DRAFT – 26 Feb 2023

Eryk Wdowiak
economics@doviak.net
www.doviak.net

Natural Language Processing

This course in Natural Language Processing will compare the performance of RNNs, Transformers, BERT and GPT to previous approaches. And it will pay particular attention to how those performance gains were achieved. Did the researchers develop a better model? Or did they train a larger model?

For example, the fluency and translation quality of neural translation models far surpasses that of phrase-based statistical models. And in low-resource cases too.

But what's important is how those performance gains were achieved. Instead of translating words or phrases, the neural approach attempts to understand context. Neural models translate better than phrase-based models because they attempt to create a sentence in the target language with the same meaning as the source language sentence.

In that spirit, this course will explore neural approaches to natural language processing. Comparing them, it will ask how we can develop models that better understand our language.

By training small comparably-sized models, we can compare approaches. Holding model size constant, we'll ask which training or fine-tuning technique performs best on a given task. Identifying the techniques that work well at small scale, we'll find techniques that work exceptionally well at large scale.

course materials

Since its introduction in 2017, the Transformer has shown the field of Natural Language Processing that "Attention is All You Need." So we'll pay attention to the Transformer and the models that it has inspired, following the most recent developments in the field, reading the original research.

I have assembled that research into the course readings listed below and at: *doviak.net/courses/nlp*, where you will find links to the readings, notes, announcements and more.

course requirements

An exam broadens your knowledge. A course project deepens your knowledge. Accordingly, this course will require an exam and a course project. And it will require you to help your classmates.

Regular and constructive class participation will be 15 percent of your final grade. The final exam will give you an opportunity to study all of the course readings. It will be worth 35 percent of your final grade.

And the course project (described below) will give you an opportunity to study and compare NLP models and methods. It will be worth 50 percent of your final grade. Please give me an opportunity to help you. Please submit a project proposal by mid-semester.

course project

In a statistics course, students learn how to test a null hypothesis. Those tests assume probability distributions that do not exist in machine learning. Nonetheless, we should still formulate null hypotheses because we can test them informally.

By focusing on comparability, we can test the null hypothesis that an innovative technique does not perform better than its predecessor. Holding model size constant and comparing techniques, we'll discover the most effective ways to train or fine-tune a model.

So for the course project, please identify an innovative technique and compare its performance to an alternative. For a good example, please see the comparison of masked-language modeling to the traditional left-to-right modeling by Devlin et al (2019).

Then please describe your comparison in a formal paper, which reviews and cites previous research and which reports the results of your comparisons.

course readings

lec 00 – context and background

- Bender and Gebru et al (2021). "On the Dangers of Stochastic Parrots"
- Koehn and Knowles (2017). "Six Challenges for Neural Machine Translation"
- Sennrich and Zhang (2019). "Revisiting Low-Resource NMT"

lec 01 – tools for our "NLP kitchen"

- Sennrich and Haddow (2016). "Linguistic Input Features Improve NMT"
- Oncevay et al (2022). "Revisiting Syllables"
- Wdowiak (2022). "A Recipe for Low-Resource NMT"

lec 02 – word embeddings

- Mikolov et al (2013). "Efficient Estimation of Word Representations in Vector Space"
- Mikolov et al (2013). "Distributed Representations of Words and Phrases"
- Pennington et al (2014). "GloVe: Global Vectors for Word Representation"

lec 03 – subword segmentation

- Sennrich, Haddow and Birch (2015). "NMT of Rare Words with Subword Units"
- Kudo and Richardson (2018). "SentencePiece"

lec 04 – recurrent neural networks

- Bahdanau, Cho and Bengio (2014). "NMT by Jointly Learning to Align and Translate"
- Wu et al (2016). "Google's NMT System"
- Neubig (2017). "NMT and Sequence-to-Sequence Models: a Tutorial"

lec 05 – the Transformer

- Vaswani et al (2017). "Attention is All You Need"
- Harvard NLP (2018, 2022). "The Annotated Transformer"

lec 06 – multilingual translation

- Johnson et al (2016). "Google's Multilingual NMT: Enabling Zero-Shot Translation"
- Fan et al (2020). "Beyond English-Centric Multilingual Machine Translation"
- Arivazhagan et al (2019). "Massively Multilingual NMT in the Wild"
- Zhang et al (2020). "Improving Massively Multilingual NMT and Zero-Shot Translation"
- Kudugunta et al (2019). "Investigating Multilingual NMT Representations at Scale"
- Sennrich, Haddow and Birch (2015). "Improving NMT Models with Monolingual Data"

lec 07 – GPT models

- Radford et al (2018). "Improving Language Understanding by Generative Pre-Training"
- Radford et al (2019). "Language Models are Unsupervised Multitask Learners"
- Brown et al (2020). "Language Models are Few-Shot Learners"

lec 08 – BERT models

- Devlin et al (2019). "BERT: Pre-Training of Deep Bidirectional Transformers"
- Pires et al (2019). "How Multilingual is Multilingual BERT?"
- Lewis et al (2019). "BART: Denoising Sequence-to-Sequence Pre-Training"
- Sanh et al (2020). "DistilBERT: a Distilled Version of BERT"

lec 09 – "Stochastic Parrots"

- Bender and Gebru et al (2021). "On the Dangers of Stochastic Parrots"