Adversarial Examples for Evaluating Reading Comprehension Systems

Robin Jia and Percy Liang
Stanford University

Reading Comprehension Task

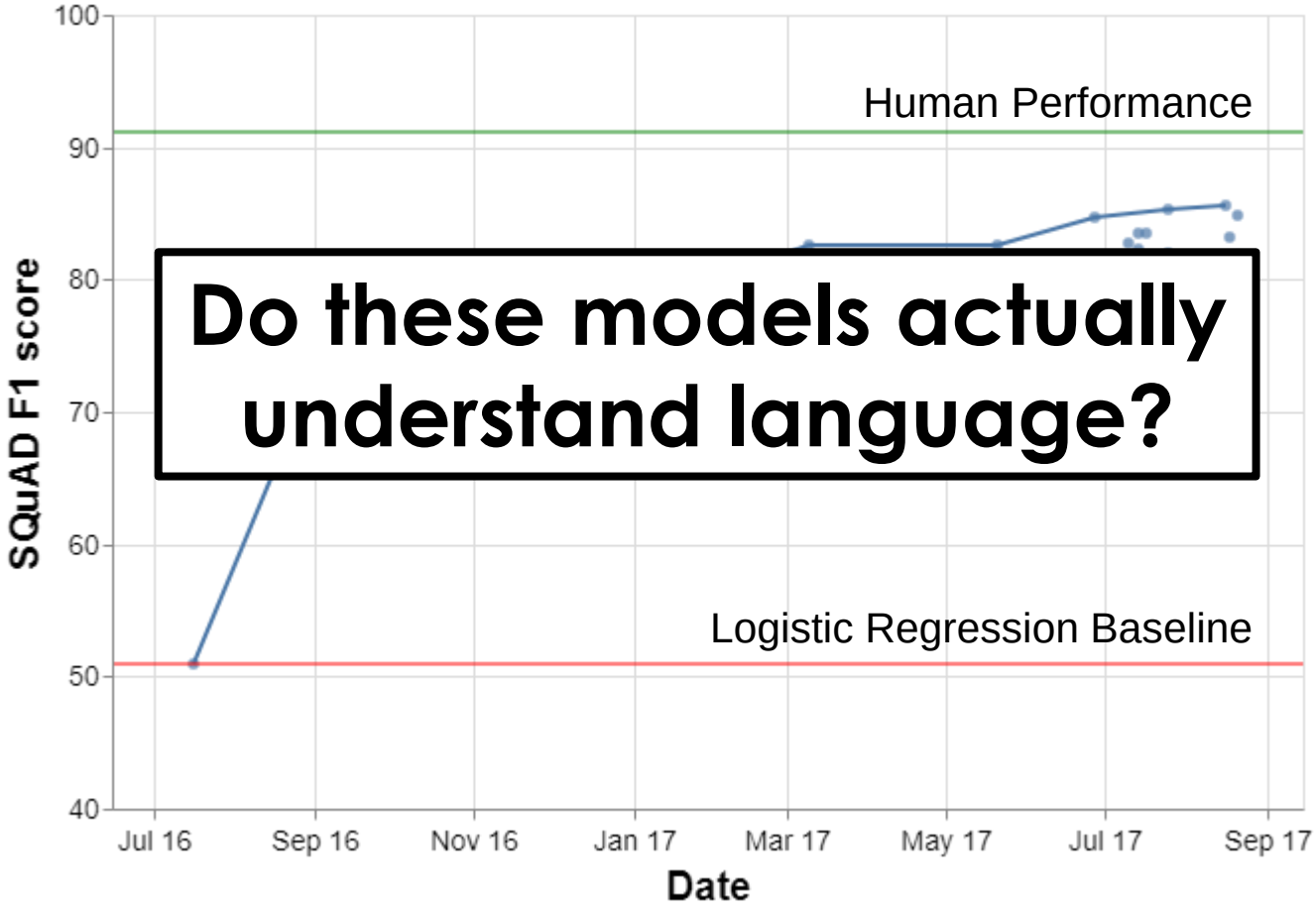
Question: "*The number of new Huguenot colonists declined after what year?*"

Paragraph: "*The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689…but quite a few arrived as late as* ***1700****; thereafter, the numbers declined…*"

Correct Answer: "***1700***"

Stanford Question Answering Dataset (Rajpurkar et al., 2016)

Progress on SQuAD



SQuAD leaderboard, https://rajpurkar.github.io/SQuAD-explorer/

Adversarial Evaluation

Question: "*The number of new Huguenot colonists declined after what year?*"

Paragraph: "*The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689…but quite a few arrived as late as* ***1700****; thereafter, the numbers declined. The number of old Acadian colonists declined after the year of* ***1675****.*"

Correct Answer: "***1700***"

Predicted Answer: "***1675***"

Model used: BiDAF Ensemble (Seo et al., 2016)

Adversarial Evaluation

Question: "*The number of new Huguenot colonists declined after what year?*"

Paragraph: "*The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689…but quite a few arrived as late as* ***1700****; thereafter, the numbers declined. expected yet later be basis need young only required* ***1961****.*"

Correct Answer: "***1700***"

Predicted Answer: "***1961***"

Model used: BiDAF Ensemble (Seo et al., 2016)

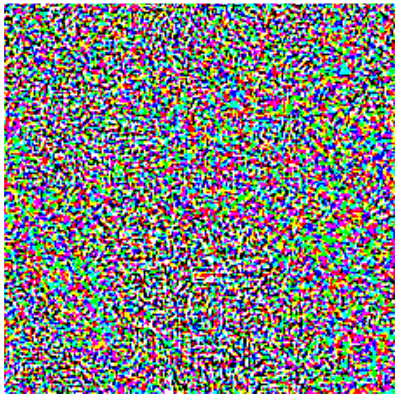
Outline

- Inspiration/Motivation
- Adding Grammatical Sentences
- Adding Word Salad
- Trying to build better systems

Outline

- Inspiration/Motivation
- Adding Grammatical Sentences
- Adding Word Salad
- Trying to build better systems

Some Inspiration


+ .007 *

=


Panda
58% confidence

Nematode
8% confidence

Gibbon
99% confidence

Local perturbations don't change semantics of image, but models are **oversensitive** to small differences!

Goodfellow et al., 2014.

Local perturbations of text

Question: "*The number of new Huguenot colonists declined after what year?*"

Paragraph: "*The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689…but quite a few arrived as late as* ***1700****; thereafter, the* ~~*numbers*~~ ***amount*** ~~*declined*~~ ***decreased****…*"

Plausible alternative answers not always present

Hard to find a lot of perturbations to try

Li et al., 2017

Preserving Semantics

- For images, most local perturbations **preserve** semantics
- For text, most local perturbations **alter** semantics
  - Even changing one word by a small amount may not preserve semantics (e.g. entity names)

Concatenative Adversaries

- Instead of locally altering the input, **append distracting text** to the paragraph
- Must ensure that added text does not actually answer the question

Distracting Text

Question: "*The number of new Huguenot colonists declined after what year ?*"

Distracting text: "*The number of new Huguenot colonists declined after the year 1675 .*"

Answer according to text: "***1675***"

Distracting Text

Question: "*The number of new Huguenot colonists declined after what year ?*"

Distracting text: "*The number of ~~new~~* ***old*** *~~Huguenot~~* ***Acadian*** *colonists declined after the year 1675 .*"

Answer according to text: ***N/A***

Local perturbations change semantics of sentence, but models are **overly stable/insensitive** to these changes!

Outline

- Inspiration/Motivation
- Adding Grammatical Sentences
- Adding Word Salad
- Trying to build better systems

Grammatical Distractors

What city did Tesla move to in 1880?

Prague

Change entities, numbers, antonyms

Generate fake answer with same NER/POS tag

What city did Tadakatsu move to in 1881?

Chicago

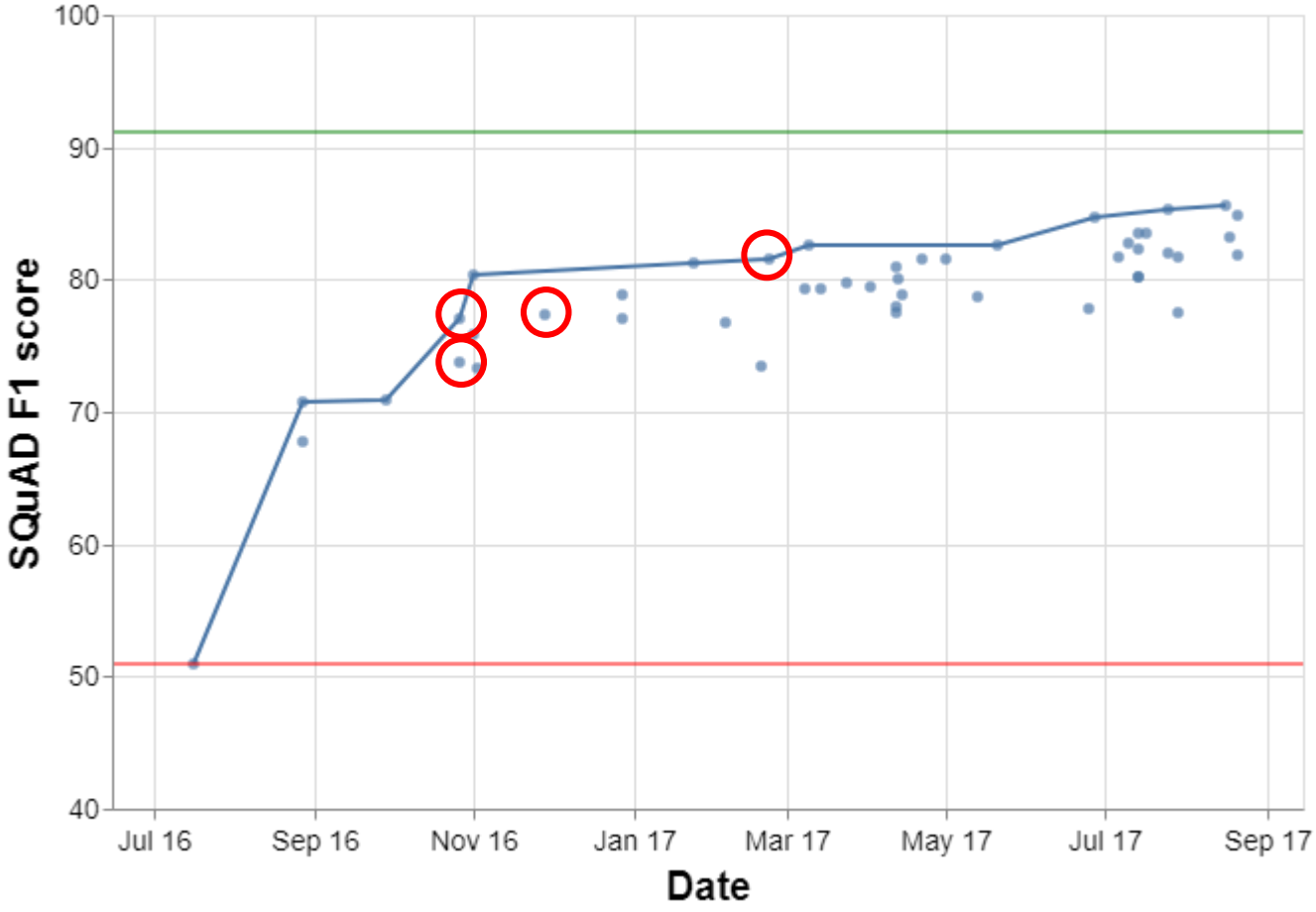
Convert to declarative sentence

Tadakatsu moved the city of Chicago to in 1881.

Have crowdworkers fix errors

Tadakatsu moved to the city of Chicago in 1881.

Four "dev" systems



SQuAD leaderboard, https://rajpurkar.github.io/SQuAD-explorer/
*Some of our results are on older versions of models than shown here

Results (4 "dev" systems)

System	Original	AddOneSent
BiDAF, ensemble (Seo et al., 2016)	80.0	**46.9**
BiDAF, single (Seo et al., 2016)	75.5	**45.7**
Match-LSTM, ensemble (Wang & Jiang, 2016)	75.4	**41.8**
Match-LSTM, single (Wang & Jiang, 2016)	71.4	**39.0**
Human Performance	92.6	**89.2**

Picking a worst-case sentence

Tadakatsu moved the city of Chicago to in 1881.

Have crowdworkers fix errors

Tadakatsu moved to the city of Chicago in 1881.

Tadakatsu moved to Chicago in 1881.

In 1881, Tadakatsu moved to the city of Chicago.

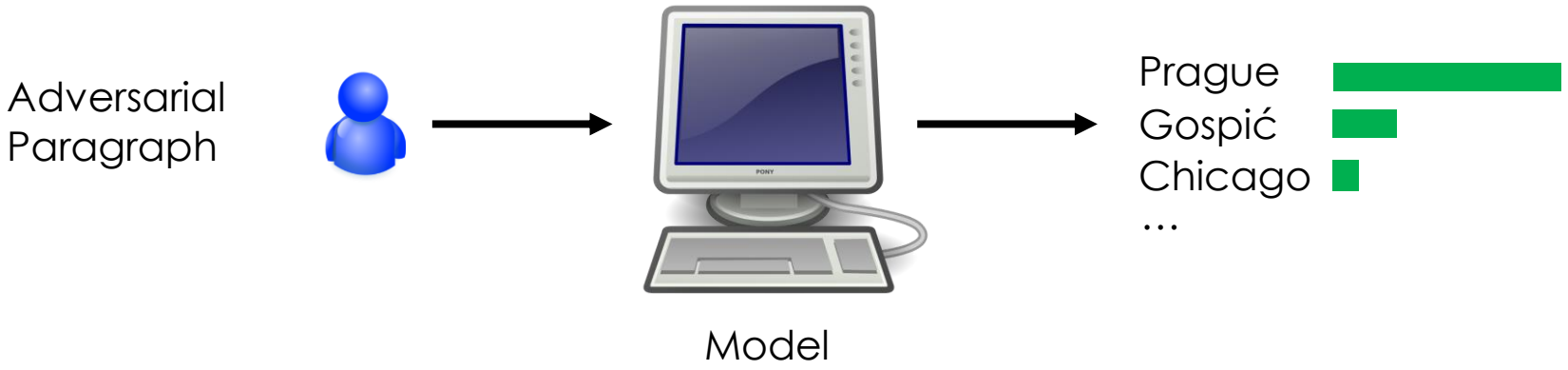
Model failed if distracted by any of these

Results (4 "dev" systems)

System	Original	AddOneSent	AddSent
BiDAF, ensemble (Seo et al., 2016)	80.0	46.9	**34.2**
BiDAF, single (Seo et al., 2016)	75.5	45.7	**34.3**
Match-LSTM, ensemble (Wang & Jiang, 2016)	75.4	41.8	**29.4**
Match-LSTM, single (Wang & Jiang, 2016)	71.4	39.0	**27.3**
Human Performance	92.6	89.2	**79.5**

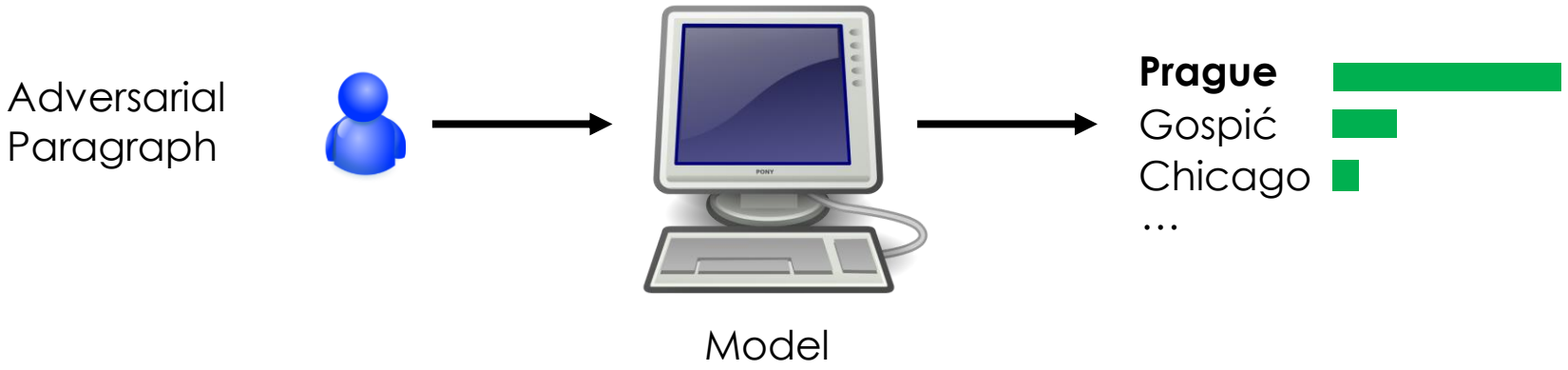
Computers on AddSent

What city did Tesla move to in 1880?



Computers on AddSent

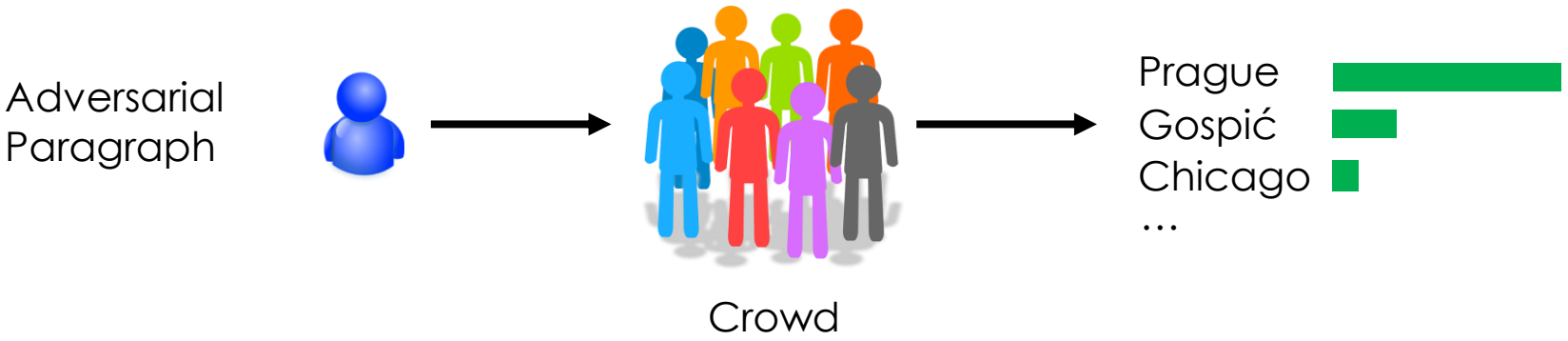
What city did Tesla move to in 1880?



Deterministically choose argmax

Humans on AddSent

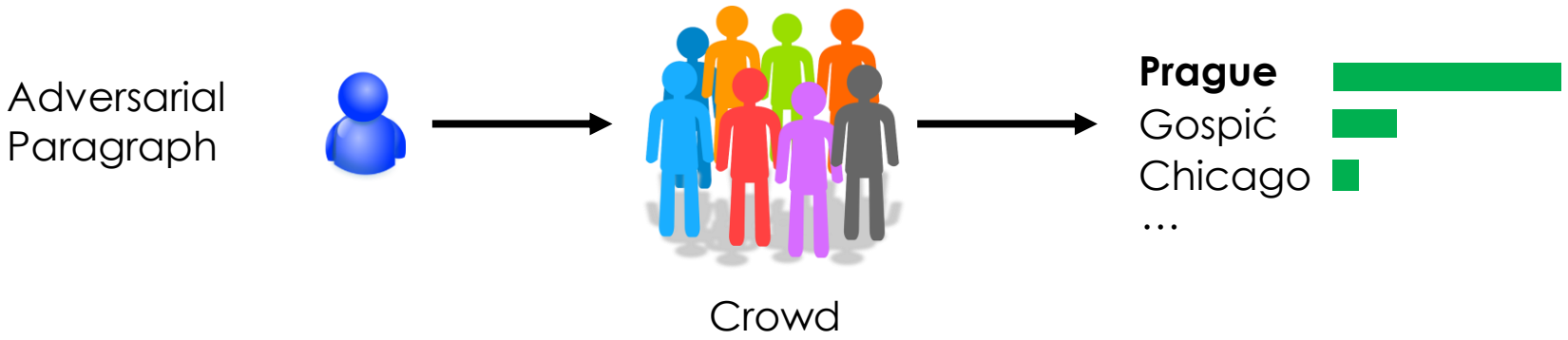
What city did Tesla move to in 1880?



Only get noisy samples!

Humans on AddSent

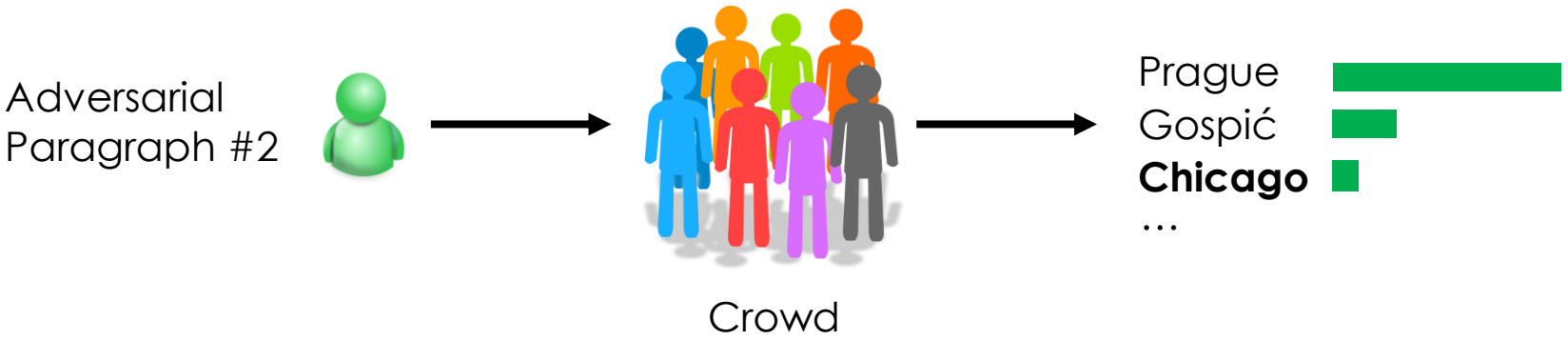
What city did Tesla move to in 1880?



Only get noisy samples!

Humans on AddSent

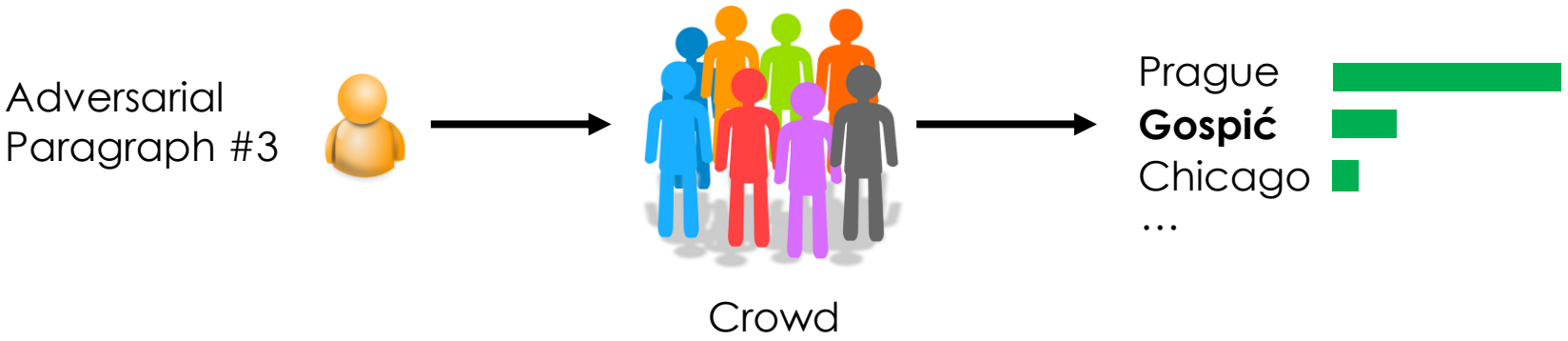
What city did Tesla move to in 1880?



Only get noisy samples!

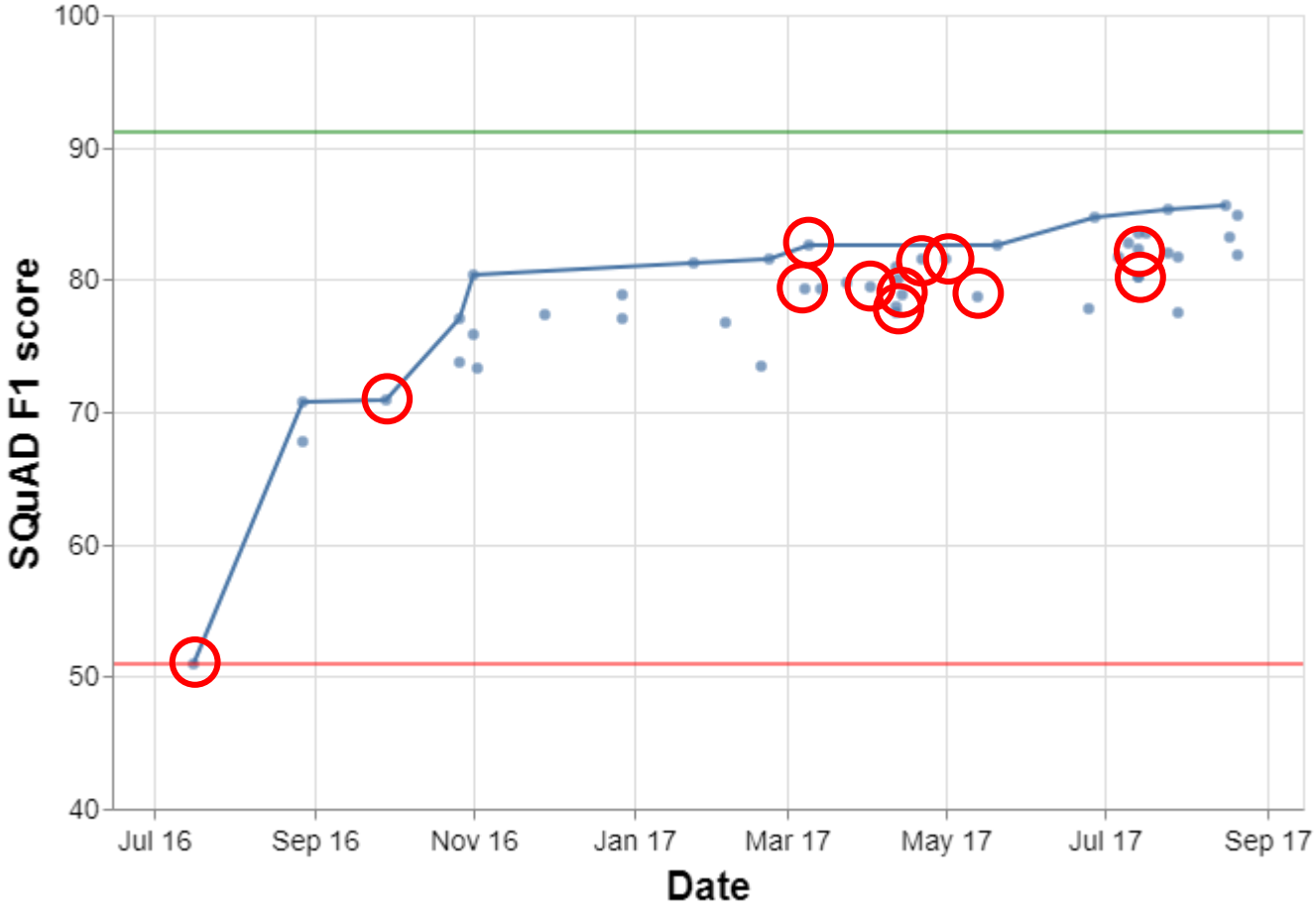
Humans on AddSent

What city did Tesla move to in 1880?



Noise augmented when picking worst-case sentence

Twelve "test" systems



SQuAD leaderboard, https://rajpurkar.github.io/SQuAD-explorer/
*Some of our results are on older versions of models than shown here

Results (12 "test" systems)

System	Original	AddOneSent	AddSent
ReasoNet, ensemble (Shen et al., 2017)	81.1	49.8	39.4
SEDT, ensemble (Liu et al., 2017)	80.1	46.5	35.0
Mnemonic Reader, ensemble (Hu et al., 2017)	79.1	55.3	46.2
Ruminating Reader (Gong and Bowman, 2017)	78.8	47.7	37.4
jNet (Zhang et al., 2017)	78.6	47.0	37.9
Mnemonic Reader, single (Hu et al., 2017)	78.5	56.0	46.6
ReasoNet, single (Shen et al., 2017)	78.2	50.3	39.4
MPCM, single (Wang et al., 2016)	77.0	50.0	40.3
SEDT, single (Liu et al., 2017)	76.9	44.8	33.9
RaSOR (Lee et al., 2016)	76.2	49.5	39.5
DCR (Yu et al., 2016)	69.3	45.1	37.8
Logistic Regression (Rajpurkar et al., 2016)	50.4	30.4	23.2

Results (12 "test" systems)

System	Original	AddOneSent	AddSent
ReasoNet, ensemble (Shen et al., 2017)	81.1	49.8	39.4
SEDT, ensemble (Liu et al., 2017)	80.1	46.5	35.0
Mnemonic Reader, ensemble (Hu et al., 2017)	**79.1**	**55.3**	**46.2**
Ruminating Reader (Gong and Bowman, 2017)	78.8	47.7	37.4
jNet (Zhang et al., 2017)	78.6	47.0	37.9
Mnemonic Reader, single (Hu et al., 2017)	**78.5**	**56.0**	**46.6**
ReasoNet, single (Shen et al., 2017)	78.2	50.3	39.4
MPCM, single (Wang et al., 2016)	77.0	50.0	40.3
SEDT, single (Liu et al., 2017)	76.9	44.8	33.9
RaSOR (Lee et al., 2016)	76.2	49.5	39.5
DCR (Yu et al., 2016)	69.3	45.1	37.8
Logistic Regression (Rajpurkar et al., 2016)	50.4	30.4	23.2

Results (12 "test" systems)

System	Original	AddOneSent	AddSent
ReasoNet, ensemble (Shen et al., 2017)	81.1	49.8	39.4
SEDT, ensemble (Liu et al., 2017)	80.1	46.5	35.0
Mnemonic Reader, ensemble (Hu et al., 2017)	79.1	55.3	46.2
Ruminating Reader (Gong and Bowman, 2017)	78.8	47.7	37.4
jNet (Zhang et al., 2017)	78.6	47.0	37.9
Mnemonic Reader, single (Hu et al., 2017)	78.5	56.0	46.6
ReasoNet, single (Shen et al., 2017)	78.2	50.3	39.4
MPCM, single (Wang et al., 2016)	77.0	50.0	40.3
SEDT, single (Liu et al., 2017)	76.9	44.8	33.9
RaSOR (Lee et al., 2016)	76.2	49.5	39.5
DCR (Yu et al., 2016)	69.3	45.1	37.8
Logistic Regression (Rajpurkar et al., 2016)	**50.4**	**30.4**	**23.2**

Partial Matches

Question: "*The number of new Huguenot colonists declined after what year?*"

Paragraph: "*The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689…but quite a few arrived as late as* ***1700****; thereafter, the numbers declined. The number of old Acadian colonists declined after the year of* ***1675****.*"

All models distracted by sentences with only **partial** match with the question

Partial Matches

Question: "*The number of new Huguenot colonists declined after what year?*"

Paragraph: "*The largest portion of the **Huguenots** to **settle** in the Cape arrived between 1688 and 1689, in seven ships as part of the organised migration, but quite a few arrived as late as **1700**; **thereafter**, the **numbers declined**, and only small groups arrived at a time.*"

Correct Answer: "***1700***"

Stanford Question Answering Dataset (Rajpurkar et al., 2016)

Outline

- Inspiration/Motivation
- Adding Grammatical Sentences
- Adding Word Salad
- Trying to build better systems

Adversarial Word Salad

- So far, only explored tiny fraction of possible distractors
- Try adding **any ungrammatical sequence of words**
  - Incoherent text cannot provide evidence for an incorrect answer

AddAny

What city did Tesla move to in 1880?

> *In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for* ***Prague****…*

Model predicts: ***"Prague"***

Model used: BiDAF Ensemble (Seo et al., 2016)

AddAny

What city did Tesla move to in 1880?

> *In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for* ***Prague****…heavy industry art countries applied design theory even medical process.*

Add random common words

AddAny

What city did Tesla move to in 1880?

> *In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for* ***Prague****…heavy industry art countries applied design* ***theory*** *even medical process.*

Pick one word at random

AddAny

What city did Tesla move to in 1880?

> *In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for* ***Prague****…heavy industry art countries applied design* ***city*** *even medical process.*

Replace with another **common word or question word**, to increase probability that model gives a wrong answer

AddAny

What city did Tesla move to in 1880?

> *In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for* ***Prague****…heavy industry art countries* ***applied*** *design city even medical process.*

Pick one word at random

AddAny

What city did Tesla move to in 1880?

> *In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for* ***Prague****…heavy industry art countries* ***what*** *design city even medical process.*

Replace with another common word or question word, to increase probability that model gives a wrong answer

AddAny

What city did Tesla move to in 1880?

> *In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for* ***Prague****…what 30 city 1880 what move city city medical move.*

Model predicts: ***"medical"***

Model used: BiDAF Ensemble (Seo et al., 2016)

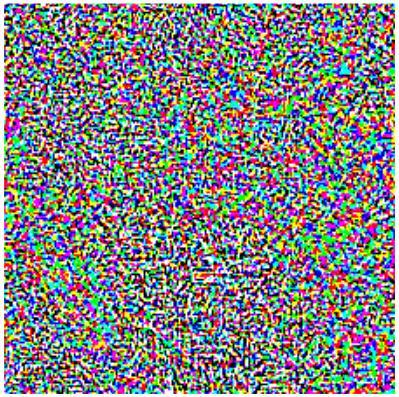
AddAny Results

System	Original	AddOneSent	AddSent	AddAny
BiDAF, ensemble	80.0	46.9	34.2	**2.7**
BiDAF, single	75.5	45.7	34.3	**4.8**
Match-LSTM, ensemble	75.4	41.8	29.4	**11.7**
Match-LSTM, single	71.4	39.0	27.3	**7.6**

Models can be fooled on almost any example

Some Inspiration

+ .007 *



=

Panda
58% confidence

Nematode
8% confidence

Gibbon
99% confidence

Goodfellow et al., 2014.

AddAny

What city did Tesla move to in 1880?

> *In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for* ***Prague****…****what*** *30* ***city 1880 what move city city*** *medical* ***move****.*

Model predicts: *"medical"*

Model used: BiDAF Ensemble (Seo et al., 2016)

AddCommon

What city did Tesla move to in 1880?

> *In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for* ***Prague****…*

Model predicts: ***"Prague"***

Model used: BiDAF Ensemble (Seo et al., 2016)

AddCommon

What city did Tesla move to in 1880?

> *In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for* ***Prague****…heavy industry art countries applied design theory even medical process.*

Add random common words

AddCommon

What city did Tesla move to in 1880?

> *In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for* ***Prague****…heavy industry art countries applied design* ***theory*** *even medical process.*

Pick one word at random

AddCommon

What city did Tesla move to in 1880?

> *In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for* ***Prague****…heavy industry art countries applied design* ***around*** *even medical process.*

Replace with another **common word**, to increase probability that model gives a wrong answer

AddCommon

What city did Tesla move to in 1880?

> *In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for* ***Prague****…finally back would move york hotel through then immediately later.*

Model predicts: ***"york hotel"***

Model used: BiDAF Ensemble (Seo et al., 2016)

AddCommon Results

System	Original	AddOneSent	AddSent	AddAny	AddCommon
BiDAF, ensemble	80.0	46.9	34.2	2.7	**52.6**
BiDAF, single	75.5	45.7	34.3	4.8	**41.7**
Match-LSTM, ensemble	75.4	41.8	29.4	11.7	**51.0**
Match-LSTM, single	71.4	39.0	27.3	7.6	**38.9**

AddCommon Errors

Question: "*What type of markets is the dwelling type below?*"

Distracting text: "*be therefore dark business business other system type feet above.*"

Predicted Answer: "***dark business***"

Model used: BiDAF Ensemble (Seo et al., 2016)

AddCommon Errors

Question: "*After the operators are warned by the escape of the steam, what may they then do?*"

Distracting text: "*came followed after then such then increased hand law may.*"

Predicted Answer: "***increased hand law***"

Model used: BiDAF Ensemble (Seo et al., 2016)

AddCommon Errors

Question: "*Where did he claim the blueprint was stored?*"

Distracting text: "*doubt was did about carried wasn't year 1961 near policy.*"

Predicted Answer: "***near policy***"

Model used: BiDAF Ensemble (Seo et al., 2016)

AddCommon Errors

Question: "*What act sets the term for judging the boundaries of sanity to which individuals wishing to sit on the SP must adhere?*"

Distracting text: "*english our programs industry religion size ran maybe leave poor.*"

Predicted Answer: "***British Nationality Act 1981***"

Model used: BiDAF Ensemble (Seo et al., 2016)

AddCommon Errors

Question: "*What act sets the term for judging the boundaries of sanity to which individuals wishing to sit on the SP must adhere?*"

Paragraph: "*As in the House of Commons, a number of qualifications apply to being an MSP. Such qualifications were introduced under the House of Commons Disqualification Act 1975 and the* ***British Nationality Act 1981****. Specifically, members must be over the age of 18 and must be a citizen of the United Kingdom, the Republic of Ireland, one of the countries in the Commonwealth of Nations, a citizen of a British overseas territory, or a European Union citizen resident in the UK. Members of the police and the armed forces are disqualified from sitting in the Scottish Parliament as elected MSPs, and similarly, civil servants and members of foreign legislatures are disqualified. An individual may not sit in the Scottish Parliament if he or she is judged to be insane under the terms of the* ***Mental Health (Care and Treatment) (Scotland) Act 2003****. english our programs industry religion size ran maybe leave poor.*"

Correct Answer: ***"Mental Health (Care and Treatment) (Scotland) Act 2003"***

Predicted Answer: "***British Nationality Act 1981***"

Outline

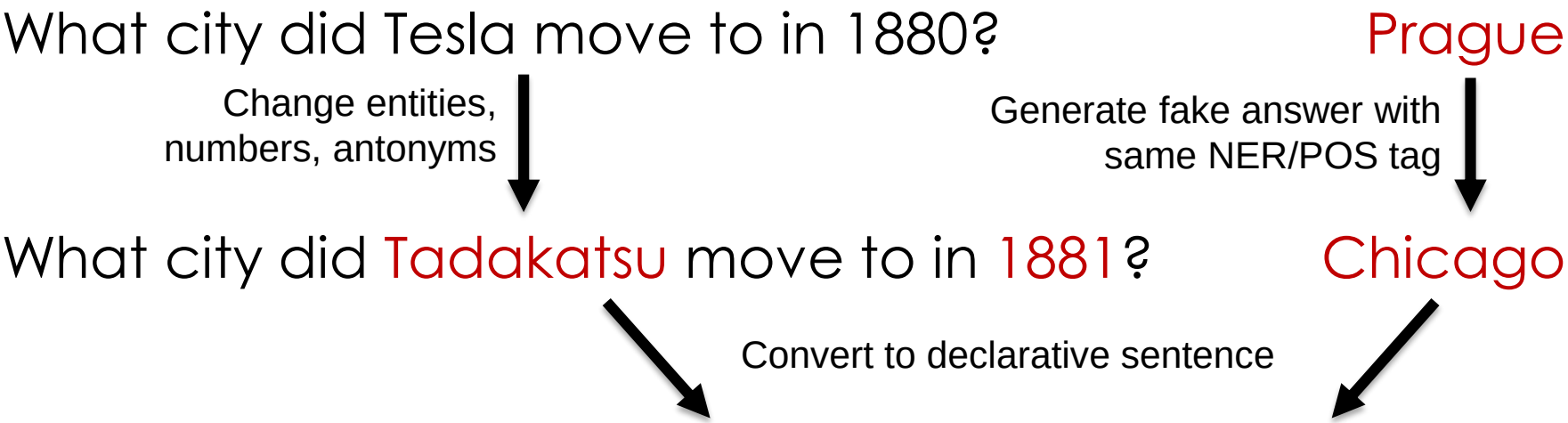
- Inspiration/Motivation
- Adding Grammatical Sentences
- Adding Word Salad
- Trying to build better systems

What can we do?

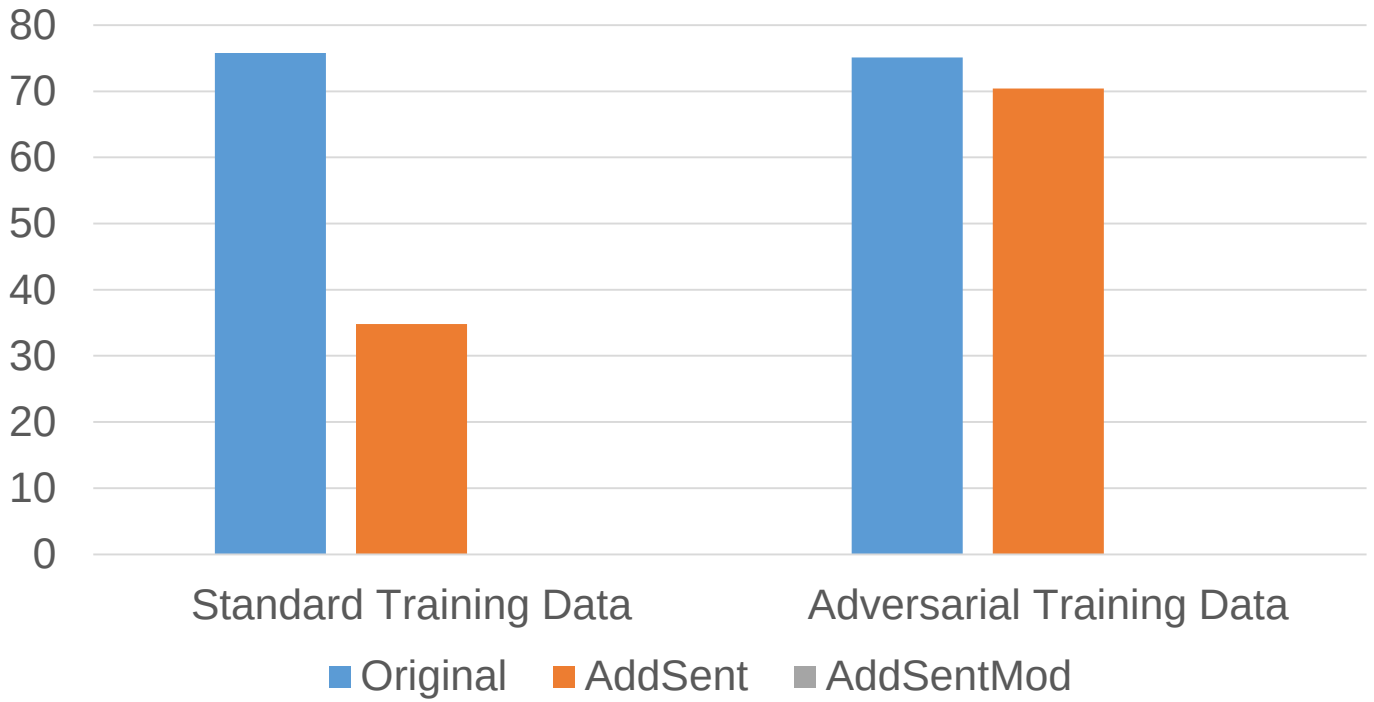
- We've identified weaknesses in existing models—how can we fix them?

Adversarial Training

- What if we train on these adversarial examples?
- Run AddSent without crowdsourcing on training data



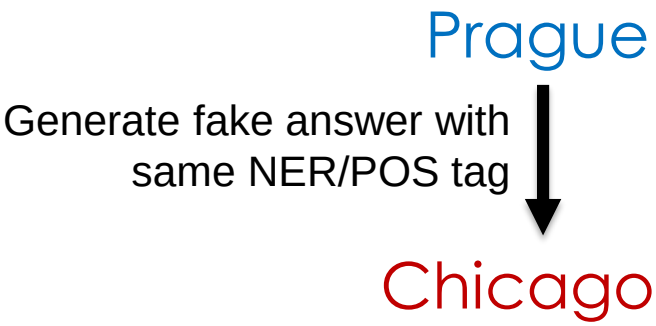
Adversarial Training



Model used: BiDAF Single (Seo et al., 2016)

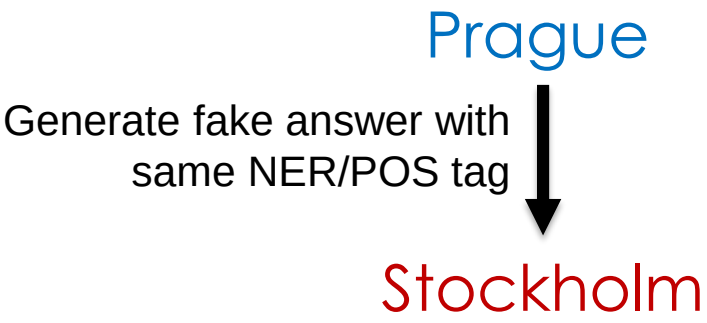
Adversarial Training

- Has the model really learned the right thing?
- Create AddSentMod, similar to AddSent
  - Add sentences to beginning instead of end
  - Use different set of fake answers

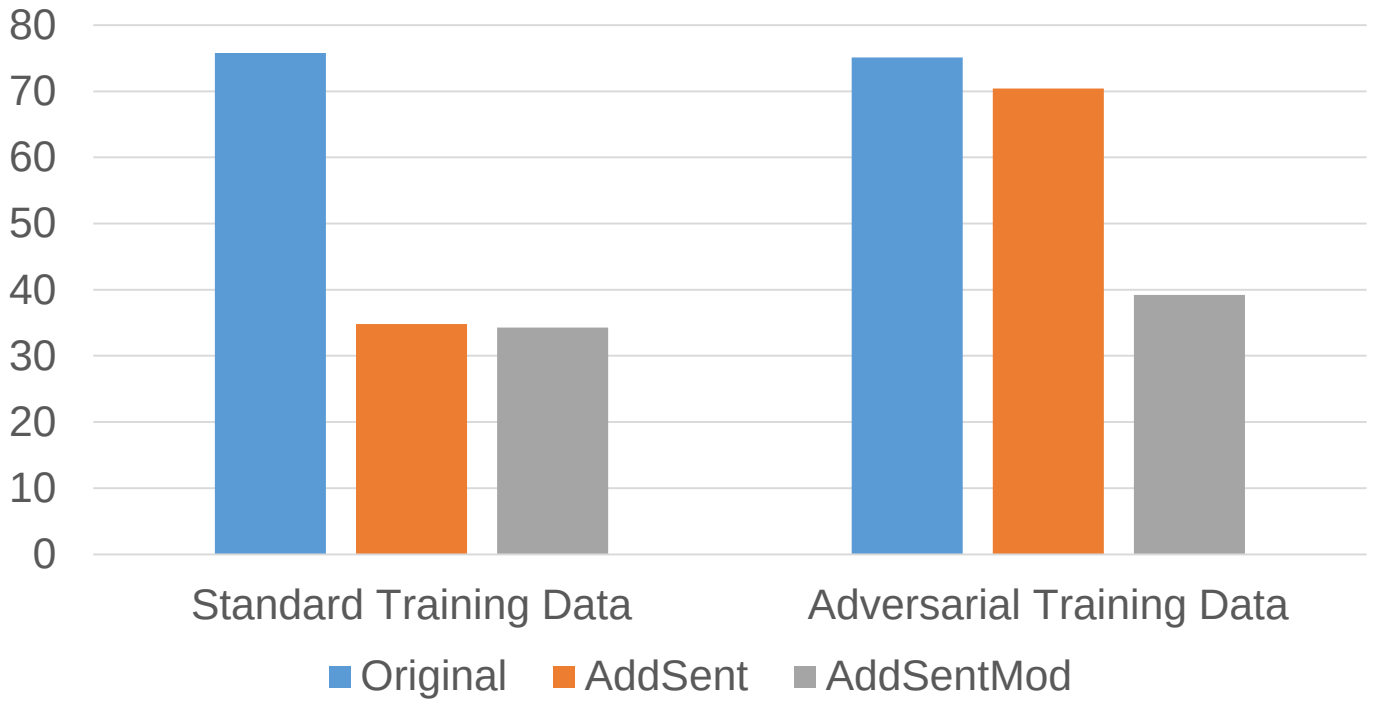


Adversarial Training

- Has the model really learned the right thing?
- Create AddSentMod, similar to AddSent
  - Add sentences to beginning instead of end
  - Use different set of fake answers



Adversarial Training



- Easy to overfit to a given adversary
  - Similar patterns observed with adversarial training in computer vision

Future Work

- Iteratively collect data that's hard for the model as it trains
- Adversary must be general enough so that overfitting not an issue

Thank you!

All code, data, and experiments available on

CodaLab

http://tiny.cc/adversarial-squad-codalab

Thanks to our funding sources!

facebook research

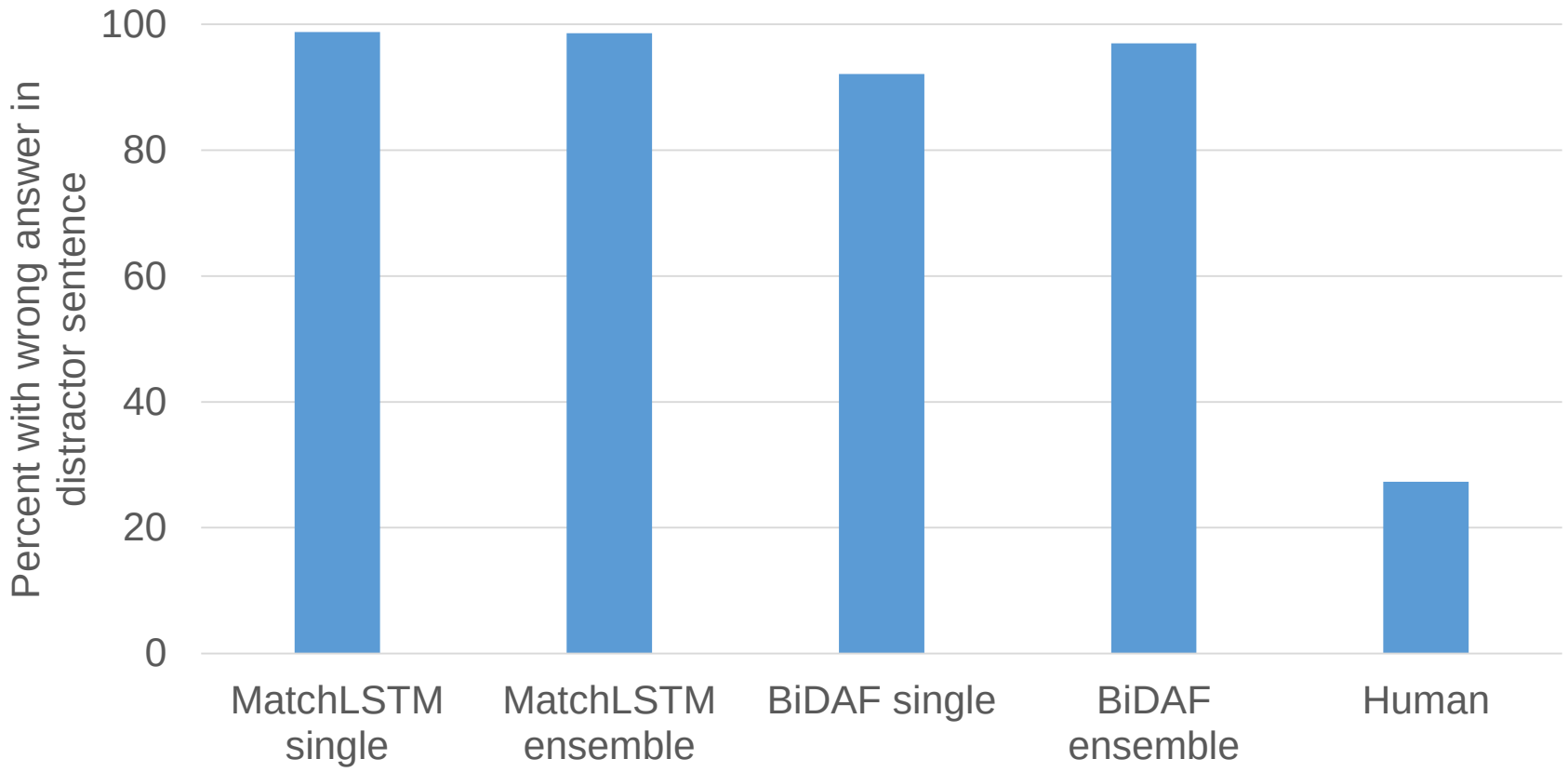
Microsoft

How good are today's systems?

System	SQuAD F1 Score
Interactive AoA Reader, ensemble (HIT + iFLYTEK)	85.3
r-net, ensemble (Microsoft Research Asia)	84.7
r-net, single (Micorosft Research Asia)	83.5
smarnet, ensemble (Eigen Technology & Zhejiang Univ.)	83.5
DCN+, single (Salesforce Research)	82.8
MEMEN, ensemble (Eigen Technology & Zhejiang Univ.)	82.7
ReasoNet, ensemble (Microsoft Research Redmond)	82.6
Mnemonic Reader, ensemble (NUDT & Fudan Univ.)	82.4
Human Performance	91.2

SQuAD leaderboard, https://rajpurkar.github.io/SQuAD-explorer/

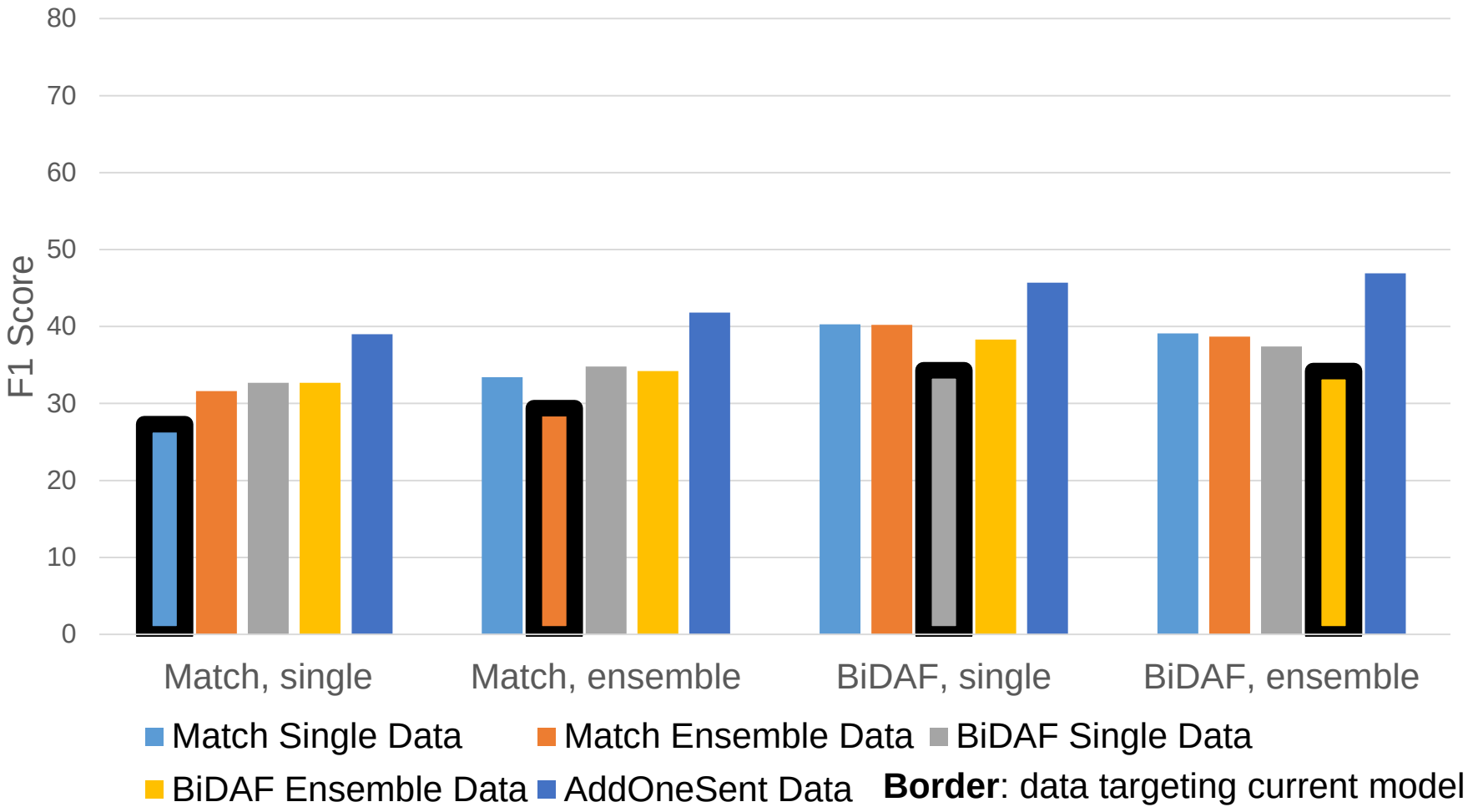
Errors due to distracting text



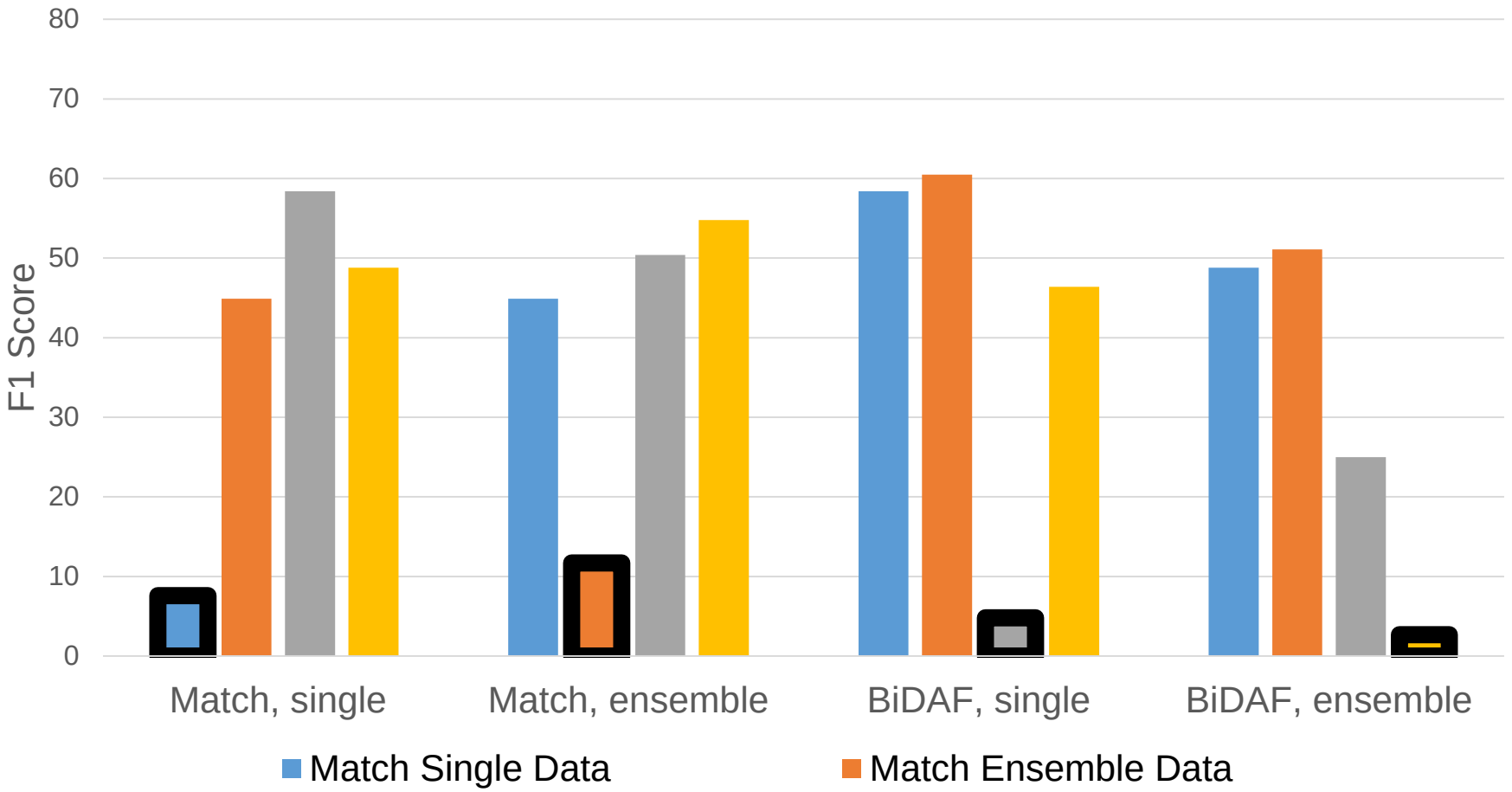
Adversary Generalization

- Do adversarial examples generated to fool one system also fool other systems?

AddSent Generalization



AddAny Generalization



Border: data targeting current model

Conclusion

- Evaluation metrics are important!
- Existing models are deficient in many ways
- Some errors can be explained; others are more unintuitive