Key Milestones in NLP

1948

The first NLP application can be considered a dictionary look-up system for automated translation developed at Birkbeck College, London. Post-World War II research focused on German-to-English translation, later shifting to Russian-to-English during the Cold War.

1957

Noam Chomsky's publication of *Syntactic Structures* revolutionized linguistics and significantly influenced NLP. His work impacted the invention of Backus-Naur Form (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. Despite this, progress was made in areas like case grammar and semantic representations, although much of the work until the late 1960s focused on syntax.

1970

NLP shifted towards AI, emphasizing world knowledge and meaningful representations, making semantics more important. 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, also found application in NLP.

1980

The increasing adoption of Machine Learning marked the beginning of statistical NLP. Annotated text corpora became crucial for training ML models and establishing evaluation standards. ML approaches to NLP gained prominence throughout the 1990s, partly driven by the success of Hidden Markov Models in speech recognition. The increased success of statistical methods over purely linguistic approaches 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, launched to simulate natural human conversations and attempt to pass the Turing Test, marked the beginning of chatbots. It won third place in the Loebner Prize in October 2003.

1998

The FrameNet project, focusing on semantic role modeling (a form of shallow semantic parsing), was introduced and continues to be an area of research.

2001

Researchers proposed using feed-forward neural networks with vector inputs (word embeddings) for language modeling, paving the way for the later use of RNNs (2010) and LSTMs (2013).

2003

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

2013

Improvements to word embeddings and efficient implementations like Word2vec increased the adoption of neural networks for NLP. RNNs and LSTMs became popular choices due to their ability to handle dynamic input sequences. CNNs from computer vision were repurposed for NLP due to their parallelizability. Recursive Neural Networks were also explored to leverage the hierarchical nature of language.

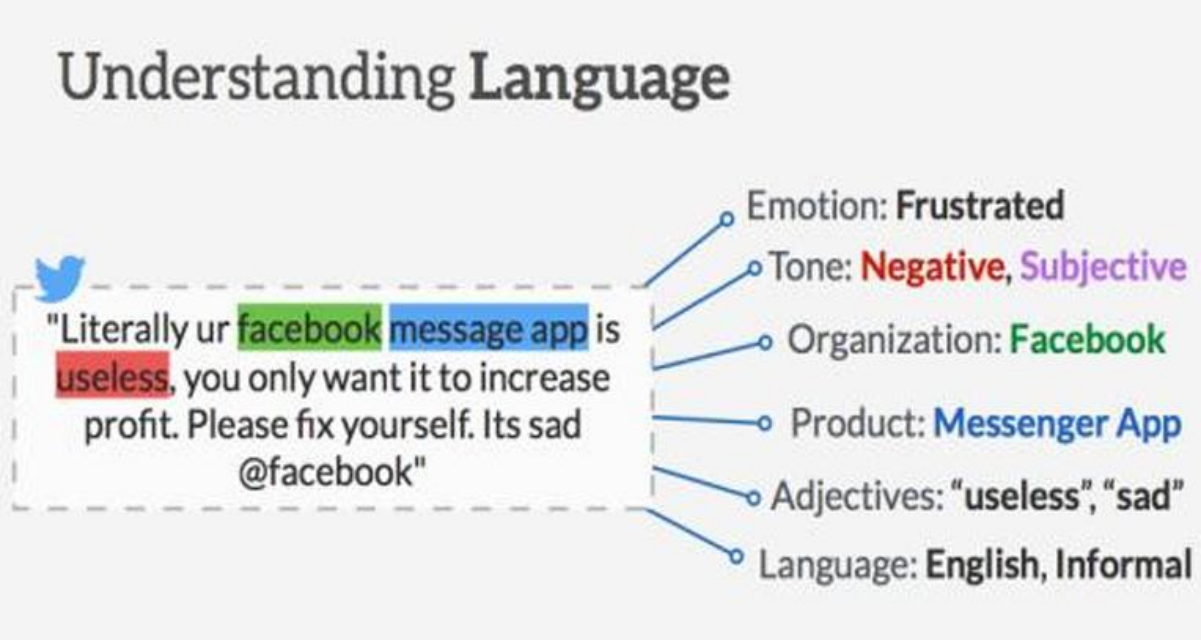
March 2016

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

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 built upon sequence-to-sequence learning (proposed in 2014), which became a preferred technique for Natural Language Generation (NLG).

Why do computers have difficulty with NLP?


Understanding Language
"Literally ur facebook message app is useless, you only want it to increase profit. Please fix yourself. Its sad @facebook"
Emotion: Frustrated
Tone: Negative, Subjective
Organization: Facebook
Product: Messenger App
Adjectives: "useless", "sad"
Language: English, Informal

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 human communication is imperfect (spelling, grammar, punctuation errors).
7. Ambiguities can be lexical, syntactic, or referential.
8. Speech adds challenges (accent, tone, noise, pronunciation, emotion, pauses).

Examples of English Language 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 happened slowly?)
3. "The complex houses married and single soldiers and their families." (Ambiguity: "complex" as noun or adjective?) This is addressed via part-of-speech tagging.
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." (Requires context: "Kiwis" refers to New Zealanders.)