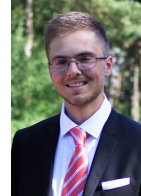


A Comparison of N-gram and Hidden Markov Model Text Generators

Akash Kumar Dhaka	Dario Vidas	Stefan Annell
1990-02-26	1992-06-24	1987-09-30
akashd@kth.se	vidas@kth.se	steann@kth.se



Abstract

Natural Language Generation (NLG) is a Natural Language Processing (NLP) task of generating natural language. The problem is well known and well researched, but the field is far from being fully explored. In this paper, we propose an idea of building a text generator based on Hidden Markov Model (HMM). HMM allows us to append words in a sentence that depend on a whole sequence until that point, unlike other methods, such as n-gram, which use only a certain number of words to predict the next one. . This type of text generators may outperform simple ones and match the performance of a more complex generators.

1 Introduction

Natural Language Generation(NLG) is the field of generating natural language for a computer from a knowledge base or some information that the system can understand. Some fields where the NLG is beeing applied are weather forecast generation, summarize statistical data extracted from a database and describing chains of reasoning performed by a system [?].

This paper is about designing, implementing and comparing different methods of generating random sentences when fed in by a specific corpus. The generated sentences should be as correct as possible grammatically and semantically. Which means that Natural Language Processing (NLP) needs to be implemented for building the corpus and tagging / structuring the words, allowing our system to form sentences that are meaningful and semantically as correct as possible. The corpora used were of different sizes to test our system properly.

This is very interesting field of research and a good topic to study because it incorporates several complicated tasks that are useful in wide variety of applications. First of all, it is very difficult for a computer to understand the content of a text, there has been a lot of work done in this field. It is also hard for a computer to construct sentences that are correct semantically because of the complexity of a natural languages, such as English as mentioned in [?]. The evaluation of how correct a sentence is also very challenging and a field in its own. This fact raises another problem, which is the infinite number of combinations of words. Furthermore, the aggravating circumstance that the language (both dictionary and the grammar) is continuously evolving, makes it even harder to define a specific set of rules that apply all the time, considering that our datasets (news) are written by numerous authors and in different time instants. NLG systems broadly have three subcomponents:text planning, sentence planning and surface realisation. In text planning, content and target of the text are determined, During sentence planning, syntactic and lexical means are ascertained whereas in realisation, actual words are strung together to give form to the basic structure which comes out of sentence planning. It is also preferred if the results are benchmarked in some systematical way, either manually (which is not preferred), semi-manually where the system might give suggestions and the operator help the computer or it is done completely automated (which is also quite hard).

1.1 Contribution

It is very hard for a computer to communicate with a human since the language is very complex and it is hard to replicate and the language is one

of the big borders between computers and humans which makes this field of research very interesting. The method that we are presenting is a new and interesting way to generate text focused on a topic using Hidden Markov Models (HMM). We used three variants of HMM for this task- unsupervised HMM, supervised HMM and chunked supervised HMM. We also generated sentences by trigrams and evaluated the performance of all the generators with human written sentences.

1.2 Outline

In Section ?? previous work is recognized and the papers we have researched to support our work is presented. Section ?? contains an overview on how we are going to solve the problem, and why we are choosing a specific method and also how we are implementing the system. The result of the implementation is presented in Section ?? and the whole paper is summarized and concluded in Section ??.

2 Related work

The field of NLG and NLP are two very large fields that are subfields of artificial intelligence and computational linguistics. A lot of research has been done in these fields, although the fields are far from fully explored. Some of the earliest work in this field was performed in the 1970's by people like Goldman and Davey, where they helped to define the main problems in NLG as explained in [?, p. 19-20], where they also mention that NLG started to be researched a lot during 1980's by people like McKeown and Appelt that had a big impact on how the subsequent research was going to be performed on NLG. The field had a big increase in the 1990's where a lot of real-world applications were made, like the weather forecast system FOG.

The paper Building Applied Natural Language Generation Systems [?] is about the architecture and how to structure your NLG system when designing it. They describe six main tasks that need to be performed when generating a text; content determination, discourse planning, sentence aggregation, lexicalization, referring expression generation and linguistic realisation. They also discuss the sub components the system needs to design and the algorithms, using which the different subproblems can be solved. The main use of this paper is the structure of the generation part of the system, but also suggestions and pointers on how the arising problems can be solved.

Usage of HMMs in text generation is mentioned in [?] where the model is used in a task-oriented environment. Authors describe HMM generation

process to be very similar to POS (part of speech) tagging. States represent words, and state sequences, word phrases. The model is designed from the corpus data using ABL algorithm and it learns using the Baum-Welch algorithm.

These papers [?] [?] researched some different methods for evaluating Natural Language Generation systems, by comparing automated methods and human evaluation methods. Where [?] propose methods like Bleu, NIST and Rogue which looks at the words of the generated sentence, and compares it to a reference. [?] Suggests a pos tagging method, that breaks up the sentences into POS tags, and compares it to a big reference corpus. They both discuss the pros and cons of the evaluation systems and suggests that a combination of automated methods should be used, since the different automated methods covers different parts of the text.

3 Hidden Markov Model

3.1 N-gram

I DIDNT MODIFY THIS n-gram models have been used for the task of NLG [?]. n-gram models use the statistical property of co-occurrences to calculate the most likely word to appear given previous (n-1) grams. Intuition would tell us that a higher n gram model will perform better than a lower one as the samples will be fewer and are more likely to be semantically correct but will be much more computationally intensive. We have used trigrams as one of the generation techniques, But there are also some problems with n-gram models. One of them is of the sparse data combinations as n becomes greater. The other problem which will be verified in the results section is that of non-randomness. Trigram tends to pick full sentences out of the corpus and that is where HMMs perform much better over trigrams. This problem with ngrams gets aggravated if the training corpus is small in size.

3.2 Hidden Markov Models

Hidden Markov Model (HMM) is a tool for modelling time series data. It can be presented as a dynamic Bayesian network. In HMM, “hidden” stands for the states which are not directly visible to the observer [?]. A certain HMM is fully defined by its:

- transition matrix A ,

- emission matrix B ,
- initial state probability distribution π .

We propose to use HMMs for the surface generalisation task. HMMs have been used with very good results in the fields of speech recognition and also in NLU (Natural Language Understanding) [?]. The idea of HMM in this work is to do something similar to the inverse of what POS Tagging does. In POS Tagging, an input string of words is mapped to a hidden sequence of semantic POS tags. The POS Tags then can serve as the hidden states in the HMM model and the words as the observations. ***WRITE SOMETHING NICE HERE*** Algorithm ??.

In our work, we have used three types of HMMs for the text generation: *supervised*, *supervised with text chunking* and *unsupervised*.

Algorithm 1 HMM Text Generation Algorithm

```

1:  $sequence \leftarrow \emptyset$ 
2:  $state \leftarrow \text{PICKRANDOMSTATE}()$ 
3: repeat
4:    $emission \leftarrow \text{GENERATEEMISSION}(state)$ 
5:    $sequence \leftarrow \text{APPEND}(sequence, emission)$ 
6:    $state \leftarrow \text{STATETRANSITION}(state)$ 
7: until enough
8:  $\text{OUTPUT}(sequence)$ 

```

3.2.1 Supervised HMMs

In the supervised HMMs, the POS tags are used as the states and the words and punctuation of the corpus are treated as the emissions. NLTK's ¹ default POS Tagger was used for evaluating the POS tags in a sentence. The state-transition matrix and emission matrix are then computed with them, using the supervised learning algorithm. The algorithm is purely statistical and is described by the following equations:

$$A_{ij} = \frac{S_{ij}}{\sum_k S_{ik}} \quad B_i(w) = \frac{E_i(w)}{\sum_{w'} E_i(w')} \quad \pi_i = \frac{S_i}{\sum_k S_k}$$

where S_{ij} is the number of occurrences of transitions from state i to state j , $E_i(w)$ is the number of occurrences of the observation w in the state i and S_i is the number of occurrences of state i .

¹Natural Language Toolkit - Open source library for NLG written in python

3.2.2 Supervised HMMs with Text Chunking

This type of HMM is an extension of a supervised HMM described in section ???. Text chunking is a method of grouping words in a sentence, in a way that they produce semantical units. For example, the sentence “*The quick brown fox jumps over the lazy dog.*” gets splitted into the following chunks: “*The quick brown fox*”, “*jumps*”, “*over*”, “*a lazy dog*”, “*.*”. It can be clearly seen now that “*The quick brown fox*” is a single semantical unit.

In the chunked HMMs, we try to enrich the emissions with more information about the semantics of a sentence and thus add chunk tags with POS tags as hidden states. We also add chunks as emissions, which have previously consisted of only words and punctuation. The learning algorithm is exactly the same as the one used for the regular supervised HMMs.

3.2.3 Unsupervised HMMs

Unsupervised HMMs differ from supervised ones in several ways. The states are implicit, the learner only knows the number of states but not what they represent. The algorithm used for learning is Baum-Welch, which is a special case of expectation maximization algorithm. Using this algorithm, we let the learner find the patterns in the data and infer the states itself. The emissions in unsupervised HMMs are the same as the ones in supervised HMMs.

3.3 Performance measures for experiments

As in other scientific fields, it is crucial to test how well our systems, modules and algorithms work. This type of performance measure is called evaluation. There are three techniques of evaluating NLG systems: task-based, human ratings and metrics [?]. In our project we are going to use human ratings and metric evaluations.

Evaluation gives as an output the “goodness” of a sentence or a text. The “goodness” is defined by three characteristics: fluency, adequateness and readability [?]. It is often a difficult task to assess each characteristic. Human ratings are based on opinions of individuals rating the given text. These types of evaluations are subjective and hard to measure. But the main advantage is the that the natural language is made for interactions between people and we can evaluate texts more accurately than machines. In our project we will spend a certain amount of time evaluating our models ourselves and with the help of others to reduce the bias.

Metric evaluations are based on software evaluations of generated text.

They are more objective than human ratings, but also less reliable. Most evaluators compare generated text with texts written by people to compute the score. In our project, we used three evaluators- BLEU, Linguistic checker and POS sequence checker [?]. ADD TEXT ABOUT BLEU HERE STEFAN!! For evaluating the grammatical correctness of the generated sentence, we used the open source tool LANGUAGE TOOL. This software is quite clean and robust. It gives all the possible errors and their description when it is fed with a sentence. The goodness of a sentence can be evaluated by using the formula:

The third test used for evaluation is the POS tag sequence checker. This measure gives us a good estimate of how correct the sentence is correct semantically.

3.4 Measure success of the project work or approach

The main task of the project is to successfully implement HMM text generator. Besides HMM text generator, we are also going to implement simple n-gram generators such as bigram and trigram. To measure success of the project approach, we are going to use performance measures listed above, as well as comparison between HMM and n-gram generator. If our hypothesis proves to be true, HMM will perform better than bigram and trigram.

3.5 Measure the progress during the course of the project

The project is divided into several stages. In the beginning, we will set the hypothesis and elaborate details regarding the same. By analyzing collected research papers, we will make decisions regarding implementation details, improvements and alterations, and also define performance expectations of our approach.

Second stage is the implementation. During this stage, we will implement bigram, trigram and HMM text generators (and genetic algorithm for HMM optimization if needed). In the same time we will also test our application to ensure algorithm correctness.

Finally, we will run experiments (evaluations) mentioned above and alter parameters accordingly to achieve better results.

During the course of the project, we will record our progress based on the project plan described above. Prediction of the time consumption will be made for each of the stages. Using those predictions, we will adjust the pace of the project progress for the purpose of completing the task in time.

3.6 Implementation

Implementation will be done shortly.

4 Experimental results

4.1 Experimental setup

When performing the tests, we first generated around 20 sentences with each of our four systems, then we ran them through the our evaluation methods which generated a score for each of the systems. We used human written sentences to have as a reference, which were either written manually or pulled from a news article. We also generated a score for the human tests. We then plotted the results to get a good overview of our results.

4.2 Experiment

As seen from the xxx the Bleu results are all very close to the human results. The human results are used as a benchmark value on what score a good sentence should get and since all our results are close to the human results we found that they are to unreliable and because of this, we disregarded the Bleu score from our results.

In the long corpus xxx, the trigrams performed the best since it received the highest LC and POS score, which co-relates with the grammatical and the semantics of the sentence. The HMM systems were all pretty similar, but the Supervised HMM performed a little bit better on the POS then the other systems.

The short corpus gave a lower score for trigrams and a higher score for all the HMM systems, where the supervised was the best. Overall the score was higher and closer to the human score with a shorter corpus. The results of the trigrams here is quite deceiving, since the sentences are not true random it is basically only a copy of the corpus. Which is also the nature of trigrams when dealing with a shorter corpus. So even tho the results look ok, the sentences aren't very good since it is not random.

The chunking HMM system were introduced to increase the performance of the supervised HMM score. But even tho the testing on the short corpus gives a better result then the long corpus the chunking HMM still performs worse then the supervised HMM. A problem that we could see when generating sentences with all our HMM systems was that it generated way to many punctuations and citations.

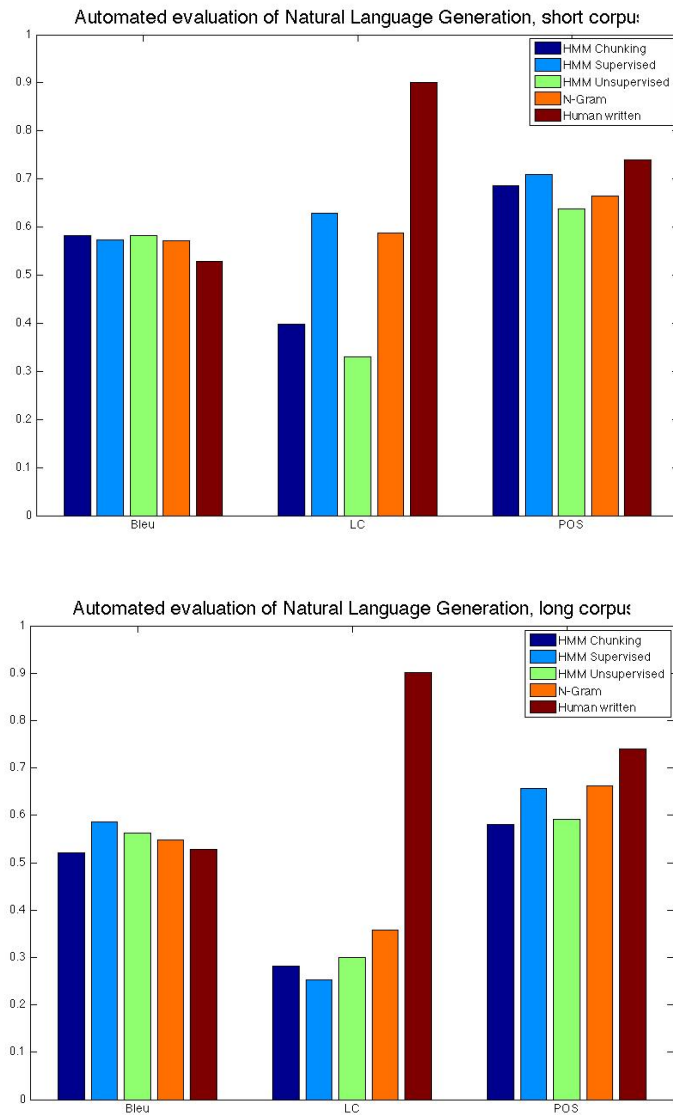


Figure 1: Experimental results when generating sentences with the four tested systems for a long corpus and a short corpus respectively, using the three automated evaluation methods, and comparing it to the human results.

5 Summary and Conclusions

The results from the Bleu