Language and Information
Natural Language Processing, NLP

INSC 702: Advanced Topics in Information Science

Shimelis Assefa

Agenda

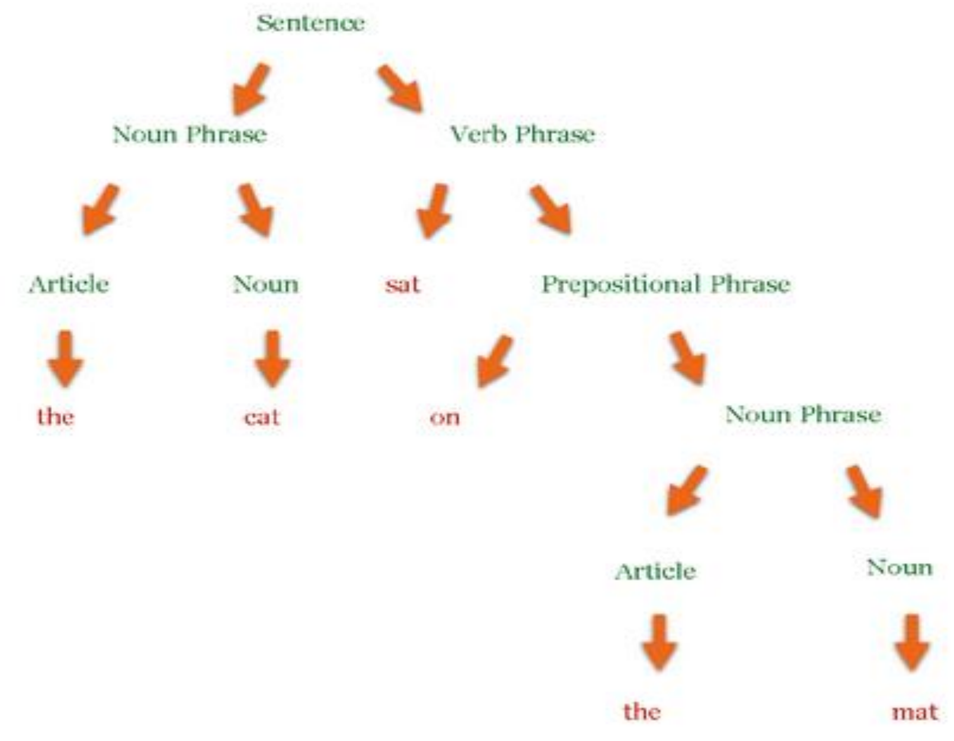
- Introduction to Natural Language Processing, NLP.
- Key Applications.
- NLP Tools
- Case Studies

Introduction

- How can we help a computer to understand natural language – that is the goal of NLP
- Some of the most common applications of NLP include:
  - machine translators,
  - speech recognition, and
  - auto spell, etc.

Introduction…

- Symbolic and statistical NLP



Introduction

- Symbolic and statistical NLP
- Started in the 1950s
- Alan Turing's 1950 work "Computing Machinery and Intelligence"
- Utilizes manually written rules based on linguistic features.
- Such techniques were inadequate for capturing the complexities of natural language as well as processing an ever-increasing amount of data

Introduction

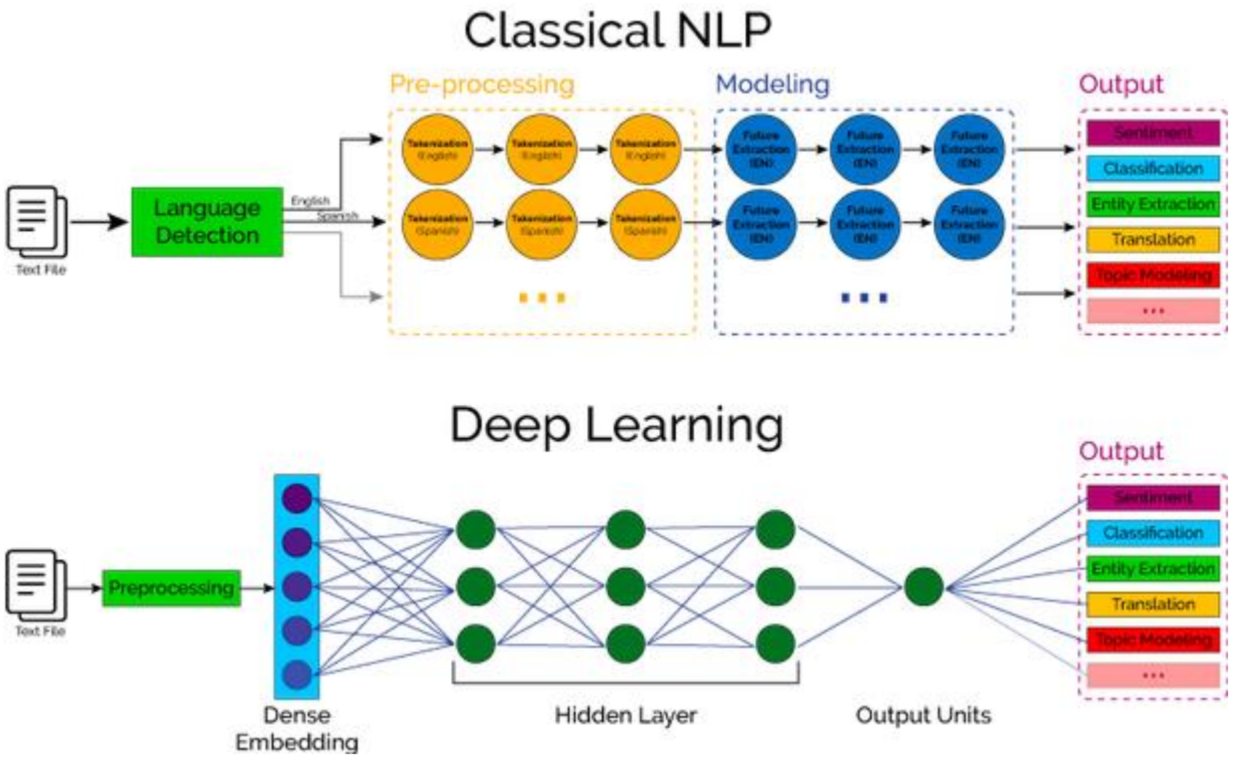
- Beginning in the 1990s, the inadequacies of symbolic NLP catalyzed a shift towards statistical models of natural language
- Statistical models combine “computer algorithms with machine learning and deep learning models to automatically extract, classify, and label elements of text and voice data'' (IBM Cloud Education, 2020).
- These models worked by outputting the statistical likelihood of every possible meaning for every aforementioned extracted element.
- The current phase of NLP is Neural NLP which began in the 2010s with the rise of the use of neural networks in NLP processes.

Introduction

- Neural NLP
- The original purpose of neural networks was to use computers to simulate the structures of the human brain so that computers could “think” like humans.
- This idea was partially correct since the performance of multi-layer models performs so well that they were implemented whenever possible.
- However, it turned out that even with the structure of the human brain, the “think flow” had to be implemented differently than humans.
- Neural networks require two conditions:
    - a large amount of data to get high accuracy.
    - adequate computing resources

Introduction …

- Neural NLP



Introduction …

- Beginning in 2010s, or big data era, the popularity of the internet brought enormous amounts of data online for scientists to train neural networks.
- Also, technological advancement saw the creation of much stronger computers,
- These modern developments have allowed deep learning to become the hottest area in computer science.
- Scientists are finally able to apply neural network models to many different areas including NLP.
- With the help of neural networks, applications like machine translation, text generation and many others perform much better than ever before.
- These improvements are possible because neural networks can extract knowledge from data layer by layer.

NLP Applications

- Since it’s impossible to really know if computers can fully “understand” natural languages like humans do, the only quantified way to evaluate it is the performance in natural-language-processing related applications and tasks.
- NLP has a lot of practical applications that can help human beings in daily life such as machine translation, Automatic spelling, and automatic summarization.

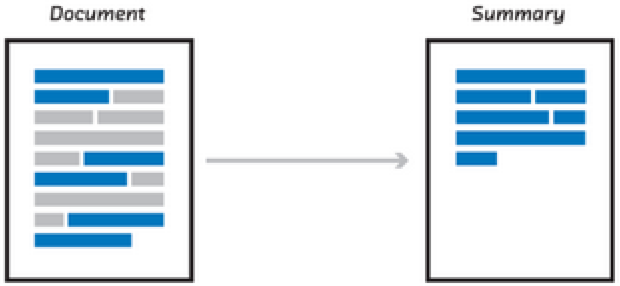
NLP Applications…

- Machine translation
- Machine translation has been a problematic area for decades.
- For example, widely used machine translation programs such as google translate are known for their inaccuracy.
- Only some of the words are translated correctly and often these programs struggle to correctly translate whole sentences.
- Overall, the accuracy rate for these programs is low.
- However, after the introduction of deep learning, machine translators have become more and more precise and accurate.
- In a study in 2020, researchers compared the performance between a machine translation program using deep learning techniques and a professional human translation agency (CUBBITT).
- The model the researchers built performed better in 52% of sentences and worse in only 26% of sentences (Popel, Martin 2020).

NLP Applications…

- Automatic spelling
- Automatic spelling is so commonly used that every website or app we input text into most likely has this feature.
- However, it does do more than checking for typos.
- It can predict what you are trying to write and suggest other choices of alternatives to you.
- Also, it has the function to correct grammatical errors automatically.
- Automatic spelling has already been integrated into apps and websites like outlook, gmail and smartphone interfaces.

NLP Applications…



- Automatic summarization
- Automatic summarization is an exciting field of NLP even though the industrial applications in this field are relatively new.
- The goal for scientists and engineers in the development of automatic summarization is to condense large quantities of information into more manageable quantities.
- Availability of large quantities of data (including textual data) presents its own challenges for humans.
- For this reason, a precise automatic summarization is useful for humans when searching for information.

Data Sources

- Quite numerous places to get textual data
  - Corpus Data -https://www.corpusdata.org/
  - Web based resources – e.g. - Wikipedia
  - Free Books – Project Gutenberg - A library of free, downloadable eBooks, includes books in English, Portuguese, German, and French
  - MIMIC IV - Medical Information Mart for Intensive Care III (**MIMIC-Iv**) dataset is a large, de-identified and publicly-available collection of medical records - https://physionet.org/content/mimiciv/2.0/
  - Physionet - https://physionet.org/

Text Analysis & NLP Tools

- NLTK - open-source toolkit for NLP using Python - https://www.nltk.org/
- General programming languages and selected libraries such as R and Python
- Open-source, web-based text analysis environment – e.g. - https://voyant-tools.org/
- Creating an alphabetical listing of all occurrences of each principal word in a text corpus along with their immediate context – e.g. - – concordance and text analysis tool - https://www.laurenceanthony.net/software/antconc/

NLP Tools …

- Open-source, Java-based tool for identifying and tagging the names of people, places, and things within a text corpus – e.g., Stanford Named Entity Recognizer - https://nlp.stanford.edu/software/CRF-NER.html
- TaggerOne in biomedical science –to recognize disease, drugs, chemicals, etc in text https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/taggerone/
- Topic modeling – e.g. - Mallet: MAchine Learning for LanguagE Toolkit - https://mimno.github.io/Mallet/topics

NLP Tools …

- For chemical–gene/protein interactions, chemical–disease and gene–disease relationships – check this too - https://ctdbase.org/
- IR, Library specific tools
- https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/

Case Studies

- IBM Watson Natural language Understanding, Classifier
  - https://www.ibm.com/cloud/watson-natural-language-understanding
- NCBI BIO NLP research
  - https://www.ncbi.nlm.nih.gov/research/bionlp/
    - https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000716
    - https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/PubReCheck/#/
- Chatbots – e.g. – Walk with Yeshi – a FB messenger bot built by Loaki in partnership with Charity:Water
  - Library Chatbot – e.g. - **T-Rex**, the new University of Calgary Library **chatbot** to get answers to quick questions **https://libguides.ucalgary.ca/c.php?g=718553**
  - Machine translation in Amharic -.e.g. https://lesan.ai/

Case Studies

- Google Biomed Explorer - an NLP tool that searches PubMed, PubMed Central, and CORD-19 (Covid-19 literature) using a question focused search feature.
- This may help those working on a quick complex health research question who are having difficulty representing the question with a search strings or keywords.
    - https://sites.research.google/biomedexplorer/

Case Studies

- Text summarization paper from ARXIV
  - https://arxiv.org/search/?query=text+summarization&searchtype=all&source=header
  - Afaan oromo - https://arxiv.org/abs/2103.02900
- KDnuggets NLP resources
  - https://www.kdnuggets.com/tag/natural-language-processing