Key Milestones in NLP

1948

Early NLP applications emerge with a dictionary look-up system for automated translation (German to English initially, then Russian to English during the Cold War) developed at Birkbeck College, London.

1957

Noam Chomsky publishes "Syntactic Structures," revolutionizing linguistics and significantly influencing NLP. His work impacts later developments such as Backus-Naur Form (1963) for representing programming language syntax and Regular Expressions (1956) for specifying text search patterns.

1966

The ALPAC Report highlights the limited success of machine translation, leading to a funding drought in the field until the 1980s. Despite this, progress continues in areas like case grammar and semantic representations, although much of the work focuses on syntax.

1970s

NLP is influenced by AI, shifting focus towards world knowledge and meaningful representations. Semantics gains importance. Key systems of this era include SHRDLU (1973) and LUNAR (1978), leading to the adoption of logic for knowledge representation and reasoning in the 1980s. The Prolog programming language, invented in 1970, is also applied to NLP.

1980s

The decade marks the rise of Machine Learning and statistical NLP. Annotated text corpora are used to train ML models, providing a gold standard for evaluation. ML approaches to NLP become prominent throughout the 1990s, partly due to the success of Hidden Markov Models in speech recognition. The increased effectiveness of statistical methods over purely linguistic approaches is famously summarized by Fred Jelinek: "Every time I fire a linguist, the performance of our speech recognition system goes up."

1982

Project Jabberwacky is launched, marking the beginning of chatbots aimed at simulating natural human conversations and passing the Turing Test. It achieved third place in the Loebner Prize in October 2003.

1998

The FrameNet project is introduced, focusing on semantic role modelling – a form of shallow semantic parsing that remains an active area of research.

Key Milestones in NLP (Continued)

2001

Word embeddings, represented by vector inputs in feed-forward neural networks, are proposed as an improvement over classical N-gram models for language modeling. This paved the way for the later use of RNNs (2010) and LSTMs (2013).

2003

Latent Dirichlet Allocation (LDA) is invented and quickly becomes a standard method for topic modeling in machine learning.

2013

Improvements in word embeddings, coupled with efficient implementation in Word2vec, lead to increased adoption of neural networks for NLP. RNNs and LSTMs become popular choices due to their ability to handle the dynamic input sequences common in NLP. CNNs, adapted from computer vision, are also used due to their parallelizability. Recursive Neural Networks are explored to leverage the hierarchical nature of language.

March 2016

Microsoft launches Tay, a chatbot on Twitter, which is shut down within 16 hours due to its adoption of racist and abusive language learned from user interactions. Microsoft later launched the Zo chatbot.

September 2016

Google replaces its phrase-based translation system with Neural Machine Translation (NMT) using a deep LSTM network. This resulted in a 60% reduction in translation errors. This work builds upon sequence-to-sequence learning (proposed in 2014), which becomes a preferred technique for Natural Language Generation (NLG).

NLU vs NLG Comparison

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, syntachic analysis, semantic analysis, discourse analysis.	3. Different levels of synthesis required: deep learning (what to say), syntachic 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 frequently deals with unstructured data (e.g., social media posts, news articles, emails).

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, sarcasm).
5. Words and text can have multiple meanings depending on context.
6. Language evolves, and human communication is imperfect (spelling, grammar, punctuation errors).
7. Ambiguities are lexical, syntactic, or referential.
8. Speech adds challenges (accent, tone, noise, pronunciation, emotion).

Examples of English Language Complexities

1. "One morning I shot an elephant in my pajamas" – ambiguity regarding who was wearing the pajamas.
2. "Listening to loud music slowly gives me a headache" – ambiguity regarding what happened slowly.
3. "The complex houses married and single soldiers and their families" – "complex" is a noun, not an adjective; this requires part-of-speech tagging.
4. "John had a card for Helga, but couldn't deliver it because he was in her way" – coreference resolution needed to understand "he".
5. "The Kiwis won the match" – requires contextual understanding of "Kiwis" as New Zealanders.