Neural NLP: Applications & Opportunities

Viktor Schlegel, University of Manchester

Mar 17th 2020

What’s the talk about

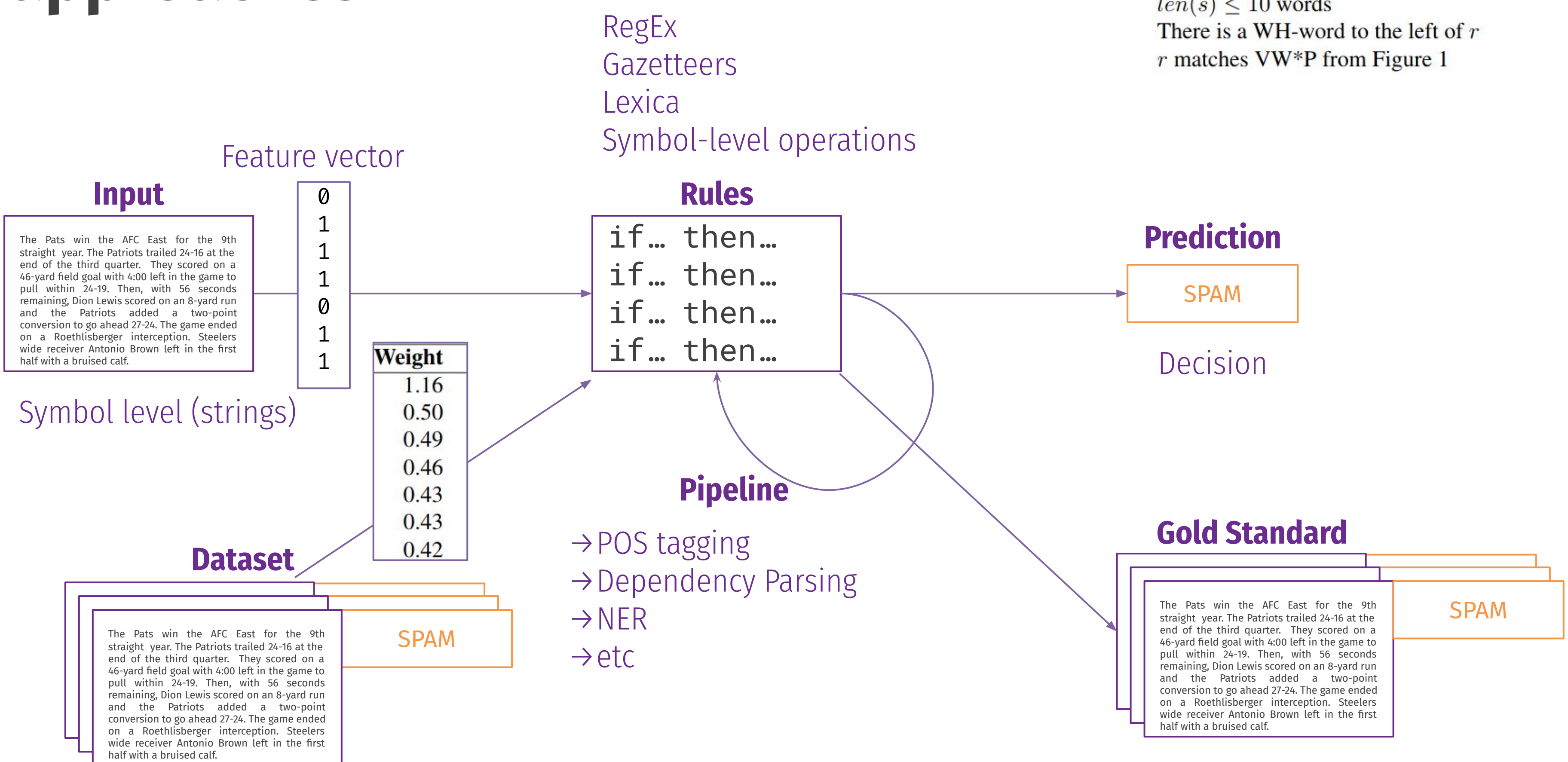
1. The past and the Present of NLP
2. Applications of state of the art NLP
3. (some) opportunities and open research questions

The University of Manchester

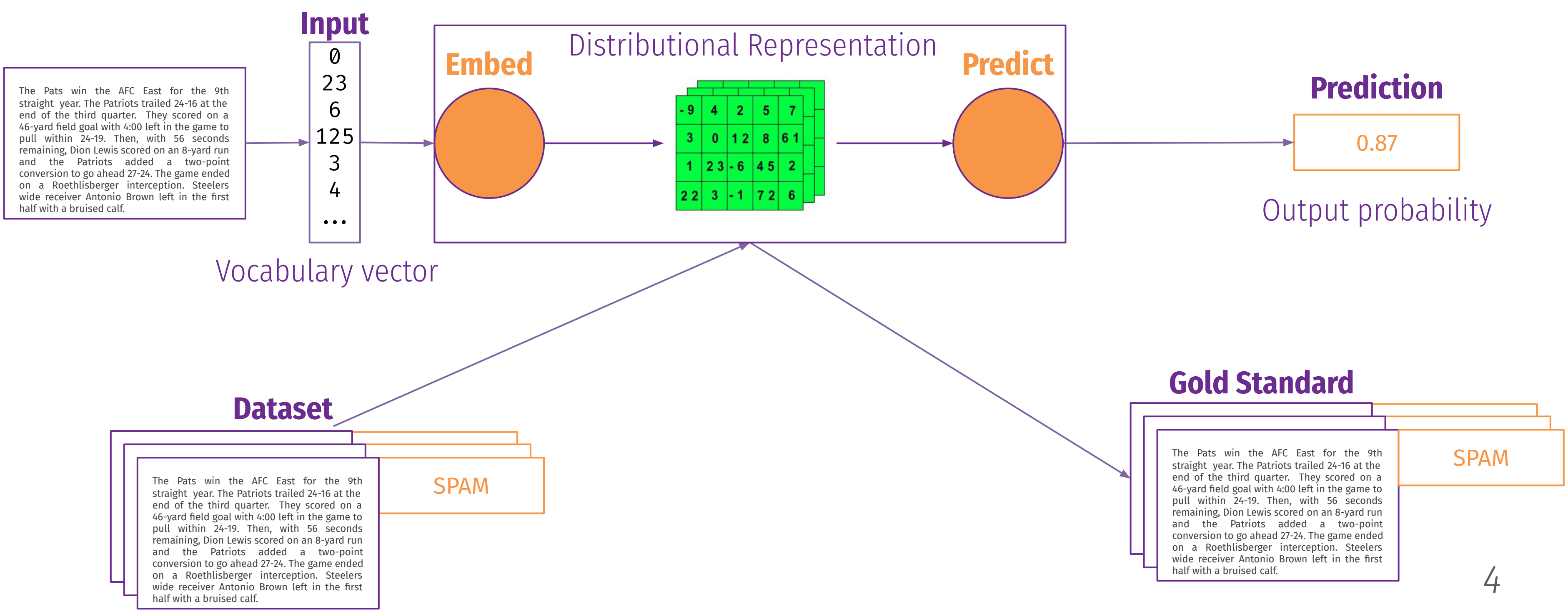
archetypal example:
Spam detection

Rule- and feature based approaches

Feature
(x, r, y) covers all words in s
The last preposition in r is *for*
The last preposition in r is *on*
The last preposition in r is *of*
$len(s) \leq 10$ words
There is a WH-word to the left of r
r matches VW*P from Figure 1



Embedding-based approaches



Distributional Representation

“Tell me with whom thou art found, and I will tell thee who thou art.” - JW Goethe

“Words that occur in similar contexts are similar.”

Place words in a high-dimensional vector space

→ Move words closer that appear in similar contexts

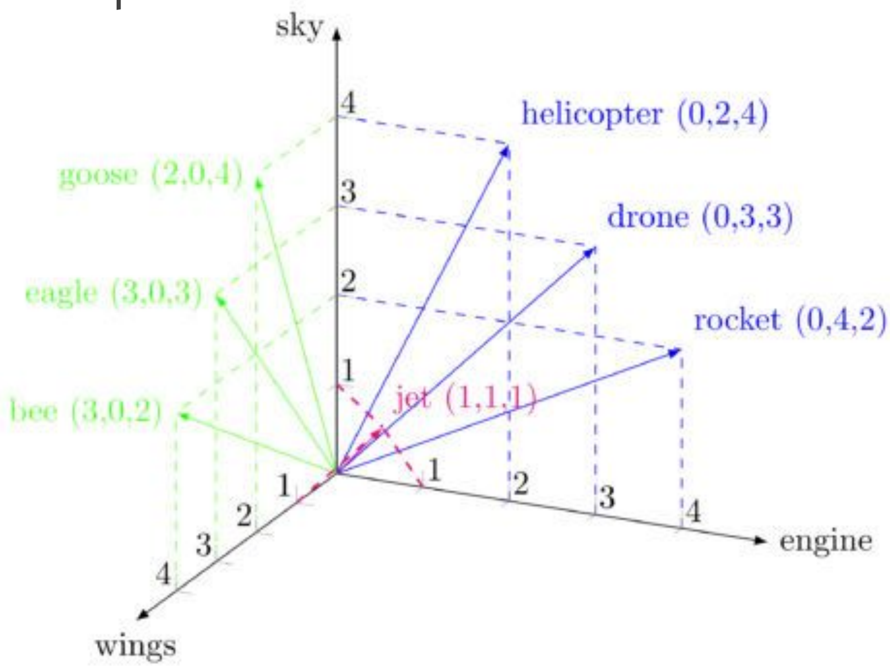
→ Move words apart that do not appear in similar contexts

As observed in some big corpus

The University of Manchester

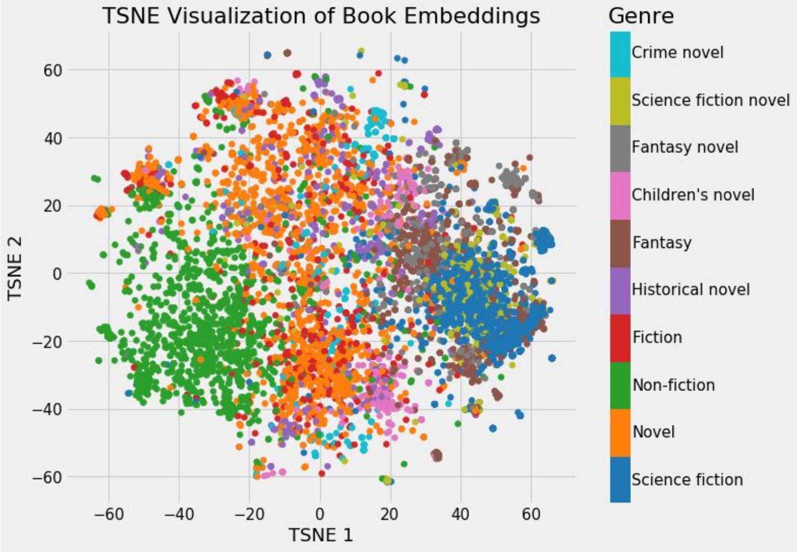
Embeddings

Static map from words to their (high dimensional) distributional representation.



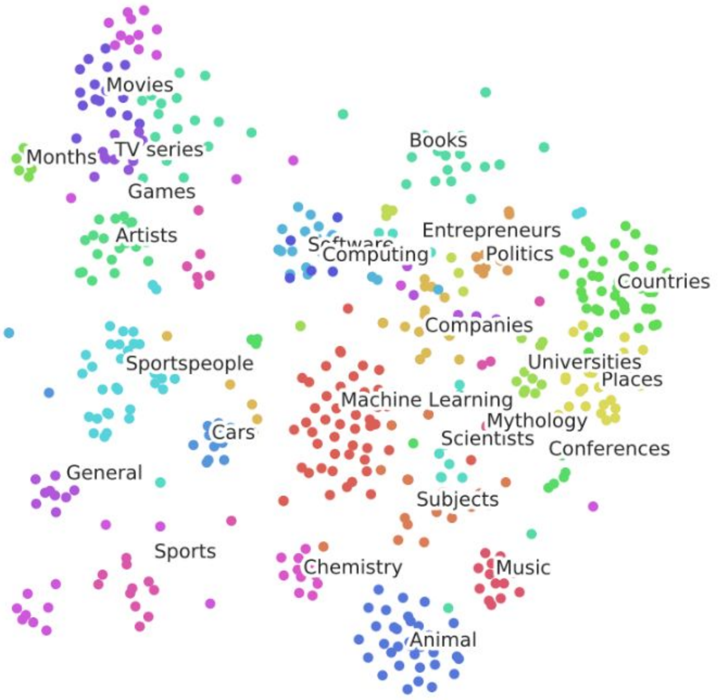
jet ~= ¼ * (bee + rocket)

Tasks



Exploit similarity (low distance) in high-dimensional vector space

- topic modelling
- visualisations
- clustering
- information retrieval (K-nearest neighbours)



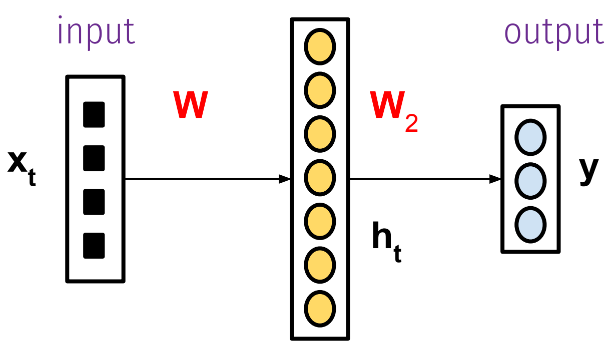
Wikipedia Articles

MANCHESTER 1824

The University of Manchester

Predict

What is a neural network



Matrix multiplication

(Non linear) activation function

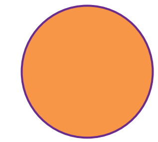
$$h_t = g(Wx_t)$$

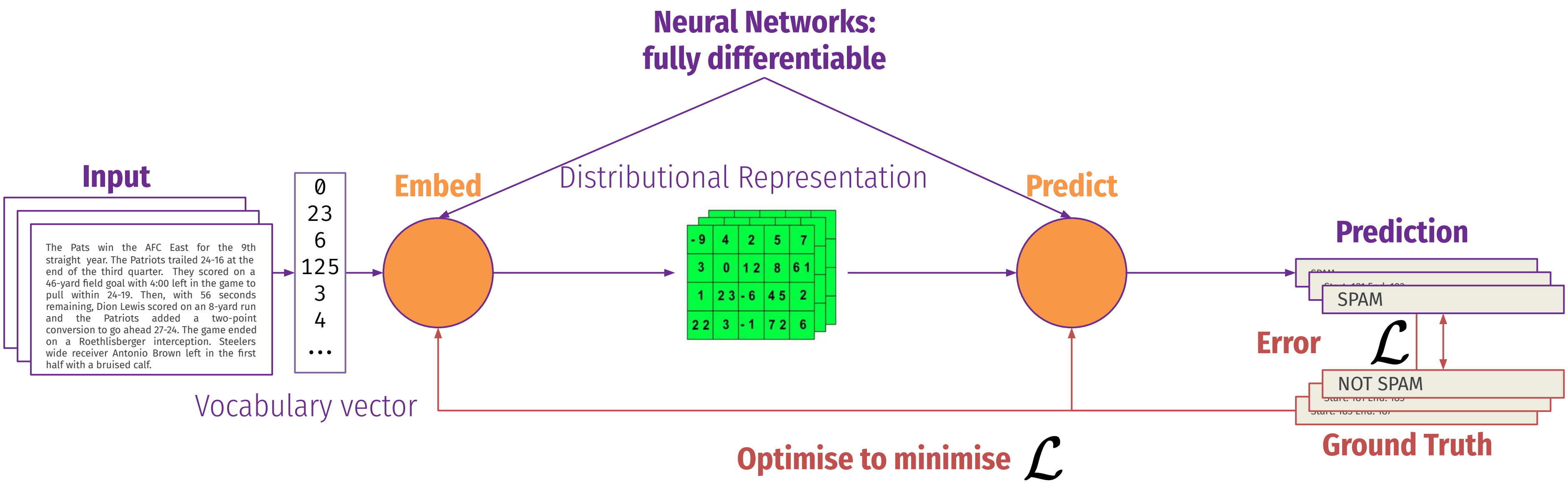
"hidden vector"

The University of Manchester

Training
Embedding-based approaches

1. embed inputs
2. obtain prediction
3. calculate error between prediction & actual value
4. attribute error to parameters in neural networks (**backpropagation**)
5. Change parameters to reduce the error (**gradient descent**)





MANCHESTER
1824
The University of Manchester

What that means

Collect examples as input/output pairs

The Pats win the AFC East for the 9th straight year. The Patriots trailed 24-16 at the end of the third quarter. They scored on a 46-yard field goal with 4:00 left in the game to pull within 24-19. Then, with 56 seconds remaining, Dion Lewis scored on an 8-yard run and the Patriots added a two-point conversion to go ahead 27-24. The game ended on a Roethlisberger interception. Steelers wide receiver Antonio Brown left in the first half with a bruised calf.

SPAM

The Pats win the AFC East for the 9th straight year. The Patriots trailed 24-16 at the end of the third quarter. They scored on a 46-yard field goal with 4:00 left in the game to pull within 24-19. Then, with 56 seconds remaining, Dion Lewis scored on an 8-yard run and the Patriots added a two-point conversion to go ahead 27-24. The game ended on a Roethlisberger interception. Steelers wide receiver Antonio Brown left in the first half with a bruised calf.

NO SPAM

The Pats win the AFC East for the 9th straight year. The Patriots trailed 24-16 at the end of the third quarter. They scored on a 46-yard field goal with 4:00 left in the game to pull within 24-19. Then, with 56 seconds remaining, Dion Lewis scored on an 8-yard run and the Patriots added a two-point conversion to go ahead 27-24. The game ended on a Roethlisberger interception. Steelers wide receiver Antonio Brown left in the first half with a bruised calf.

SPAM

Mature software exists to perform the computations

1. embed inputs
2. obtain prediction
3. calculate error between prediction & actual value
4. attribute error to parameters in neural networks (**backpropagation**)
5. Change parameters to reduce the error (**gradient descent**)

What that means

Minimal supervision

- No task-specific knowledge needs to be encoded in the neural network
- neural networks learns to perform task from input-output examples

... as long as you have the data

Tasks: Document-Level Classification

Assign (one or multiple) fixed category to a piece of text

- Paraphrasing
- Abuse (e.g. offensive language) detection
- Fact Verification
- Stance detection
- Sentiment Analysis
- Emotion Classification

Fact Verification

Claim: The Rodney King riots took place in the most populous county in the USA.

`[wiki/Los_Angeles_Riots]`
The 1992 Los Angeles riots, also known as the Rodney King riots were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.

`[wiki/Los_Angeles_County]`
Los Angeles County, officially the County of Los Angeles, is the most populous county in the USA.

Verdict: Supported

evidence

Thorne et al: "FEVER: a large-scale dataset for Fact Extraction and VERification" NAACL 2018

The University of Manchester

Stance detection

SDQC support classification. Example 1:

u1: We understand there are two gunmen and up to a dozen hostages inside the cafe under siege at Sydney.. ISIS flags remain on display #7News **[support]**

- **u2:** @u1 not ISIS flags **[deny]**
- **u3:** @u1 sorry - how do you know it's an ISIS flag? Can you actually confirm that? **[query]**
  - **u4:** @u3 no she can't cos it's actually not **[deny]**
- **u5:** @u1 More on situation at Martin Place in Sydney, AU –LINK– **[comment]**
- **u6:** @u1 Have you actually confirmed its an ISIS flag or are you talking shit **[query]**

SDQC support classification. Example 2:

u1: These are not timid colours; soldiers back guarding Tomb of Unknown Soldier after today's shooting #StandforCanada –PICTURE– **[support]**

- **u2:** @u1 Apparently a hoax. Best to take Tweet down. **[deny]**
- **u3:** @u1 This photo was taken this morning, before the shooting. **[deny]**
- **u4:** @u1 I don't believe there are soldiers guarding this area right now. **[deny]**
  - **u5:** @u4 wondered as well. I've reached out to someone who would know just to confirm that. Hopefully get response soon. **[comment]**
    - **u4:** @u5 ok, thanks. **[comment]**

Derczynski et al: "RumourEval: Determining rumour veracity and support for rumours" SemEval2017 Task 8

Sentiment Analysis

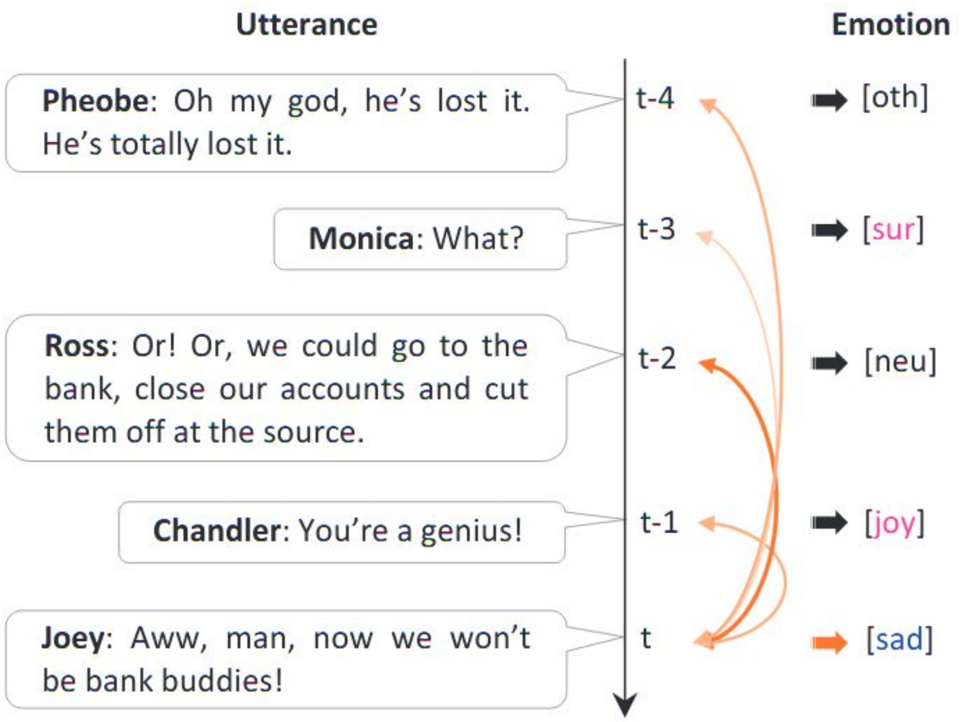
Tweet Content	Score
@ThisIsDeep_ you are about as deep as a turd in a toilet bowl. Internet culture is #garbage and you are bladder cancer.	-4
A paperless office has about as much chance as a paperless bathroom	-3
Today will be about as close as you'll ever get to a "PERFECT 10" in the weather world! Happy Mother's Day! Sunny and pleasant! High 80.	3
I missed voting due to work. But I was behind the Austrian entry all the way, so to speak. I might enter next year. Who knows?	1

negative

positive

Ghosh et al: "Sentiment Analysis of Figurative Language in Twitter" SemEval-2015 Task 11

Emotion Detection



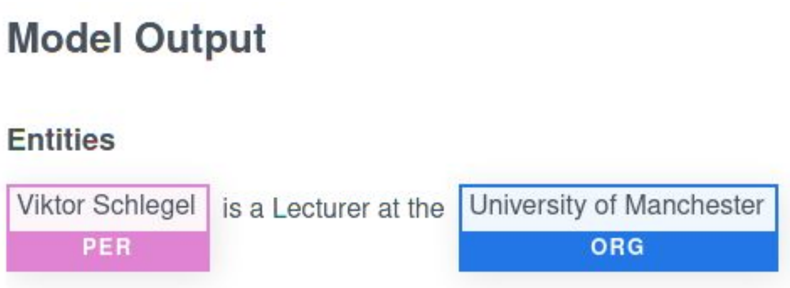
Poria, et al. "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations", ACL 2019

The University of Manchester

Tasks: Token-Level Classification

Assign (one or multiple) fixed category to each word:

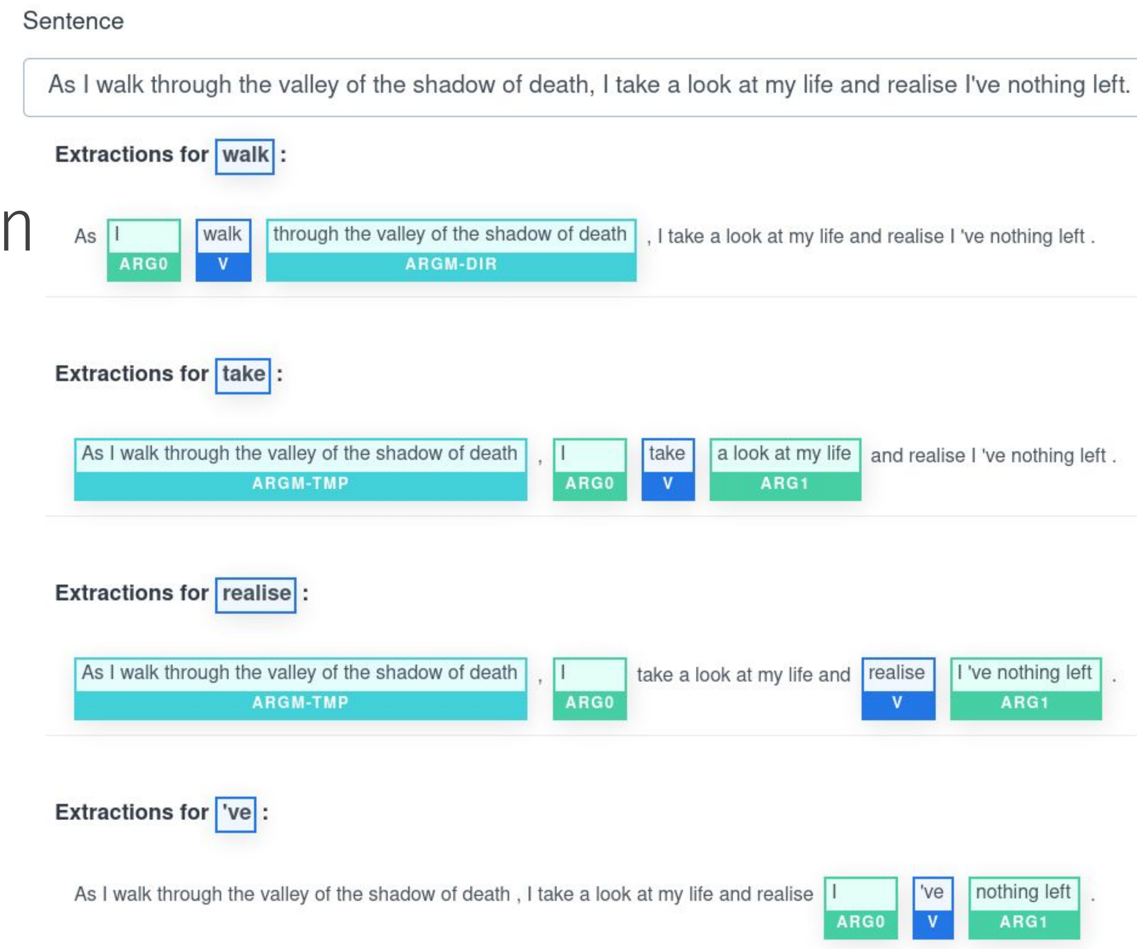
- part of speech tagging
- named entity recognition



The University of Manchester

Tasks: Token-Level Classification

- (open) information extraction



Token-Level Classification: Language modelling

Task: Predict next word given *n* previous words.

$$P(x_{n+1} | x_1, \ldots, x_n)$$

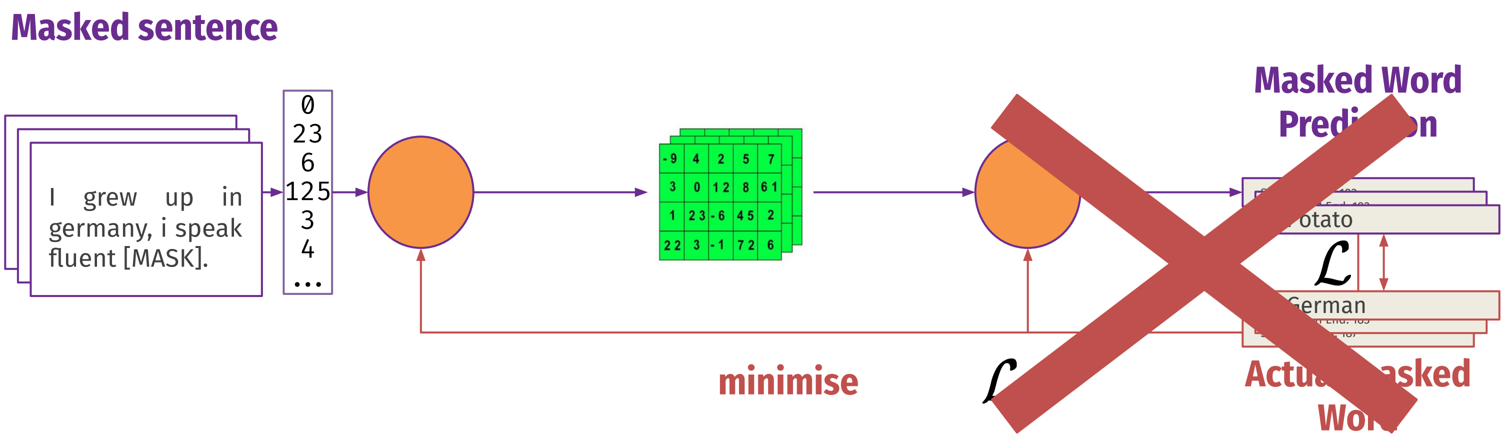
"I grew up in Germany, i speak fluent ____"

syntax information

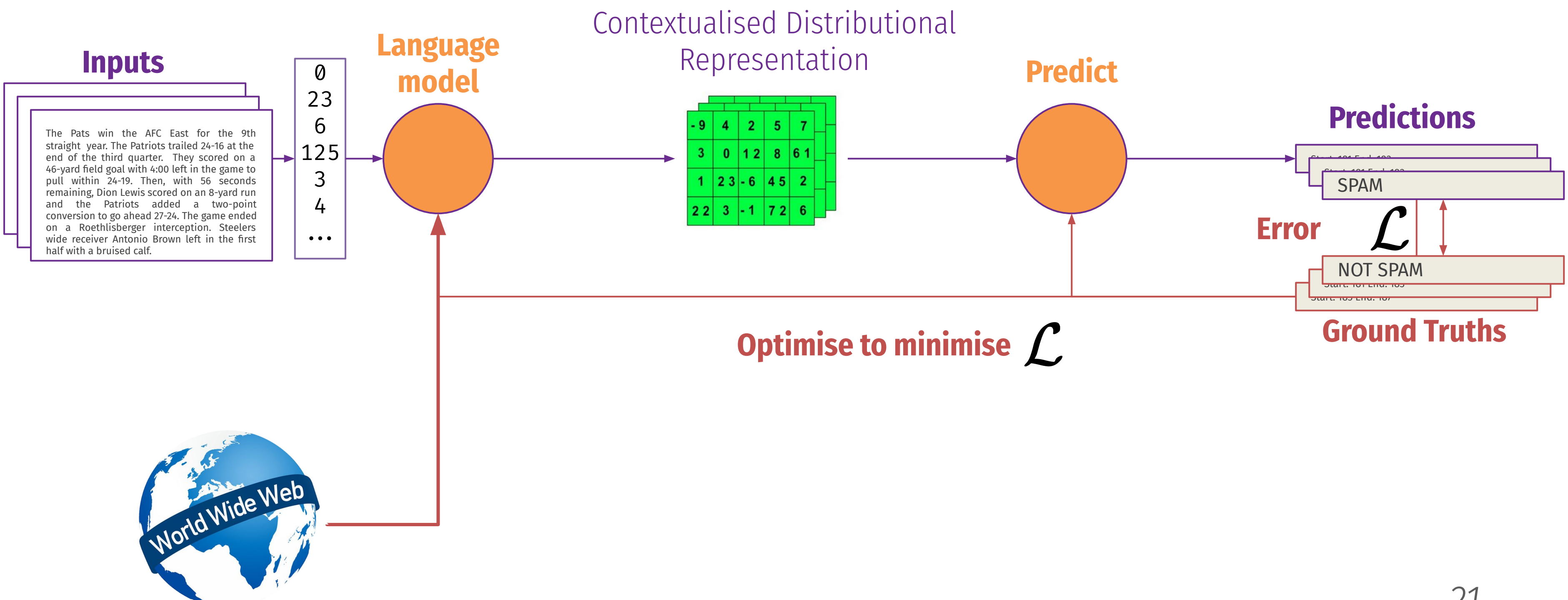
semantic information

Language models for embeddings

self supervision

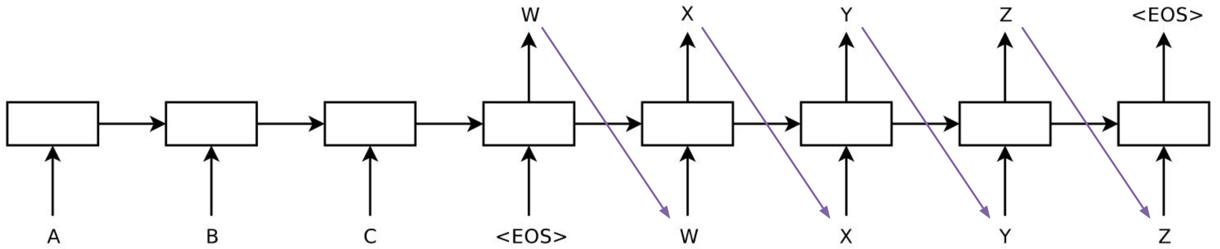


State-of-the-art NLP



0: Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." 2018

The University of Manchester



Tasks: Generation

Given a piece of text, generate the next token

- Translation
- Abstractive summarisation
- Question Answering

$$P(x_{n+1}|x_1, \ldots, x_n)$$

The University of Manchester

Abstractive summarisation

How to Help Save Rivers

Method 1 **Reducing Your Water Usage**

1 **Take quicker showers to conserve water.** One easy way to conserve water is to cut down on your shower time. Practice cutting your showers down to 10 minutes, then 7, then 5. Challenge yourself to take a shorter shower every day.

2 **Wait for a full load of clothing before running a washing machine.** Washing machines take up a lot of water and electricity, so running a cycle for a couple of articles of clothing is inefficient. Hold off on laundry until you can fill the machine.

3 **Turn off the water when you're not using it.** Avoid letting the water run while you're brushing your teeth or shaving. Keep your hoses and faucets turned off as much as possible. When you need them, use them sparingly.

Article 1:

One easy way to conserve water is to cut down on your shower time. Practice cutting your showers down to 10 minutes, then 7, then 5. Challenge yourself to take a shorter shower every day. Washing machines take up a lot of water and electricity, so running a cycle for a couple of articles of clothing is inefficient. Hold off on laundry until you can fill the machine. Avoid letting the water run while you're brushing your teeth or shaving. Keep your hoses and faucets turned off as much as possible. When you need them, use them sparingly.

...

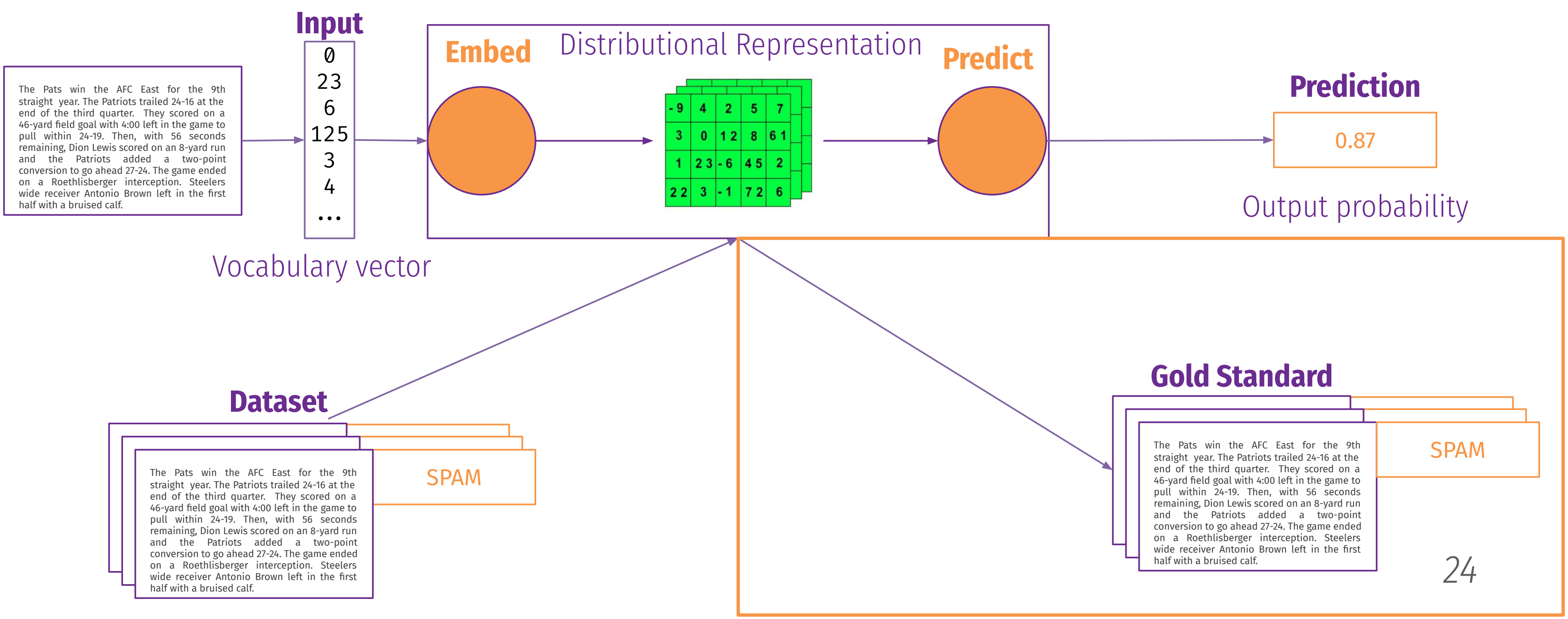
Summary 1:

Take quicker showers to conserve water. Wait for a full load of clothing before running a washing machine. Turn off the water when you're not using it.

Koupaee, Wang: "WikiHow: A Large Scale Text Summarization Dataset" arXiv:1810.09305

The University of Manchester

Open research questions



MANCHESTER 1824

The University of Manchester

Machine Reading Comprehension: “ answers for questions over passages of text”

The Pats win the AFC East for the 9th straight year. The Patriots trailed 24-16 at the end of the third quarter. They scored on a 46-yard field goal with 4:00 left in the game to pull within 24-19. Then, with 56 seconds remaining, Dion Lewis scored on an 8-yard run and the Patriots added a two-point conversion to go ahead 27-24. The game ended on a Roethlisberger interception. Steelers wide receiver Antonio Brown left in the first half with a bruised calf.

Who was injured during the match?

(a) Rob Gronkowski (b) Ben Roethlisberger (c) Dion Lewis (d) Antonio Brown

The Patriots champion the cup for _9__ consecutive seasons.

What was the final score of the game?

How many points ahead were the Patriots by the end of the game? 3

Why Question Answering/MRC

MRC/QA as interface

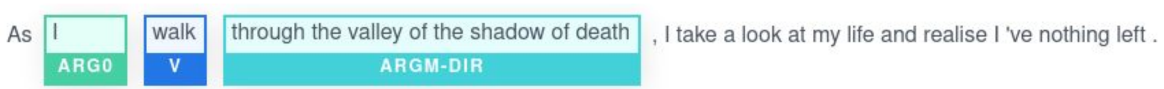
- HCI interface to exploration

```
SELECT name FROM mountains WHERE
    country = UK
ORDERBY height
LIMIT 1
```

vs

"What is the highest mountain in the UK?"

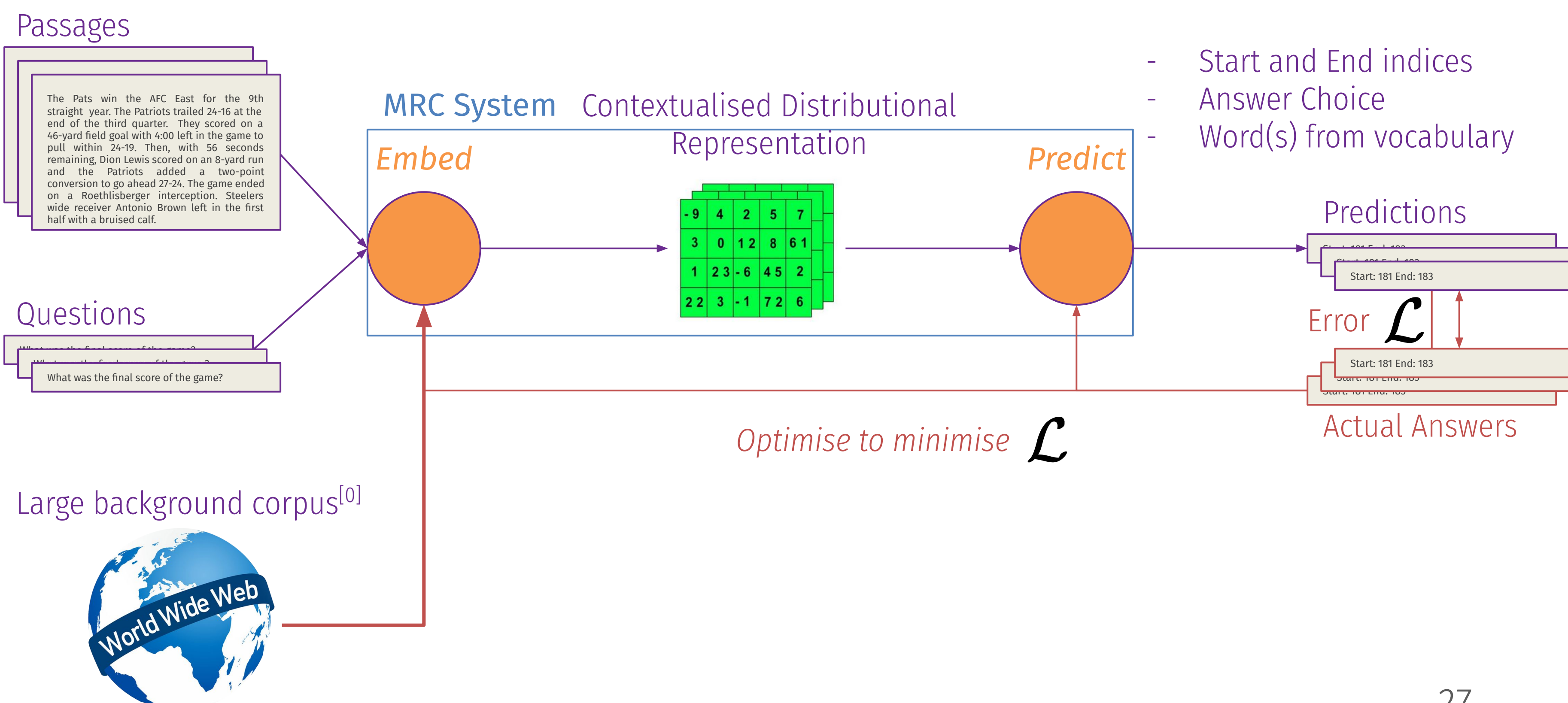
- Interface to other tasks



What is being done? → walk
Who walks through the valley? → I
Where do I walk? → through the valley of the shadow of death

The University of Manchester

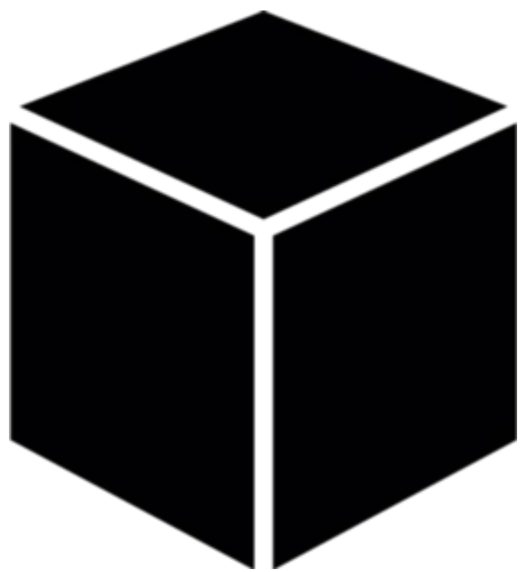
Language models for MRC



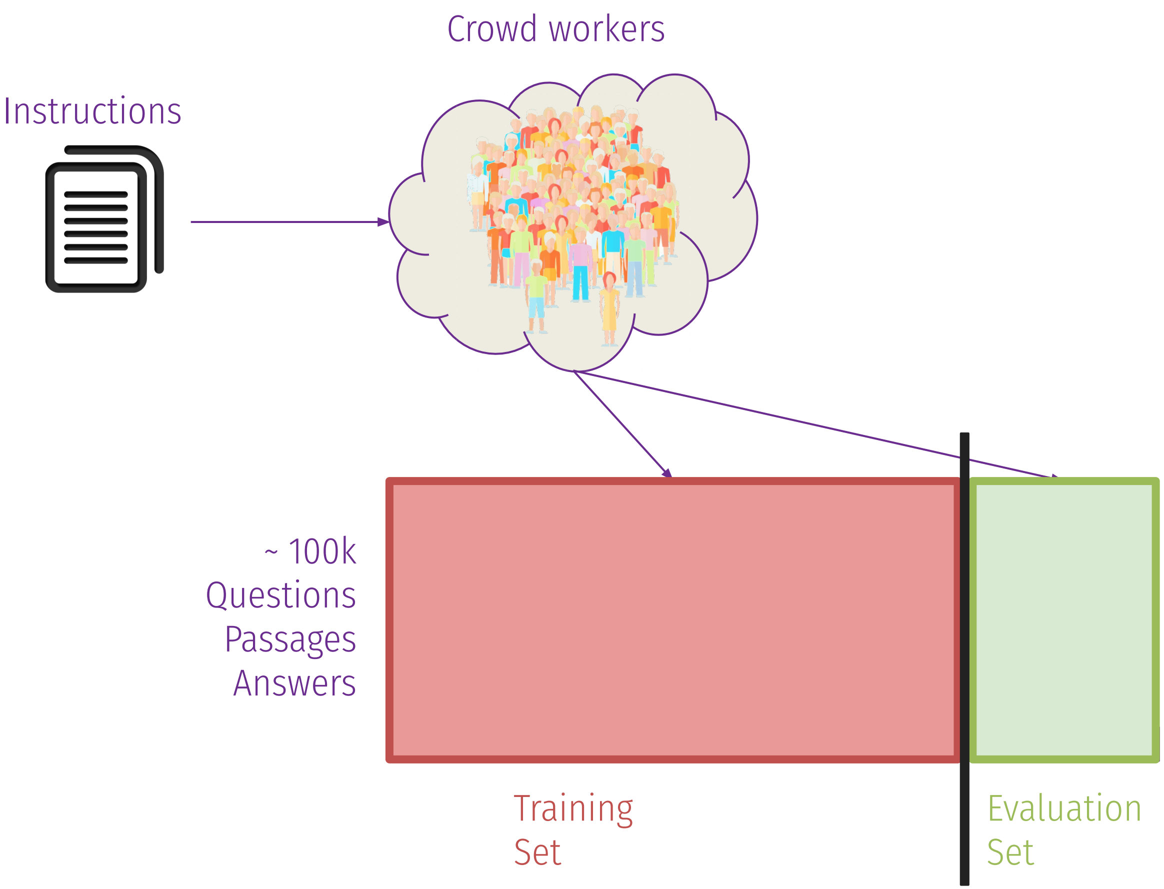
0: Devlin, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." 2018

The University of Manchester

Neural networks are black boxes



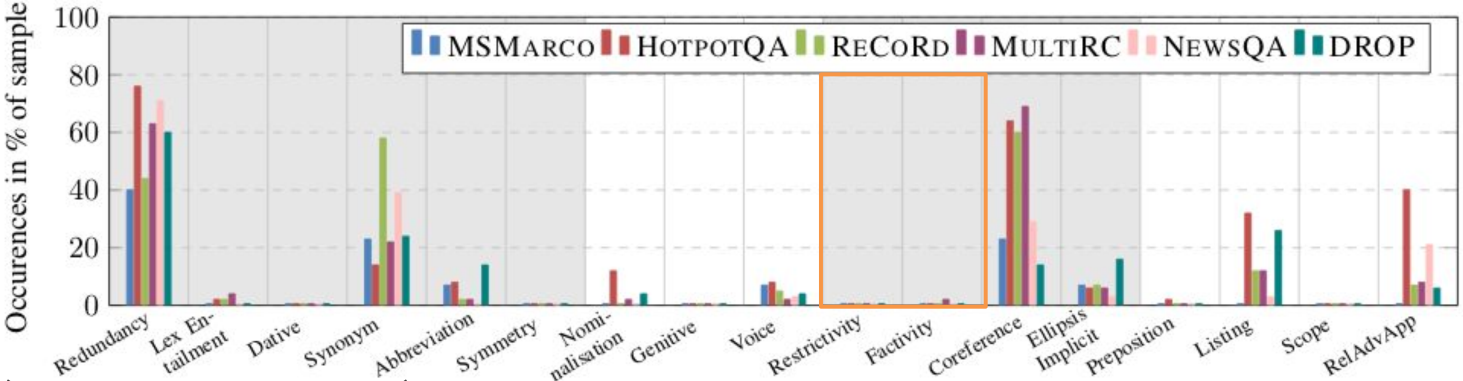
Black-box testing



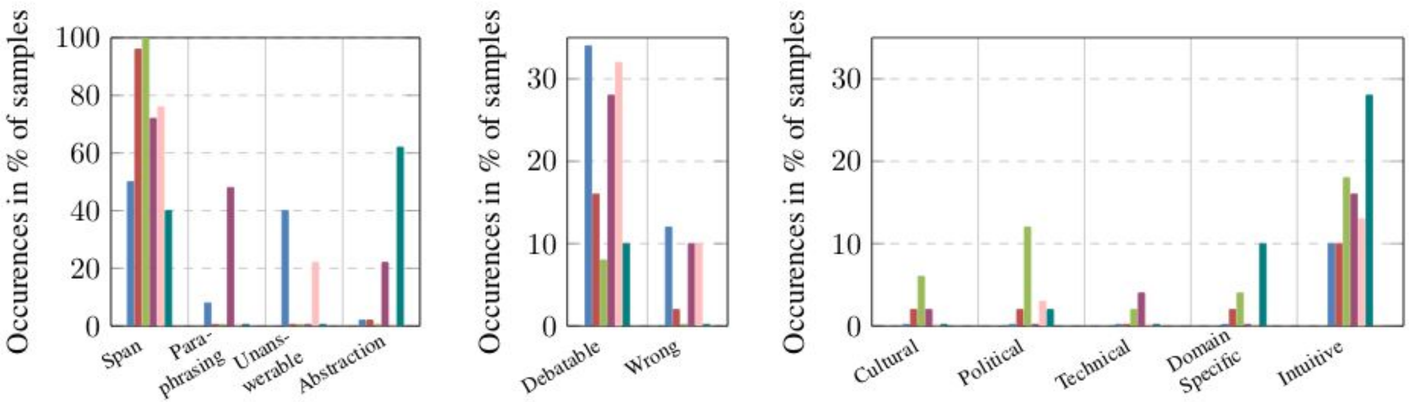
1. What do gold standards (not) evaluate?
2. Which capabilities do the models (not) acquire?

We don’t really know what these models learn.

A (systematic) look into the data...



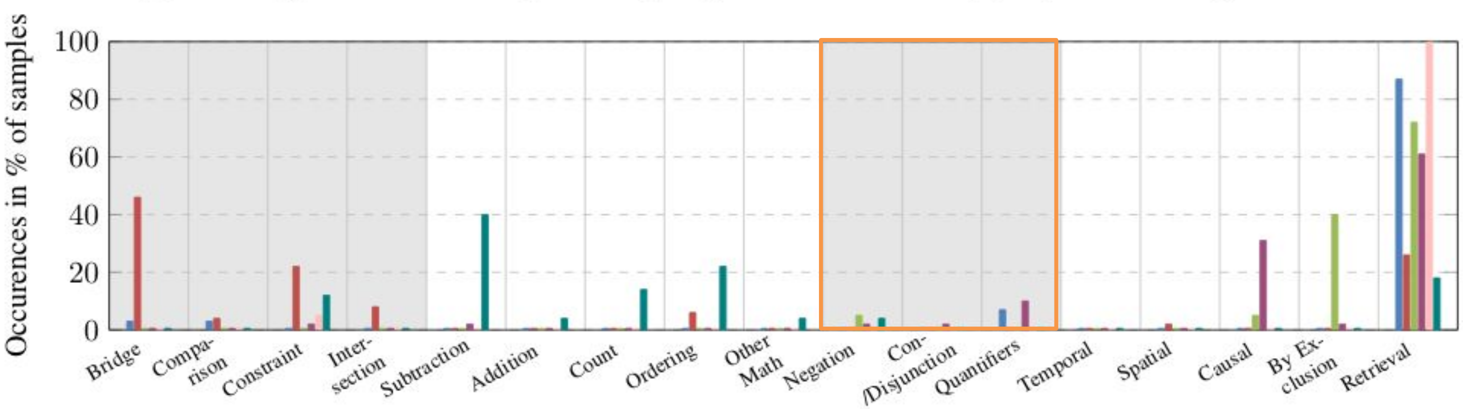
(a) Lexical (grey background) and syntactic (white background) linguistic features



(b) Answer Types

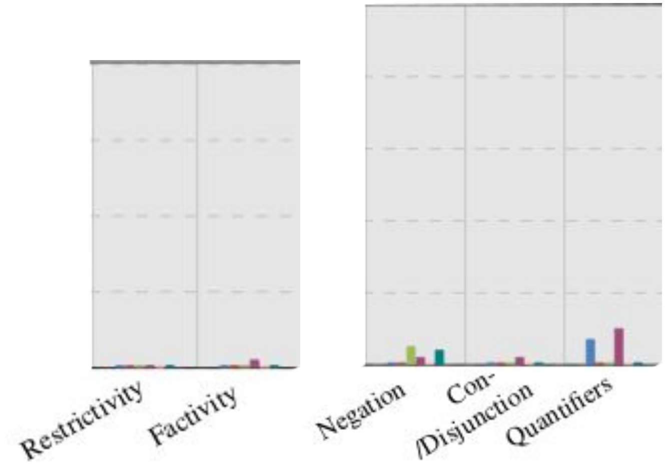
(c) Answer Quality

(d) Required Knowledge



1. What do benchmarks (not) evaluate?

The University of Manchester



What's missing?

E.g.:
Restrictivity:
"Brady almost scored a TD" vs "Brady scored a TD"

Factivity:
"He was probably involved in the plot" vs "He was definitely involved in the plot"

Discourse relations:
Conditionals:
"If they have milk, I will buy 3 bottles of milk" vs "I will buy 3 bottles of milk."

Conjunctions:
"Mary ate an apple. John ate an apple. John ate a pear."
"Who ate an apple?" vs "Who ate an apple and a pear?"

"...look similar but mean different things."

Common Theme:

Phenomena that preserve the lexical surface form while altering the meaning are missing!

⇒ cannot say, whether systems learn to process them

Schlegel, et al. A Framework for Evaluation of Machine Reading Comprehension Gold Standards. LREC 2020
Wu, Schlegel, Batista-Navarro: Is the Understanding of Explicit Discourse Relations Required in Machine Reading Comprehension?. EACL 2021

The University of Manchester

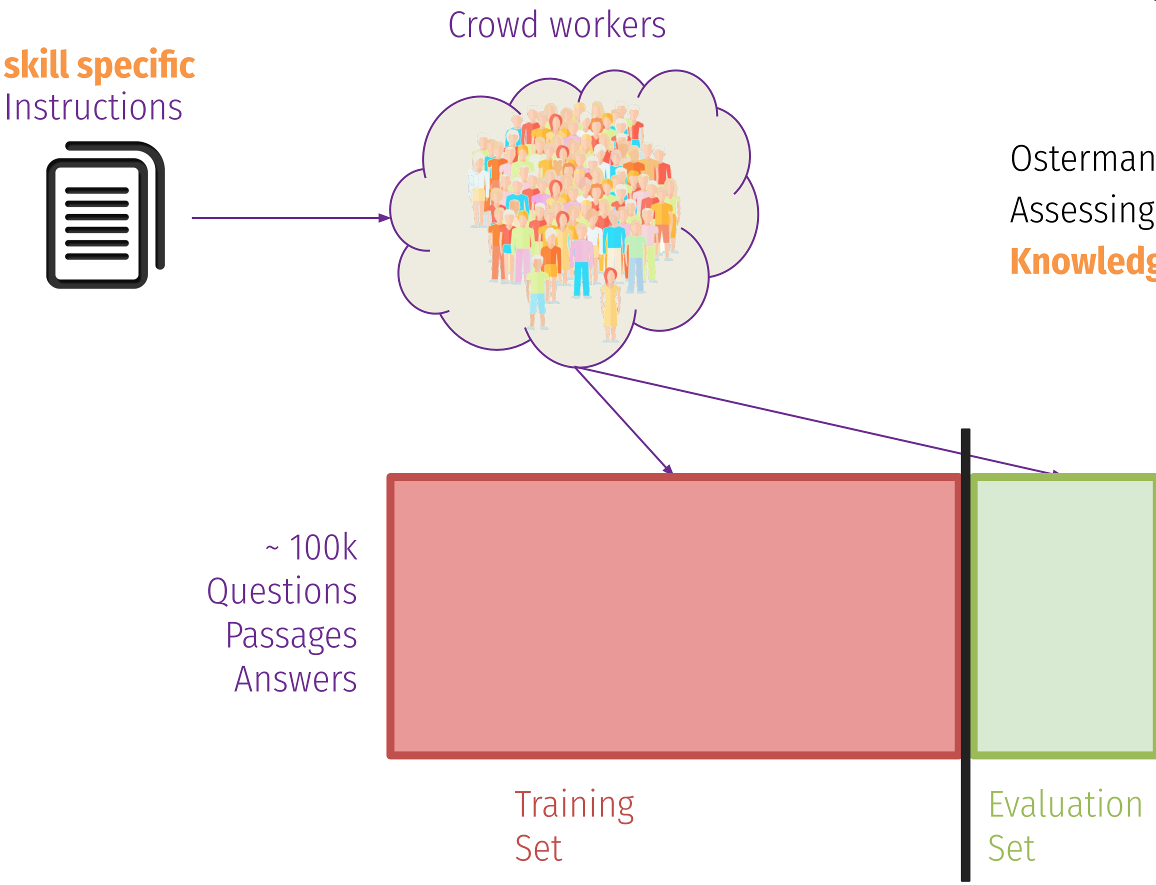
Dua et al: DROP: A Reading Comprehension Benchmark **Requiring Discrete Reasoning** Over Paragraphs. NAACL 2019

"Skill-based" Evaluation

Yang et al: HotpotQA: A Dataset for **Diverse, Explainable Multi-hop** Question Answering. EMNLP 2018

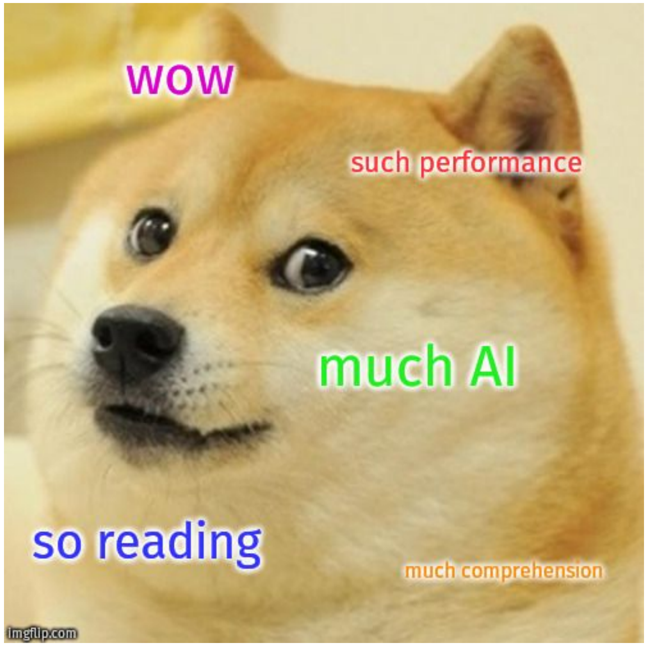
Ostermann et al: MCScript: A **Novel** Dataset for Assessing Machine Comprehension Using **Script Knowledge**. LREC 2018

Liu et al: LogiQA: A **Challenge** Dataset for Machine Reading Comprehension with **Logical Reasoning**. IJCAI 2020



MANCHESTER
1824
The University of Manchester

MRC: Expectation...



Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	**90.002**	**92.425**

1: Rajpurkar, Jia, Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD

The University of Manchester

MRC: vs reality

Answer

2

Explanation

The model decided this was a counting problem.

Passage

I have an apple.

Question

How many apples do I have?

Gardner, et al. "AllenNLP: A Deep Semantic Natural Language Processing Platform." Proceedings of Workshop for NLP Open Source Software (NLP-OSS). 2018.

- Survey 121 papers for
  - Reported Data and model "weaknesses"
  - Reported Methods to reveal them
  - Reported Methods to overcome them
- Classify, categorise, detect common themes

MANCHESTER
1824
The University of Manchester

Weaknesses...

Passage 1: Marietta Air Force Station
Marietta Air Force Station (ADC ID: M-111, NORAD ID: Z-111) is a closed United States Air Force General Surveillance Radar station. It is located 2.1 mi northeast of Smyrna, Georgia. It was closed in 1968.

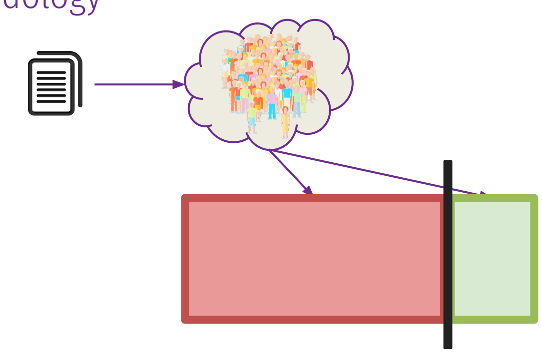
Passage 2: Smyrna, Georgia
Smyrna is a city northwest of the neighborhoods of Atlanta. [...] As of the 2010 census, the city had a population of ***51,271****. The U.S. Census Bureau estimated the population in 2013 to be 53,438. [...]*

Question: *What is the 2010 population of the city 2.1 miles southwest of Marietta Air Force Station?*

words in question colocated with answer
=> unwanted cues

low-bias models rely on strongest signal in data: learn to "pay attention" to the surface form
=> sophisticated word matching

Undetected with usual evaluation methodology



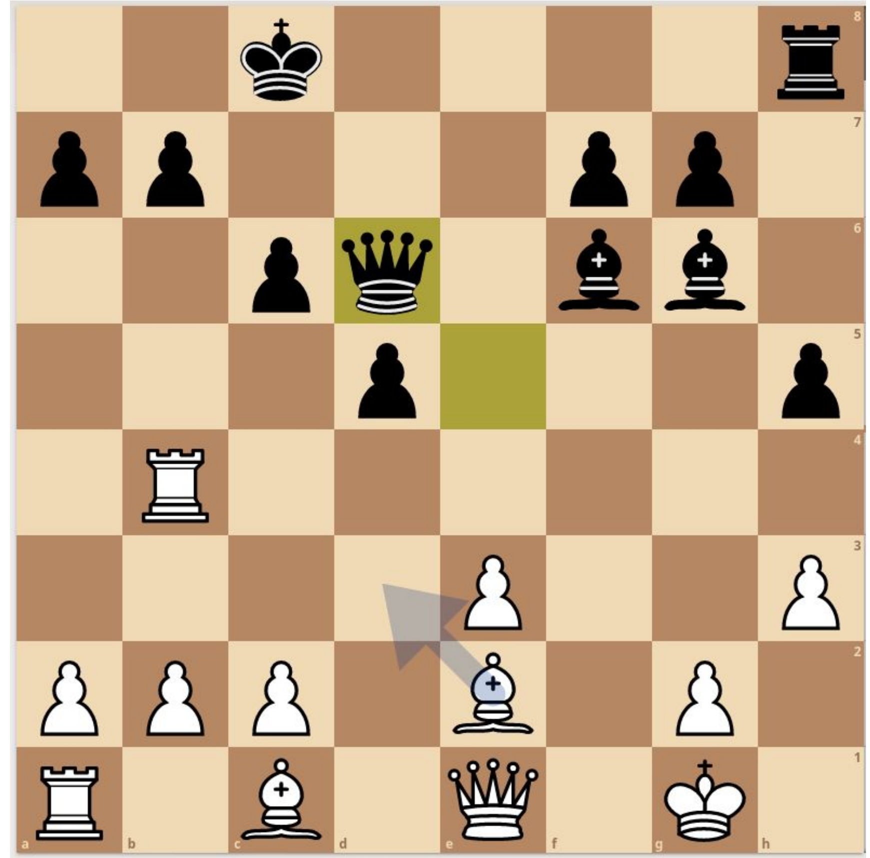
Schlegel, Nenadic, Batista-Navarro: Beyond Leaderboards: A survey of methods for revealing weaknesses in Natural Language Inference data and models arXiv preprint. 2020

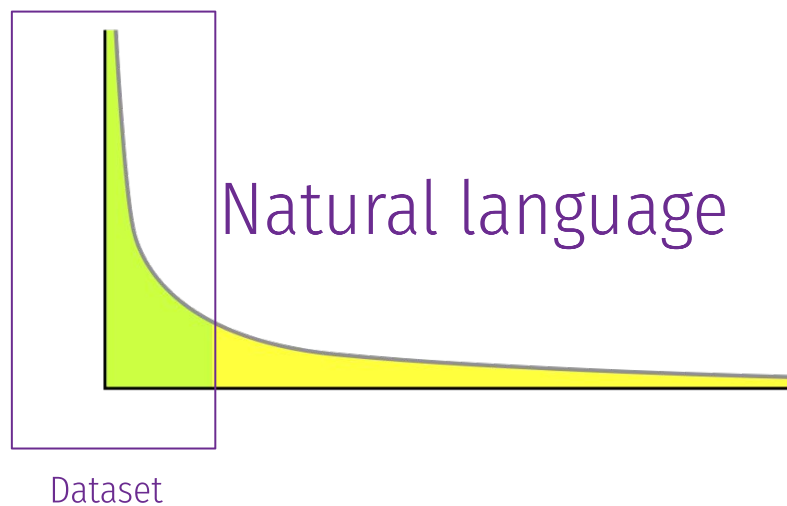
We still don’t really know what these systems learn.

All we can say “It learned to succeed at this dataset”.

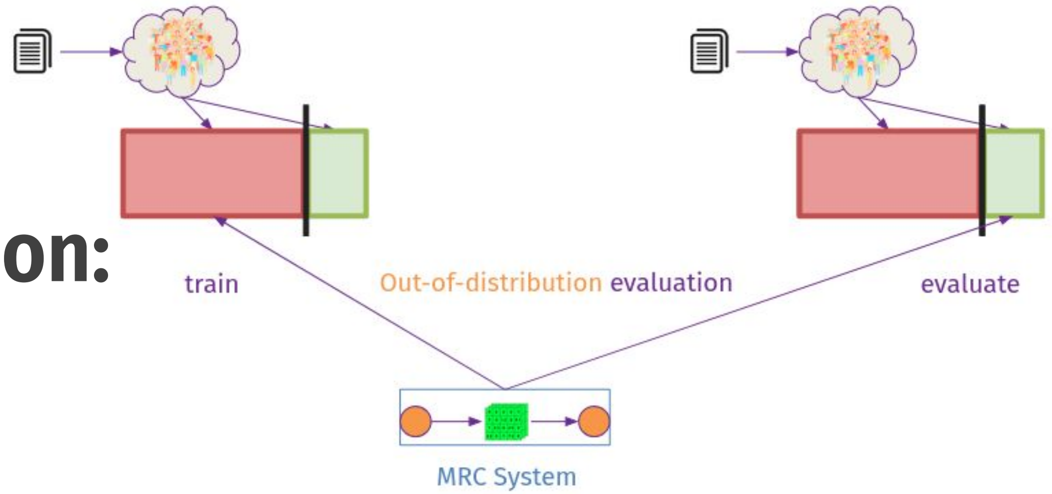
Dataset represents task?

E.g. in Chess, task: outperform humans consistently
dataset: millions of games





“Skill based” evaluation: Challenge sets



Construct challenge set:

- Hand-craft/generate examples that require skill of interest to solve
- Evaluate (optimised) MRC system on those examples
  - “Good” performance: system “learned” it?
  - “Bad” performance: need to discount

- Unknown words?
- Different topic?
- Different passage lengths?
- ???
- Doesn’t learn the skill?

“domain shift”

(B) **Original**: *curled in*

(I1) **Modal negation**: couldn't curl *in*
(I2) **Adverbial Modification**: almost *curled in*
(I3) **Implicit Negation**: was prevented from *curling in*
(I4) **Explicit Negation**: didn't succeed in *curling in*
(I5) **Polarity Reversing**: lacked the nerve to *curl in*
(I6) **Negated Polarity Preserving**: wouldn't find the opportunity to *curl in*

For example: Semantic altering Modifications (SAM)

Understanding MRC models can process SAM is interesting, because those phenomena require some sort of deeper comprehension beyond the exploitation of lexical cues.

Brady scored a 45-yard TD. Brady scored another 25-yard touchdown after 5 minutes. Brady scored 2 20-yard touchdowns in the second half.

What is the longest TD run?

Compare numbers next to 'TD/touchdown', pick highest!

Brady *almost* scored a 45-yard TD. Brady scored another 25-yard TD touchdown after 5 minutes. Brady scored 2 20-yard touchdowns in the second half.

What is the longest TD run?

Doesn't work anymore!

(B) **Original**: *curled in*
(I1) **Modal negation**: couldn't curl *in*
(I2) **Adverbial Modification**: almost *curled in*
(I3) **Implicit Negation**: was prevented from *curling in*
(I4) **Explicit Negation**: didn't succeed in *curling in*
(I5) **Polarity Reversing**: lacked the nerve to *curl in*
(I6) **Negated Polarity Preserving**: wouldn't find the opportunity to *curl in*

Baseline Passage: *Bob baked some cookies for Alice. [...]*	
Intervention Passage: *Bob baked Alice some cookies. [...]*	
Question: *What did Bob bake?*	***Answer:*** *some cookies (not Alice)*
Baseline Passage: *Bob drew a picture of mom for Alice. [...]*	
Intervention Passage: *Bob drew Alice a picture of mom. [...]*	
Question: *Who did Bob draw?*	***Answer:*** *mom (not Alice)*

Also useful for other phenomena!
(e.g. dative alteration)

Can MRC systems process SAM?

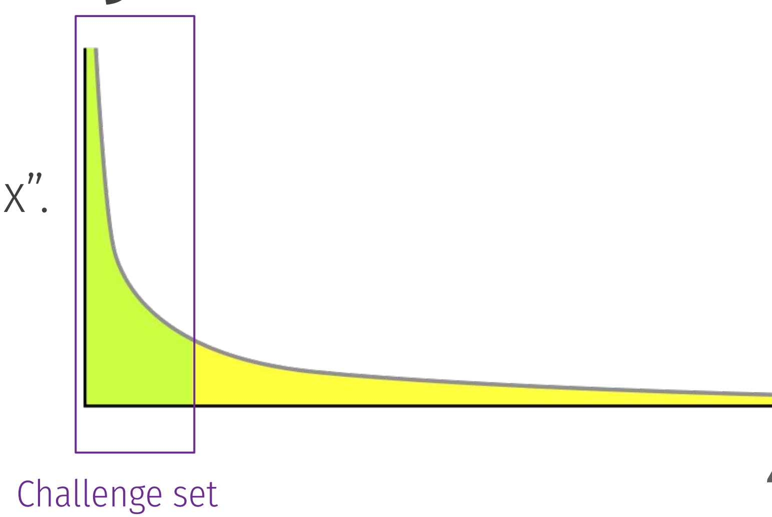
Architecture	Average *DICE*	SQuAD EM/F1	SQuAD *DICE*	HotpotQA EM/F1	HotpotQA *DICE*	NewsQA' EM/F1	NewsQA' *DICE*	DROP' EM/F1	DROP' *DICE*
bidaf	11 ± 3	67.2/76.9	12 ± 4	44.6/57.9	4 ± 3	40.0/54.3	13 ± 5	50.8/56.8	18 ± 12
bert-base	13 ± 2	76.3/84.9	13 ± 3	50.7/64.9	17 ± 4	46.6/62.5	13 ± 3	50.5/58.2	10 ± 3
bert-large	15 ± 2	81.9/89.4	15 ± 3	54.4/68.7	14 ± 3	49.1/65.7	14 ± 4	62.2/68.7	16 ± 3
roberta-base	15 ± 2	82.4/89.9	8 ± 3	51.9/66.4	17 ± 4	50.8/66.9	14 ± 3	63.5/69.3	20 ± 3
roberta-large	18 ± 1	86.4/93.3	16 ± 3	58.6/72.9	21 ± 3	54.4/71.1	15 ± 3	77.3/82.8	20 ± 2
albert-base	14 ± 2	82.8/90.3	10 ± 3	55.4/69.7	17 ± 3	49.7/65.7	11 ± 3	60.7/67.0	18 ± 4
albert-large	16 ± 1	85.4/92.1	18 ± 3	59.4/73.7	12 ± 2	52.5/68.9	17 ± 3	69.3/75.1	18 ± 3
albert-xl	27 ± 2	87.1/93.5	19 ± 2	62.4/76.2	21 ± 3	54.2/70.4	29 ± 3	76.4/81.8	40 ± 3
albert-xxl	27 ± 1	88.2/94.4	29 ± 2	65.9/79.5	29 ± 3	54.3/71.0	25 ± 3	78.4/84.5	23 ± 2
t5-small	10 ± 1	76.8/85.8	13 ± 3	51.8/65.6	10 ± 3	47.3/63.3	8 ± 2	60.4/66.1	10 ± 3
t5-base	16 ± 1	82.4/90.6	16 ± 3	61.0/74.4	20 ± 3	52.4/68.8	14 ± 3	69.0/74.9	15 ± 2
t5-large	20 ± 1	86.3/93.1	21 ± 2	65.0/78.5	29 ± 3	53.4/70.0	16 ± 3	70.1/75.3	8 ± 2
average	19 ± 0	76.4/83.2	18 ± 1	53.1/65.9	20 ± 1	47.1/62.1	17 ± 1	61.5/67.0	20 ± 1
albert-xl-comb	20 ± 2	85.3/92.2		60.6/74.3		53.6/70.4		76.9/82.4	
random	5 ± 0								
learned	98 ± 0								

11.5 GB

Schlegel, Nenadic, Batista-Navarro: "Semantics Altering Modifications for Evaluating Comprehension in Machine Reading, AAAI 2021

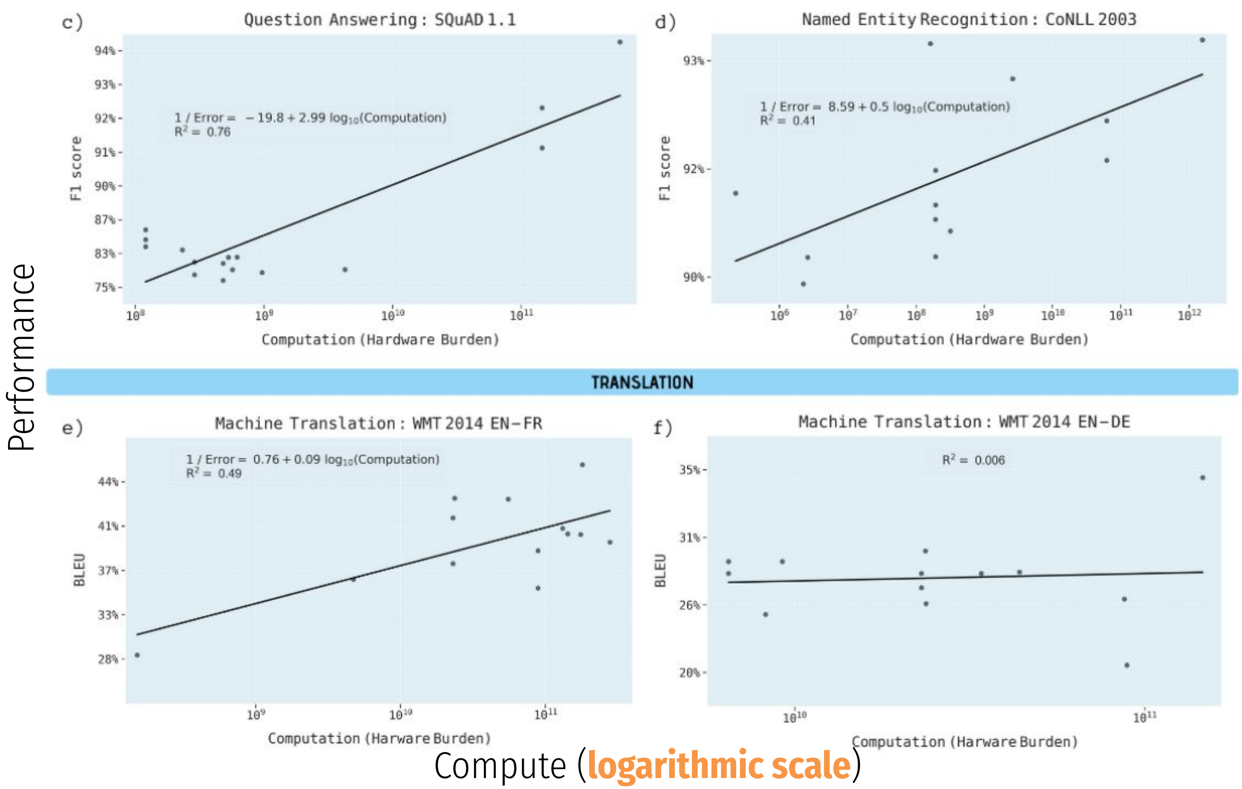
We still still don’t really know what these models learn.

All we can say “models didn’t learn skill x”.



Diminishing returns

- NLP progress is driven by size (model/data)
- bigger models require more compute
- Participation barrier



Thompson et al. The Computational Limits of Deep Learning. http://arxiv.org/abs/2007.05558

Raffel et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. JMLR 2020

The University of Manchester

Being represented

- if a group is not represented in training data, data driven solution “won’t work” for them
  - e.g. face recognition system trained on a non-diverse dataset
  - or NLP system trained only on standard english

“HCI interface”

Original	When is the suspended team scheduled to return?
Adversary	When are the suspended team schedule to returned?
Prediction	**Before:** 2018 **After:** No answer

Tan, S et al: It’s Morphin’ Time! Combating Linguistic Discrimination with Inflectional Perturbations. ACL 2020

MANCHESTER 1824
The University of Manchester

Majority isn't always right!

Neural networks excel at exploiting **statistical patterns** in data

- No device to distinguish between spurious and robust correlations
- With more training data, so the hope, robust correlations will dominate over spurious
- However: Majority is not always right!

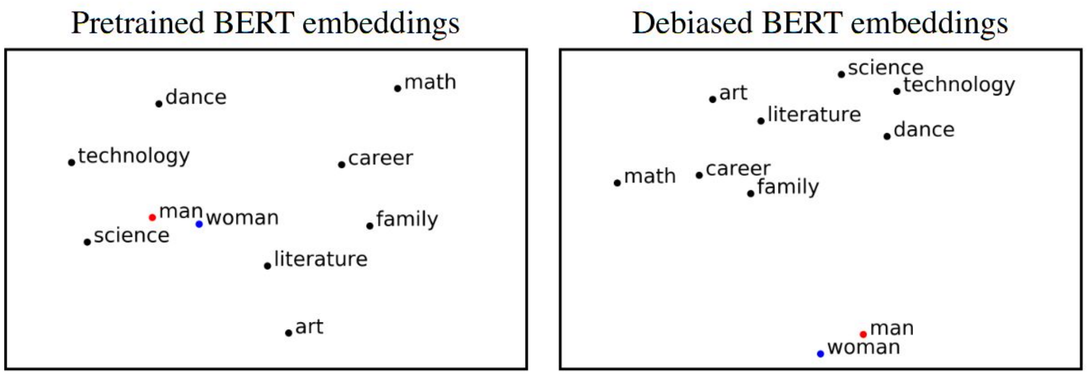
The nurse notified the patient that his shift would be ending in an hour.

Whose shift will end in an hour?

The nurse notified the patient that her shift would be ending in an hour.

Majority isn't always right!

⇒ Debiasing representations



Mehrabi et al: A survey on bias and fairness in machine learning, arXiv:1908.09635

MANCHESTER 1824
The University of Manchester

What if there's no data?

- Language models work so well because there's a lot of data to learn a rich representation
- what if a language has little (no) data available?
  ⇒ few-shot/transfer learning
  ⇒ cross-lingual representations (can help)
  ⇒ cross-lingual few-shot learning

Schick, Schütze: It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners arXiv:2009.07118
Conneau et al: Unsupervised Cross-lingual Representation Learning at Scale. ACL 2020
Nooralahzadeh et al: Zero-Shot Cross-Lingual Transfer with Meta Learning EMNLP 2020

The University of Manchester

In conclusion

- Neural network based approaches model tasks end-to-end
    - …And can learn to perform a task based on input-output examples
    - …if there's enough input-output examples
- Many tasks can be modelled as input-output examples
- More training data and bigger models usually means better performance

In conclusion

- Neural networks excel at inferring statistical patterns from training data
  - Superb performance if task is well represented
  - Unpredictable performance if application differs from training data
  - To interpret their behaviour on a fine-grained level, we can evaluate their inferred capabilities
  - We use task-specific data that requires a capability to be solved successfully to establish the capability

In conclusion

- Deep learning research has impact beyond the scientific community
  - Higher participation barrier can lead to underrepresentation
  - Bias in data can be exacerbated