Overview of NLP

Arjun Bala

a. Natural Language Processing is the set of method used for computers to interact with human languages. There are many different types of interactions such as text understanding, text generation, etc.
b. Natural Language Processing methods are a subset of Artificial Intelligence as AI refers to the computer's ability to achieve tasks typically done by humans. One task that the human brain does is read, understand, and generate new text and NLP is the branch of AI related to this.
c. Natural Language Understanding is the way a computer can use human natural text to gain some sort of understanding. For example, the Amazon Alexa uses voice to text data to understand that when someone says "Alexa, could you play music" Alexa understands and does as asked. Natural Language Generation is where the computer takes an input to generate text based upon the input. For example, DALL-E 2 can take an image as an input and generate a text description of the image.
d. Some examples of modern NLP applications are
    a. Voice assistants (i.e., Alexa, Google, Siri, etc.)
    b. Grammarly uses NLP to correct text
    c. Google's autofill for predicting the rest of your search
    d. Voice to text data to transcribe audio
    e. Services such as google translate to translate text
e. The 3 main approaches are:
    a. Rules-Based

        In rule-based learning, there is a human who creates rules for the computer to use when approaching text. In this method it is easy for people to debug as people can look at what error it is and just edit the code as all the code is more intuitive than the other methods used. One advantage rule-based learning has is that it requires little data to be trained as most rules are written by humans so there is not too much training that takes place. One disadvantage is that these methods don't generalize well to tasks other than the ones they are specifically created for. Some use cases of rule-based learning are figuring out parts of speech or categorizing text.

    b. Statistical and Probabilistic
        i. In statistical and probabilistic modelling, we use machine learning methods for the computer to interact with natural language. In this method we have training data and test data which we use to train our machine learning algorithm. The advantage of statistical and probabilistic modelling methods is that there is less training required than in Deep Learning. A disadvantage is that your data needs to be structured to be inputted into a machine learning algorithm. Some use cases are identifying aspects of text or lemmatization.
    c. Deep Learning

i. In Deep Learning we use Neural Networks for the computer to interact with natural language. In this method we use large amounts of data to train Neural Networks. We can also use different types of Neural Networks for more specific tasks such as Recurrent Neural Networks or Convolutional Neural Networks (which have NLP uses, not just image processing uses). One advantage of this method is that it can handle more complex tasks of NLP with potentially higher accuracies. The disadvantage of Deep Learning is that we usually require more data which can be expensive and that it takes longer to train than other methods. Some use cases are text classification, sentiment analysis, and chat bots.

f. Natural Language Processing is a new field with unlimited potential, we have barely scratched the surface of what we can achieve with it. I hope to learn more about NLP to have the ability to approach new problems and have the knowledge needed to start constructing models as solutions. In addition, I hope to gain a better sense of how natural language works and understand what types of intricacies modern NLP considers when interacting with natural languages.