TACKLING NATURAL LANGUAGE GENERATION CHALLENGES AT NARRATIVE SCIENCE

Mike Pham & Clayton Norris

Oct 19, 2017

Narrative Science

OVERVIEW

WHAT IS QUILL?

Quill is an Advanced Natural Language Generation (NLG) platform

NLG A form of artificial intelligence (AI) that automatically produces language from structured data.

intent-driven Advanced NLG uses intent, or what you want to know, as its guide from the very beginning.

HOW IS THIS DIFFERENT THAN OTHER NLG?

So what?

HOW IS THIS DIFFERENT THAN OTHER NLG?

So what?

- How is this different than Amazon sending me a templated email receipt of my recent purchases?

HOW IS THIS DIFFERENT THAN OTHER NLG?

So what?

- How is this different than Amazon sending me a templated email receipt of my recent purchases?
- What about all those neural nets generating facebook posts that sound eeriely like my previous posts?

HOW IS THIS DIFFERENT THAN OTHER NLG?

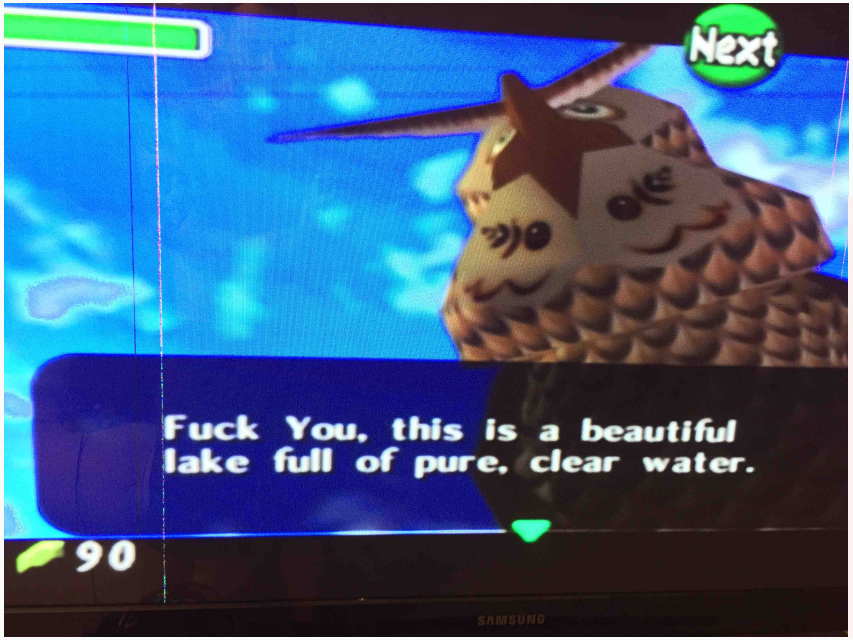
So what?

- How is this different than Amazon sending me a templated email receipt of my recent purchases?
- What about all those neural nets generating facebook posts that sound eeriely like my previous posts?

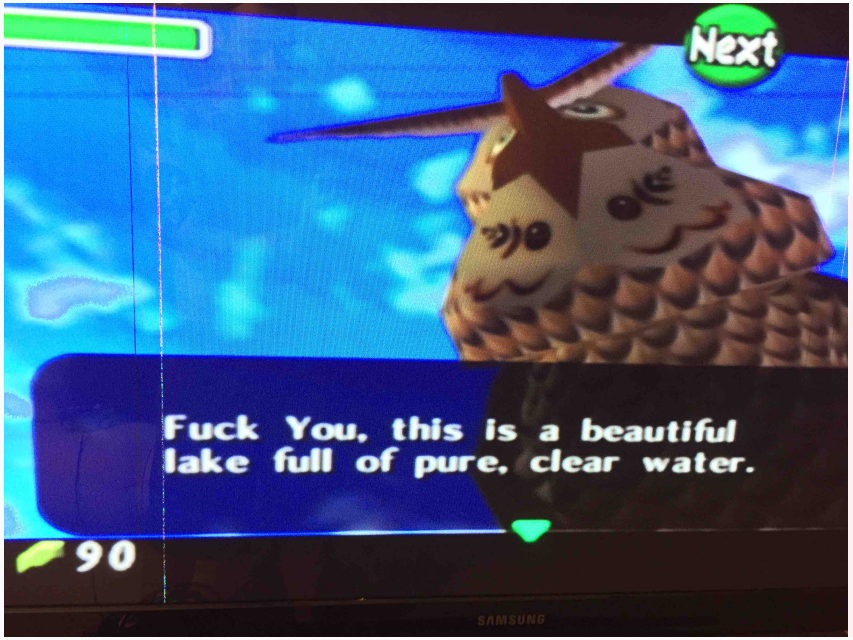
OTHER NLG

trigger warning: offensive language

trigger warning: offensive language



trigger warning: offensive language



- What seems off here?

LIMITATIONS OF NLG

- Video games conversations have complex decision trees

LIMITATIONS OF NLG

- Video games conversations have complex decision trees
  - Can result in very good and/or appropriate language
  - ...but often is mad-libby
  - Flexibility and linguistic creativity is limited and/or unscaleable in production

LIMITATIONS OF NLG

- Video games conversations have complex decision trees
  - Can result in very good and/or appropriate language
  - ...but often is mad-libby
  - Flexibility and linguistic creativity is limited and/or unscaleable in production
- Neural nets can learn from data to generate new language

LIMITATIONS OF NLG

- Video games conversations have complex decision trees
  - Can result in very good and/or appropriate language
  - ...but often is mad-libby
  - Flexibility and linguistic creativity is limited and/or unscaleable in production
- Neural nets can learn from data to generate new language
  - Can often produce highly natural and nuanced language
  - but has no idea what it's saying
  - and we have no idea why it's saying it either

LINGUISTICALLY SAVVY & INTENT-DRIVEN

HAMMERS AND NAILS

3 STRATEGIES

- hardcode/exhaustive listing
- rules/principled based
- machine learning

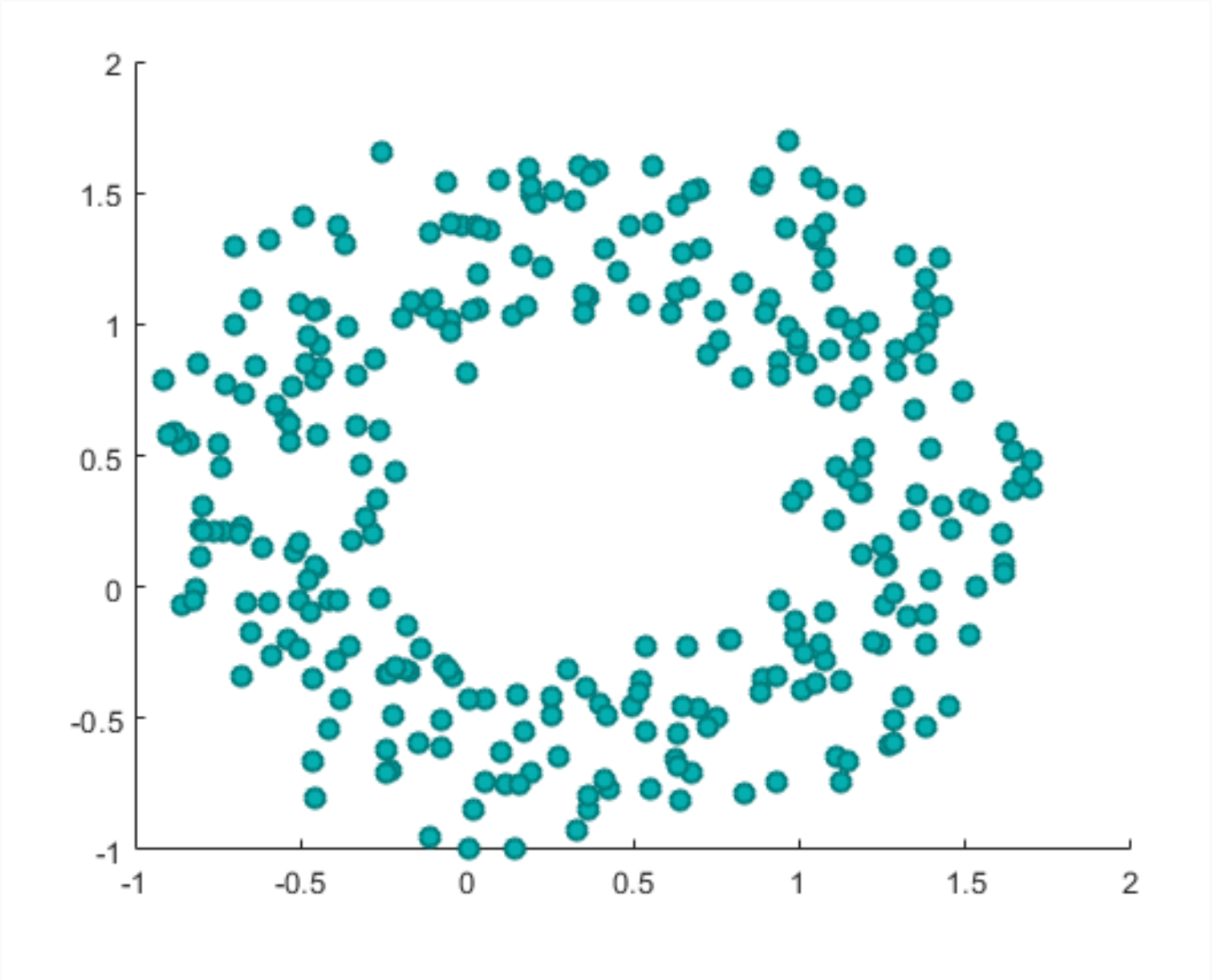
3 STRATEGIES

- hardcode/exhaustive listing
- rules/principled based
- machine learning

Where does each strategy fit best? How to combine them?

PERSPECTIVE

What do you see? How would you recreate this data distribution?



OUTLINE OF TALK

Overview

Irregular Verbs

Pronouns

Sentence Selection

Conclusion

Conclusion

IRREGULAR VERBS

VERB INFLECTION

- A single verb can have various **word forms**:

(1) CREATE

a. create, creates, created, creating

b. creator, creation, creative, creatively

- (1a) is an example of **inflectional morphology**
  - expresses grammatical features
  - (usually) doesn't change basic meaning or part of speech

GRAMMATICAL FEATURES

- **Grammatical features** are properties that the grammar of any language tracks and manifests
- Some features that English is sensitive to:
  - **number**: dog, dogs
  - **tense**: create, created
  - **gender**: he, she
  - **person**: we, yall, they
  - **mass/count**: 3 books, *3 bloods
  - **case**: I, me, my, mine

INFLECTIONAL PARADIGMS

- Word forms can track multiple features at once
- This can be tracked within an **inflectional paradigm**

CREATE

	Present	
	singular	plural
1	create	create
2	create	create
3	creates	create

	Past	
	singular	plural
1	created	created
2	created	created
3	created	created

INFLECTIONAL PARADIGMS

- Word forms can track multiple features at once
- This can be tracked within an **inflectional paradigm**

CREATE

	Present	
	singular	plural
1	create	create
2	create	create
3	creates	create

	Past	
	singular	plural
1	created	created
2	created	created
3	created	created

- Only 3rd person singular is different – this looks easy!
  - Just add -s to the 3.sg present form and -d to all past forms!

Unfortunately, we all know there are **irregular verbs** in English

BE

	Present	
	singular	plural
1	am	are
2	are	are
3	is	are

	Past	
	singular	plural
1	was	were
2	were	were
3	was	were

IRREGULARS

Unfortunately, we all know there are **irregular verbs** in English

BE

	Present	
	singular	plural
1	am	are
2	are	are
3	is	are

	Past	
	singular	plural
1	was	were
2	were	were
3	was	were

- Darn, how do we get am or was from be?

BEST STRATEGY

- There are rules for regular morphology
- Which verbs are irregular seems arbitrary
- How irregular verbs inflect also seems arbitrary
- Rules might be tough to derive
- Machine Learning may work, but do we actually want to overfit data?

FINITE PROBLEM SET

- Wikipedia lists about 200 English irregular verbs, including shrive, stave, gild
- This is finite set, and most words aren't even that relevant
- Verb dictionaries exist
- There are subgroups within the irregulars
- It is feasible to exhaustively hardcode a list of all irregulars without rules or ML
- This overfits the dataset, but by definition we don't expect irregular verb patterns to be productive

PRONOUNS

SENTENCE SELECTION

GRAMMATICALITY VS STYLE

Grammaticality: only grammatical and accurate sentences should be **generated**

Sentence selection: the stylistically best sentence from the set of grammatical candidate sentences should be **selected**

GRAMMATICALITY VS STYLE

Grammaticality: only grammatical and accurate sentences should be **generated**

Sentence selection: the stylistically best sentence from the set of grammatical candidate sentences should be **selected**

- but what determines a stylistically ‘good‘ sentence?

SUBJECTIVE AXES OF ‘GOODNESS‘

- Most native speakers will agree when a sentence is grammatical
- But style is vague and elusive, varying from person to person

SUBJECTIVE AXES OF 'GOODNESS'

- Most native speakers will agree when a sentence is grammatical
- But style is vague and elusive, varying from person to person
- Which do think is the best sentence?

(2) Aaron Young generated $3M in revenue in 2016.

(3) Aaron Young's revenue was $3M in 2016.

(4) Revenue for Aaron Young was $3M in 2016.

(5) In 2016, Aaron Young generated $3M in revenue.

(6) Aaron Young's 2016 generated revenue was $3M.

CONCLUSION

TYPOGRAPHY

```
The theme provides sensible defaults to \emph{emphasis}
text, \alert{accent} parts or show \textbf{bold} results.
```

becomes

The theme provides sensible defaults to emphasis text, accent parts or show **bold** results.

CONCLUSION

SUMMARY

Get the source of this theme and the demo presentation from

`github.com/matze/mtheme`

Questions?