Automatic Punctuation using Deep NLP

Traditional definition of NLP

- Deal with analyzing, understanding and generating the languages that humans use naturally (natural language).

- Study knowledge of language at different levels
  - Phonetics and Phonology - the study of linguistic sounds
  - Morphology - the study of the meaning of components of words
  - Syntax - the study of structural relationship between words
  - Semantics - the study of meaning
  - Discourse - the study of linguistic units larger than a single utterance

Challenge of NLP

Many to many mappings between symbolic language and semantic meaning.

Ambiguity

Example: I made her duck.

- I cooked duck for her.
- I cooked duck belonging to her.
- I caused her to quickly lower her head or body.
- I waved my magic wand and turned her into a duck!

How do we represent the meaning of word?

Definition : meaning (Oxford dictionary)

- The idea that is represented by the word, phrase etc.
- The idea that a person wants to communicate by what they say or do.
- The idea that is expressed in a work of writing or art, etc.

Linguistic way of thinking of meaning:

Signifier (symbol) ⇔ Signified (idea or thing)

Representing words as discrete symbols

- In traditional NLP words were represented as discrete symbols.
- Meaning words can be represented as one-hot vectors.

For example, car and bike can be represented as :

car = [0 0 0 0 1 0 0 0 0 ....]

bike = [0 1 0 0 0 0 0 0 ....]

Where dimensionality of the vector is the vocabulary size.

But -

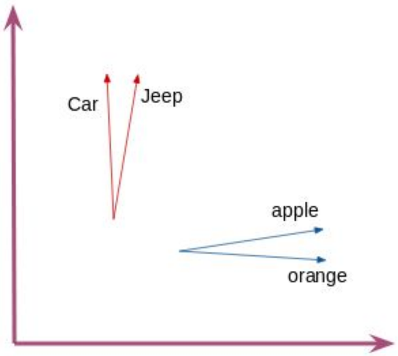
- These two vectors are orthogonal.
- There is no notion of similarity between the two.

Instead : learn to encode similarities in the vector themselves.

Distributed representation of words

“You shall know a word by the company it keeps” - J.R. Firth 1957

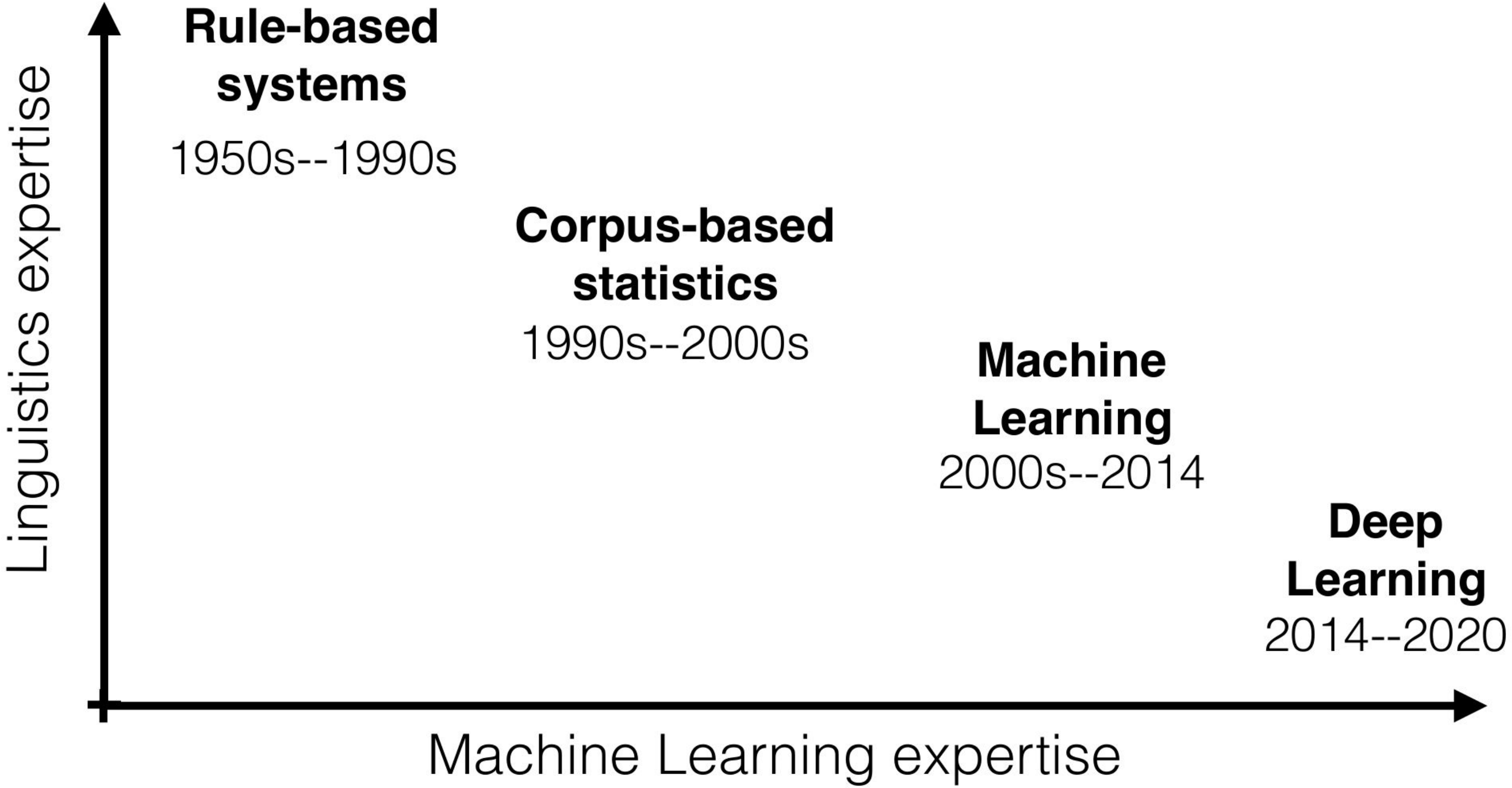
- One of the most successful ideas in modern NLP.
- A word’s meaning is dependent on its context.
- When a word *w* appears in a text, its *context* is the set of words which appear nearby.



The core idea here is that words that are used in similar ways are likely to have related meanings.

- Word2Vec - Mikolov et al. 2013
- Glove embeddings - Pennington et al. 2014

DEEP LEARNING EXPLOSION IN NLP



Linguistics expertise

Rule-based systems

1950s--1990s

Corpus-based statistics

1990s--2000s

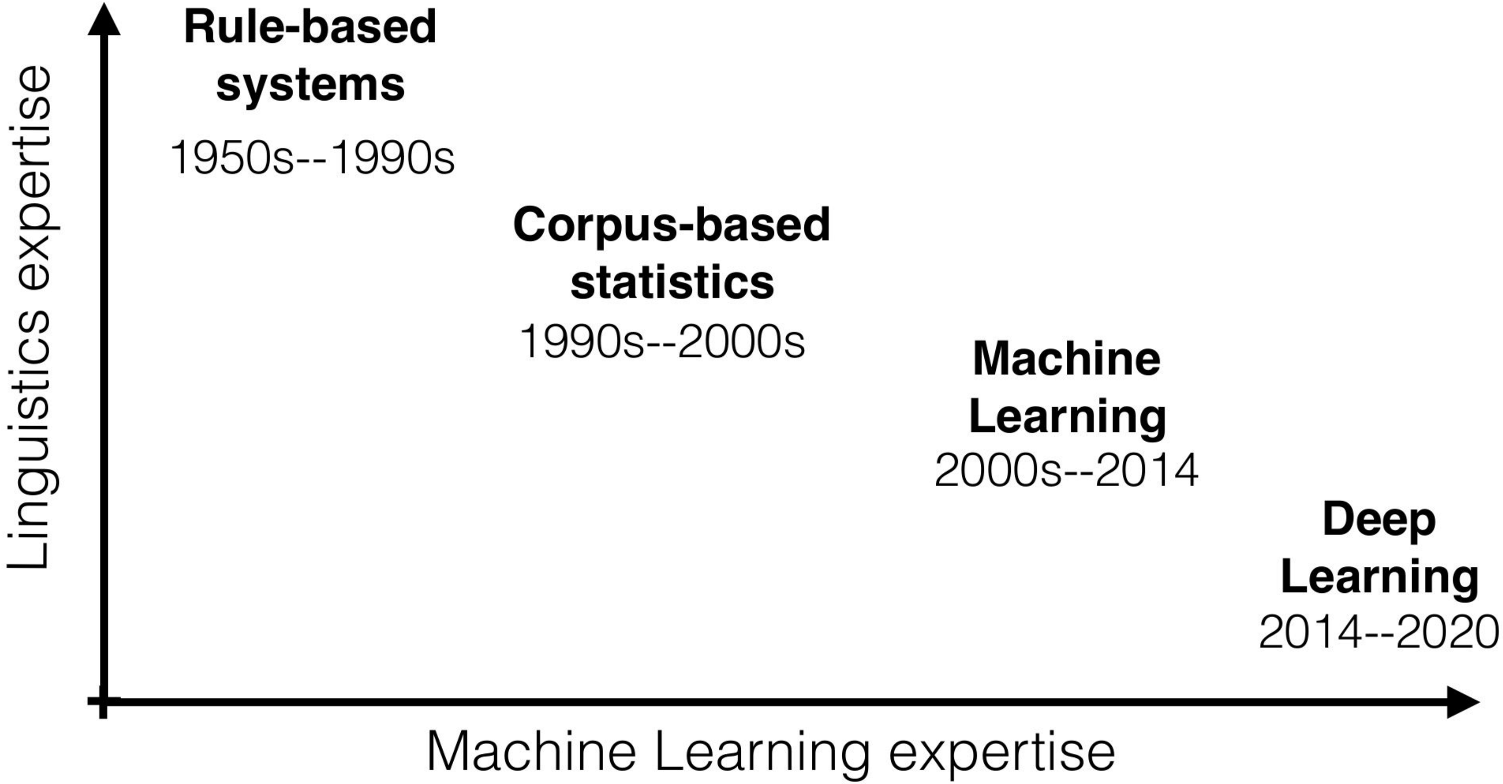
Machine Learning

2000s--2014

Deep Learning

2014--2020

Machine Learning expertise



Rule-based systems
1950s--1990s

Corpus-based statistics
1990s--2000s

Machine Learning
2000s--2014

Deep Learning
2014--2020

Linguistics expertise

Machine Learning expertise

Transparent	Debugability	**Black box**
Little	Needed labeled data	**Lots**

Automatic Punctuation

X : **hello how are you**

Y: Hello, how are you ?

Neural Seq2Seq tasks

- The sequence to sequence models is an example of **Conditional Language Model**.

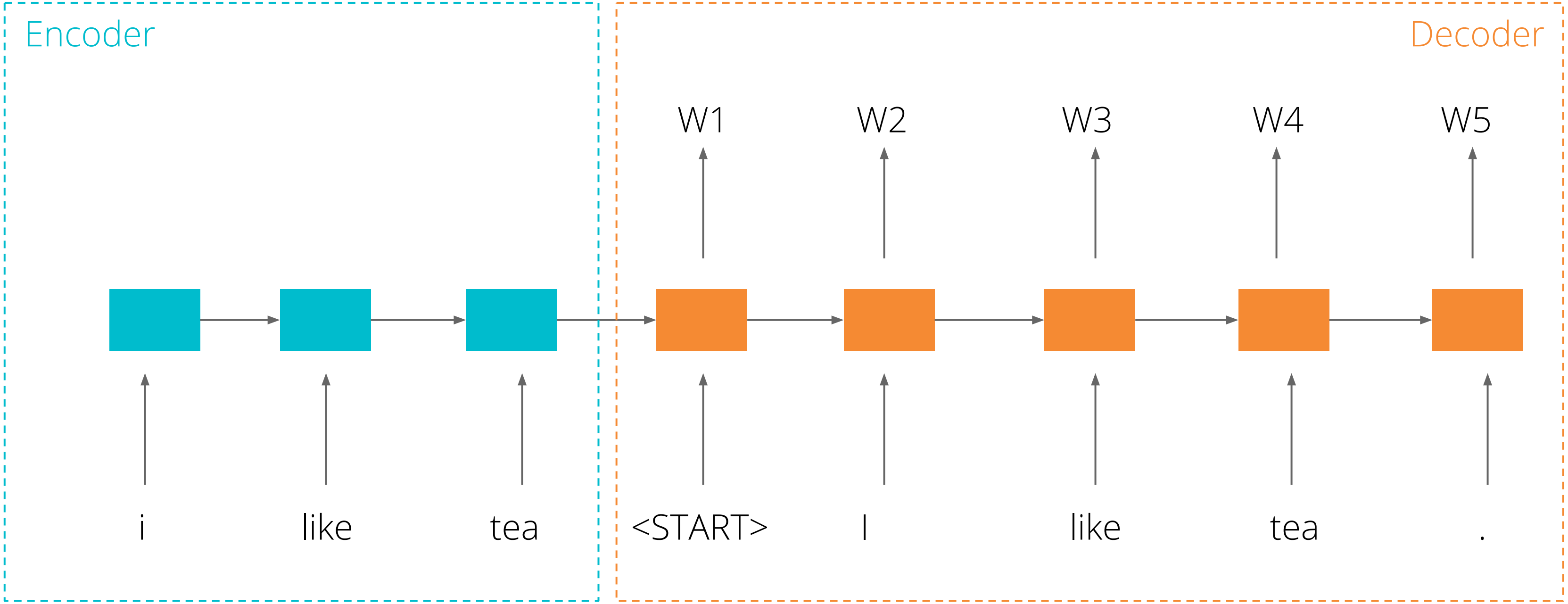
  They directly calculate :

$$P(y|x) = P(y_1|x)P(y_2|y_1, x)P(y_3|, y_1, y_2, x)...P(y_T|y_1, ..., y_{T-1}, x)$$

Encoder-Decoder Approach:

- An encoder RNN reads and encodes a source sentence into a fixed-length memory vector.
- A decoder RNN outputs a variable-length translation from the encoded memory vector.
- Encoder-decoder RNNs are jointly learned, optimizing target likelihood.

Encoder - Decoder Model (Training)



Seq2Seq is optimized as a single system.
Backpropagation operates "end-to-end"

Punctuation and inverse text normalization

Data Preparation: AI for bharat corpus for Hindi (63 million lines)

Preprocessing:

- Normalization
- Word tokenization
- Deduplication

Currently we are focusing on five punctuation marks (, ! ? - ।)

Remove all foreign characters from text apart from Hindi and punctuation.

This model will work on ASR output so some missing words in foreing languages from the original sentences will only make our data more closer to actual data in distribution.

In processed data, number of lines with no punctuation marks: 12 million lines.

Punctuation and inverse text normalization

- Training data - 1,000,000 lines
- Valid data - 10,000 lines
- Test data - 10,000 lines
- Preprocessed the data using Byte Pair Encoding to reduce the number of tokens and handle out of vocabulary words.
- Trained a seq2seq transformer model with input sequence as unpunctuated text and output sentence as punctuated text.

Eg: दक्षिण कोरिया अमेरिका और मेजबान न्यूजीलैंड की महिला टीमें हिस्सा ले रही हैं

दक्षिण कोरिया, अमेरिका और मेजबान न्यूजीलैंड की महिला टीमें हिस्सा ले रही हैं।

Overall WER on test set: 5.88

Overall CER on test set: 2.65

Language Modelling using Attention

BERT: Bidirectional Transformers for Language Understanding

How NLP cracked transfer learning

- The year 2018 ha been an inflection point for machine learning models involving text.
- Our conceptual understanding of how to represent words and sentences in a way that best captures the underlying meanings and relationships is rapidly evolving.

Model Details :

- Data : Wikipedia (2.5B words) + Book Corpus (800M words)
- Batch size: 131, 072 words
- Training time: 1 million steps (~40 epochs)
- Optimizer: AdamW, 1e - 4, learning rate, linear decay
- Trained on 4x4 or 8x8 TPU
- BERT large has around 340 million parameters.

BERT: Bidirectional Transformers for Language Understanding

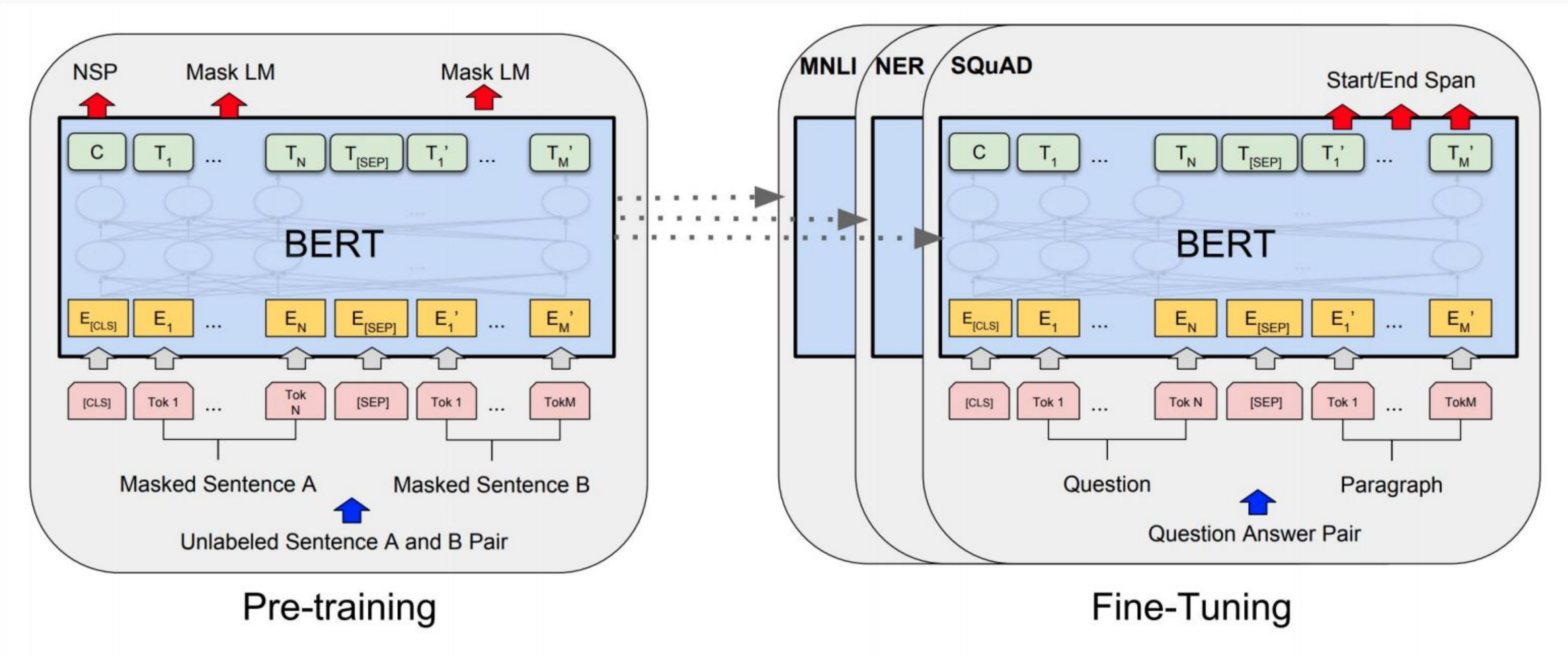
Main Ideas :

- Propose a new pre-training objective so that a deep bidirectional transformer can be trained.
  - The “masked language model” (MLM): the objective is to predict the original word of a masked word based only on its context.
  - Next sentence prediction

Merits of BERT :

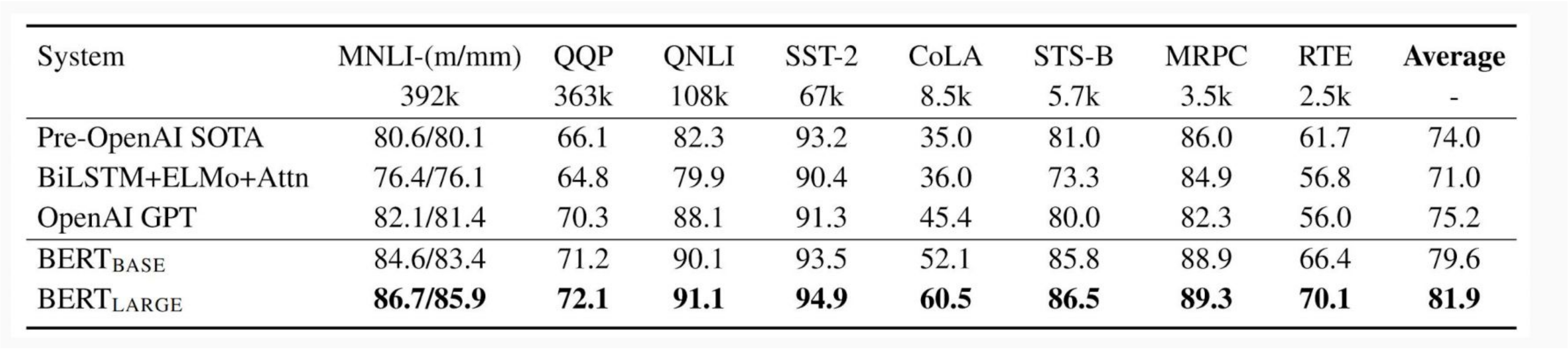
- Just fine-tune BERT model for specific tasks to achieve state of the art performance.
- BERT advances the state of the art for eleven NLP tasks

Fine Tuning Procedure



Source: https://arxiv.org/abs/1810.04805

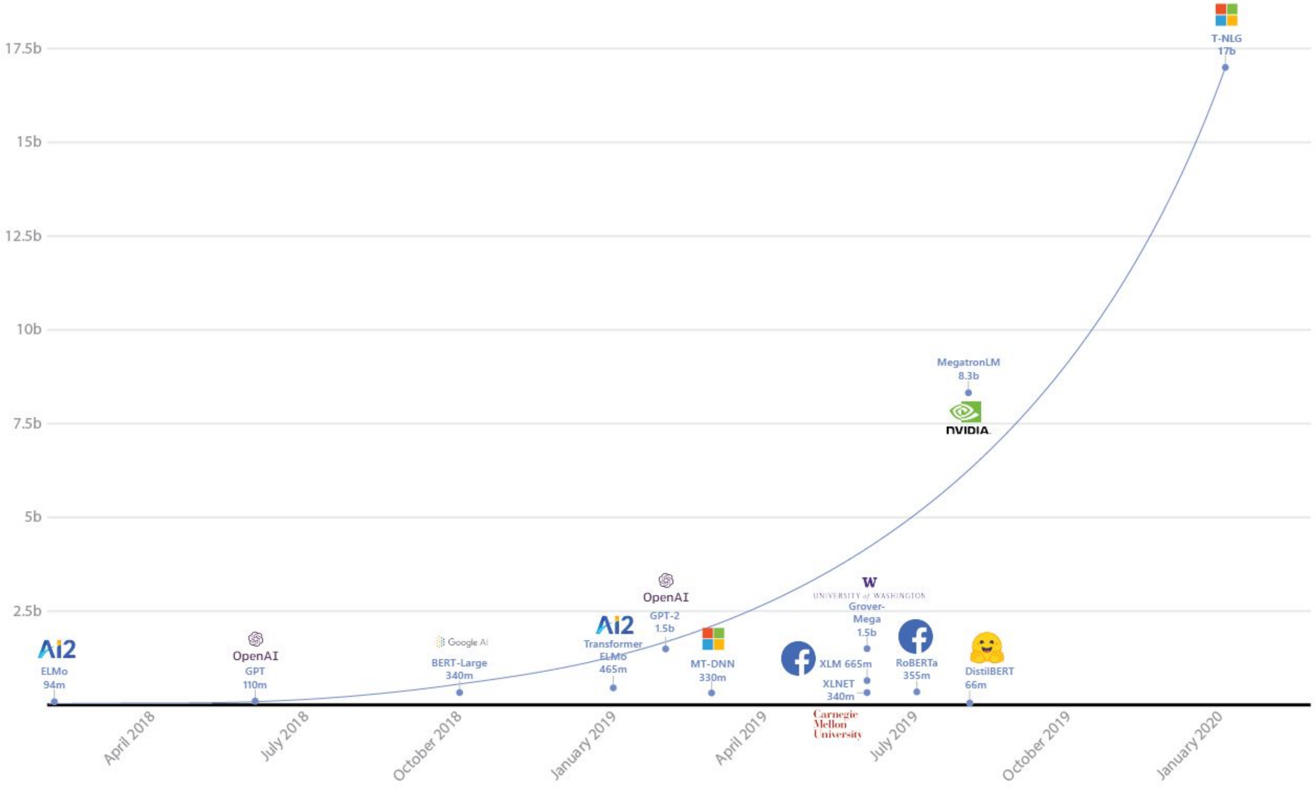
GLUE (General Language Understanding Evaluation)



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	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
BERT$_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
BERT$_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

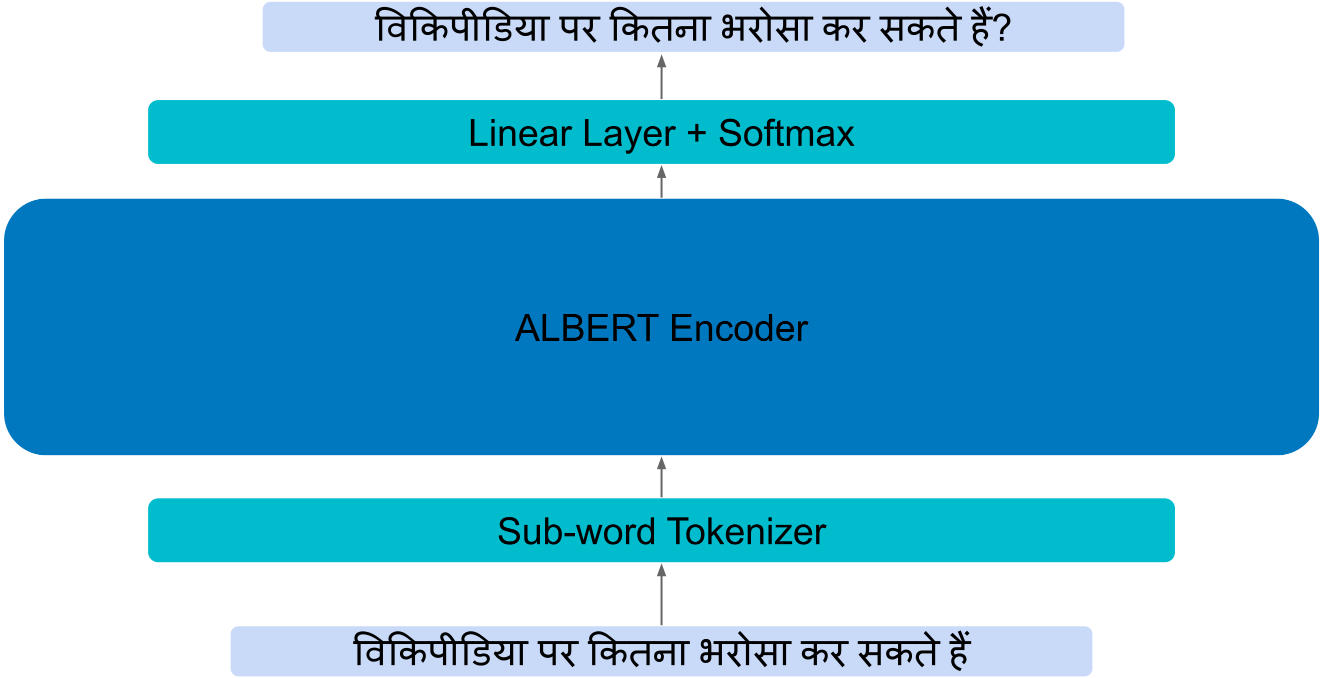
- MNLI: Multi-Genre Natural Language Inference
- QQP: Quora Question Pairs
- QNLI: Question Natural Language Inference
- SST-2: Stanford Sentiment Treebank
- CoLA: The corpus of Linguistic Acceptability
- STS-B: The Semantic Textua Similarity Benchmark
- MRPC: Microsoft Research Paraphrase Corpus
- RTE: Recognizing Textual Entailment

Increasing Complexity of Models



Source : https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

ALBert Embeddings instead of seq2seq (sequence labelling)



विकिपीडिया पर कितना भरोसा कर सकते हैं?

Linear Layer + Softmax

ALBERT Encoder

Sub-word Tokenizer

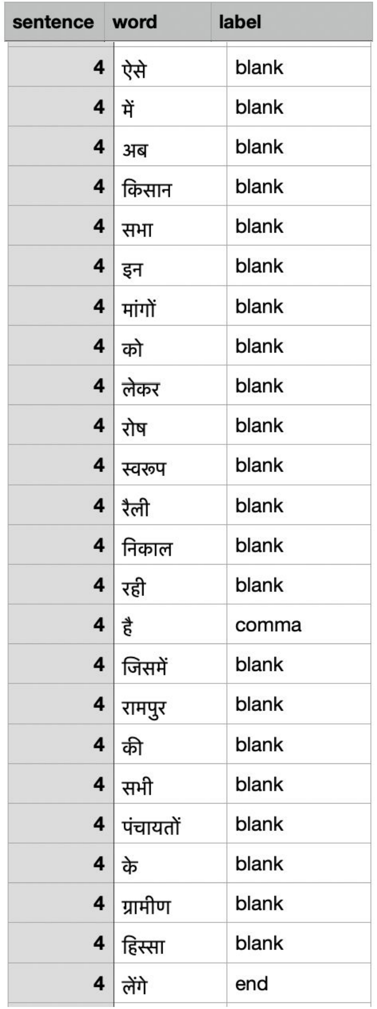
विकिपीडिया पर कितना भरोसा कर सकते हैं

Sequence labelling

Original training sentence:

ऐसे में अब किसान सभा इन मांगों को लेकर रोष स्वरूप रैली निकाल रही है , जिसमें रामपुर की सभी पंचायतों के ग्रामीण हिस्सा लेंगे ।

```
labels_map = {
    ',' :  'comma',
    '?' :  'qm',
    '!' :  'ex',
    '-' :  'hyp',
    '।' :  'end'
}
```



sentence	word	label
4	ऐसे	blank
4	में	blank
4	अब	blank
4	किसान	blank
4	सभा	blank
4	इन	blank
4	मांगों	blank
4	को	blank
4	लेकर	blank
4	रोष	blank
4	स्वरूप	blank
4	रैली	blank
4	निकाल	blank
4	रही	blank
4	है	comma
4	जिसमें	blank
4	रामपुर	blank
4	की	blank
4	सभी	blank
4	पंचायतों	blank
4	के	blank
4	ग्रामीण	blank
4	हिस्सा	blank
4	लेंगे	end

Sequence labelling with ALBERT

- Indic Bert by AI4Bharat is actually on ALBERT architecture instead of BERT.
- It is trained on 12 major Indian languages.

Inputs :
Eg. Today is a sunny day.
Tokenized input : To## ##day is a sun## ##y day

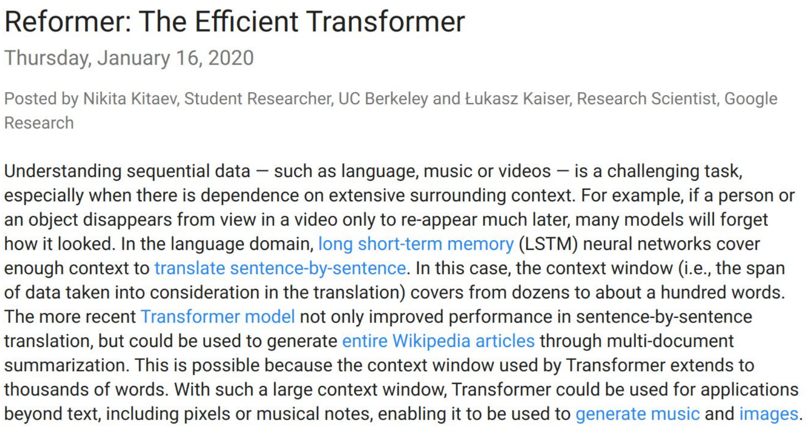
Outputs:

We make the model to predict a token after each word in a sentence and put a fully connected layer on top of it with the required number of labels (punctuation marks).

Sequence labelling with ALBERT

	Recall	Precision	F1 Score
blank	0.99	0.98	0.99
Comma (,)	0.75	0.83	0.78
End (|)	0.91	0.89	0.90
Hyphen (-)	0.57	0.69	0.63
Ex (!)	0.08	0.44	0.14
Qm (?)	0.57	0.58	0.57

CURRENT STATE OF AFFAIRS

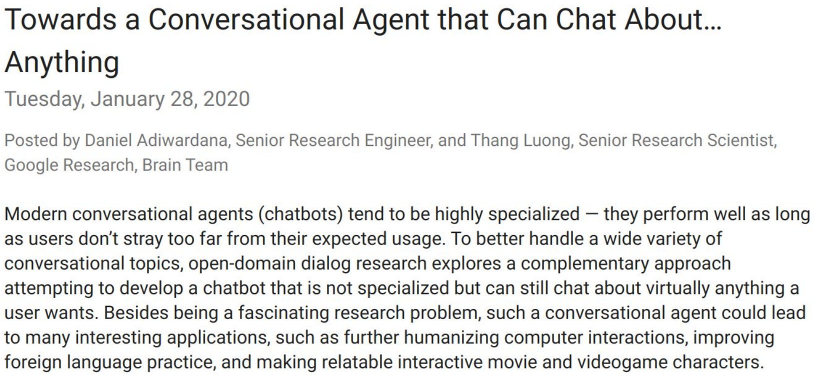


Reformer: The Efficient Transformer

Thursday, January 16, 2020

Posted by Nikita Kitaev, Student Researcher, UC Berkeley and Łukasz Kaiser, Research Scientist, Google Research

Understanding sequential data — such as language, music or videos — is a challenging task, especially when there is dependence on extensive surrounding context. For example, if a person or an object disappears from view in a video only to re-appear much later, many models will forget how it looked. In the language domain, long short-term memory (LSTM) neural networks cover enough context to translate sentence-by-sentence. In this case, the context window (i.e., the span of data taken into consideration in the translation) covers from dozens to about a hundred words. The more recent Transformer model not only improved performance in sentence-by-sentence translation, but could be used to generate entire Wikipedia articles through multi-document summarization. This is possible because the context window used by Transformer extends to thousands of words. With such a large context window, Transformer could be used for applications beyond text, including pixels or musical notes, enabling it to be used to generate music and images.

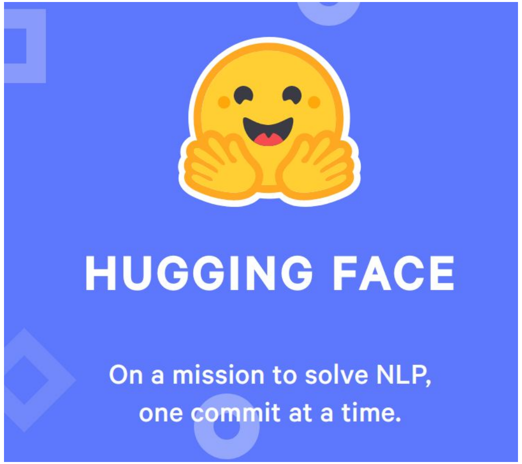


Towards a Conversational Agent that Can Chat About… Anything

Tuesday, January 28, 2020

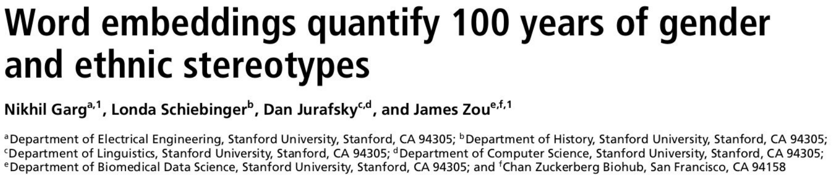
Posted by Daniel Adiwardana, Senior Research Engineer, and Thang Luong, Senior Research Scientist, Google Research, Brain Team

Modern conversational agents (chatbots) tend to be highly specialized — they perform well as long as users don't stray too far from their expected usage. To better handle a wide variety of conversational topics, open-domain dialog research explores a complementary approach attempting to develop a chatbot that is not specialized but can still chat about virtually anything a user wants. Besides being a fascinating research problem, such a conversational agent could lead to many interesting applications, such as further humanizing computer interactions, improving foreign language practice, and making relatable interactive movie and videogame characters.



CHALLENGES AND FUTURE OF DEEP NLP

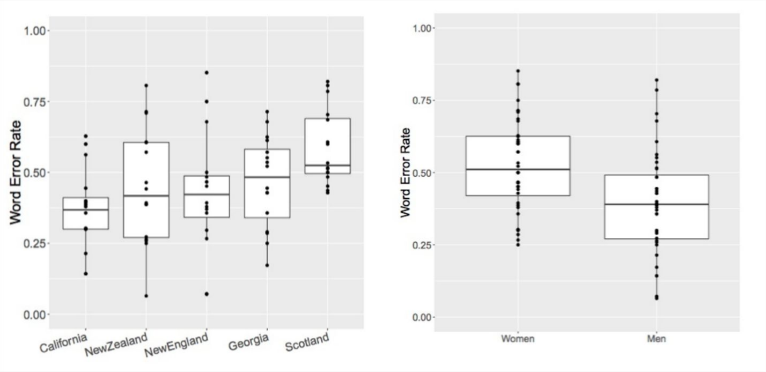
- Very diferent from Chomsky's school of linguistics. Skepticism among the community.
- Lack of interpretablity of model and lack of theoretical foundations.
- Requirement of large amount of data and powerful computing resources.
- We are not using linguistic, lexical and word knowledge in our systems yet.
- Inherent bias in data collection.
- With more data and more computational power these models are only going to get better.



Word embeddings quantify 100 years of gender and ethnic stereotypes

Nikhil Garg[a,1], Londa Schiebinger[b], Dan Jurafsky[c,d], and James Zou[e,f,1]

[a]Department of Electrical Engineering, Stanford University, Stanford, CA 94305; [b]Department of History, Stanford University, Stanford, CA 94305; [c]Department of Linguistics, Stanford University, Stanford, CA 94305; [d]Department of Computer Science, Stanford University, Stanford, CA 94305; [e]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305; and [f]Chan Zuckerberg Biohub, San Francisco, CA 94158



Details: Rachael Tatman, Gender and Dialect Bias in YouTube's Automatic Captions (2017)

Thank You!