MAI 5201: Natural Language Processing

Introduction to Natural Language Processing

Instructor: Dr. Christopher Clarke

Overview & Agenda

1. **What is Natural Language Processing?**
    - Definitions and scope
    - NLP vs. Computational Linguistics
2. **Historical Development of NLP**
    - Early symbolic approaches
    - Statistical revolution
    - Neural networks and transformers
3. **Core NLP Tasks & Applications**
    - Text processing fundamentals
    - Understanding and generation tasks

Overview & Agenda (Contd.)

4. **Challenges in NLP**
    - Ambiguity and context
    - Multilingual & Ethical considerations
5. **Course Overview & Structure**
    - Learning objectives, assessment and projects
    - Weekly roadmap
6. **Modern NLP Landscape**
    - Large Language Models
    - Real-world applications

What is Natural Language Processing?

Natural Language Processing (NLP) is:

> *"A subfield of artificial intelligence that focuses on the interaction between computers and humans using natural language."*

Key Components:

- **Understanding** text and speech
- **Generating** human-like language
- **Processing** at scale and in real-time
- **Reasoning** about linguistic content

What is Natural Language Processing? (Contd.)

NLP sits at the intersection of:

- **Computer Science** (algorithms, data structures)
- **Linguistics** (syntax, semantics, pragmatics)
- **Machine Learning** (statistical models, neural networks)
- **Cognitive Science** (how humans process language)

> *"The goal is to bridge the gap between human communication and computer understanding."*

Examples: Machine translation, sentiment analysis, chatbots, text summarization, speech recognition

Brief History of NLP

1950s - 1960s: The Symbolic Era

- First machine translation experiments (Georgetown-IBM)
- Rule-based approaches and expert systems
- Focus on syntax and formal grammars

1970s - 1980s: Linguistic Foundations

- Chomsky's influence on computational linguistics
- Development of parsing algorithms
- Early speech recognition systems

Brief History of NLP (Contd.)

1990s - 2000s: Statistical Revolution

- Shift from rules to statistical models
- Hidden Markov Models for speech
- N-gram language models

2000s - 2010s: Machine Learning Era

- Support Vector Machines for text classification
- Word embeddings (Word2Vec, GloVe)
- Recurrent Neural Networks for sequences

Brief History of NLP (Contd.)

2010s - Present: Deep Learning & Transformers

- Neural machine translation breakthrough
- Attention mechanisms and Transformers (2017)
- Large Language Models (BERT, GPT, ChatGPT)
- Multimodal AI (text + images)

We've gone from translating "The spirit is willing but the flesh is weak" to "The vodka is good but the meat is rotten"... to near-human quality!

Core NLP Tasks

Text Processing:

- Tokenization, stemming, lemmatization
- Part-of-speech tagging
- Named entity recognition

Understanding Tasks:

- Text classification (sentiment, topic)
- Question answering
- Reading comprehension
- Information extraction

Core NLP Tasks (Contd.)

Generation Tasks:

- Machine translation
- Text summarization
- Dialogue systems and chatbots
- Creative writing assistance

Speech Tasks:

- Speech recognition (ASR)
- Text-to-speech synthesis
- Speaker identification

Core NLP Tasks (Contd.)

Advanced Tasks:

- Semantic parsing
- Coreference resolution
- Discourse analysis
- Language modeling (predicting next word)

Multimodal Tasks:

- Image captioning
- Visual question answering
- Text-to-image generation

NLP Applications in the Real World

Search & Information Retrieval:

- Google Search, Elasticsearch
- Document ranking and relevance

Social Media & Communication:

- Sentiment analysis for brand monitoring
- Content moderation and spam detection
- Language translation in messaging apps

NLP Applications in the Real World (Contd.)

Business Intelligence:

- Extracting insights from customer feedback
- Automated report generation
- Legal document analysis

Finance:

- Fraud detection in transactions
- Market sentiment analysis
- Algorithmic trading strategies

NLP Applications in the Real World (Contd.)

Healthcare:

- Electronic health record processing
- Medical literature analysis
- Clinical decision support

Education:

- Automated essay scoring
- Intelligent tutoring systems
- Language learning apps

NLP Applications in the Real World (Contd.)

Entertainment & Media:

- Content recommendation
- Automated journalism
- Interactive storytelling

> *"Every time you use Google, ask Siri a question, or get a Netflix recommendation, you're experiencing NLP!"*

Challenges in NLP

1. Ambiguity

- **Lexical:** "Bank" (financial vs. river)
- **Syntactic:** "Flying planes can be dangerous"
- **Semantic:** "He saw her duck" (verb vs. noun)

2. Context Dependency

- Same words, different meanings in different contexts
- Sarcasm and irony detection
- Cultural and temporal context

Challenges in NLP (Contd.)

3. Variability

- Different ways to express the same meaning
- Spelling errors, abbreviations, slang

4. Data & Resource Issues

- Need for large, high-quality datasets
- Low-resource languages (like indigenous languages in Guyana!)
- Bias in training data

Challenges in NLP (Contd.)

5. Computational Complexity

- Real-time processing requirements
- Scaling to billions of documents
- Memory and computational constraints

6. Evaluation Challenges

- How do you measure "understanding"?
- Subjective tasks like creativity and humor

NLP in a Developing Nation Context 🇬🇾

> *"Language technology should serve everyone, not just English speakers!"*

Opportunities:

- **Multilingual Solutions:** Supporting Guyanese Creole, Indigenous languages
- **Educational Technology:** Language learning platforms
- **Digital Inclusion:** Voice interfaces for low-literacy populations
- **Cultural Preservation:** Digitizing and processing cultural texts

Challenges:

- Limited computational resources
- Lack of language data for local languages and need for culturally appropriate solutions

Why Study NLP Now?

1. Explosive Growth in Language Data

- Social media, web content, digital documents
- Need for automated processing and understanding

2. Breakthrough Technologies

- GPT, BERT, and other transformer models
- Democratization of NLP through APIs and libraries

3. Massive Economic Impact

- Trillion dollar industry (OpenAI, Google, Microsoft, etc. all rely on NLP)
- High demand for NLP engineers and researchers

Why Study NLP Now? (Contd.)

4. Societal Impact

- Breaking down language barriers
- Democratizing access to information
- Enabling new forms of human-computer interaction

Course Objectives

By the end of **MAI 5201**, you should be able to:

1. **Understand** fundamental NLP concepts and techniques
2. **Implement** core NLP algorithms from scratch
3. **Apply** modern deep learning approaches to NLP tasks
4. **Evaluate** NLP systems using appropriate metrics
5. **Design** end-to-end NLP applications
6. **Critically analyze** recent research papers in NLP
7. **Consider** ethical implications of NLP technology

Note: This course builds on MAI 5101 (AI Fundamentals) and assumes familiarity with Python and basic machine learning concepts.

Weekly Roadmap (Weeks 1-4)

Week	Topics	Key Concepts
1	*Introduction & Text Processing* - Course overview, RegEx, tokenization	Environment setup, basic text processing
2	*Text Processing Fundamentals* - Edit distance, normalization	String algorithms, preprocessing
3	*N-gram Language Models* - Statistical LMs, smoothing	Probability, maximum likelihood
4	*Text Classification* - Naive Bayes, sentiment analysis	Classification, evaluation metrics

Weekly Roadmap (Weeks 5-8)

Week	Topics	Key Concepts
5	*Logistic Regression* - Binary/multiclass classification	Optimization, regularization
6	*Vector Semantics* - TF-IDF, Word2Vec, GloVe	Distributional semantics
7	*Neural Networks for NLP* - Feedforward networks, embeddings	Backpropagation, word representations
8	*RNNs and LSTMs* - Sequence modeling	Recurrence, vanishing gradients

Weekly Roadmap (Weeks 9-12)

Week	Topics	Key Concepts
9	*Transformers* - Attention, self-attention	Transformer architecture
10	*Large Language Models* - GPT family, scaling laws	Generative models, emergence
11	*Masked Language Models* - BERT, fine-tuning	Pre-training, transfer learning
12	*Model Alignment* - Prompting, RLHF	Instruction following, safety

Weekly Roadmap (Weeks 13-15)

Week	Topics	Key Concepts
13	*Advanced Applications* - Information extraction, QA	Real-world systems
14	*Ethics & Current Research* - Bias, fairness, future directions	Responsible AI, current trends
15	*Project Presentations*	Student projects and demos

Note: Schedule may be adjusted based on class progress and interests.

Assignments & Assessment

Assessment Breakdown:

Component	Weight	Description
Homework (5 total, lowest dropped)	40%	Programming assignments
Research Paper Presentation	10%	Present one NLP paper
Class Participation	5%	Active engagement
Paper Summaries	5%	Critical analysis of papers
Course Project	40%	Applied or research project

Course Project Options

Option 1: Applied NLP Project

- Build a real-world NLP application
- Focus on implementation and evaluation

Option 2: Research Project

- Implement and extend a recent research paper
- Conduct experiments and analysis
- Write a research-style report

Course Project Timeline

Key Dates:

- Week 6: Project proposal due
- Week 10: Progress report
- Week 15: Final presentation and demo

Modern NLP: The Transformer Revolution

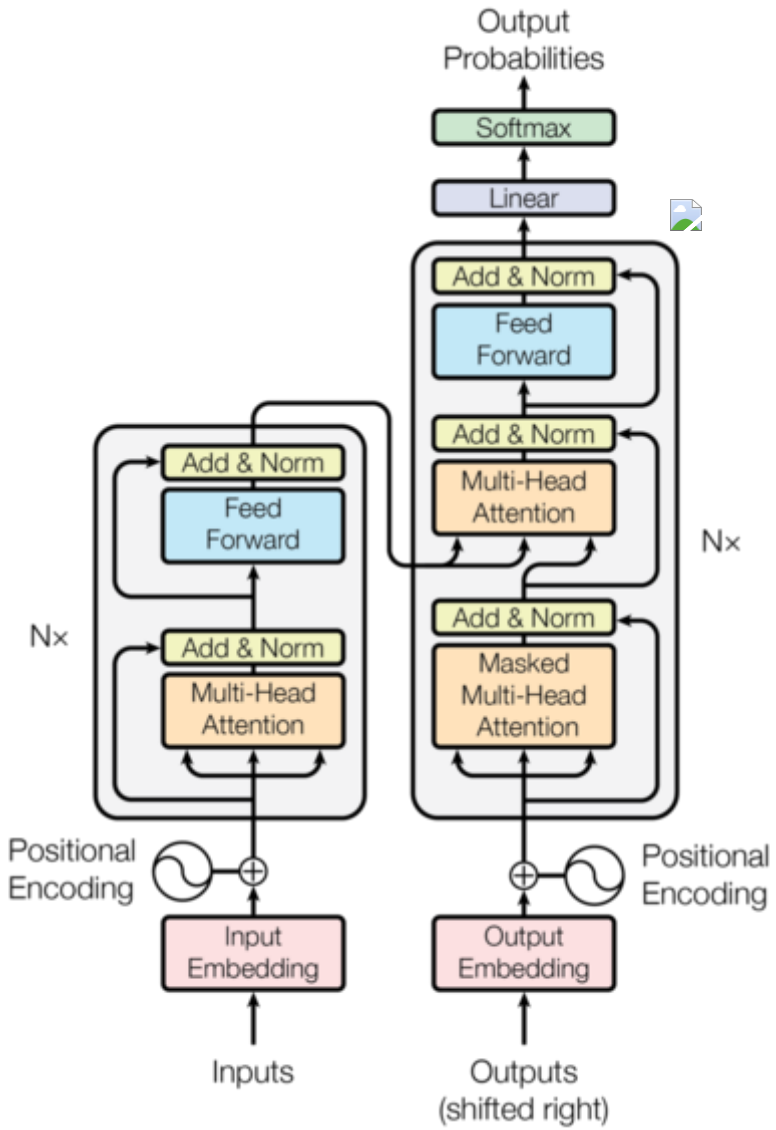
Before 2017: Rule-based → Statistical → Early Neural

After 2017: Transformer architecture changes everything

Key Innovations:

- **Attention mechanism** ("Attention is All You Need")
- **Parallelizable training** (faster than RNNs)
- **Transfer learning** (pre-train then fine-tune)
- **Scaling laws** (bigger models = better performance)

Impact: GPT, BERT, ChatGPT, and the current AI revolution



Large Language Models: A New Paradigm

What are LLMs?

- Neural networks trained on massive text corpora
- Billions or trillions of parameters
- Can perform many tasks without task-specific training

Capabilities:

- Text generation and completion / Translation, summarization, Q&A
- Code generation and debugging / Creative writing and reasoning

> *"We've gone from teaching computers specific tasks to teaching them to understand and generate language in general."*

Practical Considerations

Tools & Libraries We'll Use:

- **NLTK & spaCy** for text processing
- **scikit-learn** for traditional ML
- **PyTorch/TensorFlow** for deep learning
- **Hugging Face Transformers** for pre-trained models
- **Google Colab** for GPU computing

Programming Expectations:

- Comfortable with Python
- Basic understanding of NumPy and pandas
- Willingness to learn new libraries

Next Steps

Before Next Class:

- **Review** the course README and syllabus
- **Set up** your development environment (Python, Jupyter, etc.)
- **Read** Jurafsky & Martin Chapter 1
- **Think about** potential project ideas

This Week:

- HW 0 will be released (environment setup + basic NLP exploration)
- Start thinking about which research paper you'd like to present

Questions? Email me at christopher.clarke@uog.edu.gy