Natural Language Processing

SCHOOL OF INFOCOMM

Need to know

- **Trainers**
  - Ho Chee Wai (ho_chee_wai@rp.edu.sg)
  - Tan Poh Keam (tan_poh_keam@rp.edu.sg)

- **Assessment**
  - Written assignment

Learning Objectives

- Describe the principles, concepts and usage of NLP
- Perform text pre-processing activities
- Discuss and use suitable methods for feature engineering of text
- Discuss and use suitable approaches to analysis, classify and compare text
- Explain how deep-learning techniques are used for text generation tasks
- Consume Machine Learning services to build NLP applications
- Use Azure Bot Services to build conversation systems

Where We're Headed

When	Topics
15 Jun	NLP Introduction Regular Expression Text Pre-processing
16 Jun	Features Engineering Document Classification (supervised learning)
17 Jun	Topic Modelling (unsupervised learning) Word Embedding
18 Jun	Topic Modelling (unsupervised learning) Text Generation
19 Jun	Google Language Services

When	Topics
22 Jun	Azure Bot Services Cognitive Services
23 Jun	Extra / extended classes (if needed)
25 Jun	Extra / extended classes (if needed)

Introduction to NLP

What is Natural Language Processing (NLP)

By "natural language" we mean a language that is used for everyday communication between humans.

Two main components:

- Natural Language Understanding (NLU)
- Natural Language Generation (NLG)

NLP is an Intersection of several fields

- Computer Science
- Artificial Intelligence
- Linguistics

NLP is AI-Complete

- Requires all types of knowledge & context awareness humans possess → It's hard

Ambiguity of Language

- Synonymy – different words, same meaning

- Polysemy – same word, different meaning

- Text and speech are unstructured data

- No fixed structure – Sentence format

- No fixed schema – Grammar

- Misspell, slang, abbreviations

History of NLP

- NLP has been through (at least) 3 major eras:

  - 1950s-1980s: Linguistics Methods and Rules

  - 1980s-Now: Statistical + Machine Learning Methods

  - Now - ???: Deep Learning

- You're right near the start of a paradigm shift!

1950s - 1970s: Linguistics & Rules

- Approach focused on:

  - Linguistics: Grammar rules, sentence structure parsing

  - Handwritten Rules: Huge sets of logical (if/else) statements

  - Phase structure grammar : conversion of sentences into forms computers can understand

- Problems:

  - Too complex to maintain

  - Cannot scale

  - Cannot generalize

# 1980s - Now: Statistical / Machine Learning Methods

- Approach shifted from linguistics to data-driven

- Increasing computational power and ease to access of text

  - First web page (1991) → discussion forums, blogs, news portal

  - Digital archives

- NLP starts using statistical and probabilistic models

  - Data mining → text mining

- Generic machine learning algorithms applied to NLP tasks

  - Sentiment analysis using logistic regression

  - Language models with Markov models

Now and future: Deep Learning

- More advances in computing power with parallelization (GPU)

- Availability of large datasets becomes the norm

- Neural Network

  - Learnt word representation with finite dimensions

  - Capture semantic and relationships among words

- RNN / LSTM

  - Allows sequential processing and learning of text

  - Application into machine translation tasks and questions/answering systems

- Attention-based model

  - A way to place various degree of focus (attention) on different part of the text

  - Break-through in machine translation and text generation tasks

NLP: Speech vs Text

- Data source can refer to written text or speech

- Goal of both is the same: translate raw data (text or speech) into underlying concepts (NLU) then possibly into the other form (NLG)
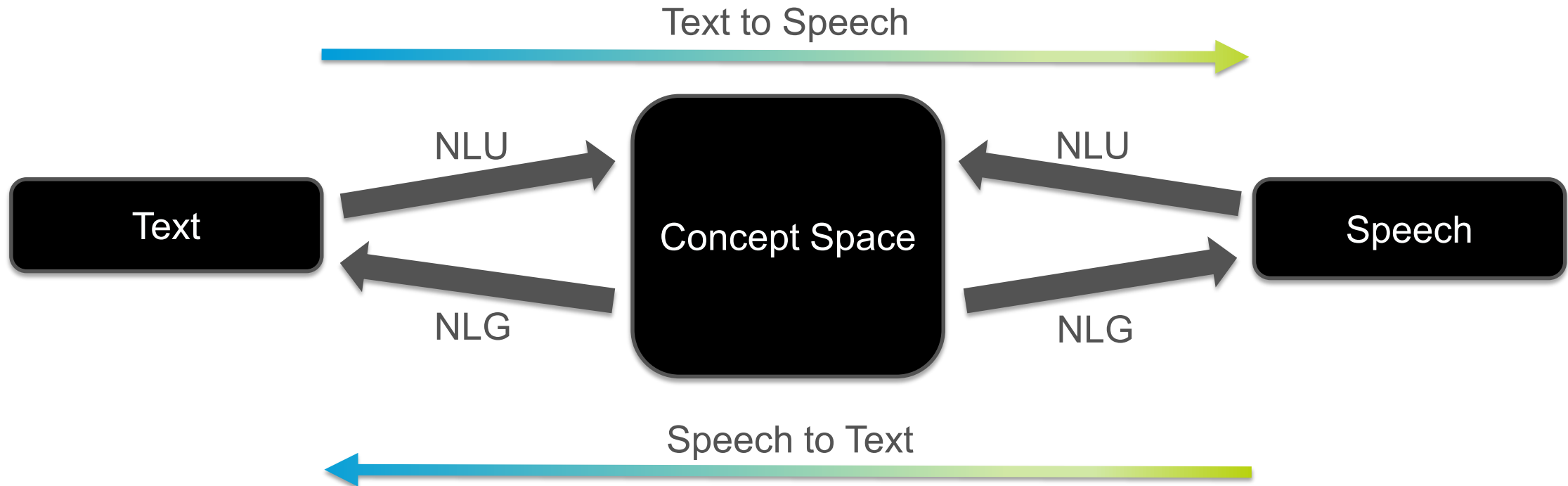
NLU Applications

- Classification

- Natural Language Search

- Document Recommendation

- Topic Modeling

- Language Identification

- Intent matching

NLU Application: Document Classification

- Classify documents (discrete collections of text) - into categories

  - Classify emails as spam vs not spam

  - Classify product reviews as positive vs negative

  - Assign labels to documents

NLU Application: Document Recommendation

- Choosing the most relevant document based on some information or "finger print":

    ▪ Show most relevant webpages based on query to search engine

    ▪ Recommend news articles based on past articles liked or read

    ▪ Recommend restaurants based on restaurant reviews

NLU Application: Topic Modeling

Breaking a set of documents into topics at the word level

- See how prevalence of certain topics covered in a magazine changes over time

- Find documents belonging to a certain topic

Election 2015: Parties row over GP out-of-hours cover

Politics or Health?

Chelsea FC reports a record £18m in annual profit

Business or Sport?

NLU Application: Intent Matching

Understanding that there are many ways to say or ask for the same thing

- Use in dialog systems

May I know the opening time for the store?

Can I come to shop now?

What time do you close?

Is the mall still open?

I can come at 8 pm, is it okay ?
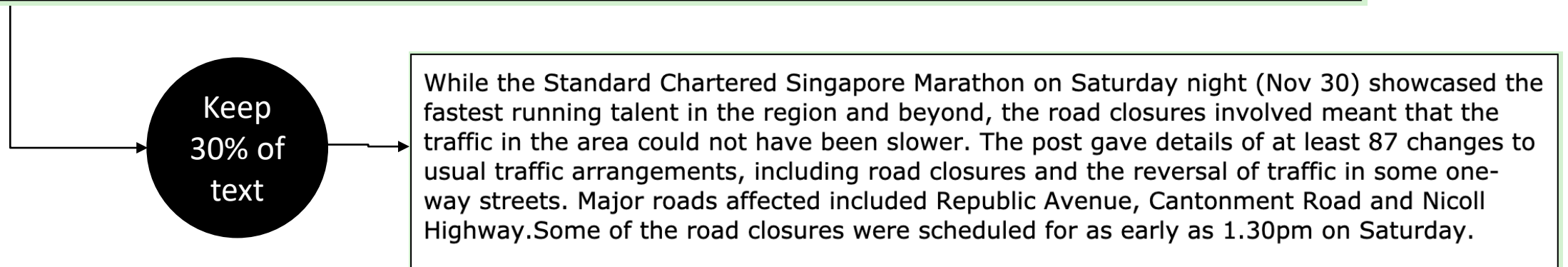
NLG Applications

- Machine Translation

- Document Summarization

- Text Generation

- Question Answering

- Image Captioning

Notice NLU is almost a prerequisite for NLG

NLG Application: Document Summarization

Read less

While the Standard Chartered Singapore Marathon on Saturday night (Nov 30) showcased the fastest running talent in the region and beyond, the road closures involved meant that the traffic in the area could not have been slower. Unlike the past 17 times when the annual event was flagged off in the morning and roads were less busy, Saturday's marathon started at 6pm as organisers sought to raise the event's profile. The changed timing and heavy traffic conditions around that hour led to traffic gridlock and left some drivers furious. While thousands of runners pounded the asphalt, motorists interviewed by TODAY were stuck in traffic for up to 2.5 hours. For some couples holding wedding banquets in town on Saturday, as many as half their guests arrived late, with many no-shows who turned back in frustration. On Nov 27, the police took to Facebook and Twitter to warn commuters of delays during the event. The post gave details of at least 87 changes to usual traffic arrangements, including road closures and the reversal of traffic in some one-way streets. Major roads affected included Republic Avenue, Cantonment Road and Nicoll Highway.Some of the road closures were scheduled for as early as 1.30pm on Saturday. Motorists who had to travel around the affected areas for work, such as wedding planners and performers, were also caught up in the bumper-to-bumper traffic.

Keep 30% of text

While the Standard Chartered Singapore Marathon on Saturday night (Nov 30) showcased the fastest running talent in the region and beyond, the road closures involved meant that the traffic in the area could not have been slower. The post gave details of at least 87 changes to usual traffic arrangements, including road closures and the reversal of traffic in some one-way streets. Major roads affected included Republic Avenue, Cantonment Road and Nicoll Highway.Some of the road closures were scheduled for as early as 1.30pm on Saturday.

NLG Application: Document Summarization

Automatically generate text summaries of documents

Input: Article 1st sentence	Model-written headline
metro-goldwyn-mayer reported a third-quarter net loss of dlrs 16 million due mainly to the effect of accounting rules adopted this year	mgm reports 16 million net loss on higher revenue
starting from july 1, the island province of hainan in southern china will implement strict market access control on all incoming livestock and animal products to prevent the possible spread of epidemic diseases	hainan to curb spread of diseases
australian wine exports hit a record 52.1 million liters worth 260 million dollars (143 million us) in september, the government statistics office reported on monday	australian wine exports hit record high in september

Source: https://ai.googleblog.com/2016/08/text-summarization-with-tensorflow.html

NLG Application: Machine Translation

Automatically translate text between languages

Example from Google Translate

NLG Application: Text Generation

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Source: https://openai.com/blog/better-language-models/

NLG Application: Question and Answering

QnA is concerned with building systems that automatically **answer questions** posed by humans in a natural language.

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%

NLG Application: Question Answering

Image Credit: Amazon

Improving the quality of voice assistants' responses to questions is of interest to tech giants like Google, Apple, and Microsoft, who seek to address shortfalls in their respective natural language processing (NLP) technologies. They've plenty in the way of motivation — more than 50% of U.S. smart speaker owners say they ask questions of their devices, according to a survey conducted last year by Adobe.

NLG Application: Image Captioning

Demo:

- https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/

Publications on Applied Areas of NLP
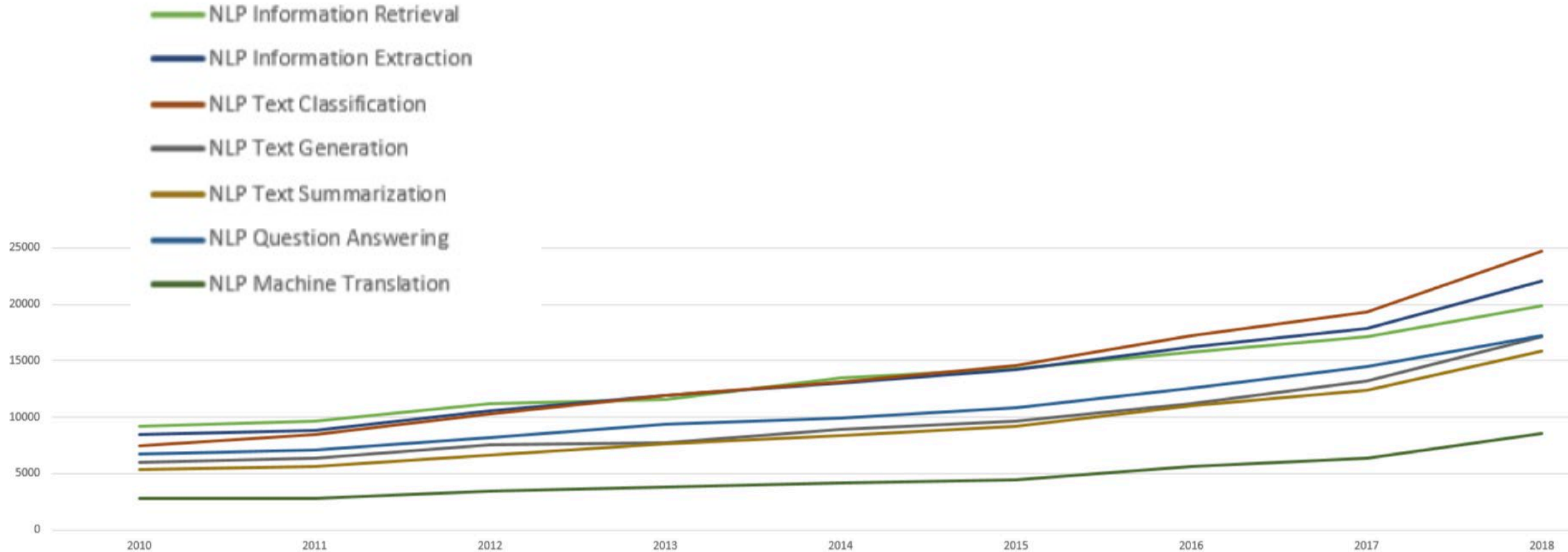

Credit: A Survey of the Usages of Deep Learning for Natural Language Processing , Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita (https://arxiv.org/pdf/1807.10854.pdf)

NLP Benchmarks

BLEU Score - BiLingual Evaluation Understudy
 - Metric to evaluate translations from one language to another

SQuAD — Stanford Question Answering Dataset
- Dataset of text and related questions with answers from the dataset. Given a question and the model needs to predict the right answer

MS MACRO — MAchine Reading COmprehension Dataset
- Large scale dataset focused on machine reading comprehension

GLUE - General Language Understanding Evaluation
- Evaluates the performance of a model on a collection of multiple NLP tasks

GLUE

- A collection of resources for training, evaluating, and analysing natural language understanding systems

- **Language understanding tasks** built on established **existing datasets** and selected to cover a diverse range **of dataset sizes, text genres, and degrees of difficulty**

- Website contains a leader board for **tracking performance of emerging models / algorithms** against benchmark (https://gluebenchmark.com/leaderboard)

Credit: https://mccormickml.com/2019/11/05/GLUE/
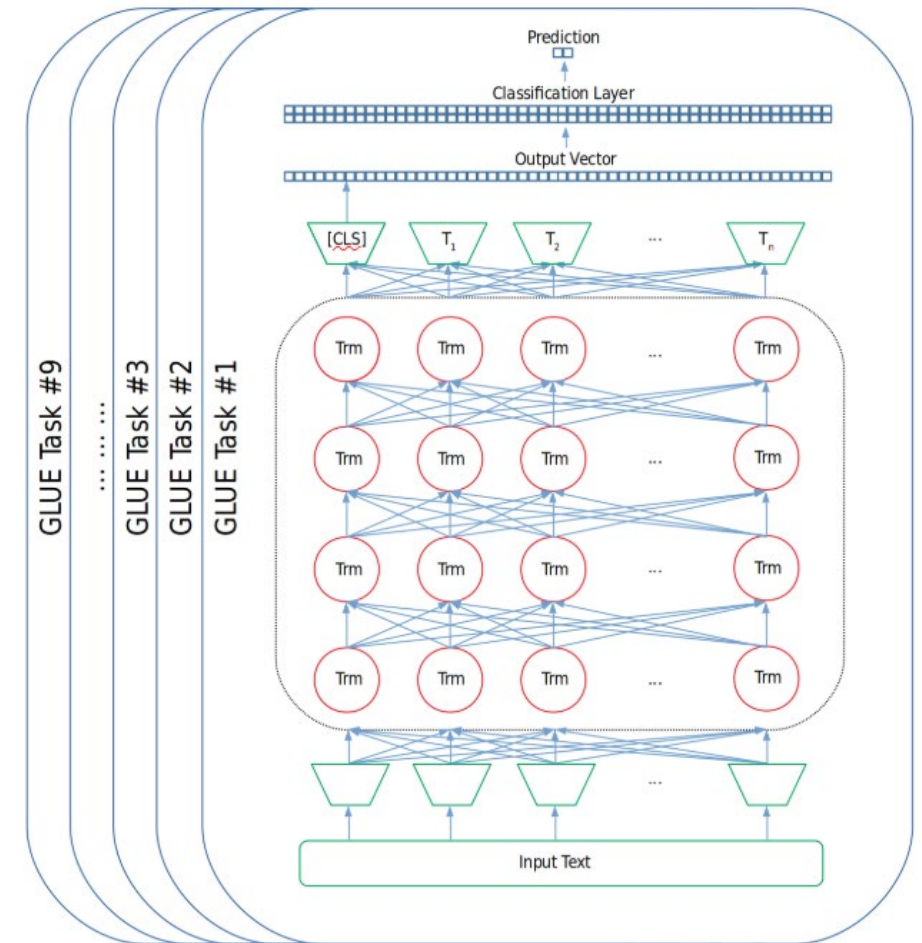
Dataset	Description	Data example
CoLA	Is the sentence grammatical or ungrammatical?	"This building is than that one." = **Ungrammatical**
SST-2	Is the movie review positive, negative, or neutral?	"The movie is funny , smart , visually inventive , and most of all , alive ." = **.93056 (Very Positive)**
MRPC	Is the sentence B a paraphrase of sentence A?	A) "Yesterday , Taiwan reported 35 new infections , bringing the total number of cases to 418 ." B) "The island reported another 35 probable cases yesterday , taking its total to 418 ." = **A Paraphrase**
STS-B	How similar are sentences A and B?	A) "Elephants are walking down a trail." B) "A herd of elephants are walking along a trail." = **4.6 (Very Similar)**
QQP	Are the two questions similar?	A) "How can I increase the speed of my internet connection while using a VPN?" B) "How can Internet speed be increased by hacking through DNS?" = **Not Similar**
MNLI-mm	Does sentence A entail or contradict sentence B?	A) "Tourist Information offices can be very helpful." B) "Tourist Information offices are never of any help." = **Contradiction**
QNLI	Does sentence B contain the answer to the question in sentence A?	A) "What is essential for the mating of the elements that create radio waves?" B) "Antennas are required by any radio receiver or transmitter to couple its electrical connection to the electromagnetic field." = **Answerable**
RTE	Does sentence A entail sentence B?	A) "In 2003, Yunus brought the microcredit revolution to the streets of Bangladesh to support more than 50,000 beggars, whom the Grameen Bank respectfully calls Struggling Members." B) "Yunus supported more than 50,000 Struggling Members." = **Entailed**
WNLI	Sentence B replaces sentence A's ambiguous pronoun with one of the nouns - is this the correct noun?	A) "Lily spoke to Donna, breaking her concentration." B) "Lily spoke to Donna, breaking Lily's concentration." = **Incorrect Referent**

Cognitive Language Services for Building Apps

Decision

Language

Speech

Vision

Web search

Extract meaning from unstructured text

Immersive Reader PREVIEW

Help readers of all abilities comprehend text using audio and visual cues.

Language Understanding

Build natural language understanding into apps, bots, and IoT devices.

QnA Maker

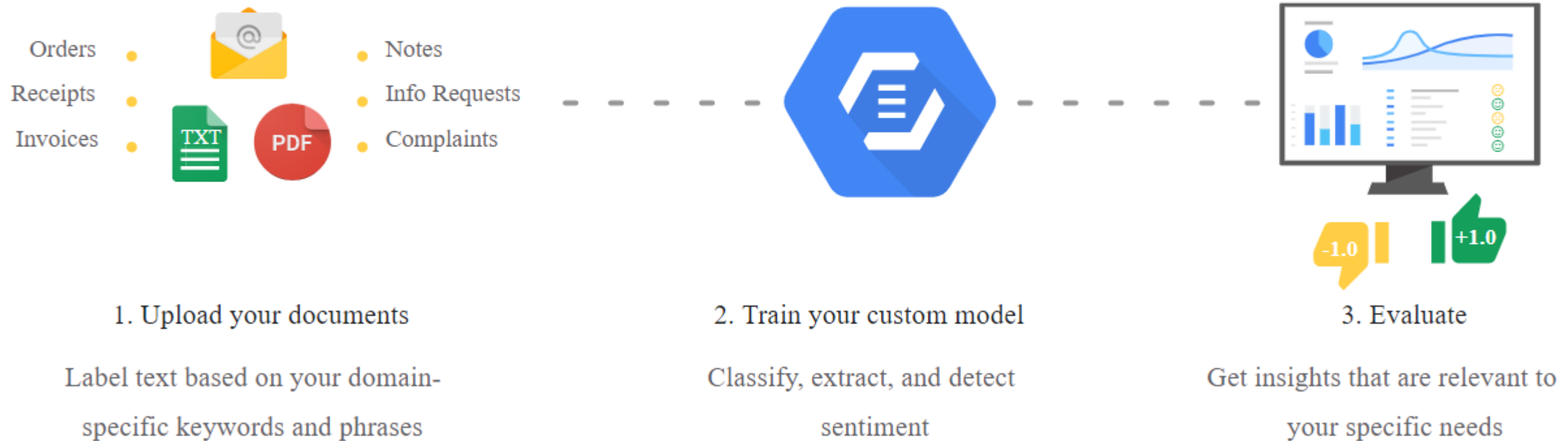Create a conversational question and answer layer over your data.

Text Analytics

Detect sentiment, key phrases, and named entities.

Translator

Detect and translate more than 60 supported languages.

Language Services for Building Apps

Orders • Notes •
Receipts • Info Requests •
Invoices • Complaints •

1. Upload your documents

Label text based on your domain-specific keywords and phrases

2. Train your custom model

Classify, extract, and detect sentiment

3. Evaluate

Get insights that are relevant to your specific needs

Key Takeaways

- It is hard to understand language, but tech and research community are discovering better way to create better algorithms and models to perform NLP tasks

- There are many NLP enabled applications around us

- You can create models or consume language API or services