Dr. Denton Bobeldyk

CIS 365 Artificial Intelligence

Natural Language Processing

Delivery Methods

Lecture

Videos

Lab Time

Small Groups

Natural Language Processing (NLP)

- Subfield of artificial intelligence and linguistics.
- Focuses on enabling computers to understand, interpret, and respond to human language in a a way that is both meaningful and useful
- Combines computational techniques and linguistic principles to process and analyze large amounts of natural language data

Text Processing

- Tokenization
  - Splitting text into words or sentences
- Stopword removal
  - Eliminating common words like “is”, “and”, and “the”
- Lemmatization/Stemming
  - reducing words to their base or root form

Text Processing

- Tokenization
  - Splitting text into words or sentences
- Stopword removal
  - Eliminating common words like “is”, “and”, and “the”
- Lemmatization/Stemming
  - reducing words to their base or root form

Applications of NLP

Applications of NLP

- Text Analysis
- Speech Processing
- Information Retrieval
- Language Translation
- Question Answering Systems
- Content Generation

Text Analysis

- Sentiment Analysis
  - Determining the emotional tone of text (e.g., positive or negative reviews)
- Topic Modeling
  - Identifying the main topics in a collection of documents
- Customer Review Summarization
  - Summarizing reviews of a product (e.g., Customers were happy with the product and tend to mention stick drift)

Speed Processing

- Speech Recognition: Converting spoken language into text (e.g., Siri, Google Assistant)
- Text-to-speech: Generating human-like speech from text

Information Retrieval

- Search Engines
  - Understanding and ranking web content based on queries
- Chatbots and virtual assistants
  - Providing automated conversational responses

Language Translation

- Machine Translation
  - Converting text or speech from one language to another (e.g., google translate)

Question Answering Systems

- Answering natural language questions (e.g., IBM Watson, OpenAI's GPT)

Name some question/answer systems

- IBM Watson
- Chat GPT
-

Name some question/answer systems

- IBM Watson
- Chat GPT
- Google Bard
- Meta's LlaMa
- X's Grok

Content Generation

- Summarization
  - Produce concise summaries of long texts
- Text generation
  - Creating human-like text (e.g., GPT based models)

Key components of NLP

Key components of NLP

- Text Processing
- Linguistic Analysis
- Statistical and Machine Learning Techniques

Text Processing

- Tokenization
  - Splitting text into words or sentences
- Stopword removal
  - Eliminating common words like “is”, “and”, “the”
- Lemmatization/Stemming
  - Reducing words to their base or root form

Linguistic Analysis

- Syntax
  - Understanding the grammatical structure of sentences (e.g., parsing)
- Semantics
  - Deriving meaning from words and sentences
- Pragmatics
  - Interpreting language in context, including implied meanings

Statistical and Machine Learning Techniques

- Used to model language patterns, predict outcomes and classify text
- Algorithms like hidden Markov models, conditional random fields and deep learning models play a significant role

Challenges in NLP

Challenges in NLP

- Ambiguity
- Context Understanding
- Low-resource languages
- Domain specific jargon
- Multilingual processing

Challenges - Ambiguity

- Words and sentences often have multiple meanings depending on context
- “I saw her duck” could refer to an action or a bird

Challenges - Context Understanding

- Understanding nuances like sarcasm, idioms or implied meanings

Challenges - Low-resource languages

- Many languages lack sufficient labeled data for effective NLP development
- Types of low resource languages:
  - Indigenous and endangered languages
    - Aymara (spoken in Bolivia, Peru, Chile)
    - Hausa (spoken in Nigeria, Niger, and neighboring areas)
  - Regional and minority languages
    - Welsh - spoken in Wales
    - Twi - spoken in Ghana

Challenges - Low-resource languages

- Types of low-resource languages continued:
  - Underrepresented official languages
    - Amharic
      - Official language of Ethiopia
    - Lao
      - Official language of Laos
    - Sinhala
      - Official language of Sri Lanka

Challenges - Domain-specific jargon

- Handling specialized language in fields like medicine or law

Challenges - Multilingual Processing

- Managing language variations, dialects, and multilingual inputs

Popular Techniques and Models

Popular Techniques and Models

- Rule based approaches
- Statistical methods
- Deep Learning
- Pretrained language models

Rule-based approaches

- Early NLP systems relied on predefined linguistic rules
- Limited flexibility but useful for specific applications

Statistical methods

- Probabilistic models like n-grams and Markov models
- Effective for text prediction and language modeling

Statistical methods Example

- Train your model on a large amount of text as our ‘corpus’

Statistical methods Example

- Train your model on a large amount of text as our ‘corpus’

What sort of ethical concerns are there for this?

Statistical methods Example

- Train your model on a large amount of text as our ‘corpus’

For example:
I love natural language processing.

I love programming.

I enjoy learning NLP.

Statistical methods Example

- Generate Bigram probabilities

$$P(w_n \mid w_{n-1}) = \frac{\text{Count}(w_{n-1}, w_n)}{\text{Count}(w_{n-1})}$$

$P(w_n \mid w_{n-1})$ = probability of w_n, given the preceding word w_{n-1}

$\text{Count}(w_{n-1}, w_n)$ = the count of the bigram

$\text{Count}(w_{n-1})$ = count of the preceding word

Statistical methods Example

- Generate Bigram counts
  - Count(I,love) = 2
  - Count(love, natural) = 2
  - Count(love, programming) = 1
  - Count(natural, language) = 1
  - Count(language, processing) = 1
  - Count(enjoy, learning) = 1
  - Count(learning, NLP) = 1

Statistical methods Example

- Generate Bigram counts
  - Count(I) = 3
  - Count(love) = 2
  - Count(natural) = 1
  - Count(language) = 1
  - Count(enjoy) = 1
  - Count(learning) = 1

Statistical methods Example

- Bigram Probabilities:
  - $P(\text{love} \mid \text{I}) = 2/3$
  - $P(\text{programming} \mid \text{love}) = 1/2$
  - $P(\text{natural} \mid \text{love}) = 1/2$
  - $P(\text{language} \mid \text{natural}) = 1$
  - $P(\text{processing} \mid \text{language}) = 1$
  - $P(\text{learning} \mid \text{enjoy}) = 1$
  - $P(\text{NLP} \mid \text{learning}) = 1$

Statistical methods Example

- Predict the next word:
  - Start with “I”
    - $P(\text{love} \mid \text{I}) = 2/3$
    - $P(\text{enjoy} \mid \text{I}) = 1/3$
    - Predict “love”
  - Current word is “love”:
    - $P(\text{natural} \mid \text{love}) = 1/2$
    - $P(\text{programming} \mid \text{love}) = 1/2$
    - Predict “natural” or “programming” (randomly choose if equal)

Deep Learning

- Neural networks like LSTM, GRUs and Transformers (e.g., BERT, GPT)
- Achieve state of the art performance in tasks like translation, summarization and question answering

Pretrained Language Models

- BERT
- GPT

Future Directions of NLP

- Explainable NLP
  - Developing systems that can explain their reasoning for decisions
- Real-Time Multimodal processing
  - Combining NLP with computer vision applications like captioning videos
- Personalized NLP
  - Tailoring models to individual user preference and behavior
- Low-resource language inclusion
  - Building models that work effectively for languages with limited data