Natural Language Processing and Transformers: A Summary

Natural Language Processing (NLP) is a field at the intersection of linguistics, computer science, and artificial intelligence that enables machines to understand, interpret, and generate human language. It powers applications such as spell check, autocomplete, voice assistants, spam filters, and machine translation. NLP systems can process vast amounts of text data, outperforming humans in scale and speed.

Key capabilities of NLP include sentiment analysis, topic modeling, named entity recognition (NER), text categorization, clustering, and information extraction. Sentiment analysis identifies emotional tone, while topic modeling (e.g., Latent Dirichlet Allocation) uncovers hidden themes in large text corpora. NER detects and classifies entities like people, locations, and organizations. Text categorization and clustering group documents by themes or topics, and information extraction pulls structured data from unstructured text.

Despite its power, NLP faces challenges. Language is inherently ambiguous, with sarcasm, irony, and domain-specific jargon complicating interpretation. For example, phrases like "no woman no cry" or "to short a stock" require contextual understanding beyond grammar.

Language modeling is central to NLP. It involves estimating the probability of word sequences, enabling applications like autocomplete, chatbots, and translation. Traditional models like n-gram language models use statistical methods to predict the next word based on previous ones. Bigrams and trigrams capture local word order, improving over simple bag-of-words (BoW) models, which ignore word order and represent text as word frequency vectors.

To better capture meaning, skip-grams and word embeddings were introduced. Skip-grams consider

non-consecutive word pairs, helping to learn semantic relationships. Word embeddings, such as those produced by Word2Vec, represent words as dense vectors based on their context. These vectors capture relationships like "king - man + woman = queen" and support analogical reasoning. However, Word2Vec assigns a single vector per word, failing to distinguish between different senses of polysemous words like "bank."

To address this, contextual embeddings were developed. Models like ELMo, BERT, and GPT generate word representations that vary with context, capturing nuanced meanings. These models are built on the transformer architecture, which uses self-attention mechanisms to model dependencies across entire sequences, enabling parallel processing and superior performance.

Transformers have revolutionized NLP. Unlike recurrent models, they process entire sequences simultaneously, making them faster and more effective. BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) are pre-trained on large corpora and fine-tuned for specific tasks, achieving state-of-the-art results in translation, summarization, and question answering.

In conclusion, NLP has evolved from simple statistical models to sophisticated transformer-based architectures. Its capabilities in understanding and generating human language are transforming industries, making it a cornerstone of modern AI applications.