do not consider this type of redundancy may over-estimate actual system quality. The percentage of results marked as redundant for the BCB-S, BR, and PD approaches were 22.6%, 52.5%, and 22.9%, respectively. Thus, the BR system, which appeared to have excellent (lenient) precision in our initial evaluation, returns a very large number of redundant paraphrases. This remarkably reduces the lenient P1 from 0.83 in our initial evaluation to just 0.05 in our redundancy-based evaluation. The BCB-S and PD approaches return a comparable number of redundant results. As with our previous evaluation, the BCB-S approach tends to perform better under the lenient evaluation, while PD is better under the strict evaluation. Estimated 95% confidence intervals show all differences between BCB-S and PD are statistically significant, except for lenient P10.

Of course, existing paraphrasing approaches do not explicitly account for redundancy, and hence this evaluation is not completely fair. However, these findings suggest that redundancy may be an important issue to consider when developing and evaluating data-driven paraphrase approaches. There are likely other characteristics, beyond redundancy, that may also be important for developing robust, effective paraphrasing techniques. Exploring the space of such characteristics in a task-dependent manner is an important direction of future work.

3.5 Discussion

In all of our evaluations, we found that the simple approaches are surprisingly effective in terms of precision, coverage, and redundancy, making them a reasonable choice for an "out of the box" approach for this particular task. However, additional task-dependent comparative evaluations are necessary to develop even deeper insights into the pros and cons of the different types of approaches.

From a high level perspective, it is also important to note that the precision of these widely used, commonly studied paraphrase generation approaches is still extremely poor. After accounting for redundancy, the best approaches achieve a precision at 1 of less than 20% using the strict criteria and less than 26% when using the lenient criteria. This suggests that there is still substantial work left to be done before the output of these systems can reliably be used to support other tasks.

4 Conclusions and Future Work

This paper examined the tradeoffs between simple paraphrasing approaches that do not make use of any NLP resources and more sophisticated approaches that use a variety of such resources. Our evaluation demonstrated that simple harvesting approaches fare well against more sophisticated approaches, achieving state-of-the-art precision, good coverage, and relatively low redundancy.

In the future, we would like to see more empirical evaluations and detailed studies comparing the practical merits of various paraphrase generation techniques. As Madnani and Dorr (Madnani and Dorr, 2010) suggested, it would be beneficial to the research community to develop a standard, shared evaluation that would act to catalyze further advances and encourage more meaningful comparative evaluations of such approaches moving forward.

Acknowledgments

The authors gratefully acknowledge the support of the DARPA Machine Reading Program under AFRL prime contract no. FA8750-09-C-3705. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the DARPA, AFRL, or the US government. We would also like to thank the anonymous reviewers for their valuable feedback and the Mechanical Turk workers for their efforts.

References

I. Androutsopoulos and P. Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 597–604, Morristown, NJ, USA. Association for Computational Linguistics.

Ken Barker, Bhalchandra Agashe, Shaw-Yi Chaw, James Fan, Noah Friedland, Michael Glass, Jerry Hobbs, Eduard Hovy, David Israel, Doo Soon Kim, Rutu Mulkar-Mehta, Sourabh Patwardhan, Bruce Porter, Dan Tecuci, and Peter Yeh. 2007. Learning by reading: a prototype system, performance baseline and lessons learned. In *Proceedings of the 22nd national conference on Artificial intelligence - Volume 1*, pages 280–286. AAAI Press.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 16–23, Morristown, NJ, USA. Association for Computational Linguistics.

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 50–57, Morristown, NJ, USA. Association for Computational Linguistics.

Rahul Bhagat and Deepak Ravichandran. 2008. Large scale acquisition of paraphrases for learning surface patterns. In *Proceedings of ACL-08: HLT*, pages 674–682, Columbus, Ohio, June. Association for Computational Linguistics.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 196–205, Morristown, NJ, USA. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*, 76(5):378–382.

Stanley Kok and Chris Brockett. 2010. Hitting the right paraphrases in good time. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 145–153, Morristown, NJ, USA. Association for Computational Linguistics.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, March.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question-answering. *Nat. Lang. Eng.*, 7:343–360, December.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Comput. Linguist.*, 36:341–387.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: extracting paraphrases and generating new sentences. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 102–109, Morristown, NJ, USA. Association for Computational Linguistics.

Marius Pasca and Pter Dienes. 2005. Aligning needles in a haystack: Paraphrase acquisition across the web. In Robert Dale, Kam-Fai Wong, Jian Su, and Oi Yee Kwong, editors, *Natural Language Processing IJCNLP 2005*, volume 3651 of *Lecture Notes in Computer Science*, pages 119–130. Springer Berlin / Heidelberg.

Shiqi Zhao, Haifeng Wang, Ting Liu, and Sheng Li. 2009. Extracting paraphrase patterns from bilingual parallel corpora. *Natural Language Engineering*, 15(Special Issue 04):503–526.