Natural Language Processing

L.Jeganathan

School of Computer Science and Engineering

September 20, 2022

VIT
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

OUTLINE

Outline

What is NLP?

To enable computers to understand the human language

- Analysing and Processing natural language data
- Tasks related to human perception

Fourth Industrial Revolution!!

NLP has become part of the so called Fourth Industrial Revolution (AI automation)

Rapid advances in NLP due to

- Advances in Algorithms for training large neural networks
- Availability of vast amounts of data through internet
- Availability of massive parallel computing capabilities through Graphical Processing Units (GPUs)
- A new paradigm : Transfer Learning

Basic Tasks of NLP

Understanding the word/sentence

To learn the meaning of a word/sentence in a language

Part-of-speech (POS) tagging

Tagging a word in text with its part of speech; potential tags include verb, adjective, and noun.

Named entity recognition (NER)

Detecting entities in unstructured text, such as PERSON, ORGANIZATION, and LOCATION.
POS tagging is part of NER

Sentence/document classification

Tagging sentences or documents with predefined categories, such as sentiments : positive, negative, entertainment, science, history, or some other predefined set of categories.

Sentiment analysis

Assigning to a sentence or document the sentiment expressed in it, for example, positive,negative.
SA is a special case of sentence/document classification.

Summarization

Summarizing the content of a collection of sentences or documents, usually in a few sentences or keywords.

Machine translation

Translating sentences/documents from one language into another language or a collection of languages.

Question answering

Determining an appropriate answer to a question posed by a human;
Question: What is the capital of India? Answer: New Delhi

Chatterbot/chatbot

Carrying out a conversation with a human convincingly, potentially aiming to accomplish some goal, such as maximizing the length of the conversation or extracting some specific information from the human.
chatbot can be formulated as a question-answering system.

A TYPICAL NEURAL NETWORK MODELS FOR NLP

- Huge data set
- Design a deep neural network for feature engineering
- Feed-forward network -supervised/unsupervised learning models

L

earn by training deep neural networks on billions of words/images

Google trains the model with Wikipedia-English (2.5 billion words)

To train an NLP model, we require

For learning from scratch, we require

- Huge amount of data
- Huge computing facility

D

oes this mean that we are locked out of being able to achieve state-of-the art results on the NLP tasks?

Transfer Learning was born!!!

Transfer Learning

We typically do not learn from scratch for any given problem. Rather, we learn from the prior knowledge gained by us related to the given problem.

L

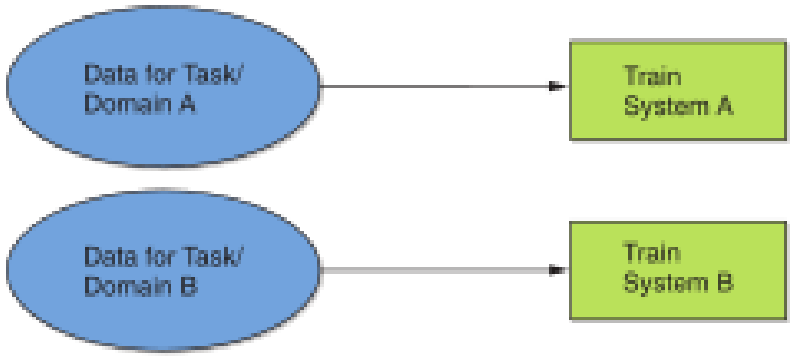
earning to play a musical instrument will be easier if we know how to play another instrument

Transfer learning

- Enables us to adapt or transfer knowledge acquired from one set of tasks/domain to a different set of tasks/domain
- TL: a model trained with massive resources (data, computing, time, cost)can be fine-tuned and re-used in a new settings bu us at a fraction of original resource

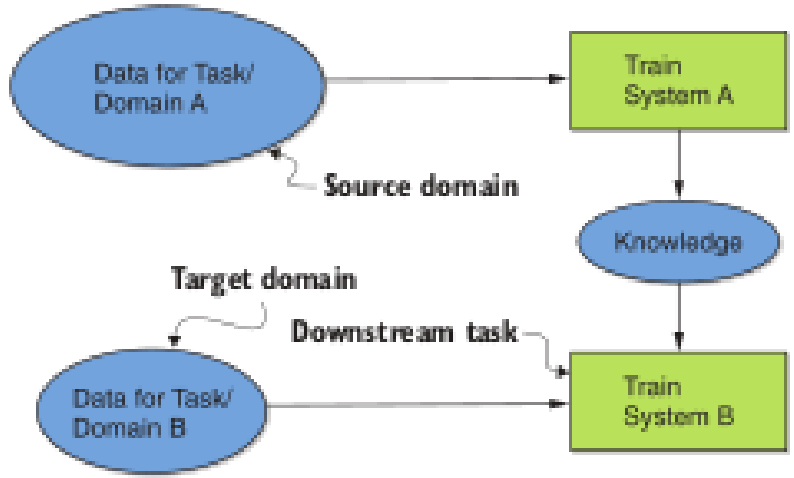
LEARNING FROM SCRATCH VS TRANSFER LEARNING

When training systems in parallel, more data is required for each task/domain.



Transfer learning paradigm: Knowledge is shared between different tasks/domains.

When information is shared between different tasks/domains, it is possible to reduce training data and computing requirements for future tasks (as indicated here by the smaller "Data for Task B").



Outline

Born around 1950's

Georgetown-IBM Experiment of 1954, in which a set of approximately 60 Russian sentences was translated into English

In 1960's

- ELIZA simulated a psychotherapist (An MIT venture)
- The vector space model for information representation was developed, where words came to be represented by vectors of real numbers, which were amenable to computation

In 1970's

Development of a number of chatterbot/chatbot concepts based on sets of handcrafted rules for processing the input

1980-1990

- Advent of the application of systematic machine learning methodologies to NLP, where rules were discovered by the computer versus being crafted by humans.
- Witnessed the application of singular value decomposition (SVD) to the vector space model (leading to latent semantic analysis) unsupervised technique for determining the relationship between words in language

2010 - Distributed semantics

- Neural network (deep learning) based NLP .
- Deep learning NLP achieved state-of-art results for tasks such as machine translation and text classification.
- Development of the word2vec model and its variants sent2vec doc2vec,
- These neural-network-based techniques vectorize words, sentences, and documents (respectively) in a way that ensures the distance between vectors in the generated vector space is representative of the difference in meaning between the corresponding entities (words, sentences, and documents)

- The metric used to train these neural-network-based models was derived from the field of linguistics (distributional semantics), and did not require labeled data
- Meaning of a word is based on the words that surround it.
- Embedding the words in avector space.
- Analysis (NLP) is just the application of Statistical/machine learning algorithms (such as classification/clustering)
- Embedding is unsupervised and ML models are supervised

2014 - Sequence-Sequence Model

- These models learn to associate an input sequence, such as a source sentence in one language, with an output sequence.
- The input sequence (source sequence) is converted into a Context vector through an Encoder.
- Decoder converts the context vector into a target sequence.
- Both the encoder and decoder are Recurrent Neural Networks.

- Various layers of Sequence-sequence model automated all the intermediate steps of machine translation : such as POS tagging, dependency parsing, and language modeling.
- Encoder-Decoders are able to encode order information (which word occurs first and which word occurs later)
- Weakness: Inability to deal with long-range dependencies. Inability to perform parallelization.

2017

Transformer model

- Vaswani et al. published their seminal paper, Attention Is All You Need.
- Attention mechanism : significantly improved the performance of machine translation sequence-to-sequence models by allowing the model to focus on the parts of the input sequence that were most relevant for the output
- Attention: replaced recurrence

Word vectorization Vs Character Vectorization

- Till 2015, words were the atomic unit for vectorization.
- Vectorization is not possible with misspelled words.
- Rise of social media changed the definition of Natural Language (due to emoticons)
- Remedy : Treat the language at the character level - character vectorization

State of the art - Transfer Learning

- In 2018, practical and scalable methods : Transformer-based Transfer learning were developed to accomplish the hardest perceptual problems in NLP
- A library of pretrained models became available for a large subset of NLP data, together with well-defined techniques for fine-tuning them to particular tasks at hand with labeled datasets of a size significantly smaller than what would be needed otherwise.

- Semantic Inference for the Modeling of Ontologies (SIMOn) employed characterlevel convolutional neural networks (CNNs) combined with bidirectional LSTMs for structural semantic text classification.
- Embeddings from Language Models? (ELMo)- an attempt to develop contextualized embeddings of words using bidirectional LSTMs.
- The Universal Language Model Fine-Tuning (ULMFiT) method - to fine-tune any neural-network-based language model for any particular task and was initially demonstrated in the context of text classification

- OpenAI Generative Pretrained Transformer (GPT) modified the encoder-decoder architecture of the transformer to achieve a fine-tunable language model for NLP- It discarded the encoders and retained the decoders and their self-attention sublayers.
- Bidirectional Encoder Representations from Transformers (BERT) did the opposite, modifying the transformer architecture by preserving the encoders and discarding the decoders and also relying on masking of words, which would then need to be predicted accurately as the training metric.

In all of these language-model-based methods : ELMo, ULMFiT, GPT, and BERT, it was shown that generated embeddings could be fine-tuned for specific downstream NLP tasks with relatively few labeled data points.

OUTLINE

Most Challenging Task: to find the meaning of the word

What do you mean by : meaning of a word

- Idea that is represented by a word/phrase
- Idea that a person wants to express by using that word.

A common-sense thought in Linguistics: Denotational Semantics

$$Signifier(symbols) \iff Signified(idea)$$

- Earlier approach : wordNet (Thesaurus) with synonyms/hypernyms
- Issues: missing new mwaning, subjective, issues with words which have more than one meaning,
- Represent words as one-hot vectors
- Hotel=[1 0 0 0], Motel =[0 1 0 0].
- vectors are orthogonal- no similarity

Instead, can we ‘Learn to encode ’similarity’ in the vectors themselves.

Distributional Semantics

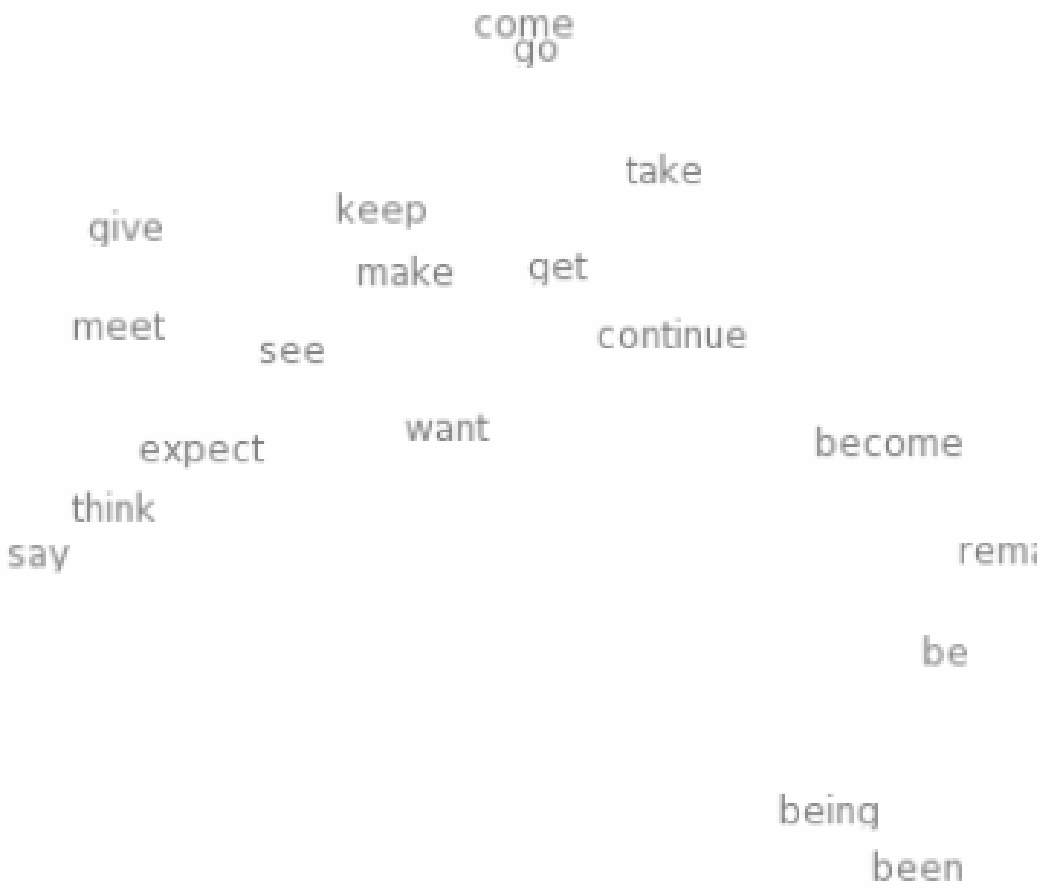
> You shall know a word by the company that it keeps- J. R. Firth

- A word's meaning is given by the words that frequently appear closeby.
- When a word appears in a text, its context is the set of words that appear closeby (fixed window)
- Use many contexts of w to build up a represnetation of w

Word vectors/Word embeddings/Word Representation

- Build a dense vectors for each word chosen so that it is similar to the vectors of words that appear in a similar context,
- Bank = $\begin{bmatrix} 0.286 \\ 0.782 \\ \dots \\ \dots \\ 0.271 \end{bmatrix}$

$$pect = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \\ 0.487 \end{pmatrix}$$



WORD2VEC- A LEARNING FRAMEWORK

- Large carpus of text
- Every word is represented by the ‘learned’ vectors
- Go through each position t in the text, which has the center word ‘c’ and the context words ‘o’
- Use the ‘similarity of vectors’ for ‘c’ and ‘o’ to calculate the probability of c given o (viceversa)
- Keep adjusting the word vectors to maximize the probability.

Outline

Outline

- AI models write Scientific Paper : https://www.scientificamerican.com/article/we-asked-gpt-3-to-write-an-academic-paper-about-itself-mdash-then-we-tried-to-get-it-published/
- Your AI pair programmer- Trained on billions of lines of code, GitHub Copilot turns natural language prompts into coding suggestions across dozens of languages
- Is LaMDA (A language model-Google) Sentient? - LaMDA is a highly advanced AI chat platform analyzing trillions of words from the internet, so it's skilled at sounding like a real person.

- OpenAI- GPT-3 language generator is a language model. With GPT-3, you can input the topic and instructions to write in the style of a particular author, and it will generate a short story or essay, for instance.
- GitHub Copilot, powered by OpenAI-codex is an AI pair programmer that helps you write code faster and with less work. It draws context from comments and code to suggest individual lines and whole functions

REFERENCES

- Paul Azunre, Transfer Learning for NLP
- Denis Rothman, Transformers for NLP
- https://web.stanford.edu/class/cs224n/