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

|  |
| --- |
| Jordan Aitken |

Contents

[What is Natural Language Processing? 2](#_Toc471997675)

[Natural Language Generation 2](#_Toc471997676)

[Language Models 2](#_Toc471997677)

[What are Language Models? 2](#_Toc471997678)

[Types of Language Models 2](#_Toc471997679)

[N-Grams 2](#_Toc471997680)

[Cache Language Model 2](#_Toc471997681)

[Continuous Space Language Model 2](#_Toc471997682)

[Factored Language Model 3](#_Toc471997683)

[N-Grams 3](#_Toc471997684)

[What are N-Grams? 3](#_Toc471997685)

[How are N-Grams used? 3](#_Toc471997686)

[Unigrams, Bigrams, Trigrams…. 3](#_Toc471997687)

[Advantages and Disadvantages 5](#_Toc471997688)

[Advantages 5](#_Toc471997689)

[Disadvantages 5](#_Toc471997690)

[SN-Grams 5](#_Toc471997691)

[Why use SN-Grams? 6](#_Toc471997692)

[Published Papers 6](#_Toc471997693)

[Summarisation 6](#_Toc471997694)

[References 7](#_Toc471997695)

# 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. E.g. 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

### 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.

### Cache Language Model

(Kuhn & De Mori, 1990)

### Continuous Space Language Model

### Factored Language Model

# 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.

“<s> The man walked to the supermarket </s>”

|  |  |
| --- | --- |
| 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)



**Example Corpus**

“I would not like them

here or there.

I would not like them

anywhere.

I do not like

green eggs and ham.

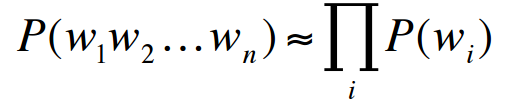
I do not like them,

Sam-I-am.”

(Seuss, 1960)

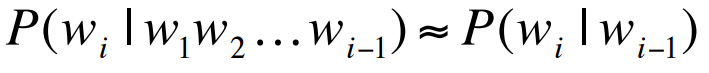
#### Unigrams

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

#### Long-Range Dependencies

Language often has long-range dependencies, which means that an n-gram only acknowledges words that are within its range. For example:

“The cup that was on top of the cupboard in the bedroom fell.”

The language model is unable to link the words “cup” and “fell” in any n-gram where n < 12, which is an unusually high value. (Jurafsky)

#### Word Order

Using n-grams, word order is important, and the model relies heavily on the order of the words. This can lead to inaccuracies and missed information. For example, “the man and the boy” and “the boy and the man” may return different probability values.

#### Corpus Reliance

N-grams only work when used with a corpus. The language model must have seen an occurrence of a certain sequence of words in the corpus, or it will be assigned a probability of zero, no matter how much sense the sentence might make.

# Semantic N-Grams

## What is the Semantic N-Gram Model?

The semantic n-gram (sngram) model will be a language model that looks at how often words occur together in a sentence, and will

## Why use SN-Grams?

Although n-grams are useful and we can often get away with using them, they are flawed, especially with long-range dependencies. For example, using a bigram model, given the text “The big orange cat fell” the model would fail to recognise that “big” and “cat” are related.

## 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

Kuhn, R., & De Mori, R. (1990, June 6). *A Cache-Based Natural Language Model for Speech Recognition.* Retrieved from http://visgraph.cs.ust.hk/biometrics/Papers/Voice/pami1990-06-01.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

Seuss, D. (1960). In D. Seuss, *Green Eggs and Ham* (p. 8).

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