Natural Language Processing

Stuff we didn't cover

- NLP applications
- Cross-task benchmarks
- Multi-task learning and transfer learning
- Using unlabeled data
- Generalization, Bias, Robustness
- Architectures

NLP applications

- We skipped many NLP applications

Coreference resolution

clustering together of expressions that refer to the same concept/entity

Michelle LaVaughn Robinson Obama is an American lawyer and writer. ***She*** *is the wife of the 44th president of the United States, Barack Obama, and the first African-American first lady of the United States*

Main sub-tasks

- Entity extraction
- Coreference resolution
- Entity linking

Entity extraction

- Find all mentions of entities in a text
  - *Michelle LaVaughn Robinson Obama*
  - *She*
  - *The wife of…*

Coreference resolution

- Clustering of the entities extracted

Michelle Obama
she
the wife of…

Barack Obama
The president…

Entity linking

Michelle Obama
she
the wife of…

Michelle Obama

Former First Lady of the United States

Michelle LaVaughn Robinson Obama is an American lawyer and writer who was First Lady of the United States from 2009 to 2017. She is married to the 44th President of the United States, Barack Obama, and is the first African-American First Lady. Wikipedia

Born: January 17, 1964 (age 53), Chicago, Illinois, United States

Height: 1.8 m

Spouse: Barack Obama (m. 1992)

Education: Harvard Law School (1985–1988), more

Parents: Marian Shields Robinson, Fraser C. Robinson III

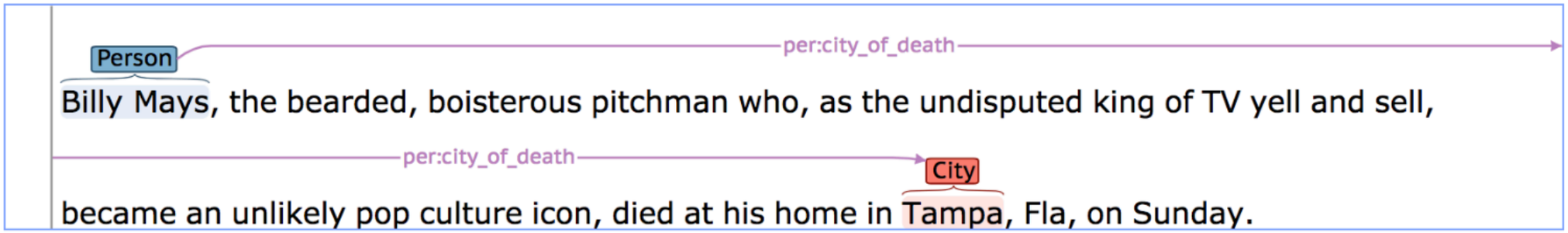
Bowman et al., 2015

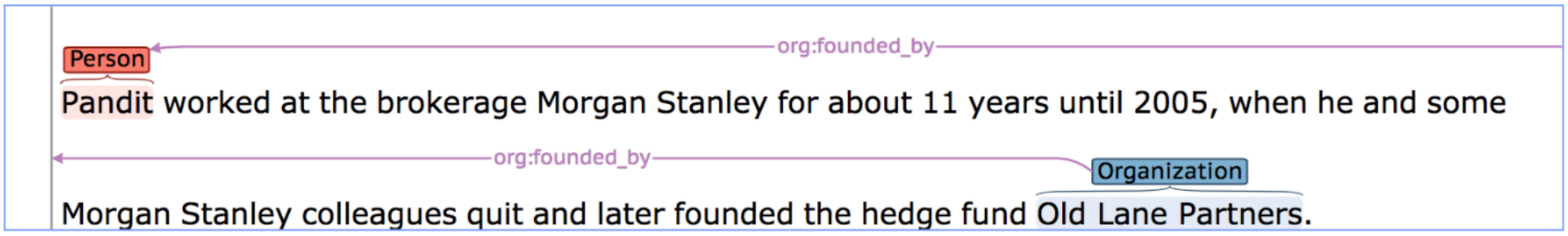
Natural Language Inference

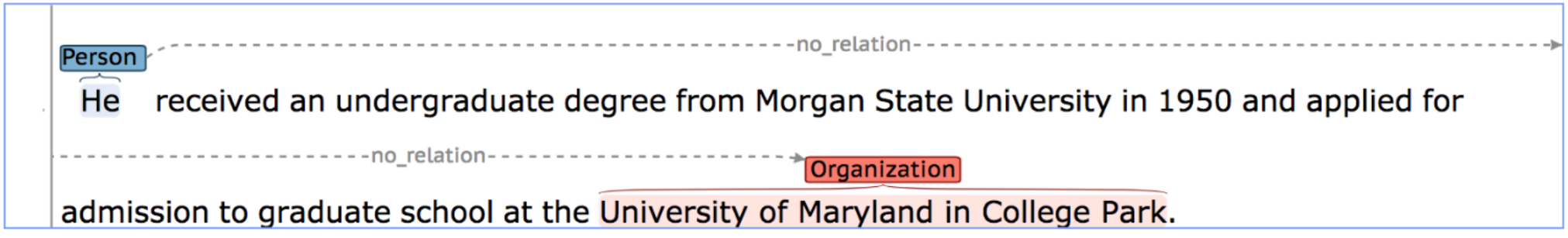
Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd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

Approaches: represent both sentences and treat as a classification problem

Relation extraction







Wang et al., 2015

Cross-task benchmarks

SuperGLUE GLUE | Paper Code Tasks Leaderboard FAQ Diagnostics Submit Login

Rank	Name	Model	URL	Score	AX	CB	COPA	MultiRC	RTE	WiC	WSC
1	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.6	76.6	95.8/98.9	100.0	81.8/51.9	93.6	80.0	100.0
2	SuperGLUE Baselines	BERT++		70.5	42.1	82.7/88.4	77.4	66.2/22.2	77.6	70.4	67.8
		BERT	BERT	68.0	19.1	80.6/84.4	69.8	66.2/22.2	73.2	70.4	67.8
		CBOW		48.6	2.4	47.6/69.2	52.2	38.8/0.5	50.4	50.0	61.0
		Most Frequent Class		46.9	0.0	21.7/48.4	50.0	61.1/0.3	50.3	50.0	65.1
		Outside Best		-	-	-	84.4	70.4/24.5	82.7	-	-

Cross-task benchmarks

Rank	Submission	DROP EM	DROP F1	DuoRC EM	DuoRC F1	Narrative B1	Narrative B4	Narrative MET.
1	**NABERT-baseline** *UCI and AI2*	0.20	0.24	0.25	0.34	0.12	0.02	0.23

Dua et al., 2019

Datasets

Link to the single format datasets are provided in to formats: MultiQA, and SQuAD2.0. The SQuAD2.0 are G JSONs that are the results of applying convert_multiqa_to_squad_format to the MultiQA dataset. To used th Pytorch-Transformers code simply unzip them. (see models Readme for an example)

Dataset	MultiQA format	SQuAD2.0 format (GZipped)
SQuAD-1.1	train , dev	train , dev
SQuAD-2.0	train , dev	train , dev
NewsQA	train , dev	train , dev
HotpotQA	train , dev	train , dev
TriviaQA-unfiltered	dev	dev
TriviaQA-wiki	train , dev	train , dev
SearchQA	train , dev	train , dev
BoolQ	train , dev	train , dev
ComplexWebQuestions	train , dev	train , dev
DROP	train , dev	train , dev
WikiHop	train , dev	train , dev
DuoRC Paraphrase	train , dev	train , dev
DuoRC Self	train , dev	train , dev
ComplexQuestions	train , dev	train , dev
ComQA	train , dev	train , dev
Natural Questions	Coming soon	Coming soon

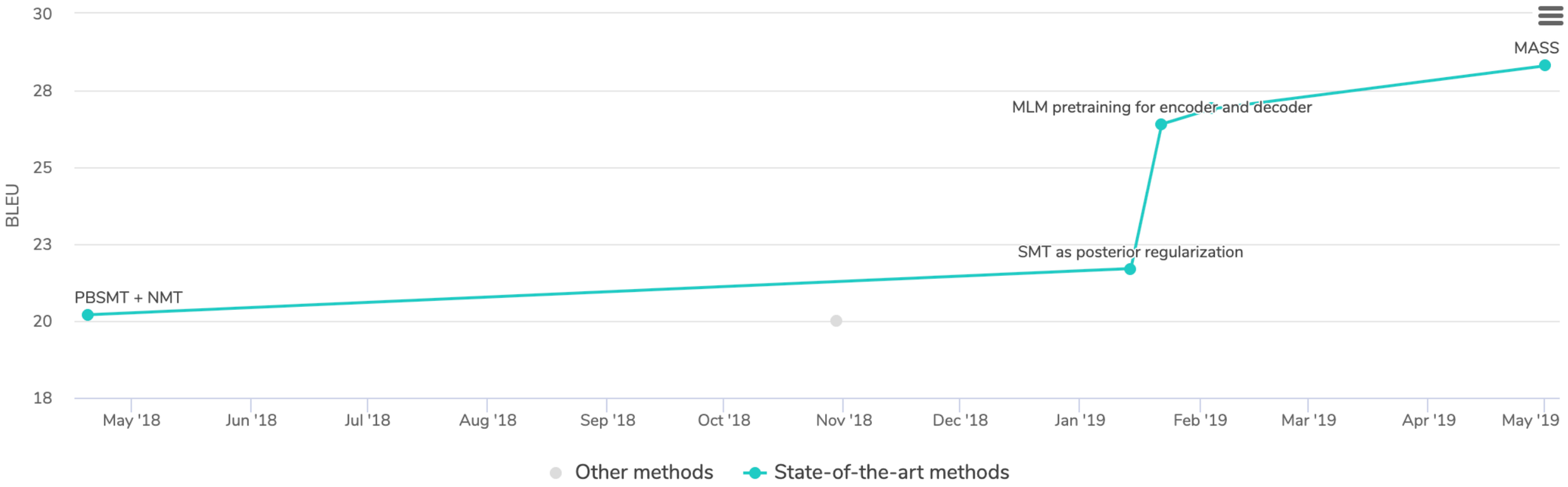
Talmor and Berant, 2019

Transfer learning

- Learning from multiple tasks and transferring knowledge from data rich problems to data poor problems
- Multi-task learning
- Recent tutorial

Unlabeled data

- More and more work on unsupervised and self-supervised learning



Generalization

- We train models assuming the test distribution is identical to the training distribution

- Often we find that any deviation from the distribution leads to low performance

Generalization

- Research threads:
    - Adversarial attacks: showing that it is easy to fail NLP models
    - Compositional generalization: if language is compositional, we expect to generalize to compositions of atoms seen at training time
    - Robustness: How to train models that are not sensitive to small changes to the input

Bias and ethics

- Representations learned from free text record and amplify existing human biases
  - doctor - man + woman = nurse
- A lot of research on these biases and how to mitigate them
- In general, NLP is now **popular**, which leads to ethically questionable research:

Architectures

- Graph neural networks

- CNNs (there is a class just on that)

Semester B

מספר הקורס	שם הקורס
0368-3711-01	סמינר בעיבוד שפה טבעית Seminar in Natural Language Processing
מדעים מדויקים/מדעי המחשב	

שם המרצה	סוג השיעור	בניין	חדר	יום	שעה	סמסטר
ד"ר ברנט יהונתן	סמינר	קפלון	324	ד'	0800 - 1000	ב'
דרישות קדם						

מספר הקורס	שם הקורס
0368-4205-01	סמינר בתרגום מכונה וייצור שפה טבעית Machine Translation and Natural Language Generation Seminar
מדעים מדויקים/מדעי המחשב	

שם המרצה	סוג השיעור	בניין	חדר	יום	שעה	סמסטר
מר לוי עמר	סמינר	קפלון	319	ב'	1400 - 1600	ב'
דרישות קדם						

מספר הקורס	שם הקורס
0368-4206-01	שיטות מתקדמות בעיבוד שפת טבעית Advanced Methods in Natural Language Processing
מדעים מדויקים/מדעי המחשב	

שם המרצה	סוג השיעור	בניין	חדר	יום	שעה	סמסטר
מר לוי עמר	שיעור	קפלון	324	ג'	1000 - 1300	ב'
דרישות קדם						

Broad takeaways

What did we cover?

- Word embeddings
- Language models
- Sequence tagging
- Syntactic parsing
- Sequence-to-sequence models

Main technical tools

- Structured prediction
- Deep learning
- General recipe:
  - Define a parameterized mapping from input to output
  - Define a loss function
  - Optimize
  - Find best output at test time

High-level observations

- Often linear models can be replaced with non-linear ones without change to the guarantees
- Burden is moving from inference to learning. Information flows between various variables in a neural network and inference becomes simple
  - This is evident in transformers
- It is still an active research area

Hopefully you

- Appreciate the complexity of building systems for natural language
- Understand the main tools used to build state-of-the-art systems nowadays
- Have solid background to read papers
- Have solid background to develop models for NLP

Thank you!