НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

Генерация скороговорок с помощью марковских цепей и LSTM

Наталья Семенова

Национальный исследовательский университет
«Высшая школа экономики»

31 марта 2018 г.

Что нужно сделать:

План

- Выкачать скороговорки на английском с сайта http://www.uebersetzung.at/twister/en.htm Всего 593 штуки.
- Сгенерировать скороговорки пословно с помощью марковской цепи
- Сгенерировать скороговорки посимвольно с помощью LSTM
- Посмотреть результаты

English Tongue Twisters

1st International Collection of Tongue Twisters
www.uebersetzung.at/twister/en.htm © 1996-2015 by Mr.Twister

Please click on the number above the tongue twister for a rough translation;
You can use this form to submit a new tongue twister.

1

Peter Piper picked a peck of pickled peppers.
A peck of pickled peppers Peter Piper picked.
If Peter Piper picked a peck of pickled peppers,
Where's the peck of pickled peppers Peter Piper picked?

2

I saw Susie sitting in a shoe shine shop.
Where she sits she shines, and where she shines she sits.

3

How many boards
Could the Mongols hoard
If the Mongol hordes got bored?

Марковские цепи:результат

примеры скороговорок

- "Craig Quinn's quick witted cricket critic Hitchcock Hawk Watch"
- 'Babcock Peggy Babcock Peggy Babcock If Kanta can cans'
- 'Seven sleazy shysters in a Hottentot tot could But'
- 'Brain Lock Which witch wishes I ever felt I'
- 'Lou slued loose the third three-toed tree toads toes'
- "I'm a tie and coldest frosts Fuzzy Wuzzy was"
- 'Sheep Sheets Cheap Sheep Association The greedy Greek grapes'
- + работает быстро
- - иногда выдает дикую чушь
- в планах еще было попробовать реализацию Markovify

LSTM:параметры

https://machinelearningmastery.com/text-generation-lstm-recurrent-neural-networks-python-keras/

```
# определяем LSTM модель (2 слоя)
model = Sequential()
model.add(LSTM(256, input_shape=(X.shape[1], X.shape[2]), return_sequences=True))
model.add(Dropout(0.2))
model.add(LSTM(256))
model.add(Dropout(0.2))
model.add(Dense(y.shape[1], activation='softmax'))
model.compile(loss='categorical_crossentropy', optimizer='adam')
```

epochs = 10, batch size = 128

LSTM:результат

примеры скороговорок и проблемы

- ban the sas the saw the saw the saw the saw the saw the saw
- the saw io a waw io a waw io a waw wood saw wood saw wood s
- she shitt she shitt she shitt she shitt she shitt she shitt
- cutter of the siett the shett she shitt she shitt she shitt
- dia bld bld bla bla bla bla bla bla bla bla bla bla bla bla
- в целом результат плохой, а все потому что у меня слишком маленький корпус для этого метода, на корпусе из 150 000 символов он работает гораздо лучше + можно было сделать больше эпох, но мой компьютер тогда, скорее всего, сгорел бы

Проблемы оценки
и как можно еще генерировать скороговорки

Существуют методы оценки сгенерированных текстов: метод шинглов, прямое сравнение, они используются для оценки текстов описаний товаров и т.д.
Идея для продолжения: можно использовать фонологические фичи для стохастической генерации (+ гласные буквы в основном не имеют такого значения для скороговорки, как согласные, на тренировку произношения которых и нацелены скороговорки). Скорее всего так результат будет значительно лучше.