Diversity-aware Evaluation for Paraphrase Patterns

Hideki Shima
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA
`hideki@cs.cmu.edu`

Teruko Mitamura
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA
`teruko@cs.cmu.edu`

Abstract

Common evaluation metrics for paraphrase patterns do not necessarily correlate with extrinsic recognition task performance. We propose a metric which gives weight to lexical variety in paraphrase patterns; our proposed metric has a positive correlation with paraphrase recognition task performance, with a Pearson correlation of 0.5~0.7 (k=10, with "strict" judgment) in a statistically significant level (p-value<0.01).

1   Introduction

We propose a diversity-aware paraphrase evaluation metric called DIMPLE[1], which boosts the scores of lexically diverse paraphrase pairs. Paraphrase pairs or patterns are useful in various NLP related research domains, since there is a common need to automatically identify meaning equivalence between two or more texts.

Consider a paraphrase pair resource that links "killed" to "assassinated" (in the rest of this paper we denote such a rule as ⟨"killed"[2], "assassinated"[3]⟩). In automatic evaluation for Machine Translation (MT) (Zhou et al., 2006; Kauchak and Barzilay, 2006; Padó et al., 2009), this rule may enable a metric to identify phrase-level semantic similarity between a system response containing "killed", and a reference translation containing "assassinated". Similarly in query expansion for information retrieval (IR) (Riezler et al., 2007), this rule may enable a system to expand the query term "killed" with the paraphrase "assassinated", in order to match a potentially relevant document containing the expanded term.

To evaluate paraphrase patterns during pattern discovery, ideally we should use an evaluation metric that strongly predicts performance on the extrinsic task (e.g. fluency and adequacy scores in MT, mean average precision in IR) where the paraphrase patterns are used.

Many existing approaches use a paraphrase evaluation methodology where human assessors judge each paraphrase pair as to whether they have the same meaning. Over a set of paraphrase rules for one source term, Expected Precision (EP) is calculated by taking the mean of precision, or the ratio of positive labels annotated by assessors (Bannard and Callison-Burch, 2005; Callison-Burch, 2008; Kok and Brockett, 2010; Metzler et al., 2011).

The weakness of this approach is that EP is an intrinsic measure that does not necessarily predict how well a paraphrase-embedded system will perform in practice. For example, a set of paraphrase pairs ⟨"killed", "shot and killed"⟩, ⟨"killed", "reported killed"⟩ ... ⟨"killed", "killed in"⟩ may receive a perfect score of 1.0 in EP; however, these patterns do not provide lexical diversity (e.g. ⟨"killed", "assassinated"⟩) and therefore may not perform well in an application where lexical diversity is important.

The goal of this paper is to provide empirical evidence to support the assumption that the proposed paraphrase evaluation metric DIMPLE correlates better with paraphrase recognition task metric scores than previous metrics do, by rewarding lexical diverse patterns.

2   DIMPLE Metric

Patterns or rules for capturing equivalence in meaning are used in various NLP applications. In a broad sense,

---

[1] DIversity-aware Metric for Pattern Learning Experiments
[2] Source term/phrase that contains "killed"
[3] Paraphrase that contains "assassinated"

the terms "paraphrase" will be used to denote pairs or a set of patterns that represent semantically equivalent or close texts with different surface forms.

Given paraphrase patterns P, or the ranked list of distinct paraphrase pairs sorted by confidence in descending order, DIMPLE$_k$ evaluates the top k patterns, and produces a real number between 0 and 1 (higher the better).

2.1 Cumulative Gain

DIMPLE is inspired by the Cumulative Gain (CG) metric (Järvelin and Kekäläinen, 2002; Kekäläinen, 2005) used in IR. CG for the top k retrieved documents is calculated as $CG_k = \sum_{i=1}^{k} gain_i$ where the gain function is human-judged relevance grade of the i-th document with respect to information need (e.g. 0 through 3 for irrelevant, marginally relevant, fairly relevant and highly relevant respectively). We take an alternative well-known formula for CG calculation, which puts stronger emphasis at higher gain:

$$CG_k = \sum_{i=1}^{k} (2^\wedge gain_i - 1).$$

2.2 DIMPLE Algorithm

DIMPLE is a normalized CG calculated on each paraphrase. The gain function of DIMPLE is represented as a product of pattern quality Q and lexical diversity D: $gain_i = Q_i \cdot D_i$. DIMPLE at rank k is a normalized CG$_k$ which is defined as:

$$DIMPLE_k = \frac{CG_k}{Z} = \frac{\sum_{i=1}^{k} \{2^\wedge (Q_i \cdot D_i) - 1\}}{Z}$$

where Z is a normalization factor such that the perfect CG score is given. Since Q takes a real value between 0 and 1, and D takes an integer between 1 and 3, $Z = \sum_{i=1}^{k} \{2^\wedge 3 - 1\}$.

Being able to design Q and D independently is one of characteristics in DIMPLE. In theory, Q can be any quality measure on paraphrase patterns, such as the instance-based evaluation score (Szpektor et al., 2007), or alignment-based evaluation score (Callison-Burch et al., 2008). Similarly, D can be implemented depending on the domain task; for example, if we are interested in learning paraphrases that are out-of-vocabulary or domain-specific, D could consult a dictionary, and return a high score if the lexical entry could not be found.

The DIMPLE framework is implemented in the following way[4]. Let Q be the ratio of positive labels

---

[4] Implementation used for this experiment is available at http://code.google.com/p/dimple/

---

averaged over pairs by human assessors given p_i as to whether a paraphrase has the same meaning as the source term or not. Let D be the degree of lexical diversity of a pattern calculated using Algorithm 1 below.

Algorithm 1. *D* score calculation

```
 Input: paraphrases {w₁, …, wₖ} for a source term s
 1: Set history1 = extractContentWords(s)
 2: Set history2 = stemWords(history1)
 3: for i=1 to k do
 4:     Set W1 = extractContentWords(wᵢ)
 5:     Set W2 = stemWords(W1) // Porter stemming
 6:     if W1==∅ || W1 ∩ history1 != ∅
 7:         D[i] = 1 // word already seen
 8:     else
 9:       if W2 ∩ history2 != ∅
10:           D[i] = 2 // root already seen
11:       else
12:           D[i] = 3 // unseen word
13:       end if
14:       history1 = W1 ∪ history1
15:       history2 = W2 ∪ history2
16:     end if
17: end for
```

3 Experiment

We use the Pearson product-moment correlation coefficient to measure correlation between two vectors consisting of intrinsic and extrinsic scores on paraphrase patterns, following previous meta-evaluation research (Callison-Burch et al., 2007; Callison-Burch et al., 2008; Tratz and Hovy, 2009; Przybocki et al., 2009). By *intrinsic* score, we mean a theory-based direct assessment result on the paraphrase patterns. By *extrinsic* score, we mean to measure how much the paraphrase recognition component helps the entire system to achieve a task. The correlation score is 1 if there is a perfect positive correlation, 0 if there is no correlation and -1 if there is a perfect negative correlation.

Using a task performance score to evaluate a paraphrase generation algorithm has been studied previously (Bhagat and Ravichandran, 2008; Szpektor and Dagan, 2007; Szpektor and Dagan, 2008). A common issue in extrinsic evaluations is that it is hard to separate out errors, or contributions from other possibly complex modules. This paper presents an approach which can predict task performance in more simple experimental settings.

3.1 Annotated Paraphrase Resource

We used the paraphrase pattern dataset "paraphrase-eval" (Metzler et al., 2011; Metzler and Hovy, 2011) which contains paraphrase patterns acquired by multiple algorithms: 1) PD (Pasca and Dienes, 2005),

which is based on the left and right n-gram contexts of the source term, with scoring based on overlap; 2) BR (Bhagat and Ravichandran, 2008), based on Noun Phrase chunks as contexts; 3) BCB (Bannard and Callison-Burch, 2005) and 4) BCB-S (Callison-Burch, 2008), which are based on monolingual phrase alignment from a bilingual corpus using a pivot. In the dataset, each paraphrase pair is assigned with an annotation as to whether a pair is a correct paraphrase or not by 2 or 3 human annotators.

The source terms are 100 verbs extracted from newswire about terrorism and American football. We selected 10 verbs according to their frequency in extrinsic task datasets (details follow in Section 3.3).

Following the methodology used in previous paraphrase evaluations (Bannard and Callison-Burch, 2005; Callison-Burch, 2008; Kok and Brockett, 2010), the labels were annotated on a pair of two sentences: an original sentence containing the source term, and the same sentence with the source term replaced with the paraphrase pattern, so that contextual information could help annotators to make consistent judgments. The judgment is based on whether the "same meaning" is present between the source term and its paraphrase. There is a lenient and a strict distinction on the "same meaning" judgments. The strict label is given when the replaced sentence is grammatically correct whereas the lenient label is given even when the sentence is grammatically incorrect.

In total, we have 10 (source terms listed in Table 1) × 4 (paraphrase generation algorithms introduced above) = 40 sets of paraphrase patterns. In each set of paraphrase patterns, there are up to 10 unique ⟨source term, paraphrase⟩ pairs.

3.2 Intrinsic Paraphrase Metrics

We will discuss the common metric EP, and its variant EPR as baselines to be compared with DIMPLE. For each metric, we used a cutoff value of k=1, 5 and 10.
EP: Our baseline is the Expected Precision at k, which is the expected number of correct paraphrases among the top k returned, and is computed as: $EP_k = \frac{1}{k}\sum_{i=1}^{k} Q_i$ where Q is the ratio of positive labels. For instance, if 2 out of 3 human annotators judged that $p_i = \langle$"killed", "fatally shot"$\rangle$ has the same meaning, $Q_i = 2/3$.
EPR: Metzler et al., (2011) extended EP with a Redundancy judgment, which we shall call EPR where lexically redundant paraphrases did not receive a credit. Unlike Metzler et al., (2011) where humans judged redundancies, we do the judgment automati-

cally with a Porter Stemmer (Porter, 1980) to extract and compare stemmed forms. In that way EPR's output become comparable to DIMPLE's, remaining redundancy scoring different (i.e. binary filtering in EPR and 3-level weighting in DIMPLE).

3.3 Extrinsic Evaluation Datasets

Ideally, paraphrase metric scores should correlate well with task performance metrics. To insulate the experiment from external, uncontrollable factors (e.g. errors from other task components), we created three datasets with slightly different characteristics, where the essential task of recognizing meaning equivalence between different surface texts can be conducted.

The numbers of positive-labeled pairs that we extracted for the three corpus, MSRPC, RTE and CQAE are 3900, 2805 and 27397 respectively. Table 1 shows the number of text pairs selected in which at least one of each pair contains a frequently occurring verb.

Src verb	MSRPC	RTE	CQAE
found	89	62	319
called	59	61	379
told	125	34	189
killed	48	109	277
accused	30	44	143
to take	21	23	63
reached	22	18	107
returned	14	20	57
turned	22	10	94
broke	10	10	35

Table 1. 10 most frequently occurring source verbs in three datasets. Numbers are positive-labeled pairs where the verb appears in at least one side of a pair.

MSRPC: The Microsoft Research Paraphrase Corpus (Dollan et al., 2005) contains 5800 pairs of sentences along with human annotations where positive labels mean semantic equivalence of pairs.
RTE: (Quasi-)paraphrase patterns are useful for the closely related task, Recognizing Textual Entailment. This dataset has been taken from the 2-way/3-way track at PASCAL/TAC RTE1-4. Positive examples are premise-hypothesis pairs where human annotators assigned the entailment label. The original dataset has been generated from actual applications such as Text Summarization, Information Extraction, IR, Question Answering.
CQAE: Complex Question Answering Evaluation (CQAE) dataset has been built from 6 past TREC QA tracks, i.e., "Other" QA data from TREC 2005 through 2007, relation QA data from TREC 2005 and ciQA from TREC 2006 and 2007 (Voorhees and Dang, 2005; Dang et al., 2006; Dang et al., 2007). We created unique pairs consisting of a system response (often sen-

37

tence-length) and an answer nugget as positive examples, where the system response is judged by human as containing or expressing the meaning of the nugget.

3.4 Extrinsic Performance Metric

Using the dataset described in Section 3.3, performance measures for each of the 40 paraphrase sets (10 verbs times 4 generators) are calculated as the ratio of pairs correctly identified as paraphrases.

In order to make the experimental settings close to an actual system with an embedded paraphrase engine, we first apply simple unigram matching with stemming enabled. At this stage, a text with the source verb "killed" and another text with the inflectional variant "killing" would match. As an alternative approach, we consult the paraphrase pattern set trying to find a match between the texts. This identification judgment is automated, where we assume a meaning equivalence is identified between texts when the source verb matches[5] one text and one of up to 10 paraphrases in the set matches the other. Given these evaluation settings, a noisy paraphrase pair such as ⟨"killed", "to"⟩ can easily match many pairs and falsely boost the performance score. We filter such exceptional cases when the paraphrase text contains only functional words.

3.5 Results

We conducted experiments to provide evidence that the Pearson correlation coefficient of DIMPLE is higher than that of the other two baselines. Table 2 and 3 below present the result where each number is the correlation calculated on the 40 data points.

	EP_k			EPR_k			$DIMPLE_k$		
k=1	5	10	1	5	10	1	5	10
MSRPC	-0.02	-0.24	-0.11	0.33	0.27	-0.12	0.32	0.20	0.25
RTE	0.13	-0.05	0.11	0.33	0.12	0.09	**0.46**	0.25	0.37
CQAE	0.08	-0.09	0.00	-0.02	-0.08	-0.13	0.35	0.25	**0.40**

Table 2. Correlation between intrinsic paraphrase metrics and extrinsic paraphrase recognition task metrics where DIMPLE's Q score is based on *lenient* judgment. Bold figures indicate statistical significance of the correlation statistics (null-hypothesis tested: "there is no correlation", p-value<0.01).

	EP_k			EPR_k			$DIMPLE_k$		
k=1	5	10	1	5	10	1	5	10
MSRPC	0.12	0.13	0.19	0.26	0.36	0.37	0.26	0.35	**0.52**
RTE	0.34	0.34	0.29	**0.43**	**0.41**	**0.38**	**0.49**	**0.55**	**0.58**
CQAE	**0.44**	**0.51**	**0.47**	0.37	**0.60**	**0.55**	0.37	**0.70**	**0.70**

Table 3. Same as the Table 2, except that the Q score is based on *strict* judgment.

---

[5] We consider word boundaries when matching texts, e.g. "skilled" and "killed" do not match.

Table 2 shows that correlations are almost always close to 0, indicating that EP does not correlate with the extrinsic measures when the Q score is calculated in lenient judgment mode. On the other hand, when the Q function is based on strict judgments, EP scores sometimes show a medium positive correlation with the extrinsic task performance, such as on the CQAE dataset.

In both tables, there is a general trend where the correlation scores fall in the same relative order (given the same cut-off value): EP < EPR < DIMPLE. This suggests that DIMPLE has a higher correlation than the other two baselines, given the task performance measure we experimented with. As we can see from Table 2, DIMPLE correlates well with paraphrase task performance, especially when the cutoff value k is 5 or 10. The higher values in Table 3 (compared to Table 2) show that the strict judgment used for intrinsic metric calculation is preferable over the lenient one.

4 Conclusion and Future Works

We proposed a novel paraphrase evaluation metric called DIMPLE, which gives weight to lexical variety. We built large scale datasets from three sources and conducted extrinsic evaluations where paraphrase recognition is involved. Experimental results showed that Pearson correlation statistics for DIMPLE are approximately 0.5 to 0.7 (when k=10 and "strict" annotations are used to calculate the score), which is higher than scores for the commonly used EP and EPR metrics.

Future works include applying DIMPLE on patterns for other tasks where lexical diversity matters (e.g. Relation Extraction) with a customized Q and D functions. If Q function can be also calculated fully automatically, DIMPLE may be useful for learning lexically diverse pattern learning when it is incorporated into optimization criteria.

References

Bannard, Colin and Chris Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora. In Proceedings of ACL 2005.

Bhagat, Rahul, Patrick Pantel, Eduard Hovy, and Marina Rey. 2007. LEDIR: An Unsupervised Algorithm for Learning Directionality of Inference Rules. In Proceedings of EMNLP-CoNLL 2007.

Bhagat, Rahul and Deepak Ravichandran. 2008. Large Scale Acquisition of Paraphrases for Learning Surface Patterns. In Proceedings of ACL-08: HLT.

Callison-Burch, Chris. 2008. Syntactic Constraints on Paraphrases Extracted from Parallel Corpora. In Proceedings of EMNLP 2008.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In Proceedings of the Second Workshop on Statistical Machine Translation - StatMT '07.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further Meta-Evaluation of Machine Translation. In Proceedings of the Third Workshop on Statistical Machine Translation - StatMT '08.

Dang, Hoa Trang, Jimmy Lin, and Diane Kelly. 2006. Overview of the TREC 2006 Question Answering Track. In Proceedings of TREC 2006.

Dang, Hoa Trang, Diane Kelly, and Jimmy Lin. 2007. Overview of the TREC 2007 Question Answering Track. In Proceedings of TREC 2007.

Dolan, William B., and Chris Brockett. 2005. Automatically Constructing a Corpus of Sentential Paraphrases. In Proceedings of the Third International Workshop on Paraphrasing (IWP2005).

Järvelin, Kalervo, Jaana Kekäläinen. 2002. Cumulated Gain-based Evaluation of IR Techniques. ACM Trans. Inf. Syst., Vol. 20, No. 4. (October 2002), pp. 422-446.

Kauchak, David, and Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation. In Proceedings of HLT-NAACL 2006.

Kekäläinen, Jaana. 2005. Binary and Graded Relevance in IR Evaluations – Comparison of the Effects on Ranking of IR Systems. Information Processing & Management, 41, 1019-1033.

Kok, Stanley and Chris Brockett. 2010. Hitting the Right Paraphrases in Good Time. In Proceedings of HLT-NAACL 2010.

Lin, Dekang, and Patrick Pantel. 2001. DIRT - Discovery of Inference Rules from Text. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '01 323-328.

Metzler, Donald, Eduard Hovy, and Chunliang Zhang. 2011. An Empirical Evaluation of Data-Driven Paraphrase Generation Techniques. In Proceedings of ACL-HLT 2011.

Metzler, Donald and Eduard Hovy. 2011. Mavuno: A Scalable and Effective Hadoop-Based Paraphrase Harvesting System. To appear in Proceedings of the KDD Workshop on Large-scale Data Mining: Theory and Applications (LDMTA 2011).

Miller, Geroge A. 1995. Wordnet: A Lexical Database for English. CACM, 38(11):39-41.

Padó, Sebastian, Michel Galley, Dan Jurafsky, and Christopher D. Manning. 2009. Robust Machine Translation Evaluation with Entailment Features. In Proceedings of ACL-IJCNLP '09.

Pasca, Marius and Pter Dienes. 2005. Aligning Needles in a Haystack: Paraphrase Acquisition Across the Web. In Processing of IJCNLP 2005.

Porter, Martin F. 1980. An Algorithm for Suffix Stripping, Program, 14(3): 130–137.

Przybocki, Mark, Kay Peterson, Sébastien Bronsart, and Gregory Sanders. 2009. The NIST 2008 Metrics for Machine Translation Challenge—Overview, Methodology, Metrics, and Results. Machine Translation, Volume 23 Issue 2-3.

Riezler, Stefan, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu Mittal, and Yi Liu. 2007. Statistical Machine Translation for Query Expansion in Answer Retrieval. In Proceedings of ACL 2007.

Szpektor, Idan and Ido Dagan. 2007. Learning Canonical Forms of Entailment Rules. In Proceedings of RANLP 2007.

Szpektor, Idan, Eyal Shnarch and Ido Dagan. 2007. Instance-based Evaluation of Entailment Rule Acquisition. In Proceedings of ACL 2007.

Szpektor, Idan and Ido Dagan. 2008. Learning Entailment Rules for Unary Templates. In Proceedings of COLING 2008.

Tratz, Stephen and Eduard Hovy. 2009. BEwT-E for TAC 2009's AESOP Task. In Proceedings of TAC-09. Gaithersburg, Maryland.

Voorhees, Ellen M., and Hoa Trang Dang. 2005. Overview of the TREC 2005 Question Answering Track. In Proceedings of TREC 2005.

Zhou, Liang, Chin-Yew Lin, and Eduard Hovy. 2006. Re-evaluating Machine Translation Results with Paraphrase Support. In Proceedings of EMNLP 2006.