FREESTYLE: A RAP BATTLE BOT THAT LEARNS TO IMPROVISE

Dekai Wu
Hong Kong University of Science & Technology
dekai@cs.ust.hk

Karteek Addanki
HKUST
vskaddanki@cs.ust.hk

ABSTRACT

We demonstrate a rap battle bot that autonomously learns to freestyle creatively in real time, via a fast new hybrid compositional improvisation model integrating symbolic transduction grammar induction with novel bilingual recursive neural networks. Given that rap and hip hop represent one of music's most influential recent developments, surprisingly little research has been done in music technology. In true rap battling—the genre's most difficult form—an appropriate output response must be improvised whenever challenged by some input line of lyrics. As with many musical improvisation tasks, modeling the creative process is complex because it requires *compositionality*: improvising a good response not only requires making salient associations with the challenge at many overlapping levels of granularity, but simultaneously satisfying contextual preferences across a wide variety of dimensions. Our new real-time system accomplishes this via an efficient, recursive, neurally guided stochastic grammar-based transducer. The neural network is a newly enhanced version of our recently developed TRAAM (transduction recursive auto-associative memory) model. We demonstrate strong connections between music and language processing and learning—the freestyle learning capability arises from exactly the same compositional structure learning model that autonomously learns semantic interpretation / translation models as well as other music improvisation models.

1. INTRODUCTION

Musical improvisation shares a great deal with linguistic expression. In challenge-response or call-response improvisation, to respond creatively requires on-the-fly composition of many small parts, each of which is related in some interesting way to the challenge, while cohering together in contextually satisfying ways—which in rap music includes fluency, salience, metrical or syntactic parallelism, rhyming at various points, and so on. It is no accident that creative construction of music is called *composition*.

Taking the first line of the following rap as a challenge, what would be involved in improvisationally composing the second line as a response?

1: *focus is formed by flaunts to the soul, souls who flaunt styles gain praises by pounds*
2: *common are speakers who are never scrolls, scrolls written daily creates a new sound*
—De La Soul, "The Magic Number"

A rapper would need to compose a response integrating many complex factors, such as:

- the response line should somehow be salient to the challenge line
- some phrases within the response line can somehow be salient to corresponding phrases within the challenge line—e.g., 'focus is formed by flaunts to the soul' is salient to 'common are speakers who are never scrolls'
- some individual words within the response line can somehow be salient to corresponding words within the challenge line—e.g., 'is' is salient to 'are', and 'who flaunt styles' is salient in a different way to 'written daily'
- the response line should flow fluently (yet sometimes may allow for stylistic ungrammaticality, disfluencies such as stuttering, or slang constructs)
- some phrases within the response line can use metrical parallelism to corresponding phrases within the challenge line—e.g., 'scrolls written daily creates a new sound' has a close meter to 'souls who flaunt styles gain praises by pounds'
- some phrases within the response line can use syntactic parallelism to corresponding phrases within the challenge line—e.g., 'focus is ...' is syntactically parallel to 'common are ...'
- the response line should typically rhyme with the challenge line—e.g., 'pounds' rhymes with 'sound'
- some words or phrases within the response line may also be made to rhyme with the challenge line—e.g., 'soul' rhymes with 'scrolls', and 'gain praises' rhymes with 'creates'

These are all soft preferences, not hard constraints. Each choice influences the others, creating combinatorial dependencies that make good improvisation exponential complex in the absence of more sophisticated models.

2. FREESTYLE: A RAP BATTLE BOT

We demonstrate a rap battle bot that—starting with zero knowledge, not even a pronunciation dictionary—learns to integrate such factors entirely by itself, guided by a newly enhanced bilingual recursive neural network model called TRAAM (transduction recursive auto-associative memory). For each challenge-response pair it sees in training, the model learns relationships between what it suspects are associated fragments at all these various levels of granularity between the the challenge and response. The hierarchically *compositional* relationship can be represented as a tree whose leaves are the individual words or phrases associated with each other by dint of salience, syntactic function, or rhyme, and whose internal nodes are progressively longer compositions of the shorter chunks:

```
[ [ [ 'focus'/'common'
      [ 'is'/'are'
         [ [ 'formed by flaunts'/'speakers'
             'to the'/'who are never' ]
           'soul'/'scrolls' ] ] ]
      ','/',' ]
  [ [ 'souls'/'scrolls'
     'who flaunt styles'/'written daily' ]
    [ 'gain praises'/'creates'
      'by'/'a new' 'pounds'/'sound' ] ] ] ]
```

Relating one's response to any given challenge using bilingual parse trees like these makes clear the many constituent relationships between associated fragments. The bilingual formalism represents a relation between a challenge language and a response language. We have developed transduction grammar induction models that learn symbolic transduction rules of this form for rap battling [7], by gereralizing the syntax directed transduction grammars (SDTGs) of classic formal language theory [2] to be stochastic. We restrict induction to the subclass of SDTGs known as inversion transduction grammars or ITGs [4], because a wide range of polynomial time learning and prediction algorithms exist (unlike SDTGs), and yet ITGs have long been empirically demonstrated to possess attractive language universal properties for machine translation [6].

However, the computational complexity of explicitly enumerating hypotheses rapidly becomes exponential for symbolic transduction grammar induction, when nonterminal category induction is required to take into account so many contextual factors. We therefore have developed TRAAM [1] to be a recursive neural network encoding of ITGs, that still models the compositionality of associations between challenge fragments and response fragments—but using distributed vector representations that can be efficiently trained via recursive backpropagation to better implicitly generalize over soft regions or neighborhoods of hypotheses. TRAAM can be seen as a bilingual generalization of Pollack's [3] RAAM (recursive auto-associative memory) model—where RAAM approximates monolingual context-free grammars, TRAAM approximates bilingual transduction grammars.

We demonstrate for the first time a real-time version of FREESTYLE that exploits speed and response quality gains from a new generation of our TRAAM based model. Training convergence has been accelerated by an order of magnitude, leading to improved optimization accuracy. Analysis of the trained TRAAM networks indicates excellent formation of clusters representing meaningful abstract patterns. This in turn now allows TRAAM to guide the ITG based rap battle bot very heavily, producing significantly improved responses to challenges. The encouraging results suggest our new model may apply to improving musical improvisation in other genres, for example flamenco [5], as well.

3. ACKNOWLEDGMENTS

This material is based upon work supported in part 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; by the European Union under the FP7 grant agreement no. 287658; and by the Hong Kong Research Grants Council (RGC) research grants GRF620811, GRF621008, GRF612806, FSGRF13EG28, FSGRF14EG35. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA, the EU, or RGC.

4. REFERENCES

[1] Karteek Addanki and Dekai Wu. Transduction Recursive Auto-Associative Memory: Learning bilingual compositional distributed vector representations of Inversion Transduction Grammars. In *Proceedings of SSST-8, Eigth Workshop on Syntax, Semantics and Structure in Statistical Translation (at EMNLP 2014)*, pages 112–121, Doha, Oct 2014.

[2] Philip M. Lewis and Richard E. Stearns. Syntax-directed transduction. *Journal of the Association for Computing Machinery*, 15(3):465–488, 1968.

[3] Jordan B Pollack. Recursive distributed representations. *Artificial Intelligence*, 46(1):77–105, 1990.

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

[5] Dekai Wu. Simultaneous unsupervised learning of flamenco metrical structure, hypermetrical structure, and multipart structural relations. In *14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, Curibata, Brazil, Nov 2013.

[6] Dekai Wu. The magic number 4: Evolutionary pressures on semantic frame structure. In *10th International Conference on the Evolution of Language (Evolang X)*, Vienna, Apr 2014.

[7] Dekai Wu, Karteek Addanki, and Markus Saers. Modeling hip hop challenge-response lyrics as machine translation. In *Machine Translation Summit XIV (MT Summit 2013)*, Nice, France, Sep 2013.