Markov chain text generation with inferred parts of speech

Travis Mick

June 29, 2018

Abstract

Markov chains have commonly been used in text generators for applications such as chat bots, typically implemented by utilizing the likelihood of k-word sequences to inform the iterative production of words. In this document, we outline a technique to improve the quality of such systems' outputs by leveraging inferred knowledge about the parts of speech for each word in a sequence.

1 Introduction

Traditional Markov chaining text generation leverages a database indexing the frequency $F^W(w)$ of each observed k-word sequence $w = (w_0, w_1, \ldots, w_{k-1})$. The database can be queried for a likely sequence given some prefix. Given a prefix $w' = (w_0, w_1, \ldots, w_{j-1})$ of length $j < k$, a candidate sequence c with prefix w' can be assigned a score as

$$S^W(c) = \frac{F^W(c)}{\sum_{c' \in C^W(w')} F^W(c')} \tag{1}$$

where $C^W(w')$ is the set of all possible sequences with the prefix w'. Ostensibly, the highest scored sequence is returned by the text generation algorithm, and an output string can be obtained by iterating this process while maintaining a rolling window of k words.

Unfortunately, the choice of k introduces a trade-off between overfitting and underfitting. Too low of k would result in unnatural sequences of words being produced, and a k too large would yield few novel results. Therefore, we suggest augmenting the k-word Markov chain with additional information about the parts of speech of the constituent words to increase output quality given small k.

2 Generation model

In addition to a word-sequence frequency database, we will utilize a database of frequencies for part of speech sequences, as well as a database of frequencies of observations of a word being used as a particular part of speech.

First, let $F^P(p)$ give the frequency of observations of k-length part of speech sequences $p = (p_0, p_1, \ldots, p_{k-1})$. Then, let $P(w_i, p_i)$ give the frequency of observations of word w_i used as part of speech p_i. We can now define the likelihood of a word sequence w corresponding to a part of speech sequence p as

$$P^A(w, p) = \frac{F^P(p)}{\sum_{p' \in X^k} F^P(p')} \cdot \prod_{0 \leq i < k} \frac{P(w_i, p_i)}{\sum_{p'_i \in X} P(w_i, p'_i)} \tag{2}$$

Where X is the set of all parts of speech and X^k is the set of all possible part of speech sequences of length k. Note that we have utilized information about the correspondence between parts-of-speech and words, as well as the likelihood of the part of speech sequence itself. We can leverage this to produce an augmented score for a candidate word sequence c with prefix w'.

$$S^A(c) = S^W(c) \cdot \max_{p \in X^k} \Big(P^A(c, p) \Big) \tag{3}$$

Note that we score the word sequence based on the best possible part of speech assignment as informed by Eqn. 2.

3 Learning model

In many applications, Markov chain text generators attempt to learn from observations in real time. While the learning process to inform F^W is straightforward, we must define a mechanism to label observed word sequences with parts of speech in order to update F^P.

When attempting to assign a part of speech sequence to a word sequence, we must come up with a scoring mechanism to implement inference, as was done with word sequences in Eqn. 1. Given a part of speech sequence prefix $p' = (p_0, p_1, \ldots, p_{j-1})$ of length $j < k$ and a corresponding word sequence w, a candidate part of speech sequence c with prefix p' can be assigned a score as

$$S^P(c) = P^A(w, c) \cdot \frac{F^P(c)}{\sum_{c' \in C^P(p')} F^P(c')} \tag{4}$$

where $C^P(p')$ is the set of all possible sequences with the prefix p'. We have again leveraged Eqn. 2 to provide weight to both observed part of speech

sequences and the correspondence between parts of speech and words. Note that because our learning model requires prior observations in order to infer parts of speech for new observations, the system must be trained prior to being exposed to input.