Re-examining Machine Translation Metrics for Paraphrase Identification

Nitin Madnani **Joel Tetreault**
Educational Testing Service
Princeton, NJ, USA
{nmadnani,jtetreault}@ets.org

Martin Chodorow
Hunter College of CUNY
New York, NY, USA
martin.chodorow@hunter.cuny.edu

Abstract

We propose to re-examine the hypothesis that automated metrics developed for MT evaluation can prove useful for paraphrase identification in light of the significant work on the development of new MT metrics over the last 4 years. We show that a meta-classifier trained using nothing but recent MT metrics outperforms all previous paraphrase identification approaches on the Microsoft Research Paraphrase corpus. In addition, we apply our system to a second corpus developed for the task of plagiarism detection and obtain extremely positive results. Finally, we conduct extensive error analysis and uncover the top systematic sources of error for a paraphrase identification approach relying solely on MT metrics. We release both the new dataset and the error analysis annotations for use by the community.

1 Introduction

One of the most important reasons for the recent advances made in Statistical Machine Translation (SMT) has been the development of automated metrics for evaluation of translation quality. The goal of any such metric is to assess whether the translation hypothesis produced by a system is semantically equivalent to the source sentence that was translated. However, cross-lingual semantic equivalence is even harder to assess than monolingual, therefore, most MT metrics instead try to measure whether the hypothesis is semantically equivalent to a human-authored reference translation of the same source sentence. Using such automated metrics as proxies for human judgments can provide a quick assessment of system performance and allow for short feature and system development cycles, which are important for evaluating research ideas.

In the last 5 years, several shared tasks and competitions have led to the development of increasingly sophisticated metrics that go beyond the computation of n-gram overlaps (BLEU, NIST) or edit distances (TER, WER, PER etc.). Note that the task of an MT metric is essentially one of identifying whether the translation produced by a system is a paraphrase of the reference translation. Although the notion of using MT metrics for the task of paraphrase identification is not novel (Finch et al., 2005; Wan et al., 2006), it merits a re-examination in the light of the development of these novel MT metrics for which we can ask "How much better, if at all, do these newer metrics perform for the task of paraphrase identification?"

This paper describes such a re-examination. We employ 8 different MT metrics for identifying paraphrases across two different datasets - the well-known Microsoft Research paraphrase corpus (MSRP) (Dolan et al., 2004) and the plagiarism detection corpus (PAN) from the 2010 Uncovering Plagiarism, Authorship and Social Software Misuse shared task (Potthast et al., 2010). We include both MSRP and PAN in our study because they represent two very different sources of paraphrased text. The creation of MSRP relied on the massive redundancy of news articles on the web and extracted sentential paraphrases from different stories written about the same topic. In the case of PAN, humans consciously paraphrased existing text to generate new,

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 182–190,
Montréal, Canada, June 3-8, 2012.

plagiarized text.

In the next section, we discuss previous work on paraphrase identification. In §3, we describe our approach to paraphrase identification using MT metrics as features. Our approach yields impressive results – the current state of the art for MSRP and extremely positive for PAN. In the same section, we examine whether each metric's purported strength is demonstrated in our datasets. Next, in §4 we conduct an analysis of our system's misclassifications for both datasets and outline a taxonomy of errors that our system makes. We also look at annotation errors in the datasets themselves. We discuss the findings of the error analysis in §5 and conclude in §6.

2 Related Work & Our Contributions

Our goal in this paper is to examine the utility of a paraphrase identification approach that relies solely on MT evaluation metrics and no other evidence of semantic equivalence. Given this setup, the most relevant previous work is by Finch et al. (2005) which uses BLEU, NIST, WER and PER as features for a supervised classification approach using SVMs. In addition, they also incorporate part-of-speech information as well as the Jiang-Conrath WordNet-based lexical relatedness measure (Jiang and Conrath, 1997) into their edit distance calculations. In the first part of our paper, we present classification experiments with newer MT metrics not available in 2005, a worthwhile exercise in itself. However, we go much further in our study:

- We apply our approach to two different paraphrase datasets (MSRP and PAN) that were created via different processes.
- We attempt to find evidence of each metric's purported strength in both datasets.
- We conduct an extensive error analysis to find types of errors that a system based solely on MT metrics is likely to make. In addition, we also discover interesting paraphrase pairs in the datasets.
- We release our sentence-level PAN dataset (see §3.3.2) which contains more realistic examples of paraphrase and can prove useful to the community for future evaluations of paraphrase identification.

BLEU-based features were also employed by Wan et al. (2006) who use them in combination with several other features based on dependency relations and tree edit-distance inside an SVM.

There are several other supervised approaches to paraphrase identification that do not use any features based on MT metrics. Mihalcea et al. (2006) combine pointwise mutual information, latent semantic analysis and WordNet-based measures of word semantic similarity into an arbitrary text-to-text similarity metric. Qiu et al. (2006) build a framework that detects dissimilarities between sentences and makes its paraphrase judgment based on the significance of such dissimilarities. Kozareva and Montoyo (2006) use features based on LCS, skip n-grams and WordNet with a meta-classifier composed of SVM, k-nearest neighbor and maximum entropy classifiers. Islam and Inkpen (2007) measure semantic similarity using a corpus-based measure and a modified version of the Longest Common Subsequence (LCS) algorithm. Rus et al. (2008) take a graph-based approach originally developed for recognizing textual entailment and adapt it for paraphrase identification. Fernando and Stevenson (2008) construct a matrix of word similarities between all pairs of words in both sentences instead of relying only on the maximal similarities. Das and Smith (2009) use an explicit model of alignment between the corresponding parts of two paraphrastic sentences and combine it with a logistic regression classifier built from n-gram overlap features. Most recently, Socher et al. (2011) employ a joint model that incorporates the similarities between both single word features as well as multi-word phrases extracted from the parse trees of the two sentences.

We compare our results to those from all the approaches described in this section later in §3.4.

3 Classifying with MT Metrics

In this section, we first describe our overall approach to paraphrase identification that utilizes only MT metrics. We then discuss the actual MT metrics we used. Finally, we describe the datasets on which we evaluated our approach and present our results.

MSRP	They had published an advertisement on the Internet on June 10, offering the cargo for sale, he added.
	On June 10, the ship's owners had published an advertisement on the Internet, offering the explosives for sale.
	Security lights have also been installed and police have swept the grounds for booby traps.
	Security lights have also been installed on a barn near the front gate.
PAN	Dense fogs wrapped the mountains that shut in the little hamlet, but overhead the stars were shining in the near heaven.
	The hamlet is surrounded by mountains which is wrapped with dense fogs, though above it, near heaven, the stars were shining.
	In still other places, the strong winds carry soil over long distances to be mixed with other soils.
	In other places, where strong winds blow with frequent regularity, sharp soil grains are picked up by the air and hurled against the rocks, which, under this action, are carved into fantastic forms.

Table 1: Examples of paraphrases and non-paraphrases (in italics) from the MSRP and PAN corpora.

3.1 Classifier

Our best system utilized a classifier combination approach. We used a simple meta-classifier that uses the average of the unweighted probability estimates from the constituent classifiers to make its final decision. We used three constituent classifiers: Logistic regression, the SMO implementation of a support vector machine (Platt, 1999; Keerthi et al., 2001) and a lazy, instance-based classifier that extends the nearest neighbor algorithm (Aha et al., 1991). We used the WEKA machine learning toolkit to perform our experiments (Hall et al., 2009). [1]

3.2 MT metrics used

1. **BLEU** (Papineni et al., 2002) is the most commonly used metric for MT evaluation. It is computed as the amount of n-gram overlap—for different values of n—between the system output and the reference translation, tempered by a penalty for translations that might be too short. BLEU relies on exact matching and has no concept of synonymy or paraphrasing. We use BLEU1 through BLEU4 as 4 different features for our classifier (hereafter BLEU(1-4)).

2. **NIST** (Doddington, 2002) is a variant of BLEU that uses the arithmetic mean of n-gram overlaps, rather than the geometric mean. It also weights each n-gram according to its informativeness as indicated by its frequency. We use NIST1 through NIST5 as 5 different features for our classifier (hereafter NIST(1-5)).

3. **TER** (Snover et al., 2006) is defined as the number of edits needed to "fix" the translation output so that it matches the reference. TER differs from WER in that it includes a heuristic algorithm to deal with shifts in addition to insertions, deletions and substitutions.

4. **TERp** (TER-Plus) (Snover et al., 2009) builds upon the core TER algorithm by providing additional edit operations based on stemming, synonymy and paraphrase.

5. **METEOR** (Denkowski and Lavie, 2010) uses a combination of both precision and recall unlike BLEU which focuses on precision. Furthermore, it incorporates stemming, synonymy (via WordNet) and paraphrase (via a lookup table).

6. **SEPIA** (Habash and El Kholy, 2008) is a syntactically-aware metric designed to focus on

[1] These constituent classifiers were chosen since they were the top 3 performers in 5-fold cross-validation experiments conducted on both MSRP and PAN training sets. The meta-classifier was chosen similarly once the constituent classifiers had been chosen.

structural n-grams with long surface spans that cannot be captured efficiently with surface n-gram metrics. Like BLEU, it is a precision-based metric and requires a length penalty to minimize the effects of length.

7. **BADGER** (Parker, 2008) is a language independent metric based on compression and information theory. It computes a compression distance between the two sentences that utilizes the Burrows Wheeler Transformation (BWT). The BWT enables taking into account common sentence contexts with no limit on the size of these contexts.

8. **MAXSIM** (Chan and Ng, 2008) treats the problem as one of bipartite graph matching and maps each word in one sentence to at most one word in the other sentence. It allows the use of arbitrary similarity functions between words.[2]

Our choice of metrics was based on their popularity in the MT community, their performance in open competitions such as the NIST MetricsMATR challenge (NIST, 2008) and the WMT shared evaluation task (Callison-Burch et al., 2010), their availability, and their relative complementarity.

3.3 Datasets

In this section, we describe the two datasets that we used to evaluate our approach.

3.3.1 Microsoft Research Paraphrase Corpus

The MSRP corpus was created by mining news articles on the web for topically similar articles and then extracting potential sentential paraphrases using a set of heuristics. Extracted pairs were then shown to two human judges with disagreements handled by a third adjudicator. The kappa was reported as 0.62, which indicates moderate to high agreement. We used the pre-stipulated train-test splits (4,076 sentence pairs in training and 1,725 in test) to train and test our classifier.

[2]We also experimented with TESLA—a variant of MAXSIM that performs better for MT evaluation—in our preliminary experiments However, both MAXSIM and TESLA performed almost identically in our cross-validation experiments. Therefore, we only retained MAXSIM in our final experiment since it was significantly faster to run than the version of TESLA we had.

3.3.2 Plagiarism Detection Corpus (PAN)

We wanted to evaluate our approach on a set of paraphrases where the semantic similarity was not simply an accidental by-product of topical similarity but rather consciously generated. We used the test collection from the PAN 2010 plagiarism detection competition. This dataset consists of 41,233 text documents from Project Gutenberg in which 94,202 cases of plagiarism have been inserted. The plagiarism was created either by using an algorithm or by explicitly asking Turkers to paraphrase passages from the original text. We focus only on the human-created plagiarism instances.

Note also that although the original PAN dataset has been used in plagiarism detection shared tasks, those tasks are generally formulated differently in that the goal is to find all potentially plagiarized passages in a given set of documents along with the corresponding source passages from other documents. In this paper, we wanted to focus on the task of identifying whether two given *sentences* can be considered paraphrases.

To generate a sentence-level PAN dataset, we wrote a heuristic alignment algorithm to find corresponding pairs of sentences within a passage pair linked by the plagiarism relationship. The alignment algorithm utilized only bag-of-words overlap and length ratios and no MT metrics. For our negative evidence, we sampled sentences from the same document and extracted sentence pairs that have at least 4 content words in common. We then sampled randomly from both the positive and negative evidence files to create a training set of 10,000 sentence pairs and a test set of 3,000 sentence pairs.

Table 1 shows examples of paraphrastic and non-paraphrastic sentence pairs from both the MSRP and PAN datasets.

3.4 Results

Before presenting the results of experiments that used multiple metrics as features, we wanted to determine how well each metric performs on its own when used for paraphrase identification. Table 2 shows the classification results on both the MSRP and PAN datasets using each metric as the only feature. Although previously explored metrics such as BLEU and NIST perform reasonably well, they are

	MSRP		PAN	
Metric	**Acc.**	**F1**	**Acc.**	**F1**
MAXSIM	67.2	79.4	84.7	83.4
BADGER	67.6	79.9	88.5	87.9
SEPIA	68.1	79.8	87.7	86.8
TER	69.9	80.9	85.7	83.8
BLEU(1-4)	72.3	80.9	87.9	87.1
NIST(1-5)	72.8	81.2	88.2	87.3
METEOR	73.1	81.0	89.5	88.9
TERp	74.3	81.8	91.2	90.9

Table 2: Classification results for MSRP and PAN with individual metrics as features. Entries are sorted by accuracies on MSRP.

clearly outperformed by some of the more robust metrics such as TERp and METEOR.

Table 3 shows the results of our experiments employing multiple metrics as features, for both MSRP and PAN. The final row in the table shows the results of our best system. The remaining rows of this table show the top performing metrics for both datasets; we treat BLEU, NIST and TER as our baseline metrics since they are not new and are not the primary focus of our investigation. In terms of novel metrics, we find that the top 3 metrics for both datasets were TERp, METEOR and BADGER respectively as shown. Combining all 8 metrics led to the best performance for MSRP but showed no performance increase for PAN.

	MSRP		PAN	
Features	**Acc.**	**F1**	**Acc.**	**F1**
Base Metrics	74.1	81.5	88.6	87.8
+ TERp	75.6	82.5	91.5	91.2
+ METEOR	76.6	83.2	92.0	91.8
+ BADGER	77.0	83.7	**92.3**	**92.1**
+ Others	**77.4**	**84.1**	**92.3**	**92.1**

Table 3: The top 3 performing MT metrics for both MSRP and PAN datasets as identified by ablation studies. BLEU(1-4), NIST(1-5) and TER were used as the 10 base features in the classifiers.

Our results for the PAN dataset are much better than those for MSRP since:

(a) It is likely that our negative evidence is too easy for most MT metrics.

(b) Many plagiarized pairs are linked simply via lexical synonymy which can be easily captured by metrics like METEOR and TERp, e.g., the sentence "*Young's main contention is that in literature genius must make rules for itself, and that imitation is suicidal*" is simply plagiarized as "*Young's major argument is that in literature intellect must make rules for itself, and that replication is dangerous.*" However, the PAN corpus does contains some very challenging and interesting examples of paraphrases—even more so than MSRP—which we describe in §4.

Finally, Table 4 shows that the results from our best system are the best ever reported on the MSRP test set when compared to all previously published work. Furthermore, the single best performing metric (TERp)—also shown in the table—outperforms, by itself, many previous approaches utilizing multiple, complex features.

Model	**Acc.**	**F1**
All Paraphrase Baseline	66.5	79.9
(Mihalcea et al., 2006)	70.3	81.3
(Rus et al., 2008)	70.6	80.5
(Qiu et al., 2006)	72.0	81.6
(Islam and Inkpen, 2007)	72.6	81.3
(Fernando and Stevenson, 2008)	74.1	82.4
TERp	74.3	81.8
(Finch et al., 2005)	75.0	82.7
(Wan et al., 2006)	75.6	83.0
(Das and Smith, 2009)	76.1	82.7
(Kozareva and Montoyo, 2006)	76.6	79.6
(Socher et al., 2011)	76.8	83.6
Best MT Metrics	**77.4**	**84.1**

Table 4: Comparing the accuracy and F-score for the single best performing MT metric TERp (in gray) as well as the best metric combination system (in gray and bold) with previously reported results on the MSRP test set ($N = 1,752$). Entries are sorted by accuracy.

3.5 Metric Contributions

In addition to quantitative results, we also wanted to highlight specific examples from our datasets that can demonstrate the strength of the new metrics over simple n-gram overlap and edit-distance based metrics. Below we present examples for the 4 best

metrics across both datasets:

- **TERp** uses stemming and phrasal paraphrase recognition to accurately classify the sentence pair "*For the weekend, the top 12 movies grossed $157.1 million, up 52 percent from the same weekend a year earlier.*" and "*The overall box office soared, with the top 12 movies grossing $157.1 million, up 52 percent from a year ago.*" from MSRP as paraphrases.

- **METEOR** uses synonymy and stemming to accurately classify the sentence pair "*Her letters at this time exhibited the two extremes of feeling in a marked degree.*" and "*Her letters at this time showed two extremes of feelings.*" from PAN as plagiarized.

- **BADGER** uses unsupervised contextual similarity detection to accurately classify the sentence pair "*Otherwise they were false or mistaken reactions*" and "*Otherwise, were false or wrong responses*" from PAN as plagiarized.

- **SEPIA** uses structural n-grams via dependency trees to accurately classify the sentence pair "*At his sentencing, Avants had tubes in his nose and a portable oxygen tank beside him.*" and "*Avants, wearing a light brown jumpsuit, had tubes in his nose and a portable oxygen tank beside him.*" from MSRP as paraphrases.

4 Error Analysis

In this section, we conduct an analysis of the misclassifications that our system makes on both datasets. Our analyses consisted of finding the sentences pairs from the test set for each dataset which *none* of our systems (not just the best one) ever classified correctly and inspecting a random sample of 100 of these. This inspection yields not only the top sources of error for an approach that relies solely on MT metrics but also uncovers sources of annotation errors in both datasets themselves.

4.1 MSRP

In their paper describing the creation of the MSRP corpus, Dolan et al. (2004) clearly state that "the degree of mismatch allowed before the pair was judged non-equivalent was left to the discretion of the individual rater" and that "many of the 33% of sentence pairs judged to be not equivalent still overlap significantly in information content and even wording". We found evidence that the raters were not always consistent in applying the annotation guidelines. For example, in some cases the lack of attribution for a quotation led the raters to label a pair as paraphrastic whereas in other cases it did not. For example, the pair "*These are real crimes that hurt a lot of people.*" and "*'These are real crimes that disrupt the lives of real people,' Smith said.*" was not marked as paraphrastic. Furthermore, even though the guidelines instruct the raters to "treat anaphors and their full forms as equivalent, regardless of how great the disparity in length or lexical content between the two sentences", we found pairs of sentences marked as non-paraphrastic which only differed in anaphora. However, the primary goal of this analysis is to find sources of errors in an MT-metric driven approach and below we present the top 5 such sources:

1. **Misleading Lexical Overlap**. Non-paraphrastic pairs where there is large lexical overlap of secondary material between the two sentences but the primary semantic content is different. For example, "*Gyorgy Heizler, head of the local disaster unit, said the coach had been carrying 38 passengers.*" and "*The head of the local disaster unit, Gyorgy Heizler, said the coach driver had failed to heed red stop lights.*".

2. **Lack of World Knowledge**. Paraphrastic pairs that require world knowledge. For example, "*Security experts are warning that a new mass-mailing worm is spreading widely across the Internet, sometimes posing as e-mail from the Microsoft founder.*" and "*A new worm has been spreading rapidly across the Internet, sometimes pretending to be an e-mail from Microsoft Chairman Bill Gates, antivirus vendors said Monday.*".

3. **Tricky Phrasal Paraphrases**. Paraphras-

tic pairs that contain domain-dependent semantic alternations. For example, "*The leading actress nod went to energetic newcomer Marissa Jaret Winokur as Edna's daughter Tracy.*" and "*Marissa Jaret Winokur, as Tracy, won for best actress in a musical.*".

4. **Date, Time and Currency Differences**. Paraphrastic pairs that contain different temporal or currency references. These references were normalized to generic tokens (e.g., `$NUMBER`) before being shown to MSRP raters but are retained in the released dataset. For example, "*Expenses are expected to be approximately $2.3 billion, at the high end of the previous expectation of $2.2-to-$2.3 billion.*" and "*Spending on research and development is expected to be $4.4 billion for the year, compared with the previous expectation of $4.3 billion.*".

5. **Anaphoric References**. Paraphrastic pairs wherein one member of the pair contains anaphora and the other doesn't (these are considered paraphrases according to MSRP guidelines). For example, "*They certainly reveal a very close relationship between Boeing and senior Washington officials.*" and "*The e-mails reveal the close relationship between Boeing and the Air Force.*".

Note that most misclassified sentence pairs can be categorized into more than one of the above categories.

4.2 PAN

For the PAN corpus, the only real source of error in the dataset itself was the sentence alignment algorithm. There were many sentence pairs that were erroneously linked as paraphrases. Leaving aside such pairs, the 3 largest sources of error for our MT-metric based approach were:

1. **Complex Sentential Paraphrases**. By far, most of the misclassified pairs were paraphrastic pairs that could be categorized as real world plagiarism, i.e., where the plagiarizer copies the idea from the source but makes several complex transformations, e.g., sentence splitting, structural paraphrasing etc. so as to render an MT-metric based approach powerless. For example, consider the pair "*The school bears the honored name of one who, in the long years of the anti-slavery agitation, was known as an uncompromising friend of human freedom.*" and "*The school is named after a man who defended the right of all men and women to be free, all through the years when people campaigned against slavery.*" Another interesting example is the pair "*The most unpromising weakly-looking creatures sometimes live to ninety while strong robust men are carried off in their prime.*" and "*Sometimes the strong personalities live shorter than those who are unexpected.*".

2. **Misleading Lexical Overlap**. Similar to MSRP. For example, "*Here was the second period of Hebraic influence, an influence wholly moral and religious.*" and "*This was the second period of Hellenic influence, an influence wholly intellectual and artistic.*".

3. **Typographical and Spelling Errors**. Paraphrastic pairs where the Turkers creating the plagiarism also introduced other typos and spelling errors. For example, "*The boat then had on board over 1,000 souls in all*" and "*1000 people where on board at that tim*".

5 Discussion

The misses due to "Date, Time, and Currency Differences" are really just the result of an artifact in the testing. It is possible that an MT metrics based approach could accurately predict these cases if the references to dates etc. were replaced with generic tokens as was done for the human raters. In a similar vein, some of the misses that are due to a lack of world knowledge might become hits if a named entity recognizer could discover that "*Microsoft founder*" is the same as "*Microsoft Chairman*". Similarly, some of the cases of anaphoric reference might be recognized with an anaphora resolution system. And the problem of misspelling in PAN could be remedied with automatic spelling correction. Therefore, it is possible to improve the MT metrics based approach further by utilizing certain NLP systems as pre-processing modules for the text.

The only error category in MSRP and PAN

that caused false positives was "Misleading Lexical Overlap". Here, the take-away message is that not every part of a sentence is equally important for recognizing semantic equivalence or non-equivalence. In a sentence that describes what someone communicated, the content of what was said is crucial. For example, despite lexical matches everywhere else, the mismatch of "*the coach had been carrying 38 passengers*" and "*the driver had failed to heed the red stop lights*" disqualifies the respective sentences from being paraphrases. Along the same line, differences in proper names and their variants should receive more weight than other words. A sentence about "*Hebraic influence*" on a period in history is not the same as a sentence which matches in every other way but is instead about "*Hellenic influence*". These sentences represent a bigger challenge for an approach based solely on MT metrics. Given enough pairs of "near-miss" non-paraphrases, our system might be able to figure this out, but this would require a large amount of annotated data.

6 Conclusions

In this paper, we re-examined the idea that automatic metrics used for evaluating translation quality can perform well explicitly for the task of paraphrase recognition. The goal of our paper was to determine whether approaches developed for the related but different task of MT evaluation can be as competitive as approaches developed specifically for the task of paraphrase identification. While we do treat the metrics as black boxes to an extent, we explicitly chose metrics that were high performing but also complementary in nature.

Specifically, our re-examination focused on the more sophisticated MT metrics of the last few years that claim to go beyond simple n-gram overlap and edit distance. We found that a meta-classifier trained using only MT metrics outperforms all previous approaches for the MSRP corpus. Unlike previous studies, we also applied our approach to a new plagiarism dataset and obtained extremely positive results. We examined both datasets not only to find pairs that demonstrated the strength of each metric but also to conduct an error analysis to discover the top sources of errors that an MT metric based approach is susceptible to. Finally, we discovered that using the TERp metric by itself provides fairly good performance and can outperform many other supervised classification approaches utilizing multiple, complex features.

We also have two specific suggestions that we believe can benefit the community. First, we believe that binary indicators of semantic equivalence are not ideal and a continuous value between 0 and 1 indicating the degree to which two pairs are paraphrastic is more suitable for most approaches. However, rather than asking annotators to rate pairs on a scale, a better idea might be to show the sentence pairs to a large number of Turkers ($\geq$ 20) on Amazon Mechanical Turk and ask them to classify it as either a paraphrase or a non-paraphrase. A simple estimate of the degree of semantic equivalence of the pair is simply the proportion of the Turkers who classified the pair as paraphrastic. An example of such an approach, as applied to the task of grammatical error detection, can be found in (Madnani et al., 2011).[3] Second, we believe that the PAN corpus—with Turker simulated plagiarism—contains much more realistic examples of paraphrase and should be incorporated into future evaluations of paraphrase identification. In order to encourage this, we are releasing our PAN dataset containing 13,000 sentence pairs.

We are also releasing our error analysis data (100 pairs for MSRP and 100 pairs for PAN) since they might prove useful to other researchers as well. Note that the annotations for this analysis were produced by the authors themselves and, although, they attempted to accurately identify all error categories for most sentence pairs, it is possible that the errors in some sentence pairs were not comprehensively identified.[4]

Acknowledgments

We would like to thank Aoife Cahill, Michael Heilman and the three anonymous reviewers for their useful comments and suggestions.

[3] A good approximation is to use an ordinal scale for the human judgments as in the Semantic Textual Similarity task of SemEval 2012. See `http://www.cs.york.ac.uk/semeval-2012/task6/` for more details.

[4] The data is available at `http://bit.ly/mt-para`.

References

D. W. Aha, D. Kibler, and M. K. Albert. 1991. Instance-based learning algorithms. *Mach. Learn.*, 6:37–66.

C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, and O. Zaidan, editors. 2010. *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*.

Y. S. Chan and H. T. Ng. 2008. MAXSIM: A maximum similarity metric for machine translation evaluation. In *Proceedings of ACL-HLT*, pages 55–62.

D. Das and N.A. Smith. 2009. Paraphrase Identification as Probabilistic Quasi-synchronous Recognition. In *Proceedings of ACL-IJCNLP*, pages 468–476.

M. Denkowski and M. Lavie. 2010. Extending the METEOR Machine Translation Metric to the Phrase Level. In *Proceedings of NAACL*.

G. Doddington. 2002. Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics. In *Proceedings of HLT*, pages 138–145.

B. Dolan, C. Quirk, and C. Brockett. 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In *Proceedings of COLING*, pages 350–356, Geneva, Switzerland.

S. Fernando and M. Stevenson. 2008. A Semantic Similarity Approach to Paraphrase Detection. In *Proceedings of the Computational Linguistics UK (CLUK) 11th Annual Research Colloquium.*

A. Finch, Y.S. Hwang, and E. Sumita. 2005. Using Machine Translation Evaluation Techniques to Determine Sentence-level Semantic Equivalence. In *Proceedings of the Third International Workshop on Paraphrasing*, pages 17–24.

N. Habash and A. El Kholy. 2008. SEPIA: Surface Span Extension to Syntactic Dependency Precision-based MT Evaluation. In *Proceedings of the Workshop on Metrics for Machine Translation at AMTA*.

M. Hall, E. Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11.

A. Islam and D. Inkpen. 2007. Semantic Similarity of Short Texts. In *Proceedings of RANLP*, pages 291–297.

J. J. Jiang and D. W. Conrath. 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *CoRR*, cmp-lg/9709008.

S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy. 2001. Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Comput.*, 13(3):637–649.

Z. Kozareva and A. Montoyo. 2006. Paraphrase Identification on the Basis of Supervised Machine Learning Techniques. In *Proceedings of FinTAL*, pages 524–233.

N. Madnani, J. Tetreault, M. Chodorow, and A. Rozovskaya. 2011. They Can Help: Using Crowdsourcing to Improve the Evaluation of Grammatical Error Detection Systems. In *Proceedings of ACL (Short Papers)*.

R. Mihalcea, C. Corley, and C. Strapparava. 2006. Corpus-based and Knowledge-based Measures Of Text Semantic Similarity. In *Proceedings of AAAI*, pages 775–780.

NIST. 2008. NIST MetricsMATR Challenge. Information Access Division. `http://www.itl.nist.gov/iad/mig/tests/metricsmatr/`.

K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*.

S. Parker. 2008. BADGER: A New Machine Translation Metric. In *Proceedings of the Workshop on Metrics for Machine Translation at AMTA*.

John C. Platt. 1999. Advances in kernel methods. chapter Fast Training of Support Vector Machines using Sequential Minimal Optimization, pages 185–208. MIT Press.

M. Potthast, B. Stein, A. Barrón-Cedeño, and P. Rosso. 2010. An Evaluation Framework for Plagiarism Detection. In *Proceedings of COLING*, pages 997–1005.

L. Qiu, M. Y. Kan, and T. S. Chua. 2006. Paraphrase Recognition via Dissimilarity Significance Classification. In *Proceedings of the EMNLP*, pages 18–26.

V. Rus, P.M. McCarthy, M.C. Lintean, D.S. McNamara, and A.C. Graesser. 2008. Paraphrase Identification with Lexico-Syntactic Graph Subsumption. In *Proceedings of FLAIRS*, pages 201–206.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA*.

M. Snover, N. Madnani, B. Dorr, and R. Schwartz. 2009. TER-Plus: Paraphrase, Semantic, and Alignment Enhancements to Translation Edit Rate. *Machine Translation*, 23(2–3):117–127.

R. Socher, E.H. Huang, J. Pennington, A.Y. Ng, and C.D. Manning. 2011. Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. In *Advances in Neural Information Processing Systems 24 (NIPS)*.

S. Wan, R. Dras, M. Dale, and C. Paris. 2006. Using Dependency-based Features to Take the "para-farce" Out of Paraphrase. In *Proceedings of the Australasian Language Technology Workshop (ALTW)*, pages 131–138.