Evaluating Improvised Hip Hop Lyrics – Challenges and Observations

Karteek Addanki Dekai Wu
Hong Kong University of Science and Technology
Human Language Technology Center
Department of Computer Science
HKUST, Clear Water Bay, Hong Kong
{vskaddanki,dekai}@cs.ust.hk

Abstract

We investigate novel challenges involved in comparing model performance on the task of improvising responses to hip hop lyrics and discuss observations regarding inter-evaluator agreement on judging improvisation quality. We believe the analysis serves as a first step toward designing robust evaluation strategies for improvisation tasks, a relatively neglected area to date. Unlike most natural language processing tasks, improvisation tasks suffer from a high degree of subjectivity, making it difficult to design discriminative evaluation strategies to drive model development. We propose a simple strategy with fluency and rhyming as the criteria for evaluating the quality of generated responses, which we apply to both our inversion transduction grammar based Freestyle hip hop challenge-response improvisation system, as well as various contrastive systems. We report inter-evaluator agreement for both English and French hip hop lyrics, and analyze correlation with challenge length. We also compare the extent of agreement in evaluating fluency with that of rhyming, and quantify the difference in agreement with and without precise definitions of evaluation criteria.

Keywords: ITG induction, rap lyrics, computational musicology

1. Introduction

Although cheap and efficient evaluation methods exist for comparing the model performance on conventional NLP tasks, few efforts have been made toward identifying the problems and establishing robust evaluation strategies for comparing the system performance on non-conventional NLP tasks. Evaluating model performance is trivial for problems where the reference solution exists and is easily available. However, this is not true of many NLP tasks particularly for those that apply statistical learning methods to realize creative tasks. In this paper, we discuss the problems in evaluation wherein statistical machine translation (SMT) algorithms are used to improvise hip hop lyrics. Unlike the conventional task of machine translation where a set of reference translations can be generated, improvisation tasks have a much larger space of possibly "correct answers" which cannot be enumerated. Human evaluators are therefore necessary to compare the performance of different models. In order to be able to design a robust evaluation strategy, it is important to identify the task-specific issues that affect the human evaluation. This would also hopefully serve as a first step toward developing automatic or semi-automatic methodologies for evaluating performance of machine learning systems on tasks involving improvisation.

The genre of lyrics in music has been woefully understudied from the view of computational linguistics, despite being a form of language that has perhaps had the most impact across almost all human cultures. Hip hop provides an ideal genre due to its lack of well-defined structure in terms of rhyme scheme, meter or overall meaning to bring to light some of the less studied modeling and evaluation issues. Freestyle rapping is a prominent part of hip hop culture in which rap lyrics are improvised, commonly realized as a freestyle battle in which rappers compete using improvised lyrics (Fellowship *et al.*, 2006). We report observations and discuss some of the challenges involved in comparing performance of systems on the task of improvising a hip hop lyrical challenge. The task is extremely subjective even in comparison with other fairly subjective NLP tasks such as evaluating translation adequacy. Furthermore, the evaluation scheme should be able to quantify the differences in model performance even when the model improvisations are significantly inferior to a potential human improvisation.

We describe the evaluation methodologies adopted by previous efforts on applying statistical NLP methods to unconventional domains and highlight the differences with our current evaluation task in Section 2. Section 3 describes our ITG based Freestyle system for improvising a response given a hip hop lyrical challenge, along with several contrastive systems, and Section 4 discusses our specific experimental setup. Section 5 describes the evaluation scheme used for comparing the performance, contrasts our task with evaluating translation quality and report for the first time, the inter-evaluator agreement (IEA) on this task. We investigate the relationship between the IEA with the number of categories in our evaluation scale and the precision of evaluation instructions. Finally, the effect of challenge length on the IEA is analyzed and IEA across different language pairs are compared. Conclusions are presented in Section 6.

2. Related work

Our work appears to be one of the first to attempt to study evaluation of automatically generated improvised lyrics in music. The lyrics genre has been severely under-studied from the perspective of computational linguistics, in spite of the fact that it is a form of language that has perhaps had the most impact across all human cultures. Our analysis is a result of our modeling efforts on the complex and highly unstructured domain of hip hop lyrics (Addanki and Wu, 2013; Wu *et al.*, 2013a,b). As an attempt to highlight the challenges in evaluating the performance on our novel task,

we discuss the evaluation methodologies adopted by earlier attempts that apply statistical NLP learning methods to unconventional domains such as poetry and lyrics.

Most previous approaches to evaluation in tasks that can be seen in some sense to be loosely related to improvising hip hop lyrics use some sort of domain specific constraint—for example, on the number of words in a line, or the meter in a verse—or other prior linguistic knowledge, in order to define how "good" output candidates may be identified. Consequently, most of these approaches measure their performance by evaluating this "goodness" criterion, as opposed to objectively evaluating against the original task. However, the domain of hip hop lyrics enforces very few structural and linguistic constraints, which makes it hard to come up with a similar "goodness" criterion. For example, hip hop lyrics do not necessitate a set number of syllables per line, and words that are not a part of any standard lexicon such as sho, flo, holla frequently appear. Even assuming such a criterion were possible, human evaluation would still be needed to meta-evaluate alternative criterion, making human evaluation of unconventional tasks like ours an interesting area of study.

Jiang and Zhou (2008) trained a phrase based SMT system to "translate" the first line of a Chinese couplet or *duilian* into the second by applying linguistic constraints to the n best output of the SMT system. They evaluated their output by comparing the BLEU (Papineni *et al.*, 2002) score of the generated couplet lines against a set of manually collected references as they observed that BLEU score correlated highly with human evaluation. Unfortunately, we cannot adopt similar strategies for the domain of hip hop lyrics which, unlike their couplets, do not enforce an identical number of characters in each line and one-to-one correspondence in metrical length. This explodes the number of possible references and hence prohibits us from using BLEU or other off-the-shelf automatic MT evaluation metrics (Doddington, 2002; Leusch *et al.*, 2006; Snover *et al.*, 2006).

Barbieri *et al.* (2012) used controlled Markov processes to semi-automatically generate lyrics that satisfy the structural constraints of rhyme and meter and measure syntactic correctness and semantic similarity through human evaluation. We cannot merely adopt their evaluation criterion since, as discussed above, our domain unlike theirs does not have a well-defined syntactic structure. Further, their measure of semantic similarity is tightly coupled with the topic model they used to generate the lyrics, and does not extend easily to our model.

Tamil lyrics represented as a sequence of labels using the *KNM* system (where *K*, *N* and *M* represented the long vowels, short vowels and consonants respectively) were automatically generated given a melody using conditional random fields by A. *et al.* (2009). They did not address the problem of evaluating the quality of generated lyrics, instead simply reporting performance on the tasks of syllable, word boundary and sentence boundary identification.

Poems were translated through SMT algorithms in conjunction with stress patterns and rhymes found in a pronunciation dictionary by (Genzel *et al.*, 2010). However, they did not evaluate the syntactic or semantic quality of the generated generalizations, but merely reported a reduction in the BLEU scores relative to a baseline in which no restrictions are imposed on the output. While their results indicated how simple surface-based metrics such as BLEU are not useful in measuring the performance on non-conventional SMT tasks, they revealed very little about the quality of the generated output. In a similar effort, Greene *et al.* (2010) attempted to translate Dante's divine comedy by assigning stress patterns to words given the meter of a line. While they acknowledged that evaluating the quality of generated translations is an open problem, they relied on correctly identifying stress patterns in the output to measure their model performance.

Sonderegger (2011) attempted to infer the pronunciation of words in old English by identifying the rhyming patterns using graph theory and reported the f-score on the inferred pronunciations against a gold standard. A language-independent generative model for stanzas in poetry was proposed by Reddy and Knight (2011) via which they could discover rhyme schemes in French and English poetry. Similar to Sonderegger (2011), they also reported the f-score on the discovered rhyme schemes.

3. Improvisation via transduction grammar induction

Our Freestyle model improvises rhyming and fluent responses to hip hop lyric challenges by learning a stochastic ITG (inversion transduction grammar) similar to those used for statistical machine translation (Wu *et al.*, 2013a,b). Although similar problems in other domains such as poetry have been tackled in the past, none of them dealt with a highly unstructured domain such as hip hop. We improvise responses to hip hop lyrical challenges by transducing or "translating" the challenge into a response. In this section, we briefly describe our transduction grammar based model and the phrase based SMT (Pbsmt) baseline, along with two disfluency handling strategies that we employ on both these models.

Examples of challenge-response pairs generated by Freestyle can be seen in Table 1 (the fluency and rhyming judgments also shown will be discussed later).

A brief terminological note: "stanza" and "verse" are frequently confused and sometimes conflated. Worse yet, their usage for song lyrics is often contradictory to that for poetry. To avoid ambiguity we consistently follow these technical definitions for segments in decreasing size of granularity:

verse a large unit of a song's lyrics. A song typically contains several verses interspersed with choruses. In the present work, we do not differentiate choruses from verses. In song lyrics, a verse is most commonly represented as a separate paragraph.

stanza a segment within a verse which has a meter and rhyme scheme. Stanzas often consist of 2, 3, or 4 lines, but stanzas of more lines are also common. Particularly in hip hop, a single verse often contains many stanzas with different rhyme schemes and meters.

line a segment within a stanza consisting of a single line. In poetry, strictly speaking this would be called a

Challenge	Response	Fluency	Rhyming
cause you and your friends ain't nothing but fiends	you know that you me can you a dream	*acceptable*	*bad*
cause you ain't going home till the early **morn**	and the you this alone i i gotta **on**	*bad*	*acceptable*
created different elements and they **travelled** abroad	so you rather to the **well** land	*good*	*acceptable*
man i die to see em all thun i just don't **care**	in the sky and me the in polla and the you **there**	*acceptable*	*good*
what would i **do**	just me and **you**	*good*	*good*
almost a thought	what is this	*good*	*bad*
we fell off into a club to try to **pop** it to daz	you was a of the love i like a **stop** the the i	*bad*	*good*

Table 1: Examples of hip hop lyrical challenges and the responses generated by our models along with the scores on fluency and rhyming given by the human evaluators. Rhyming tokens in challenge and the corresponding responses are shown in bold face.

"verse", which however conflicts with the conventional use of "verse" in song lyrics.

3.1. Inversion transduction grammars

To serve as the transduction model for "translating" hip hop lyrical challenges into responses, Freestyle employs a token-based BITG or bracketing ITG (Wu, 1995a,b, 1997) due to its empirically demonstrated representational capacity across a wide spectrum of natural language tasks including textual entailment (Wu, 2006), mining parallel sentences (Wu and Fung, 2005) and machine translation (Zens and Ney, 2003; Haghighi *et al.*, 2009). We rely on the expressiveness of the ITG models to identify potential rhyming candidates through token level correspondences. ITG models, unlike monotonic finite-state transduction grammar models, are capable of learning token level correspondences involving non-monotonic alignments.

Model parameters for the ITG are estimated via expectation maximization (Dempster *et al.*, 1977) using the generalized inside-outside algorithm of Wu (1995c). As the hip hop lyrics training corpora (described below) are fairly large, beam pruning is used to make the training faster. Further details of the transduction grammar induction can be found in (Saers and Wu, 2011; Saers *et al.*, 2012).

3.2. Decoding heuristics

Once an ITG has been induced, an ITG based decoder Wu (1996) generates responses to challenges by "translating" the challenges using the trained transduction grammar together with an n-gram language model. The decoder uses a CKY-style parsing algorithm (Cocke, 1969) along with cube pruning (Chiang, 2007) to find the optimal translation candidate according to the induced grammar and the language model. The language model is trained on the entire training corpus using SRILM (Stolcke, 2002). The weights for the language model and the transduction grammar are empirically determined using a small development set.

We restrict the output to follow the same rhyming order as the challenge, as interleaved rhyming order is harder to evaluate without the larger context of the song. Therefore, we restrict the reordering to only be monotonic during decoding. Further, we penalize singleton rules to produce responses of similar length as the challenges because successive lines in a stanza are typically of similar length. Finally, *reflexive* translation rules that map the same surface form to itself such as $A \rightarrow$ yo/yo are penalized as they have unusually high probability due to presence of repeated chorus lines with identical surface form as training examples.

3.3. Phrase based SMT baseline

For comparative purposes, we also constructed a baseline applying an off-the-shelf phrase-based SMT (Pbsmt) system to our novel task of generating rhyming and fluent responses. We trained a standard Moses baseline (Koehn *et al.*, 2007) on the same training data and used the same 4-gram language model to generate responses. Since automatic quality evaluation metrics of the kind used in SMT like BLEU are not applicable to hip hop responses (as discussed earlier), the Pbsmt model weights cannot be tuned using conventional methods such as MERT (Och, 2003). Hence, we use a slightly higher than typical language model weight, chosen empirically by manually evaluating the output on a small development set.

3.4. Disfluency handling: correction vs. filtering

An inspection of the output from the initial runs of our model showed a disturbingly high proportion of responses containing disfluencies with successive repetitions of words such as the and I. Further, error analysis revealed that lyrics in the training data contained such disfluencies and backing vocal lines, amounting to 10% of our training data. We propose and compare the following two disfluency handling strategies: (1) filtering out all lines from our training corpus which contained such disfluencies, and (2) implementing a disfluency detection and correction algorithm (for example, the the the, which frequently occurred in the training corpus, was corrected to simply the). The Pbsmt baseline and the Freestyle model were trained on both the filtered and corrected versions of the training corpus.

4. Experimental setup

In this section, we briefly describe datasets used in our experiments comparing the performance of our transduction grammar based model and our Pbsmt baseline on English hip hop lyric improvisation. Taking advantage of the language independence and linguistics-light approach of our unsupervised transduction grammar induction methods, we also apply our models to rap in Maghrebi French and describe the datasets for French hip hop experiments. We also review the rhyme scheme detection module used to select our training data as described in Addanki and Wu (2013).

4.1. English dataset

About 800Mb (raw HTML data) of freely available user generated content amounting to lyrics of approximately 52,000 hip hop songs was crawled from the Internet and preprocessed by stripping HTML tags, metadata and normalizing for special characters and case differences. The entire corpus contained 22 million tokens with 260,000 verses and 2.7 million lines of hip hop lyrics. A randomly chosen subset of 85 lines was used as a test set to provide the hip hop lyrical challenges to the systems. In order to train the rhyme scheme detector module, we extracted the end-of-line words and words before all the commas from each verse. We obtained a corpus containing 4.2 million tokens corresponding to potential rhyming candidates with around 153,000 unique token types.

4.2. French dataset

Approximately lyrics corresponding to 1300 songs from the genre French hip hop were downloaded from the Internet. About 85% of these songs were by Maghrebi French artists of Algerian, Moroccan, or Tunisian cultural backgrounds, while the remaining were by artists from the rest of Francophonie. For training the rhyme scheme detection module, we obtained a corpus from the end-of-line tokens in the lyrics amounting to 120,000 tokens corresponding to potential rhyming candidates with around 29,000 unique token types. Training data of about 47,000 sentence pairs for transduction grammar induction was selected using rhyme scheme detection module.

4.3. Rhyme scheme detection module

Due to variance in hip hop rhyme schemes, it is not desirable to train our SMT system on successive lines of hip hop lyrics as described by Jiang and Zhou (2008). For instance, it is very common for a stanza to follow the **ABAB** rhyme scheme and therefore adding successive lines to the SMT system would drive the system to learn incorrect rhyme correspondences. Further, a verse may contain multiple stanzas and rhyme schemes which further exacerbates the problem. The naive solution of adding all possible pairs of lines in a verse explodes the training data size making it impractical, in addition to adding a number of noisy training examples.

We employ a rhyme scheme detection module as described in Addanki and Wu (2013) to select training instances that are likely to rhyme. Lines in a stanza that are marked as rhyming according to the rhyme scheme detection module are added as training examples to the SMT systems thereby biasing thereby biasing the model towards learning the correct rhyme associations.

A generative model for a verse of hip hop lyrics based on a hidden Markov model (HMM) is used to learn the rhyme schemes in an unsupervised fashion from the training data. Each state in the HMM corresponds to a stanza with a particular rhyme scheme such as **AA**, **ABAB**, **AAAA** and the emissions correspond to the final words in the stanza. As opposed to segmenting the verse into stanzas manually, each path through the lattice of the HMM corresponds to a *soft-segmentation* of the verse. The maximum length of a stanza is restricted to four as exhaustively considering the exponential number of partitions (Sloane, 2013) is expensive. Further, rhyme schemes that cannot be partitioned into a sequence of two smaller rhyme schemes is explicitly represented as a state in the HMM. For example, a rhyme scheme of length 3 **AAB** can be represented via a sequence of two smaller rhyme schemes **AA** and **B** without losing any rhyme correspondences and hence not represented in the HMM explicitly. Our HMM model is fully connected with the following 9 states: **A**, **AA**, **ABA**, **AAA**, **ABAB**, **AABA**, **ABAA**, **BAAA**, **AAAA**.

The parameters of the HMM are estimated using the forward-backward algorithm (Baum *et al.*, 1970; Devijer, 1985) on a training corpus generated from the end-of-line tokens in the lyrics. Each verse is segmented into stanzas along with their corresponding rhyme schemes according to the Viterbi parse of the trained model. All pairs of lines in a stanza that rhyme with each other are added as training examples to the SMT system. Each selected pair generates two training instances: a challenge-response and a response-challenge pair as the source and target languages are identical.

5. Evaluating improvisation

Since evaluating the quality of improvised responses is a highly subjective task, simple and well-defined criteria are necessary to distinguish the performance of different models. We choose fluency and rhyming as our criterion for evaluating the system performance. Fluency ensures that the responses appear similar to natural language output and rhyming is a domain requirement of hip hop. We note that there exists no standard criterion for judging the merit of improvised responses and our evaluation scheme is targeted at discriminating good models from the bad ones.

The output of all the systems on the test set was given to three independent frequent hip hop listeners for manual evaluation. They were asked to evaluate the system outputs according to fluency and the degree of rhyming. They were free to choose the tune to make the lyrics rhyme as the beats of the song were not used in the training data. Each evaluator was asked to score the response of each system on the criterion of fluency and rhyming as being *good*, *acceptable* or *bad*. Table 1 shows examples of challenge-response pairs generated by the system along with the corresponding rating provided by the human evaluators on the criterion both fluency and rhyming.

One might be tempted to contrast the trends observed as a part of our evaluation methodology with those commonly observed as a part of human evaluation in translation task because (1) our problem requires our challenge to be "transduced" into a response not unlike translating from one language to another and (2) due to the similarity of our underlying model to those used in SMT. However, one must be wary of drawing an equivalence between the problem of improvisation and translation although the former can be conveniently modeled as latter. The space of possible *correct* outputs in an improvisation task is very much larger than in translation. Evaluating the quality of responses warrants a greater deal of subjectivity compared to evaluating translations which yield moderate agreements (Landis and Koch, 1977) at best in shared SMT tasks (Federico *et al.*, 2011).

Model	Scale	Fluency	Rhyming
Freestyle+correction	3-scale	0.138	0.124
Freestyle+correction	2-scale	0.209	0.177
Pbsmt+correction	3-scale	0.282	0.107
Pbsmt+correction	2-scale	0.457	0.145
Freestyle+filtering	3-scale	0.147	0.121
Freestyle+filtering	2-scale	0.164	0.153
Pbsmt+filtering	3-scale	0.216	0.208
Pbsmt+filtering	2-scale	0.338	0.283

Table 2: Inter-evaluator agreement (as measured by Fleiss' kappa) for different models on the criterion of fluency and rhyming. The instructions to evaluators did not contain examples of machine-generated responses or precise definitions of fluency and rhyming.

Model	Scale	Fluency	Rhyming
Freestyle+correction	3-scale	0.082	0.122
Freestyle+correction	2-scale	0.049	0.183
Pbsmt+correction	3-scale	-0.154	0.131
Pbsmt+correction	2-scale	-0.101	0.188
Freestyle+filtering	3-scale	0.176	0.117
Freestyle+filtering	2-scale	0.153	0.132
Pbsmt+filtering	3-scale	-0.142	0.094
Pbsmt+filtering	2-scale	-0.083	0.157

Table 3: Inter-evaluator agreement (as measured by Fleiss' kappa) for different models on the criterion of fluency and rhyming. The instructions to evaluators did not contain examples of machine-generated responses or precise definitions of fluency and rhyming.

5.1. Inter-evaluator agreement is lower

The inter-evaluator agreement is low on our current task in comparison with other common NLP tasks such as translation adequacy judgement. Unlike the conventional translation tasks where the human evaluator is tasked with estimating the degree to which the semantic content of the input sentence the translate communicates, improvisation tasks have no well-defined criterion for judging the merit of the outputs. Such a challenge is inherent to any creative task and commonly manifests in varying levels of appreciation experienced by works of art created by humans let alone an improvised response by a machine. The fact that the responses do not match the levels of semantic coherence and creativity exhibited by artist generated lyrics make it harder for the evaluators to agree on the quality of the output.

Table 2 shows the inter-evaluator agreement as measured by Fleiss' kappa (Fleiss, 1971) for the four systems on both the criteria of fluency and rhyming. Although the inter-evaluator agreement appear low at first blush, given the subjectivity of the task and the fact that the output was generated by a language independent learning method fair to moderate agreement (Landis and Koch, 1977) among annotators is very encouraging.

5.2. 2 category scale vs. 3 category scale

Table 2 also indicates that the inter-evaluator agreement improves when two categories (*acceptable* and *bad*) are used instead of three (*good*, *acceptable* and *bad*). Using the two category scale, we remain agnostic to the judgement of evaluators about the extent of *goodness* of the responses and quantify their agreement on the *bad* responses and the cumulative fraction of sentences that were rated *acceptable* was also used to compare the performance of models. A theoretical reason for computing these numbers is that Fleiss' kappa cannot be used for ordered-categorical ratings yielding more robust numbers on the two category scale. For all the models on the criteria of both fluency and rhyming, the inter-evaluator agreement on the two category scale is higher than the three category scale indicating that the evaluators agree more on about the responses being *bad* (or not) than the degree of *goodness*.

Model	Scale	Fluency	Rhyming
Freestyle+correction	3-scale	-0.128	-0.065
Freestyle+correction	2-scale	-0.059	-0.092
Pbsmt+correction	3-scale	-0.067	0.000
Pbsmt+correction	2-scale	-0.171	-0.022
Freestyle+filtering	3-scale	-0.009	0.054
Freestyle+filtering	2-scale	-0.002	-0.063
Pbsmt+filtering	3-scale	-0.005	0.043
Pbsmt+filtering	2-scale	-0.066	-0.011

Table 4: Correlation of inter-evaluation agreement and the challenge length for different models on the criterion of fluency and rhyming.

5.3. Fluency has higher agreement than rhyming

We can also observe from the results in Table 2 that the inter-evaluator agreement is lower for rhyming compared to fluency for all the models. This observation is not surprising as evaluating the fluency of a response is much less subjective than evaluating the degree of rhyming. In the domain of hip hop evaluating rhyming is even more subjective as rhyming is frequently achieved by alliteration and intonation. Similar trends are observed when the two-category scale is used.

5.4. Inter-evaluator agreement is independent of challenge length

We hypothesized that responses to longer challenges might be more prone to inter-evaluator disagreement than to shorter challenges because response length is proportional to challenge length and therefore provide more instances for the evaluators to disagree about fluency and rhyming. However, results in Table 4 show that the length of the challenge (and therefore the response) have no significant correlation with the inter-evaluator agreement about fluency and rhyming. It is interesting to note that most correlation coefficients are slightly negative. The challenge lengths in our evaluation fall within a narrow range (typical in the domain of hip hop) and one might find stronger correlation in tasks which involve challenge lengths spanning a larger range.

Model	Scale	Fluency	Rhyming
Model 1	3-scale	0.301	0.317
Model 1	2-scale	0.330	0.334
Model 2	3-scale	0.326	0.282
Model 2	2-scale	0.361	0.286
Model 3	3-scale	0.387	0.295
Model 3	2-scale	0.464	0.331
Model 4	3-scale	0.360	0.293
Model 4	2-scale	0.366	0.298

Table 5: Inter-evaluator agreement (as measured by Fleiss' kappa) for French hip hop lyrics generated by four different models.

5.5. Precise instructions improve agreement

We noticed that for highly subjective evaluation tasks such as ours, the instructions to evaluators play an important role in the quality of the evaluations. Table 3 shows the inter-evaluator agreement on one of our preliminary evaluation run where the evaluators were not provided with examples of machine generated responses or precise definitions of fluency and rhyming criteria. We can observe that the inter-evaluator agreement is lower for most of the models compared to the numbers in Table 2. Although our observations are by no means surprising, it is still interesting to quantify the improvement resulting from precise instructions.

5.6. Evaluation of hip hop responses in French

Table 5 shows the inter-evaluator agreement for French hip hop lyrics on the criteria of fluency and rhyming for four different models. Although the training data was significantly smaller, the inter-evaluator agreement is significantly higher for the French lyrics compared to English hip hop lyrics. A possible explanation for this improvement in inter-evaluator agreement could be that French responses are less ambiguous compared to English. Upon observing that a significantly smaller fraction of responses were labeled *acceptable* (compared to English), we speculate that the poor quality of responses caused the annotators to agree more often than in English. Further experimentation is necessary to determine the extent to which the inter-evaluator agreement depends on the language of the task.

6. Conclusions

We have discussed observations and raised issues on the novel challenges of evaluating the performance of models for improvising music lyrics—specifically on the task of improvising responses to hip hop lyrical challenges—and proposed specific evaluation methodologies that we applied to our ITG based Freestyle system and various contrastive systems. Despite being a more subjective task compared to conventional NLP evaluation tasks, we defined simple and well-defined criterion for discriminating model performance and observed encouraging inter-evaluator agreement. We compared inter-evaluator agreement for scales based on two versus three categories, and confirmed that fluency is easier to agree upon than rhyming. We observed that the inter-evaluator agreement is independent of the length of the challenge, and quantified the degree of disagreement caused by lack of examples and precise definitions in evaluation criterion. Finally, we reported agreement statistics on the evaluation of French hip hop responses, and detected stronger correlations. Further experimentation is needed to confirm whether the evaluator agreement is language dependent.

7. Acknowledgements

We would like to thank Anik Dey, Belsy Yuen, Len Foong Koong, Meriem Beloucif, Nicolas Auguin, Tyler Barth and Vineet Pandey for their help with human evaluation. This material is based upon work supported in part by the Hong Kong Research Grants Council (RGC) research grants GRF620811, GRF621008, GRF612806; by the Defense Advanced Research Projects Agency (DARPA) under BOLT contract nos. HR0011-12-C-0014 and HR0011-12-C-0016, and GALE contract nos. HR0011-06-C-0022 and HR0011-06-C-0023; and by the European Union under the FP7 grant agreement no. 287658. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the RGC, EU, or DARPA.

8. References

Ananth Ramakrishnan A., Sankar Kuppan, and Lalitha Devi Sobha. Automatic generation of Tamil lyrics for melodies. In *Workshop on Computational Approaches to Linguistic Creativity (CALC-09)*, pages 40–46, 2009.

Karteek Addanki and Dekai Wu. Unsupervised rhyme scheme identification in hip hop lyrics using hidden Markov models. In *1st International Conference on Statistical Language and Speech Processing (SLSP 2013)*, Tarragona, Spain, 2013.

Gabriele Barbieri, François Pachet, Pierre Roy, and Mirko Degli Esposti. Markov constraints for generating lyrics with style. In *20th European Conference on Artificial Intelligence, (ECAI 2012)*, pages 115–120, 2012.

Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

David Chiang. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228, 2007.

John Cocke. *Programming languages and their compilers: Preliminary notes*. Courant Institute of Mathematical Sciences, New York University, 1969.

Arthur Pentland Dempster, Nan M. Laird, and Donald Bruce Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

P.A. Devijer. Baum's forward-backward algorithm revisited. *Pattern Recognition Letters*, 3(6):369–373, 1985.

George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *The second international conference on Human Language Technology Research (HLT '02)*, pages 138–145, San Diego, California, 2002.

Marcello Federico, Luisa Bentivogli, Michael Paul, and Sebastian Stüker. Overview of the iwslt 2011 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 8–9, 2011.

Freestyle Fellowship, MC Supernatural, G Craig, et al. Freestyle: The art of rhyme kevin fitzgerald, aka dj organic sony bmg music (canada) inc., 2004 75 minutes. *Popular Music and Society*, 29(5):617–619, 2006.

Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

D. Genzel, J. Uszkoreit, and F. Och. Poetic statistical machine translation: rhyme and meter. In *2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 158–166. Association for Computational Linguistics, 2010.

E. Greene, T. Bodrumlu, and K. Knight. Automatic analysis of rhythmic poetry with applications to generation and translation. In *2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 524–533. Association for Computational Linguistics, 2010.

Aria Haghighi, John Blitzer, John DeNero, and Dan Klein. Better word alignments with supervised ITG models. In *Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP 2009)*, pages 923–931, Suntec, Singapore, August 2009.

Long Jiang and Ming Zhou. Generating Chinese couplets using a statistical MT approach. In *22nd International Conference on Computational Linguistics (COLING 2008)*, 2008.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Interactive Poster and Demonstration Sessions of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 177–180, Prague, Czech Republic, June 2007.

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

Gregor Leusch, Nicola Ueffing, and Hermann Ney. CDer: Efficient MT evaluation using block movements. In *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, 2006.

Franz Josef Och. Minimum error rate training in statistical machine translation. In *41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, pages 160–167, Sapporo, Japan, July 2003.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 311–318, Philadelphia, Pennsylvania, July 2002.

S. Reddy and K. Knight. Unsupervised discovery of rhyme schemes. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, volume 2, pages 77–82. Association for Computational Linguistics, 2011.

Markus Saers and Dekai Wu. Reestimation of reified rules in semiring parsing and biparsing. In *Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-5)*, pages 70–78, Portland, Oregon, June 2011. Association for Computational Linguistics.

Markus Saers, Karteek Addanki, and Dekai Wu. From finite-state to inversion transductions: Toward unsupervised bilingual grammar induction. In *24th International Conference on Computational Linguistics (COLING 2012)*, pages 2325–2340, Mumbai, India, December 2012.

Neil James Alexander Sloane. Bell or exponential numbers: ways of placing n labeled balls into n indistinguishable boxes. *The On-Line Encyclopedia of Integer Sequences*, `http://oeis.org/A000110`, 2013. Accessed: 2013-06-30.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *7th Biennial Conference Association for Machine Translation in the Americas (AMTA 2006)*, pages 223–231, Cambridge, Massachusetts, August 2006.

M. Sonderegger. Applications of graph theory to an English rhyming corpus. *Computer Speech & Language*, 25(3):655–678, 2011.

Andreas Stolcke. SRILM – an extensible language modeling toolkit. In *7th International Conference on Spoken Language Processing (ICSLP2002 - INTERSPEECH 2002)*, pages 901–904, Denver, Colorado, September 2002.

Dekai Wu and Pascale Fung. Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. In *Second International Joint Conference on Natural Language Processing (IJCNLP 2005)*, pages 257–268. Springer, 2005.

Dekai Wu, Karteek Addanki, and Markus Saers. FREESTYLE: A challenge-response system for hip hop lyrics via unsupervised induction of stochastic transduction grammars. In *14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, pages 3478–3482, 2013.

Dekai Wu, Karteek Addanki, and Markus Saers. Modeling hip hop challenge-response lyrics as machine translation. In *14th Machine Translation Summit (MT Summit XIV)*, 2013.

Dekai Wu. An algorithm for simultaneously bracketing parallel texts by aligning words. In *33rd Annual Meeting of the Association for Computational Linguistics (ACL 95)*, pages 244–251, Cambridge, Massachusetts, June 1995.

Dekai Wu. Stochastic inversion transduction grammars, with application to segmentation, bracketing, and alignment of parallel corpora. In *14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, volume 95,

pages 1328–1335, 1995.

Dekai Wu. Trainable coarse bilingual grammars for parallel text bracketing. In *Third Annual Workshop on Very Large Corpora (WVLC-3)*, pages 69–81, Cambridge, Massachusetts, Jun 1995.

Dekai Wu. A polynomial-time algorithm for statistical machine translation. In *34th Annual Meeting of the Association for Computational Linguistics (ACL96)*, pages 152–158, Morristown, NJ, USA, 1996.

Dekai Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403, 1997.

Dekai Wu. Textual entailment recognition using inversion transduction grammars. In Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc, editors, *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop (MLCW 2005)*, volume 3944 of *Lecture Notes in Computer Science*, pages 299–308. Springer, Berlin, 2006.

Richard Zens and Hermann Ney. A comparative study on reordering constraints in statistical machine translation. In *41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, pages 144–151, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.