香港城市大學（東莞）
City University of Hong Kong
(Dongguan)

Celebrating the Establishment in Year 2024

Lecture 5: NLP tasks (1) Understanding tasks

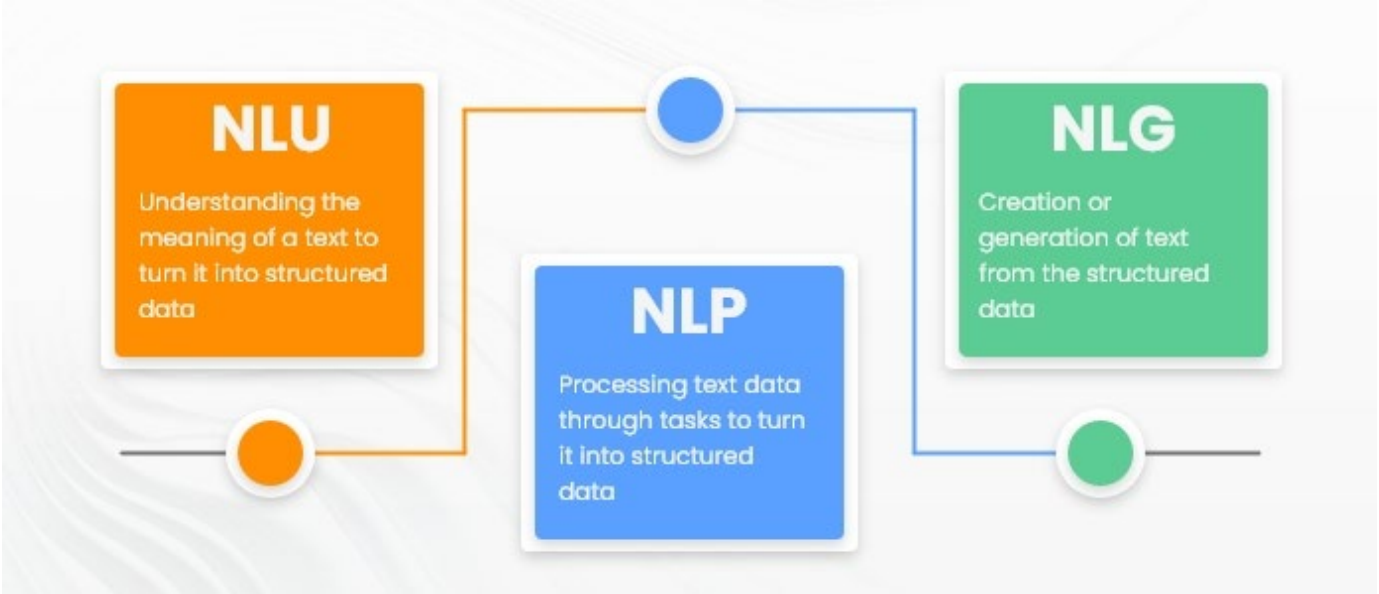
CS6493 Natural Language Processing

Instructor: Linqi Song

Outline

- NLU vs. NLG
- NLU task 1: text classification
- NLU task 2: question answering

NLP tasks – NLU vs. NLG (1)



NLU vs. NLG

- NLU focuses on comprehending and extracting meaning from natural language input. It involves tasks such as text classification and question answering (reading comprehension).
- NLG focuses on generating human-like text that conveys information or communicates effectively. It involves tasks such as machine translation and dialogue generation.
- For today's lecture, we will focus on two classic NLU tasks, i.e., text classification and question answering.

NLP tasks – NLU vs. NLG (2)

Word Tagging	**Sentence Parsing**	**Text Classification**	**Text Pair Matching**	**Text Generation**
Word segmentation	Constituency parsing	Sentiment analysis	Semantic textual similarity	Language modeling
Shallow syntax-chunking	Semantic parsing	Text classification	Natural language inference	Machine translation
Named entity recognition	Dependency parsing	Temporal processing	Relation prediction	Simplification
Part-of-speech tagging		Coreference resolution		Summarization
Semantic role labeling				Dialogue
Word sense disambiguation				Question answering

NLU tasks – GLUE

General Language Understanding Evaluation (**GLUE**) benchmark is a collection of nine NLU tasks by scholars from NYU, U. Washington, etc.

Dataset	Description	Data example	Metric
CoLA	Is the sentence grammatical or ungrammatical?	"This building is than that one." = **Ungrammatical**	Matthews
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)**	Accuracy
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**	Accuracy / F1
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)**	Pearson / Spearman
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**	Accuracy / F1
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**	Accuracy
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**	Accuracy
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**	Accuracy
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**	Accuracy

NLP tasks - hardness

Easy

- Part-of-Speech Tagging
- Named Entity Recognition
- Spam Detection
- Thesaurus

Medium

- Syntactic Parsing
- Word Sense Disambiguation
- Sentiment Analysis
- Topic Modeling
- Information Retrieval

Hard

- Machine Translation
- Text Generation
- Automatic Summarization
- Question Answering
- Conversational Interfaces

Text classification - What is text classification?

- Classification (a.k.a. “categorization”): a ubiquitous enabling technology in data science; studied within pattern recognition, statistics, and machine learning.
- Text classification, also known as text categorization, is a classical problem in natural language processing (NLP), which aims to assign labels or tags to textual units such as sentences, queries, paragraphs, and documents.
- Formulated as the task of generating a hypothesis (or “classifier”, or “model”)

$$h: \mathcal{D} \rightarrow C,$$

where $\mathcal{D} = \{x_1, x_2, \ldots\}$ is a domain of textual data items and $C = \{c_1, c_2, \ldots, c_n\}$ is a finite set of classes

Text classification tasks

- Textual data sources
  - Textual data can come from different sources, including web data, emails, chats, social media, tickets, insurance claims, user reviews, and questions and answers from customer services
- Task examples
  - spam detection, sentiment analysis, news categorization, user intent classification, content moderation

Text classification types

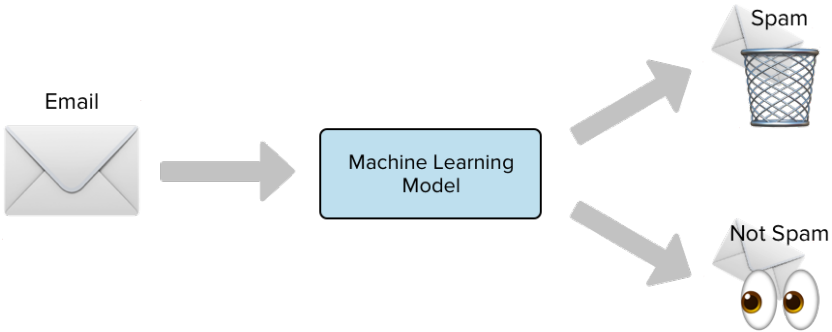
- Binary classification: each item belongs to exactly one class among two
  - E.g., assigning emails to one of {Spam, Legitimate}
- Single-Label Multi-Class (SLMC) classification: each item belongs to exactly one class among many
  - E.g., assigning news articles to one of {HomeNews, International, Entertainment, Lifestyles, Sports}
- Multi-Label Multi-Class (MLMC) classification: each item may belong to zero, one, or several classes
  - E.g., assigning computer science articles to classes in the ACM Classification System
  - May be solved as n independent binary classification problems
- Ordinal classification (OC): as in SLMC, but for the fact that there is a total order among the classes
  - E.g., assigning product reviews to one of {Disastrous, Poor, SoAndSo, Good, Excellent)

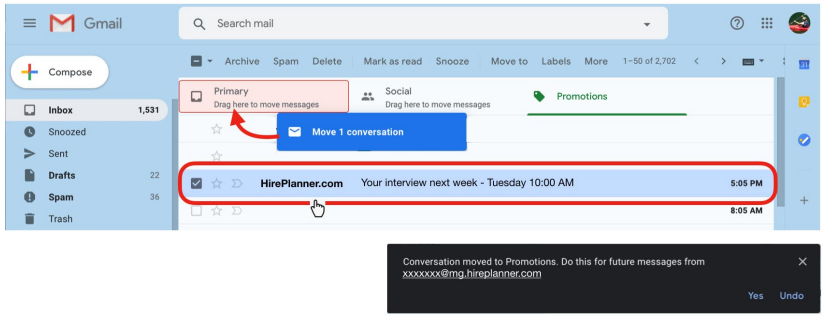
Text classification: hard classification vs. soft classification

- Hard classification (HC): determine which class(es) an item belongs to. Results are categories.
- Soft classification (SC) denotes the task of predicting a (d,c) score for each item-class pair, where the score denotes the probability / strength of evidence / confidence that d belongs to c
  - E.g., a probabilistic classifier outputs posterior probabilities
  - E.g., the AdaBoost classifier outputs scores s(d; c) that represents its confidence that d belongs to c
  - When scores are not probabilities, they can be converted into probabilities via the use of a sigmoidal function; e.g., the logistic function, softmax operations.

Text classification tasks: spam detection

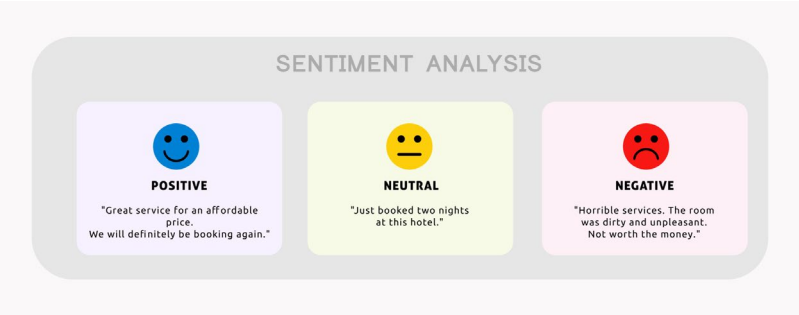
- Spam detection
  - by detecting unsolicited and unwanted emails, we can prevent spam messages from creeping into the user's inbox, thereby improving user experience.
  - Binary: spam, not spam. Multiple classes: primary, social, promotion, different tags

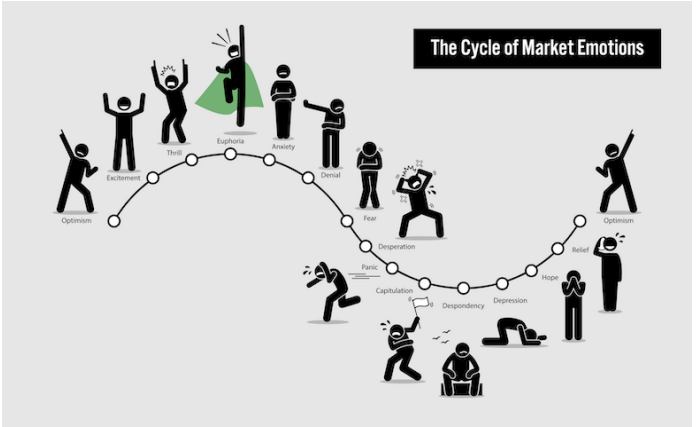




Text classification tasks: sentiment analysis

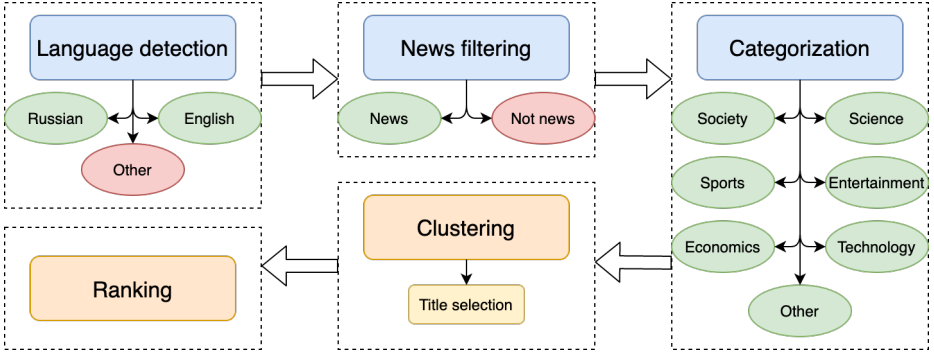
- Sentiment analysis
  - Analyzing people's opinions in textual data (e.g., product reviews, movie reviews, or tweets, emotions towards stock market), and extracting their polarity and viewpoint.
  - Binary: positive, negative; Multiple classes: fine-grained labels or multi-level intensities, e.g., review stars, {happy, sad, surprise, angry}





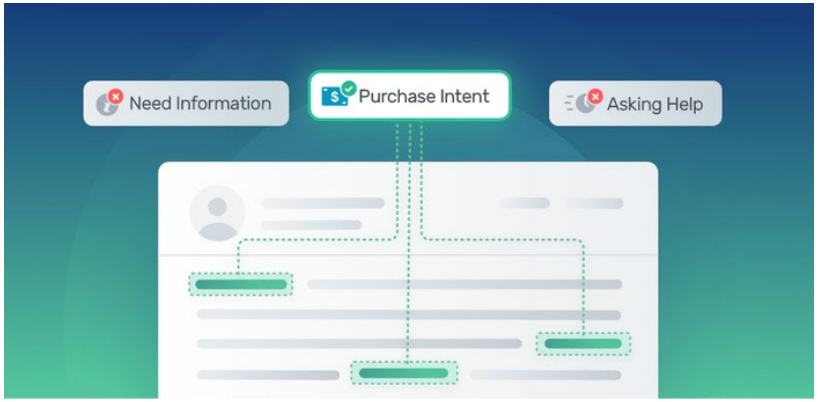
Text classification tasks: news categorization

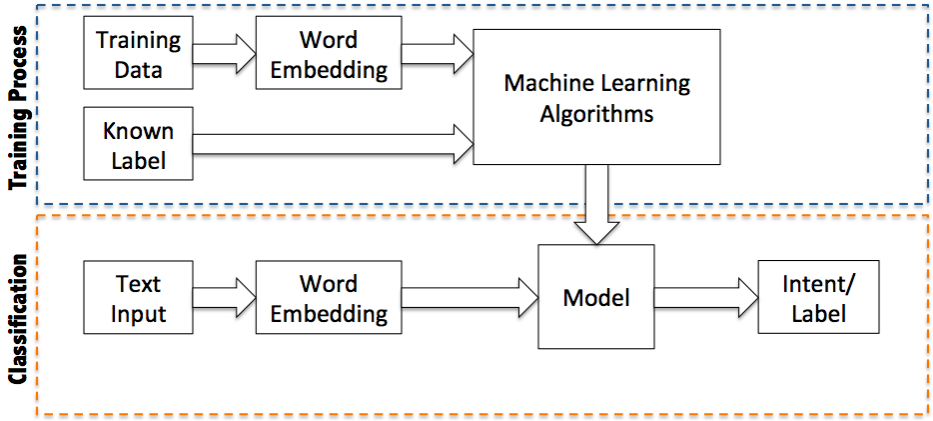
- News categorization
  - News contents are among the most important information sources. A news classification system helps users obtain information of interest in real-time by e.g., identifying emerging news topics or recommending relevant news based on user interests.
  - E.g., assigning news articles to one of {HomeNews, International, Entertainment, Lifestyles, Sports}, {popular, not popular}



Text classification tasks: user intent classification

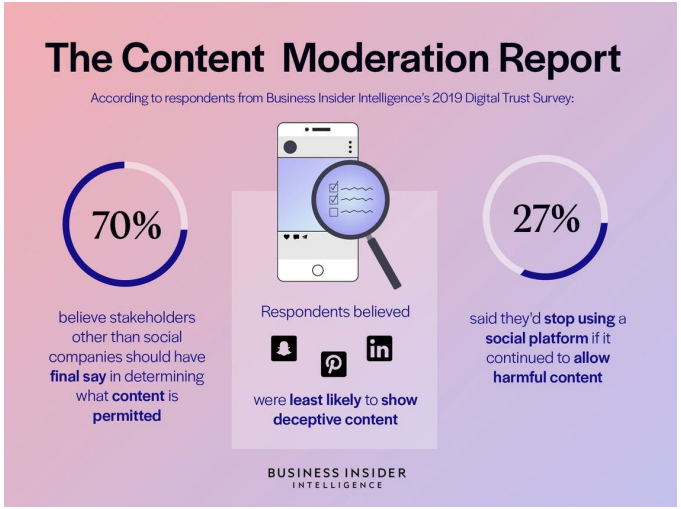
- User intent classification
  - Analyze texts and categorize them into intents. This is useful to understand the intentions behind customer queries, automate processes, and gain valuable insights.
  - E.g., *Purchase, Downgrade, Unsubscribe*, and *Demo Request*





Text classification tasks: content moderation

- Content moderation
  - On Internet websites that invite users to post comments, a moderation system is the method the webmaster chooses to sort contributions that are irrelevant, obscene, illegal, harmful, or insulting with regards to useful or informative contributions.



Text classification tasks: topic analysis

- Topic analysis
  - Aim to identify the theme or topics of a text (e.g., whether a product review is about “customer support” or “ease of use”).

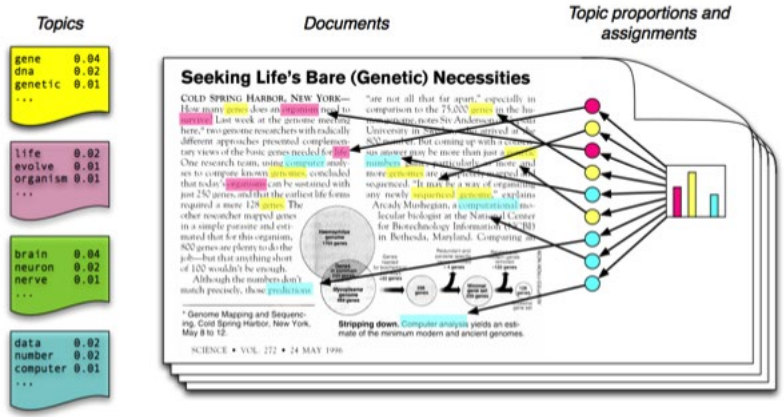
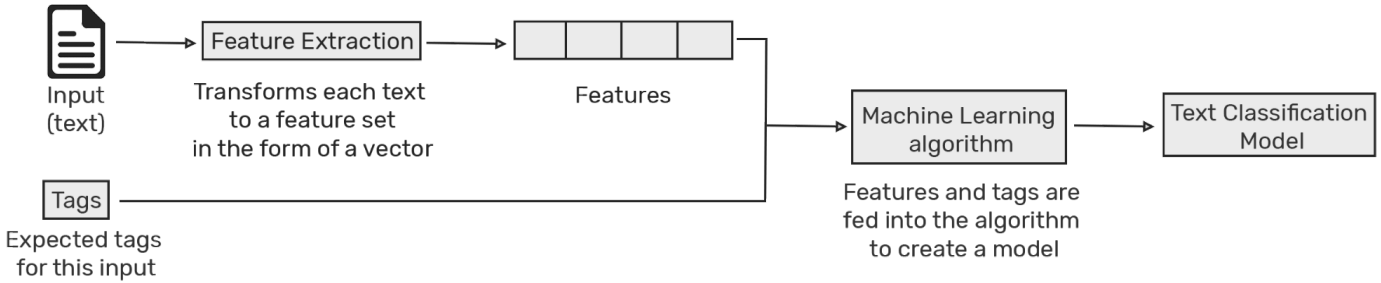


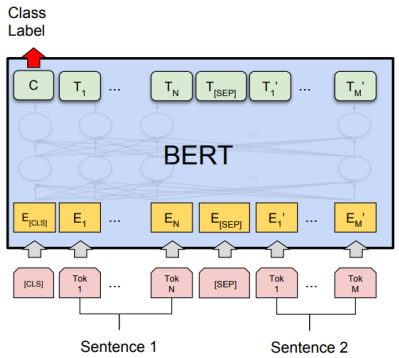
Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Text classification methods

- Feature extraction + classification



- End-to-end model

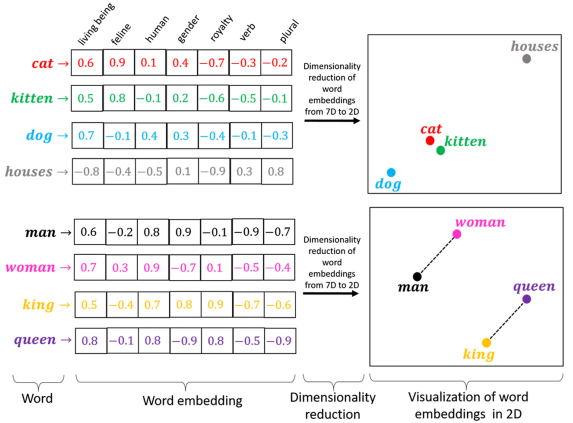


Text classification methods - preprocessing

- Preprocessing
  - Removing (or not) punctuations like . , ! $( ) * % @
  - Removing URLs
  - Removing stop words
  - Lower casing
  - Tokenization
  - Stemming
  - Lemmatization

Text classification methods - feature extraction

- Feature extraction
  - In general: dense vectors (embeddings) using subword embeddings, word2vec, GloVe, or pretrained contextualized language models like ELMo, BERT, etc.
  - In classification by topic, a typical choice is to make the set of features coincide with the set of words that occur in the training set (unigram model, a.k.a. “bag-of-words”).
  - In classification by author/title, features such average word length, average sentence length, punctuation frequency, frequency of subjunctive clauses, etc., are used.
  - Other features: syntactic information, pragmatic information, etc.



The Bag of Words Representation

Text classification methods - feature selection

- Feature selection
  - Feature selection (FS) has the goal of identifying the most discriminative features, so that the others may be discarded.
  - The "filter" approach to FS consists in measuring (via a function like mutual information) the discriminative power of each feature t_k and retaining only the top-scoring features.
  - Matrix decomposition techniques (e.g., PCA, SVD, LSA) can be used to synthesize new features that replace the features discussed above with ones not suffering from ambiguity and polysemy.

Text classification methods - classifier selection

- Classifier selection
  - Support vector machines (SVMs)
  - Boosted decision stumps
  - Logistic regression
  - Naive Bayesian methods
  - Lazy learning methods (e.g., k-NN)
  - Neural network based methods (often support end-to-end classification)

Text classification methods - end2end neural text classifiers

- TextRNN: Recurrent Neural Network for Text Classification with Multi-Task Learning

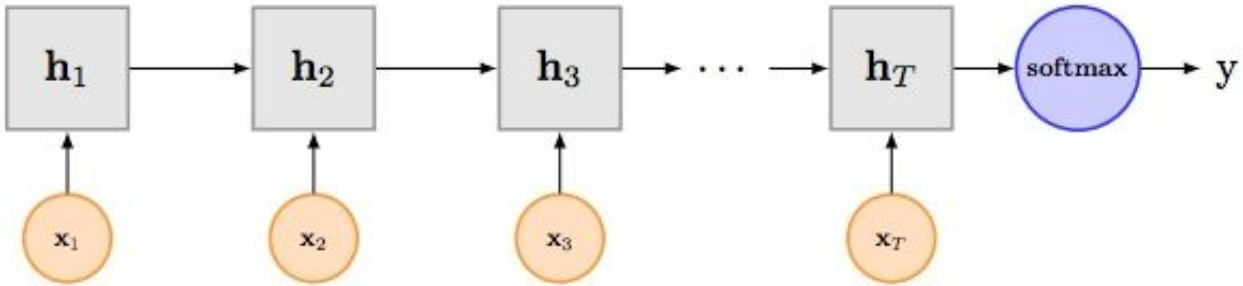
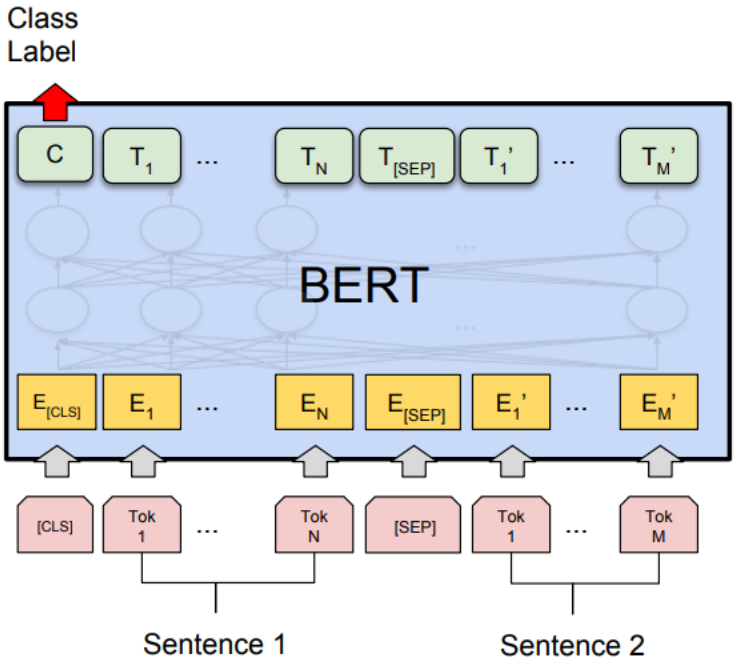
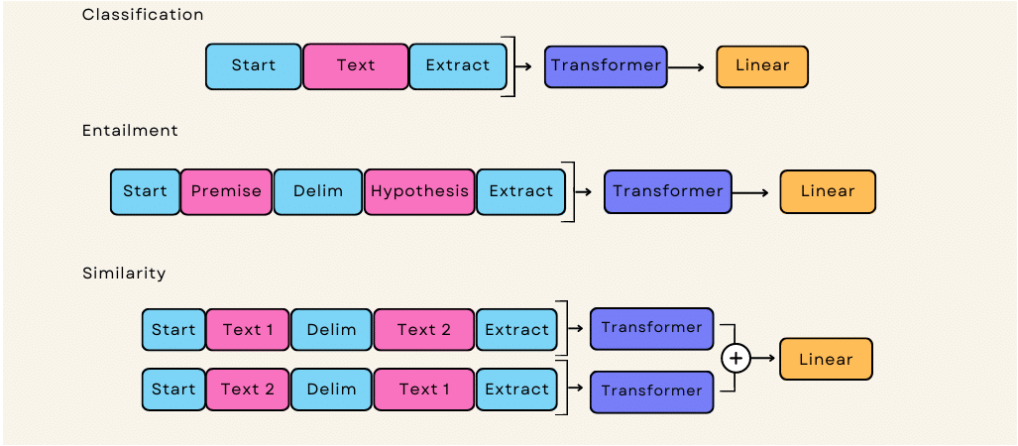


Figure 1: Recurrent Neural Network for Classification

Text classification methods - end2end neural text classifiers (8)

- BERT, GPT, etc., pretrained models

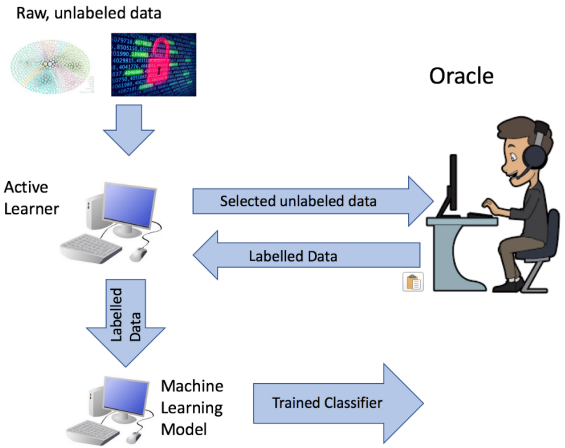




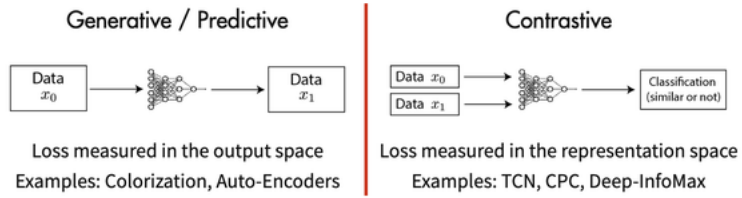
Text classification - labeling

- Labeling (annotation) is usually costly.
  - Active learning for classification, when the items to label for training purposes are suggested by the system. Active learning algorithm can interactively query a user to label new data points with the desired outputs.
  - Self-supervised learning: constructing positive/negative samples

Active Machine Learning



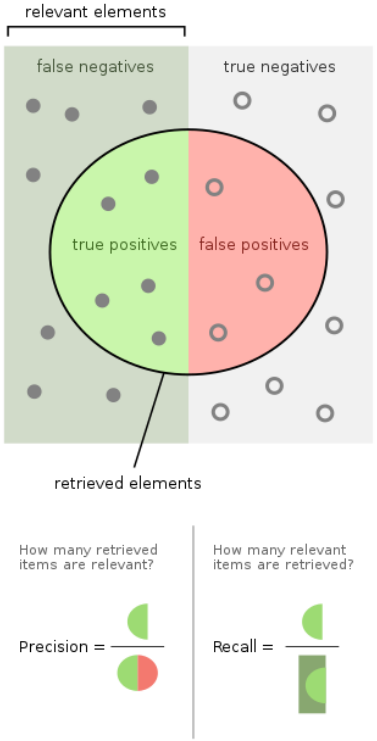
Contemporary self-supervised learning methods can roughly be broken down into two classes of methods:



Text classification - evaluation

- Classification
  - Precision-recall, F1 score

$$\textbf{Precision} = \frac{TP}{TP + FP}, \quad \textbf{Recall} = \frac{TP}{TP + FN}, \quad \textbf{F1-score} = \frac{2\,Prec\,Rec}{Prec + Rec}$$



Text classification - network selection

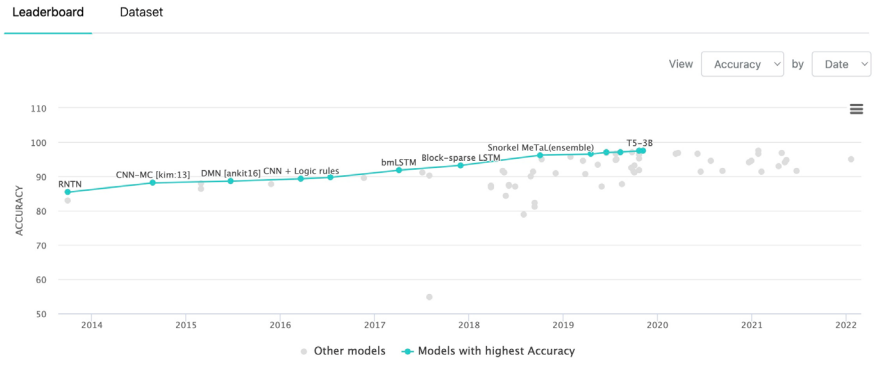
- How to select the appropriate network for your task?
  - Select a pretrained language model (PLM), such as BERT, or GPT.
  - Domain adaptation. Adapting the PLM using in-domain data by continual pre-training the selected general-domain PLM. For domains with abundant unlabeled text, such as biomedicine, pretraining language models from scratch might also be a good choice.
  - Task-specific model design. One or more task-specific layers are added on the top to generate the final output for the target task.
  - Task-specific fine-tuning. The task-specific layers can be either trained alone with the PLM fixed or trained together with the PLM, sometimes a multi-task training may be a good choice.
  - Model compression. PLMs are expensive to serve. They often need to be compressed via e.g., knowledge distillation to meet the latency and capacity constraints in real-world applications.

Text classification - sentiment analysis datasets

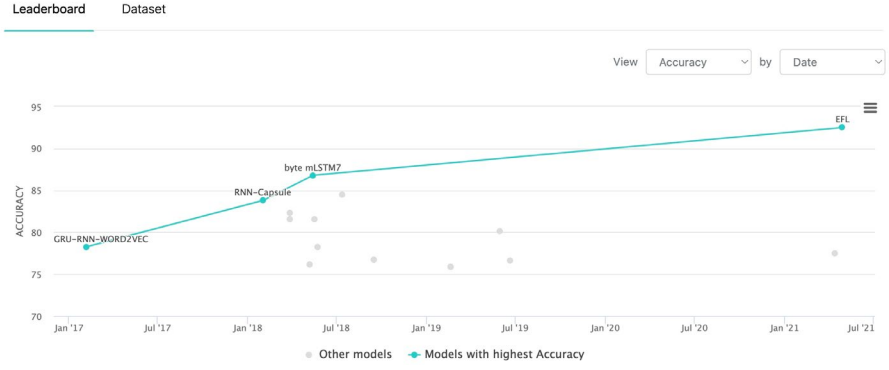
- Yelp. For two sentiment classification tasks. One is to detect fine-grained sentiment labels and is called Yelp-5. The other predicts the negative and positive sentiments, and is known as Yelp Review Polarity or Yelp-2. Yelp-5 has 650,000 training samples and 50,000 test samples for each class, and Yelp-2 includes 560,000 training samples and 38,000 test samples for negative and positive classes.
- IMDb. Binary sentiment classification of movie reviews. Equal number of positive and negative reviews. It is evenly divided between training and test sets with 25,000 reviews for each.
- Movie Review (MR). Detecting the sentiment associated with a particular review and determining whether it is negative or positive. 10,662 sentences with even numbers of negative and positive samples.
- Stanford Sentiment Treebank (SST). Extended version of MR. Two versions are available, SST-1 with fine-grained labels (5-class) and SST2 with binary labels. SST-1 consists of 11,855 movie reviews which are divided into 8,544 training samples, 1,101 development samples, and 2,210 test samples. SST-2 is with 6,920, 872 and 1,821 as training, development and test samples, respectively.
- Multi-Perspective Question Answering (MPQA). An opinion corpus with two class labels. 10,606 sentences extracted from news articles related to a wide variety of news sources. This is an imbalanced dataset with 3,311 positive documents and 7,293 negative documents.
- Amazon. Product reviews from the Amazon website with labels for both binary classification and multi-class (5-class) classification. The Amazon binary classification dataset consists of 3,600,000 and 400,000 reviews for training and test, respectively. The Amazon 5-class classification dataset (Amazon-5) consists of 3,000,000 and 650,000 reviews for training and test, respectively.

Text classification - sentiment analysis performance

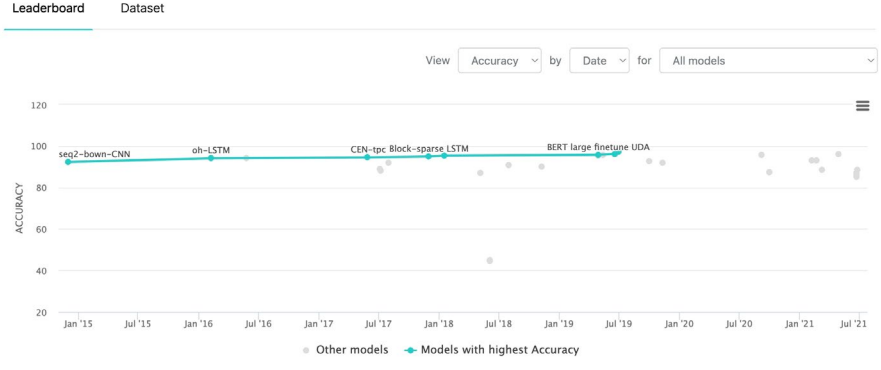
Sentiment Analysis on SST-2 Binary classification



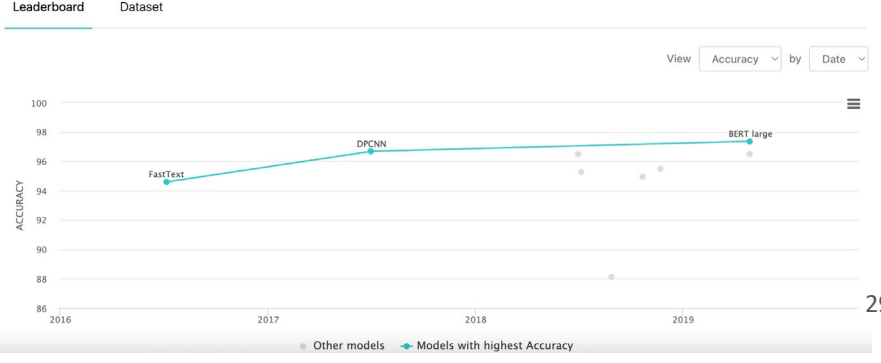
Sentiment Analysis on MR



Sentiment Analysis on IMDb



Sentiment Analysis on Amazon Review Polarity

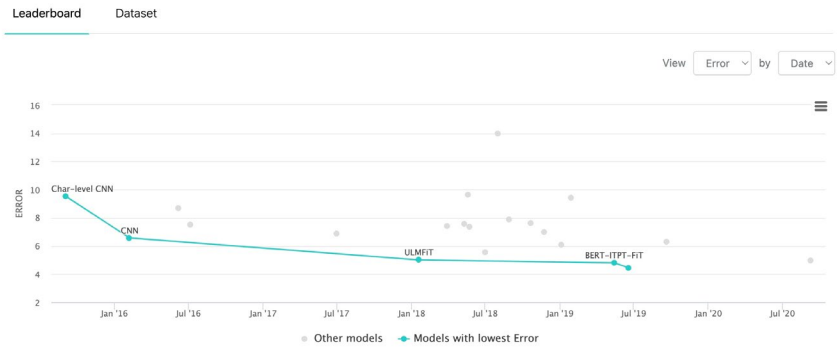


Text classification - news classification datasets

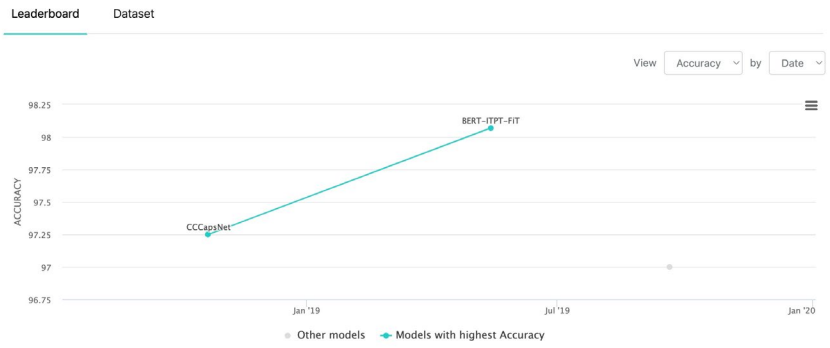
- AG News. News articles collected from more than 2,000 news sources by ComeToMyHead, an academic news search engine. 120,000 training samples and 7,600 test samples. Each sample is a short text with a four-class label (“World”, “Sports”, “Business”, “Sci/Tech”).
- 20 Newsgroups. Newsgroup documents posted on 20 different topics. One of the most popular versions contains 18,821 documents that are evenly classified across all topics.
- Sogou News. 2,909,551 news articles from the SogouCA and SogouCS news corpora, in 5 categories. The classification labels of the news are determined by their domain names in the URL. For example, the news with URL http://sports.sohu.com is categorized as a sport class.
- Reuters news. Collected from the Reuters financial newswire service in 1987. ApteMod is a multi-class version with 10,788 documents. It has 90 classes, 7,769 training documents and 3,019 test documents. Other datasets derived from a subset of the Reuters dataset include R8, R52, RCV1, and RCV1-v2.

Text classification - news classification performance

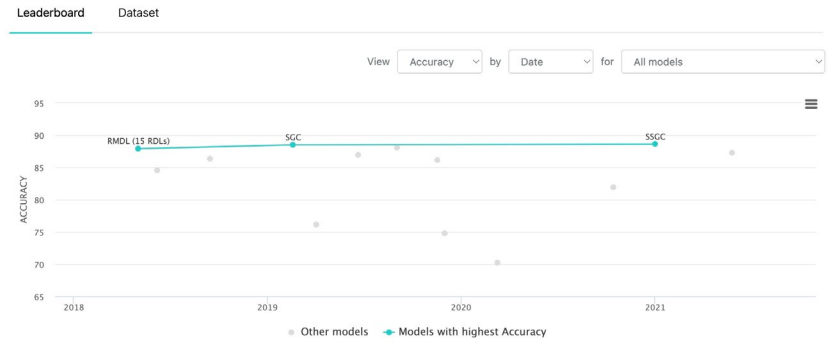
Text Classification on AG News



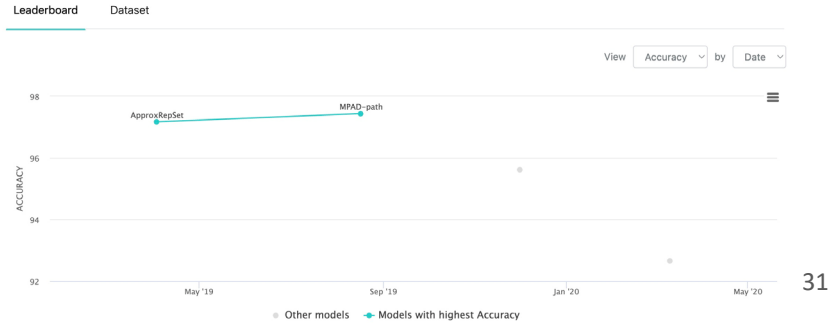
Text Classification on Sogou News



Text Classification on 20NEWS



Document Classification on Reuters-21578



Text classification - topic classification datasets

- Ohsumed. A subset of the MEDLINE database. Ohsumed contains 7,400 documents. Each document is a medical abstract that is labeled by one or more classes selected from 23 cardiovascular diseases categories.

- DBpedia. A large-scale, multilingual knowledge base that has been created from the most commonly used info boxes within Wikipedia. DBpedia is published every month and some classes and properties are added or removed in each release. The most popular version of DBpedia contains 560,000 training samples and 70,000 test samples, each with a 14-class label.

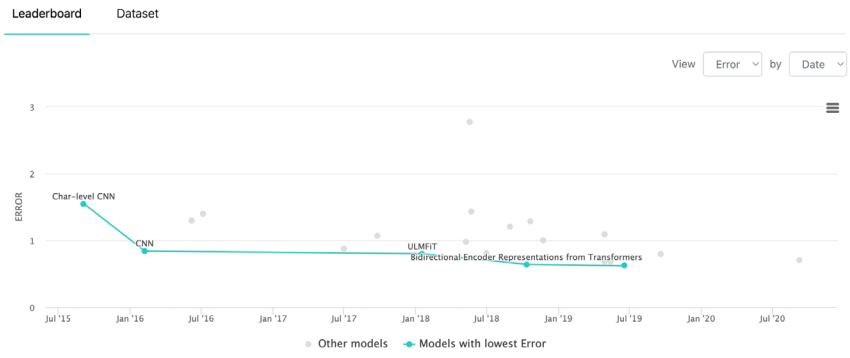
- EUR-Lex. Different types of documents are indexed according to several orthogonal categorization schemes to allow for multiple search facilities. The most popular version of this dataset is based on different aspects of European Union law and has 19,314 documents and 3,956 categories.

- Web Of Science (WOS). A collection of data and meta-data of published papers available from the Web of Science, which is the world's most trusted publisher-independent global citation database. WOS has been released in three versions: WOS-46985, WOS-11967 and WOS-5736. WOS-46985 is the full dataset of 46,985 documents with 134 categories which include 7 parents categories. WOS-11967 and WOS-5736 are two subsets of WOS-46985.

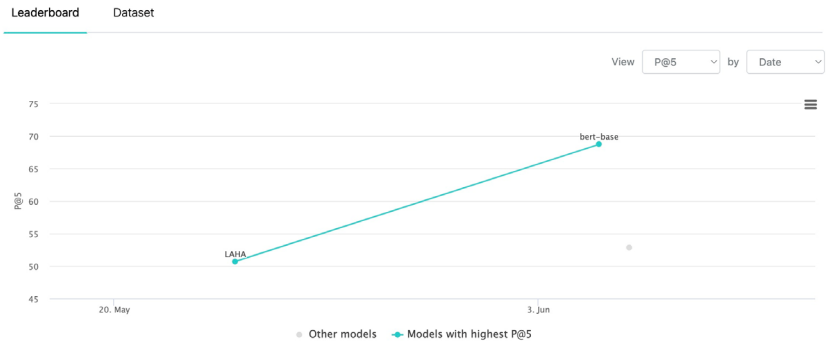
- PubMed. A search engine developed by the National Library of Medicine for medical and biological scientific papers, which contains a document collection. The dataset contains 19,717 scientific publications. Each document has been labeled with the classes of the MeSH set which is a label set used in PubMed. Each sentence in an abstract is labeled with its role in the abstract using one of the 5 following classes: background, objective, method, result, or conclusion.

Text classification - topic classification performance

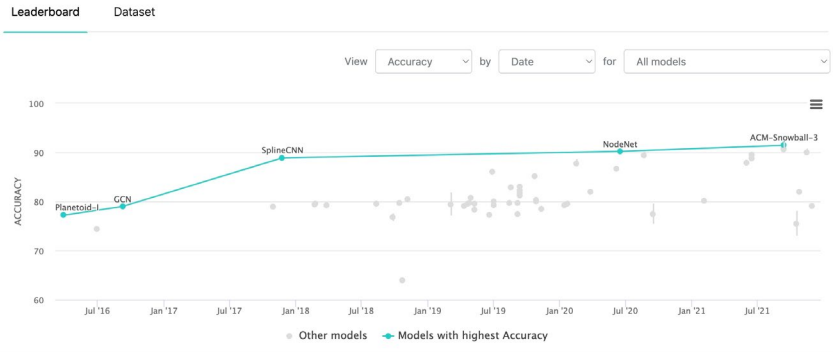
Text Classification on DBpedia



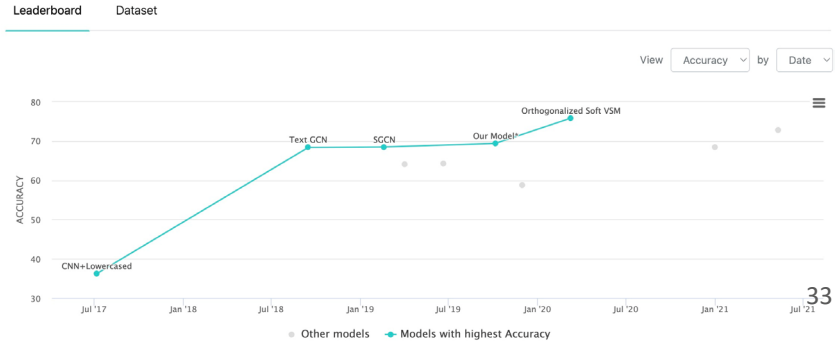
Multi-Label Text Classification on EUR-Lex



Node Classification on Pubmed



Text Classification on Ohsumed



Question-answering (QA)

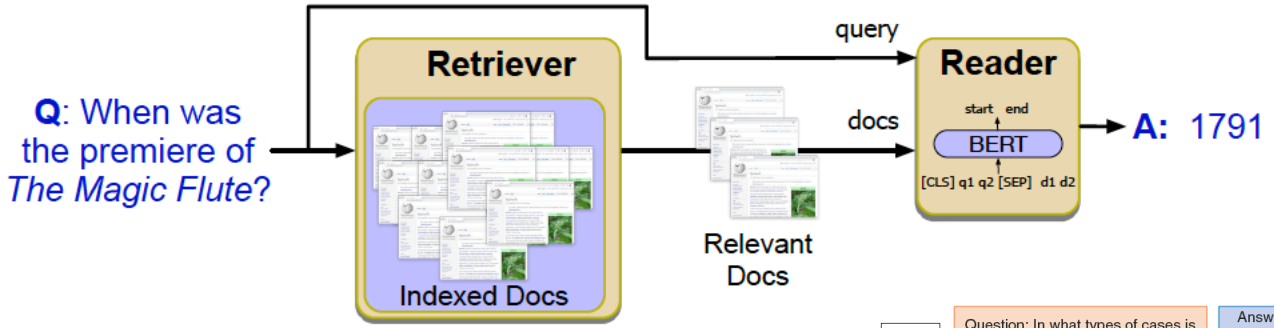
- Idea: extract information from materials (documents, conversations, online searches, etc.) and give a short and concise answer that will meet user's information needs.
- Type of questions:
  - factoid: simplest, most common
    - The symbol for mercuric oxide is?
    - Which NFL team represented the AFC at Super Bowl 50?
  - mathematical
    - 2+3=?

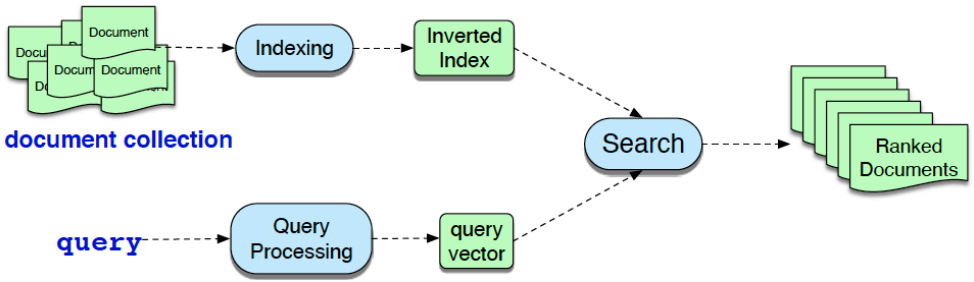
Factoid QA

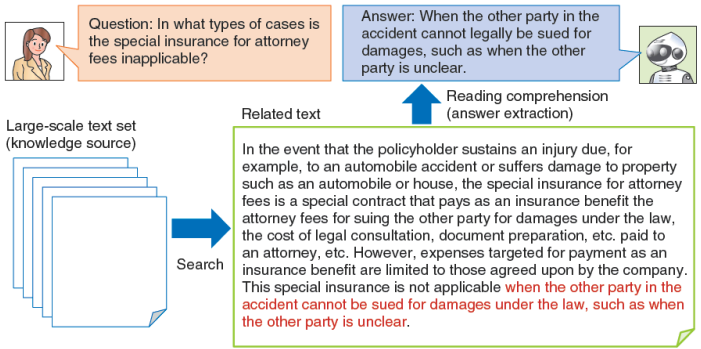
- Factoid questions: questions that can be answered with simple facts expressed in short texts, like the following:
  - Where is the Louvre Museum located?
  - What is the average age of the onset of autism?
- Information-retrieval (IR) based QA (open domain question QA)
  - Given a user question, information retrieval is used to find relevant passages (vast amount of text on the web or in collections of scientific papers like PubMed). Then neural reading comprehension algorithms read these retrieved passages and draw an answer directly from spans of text.
- Knowledge-based QA
  - To build a semantic representation of the query (logic query). These meaning representations are then used to query databases of facts.
  - E.g., What states border Texas? -> λ x.state(x)$\wedge$borders(x; Texas).

IR-based QA

- Two steps: IR + RC



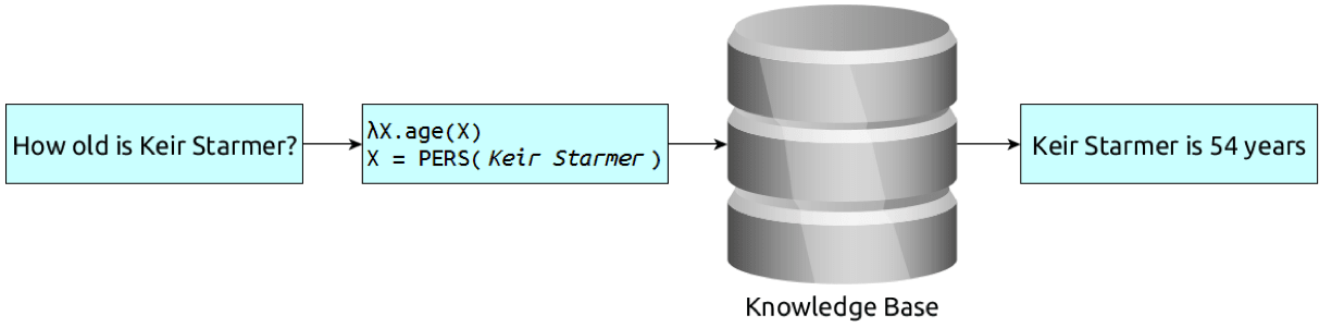




IR RC

Knowledge-based QA

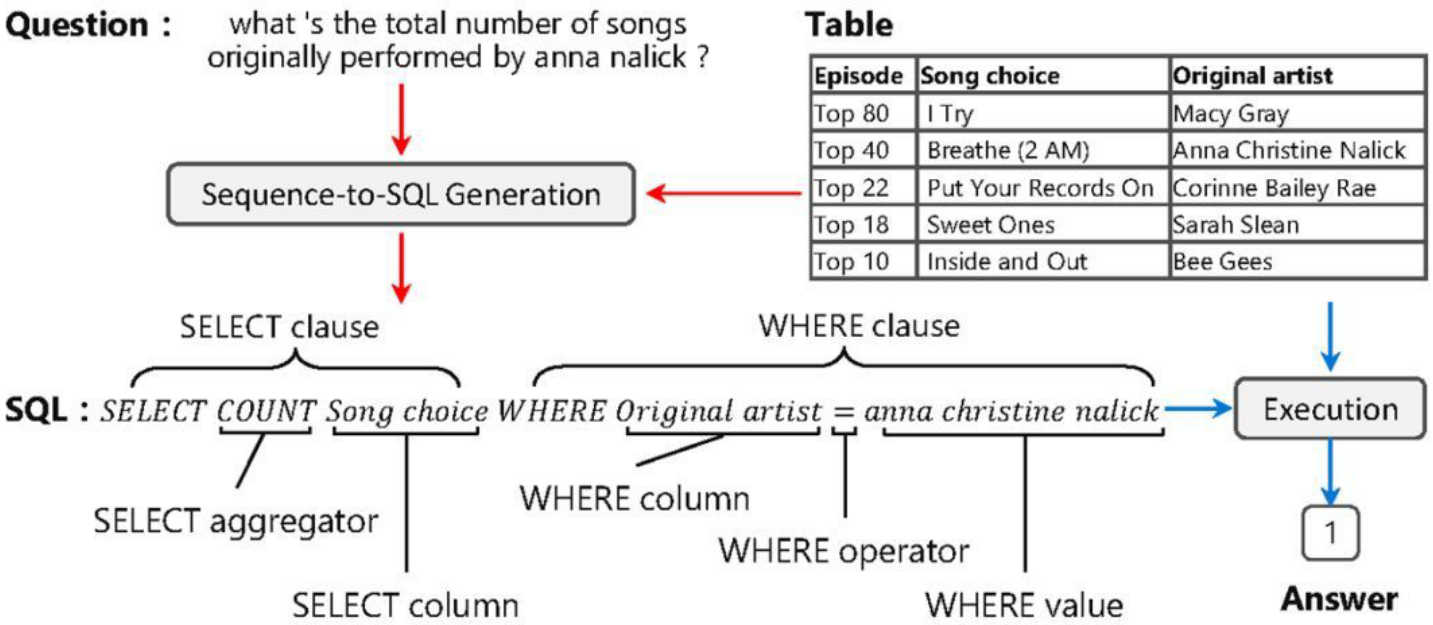
- Generate logical form expression to query the knowledge base



Question → Logical Form → KB Query → Answer

Text2SQL

- Seq2Seq model for text2SQL
- Use examples: booking restaurants, play music, etc.



Reading comprehension (RC)

- The ability to read and understand unstructured text and then answer questions about it.
- Probelm:
  - Input: context (passage of text) and query
  - output: answer
    - abstractive: free-form answer (more generation)
    - extractive: substring of the content (more understanding)
- Connection to Question-Answering
  - QA: a task
  - RC: a possible approach to solve QA
  - Other possible solutions to QA: knowledge-based intormation retrieval, keywords detection mechanism, etc.

RC necessitates language understanding

- Coreference resolution:

  understand that “she”=Alyssa

- Inferring that “special” = catfish so this must be what Alyssa ate

- Identify which entities in the text are people and among these which are Alyssa’s friends

	Alyssa got to the beach after a long trip. She’s from Charlotte. She traveled from Atlanta. She’s now in Miami. She went to Miami to visit some friends. But she wanted some time to herself at the beach, so she went there first. After going swimming and laying out, she went to her friend **Ellen**’s house. **Ellen** greeted **Alyssa** and they both had some lemonade to drink. **Alyssa** called her friends **Kristen** and **Rachel** to meet at **Ellen**’s house. The girls traded stories and caught up on their lives. It was a happy time for everyone. The girls went to a restaurant for dinner. The restaurant had a special on catfish. **Alyssa** enjoyed the restaurant’s special. **Ellen** ordered a salad. **Kristen** had soup. **Rachel** had a steak. After eating, the ladies went back to **Ellen**’s house to have fun. They had lots of fun. They stayed the night because they were tired. **Alyssa** was happy to spend time with her friends again.
(a)	**Question:** What city is Alyssa in? **Answer**: Miami
(b)	**Question**: What did Alyssa eat at the restaurant? **Answer**: catfish
(c)	**Question**: How many friends does Alyssa have in this story? **Answer**: 3

Stanford question answering dataset (SQuAD 1.1)

- SQuAD dataset motivation (problems in previous datasets):
  - High quality human-written databases not very large (on the order 10^3 in size)
  - Cloze-form questions better, but not very natural
    - Semi-synthetic (As in Cloze)
    - Not explicit question answering
  - Heuristically created → noisy

Stanford question answering dataset (SQuAD 1.1)

- Rajpurkar et al., 2016, SQuAD 1.1
- SQuAD features:
  - Questions posed by crowdworkers on a set of Wikipedia articles
  - 107,785 question-context-answer triples on 536 articles
  - Extractive question answering: the answer to each question is a segment of text (span) from the corresponding reading passage
  - Mostly, 3 ground-truth answers were given for each question by different crowdworkers
- Why is SQuAD better?
  - Human-written, human curated → less noisy than CNN/DM
  - Not cloze-form
  - Step towards better language understanding

SQuAD examples

Question: Which team won Super Bowl 50?

Passage

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.

Computational complexity theory is a branch of the theory of computation in theoretical computer science that focuses on classifying computational problems according to their inherent difficulty, and relating those classes to each other. A computational problem is understood to be a task that is in principle amenable to being solved by a computer, which is equivalent to stating that the problem may be solved by mechanical application of mathematical steps, such as an algorithm.

By what main attribute are computational problems classified utilizing computational complexity theory?

Ground Truth Answers: inherent difficulty | their inherent difficulty | inherent difficulty

Prediction: inherent difficulty

3 gold answers are collected for each answer

SQuAD 2.0

- SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones.
- To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering.

SQuAD 2.0 no answer example

Genghis Khan united the Mongol and Turkic tribes of the steppes and became Great Khan in 1206. He and his successors expanded the Mongol empire across Asia. Under the reign of Genghis' third son, Ögedei Khan, the Mongols destroyed the weakened Jin dynasty in 1234, conquering most of northern China. Ögedei offered his nephew Kublai a position in Xingzhou, Hebei. Kublai was unable to read Chinese but had several Han Chinese teachers attached to him since his early years by his mother Sorghaghtani. He sought the counsel of Chinese Buddhist and Confucian advisers. Möngke Khan succeeded Ögedei's son, Güyük, as Great Khan in 1251. He

When did Genghis Khan kill Great Khan?

Gold Answers: <No Answer>

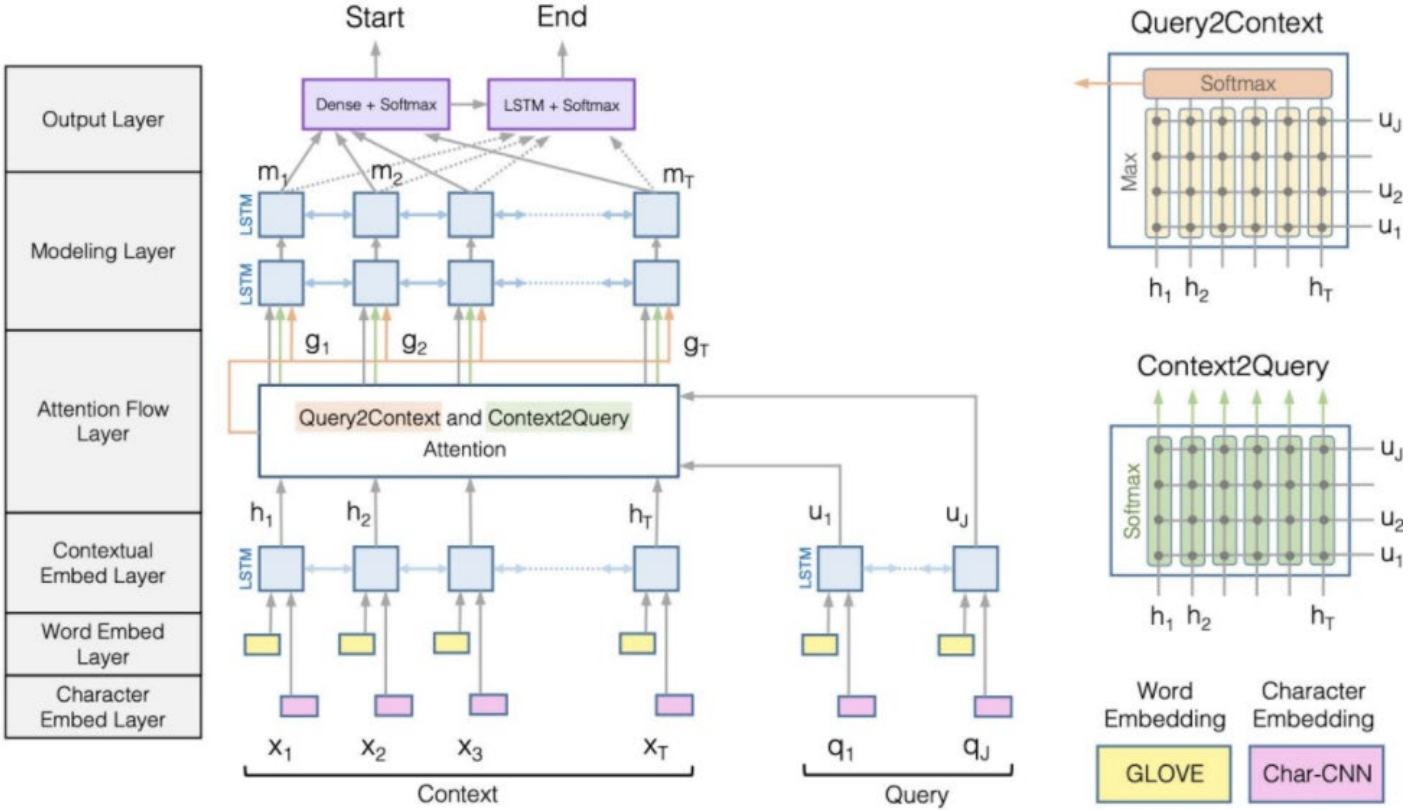
Prediction: 1234 [from Microsoft nlnet]

BiDAF (Seo et al., 2017)

- BiDAF (Seo et al., 2017), Bidirectional Attention Flow for Machine Comprehension, U Washington & Allen AI

- Motivation
  - Incorporating attention better into QA
  - Some key features
    - Bidirectional attention: query-to-context and context-to-query
    - Includes character-level, word-level, and contextual embeddings
    - Attended vectors are passed along together with original embeddings

BiDAF architecture (Seo et al., 2017)

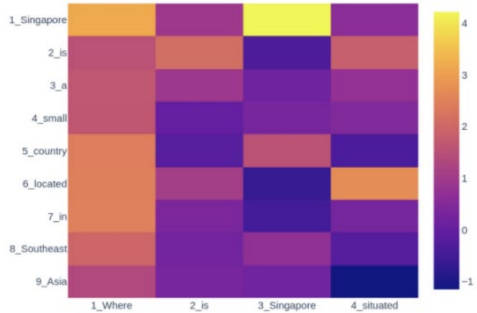


Basic components of the model

- Character embedding layer
  - Embeds each word using character-level CNNs.
- Word embedding layer
  - GloVe
- Contextual embedding layer
  - Character and word embeddings passed through bi-LSTM to obtain contextual embeddings for query and context.
- Attention flow layer
  - Produces a set of query-aware feature vectors for each word in the context (C2Q) and a context-aware vector for the query (Q2C).
- Modeling layer
  - Contextual embeddings and attended vectors passed through two-layer bi-LSTM for even more refined epresentation.
- Output layer
  - Linear layer then softmax to obtain a start probability distribution and an end probability distribution over the indices.

BiDAF closer look: attention

- Compute a similarity matrix S from context embeddings H and query embeddings U

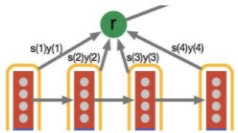


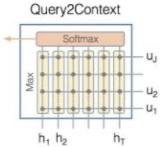
$$S_{tj} = \alpha(H_{:t}, U_{:j}) \in \mathbb{R}$$
$$\alpha(h, u) = w_{(S)}^T [h; u; h \circ u]$$

Attention (similarity) score between token *t* in context and token *j* in query

Q2C: which token in context is more related to the query?
C2Q: which token in query is more related to the context?

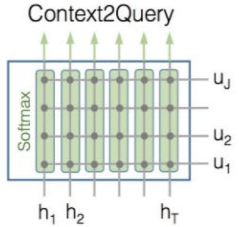
- Q2C: query → which tokens in the context to attend to





$$\tilde{h} = \sum_t b_t H_{:t}$$
$$b_t \propto \exp(\max_j S_{tj})$$

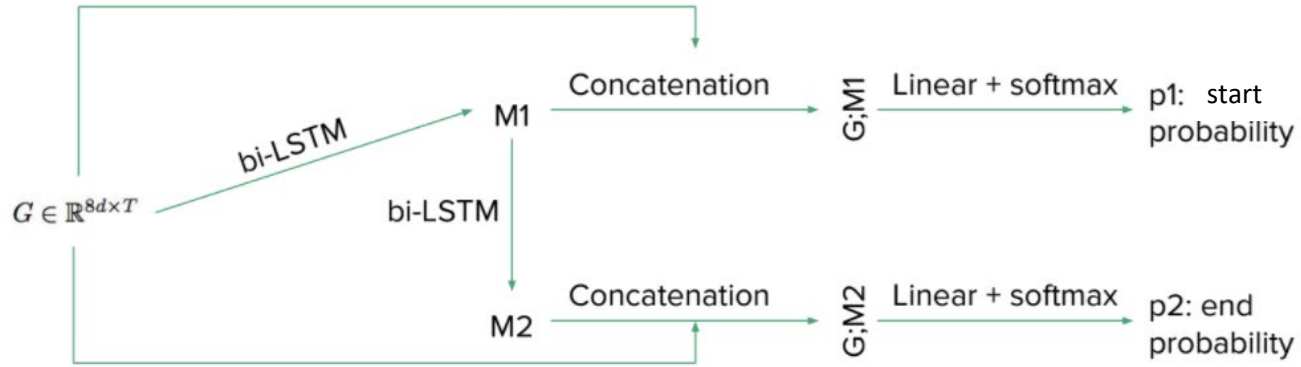
- **C2Q**: each context token → which tokens in the query it should attend to



$$\widetilde{U}_{:t} = \sum_{j=1}^{J} a_{tj} U_{:j}$$
$$a_{tj} \propto \exp(S_{tj})$$

BiDAF closer look: output flow

output flow:



Performance metrics

- Training: log likelihood of correct start/end indexes $L(\theta) = -\frac{1}{N}\sum_{i}^{N} \log(\mathbf{p}^1_{y^1_i}) + \log(\mathbf{p}^2_{y^2_i})$
- Testing: choose start-end index pair (i, j with i < j maximizing p1(i) * p2(j)
  - Remove all articles (a, an, the)
  - Exact Match (EM): choosing exactly the same start and end index as some gold answer
  - F1: treat predicted and gold answers as bags of tokens, then take harmonic mean of precision and recall

$$\text{precision} = \frac{\#\text{ of correctly predicted tokens}}{\#\text{ of predicted tokens}} \qquad \text{recall} = \frac{\#\text{ of correctly predicted tokens}}{\#\text{ of gold tokens}}$$

$$F_1 = \left(\frac{2}{\text{recall}^{-1} + \text{precision}^{-1}}\right) = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Results on SQuAD

	Single Model		Ensemble	
	EM	F1	EM	F1
Logistic Regression Baseline[a]	40.4	51.0	-	-
Dynamic Chunk Reader[b]	62.5	71.0	-	-
Fine-Grained Gating[c]	62.5	73.3	-	-
Match-LSTM[d]	64.7	73.7	67.9	77.0
Multi-Perspective Matching[e]	65.5	75.1	68.2	77.2
Dynamic Coattention Networks[f]	66.2	75.9	71.6	80.4
R-Net[g]	**68.4**	**77.5**	72.1	79.7
BIDAF (Ours)	68.0	77.3	**73.3**	**81.1**

Ensemble: train 12 models, choose start and end indices with the highest sum of confidence scores.

Ablation study

	EM	F1
No char embedding	65.0	75.4
No word embedding	55.5	66.8
No C2Q attention	57.2	67.7
No Q2C attention	63.6	73.7
Dynamic attention	63.5	73.6
BIDAF (single)	67.7	77.3
BIDAF (ensemble)	72.6	80.7

Character-level embedding:

- effective in handling out-of-vocab or rare words

Word-level embedding:

- better at capturing the overall semantics of words

Ablation study (cont.)

	EM	F1
No char embedding	65.0	75.4
No word embedding	55.5	66.8
No C2Q attention	57.2	67.7
No Q2C attention	63.6	73.7
Dynamic attention	63.5	73.6
BIDAF (single)	67.7	77.3
BIDAF (ensemble)	72.6	80.7

C2Q ablation:

attended query vector for each context word is a uniform average over the word vectors

Q2C ablation:

remove any terms incorporating attended context vectors for each query word

Ablation study (cont.)

	EM	F1
No char embedding	65.0	75.4
No word embedding	55.5	66.8
No C2Q attention	57.2	67.7
No Q2C attention	63.6	73.7
Dynamic attention	63.5	73.6
BIDAF (single)	67.7	77.3
BIDAF (ensemble)	72.6	80.7

Dynamic attention:

Update attention throughout the modelling layer

Intuition:

Separating out the attention layer gives a richer set of features to feed into the modelling layer

BiDAF result on CNN/Daily Mail

	CNN		DailyMail	
	val	test	val	test
Attentive Reader (Hermann et al., 2015)	61.6	63.0	70.5	69.0
MemNN (Hill et al., 2016)	63.4	6.8	-	-
AS Reader (Kadlec et al., 2016)	68.6	69.5	75.0	73.9
DER Network (Kobayashi et al., 2016)	71.3	72.9	-	-
Iterative Attention (Sordoni et al., 2016)	72.6	73.3	-	-
EpiReader (Trischler et al., 2016)	73.4	74.0	-	-
Stanford AR (Chen et al., 2016)	73.8	73.6	77.6	76.6
GAReader (Dhingra et al., 2016)	73.0	73.8	76.7	75.7
AoA Reader (Cui et al., 2016)	73.1	74.4	-	-
ReasoNet (Shen et al., 2016)	72.9	74.7	77.6	76.6
BIDAF (Ours)	**76.3**	**76.9**	**80.3**	**79.6**
MemNN* (Hill et al., 2016)	66.2	69.4	-	-
ASReader* (Kadlec et al., 2016)	73.9	75.4	78.7	77.7
Iterative Attention* (Sordoni et al., 2016)	74.5	75.7	-	-
GA Reader* (Dhingra et al., 2016)	76.4	77.4	79.1	78.1
Stanford AR* (Chen et al., 2016)	77.2	77.6	80.2	79.2

- Only predict start index
- Mask out non-entity words in classification layer
- For loss function: sum probability over all instances of the correct entity

BiDAF Summary

- Embeddings on multiple levels of granularity
- SQuAD: facilitated much more natural Q&A
- Bi-directional attention was new: C2Q + Q2C
- Query aware context representation without early summarization
- SOTA performance at the time

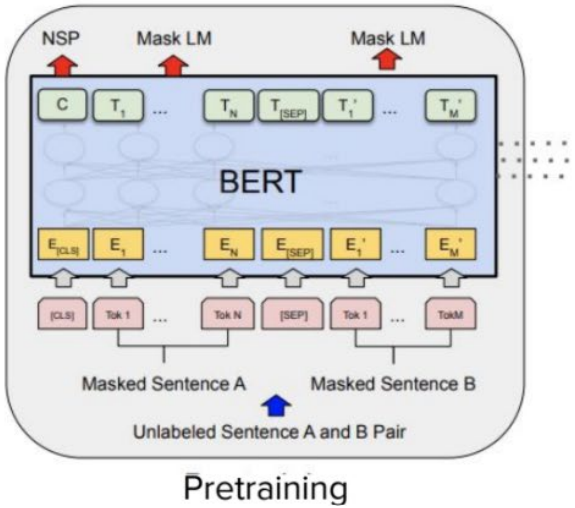
SOTA:

A SOTA score may signal to the community that you have “solved” a task that was previously unsolved, or it may signal to the community that your new method is the “best” method to solve the task, and that the rest of the community (in both academia and industry) should adopt your method as the new standard.

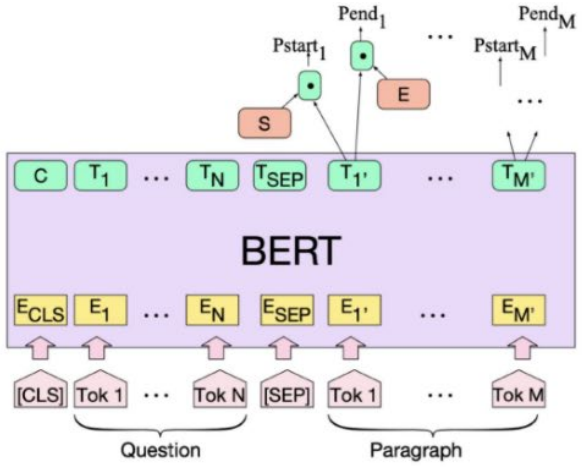
http://mitchgordon.me/ml/2021/08/31/sota.html

Current SOTA: pre-trained models

- Fine-tuning BERT
  - Two vectors S, E (fine-tuning parameters) to generate probabilities of start, end of each token.
  - SQuAd 2.0, both point to CLS.



Pretraining



Finetuning

$$Pstart_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}}$$

$$Pend_i = \frac{e^{E \cdot T_i}}{\sum_j e^{E \cdot T_j}}$$

SQuAD leader board after BERT appeared

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	**90.939**	**93.214**
2 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	90.871	93.183
3 May 16, 2021	IE-NetV2 (ensemble) *RICOH_SRCB_DML*	90.860	93.100
4 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011
5 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
5 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
5 Feb 05, 2021	FPNet (ensemble) *YuYang*	90.600	92.899
6 Apr 18, 2021	TransNets + SFVerifier + SFEnsembler (ensemble) *Senseforth AI Research* https://www.senseforth.ai/	90.487	92.894
6 Dec 01, 2020	EntitySpanFocusV2 (ensemble) *RICOH_SRCB_DML*	90.521	92.824
6 Jul 31, 2020	ATRLP+PV (ensemble) *Hithink RoyalFlush*	90.442	92.877
7 Mar 12, 2020	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.386	92.777
8 Feb 05, 2021	MixEnsemble (ensemble) *Anonymous*	90.194	92.594
9 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.115	92.580
10 Jan 12, 2021	Answer Dependent Classify (single model) *YITU*	90.059	92.517
10 Jan 30, 2021	ANet *ensemble*	90.081	92.457

Discussion: future improvements in QA

Can you think of any limitations of SQuAD (which was constructed one year after the CNN/DM work and consisting of 100,000+ questions annotated by crowd-workers)?

Future improvements in QA

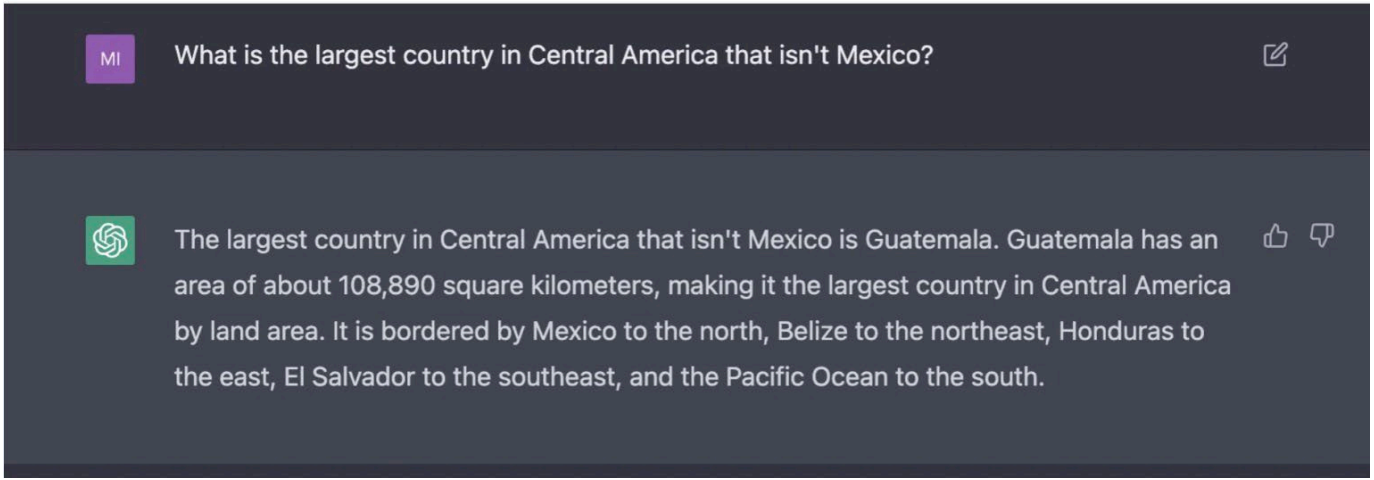
Can you think of any limitations of SQuAD (which was constructed one year after the CNN/DM work and consisting of 100,000+ questions annotated by crowd-workers)?

Possible answers:

- Only span-based answers (no yes/no, counting, implicit why). -> Explainability is needed.
- Questions were constructed looking at passages, not genuine information needs (question bias). -> Knowledge grounded QA.
- Generally greater lexical and syntactic matching between question and answer span. -> Barely any multi-fact/sentence inference.

Factoid

- Factoid: a piece of unreliable information believed to be true because of the way it is presented or repeated in print.



The easy-to-find answer here is Honduras, but for no obvious reason, I can discern, ChatGPT said Guatemala.

Summary

- QA is an important task that draws on several other components of language understanding
- Datasets are critical for reading comprehension
  - Hard to create large datasets
  - Hard to create datasets on which high performance requires “true” language understanding
- We can do well on the easier datasets but not the tougher ones yet
- The more attention, the better
  - Attentive reader < BiDAF < BERT
- Pre-training helps A LOT!