University of Northampton

Week 1

Introduction to NLP

CSY3055 – Natural Language Processing

Dr Oluseyi Oyedeji

Learning Goals for Week 1

By the end of week one, you should be able to:

- **Understand** the *scope of NLP*
- **Recognise** *challenges* in processing human language
- **Describe** the *NLP pipeline* and levels of analysis
- **Summarise** the *history and evolution* of NLP
- *Get **hands-on experience** with some basic NLP tools*
- *Reflect on ethics in NLP (introductory)*

Today's Session

- **What is NLP?**
- **Relevance** of NLP
- NLP in AI **hierarchy**
- **History** of NLP
- Challenges with **Human Languages**
- **Level** of NLP Analysis
- NLP **Pipeline**
- **Applications** of NLP
- Recent Advancements
- NLP **Tools**
- Ethics in NLP

What is NLP?

- **NLP** is a *Subfield of AI & computational linguistics*
- Enables machines to **process**, **understand**, **generate** human **language**

Input = *unstructured text/speech* → Output = **structured meaning/action**

Relevance of NLP

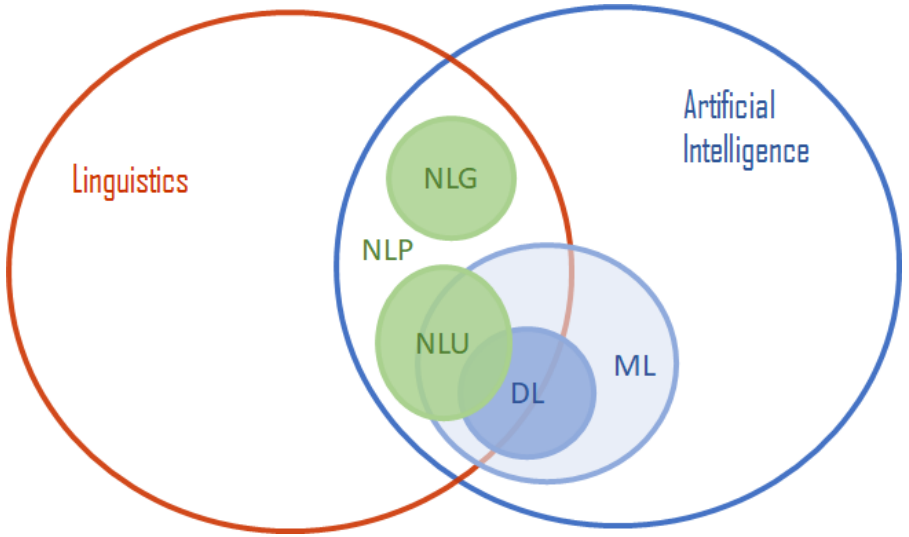
- **Assistants & Translation**
  - ✓ Digital assistants: Siri, Alexa, Google Assistant
  - ✓ Machine translation: Google Translate, DeepL
- **Search & Chatbots**
  - ✓ Information retrieval: Google Search
  - ✓ Chatbots & customer support
- **Sentiment & Summarisation**
  - ✓ Sentiment analysis: reviews, tweets
  - ✓ Automatic summarisation, news aggregation

Relevance of NLP contd.

- **Q&A and Misinformation**
  - ✓ Question answering: ChatGPT, WolframAlpha
  - ✓ Misinformation detection and prevention

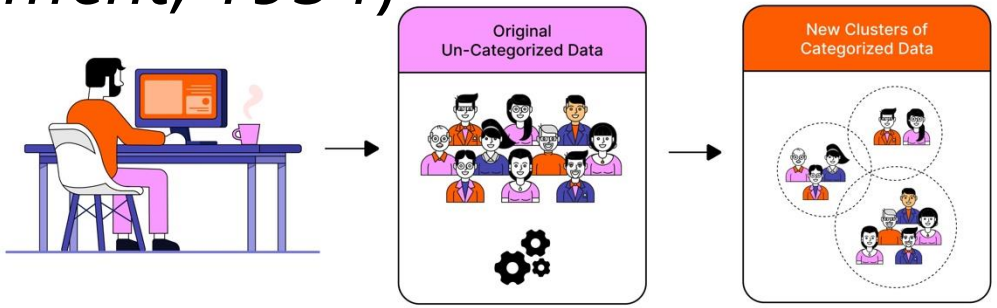
NLP in the AI Hierarchy

- **AI** = broad goal of intelligent machines
- **ML** = learning from data
- **DL** = neural networks
- **NLP** = intersection of AI, ML, and linguistics



History of NLP

- **1950s–60s**
  - ✓ Rule-based MT (*Georgetown experiment, 1954*)
  - ✓ ELIZA chatbot (1966)

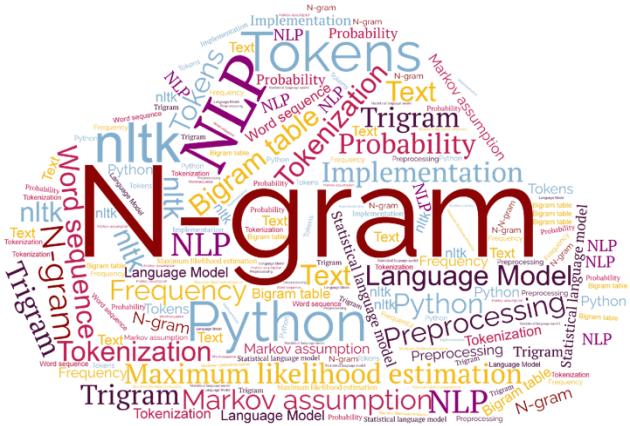


- **1970s**
  - ✓ AI-inspired Q&A (*BASEBALL system*)
  - ✓ Shift towards linguistics + AI approaches

- **1980s**
  - ✓ *Grammar-based systems*
  - ✓ Formal parsing techniques, discourse analysis

History of NLP

- **1990s**
  - ✓ Statistical revolution
  - ✓ n-grams, HMMs, probabilistic models



- **2000s**
  - ✓ Rise of ML algorithms: SVMs, CRFs, TF-IDF
  - ✓ Growth of large annotated corpora

- **2010s**
  - ✓ Word embeddings: Word2Vec, GloVe
  - ✓ Deep learning: RNNs, LSTMs, GRUs

History of NLP

- **2017 +**
  - ✓ **Transformers**: 'Attention is All You Need'
  - ✓ **BERT, GPT, T5**
  - ✓ Current era: Large Language Models (GPT-4, GPT-5, Claude, LLaMA)

Challenges with Human Language

- **Lexical Ambiguity**
  - One-word, multiple meanings
  - Example: 'bank' → riverbank or financial institution

- **Syntactic Ambiguity**
  - Sentence: 'I saw the man with the telescope'
  - Ambiguity: **who has the telescope?**

Challenges with Human Language

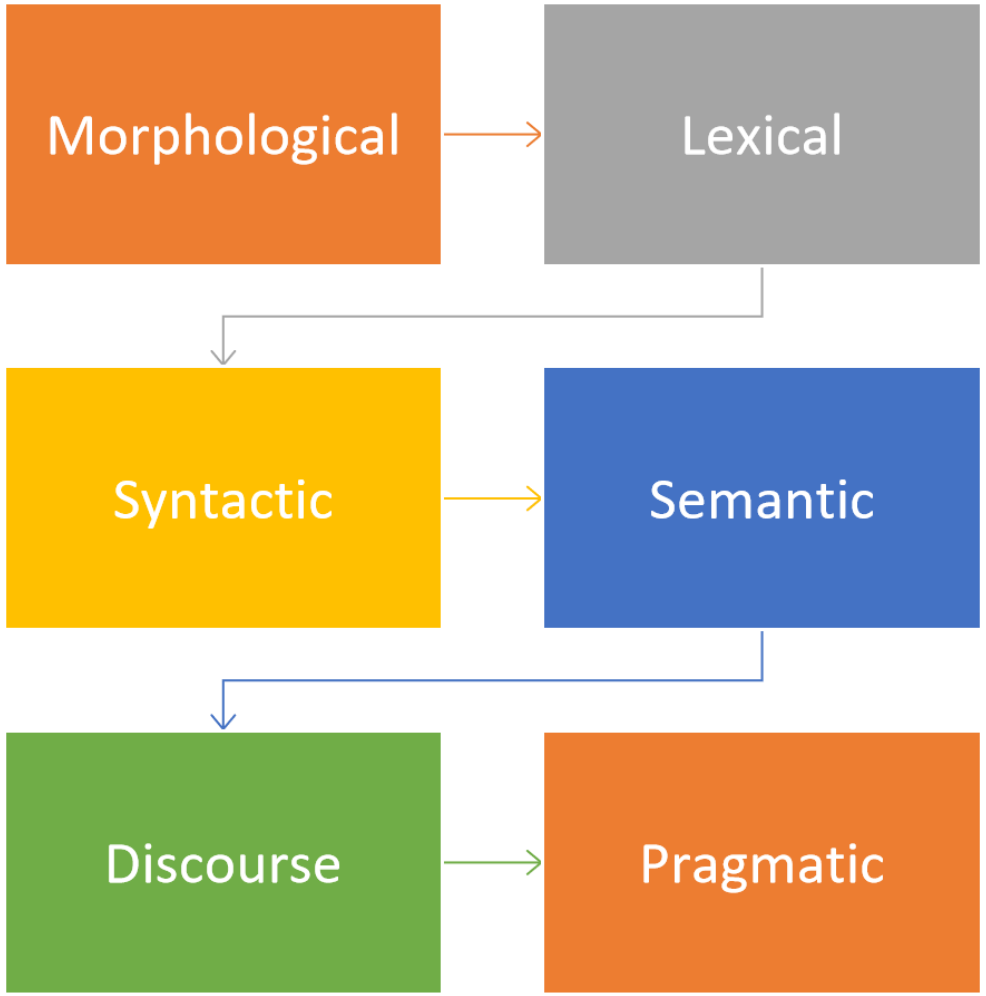
- **Semantic Ambiguity**
  - Contradictory meanings: 'Hot ice-cream'
  - Difficult for machines to resolve

- **Pragmatic Ambiguity**
  - Language depends on context
  - 'Can you pass the salt?' = request, not question

Other Language Challenges

- **Context** dependence
- Need for world knowledge
- Low-resource languages = *lack of data/tools*

Levels of NLP Analysis



Levels of NLP Analysis – Morphology

- **Morphology** is the study of **word structure**
- *Tokens, morphemes, stemming, lemmatization*
- **Tokenisation**
  - Splitting text into words/tokens
  - Example: 'NLP is amazing!' → ['`NLP`','`is`','`amazing`',’`!`’]
- **Stemming and Lemmatization**
  - *Stemming*: `relational` → **`relate`** (rule-based cut)
  - *Lemmatization*: `running` → **`run`** (dictionary-based)

Levels of NLP Analysis - Syntax

- **Syntax** is the *grammatical structure* of sentences
- **Parsing methods**
  - Constituency –> sub-phrases
  - Dependency - direct relationship
- Sentence: '`The cat sat on the mat`'
- Constituency parse: [S [NP `The cat`] [VP `sat` [PP `on` [NP `the mat`]]]]
- Dependency parse: subject=`cat`, verb=`sat`, object=`mat`

Levels of NLP Analysis – Semantic, Discourse, Pragmatics

- **Semantics** is the *study of meaning*
  - Word sense disambiguation, synonymy, polysemy
- **Discourse** refers to *meaning across sentences*
  - Coreference resolution, coherence
- **Pragmatics** refers to *language in context*
  - Speaker intention, world knowledge

NLP Pipeline

- **Input text** → *Preprocessing* → *Feature extraction* → *Modelling* → *Evaluation* → **Deployment**
- **Roadmap**: raw language → structured meaning → useful applications
- Helps connect theory to real-world systems

NLP Pipeline - Preprocessing

- **Tokenisation**
  - *Split text into words/sentences*
- **Stop word removal**
  - *Remove frequent low-value words*
- **Stemming & Lemmatization**
  - *Reduce words to root forms*
- **Normalisation**
  - *lowercase, punctuation handling, cleaning*

NLP Pipeline - Feature Extraction

- **Bag-of-Words (BoW)**
  - *Count words*
- TF-IDF
  - Adjust counts for importance
- Word embeddings: *dense vectors (Word2Vec, GloVe)*
- Contextual embeddings: *BERT, GPT*

NLP Pipeline - Deployment

- Integrating NLP into **apps and services**
- Chatbots, search engines, recommendation systems
- Challenges
  - *Scalability, fairness, bias, ethics*
- Tools
  - **Flask**, **FastAPI**, **cloud APIs**

Applications of NLP

- Machine translation
- Speech recognition & synthesis
- **Sentiment analysis**
- Information retrieval & search
- **Summarisation**
- Question answering
- Chatbots & virtual assistants
- Text generation (LLMs)

Recent Advancements

- **Word embeddings (Word2Vec, GloVe)**
  - Distributed representations
  - Semantic similarity capture
- Neural LMs: RNN, LSTM, GRU
  - Sequence modelling
- **Transformers**: Attention, BERT, GPT - Self-attention mechanism
- LLMs: GPT-4, Claude, LLaMA → Large scale, zero/few-shot learning

NLP Toolkits

- **Classical tools**
  - NLTK (Natural language toolkit)
  - Stanford CoreNLP (Java-based parser)
- **Modern Tools**
  - spaCy (efficient NLP pipelines)
  - Gensim (topic modelling, embeddings)
  - Hugging Face Transformers (pre-trained DL models)

NLP Toolkits contd.

- **Programming Language**
  - **Python**
  - R and Java exist, but Python dominates

- **Deep Learning Frameworks (for building models)**
  - **PyTorch** → dominant in research & LLMs.
  - **TensorFlow / Keras** → popular in industry & ML pipelines.
  - **JAX** → rising for high-performance training (used by Google).

Ethics in NLP

- **Bias in embeddings** (gender/race stereotypes)
  - "`doctor → male`" and "`nurse → female`" associations.
  - fairness and discrimination.
- Data privacy issues (memorising sensitive data)
- Misinformation (deepfakes, fake news)
- *Environmental cost* (training large LLMs)

Discussion

- If a spam filter wrongly blocks scholarship emails for international students, what are the consequences? How could this be avoided?

Summary

- **NLP** = AI + ML + linguistics
- Human language is complex: ambiguity, context
- **History**: rules → statistics → ML → DL → transformers
- **Applications**: translation, chatbots, search
- **Tools**: NLTK, spaCy, HuggingFace
- Ethics: bias, privacy, environment

Thank you