

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)

,

This paper aims to develop an understanding on refining natural language text generation. Natural Language Generation (NLG) is an area of research within the field of Artificial Intelligence. The aim is to generate text that is semantically correct in order to make communication with computer systems more natural and understandable for users.

Within this paper we show the difference in quality of two different approaches to text generation. 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. We will contrast this method with using constrained Markov chains. The main constraint of the Markov chains is that two words following each other will have the same sequence of part-of-speech as the corpus. Part-of-speech is a concept within NLG that divides a text into the different linguistic categories of the words within it.

We aim to show that using this constraint upon the Markov chain, sentences will have a greater diversity but still be as semantically correct as just using a n-gram. Since the problem with larger n:s within n-grams is that text is copied straight from the corpus, the constraint will help us create semantically correct sentences without taking word sequences straight from the corpus.

To be able to show differences in these two approaches we generate Twitter messages, so called tweets. We build upon the work by Barbieri et al., 2012 to implement our own constraints.

1.1 Contribution

We have implemented a unigram of part-of-speech on to a bigram in order to observe the difference in the result. As mentioned above a problem with n-grams with too high n:s is that they will simply copy parts of the corpus if said corpus does not contain a large variation of similar sentences, so that a given start of n words does not automatically lead to one sentence finish. i.e. a larger corpus is needed for larger n:s. To be able to keep a smaller

and diverse corpus we applied the unigram ontop of the n-gram to allow the program to select words of a common word type order but perhaps not words that have occurred after each other naturally in the corpus.

Our main contribution to the field is that we have tried putting this unigram constraint upon the regular Markov chain and doing it for short message generation. The research area we base this paper on focus on generating text that fits a theme, or rhythm whereas we generalize the concept.

A lot new research involves the constraining of Markov chains to produce text that fits different molds than just text generated from a corpus. It is important to be able to generate text that is

1.2 Outline

We bring up the relation of our work to some other work in the field Barbieri et al. [2012] in section 2 and explain how that research has affected ours. In section 3 we then go through the details on how the algorithm works, we also give examples to explain in detail what the difference between the two methods we are comparing. This is complemented with details regarding our specific implementation of the algorithm in section 3.1. The data from running the algorithm is explained, reviewed and evaluated in section 4.

2 Related work (1–3 pages)

Our work on the constrained Markov models are built upon Barbieri et al. [2012] 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, to simulate real lyrics from a large span of artists. They start in a standard fashion by building up a corpus of content that they want to imitate. In their case they built up a corpus consisting of the collected work of each artist. So when generating lyrics for bob dylan as an example, the corpus consisted of 7600 unique words, a total of 96 089 words and 12 408 verses. They proceeded with applying constraints on a bigram to meet their demands on the generated text to have a certain style and to rhyme. Some of the constraints they apply on their bigram is that some words need to have the same meter as in the original song, but also that the text generated needs to fit a part-of-speech template and a rhythmic template that comes from the song the newly generated lyrics is to be based on. The templates are not generated by their algorithm, but rather handcrafted to get a proper result. Our take from this paper was to be able to apply constraints upon a bigram, or rather a Markov chain in general. Similarly we utilize a part-of-speech template of sorts, but we diverge a bit from this paper.

In Shannon [1948] Shannon goes through the first basic steps in how to naively generate text from a corpus with Markov chains. He explains in depth how Markov chains can be used to generate real sounding English words from letters and also sentences from words. His approach was novel at the time but it is a staple in modern text generation, and also in this project it is used as one of the corner stones in the algorithm.

3 My method (1–4 pages)

Our method that forms the basis of this paper is generating Twitter messages with the use of both Markov chains and constrained Markov chains and to compare the two methods with each other. The theory is that constraining an Markov chain will yield better and more diverse text being generated. In section 4 where we explain our experiments and data we try out different orders of the Markov chain, for the sake of explanation we assume a first order Markov chain is used to explain how the method works. In a limited corpus such as this one:

| | |
|------------------------------|---|
| “Rolf has a dog.” | Noun → verb → singular quantifier → noun |
| “Rolf owns a dog.” | Noun → verb → singular quantifier → noun |
| “Rolf can not walk his dog.” | Noun → modal → adverb → verb → pronoun → noun |

This corpus generates a Markov chain that looks like figure 1a, with the edges being the probabilities for the specific transitions. In the figure the specific part-of-speech is included in the states beneath the words from the corpus. If we then add constraints from a transition matrix, built up from POS-analysis of the same corpus we are given the new chain that is seen in figure 1b.

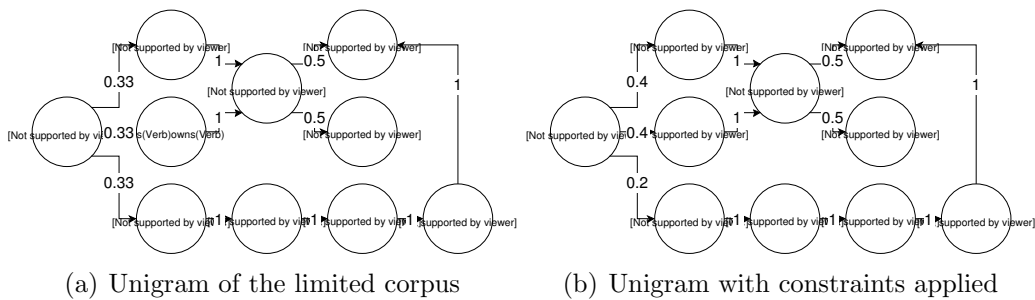


Figure 1: Unigrams

We can see in figure 1b that since both “has” and “owns” are verbs they are more probable to occur than “can”, this is reflected in the new edges. This happens because the part-of-speech “verb” is more likely to follow NNP according to our transition matrix, and thus “has” and “owns”, who are both verbs become even more likely to follow NNP (Rolf). This method can then be further applied to a bigram, a trigram or any n-grams that follows. Our method however will only have an unigram for the transition matrix, even if the Markov chain of words from the corpus is longer, the transitions are only

observed with one previous state in consideration.

Corpus:

In order to generate tweets, we first discussed the problem regarding the semantical corectness of a generalized tweet. The average level is according to our own experience far from sematically correct, which isn't a problem in understadability for an experienced Twitter user, but is a problem for our POS who would not recognize the words. Even if it would give it an "unknown"-tag, we would not be able to predict any kind of results.

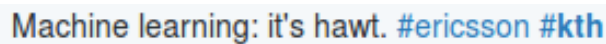
The image shows a single line of text, a tweet, with a light blue background. The text is "Machine learning: it's hawt. #ericsson #kth". The words "Machine learning:" are in a dark blue font, "it's hawt." is in a lighter blue font, and the hashtags "#ericsson" and "#kth" are in a darker blue font.

Figure 2: An example tweet

To solve this problem approximatily, we decided to generate tweets for a specific user who mostly uses correct grammar and semantics when tweeting and tweets in english. Our option fell on Carl Bildt, former foreight minister of Sweden, because of his active use of Twitter, that he tweets in english and that most of his tweets are in gramatically correct english.

3.1 Implementation (0–2 pages)

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

We start by collecting data, in the form of tweets, by a target person, Carl Bildt, to be used as the corpus of our tweet generator.

Each tweet is run through Stanford's speech tagger to get the lexical class of each word depending on context. We then use these lexical classes as states for the transition matrix, more specifically they're used to see what lexical class usually follows any given lexical class or terminates a sentence.

While iterating through the gathered data we store what word follows any given two words, including sentence terminating symbols such as '.' or '!' and store it as a bigram.

When generating tweets two methods are used, first the bigram gets to select which words to use entirely on its own digression (Markov chain). Then the process is repeated but this time the bigram is constrained by which lexical class of words it is allowed to choose from, that class being whichever

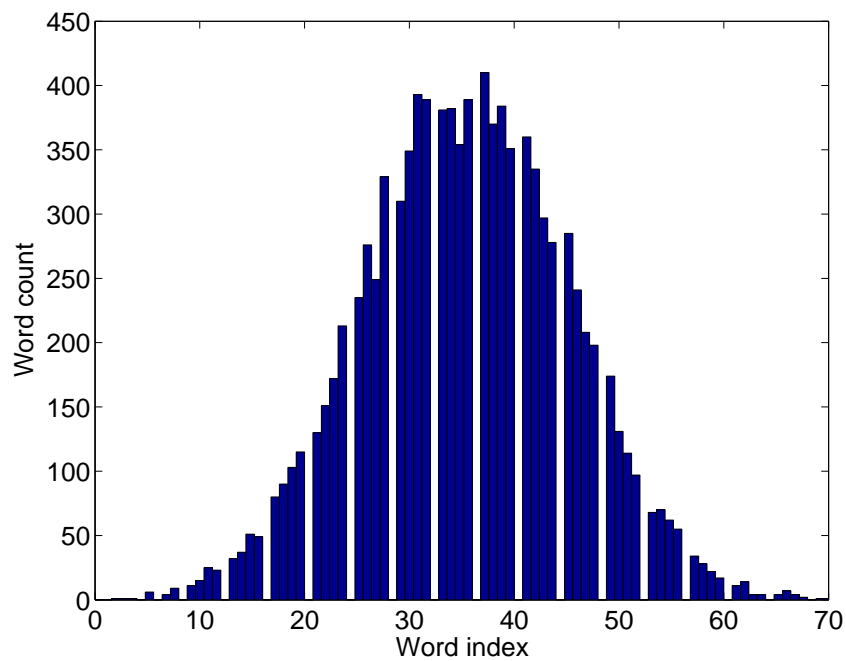


Figure 3: A description that makes browsing the paper easy and clearly describes what is in the picture. Make sure that the text in the figure is large enough to read and that the axes are labelled.

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

Gabriele Barbieri, François Pachet, Pierre Roy, and Mirko Degli Esposti. Markov constraints for generating lyrics with style. 242:115–120, 2012. URL <http://dblp.uni-trier.de/db/conf/ecai/ecai2012.html#BarbieriPRE12>.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, July, October 1948. URL <http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf>.