Text Preprocessing in NLP

Dr. Sambit Praharaj

Assistant Professor (II)

Text Preprocessing

- Tokenization and Normalization
- Sentence, Word, Subword Tokenization
- Stemming, Lemmatization, Stop Words

What is Tokenization?

- Process of breaking text into smaller units called tokens.
- Text → Tokens → Numbers → Model

Why Tokenization is Needed

- Machines cannot understand raw text
- Tokenization enables feature extraction
- Foundation of all NLP pipelines

Types of Tokenization

- Sentence Tokenization
- Word Tokenization
- Subword Tokenization
- Character Tokenization

Sentence Segmentation

- Splits text into sentences
- Important for translation, summarization
- Challenges: abbreviations, decimals

Sentence Segmentation Example

- Text: I love NLP. It is interesting!
- Output:
- 1. I love NLP.
- 2. It is interesting!

Word Tokenization

- Splits sentences into words and punctuation
- Handles contractions and symbols

Word Tokenization Example

- Sentence: I can't believe this works!
- Tokens: I, ca, n't, believe, this, works, !

Limitations of Word Tokenization

- Vocabulary explosion
- Out-of-vocabulary (OOV) words
- Morphological variations

Need for Subword Tokenization

- Handles unseen words
- Smaller vocabulary
- Better generalization

What is Subword Tokenization?

- Breaks words into smaller meaningful units
- Learned automatically from data

Byte Pair Encoding (BPE)

- Data-driven subword algorithm
- Starts with characters
- Merges frequent pairs

BPE Training Example

- Corpus:
- low
- lower
- newest
- widest
- Characters with end symbol (_)

BPE Merging Process

- Find most frequent pair
- Merge the pair
- Repeat until vocab size reached

BPE Tokenization Example

- lowest → low + est
- No unknown word

BPE Properties

- Frequency-based
- No linguistic rules
- Used in GPT models

WordPiece Tokenization

- Improved version of BPE
- Probability-based merging
- Uses continuation marker ##

WordPiece Vocabulary

- Whole words: play
- Prefixes: un
- Suffixes: ##ing, ##er

WordPiece Example

- playing → play + ##ing
- unbelievable → un + believe + ##able

Longest Match First Rule

- Choose longest valid subword
- Add ## if continuation

BPE vs WordPiece

- BPE: frequency-based
- WordPiece: probability-based
- WordPiece preserves word boundaries

Stemming

- Reduces words to base form using rules
- May not produce valid words

Porter Stemmer

- Rule-based suffix stripping
- running → run
- university → univers

Stemming Limitations

- No grammatical understanding
- May distort word meaning

Lemmatization

- Uses vocabulary and POS
- Produces dictionary words

Lemmatization Example

- better → good
- mice → mouse

Stemming vs Lemmatization

- Stemming: fast, inaccurate
- Lemmatization: slow, accurate

Stop Words

- Common frequent words
- the, is, am, are, of

Stop Words Removal Example

- This is a very good book
- After removal: very, good, book

When Not to Remove Stop Words

- Sentiment analysis
- Question answering

Complete NLP Pipeline

- Raw Text
- → Sentence Segmentation
- → Tokenization
- → Stop Word Removal
- → Stemming/Lemmatization
- → Feature Extraction

Summary

- Tokenization is the foundation of NLP
- Subword methods solve OOV
- Stemming and Lemmatization reduce variants