Evolution of Natural Language Processing (NLP) The Past, Present and Future

Dheeraj Patta (*npatta2@illinois.edu*)

ABSTRACT

The goal of this technical review is to highlight the ever evolving and constantly advancing landscape of an exciting field of Artificial Intelligence – Natural Language Processing (NLP). This review aims at shedding light into the origins of NLP, its journey and development as a field and advancements till date. This is not an extensive review of each of the technologies or methodologies or methods instead this should be treated as a brief introduction to history and evolution of Natural Language Processing (hereby referred as "NLP"). This is a consolidation of various articles, papers and research work done by respective authors and dignitaries.

INTRODUCTION

Natural language processing (NLP) is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis *(Liddy, 2001)*. The purpose of these techniques is to achieve human-like language processing for a range of tasks or applications. NLP is a research field dedicated to helping machines recognize, understand language, translate and better communicate with humans. There are wide range of applications and advancements in NLP including conversational agents, sentiment analysis, recommendations, behavioral analysis, speech recognition, smart assistants, machine translation, language understanding and generation etc.

THE PAST (1940 – 2013)

Although the prominence of NLP rose around 2018 with incredible contributions from Google (BERT) or the Transformer paper titled "Attention is all you need", the origins of NLP dates back to World War II era.

Considered one of the foundational research pieces ever about Computer Science and Artificial Intelligence titled *"Computing Machinery and Intelligence"* (*Alan Turing,1950*), it poses a simple and seminal question "Can Machines Think?" and a means to answer it by "The Turing Test". Simply put, if a machine could be part of a conversation and imitated a human so that there are no noticeable differences, the machine is capable of thinking. *The Hodgkin-Huxley* model *(1952)* demonstrated how brain uses neurons in forming an electrical network. These events helped inspire the idea of Artificial Intelligence (AI), Natural Language Processing (NLP), and the evolution of computers.

Weaver's memorandum *(Shannon and Weaver, 1949)* brought the idea of the first computer-based application related to natural language: Machine Translation (MT). It subsequently inspired many

projects, notably the Georgetown experiment *(Dostert, 1955)*, a joint project between IBM and Georgetown University that successfully demonstrated the machine translation of more than 60 Russian sentences into English albeit a small sample size of only 250 words and six grammar rules. However, it quickly failed to generalize language generation as it is based on hand-coded language rules and used dictionary lookups for words for translation and reordered the words to fit the word-order rules of the target language. It took until 1957 to introduce the idea of Generative Grammar *(Chomsky, 1957)*, a rule-based system of syntactic structures and parsing algorithms that brought insight into how mainstream linguistics could help machine translation.

The Automated Language Processing Advisory Committee (ALPAC) report *(1966)* concluded that Machine Translation (MT) was not immediately achievable and resulted in substantially slowing down NLP research. Despite setbacks, there were interesting works in both theoretical and construction of prototype systems in late *1960s* and *1970s*. Researchers focused on the representing Meaning in a language instead of defining language as a set of hard coded rules. After the introduction of transformational generative grammars *(Chomsky, 1965)*, many new theories of grammar were developed to explain syntactic anomalies and provide semantic representations, such as Case Grammar *(Fillmore, 1968)*, Semantic Networks *(Collins, 1969)*, Augmented Transition Networks *(Woods, 1970)*, and Conceptual Dependency Theory *(Schank, 1972)*. Prototypes like ELIZA *(Weizenbaum, 1966)* to replicate the conversation between a psychologist and a patient, LUNAR *(Woods, 1972)* as an interface system to a database that consisted of information about lunar rock samples using augmented transition network and PARRY *(Colby, 1974)* which attempted to simulate a person with paranoid schizophrenia based on concepts, conceptualizations, and beliefs etc. were developed.

The 1970s brought new ideas into NLP, such as building conceptual ontologies which structured real-world information into computer-understandable data. Examples are MARGIE *(Schank and Abelson, 1975)*, TaleSpin *(Meehan, 1976)*, QUALM *(Lehnert, 1977)*, SAM *(Cullingford, 1978)*, PAM *(Schank and Wilensky, 1978)* and Politics *(Carbonell, 1979)*. In the 1980s, many significant problems in NLP were addressed using symbolic approaches *(Charniak, 1983; Dyer, 1983; Riesbeck and Martin, 1986; Grosz, 1987; Hirst, 1987)*, i.e., complex hard-coded rules and grammars to parse language. It wasn't until the late 1980s and early 1990s that statistical models came as a revolution in NLP *(Bahl, 1989; Brill, 1990; Chitrao and Grishman, 1990; Brown, 1991)*, replacing most natural language processing systems based on complex sets of hand-written rules. This progress was the result of both the steady increase of computational power, and the shift to machine learning algorithms. While some of the earliest-used machine learning algorithms, such as decision trees *(Tanaka, 1994; Allmuallim, 1994)*, produced systems similar in performance to the old school hand-written rules, statistical models broke through the complexity barrier of hand-coded rules by creating them through automatic learning, which led research to increasingly focus on these models.

The first neural language model (*Bengio, 2003)* consisting of one-hidden layer and feed-forward neural network was proposed which was premise for word embeddings. In *2008, Collobert and Weston* applied multi-task learning, a sub-field of machine learning in which multiple learning tasks are solved at the same time, to neural networks for NLP. They used a single convolutional neural network architecture (CNN; *LeCunn, 1999*) that, given a sentence, was able to output many language processing predictions such as part-of-speech tags, named entity tags and semantic roles. Multi-task learning has gained in importance and is now used across a wide range of NLP tasks. Also, their paper turned out to be a discovery that went beyond multi-task learning which spearheaded ideas such as pre-training word embeddings and using CNNs.

THE PRESENT (2013– 2020)

Word2Vec (*Mikolov, 2013),* arguably the most popular word embedding model was introduced with an improvement in the training procedure. This enabled large-scale training of word embeddings on huge corpus of unstructured text and captured certain relationships between words such as gender, verb tense, and country-capital relations, which initiated a lot of interest in word embeddings as well as in the origin of these linear relationships *(Mimno and Thompson, 2017; Arora, 2018; Antoniak and Mimno, 2018; Wendlandt, 2018).*It presented the evidence that using pre-trained embeddings as initialization improved performance across a wide range of downstream tasks. Word2Vec was among the first prediction-based modeling techniques in NLP. It included techniques which are frequently used today such as Skip Gram and Continuous Bag of Words. These techniques leverage neural networks as their foundation and consider the semantics of the text. They also use a FastText algorithm (developed by Facebook), that uses character level information to generate the text representation. The word is considered as a bag of character n-grams in addition to the word itself. Later, ELMo came into existence with a key idea of solving the problem of homonyms in word representation.

This year also marked the adoption of neural network models in NLP namely recurrent neural networks (RNNs; *Elman, 1990*), convolutional neural networks (CNNs), and recursive neural networks *(Socher, 2013)*. RNNs were quickly replaced with the classic long-short term memory networks (LSTMs; *Hochreiter and Schmidhuber, 1997*), as they proved to be more resilient to the vanishing and exploding gradient problem. CNNs which are widely adopted by computer vision community started to get applied to natural language *(Kalchbrenner, 2014; Kim, 2014)*. *Sutskever (2014)* proposed sequence-to-sequence learning, a general end-to-end approach for mapping one sequence to another using a neural network. An encoder-decoder combination is used to process and predict the output sequences. Recent models include deep-LSTMs *(Wu, 2016)*, convolutional encoders *(Kalchbrenner, 2016; Gehring, 2017)*, the Transformer *(Vaswani, 2017)*, and a combination of an LSTM and a Transformer *(Chen, 2018)*. Machine translation turned out to be the perfect application for sequence-to-sequence learning.

One of the core innovations in Neural Machine Translation (NMT), the Attention (*Bahdanau, 2015)* principle was introduced which removes the bottleneck of sequence-to-sequence learning by eliminating the need for compressing the input sequence into a fixed-size vector. Attention allows decoder to refer to hidden states. It has been applied to constituency parsing *(Vinyals, 2015)*, reading comprehension *(Hermann, 2015)*, and one-shot learning *(Vinyals, 2016)*. More recently, self-attention was introduced which looks at the surrounding words in a sentence or paragraph to obtain more contextually sensitive word representations. BERT (Bidirectional Encoder Representations from Transformers) introduced by Google, uses encoder representations of the transformer network and has marked a new era in NLP by breaking several records in handling language-based tasks and topping GLUE benchmarks when released. XLNet *(by Carnegie Mellon & Google, 2019)* outperformed BERT in 20 different NLP tasks. Other notable advancements include Baidu's ERNIE, Google's T5, Alibaba's DAMO which are variations of BERT heavily focusing on attention and pretrained weights.

The latest major breakthrough in NLP is undoubtedly large pretrained language models. Pre-trained language model embeddings can be used as features in a target model *(Peters, 2018)*, or a

pre-trained language model can be fine-tuned on target task data *(Devlin, 2018; Howard and Ruder, 2018; Radford, 2019; Yang, 2019)*, which have shown to enable efficient learning with significantly less data. The main advantage comes from their ability to learn word representations from large unannotated text corpus, which is particularly beneficial for low-resource languages where labelled data is scarce. GPT-2 and GPT-3 by OpenAI are extreme examples of large pretrained models with billions of parameters and shown incredible results in NLP tasks like text generation, question and answering systems etc.

THE FUTURE (2020 -)

In order to better process language data, be it text or audio we need to develop systems that have a memory element, so it can understand the context. We will continue to see advancements in the areas of intelligent systems evolving from human-computer interaction to human-computer conversation. There will also be continuous improvements in non-verbal communication which includes body language, touch, gestures, and facial expressions. These different micro expressions can be used to show the feelings in a conversation. These micro-expressions are keys to identifying the difference between different sentiments and emotions and coupled with Natural language processing units can unlock a whole new level of interactions. We will see the rise of intelligent agents capable of understanding and integrating common knowledge (pun, sarcasm, common sense etc.) into its language interpretation and conversations with humans.

REFERENCES

https://gluebenchmark.com/leaderboard

https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-1-ffbcb937ebce

https://en.wikipedia.org/wiki/History_of_natural_language_processing

https://towardsdatascience.com/evolution-of-natural-language-processing-8e4532211cfe

https://ruder.io/a-review-of-the-recent-history-of-nlp/

https://developpaper.com/14-breakthroughs-in-nlp-research-that-can-be-applied-in-practice-1/

https://www.analyticsvidhya.com/blog/2019/08/complete-list-important-frameworks-nlp/

https://www.dataversity.net/advances-in-natural-language-processing/

https://analyticsweek.com/content/2020-trends-in-natural-language-processing/

https://xaltius.tech/latest-developments-in-natural-language-processing/