NLP for Fake News Detection

Chieu Hai Leong (chaileon@dso.org.sg)
Distinguished Member of Technical Staff
DSO National Laboratories

Brief Introduction about myself

Worked in NLP for over 20 years

- PhD in machine learning (NUS, Singapore MIT Alliance, 2009)
- Part of an AI Lab in DSO with around 80 people
    - we do machine learning, nlp, computer vision, reinforcement learning etc.

My Research Interests:

- Natural language processing
    - Information extraction
    - Sentiment analysis
    - Fake news detection
- Machine Learning applied to
    - Chemistry, Cyber Security

Introduction to NLP

NLP: program computers to process and analyze large amounts of natural language data.

News,
Social media

Other sources of information

Read:
- Retrieval (search, rank)
- Recommendation
- Translate

Analyse
- Summarization
- Aggregation
- Inference

USER

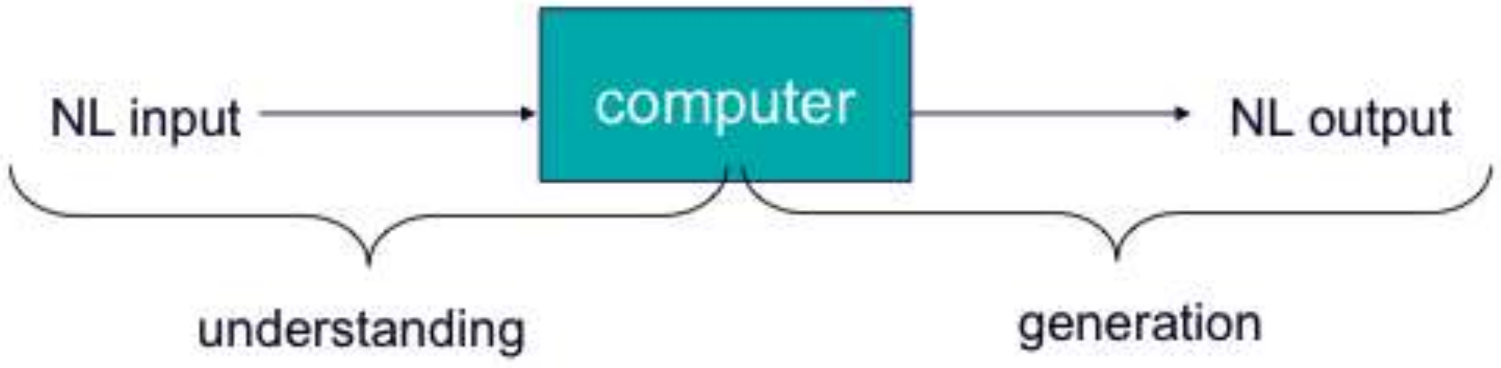
Guidelines for Writing Laboratory Reports for
The Department of Electrical Engineering and Computer Science
Approved by EE Faculty
August 30, 2004

Choose the Proper Report Format

Guidelines/Tips for your Report

Report:
- Text Generation

Modern applications of NLP

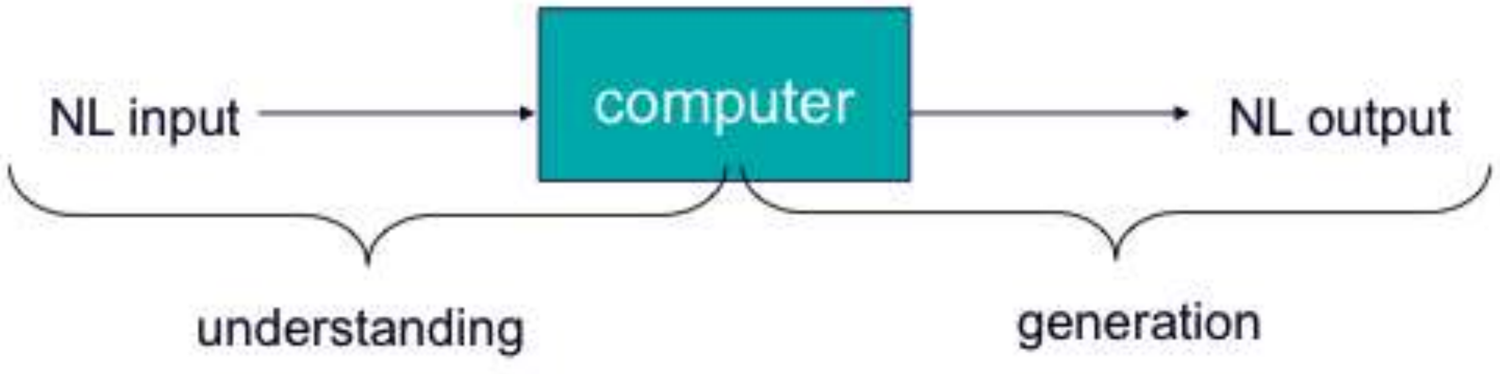


Dialogue based system, e.g. “I want a flight ticket to Washington”.

- Question answering, e.g.,
  - NL Input: “Who is the original voice of Miss Piggy?”
  - NL Output: “Frank Oz”.
- Machine translation, e.g.,
  - NL Input: Sentence in English
  - NL Output: Sentence in French
- Summarization
  - NL Input: Documents
  - NL Output: Summary

- Information retrieval
  - Google, Bing, Yahoo search.
  - Classifying documents into pre-defined topics
  - Clustering documents into topics

Fake News Detection, e.g., fact checking



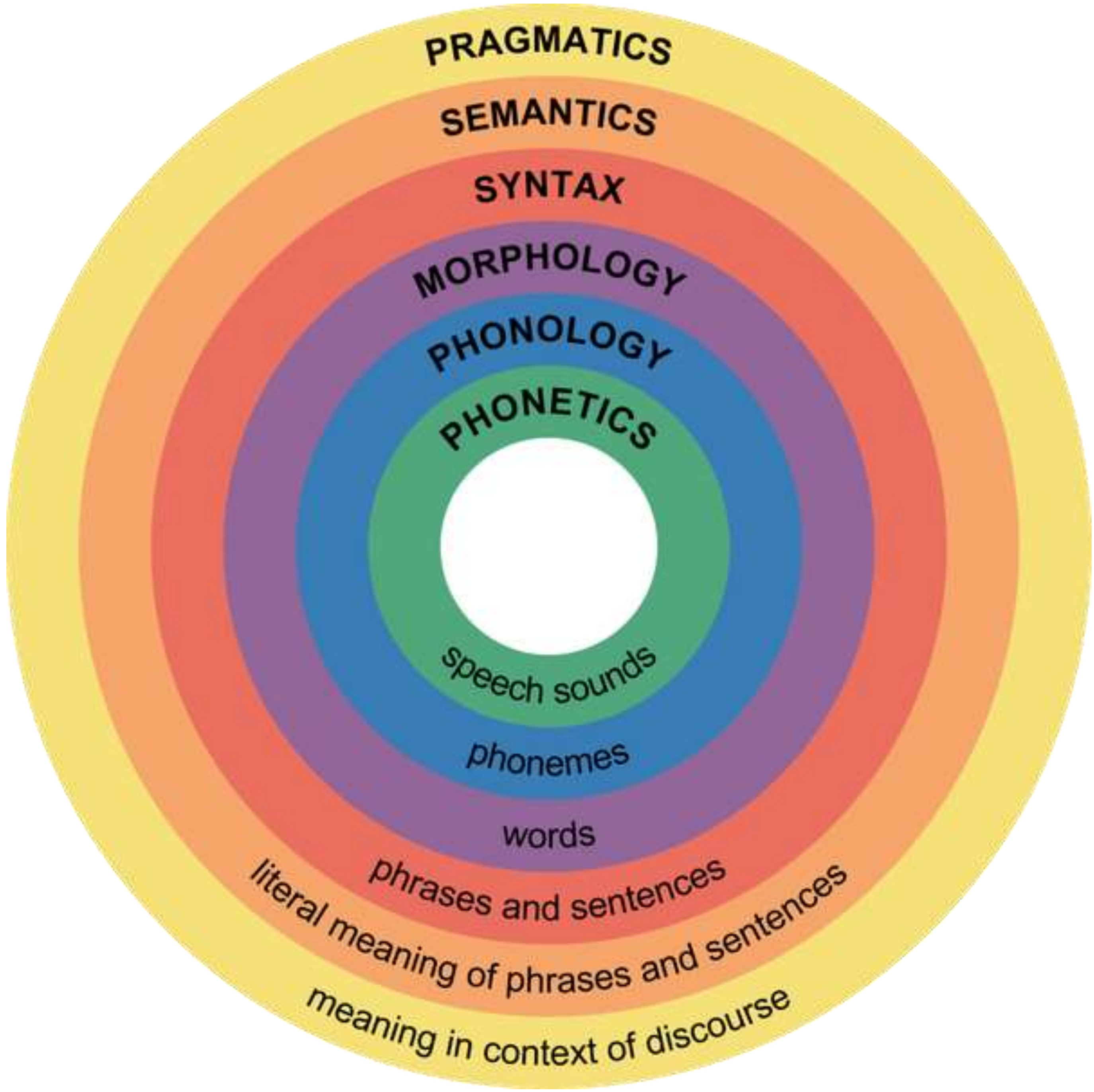
Claim, e.g.,
“Drinking bleach cures covid”

Computer

- True or Fake, or not enough information
- Explanation:
  - Evidence support or refuting the claim

NLP is hard

- Ambiguities at all levels
  - Syntax
  - Semantic
  - Discourse
  - Pragmatics



Syntax, Semantics, Discourse, Pragmatics

Syntax are rules and principles that govern the sentence structure.

- Q: What have four wheels and flies?
- A: Garbage Truck.

- Part-of-speech tagging and parsing
  - Tags:
    - https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
  - State-of-the-art:
    - http://nlpprogress.com/english/part-of-speech_tagging.html

Part-of-speech tagging

Alphabetical list of part-of-speech tags used in the Penn Treebank Project:

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Example:

Vinken	,	61	years	old
NNP	,	CD	NNS	JJ

Penn Treebank

A standard dataset for POS tagging is the Wall Street Journal (WSJ) portion of the Penn Treebank, containing 45 different POS tags. Sections 0-18 are used for training, sections 19-21 for development, and sections 22-24 for testing. Models are evaluated based on accuracy.

Model	Accuracy	Paper / Source	Code
Meta BiLSTM (Bohnet et al., 2018)	97.96	Morphosyntactic Tagging with a Meta-BiLSTM Model over Context Sensitive Token Encodings	
Flair embeddings (Akbik et al., 2018)	97.85	Contextual String Embeddings for Sequence Labeling	Flair framework
Char Bi-LSTM (Ling et al., 2015)	97.78	Finc Cha Wo	
Adversarial Bi-LSTM (Yasunaga et al., 2018)	97.59	Rob Tag	
Yang et al. (2017)	97.55	Trai witl	
Ma and Hovy (2016)	97.55	End dire	

Social media

The Ritter (2011) dataset has become the benchmark for social media part-of-speech tagging. This is comprised of some 50K tokens of English social media sampled in late 2011, and is tagged using an extended version of the PTB tagset.

Model	Accuracy	Paper
GATE	88.69	Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data
CMU	90.0 ± 0.5	Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters

<u>Syntax</u>, Semantics, Discourse, Pragmatics

<u>Syntax</u> are rules and principles that govern the sentence structure.

- Parsing:
  - I saw the man with the telescope/gun.

Syntax, Semantics, Discourse, Pragmatics

Semantics concern what words mean and how these meanings combine to form sentence meanings.

- Word level
  - Word sense disambiguation: The fisherman went to the bank.
- Sentence level
  - Almost all applications need to solve semantics
    - Sentiment analysis
    - Information extraction (names, relations, events)
    - Question answering
    - Etc.

Syntax, Semantics, Discourse, Pragmatics

Discourse concerns how the immediately preceding phrases or sentences affect the interpretation of the next phrase or sentence

Example: co-reference resolution

- Jack drank the wine on the table. It was brown and round.
- We gave the monkeys the bananas because they were hungry.
- We gave the monkeys the bananas because they were ripe.

Syntax, Semantics, Discourse, Pragmatics

Pragmatics concerns how sentences are used in different situations and how use affects the interpretation of the sentence.

"You have the green light" is ambiguous. It could mean

- you have green ambient lighting.
- you have a green light while driving your car.
- you can go ahead with the project.
- your body has a green glow.
- you have in your possession a light bulb that is tinted green.

One-hot encoding, BoW

How do you represent a document?

Vocabulary:
Man, woman, boy, girl, prince, princess, queen, king, monarch

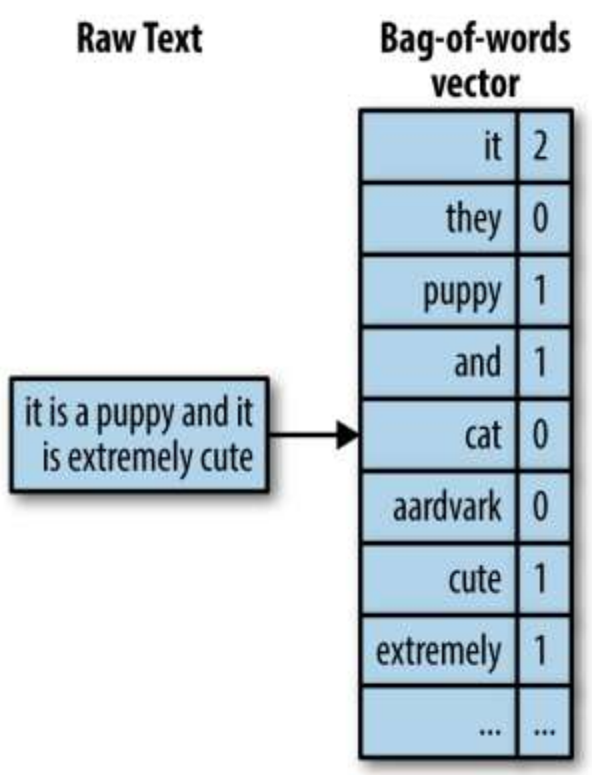
	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

One-hot encoding

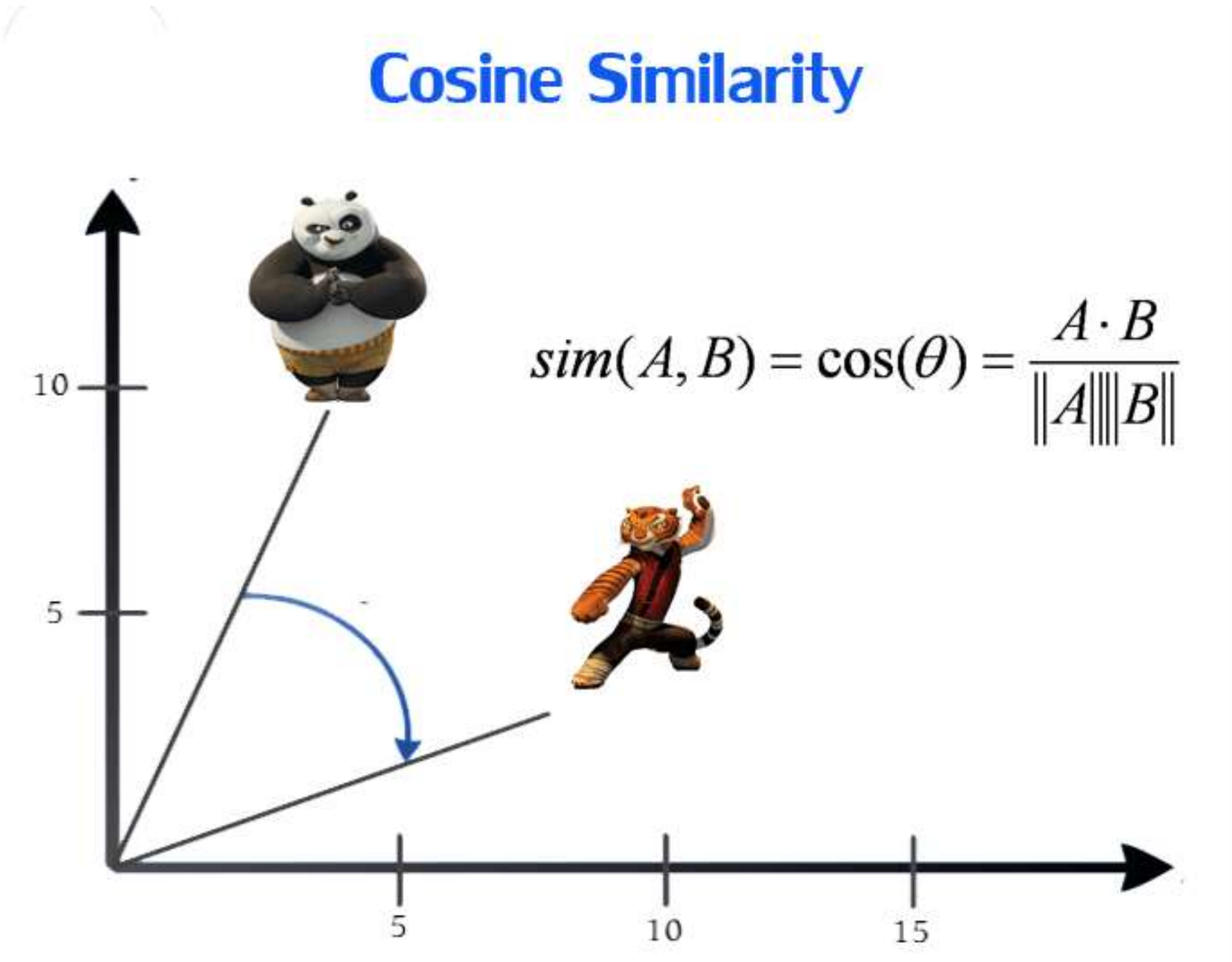
Man

1
0
0
0
0
0
0
0
0

Bag of words



Cosine similarity



1. Julie loves me more than Linda loves me
2. Jane likes me more than Julie loves me

Vocab	S2	S2
me	2	2
Jane	0	1
Julie	1	1
Linda	1	0
likes	0	1
loves	2	1
more	1	1
than	1	1

Sim(S1,S2) = 0.822

Cosine similarity on bag of words are often used in information retrieval for comparing or clustering documents.

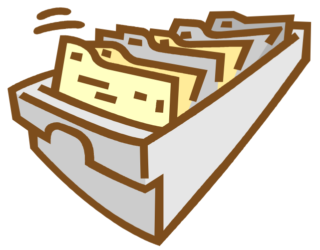
Learning Representation (~2014)

How do you represent a document?

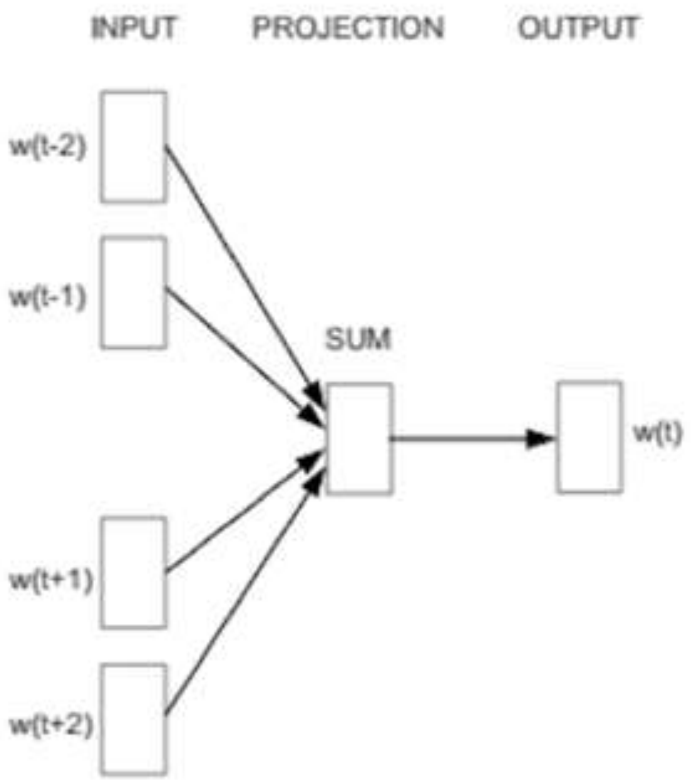
Word embedding

(learns a high dimensional vector representation for each word)

Lots of documents



Learning Representation



CBOW

Word embedding for each word in vocab

man =

1
0
0
0
0
0
0
0
0

man =

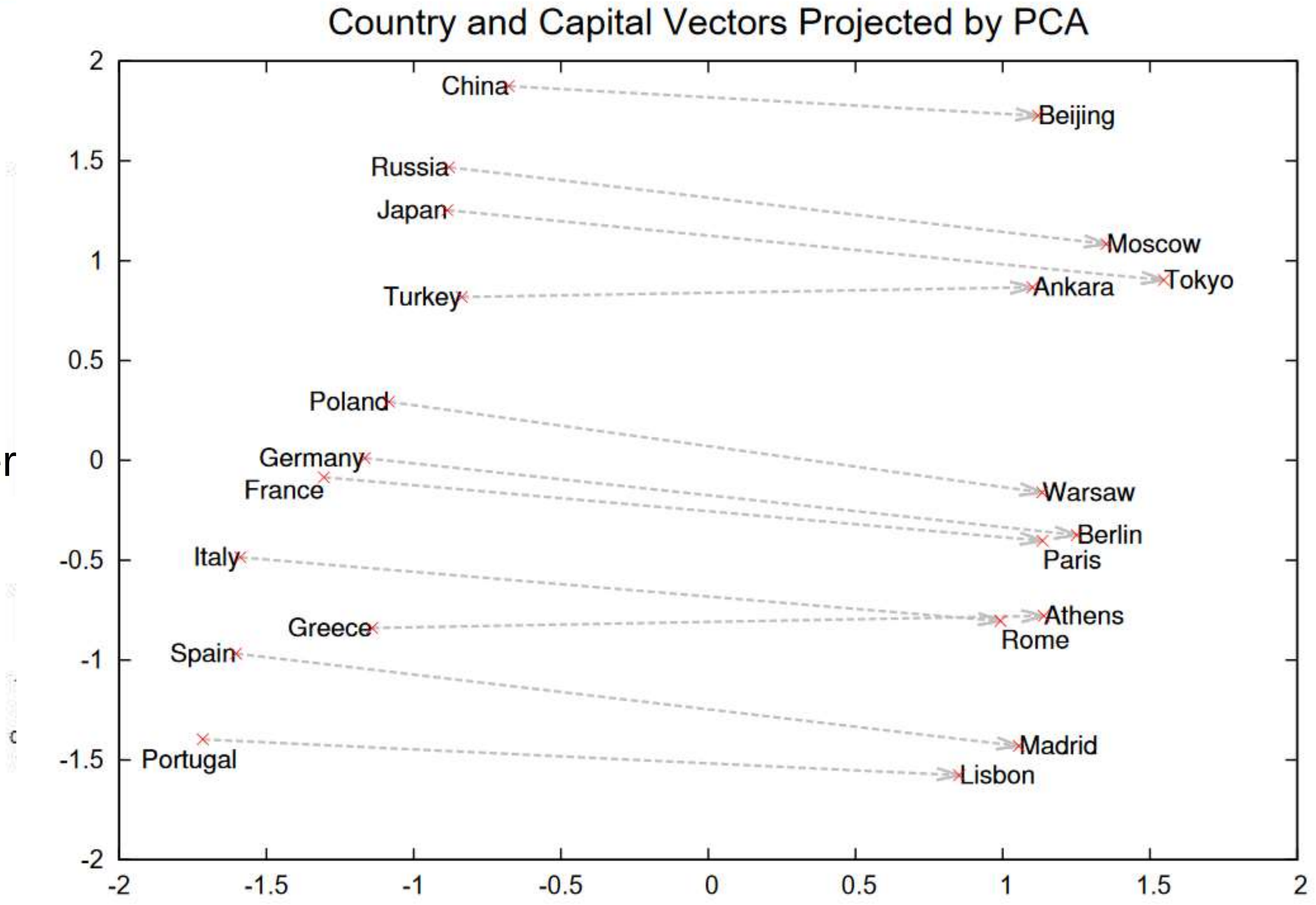
$$\begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \\ \cdots \end{pmatrix}$$

Dot product(man, language) = 0
Dot product(man, woman) = 0

Dot product(man, language)
is lower than
Dot product(man, woman)

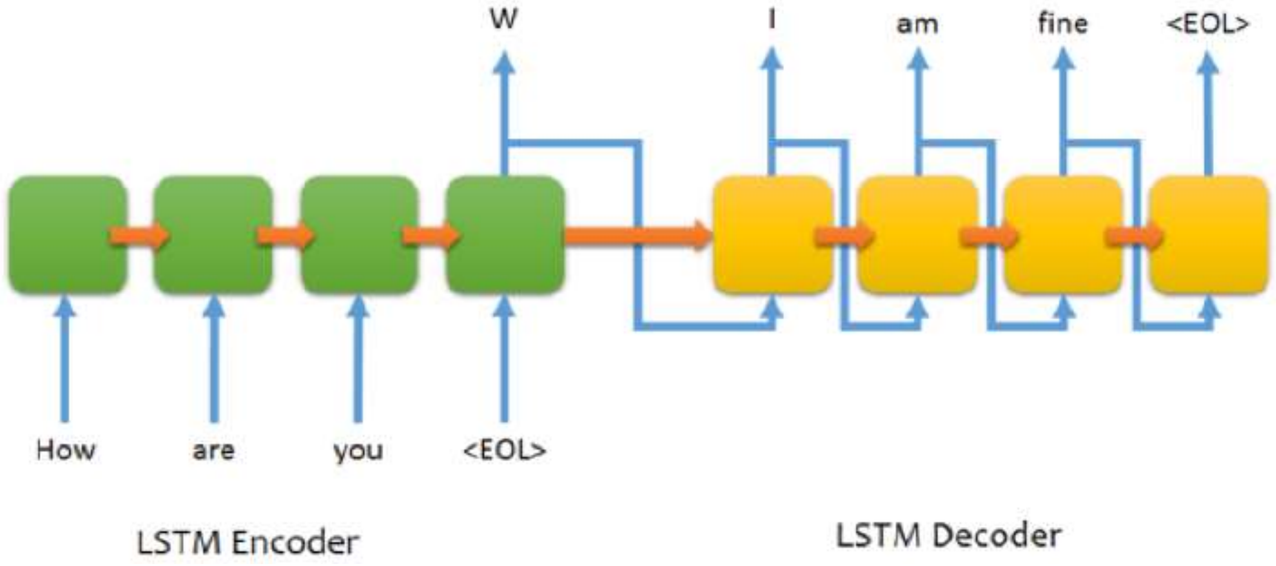
Word2Vec

tiger

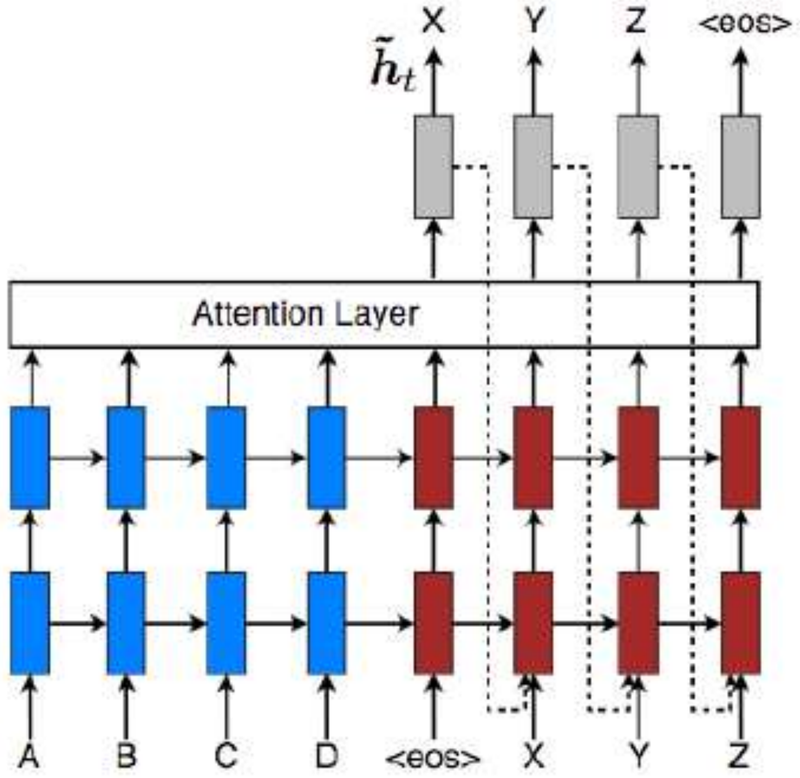


- Word embedding
  - Word2vec (Mikolov, 2013)
  - Glove (Stanford, 2014)
  - FastText (Facebook, 2016)

Seq2seq models (~2014)



Recurrent neural network models

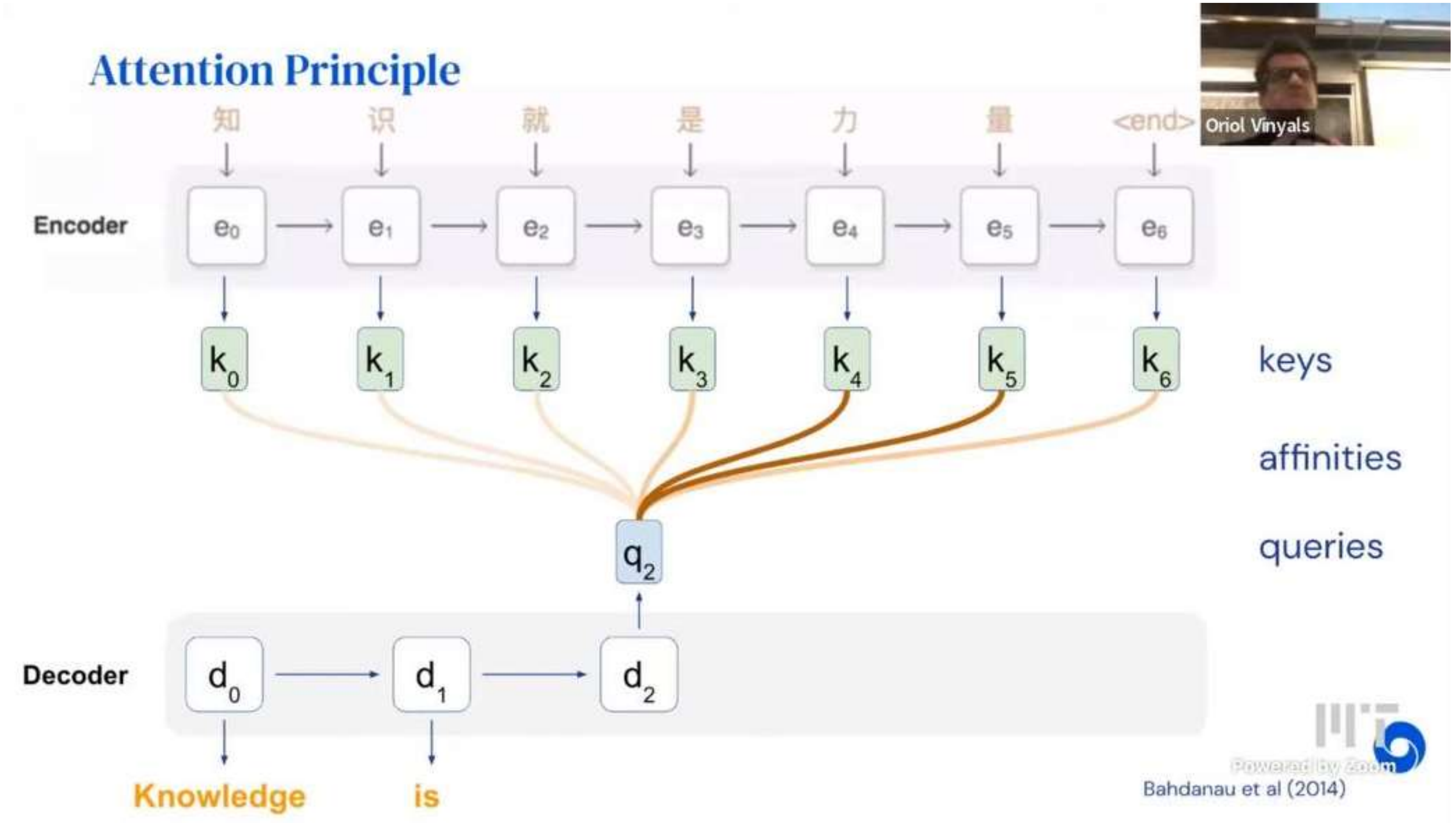


	Training Data
Input	English Text
Output	French Text
Parameters	380M
Data Size	6M Sentence Pairs, 340M Words

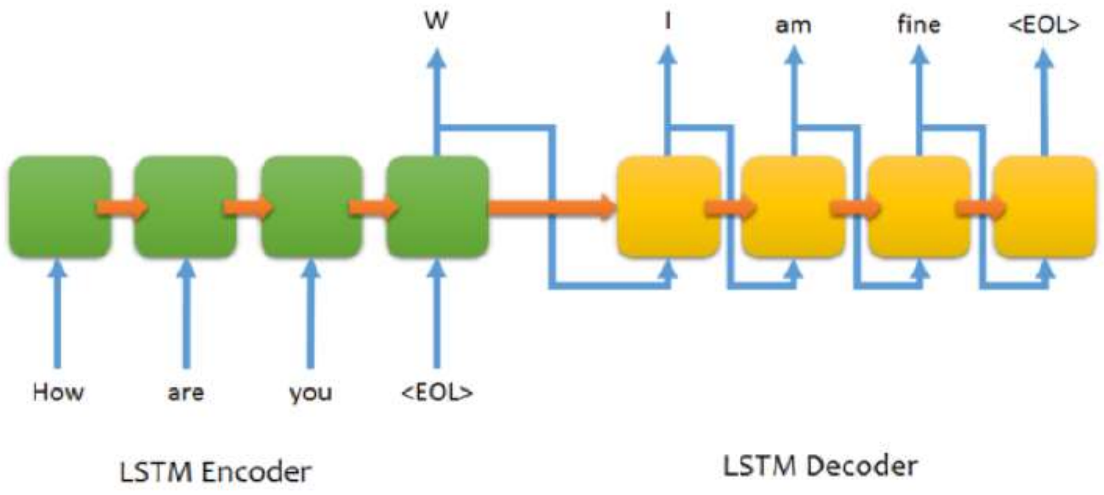
Transformer networks: the output words have direct connections (called "attention") to the input words

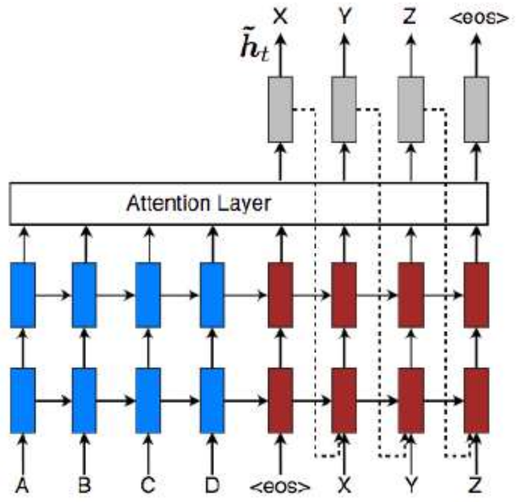
Attention mechanism

Screenshot from youtube talk by Oriol Vinyals

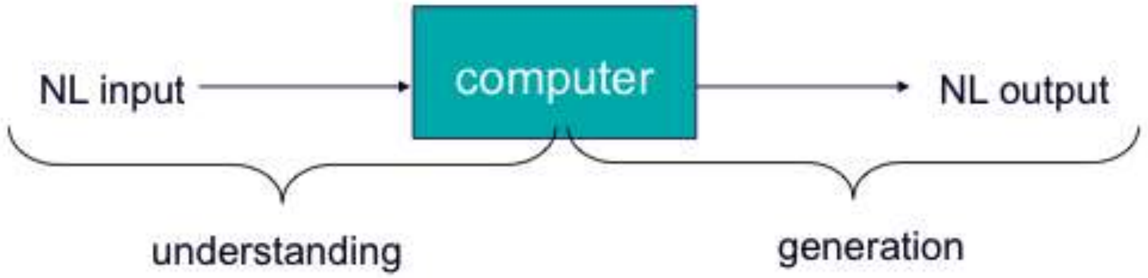


Sequence-to-sequence neural networks





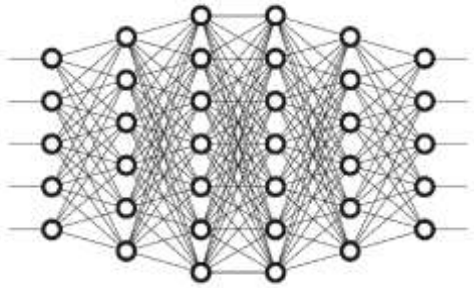
The same models apply to any problems in this form:



- Machine translation
- Chat bots (e.g., trained with open subtitles conversation)
- Summarization
  - However, it was found that such models are unable to learn to “copy” very well
  - Variants include seq2seq models with “pointers” for copying

Deep Learning Tsunami

Inputs, (e.g., image)



Outputs, (e.g., object labels in images)

deep neural networks

Imagenet competition:
classification into1000 categories

<u>2018 ACM A.M. Turing Award</u>
Citation: For conceptual and engineering breakthroughs that have made **deep neural networks** a critical component of computing.

<u>Imagenet 2012</u>: Geoff Hinton & students achieved 15.3% error, 2^{nd} place at 26.2%!

Today, it's < 5%!

Deep Learning and Natural Language Processing

"NLP is kind of like a rabbit in the headlights of the deep learning machine, waiting to be flattened." (Neil Lawrence, Deepmind Professor at Cambridge, 2015)

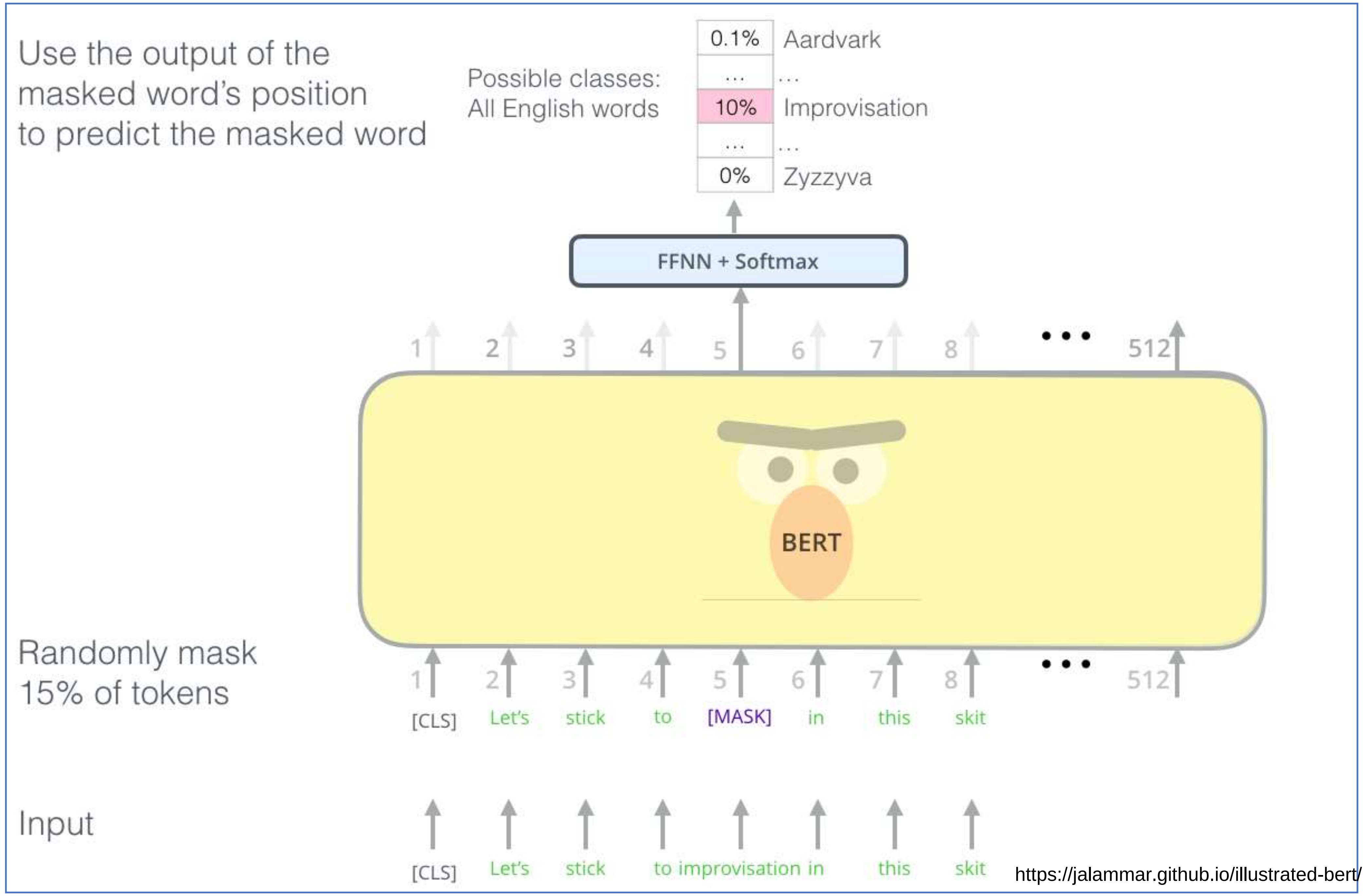
"I think that the most exciting areas over the next five years will be really understanding text and videos."

"The next big step for Deep Learning is natural language understanding"

Above quotes (and more) summarized in
Last Words: Computational Linguistics and Deep Learning, *Computational Linguistics*, Chris Manning, 2015.

Self-supervised Learning

Masked word prediction in BERT (Bidirectional Encoder Representations from Transformers).



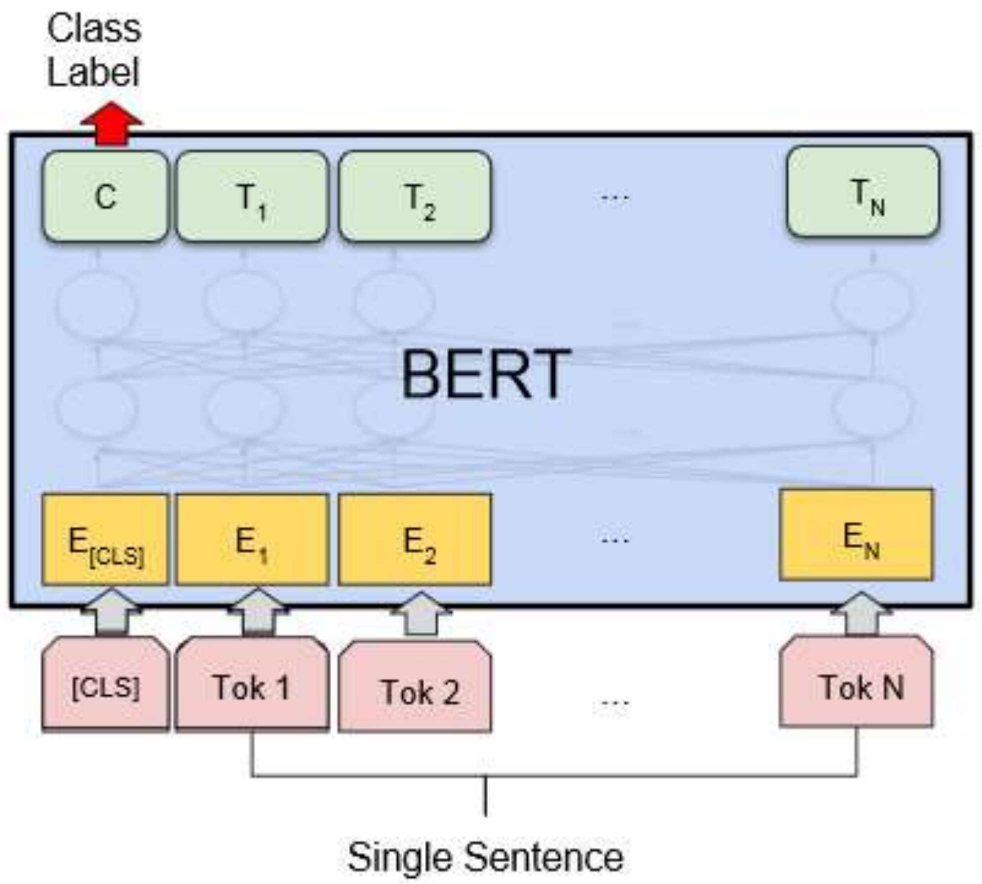
https://jalammar.github.io/illustrated-bert/

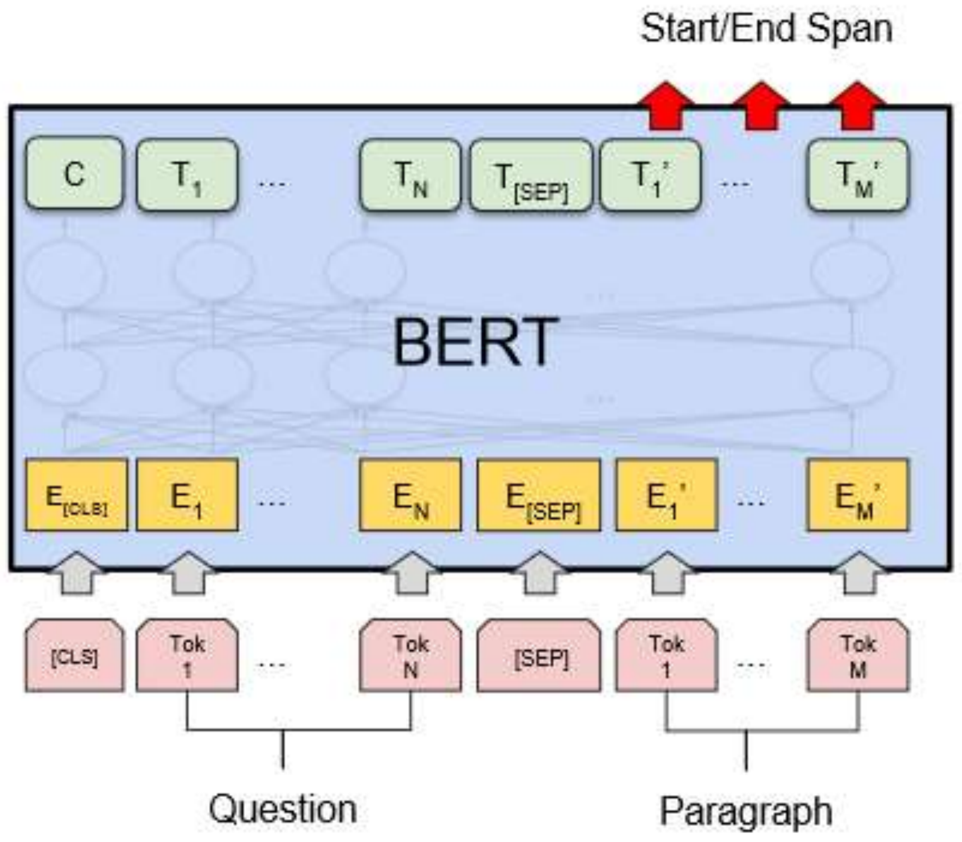
Contextual embedding: BERT

BERT is pre-trained on a huge data set.
BERT was applied to

- Single sentence classification (e.g., sentiment analysis)
- Sentence pair classification (e.g., textual entailment)
- Question answering (extract answers in text)
- Sequence labelling (e.g., extract names in text)

We can fine-tune BERT for specific task (e.g., sentiment analysis)

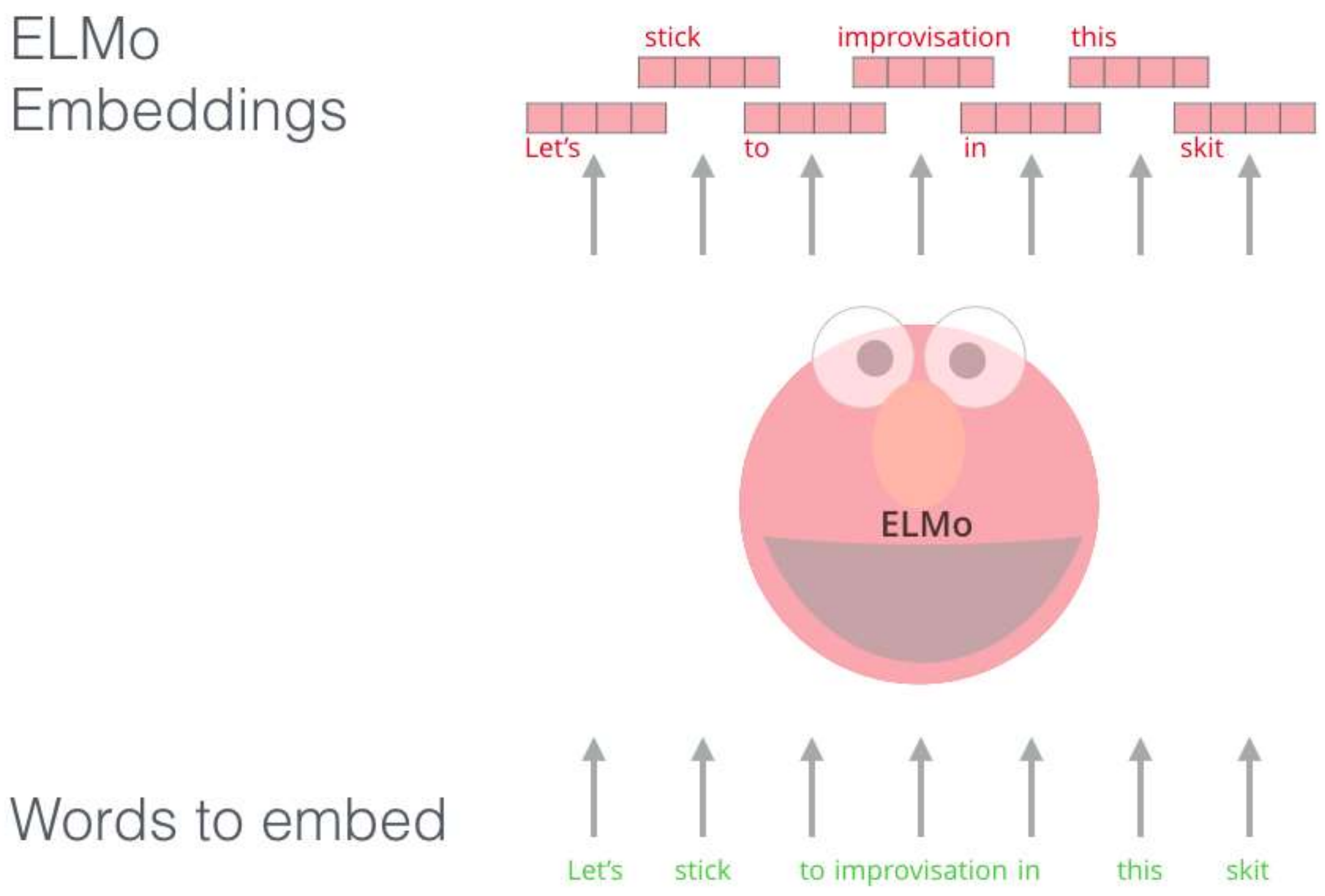




Contextual embedding

- Context independent word embedding
  - Word2vec (Mikolov, 2013)
  - Glove (Stanford, 2014)
  - FastText (Facebook, 2016)

- Context dependent word embedding
  - Elmo (UW and AllenAI, 2018)
  - OpenAI GPT (OpenAI, 2018)
  - Bert (Google, 2019)



Language Representation

History of learning language representation

- Non contextualized word embedding
  - **Word2vec** [MCCD13], **Glove** [PSM14]
- **Contextualized word embedding**
  - RNN, e.g., **ELMO** [PNZtY18]
  - **Transformers (attention networks)**
    - **GPT** [RNSS18], **GPT2** [RWC+19]
    - **BERT** [DCLT18]
    - **XLNET** [YDY+19]
    - …

ELMO
Embeddings from Language Models

Feb 2018

BERT
Bidirectional Encoder Representations from Transformers

Oct 2018

Generative pre-training

Feb 2019

XLNET

Post-BERT

May 2020

General Language Understanding Evaluation (GLUE)

Researchers designed a suite of tasks as a yardstick for general purpose language processing

- Grammatical correctness
- Sentiment analysis
- Semantic similarity
- Textual entailment

GLUE

SuperGLUE

Textual entailment examples

Hypothesis	Text	Judgement
Some men are playing a sport.	A soccer game with multiple males playing.	Entailment
Two men are smiling and laughing at the cats playing on the floor.	An older and younger man smiling.	Neutral
The man is sleeping	A man inspects the uniform of a figure in some East Asian country.	Contradiction

BERT and Transformers

BERT: Bidirectional Encoder Representations from Transformers
Google, Oct 2018

Trained on BooksCorpus (800M words) and English Wikipedia (2,500M words)

GLUE

Improved GLUE score to 80.5% (7.7% over the second place)
Today, Google's T5 achieved 90.3, while human was scoring 87.1

SuperGLUE

Google's T5 achieved 89.3, while human scored 89.8

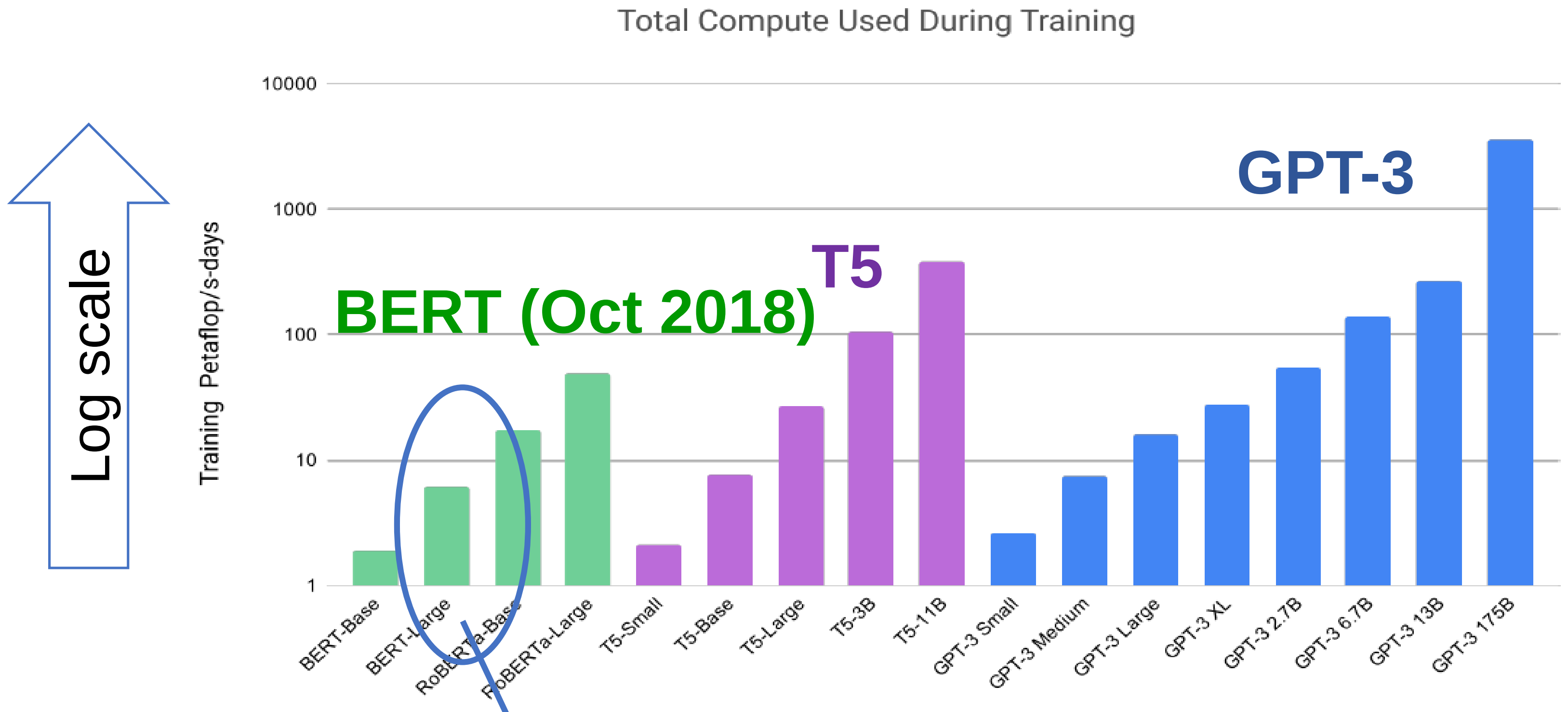
BERT Base Model ~ 1438 CO_2e
- Nearly 1 person flight NY to SF
- (to train one model)

Consumption	CO_2e (lbs)
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Energy and Policy Considerations for Deep Learning in NLP, Strubell et al., Aug 2019

Model and Training Data (Pg. 9)

Figure 2.2



“GPT-3 175B”
3.14 x 10^{23} FLOPS

Lambdalabs:
355 GPU-years and $4.6M for a single training run.

- Oct 2018 (Google): 64 TPUs in 4 days
- May 2020 (Microsoft) 1,024 GPU in 44min

1000x

Training Data Table 2.2:

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

GPT3 works by generating text

Prompt GPT3 with some text (e.g., a question), and GPT3 will "complete the story", within 2048 characters.

```
Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church
has agreed to a historic split - one that is expected to end in the
creation of a new denomination, one that will be "theologically and
socially conservative," according to The Washington Post. The majority of
delegates attending the church's annual General Conference in May voted to
strengthen a ban on the ordination of LGBTQ clergy and to write new rules
that will "discipline" clergy who officiate at same-sex weddings. But
those who opposed these measures have a new plan: They say they will form a
separate denomination by 2020, calling their church the Christian Methodist
denomination.
The Post notes that the denomination, which claims 12.5 million members, was
in the early 20th century the "largest Protestant denomination in the U.S.,"
but that it has been shrinking in recent decades. The new split will be the
second in the church's history. The first occurred in 1968, when roughly
10 percent of the denomination left to form the Evangelical United Brethren
Church. The Post notes that the proposed split "comes at a critical time
for the church, which has been losing members for years," which has been
"pushed toward the brink of a schism over the role of LGBTQ people in the
church." Gay marriage is not the only issue that has divided the church. In
2016, the denomination was split over ordination of transgender clergy, with
the North Pacific regional conference voting to ban them from serving as
clergy, and the South Pacific regional conference voting to allow them.
```

GPT2 demo: Writing with Transformers

Write With Transformer

Get a modern neural network to
auto-complete your thoughts.

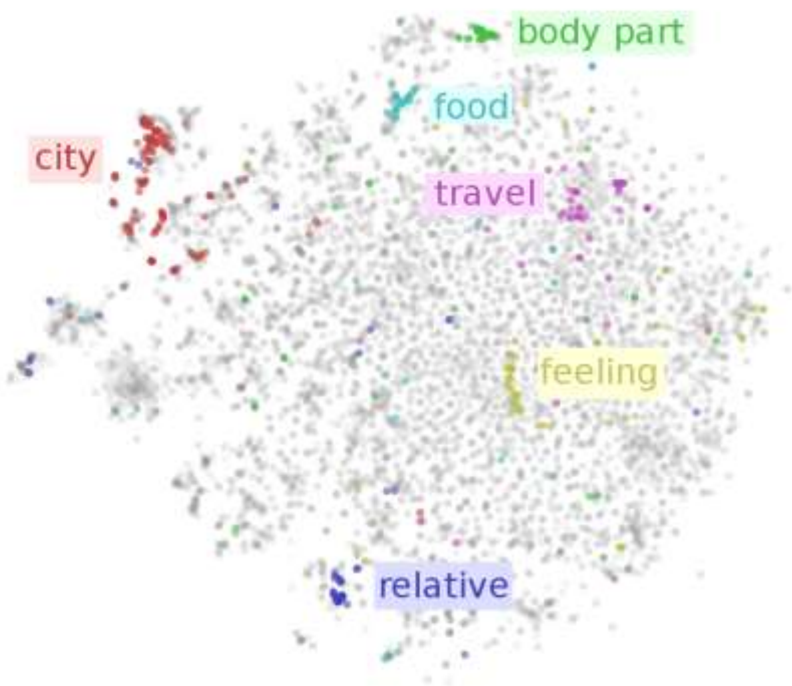
This web app, built by the Hugging Face team, is the official demo of the
🤗/transformers repository's text generation capabilities.

Star 66,194

Multi-lingual NLP

Monolingual Embedding

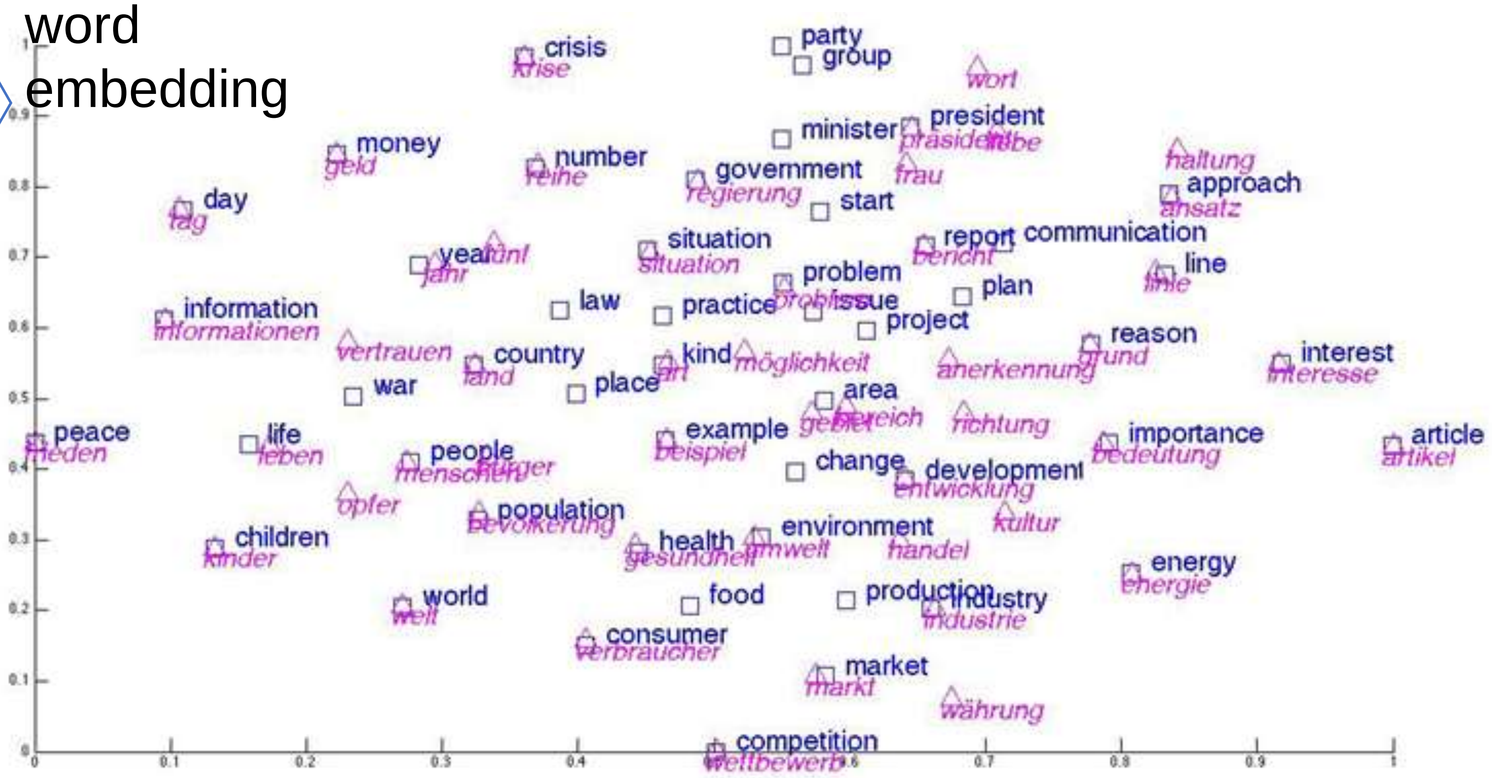
Corpus of raw textual data → Neural Networks → High dimensional (e.g., 300 dimension) word embedding



Multilingual Embedding

English
German
…
Indonesian

→ Neural Networks → Multi-lingual word embedding



Data Bias

Data Bias:

We could be solving the problem *right* for the *wrong* reasons!

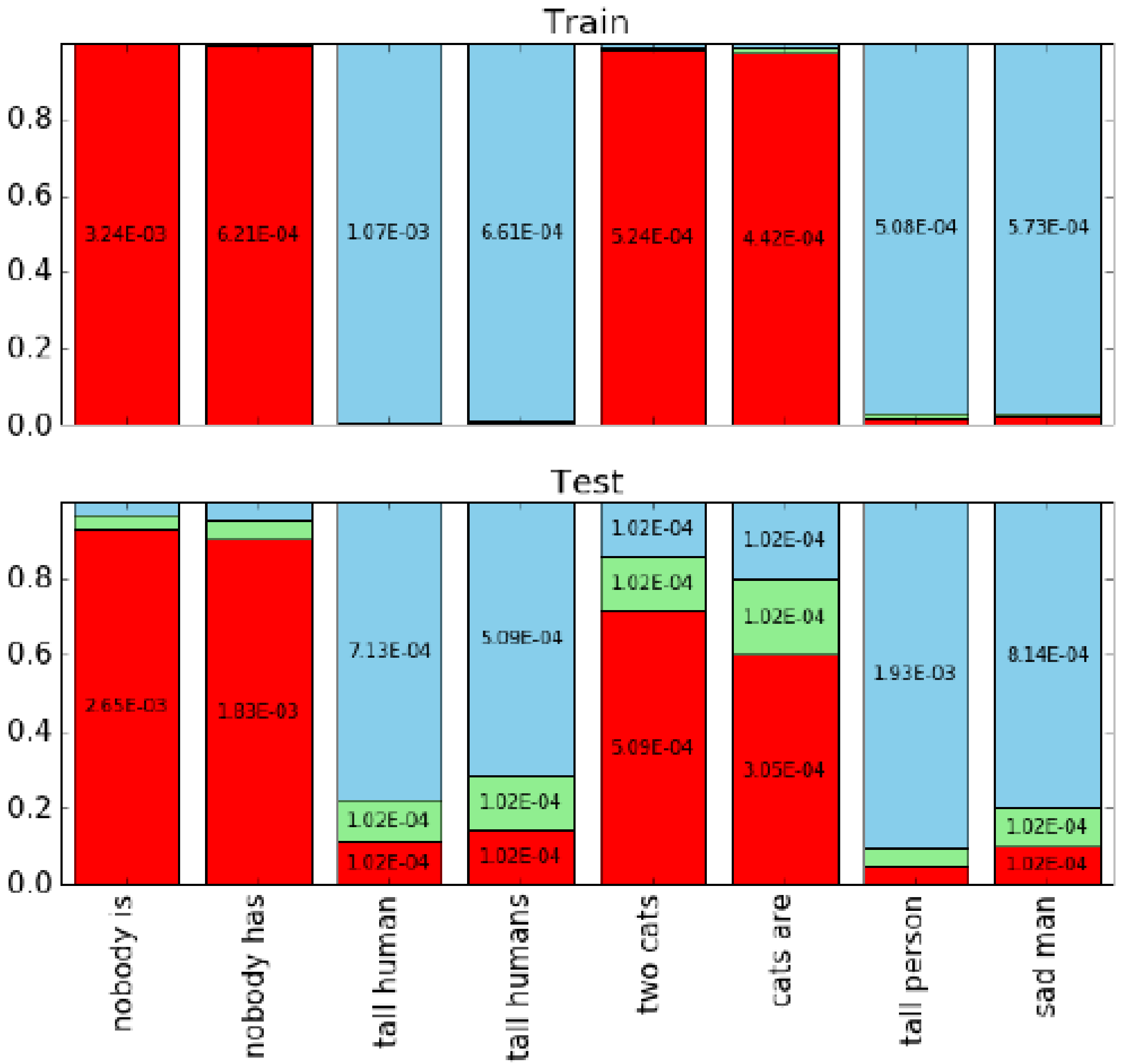
Textual Entailment: classification just on the hypothesis alone achieved 64% accuracy, well above random (33%).

Label	Premise	Hypothesis
Contradict	Black man in a nice suite that matches the rest of the choir he's singing with near a piano.	**Nobody is** singing
Neutral	An excited, smiling woman stands at a red railing as she holds a boombox to one side.	A **tall human** standing.
Entail	A group of people are walking across the street.	**Some humans** walking

Tell-tale phrases

Red: contradict,

blue: entail, green: neutral



Let’s get into Fake News Detection

Fake news can be a threat to national security

Local examples of fake news

April 2015

THE STRAITS TIMES

Student who posted fake PMO announcement on Mr Lee Kuan Yew's death given stern warning

May 2020

cna

Singapore

Cabby jailed for posting fake COVID-19 'intel' on food outlet closures, urging panic buying

April 2020

THE STRAITS TIMES

Coronavirus pandemic

Coronavirus: Fake news used to stir up unhappiness in dorms, says Shanmugam

The authorities will take action against those who deliberately spread falsehoods, says minister

May 2021

"**The new form of coronavirus in Singapore** is said to be **very dangerous for children**. It could reach Delhi in the form of a third wave. My appeal to the Central government: 1. **Cancel all air services with Singapore** with immediate effect 2. Work on vaccine alternatives for children on a priority basis,"
Tweet from Delhi Chief Minister Arvind Kejriwal

NLP for fake news detection

Spot Claims → Check Claims

Rumor Detection — Fact-Checking

- Rumor Detection (collaboration with SMU)
  - Given a social media thread, determine if it is rumor, and if so, if it is True, Fake, or Unverified.

- Fact Checking (collaboration with MIT)
  - Given a claim, check whether it is true or fake based on evidence retrieved from the web (or Wikipedia).

Serene Yeo did the fact checking work

Serena Khoo did the rumor detection work

Photo: Attending Neurips in Dec 2019 at Vancouver, just before COVID broke out.

Rumor Detection

"walmart donates $10,000 to support darren wilson and the on going racist police murders #ferguson #boycottwalmart URL"

Is this a rumor?
Is it real, fake, or unverified?

Shooting of Michael Brown

From Wikipedia, the free encyclopedia

"Michael Brown Jr." redirects here. For other people with the name, see Michael Brown (disambiguation).

On August 9, 2014, **Michael Brown Jr.**, an 18-year-old black man, was fatally shot by 28-year-old white Ferguson police officer **Darren Wilson** in the city of Ferguson, Missouri, a suburb of St. Louis.[2]

Rumor Detection

- Controversy detection from community response → Looking for claims that have high tendency to be fake by analysing content posted by the community
- Example:

Source

"walmart donates $10,000 to support darren wilson and the on going racist police murders #ferguson #boycottwalmart URL"

Reply 1

"@70torinoman @TefPoe Though it wouldn't surprise me, is there proof of this? Just googled and couldn't find."

Reply 2

"@70torinoman @TefPoe Walmart donates $10k to Darren Wilson?! No way this is true, right? Source? Link? #Ferguson #sayitaintso"

Reply 5

"@70torinoman #thanksWalmart donates $10,000 to defend #Wilson after beaten in struggle for gun w/250LN man #Ferguson URL"

Reply 7

"@70torinoman … Well I will be paying off that credit card and never returning!! I will not shop somewhere that condones murder!!"

Reply 3

"@lyburtus1 @70torinoman @TefPoe yeah I googled it to I'm the the type to jump all over what's on social media"

Reply 4

"@davidgrelle @70torinoman @TefPoe not true WM can't support this due to laws regulating public cos"

Reply 6

"@SAGandAFTRA @70torinoman That pic is not of #MikeBrown. Lie debunked first week after shooting."

Previous work: [Ma et al., 2018]

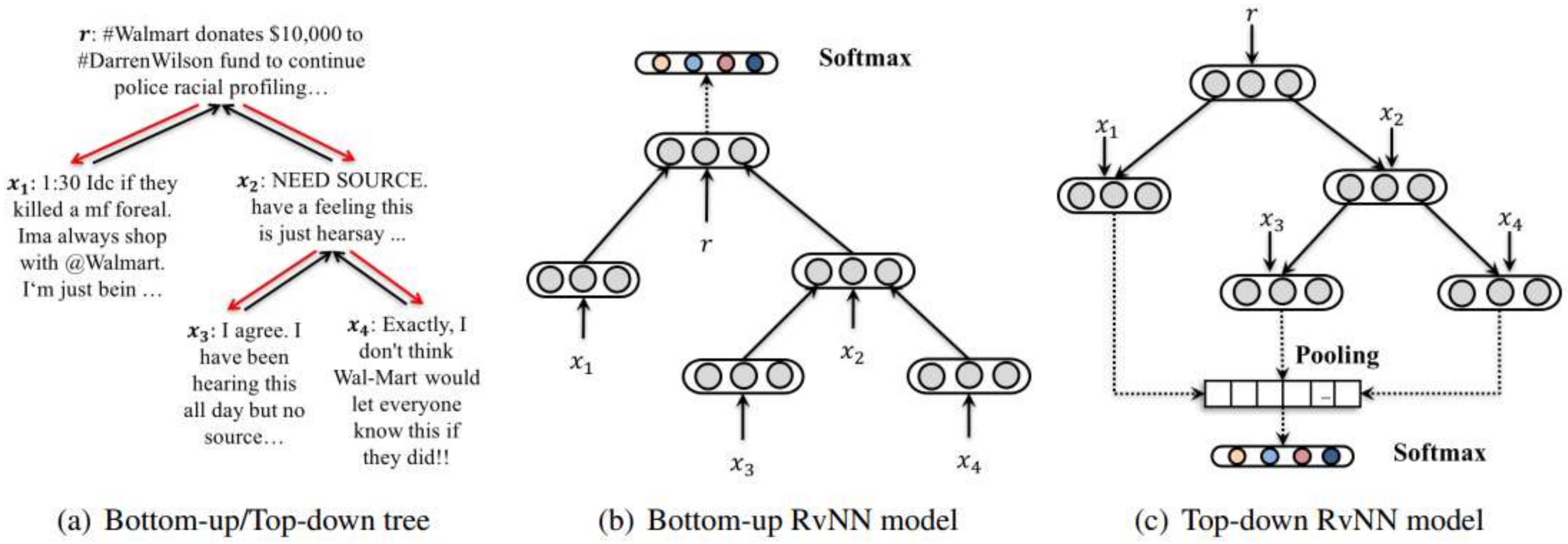


Figure from "Rumor Detection on Twitter with Tree-structured Recursive Neural Networks. Ma et al., 2018."

Models a thread in a tree structure with recursive neural networks.

Our approach: post level attention network

Ling Min Serena Khoo, Hai Leong Chieu, Zhong Qian, Jing Jiang: **Interpretable Rumor Detection in Microblogs by Attending to User Interactions.** AAAI 2020: 8783-8790

Contributions:

- Post and word level attention for interpretable results
- Structure aware methods do not always perform better
    - Is tree structure really important in twitter? Twitter conversations are mostly flat in nature. Each user sees the entire thread before replying.

Attention mechanism (is interpretable)

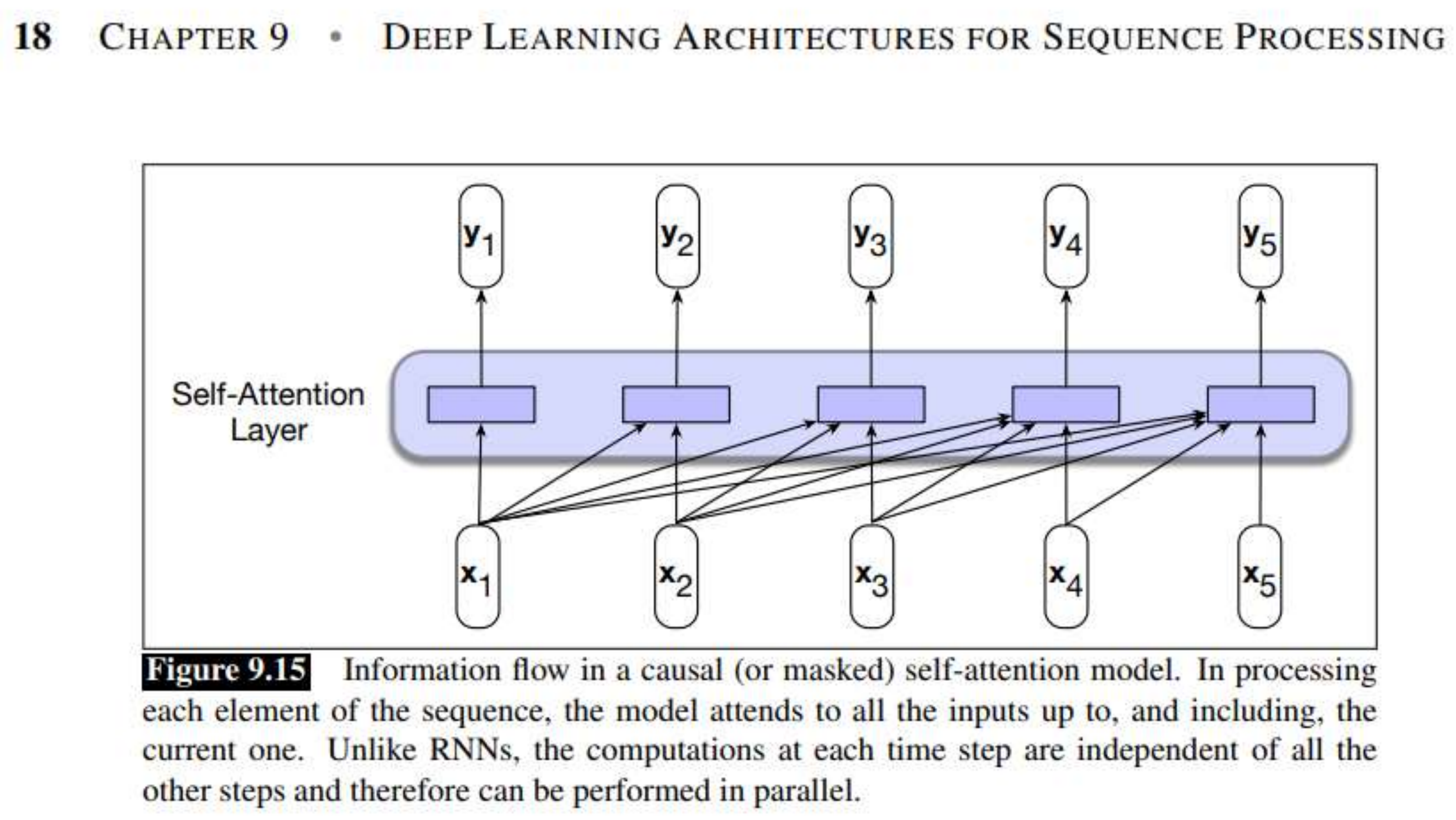
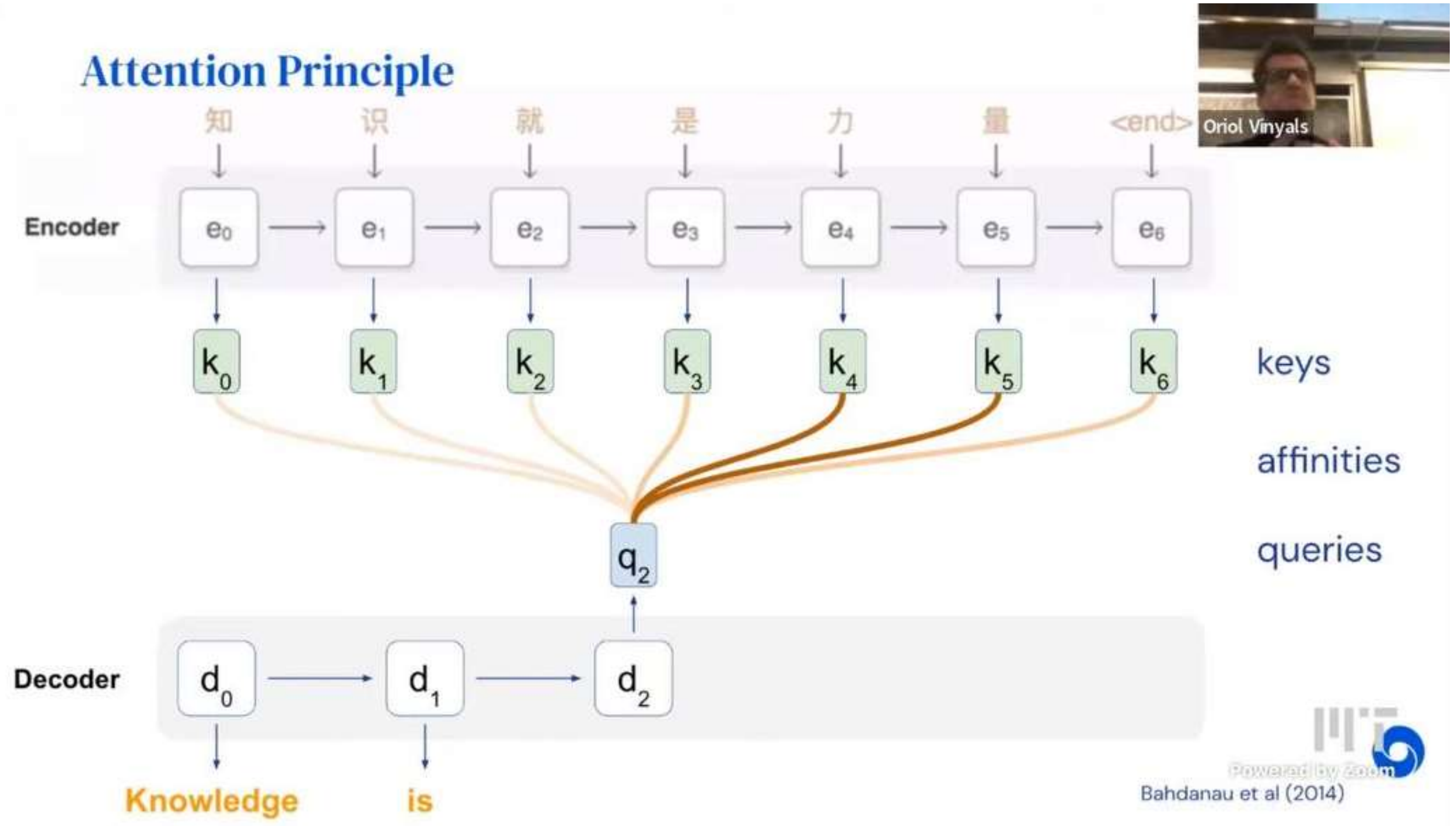

18 CHAPTER 9 • DEEP LEARNING ARCHITECTURES FOR SEQUENCE PROCESSING

Figure 9.15 Information flow in a causal (or masked) self-attention model. In processing each element of the sequence, the model attends to all the inputs up to, and including, the current one. Unlike RNNs, the computations at each time step are independent of all the other steps and therefore can be performed in parallel.

Screenshot from youtube talk by Oriol Vinyals



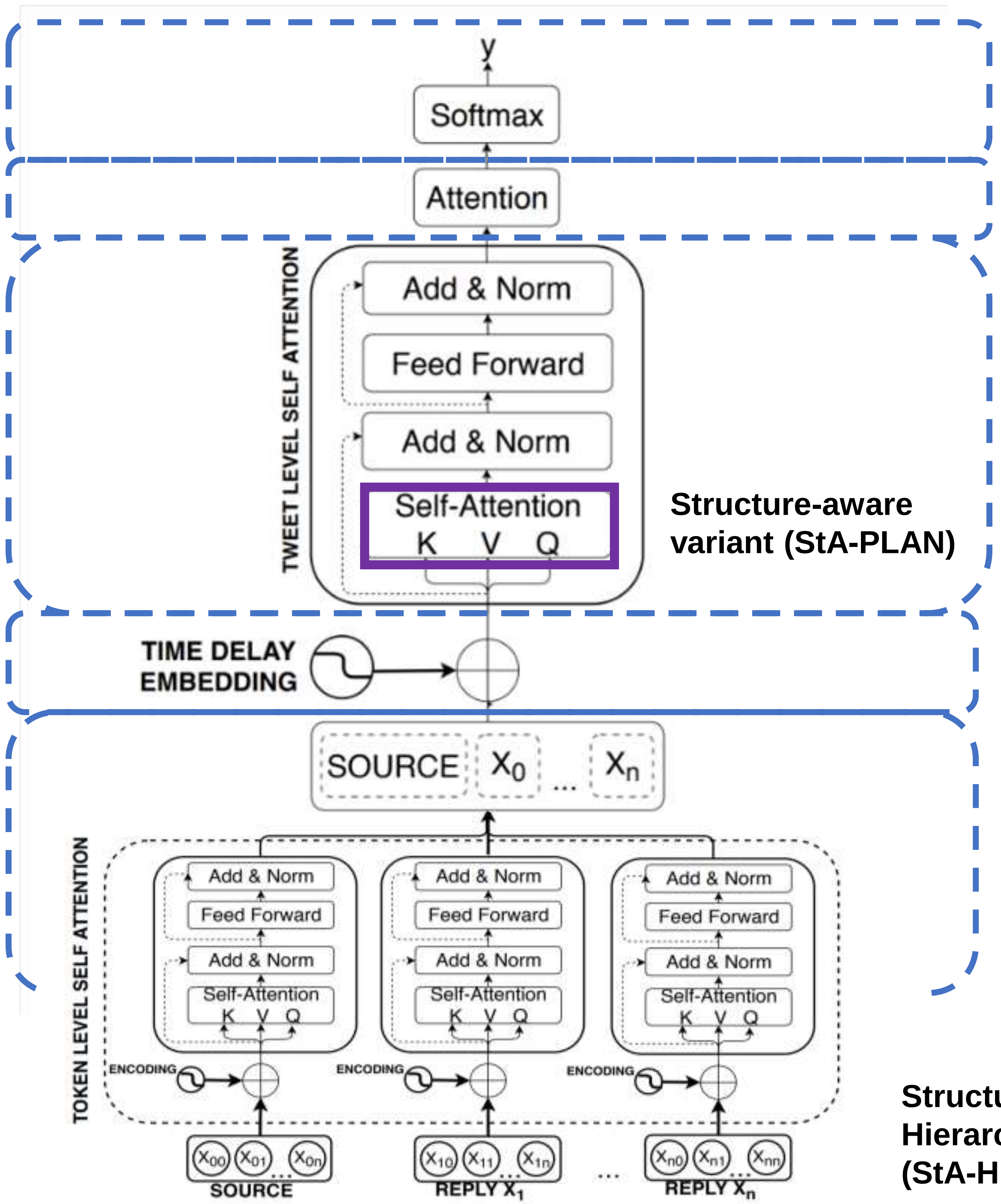
Attention in deep learning can be broadly interpreted as a vector of importance weights.

Interpretability: the features with more important weights (more heavily attended to) are the "main reasons" for the output of the model.

- Attention is not Explanation. Jain and Wallace. NAACL 2019.
- Attention is not not Explanation. Wiegreffe and Pinter. EMNLP 2019.

Hierarchical Post Level Attention Network (AAAI 2020)

PLAN + Time Delay
StA-PLAN + Time Delay
StA-HiTPLAN + Time Delay



Structure-aware variant (StA-PLAN)

Structure-aware with Hierarchical Token variant (StA-HiTPLAN)

Predict as True/ False/ Unverified/ Non-rumour

Compute attention weights on each transformed tweet embedding to get one representation of all tweets

Propagate and aggregate information between tweets

Append time delay embedding to sentence embedding of tweet

Obtain sentence representation of each tweet by token-level self attention mechanism

Post level attention

(Label) Claim	Important Tweets		#Tweets
(UNVERIFIED) Surprising number of vegetarians secretly eat meat	1	@HuffingtonPost .... then they aren't vegetarians.	33
	2	@HuffingtonPost this article is stupid. If they ever eat meat, they are not vegetarian.	
	3	@HuffingtonPost @laurenisaslayer LOL this could be a The Onion article	
(TRUE) Officials took away this Halloween decoration after reports of it being a real suicide victim. It is still unknown. URL	1	@NotExplained how can it be unknown if the officials took it down...... They have to touch it and examine it	46
	2	@NotExplained did anyone try walking up to it to see if it was real or fake? this one seems like an easy case to solve	
	3	@NotExplained thats from neighbours	
(FALSE) CTV News confirms that Canadian authorities have provided US authorities with the name Michael Zehaf-Bibeau in connection to Ottawa shooting	1	@inky_mark @CP24 as part of a co-op criminal investigation one would URL doesn't need facts to write stories it appears.	5
	2	@CP24 I think that soldiers should be armed and wear protective vests when they are on guard any where.	
	3	@CP24 That name should not be mentioned again.	

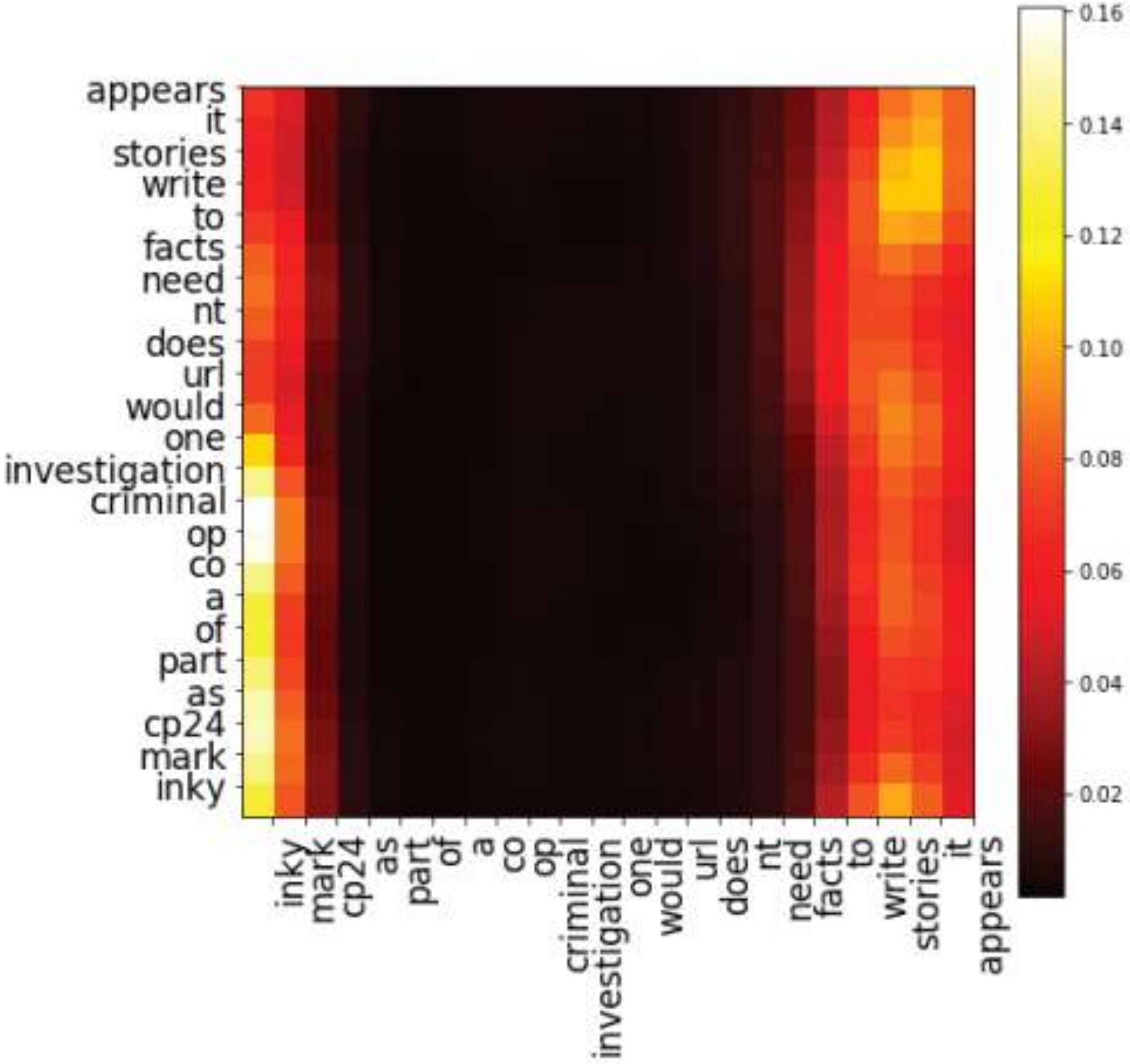
@inky_mark @CP24 as part of a co-op criminal investigation one would assume.Media doesn't need facts to write stories it appears.

In predicting the claims in the first column, the top comments with the most attention weights are listed in the second column. The third column #tweets show the total number of tweets in each thread.

Word level self-attention

Re-tweet:
@inky mark @CP24 as part of a co-op criminal investigation one would <URL> doesn’t need facts to write stories it appears.

High (attention) weights were placed on the phrase “*facts to write stories it appears*” to classify the claim as a rumor.



Higher attention weights

Results on 3 data sets: outperformed RvNN

Great Results (>80%) on two data sets!

Problem solved?

Method	Twitter15					Twitter16				
	Accuracy	F	T	U	NR	Accuracy	F	T	U	NR
BU-RvNN (Original)	70.8	72.8	75.9	65.3	69.5	71.8	71.2	77.9	65.9	72.3
TD-RvNN (Original)	72.3	75.8	82.1	65.4	68.2	73.7	74.3	83.5	70.8	66.2
BU-RvNN (Ours)	70.5	71.0	72.1	73.0	65.5	80.6	75.5	89.3	83.0	73.4
TD-RvNN (Ours)	65.9	66.1	68.9	71.4	55.9	76.7	69.8	87.2	81.3	66.1
PLAN	84.5	**85.8**	**89.5**	80.2	82.3	**87.4**	**83.9**	91.7	**88.8**	**85.3**
StA-PLAN	**85.2**	84.6	88.4	**83.7**	84.0	86.8	83.3	**92.7**	**88.8**	82.6
StA-HiTPLAN	80.8	80.2	85.1	76.0	81.7	80.7	76.5	88.8	82.0	74.9
PLAN + time-delay	84.1	84.2	87.3	80.3	84.2	84.8	77.6	89.7	85.6	84.9
StA-PLAN + time-delay	85.0	85.7	88.3	81.4	**84.4**	86.6	83.3	92.3	86.6	84.2

Not so good on the third data set (PHEME).

Except when we re-split the train/test split (last row: random split).

Method	Macro F-Score
Branch LSTM - Multitask	35.9
Tree LSTM - Multitask	37.9
BCTree LSTM - Multitask	37.1
PLAN	36.0
StA-PLAN	34.9
StA-HiTPLAN	37.9
PLAN + Time Delay	38.6
StA-PLAN + Time Delay	36.9
StA-HiTPLAN + Time Delay	**39.5**
StA-HiTPLAN + Time Delay (Random split)	77.4

PHEME data

Table 1. Outcome of the annotation of rumours.

Event name	Rumour stories	Annotated threads	Rumour threads	Non-rumour threads
Sydney Siege	61	1321	535	786
Ottawa Shooting	51	901	475	426
Charlie Hebdo	61	2169	474	1695
Germanwings	19	1022	332	690
Ferguson	42	1183	291	892
Prince to play in Toronto	6	241	237	4
Gurlitt	3	386	190	196
Putin missing	6	266	143	123
Essien has Ebola	1	18	18	0
TOTAL	250	7507	2695	4812

Twitter threads mined for 9 stories.

PHEME train/test split based on events.

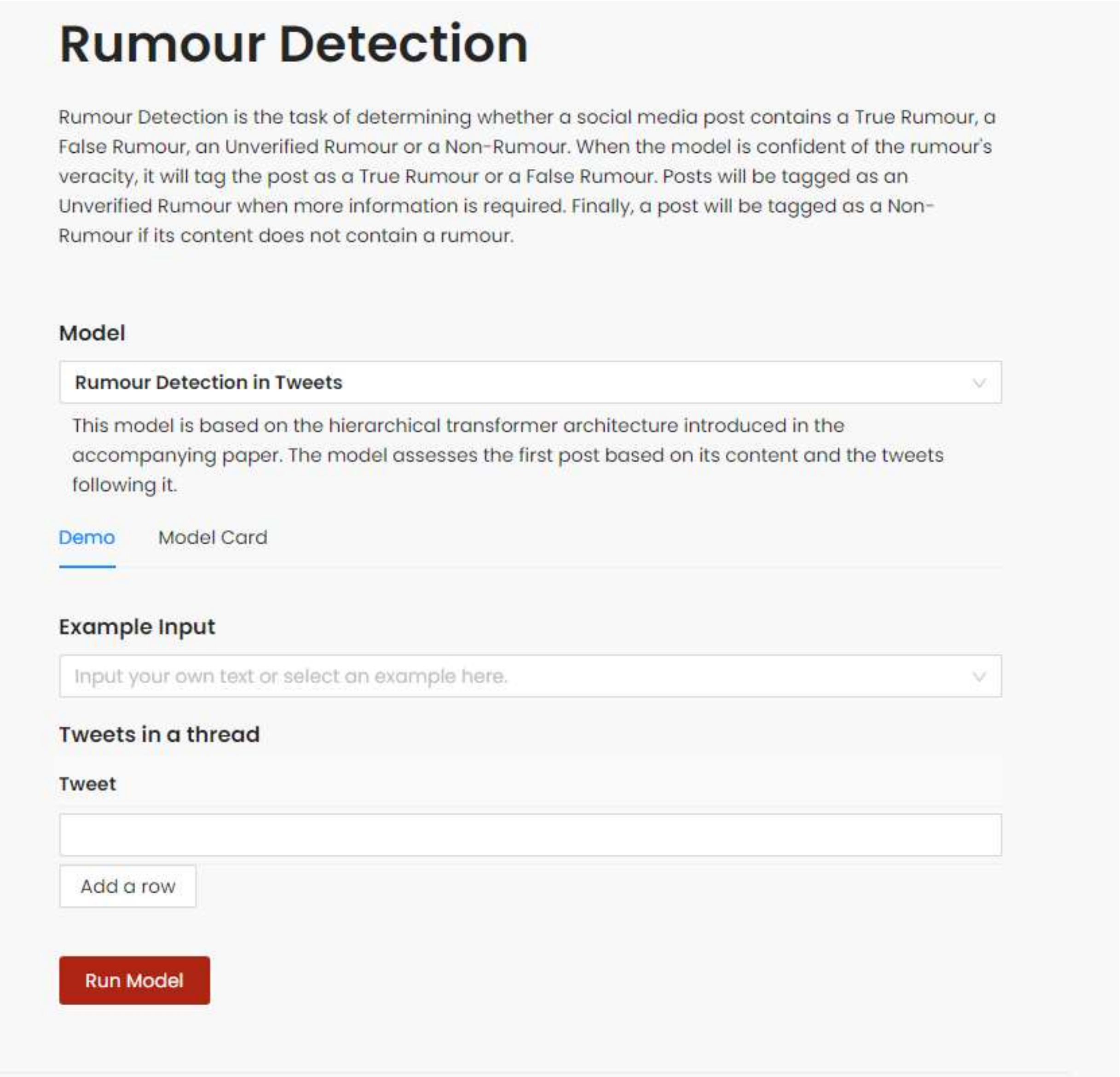
When we do random splits on PHEME, performance improved from <40% to nearly 80%.

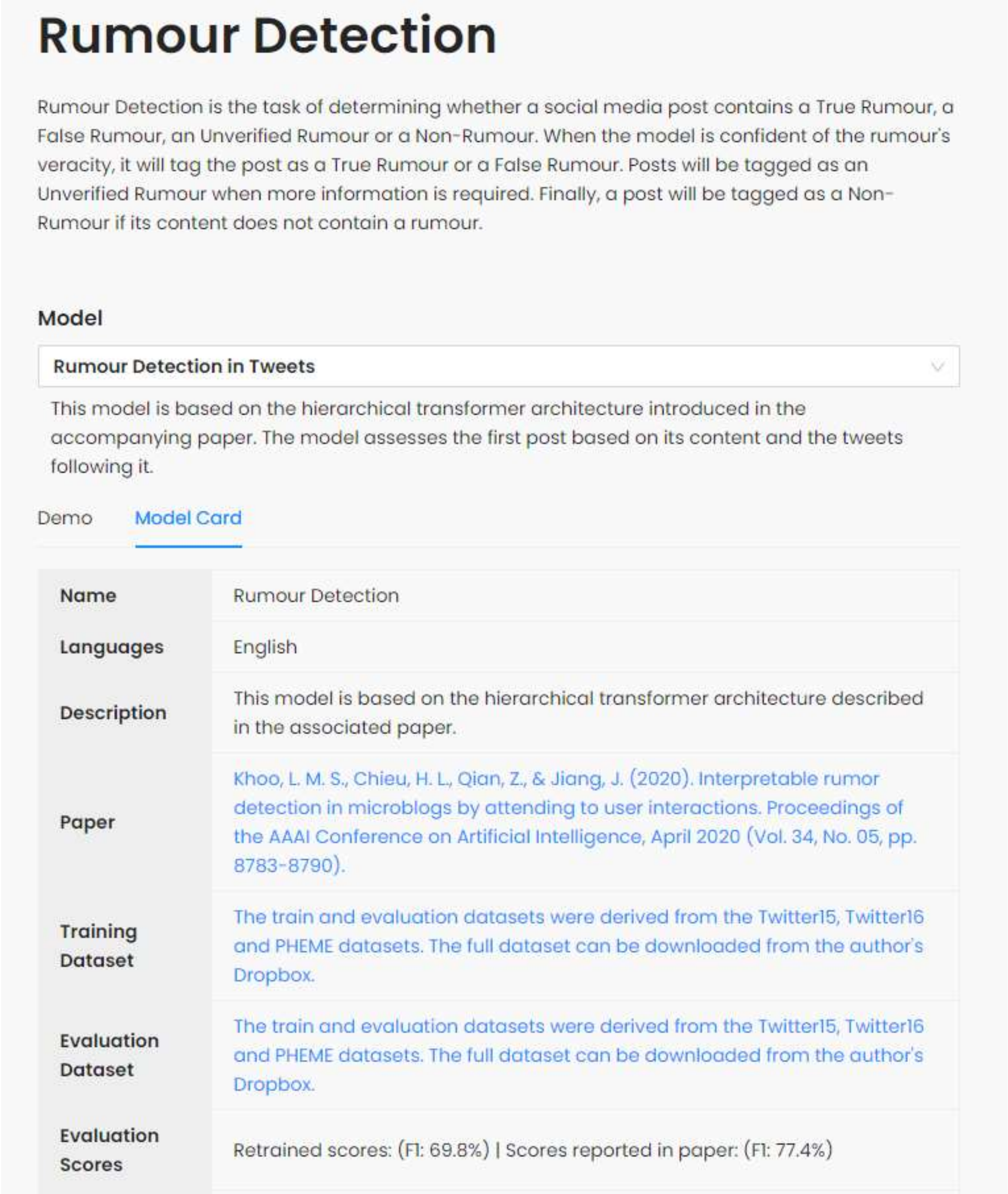
In random splits, you see threads from each event during training, and
you test on the same events. The machine just need to learn that Essien has Ebola is fake, and Charlie Hebdo is true to get high accuracy.

Demo

Demo has been set up on AISG NLP Hub:

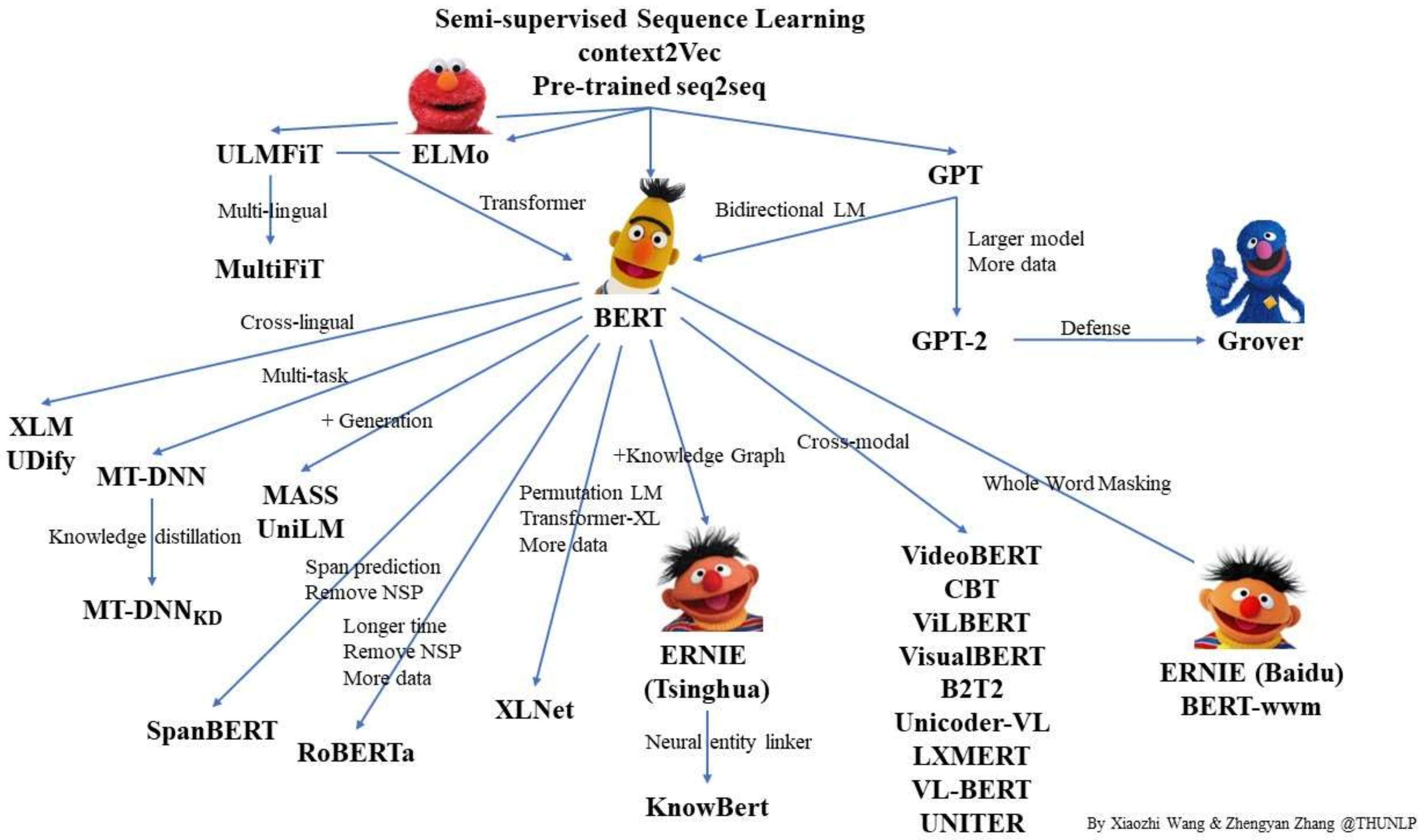
- https://sgnlp.aisingapore.net/rumour-detection-twitter





Why we didn't use BERT

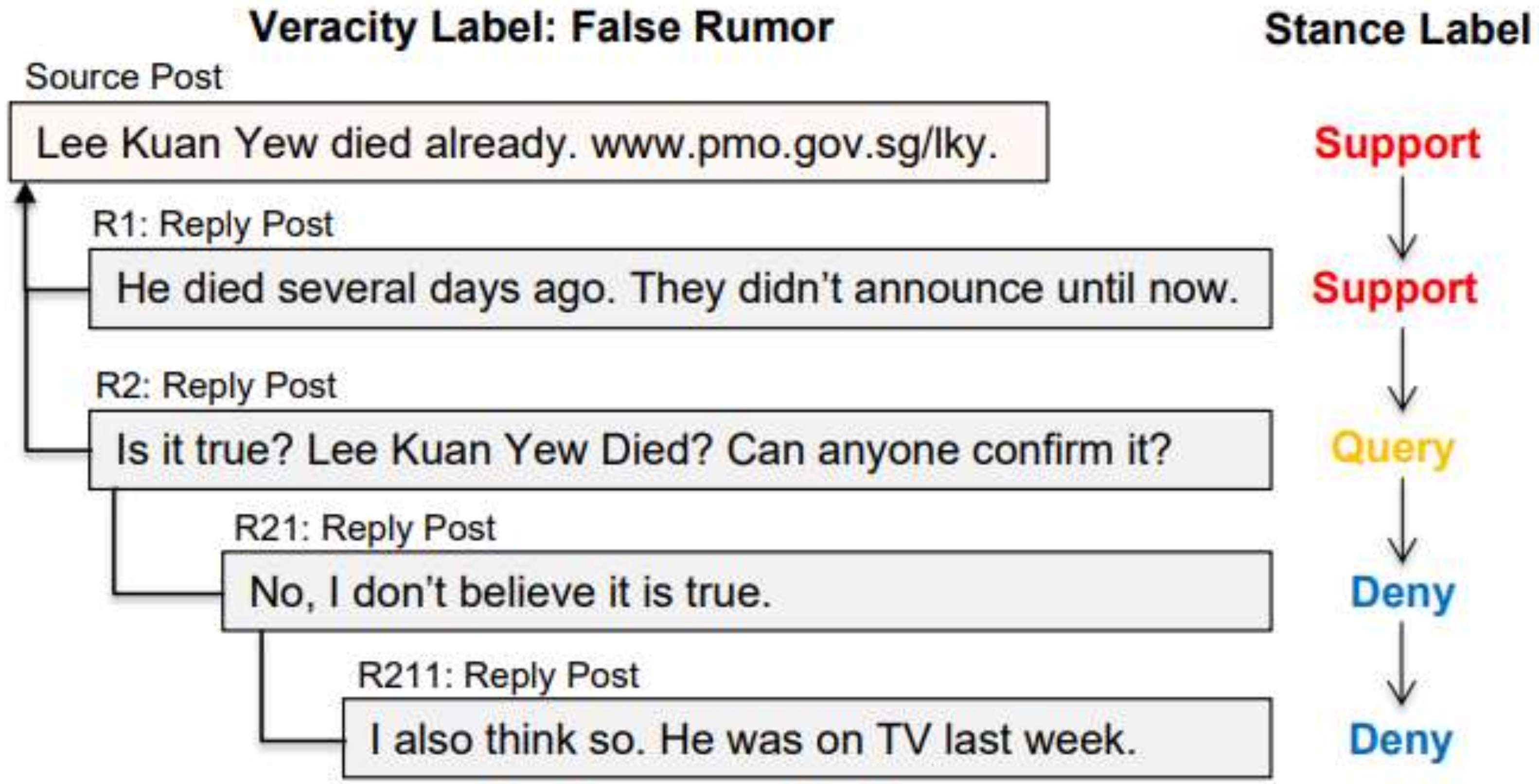
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al., NAACL 2019)



- BERT has a token limit of 512 tokens. Threads are longer than that, so we cannot fit into BERT.
- Computationally expensive.
- But we will use it in the next attempt at the problem (next slide)

Using Stance Information (and BERT)

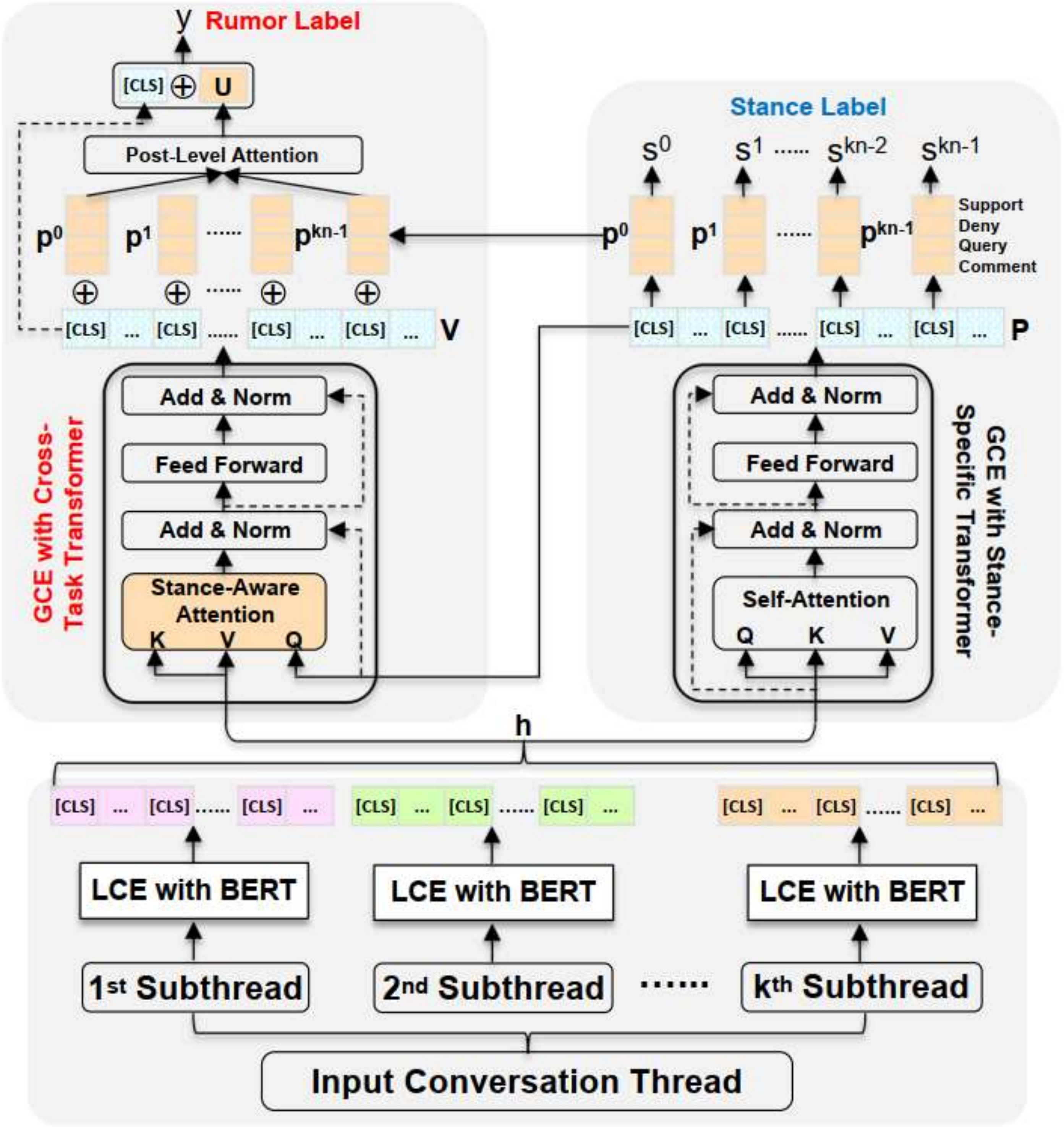
Jianfei Yu, Jing Jiang, Ling Min Serena Khoo, Hai Leong Chieu, Rui Xia:
Coupled Hierarchical Transformer for Stance-Aware Rumor Verification in Social Media Conversations. EMNLP (1) 2020: 1392-1401



			Stance Labels				Rumor Veracity Labels		
Dataset	#Threads	#Tweets	*#Support*	*#Deny*	*#Query*	*#Comment*	*#True*	*#False*	*#Unverified*
SemEval-17	325	5,568	1,004	415	464	3,685	145	74	106
PHEME	2,402	105,354			-		1,067	638	697

Stance-Aware Rumor Detection (EMNLP 2020)

- Dual Attention BERT
  - Breaks up thread into sub-threads to fit into BERT's size (512 tokens)
  - Multi-task learning: learns stance and rumor prediction at the same time



Experimental Results on stance and rumor prediction

Stance Prediction

Method	Single Stance Type Evaluation				Overall Evaluation	
	Support-F_1	Deny-F_1	Query-F_1	Comment-F_1	Macro-F_1	Accuracy
SVM (Pamungkas et al., 2018)	0.410	0.000	0.580	**0.880**	0.470	0.795
BranchLSTM (Kochkina et al., 2018)	0.403	0.000	0.462	0.873	0.434	0.784
Temporal ATT (Veyseh et al., 2017)	-	-	-	-	0.482	**0.820**
Conversational-GCN (Wei et al., 2019)	0.311	0.194	**0.646**	0.847	0.499	0.751
Hierarchical Transformer (Ours)	**0.421**	**0.255**	0.520	0.841	**0.509**	0.763

Rumor Prediction

Setting	Method	SemEval-2017 Dataset		PHEME Dataset	
		Macro-F_1	Accuracy	Macro-F_1	Accuracy
Single-Task	BranchLSTM (Kochkina et al., 2018)	0.491	0.500	0.259	0.314
	TD-RvNN (Ma et al., 2018b)	0.509	0.536	0.264	0.341
	Hierarchical GCN-RNN (Wei et al., 2019)	0.540	0.536	0.317	0.356
	HiTPLAN (Khoo et al., 2020)	0.581	0.571	0.361	0.438
	Hierarchical Transformer (Ours)	**0.592**	**0.607**	**0.372**	**0.441**
Multi-Task	BranchLSTM+NileTMRG (Kochkina et al., 2018)	0.539	0.570	0.297	0.360
	MTL2 (Veracity+Stance) (Kochkina et al., 2018)	0.558	0.571	0.318	0.357
	Hierarchical PSV (Wei et al., 2019)	0.588	0.643	0.333	0.361
	MTL2-Hierarchical Transformer (Ours)	0.657	0.643	0.375	0.454
	Dual Hierarchical Transformer (Ours)	**0.680**	**0.678**	**0.396**	**0.466**

Outperforms previous work, including our own previous work
(denoted as HitPLAN in this table).

PHEME data annotation problem

Rumor that is fake (or inaccurate):

- In the Germanwings Flight 9525, 150 died, not 148.

Thread 1

"Reports: Crashed #Germanwings plane was carrying 148 people, including 142 passengers, two pilots and four flight attendants."

> “@SPIEGEL_English: Reports:Crashed #Germanwings plane. 148 people, including 142 passengers, 2 pilots and 4 flight attendants.” Schon wieder"

> "@SPIEGEL_English BREAKING - Germanwings plane crashes in France, up to **150 believed** dead\nhttp://t.co/HWyOPGobie"

Thread 2

"BREAKING:148 passengers were on board #GermanWings Airbus A320 which has crashed in D southern French Alps.May اللّهُ protect them.AME☼♥"

> "@AbedaDocrat Ameen"

These threads are all annotated as fake, but the one on the right has no denials in the comments.

Given these examples, the machine might learn that the topic (or “148 died”) is fake, but might not learn that **this is because there is a correction in the comments.**

Further reading

Ling Min Serena Khoo, Hai Leong Chieu, Zhong Qian, Jing Jiang:
Interpretable Rumor Detection in Microblogs by Attending to User Interactions. AAAI 2020: 8783-8790

Jianfei Yu, Jing Jiang, Ling Min Serena Khoo, Hai Leong Chieu, Rui Xia:
Coupled Hierarchical Transformer for Stance-Aware Rumor Verification in Social Media Conversations. EMNLP (1) 2020: 1392-1401

Xiaoying Ren, Jing Jiang, Ling Min Serena Khoo, Hai Leong Chieu:
Cross-Topic Rumor Detection using Topic-Mixtures. *EACL 2021: 1534-1538*

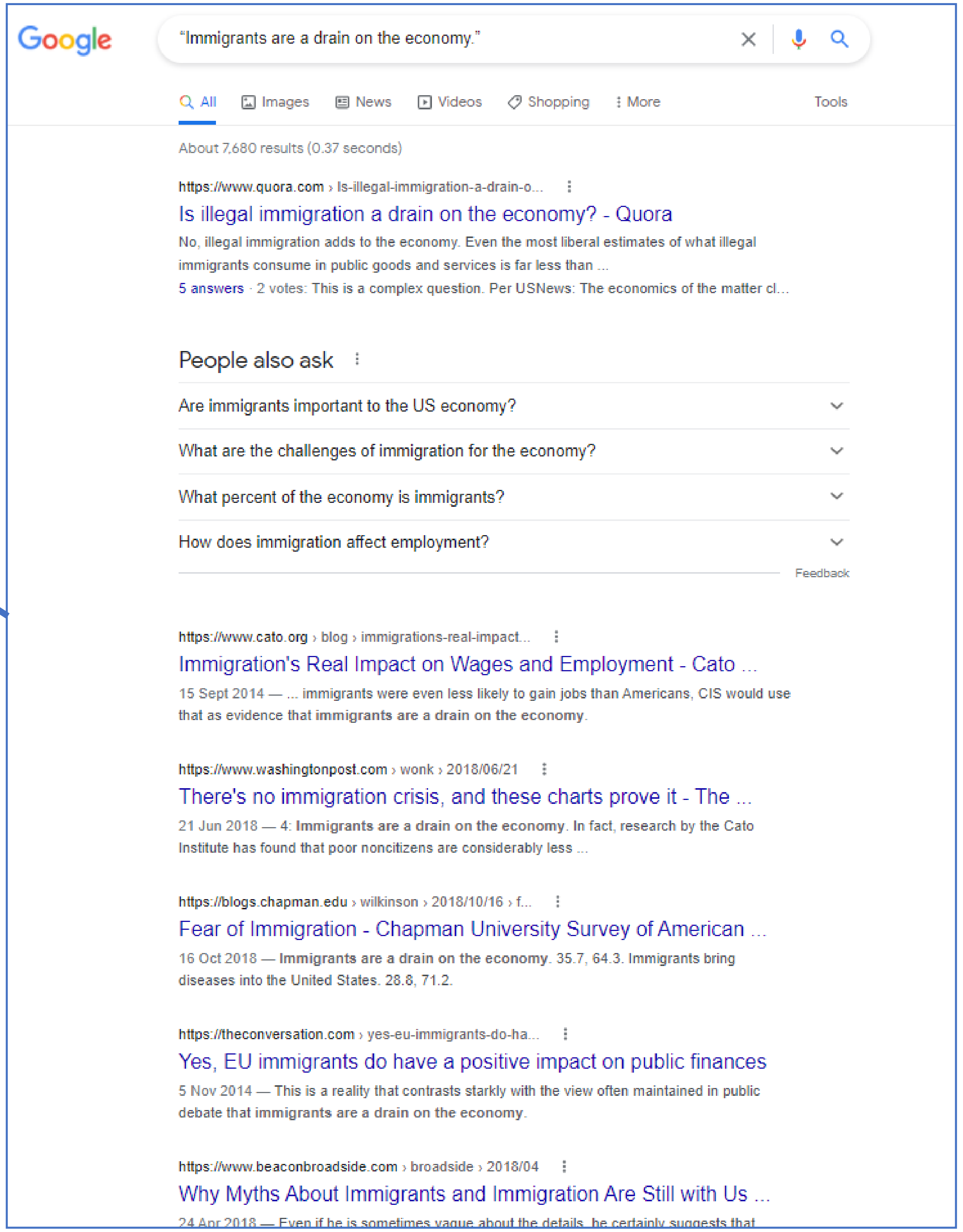
What is Fact Verification?

Input claim: *"Immigrants are a drain on the economy."*

Retrieve relevant articles from the Web (or Wikipedia)

Find supporting or refuting evidence, e.g., *Immigrants are a net gain to the economy, and several American cities…*

Predict (1) True, (2) Fake or (3) Not Enough Information.



The FEVER 2018 data set

The Fact Extraction and VERification (FEVER) Shared Task

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, Arpit Mittal

Abstract

We present the results of the first Fact Extraction and VERification (FEVER) Shared Task. The task challenged participants to classify whether human-written factoid claims could be SUPPORTED or REFUTED using evidence retrieved from Wikipedia. We received entries from 23 competing teams, 19 of which scored higher than the previously published baseline. The best performing system achieved a FEVER score of 64.21%. In this paper, we present the results of the shared task and a summary of the systems, highlighting commonalities and innovations among participating systems.

- **FEVER (synthetically created by a UK group)**
  - 185k claims manually **generated** by altering Wikipedia sentences and verified True or False
  - Annotated evidence that
    - supports real claims, or
    - refutes generated fake claims

FEVER Examples

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

Claim	Evidence	Label
Tim Roth is an English actor.	Timothy Simon Roth (born 14 May 1961) is an English actor and director.	SUPPORTS
Aristotle spent time in Athens.	At seventeen or eighteen years of age, he joined Plato's Academy in Athens and remained there until the age of thirty-seven (c. 347 BC).	SUPPORTS
Telemundo is a English-language television network.	Telemundo (telemundo) is an American Spanish-language terrestrial television network owned by Comcast through the NBCUniversal division NBCUniversal Telemundo Enterprises.	REFUTES
Magic Johnson did not play for the Lakers.	He played point guard for the Lakers for 13 seasons.	REFUTES

In FEVER, 83.2% of the claims require one sentence[1].

[1]Multi-hop fact checking of political claims, Ostrowski et al., 2020.

Data bias

Tal Schuster, Darsh J. Shah, **Yun Jie Serene Yeo**, Daniel Filizzola, Enrico Santus, Regina Barzilay:
Towards Debiasing Fact Verification Models. EMNLP/IJCNLP 2019
https://github.com/TalSchuster/FeverSymmetric

Claim-only classifiers perform competitively with top evidence aware models.
Top Fever team: 64.2%
Claims only classifier: 61.7%
Why?
Possible reasons:
- Embedding (e.g., GLOVE, BERT) contain world knowledge? Not the reason because:
  - Even without pre-trained embedding, claims-only classifier can achieve 54.1% (far above 33% random baseline)
- **Idiosyncrasies in the data?**

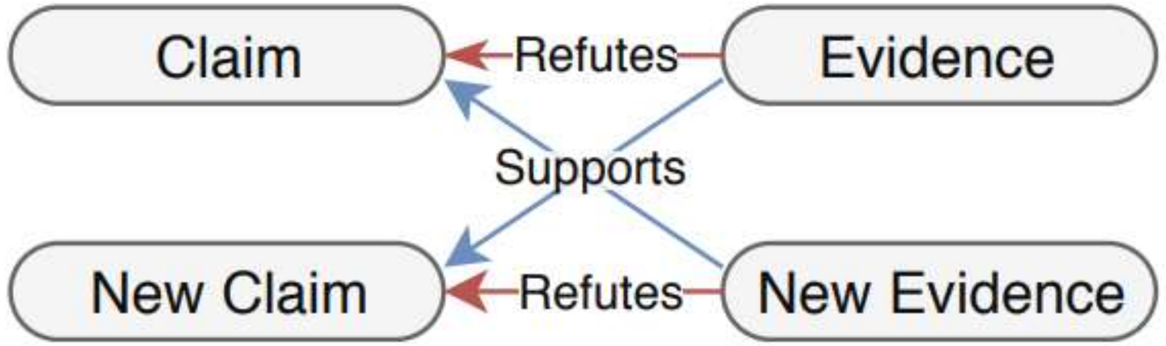
Idiosyncrasies in data construction

Bigrams that are most correlated with fake claims.

Bigram	Train LMI$\cdot 10^{-6}$	Train $p(l\|w)$	Development LMI$\cdot 10^{-6}$	Development $p(l\|w)$
did not	1478	0.83	1038	0.90
yet to	721	0.90	743	0.96
does not	680	0.78	243	0.68
refused to	638	0.87	679	0.97
failed to	613	0.88	220	0.96
only ever	526	0.86	350	0.82
incapable being	511	0.89	732	0.96
to be	438	0.50	454	0.65
unable to	369	0.88	346	0.95
not have	352	0.78	211	0.92

"Most of the n-grams **express strong negations**, which, in hindsight, is not surprising as these idiosyncrasies are **induced by the way annotators altered the original claims to generate fake claims**."

Debiasing the data



For an original claim-evidence pair, we manually generate a synthetic pair that holds the same relation (i.e. SUPPORTS or REFUTES) while expressing a fact that contradicts the original sentences.

Combining the ORIGINAL and GENERATED pairs, this new test set completely eliminates the ability of models to rely on cues from claims.

Source	Claim	Evidence	Label
ORIGINAL	Tim Roth is an English actor.	Timothy Simon Roth (born 14 May 1961) is an English actor and director.	SUPPORTS
GENERATED	Tim Roth is an American actor.	Timothy Simon Roth (born 14 May 1961) is an American actor and director.	SUPPORTS
ORIGINAL	Magic Johnson did not play for the Lakers.	He played point guard for the Lakers for 13 seasons.	REFUTES
GENERATED	Magic Johnson played for the Lakers.	He played for the Giants and no other team.	REFUTES

Towards unbiased training

- Objective:
  - Reweigh training samples to minimize bias on n-grams

- Formulation
  - Assign additional positive weight $\alpha^{(i)}$ to each training sample
  - How do we set $\alpha^{(i)}$?
    - Define bias as

$$b_j^c = \frac{\sum_{i=1}^{n} I_{[w_j^{(i)}]}(1+\alpha^{(i)}) I_{[y^{(i)}=c]}}{\sum_{i=1}^{n} I_{[w_j^{(i)}]}(1+\alpha^{(i)})}, \qquad (2)$$

    - Set α to minimize max bias (α is L2-regularized):

$$\min\left(\sum_{j=1}^{|V|} \max_c(b_j^c) + \lambda \|\vec{\alpha}\|_2\right). \qquad (3)$$

Results on the SYMMETRIC test set

Model	FEVER DEV		GENERATED	
	BASE	R.W	BASE	R.W
NSMN	81.8	-	58.7	-
ESIM	80.8	76.0	55.9	59.3
BERT	**86.2**	84.6	58.3	**61.6**

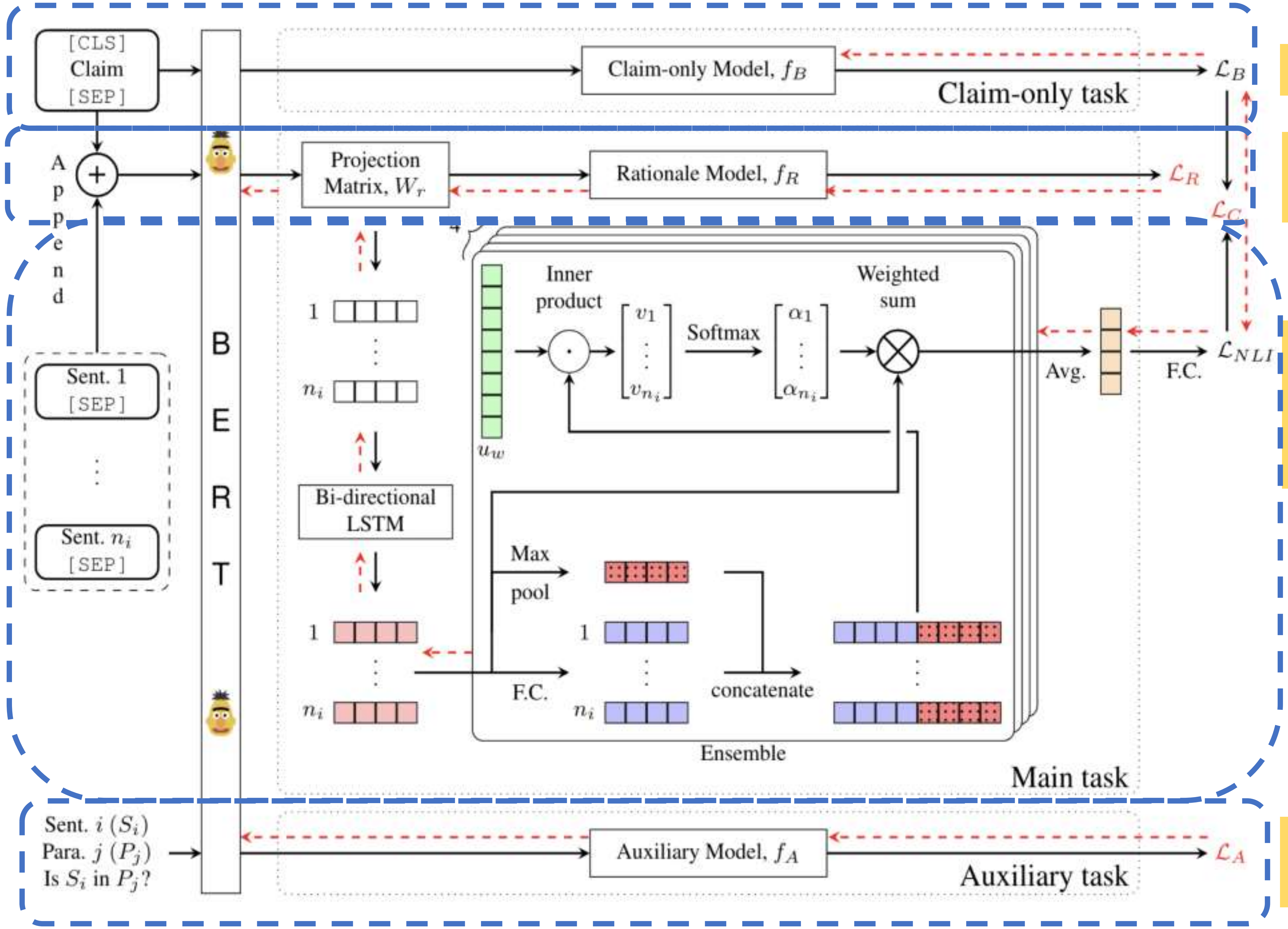
Table 3: Classifiers' accuracy on the SUPPORTS and REFUTES cases from the FEVER DEV set and on the GENERATED pairs for the SYMMETRIC TEST SET in the setting of without (BASE) and with (R.W) re-weight.

NSMN, the leading system in FEVER, the NSMN achieves only 58.7% accuracy on the symmetric test set compared to 81.8% on the original dataset.

Bigram	**Train**		**Development**	
	LMI$\cdot 10^{-6}$	$p(l\|w)$	**LMI**$\cdot 10^{-6}$	$p(l\|w)$
did not	1478	0.83	1038	0.90
yet to	721	0.90	743	0.96
does not	680	0.78	243	0.68
refused to	638	0.87	679	0.97
failed to	613	0.88	220	0.96
only ever	526	0.86	350	0.82
incapable being	511	0.89	732	0.96
to be	438	0.50	454	0.65
unable to	369	0.88	346	0.95
not have	352	0.78	211	0.92

Bigram	R.W **LMI**$\cdot 10^{-6}$	R.W $p(l\|w)$
did not	144	0.35
yet to	30	0.33
does not	67	0.35
refused to	55	0.35
failed to	31	0.33
only ever	9	0.31
incapable being	32	0.33
to be	8	0.30
unable to	10	0.32
not have	41	0.35

BERT-LSTM Model (unpublished)



To model biases

To learn how to extract evidence

To learn attention weights between snippets to aggregate and predict a final label

To improve learning of representations

[Unpublished work, Serene Yeo et al.]

Datasets

- **Training Data (synthetically created)**
  - **FEVER**
    - 185k claims **generated** by altering Wikipedia sentences and verified True or False
    - Annotated evidence that Supports and Refutes the claim
    - Used for **pre-training (to learn how to extract evidence)**
- **Test Data (real fake news on the web)**
  - **Politifact**
    - 3.6k claims made by politicians in US with 30k articles retrieved from 336 sources
    - Six fine-grained labels remapped into True and False
    - 10% hold out for evaluation, remaining do 5-fold CV
    - Used for fine-tuning and testing
  - **Snopes**
    - 4.3k claims made by general public with 29k articles retrieved from 336 sources
    - True or False labels
    - 10% hold out for evaluation, remaining do 5-fold CV
    - Used for fine-tuning and testing
  - **LIAR-PLUS**
    - 12.8k statements from Politifact with human-written justifications
    - 1.3k each for dev and test
    - Used for fine-tuning and testing

Model performance

- Performance on Politifact & Snopes:

Dataset	Configuration	*True* Claims Accuracy (%)	*False* Claims Accuracy (%)	Macro F1-Score
Snopes	DeClarE (Full)	79.0	78.3	0.82
	HAN	66.5	86.0	0.76
	Ours	**95.5**	**98.3**	**0.97**
Politifact	DeClarE (Full)	67.3	69.6	0.68
	Ours	**95.4**	**92.8**	**0.94**

State-of-the-art published results

State-of-the-art published results

- Performance on LIAR-PLUS:

Dataset	Configuration	*Validation* Accuracy (%)	*Test* Accuracy (%)
LIAR-PLUS	biLSTM	70.0	68.0
	Ours	**78.9**	**78.5**

State-of-the-art published results

Examples on local data

Queried claim: Seven countries have since banned travel to Singapore, citing lack of confidence in the Singapore government's public health measures
Overall prediction: FALSE (probability of 0.996)

Snippet rank	URL	Snippet description	Importance
0	https://www.gov.sg/article/factually-clarifications-on-falsehoods-posted-by-str-on-covid-19-situation	Seven countries have since banned travel to Singapore, citing lack of confidence in the Singapore government's public health measures; **The above are entirely false**, for the following reasons: First, as of 12:00 pm on 13 Feb 2020, the Ministry of Health ("MOH") has established through epidemiological investigation and contact tracing that 51 ..	0.7985
4	https://statestimesreview.com/2020/02/13/minister-josephine-teo-600-china-workers-have-entered-singapore-more-are-coming/	The Singapore government was unable to trace the source of any of the infected. 7 countries including China and South Korea has since banned travel to Singapore, citing lack of confidence in the Singapore government's public health measures. The Singapore government is also the only one telling the public not to wear a mask.	0.0653
3	https://www.intellasia.net/mci-slaps-declared-online-location-tag-on-states-times-review-page-762780	Minister for Communications and Information S Iswaran has on Saturday (15 February) declared the States Times Review (STR) Facebook page a Declared Online Location (DOL) under the Protection from Online Falsehoods and Manipulation Act (POFMA).	0.0486
2	https://en.wikipedia.org/wiki/2019%E2%80%9320_Wuhan_coronavirus_outbreak_by_country_and_territory	According to public health officials, Vilnius Airport had a medical exercise in December and is ready to handle infected passengers and contain the spread of the virus. Malta. Maltese local authorities have taken preventive measures, and advised the public and health workers to uphold sanitary regulation to not spread illnesses.	0.0459
1	https://www.reddit.com/r/singapore/comments/f3ornf/corrections_and_clarifications_regarding/	Health authorities in other countries such as the US and Australia have also expressly advised that they do not recommend that masks be worn by people who are well. As a good hygiene practice, people who are unwell and who have respiratory symptoms should wear a mask so that they minimise the risk of them infecting others.	0.0213
…			

Examples on local data

Queried claim: Woodlands MRT was closed for disinfection due to a suspected case of the 2019 novel coronavirus infection.
Prediction: FALSE (probability of 0.999)

Snippet rank	URL	Snippet description	Importance
0	https://www.gov.sg/article/factually-clarifications-on-falsehoods-on-woodlands-mrt-closure	**There was a false statement** contained in several Facebook posts on the 2019 novel coronavirus infection. Falsehoods On 28 Jan 2020, there were posts by several Facebook users claiming that Woodlands MRT was closed for disinfection due to a suspected case of the 2019 novel coronavirus infection.	**0.503**
1	https://www.gov.sg/article/covid-19-clarifications	Woodlands MRT was closed for disinfection - 28 Jan 2020 . Several Facebook posts claimed that Woodlands MRT was closed for disinfection due to a suspected case of the Wuhan coronavirus infection. The posts also urged members of the public not to go to Woodlands MRT. **This is not true**. Woodlands MRT was not closed on 28 Jan 2020; it was fully ...	0.234
2	https://factcheck.afp.com/china-coronavirus-singapore-denies-it-closed-subway-station-after-novel-coronavirus-discovery	https://factcheck.afp.com/china-coronavirus-singapore-denies-it-closed-subway-station-after-novel-coronavirus-discovery || A Facebook post claims Singapore closed a subway station in January 2020 after discovering a case of novel coronavirus**. The claim is false;** Singapore's Ministry of Health and Ministry of Transport **denied** that any part of its mass rapid transit (MRT) network had been shut down for disinfection.	0.182
8	https://www.facebook.com/ZainalBinSapari/posts	Zainal Bin Sapari. 8.2K likes. Father, Husband, Unionist, Teacher and a Servant Leader. ... MOT is aware of rumours circulating online that Woodlands MRT was closed for disinfection due to a suspected case of the Wuhan coronavirus infection. ... **False claims** that Woodlands MRT closed due to Wuhan coronavirus infection. gov.sg.	0.068
6	https://blackdotresearch.sg/wuhan-virus-singapore-factcheck/	In relation to Woodlands, there were also several posts on social media on 28 January urging the public not to go to Woodlands MRT station as it was closed for disinfection due to a suspected case. **MOH has come forward to address this, stating that Woodlands MRT station wasn't closed** on 28 January and was fully operation.	0.004
...			

Examples on local data

Queried claim: Two LRT trains collide between Sengkang and Renjong stations
Prediction: TRUE (probability of 0.764)

Snippet rank	URL	Snippet description	Importance
4	https://landtransportguru.net/sengkang-station/	Sengkang LRT station is overground with two platforms in an island platform arrangement, utilized alternatively by East and West loop services. At each platform, East and West LRT services are staggered one after the other and operate throughout the day. From Platform 1, LRT Routes A and D head out to the Outer West Loop and Inner East Loop via Renjong and Ranggung respectively.	**0.172**
9	https://www.sgtrains.com/network-sklrt.html	The two-car system was tested during off-peak hours from 21 December 2015, and the modification was completed on 5 January 2016, with eight two-car trains on the Sengkang LRT during peak hours. From 1 April 2017, two-car trains were also deployed on the west loop throughout the day on weekends and public holidays.	0.152
7	https://www.sgcarmart.com/news/article.php?AID=17213	According to citizen journalism site Stomp, **two LRT trains reportedly collided** on the Sengkang Line at 7:08pm last night.	0.143
8	https://mustsharenews.com/lrt-incident-sengkang/	On Monday, a passenger by the name of "Hong" was reported by Stomp as saying that the Light Rail Transit (LRT) **train she was on collided with another train in front of it**. Source. No Updates. The incident occurred between the Sengkang and Renjong LRT stations at 7.08pm, and was reported by Stomp in an article that has since been taken down.	0.136
3	https://newscollection.net/asia-pacific/singapore/two-lrt-trains-collide-between-sengkang-and-renjong-stations/	Stomp contributor Hong was **on board a Light Rail Transit (LRT) train that collided with the train in front of it** at 7.08pm today (July 3) on the Sengkang LRT Line. The Stomp contributor said	0.124
...			

- Bulk of the retrieved snippets had information supporting the claim, resulting in prediction being True.

Examples on local data

- Only snippet 5 was refuting the claim and if we were able to use it alone,

Queried claim: Two LRT trains collide between Sengkang and Renjong stations
Prediction: FALSE (probability of 0.978)

Snippet rank	URL	Snippet description (# … # denotes title)	Importance
5	https://goodyfeed.com/lrt-trains-did-not-collide-but-merely-stalled-according-to-lta-sbs/	# **LRT Trains Did Not Collide, But Merely Stalled**, According … # Yesterday, it was reported in Stomp that two train-cars on the Sengkang LRT line had "collided". According to Stomper Hong, it occurred around 7:00 p.m. yesterday (3 July 2017). That article has since been removed from the Stomp website, but the Straits Times has updated its report.. A train-car had stopped between Sengkang Town Centre and Renjong stations.	**1.0**
…			

Publications for more information

Tal Schuster, Darsh J. Shah, **Yun Jie Serene Yeo**, Daniel Filizzola, Enrico Santus, Regina Barzilay:
Towards Debiasing Fact Verification Models. EMNLP/IJCNLP 2019

Darsh J. Shah, Tal Schuster, Regina Barzilay:
Automatic Fact-Guided Sentence Modification. *AAAI 2020*

Tal Schuster, Adam Fisch, Regina Barzilay:
Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence.
NAACL 2021

Internship

If you are a Singapore citizen and are interested in joining DSO, a great way to start would be to do an internship with us!

A few available NLP projects include

- Fact checking for fake news detection
- Sentiment analysis and opinion summarization
- Multi-lingual NLP
- Style Transfer

Other machine learning projects

- Chemical toxicity classification

For more info on DSO: https://www.dso.org.sg/join-us/career-seekers