|  |
| --- |
| Edinburgh Napier University |
| Literature Review |
| Natural Language Processing |

|  |
| --- |
| Jordan Aitken |

Contents

[What is Natural Language Processing? 2](#_Toc471607184)

[Natural Language Generation 2](#_Toc471607185)

[Language Models 2](#_Toc471607186)

[What are Language Models? 2](#_Toc471607187)

[Types of Language Models 2](#_Toc471607188)

[Finite State Automata 2](#_Toc471607189)

[N-Grams 2](#_Toc471607190)

[N-Grams 3](#_Toc471607191)

[What are N-Grams? 3](#_Toc471607192)

[How are N-Grams used? 3](#_Toc471607193)

[Advantages and Disadvantages 3](#_Toc471607194)

[Advantages 3](#_Toc471607195)

[Disadvantages 3](#_Toc471607196)

[SN-Grams 3](#_Toc471607197)

[Published Papers 3](#_Toc471607198)

[Summarisation 4](#_Toc471607199)

[References 5](#_Toc471607200)

# What is Natural Language Processing?

Natural Language Processing (NLP) is investigation into how computers “understand” the written or spoken language used by humans. In this project, we will be using the English language.

# Natural Language Generation

Natural Language Generation (NLG) is a subfield of artificial intelligence and computational linguistics that is used in producing understandable text in English or other human languages. (Dale & Reiter)

# Language Models

Language models play a key part in NLP; computers do not have a grammatical understanding of the English (or other) language like humans do. Therefore, a language model is used to evaluate the probability of a sentence, and assign a value as to how probable it is that that sentence would occur. Eg. Does it make sense?

## What are Language Models?

A language model is a way to evaluate the probability of a sentence, or sequence of words. It assigns a value based on the relative likelihood of the sentence or phrase occurring. This is very useful for many language constructs such as speech recognition, machine translation, part-of-speech tagging, parsing, handwriting recognition and information retrieval. (Language Models, 2016)

## Types of Language Models

### Finite State Automata

Finite State Automata, or Finite State Machines (FSM) can potentially be a language model that could be used (Manning, 2009)

### N-Grams

An n-gram is a language model that counts the occurrence of words in a sentence, or sequence of words. It is represented by a value of n. For example, unigrams, bigrams and trigrams.

# N-Grams

## What are N-Grams?

N-gram based techniques are predominant in modern natural language processing (NLP) and its applications. Traditional n-grams are sequences of elements as they appear in texts. These elements can be words, characters, POS tags, or any other elements as they encounter one after another in texts. Common convention is that “n” corresponds to the number of elements in a sequence. (Sidorov, Velasquez, Stamatatos, Gelbukh, & Chanona-Hernández)

## How are N-Grams used?

An n-gram counts the occurrence, and calculates the probability of a set of n elements, where n is a positive integer.

### Unigrams, Bigrams, Trigrams….

Some n-grams have names that they are commonly referred to as:

Unigram: n =1,

Bigram: n = 2

Trigram: n = 3,

Four-gram: n = 4

…etc

<s> represents the start of a sentence

</s> represents the end of a sentence.

“The man walked to the supermarket”

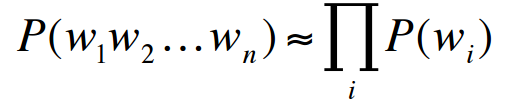
|  |  |
| --- | --- |
| 1-gram (Unigram) | <s>,  The,  man,  walked,  to,  the,  supermarket,  </s> |
| 2-gram (Bigram) | <s> The,  The man,  man walked,  walked to,  to the,  the supermarket,  supermarket </s> |
| 3-gram (Trigram) | <s> The man,  The man walked,  man walked to,  walked to the,  to the supermarket,  the supermarket </s> |

(Chambers, Tetreault, & Allen)



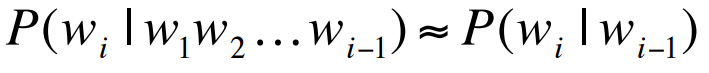
#### Unigrams

A unigram is an n-gram model where n is equal to one. This is useful for counting the occurrence of a single element in a corpus. For example, a word, punctuation mark or a part of a word.



#### Bigrams

A bigram is an n-gram model where n is equal to two. This can be used to count how often two words appear next to each other. For example, the number of times the words “man walked” appear next to each other can be counted.



#### Trigrams

## Advantages and Disadvantages

### Advantages

### Disadvantages

N-grams do not count or acknowledge words that occur outside of the n-gram range. For example,

“The cup that was on top of the book fell”

If a trigram was used, the language model would not recognise that it was the cup which fell. This is a huge issued because language has long-range dependencies.

(Jurafsky)

# SN-Grams

Semantic N-Grams (sn-grams) is used for evaluating the similarity between

## Published Papers

Introduction to Information Retrieval (2009)

Christopher D. Manning

Semantic Text Similarity using n-grams, WordNet, Syntactic Analysis, ESA and Information Retrieval based Features

Davide Buscaldi, Joeseph Le Roux, Jorge J. García Flores

Manning, C. (2009). *Introduction to Information Retrieval*. Retrieved from http://nlp.stanford.edu/IR-book/pdf/12lmodel.pdf

## Summarisation

# References

Chambers, N., Tetreault, J., & Allen, J. (n.d.). *Approaches for Automatically Tagging Affect.* Retrieved from University of Rochester: https://www.cs.rochester.edu/

Dale, R., & Reiter, E. (n.d.). *Building Natural Language Generation Systems.* Retrieved from http://assets.cambridge.org/97805216/20369/sample/9780521620369wsn01.pdf

Jurafsky, D. (n.d.). *Language Modelling.* Retrieved from http://spark-public.s3.amazonaws.com/nlp/slides/languagemodeling.pdf

*Language Models*. (2016, 10 27). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Language\_model

Manning, C. (2009). *Introduction to Information Retrieval*. Retrieved from http://nlp.stanford.edu/IR-book/pdf/12lmodel.pdf

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (n.d.). *Syntactic N-grams as Machine Learning Features.* Retrieved from http://www.g-sidorov.org: http://www.g-sidorov.org/Synt\_n\_grams\_ESWA\_FINAL.pdf