You can be Shakespeare!
A Case Study in Paraphrase Targeting Writing Styles

*Author*1[1,2] *Author*2[1,3]
(1) INSTITUTE_1, address 1
(2) INSTITUTE_2, address 2
(3) INSTITUTE_3, address 3
author1@institute1, author@institute2

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 available, it is possible to learn paraphrase models which capture stylistic phenomenon, 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

Identical 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(**?**), so much previous work has investigated methods for automatic paraphrasing(**?????**).

Although two utterances might be semantically equivalent, they can still be stylistically quite different. For example, the same information is likely to be conveyed using very different lexical and grammatical patterns in advertising materials v.s. technical manuals, or in Shakespearean plays v.s. Hollywood movies.

Systems capable of paraphrasing text targeting to 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 average people to better understand information in legal documents by translating "legalese" into ordinary English.
3. Benefit educational applications:
   (a) Students studying a specific style of writing may benefit from access to *modern English* versions of works by authors they are studying.

(b) Automatic paraphrases could help students to experiment with writing in the style of the authors whose works they are reading.

In this paper, we investigate the task of automatic paraphrasing when targeting a writing style, focusing specifically on the style of Early Modern English employed by William Shakespeare. We exploit modern translations of 17 plays written to help students better understand Shakespeare's work. A parallel corpus is extracted from these modern translations, which is then used to train phrase-based translation models which are capable of automatically paraphrasing ordinary sentences into Shakespearean English. In addition we develop several baseline systems which do not make use of this source of parallel text and instead rely on manually compiled dictionaries of expressions commonly found in Shakespearean English, or out of domain parallel monolingual text gathered through Amazon's Mechanical Turk (**?**).

We evaluate these models both through human judgments and standard evaluation metrics from the Machine Translation 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 automatic evaluation metrics do not give us a complete picture of a system's performance when the task is to generate paraphrases targeting a specific style of writing. We propose several new metrics for evaluating paraphrases targeting a specific style, and measure correlation with human judgments showing promising preliminary results.

Systems which are capable of automatically paraphrasing literary writing styles could be directly beneficial for educational applications. For example, our systems which generate paraphrases targeting a Shakespearean English could help students to experiment with writing literature in this style. In addition, out of the 37 surviving plays written by William Shakespeare modern translations are currently only available for 17. Though we have not yet formally evaluated paraphrasing Shakespeare's plays into modern English, we believe it should be possible to help make the other 20 plays more accessible to students of Shakespeare by automatically generating modern translations.

2 Shakespearean Paraphrasing

We propose to 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 highest regarded examples of English literature and are also very unique in style, many linguistic resources are available such as parallel corpora of modern translations and dictionaries of stylistically representative words and their modern equivalents.

We compare 3 different stylistic paraphrase systems targeting Shakespearean English. One which leverages parallel corpora of modern translations, another which makes use of dictionaries of stylistically representative expressions, and another which leverages out-of-domain monolingual parallel data.

2.1 Modern Translations

Having access to parallel text in the target style allows us to train statistical models for generating paraphrases, and also perform automatic evaluation of semantic adequacy using BLEU, which requires access to a set of reference translations. For this purpose we scraped modern translations of 17 Shakespeare plays from `http://nfs.sparknotes.com`, and an additional 8 translations of overlapping plays from `http://enotes.com`, giving us two reference translations for 8 out of the 17 plays.

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

After tokenizing and lowercasing, the plays were aligned using Bob Moore's bilingual sentence aligner (**?**), which produced about 21,079 alignments out of 31,718 sentences in the Sparknotes data, and 10,365 sentence pairs out of 13,640 sentences in the Enotes data. The modern translations from each source 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 unchanged modern translations against Shakespeare's original text; the Sparknotes paraphrases result in a BLEU score of 24.67, whereas the Enotes paraphrases produce a much higher BLEU of 52.30 indicating their strong similarity to the original plays. These corpus statistics are summarized in table 1.

To generate paraphrases, we apply a typical phrase-based Statistical Machine Translation pipeline, performing word alignment on the data described in table 1 using GIZA++ (**?**), then extracting phrase pairs and performing decoding using Moses (**?**).

2.2 Baselines

Phrase-based translation has been demonstrated to be an effective approach to generating paraphrases (**??**), however this approach does require the existence of parallel corpora which may not be available for many writing styles. For this reason we were motivated to investigate alternative approaches.

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 is used in combination with a language model as in standard phrase-based MT (**?**).

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 (**?**), 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 conditional probabilities for each translation according to the n-gram back off model built on Shakespeare's 36 plays by SRILM toolkit (**?**) 3, normalizing the probabilities for each source phrase, for example $p(\text{PERCHANGE}|\text{maybe}) = \frac{0.0000790755}{0.00026264791} = 0.30107035689$. This method allows us to estimate reasonable translation probabilities for use in a phrase table, which is used in combination with a language model consisting of Shakespeare's 36 plays (holding out *Romeo and Juliet*), which are then fed into the Moses decoder (**?**).

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

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 (**?**) to describe videos (**?**). We combine a phrase table extracted from this out of domain parallel text, with an in-domain language model consisting of Shakespeare's 36 plays (holding out *Romeo and Juliet* for testing), applying the Moses decoder (**?**) 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.

2.3 Comparison Using Existing Automatic Evaluation Metrics

Figure 1 compares a variety of systems targeting Shakespearean English using the previously proposed BLEU (**?**) and PINC (**?**) 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. Notice that the Enotes paraphrases are quite similar to the original text, obtaining 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 emerge from Figure 1. First, notice that all of the systems which are trained using parallel text achieve higher BLEU scores than the baseline of not making any changes to the modern translations. While the dictionary baseline achieves a competitive PINC score, indicating it is making a significant number of changes to the input, it's BLEU is lower than the *no changes* baseline. Secondly, it seems apparent that the systems whose parameters are learned using Minimum Error Rate Training (**?**) tend to be more conservative, making fewer changes to the input and thus achieving lower PINC scores, however also not seeing any BLEU improvements on the test data. Finally we note that using the larger language model seems to yield a slight improvement in BLEU score.

2.4 Examples

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

System	Description
16and7plays_36LM	Phrase table learned from all 16 Sparknotes plays (other than R&J) and language model built from all 36 of Shakespeare's plays, excluding R&J. Uses default Moses parameters.
16and7plays_36LM_MERT	Same as 16and7plays_36LM except parameters are tuned using Minimum Error Rate Training (**?**) instead of using the default Moses parameters.
16and7plays_16LM	Phrase table is built from both Sparknotes and Enotes data, and Language model is built from the 16 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 (**?**) with an in-domain language model. Described in detail in section 2.2.2.
modern (no change)	No changes are made to the input, modern translations are left unchanged.
Dictionary	Dictionary baseline described in section 2.2.1

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

3 Human Evaluation

Figure 1 provides some insight into the performance of the various systems, however it is initially unclear how well the BLEU and PINC automatic evaluation metrics perform when applied to evaluating paraphrases targeting 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, however no previous work has evaluated automatically generated paraphrases which target a specific style of writing (**?**).

While BLEU is typically used to measure semantic adequacy, it seems reasonable to assume that it may also depend on style, since utterances are more likely to contain overlapping ngrams if they are both semantically and stylistically similar. What BLEU doesn't tell us, however is what portion of it's 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.

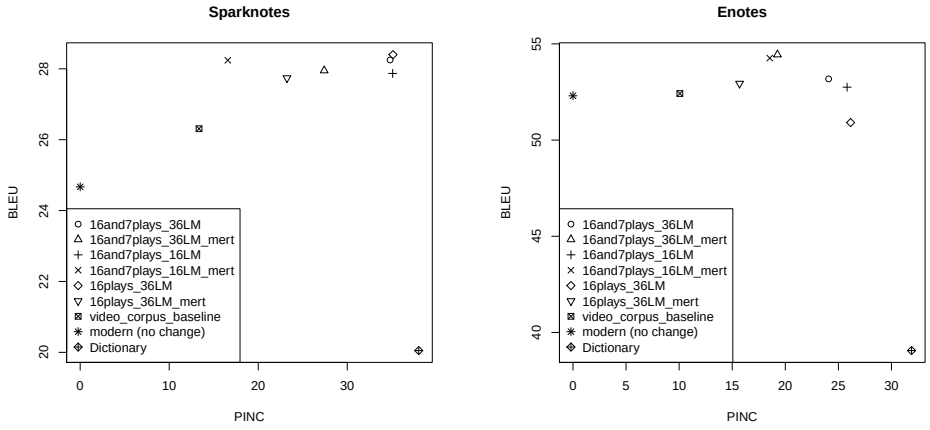

Figure 1: Various Shakespearean Paraphrase systems compared using BLEU and PINC. A brief description of each system is presented in table 4.

For this purpose, we randomly sampled 100 lines from *Romeo and Juliet*, then two of the authors annotated each sentence and it's Shakespearean translation with 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 (**?**) with a Shakespearean language model has the highest semantic adequacy, however it is also fairly conservative making few changes to the input.

The dictionary baseline, and paraphrase system trained on parallel modern translations are roughly comparable when it comes to the number of changes made to the input, however 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 evaluation metrics presented in figure 1, however we see several important trends which are not apparent based on the automatic evaluation metrics.

Although the video baseline achieves the highest semantic adequacy in the human evaluation, it's 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.

[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 Automatic Metrics Evaluating Writing Style

While PINC and BLEU do seem useful for automatically evaluating paraphrases targeting a writing style, 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 effects based on these two criteria. We would like our automatic evaluation metrics to be able to distinguish between a system which generates perfect paraphrases which don't match the target style of writing and a system which generates sentences in the correct style, but which convey different meaning.

To help address this issue we propose two new automatic evaluation metrics whose goal is to measure the degree to which automatic paraphrases match the target style. Both 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 these metrics by measuring their correlation with human judgments, however we emphasize that we are only evaluating these metrics with respect to one specific style of writing. We are cautiously optimistic that our results will generalize across

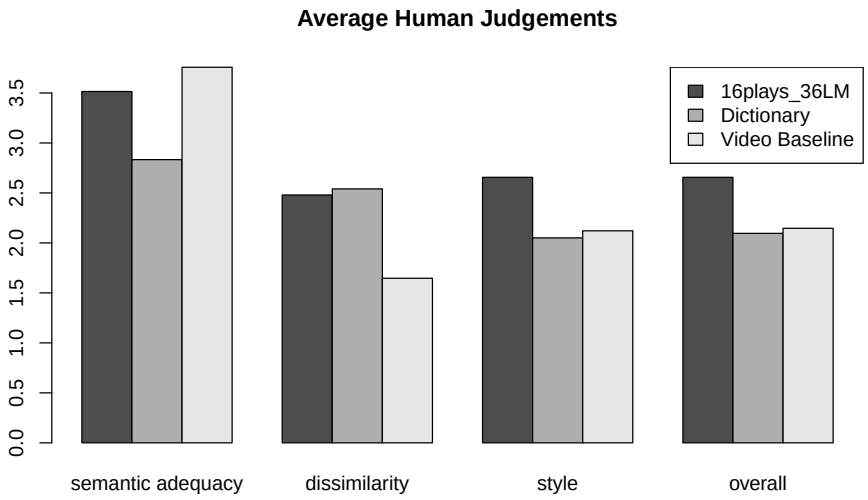

Figure 2: Average human judgments evaluating semantic adequacy, lexical dissimilarity, stylistic similarity, and overall quality of Shakespearean paraphrase systems

writing styles, however, as they are based entirely on ngram statistics.

4.1 Cosine Similarity Style Metric

As a first approach automatic evaluation of writing style, we propose 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 is that if there is a large ngram overlap between the system's output and a large corpus of text in the target style, then the output is more likely to be stylistically appropriate.

More concretely, we extract ngrams from both the source and target corpus which are then 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:

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

4.2 Logistic Regression Style Metric

We also consider a Logistic Regression based approach as an alternative to the Cosine-Similarity based style metric. 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 and out-of domain sentences to learn parameters of a logistic regression model.

		Pearson's ρ
semantic adequacy	BLEU	0.35
dissimilarity	PINC	0.78
style	BLEU	0.07
style	PINC	0.20
style	Cosine	0.37
style	Maximum Entropy	0.47

Table 7: Correlation between various human judgments and automatic evaluation metrics

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

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

Where $f(\vec{\text{sentence}})$ is a vector of ngrams, and $\vec{\theta}$ is a vector of weights.

The parameters, $\vec{\theta}$, are optimized 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.[2]

4.3 Evaluation

We trained both Logistic Regression and Cosine Similarity evaluation metrics using the original Shakespeare plays and modern translations as the source and target corpus respectively (holding out *Romeo and Juliet*), then measured Pearson's Correlation Coefficient between the automatic evaluation metrics and human judgments reported in section 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 (**?**). Presumably this is due to the conflation of stylistic differences and semantic adequacy discussed discussed in section 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 moderately correlated with human style judgments. We believe PINC has some 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 Cosine and Maximum Entropy style metrics achieve the highest overall correlation with human writing style judgments, with the Maximum Entropy style score performing significantly better.

Finally we note that overall the automatic metrics tend to agree with agree with human judgments as displayed in figure 3.

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

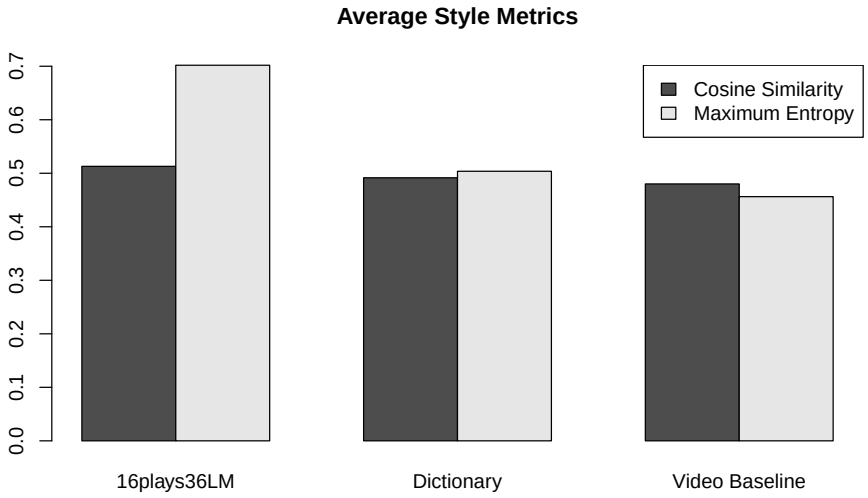

Figure 3: Overall results from comparing automatic style metrics. Note that the automatic metrics agree with results from human judgments in figure 2.

5 Related Work

Although no previous work has investigated paraphrasing modern text into Shakespearean English, or the more general task of paraphrasing while targeting a specific writing style, we highlight several strands of related work.

Perhaps most relevant is recent work generating and translating rhythmic poetry (**?**). This work focuses on automatically generating text in an appropriate meter (e.g. iambic pentameter) using finite-state transducers, however doesn't address the issue of paraphrase. 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.

Much previous work has addressed the task of automatically generating paraphrases (**???????**). In addition several authors have previously proposed automatic paraphrase evaluation metrics (**????**) for paraphrase. We are not aware, however, of any previous work that has addressed the task of generating or evaluating paraphrases targeting a specific style of writing.

Finally we highlight related work on authorship classification which can be seen as detecting a specific style of writing (**??**). This work has not specifically addressed the task of automatically generating or evaluating paraphrases in a specific style.

6 Conclusions

We have presented initial investigation into the task of automatic paraphrasing while targeting a specific style of writing. We proposed Shakespeare's plays and their modern translations as a tesbed for this task, and 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 giving an incomplete picture of which systems perform better according to each of these criteria.

To address the problems we have demonstrated with previous evaluation metrics applied to this task, we have introduced two new metrics for evaluating writing style. We measured correlation between automatic metrics and human judgments in the context of paraphrasing writing style, and showed that our new metrics have better correlation with human judgments than existing metrics in the context of our data. 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 other writing styles.

Future work could include automatically translating the remaining 20 of Shakespeare's plays into modern English, which could be beneficial to students of Shakespeare's plays and also to future human translators.