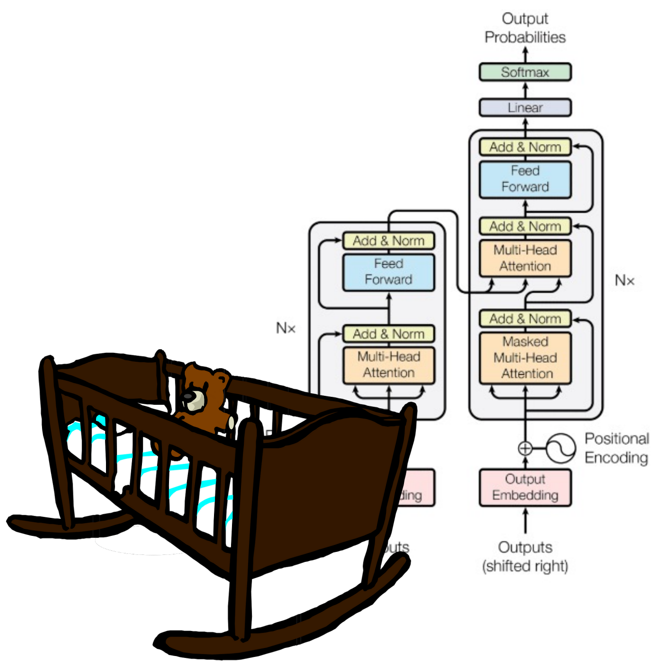
CS 505: Introduction to Natural Language Processing

Wayne Snyder
Boston University

Lecture 20 – Presentation of LX 360/660 with Najoung Kim; The Transformer Family; BERT



Lecture Plan

- Transfer Learning Concluded: LoRa and alternatives
- The Transformer Family:
  - Encoder-based models
  - Decoder-based models
  - Encoder+Decoder-based models
- BERT in detail

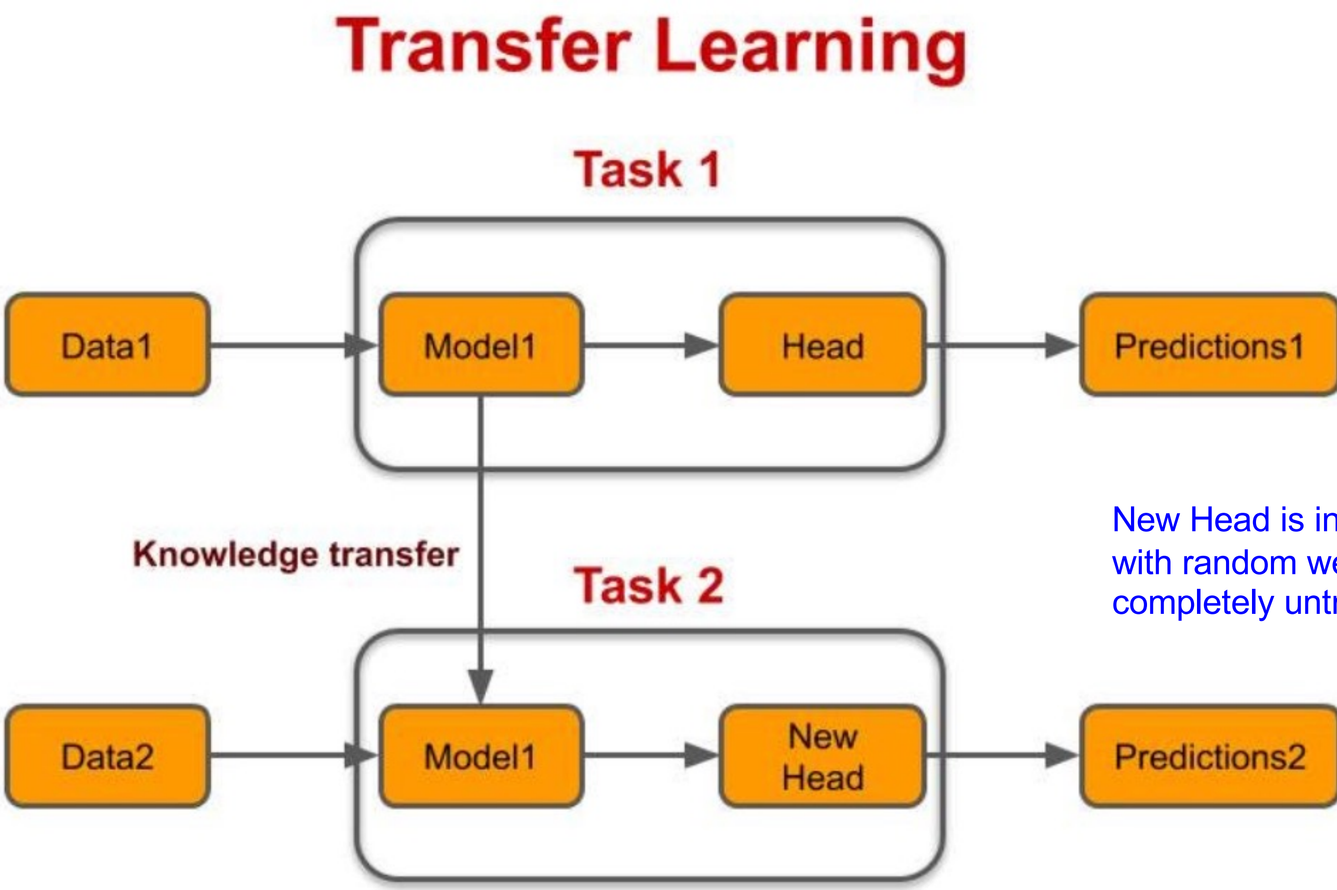
(After break we'll consider GPT and T5)

- Survey of Tasks in NLP (if time)

Paradigms for Transfer Learning

Large:

Typically smaller:



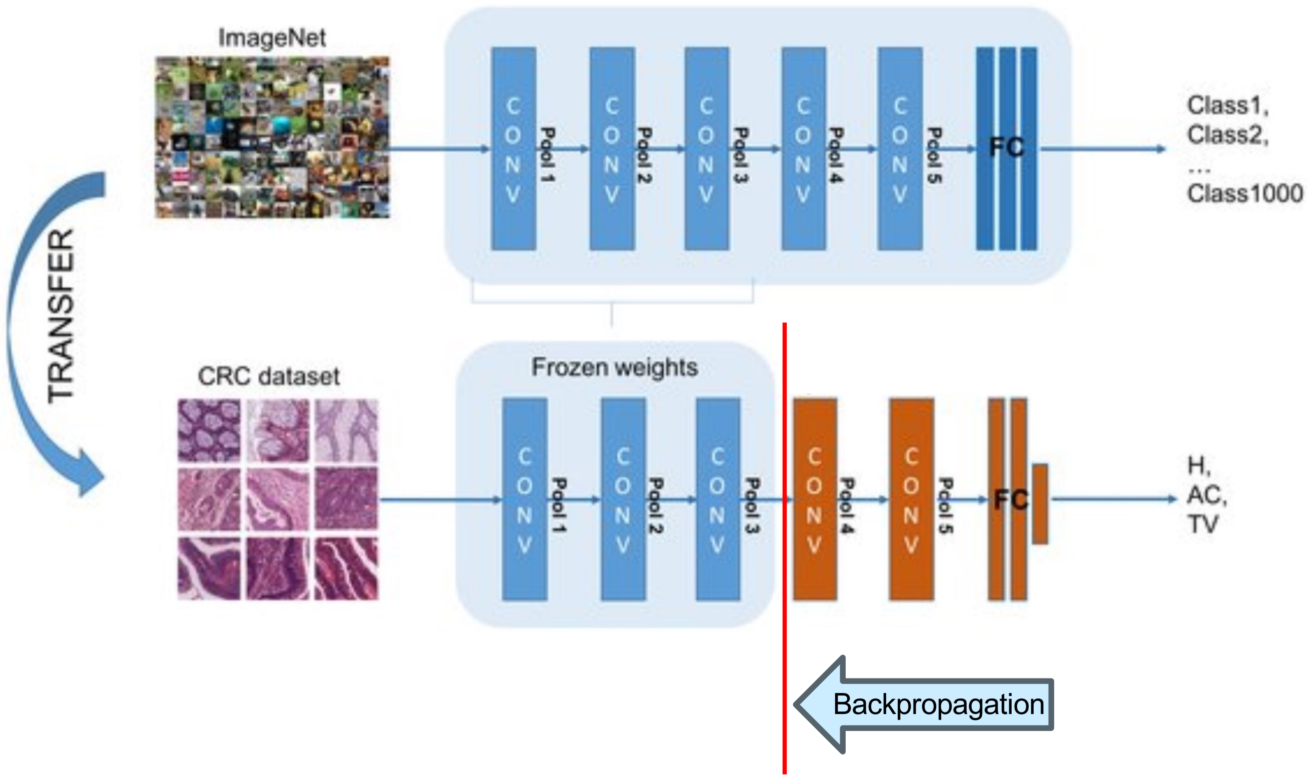
New Head is initialized with random weights: completely untrained!

Big question: How does backpropagation work? Do you freeze the pretrained model or allow it to be retrained?

Paradigms for Transfer Learning

Big question: How does backpropagation work? Do you freeze the pretrained model or allow it to be retrained?

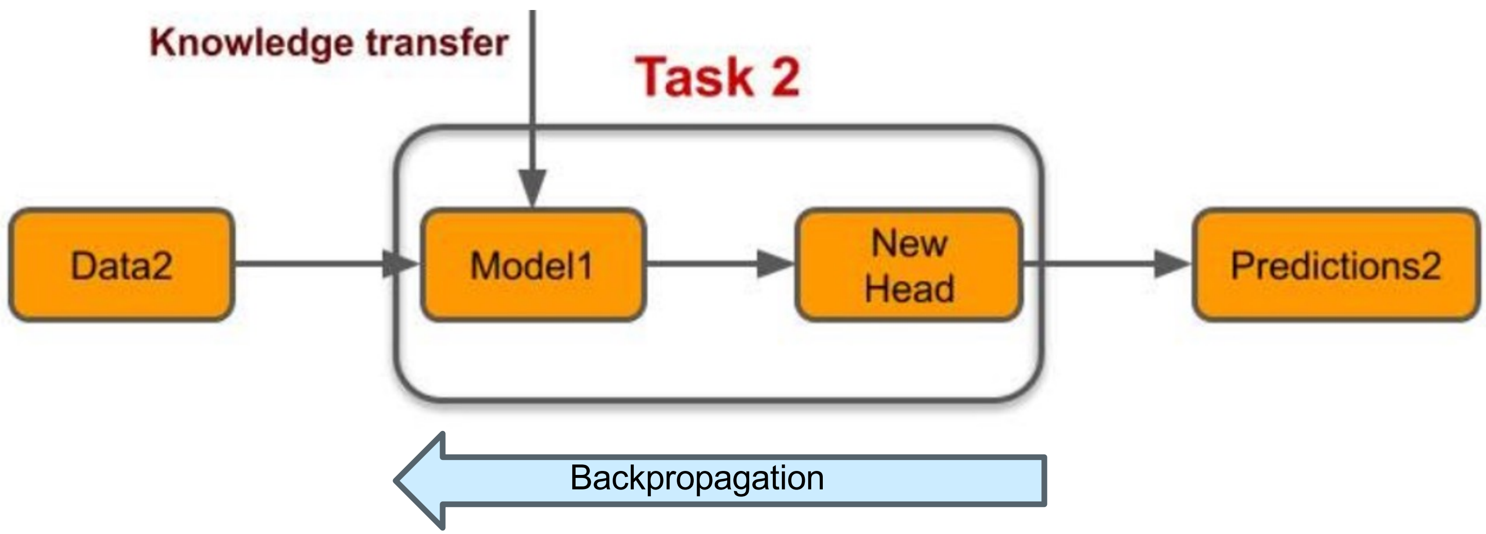
If frozen: Then you are treating the pretrained model as a feature extractor for the downstream layers:



Paradigms for Transfer Learning

Big question: How does backpropropagation work?

If you let the model be retrained with no restrictions, e.g.,



Then the new head will be trained, but the new model may take a long time to train, or – much worse -- suffer from catastrophic forgetting, where the original training is disrupted and performance gets worse!

Paradigms for Transfer Learning

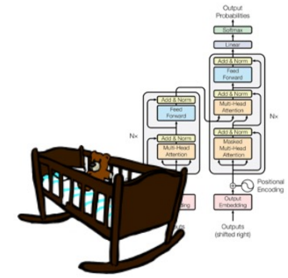
How to prevent catastrophic forgetting?

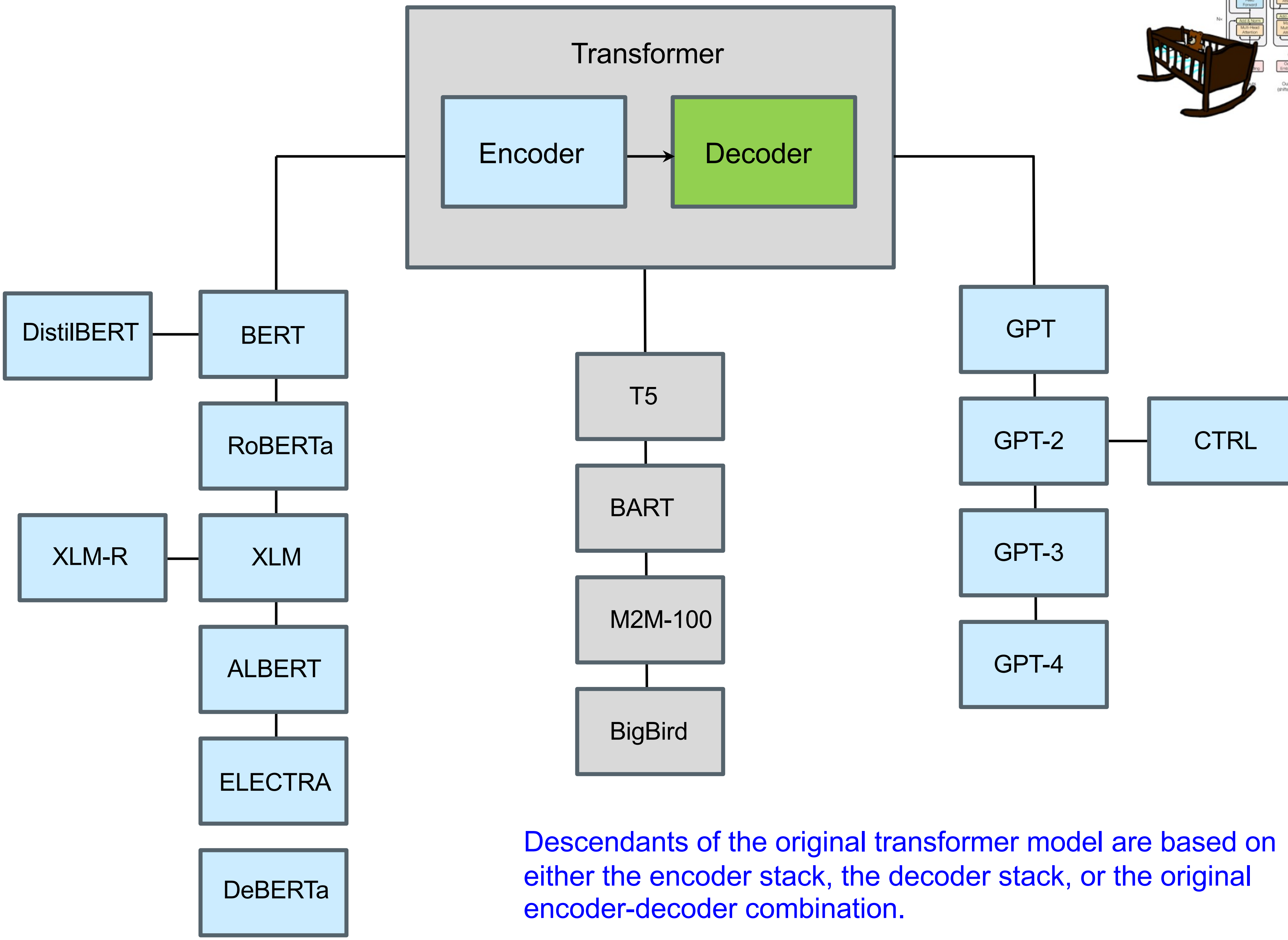
A popular solution is Layer-wise Relevance Adjustment (LoRa)

LoRa uses Layer-wise Relevance Estimation to adjust the learning rates of the pretrained model:

- During pretraining:
  - Estimate the relevance of each layer to the original task by measuring the gradients: layers with larger gradients were more relevant to the task;
- During retraining:
  - Continue with LRE with respect to the new task;
  - Use the LRE estimates to adjust how much updating should be done in these layers, or how much regularization should be performed.

The Transformer Family

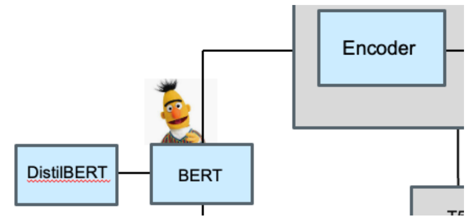




Descendants of the original transformer model are based on either the encoder stack, the decoder stack, or the original encoder-decoder combination.

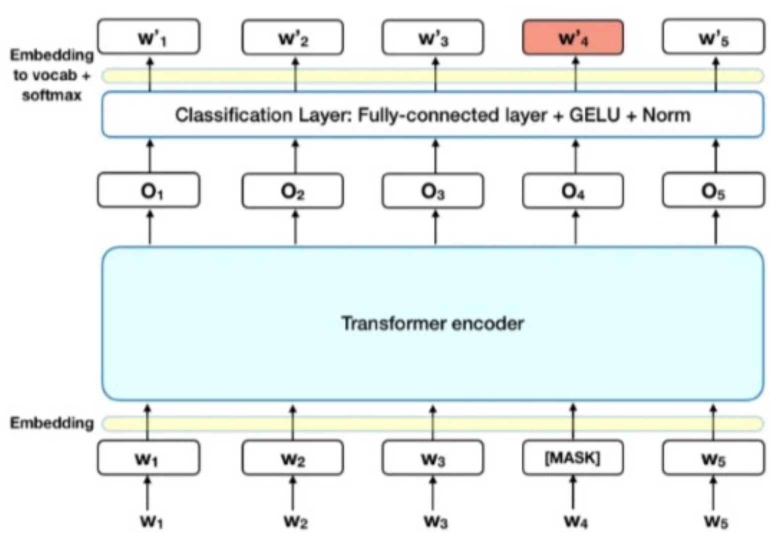
BERT: Bidirectional Encoder Representations from Transformers

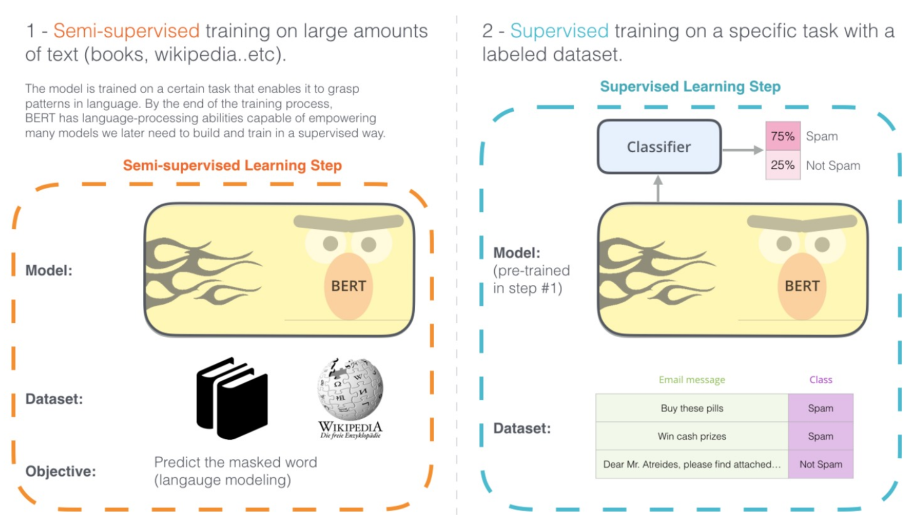
The most significant difference in the models is how they process the input sequence:



BERT consists of:

- The stacked-encoder part of the full transformer model, with
- A single linear layer on top, acting as a classifier (depending on the task);
- Pretraining based on Cloze tasks and next-sentence prediction; and
- Transfer learning to adapt to a new task.



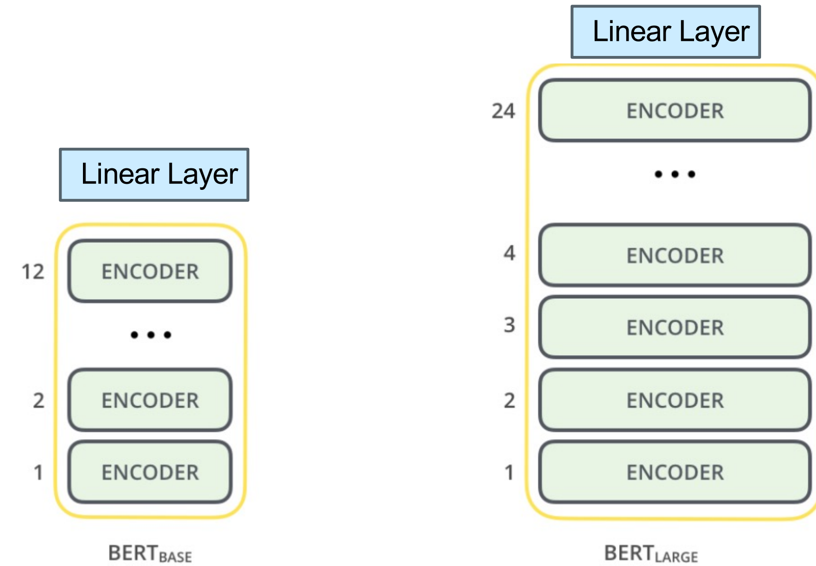


BERT

Bert has two implementations:

- ○ Bert base:
  - 12 layers; 12 attention heads per layer;
  - 768 hidden units; 110 M parameters

- ○ Bert large:
  - 24 layers; 16 attention heads per layer;
  - 1024 hidden units; 340 M parameters



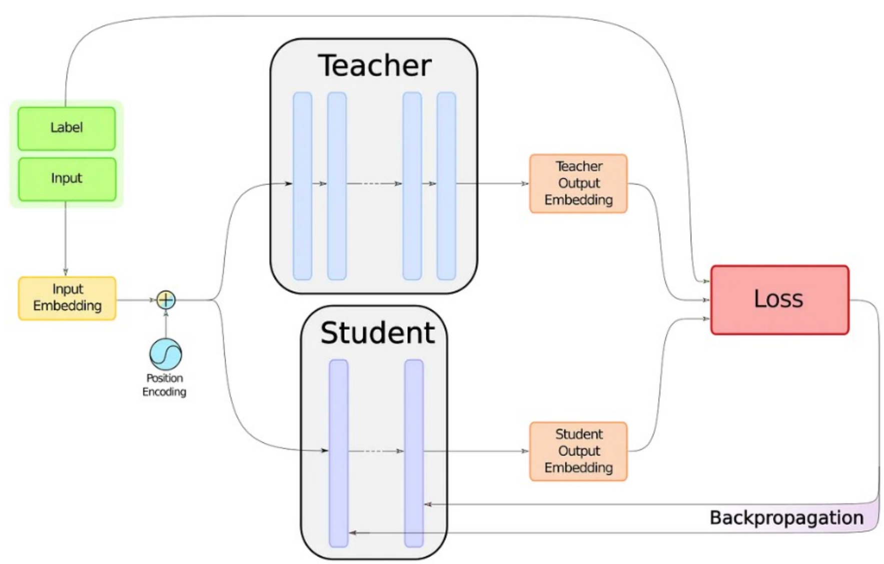
System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

BERT and DistilBERT

Another version, created by HuggingFace, is DistilBERT:

Created using Knowledge Distillation, or Student Teacher Transfer Learning:
A small network is trained not on data, but to imitate another network.

- BERT is the "teacher" network;
- DistilBERT is a smaller "student" network:
  - Only 6 encoder layers
  - 40% fewer parameters
  - 60% faster
  - Achieves 97% of BERT's performance.



BERT: Training

BERT added two significant ideas for training which allowed it to achieve SOTA performance on significant tasks:

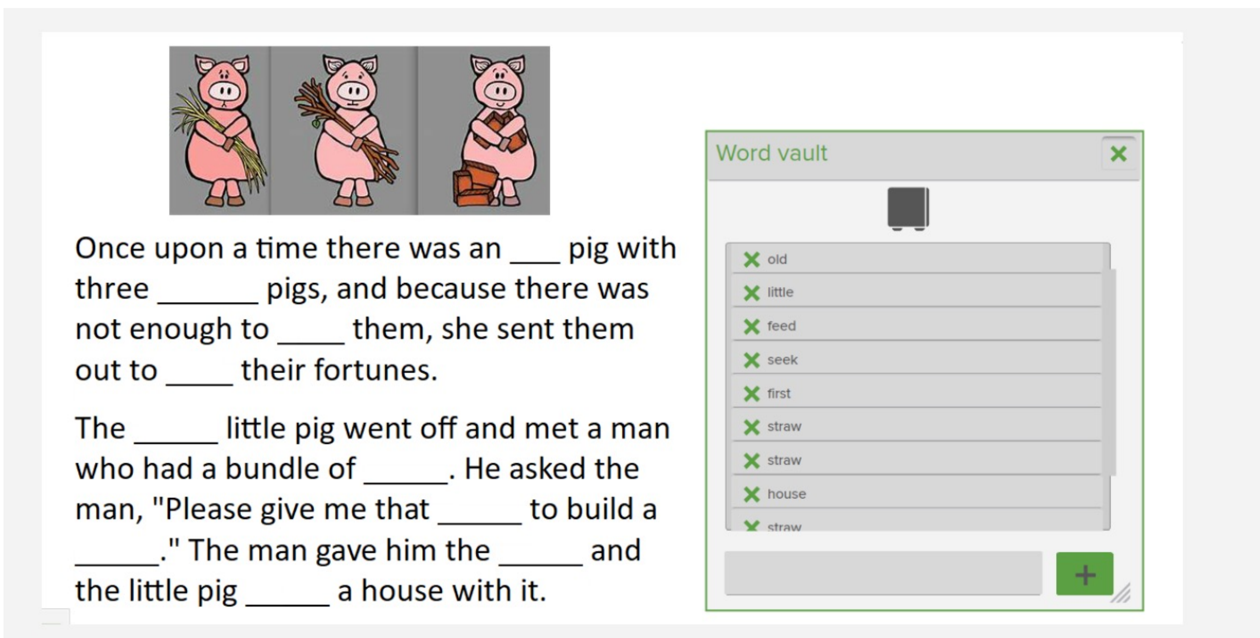
Training using a Masked Language Model (MLM); and

Focus on a "next sentence prediction" task with two sentences as input.

Note: The first versions of BERT used the 800 million word BooksCorpus and a 2.5 B word English Wikipedia corpus.

BERT: Training

The Masked Language Model is based on an educational theory/testing paradigm known as the Cloze Task, where students learn a language by filling in blanks in a story or piece of text:

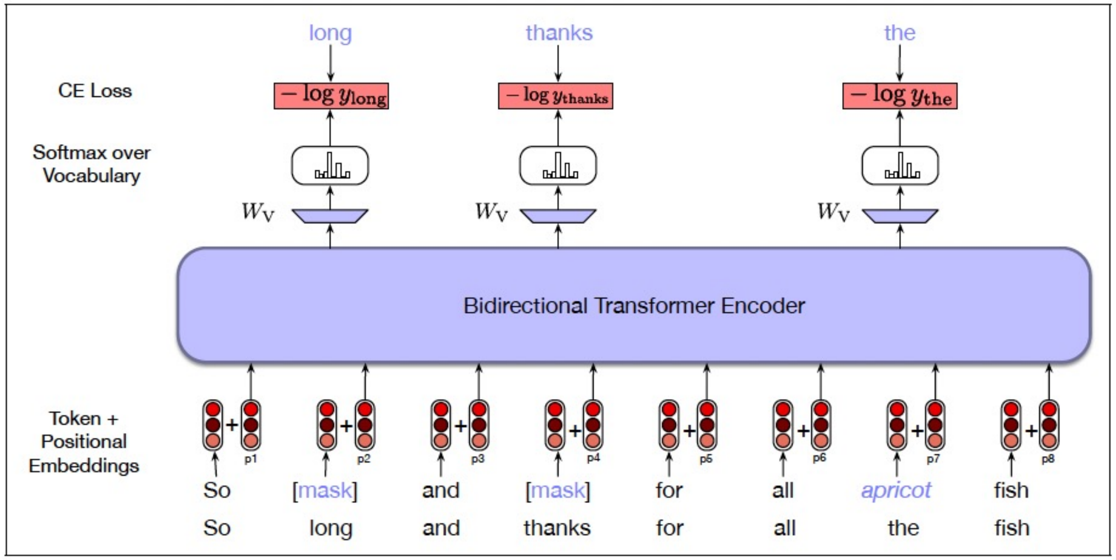


BERT: Training

Masked Language Modeling uses unannotated text from a large corpus. 15% of the words in the corpus are selected for the training phase: of these,

- 80% of replaced with the token [MASK]
- 10% are replaced with randomly-selected tokens
- 10% are left unchanged.

The model is trained to predict the missing tokens.



BERT: Training

A variation of MLM uses spans (subsequences of the input sequence); all of the words in the span are replaced as before:

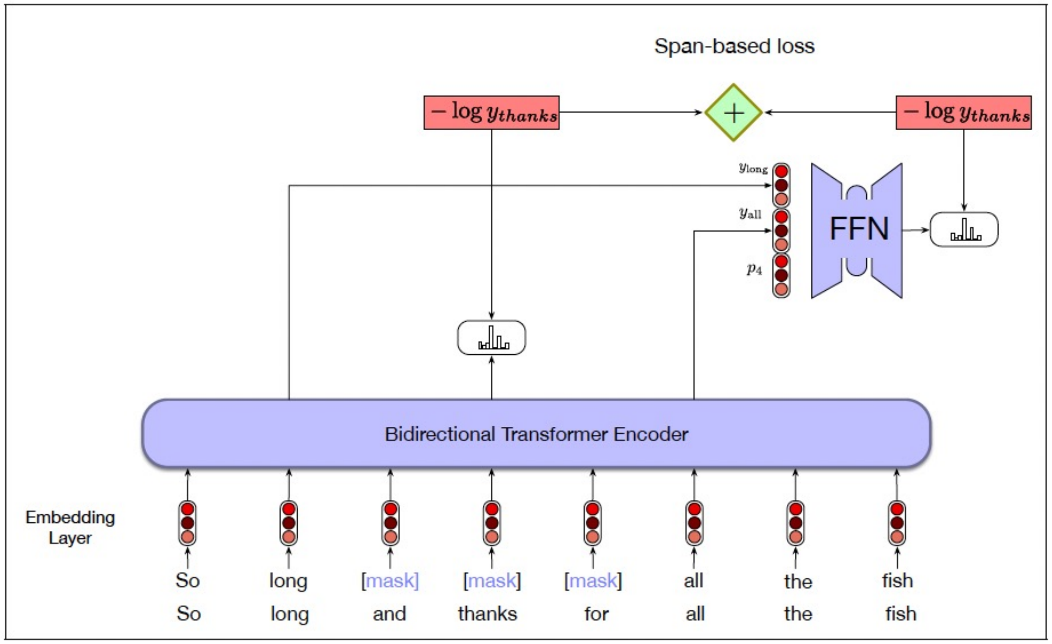


Figure 11.6 Span-based language model training. In this example, a span of length 3 is selected for training and all of the words in the span are masked. The figure illustrates the loss computed for word *thanks*; the loss for the entire span is based on the loss for all three of the words in the span.

BERT: Training

The Next Sentence Prediction task is to input TWO sentences starting with the token [CLS] and separated by the token [SEP]. The training set has 50% sentences that are next to each other in the corpus, and 50% random sentences.

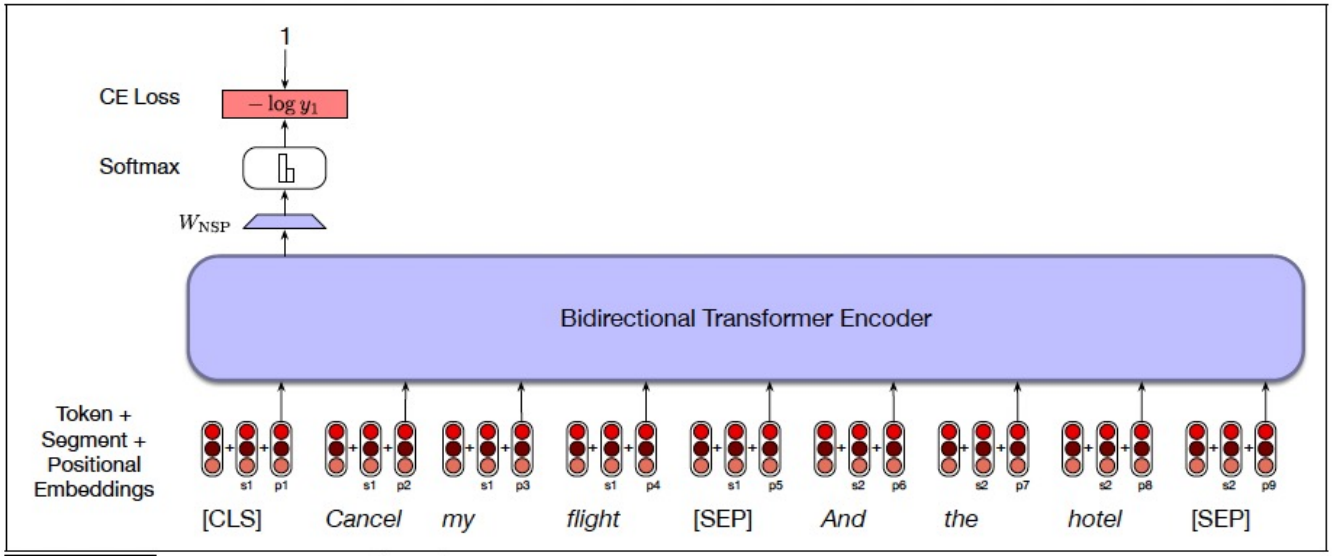


Figure 11.7 An example of the NSP loss calculation.

BERT: Training

Bert can be used for sentence classification if a single sentence is input:

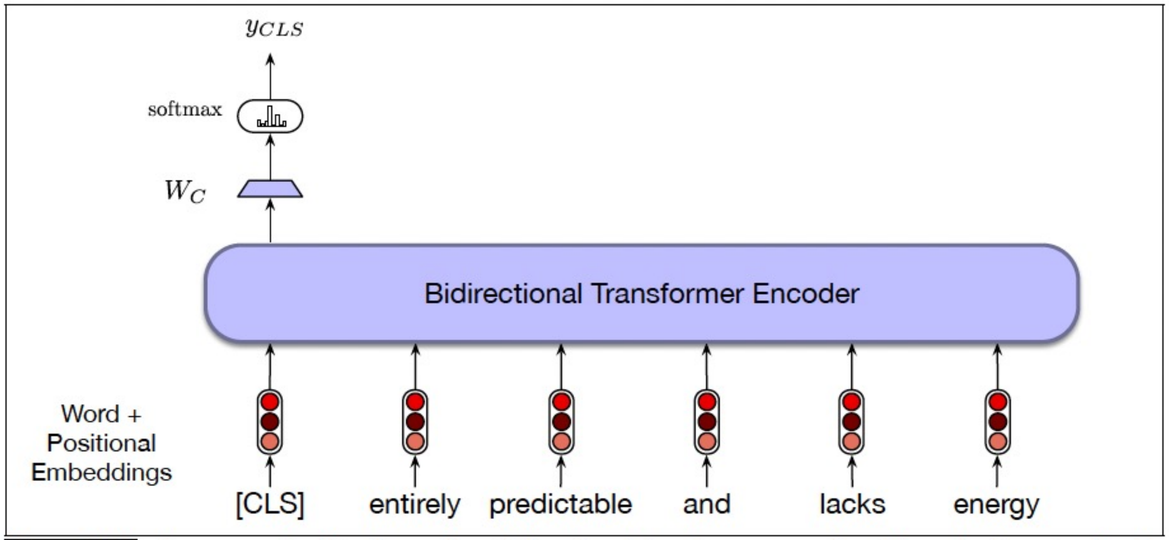


Figure 11.8 Sequence classification with a bidirectional transformer encoder. The output vector for the [CLS] token serves as input to a simple classifier.

BERT: Training

BERT can classify the relationship between two sentences:

- Neutral
  - a: Jon walked back to the town to the smithy.
  - b: Jon traveled back to his hometown.
- Contradicts
  - a: Tourist Information offices can be very helpful.
  - b: Tourist Information offices are never of any help.
- Entails
  - a: I'm confused.
  - b: Not all of it is very clear to me.

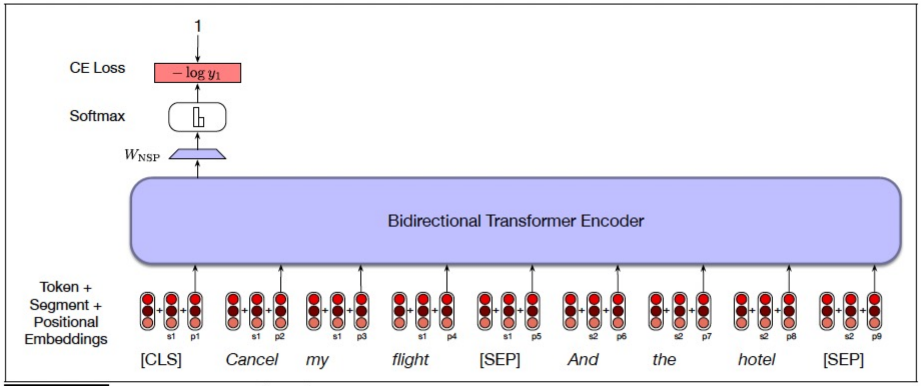


Figure 11.7 An example of the NSP loss calculation.

BERT: Training

BERT can generate the most likely sentence to follow a given sentence:

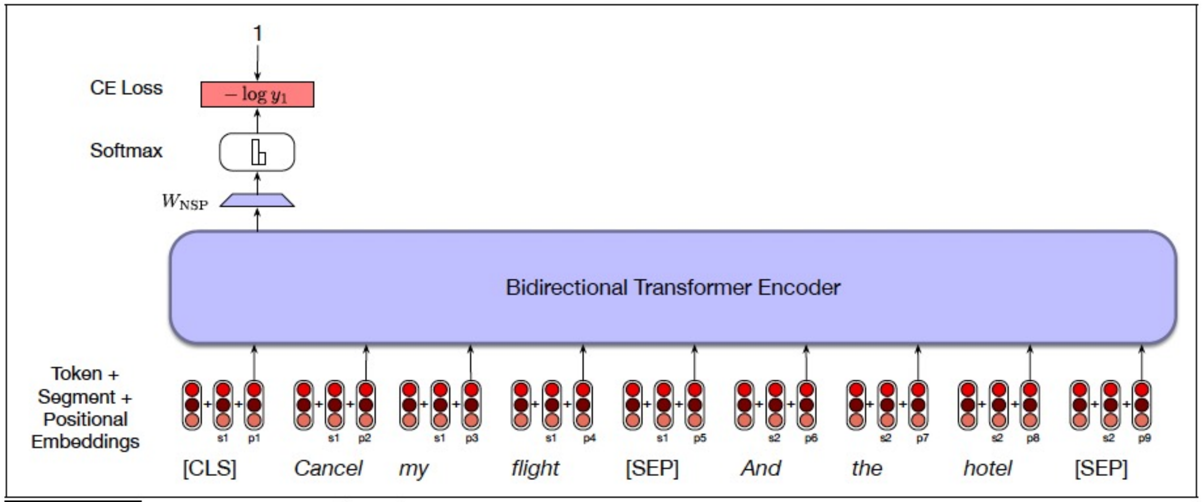


Figure 11.7 An example of the NSP loss calculation.

Use Beam Search to find most likely sentence to follow.

Using BERT

Bert can be used for sequence labelling if all of the outputs are used:

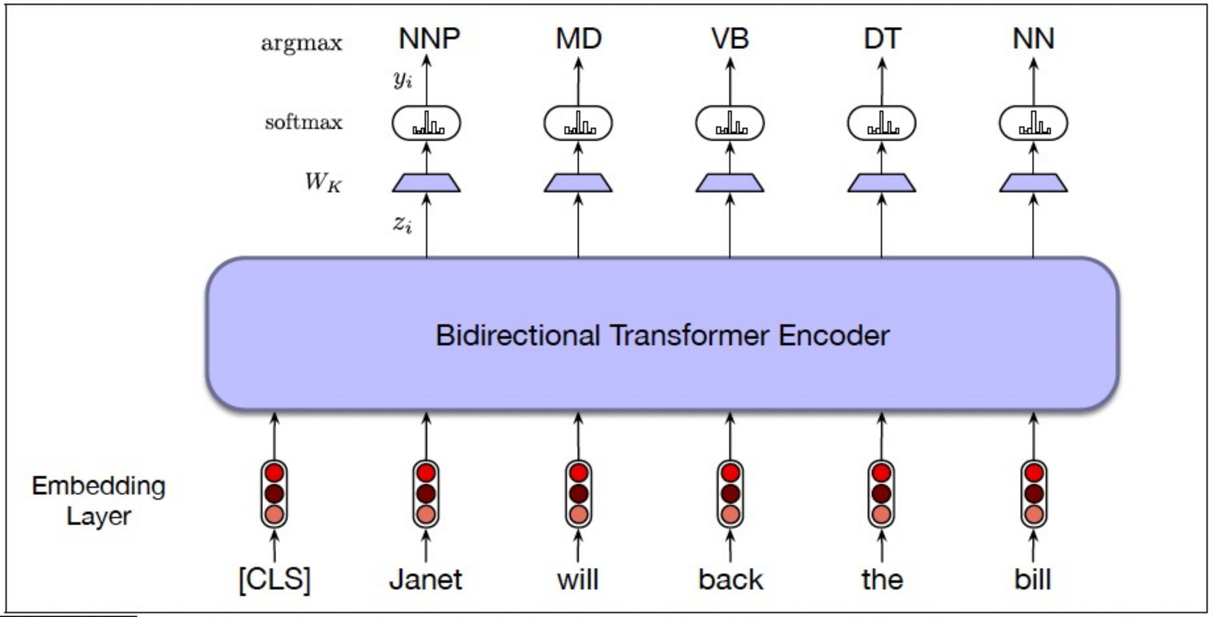
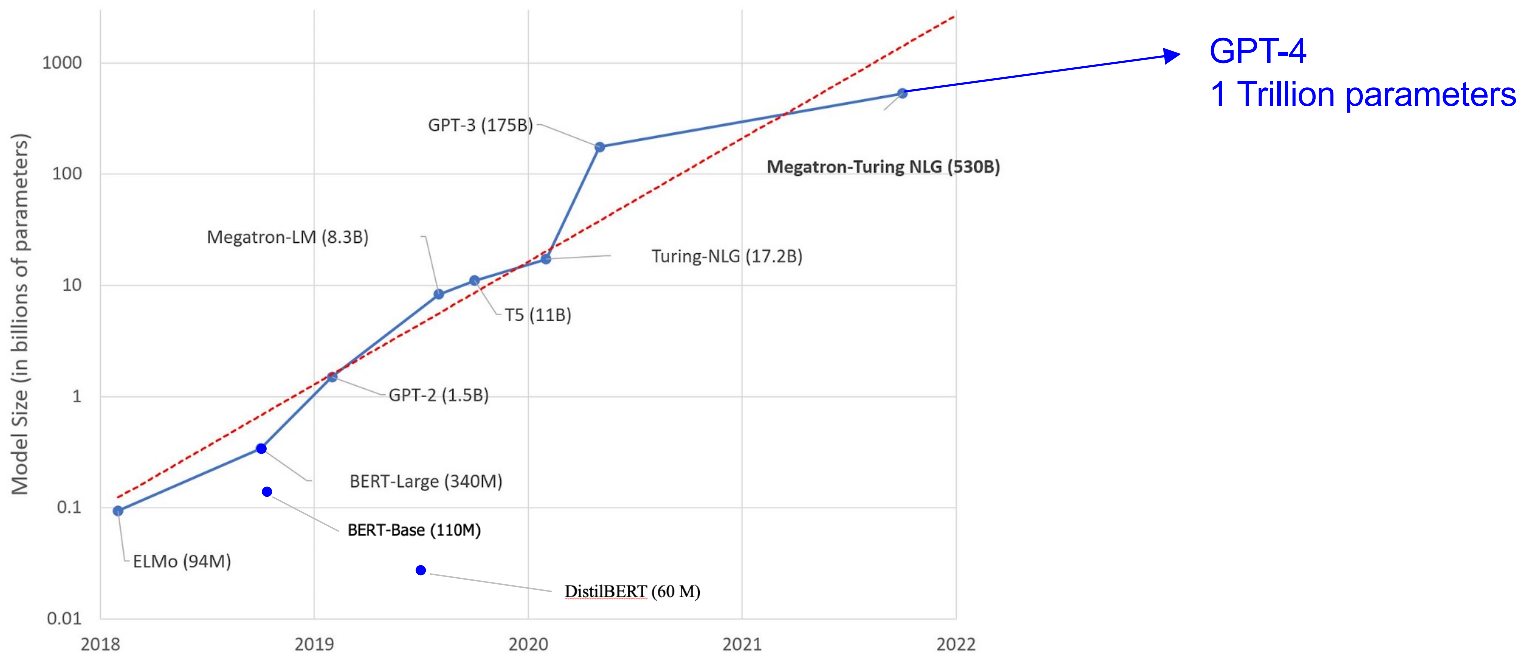


Figure 11.9 Sequence labeling for part-of-speech tagging with a bidirectional transformer encoder. The output vector for each input token is passed to a simple k-way classifier.

BERT Punches Above Its Weight!

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$BERT_{BASE}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$BERT_{LARGE}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**



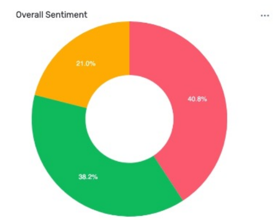
Most Important NLP Tasks: Classification

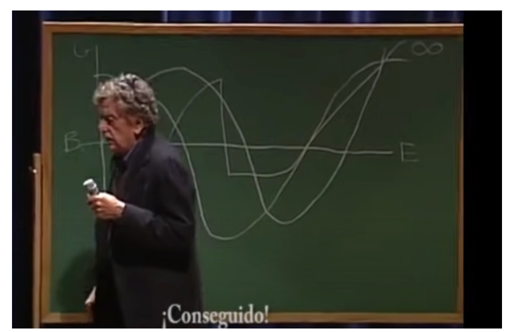
Sentiment Analysis: identifying the position of a piece of text in some scale of sentiment.

Position may be categorical (2 stars out of 5) or continuous in some range (2.3 on a scale 0 .. 10)

Types of sentiment:

- Positive – Negative
- Aspect or point of view or bias (e.g., political)
- Intent detection
- Emotion Detection
    - Happiness
    - Excited/enthusiastic
    - Frustration or Anger
- Friendship, affection, love or sexual attraction
- Humorous
- Irony
- Hate speech and Fake News detection (next slide)





Source: https://monkeylearn.com/sentiment-analysis/

Fake News and Hate Speech Detection

Fake News Detection: detecting and filtering out texts containing false and misleading information.

Stance Detection: determining an individual's reaction to a primary actor's claim. It is a core part of a set of approaches to fake news assessment.

Hate Speech Detection: detecting if a piece of text contains hate speech.

Help Center   Using Twitter   Managing your account   Safety and security   Rules and pc

Help Center › Safety and cybercrime › Hateful conduct policy

Hateful conduct policy

Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.

Hateful imagery and display names: You may not use hateful images or symbols in your profile image or profile header. You also may not use your username, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate towards a person, group, or protected category.

Facebook/Meta:

Policy Rationale

We believe that people use their voice and connect more freely when they don't feel attacked on the basis of who they are. That is why we don't allow hate speech on Facebook. It creates an environment of intimidation and exclusion, and in some cases may promote offline violence.

We define hate speech as a direct attack against people — rather than concepts or institutions— on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. We also prohibit the use of harmful stereotypes, which we define as dehumanizing comparisons that have historically been used to attack, intimidate, or exclude specific groups, and that are often linked with offline violence. We consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants and asylum seekers from the most severe attacks, though we do allow commentary and criticism of immigration policies. Similarly, we provide some protections for characteristics like occupation, when they're referenced along with a protected characteristic. Sometimes, based on local nuance, we consider certain words or phrases as frequently used proxies for PC groups.

Fake News and Hate Speech Detection

Why is NLP so necessary in this?

- Users watch 4,146,600 YouTube videos
- 456,000 tweets are sent on Twitter
- Instagram users post 46,740 photos

With 2 billion active users Facebook is still the largest social media platform. Let that sink in a moment—more than a quarter of the world's 7 billion humans are active on Facebook! Here are some more intriguing Facebook statistics:

- 32 billion people are active on Facebook **daily**
- Europe has more than 307 million people on Facebook
- There are five new Facebook profiles created every second!
- More than 300 million photos get uploaded per day
- Every minute there are 510,000 comments posted and 293,000 statuses updated

1.836 Billion Facebook Posts each day....

Even though Facebook is the largest social network, Instagram (also owned by Facebook) has shown impressive growth. Here's how this photo-sharing platform is adding to our data deluge:

- There are 600 million Instagrammers; 400 million who are active every day
- Each day 95 million photos and videos are shared on Instagram
- 100 million people use the Instagram "stories" feature daily

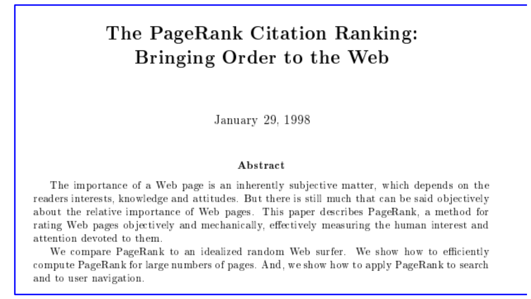
Information Retrieval

(An old subject, even before Google made it the the most popular text-processing task.)

- Resource Retrieval from text queries/questions
  - Resource could be
    - Highly structured (relational database, code)
    - Semi-structured (Markup Languages (XML), labeled documents)
    - Unstructured (documents)
  - Database search from keywords
  - Google search
  - Backend to Speech to Text systems (siri)
  - Question Answering (next slide)

- Sentence/document similarity: determining how "similar" two texts are
  - Notion of "similar" is variable (similar topic, similar sentiment, ...)
  - Relationship to IR:
    - How similar is text query to a document?
    - "Retrieve more documents similar to this one"
  - Create a map/graph of documents similar to given sentence/document
  - Plagiarism/copyright infringement

- Document Ranking: Rank documents as to some criterion (e.g., PageRank)
  - How well does this document satisfy my query?
  - How important/authoritative is this document?


The PageRank Citation Ranking: Bringing Order to the Web

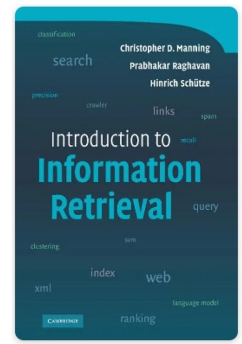
January 29, 1998

Abstract

The importance of a Web page is an inherently subjective matter, which depends on the readers interests, knowledge and attitudes. But there is still much that can be said objectively about the relative importance of Web pages. This paper describes PageRank, a method for rating Web pages objectively and mechanically, effectively measuring the human interest and attention devoted to them.

We compare PageRank to an idealized random Web surfer. We show how to efficiently compute PageRank for large numbers of pages. And, we show how to apply PageRank to search and to user navigation.

1999



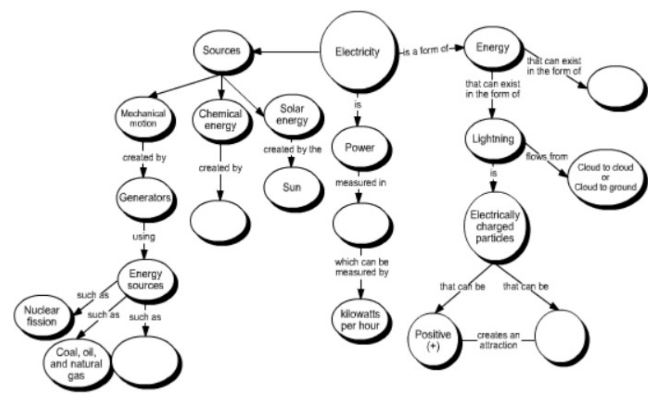
2008

Net worth of Google in 2022: $1.135 Trillion.

Entities, Relations, and Knowledge Graphs

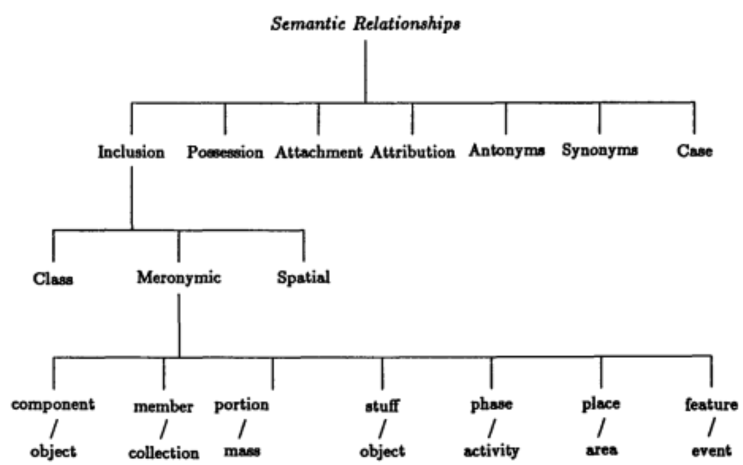
- **Named Entity Recognition:** tagging entities in text with their corresponding
- type, typically in BIO notation.)
- **Coreference Resolution:** clustering mentions in text that refer to the same underlying real-world entities.
- **Relation extraction:** extracting semantic relationships from a text, e.g.,
  - Is-A
  - Has-A
  - Son-Of
  - Part-Of
  - Size-of
  - etc., etc., etc.
- **Build a graph structure:**
  - Knowledge Graph
  - Concept Map
  - Mind Map

Concept Map about Electricity:

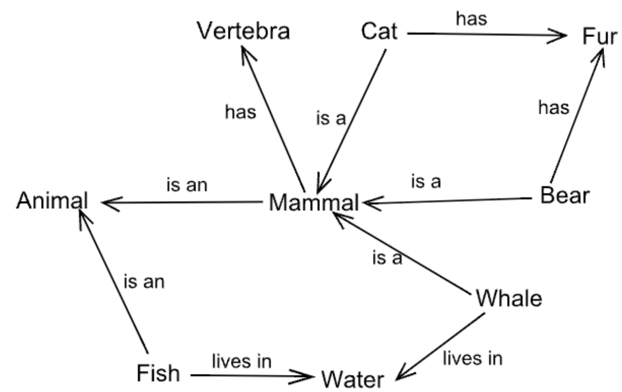


- Graphs can be used to enhance other NLP tasks: search, similarity, question answering, etc
- **Entity Linking:** recognizing and disambiguating named entities to
- a knowledge base (e.g., Wikidata).
- Relation prediction: identifying a named relation between two named
- semantic entities.

Entities, Relations, and Knowledge Graphs



Based on Landis et al. (1987); Winston et al. (1987); Chaffin et al. (1988).



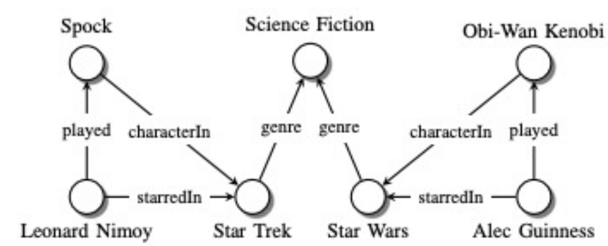


Fig. 1. Sample knowledge graph. Nodes represent entities, edge labels represent types of relations, edges represent existing relationships.

Text-to-Text Generation

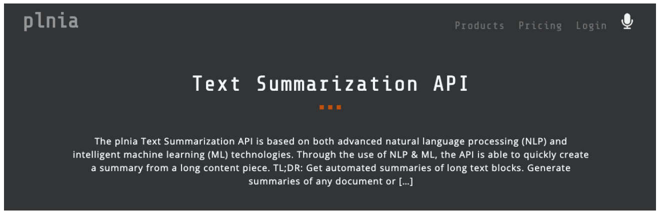
- **Machine Translation:** translating from one language to another.
  - Covered in lecture – Transformer technology transformed this task
- **Text Generation:** creating text from a prompt or subject phrase that appears indistinguishable from human-written text.
  - Covered in lecture – Use language models, Large Language Models (GPT) have transformed this task
- **Lexical Normalization:** translating/transforming a non-standard text to a standard register.
- **Paraphrase Generation:** creating an output sentence that preserves the meaning of input but includes variations in word choice and grammar.
- **Text Simplification:** making a text easier to read and understand, while preserving its main ideas and approximate meaning.
- Text Summarization (next slide)

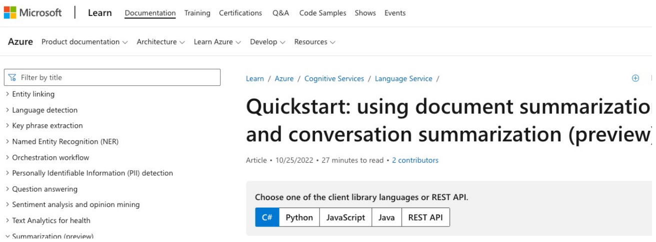
How Large Language Models are Transforming Machine-Paraphrased Plagiarism

Jan Philip Wahle[☾,*], **Terry Ruas**[*], **Frederic Kirstein**[♦,*], **Bela Gipp**[*]
[*]Georg-August-Universität Göttingen, Germany
[♦]Mercedes-Benz Group AG, Germany
[☾]wahle@gipplab.org

Topics and Keywords; Text Summarization

- **Topic Modeling:** identifying abstract "topics" underlying a collection of documents.
- **Keyword Extraction:** identifying the most relevant terms to describe the subject of a document
- **Text Summarization:** Reducing size of document while preserving the most important information
  - **Extractive:**
    - Identify the most important sentences in a document and construct the summary from these exact sentences
    - TextRank, LexRank (implements PageRank on sentences in a document)
    - Latent Semantic Analysis (Singular Value Decomposition on a word-sentence matrix)
  - **Abstractive:**
    - Create new text summarizing main points
    - Use of Large Language Models: GPT, BERT, etc...
  - **Use cases for Text Summarization:**
    - Summaries for busy executives or (students!)
    - Summaries of articles, books, chapters
    - Automatic Table of Contents or Indices
    - Downstream from Speech-to-Text systems:
      - Notetaking of meetings, lectures
      - Abstracts of podcasts, YouTube videos
      - Automatic summary of customer phone calls





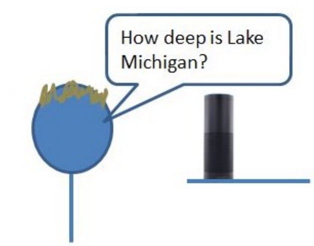
Chatbots and Question Answering

- **Slot Filling or Cloze Task:** aims to extract the values of certain types of attributes (or slots, such as cities or dates) for a given entity from texts.
- **Chatbots:** Conversation agents (started with Eliza in early 1060's!)
- **Dialog Management:** managing of state and flow of conversations.
- **Question Answering:** Responding to textual queries with textual answers
  - **Extractive QA**: The model extracts the answer from a knowledge source, such as a knowledge graph, database, or document (next slide).
  - **Open Generative QA:** The model generates free text directly based on the (global) context.
  - **Closed Generative QA:** The model generates free text directly based only on the question.

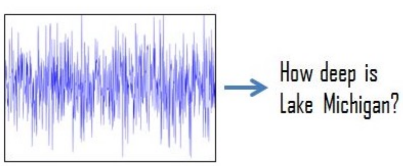
Question Answering using Knowledge Graphs

.

Step #1: I randomly shout out "Alexa, how deep is Lake Michigan?"

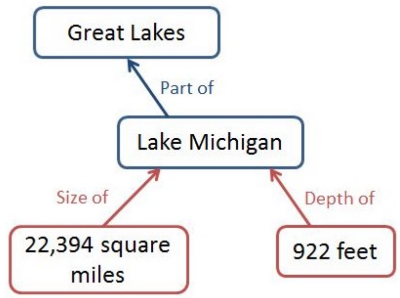


Step #2: Alexa uses voice-to-text processing to parse the noise I made into text.

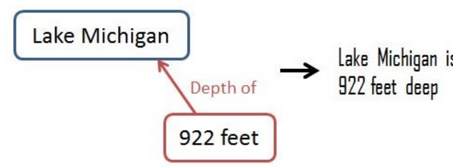


Step #3: Alexa uses Natural Language Processing (NLP) to figure out what I want.

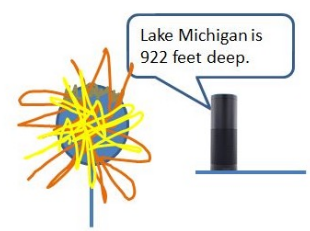
Step #4: Alexa searches its semantic graph for the answer to my question.



Step #5: Alexa uses Natural Language Generation (NLG) to construct a textual answer.



Step #6: Alexa uses text-to-voice processing to calmly blow my mind.



Reasoning with Text

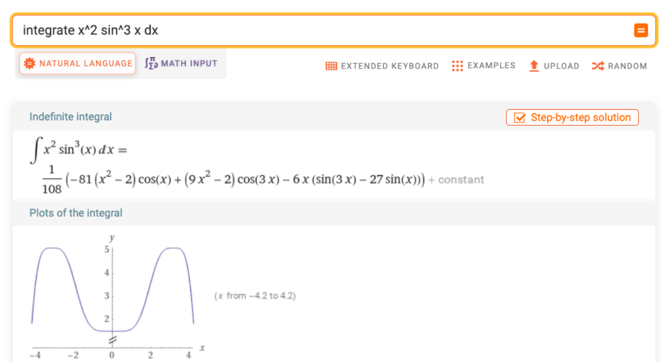
- Logical Relationship of two sentences/documents:
  - Entailment
  - Temporal sequence
  - Specialization
- Subsystem of text generation at scale

- Text-to/from-First Order Logic: Translate between text and expressions in first-order logic:

No student failed Chemistry, but at least one student failed History.
$\neg\exists x\,(Student(x) \land Failed(x,Chemistry)) \land \exists x\,(Student(x) \land Failed\,(x,History))$

10. Logic Puzzle: A farmer wants to cross a river and take with him a wolf, a goat and a cabbage. He has a boat, but it can only fit himself plus either the wolf, the goat or the cabbage. If the wolf and the goat are alone on one shore, the wolf will eat the goat. If the goat and the cabbage are alone on the shore, the goat will eat the cabbage. How can the farmer bring the wolf, the goat and the cabbage across the river without anything being eaten?

- Use cases:
  - Teaching logic
  - Game/puzzle solving
  - Interface to automated theorem prover
    - Prolog
    - Planner
    - Wolfram Alpha



Text-to-Data and Data-to-Text

- Text-to-Image: generating photo-realistic images which are semantically consistent with the text descriptions.
- Image captioning: Generate captions for input images
- Video-to-Text: Generating text describing a sequence of images
- Text-to-Speech: Human-like reading of input text.
- Speech-to-Text: transcribing speech to text

An example of some of the images created by Imagen, Google's text-to-image AI generator.