THE UNIVERSITY OF MELBOURNE

Not just another game: Hangman

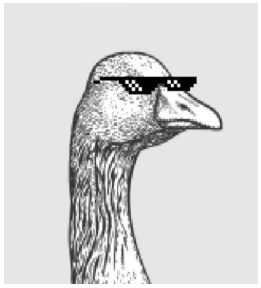
Importance of Language Models

By:
Daniel Gill
Kartik Kishore
Kaushik Ramesh

About Us

Daniel Gil

- Pursuing Masters of data science since 2018
- Extensive business process streamlining & BI experience across banking, retail, manufacturing & healthcare domains.
- Expert in Data Analytics, Business Intelligence & Natural Language Processing.



Kartik Kishore

- Ms. Data Science, University of Melbourne, Batch of 2019
- Have worked as an Identity Admin at *Cognizant* in Cyber security projects, which has been the driving force behind my interest in data analytics
- Hands-on experience in Python, R & shell. Automation & computer networks catch my fancy
- Likes pineapple on Pizza

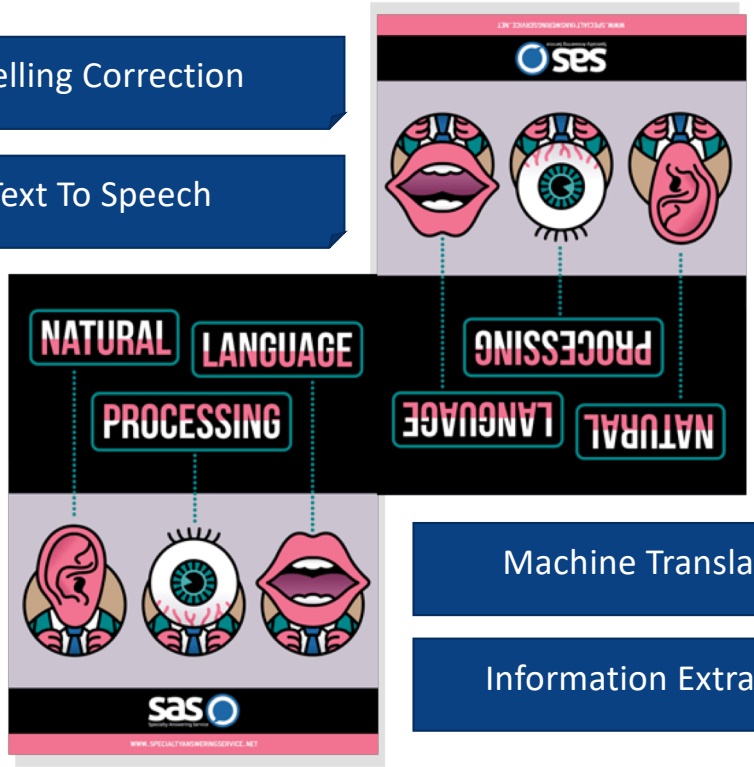
Kaushik Ramesh

- Pursuing Masters of data science since 2018
- Developed automation tools to eliminate manual interference during my tenure at Infosys limited in Bangalore, India
- Hold expertise in python, Java, R and Unix shell scripting

Introduction

Spelling Correction

Text To Speech



Machine Translation

Information Extraction

Image Source: https://www.specialtyansweringservice.net

Introduction

- These are the techniques and methods involved in manipulation of any language used by us humans.
- Natural Language Toolkit (NLTK) is a package of algorithms for symbolic and statistical natural language processing for English written in the Python programming language.
- At the core of these algorithms is hardcore statistics.

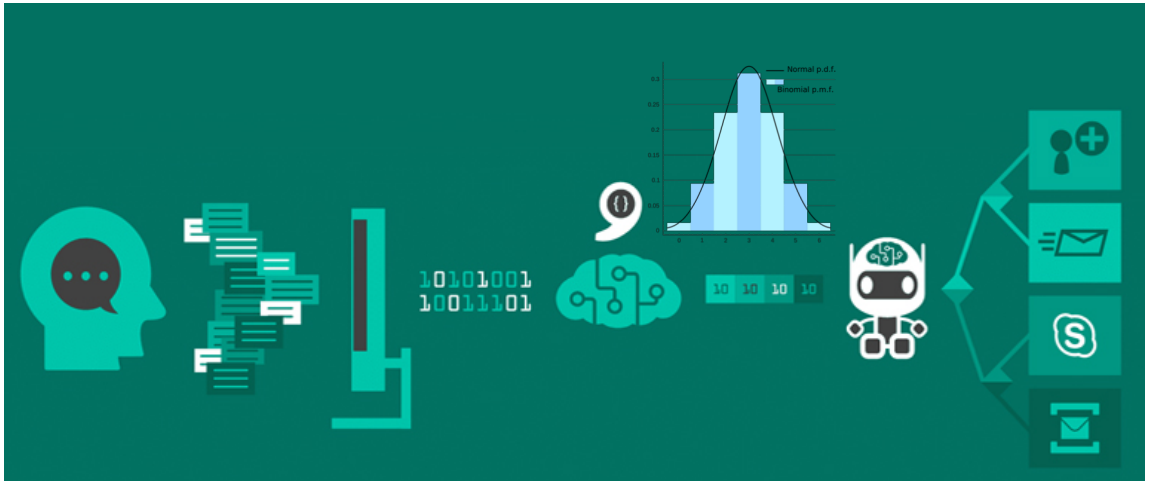
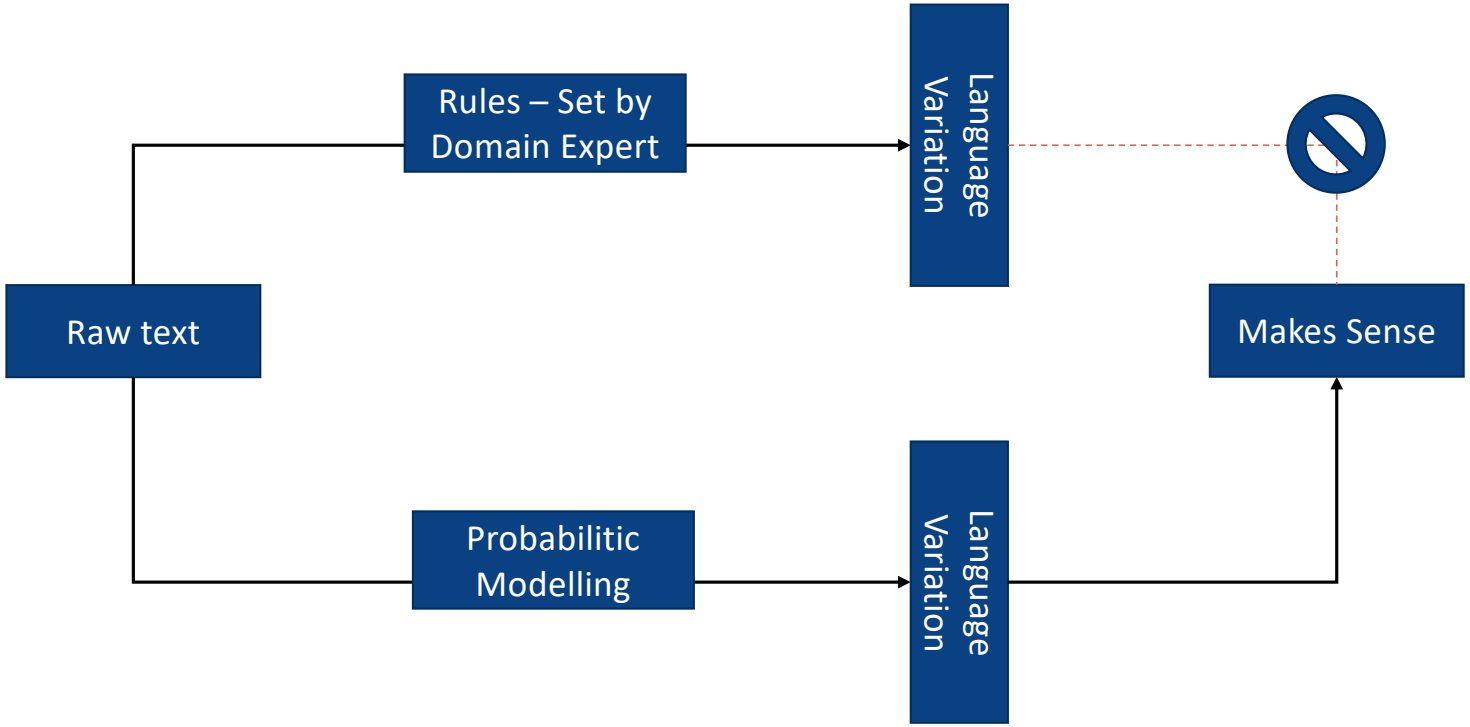


Image Source: 1. TechFunnel 2. By Cflm001 (talk) - Own work. Derived from File:BinDistApprox large.png by Xiao Fei, released under GFDL/CC-BY-SA-3.0., CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=7689435

Natural Language Processing



Basic Language Models

What is a language model?

A model created by following a probability distribution over strings of text

- How likely is a given string (observation) in a given corpus.
- It creates involve making an **nth order Markov assumption** and **estimating n-gram probabilities** via counting and subsequent smoothing.

MARKOV ASSUMPTION: Calculations on state '**k**' are influenced by output of calculation from state '**k-1**'

ESTIMATING N-GRAM PROBABILITIES: probability **P**(*w1,w2,…,wn*) is a product of word probabilities based on a history of preceding words, whereby the history is limited to m words

$$p(w_n|w_1, w_2, \cdots, w_{n-1}) \approx p(w_n|w_{n-m}, \cdots, w_{n-2}, w_{n-1})$$

Contemplations

Garbage in, Garbage out

Image Source: Saleshacker.com

Above & Beyond

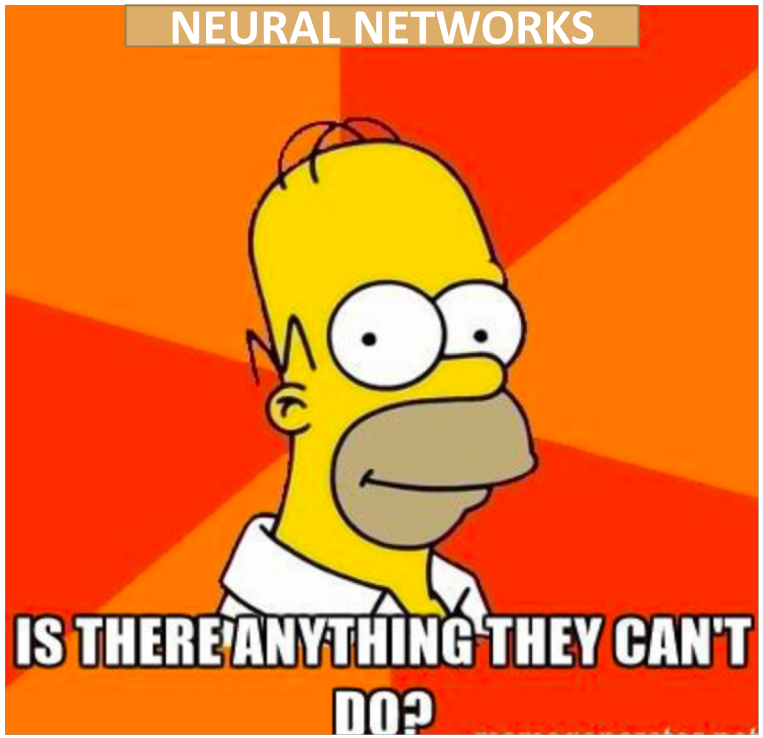


Image-source: https://blog.goodaudience.com/introduction-to-convolution-neural-networks-ef4d03a5ca69

State of the art

Continuous-space LM, also known as neural language model (NLM) have replaced traditional n-gram based models and have recently proved to give state of the art performance.

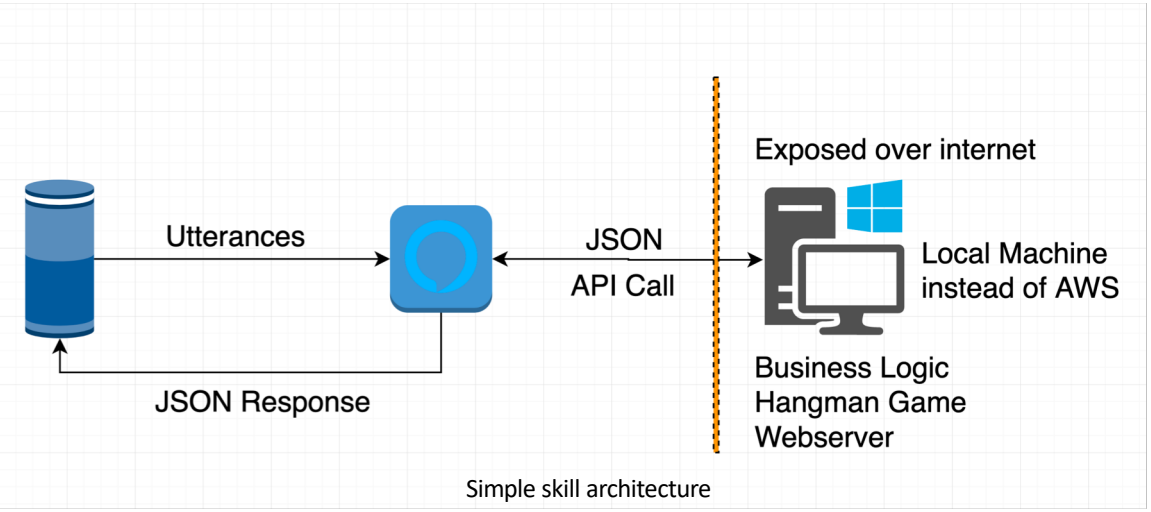
Varieties:

- **Feed-forward neural network based LM**
  - Which was proposed to tackle the problems of data sparsity
- **Recurrent neural network based LM**
  - Which was proposed to address the problem of limited context.

DEMO

amazon alexa

Image Source: developer.amazon.com/console/alexa/ask



Simple skill architecture

Applications in Research

- Politics researcher
- Journalism
- Human behavior analysis

Q&A

Any questions?

THE UNIVERSITY OF MELBOURNE

Thank You