COS 598C Advanced Topics in Computer Science: Deep Learning for Natural Language Processing

Introduction

Winter 2020

Logistics

- Instructor: Danqi Chen
- Where: Engineering Quad A-Wing A224
- When: Tuesday and Thursday 1:30-2:50pm
- Office hours: Tuesday 3-4pm (in addition to lecture feedback)
    - By appointment: https://calendly.com/danqic/cos598c

- Sign up for Piazza today!

Course goals

- This course is to prepare you for performing **cutting-edge research** in natural language processing
    - It is a graduate seminar — I'd assume you already know the basics of NLP (COS484 or equivalent). Be self-motivated.
    - Learn about the most influential ideas in each sub-area of NLP and the major problems we are facing *today*

- Practice your **research skills**
    - Comfortable with reading research papers and conducting literature survey
    - Improving both oral and written presentations
    - Providing constructive peer feedback

Course structure

- I already chose 20 NLP topics and each topic has (Google doc is here)
    - **Required**: 2 papers to read. S*uggestions are still welcome*.
    - Background reading
    - Recommended reading
- You are required to give lectures and lead the discussion of these papers! I will cover the first two topics: word embeddings and contextualized word embeddings

- Two students will sign up for one lecture — You are expected to give 1-2 lectures throughout the semester (depending on the enrollment)
    - It is your job to decide how to coordinate with your partner

- Come meet with me with your draft slides before your lecture
    - Monday 5:00-5:30pm (for Thursday lectures)
    - Thursday 5:00-5:30pm (for Tuesday lectures)

Course structure

- Everyone else is expected to read the *2* papers beforehand
    - I will send out a Google form ~2 days before the lecture and you will expect to write a small paragraph/a few sentences to answer the questions
    - Due by 12pm of the lecture day

- For each lecture, we will also ask ~3 students to sign up for providing feedback on the lectures
    - Feedback emailed to the presenters and cc'ing me, within a day of the presentation
    - Comments on clarity, structure, completeness, slides ..
    - Offer constructive criticism but also suggestions

Course structure

- There is a final project at the end
  - **Option 1**: Pick a topic and write a survey paper (at least 5-10 papers)
  - **Option 2**: Pick an NLP task and 1-3 papers with code on the task. Run the code and analyze/understand the limitations of existing systems. *paperswithcode.com*
  - **Option 3**: Come up with your own project. You can work as a team of 1 or 2.

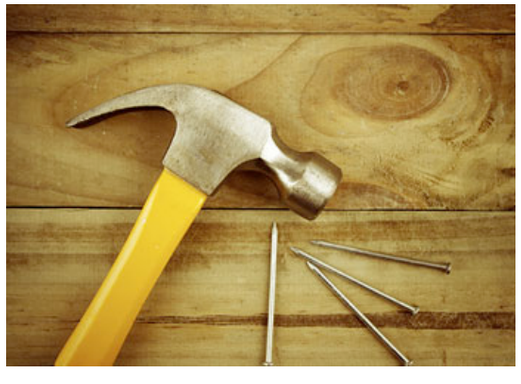
- Proposal deadline: March 10th, 11:59pm

What is this course about

- **Deep Learning for NLP**
  - No symbolic and statistical NLP methods **[COS 484]**
  - All the papers we chose were written in 2013-2020. Most papers were published in ACL/EMNLP/NAACL and some were published in ICML/NeurIPS/ICLR.

- The course is mostly **problem-driven**: make sure you have all the hammers you need (next slide)!

- Focused on **English** NLP
  - Many techniques can be directly transferred to other languages
  - We don't cover multi-lingual or cross-lingual or low-resource NLP research, and also *machine translation* **[COS 401]**



#Benderrule

CONFERENCE	% ENGLISH	NEXT MOST COMMON LANGUAGE(S)	% NEXT MOST COMMON LANGUAGE(S)	SOURCE
ACL 2004	87	Chinese	9	Mielke 2016 [6]
ACL 2008	63	German, Chinese	4	Bender 2009 [7]
ACL 2008	87	Chinese	16	Mielke 2016 [8]
EACL 2009	55	German	7	Bender 2011 [9]
ACL 2012	86	Chinese	23	Mielke 2016 [10]
ACL 2015	75	Chinese	5	Munro 2015 [11]
ACL 2016	90	Chinese	13	Mielke 2016 [12]

https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/

What is this course NOT about

- Neural networks basics **[COS 485]**
  - ML Basics: supervised learning, unsupervised learning, reinforcement learning
  - Feedforward NNs, ConvNet, RNNs (LSTMs/GRUs), Transformers…
  - Optimization
  - Sequence-to-sequence models and attention

- Hands-on experience: PyTorch, TensorFlow etc.

I like your feedback

- Feedback form: https://forms.gle/yrWbVqgjysLnDEr58

Sign up for lectures

- I will send out a form by tonight (Feb 4th):
    - Your priority of topics
    - Your blackout dates

- Complete it by Thursday noon and I will send out the schedule by the weekend. I will give the first team a bonus :)

Grading structure

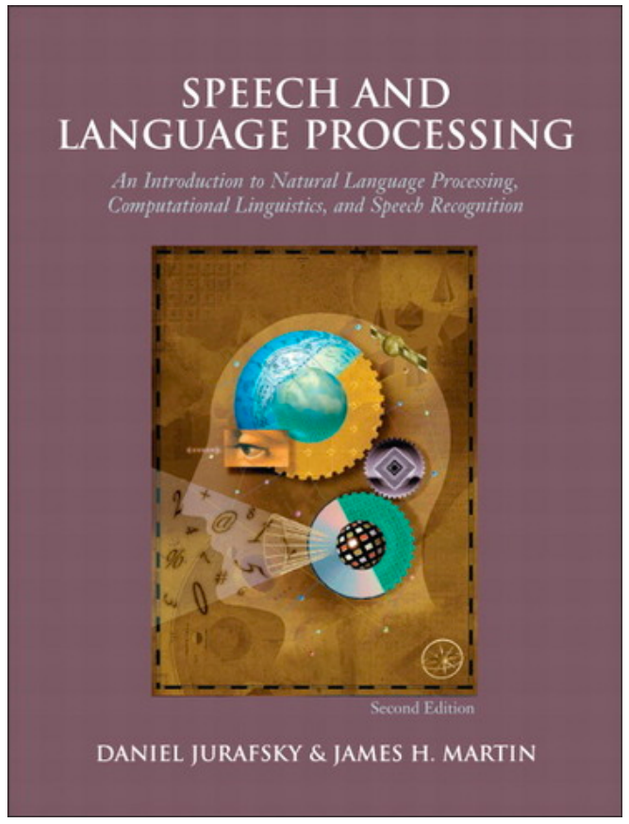
- Participation: 20%
    - Pre-class questionnaire 15%
    - In-class participation 5%
- Presentations: 45%
    - Lectures 40%
    - Peer feedback 5%
- Final paper: 35%

Textbook

Speech and Language Processing (3rd ed. draft)

Dan Jurafsky and James H. Martin

Draft chapters in progress, October 16, 2019

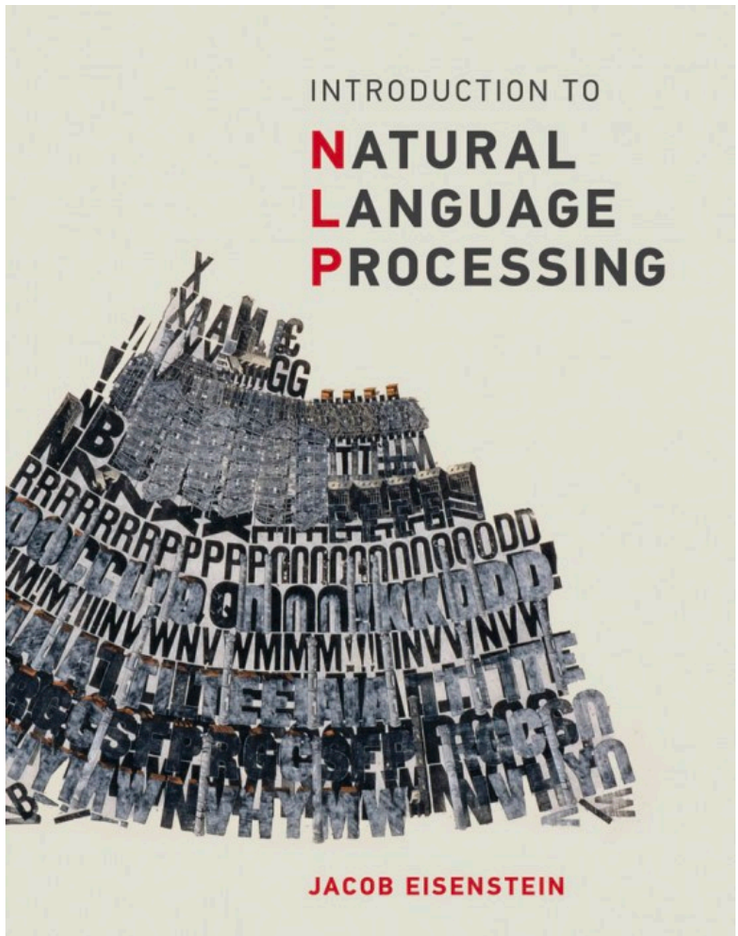


https://web.stanford.edu/~jurafsky/slp3/

Textbook

Introduction to Natural Language Processing

By Jacob Eisenstein



https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf

How to read papers

- Read the papers in context
    - All the papers are built on top of other papers!

- Grasp the key ideas
    - What is the biggest contribution of this paper?
    - Why is this paper important?

- Pay attention to the details (both methodology and experiments)

How to present papers

- Motivation & problem formulation
- Data and evaluation
    - Look at data if possible!
- Related work
- Method
- Experiments
- Analysis

Deep Learning for NLP

“The general approach to building Deep Learning systems is compelling and powerful: The researcher defines a **model architecture** and **a top-level loss function** and then both the **parameters** and the **representations** of the model self-organize so as to minimize this loss, in and **end-to-end** learning framework.”

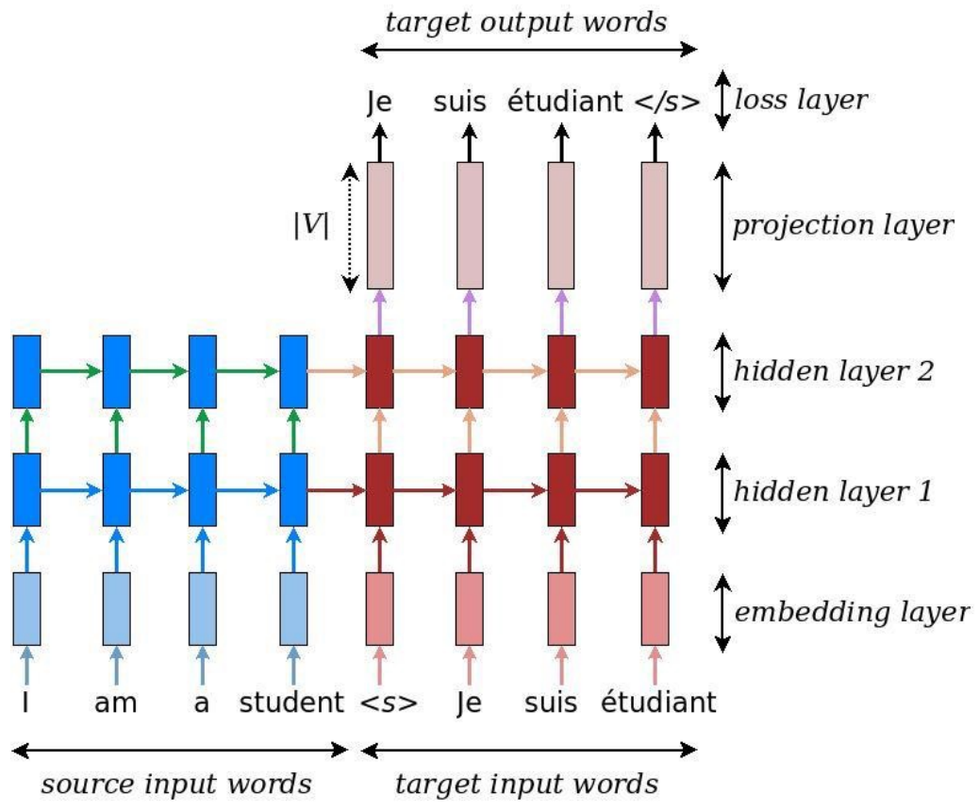
—— Chris Manning (2015)

Deep Learning for NLP

Distributed representations

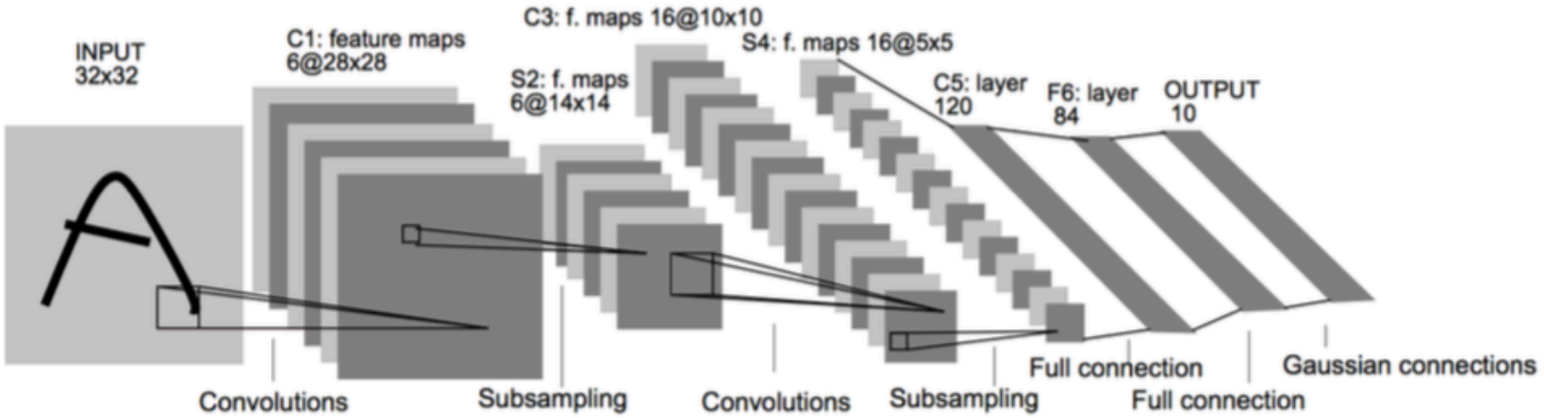
Word Vectors

End-to-end learning

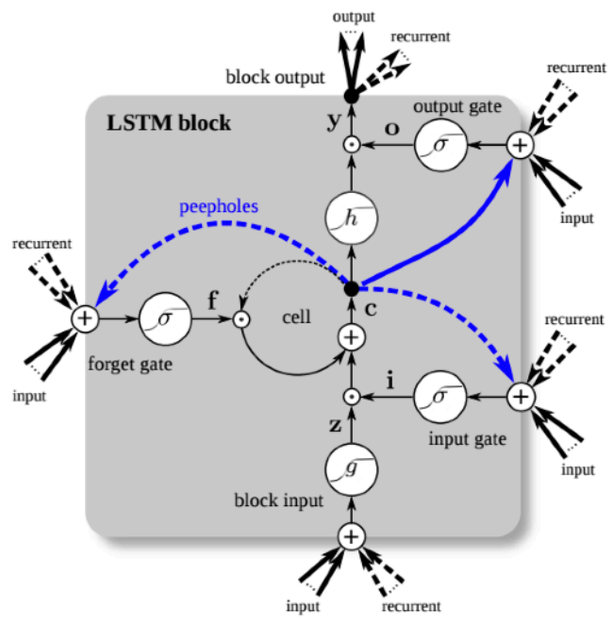


NN "dark ages"

- Neural network algorithms date from the 80s
- ConvNets: applied to MNIST by LeCun in 1998



- Long Short-term Memory Networks (LSTMs): Hochreiter and Schmidhuber 1997



- Henderson 2003: neural shift-reduce parser, not SOTA

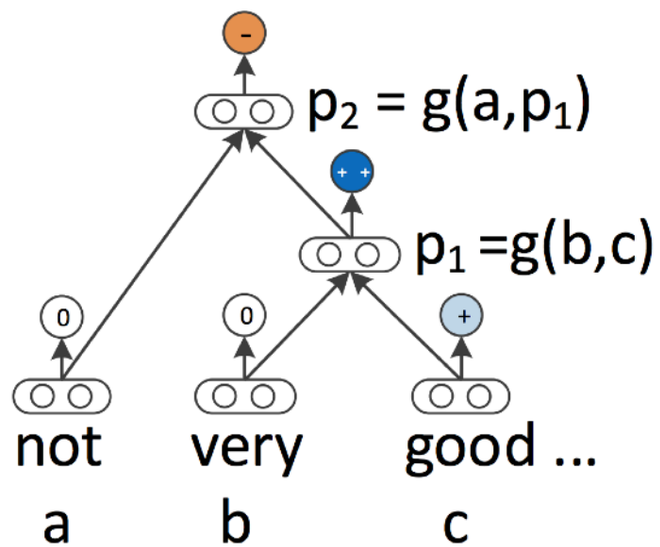
Credits: Greg Durrett

2008-2013: A glimmer of light

- Collobert and Weston 2011: "**NLP (almost) from Scratch**"
  - Feedforward NNs can replace "feature engineering"
  - 2008 version was marred by bad experiments, claimed SOTA bu wasn't, 2011 version tied SOTA

- Krizhevskey et al, 2012: AlexNet for ImageNet Classification

- Socher 2011-2014: tree-structured RNNs working okay



Credits: Greg Durrett

2014: Stuff starts working

- Kim (2014) + Kalchbrenner et al, 2014: sentence classification
  - ConvNets work for NLP!

- Sutskever et al, 2014: sequence-to-sequence for neural MT
  - LSTMs work for NLP!

- Chen and Manning 2014: dependency parsing
  - Even feedforward networks work well for NLP!

- 2015: explosion of neural networks for everything under the sun

Credits: Greg Durrett

Why didn't they work before?

- **Datasets too small**: for MT, not really better until you have 1M+ parallel sentences (and really need a lot more)

- **Optimization not well understood**: good initialization, per-feature scaling + momentum (Adagrad/Adam) work best out-of-the-box
    - Regularization: dropout is pretty helpful
    - Computers not big enough: can't run for enough iterations

- Inputs: need **word embeddings** to represent continuous semantics

Credits: Greg Durrett

The "Promise"

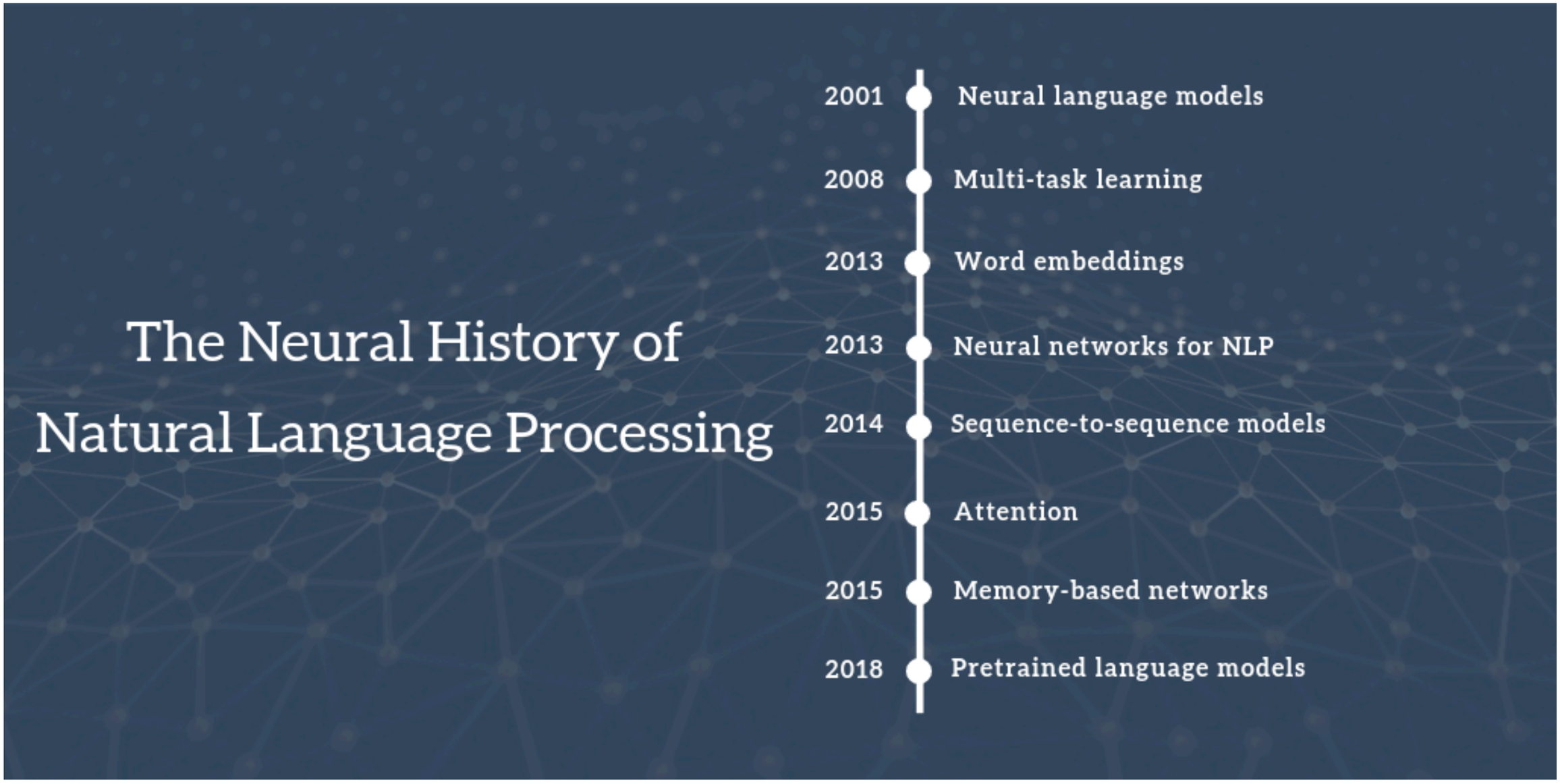
- Most NLP works in the past focused on human-designed representations and input features

Var	Definition	Value in Fig. 5.2
x_1	count(positive lexicon) $\in$ doc)	3
x_2	count(negative lexicon) $\in$ doc)	2
x_3	$\begin{cases} 1 & \text{if "no"} \in \text{doc} \\ 0 & \text{otherwise} \end{cases}$	1
x_4	count(1st and 2nd pronouns $\in$ doc)	3
x_5	$\begin{cases} 1 & \text{if "!"} \in \text{doc} \\ 0 & \text{otherwise} \end{cases}$	0
x_6	log(word count of doc)	$\ln(64) = 4.15$

- **Representation learning** attempts to automatically learn good features and representations

- **Deep learning** attempts to learn multiple levels of representation on increasing complexity/abstraction

A Neural History of NLP



https://ruder.io/a-review-of-the-recent-history-of-nlp/

Topics of Interest

NLP problems at different levels:

- **Linguistic levels**: (speech), words, syntax, semantics
- **Intermediate tasks/tools**: parts-of-speech, entities, parsing, coreference
- **Full applications**: sentiment analysis, question answering, dialogue, text summarization, machine translation

Prominent problems in existing NLP systems:

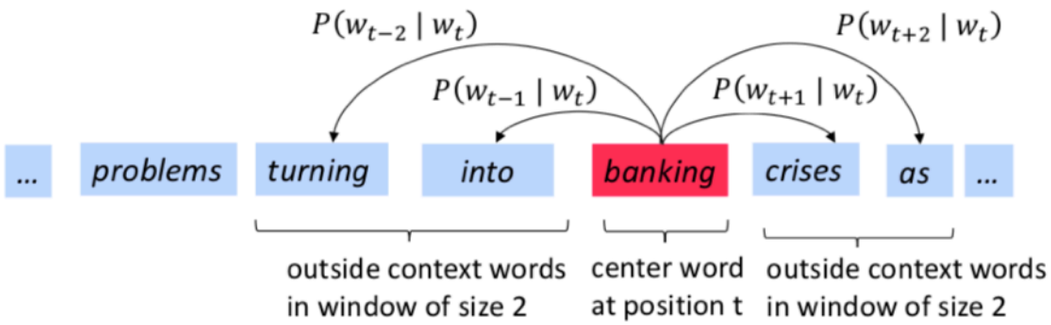
- Bias in language data
- Annotation artifacts
- Interpretability
- Adversarial examples
- General linguistic intelligence

Lecture 2: word embeddings

Represent words using continuous vectors!

(Mikolov et al, 2013)

word2vec



Word	Cosine distance
norway	0.760124
denmark	0.715460
finland	0.620022
switzerland	0.588132
belgium	0.585835
netherlands	0.574631
iceland	0.562368
estonia	0.547621
slovenia	0.531408

(Baroni et al, 2014)

***Don't count, predict!* A systematic comparison of context-counting vs. context-predicting semantic vectors**

Marco Baroni and **Georgiana Dinu** and **Germán Kruszewski**

(Levy et al, 2015)

Improving Distributional Similarity with Lessons Learned from Word Embeddings

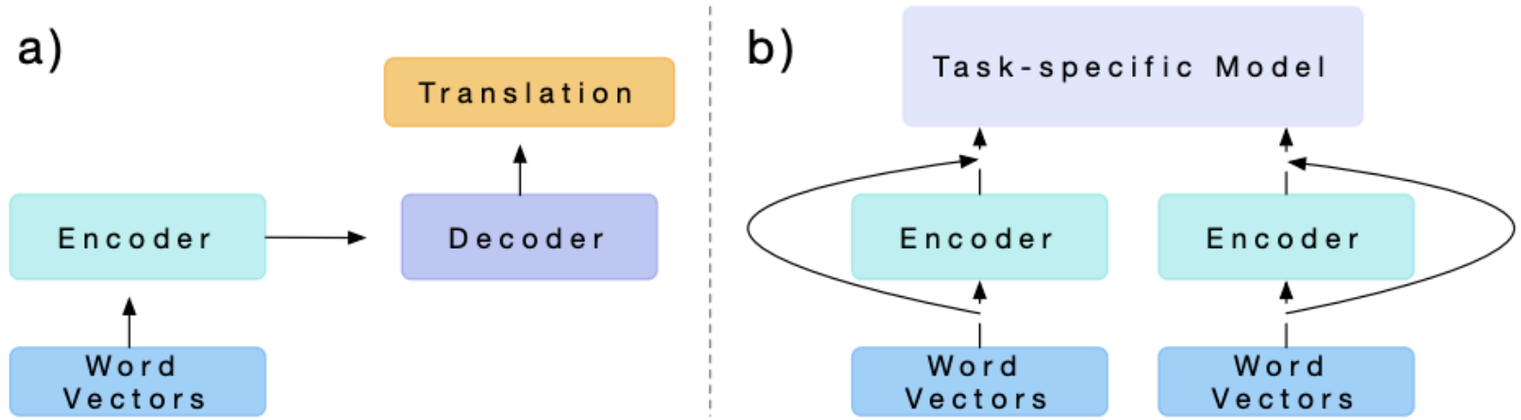
Omer Levy **Yoav Goldberg** **Ido Dagan**

Lecture 3: contextualized word embeddings

Represent words based on their context!

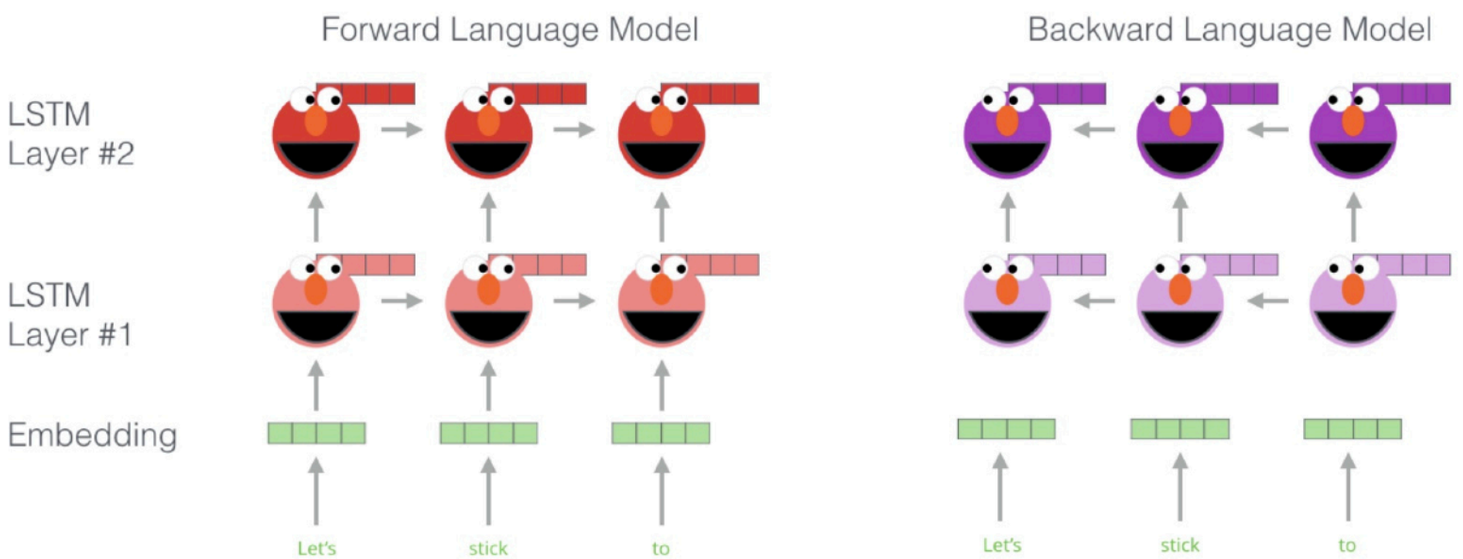
(McCann et al, 2017)

CoVe



(Peters et al, 2018)

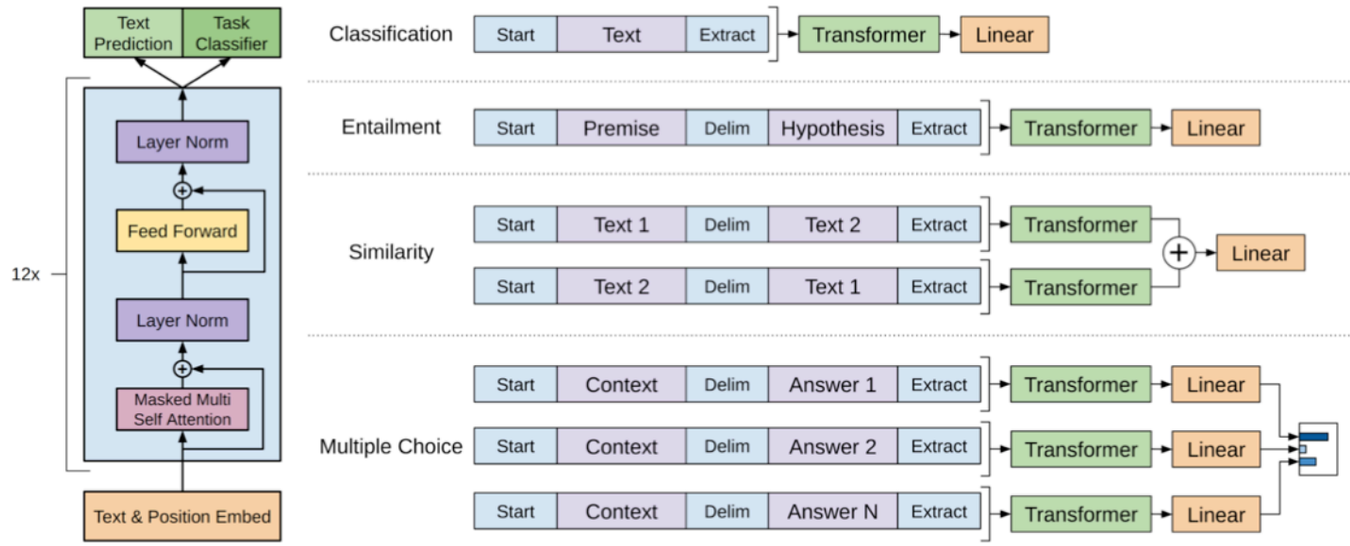
ELMo



Lecture 4: Pre-training and fine-tuning

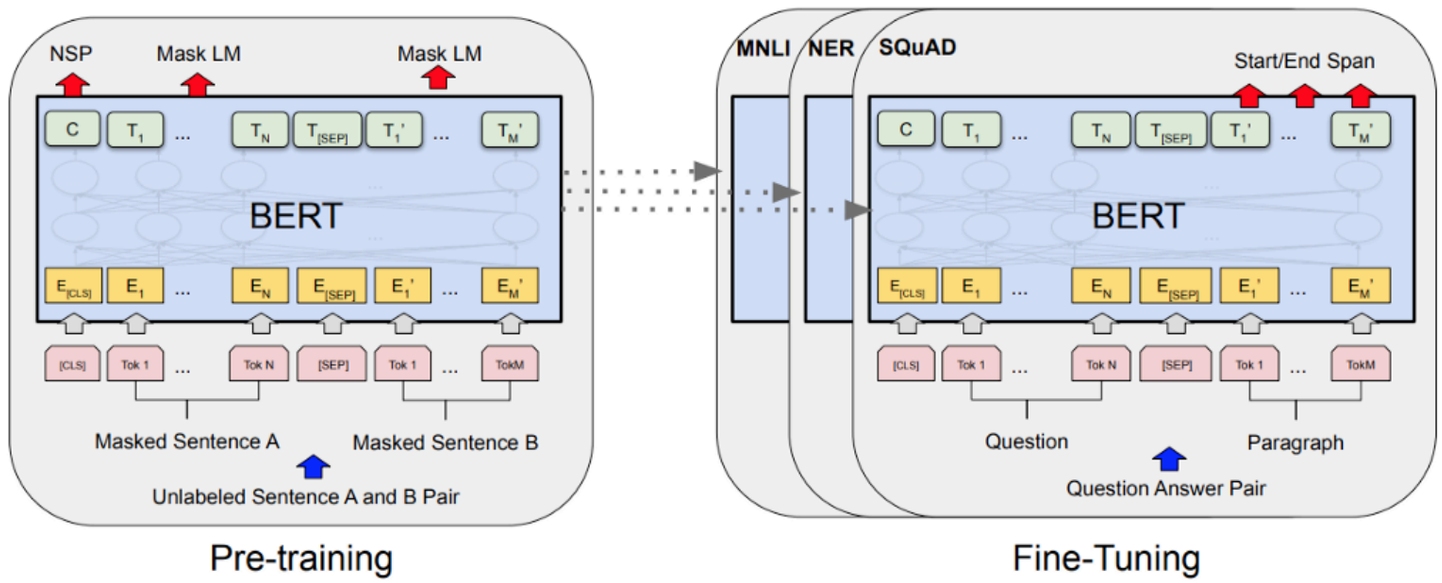
(Radford et al, 2018)

OpenAI GPT



(Devlin et al, 2019)

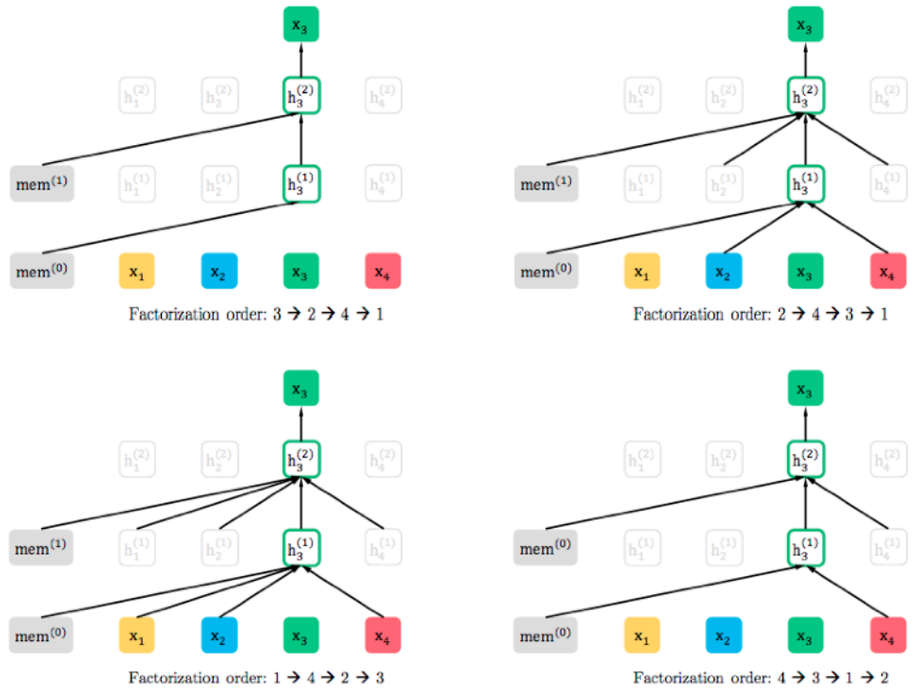
BERT



Lecture 5: Pre-training and fine-tuning (cont'd)

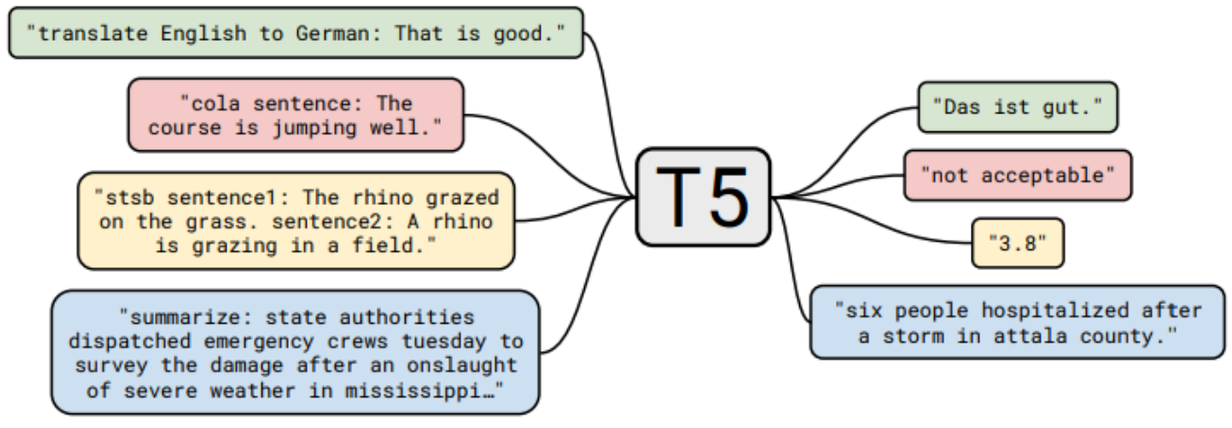
(Yang et al, 2019)

XLNet



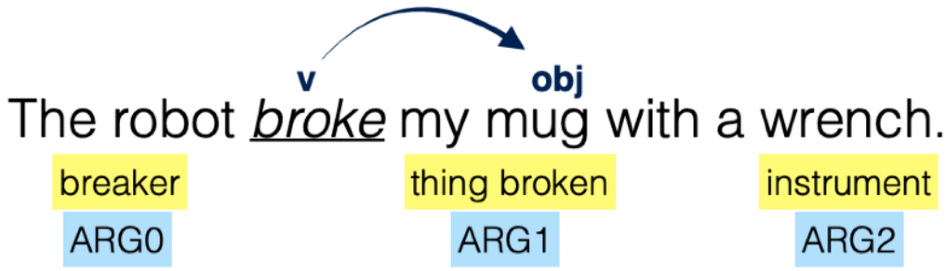
(Raffel et al, 2019)

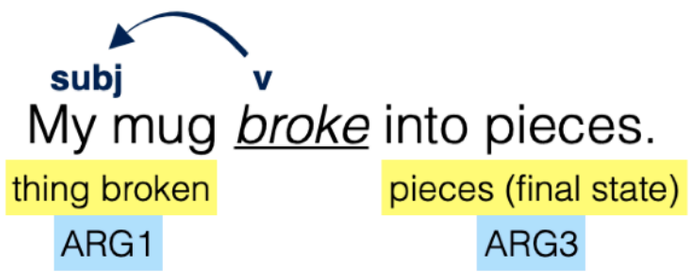
T5



Lecture 6: semantic role labeling

Who did what to whom

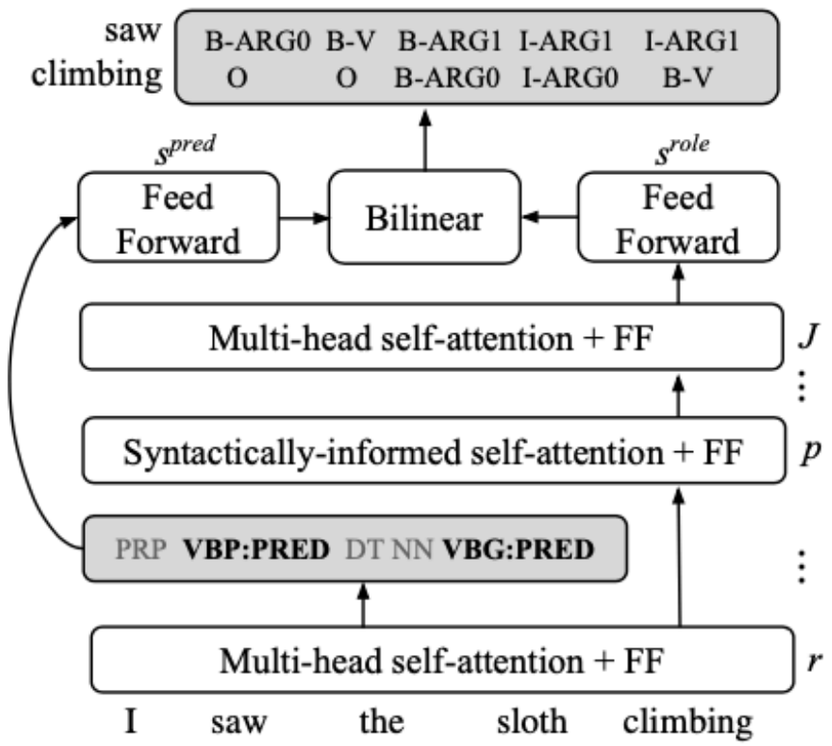




Frame: *break.01*

role	description
ARG0	breaker
ARG1	thing broken
ARG2	instrument
ARG3	pieces
…	…

(He et al, 2017)

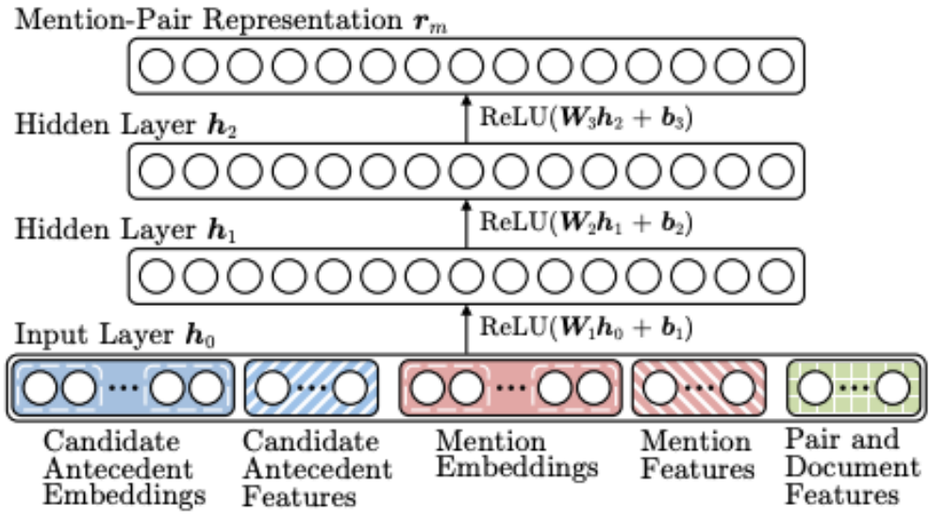


(Strubell et al, 2018)

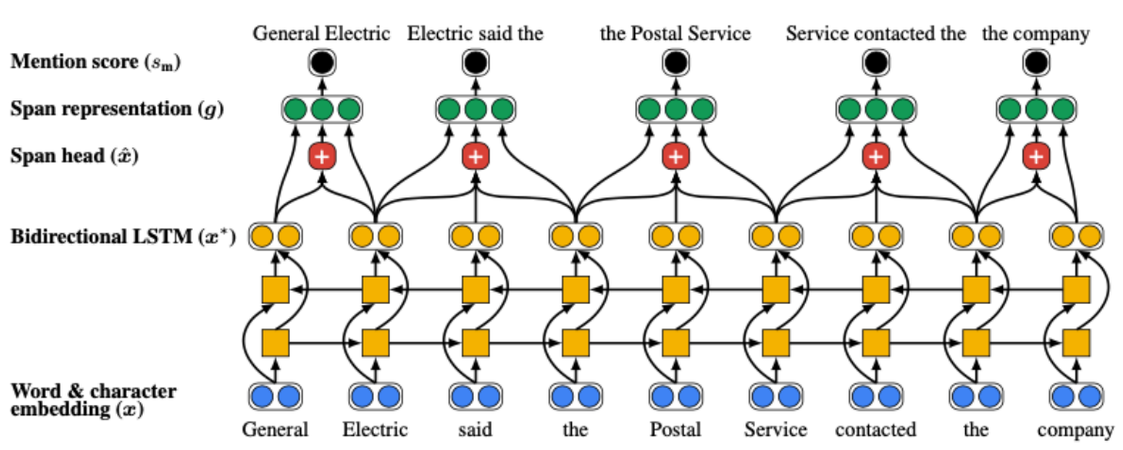
Lecture 7: coreference resolution

Cluster mentions in text which refer to the same real-world entities

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affairs experience as a former First Lady.



(Clark and Manning, 2016)

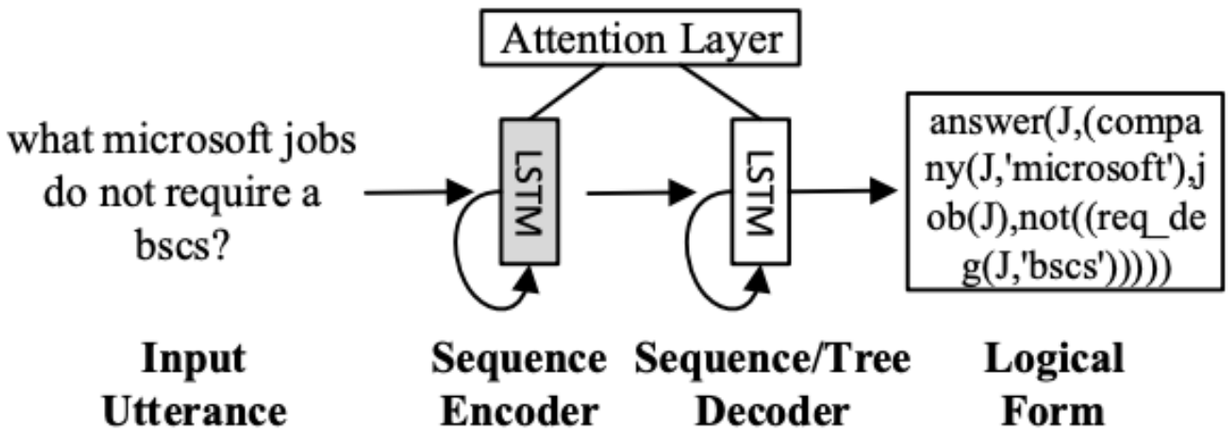


(Lee et al, 2017)

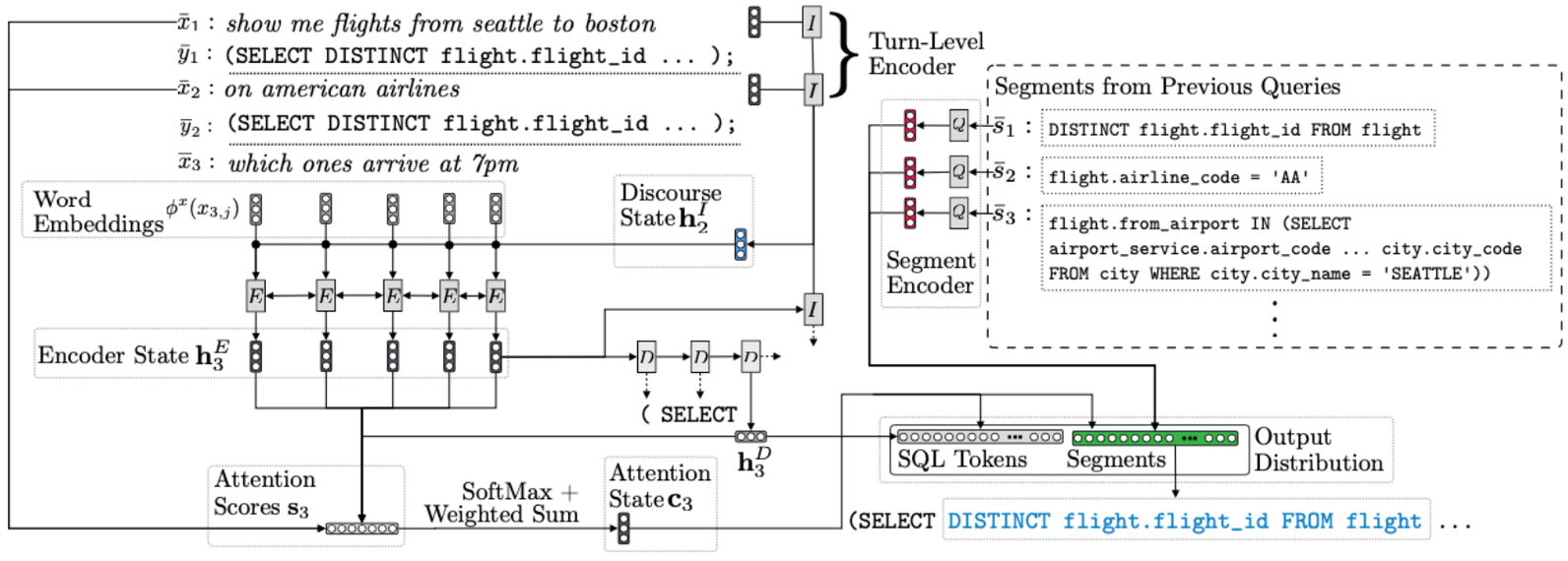
Lecture 8: semantic parsing

Transform text into logical form

(Dong and Lapata, 2016)



(Suhr et al, 2018)



Lecture 9: reading comprehension

Read a passage of text and answer comprehension questions

Passage

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

Question: Which NFL team won Super Bowl 50?
Answer: Denver Broncos

Question: What does AFC stand for?
Answer: American Football Conference

Question: What year was Super Bowl 50?
Answer: 2016

Q: Where did Charles travel to first, Castile or Barcelona?

In 1517, the seventeen-year-old King sailed to Castile, where he was formally recognised as King of Castile.There, his Flemish court .... In May 1518, Charles traveled to Barcelona in Aragon.

(Hermann et al, 2015)

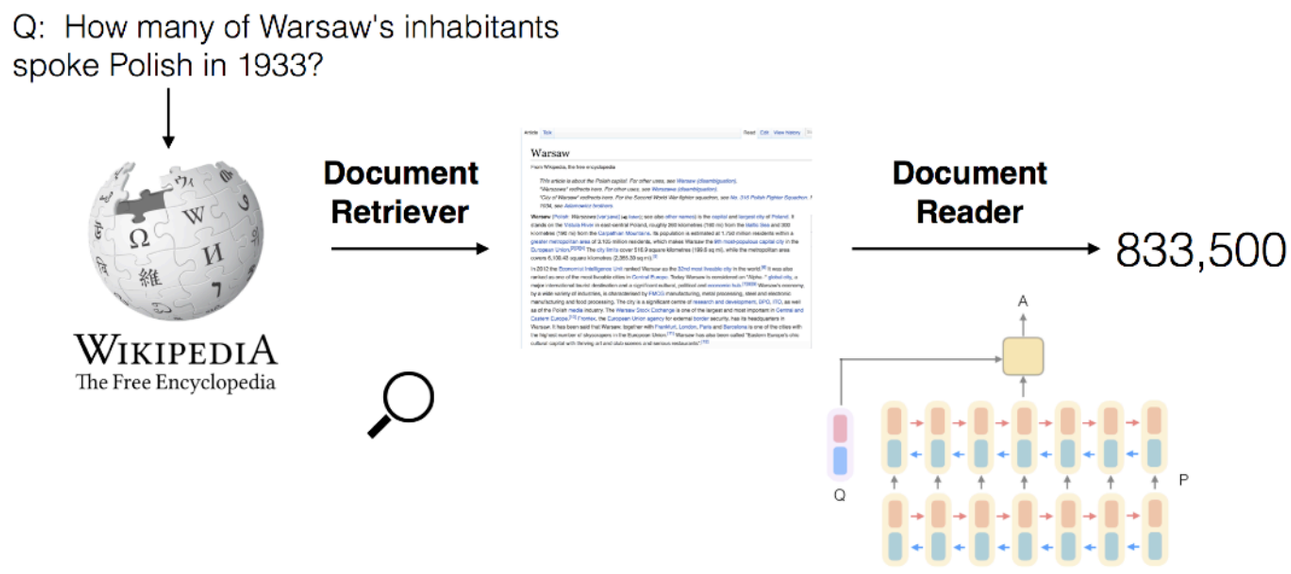
(Seo et al, 2017)

Lecture 10: open-domain QA

Answer any factoid questions over large collection of documents

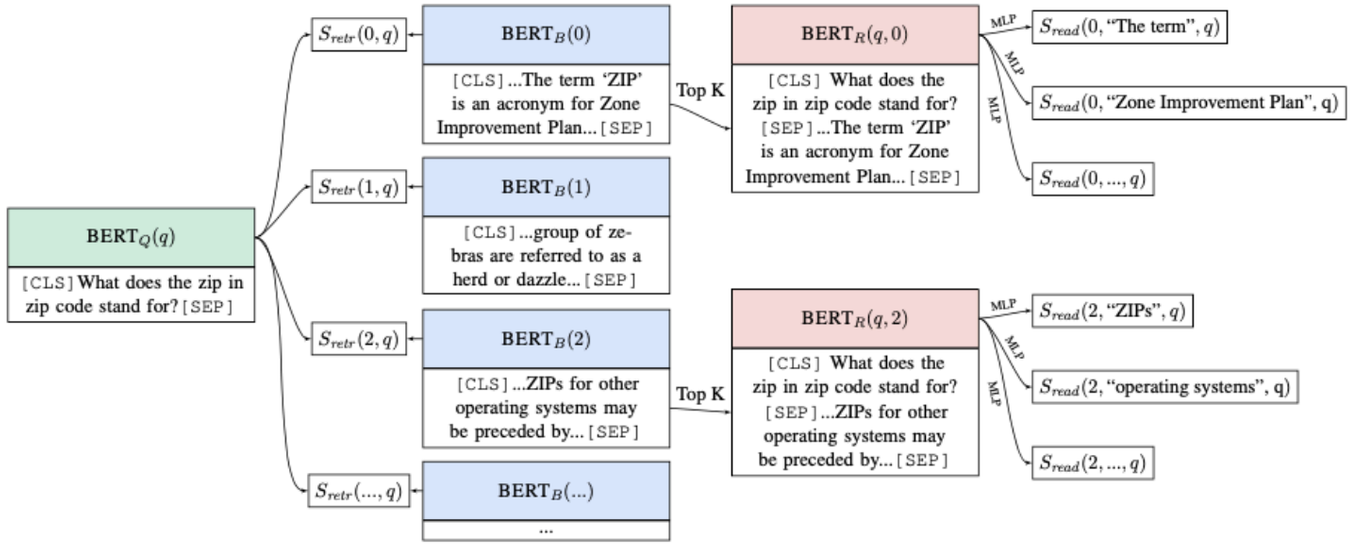
(Chen et al, 2017)

DrQA



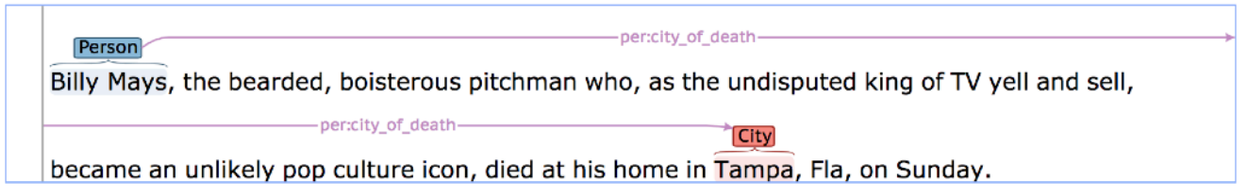
(Lee et al, 2019)

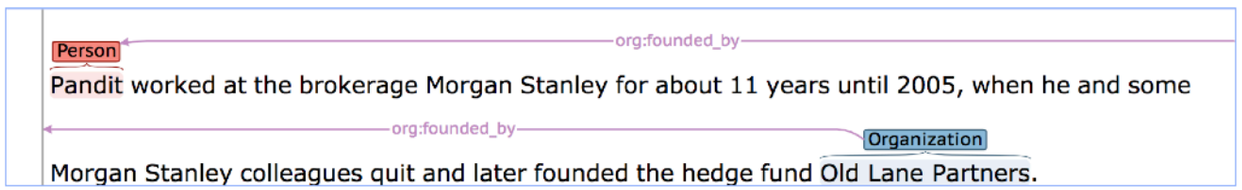
ORQA



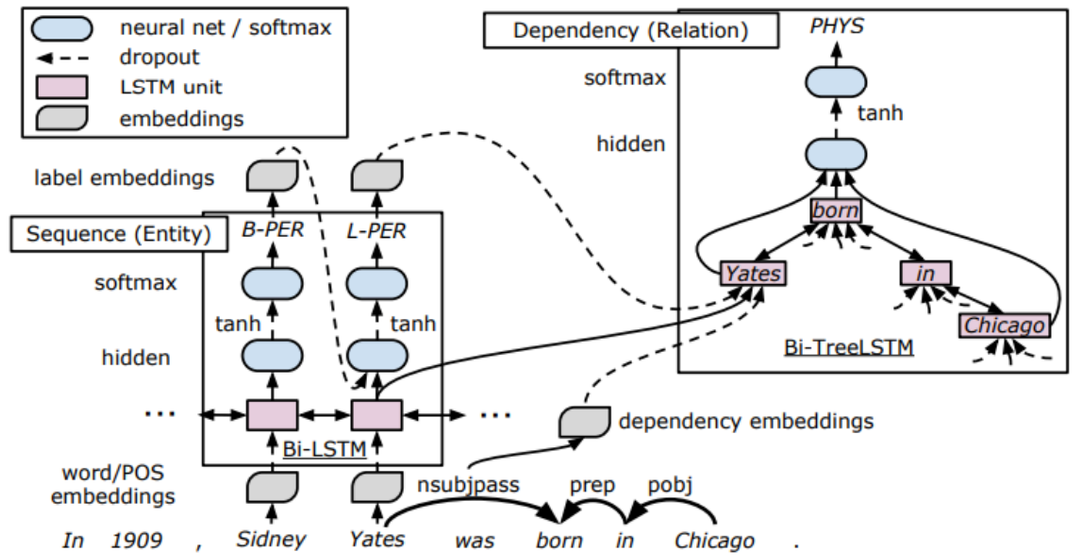
Lecture 11: relation extraction

Identify relations between two mentions in the text

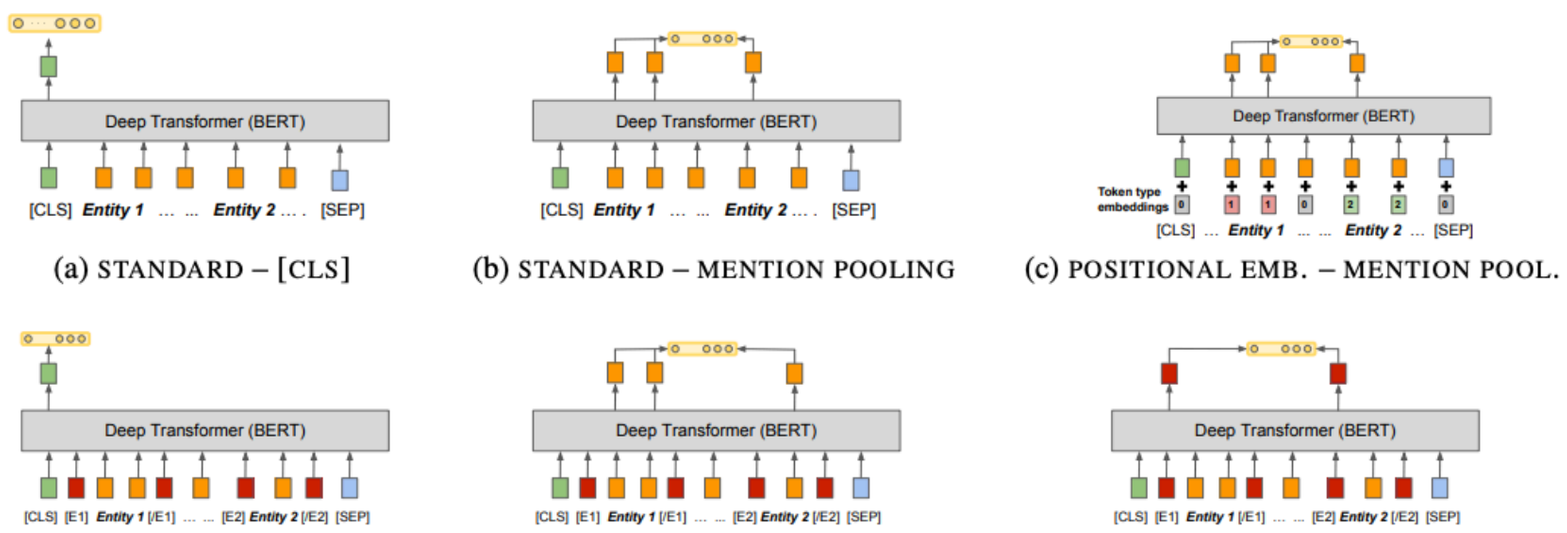




(Miwa and Bansal, 2016)



(Soares et al, 2019)

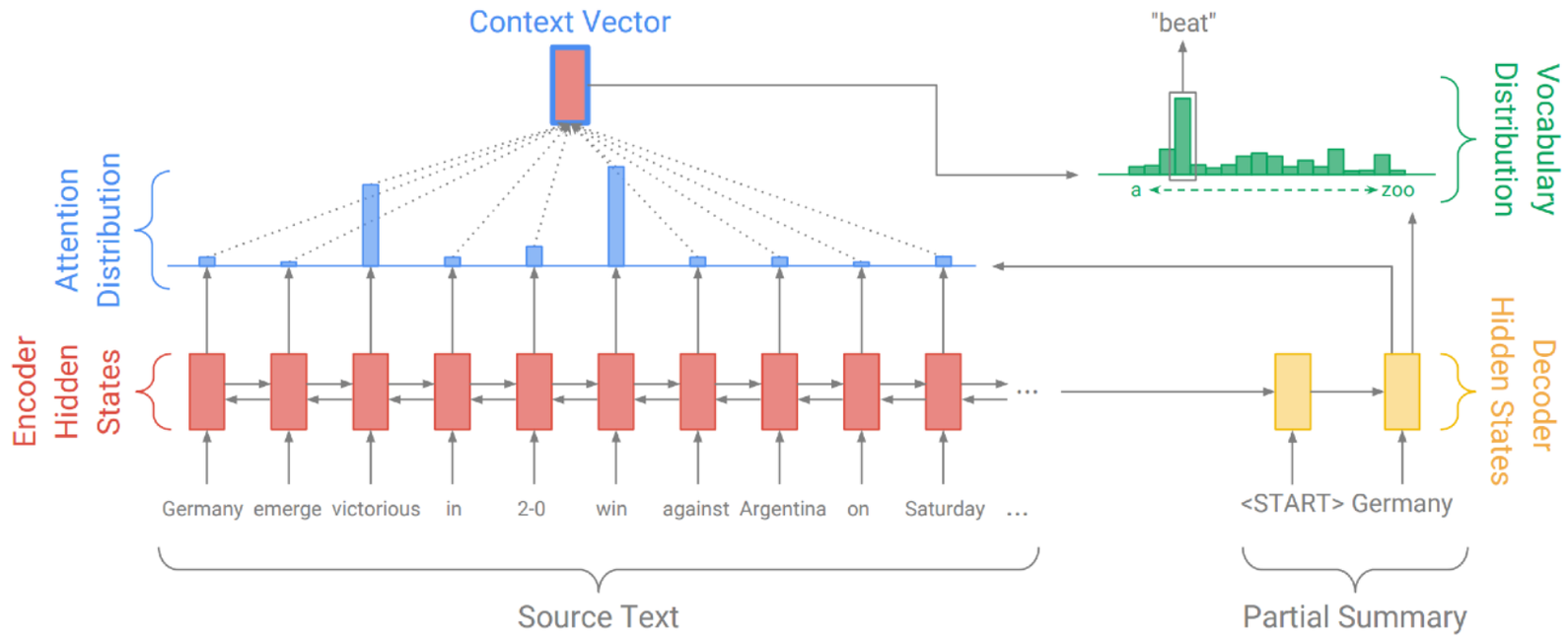


(a) STANDARD – [CLS] (b) STANDARD – MENTION POOLING (c) POSITIONAL EMB. – MENTION POOL.

Lecture 12: summarization

Condense a piece of text into a short summary

(See et al, 2017)



(Paulus et al, 2018)

A Deep Reinforced Model for Abstractive Summarization

Romain Paulus, Caiming Xiong*& Richard Socher
Salesforce Research
575 High Street
Palo Alto, CA 94301, USA
{rpaulus,cxiong,rsocher}@salesforce.com

Lecture 13: summarization (cont'd)

(Cheng and Lapata, 2016)

AFL star blames vomiting cat for speeding

Adelaide Crows defender Daniel Talia has kept his driving license, telling a court he was speeding 36km over the limit because he was distracted by his sick cat.

The 22-year-old AFL star, who drove 96km/h in a 60km/h road works zone on the South Eastern expressway in February, said he didn't see the reduced speed sign because he was so distracted by his cat vomiting violently in the back seat of his car.

In the Adelaide magistrates court on Wednesday, Magistrate Bob Harrap fined Talia $824 for exceeding the speed limit by more than 30km/h.

He lost four demerit points, instead of seven, because of his significant training commitments.

- *Adelaide Crows defender Daniel Talia admits to speeding but says he didn't see road signs because his cat was vomiting in his car.*
- *22-year-old Talia was fined $824 and four demerit points, instead of seven, because of his 'significant' training commitments.*

(Liu et al, 2018)

GENERATING WIKIPEDIA BY SUMMARIZING LONG SEQUENCES

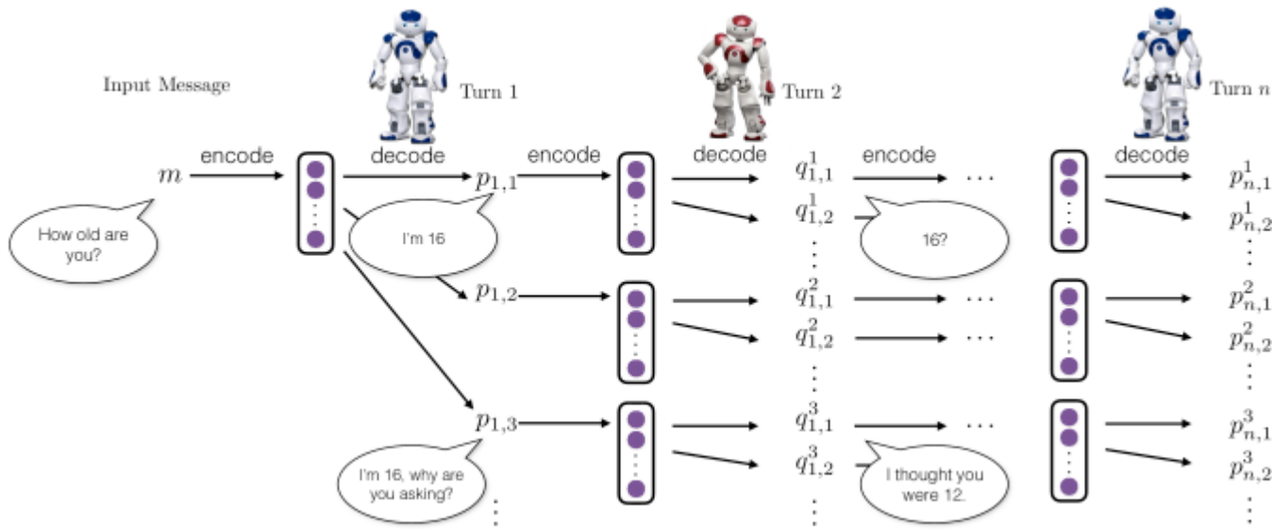
Peter J. Liu*, **Mohammad Saleh***,
Etienne Pot†, **Ben Goodrich, Ryan Sepassi, Łukasz Kaiser, Noam Shazeer**
Google Brain
Mountain View, CA
{peterjliu,msaleh,epot,bgoodrich,rsepassi,lukaszkaiser,noam}@google.com

Lecture 14: dialogue

“Chatbot”

Machine: *hi*
Human: *hello*
Machine: *could you please let me know what are the operating systems you are using ?*
Human: *i am using linux*
Machine: *ok it was in the ssl vpn on which you stated it was broken*
Machine: *please check if you are logged in to corp network*
Machine: *what is the error that you are running please*
Human: *i am seeing an error related to vpn*
Machine: *what is the error message that you are getting when connecting to vpn using network connect ?*
Human: *connection refused or something like that*
Machine: *may i know the version of network connect you connect ?*
Human: *i am not sure i know that*
Machine: *is the network connect prompting for any user certificate*

(Vinyals and Le, 2015)



(Li et al, 2016)

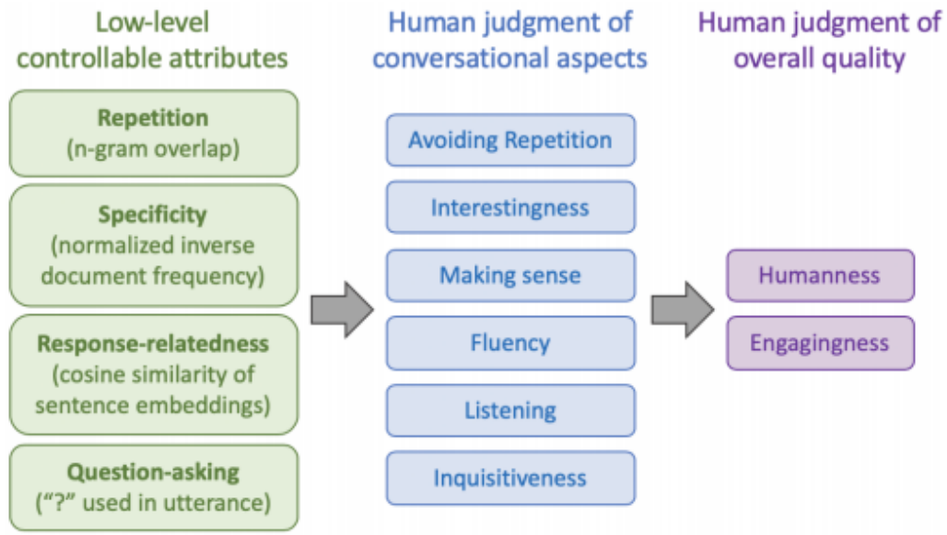
Lecture 15: dialogue (cont'd)

Chatbot with controllable attributes

Persona 1	Persona 2
I like to ski	I am an artist
My wife does not like me anymore	I have four children
I have went to Mexico 4 times this year	I recently got a cat
I hate Mexican food	I enjoy walking for exercise
I like to eat cheetos	I love watching Game of Thrones

[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

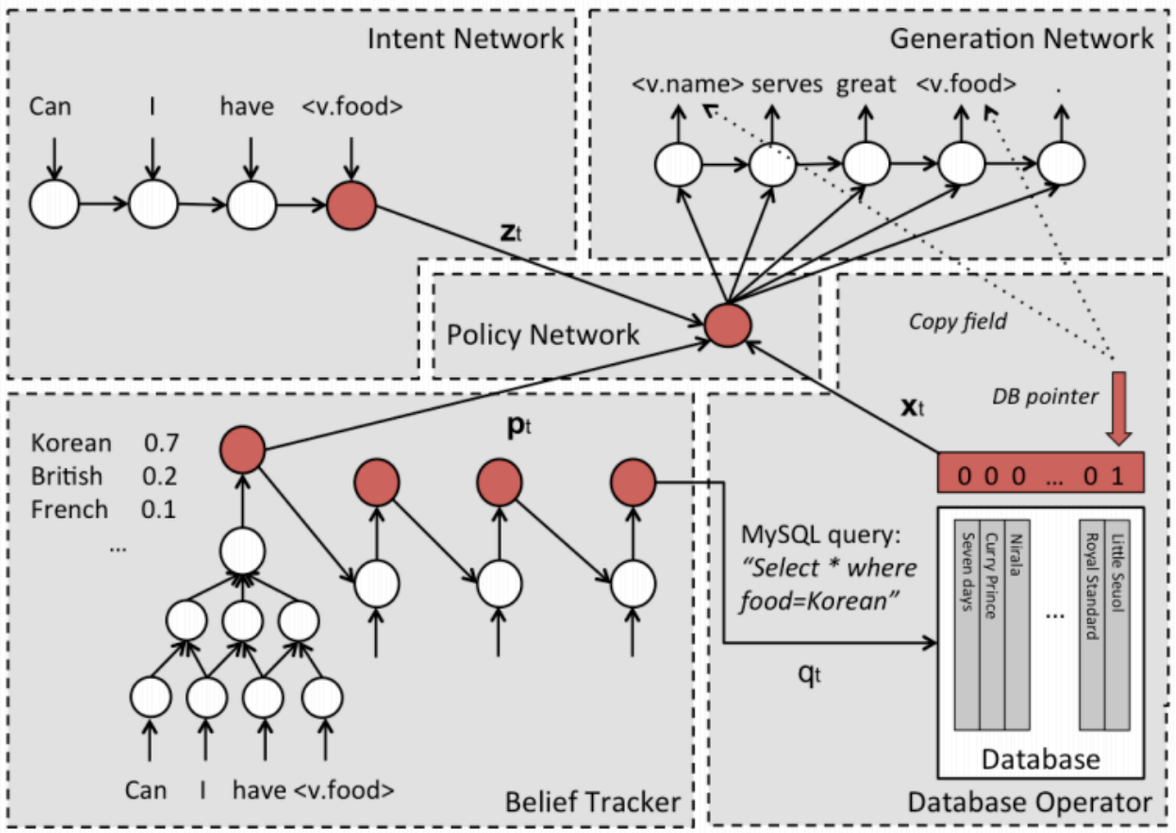
(Zhang et al, 2018)



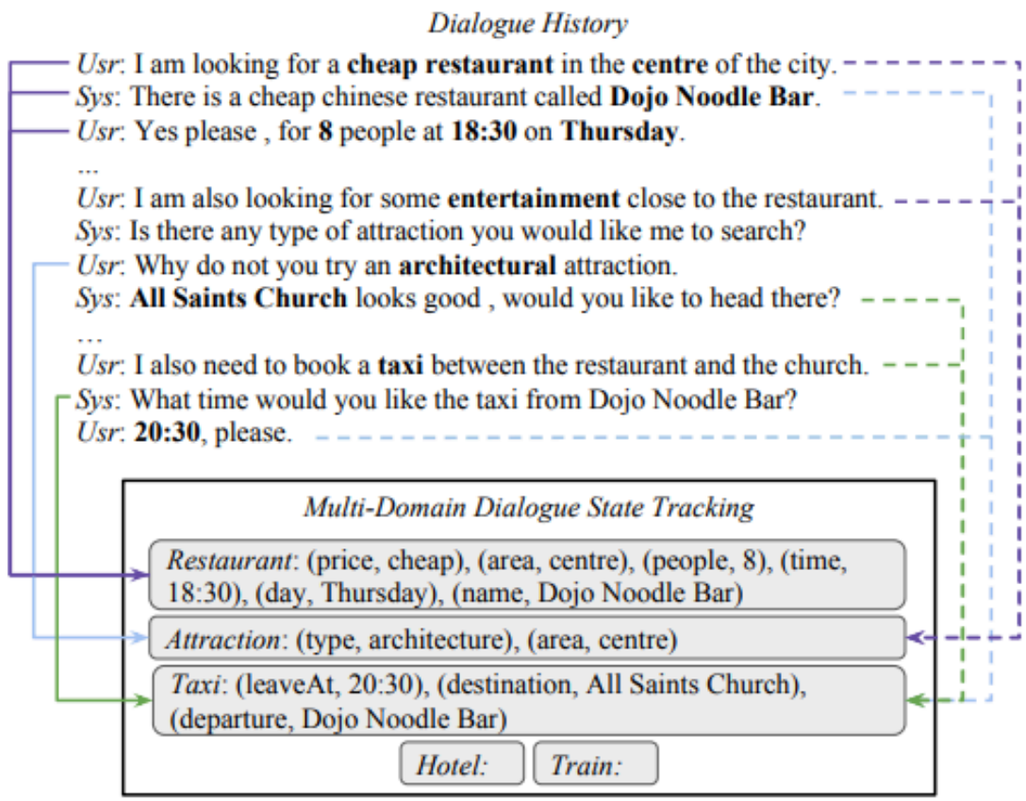
What makes a good conversation?
How controllable attributes affect human judgments

(See et al, 2019)

Lecture 16: task-oriented dialogue



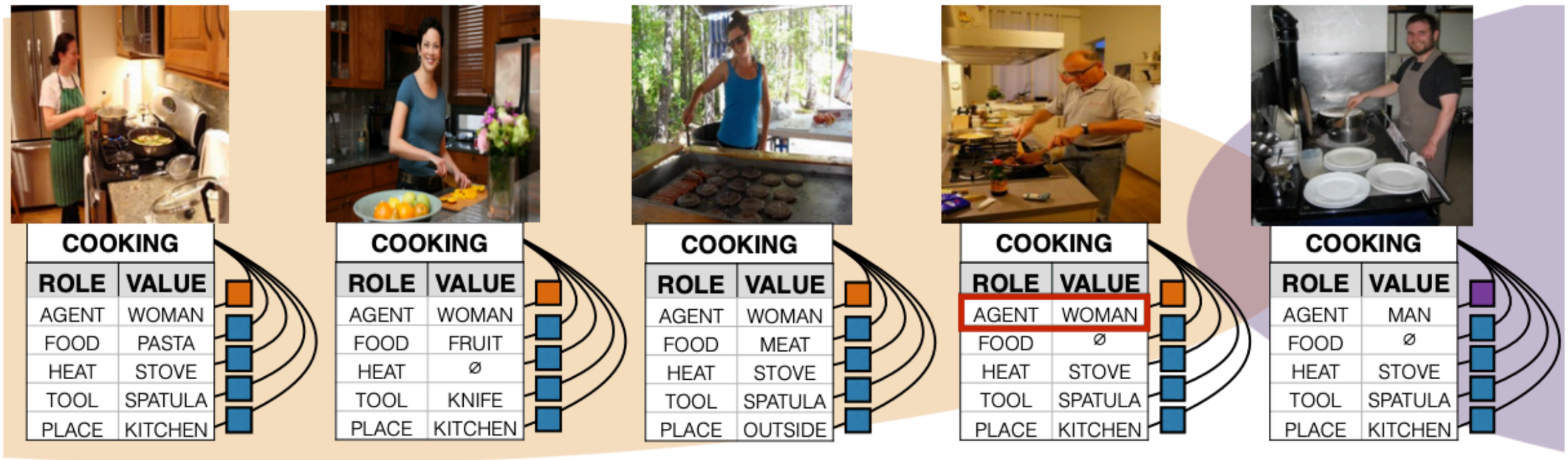
(Wen et al, 2017)



(Wu et al, 2019)

Lecture 17: Bias in language

(Zhao et al, 2017)



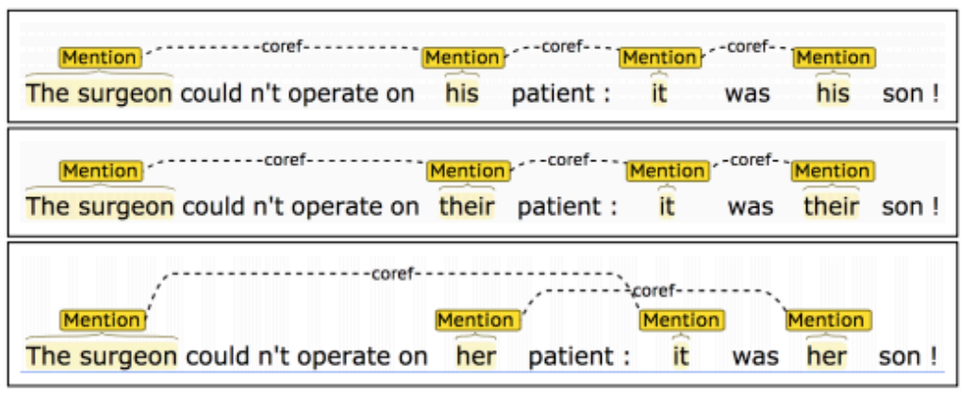


Figure 1: Stanford CoreNLP rule-based coreference system resolves a male and neutral pronoun as coreferent with "The surgeon," but does not for the corresponding female pronoun.

(May et al, 2019)

On Measuring Social Biases in Sentence Encoders

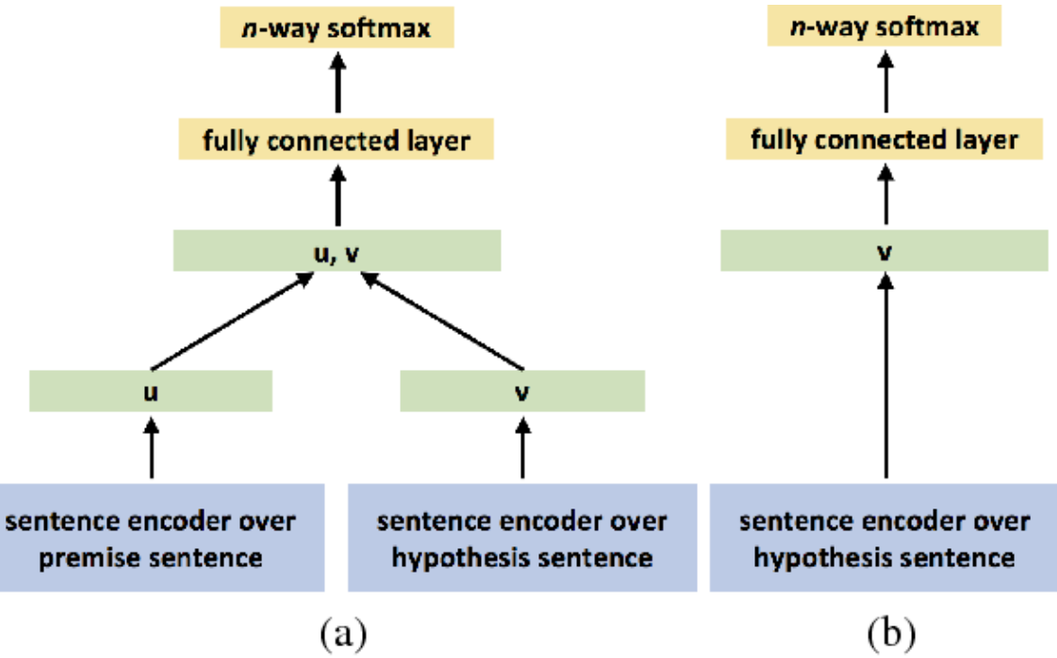
Chandler May[1] **Alex Wang**[2] **Shikha Bordia**[2]
Samuel R. Bowman[2] **Rachel Rudinger**[1]
[1]Johns Hopkins University [2]New York University
{cjmay,rudinger}@jhu.edu {alexwang,sb6416,bowman}@nyu.edu

Lecture 18: annotation artifacts in NLP

Premise: ***The brown cat ran***

Hypothesis: ***The animal moved***

entailment neutral contradiction



(Kaushik and Lipton, 2018)

(Belinkov et al, 2019)

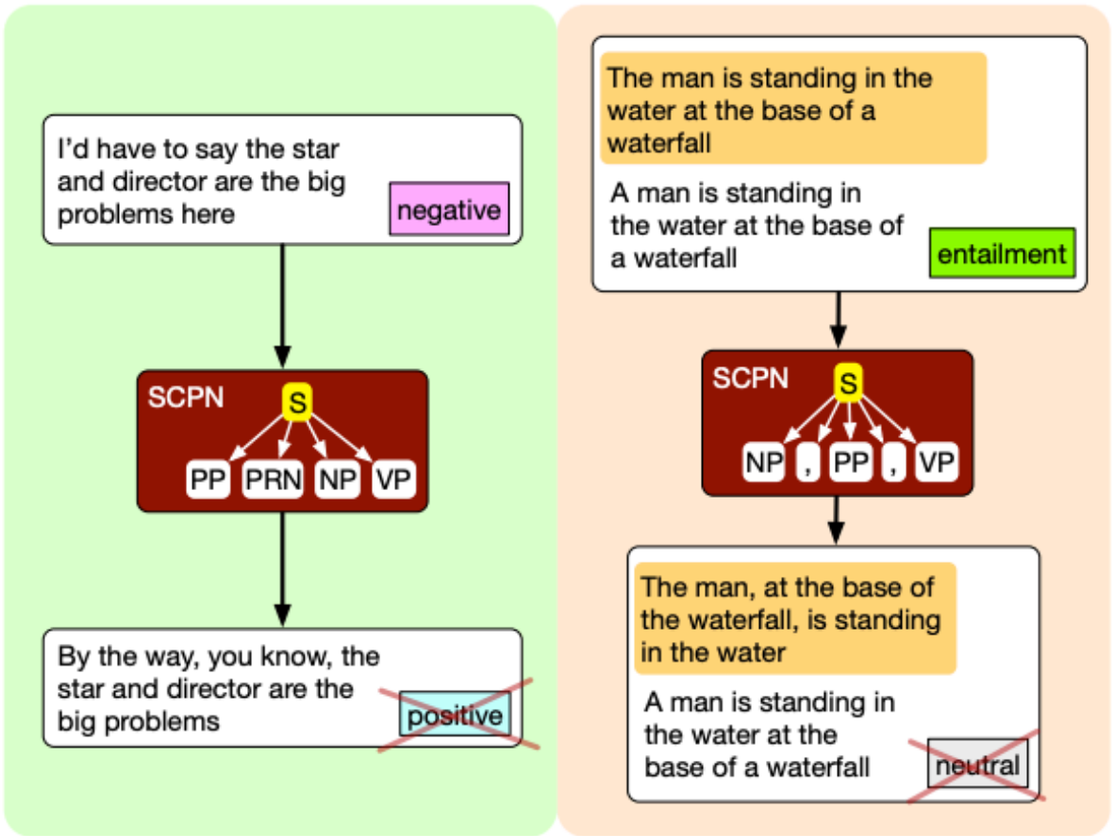
Don't Take the Premise for Granted: Mitigating Artifacts in Natural Language Inference

Yonatan Belinkov[13*] **Adam Poliak**[2*]
Stuart M. Shieber[1] **Benjamin Van Durme**[2] **Alexander M. Rush**[1]
[1]Harvard University [2]Johns Hopkins University [3]Massachusetts Institute of Technology
{belinkov,shieber,srush}@seas.harvard.edu
{azpoliak,vandurme}@cs.jhu.edu

Lecture 19: adversarial examples

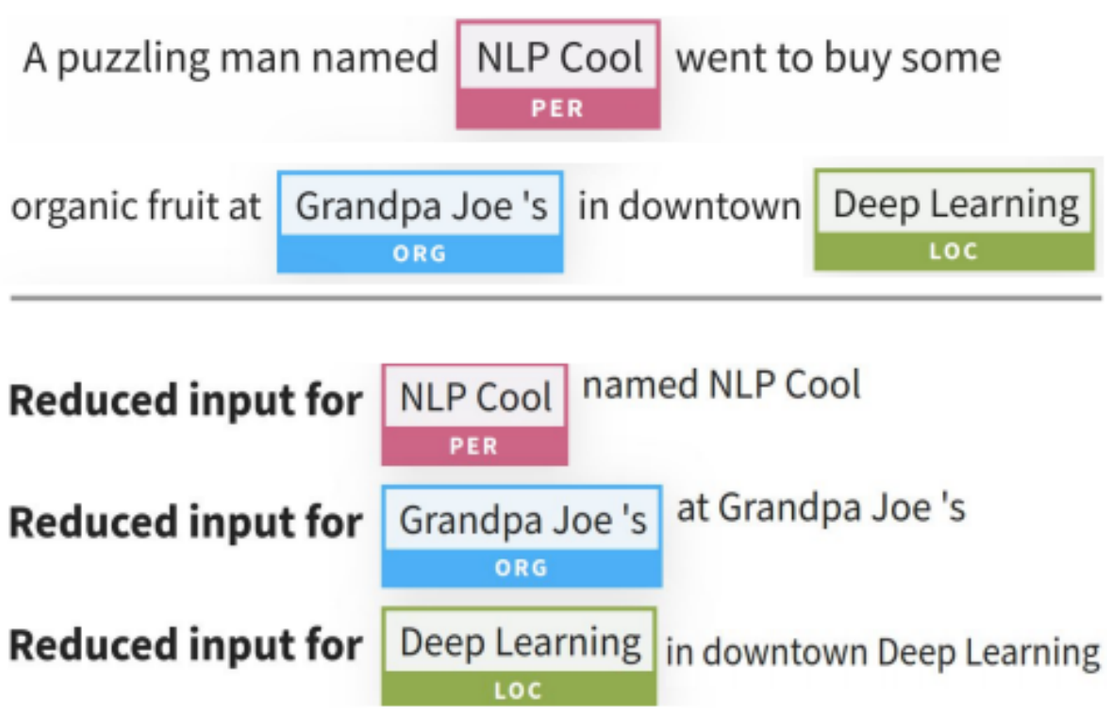
Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

(Jia and Liang, 2017)

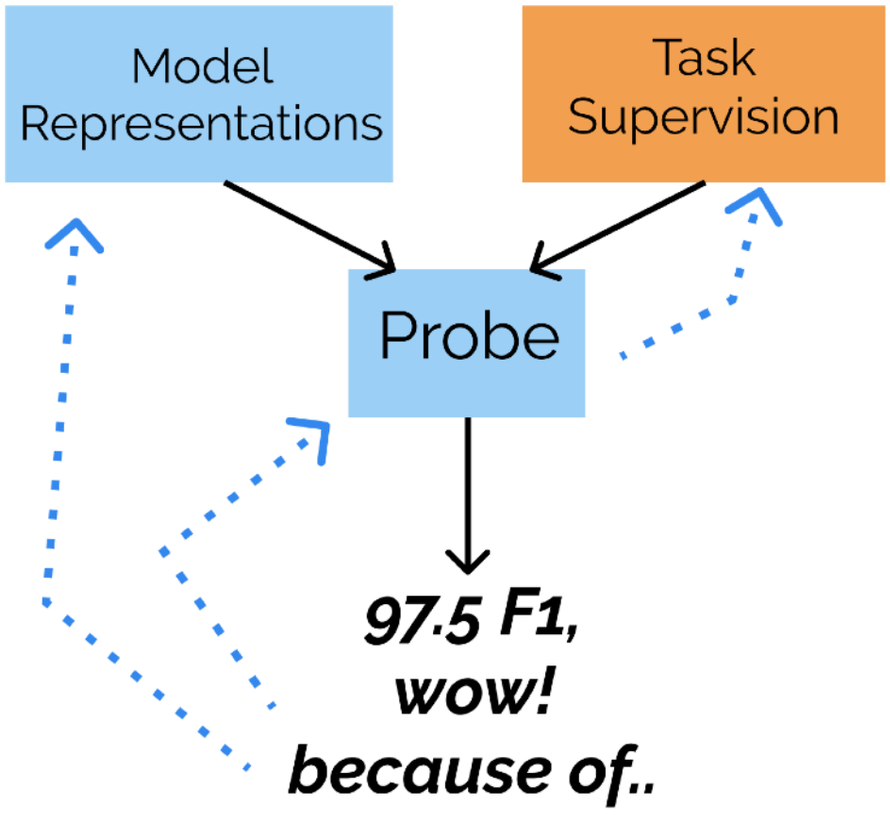


(Iyyer et al, 2018)

Lecture 20: Interpretability



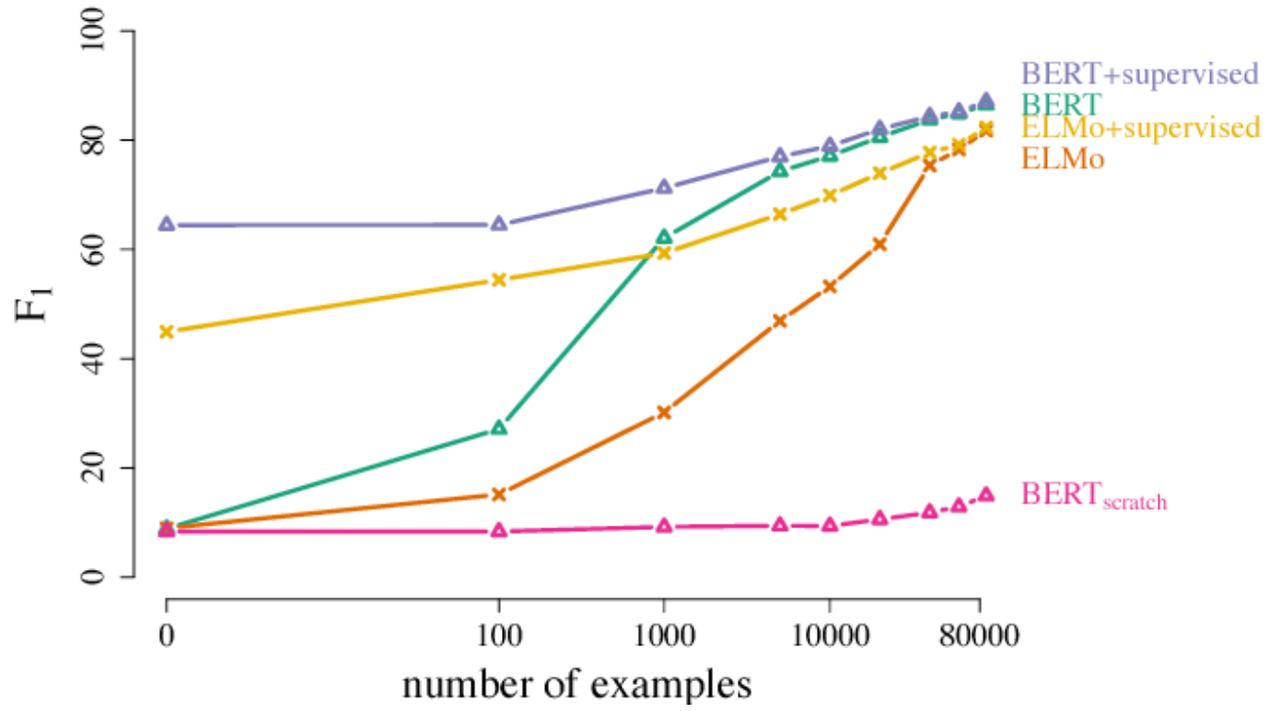
(Wallace et al, 2019)



(Hewitt and Liang, 2019)

Lecture 21: general linguistic intelligence

(Yogatama et al, 2019)



(Wang et al, 2019)

SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems

Alex Wang* New York University
Yada Pruksachatkun* New York University
Nikita Nangia* New York University
Amanpreet Singh* Facebook AI Research
Julian Michael University of Washington
Felix Hill DeepMind
Omer Levy Facebook AI Research
Samuel R. Bowman New York University

Next lecture: word embeddings

- (Mikolov et al, 2013): Distributed Representations of Words and Phrases and their Compositionality

- (Baroni et al, 2014): Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors

- [Optional] (Levy et al, 2015): Improving Distributional Similarity with Lessons Learned from Word Embeddings