The potential of Language Technology and AI

(Manuscript to the slides)

Invited talk

European Language Resource Coordination (ELRC) Workshop
https://lr-coordination.eu

11 December 2020

Prof. Dr. Christoph Schommer
Department of Computer Science
University of Luxembourg
Email: Christoph.Schommer@uni.lu
Phone: +352-466644-5228

Good morning!

In recent years, the term AI has taken on a new meaning and is now mainly seen in the border area between "Intelligent Systems", "Computer Engineering" and "Data Science". This interdisciplinarity is very challenging per se and also triggers fears, raises questions and increasingly worries society.

For example, the ethical and moral behaviour of robots as well as that of humans, the topic of "autonomy of machines", the need for appropriate education and training, and so on. On the other hand, the current innovations and investments in the field of artificial intelligence are exemplary. In combination with "machine learning" and "human-machine interfaces", a wide variety of fields in the humanities, financial sciences, life sciences - just to name a few - are being promoted.

Of perhaps greatest interest are the fields of natural language processing (NLP) and natural language understanding (NLU). And indeed, both fields are again a very focused application area with great potential.

One not only deals with natural languages such as English, French, German, Luxembourgish, etc. in particular, but also targets their use. For example, conversations in social networks such as Twitter or WhatsApp. Not to be forgotten are the numerous applications that involve a more advanced simulation of human cognitive performance: the recognition of emotions in texts, the automatic summarisation of content, the tracking of topics, machine translation, and many more.

Deep Learning.

The concept of "Deep learning" is again playing an increasingly decisive role (after the "golden 90s") and there is a kind of "resurrection" of various neural architectures such as traditional feedforward networks, but also newer ones such as recurrent networks, convolutional neural networks, Deep Boltzmann Machine, AutoEncoder, GANs and many others.

I would like to briefly introduce some of them here.

Convolutional Neural Networks (or CNNs) are perhaps the most popular choice among neural network architectures today. The name "convolution" is derived from a mathematical operation, namely the convolution of different neuronal layers: for example, the input signal is received in the "convolution" layer and smoothed in another layer. This is connected with the aim of reducing the sensitivity of the filters to noise and other fluctuations. A subsequent layer determines how the signal flows from one layer to another. This is similar to the neurons in our brain. Last but not least, there is a layer that connects all the layers of the network with each neuron of a preceding layer and the neurons of the following layer.

A second architectural model are recurrent neural networks (RNNs). These are very popular in areas where the order, in which information is presented, is crucial. As a result, there are many applications in real-world domains such as speech synthesis or machine translation. RNNs are called "recurrent" mainly because a uniform task is performed for each individual element of a sequence. This output also depends on the previous computations.

Autoencoders apply the principle of "back-propagation" in an unsupervised environment. Some of the popular applications of autoencoders are anomaly detection and "deep fakes", respectively. The core task of autoencoders is to learn the process from input signal to output signal. Once this is done, what is learned can be applied to other images and, for example, spoken language can not only be imitated vocally, but also transferred to other people.

The basic premise of Generative Adversarial Networks (GANs) is the simultaneous training of two neural models. These networks basically compete with each other: one model (the so-called "generator") tries to generate new instances or examples. The other model (the so-called "discriminator") tries to discriminate between training data and the data coming from the "generator".

In addition to the above, some other deep learning models are becoming increasingly popular, such as LSTMs (Long-Short-Term-Memory). LSTMs are a special type of recurrent neural network that contain a special memory cell. This memory cell can store information over longer periods of time. A series of so-called "gates" determine when a certain piece of information enters the memory and when it can be forgotten.

Finally, "Seq2Seq models" are becoming increasingly important, especially for machine translation and the construction of efficient chatbots.

Jared Peterson, Director of Advanced Analytics at SAS, recently mentioned on AnalyticsWeek.com that "NLP is significant because the NLP trend over the last decade has increasingly moved from symbolic-working and also rule-based models towards deep learning", i.e., sub-symbolic. In my opinion, current developments in the field of "natural language processing" increasingly include a use of deep learning techniques, but rule-based (and symbolic) models are still used for text analysis, argument mining, topic modelling, and so on. An important reason is the explainability of the results. This is demonstrated by current products such as Amazon Alexa or Mitsuku ("Kuki"), a chatbot from Pandorabots.

NLP and Poetry Production.

In a recent newspaper article in “The New Yorker” magazine titled "What Happens When Machines Learn to Write", author Dan Rockmore writes: "Programming is an art form of words and punctuation, thoughtfully placed and purposeful, even if not necessarily used to elicit surprise or laughter. Arranged on a page, each programme uses indentations, stanzas and a distinct visual hierarchy to convey meaning."

Ranjit Bhatnagar is one such programmer. But he is also an artist. In 2012, he invented "Pentametron", which scours Twitter data for tweets in iambic pentameter. First, using a pronunciation dictionary created by Carnegie Mellon University, he built a programme to count syllables and recognise the metre. Then, using a separate piece of code, he began identifying rhymes and compiling sonnets. For the first National Novel Generation Month competition in 2013, Bhatnagar submitted "I got an alligator for a pet!", a collection of 504 sonnets created with Pentametron.

Bhatnagar's programming code required that each line be an entire tweet, or essentially a complete thought (or at least what counts as a thought on Twitter). He also "did his best" to adhere to strict rules of metre and rhyme. Thus the piece "Good night! Tomorrow is another day :)" was "machine-written":

I wanna be a little kid again.
I’m feeling kinda empty on the low.
You should unwind a little now and then.
Team Stacie looking like a sleepy hoe.

Back to the Sunshine State. The devil is
a lie. I hate myself a lot sometimes,
I mean, possessive, holy shit, this is
the second time. I’m always catching dimes.

I’m not the only one, I’m pinning this
again. I love a windy sunny day.
Not coming out until tonight. I miss
the happy me. I gotta find a way.

I always fall into the bullshit. Why?
Socks on in bed—the devil is a lie.

For a long time, the Turing Test has been a standard for assessing artificial intelligence. But, in the context of creating poetry - rather than simulating consciousness/cognitive ability - it may not be the most valuable or interesting metric (anymore).

At least that's what Mary Flanagan, professor of digital humanities at Dartmouth College, thinks. She finds the idea that machine-generated poems should be expected to pass the Turing test at some point just "boring" and "uninteresting". She says, "Humans are already good at producing human-sounding sonnets, so why make a computer do the same? She argues that a computer should instead do something new or rather assist the artist".

The project DEEPHOUSE.

At a time when COVID-19 is attracting worldwide attention, the amount and variety of data is increasing dramatically. The result is "data lakes" in which (raw) data appears in different formats and qualities. For example, the data of Johns Hopkins University and also the data of Twitter; or also that of the "Open Research Dataset Challenge" (CORD-19) - a data resource of almost 60000 scientific articles, of which about 75% are full-text articles for Covid19.

DeepHouse is a project funded by the "Fonds National de Recherche" (FNR). We have two main goals in mind: a) we want to consolidate the available text data and time series data in a Covid19 data warehouse, e.g. along multidimensional axes by applying appropriate data integration techniques. And b): we want to build a web-based platform that is extensible and demonstrates successful discovery of time-related sequences or time series. This is achieved, for example, by visualising or tracking themes.

After the first 5 months of project work, we have now pre-processed the text data, consolidated it and visualised it with a user interface. Handling the data is generally very time-consuming because different data sources have different standards. Finding homogeneity by merging heterogeneous data sources takes time - and a lot of decisions. As for the user interface, a first version offers the possibility to set time and place and to list Twitter messages related to Covid19.

The project STRIPS.

In a newspaper article of the Luxemburger Wort of 25 April 2019, the STRIPS project was described as follows:

a) STRIPS is an interdisiplinary project of computer science and linguistics
b) STRIPS's aim is to develop a kind of "toolbox" of semantic search algorithms for Luxembourgish, using texts available in Luxembourgish - more precisely, user comments from online articles of RTL.lu.

Peter Gilles, professor of linguistics, mentions that "the algorithms are programmed in such a way that they should first help to assign the comments in an evaluative way: namely into the categories "positive", "negative" and "neutral". The aim is to filter out moods in this way.

To achieve this, the STRIPS computer model used has been manually "fed" with comments for about 1 year and the Luxembourgish sentences are first labelled. The aim is to obtain a sufficiently large training set and thus create the basis for learning and associated prediction. Among others, students from some Luxembourgish schools, such as the Nic-Biever School in Dudelange, help us with the labelling (or annotation).

NLP is not always “successful”

NLP refers in particular to a series of processing steps that are traditionally found in linguistic realms and, for example, in text processing and its analysis. These include processes such as parsing, tokenisation, part-of-speech tagging, named-entity recognition, and above all disambiguation, i.e., making syntactically ambiguous sentences unambiguous (example: Peter sees Susanne with the telescope) or ambiguous word meanings.

But be careful: even this doesn't always work in our highly technical and AI-savvy world! Because if you enter sentences like "*The old man the boat*" (note: here "*man*" is the verb and "*old*" is a noun) or "*The astronomer marries a star*" (here "*star*" is neither a geometric object nor a space object, but a TV, film or radio star) into *Deepl.com* or *Google.Translate.com*, these systems still translate these sentences incorrectly.

Next Generations.

Finally, let's come back to the topic of "Deep Learning". Recently, a number of new approaches have emerged, of which I would like to mention three very briefly:

a) BERT (= Bidirectional Encoder Representations from Transformers) relies on transformers that pre-train the learning models based on mass data. The result is not only a lower training effort, but also a better understanding of the context and its meaning.

b) Considering speech recognition as a kind of language technology, wav2vec uses self-supervision to learn from unlabelled training data without transcriptions of audio signals. wav2vec enables the development of high-quality speech recognition systems for languages such as Luxembourgish, for which only a small amount of transcribed data is available. The model is trained to predict the correct speech unit for masked parts of the audio while learning how the speech units should sound.

c) GPT-3 was developed by OpenAI and stands for Generative Pre-trained Transformer. GPT-3 is able to generate text using pre-trained algorithms. Strictly speaking, GPT-3 is a "language prediction model" that "pre-trains" based on a huge corpus of texts (currently 570GB of texts, including Wikipedia). In order to learn how language constructs, such as sentences, are built, GPT-3 uses semantic analysis and not only examines words and their meanings, but also gathers an understanding of how word usage differs depending on collocations. GPT-3 is probably the largest artificial neural network ever built.

Thank you very much for your attention!