Key Milestones in NLP

1948: The first NLP application, a dictionary look-up system for automated translation (German to English, later Russian to English), was developed at Birkbeck College, London.

1957

Noam Chomsky's publication of "Syntactic Structures" revolutionized linguistics and significantly influenced NLP. His work influenced the invention of Backus-Naur Form (BNF) notation (1963) for representing programming language syntax and Regular Expressions (1956) for specifying text search patterns.

1966

The ALPAC Report highlighted the limited success of machine translation, leading to a funding drought until 1980. However, progress continued in areas like case grammar and semantic representations, although much of the work focused on syntax.

1970

NLP shifted towards AI, emphasizing world knowledge and meaningful representations. Semantics gained prominence, exemplified by systems like SHRDLU (1973) and LUNAR (1978). This led to the adoption of logic for knowledge representation and reasoning in the 1980s, along with the invention of the Prolog programming language for NLP applications.

1980

Machine learning became increasingly important, marking the birth of statistical NLP. Annotated text corpora were used to train models and evaluate performance. ML approaches dominated through the 1990s, partly due to the success of Hidden Markov Models in speech recognition. The shift towards statistical methods is famously summarized by Fred Jelinek's quote: "Every time I fire a linguist, the performance of our speech recognition system goes up."

1982

Project Jabberwacky, aimed at simulating natural conversations to pass the Turing Test, was launched, initiating the development of chatbots. It won third place in the Loebner Prize in 2003.

1998

The FrameNet project, focused on semantic role modeling (a type of shallow semantic parsing), was introduced and continues to be researched.

Key Milestones in NLP (Continued)

2001: Word embeddings (vector inputs in feed-forward neural networks) were proposed for language modeling, later leading to the use of RNNs (2010) and LSTMs (2013).

2003: Latent Dirichlet Allocation (LDA) was invented and became a standard method for topic modeling.

2013: Improved word embeddings and word2vec's efficient implementation spurred the wider adoption of neural networks in NLP. RNNs and LSTMs became popular due to their ability to handle dynamic input sequences, while CNNs from computer vision were adapted for their parallelizability. Recursive Neural Networks were also explored to leverage the hierarchical nature of language.

March 2016: Microsoft launched Tay, a chatbot on Twitter, which was shut down within 16 hours due to its adoption of racist and abusive language. Zo, a replacement chatbot, was launched months later.

September 2016

Google replaced its phrase-based translation system with Neural Machine Translation (NMT) using a deep LSTM network, reducing translation errors by 60%. This was based on sequence-to-sequence learning (proposed in 2014), which became a preferred technique for NLG.

NLU vs NLG Analysis

NLU vs NLG

NLU	NLG
1. NLU is taking some spoken / typed sentence and working out what it means.	1. NLG is taking some formal representation of what you want to say & working out a way to express it in a natural language.
2. In NLU the system needs to disambiguate the input sentence to produce the machine representation language	2. In NLG the system needs to make decisions about how to put a concept into words.
3. Different levels of analysis required: morpological analysis, syntactic analysis, semantic analysis, discourse analysis.	3. Different levels of synthesis required: deep learning (what to say), syntactic generation.
4. NLU is most harder than NLG.	4. NLG is less harder than NLU.

Why do computers have difficulty with NLP?

1. Computers traditionally handle structured data (organized, indexed, and referenced, often in databases).
2. NLP deals with unstructured data (e.g., social media posts, news articles).

3. NLP must learn the structure and grammar of natural language (80% of enterprise data is unstructured).
4. Human language is complex (ambiguous phrases, colloquialisms, metaphors, etc.).
5. Words and text can have multiple meanings depending on context.
6. Language evolves, and humans communicate imperfectly (spelling, grammar, punctuation errors).
7. Ambiguities are lexical, syntactic, or referential.
8. Speech adds further challenges (accent, tone, noise, etc.).

Examples of English Complexities:

1. "One morning I shot an elephant in my pajamas." (Ambiguity: who was in pajamas?)
2. "Listening to loud music slowly gives me a headache." (Ambiguity: what was slow?)
3. "The complex houses married and single soldiers and their families." (Ambiguity: "complex" as noun or adjective?)
4. "John had a card for Helga, but couldn't deliver it because he was in her way." (Coreference resolution: who is "he"?)
5. "The Kiwis won the match." (Contextual understanding: "Kiwis" as New Zealanders).