Ethics in NLP

Katharina Kann — CSCI/LING5832

Creating and Running Annotation Efforts on MTurk

Creating a Human Intelligence Task

- Specify parameters (specify requirements for which Turkers can complete your HIT)

Creating a Human Intelligence Task

- Specify parameters (specify requirements for which Turkers can complete your HIT)
- Design an HTML template

Creating a Human Intelligence Task

- Specify parameters (specify requirements for which Turkers can complete your HIT)
- Design an HTML template
- Upload a (Microsoft) CSV file to populate the variables

Creating a Human Intelligence Task

- Specify parameters (specify requirements for which Turkers can complete your HIT)
- Design an HTML template
- Upload a (Microsoft) CSV file to populate the variables
- Pre-pay Amazon for the work

Creating a Human Intelligence Task

- Specify parameters (specify requirements for which Turkers can complete your HIT)
- Design an HTML template
- Upload a (Microsoft) CSV file to populate the variables
- Pre-pay Amazon for the work
- Approve/reject work from Turkers

Creating a Human Intelligence Task

- Specify parameters (specify requirements for which Turkers can complete your HIT)
- Design an HTML template
- Upload a (Microsoft) CSV file to populate the variables
- Pre-pay Amazon for the work
- Approve/reject work from Turkers
- Analyze results

Writing Instructions

- Be sure to be as specific as possible in your instructions so that there's no confusion.
    - For example, when asking workers to extract text from an image, ask workers to type the text exactly as shown in the image including capitalizations, spaces and punctuation.

Writing Instructions

- Be sure to be as specific as possible in your instructions so that there's no confusion.
  - For example, when asking workers to extract text from an image, ask workers to type the text exactly as shown in the image including capitalizations, spaces and punctuation.
- Include an example of a right answer, and a wrong answer.

Writing Instructions

- Be sure to be as specific as possible in your instructions so that there's no confusion.
  - For example, when asking workers to extract text from an image, ask workers to type the text exactly as shown in the image including capitalizations, spaces and punctuation.
- Include an example of a right answer, and a wrong answer.
- Clarify what you expect if the HIT is not doable because of missing data or other problems.

Writing Instructions

Reading comprehension test

Please do the following:

- Read the short newspaper article or blog post linked below (if it doesn't load in the frame then open the link in another window)
- Write reading comprehension questions about it
- Give sample answers for each of your questions.

Good reading comprehension questions:

- Ask about why something happened or why someone did something.
- Ask about relationships between people or things.
- Should be answerable in a few words.

Poor reading comprehension questions:

- Ask about numbers or dates.
- Only require a yes/no answer.

http://blogs.wsj.com/digits/2012/11/04/how-the-journal-tested-googles-search-results/

WSJ | WSJ LIVE Live | MARKETWATCH | BARRON'S | PORTFOLIO | DJX | MORE

THE WALL STREET JOURNAL.

Credit: Chris Callison-Burch

In-Class Exercise

- Open the following link and complete the annotation task:
  - Task
- Discuss the following questions in a breakout room:
  - Did you think the task was difficult? Why?
  - How could the instructions be improved?

Crowdsourcing Issues and Ethics

Requesters' Concerns

- Workers may do substandard work or blatantly cheat
  - Cheating by randomly clicking or typing, using scripts to enter useless input, or giving mediocre answers that are not useful

Requesters' Concerns

- Workers may do substandard work or blatantly cheat
  - Cheating by randomly clicking or typing, using scripts to enter useless input, or giving mediocre answers that are not useful
- Can't judge workers' skills or qualifications in advance

Requesters’ Concerns

- Workers may do substandard work or blatantly cheat
  - Cheating by randomly clicking or typing, using scripts to enter useless input, or giving mediocre answers that are not useful
- Can’t judge workers’ skills or qualifications in advance
- Often difficult to judge the quality of work automatically

Crowdworkers’ Concerns

- Pay is not enough

Crowdworkers’ Concerns

- Pay is not enough
- No employment stability or benefits

Crowdworkers’ Concerns

- Pay is not enough
- No employment stability or benefits
- Requesters can exploit workers
  - Not accept work (no pay)
  - Make them do stressful or unethical work

What You Should Do

- Pay at or above minimum wage (currently $11.10/hour in Colorado)

What You Should Do

- Pay at or above minimum wage (currently $11.10/hour in Colorado)
- Disclose your payment practices, your affiliation, the goals of the task, and flag any potentially emotionally distressing portions of your task(s)

What You Should Do

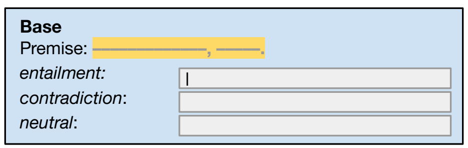
- Pay at or above minimum wage (currently $11.10/hour in Colorado)
- Disclose your payment practices, your affiliation, the goals of the task, and flag any potentially emotionally distressing portions of your task(s)
- Choose platforms that have higher labor standards, and/or enable discussion with workers

Research on Crowdsourcing for NLP

Proposed Strategy: New Protocols and Negative Results for Textual Entailment Data Collection (Bowman et al., 2020)

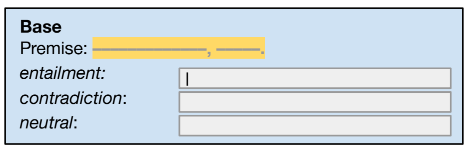
- Goals:
    1. Improving the ease with which annotators can produce sound training examples; or
    2. Improving the quality and diversity of those examples
- Collecting 8.5k-example training sets, 3k validation sets
- Compare to a baseline performance

Proposed Strategy: New Protocols and Negative Results for Textual Entailment Data Collection (Bowman et al., 2020)

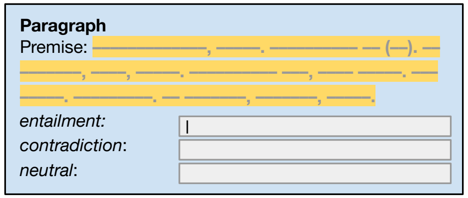


Original

Proposed Strategy: New Protocols and Negative Results for Textual Entailment Data Collection (Bowman et al., 2020)

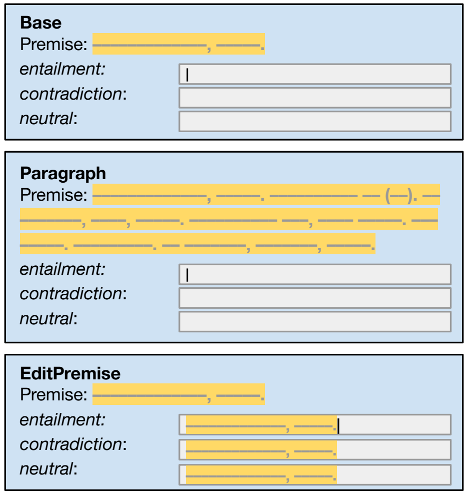


Original



Use a paragraph as premise (more difficult and diverse?)

Proposed Strategy: New Protocols and Negative Results for Textual Entailment Data Collection (Bowman et al., 2020)

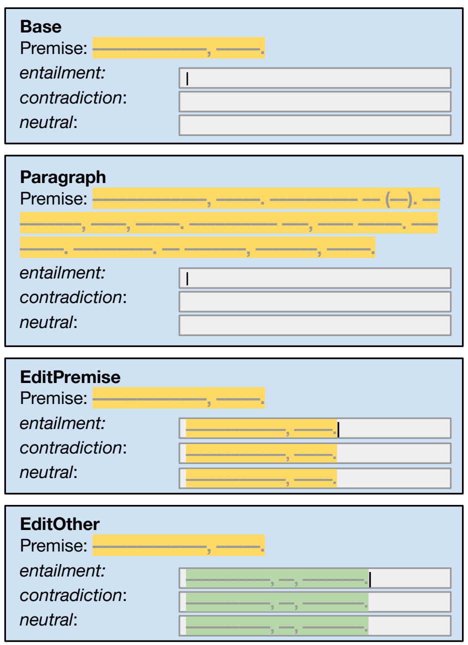


Original

Use a paragraph as premise (more difficult and diverse?)

Prefilling a single seed text (quicker and fewer artifacts?)

Proposed Strategy:
New Protocols and Negative Results for Textual Entailment Data Collection (Bowman et al., 2020)



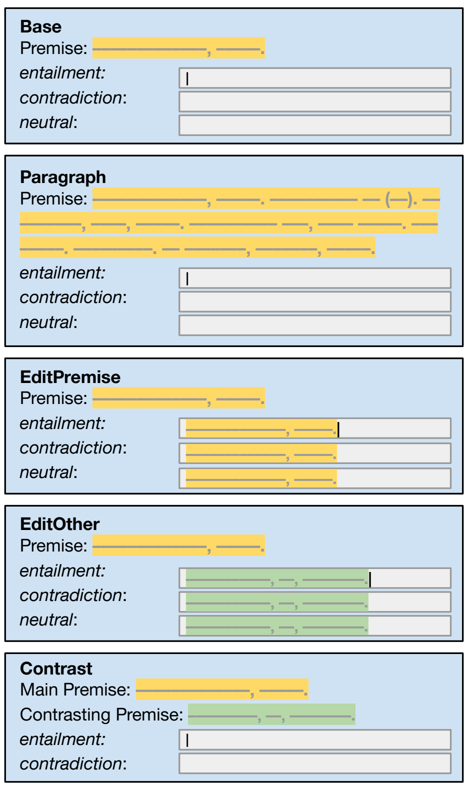
Original

Use a paragraph as premise (more difficult and diverse?)

Prefilling a single seed text (quicker and fewer artifacts?)

Prefilling a single seed text (from a corpus but similar)

Proposed Strategy: New Protocols and Negative Results for Textual Entailment Data Collection (Bowman et al., 2020)



Original

Use a paragraph as premise (more difficult and diverse?)

Prefilling a single seed text (quicker and fewer artifacts?)

Prefilling a single seed text (from a corpus but similar)

Relationship with premise, but not with contrast

Proposed Strategy: New Protocols and Negative Results for Textual Entailment Data Collection (Bowman et al., 2020)

Intermediate-Training Data	Avg. μ (σ)
None	67.3 (1.2)
BASE	**72.2** (0.1)
PARAGRAPH	70.3 (0.1)
EDITPREMISE	69.6 (0.6)
EDITOTHER	70.3 (0.1)
CONTRAST	69.2 (0.0)
MNLI8.5k	71.0 (0.6)
MNLIGov8.5k	70.9 (0.5)
ANLI8.5k	70.5 (0.3)
MNLI	70.0 (0.0)
ANLI	70.4 (0.9)

From Bowman et al. (2020)

Proposed Strategy: New Protocols and Negative Results for Textual Entailment Data Collection (Bowman et al., 2020)

Intermediate-Training Data	Avg. μ (σ)
None	67.3 (1.2)
BASE	**72.2** (0.1)
PARAGRAPH	70.3 (0.1)
EDITPREMISE	69.6 (0.6)
EDITOTHER	70.3 (0.1)
CONTRAST	69.2 (0.0)
MNLI8.5k	71.0 (0.6)
MNLIGov8.5k	70.9 (0.5)
ANLI8.5k	70.5 (0.3)
MNLI	70.0 (0.0)
ANLI	70.4 (0.9)

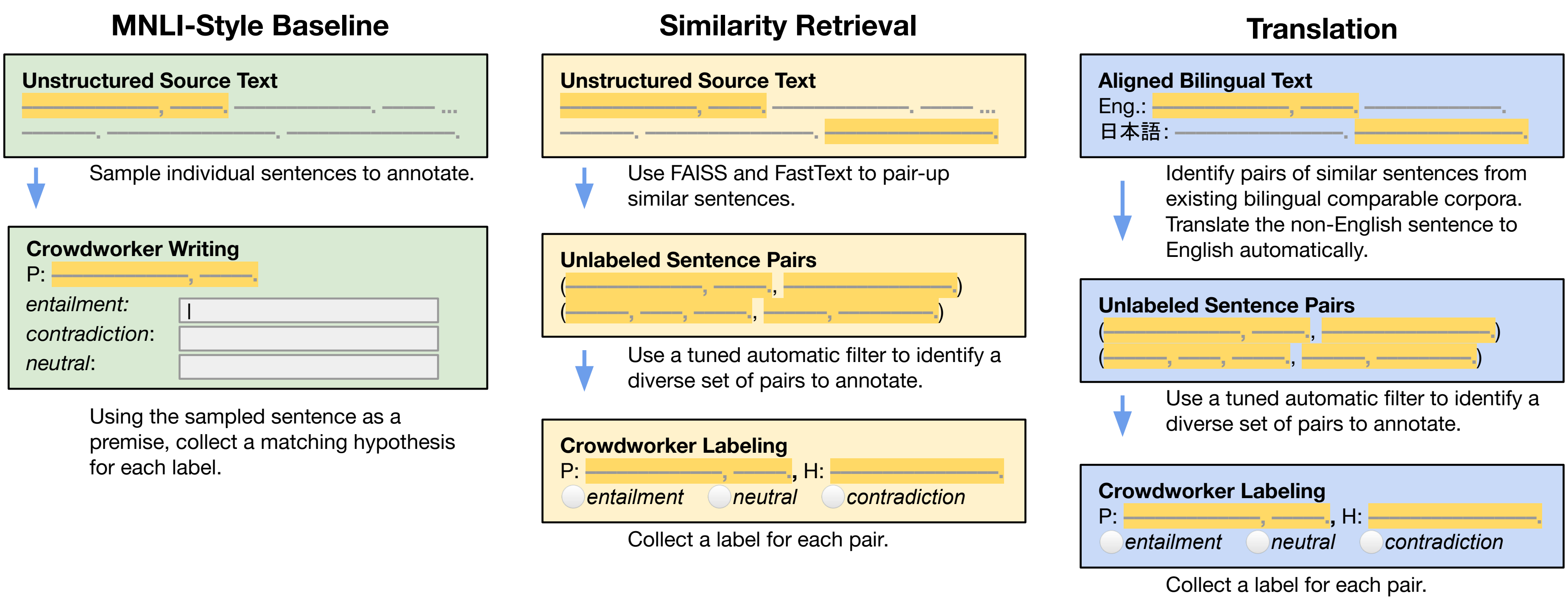
From Bowman et al. (2020)

- **None** is better for transfer learning (intermediate-task training) or generalization
- …but reduce previously known issues with data (hypothesis only)

Proposed Strategy: Asking Crowdworkers to Write Entailment Examples: The Best of Bad Options (Vania et al., 2020)

- Goals:
    1. Increase the speed of annotations; and
    2. Reduce annotation artifacts
- Collecting 3k-example datasets/~6k examples
- Compare to a baseline performance

Proposed Strategy: Asking Crowdworkers to Write Entailment Examples: The Best of Bad Options (Vania et al., 2020)



From Vania et al. (2020)

Proposed Strategy: Asking Crowdworkers to Write Entailment Examples: The Best of Bad Options (Vania et al., 2020)

- **Negative** results on NLI generalization
- **Negative** results for transfer learning
- **Mixed results** with regards to annotation artifacts

Proposed Strategy:
OCNLI: Original Chinese Natural Language Inference
(Hu et al., 2020)

- Goals:
  - Collect a large-scale Chinese NLI dataset
  - Diverse hypotheses
- Collecting ~56k sentence pairs
- No automatic translation!
  - “first human-elicited MNLI-style corpus for a non-English language”

Proposed Strategy: OCNLI: Original Chinese Natural Language Inference (Hu et al., 2020)

- 3 hypotheses per label and premise
    - *Easy*, *medium*, and *hard*
- Explicit instructions and monetary bonus; desiderata: 1) diverse ways of making inferences, and 2) contradictions that do not contain a negator
- Constraint on hypothesis generation: only one out of the three contradictions can contain a negator

Proposed Strategy: OCNLI: Original Chinese Natural Language Inference (Hu et al., 2020)

- 3 hypotheses per label and premise
  - *Easy*, *medium*, and *hard*
- Explicit instructions and monetary bonus; desiderata: 1) diverse ways of making inferences, and 2) contradictions that do not contain a negator
- Constraint on hypothesis generation: only one out of the three contradictions can contain a negator
- The sentences are more challenging
  - …but: more hypothesis-only bias!

Proposed Strategy: OCNLI: Original Chinese Natural Language Inference (Hu et al., 2020)

Subsets	Instructions	# Pairs / Mean length of hypothesis *H* in characters			
		Total	`easy`	`medium`	`hard`
SINGLE	same as MNLI; one *H* per label	11,986 / 10.9	n.a.	n.a.	n.a.
MULTI	three *H*s per label	12,328 / 10.4	4,836 / 9.9	4,621 / 10.6	2,871 / 11.0
MULTIENCOURAGE	MULTI + encouraging annotators to use fewer negators and write more diverse hypotheses	16,584 / 12.2	6,263 / 11.5	6,092 / 12.5	4,229 / 12.7
MULTICONSTRAINT	MULTI + constraints on the negators used in contradictions	15,627 / 12.0	5,668 / 11.6	5,599 / 12.2	4,360 / 12.4
total		56,486 / 11.5			

From Hu et al. (2020)

Proposed Strategy: OCNLI: Original Chinese Natural Language Inference (Hu et al., 2020)

	SINGLE	MULTI	MULTIENC	MULTICON
BERT: fine-tune on XNLI				
dev_full	77.3	73.6	68.6	65.8
easy	na.	74.0	70.1	68.4
medium	na.	74.3	69.6	65.9
hard	na.	72.5	66.2	63.1
RoBERTa: fine-tune on XNLI				
dev_full	78.9	77.3	71.3	70.8
easy	na.	77.2	72.8	73.5
medium	na.	78.6	71.7	70.2
hard	na.	76.2	69.4	68.7

From Hu et al. (2020)

Proposed Strategy: OCNLI: Original Chinese Natural Language Inference (Hu et al., 2020)

	SINGLE	MULTI	MULTIENC	MULTICON
BERT: fine-tune on XNLI				
dev_full	77.3	73.6	68.6	65.8
easy	na.	74.0	70.1	68.4
medium	na.	74.3	69.6	65.9
hard	na.	72.5	66.2	63.1
RoBERTa: fine-tune on XNLI				
dev_full	78.9	77.3	71.3	70.8
easy	na.	77.2	72.8	73.5
medium	na.	78.6	71.7	70.2
hard	na.	76.2	69.4	68.7

From Hu et al. (2020)

Test data	BERT	RoBERTa
OCNLI_dev	65.3	65.7
OCNLI_test	64.3	65.0
OCNLI_test_easy	63.5	64.0
OCNLI_test_medium	63.9	65.6
OCNLI_test_hard	65.5	65.5
MNLI	na.	62.0

Hypothesis only; from Hu et al. (2020)

Ethics in NLP

Bias in Word Embeddings

Pitfalls of Unsupervised Learning

Word similarity:

- Occupations most similar to *she*:
  - *nurse*, *librarian, nanny, stylist, dancer*
- Occupations most similar to *he*:
  - *architect, captain, philosopher, legend, hero*

Source: Bolukbasi et al. '16, Quantifying and Reducing Stereotypes in Word Embeddings

Pitfalls of Unsupervised Learning

Word analogy:

- doctor - father + mother: nurse

Source: [Bolukbasi et al. ‘16](), Quantifying and Reducing Stereotypes in Word Embeddings

Pitfalls of Unsupervised Learning

Additionally:

- African American names have a higher cosine similarity with unpleasant words.
- European American names (‘Brad’, ‘Greg’, ‘Courtney’) have a higher cosine similarity with pleasant words.

Source: Bolukbasi et al. ‘16, Quantifying and Reducing Stereotypes in Word Embeddings

Pitfalls of Unsupervised Learning

Impossible to avoid these issues altogether when learning from naturally occurring text.

Mitigating bias will usually require identifying explicitly, and the best method will depend on the task at hand.

Source: Bolukbasi et al. '16, Quantifying and Reducing Stereotypes in Word Embeddings

Ethical Issues for Publishing in NLP

A General Issue in Machine Learning

“Instead of relying on algorithms, which we can be accused of manipulating for our benefit, we have turned to machine learning, an ingenious way of disclaiming responsibility for anything. Machine learning is like money laundering for bias. It's a clean, mathematical apparatus that gives the status quo the aura of logical inevitability. The numbers don't lie.”

- Maciej Cegłowski

Data and Unwanted Biases

- Machine learning methods generally pick up on biases that are present in training data.

Data and Unwanted Biases

- Machine learning methods generally pick up on biases that are present in training data.
- Deploying biased models in the wrong places can lead to harms far worse than bad user experiences.
    - Résumé screening, exam scoring, predictive policing...

Data and Unwanted Biases

- Machine learning methods generally pick up on biases that are present in training data.
- Deploying biased models in the wrong places can lead to harms far worse than bad user experiences.
    - Résumé screening, exam scoring, predictive policing...
- Some ML techniques can amplify biases in data.
    - Zhao et al. reading on multi-label image classifiers:
        - In training data, women appear in cooking scenes 33% more often than men.
        - In model's labeling of similar test data, women are detected in cooking scenes 68% more often than men.

Data and Unwanted Biases

10.10.18

Amazon's hiring AI may have weeded out women: Report

[Photo: Flickr user Tony Webster]

Data and Unwanted Biases

- Model de-biasing can be complex, political, and impossible to do fully…

Data and Unwanted Biases

- Model de-biasing can be complex, political, and impossible to do fully…
- … and it may harm performance on reasonable metrics.
  - Why?

Data and Unwanted Biases

- Model de-biasing can be complex, political, and impossible to do fully…
- … and it may harm performance on reasonable metrics.
    - Why?
- Many issues won't be obvious at first: Look to domain experts in the application areas you're working on for advice on issues to watch for.

A Related Issue: Exclusion

- Colloquial African-American English isn't well represented in training data for language identification, parsing, etc., so technologies like translation and intelligent assistants aren't as usable for its speakers.
    - Users are forced to choose between avoiding their preferred dialect or missing out on the benefits of the technology.

A Related Issue: Exclusion

- Colloquial African-American English isn't well represented in training data for language identification, parsing, etc., so technologies like translation and intelligent assistants aren't as usable for its speakers.
  - Users are forced to choose between avoiding their preferred dialect or missing out on the benefits of the technology.
- Similar situation for English varieties in India, Nigeria, Philippines, Singapore, Caribbean, etc., and for regional/minority languages in general.
  - Note: Regional or informal dialects of a language are generally just as standardized, just as complex, and just as easy or hard to model as the standard forms of the language.
- See Blodgett and O'Connor (2017)

Talk About Data

Easy steps to avoid problems stemming from biased or unrepresentative data:

- When writing up NLP research, be clear about:

Talk About Data

Easy steps to avoid problems stemming from biased or unrepresentative data:

- When writing up NLP research, be clear about:
    - What your data looks like, why it was collected, and what kind of information your system learns from it.

Talk About Data

Easy steps to avoid problems stemming from biased or unrepresentative data:

- When writing up NLP research, be clear about:
    - What your data looks like, why it was collected, and what kind of information your system learns from it.
    - Who (country, region, gender, native language, etc.) produced the text and labels in your dataset(s).

Talk About Data

Easy steps to avoid problems stemming from biased or unrepresentative data:

- When writing up NLP research, be clear about:
    - What your data looks like, why it was collected, and what kind of information your system learns from it.
    - Who (country, region, gender, native language, etc.) produced the text and labels in your dataset(s).
    - Any known biases in your dataset(s) (including obvious ones).

Talk About Data

Easy steps to avoid problems stemming from biased or unrepresentative data:

- When writing up NLP research, be clear about:
    - What your data looks like, why it was collected, and what kind of information your system learns from it.
    - Who (country, region, gender, native language, etc.) produced the text and labels in your dataset(s).
    - Any known biases in your dataset(s) (including obvious ones).
- This is especially important when writing for nontechnical potential users/clients.

Talk About Data

Easy steps to avoid problems stemming from biased or unrepresentative data:

- When writing up NLP research, be clear about:
  - What your data looks like, why it was collected, and what kind of information your system learns from it.
  - Who (country, region, gender, native language, etc.) produced the text and labels in your dataset(s).
  - Any known biases in your dataset(s) (including obvious ones).
- This is especially important when writing for nontechnical potential users/clients.
- When possible, build useful confidence metrics.
  - Notify the user when the system is out of its comfort zone.

More Issues

- We’ve been talking about ways to avoid unintentional harms. NLP technologies can also be used intentionally for ethically problematic applications:
  - Communication monitoring by repressive governments
  - Removal of political speech on online platforms
  - Filtering of communication in minority dialects/ languages

More Issues

- We’ve been talking about ways to avoid unintentional harms. NLP technologies can also be used intentionally for ethically problematic applications:
  - Communication monitoring by repressive governments
  - Removal of political speech on online platforms
  - Filtering of communication in minority dialects/ languages
- Most NLP technologies have some potential to do harm, even if their most obvious use case is harmless and worthwhile. These are dual use technologies. Be aware of what the harmful uses are, and do what you can to avoid supporting them.

Important Remarks

- If you’re presented with a clearly unethical (or illegal) application: Just don’t do it.
    - The NLP job market is remarkably strong: There are plenty of other employers (or research funders) out there.
    - A public scandal can easily end a career.

Crowdsourcing Issues and Ethics

Requesters' Concerns

- Workers may do substandard work or blatantly cheat
  - Cheating by randomly clicking or typing, using scripts to enter useless input, or giving mediocre answers that are not useful

Requesters’ Concerns

- Workers may do substandard work or blatantly cheat
  - Cheating by randomly clicking or typing, using scripts to enter useless input, or giving mediocre answers that are not useful
- Can’t judge workers’ skills or qualifications in advance

Requesters’ Concerns

- Workers may do substandard work or blatantly cheat
  - Cheating by randomly clicking or typing, using scripts to enter useless input, or giving mediocre answers that are not useful
- Can’t judge workers’ skills or qualifications in advance
- Often difficult to judge the quality of work automatically

Crowdworkers’ Concerns

- Pay is not enough

Crowdworkers’ Concerns

- Pay is not enough
- No employment stability or benefits

Crowdworkers’ Concerns

- Pay is not enough
- No employment stability or benefits
- Requesters can exploit workers
  - Not accept work (no pay)
  - Make them do stressful or unethical work

What You Should Do

- Pay at or above minimum wage (currently $11.10/hour in Colorado)

What You Should Do

- Pay at or above minimum wage (currently $11.10/hour in Colorado)
- Disclose your payment practices, your affiliation, the goals of the task, and flag any potentially emotionally distressing portions of your task(s)

What You Should Do

- Pay at or above minimum wage (currently $11.10/hour in Colorado)
- Disclose your payment practices, your affiliation, the goals of the task, and flag any potentially emotionally distressing portions of your task(s)
- Choose platforms that have higher labor standards, and/or enable discussion with workers

IRB and Human Subjects (Slides inspired by Tsvetkov and Black's slides)

History of using Human Subjects

- WWII Nazi and Japanese prisoners in concentration camps
  - Medical science did learn things
  - But even at the time this was not considered acceptable
- Tuskegee Syphilis Experiments
- Stanford Prison Experiment
- National Research Act of 1974

Tuskegee Syphilis Experiment

- Understand how untreated syphilis develops
- US Public Health System 1932-1972
- Rural African-American sharecroppers, Macon Co, Alabama
  - 399 already had syphilis
  - 201 not infected
- Given free health care, meals and burial service
- Not provided with penicillin when it would have helped
  - (Though not known at the start of the experiment)
- Peter Buxton, whistleblower, 1972

Stanford Prison Experiment

- Philip Zimbardo, Stanford University, August 1971
- Test how perceived power affects subjects
- Groups arbitrarily split in two
  - One group were defined “prisoners”
  - One group were defined “guards”
- “Guards” selected uniforms, and defined discipline

Ethics in Human Subject Use

- These experiments (especially the Tuskegee Experiment) led to the National Research Act 1974
  - Requiring “Informed Consent” from participants
  - Requiring external review of experiments
  - For all federal funded experiments

IRB (Ethical Review Board)

- Institutional Review Board
  - Internal to institution
  - Independent of researcher

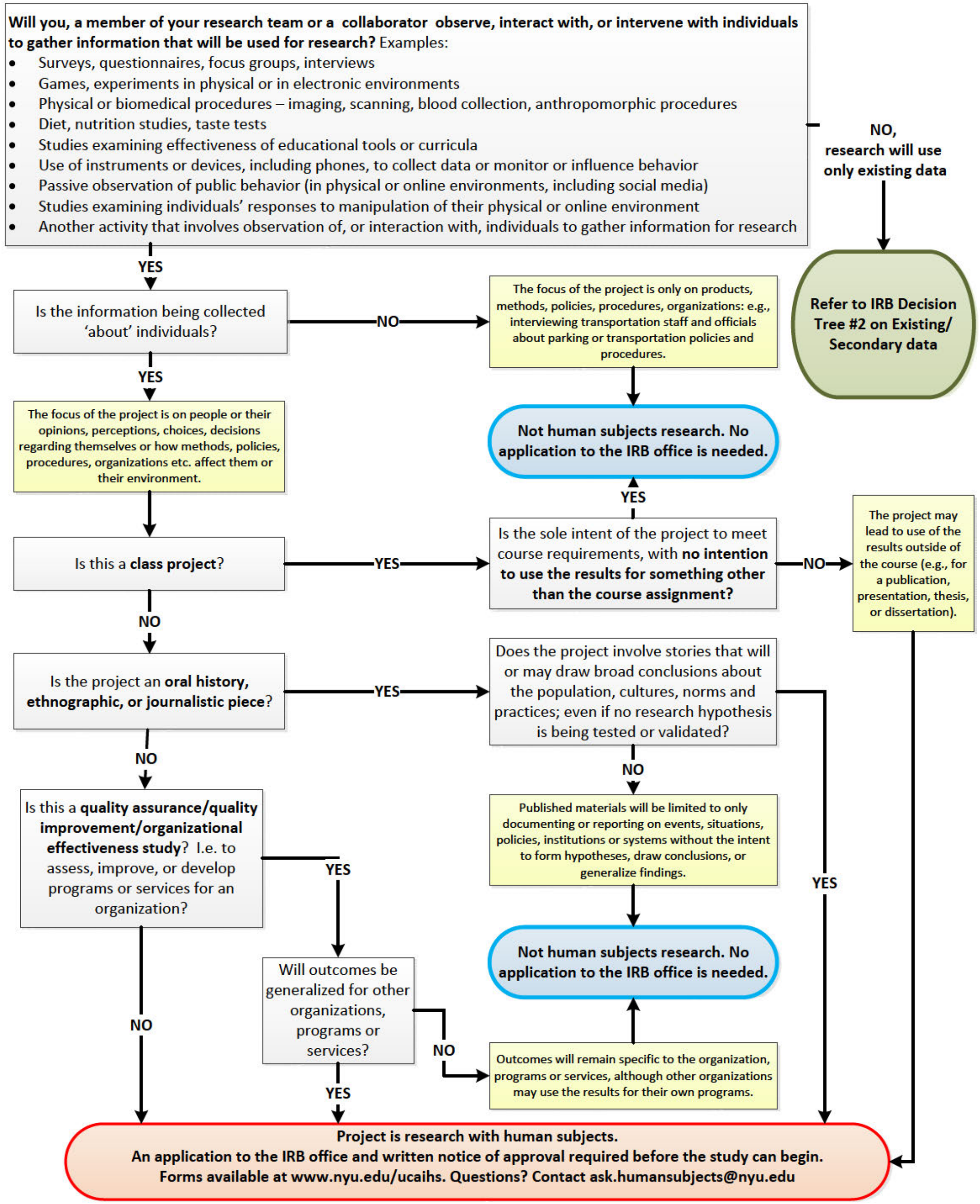
IRB (Ethical Review Board)

- Institutional Review Board
  - Internal to institution
  - Independent of researcher
- Reviews all human experimentation
  - Assesses instructions
  - Compensation
  - Contribution of research
  - Value to the participant
  - Protection of privacy

IRB (Ethical Review Board)

- Most IRB have special requirements for involving
  - Minors, pregnant women, disabled
- So most experiments exclude these
- Protected or hard to access groups are underrepresented (we will discuss this later)

Does Your Project Require an Application to the NYU IRB Office?
Decision Tree #1



Ethical Questions

- Can you lie to a human subject?
- Can you harm a human subject?
- Can you mislead a human subject?

Ethical Questions

- Can you lie to a human subject?
- Can you harm a human subject?
- Can you mislead a human subject?

- What about Wizard of Oz experiments?
- What about gold standard data?

Ethical Issues in NLP

Terminology from Ethics of Technology (Hovy and Spruit, 2016)

- Exclusion
- Overgeneralization
- Topic under- and overexposure
- Dual use

Exclusion (Hovy and Spruit, 2016)

As a result of the situatedness of language, any data set carries a **demographic bias**, i.e., latent information about the demographics in it. Overfitting to these factors can have have severe effects on the applicability of findings. In psychology, where most studies are based on western, educated, industrialized, rich, and democratic research participants (so-called WEIRD, Henrich et al. (2010)), the tacit assumption that human nature is so universal that findings on this group would translate to other demographics has led to a heavily biased corpus of psychological data. In NLP, overfitting to the demographic bias in the training data is due to the *i.i.d.* assumption. I.e., models implicitly assume all language to be identical to the training sample. They therefore perform worse or even fail on data from other demographics.

Overgeneralization (Hovy and Spruit, 2016)

Exclusion is a side-effect of the data. **Overgeneralization** is a modeling side-effect.

As an example, we consider automatic inference of user attributes, a common and interesting NLP task, whose solution also holds promise for many useful applications, such as recommendation engines and fraud or deception detection (Badaskar et al., 2008; Fornaciari and Poesio, 2014; Ott et al., 2011; Banerjee et al., 2014).

The cost of false positives seems low: we might be puzzled or amused when receiving an email addressing us with the wrong gender, or congratulating us to our retirement on our 30th birthday.

In practice, though, relying on models that produce false positives may lead to bias confirmation and overgeneralization. Would we accept the same error rates if the system was used to predict sexual orientation or religious views, rather than age or gender? Given the right training data, this is just a matter of changing the target variable.

Topic Under- and Overexposure

- A problem with the research design

Topic Under- and Overexposure

- A problem with the research design
- If language by certain minority groups was harder to process, this group could be perceived as difficult or abnormal

Topic Under- and Overexposure

- A problem with the research design
- If language by certain minority groups was harder to process, this group could be perceived as difficult or abnormal
- Underexposure to certain language’s data makes working on those even more difficult

Dual Use

- Developed NLP technologies can negatively affect people's lives without being made for that intentionally

Dual Use

- Developed NLP technologies can negatively affect people's lives without being made for that intentionally
- Examples:
  - NLP can be used to detect fake news, but also to create them
  - Text classification can help understand slang, but also be used for censorship

Ethics in NLP

- In the last years, ethical issues in NLP have started to receive a growing amount of attention

Ethics in NLP

- In the last years, ethical issues in NLP have started to receive a growing amount of attention
- Multiple workshops on the topic:
  - Ethics in NLP
  - Workshop on Abusive Language Online
  - Fairness, Accountability, and Transparency in Machine Learning
  - …

In-Class Exercise 2

- You will be randomly assigned to breakout rooms; each room will be assigned to one of the before mentioned workshops
- Find the proceedings of your workshop
- In your groups, discuss which paper you find the most interesting
  - Read and discuss it in your rooms
  - Prepare to present it to the entire class afterwards

In-Class Exercise 2

20 minutes!

- You will be randomly assigned to breakout rooms; each room will be assigned to one of the before mentioned workshops
- Find the proceedings of your workshop
- In your groups, discuss which paper you find the most interesting
  - Read and discuss it in your rooms
  - Prepare to present it to the entire class afterwards

Bias in NLP

- One of the first papers on bias in NLP was Hovy and Spruit (2016)
- Goal was to get researchers to discuss the topic more (as opposed to the media)
- They identified 3 sources of bias
  - We will see them in the next slides
  - (This is only one possible categorization)

Problems in the Data

- Certain groups are not represented in the data
- Raw text contains all sorts of biases
  - Leads to biases in word embeddings
  - Leads to biases in language models
  - …

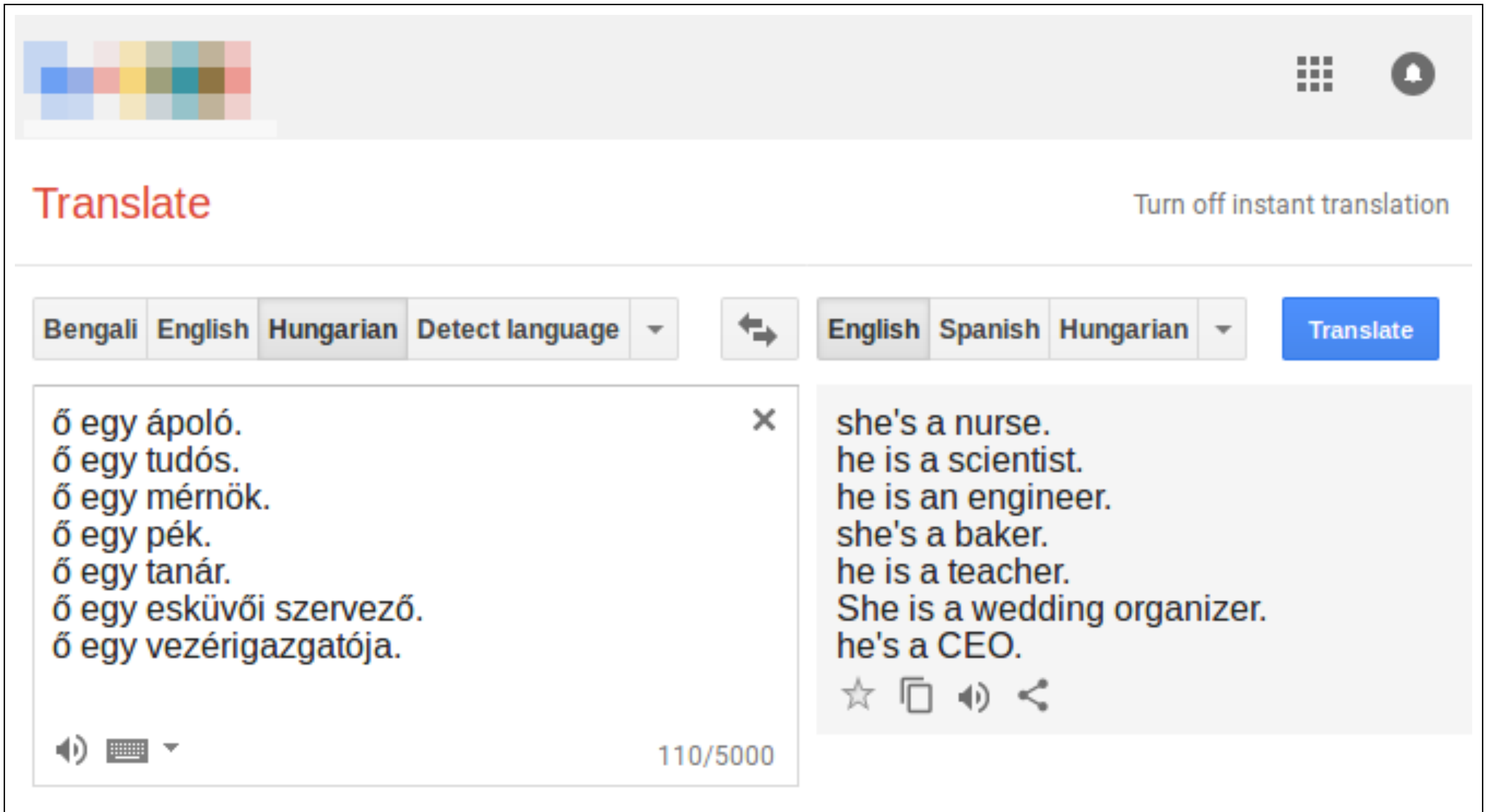
Problems in the Models

- Models exaggerate existing bias in data
- Models might work better or worse for certain datasets/languages

Problems in the Research Design

- Certain languages are studied more than others
- Certain groups are studied more than others

Bias in Machine Translation



Prates et al., 2019

Other Possible Issues

- Privacy issues (using of data, storing of data, sending of data, etc.)

Other Possible Issues

- Privacy issues (using of data, storing of data, sending of data, etc.)
- What information is ethical to infer from user data?

Debiasing

How to Reduce Bias in Data

- Controlling for biases of the annotators:
  - Age, gender, etc.
  - Spoken languages
  - Native language
  - …

How to Reduce Bias in Data

- Downsampling overly represented classes
    - However, this reduces the amount of available training instances!

How to Reduce Bias in Data

- Downsampling overly represented classes
  - However, this reduces the amount of available training instances!
- Reweighting of training instances
  - Based on features like gender or age

How to Reduce Bias in Data

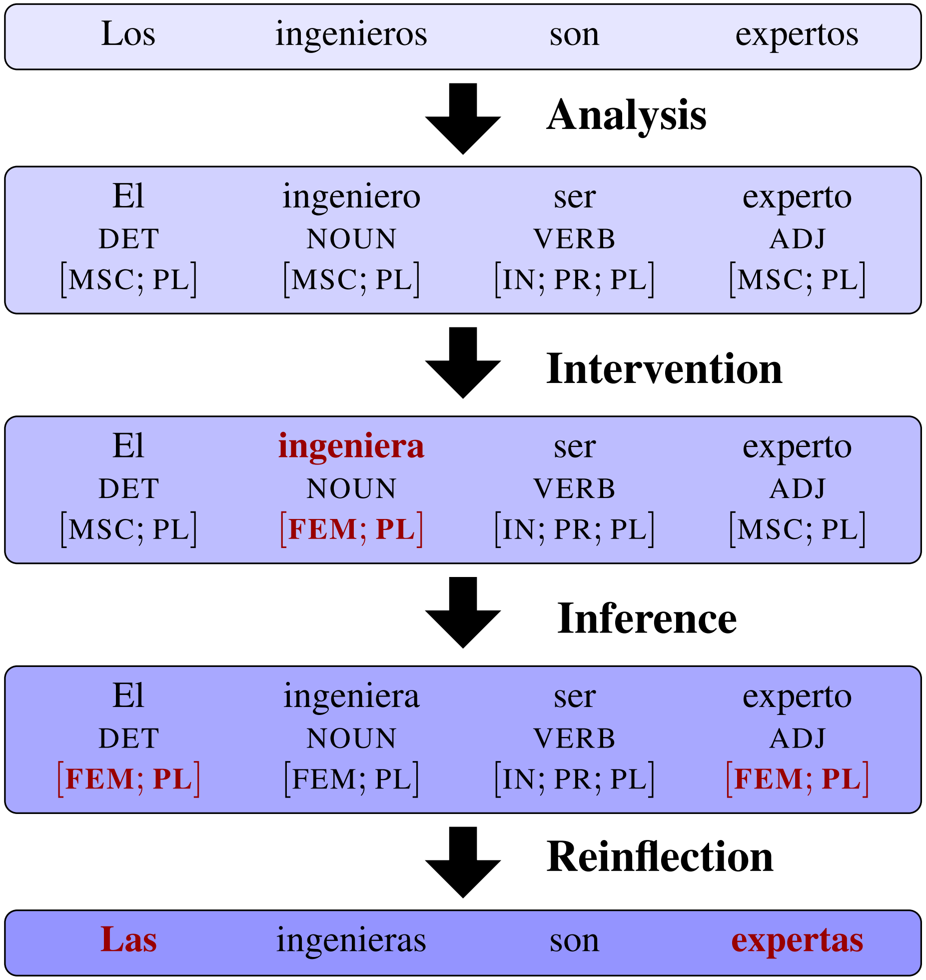
- Downsampling overly represented classes
  - However, this reduces the amount of available training instances!
- Reweighting of training instances
  - Based on features like gender or age
- Combinations are possible, too

How to Reduce Bias in Data

- Data augmentation (Zmigrod et al., 2019):
  - Add slightly altered examples to data to counteract bias

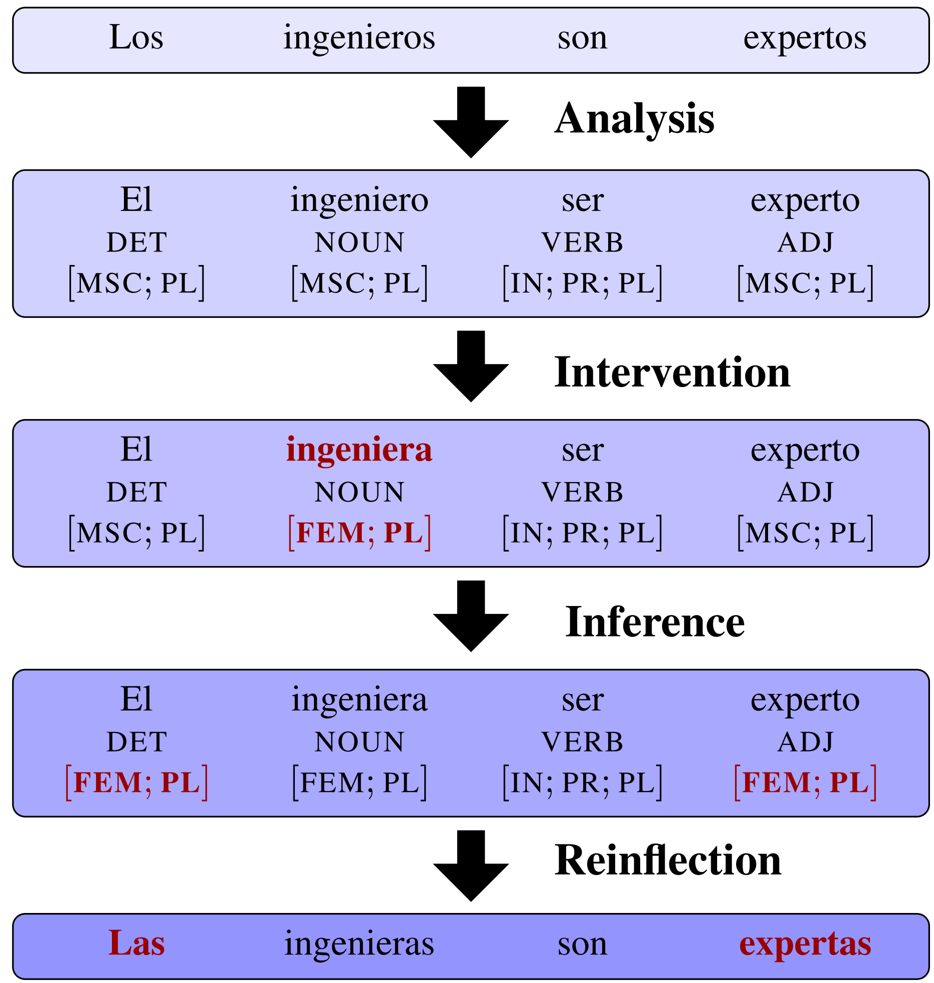
How to Reduce Bias in Data

- Data augmentation (Zmigrod et al., 2019):
  - Add slightly altered examples to data to counteract bias



How to Reduce Bias in Data

- Data augmentation (Zmigrod et al., 2019):
    - Add slightly altered examples to data to counteract bias
- Similar: Rudinger et al. (2018); Zhao et al. (2018)



How to Reduce Bias in Models

- Li et al. (2018):
  - Use an adversarial multi-task learning setup
  - Reverse gradients for tasks that consist of predicting demographics
  - Model learns to ignore demographics

NLP against Unethical Behavior

Fake News Detection

- From the FEVER workshop:
  - “With billions of individual pages on the web providing information on almost every conceivable topic, we should have the ability to collect facts that answer almost every conceivable question. However, only a small fraction of this information is contained in structured sources (Wikidata, Freebase, etc.) – we are therefore limited by our ability to transform free-form text to structured knowledge. There is, however, another problem that has become the focus of a lot of recent research and media coverage: false information coming from unreliable sources.”

Fake News Detection

- Fake news detection (FEVER task definition)
  - Verify information using evidence, e.g., from Wikipedia

Fake News Detection

- Fake news detection (FEVER task definition)
  - Verify information using evidence, e.g., from Wikipedia
  - Given a factual claim involving one or more entities (resolvable to Wikipedia pages), extract textual evidence (sets of sentences from Wikipedia pages) that support or refute the claim

Fake News Detection

- Fake news detection (FEVER task definition)
  - Verify information using evidence, e.g., from Wikipedia
  - Given a factual claim involving one or more entities (resolvable to Wikipedia pages), extract textual evidence (sets of sentences from Wikipedia pages) that support or refute the claim
  - Using this evidence, label the claim as Supported, Refuted given the evidence or NotEnoughInfo

Fake News Detection

- Fake news detection (FEVER task definition)
    - Verify information using evidence, e.g., from Wikipedia
    - Given a factual claim involving one or more entities (resolvable to Wikipedia pages), extract textual evidence (sets of sentences from Wikipedia pages) that support or refute the claim
    - Using this evidence, label the claim as Supported, Refuted given the evidence or NotEnoughInfo
    - A claim's evidence may consist of multiple sentences that only if examined together provide the stated label

Fake News Detection

Data format (*=appears also in test data):

- **id***: The ID of the claim
- **label**: The annotated label for the claim. Can be one of SUPPORTS | REFUTES | NOT ENOUGH INFO
- **claim***: The text of the claim
- **evidence**: A list of evidence sets (lists of [Annotation ID, Evidence ID, Wikipedia URL, sentence ID] tuples) or a [Annotation ID, Evidence ID, null, null] tuple if the label is NOT ENOUGH INFO

Hate Speech Detection

- Hate speech (e.g., racism) can be frequently found online
  - There has been an interest from both academia and industry into automatic detection of hate speech!

Hate Speech Detection

- Hate speech (e.g., racism) can be frequently found online
  - There has been an interest from both academia and industry into automatic detection of hate speech!
- (It’s not very easy to annotate!)

Hate Speech Detection

- Task definition:
  - Given some text (e.g., tweets, social media comments, etc.), label it as either containing hate speech or not

Hate Speech Detection

- Task definition:
  - Given some text (e.g., tweets, social media comments, etc.), label it as either containing hate speech or not
  - Labels could be, for instance, RACISM, SEXISM, or NEITHER (Waseem, 2016)

Ethics Statements

Transactions of the Association for Computational Linguistics (TACL)

- **Authors**. The corresponding author is responsible for the appropriateness and completeness of the authorship list, for certifying the article's originality as described below, and for securing the agreement of all authors to the journal's open access and ethics policies.

Transactions of the Association for Computational Linguistics (TACL)

- **Originality.** All articles must represent original work: when submitted, the submission must not have been previously published, and the material in it must not have been under review by another journal or conference; further, it must be that no material in it was or is submitted for review at another conference or journal while under review by TACL. For each submission, the submitting author must affirm the following: "The submission does not contain any instances of research fabrication or plagiarism --- the use of the ideas or language of others without attribution. Note that rephrasing the language or wording of others without acknowledgment of the original source is still plagiarism, that is, plagiarism extends beyond word-for-word copying."

Transactions of the Association for Computational Linguistics (TACL)

- **Reviewers and editors.** All such parties, including the Editors-in-Chief (EiCs), must keep submissions confidential except for the purposes of investigating possible misconduct (such as plagiarism) or checking compliance with TACL's policies barring resubmissions from other conferences or with other organizations' multiple submission polices. They must also prevent or undo assignment to submissions with which they have a conflict of interest (COI, defined below) by reporting the COI to the person assigning them to the submission and/or to the (other) EiCs. Knowledge of or guesses as to author identity must not influence judgment of a submission’s merit.

Transactions of the Association for Computational Linguistics (TACL)

- **Conflicts of Interest (COIs).** TACL uses the definition of COI set forth by the Association for Computational Linguistics (ACL), namely, a person has a COI with a submitted paper if that person: (1) is a co-author of the paper; or (2) has been a student or supervisor of one of the authors in the previous five years; or (3) has co-authored a paper or collaborated with one of the authors in the previous five years; or (4) is employed at the same company or institution as an author; or (5) has any other circumstances that could cause a bias in evaluating the paper.

ACL Anti-Harassment Policy

[…] Harassment and hostile behavior are unwelcome at any ACL conference, associated event, or in ACL-affiliated on-line discussions. This includes: speech or behavior that intimidates, creates discomfort, or interferes with a person's participation or opportunity for participation in a conference or an event. We aim for ACL-related activities to be an environment where harassment in any form does not happen, including but not limited to: harassment based on race, gender, religion, age, color, appearance, national origin, ancestry, disability, sexual orientation, or gender identity. Harassment includes degrading verbal comments, deliberate intimidation, stalking, harassing photography or recording, inappropriate physical contact, and unwelcome sexual attention. The policy is not intended to inhibit challenging scientific debate, but rather to promote it through ensuring that all are welcome to participate in shared spirit of scientific inquiry. Vexatious complaints and willful misuse of this procedure will render the complainant subject to the same sanctions as a violation of the anti-harassment policy. […]

Wrapping up

- Discussed today:
  - Ethical issues we had discussed previously
  - IRB and human subjects
  - Ethical issues in NLP
  - Debiasing
  - NLP against unethical behavior
  - Ethics Statements

- On Wednesday: Information extraction