Paraphrasing for Style

Wei Xu[1] *Alan Ritter*[2] *William B. Dolan*[3] *Ralph Grishman*[1] *Colin Cherry*[4]
(1) New York University
(2) University of Washington
(3) Microsoft Research
(4) National Research Council, Canada
xuwei@cims.nyu.edu, aritter@cs.washington.edu, billdol@microsoft.com,
grishman@cs.nyu.edu, Colin.Cherry@nrc-cnrc.gc.ca

ABSTRACT
We present initial investigation into the task of paraphrasing language while targeting a particular writing style. The plays of William Shakespeare and their modern translations are used as a testbed for evaluating paraphrase systems targeting a specific style of writing. We show that even with a relatively small amount of parallel training data, it is possible to learn paraphrase models which capture stylistic phenomena, and these models outperform baselines based on dictionaries and out-of-domain parallel text. In addition we present an initial investigation into automatic evaluation metrics for paraphrasing writing style. To the best of our knowledge this is the first work to investigate the task of paraphrasing text with the goal of targeting a specific style of writing.

KEYWORDS: Paraphrase, Writing Style.

1 Introduction

The same meaning can be expressed or *paraphrased* in many different ways; automatically detecting or generating different expressions with the same meaning is fundamental to many natural language understanding tasks(Giampiccolo et al., 2007), so much previous work has investigated methods for automatic paraphrasing(Madnani and Dorr, 2010).

Paraphrases can differ along many dimensions, including utterance length, diction level, and speech register. There is a significant literature in sentence compression aimed at modeling the first of these, length: producing meaning-preserving alternations that reduce the length of the input string (Chandrasekar et al., 1996; Vanderwende et al., 2007; Clarke and Lapata, 2008; Cohn and Lapata, 2009; Yatskar et al., 2010). However, we know of no previous work aimed at modeling meaning-preserving transformations that systematically transform the register or style of an input string. Can we learn to reliably map from one form of language to another, transforming formal prose into a more colloquial form, or a casual email into a more formal equivalent?

Systems capable of paraphrasing text targeting a specific writing style could be useful for a variety of applications. For example, they could:

1. Help authors of technical documents to adhere to appropriate stylistic guidelines.

2. Enable non-experts to better consume technical information, for example by translating legalese or medical jargon into nontechnical English.

3. Benefit educational applications, allowing students to:

   (a) Access *modern English* versions of works by authors they are studying.

   (b) Experiment with writing in the style of an author they are studying.

In this paper, we investigate the task of automatic paraphrasing while targeting a particular writing style, focusing specifically on the style of Early Modern English employed by William Shakespeare. We explored several different methods, all of which rely on techniques from phrase-based MT, but which were trained on different types of parallel monolingual data. The first system was trained on the text of Shakespeare's plays, along with parallel modern English "translations" that were written to help students better understand Shakespeare's work. We also developed several baselines which do not make use of this parallel text and instead rely on manually compiled dictionaries of expressions commonly found in Shakespearean English, or existing corpora of out-of-domain parallel monolingual text.

We evaluate these models both through human judgments and standard evaluation metrics from the Machine Translation (MT) and Paraphrase literature, however no previous work has investigated the ability of automatic evaluation metrics to capture the notion of writing style. We show that previously proposed metrics do not provide the complete picture of a system's performance when the task is to generate paraphrases targeting a specific style of writing. We therefore propose three new metrics for evaluating paraphrases targeting a specific style, and show that these metrics correlate well with human judgments.

corpus	initial size	aligned size	No-Change BLEU
http://nfs.sparknotes.com	31,718	21,079	24.67
http://enotes.com	13,640	10,365	52.30

Table 1: Parallel corpora generated form modern translations of Shakespeare's plays

2 Shakespearean Paraphrasing

We use Shakespeare's plays as a testbed for the task of paraphrasing while targeting a specific writing style. Because these plays are some of the most highly-regarded examples of English literature and are written in a style that is now 400 years out of date, many linguistic resources are available to help modern readers pick their way through these Elizabethan texts. Among these are "translations" of the plays into colloquial English, as well as dictionaries that provide modern equivalents for archaic words and phrases.

We compare 3 different stylistic paraphrase systems targeting Shakespearean English which rely on different types of linguistic resources. One leverages parallel "translations", another exploits dictionary resources, and a third relies on modern, out-of-domain monolingual parallel data and an in-domain language model.

2.1 Modern Translations

Access to parallel text in the target style allows us to train statistical models that generate paraphrases, and also perform automatic evaluation of semantic adequacy using BLEU, which requires availability of reference translations. For this purpose we scraped modern translations of 17 Shakespeare plays from `http://nfs.sparknotes.com`, and additional translations of 8 of these plays from `http://enotes.com`.

After tokenizing and lowercasing, the plays were sentence aligned (Moore, 2002), producing 21,079 alignments from the 31,718 sentence pairs in the Sparknotes data, and 10,365 sentence pairs from the 13,640 original pairs in the Enotes data. The modern translations from the two sources are qualitatively quite different. The Sparknotes paraphrases tend to differ significantly from the original text, whereas the Enotes translations are much more conservative, making fewer changes. To illustrate these differences empirically and provide an initial paraphrase baseline, we computed BLEU scores of the modern translations against Shakespeare's original text; the Sparknotes paraphrases yield a BLEU score of 24.67, whereas the Enotes paraphrases produce a much higher BLEU of 52.30 reflecting their strong similarity to the original texts. These results are summarized in Table 1.

To generate paraphrases, we applied a typical phrase-based statistical MT pipeline, performing word alignment on the data described in table 1 using GIZA++ (Och and Ney, 2003), then extracting phrase pairs and performing decoding using Moses (Koehn et al., 2007).

For evaluation purposes, the parallel text of one play, Romeo and Juliet, was held out of the training corpus for this system and the baseline systems described in the following section.

2.2 Baselines

Phrase-based translation has been demonstrated as an effective approach to paraphrasing (Quirk et al., 2004; Chen and Dolan, 2011). However, this approach does require the existence

target	source	target	source
ABATE	shorten	AYE	always
CAUTEL	deceit	GLASS	mirror
SUP	have supper	VOICE	vote

Table 2: Example dictionary entries

Smoothed Probability Estimate	target	source
0.0000790755	PERCHANCE	maybe
0.00003691883	PERADVENTURE	maybe
0.00007524298	HAPLY	maybe
0.00007141065	HAPPILY	maybe
total 0.00026264791		

Table 3: Example ngram probabilities in target language

of parallel corpora of aligned phrases and sentences, resources which may not be available for many writing styles that we might wish to target. For this reason we were motivated to investigate alternative approaches in order to help quantify how critical this type of parallel data is for the task of stylistic paraphrasing.

2.2.1 Dictionary Based Paraphrase

Several dictionaries of stylistically representative words of Shakespearean English and their modern equivalents are available on the web. These dictionaries can be used to define a translation model which can be used in combination with a language model as in standard phrase-based MT.

To build a phrase table, we scraped a set of 68,709 phrase/word pairs from `http://www.shakespeareswords.com/`; example dictionary entries are presented in table 2. As described in (Koehn and Knight, 2000), we estimate phrase translation probabilities based on the frequencies of the translation words/phrases in the target language (Shakespearean English). For instance, if we look at the modern English word *maybe*, our dictionary lists 4 possible Shakespearean translations. We obtained the probabilities for each translation according to the n-gram back-off model built from 36 of Shakespeare's plays using the SRILM toolkit (Stolcke, 2002), normalizing the probabilities for each source phrase, for example $p(\text{PERCHANGE}|\text{maybe}) = \frac{0.0000790755}{0.00026264791} = 0.30107035689$. An example is presented in Table 3. This method allows us to estimate reasonable translation probabilities for use in a phrase table, which is used in combination with a language model built from the 36 plays, which are then fed into the Moses decoder (Koehn et al., 2007).

2.2.2 Out of Domain Monolingual Parallel Data

As a final baseline we consider a paraphrase system which is trained on out-of-domain data gathered by asking users of Amazon's Mechanical Turk Service (Snow et al., 2008) to caption the action in short video segments (Chen and Dolan, 2011). We combined a phrase table extracted from this modern, out of domain parallel text, with an in-domain language model consisting of Shakespeare's 36 plays, applying the Moses decoder (Koehn et al., 2007) to find the best

paraphrases. Although this monolingual parallel data does not include text in the target writing style, the in-domain language model does bias the system's output towards Shakespeare's style of writing. We found that performing Minimum Error Rate Training (Och, 2003) using a small set of held out parallel text from Romeo and Juliet was necessary in order to tune the video corpus baseline to generate reasonable paraphrases.

2.3 Comparison Using Existing Automatic Evaluation Metrics

Figure 1 compares a variety of systems targeting Shakespearean English using the previously proposed BLEU (Papineni et al., 2002) and PINC (Chen and Dolan, 2011) automatic evaluation metrics which have been demonstrated to correlate with human judgments on semantic adequacy and lexical dissimilarity with the input. A description of each of the systems compared in this experiment is presented in Table 4. As mentioned in §2.1, the Enotes paraphrases diverge little from the original text, resulting in a BLEU score of 52.3 when compared directly to the original lines from Shakespeare's plays. Because our goal is to produce paraphrases which make more dramatic stylistic changes to the input, in the remainder of this paper, we focus on the Sparknotes data for evaluation.

2.3.1 Discussion

Two main trends are evident in Figure 1. First, notice that all of the systems trained using parallel text achieve higher BLEU scores than the unmodified modern translations. While the dictionary baseline achieves a competitive PINC score, indicating it is making a significant number of changes to the input, its BLEU is lower than that of the modern translations. Secondly, it seems apparent that the systems whose parameters are tuned using Minimum Error Rate Training tend to be more conservative, making fewer changes to the input and thus achieving lower PINC scores, while not improving BLEU on the test data. Finally we note that using the larger target language model seems to yield a slight improvement in BLEU score.

2.4 Examples

Example paraphrases of lines from *Romeo and Juliet* and several Hollywood movies, generated by the top performing system according to BLEU and PINC, are presented in table 5.

3 Human Evaluation

Figure 1 provides some insight into the performance of the various systems, but it is initially unclear how well the BLEU and PINC automatic evaluation metrics perform when applied to paraphrases that target a specific style of writing. BLEU and PINC have previously been shown to have high correlation with human judgments of semantic adequacy and lexical dissimilarity of paraphrase candidates, but the implications of this for the more specialized task of stylistic paraphrasing are unclear.

While BLEU is typically used to measure semantic adequacy, it seems reasonable to assume that it could also be useful for measuring stylistic alternations, since utterances are more likely to contain overlapping ngrams if they are both semantically and stylistically similar. What BLEU cannot tell us, however is what portion of its improvements are due to stylistic similarity or semantic equivalence. For this reason, we were motivated perform an evaluation based on human judgments of semantic adequacy, lexical dissimilarity and stylistic similarity.

For this purpose, we randomly sampled 100 lines from *Romeo and Juliet*, then two of the authors

System	Description
16and7plays_36LM	Phrase table extracted from all 16 Sparknotes plays (other than R&J) and language model built from all 36 of Shakespeare's plays, again excluding R&J. Uses default Moses parameters.
16and7plays_36LM_MERT	Same as 16and7plays_36LM except parameters are tuned using Minimum Error Rate Training (Och, 2003).
16and7plays_16LM	Phrase table is built from both Sparknotes and Enotes data, and Language model is built from the 16 plays with modern translations
16and7plays_16LM_MERT	Same as 16and7plays_16LM except parameters are tuned using MERT.
16plays_36LM	Only Sparknotes modern translations are used. All 36 plays are used to train Shakespearean language model.
16plays_36LM_MERT	Same as 16plays_36LM except parameters are tuned using MERT.
video_corpus_baseline	Paraphrase system combining out of domain parallel text (Chen and Dolan, 2011) with an in-domain language model. Described in detail in §2.2.2.
modern (no change)	No changes are made to the input, modern translations are left unchanged.
Dictionary	Dictionary baseline described in §2.2.1

Table 4: Descriptions of various systems for Shakespearean paraphrase. *Romeo and Juliet* is held out for testing.

annotated each sentence and its Shakespearean translation to indicate semantic adequacy, lexical dissimilarity, stylistic similarity, and overall quality. The aggregate results of the human evaluation are displayed in Figure 2. Agreement between annotators measured using Pearson's ρ is displayed in Table 6.

Based on the human evaluation, it appears that the baseline combining paraphrases collected from Mechanical Turk (Chen and Dolan, 2011) with a Shakespearean language model has the highest semantic adequacy, yet this approach is also fairly conservative in that it makes few changes to the input.

The dictionary baseline, and the paraphrase system trained on parallel modern translations are roughly comparable in terms of the number of changes made to the input, but the system trained on modern translations achieves higher semantic adequacy, while also being rated higher on style and overall.

These results are roughly in line with the automatic metrics presented in Figure 1. However we also see several important trends which are not apparent from the automatic evaluation. Although the video baseline achieves the highest semantic adequacy in the human evaluation,

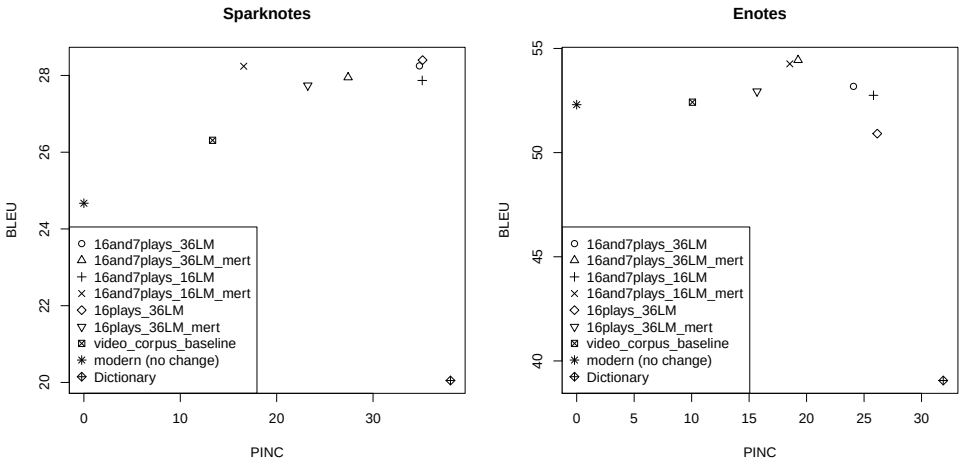

Figure 1: Various Shakespearean paraphrase systems compared using BLEU and PINC. A brief description of each system is presented in table 4.

its BLEU score is significantly lower than 16plays_36LM on the Sparknotes data.[1] It would appear that in this case BLEU is conflating semantic adequacy with writing style. Although the paraphrases produced by the video baseline have high semantic adequacy, their style tends to differ substantially from the reference translations resulting in fewer ngram matches, and thus a lower BLEU score.

4 Automatic Metrics Evaluating Writing Style

While PINC and BLEU do seem useful for automatically evaluating stylistic paraphrases, BLEU tends to conflate the notions of semantic adequacy with writing style. When comparing various systems using automatic metrics, it would seem useful to separate the effects caused by these two distinct criteria. We would like our automatic evaluation metrics to distinguish between a system which generates perfect paraphrases which do not match the target style of writing versus a system which generates sentences in the correct style, but which convey different meaning.

To help address this issue we propose three new automatic evaluation metrics whose goal is to measure the degree to which automatic paraphrases match the target style. These metrics assume existence of large corpora in both the source and target style, but do not require access to any parallel text, or human judgments.

We present a preliminary evaluation of the proposed metrics by measuring their correlation with human judgments, but it should be emphasized that we are only evaluating these metrics with respect to one specific style of writing. We are optimistic that these results will generalize across writing styles, however, since they are based entirely on ngram statistics.

[1] Note that the BLEU score of 16plays_36LM is significantly lower when evaluated on the Enotes data. This makes sense, because the 16 plays come from Sparknotes. This system is not trained on the 7 Enotes plays which, whose modern translations tend to be slightly different in style.

Source	Speaker	Input	Output
Romeo & Juliet	Benvolio	He killed your relative, brave Mercutio, and then young Romeo killed him.	he slew thy kinsman , brave mercutio , and then young romeo kill him .
Romeo & Juliet	Romeo	I can read my own fortune in my misery.	i can read mine own fortune in my woes .
Star Wars	Palpatine	If you will not be turned, you will be destroyed!	if you will not be turn 'd , you will be undone !
Star Wars	Luke	Father, please! Help me!	father , i pray you , help me !
The Matrix	Agent Smith	Good bye, Mr. Anderson.	fare you well , good master anderson .
The Matrix	Morpheus	I'm trying to free your mind, Neo. But I can only show you the door. You're the one that has to walk through it.	i 'll to free your mind , neo. but i can but show you the door. you 're the one that hath to tread it .
Raiders of the Lost Ark	Belloq	Good afternoon, Dr. Jones.	well met , dr. jones .
Raiders of the Lost Ark	Jones	I ought to kill you right now.	i should kill thee straight .

Table 5: Example Shakespearean paraphrases generated by the best overall system.

Semantic Adequacy	Lexical Dissimilarity	Style	Overall
0.73	0.82	0.64	0.62

Table 6: Agreement between annotators measured using Pearson's ρ.

4.1 Cosine Similarity Style Metric

As a first approach to automatic evaluation of writing style, we present a vector-space model of similarity between the system output and a large corpus of text in both the source and target style. The intuition behind this metric is that a large ngram overlap between the system's output and a corpus of text in the target style should indicate that the output is likely to be stylistically appropriate.

More concretely, we extract ngrams from both the source and target corpus which are represented as binary vectors $\vec{s}$, and $\vec{t}$; similarly the output sentence is represented using a vector of ngrams $\vec{o}$. The proposed metric is the normalized cosine similarity between the source and target corpora:

$$S_{\text{Cosine}}(\vec{o}) = \frac{\frac{\vec{o}\cdot\vec{t}}{\|\vec{o}\|\times\|\vec{t}\|}}{\frac{\vec{o}\cdot\vec{t}}{\|\vec{o}\|\times\|\vec{t}\|} + \frac{\vec{o}\cdot\vec{s}}{\|\vec{o}\|\times\|\vec{s}\|}}$$

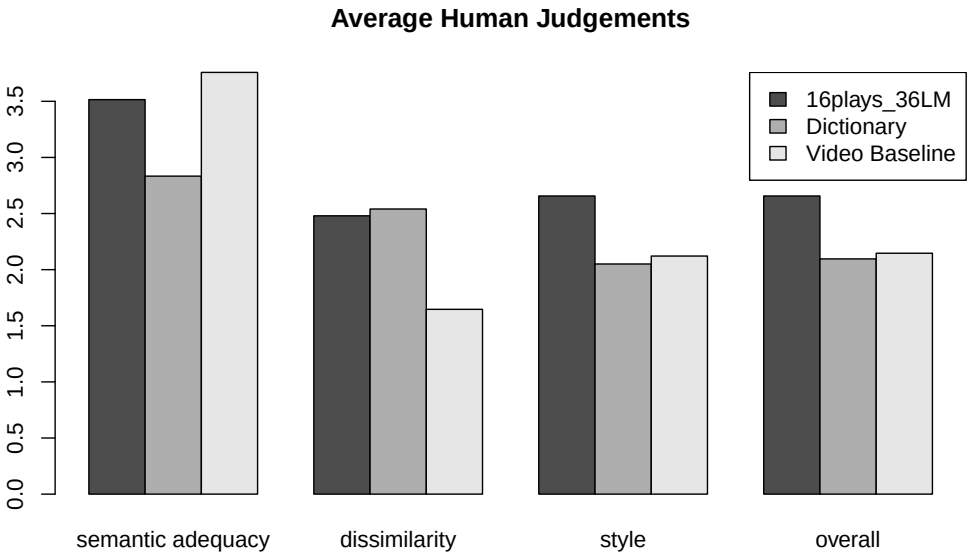

Figure 2: Average human judgments evaluating semantic adequacy, lexical dissimilarity, stylistic similarity, and overall quality of Shakespearean paraphrase systems

4.2 Language Model Style Metric

Another approach is to build a language model from a corpus of text in the target style and a background language model from text outside the style, then apply Bayes' rule to estimate the posterior probability that a sentence was generated from the target language model[2]:

$$\begin{aligned} P(\text{style} = \text{target}|\text{sentence}) &= \frac{P_{\text{LM}}(\text{sentence}|\text{target})P(\text{target})}{P(\text{sentence})} \\ &= \frac{P_{\text{LM}}(\text{sentence}|\text{target}) \times 0.5}{P_{\text{LM}}(\text{sentence}|\text{target}) \times 0.5 + P_{\text{LM}}(\text{sentence}|\text{source}) \times 0.5} \\ &= \frac{P_{\text{LM}}(\text{sentence}|\text{target})}{P_{\text{LM}}(\text{sentence}|\text{target}) + P_{\text{LM}}(\text{sentence}|\text{source})} \end{aligned}$$

4.3 Logistic Regression Style Metric

We also consider an approach to measuring style which is based on logistic regression. Here the idea is to estimate the probability that each sentence belongs to the target style based on the ngrams it contains, using large corpora of *in domain* and *out-of domain* sentences to learn parameters of a logistic regression model.

The probability that a sentence belongs to the target style is estimated as follows:

$$P(\text{style} = \text{target}|\text{sentence}) = \frac{1}{1 + e^{-\left(\vec{\theta} \cdot f(\vec{\text{sentence}})\right)}}$$

[2] Here we assume an uninformative prior, that is $P(\text{source}) = P(\text{target}) = 0.5$.

		ρ (Annotator 1)	ρ (Annotator 2)
semantic adequacy	BLEU	0.35	0.31
dissimilarity	PINC	0.78	0.82
style	BLEU	0.07	0.06
style	PINC	0.20	0.45
style	Cosine	0.37	0.41
style	LM	0.46	0.51
style	Logistic regression	0.47	0.47

Table 7: Correlation between various human judgments and automatic evaluation metrics. Pearson's correlation coefficient is displayed between the automatic metrics and human judgments from each annotator.

Where $f(\vec{\text{sentence}})$ is a vector of ngrams contained by the sentence, and $\vec{\theta}$ is a vector of weights corresponding to each possible ngram.

The parameters, $\vec{\theta}$, are optimized to maximize conditional likelihood on the source and target corpus, where the assumption is that the target corpus is in the target style, whereas the source corpus is not.[3]

4.4 Evaluation

We trained the logistic regression, language model and cosine similarity evaluation metrics using the original Shakespeare plays and modern translations as the source and target corpus respectively, then measured Pearson's Correlation Coefficient between the automatic evaluation metrics and human judgments described in §3. These results are reported in table 7.

As can be seen in table 7, the correlation between semantic adequacy and BLEU appears smaller than that reported in previous work (Chen and Dolan, 2011). Presumably this is due to the conflation of stylistic differences and semantic adequacy discussed in §3. However it also appears that the correlation between BLEU and human style judgments is too low to be of practical use for evaluating style.

PINC, on the other hand has high correlation with judgments on dissimilarity, and is also correlated with human style judgments. We believe PINC has correlation with writing style, because the systems we are evaluating all target Shakespearean English, so whenever changes are made to the input, they are likely to make it similar to the target style. Although PINC has relatively high correlation with human judgments, it is likely not a very useful measure of writing style in practice. For example, consider a paraphrase system which makes many changes to the input and thus gets a high PINC score, but targets a completely different writing style.

Both the language model and logistic regression style metrics achieve the highest overall correlation with human writing style judgments, achieving comparable performance.

We note that overall the automatic metrics tend to agree with agree with human judgments as displayed in Figure 3.[4]

[3] Parameters were optimized using MEGAM `http://www.cs.utah.edu/~hal/megam/`.

[4] Although the automatic style metrics rate the dictionary system higher than the video corpus baseline, both systems have very comparable style scores in the automatic and human evaluations.

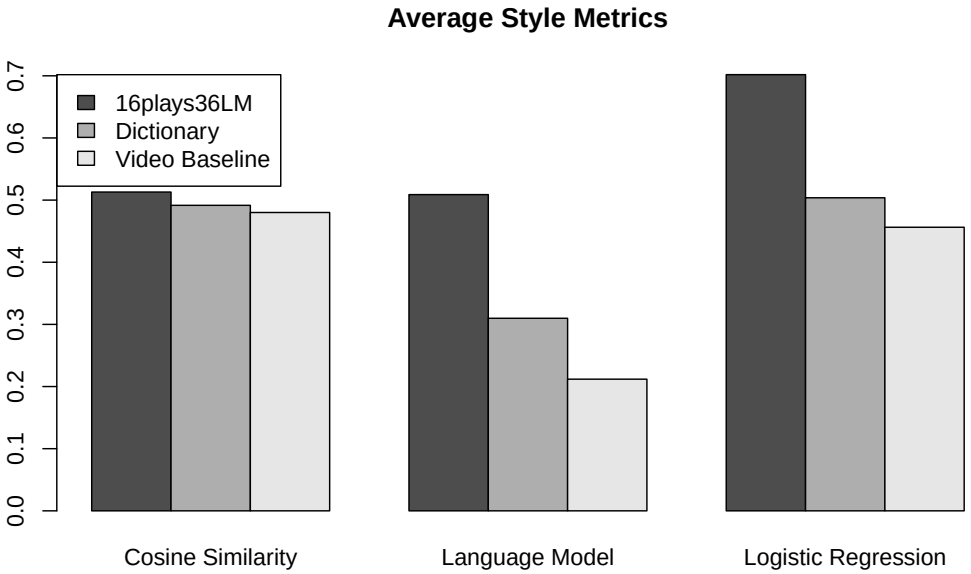

Figure 3: Results comparing the 3 systems using the automatic style metrics.

		ρ (Annotator 1)
semantic adequacy	BLEU	0.27
dissimilarity	PINC	0.79
style	BLEU	0.12
style	PINC	0.41
style	Cosine	0.37
style	LM	0.45
style	Logistic regression	0.46

Table 8: Correlation between human judgments and automatic evaluation metrics when paraphrasing Shakespeare's plays into modern prose.

5 Translating Shakespeare's Plays to Modern English

Finally we perform an evaluation on the task of automatically translating Shakespeare's plays into modern English.

For the purposes of this evaluation, we make use of the same paraphrase systems previously described, but swap the source and target languages. Additionally, each system makes use of a language model constructed from the 16 modern translations, with *Romeo and Juliet* held out for testing. 100 lines from *Romeo and Juliet* were automatically translated into modern English using each system, and the aligned modern translations were used as a reference when computing BLEU. The results of evaluating each of the automatic evaluation metrics on this data are presented in Figure 5, correlation of the automatic metrics with with human judgments are presented in Table 8 and average human judgments are presented in Figure 4.

These results suggest that in comparison to the dictionary and video corpus baselines, our

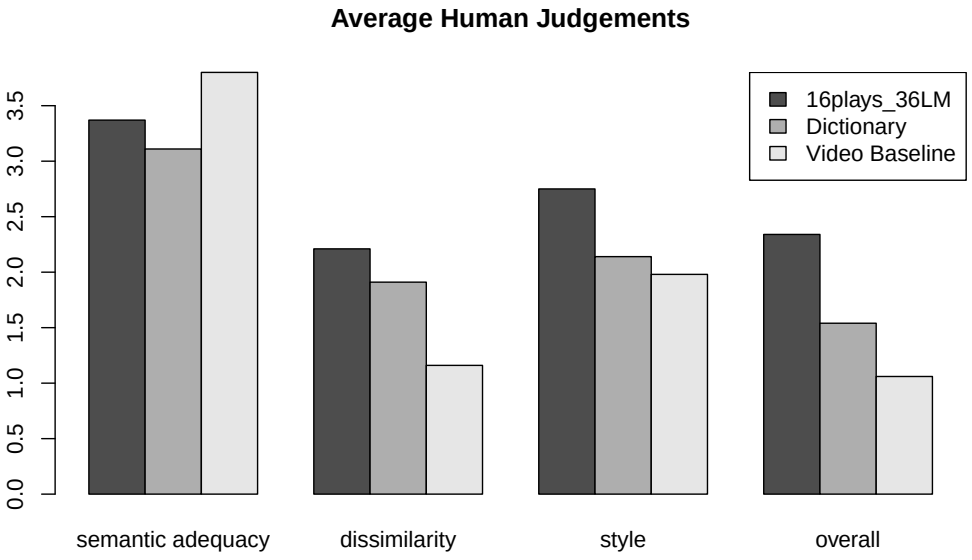

Figure 4: Average human judgments translating Shakespeare's plays into modern English.

system trained on modern translations generates a large number of paraphrases which match the target style. Note that the paraphrase system based on the out-of-domain video corpus makes very few changes to the input, and thus achieves a very low PINC score. This is due to the many out of vocabulary words in Shakespeare's plays which result in very few matching source phrases in the video baseline's phrase table. Several automatic paraphrases into modern English are presented in Table 9.

6 Related Work

Much previous work has addressed the task of automatically generating paraphrases (Barzilay and Lee, 2003; Dolan et al., 2004; Shinyama and Sekine, 2003; Das and Smith, 2009; Bannard and Callison-Burch, 2005; Callison-Burch, 2008; Kok and Brockett, 2010). In addition several authors have previously proposed automatic metrics specifically for evaluating paraphrases (Chen and Dolan, 2011; Callison-Burch et al., 2008; Liu et al., 2010). We are not aware, however, of any work that has addressed the task of generating or evaluating paraphrases targeting a specific style of writing.

Perhaps most relevant, however, is recent work automatic generation of rhythmic poetry (Greene et al., 2010). This work focuses on automatically generating and translating poetry in an appropriate meter (e.g. iambic pentameter) using finite-state transducers, but does not investigate the task of paraphrase. Their generation system is trained on Shakespeare's sonnets, and they investigate the task of automatically translating Dante's Divine Comedy from Italian to English. While our work does not address the issue of meter, it should be possible to combine our translation models with their weighted finite state transducers to produce Shakespearean paraphrase models which produce output in an appropriate meter.

Finally we highlight related work on authorship classification which can be seen as detecting a specific style of writing (Gamon, 2004; Raghavan et al., 2010). This work has not specifically

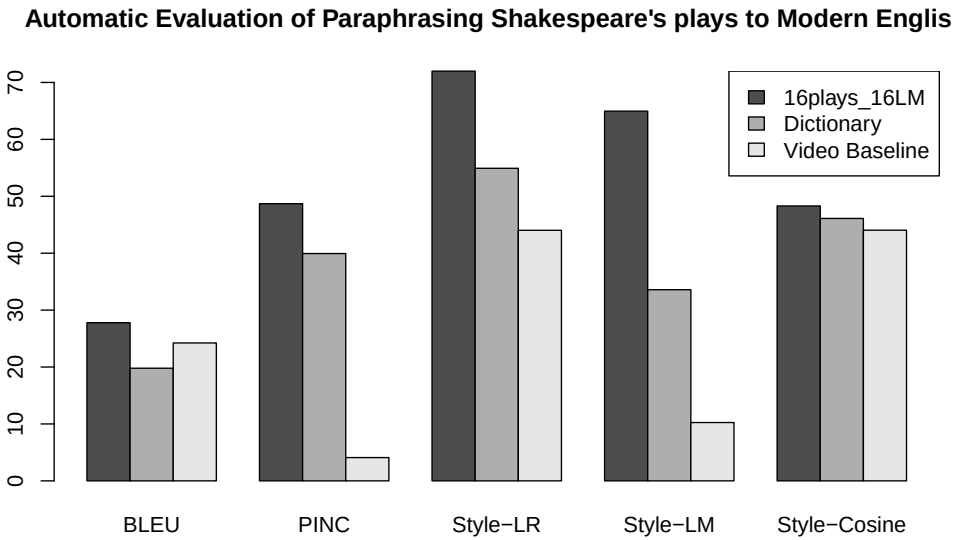

Figure 5: Automatic evaluation of paraphrasing Shakespeare's plays into modern English comparing a system based on parallel text (16plays_16LM), a Dictionary baseline, and a system trained on out of domain parallel monolingual text. Note that the video corpus baseline achieves low overall PINC score, as few phrases in the input match phrases found in its phrase table, resulting in a small number of changes to the input.

Speaker	Input	Output
MERCUTIO	i will bite thee by the ear for that jest .	i ' ll bite you by the ear for that joke .
MONTAGUE	what further woe conspires against mine age ?	what ' s true despair conspires against my old age ?
ROMEO	how doth my lady ?	how is my lady ?
FRIAR LAURENCE	hast thou slain tybalt ?	have you killed tybalt ?
NURSE	an i might live to see thee married once , i have my wish .	if i could live to see you married , i ' ve my wish .
PRINCE	benvolio , who began this bloody fray ?	benvolio , who started this bloody fight itself ?
JULIET	what is your will ?	what do you want ?
LADY CAPULET	call her forth to me .	bring her out to me .

Table 9: Example modern paraphrases of lines from Romeo and Juliet generated using our system.

addressed the task of automatically generating or evaluating paraphrases in a specific style, however.

7 Conclusions

We have presented the first investigation into the task of automatic paraphrasing while targeting a specific writing style. Using Shakespeare's plays and their modern translations as a testbed for this task, we developed a series of paraphrase systems targeting Shakespearean English. We showed that while existing evaluation metrics are useful for evaluating paraphrases in this context, BLEU tends to conflate semantic equivalence with writing style and thus gives an incomplete picture of system performance on these different dimensions.

To address this problem, we introduced three new metrics for evaluating writing style, one based on cosine similarity one based on language models, and the third based on logistic regression. We measured correlation between automatic metrics and human judgments, and showed that our new metrics have better correlation with human judgments than existing metrics in the context of our task. While this evaluation is limited to one specific style of writing, we are optimistic that these or similar metrics will also perform well when evaluating paraphrase systems targeting other writing styles.

We have shown that access to even a small amount of parallel text produces paraphrase systems capable of generating a large number of stylistically appropriate paraphrases while preserving the meaning of the input text. Our paraphrase systems targeting Shakespearean English could be beneficial for educational applications, for example helping to make Shakespeare's work accessible to a broader audience. Future work could investigate stylistic paraphrasing in other domains, such as paraphrasing emails into formal documents, or translating legal documents into nontechnical English.

References

Bannard, C. and Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*.

Barzilay, R. and Lee, L. (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.

Callison-Burch, C. (2008). Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Callison-Burch, C., Cohn, T., and Lapata, M. (2008). Parametric: an automatic evaluation metric for paraphrasing. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*.

Chandrasekar, R., Doran, C., and Srinivas, B. (1996). Motivations and methods for text simplification. In *Proceedings of the 16th conference on Computational linguistics - Volume 2*.

Chen, D. L. and Dolan, W. B. (2011). Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011)*, Portland, OR.

Clarke, J. and Lapata, M. (2008). Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research*, 31(1):399–429.

Cohn, T. and Lapata, M. (2009). Sentence compression as tree transduction. *Journal of Artificial Intelligence Research*, 34:637–674.

Das, D. and Smith, N. A. (2009). Paraphrase identification as probabilistic quasi-synchronous recognition. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09.

Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of Coling 2004*.

Gamon, M. (2004). Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of the 20th international conference on Computational Linguistics*.

Giampiccolo, D., Magnini, B., Dagan, I., and Dolan, B. (2007). The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07.

Greene, E., Bodrumlu, T., and Knight, K. (2010). Automatic analysis of rhythmic poetry with applications to generation and translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*.

Koehn, P. and Knight, K. (2000). Estimating word translation probabilities from unrelated monolingual corpora using the em algorithm. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*.

Kok, S. and Brockett, C. (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*.

Liu, C., Dahlmeier, D., and Ng, H. T. (2010). PEM: A paraphrase evaluation metric exploiting parallel texts. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

Madnani, N. and Dorr, B. (2010). Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, AMTA '02.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Comput. Linguist.*

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*.

Quirk, C., Brockett, C., and Dolan, W. (2004). Monolingual machine translation for paraphrase generation. In *Proceedings of EMNLP 2004*.

Raghavan, S., Kovashka, A., and Mooney, R. (2010). Authorship attribution using probabilistic context-free grammars. In *Proceedings of the ACL 2010 Conference Short Papers*.

Shinyama, Y. and Sekine, S. (2003). Paraphrase acquisition for information extraction. In *Proceedings of the second international workshop on Paraphrasing - Volume 16*, PARAPHRASE '03.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08.

Stolcke, A. (2002). Srilm - an extensible language modeling toolkit. In *INTERSPEECH*.

Vanderwende, L., Suzuki, H., Brockett, C., and Nenkova, A. (2007). Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Inf. Process. Manage.*

Yatskar, M., Pang, B., Danescu-Niculescu-Mizil, C., and Lee, L. (2010). For the sake of simplicity: unsupervised extraction of lexical simplifications from wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.