Deep Learning for NLP with PyTorch

Oct 29, 2020

Topics

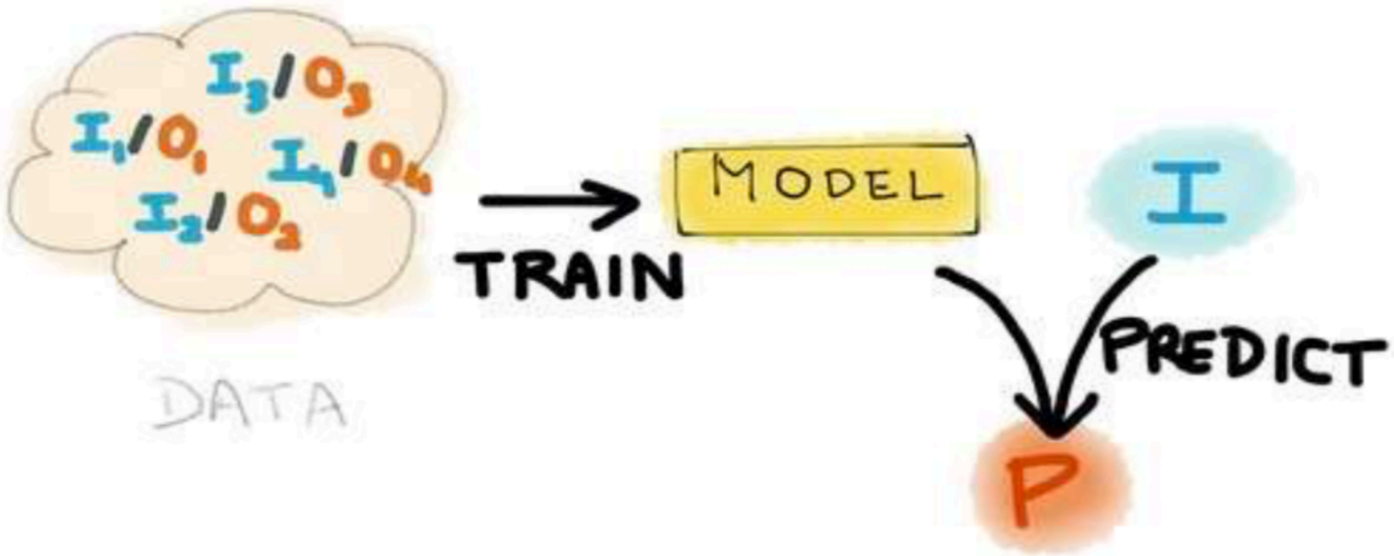
- AI and NLP Introduction
- NLP Tools and Pipeline
    - Word vectors, Tranformers, BERT
- Text Classification, Text Summarization
    - Labs using Pytorch, SpaCy, Gensim, BERT/XLNet

Bio

- Ravi Ilango
- Lead Data Scientist, Stealth Startup, SantaClara
- 10+ Years at Apple, Sr Data Scientist at FogHorn Systems, Sr DataScietist at StatesTitle
- Education:
    - BE Mech (Madras University, India)
    - Masters Program in Aero and Production (IIT Madras, India)
    - MBA (Santa Clara University)
    - Graduate Certificate in Data Mining and Machine Learning (Stanford)

What is machine learning

- Subfield of computer science that explores how machines can learn to perform certain task without explicit programming

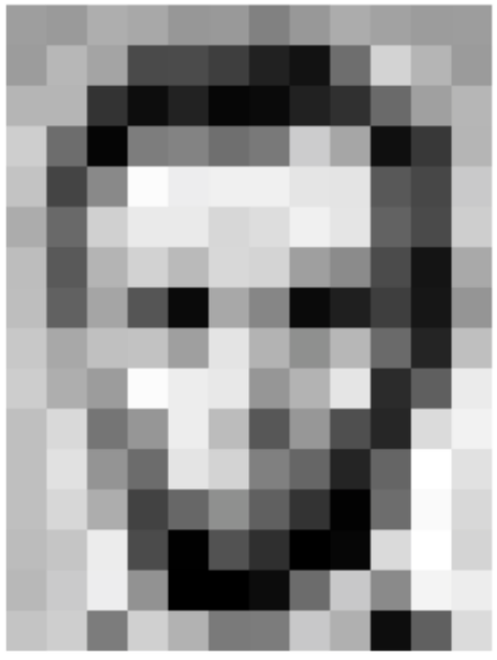


Supervised

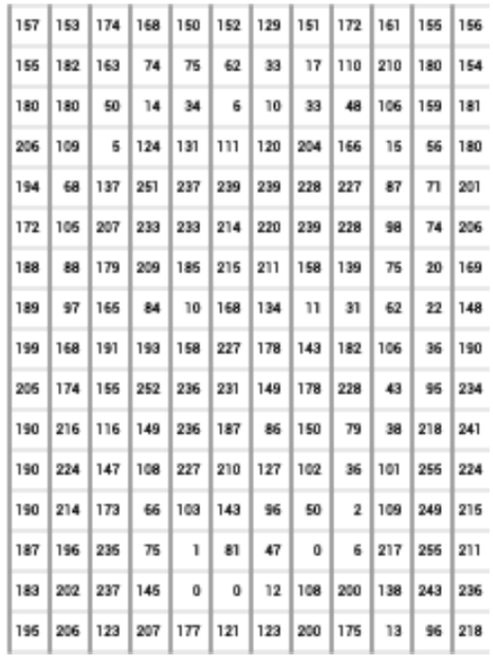

Monday
stock prices
supervised
learning
Tuesday
stock prices


what we
know
supervised
learning
what we want
to know

Tasks in Computer Vision



Input Image



157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
196	206	123	207	177	121	123	200	175	13	96	218

Pixel Representation

classification

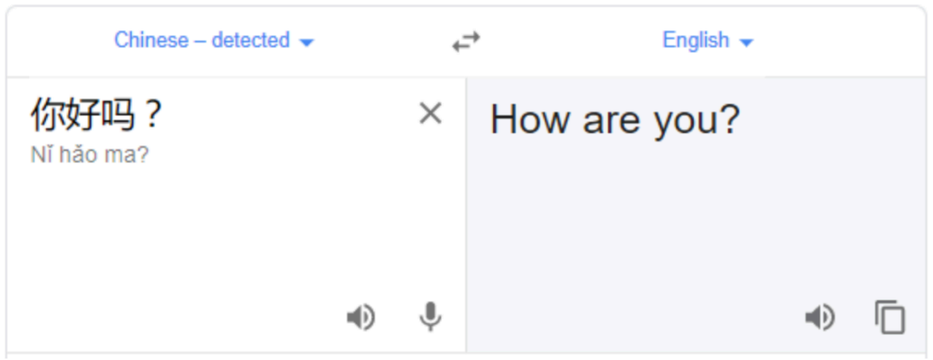
Lincoln	0.8
Washington	0.1
Jefferson	0.05
Obama	0.05

- **Regression**: output variable takes continuous value
- **Classification**: output variable takes class label. Can produce probability of belonging to a particular class

NLP Everyday

Google Translate

used by 500 million people every day



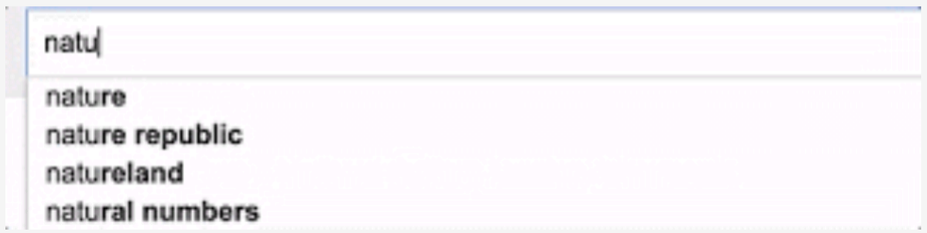
Google Translate from Chinese to English

Chatbot

*Our Financial Consultants use the ask**PRU** chatbot to provide a better service experience to customers.*

Email assistant

Predictive typing suggests the next word in the sentence.



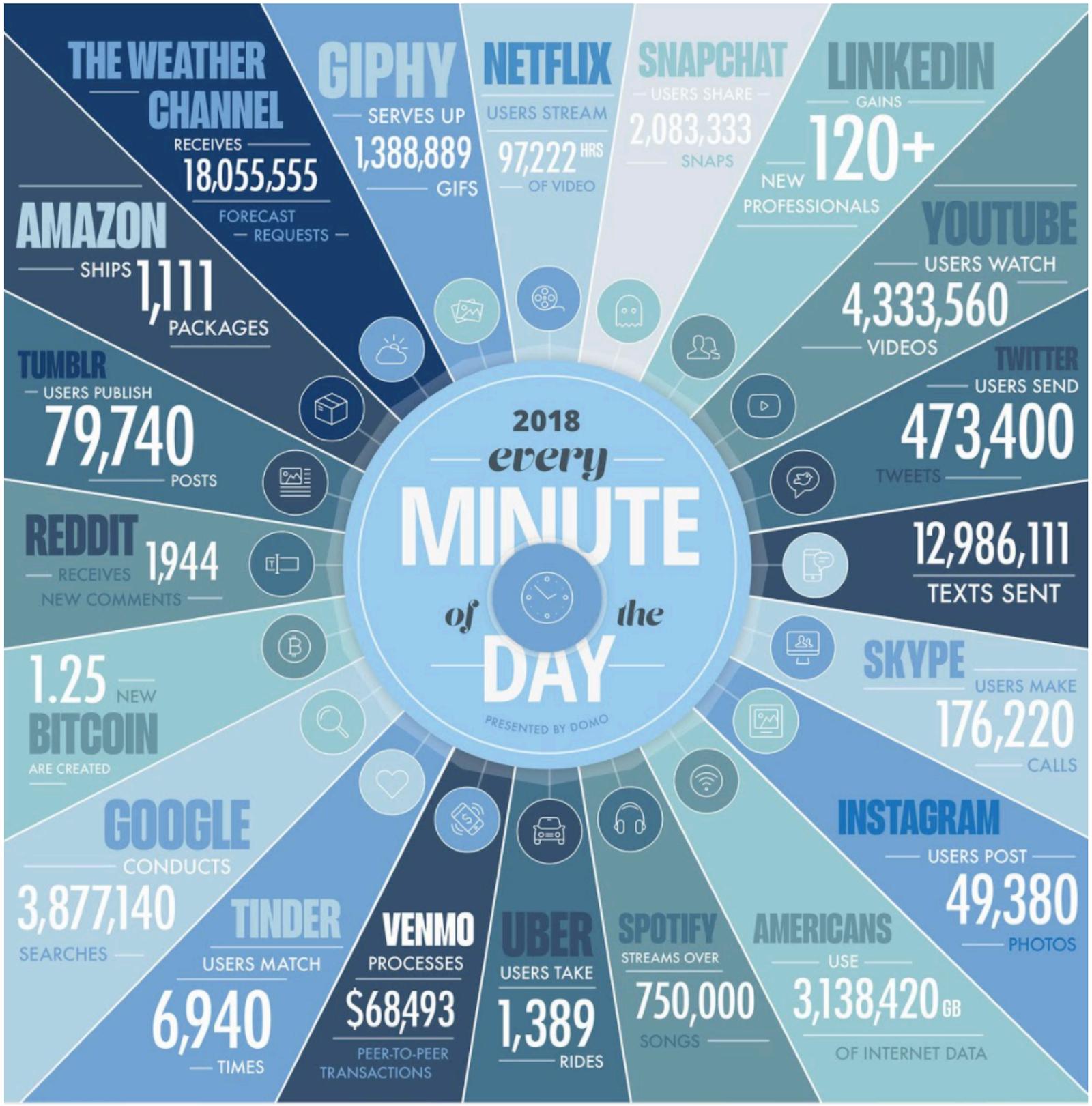
3.5B searches/daily

NLP in business operation

- Achieve efficiencies using machine processing of unstructured tect data

Extract and Summarization

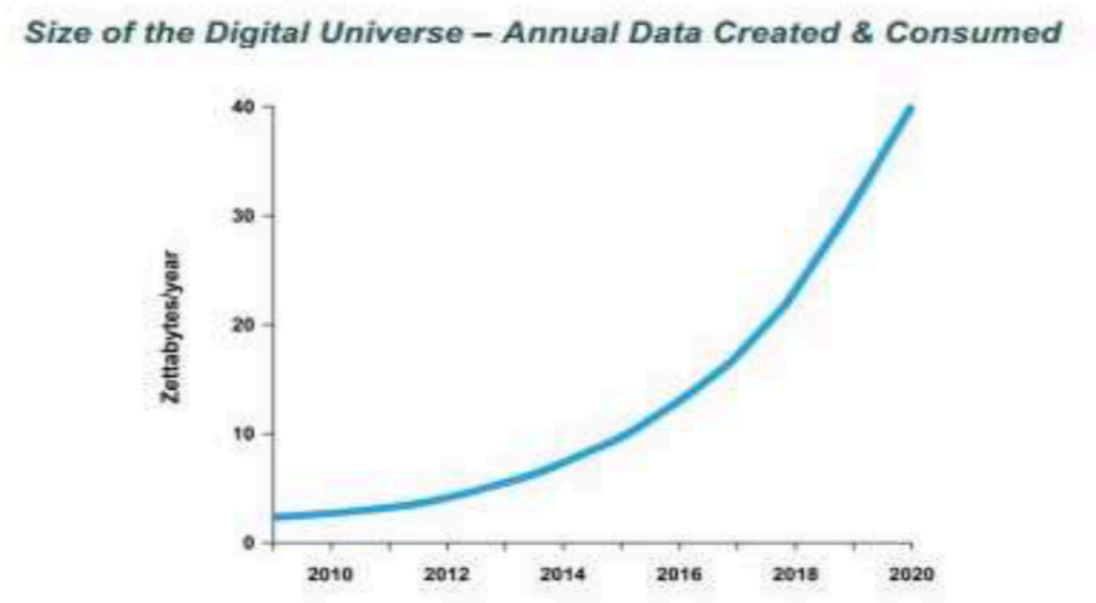
Data production rate



https://www.digitalinformationworld.com/2018/06/infographics-data-never-sleeps-6.html

Text data

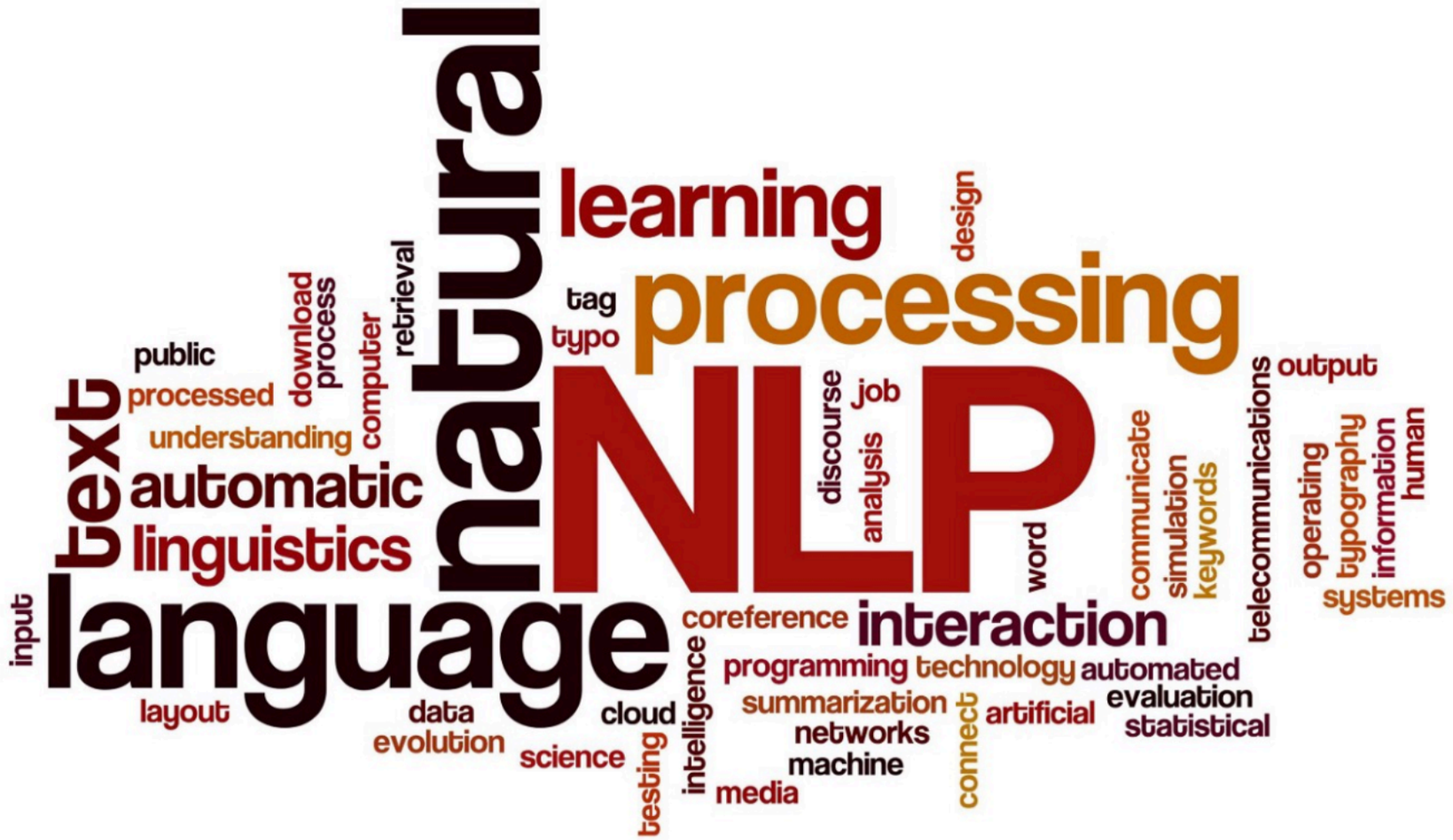
- About 80% of data in organizations are in text format
- Harder to analyse than structured data
- Huge amount of textual documents
  - Only in biomedicine 2200 scientific papers are published every day
- Growing exponentially



Natural Language Processing

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

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


natural
learning
processing
NLP
language
text
automatic
linguistics
interaction
public
processed
understanding
retrieval
tag
typo
design
output
discourse
job
analysis
word
communicate
simulation
keywords
telecommunications
operating
typography
information
human
systems
input
coreference
programming
technology
automated
evaluation
statistical
summarization
artificial
networks
machine
media
connect
intelligence
testing
science
evolution
data
cloud
layout

NLP Use Cases in Enterprises

- Text Classification- It is the process of categorizing or tagging text based on its content
  - organizing customer support tickets, customer reviews, emails

- Text Extraction- It is the process of extracting specific information from documents
  - extract key info from business documents (invoices, agreements), emails

Spam Detection

Spam Detection

- data from 4601 emails sent to an individual (named George, at HP labs, before 2000). Each is labeled as *spam* or *email*.
- goal: build a customized spam filter.
- input features: relative frequencies of 57 of the most commonly occurring words and punctuation marks in these email messages.

	george	you	hp	free	!	edu	remove
spam	0.00	2.26	0.02	0.52	0.51	0.01	0.28
email	1.27	1.27	0.90	0.07	0.11	0.29	0.01

Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between spam *and* email.

Approaches

- Rule based
  - Human defined rules to extract information
  - Needs expert humans who know how people express certain things
  - Is quite laborious
- Machine learning based
  - Machine tries to learn what to extract guided by human
  - Needs annotated corpora (usually fairly large)
    - This is expensive to create and quite laborious

Why is language interpretation hard?

Consider the following sententces

“I enjoy working in a bank.”

“I enjoy working near a river bank.”

Understanding language is tough

- Syntax - grammer
- Semantics - meaning
- Pragmatics - what the text is trying to achieve

Ambiguity

Resolving ambiguity is hard

Ambiguity

Find at least 6 meanings of this sentence:

```
I made her duck
```

Ambiguity

Find at least 6 meanings of this sentence:

`I made her duck`

I cooked waterfowl for her benefit (to eat)

I cooked waterfowl belonging to her

I created the (plaster?) waterfowl she owns

I caused her to quickly lower her head or body

I recognized the true identity of her spy waterfowl

I waved my magic wand and turned her into undifferentiated waterfowl

Ambiguity is Pervasive

I caused her to quickly lower her head or body

Part of speech: “duck” can be a Noun or Verb

I cooked waterfowl belonging to her.

Part of speech:

“her” is possive pronoun (“of her”)

“her” is dative pronoun (“for her”)

I made the (plaster) duck statue she owns

Word Meaning : “make” can mean “create” or “cook”

Ambiguity is Pervasive: Phonetics!!!!!

Aye mate, her duck

I mate or duck

I'm eight or duck

Eye maid; her duck

I maid her duck

I'm aid her duck

I mate her duck

I'm ate her duck

I'm ate or duck

I mate or duck

Extracting Sentiment and Social Meaning

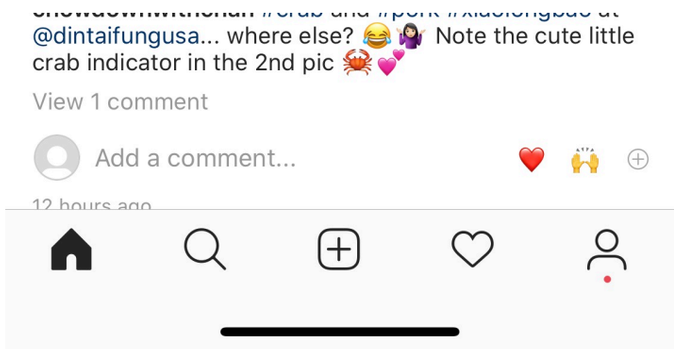
Lots of meaning is in **connotation**

"connotation: an idea or feeling that a word invokes in addition to its literal or primary meaning."

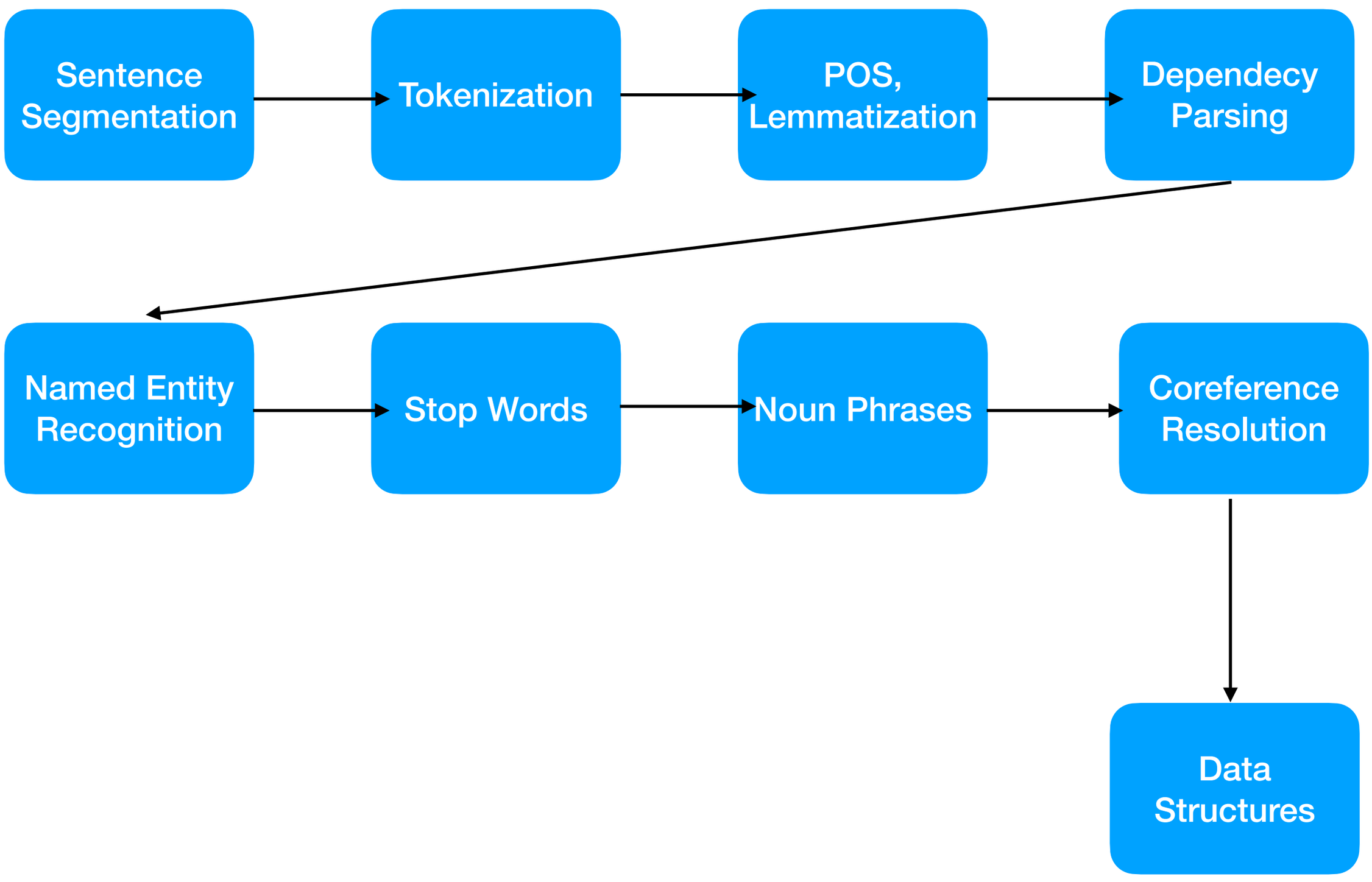
Extracting connotation is generally called **sentiment analysis**

More difficulties: Non-standard language, emojis, hashtags, names

chowdownwithchan #crab and #pork #xiaolongbao at @dintaifungusa... where else? 😂🤷‍♀️ Note the cute little crab indicator in the 2nd pic 🦀💕



NLP Pipeline to produce representative data structures



NLP Data Scientist - (some) tools

python™

Transformers

PyTorch

scikit learn

NLTK

TensorFlow

spaCy

CoreNLP

PyText

gensim

NLP Data Scientist - (some) concepts & techniques

- Word vectors, tokenization, document vecetorization
- POS (parts of speech), NER
- Document parsing techniques, Regex
- Supervised ML, Classifiers, Deep Learning and how to use word embeddings
- Using Pretrained models

Word Embedding

"You shall know a word by the company it keeps"
(J. R. Firth 1957)

One of the most successful ideas of modern statistical NLP!

government debt problems turning into **banking** crises as has happened in

saying that Europe needs unified **banking** regulation to replace the hodgepodge

these words represent banking

Vector Space Model

- Neural word embeddings
- Combine vector space semantics with the prediction of probabilistic models
- Words are represented as a **dense** vector:

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

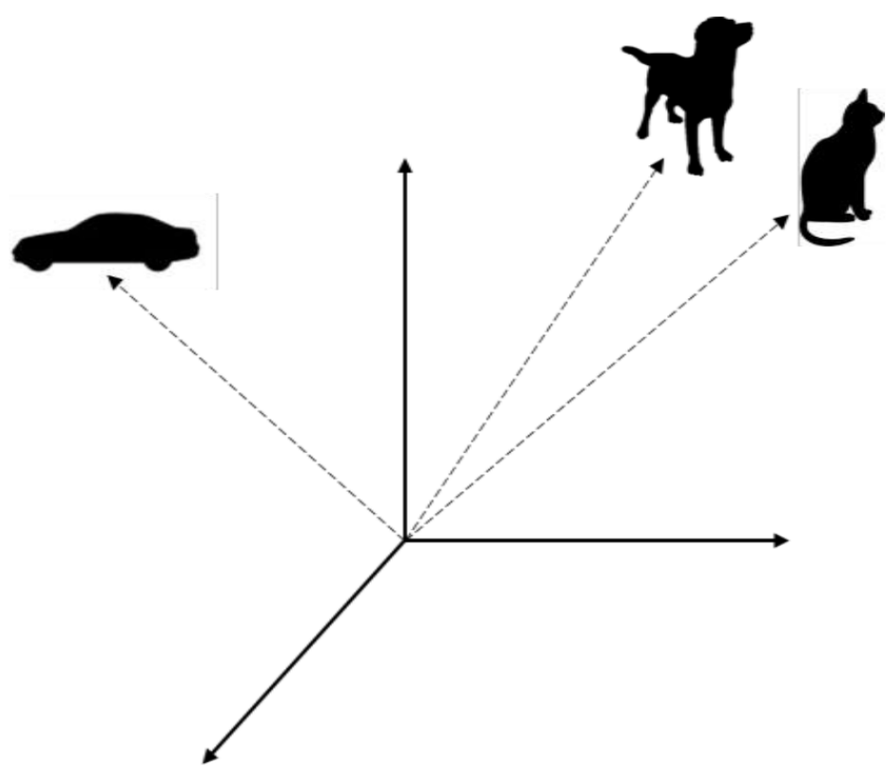
Vector Space Model

Embeddings

The vectors we have been discussing so far are very high-dimensional (thousands, or even millions) and sparse.

But there are techniques to learn lower-dimensional dense vectors for words using the same intuitions.

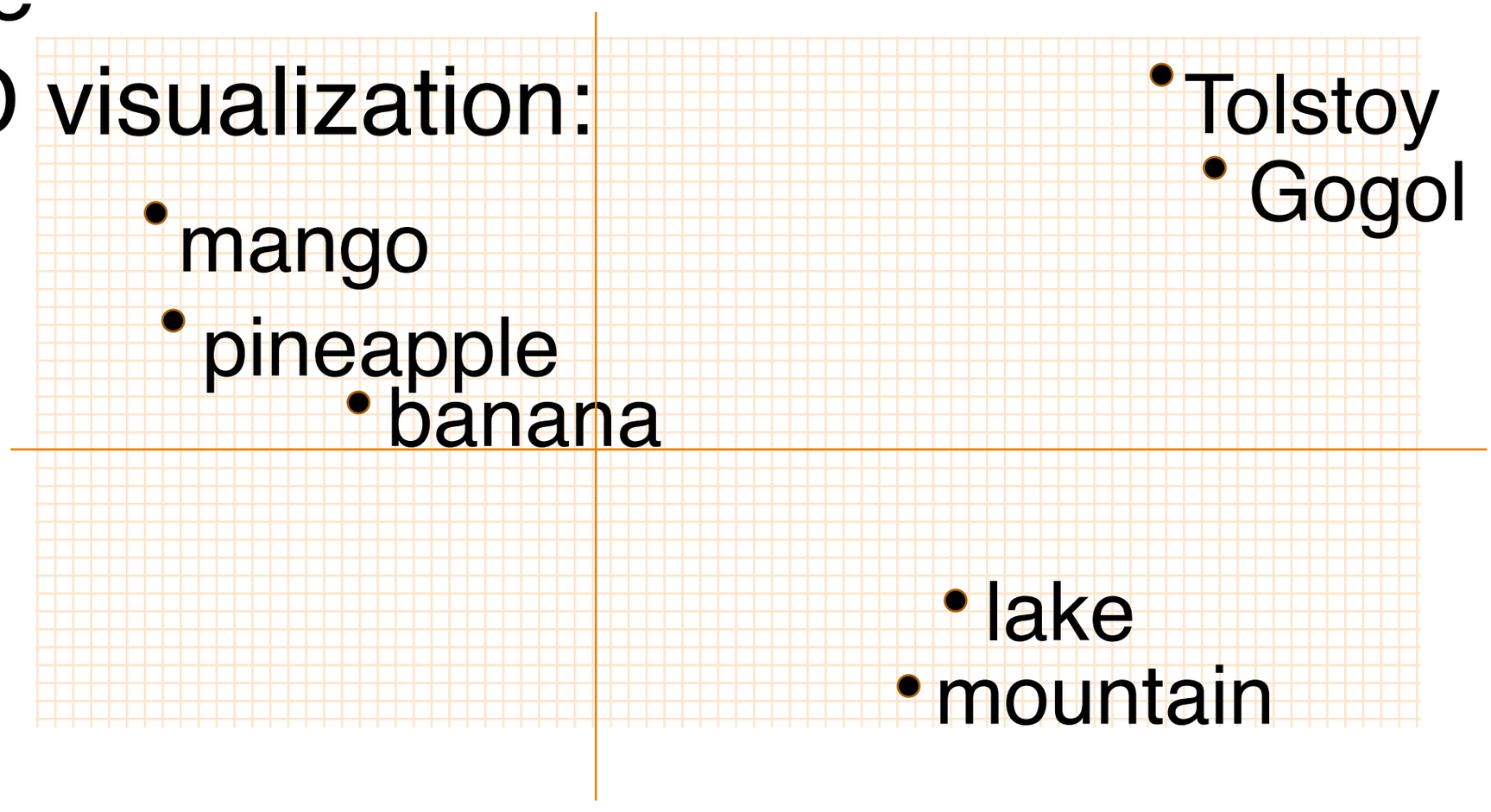
These dense vectors are called embeddings.



Neural word embeddings

A word's meaning is a point in 300-dimensional space

A 2-D visualization:



Embeddings are the core of NLP

Core technology for any NLP task (question answering, machine translation, information retrieval, etc)

- Finding synonyms for words
- Deciding if two sentences have similar meaning

How to learn these "embeddings"?

Push words **together in space** they occur **together in text**

Read millions of words.

When you see:

Banana, mango, or pineapple are all delicious…

Move **banana** closer to **mango**

Move **banana** further from **Tolstoy**

Problem: Embeddings reflect cultural bias!

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings." In *NeurIPS 2016*, pp. 4349-4357.

Ask “Paris : France :: Tokyo : x”

- x = Japan

Ask “father : doctor :: mother : x”

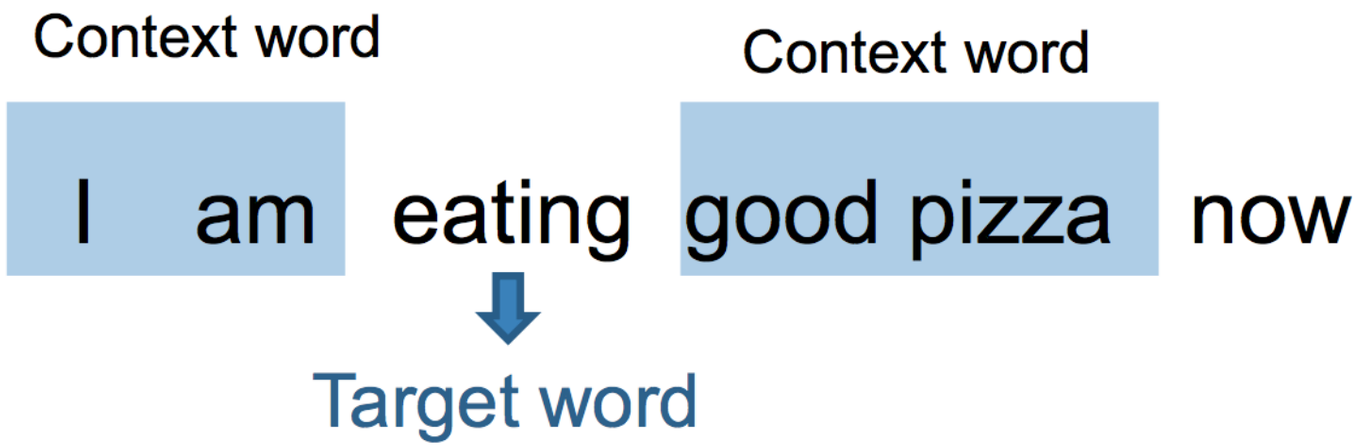
- x = nurse

Ask “man : computer programmer :: woman : x

- x = homemaker

Continuous Bag-of-Words Model

- Continuous Bag-of-Words Model
  - Predict target word by the context words
  - Eg: Given a sentence and window size 2

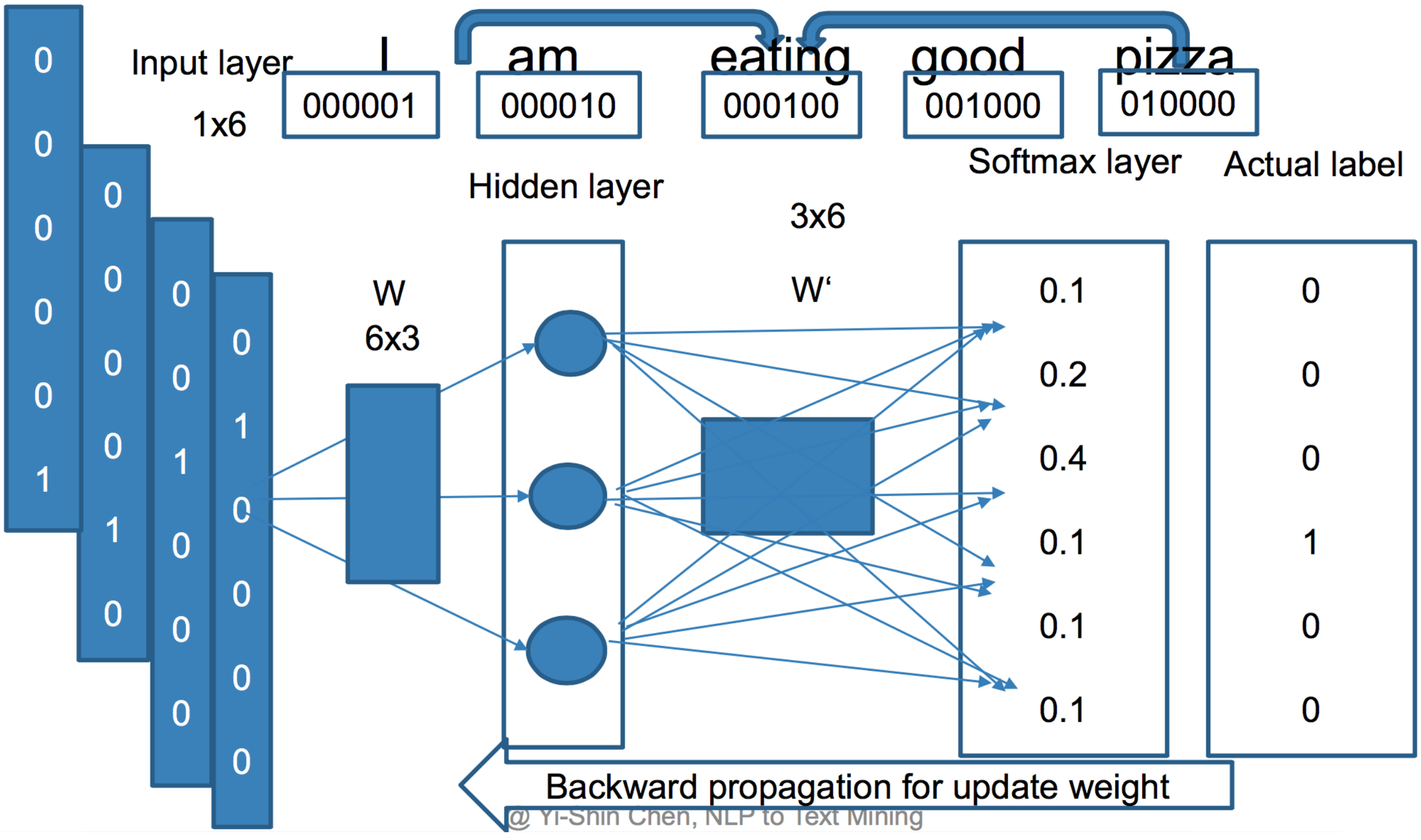


Ex: ([features], label)
([I, am, good, pizza], eating), ([am, eating, pizza, now], good) and so on and so forth.

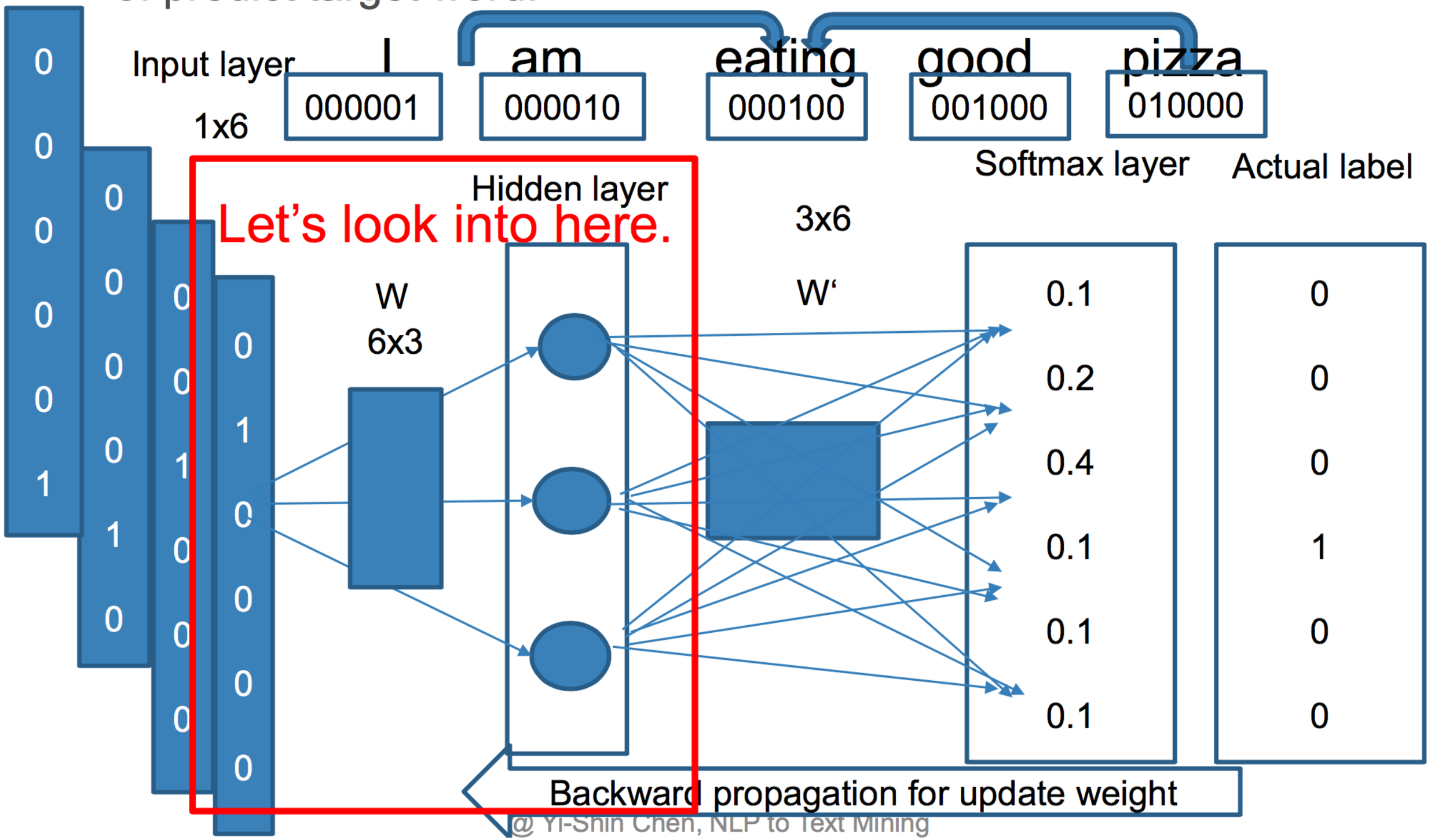
Word Embedding in a Deep Learning network

Continuous Bag-of-Words Model (Contd.)

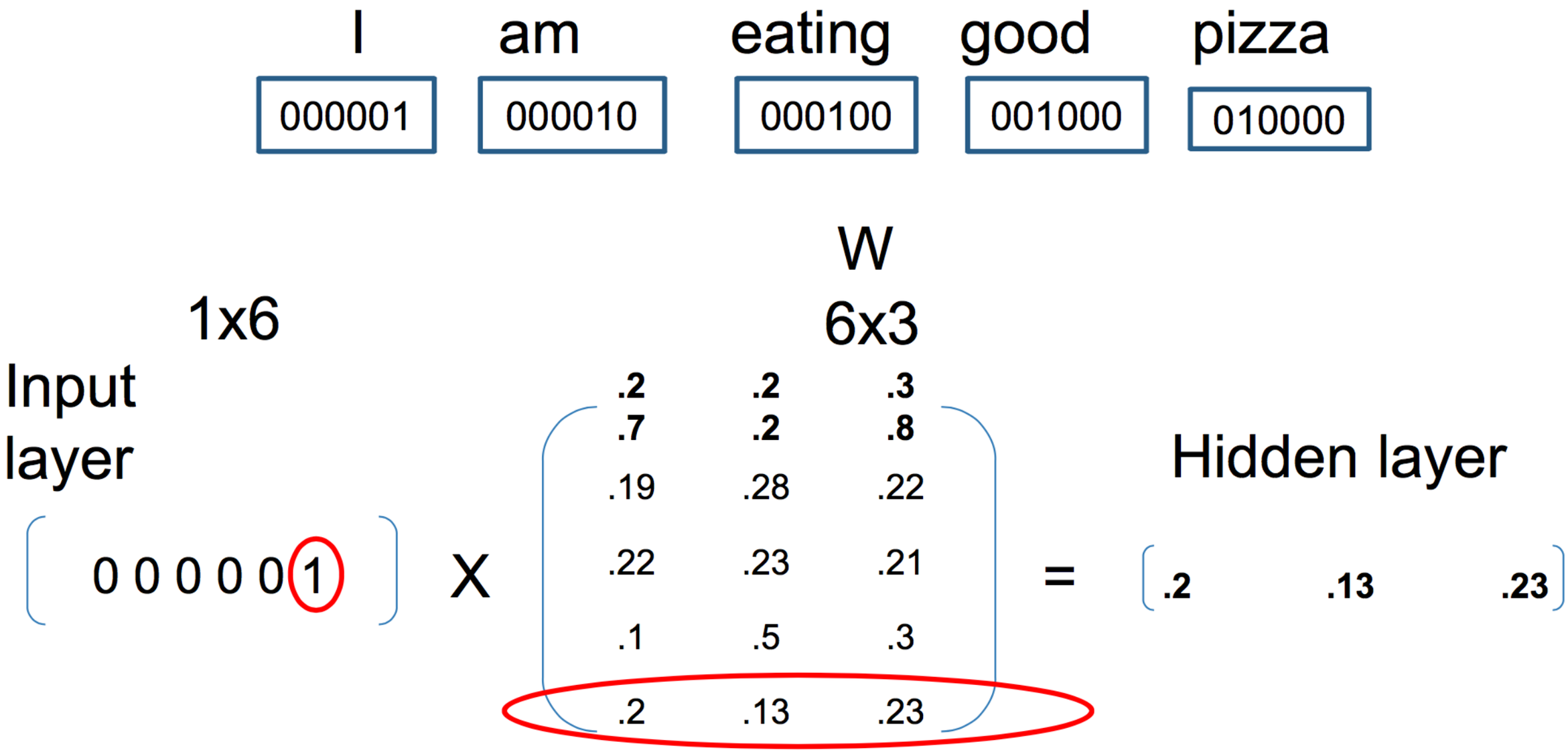
- We train one depth neural network, set hidden layer's width 3.



- The goal is to extract word representation vector instead of prob of predict target word.



Continuous Bag-of-Words Model (Contd.)



- The output of the hidden layer is just the "word vector" for the input word

Lab 1: Parts Of Speech

Python, SpaCy

Lab 2: Text Classification

Python, SkLearn, TFIDF

Lab 3: Text Classification

Python, Tokenizer, LSTM

Tranformers

- **Transformer:** It applies attention mechanisms to gather information about the relevant context of a given word, and then encode that context in a rich vector that smartly represents the word.

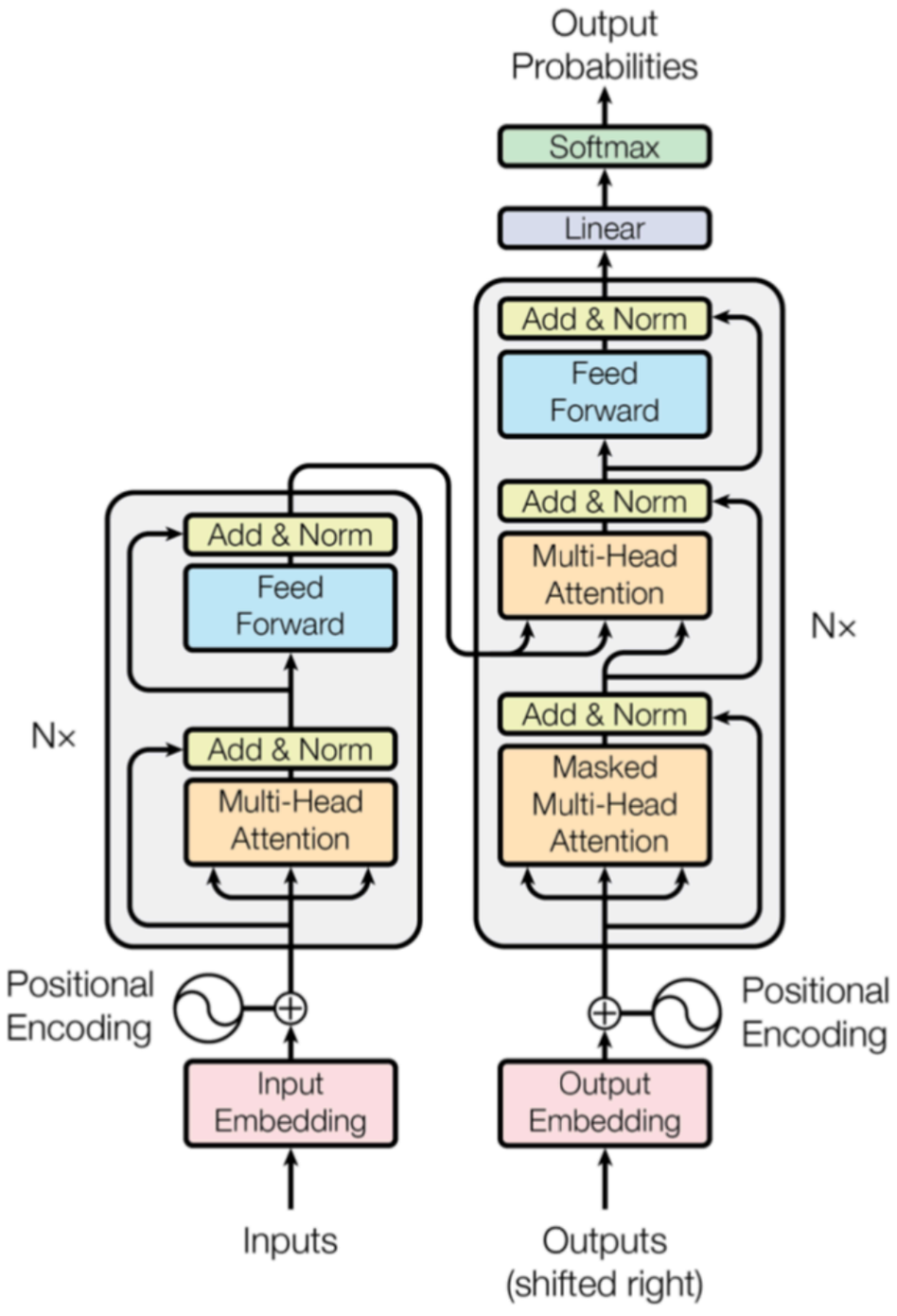


Figure 1: The Transformer - model architecture.

Source: Attention Is All You Need

- https://medium.com/analytics-vidhya/transformer-vs-rnn-and-cnn-18eeefa3602b

PyTorch-Transformers

- PyTorch-Transformers is a library of state-of-the-art pre-trained models for Natural Language Processing (NLP)

- Contains PyTorch implementations, pre-trained model weights for BERT (Google), GPT (OpenAI), Transformer-XL, XLNet, XML (Facebook) and more …

Transfer Learning

In practice, very few people train an entire Convolutional Network from scratch (with random initialization), because it is relatively rare to have a dataset of sufficient size. Instead, it is common to pretrain a ConvNet on a very large dataset (e.g. ImageNet, which contains 1.2 million images with 1000 categories), and then use the ConvNet either as an initialization or a fixed feature extractor for the task of interest

Source: http://cs231n.github.io/transfer-learning/

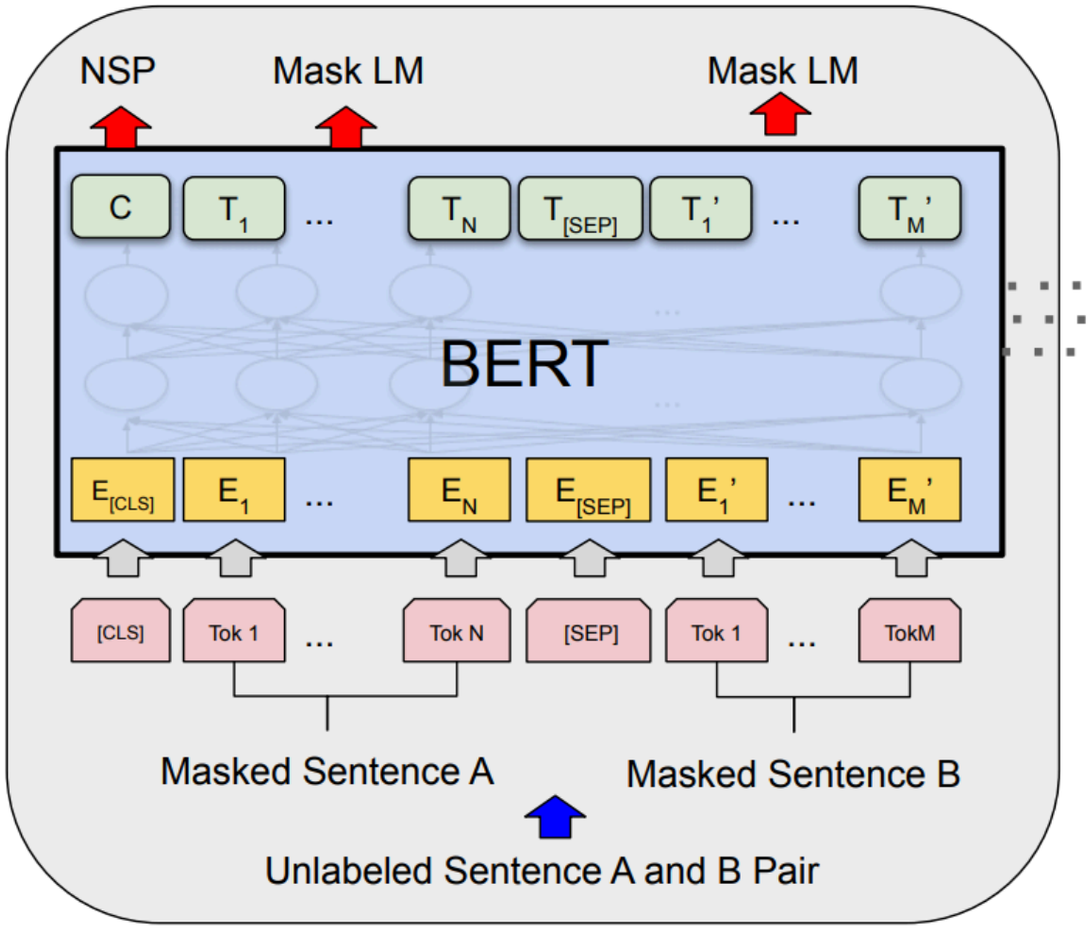
BERT - Bidirectional Encoder Representations from Transformers

Model Details

- Data: Wikipedia (2.5B words) + BookCorpus (800M words)
- Batch Size: 131,072 words (1024 sequences * 128 length or 256 sequences * 512 length)
- Training Time: 1M steps (~40 epochs)
- Optimizer: AdamW, 1e-4 learning rate, linear decay
- `BERT-Base`: 12-layer, 768-hidden, 12-head
- `BERT-Large`: 24-layer, 1024-hidden, 16-head
- Trained on 4x4 or 8x8 TPU slice for 4 days

BERT - Bidirectional Encoder Representations from Transformers

- Pretraining
  - Masked LM
  - Next Sentence Prediction (NSP)
  - Books Corpus (800M words) and English Wikipedia (2,500M words).



Pre-training

Source: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Lab 4: Text Classification

Python, XLNet, SkLearn

Summarization

- Extractive summarization
  - Extract parts of text from the corpus and arrange them to form a summary
  - NLTK Summarizer, Gensim Summarizer
- Abstractive summarization
  - Generate new text using natural language generation

Summarization

- NLTK Summarizer
- Gensim Summarizer

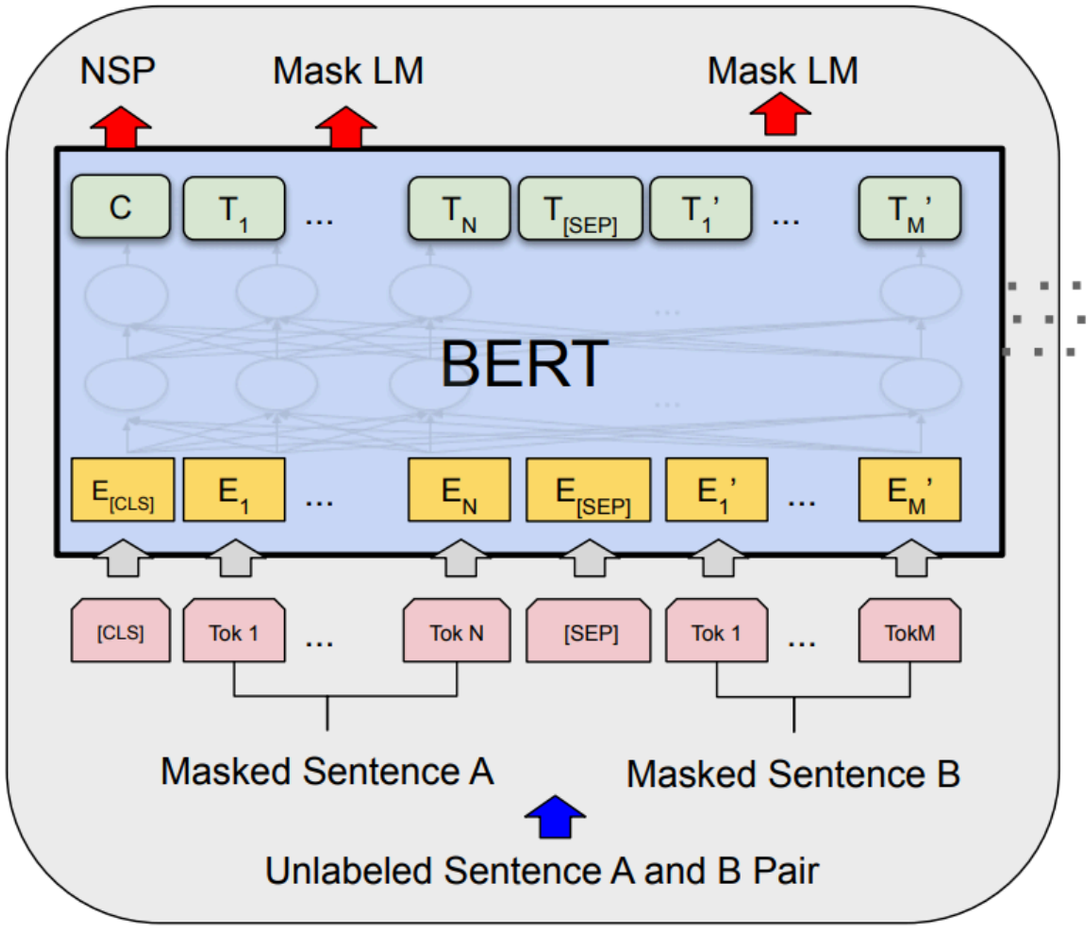
Lab 5: Text Summarization

Python, NLTK, Gensim

Additional Slides

BERT - Bidirectional Encoder Representations from Transformers

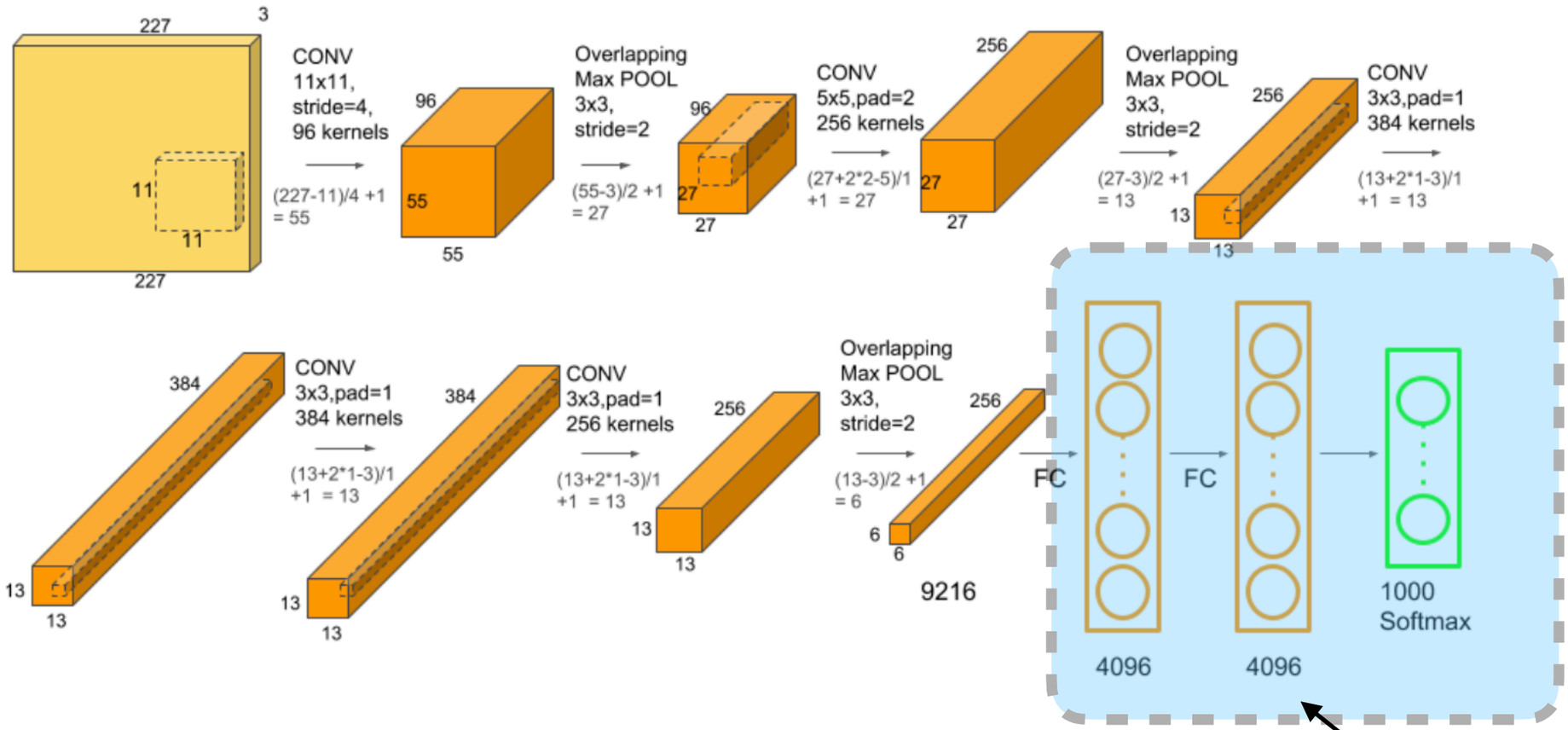
- Pretraining
  - Masked LM
  - Next Sentence Prediction (NSP)
  - Books Corpus (800M words) and English Wikipedia (2,500M words).



Pre-training

Source: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Transfer Learning - Fixed Feature Extractor



AlexNet Architecture

Train on new data

Typically small dataset, different from original dataset

- https://medium.com/analytics-vidhya/transformer-vs-rnn-and-cnn-18eeefa3602b

Sequence Model

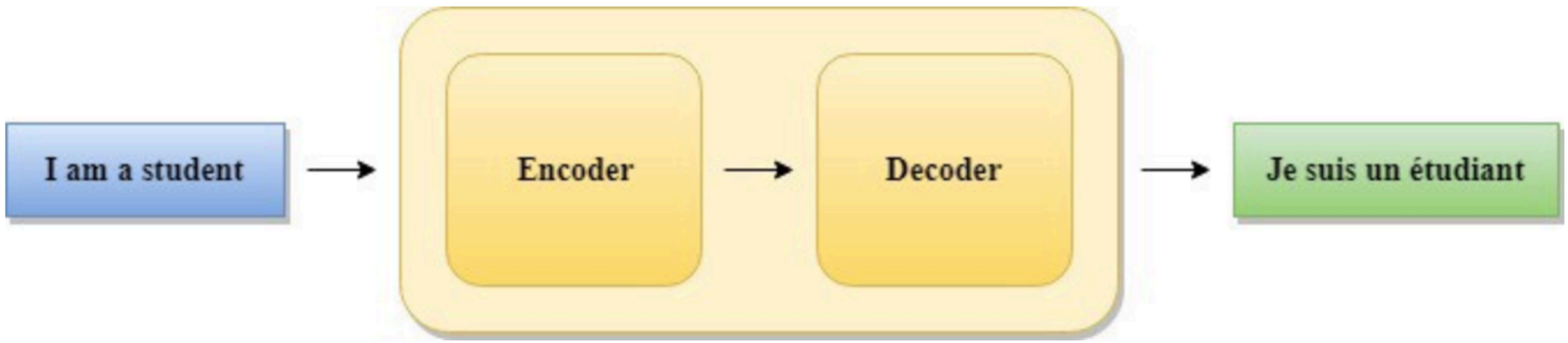


Figure: Sequence to Sequence model (Seq2Seq)

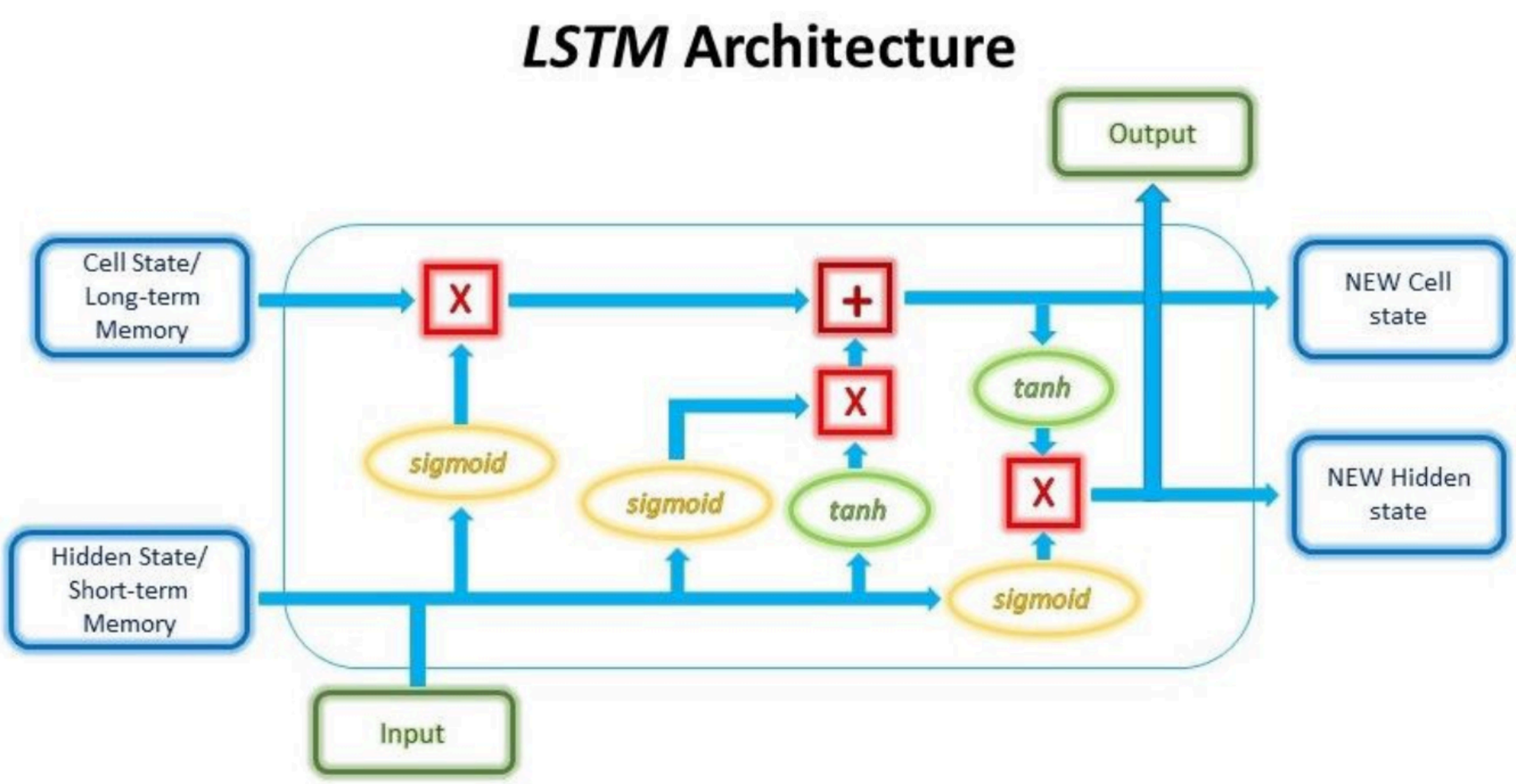
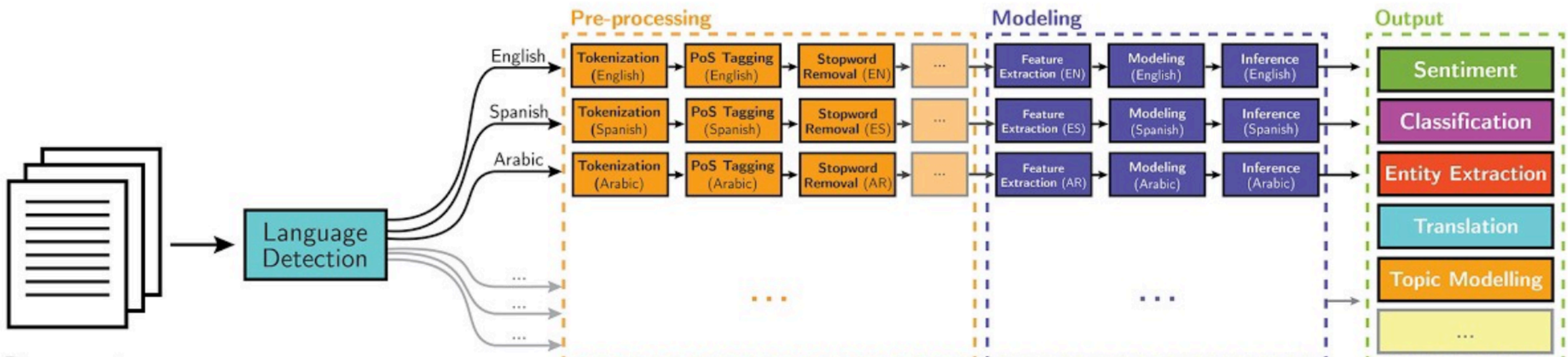
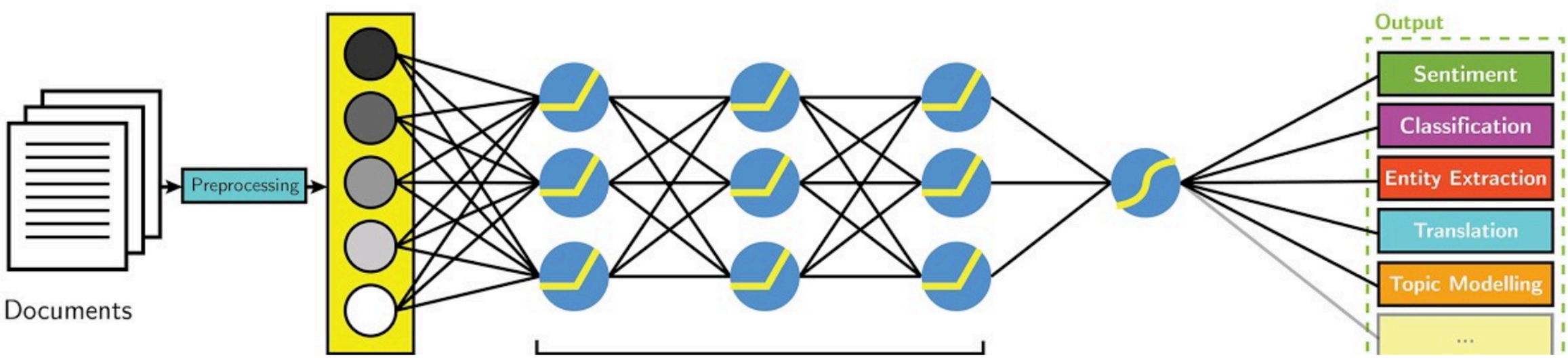


Figure: LSTM Architecture (https://blog.floydhub.com/long-short-term-memory-from-zero-to-hero-with-pytorch/)

NLP


Pre-processing
Modeling
Output
English
Spanish
Arabic
Tokenization (English)
PoS Tagging (English)
Stopword Removal (EN)
Tokenization (Spanish)
PoS Tagging (Spanish)
Stopword Removal (ES)
Tokenization (Arabic)
PoS Tagging (Arabic)
Stopword Removal (AR)
Feature Extraction (EN)
Modeling (English)
Inference (English)
Feature Extraction (ES)
Modeling (Spanish)
Inference (Spanish)
Feature Extraction (AR)
Modeling (Arabic)
Inference (Arabic)
Language Detection
Sentiment
Classification
Entity Extraction
Translation
Topic Modelling
Documents


Deep Learning-based NLP
Preprocessing
Output
Sentiment
Classification
Entity Extraction
Translation
Topic Modelling
Documents