CSI2121: Big Data

Ch2. Text Understanding (part 1)

Jinyoung Yeo

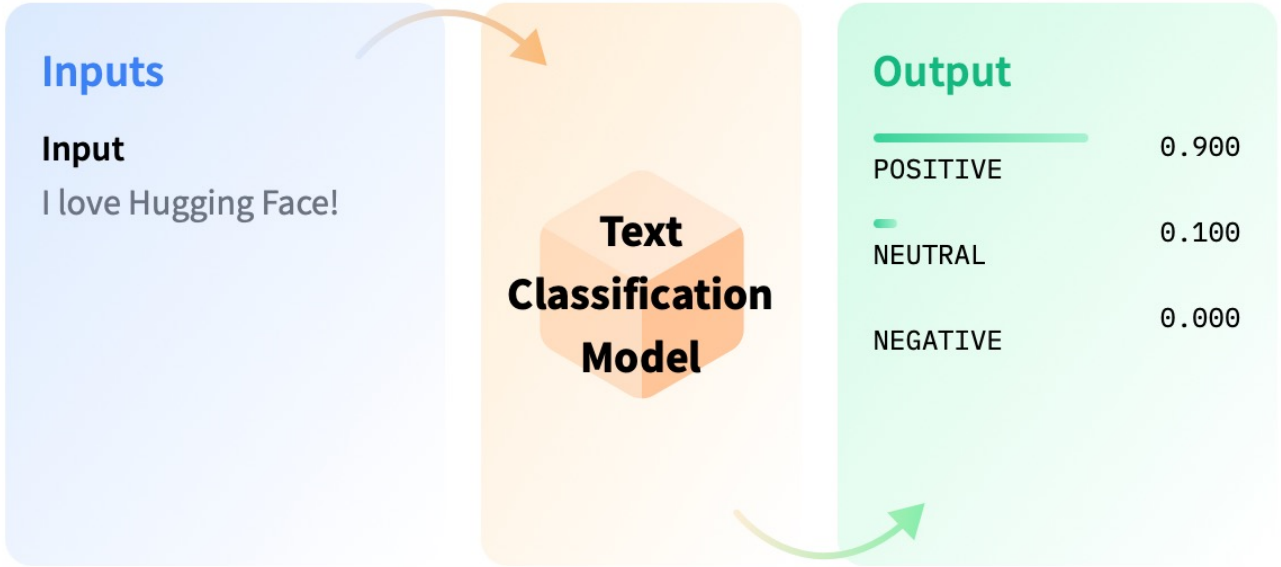
Yonsei AI

Outline

- Text classification
- Metrics
- Sentiment analysis
- Grammatical correctness
- Textual similarity
- Natural language inference
- GLUE benchmark
- KLUE benchmark

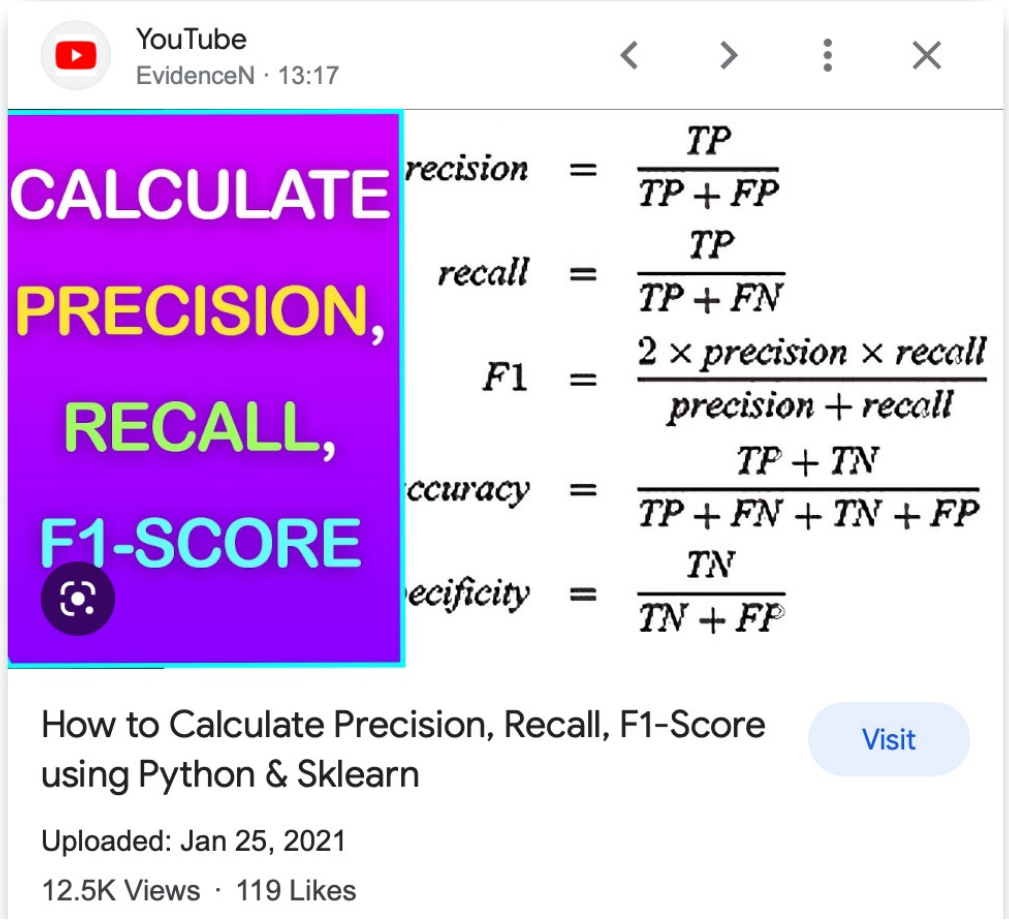
Text Classification

- Text classification is the task of assigning a label or class to a give text.
- Some use cases are sentiment analysis , natural language inference, and assessing gramamatical correctness.



Metrics

- Accuracy
- F1-Score



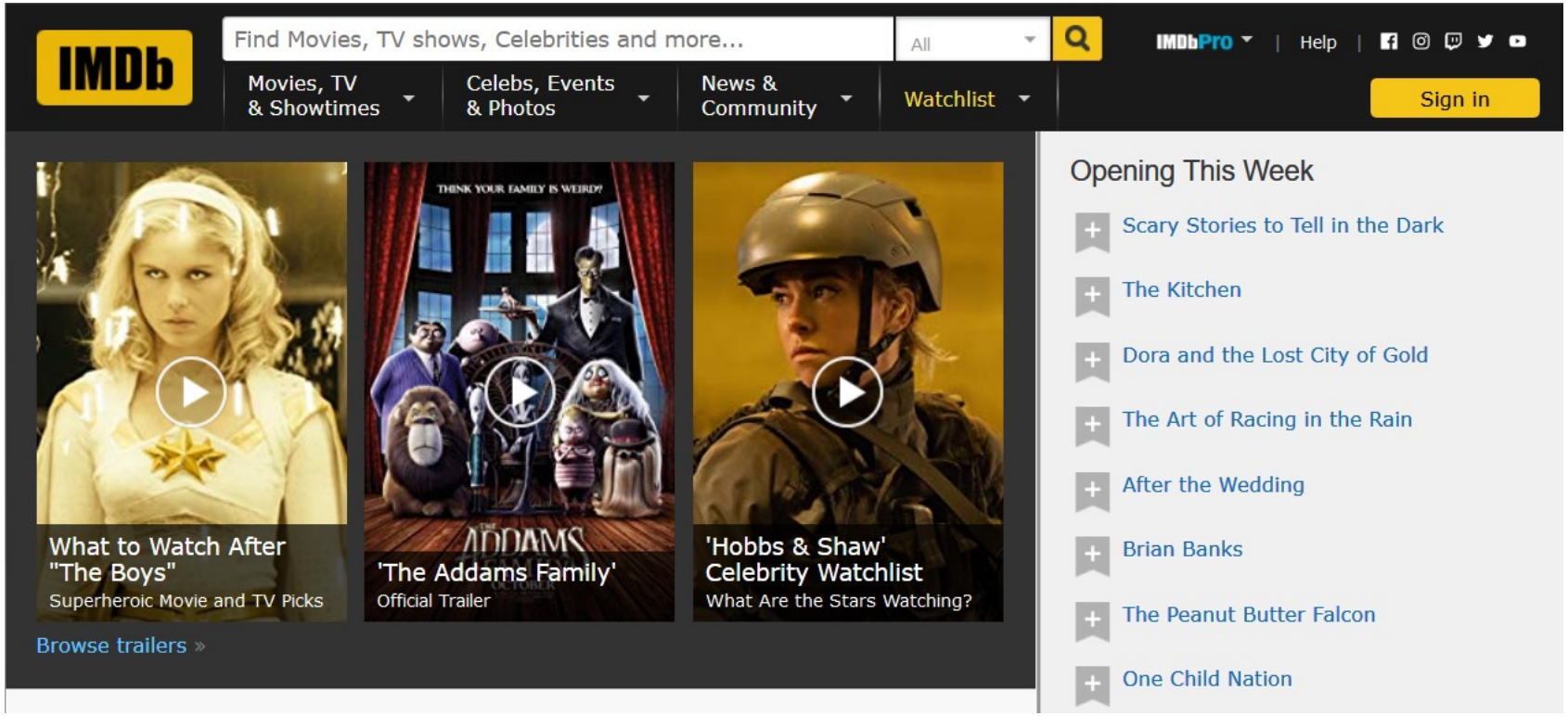
https://www.youtube.com/watch?v=9w4HJ0VUy2g

Sentiment Analysis

- In Sentiment Analysis, the classes can be polarities like positive, negative, neutral, or sentiments such as happiness or anger.
- You can track the sentiments of your customers from the product reviews using sentiment analysis models.
- This can help understand churn and retention by grouping reviews by sentiment, to later analyze the text and make strategic decisions based on this knowledge.

IMDB Large Movie Reviews Corpus

- Binary sentiment classification dataset containing 50,000 polarized (positive or negative) movie reviews.



IMDB Large Movie Reviews Corpus

- Comparably long text such as document.

```
If you like adult comedy cartoons, like South Park, then this is nearly a similar
format about the small adventures of three teenage girls at Bromwell High.
Keisha, Natella and Latrina have given exploding sweets and behaved like bitches,
I think Keisha is a good leader. There are also small stories going on with the
teachers of the school. There's the idiotic principal, Mr. Bip, the nervous
Maths teacher and many others. The cast is also fantastic, Lenny Henry's Gina
Yashere, EastEnders Chrissie Watts, Tracy-Ann Oberman, Smack The Pony's Doon
Mackichan, Dead Ringers' Mark Perry and Blunder's Nina Conti. I didn't know this
came from Canada, but it is very good. Very good!
```

Stanford Sentiment Treebank (SST)

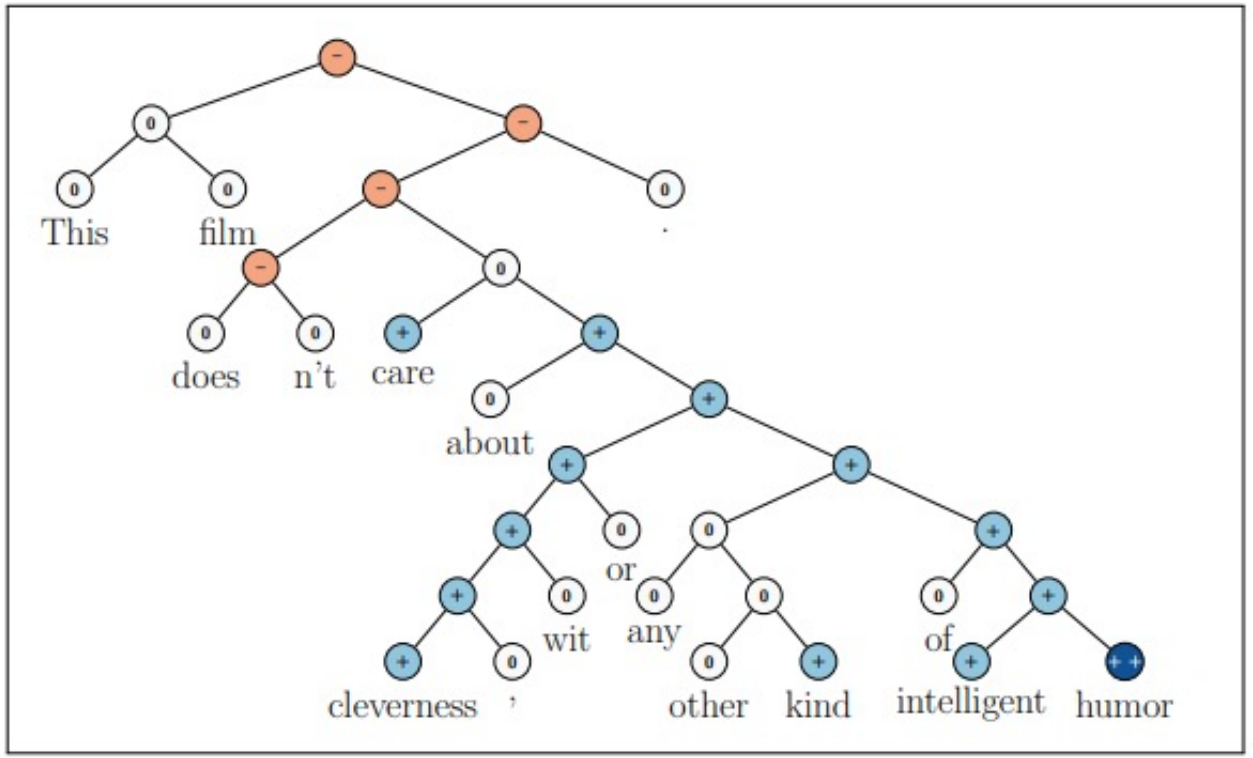


Figure 1: Example of the Recursive Neural Tensor Network accurately predicting 5 sentiment classes, very negative to very positive (– –, –, 0, +, + +), at every node of a parse tree and capturing the negation and its scope in this sentence.

Stanford Sentiment Treebank (SST)

- Fine-grained sentiment labels for 215,154 phrases in the parse trees of 11,855 sentences.
- Allows for a complete analysis of the compositional effects of sentiment in language.
- Can be reduced to predicting only the sentiment of a single sentence.

Stanford Sentiment Treebank (SST)

Model	Fine-grained		Positive/Negative	
	All	Root	All	Root
NB	67.2	41.0	82.6	81.8
SVM	64.3	40.7	84.6	79.4
BiNB	71.0	41.9	82.7	83.1
VecAvg	73.3	32.7	85.1	80.1
RNN	79.0	43.2	86.1	82.4
MV-RNN	78.7	44.4	86.8	82.9
RNTN	**80.7**	**45.7**	**87.6**	**85.4**

Table 1: Accuracy for fine grained (5-class) and binary predictions at the sentence level (root) and for all nodes.

Grammatical Correctness

- Linguistic Acceptability is the task of assessing the grammatical acceptability of a sentence.
  - The classes in this task are “acceptable” and “unacceptable”.
- The benchmark dataset used for this task is Corpus of Linguistic Acceptability (CoLA).
  - The dataset consists of texts and their labels.

Example: Books were sent to each other by the students.
Label: Unacceptable

Example: She voted for herself.
Label: Acceptable.

Quora Question Pairs

- Quora Question Pairs models assess whether two provided questions are paraphrases of each other.
  - The model takes two questions and returns a binary value, with 0 being mapped to “not paraphrase” and 1 to “paraphrase".
- The benchmark dataset is Quora Question Pairs inside the GLUE benchmark.
  - The dataset consists of question pairs and their labels.
  - Over 400,000 lines of potential question duplicate pairs.
  - The ground-truth labels contain some amount of noise: they are not guaranteed to be perfect.

Quora Question Pairs

Question1: “How can I increase the speed of my internet connection while using a VPN?”
Question2: How can Internet speed be increased by hacking through DNS?
Label: Not paraphrase

Question1: “What can make Physics easy to learn?”
Question2: “How can you make physics easy to learn?”
Label: Paraphrase

Microsoft Research Paraphrase Corpus (MRPC)

The euro rose above US$1.18, the highest price since its January 1999 launch.

The euro rose above $1.18 the highest level since its launch in January 1999.

However, without a carefully controlled study, there was little clear proof that the operation actually improves people's lives.

But without a carefully controlled study, there was little clear proof that the operation improves people's lives.

David Gest has sued his estranged wife Liza Minelli for %MONEY% million for beating him when she was drunk

Liza Minelli's estranged husband is taking her to court for %MONEY% million after saying she threw a lamp at him and beat him in drunken rages

Microsoft Research Paraphrase Corpus (MRPC)

- 5801 pairs of sentences, each accompanied by a binary judgment indicating whether human raters considered the pair of sentences to be similar enough in meaning to be considered close paraphrases.
- Small and unbalanced class distribution, for examples, 68% of pairs are labeled as positive.
  - But enable to test data noise/scarcity scenarios.

Semantic Textual Similarity (STS)

- Collection of sentence pairs drawn from news headlines, video and image captions, and NLI data.
- Unlike MRPC, the assessment of pairs of sentences is based on their degree of semantic similarity (i.e., regression).

5	*The two sentences are completely equivalent, as they mean the same thing.*
	The bird is bathing in the sink. Birdie is washing itself in the water basin.
4	*The two sentences are mostly equivalent, but some unimportant details differ.*
	Two boys on a couch are playing video games. Two boys are playing a video game.
3	*The two sentences are roughly equivalent, but some important information differs/missing.*
	John said he is considered a witness but not a suspect. "He is not a suspect anymore." John said.
2	*The two sentences are not equivalent, but share some details.*
	They flew out of the nest in groups. They flew into the nest together.
1	*The two sentences are not equivalent, but are on the same topic.*
	The woman is playing the violin. The young lady enjoys listening to the guitar.
0	*The two sentences are completely dissimilar.*
	The black dog is running through the snow. A race car driver is driving his car through the mud.

Table 1: Similarity scores with explanations and English examples from Agirre et al. (2013).

- Each pair is human-annotated with a similarity score from 1 to 5; the task is to predict theses scores.
- Evaluation metrics: Pearson and Spearman correlation coefficients.

Natural Language Inference (NLI)

- In NLI the model determines the relationship between two given texts. Concretely, the model takes a premise and a hypothesis and returns a class that can either be:
  - entailment, which means the hypothesis is true.
  - contraction, which means the hypothesis is false.
  - neutral, which means there's no relation between the hypothesis and the premise.
- The benchmark dataset for this task is GLUE (General Language Understanding Evaluation). NLI models have different variants, such as Multi-Genre NLI, Question NLI and Winograd NLI.

Stanford NLI (SNLI)

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

Stanford NLI (SNLI)

- 570k human-written English sentence pairs manually labeled for balanced classification with the labels entailment, contradiction, and neutral.
- Neural model also must handle phenomena like lexical entailment, quantification, coreference, tense, belief, modality, and lexical and syntactic ambiguity.
- SNLI is not sufficiently demanding to serve as an effective benchmark for NLU because of its limited domain (image captions).

Multi-Genre NLI (MNLI)

- MNLI is used for general NLI. Here are some examples:

Example 1:
Premise: A man inspects the uniform of a figure in some East Asian country.
Hypothesis: The man is sleeping.
Label: Contradiction

Example 2:
Premise: Soccer game with multiple males playing.
Hypothesis: Some men are playing a sport.
Label: Entailment

Met my first girlfriend that way.	FACE-TO-FACE **contradiction** C C N C	I didn't meet my first girlfriend until later.
8 million in relief in the form of emergency housing.	GOVERNMENT **neutral** N N N N	The 8 million dollars for emergency housing was still not enough to solve the problem.
Now, as children tend their gardens, they have a new appreciation of their relationship to the land, their cultural heritage, and their community.	LETTERS **neutral** N N N N	All of the children love working in their gardens.
At 8:34, the Boston Center controller received a third transmission from American 11	9/11 **entailment** E E E E	The Boston Center controller got a third transmission from American 11.
I am a lacto-vegetarian.	SLATE **neutral** N N E N	I enjoy eating cheese too much to abstain from dairy.
someone else noticed it and i said well i guess that's true and it was somewhat melodious in other words it wasn't just you know it was really funny	TELEPHONE **contradiction** C C C C	No one noticed and it wasn't funny at all.

Table 1: Randomly chosen examples from the development set of our new corpus, shown with their genre labels, their selected gold labels, and the validation labels (abbreviated E, N, C) assigned by individual annotators.

- Longer text
- Written and Spoken language
- Additional domains (genre)
- Hypothesis sentences in MultiNLI generally cannot be derived from their premise sentences

Genre
SNLI
FICTION
GOVERNMENT
SLATE
TELEPHONE
TRAVEL
9/11
FACE-TO-FACE
LETTERS
OUP
VERBATIM

Question Natural Language Inference

- QNLI is the task of determining if the answer to a certain question can be found in a given document.
  - If the answer can be found the label is “entailment”.
  - If the answer cannot be found the label is “not entailment".

Question: What percentage of marine life died during the extinction?
Sentence: It is also known as the “Great Dying” because it is considered the largest mass extinction in the Earth’s history.
Label: not entailment

Question: Who was the London Weekend Television’s Managing Director?
Sentence: The managing director of London Weekend Television (LWT), Greg Dyke, met with the representatives of the "big five" football clubs in England in 1990.
Label: entailment

GLUE Benchmark

- The General Language Understanding Evaluation (GLUE) benchmark is a collection of resources for training, evaluating, and analyzing natural language understanding systems.

GLUE Benchmark

- A benchmark of nine sentence- or sentence-pair language understanding tasks built on established existing datasets and selected to cover a diverse range of dataset sizes, text genres, and degrees of difficulty,
- A diagnostic dataset designed to evaluate and analyze model performance with respect to a wide range of linguistic phenomena found in natural language, and
- A public leaderboard for tracking performance on the benchmark and a dashboard for visualizing the performance of models on the diagnostic set.

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

Corpus	|Train|	|Test|	Task	Metrics	Domain
			Single-Sentence Tasks		
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks		
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks		
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

Table 1: Task descriptions and statistics. All tasks are single sentence or sentence pair classification, except STS-B, which is a regression task. MNLI has three classes; all other classification tasks have two. Test sets shown in bold use labels that have never been made public in any form.

As of Mar 20, 2023

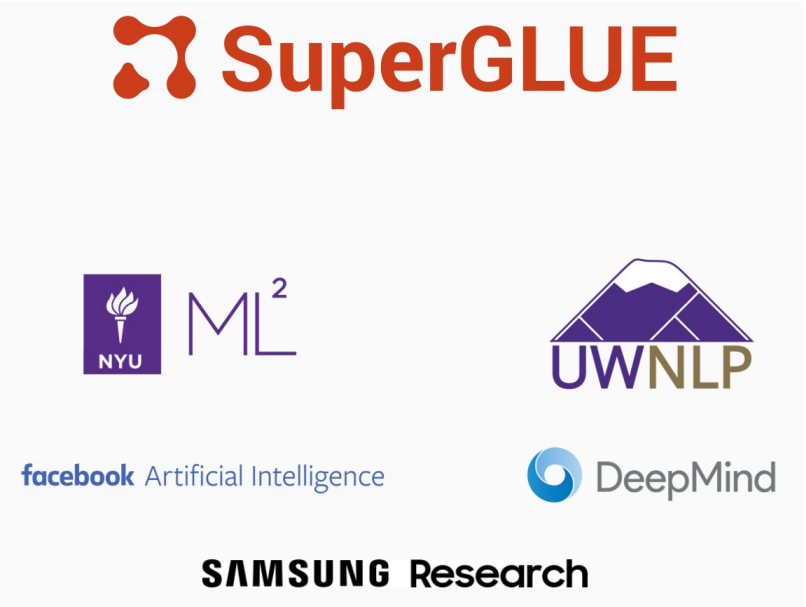
Rank	Name	Model	URL	Score
1	Microsoft Alexander v-team	Turing ULR v6		91.3
2	JDExplore d-team	Vega v1		91.3
3	Microsoft Alexander v-team	Turing NLR v5		91.2
4	DIRL Team	DeBERTa + CLEVER		91.1
5	ERNIE Team - Baidu	ERNIE		91.1
6	AliceMind & DIRL	StructBERT + CLEVER		91.0
7	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.8
8	HFL iFLYTEK	MacALBERT + DKM		90.7
9	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6
10	T5 Team - Google	T5		90.3

As of Mar 20, 2023

11	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9
12	Huawei Noah's Ark Lab	NEZHA-Large		89.8
13	LG AI Research	ANNA		89.8
14	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)		89.7
15	ELECTRA Team	ELECTRA-Large + Standard Tricks		89.4
16	David Kim	2digit LANet		89.3
17	倪仕文	DropAttack-RoBERTa-large		88.8
18	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4
19	Junjie Yang	HIRE-RoBERTa		88.3
20	Shiwen Ni	ELECTRA-large-M (bert4keras)		88.3
21	Facebook AI	RoBERTa		88.1
22	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6
23	GLUE Human Baselines	GLUE Human Baselines		87.1

SuperGLUE Benchmark

- We take into account the lessons learnt from original GLUE benchmark and present SuperGLUE, a new benchmark styled after GLUE with a new set of more difficult language understanding tasks, improved resources, and a new public leaderboard.



SuperGLUE Benchmark

As of Mar 20, 2023

Rank	Name	Model	URL	Score
1	JDExplore d-team	Vega v2		91.3
2	Liam Fedus	ST-MoE-32B		91.2
3	Microsoft Alexander v-team	Turing NLR v5		90.9
4	ERNIE Team - Baidu	ERNIE 3.0		90.6
5	Yi Tay	PaLM 540B		90.4
6	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4
7	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3
8	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8
9	T5 Team - Google	T5		89.3
10	SPoT Team - Google	Frozen T5 1.1 + SPoT		89.2

SuperGLUE Tasks

Name	Identifier	Download	More Info	Metric
Broadcoverage Diagnostics	AX-b			Matthew's Corr
CommitmentBank	CB			Avg. F1 / Accuracy
Choice of Plausible Alternatives	COPA			Accuracy
Multi-Sentence Reading Comprehension	MultiRC			F1a / EM
Recognizing Textual Entailment	RTE			Accuracy
Words in Context	WiC			Accuracy
The Winograd Schema Challenge	WSC			Accuracy
BoolQ	BoolQ			Accuracy
Reading Comprehension with Commonsense Reasoning	ReCoRD			F1 / Accuracy
Winogender Schema Diagnostics	AX-g			Gender Parity / Accuracy

DOWNLOAD ALL DATA

KLUE Benchmark

- Korean Language Understanding Evaluation (KLUE) benchmark is a series of datasets to evaluate natural language understanding capability of Korean language models.
- KLUE consists of 8 diverse and representative tasks, which are accessible to anyone without any restrictions.

KLUE Benchmark

- KLUE benchmark is composed of 8 tasks:
  - Topic Classification (TC)
  - Sentence Textual Similarity (STS)
  - Natural Language Inference (NLI)
  - Named Entity Recognition (NER)
  - Relation Extraction (RE)
  - (Part-Of-Speech) + Dependency Parsing (DP)
  - Machine Reading Comprehension (MRC)
  - Dialogue State Tracking (DST)

As of Mar 20, 2023

KLUE Leaderboard

Unlike other benchmarks, klue benchmarks do not provide total scores and leaderboards for the entire task. On the leaderboard, you can check each score for one model and sort by each evaluation metric.

All | Small Size | Base Size | Large Size

#	Team	Model	Description	YNAT	KLUE-STS		KLUE-NLI	KLUE-NER		KLUE-RE		KLUE-DP		KLUE-MRC		WOS	
				F1	R^P	F1	ACC	$F1^E$	$F1^C$	$F1^{mic}$	AUC	UAS	LAS	EM	ROUGE	JGA	$F1^S$
1	KLUE-team	KLUE-BERT-base	More	85.73	90.85	82.84	81.63	83.97	91.39	66.44	66.17	89.96	88.05	62.32	68.51	46.64	91.61
2	KLUE-team	KLUE-RoBERTa-large	More	85.69	93.35	86.63	89.17	85	91.86	71.13	72.98	93.48	88.36	75.58	80.59	50.22	92.23
3	KLUE-team	KLUE-RoBERTa-base	More	85.07	92.5	85.4	84.83	84.6	91.44	67.65	68.55	93.04	88.32	68.67	73.98	47.49	91.64
4	KLUE-team	KLUE-RoBERTa-small	More	84.98	91.54	85.16	79.33	83.65	91.14	60.89	58.96	90.04	88.14	57.32	62.7	46.62	91.44
5	KLUE-tester		More	79.63	88.51	81.22	67.03	81.07	89.39	44.86	31.99	89.58	88.03	40.74	45.86	2.44	48.04

Rows per page: 5 1-5 of 5

Other Tasks

- Assigning subject categories, topics, or genres
- Spam detection
- Age/gender identification
- Language identification

How to solve?

Positive Seeds

achieve
advantage
better
creative
efficiency
efficiently
enhance
greater
...

Negative Seeds

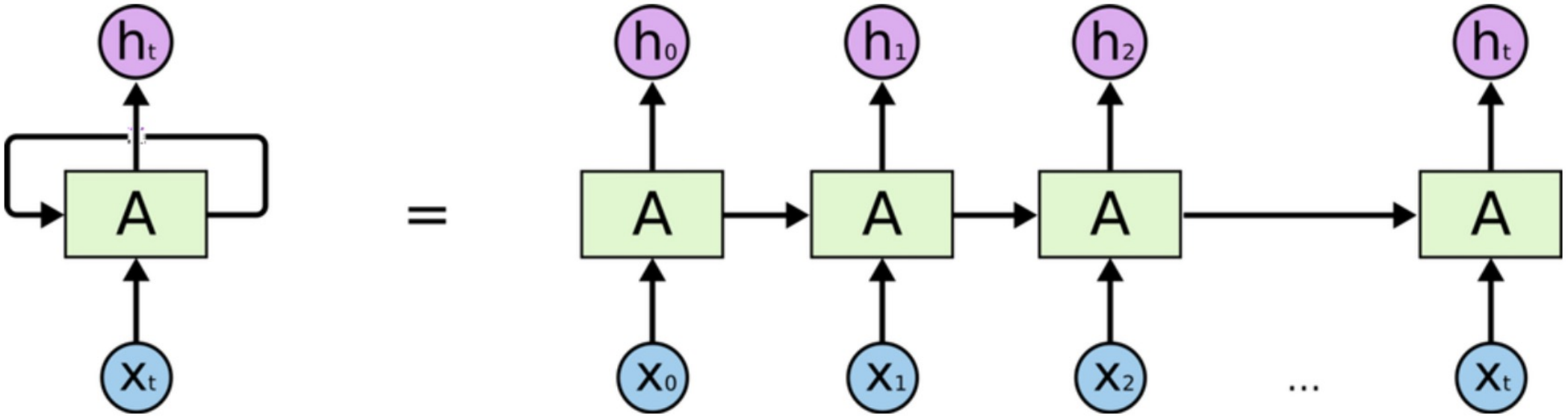
adverse
adversely
against
complaint
concern
damages
default
deficiencies
...

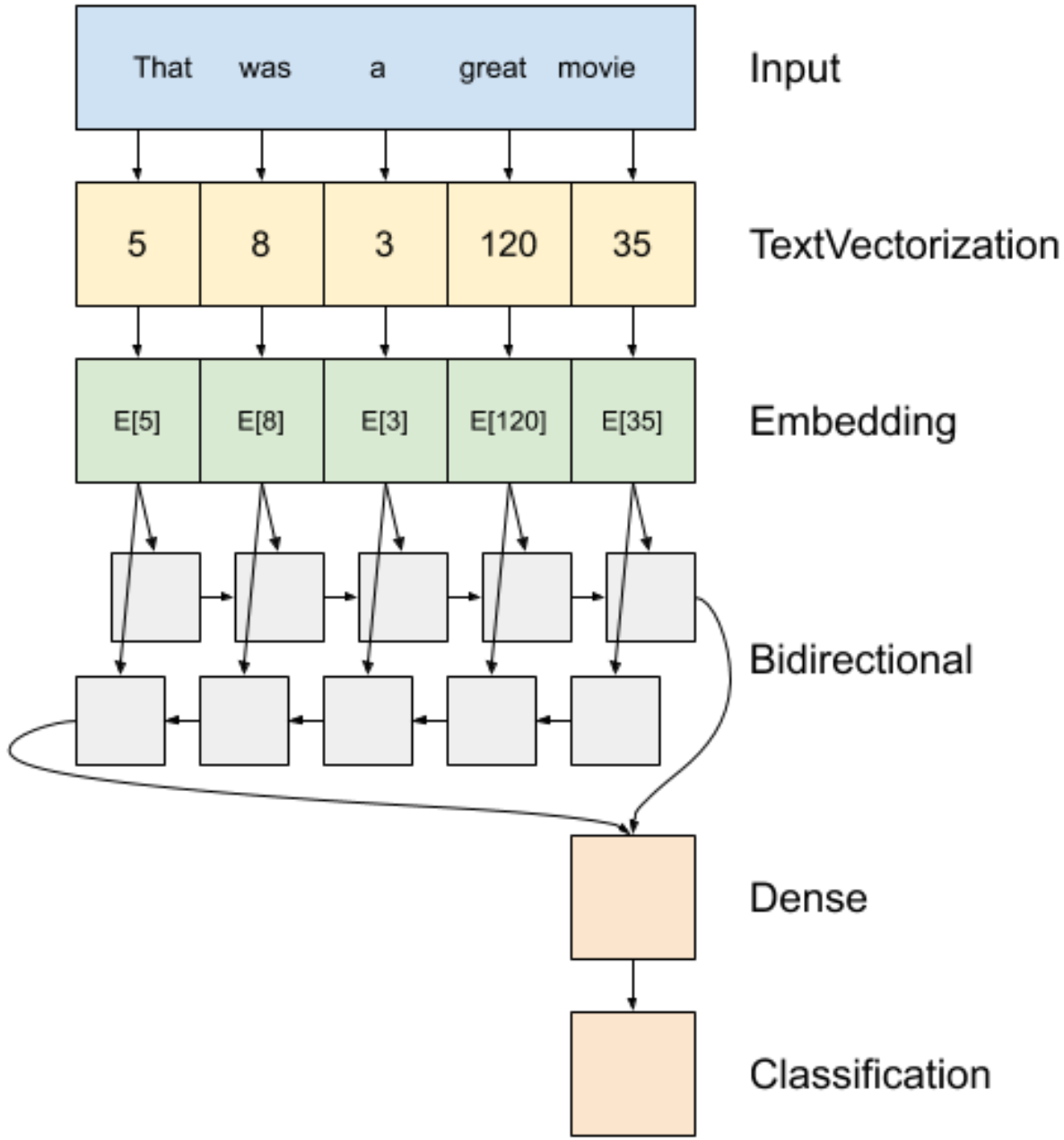
Generate

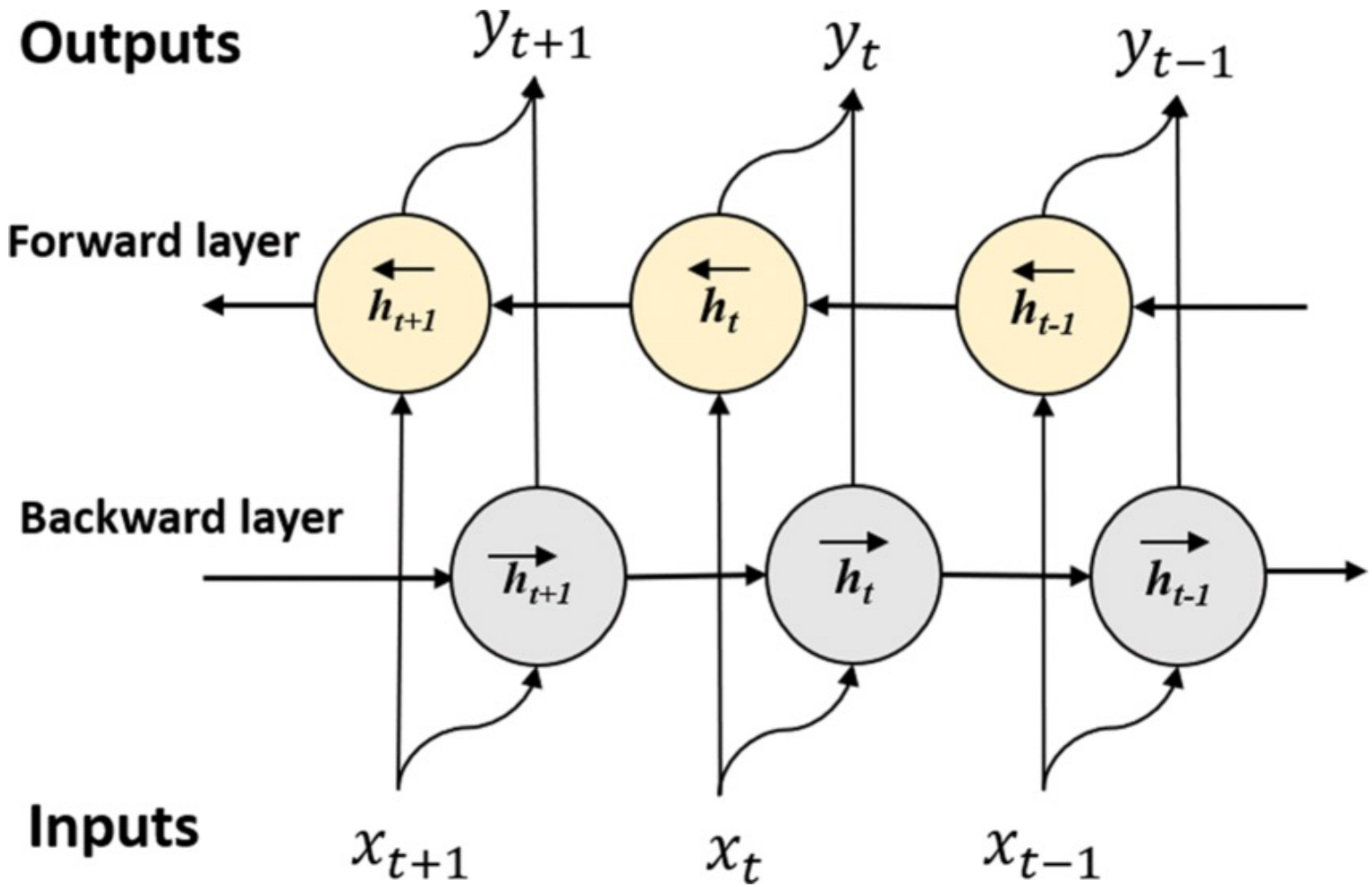
Positive Words

Negative Words

How to solve?




That was a great movie
Input
5 8 3 120 35
TextVectorization
E[5] E[8] E[3] E[120] E[35]
Embedding
Bidirectional
Dense
Classification


Outputs
y_{t+1}
y_t
y_{t-1}
Forward layer
$\overleftarrow{h}_{t+1}$
$\overleftarrow{h}_t$
$\overleftarrow{h}_{t-1}$
Backward layer
$\overrightarrow{h}_{t+1}$
$\overrightarrow{h}_t$
$\overrightarrow{h}_{t-1}$
Inputs
x_{t+1}
x_t
x_{t-1}

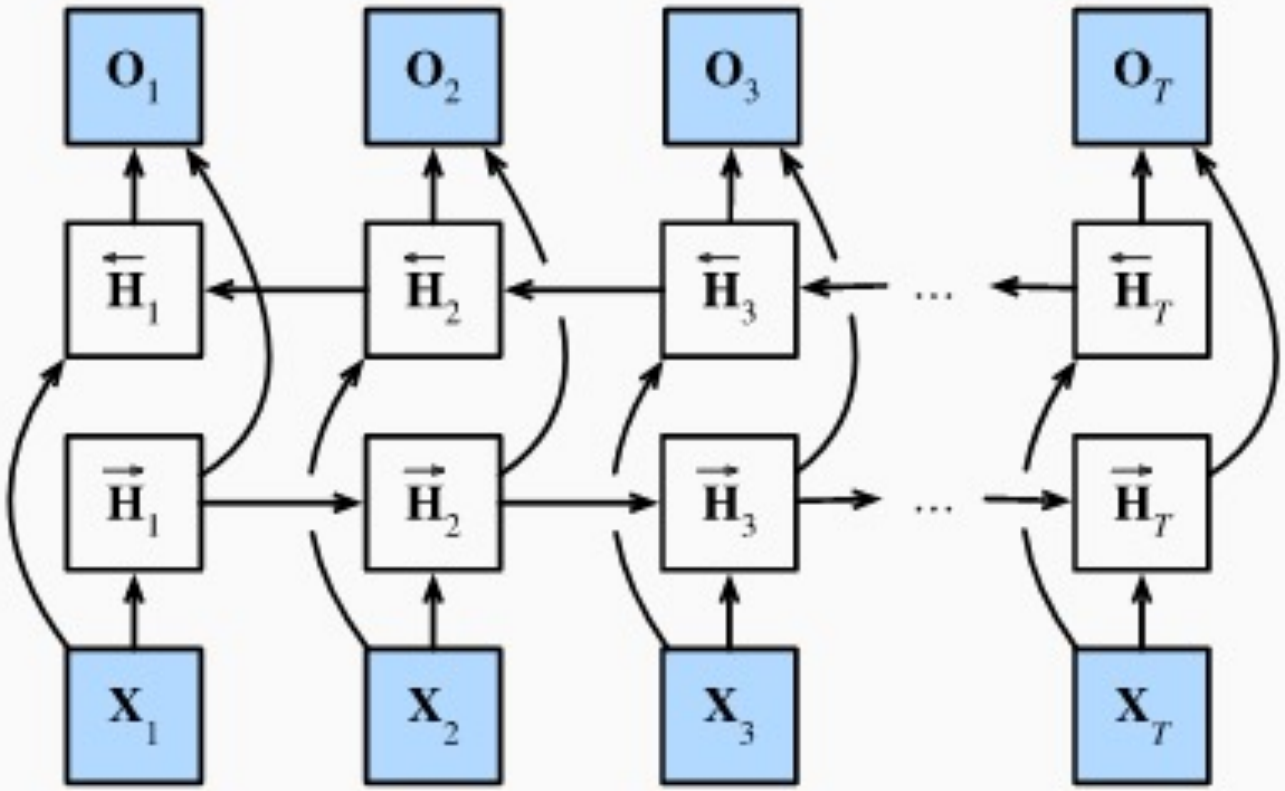
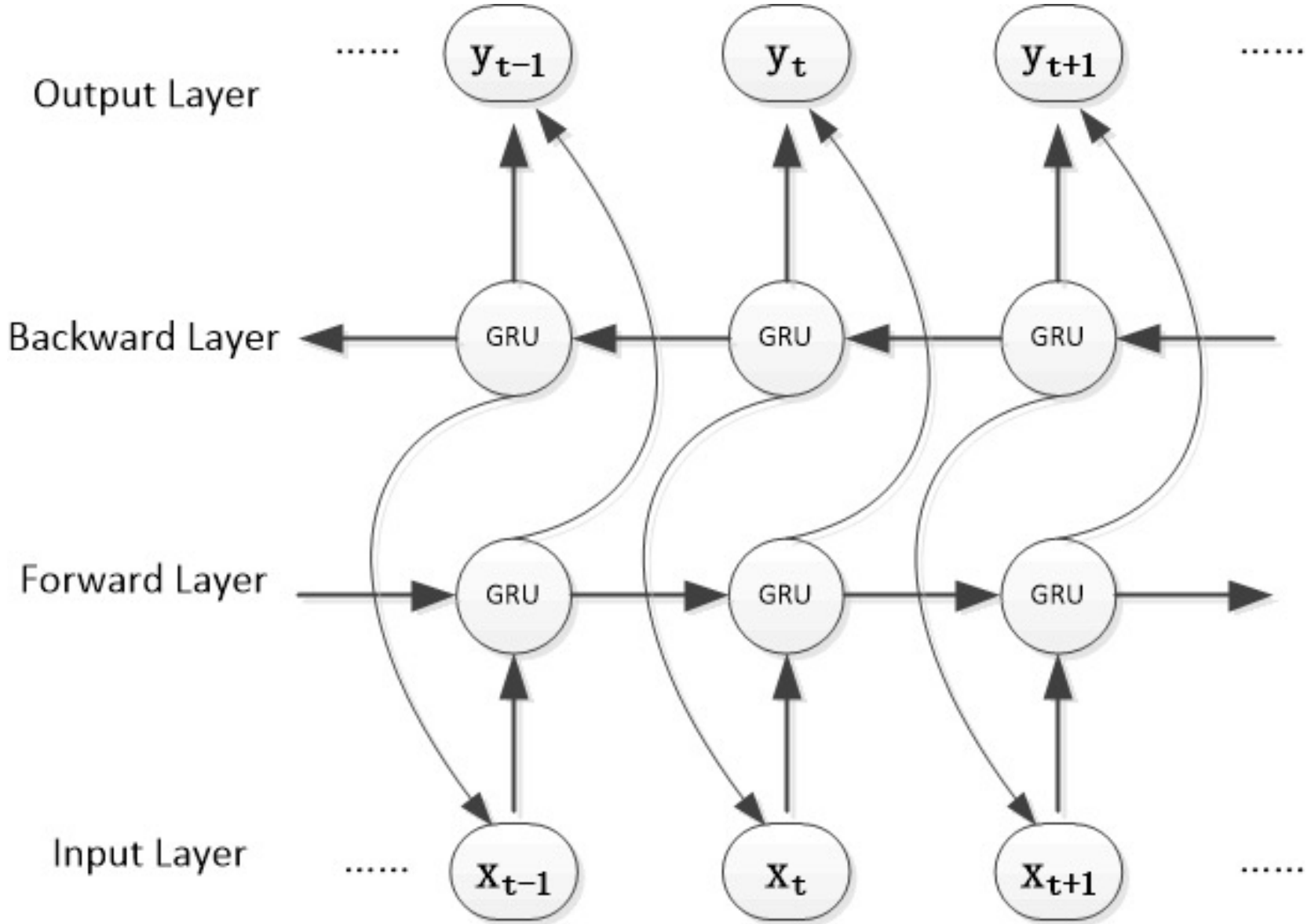
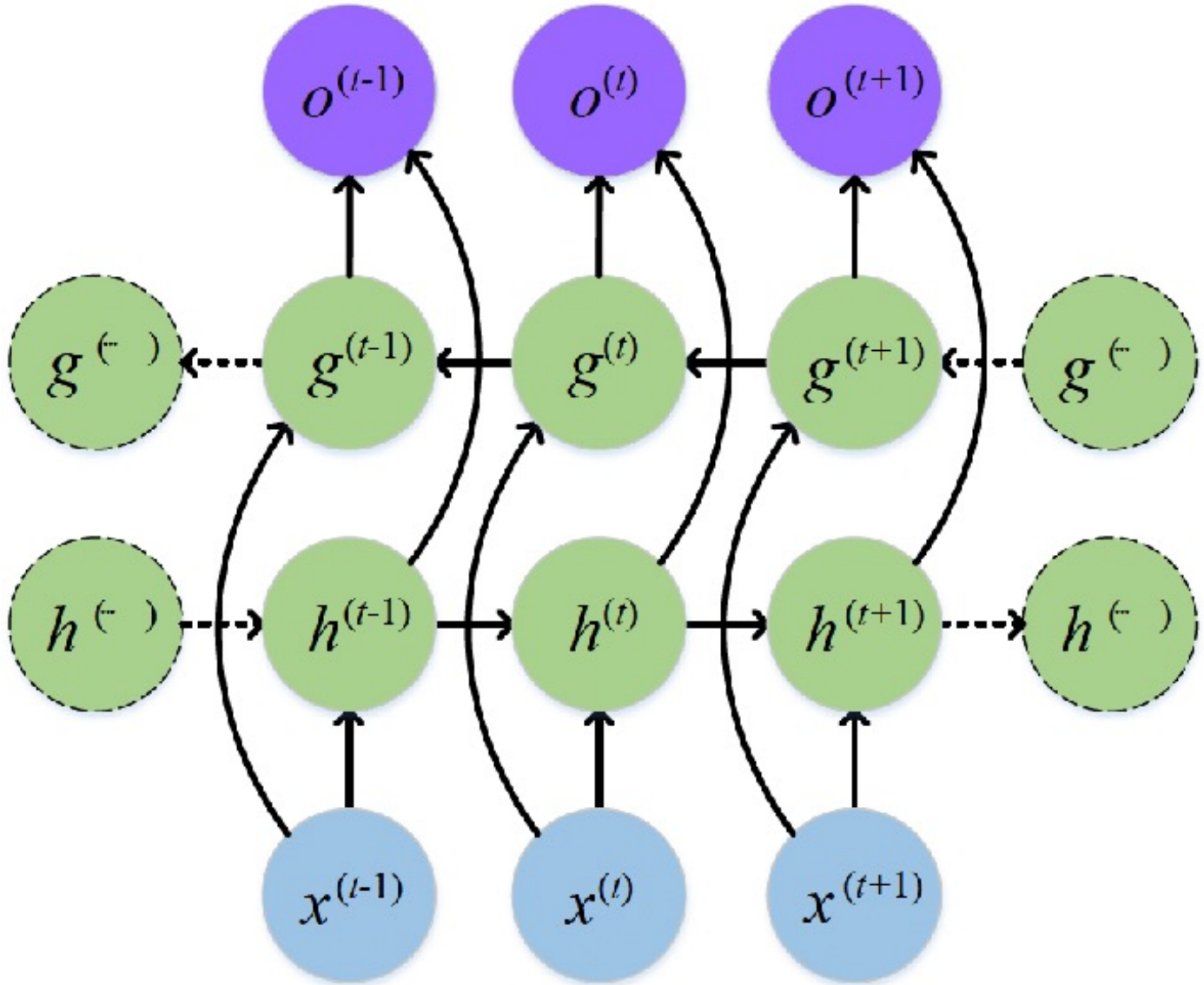
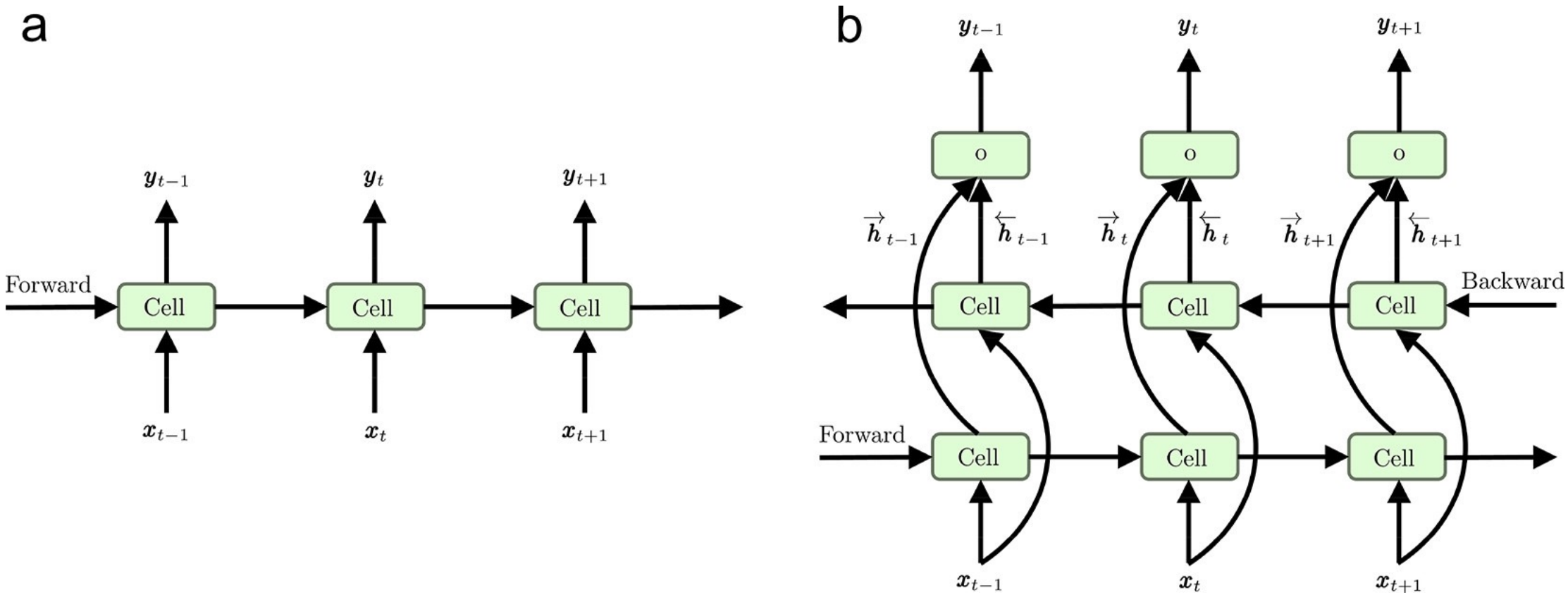


Fig. 9.4.2 Architecture of a bidirectional RNN.


Output Layer
Backward Layer
Forward Layer
Input Layer
GRU
GRU
GRU
GRU
GRU
GRU
y_{t-1}
y_t
y_{t+1}
x_{t-1}
x_t
x_{t+1}


$o^{(t-1)}$
$o^{(t)}$
$o^{(t+1)}$
$g^{(\leftarrow)}$
$g^{(t-1)}$
$g^{(t)}$
$g^{(t+1)}$
$g^{(\leftarrow)}$
$h^{(\leftarrow)}$
$h^{(t-1)}$
$h^{(t)}$
$h^{(t+1)}$
$h^{(\leftarrow)}$
$x^{(t-1)}$
$x^{(t)}$
$x^{(t+1)}$


a
y_{t-1}
y_t
y_{t+1}
Forward
Cell
Cell
Cell
x_{t-1}
x_t
x_{t+1}
b
y_{t-1}
y_t
y_{t+1}
o
o
o
$\overrightarrow{h}_{t-1}$
$\overleftarrow{h}_{t-1}$
$\overrightarrow{h}_t$
$\overleftarrow{h}_t$
$\overrightarrow{h}_{t+1}$
$\overleftarrow{h}_{t+1}$
Backward
Cell
Cell
Cell
Forward
Cell
Cell
Cell
x_{t-1}
x_t
x_{t+1}

Reference

- https://huggingface.co/tasks/text-classification
- https://gluebenchmark.com
- https://huffon.github.io/2019/11/16/glue/

CSI2121: Big Data

Ch3. Text Understanding (part 2)

Jinyoung Yeo

Yonsei AI

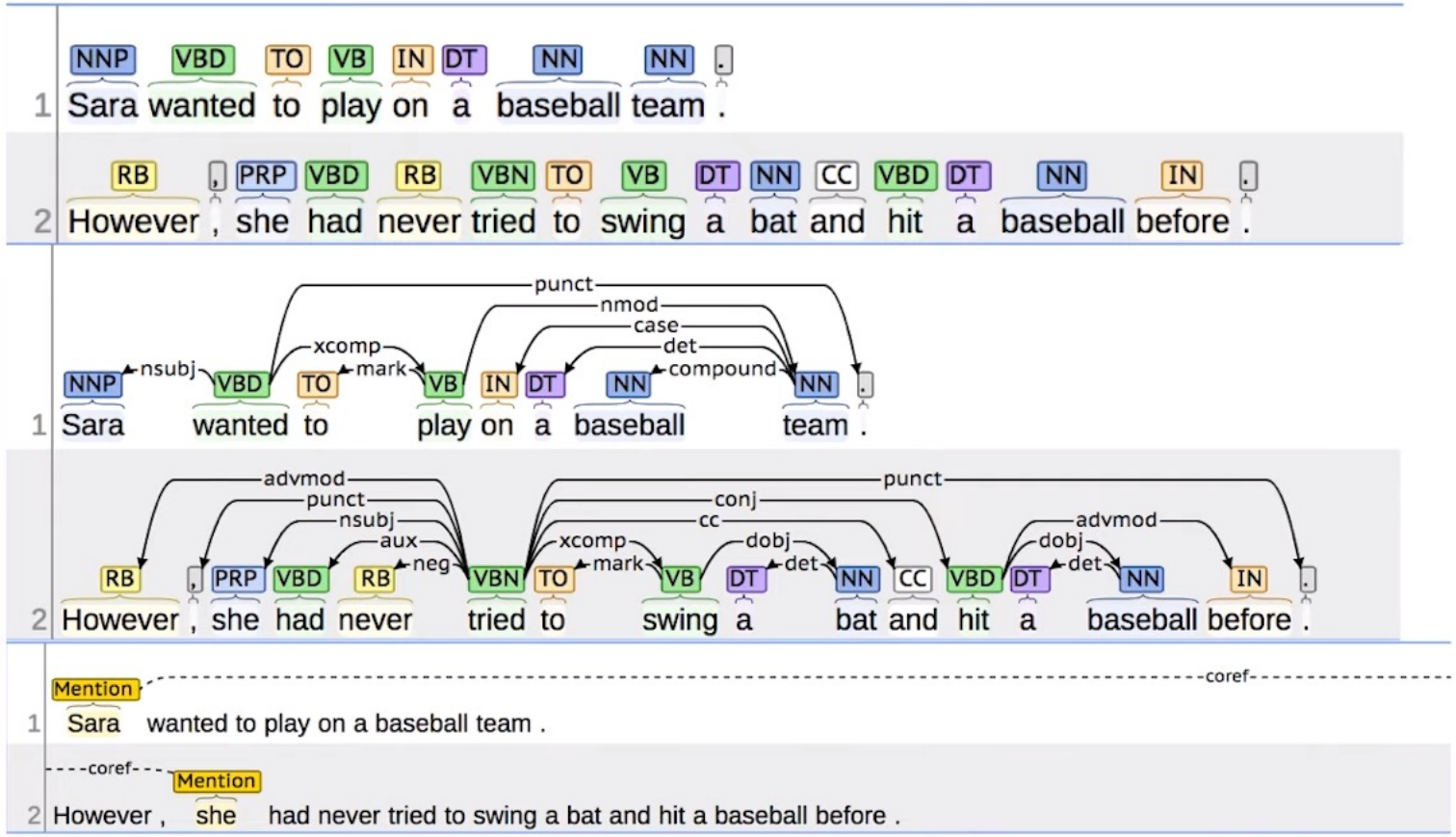
Outline

- Reading comprehension
- Question answering
- CNN/Daily mail dataset
- Attentive reader
- Stanford question answering dataset
- Open-domain QA

Teaching Machines to Read

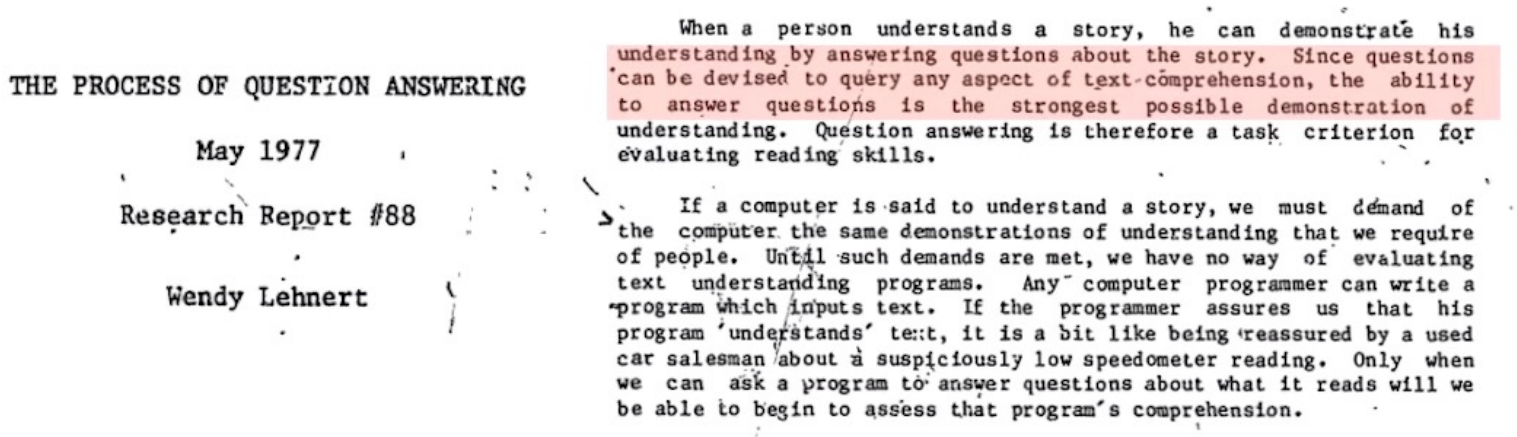
Sara wanted to play on a baseball team . However, she had never tried to swing a bat and hit a baseball before .

Teaching Machines to Read



Reading Comprehension as Question Answering

Wendy Lehnert. 1977. "The Process of Question Answering"



THE PROCESS OF QUESTION ANSWERING

May 1977

Research Report #88

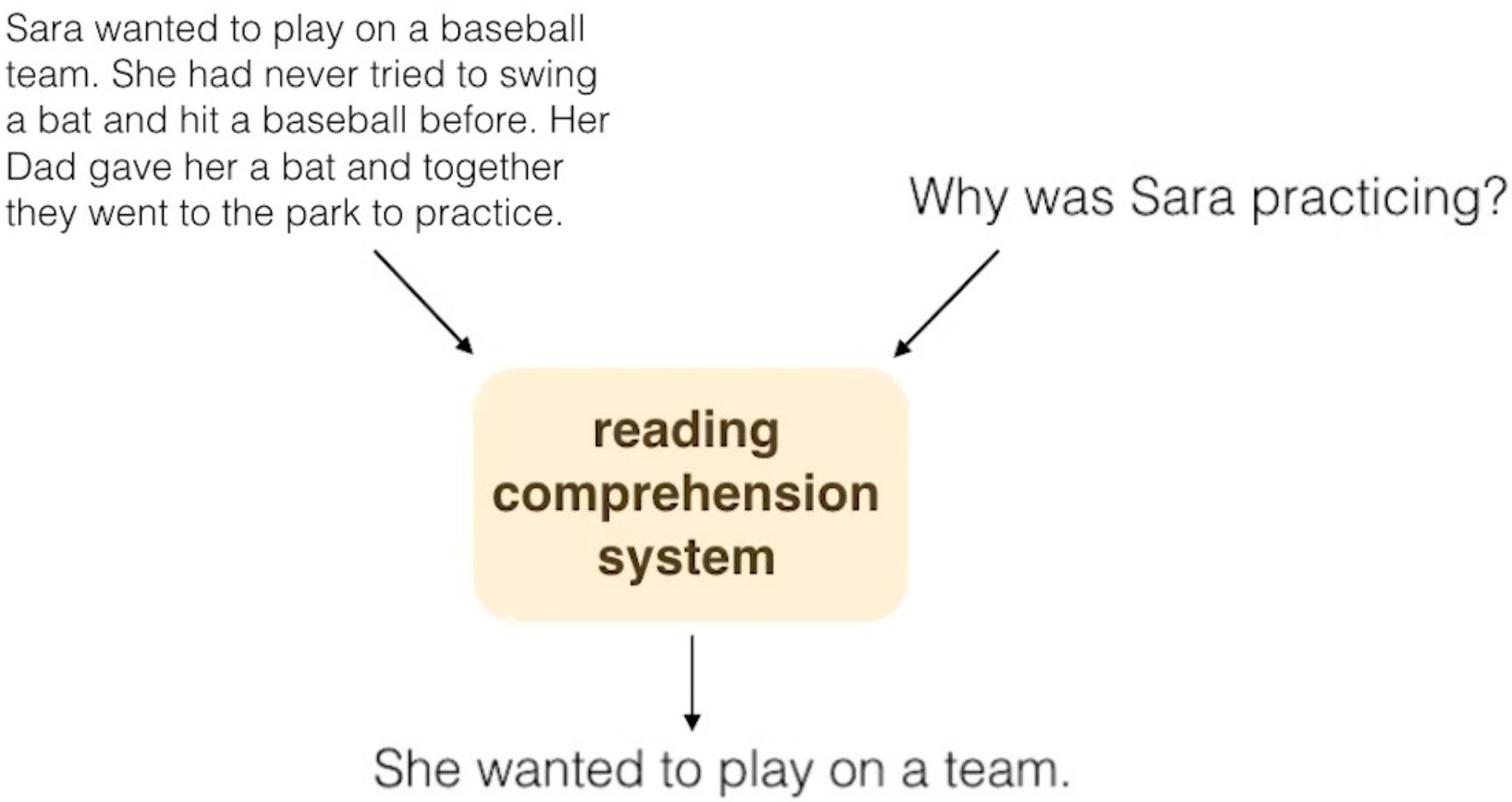
Wendy Lehnert

When a person understands a story, he can demonstrate his understanding by answering questions about the story. Since questions can be devised to query any aspect of text-comprehension, the ability to answer questions is the strongest possible demonstration of understanding. Question answering is therefore a task criterion for evaluating reading skills.

If a computer is said to understand a story, we must demand of the computer the same demonstrations of understanding that we require of people. Until such demands are met, we have no way of evaluating text understanding programs. Any computer programmer can write a program which inputs text. If the programmer assures us that his program 'understands' text, it is a bit like being reassured by a used car salesman about a suspiciously low speedometer reading. Only when we can ask a program to answer questions about what it reads will we be able to begin to assess that program's comprehension.

"Since questions can be devised to query **any aspect** of text comprehension, the ability to answer questions is the **strongest possible demonstration of understanding**."

Reading Comprehension as Question Answering



Reading comprehension is a “new” field

Before **2015**, we hadn’t had any statistical NLP systems which are capable of reading a simple passage and answering questions.

Datasets (passage, question, answer)

- MCTest: 2600 questions
- ProcessBank: 500 questions

Systems

- Hand-built systems
- Classifier with linguistic features

Reading comprehension is a “new” field

Since 2015…

Datasets (passage, question, answer)

CNN/Daily Mail

SQuAD

LAMBADA

WDW

CBT

MSMARCO

Maluuba NewsQA

WikiReading

TriviaQA

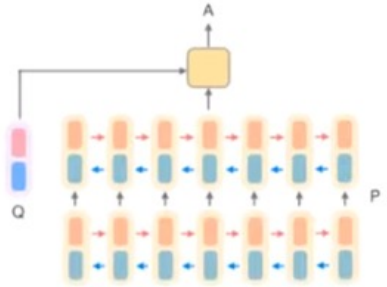
SearchQA

RACE

More than 100k examples!

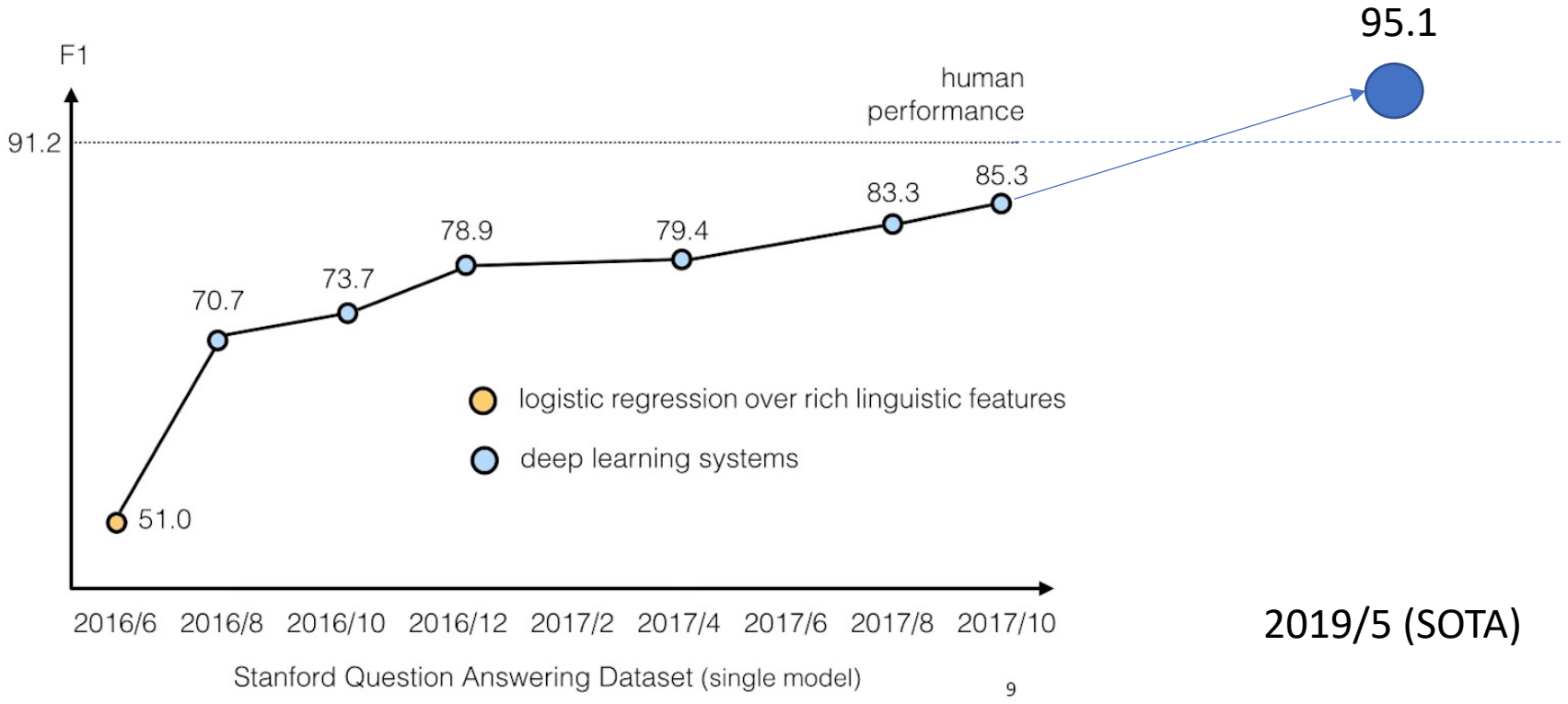
Systems

End-to-end neural networks!



8

Progress is rapid!



2019/5 (SOTA)

SQuAD1.1 Leaderboard

Here are the ExactMatch (EM) and F1 scores evaluated on the test set of SQuAD v1.1.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Jul 24, 2021	{ANNA} (single model) *LG AI Research*	**90.622**	**95.719**
2 Apr 10, 2020	LUKE (single model) *Studio Ousia & NAIST & RIKEN AIP* https://arxiv.org/abs/2010.01057	90.202	95.379
3 May 21, 2019	XLNet (single model) *Google Brain & CMU*	89.898	95.080
4 Dec 11, 2019	XLNET-123++ (single model) *MST/EOI* http://tia.today	89.856	94.903
4 Aug 11, 2019	XLNET-123 (single model) *MST/EOI*	89.646	94.930
5 Jul 21, 2019	SpanBERT (single model) *FAIR & UW*	88.839	94.635
6 Jul 03, 2019	BERT+WWM+MT (single model) *Xiaoi Research*	88.650	94.393
7 Jul 21, 2019	Tuned BERT-1seq Large Cased (single model) *FAIR & UW*	87.465	93.294

SQuAD2.0

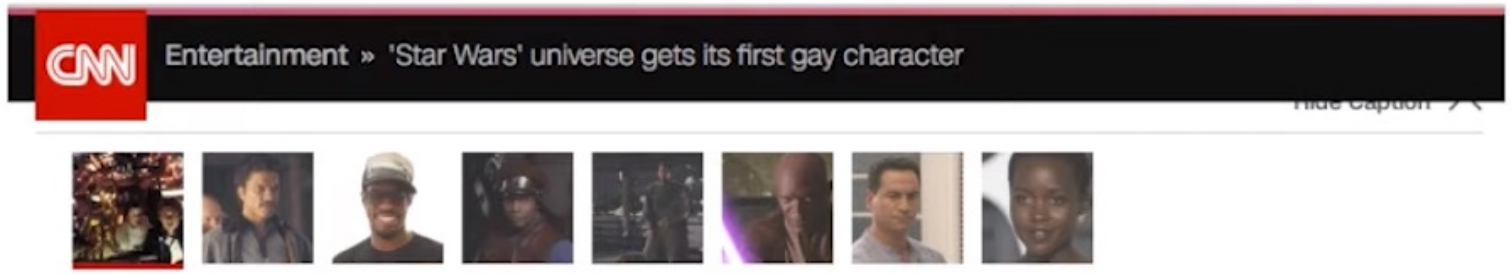
The Stanford Question Answering Dataset

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

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

CNN/Daily Mail Datsets



CNN Entertainment » 'Star Wars' universe gets its first gay character

Hide Caption

Story highlights

Official "Star Wars" universe gets its first gay character, a lesbian governor

The character appears in the upcoming novel "Lords of the Sith"

Characters in [redacted] movies have gradually become more diverse

(CNN) — If you feel a ripple in the Force today, it may be the news that the official Star Wars universe is getting its first gay character.

According to the sci-fi website Big Shiny Robot, the upcoming novel "Lords of the Sith" will feature a capable but flawed Imperial official named Moff Mors who "also happens to be a lesbian."

The character is the first gay figure in the official Star Wars universe -- the movies, television shows, comics and books approved by Star Wars franchise owner Disney -- according to Shelly Shapiro, editor of "Star Wars" books at Random House imprint Del Rey Books.

CNN/Daily Mail Datsets

passage

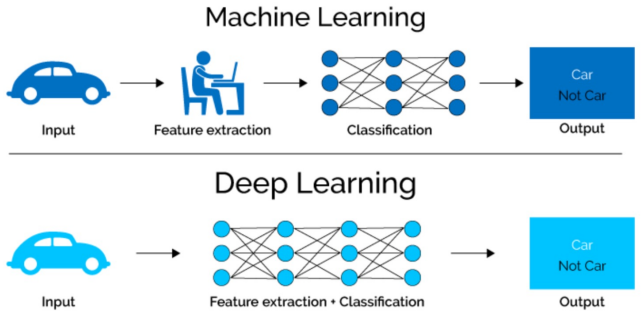
(@entity4) if you feel a ripple in the force today , it may be the news that the official @entity6 is getting its first gay character . according to the sci-fi website @entity9 , the upcoming novel " @entity11 " will feature a capable but flawed @entity13 official named @entity14 who " also happens to be a lesbian . " the character is the first gay figure in the official @entity6 -- the movies , television shows , comics and books approved by @entity6 franchise owner @entity22 -- according to @entity24 , editor of " @entity6 " books at @entity28 imprint @entity26 .

question characters in " @placeholder " movies have gradually become more diverse

answer @entity6

CNN: 380k, Daily Mail: 879k training - free!

A Cateogrical-feature Classifier



Can we build a simple, end-to-end neural network to tackle this problem?

How is it different from feature-based linear classifiers?

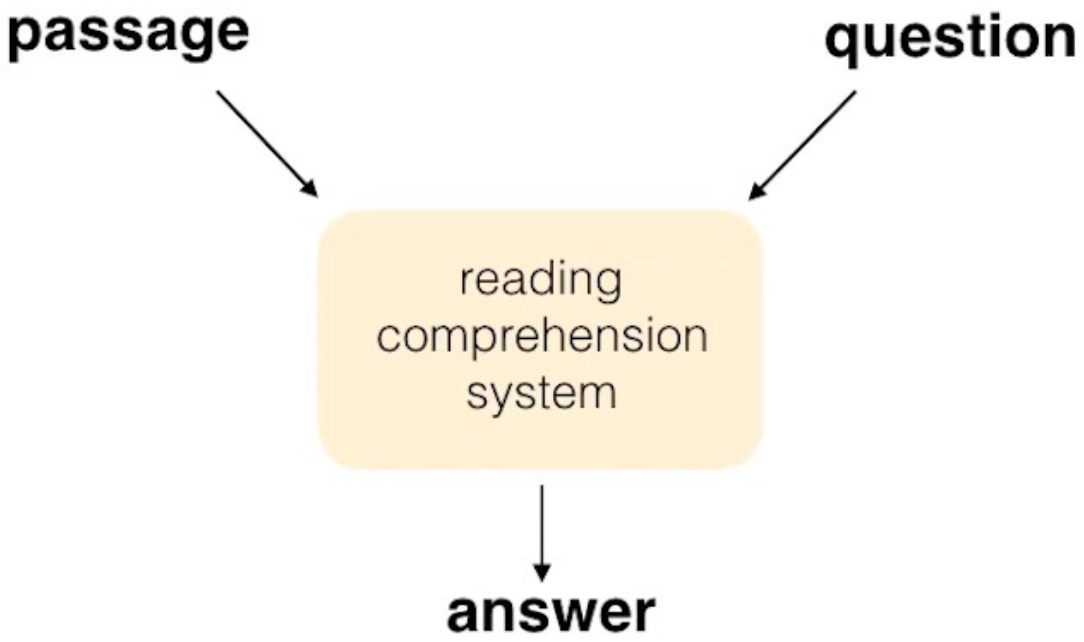
For each candidate entity *e*, we build a symbolic feature vector:

$$f_{P,Q}(e) > f_{P,Q}(e')$$

1. Whether *e* occurs in *P*
2. Whether *e* occurs in *Q*
3. Frequency of *e* in *P*
4. First position of *e* in *P*

5. Whether e co-occurs with another *Q* word in *P*.
6. word distance
7. **n-gram** exact match
8. **dependency parse** match

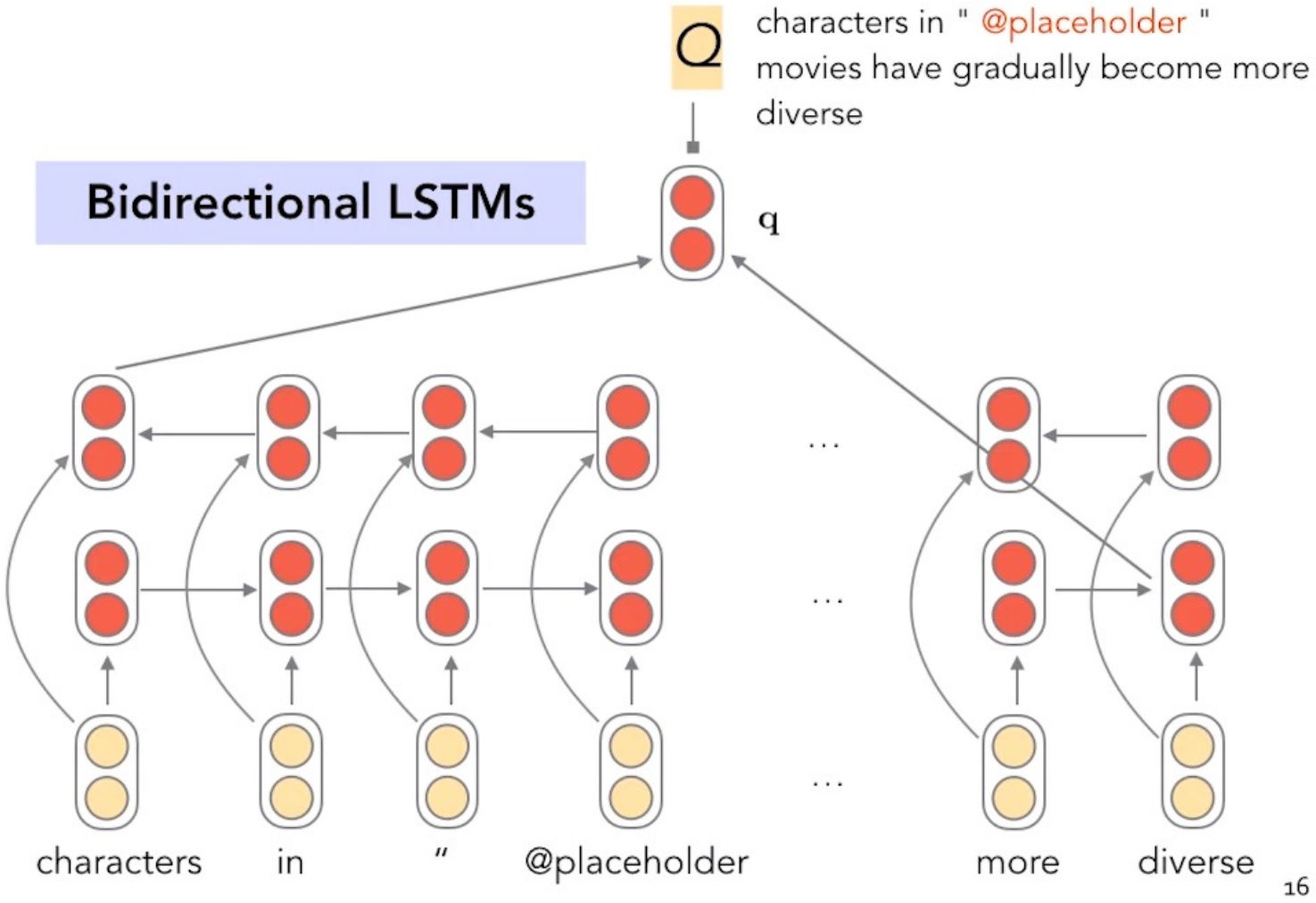
Attentive Reader



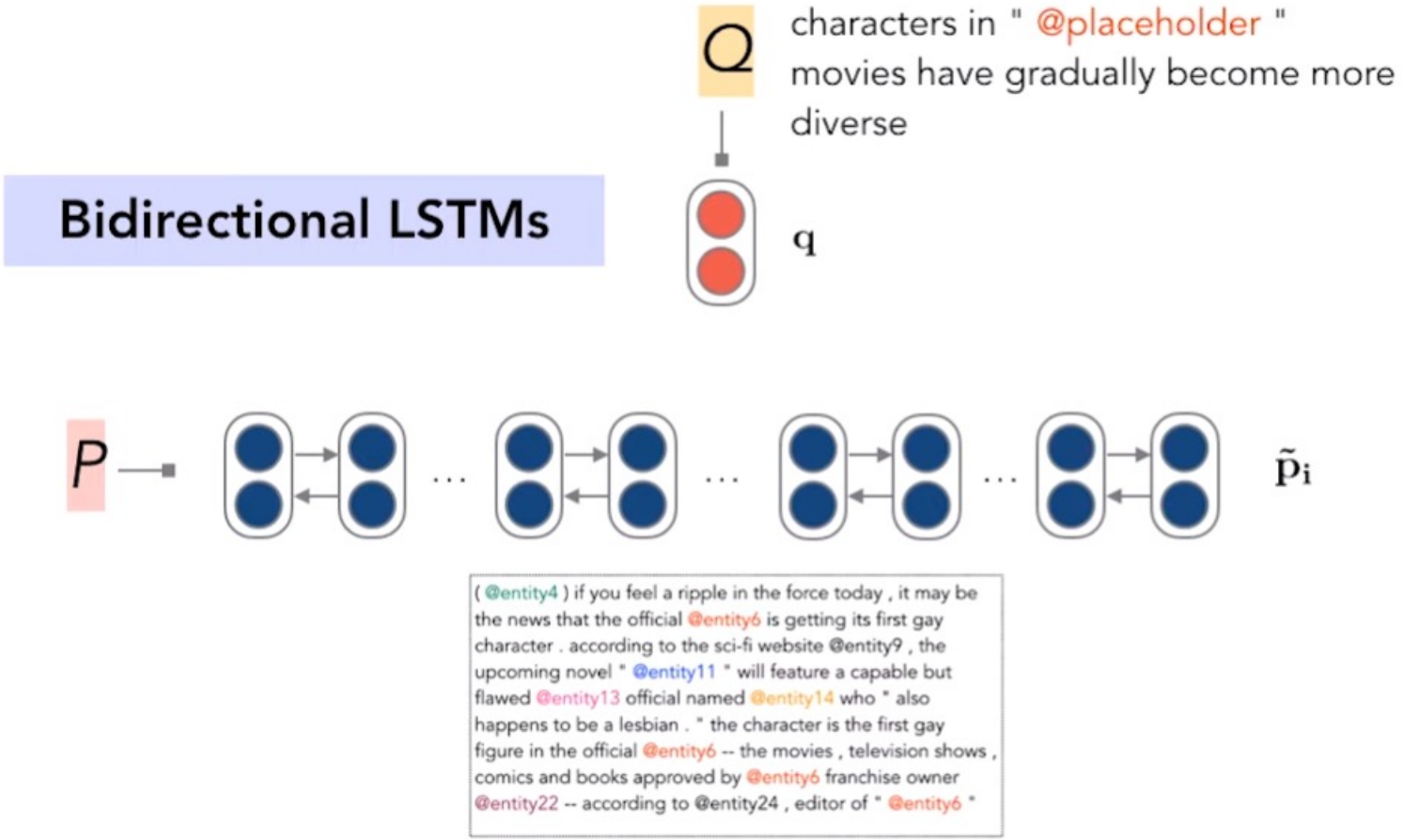
1. Encode the question
2. Encode the passage
3. Model the interaction between passage and question
4. Infer the answer

… all in vector space!

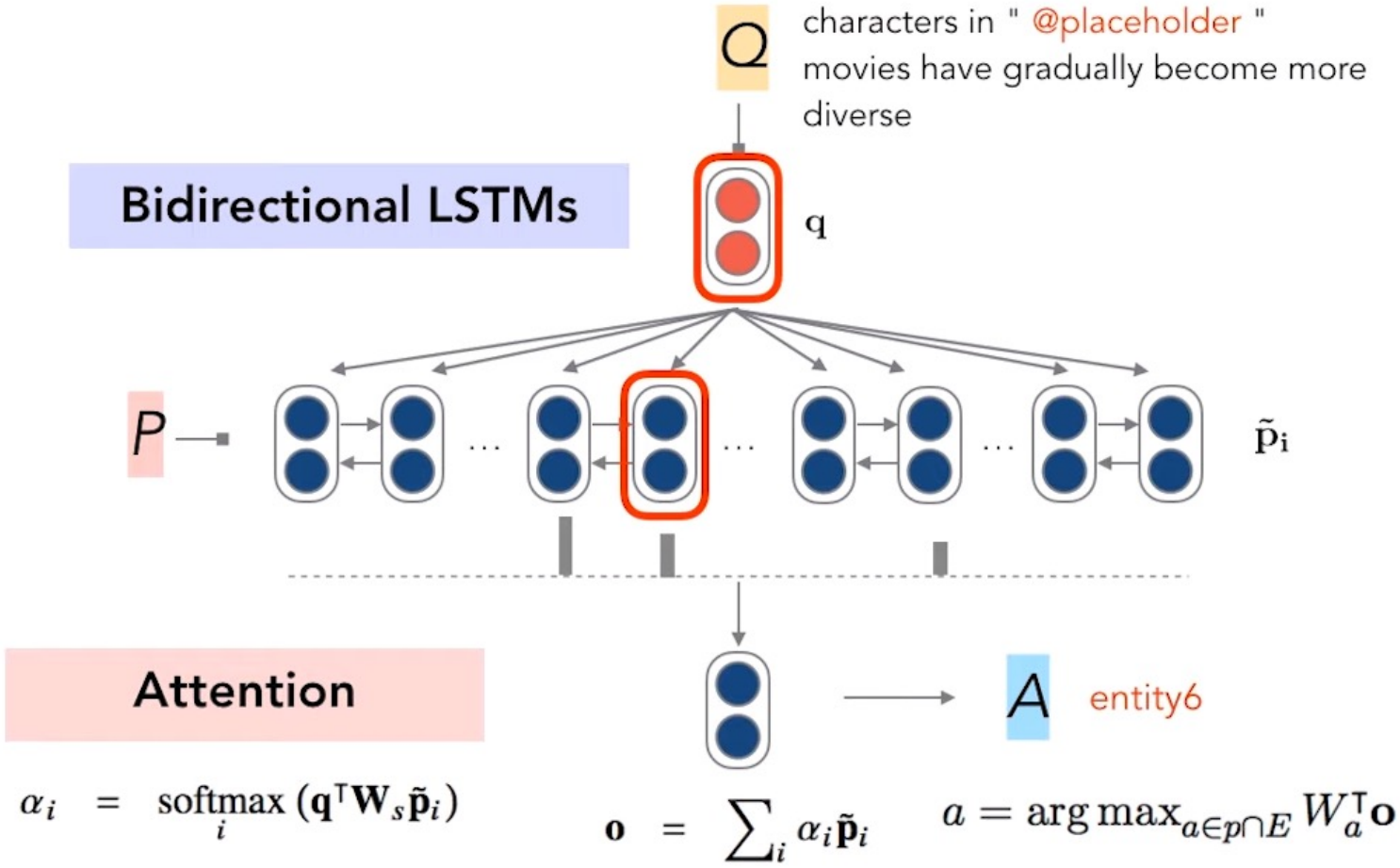
Attentive Reader



Attentive Reader



“Stanford” Attentive Reader (2016)



Analysis

Categorical feature classifier	67.9	68.3
Stanford: Attentive Reader	**73.6**	**76.6**

What level of language understanding is needed?
What have current models learned?

1. exact match
2. paraphrasing
3. partial clue
4. multiple sentence
5. coreference errors
6. ambiguous/hard cases

Analysis

exact match

@placeholder and @entity8 were tried together and convicted of murder , but now cleared

@entity12 and @entity8 were tried together and convicted of murder by two separate courts .

paraphrasing

@placeholder says he understands why @entity0 won't play at this tournament

"entity0 called me personally to let me know that he wouldn't be be playing here at @entity23" **entity3** said.

Analysis

partial clue

a tv movie based on @entity12 's book " **@placeholder** " casts a @entity76 actor as @entity5

@entity12 professed that his "**entity11**" is not a religious book…

multiple sentences

" he is doing a his - and - her duet all by himself, " @entity6 said of **@placeholder**

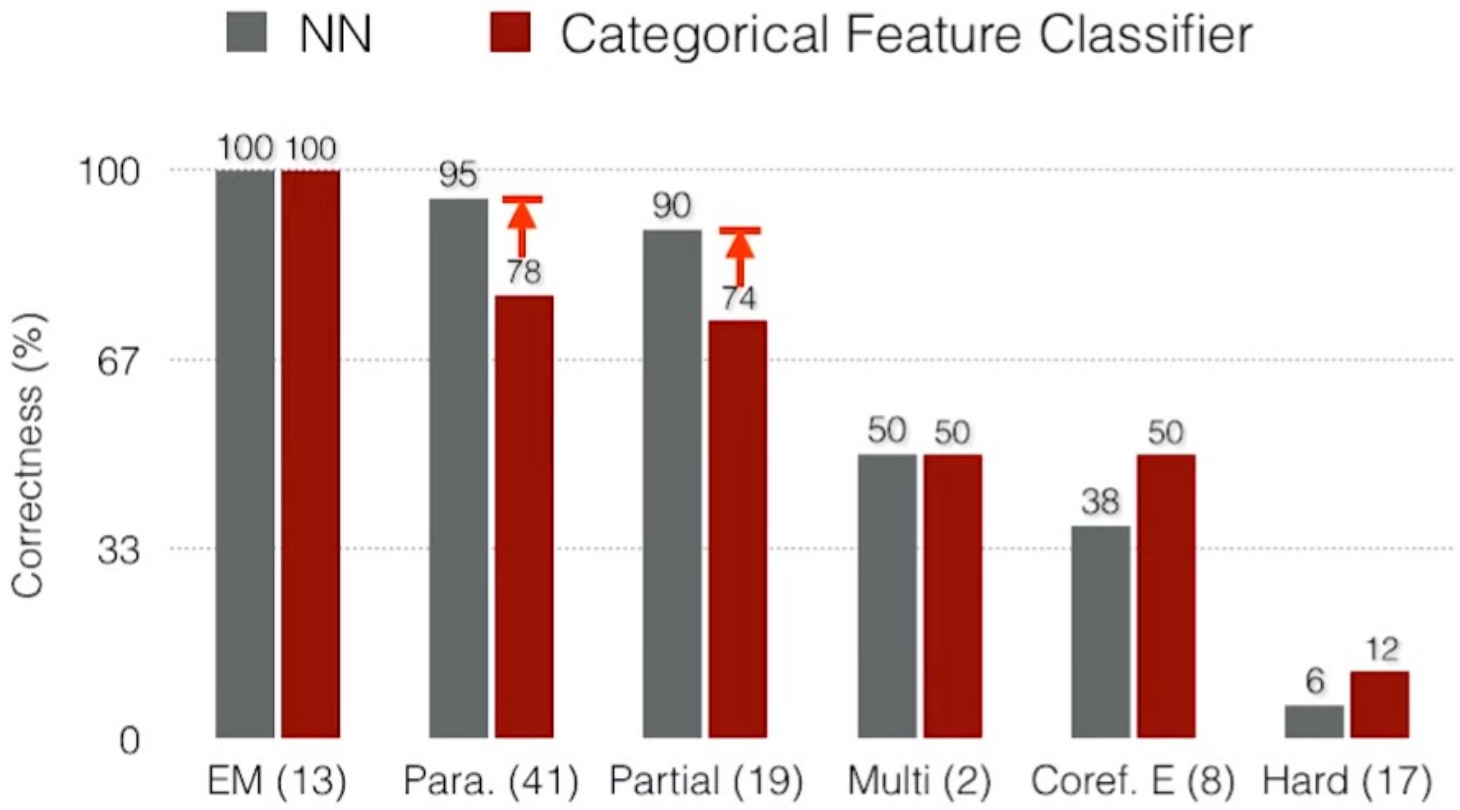
" we got some groundbreaking performances , here too , tonight , " @entity6 said . " we got **@entity17** , who will be doing some musical performances . he 's doing a his - and - her duet all by himself . "

Analysis

What level of language understanding is needed?
What have current models learned?

1. exact match	13%
2. paraphrasing	**41%**
3. partial clue	**19%**
4. multiple sentence	2%
5. coreference errors	8%
6. ambiguous/hard cases	17%

Analysis


NN
Categorical Feature Classifier
Correctness (%)
100
67
33
0
100 100
95
78
90
74
50 50
38
50
6
12
EM (13)
Para. (41)
Partial (19)
Multi (2)
Coref. E (8)
Hard (17)

BiLSTMs + attention models are really good at learning semantic matching.

Does it work in a more real QA setup?

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

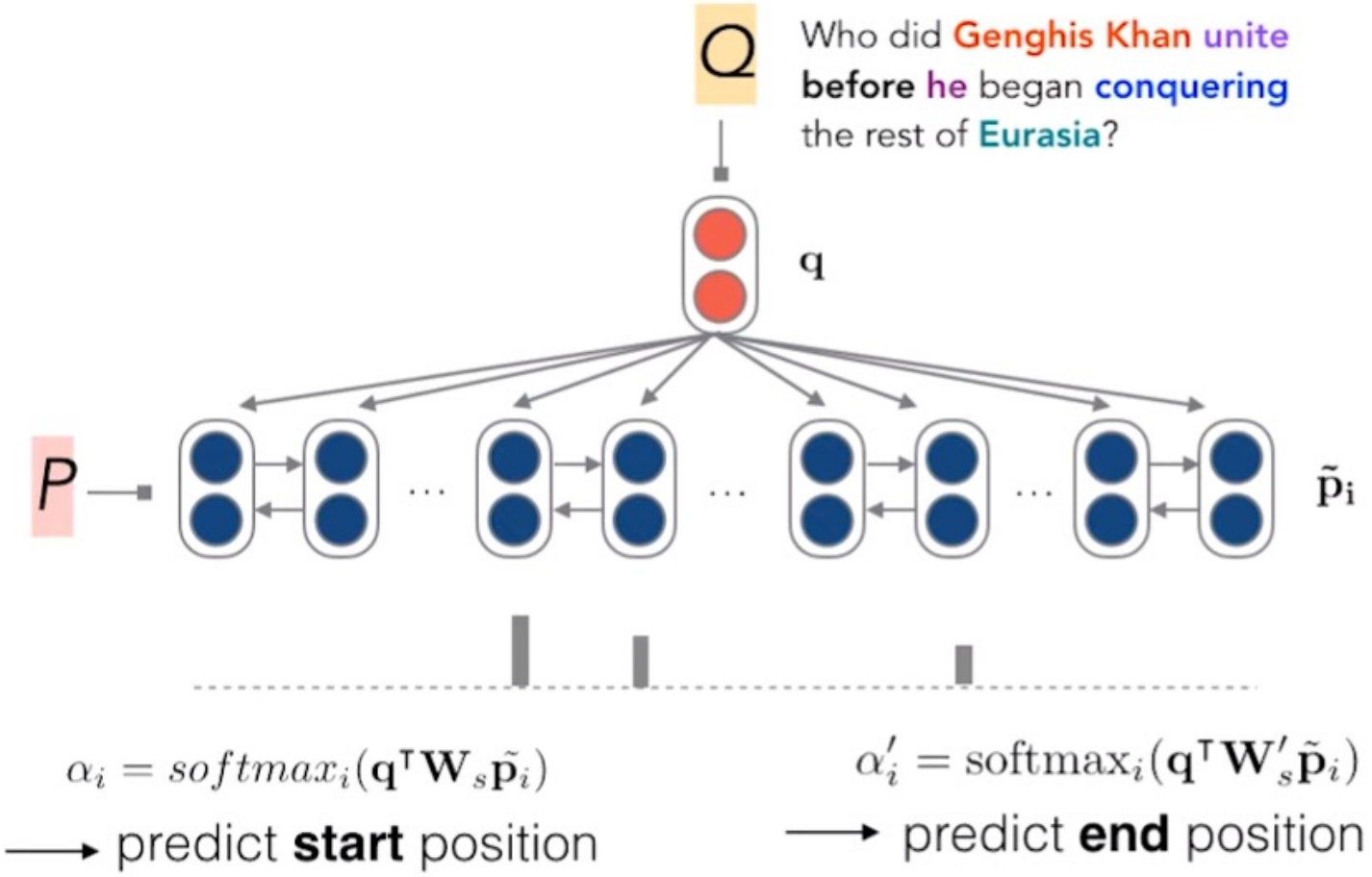
Stanford Question Answering Dataset (SQuAD)

- **Passage**: selected from Wikipedia
- **Question**: crowdsourced
- **Answer**: must be a span in the passage

Who did **Genghis Khan** **unite** **before** **he** began **conquering** the rest of **Eurasia**?

He came to power by **uniting** many of the nomadic tribes of Northeast Asia. **After** founding the Mongol Empire and being proclaimed "**Genghis Khan**", he started the Mongol invasions that resulted in the **conquest** of most of **Eurasia**. These included raids or invasions of the Qara Khitai, Caucasus, Khwarezmid Empire, Western Xia and Jin dynasties. These campaigns were often accompanied by wholesale massacres of the civilian populations – especially in the Khwarezmian and Xia controlled lands. By the end of his life, the Mongol Empire occupied a substantial portion of Central Asia and China.

Attentive Reader for SQuAD



Common Failure Cases

Question: What is the total number of professors, instructors, and lecturers at Harvard?

Harvard's 2,400 professors, lecturers, and instructors instruct 7,200 undergraduates and 14,000 graduate students. The school color is crimson, which is also the name of the Harvard sports teams and the daily newspaper, The Harvard Crimson. The color was unofficially adopted (in preference to magenta) by an 1875 vote of the student body, although the association with some form of red can be traced back to 1858, when Charles William Eliot, a young graduate student who would later become Harvard's 21st and longest-serving president (1869–1909), bought red bandanas for his crew so they could more easily be distinguished by spectators at a regatta.

Correct Predicted

Common Failure Cases

Question: What is the population of the second largest city in California ?

Los Angeles (at 3.7 million people) and San Diego (at 1.3 million people), both in southern California, are the two largest cities in all of California (and two of the eight largest cities in the United States). In southern California there are also twelve cities with more than 200,000 residents and 34 cities over 100,000 in population. Many of southern California's most developed cities lie along or in close proximity to the coast, with the exception of San Bernardino and Riverside.

Correct Predicted

Open-domain QA

- SQuAD is still a restricted QA setup:
  - Questions that can be answered by **span selection**
  - Annotators can see the paragraph when writing questions

$\Longrightarrow$ high **lexical overlap** between question and paragraph

Can we leverage reading comprehension systems for even broader open-domain question answering?

Reading Wikipedia to Answer Open-Domain Question

Reading Wikipedia to Answer Open-Domain Question

DrQA = Information Retrieval + Reading Comprehension

Thank you for listening. Any question?

Reference

- https://www.youtube.com/watch?v=1RN88O9C13U&t=2055s