DEIC 2023 conference

Hotel Comwell
Kolding, Denmark
November 7th, 2023

Research area:
Natural Language Processing

Manex Agirrezabal
Associate Professor
University of Copenhagen

Outline

- Me and myself
- Artificial Intelligence
- Natural Language Processing
- Corpora size and model size
- Model size and training time

Outline

- **Me and myself**
- Artificial Intelligence
- Natural Language Processing
- Corpora size and model size
- Model size and training time

About myself

- Computer Engineering at the University of the Basque Country (UPV/EHU) (2006-2011)
- M. Sc. and PhD in Analysis and Processing of Language at the same place (2011-2012, 2013-2017)
- PostDoc at the University of Copenhagen (2017-2019)
- Assistant Professor (2019-2022) at “
- Associate Professor (2023-) at “

About myself

- Natural Language Processing
- Computational creativity
- Multimodal communication
- Computational morphology and phonology
- Finite-State methods
- Poetry

About myself

I teach in a Master Program called IT & Cognition (M. of Sc.)

- Language Processing 1 (together with a colleague)
- Language Processing 2 (together with a colleague)

I taught in previous years:

- Scientific Programming

Outline

- Me and myself
- **Artificial Intelligence**
- Natural Language Processing
- Corpora size and model size
- Model size and training time

Artificial Intelligence

- What is it? (From Wikipedia & Oxford English Dictionary)
  - The theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages

- Goals (Wikipedia):
  - Reasoning/Problem solving
  - Knowledge representation
  - Learning (Supervised learning, Unsupervised learning, Grammatical Inference)
  - Natural Language Processing
  - Perception
  - Social Intelligence / Affective computing
  - General Intelligence

Artificial Intelligence

- Artificial Intelligence has had periods of interest and reluctance
- 1954: Optimism, Chomsky's grammars, expected breakthroughs
  - But, word sense disambiguation was underestimated
- Example from that time (backtranslation):
  - (ENGLISH) the spirit is willing but the flesh is weak
  - (RUSS-ENGLISH) the vodka is good but the meat is rotten

Outline

- Me and myself
- Artificial Intelligence
- **Natural Language Processing**
- Corpora size and model size
- Model size and training time

Natural Language Processing

- What is it?
  - Natural language processing (NLP) is a subfield of linguistics, computer science and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

- Goals:
  - Make computers understand language (spoken, written and further)
  - Make computers produce language (spoken, written, and further)

Natural Language Processing

Brief history of NLP (Wikipedia):

- 1950-1990: Symbolic NLP
  - 1954: MT experiments (English-Russian)
  - 1960s: Chatbots, such as, ELIZA (test eliza)
  - 1970s: Beginning of ontologies, representing knowledge
  - 1980s: HPSG, Word Sense Disambiguation, Rhetorical Structure Theory
- 1990-2010: Statistical NLP
  - Statistical Machine Translation
  - Word-based language models were crucial
  - IBM Watson[1]
  - https://www.youtube.com/watch?v=WFR3lOm_xhE

https://en.wikipedia.org/wiki/IBM_Watson

Natural Language Processing

- Neural Language Processing
  - 2013: Irruption of Word embeddings (word2vec)
  - 2014: Neural Networks for Machine Translation
  - 2015: Attention Mechanisms for Machine Translation
  - 2017: Attention is the only thing we need
  - 2018-: BERT/ELMO/GPT1-3

- Current paradigm:
  - Train a massive language model
  - Finetune on a specific task

Outline

- Me and myself
- Artificial Intelligence
- Natural Language Processing
- **Corpora size and model size**
- Model size and training time

Size matters

- Corpora size (amount of data to process)
- Model size (ability of your model to save information)
  - Larger model does not always imply better results
  - But you do need a certain size to get to some results

Corpora size

- CoNLL-2003
  - Named Entity Recognition dataset (training data)
  - 15,000 sentences
- GLUE (2018)
  - General Language Understanding Evaluation
  - 9 different tasks
  - ~800,000 sentences

Sang, Erik F., and Fien De Meulder. "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition." *arXiv preprint cs/0306050* (2003).

Wang, Alex, et al. "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding." *International Conference on Learning Representations*. 2018.

Model size

- Sentiment analysis
  - Positive vs. Negative
  - Logistic Regression
  - 50,000 words

50,000 parameters

Model size

- Sentiment analysis
  - Positive vs. Negative
  - ~~Logistic Regression~~ Multilayer Perceptron (3 layers, 1,000 cells in each layer)
  - 50,000 words

52,001,000 parameters

Model size

- Sentiment analysis
  - Positive vs. Negative
  - ~~Logistic Regression~~ ~~Multilayer Perceptron~~

  Recurrent Neural Network (3 layers, 1,000 cells in each layer)
  - 50,000 words

55,001,000 parameters (but the model is recurrent)

Model size

- Sentiment analysis
  - Positive vs. Negative
  - Logistic Regression
  - 50,000 words

50,000 parameters

Model size

- ~~Sentiment analysis~~ Authorship attribution
  - ~~Positive vs. Negative~~ 50 authors
  - Logistic Regression
  - 50,000 words

2,500,000 parameters

Model size

- Specific model architecture
- Vocabulary size
- Number of classes

Outline

- Me and myself
- Artificial Intelligence
- Natural Language Processing
- Corpora size and model size
- **Model size and training time**

Model size and training time

- The larger the model, the longer it takes to train
- By default, in a normal scenario:
  - Double size, double time to train
- Unless you have clever techniques

Scenario no. 1

- Imagine you have a dataset and a model archt.
- You have one computer
- It takes 10 days to process

Scenario no. 2

- You have two computers.
- You process half of the data in one computer, half of the data in another computer.
- Let's suppose it takes one day to combine their output.
- In the end, it takes 6 days to process the whole data.

Scaling it up…

- Now let's scale this up to more computing resources and more complex problem scenarios.
- What do we need for that?
  - HPCs and GPUs
  - Clever parallelization

In conclusion

- Models are getting more complex and larger
- We need HPCs and GPUs to keep up
- We have used many of DEIC's resources (thanks!)
- We want to keep using them (thanks in advance!)

Thank you very much for your attention

Questions?

DEIC 2023 conference

Hotel Comwell
Kolding, Denmark
November 7th, 2023

Research area:
Natural Language Processing

Manex Agirrezabal
Associate Professor
University of Copenhagen