Carl Bildt Tweets: A comparison of regular and constrained Markov chain for text generation

Group Ain't intelligent

Viktor Björkholm	Jesper Bränn	Daniel Duberg	Jakob Tideström
92-11-17	92-09-30	93-01-13	90-10-04
viktorbj@kth.se	jbrann@kth.se	dduberg@kth.se	jakobti@kth.se




Abstract

Skriv sist, nr vi vet vad vi har skrivit om (y)

1 Introduction (1–2 pages)

There are a few approaches to generating text to make it seem like it was written by a human. One of these approaches is using Markov chains or more commonly known as n-grams. These n-grams take n words in sequence and uses a corpus of text to guess what the most probable next word is. Using a larger n means that more text is copied straight from the corpus, however this also means that there is a higher likelihood that the text being generated is meaningful.

In this paper we will generate twitter messages, tweets, using Markov chains. We explore the possiblity of using a unigram, that is a 1-gram, with instead of words using parts-of-speeches to keep the grammatical structure of the corpus as well as keeping a **trigram** of words from the corpus to ensure that the semantical structure is also kept intact. This will be compared to the same model which is then constrained to a few predetermined criteria that we have selected, those being: The length and end state of the Markov chain. These criteria allows us to have a finer control over the Markov chain.

The concepts brought up in research by [1] are used and tweaked for our use case. Since twitter restricts messages to 140 characters we need to ensure that while generating messages long enough for the text to have a meaning, they should not go over the limit.

In this paper we compare the differences in result when using an n-gram on a corpus versus the result of using a constrained Markov model (uni- or bi-gram?). For the second implementation, basing our research on the work described in [1], We opted for generating tweets with the same structure in the form of Parts-of-Words as our corpus dictated, in the second implementation basing our work on the research by .

– Lite info om vra resultat –

1.1 Contribution

Varfoer aer det vi gjort relevant?

This paper contributes to the research field by Putting constraints on markov models seems to be a relatively recent approach. Our goal is to explore the **impact** of using constrained markov models to generate natural language.

1.2 Outline

Bla bla bla bla bla bla bla Section 2 bla bla bla bla bla bla bla bla bla Section 3 bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla

Section 4 bla bla bla bla bla bla Section 5 bla bla bla bla bla

2 Related work

Our work on the constrained Markov models are built upon [1] work, where the authors generated lyrics from different artists using Markov models with constraints. These were to be generated in a specific style and with a rhythm. However those criteria were not necessary for our work but we utilized the knowledge of constrained Markov models from them.

3 My method

Our method consisted of building a transition matrix as a unigram of parts-of-speech (POS) together with a bigram with actual words from our corpus of text. From the transition matrix we generated a constrained transition matrix based on the amount of number of words we wanted the tweet to contain **(and other criteria)**. The constrained matrix was generated using the method described in the work by [1]. We generalized the method and made it work for our transition matrix even though it consisted only of parts-of-speech.

The different constrains for creating a tweet that we had to consider were that they can not be longer than 140 characters, they have to end with an end-symbol (dot, exclamation mark or question mark) and they should probably have a reasonable minimum length. When iterating through a corpus we are using a POS-tagger to identify the different types of words and coding to be able to build a transition matrix for the sequences for the different word types (the probability for a noun to be followed by a word for an example). The next step is to implement our constraints on the transition matrix. The constraint for the length of the tweet caused us problems since we only know the types of the words and would not be in touch with the length of the specific words that are chosen for our tweet. We decided to approximate this constraint to a limited number of words in our tweet.

#yolo

3.1 Implementation

The implementation relied on a Part-of-Speech tagger from Stanford's Natural Language Processsing group.

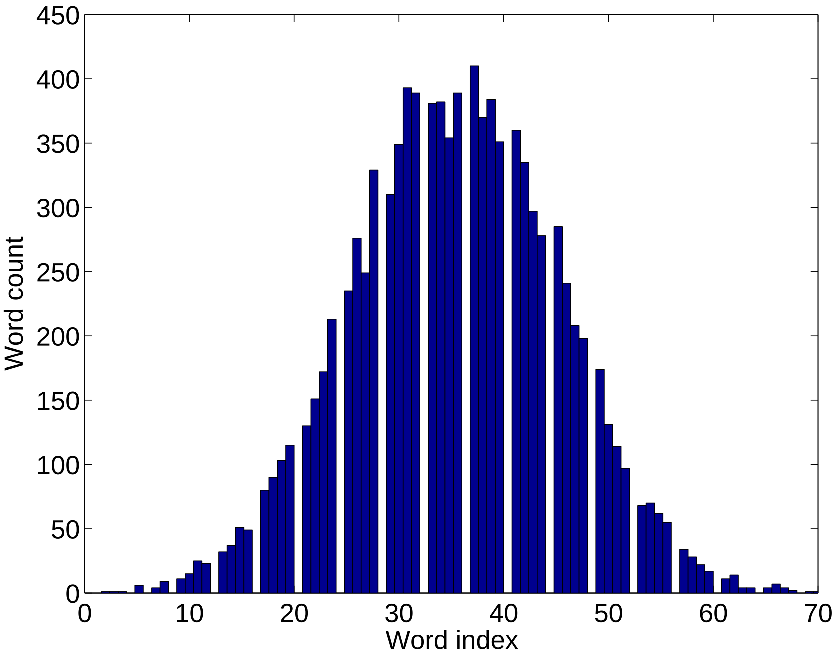

Figure 1: A description that makes browsing the paper easy and clearly describes what is in the picture. Make sure that the text in the figure is large enough to read and that the axes are labelled.

4 Experimental results

Some images here and stuff would be nice.

4.1 Experiemntal setup

Bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla

4.2 Experiment ...

Bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla

Bla bla bla bla bla Figure 1 bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla

Bla bla	Bla bla	Bla bla
42	42	42
42	42	42

Table 1: A description that makes browsing the paper easy and clearly describes what is in the table.

Bla bla bla bla bla Table 1 bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla

5 Summary and Conclusions

Bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla

References

[1] Pierre Roy Gabriele Barbieri, Francois Pachet and Mirko Degli Esposito. Markov constraints for generating lyrics with style. 2012.