Final Essay:
Natural Language Processing

Eva Richter

31 March 2022

Abstract

Today, the NLP industry is booming due to the increase of new technologies and improved access to digitally available data. *Siri*, *Google Translate* or *Grammarly* are just a few examples of NLP applications that are being increasingly used and have become an integral part of most people's lives. In view of this, the following essay offers an overview of the most important concepts, developments and challenges in the field of NLP.

1 Definitions and Applications of NLP

Natural Language Processing (NLP) deals with the interaction between computers and humans, or rather human language. More precisely, NLP aims to recognise, understand (Natural Language Understanding) and generate (Natural Language Generation) natural language based on computational techniques and thus forms an intersection between computer science, artificial intelligence and linguistics Hirschberg and Manning [2015]. Especially the ambiguity and diversity of natural language pose a challenge for NLP tasks on many levels. Overall, natural language can be studied at the morphological, syntactic, semantic, discourse and pragmatic level, with syntactic (arranging sentence structures in a grammatical way) and semantic analysis (discovering the meaning behind words) being considered the major tasks[Johri et al., 2021].

The number of available NLP applications has increased to such a degree that only some of them can be mentioned here. These include classification tasks, which work particularly well for big data, such as filtering spam e-mails from important e-mails. Another popular application of NLP is sentiment classification, for which information such as reviews or tweets are extracted from websites and social media to provide companies with information about customer preferences [Johri et al., 2021]. A more controversial application of NLP takes place in the healthcare industry (e.g. *Amazon Comprehend Medical*) in which monitoring electronic health records or patient notes based on pattern recognition methods can help to detect and predict diseases [Yse, 2019]. Finally, among the best-known applications are search engines such as *Google* or *Bing* and their online machine translation services, which are now used by virtually everyone. *Siri* and *Alexa* are conversational chatbots that can interact with humans based on text-to-speech synthesis [Hirschberg and Manning, 2015].

Despite the variety of applications, the underlying NLP pipeline consists usually of the same steps (Figure 1): a preprocessing step, in which the data is cleaned, normalised and tokenised, a subsequent feature extraction step, in which the features relevant to the respective NLP task are extracted and a modeling step, in which a machine learning model is created to make predictions on the unseen data [Gyanetsu, 2022].

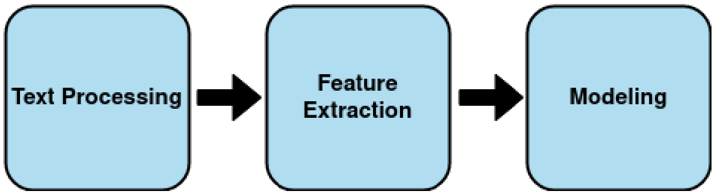

Figure 1: Three Stages of an NLP Pipeline (adapted from [Gyanetsu, 2022]

2 The Evolution of NLP

Even though NLP has found a wide range of applications, its beginnings are inextricably linked to the pursuit of machine translation (MT), one of the first non-numeric applications of computers [Johri et al., 2021], which therefore will be described in particular in this section. After the name *machine translation* first appeared in the 1930s in the context of the patents of Georges Artsrouni and Peter Troyanskii [Johri et al., 2021], the Georgetown Experiment in 1954 was the first milestone in the history of MT and NLP, as it presented the first fully automatic translation of 60 sentences from Russian into English [Pestov, 2018]. In the first decades, the underlying technology of all MT and NLP systems was the so-called rule-based approach, which is based on a linguistic analysis of language and the use of manually created bilingual dictionaries and grammatical rules [Werthmann and Witt, 2014]. Specifically for MT, three different variants have emerged within this rule-based approach, which differ regarding the depth of analysis (Figure 2).

Due to the manual elaboration of fine-grained rules, rule-based systems are transparent and can even achieve notable accuracy given a reasonable size of the set of rules and the dictionary. However, there are major limitations since the elaboration of such rules and dictionaries is very time-consuming and costly due to the complexity and diversity of natural languages[Werthmann and Witt, 2014]. Despite the disillusionment that followed the ALPAC report, which questioned MT's prospects in 1966, there was significant progress in NLP during this period. Examples are the machine *SHRDLU*, developed by Terry Winograd, which could perform various tasks in a *block worlds* environment [Johri et al., 2021], and *ELIZA*, a chatbot program created by Joseph Weizenbaum as a simulation of a Rogerian psychotherapist, which was able to mimic human conversation remarkably well based on simple pattern-matching methods (Jurafsky and Martin [2009]). The latter became the first great achievement of A.I., the idea of which eventually gave new life to NLP and MT in the 1960s [Gyanetsu, 2022].

Since the 1980s and especially from the 1990s onwards, NLP was revolutionised and statistical NLP procedures have gradually replaced the rule-based procedures or have been combined with them in hybrid systems. Thus, NLP became a success in the use of *big data* even before the power of machine learning (ML) or the term *big data* achieved a breakthrough [Hirschberg and Manning, 2015]. This breakthrough is based in particular on the growing availability of linguistic data in digital form (especially through the work of the *Linguistic Data Consortium*) as well as the increase in storage and computing power [Johri et al., 2021]. The methods for statistical NLP come mainly from the ML domain, where the goal is *learning*, i.e. to find patterns, extract information and create probabilistic models. A distinction can be made between supervised and unsupervised ML techniques, the former aiming at predicting information based on labelled training data (e.g. parts of speech (POS)) and the latter at grouping data into clusters [Mishra, 2017]. In the field of SMT, word-based MT systems were largely replaced by phrase-based MT systems (PBMT), for which algorithms learn sequences (so-called *n-grams*) rather than individual words, and can thus provide more accurate translations. Such a PBMT system was introduced in 2006 by *Google Translate* followed by *Yandex*, *Bing* and many other translation service providers [Pestov, 2018].

For some years now, there has been a paradigm shift towards neural approaches, so that almost most NLP tasks such as image classification, text generation or MT are now carried out based on neural networks (NNs) for ML. The dominance of neural approaches results from being able to perform tasks better than is possible with rule-based or statistical approaches within the framework of rules and fixed criteria [Johri et al., 2021]. The underlying technology is so-called *Deep Learning*, a specific form of ML in which NNs independently recognise patterns from growing amounts of training data in various procedures, develop these patterns further and can constantly improve themselves via feedback. The underlying NNs are learning algorithms inspired by the structures of the biological networks in the human brain and consist of the linking of neurons, so-called units, which can lie in several layers on top of each other [Rey and Wender, 2018]. This complexity gives deep learning models a higher level of abstraction than conventional machine learning approaches and gives *Deep* (*Learning*) its meaning. In 2016, Google announced that it was switching to neural machine translation for some languages after developing its own *Google Neural Machine Translation* (GNMT) system. Other languages and translation service providers followed, so that neural MT has now almost completely replaced statistical MT [Pestov, 2018].

NMT models were initially presented as sequence-to-sequence architectures consisting of an encoder and a decoder, each forming a recurrent NN with long short-term memory [Pestov, 2018]. However, in 2017, the NLP landscape was revolutionised once again with the release of the GMT paper *Attention Is All You Need*, a year before *BERT*, a bidirectional encoder representations from transformers was first released. The success of the transformer architectures lies in the self-attention, i.e. the many attention layers through which the transformer passes on the input, so that they are now used far beyond MT tasks [Horev, 2018].

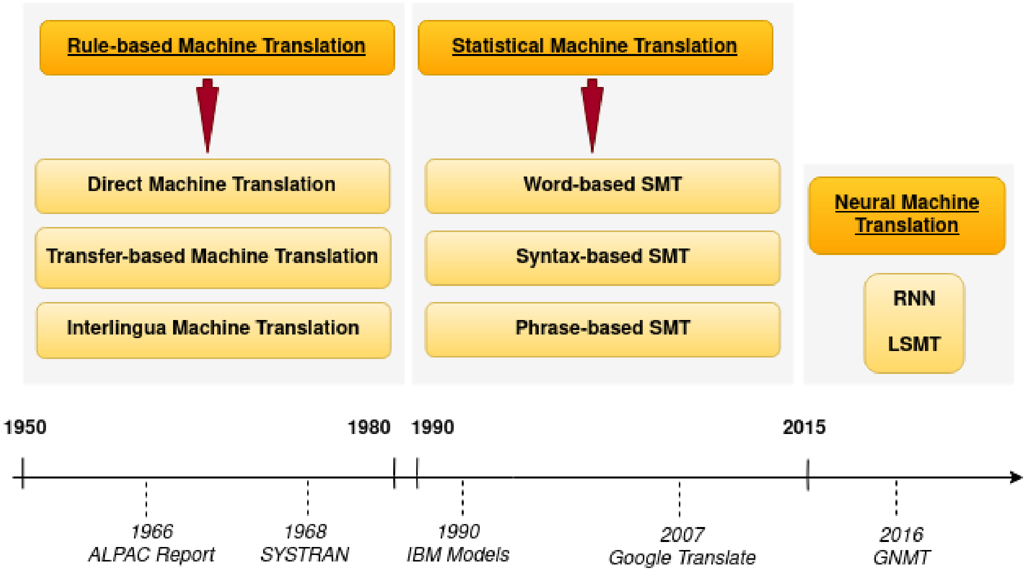
Rule-based Machine Translation
Direct Machine Translation
Transfer-based Machine Translation
Interlingua Machine Translation
Statistical Machine Translation
Word-based SMT
Syntax-based SMT
Phrase-based SMT
Neural Machine Translation
RNN
LSMT
1950
1980
1990
2015
1966
ALPAC Report
1968
SYSTRAN
1990
IBM Models
2007
Google Translate
2016
GNMT

Figure 2: The Evolution of NLP and MT, adapted from [Pestov, 2018]

3 The Importance of Language Data

As further explained in section 4, language data is the essence of statistical and neural ML algorithms. Due to the abundance of available data, current deep learning models can be trained on larger data sets with greater accuracy than ever before. At the same time, however, the main problem with neural (and statistical) approaches is that the quality of the output depends almost only on the size and on how balanced the data is. Too little data or too unspecific data leads to poor results, given the property of these models to only learn based on the data they have seen [Schmalz, 2019].

Since statistical and neural methods only require large amounts of data and do not require the elaborate modeling of linguistic rules, it is often assumed that they do not require linguistic knowledge [Hirschberg and Manning, 2015]. However, there are also assumptions that statistical approaches also need linguistic knowledge in that they require manual annotation of data sets. Furthermore, Johnson [2009] assumes that statistical parsing requires "a greater level of linguistic sensitivity", since the probability of a parse is characterized by its statistical features and a parser designer has to choose which features to include. However, it is precisely this *missing* linguistic knowledge that makes statistical methods far more time and cost effective than rule-based approaches [Hirschberg and Manning, 2015].

The annotations can be for instance POS tags or sentiments. The fine-grainedness of the annotations may depend on the task, where of course a more fine-grained analysis (for instance not only labeling a POS tag as verb, but also specifying the tense) is more time-consuming than a more superficial annotation. To reduce the time and cost associated with supervised approaches, semi-supervised and finally unsupervised approaches were introduced, aiming at finding patterns based on the data itself without knowing what they could actually look like [Johri et al., 2021]. So-called *cluster algorithms* find structures in the data, so that elements that are more similar to each other are grouped into the same cluster, as opposed to less similar elements [Mishra, 2017]. In addition, data is often biased. Unsupervised training approaches can detect such patterns of discrimination hidden in the language. An example is Amazon's automated resume screening for selecting the top job candidates, which gave higher scores to men because women were underrepresented in the training set. The bias towards certain social groups can be further amplified by the design, sampling, and processing of data [Krishnamurthy, 2019].

4 Challenges in NLP

The main challenge that researchers in NLP and MT have struggled with for most of its history is the resolution of ambiguities in language due to its complexity. While this problem seemed almost unsolvable, especially using rule-based approaches, neural approaches seem to have largely solved the problem, since no modeling of rules is required and instead algorithms themselves derive the process of mapping an input to an output [Johri et al., 2021]. However, there are other problems regarding the use of neural approaches such as the lack of transparency due to which NNs are often described as *black box* since it is not possible to trace the decision-making process of NNs, i.e. to find out the importance of the input features based on the output. Another problem is *overfitting*, which happens when the model learns many details during training and cannot generalise well, resulting in poor performance on unseen data [Mangelsdorf, 2019].

As already mentioned in section 3 the dependence on training data is probably the biggest advantage and disadvantage of neural and statistical NLP at the same time. The need for only training data makes machine learning algorithms, as opposed to rule-based approaches, which require costly and time-consuming modelling of linguistic rules, robust and popular [Hirschberg and Manning, 2015]. However, this "simplicity" also poses a number of challenges, especially since the quality of the output depends solely on the appropriateness of the data for a specific task and the quantity of the training data. In most cases, the data is simply too sparse or not balanced enough to capture specific phenomena [Johnson, 2009]. The availability of many NLP resources and systems is strongly language-dependent. While large amounts of data are available for high-resource languages such as German, English or Spanish, there is often little or no training data for low-resource languages such as Punjabi or Swahili available. In the field of MT, in cases where no bilingual corpora are available, translations have to be indirectly derived from other language pairs in *Zero-Shot-Translations*, which, however, considerably reduces the quality of the translations. In this respect, one of the main challenges for the future is the development of resources and tools for these languages [Hirschberg and Manning, 2015]. However, just as important as the amount of training data is the type of training corpus. As with some languages, there is not always adequate training material for different domains, which can have a negative impact on terminology and result in out-of-vocabulary words. Another challenge resulting from the amount of training data is of a more practical nature and concerns the processing power of computers. Despite all the advances, development runtime is still a challenge even using GPUs, since training with millions of data points leads to incredibly long runtimes [Johri et al., 2021].

Although the history of NLP has been marked by numerous setbacks, recent developments give a positive outlook for the future. While sarcasm and idioms still pose difficulties for NLP systems, textual and audio files in particular can now be processed quickly and with almost no problems [Johri et al., 2021]. Nevertheless, the need for improvements and new developments in all areas of NLP is always present.

References

Gyanetsu. What is natural language processing? intro to nlp in machine learning, 2022. URL `https://www.gyansetu.in/what-is-natural-language-processing#`.

J. Hirschberg and C. D. Manning. Advances in natural language processing. *Science*, 41(6):261–266, 2015.

R. Horev. Towards data science: Bert explained: State of the art language model for nlp, 2018. URL `https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270`.

M. Johnson. How the statistical revolution changes (computational) linguistics. *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics*, pages 3–11, 2009.

P. Johri, S.K. Khatri, M. Suvanov S. Al-Taani, A. Sabharwal, and A. Chauhan. Natural language processing: History, evolution, application and future work. *Proceedings of 3rd International Conference on Computing Informatics and Networks*, pages 365–375, 2021.

D. Jurafsky and J.H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Prentice Hall, Upper Saddle River, N.J., 2009.

P. Krishnamurthy. Towards data science: Understanding data bias, 2019. URL `https://towardsdatascience.com/survey-d4f168791e57`.

A. Mangelsdorf. *Künstliche Intelligenz.Technologien, Anwendung, Gesellschaft*, chapter 2. Wittpahl, V., 2019. URL `https://link.springer.com/content/pdf/10.1007%2F978-3-662-58042-4.pdf`.

S. Mishra. Towards data science: Unsupervised learning and data clustering, 2017. URL `https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a`.

I. Pestov. A history of machine translation from the cold war to deep learning, 2018. URL `https://www.freecodecamp.org/news/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5/`.

G. Rey and K. Wender. *Neuronale Netze: Eine Einführung in die Grundlagen, Anwendungen und Datenauswertung.* Hogrefe AG, Bern, 2018. URL `http://www.neuronalesnetz.de/index.html`.

A. Schmalz. *Künstliche Intelligenz.Technologien, Anwendung, Gesellschaft*, chapter 10. Wittpahl, V., 2019. URL https://link.springer.com/content/pdf/10.1007%2F978-3-662-58042-4.pdf.

A. Werthmann and A. Witt. Maschinelle Übersetzung – gegenwart und perspektiven. *Translation and Interpretation in Europe. Contributions to the Annual Conference 2013 of EFNIL in Vilnius*, pages 73–103, 2014.

D. L. Yse. Towards data science: Your guide towards natural language processing (nlp), 2019. URL https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1.