Notes on NLP

The next cool thing: NVIDIA RIVA + RASA ChatBot (2022):

- https://docs.nvidia.com/deeplearning/riva/user-guide/docs/samples/sample-apps/virtual-assistant-rasa/README.html

Two architectural options because of overlapping services:

- Option 1: Riva ASR + Riva TTS + Riva NLP + Rasa dialog manager (“more RIVA”)
- Option 2: Riva ASR + Riva TTS + Rasa NLU + Rasa dialog manager (“more RASA”)

Abbreviations:

- ASR: Automatic Speech Recognition
- TTS: Text to Speech
- NLU: Natural Language Understanding
- NLG: Natural Language Generation
- NLP: Natural Language Processing

Papers / Websites

- The Annotated Transformer (Harvard): http://nlp.seas.harvard.edu/2018/04/03/attention.html
- Attention Is All You Need (original paper): https://arxiv.org/abs/1706.03762
- Distilling the Knowledge in a Neural Network: https://arxiv.org/abs/1503.02531
- Well-Read Students Learn Better: On the Importance of Pre-training Compact Models: https://arxiv.org/abs/1908.08962
- TensorFlow Hub: https://tfhub.dev/
- Google Research, BERT - Smaller Models on Git: https://github.com/google-research/bert
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding: https://arxiv.org/pdf/1810.04805.pdf

Websites

- Getting meaning from text: self-attention step-by-step video (Romain Futrzynski): https://peltarion.com/blog/data-science/self-attention-video
- The Illustrated Transformer (Jay Alammar): http://jalammar.github.io/illustrated-transformer/
- Paper Dissected: “Attention is All You Need” Explained: https://mlexplained.com/2017/12/29/attention-is-all-you-need-explained/
- Speech and Language Processing, Dan Jurafsky and James H. Martin (3rd ed. draft): https://web.stanford.edu/~jurafsky/slp3/

Videos

- RNN W3L08 - Attention Model (Andrew Ng): https://www.youtube.com/watch?v=FMXUkEbjf9k&feature=youtu.be
- Attention is all you need; Attentional Neural Network Models | Łukasz Kaiser | Masterclass: https://www.youtube.com/watch?v=rBCqOTEfxvg&feature=youtu.be
- Transformer Neural Networks - EXPLAINED! (Attention is all you need) (CodeEmporium): https://www.youtube.com/watch?v=TQQlZhbC5ps&feature=youtu.be
- Attention in Neural Networks (CodeEmporium): https://www.youtube.com/watch?v=W2rWgXJBZhU&t
-  [Transformer] Attention Is All You Need | AISC Foundational (Joseph Palermo (Dessa)): https://www.youtube.com/watch?v=S0KakHcj_rs&feature=youtu.be
- Ivan Bilan: Understanding and Applying Self-Attention for NLP | PyData Berlin 2018: https://www.youtube.com/watch?v=OYygPG4d9H0&feature=youtu.be
-  Transformer (Attention is all you need)(Minsuk Heo): https://www.youtube.com/watch?v=z1xs9jdZnuY&feature=youtu.be
-  Self-attention step-by-step | How to get meaning from text | Peltarion Platform (Romain Futrzynski): https://www.youtube.com/watch?v=-9vVhYEXeyQ&feature=emb_logo

Literature overview on NLP

These tables should give an overview over recent and influential literature in the field of Natural Language Processing from the past few years.

General overview

NLP, transfer learning, language models.

Author	Title	Link to code	Abstract (short)
Vaswani et al. (2017)	Attention Is All You Need	Code used for training and evaluation: https://github.com/tensorflow/tensor2tensor	Introduction of a new simple network architecture, the **Transformer**, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely.

Author	Title	Link to code	Abstract (short)
Kim et al. (2017)	Structured Attention Networks	https://github.com/harvardnlp/struct-attn	In this work, we experiment with incorporating richer structural distributions, encoded using graphical models, within deep networks. We show that these structured **attention** networks are simple extensions of the basic attention procedure, and that they allow for extending attention beyond the standard soft-selection approach, such as attending to partial segmentations or to subtrees.
Radford et al. (2018)	Improving Language Understanding by Generative Pre-Training	https://github.com/openai/finetune-transformer-lm	Natural language understanding comprises a wide range of diverse tasks such as textual entailment, question answering, semantic similarity assessment, and document classification. Although large unlabeled text corpora are abundant, labeled data for learning these specific tasks is scarce, making it challenging for discriminatively trained models to perform adequately. We demonstrate that large gains on these tasks can be realized by generative pre-training of a language model on a diverse corpus of unlabeled text, followed by discriminative fine-tuning on each specific task. ***GPT-1***
Devlin et al. (2018)	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	https://github.com/openai/finetune-transformer-lm	Introduction of a new language representation model called **BERT**, which stands for Bidirectional Encoder Representations from Transformers. BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers.

Author	Title	Link to code	Abstract (short)
Radford et al. (2019)	Language Models are Unsupervised Multitask Learners	https://github.com/openai/gpt-2	Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on taskspecific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText.(...) Our largest model, **GPT-2**, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits Web-Text.
Ruder (2019)	Neural Transfer Learning for Natural Language Processing	https://github.com/sebastianruder	Multiple novel methods for different **transfer learning** scenarios were presented and evaluated across a diversity of settings where they outperformed single-task learning as well as competing transfer learning methods.
Kovaleva et al. (2019)	Revealing the Dark Secrets of BERT	-	BERT-based architectures currently give state-of-the-art performance on many NLP tasks, but little is known about the exact mechanisms that contribute to its success. In the current work, we focus on the **interpretation of self-attention**, which is one of the fundamental underlying components of BERT.
Rogers et al. (2020)	A Primer in BERTology: What We Know About How BERT Works	-	This paper is the first survey of over 150 studies of the popular BERT model. We review the current state of knowledge about **how BERT works**, what kind of information it learns and how it is represented, common modifications to its training objectives and architecture, the overparameterization issue and approaches to compression.

Author	Title	Link to code	Abstract (short)
Brown et al. (2020)	Language Models are Few-Shot Learners	https://github.com/openai/gpt-3	Demonstration that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train **GPT-3**, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting.
Schick and Schütze (2020)	It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners	https://github.com/timoschick/pet	We show that **performance similar to GPT-3** can be obtained with language models that are much "greener" in that their parameter count is several orders of magnitude smaller. This is achieved by converting textual inputs into cloze questions that contain a task description, combined with gradient-based optimization; exploiting unlabeled data gives further improvements.
Jaegle et al. (2021)	Perceiver IO: A General Architecture for Structured Inputs & Outputs	https://github.com/deepmind/deepmind-research/tree/master/perceiver	The recently-proposed Perceiver model obtains good results on several domains (images, audio, multimodal, point clouds) while scaling linearly in compute and memory with the input size. While the Perceiver supports many kinds of inputs, it can only produce very simple outputs such as class scores. **Perceiver IO** overcomes this limitation without sacrificing the original's appealing properties by learning to flexibly query the model's latent space to produce outputs of arbitrary size and semantics.

Author	Title	Link to code	Abstract (short)
Callaghan et al. (2021)	Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies	-	Increasing evidence suggests that climate change impacts are already observed around the world. Global environmental assessments face challenges to appraise the growing literature. Here the language model **BERT was used** to identify and classify studies on observed climate impacts, producing a comprehensive machine-learning-assisted evidence map.

Specific overview

Speech recognition

Author	Title	Link to code	Abstract (short)
Amodei et al. (2015)	Deep Speech 2: End-to-End Speech Recognition in English and Mandarin	-	We show that an end-to-end deep learning approach can be used to recognize either English or Mandarin Chinese speech—two vastly different languages. Because it replaces entire pipelines of hand-engineered components with neural networks, end-to-end learning allows us to handle a diverse variety of speech including noisy environments, accents and different languages.
Agarwal and Zesch (2019)	German End-to-end Speech Recognition based on DeepSpeech	https://github.com/AASHISHAG/deepspeech-german	Description of the process of training German models based on the Mozilla DeepSpeech architecture using publicly available data.

Information Extraction

Named Entity Recognition

Author	Title	Link to code	Abstract (short)
Anthofer (2017)	A Neural Network for Open Information Extraction from German Text	https://github.com/danielanthofer/nnoiegt	Systems that extract information from natural language texts usually need to consider language-dependent aspects like vocabulary and grammar. Compared to the develop ment of individual systems for different languages, development of multilingual information extraction (IE) systems has the potential to reduce cost and effort. One path towards IE from different languages is to port an IE system from one language to another. PropsDE is an open IE (OIE) system that has been ported from the English system PropS to the German language.
Riedl and Padó (2018)	A Named Entity Recognition Shootout for German	https://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/german-ner/	We ask how to practically build a model for German named entity recognition (NER) that performs at the state of the art for both contemporary and historical texts, i.e., a big-data and a small-data scenario.
Torge et al. (2021)	Transfer Learning for Domain-Specific Named Entity Recognition in German	-	Investigation of different transfer learning approaches to recognize unknown domain-specific entities, including the influence on varying training data size.

Links to Websites and Videos

Author	Title	Link	Information
Manning et al. (2008)	Introduction to Information Retrieval	https://nlp.stanford.edu/IR-book/html/htmledition/irbook.html	Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

Author	Title	Link	Information
Olah and Carter (2016)	Attention and Augmented Recurrent Neural Networks	https://distill.pub/2016/augmented-rnns/	Recurrent neural networks are one of the staples of deep learning, allowing neural networks to work with sequences of data like text, audio and video. They can be used to boil a sequence down into a high-level understanding, to annotate sequences, and even to generate new sequences from scratch.
Alexander Rush	The Annotated Transformer	https://nlp.seas.harvard.edu/2018/04/03/attention.html	In this post Alexander Rush presents an “annotated” version of the paper in the form of a line-by-line implementation. He has reordered and deleted some sections from the original paper and added comments throughout. This document itself is a working notebook, and should be a completely usable implementation. In total there are 400 lines of library code which can process 27,000 tokens per second on 4 GPUs.
Ruder (2018)	NLP’s ImageNet moment has arrived	https://thegradient.pub/nlp-imagenet/	Big changes are underway in the world of Natural Language Processing (NLP). The long reign of word vectors as NLP’s core representation technique has seen an exciting new line of challengers emerge: ELMo, ULMFiT, and the OpenAI transformer. These works made headlines by demonstrating that pretrained language models can be used to achieve state-of-the-art results on a wide range of NLP tasks.
Garbade (2018)	A Simple Introduction to Natural Language Processing	https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32	This post gives a simple introduction to Natural Language Processing.
Jay Alammar	Visualizing A Neural Machine Translation Model (Mechanics of Seq2seq Models With Attention)	https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/	Sequence-to-sequence models are deep learning models that have achieved a lot of success in tasks like machine translation, text summarization, and image captioning.
Jay Alammar	The Illustrated Transformer	http://jalammar.github.io/illustrated-transformer/	In this post, we will look at The Transformer – a model that uses attention to boost the speed with which these models can be trained. The Transformers outperforms the Google Neural Machine Translation model in specific tasks.
Jay Alammar	The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning)	https://jalammar.github.io/illustrated-bert/	This post gives an introduction and overview of the BERT model and Transfer Learning.

Author	Title	Link	Information
Jay Alammar	A Visual Guide to Using BERT for the First Time	https://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/	This post is a simple tutorial for how to use a variant of BERT to classify sentences. This is an example that is basic enough as a first intro, yet advanced enough to showcase some of the key concepts involved.
Schreiner (2018)	Deepmind: Mit Perceiver IO auf dem Weg zur Multi-KI	https://mixed.de/deepmind-mit-perceiver-io-auf-dem-weg-zur-multi-ki/	Deepmind stellt Perceiver IO vor, ein echtes Multitalent unter den neuronalen Netzen. Es könnte die weit verbreitete Transformer-Architektur ablösen.
Sanagapati (2020)	Knowledge Graph & NLP Tutorial - (BERT, spaCy, NLTK)	https://www.kaggle.com/pavansanagapati/knowledge-graph-nlp-tutorial-bert-spacy-nltk	This post is an introduction to NLP and Knowledge Graphs and also a tutorial how to use BERT, spaCy and NLTK.
Sebastian Raschka	Transformers from the Ground Up - Sebastian Raschka at PyData Jeddah	https://www.youtube.com/watch?v=OGGhpLBeCuI	VIDEO - This talk will explain how transformers work. Then, some popular transformers like GPT and BERT will be examined and their differences will be outlined. Equipped with this understanding, it will be explained how fine-tuning of a BERT model for sentiment classification in Python works.
Komarraju (2021)	DeepMind’s Perceiver IO is Now an Open-Source Deep Learning Model	https://www.analyticsinsight.net/deepminds-perceiver-io-is-now-an-open-source-deep-learning-model/	To leverage developments in deep learning, DeepMind has open-sourced Perceiver IO. It’s a general-purpose deep learning model architecture for various types of inputs and outputs. As described on DeepMind’s blog, Perceiver IO can serve as a replacement for transformers, using attention to map inputs into a latent representation space. Eliminating the drawbacks of a transformer, Perceiver IO facilitates longer input sequences without incurring quadratic compute and memory loss.
Bastian (2021)	KI-Start-up Cohere will Sprach-KI zum Massenmarkt machen	https://mixed.de/sprach-ki-millionen-invest-fuer-gpt-3-konkurrenz/	Das US-Start-up Cohere widmet sich der Entwicklung fortschrittlicher Sprach-KI und geht in den Wettbewerb mit etablierten großen Playern wie OpenAI. Es startet mit reichlich Rückenwind.
Akash (2021)	“Ok, Google!”— Speech to Text in Python with Deep Learning in 2 minutes	https://www.analyticsvidhya.com/blog/2021/09/ok-google-speech-to-text-in-python-with-deep-learning-in-2-minutes/	This blog post is a tutorial to build a very simple speech recognition system that takes our voice as input and produces the corresponding text by hearing the input.
Hugging Face	Transformers	https://huggingface.co/transformers/	This page gives an overview about the transformer architecture and the models provided by Hugging Face.

Author	Title	Link	Information
KDnuggets	Text Preprocessing Methods for Deep Learning	https://www.kdnuggets.com/2021/09/text-preprocessing-methods-deep-learning.html	This post focuses on the pre-processing pipeline for NLP tasks like classification.
Wiggers (2021)	Microsoft and Nvidia team up to train one of the world's largest language models	https://venturebeat.com/2021/10/11/microsoft-and-nvidia-team-up-to-train-one-of-the-worlds-largest-language-models/	Microsoft and Nvidia announced that they trained what they claim is the largest and most capable AI-powered language model to date: Megatron-Turing Natural Language Generation (MT-NLP). The successor to the companies' Turing NLG 17B and Megatron-LM models, MT-NLP contains 530 billion parameters and achieves "unmatched" accuracy in a broad set of natural language tasks, Microsoft and Nvidia say — including reading comprehension, commonsense reasoning, and natural language inferences.
Tang (2021)	DeepSpeech for Dummies - A Tutorial and Overview	https://www.assemblyai.com/blog/deepspeech-for-dummies-a-tutorial-and-overview-part-1/	This post shows basic examples of how to use DeepSpeech for asynchronous and real time transcription.
Dickson (2021)	What are graph neural networks (GNN)?	https://bdtechtalks.com/2021/10/11/what-is-graph-neural-network/	Basically, anything that is composed of linked entities can be represented as a graph. Graphs are excellent tools to visualize relations between people, objects, and concepts. Beyond visualizing information, however, graphs can also be good sources of data to train machine learning models for complicated tasks. This article gives an overview of how graph neural networks (GNN) can be used to extract important information from graphs and make useful predictions.