On the Hunt for Data

Mariana Romanyshyn
Computational Linguist, Tech Lead at Grammarly

Data Annotation

Data Annotation: who?

- Own annotators
- Volunteers
- Crowdsourcing platforms
  - *AMT, CrowdFlower*
- Expert linguists
  - *Appen, Leapforce, iSoftStone*

Data Annotation: who?

Crowdsourcing

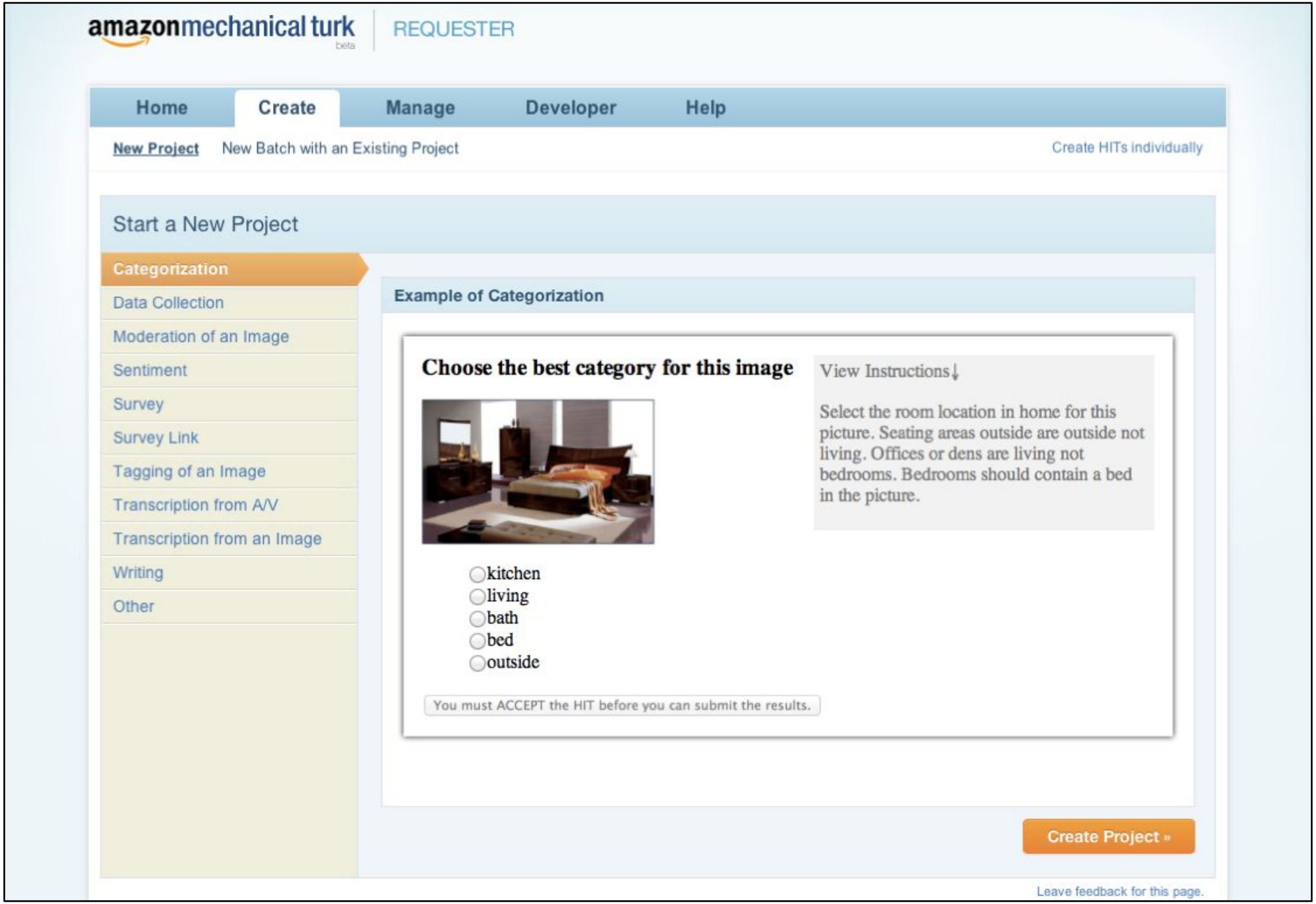
+ cheap and fast
- little control over quality

Expert Linguists

- expensive and time-consuming
+ easier to control quality

Crowdsourcing: Amazon Mechanical Turk

- mturk.com - a platform for work that requires human intelligence
- Requesters vs. Workers
- Human Intelligence Task (HIT)
- Provides a sandbox: requestersandbox.mturk.com



Start a New Project

Example of Categorization

Choose the best category for this image

View Instructions↓

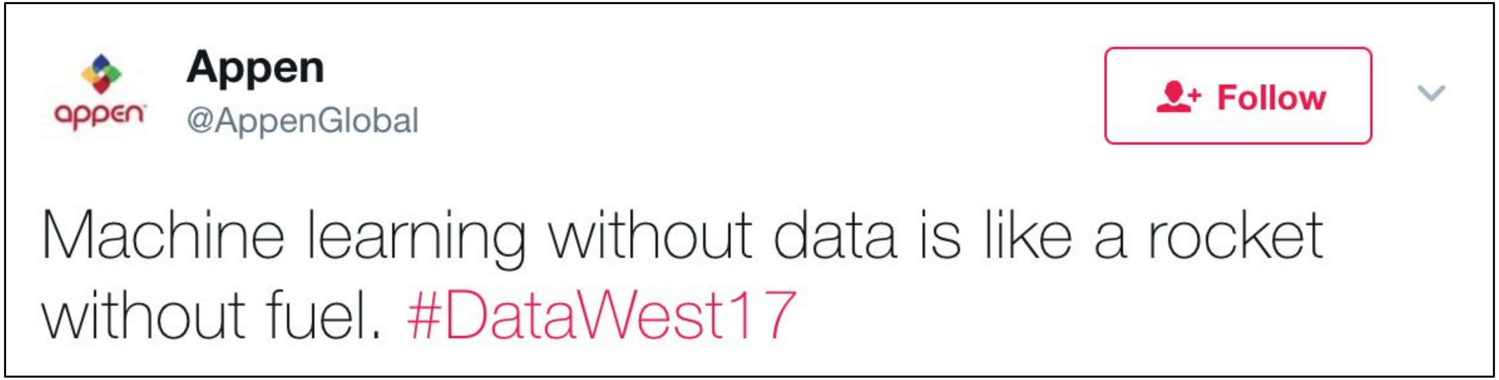
Select the room location in home for this picture. Seating areas outside are outside not living. Offices or dens are living not bedrooms. Bedrooms should contain a bed in the picture.

○kitchen
○living
○bath
○bed
○outside

You must ACCEPT the HIT before you can submit the results.

Expert Linguists: Appen

- appen.com - development of high-quality, human annotated datasets for ML
- 180 languages
- 400,000 annotators

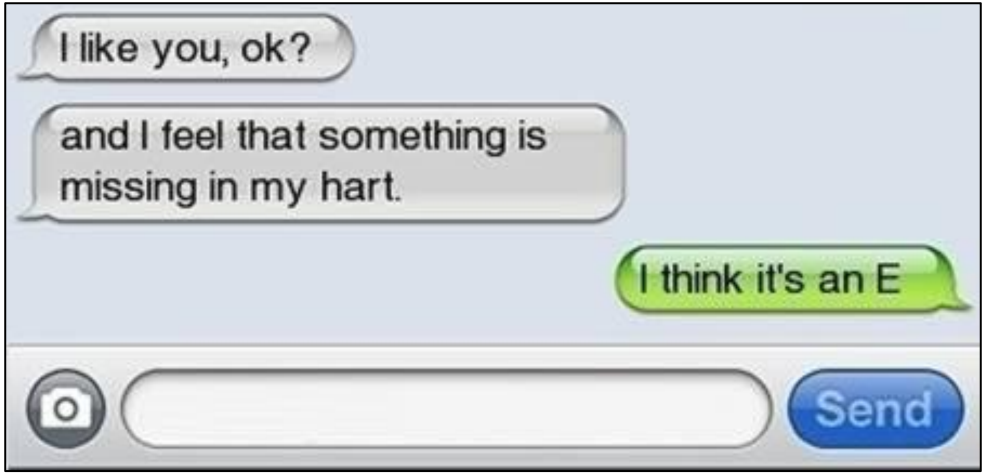


Appen
@AppenGlobal

Machine learning without data is like a rocket without fuel. #DataWest17

Use case: mobile spelling corrections

- What we need?
  - Spelling error annotations in mobile phone messages.



I like you, ok?

and I feel that something is missing in my hart.

I think it's an E

Use case: mobile spelling corrections

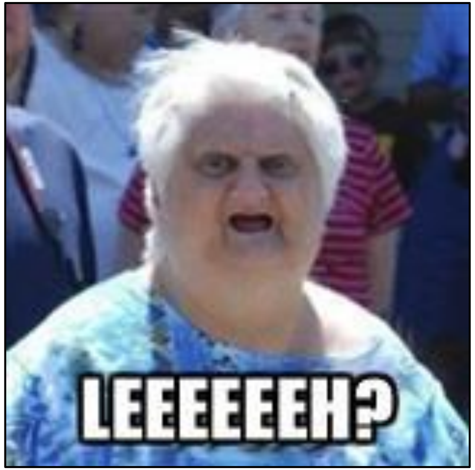
- What's available?
  - *NUS SMS Corpus*
    - *55,000 messages*
  - *Mobile Forensics corpus*
    - *4,934 messages*
  - *The Enron Mobile Email Dataset*
    - *2,600 messages*
  - *SMS Spam Collection v. 1*
    - *425 spam messages + NUS SMS*

NUS SMS Corpus: Singlish

Waiting in a car 4 my mum lor. U leh? Reach home already?

Lor - expresses general agreeability

Leh - expresses negativity



The data says we need more data.

Attack plan

- Collect data
  - Scrape Twitter
  - Use Amazon Mechanical Turk
- Annotate data
  - Automatic annotation
  - Annotation with expert linguists

Twitter

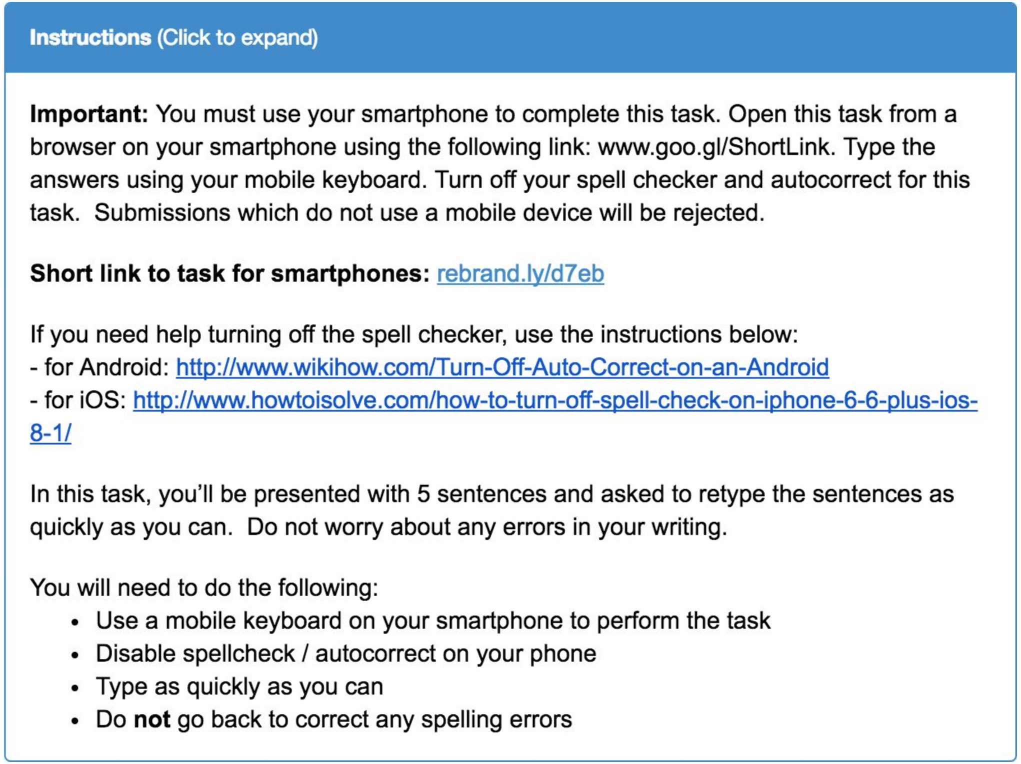
- Twython
- 1,000 from 2011-2013 and another 1,000 from 2016
- “source” in [“IPhone”, “Android”, “Mobile”]
- Data quality:
    - language filter
    - profanities
    - too short or just hashtags
    - average word length < 3, etc.

AMT data collection: idea 1

- What if we ask the turkers to ***retype*** some short messages?
    - How to set up AMT on the phone?
    - What messages to retype?
    - How do we know…
        - they are not copy-pasting?
        - they are not typing some other text instead?
        - they are using a mobile phone?
        - they are not using autocorrect?

Results

- 10,000 sentences
- 2 days
- $0.05 per HIT
- 33,000 misspellings



Instructions (Click to expand)

Important: You must use your smartphone to complete this task. Open this task from a browser on your smartphone using the following link: www.goo.gl/ShortLink. Type the answers using your mobile keyboard. Turn off your spell checker and autocorrect for this task. Submissions which do not use a mobile device will be rejected.

Short link to task for smartphones: rebrand.ly/d7eb

If you need help turning off the spell checker, use the instructions below:
- for Android: http://www.wikihow.com/Turn-Off-Auto-Correct-on-an-Android
- for iOS: http://www.howtoisolve.com/how-to-turn-off-spell-check-on-iphone-6-6-plus-ios-8-1/

In this task, you'll be presented with 5 sentences and asked to retype the sentences as quickly as you can. Do not worry about any errors in your writing.

You will need to do the following:

- Use a mobile keyboard on your smartphone to perform the task
- Disable spellcheck / autocorrect on your phone
- Type as quickly as you can
- Do **not** go back to correct any spelling errors

Example

Pack my box with five dozen liquor jugs.

Pack my box with five dozen liquor jugs.

Pack my box with five dozen liquour jugs.

Pack my box with five dozen liquir jugs.

Paxk my box with guve dozen liquorr jugs.

Pack my box with five dozen liquor jugs.

AMT data collection: idea 2

- What if we ask the turkers to ***give short answers***?
    - How to set up AMT on the phone?
    - What questions to ask?
    - How do we know…
        - they are not copy-pasting random text?
        - they are using a mobile phone?
        - they are not using autocorrect?

Results

- 2,000 answers to 200 questions
- 4 days
- $0.15 per HIT

Issues:

- Misspellings cannot be extracted
- Some data bias

Bias

Saree. Attried in saree looks gorgeous. Its neet beautiful and sexy dress.

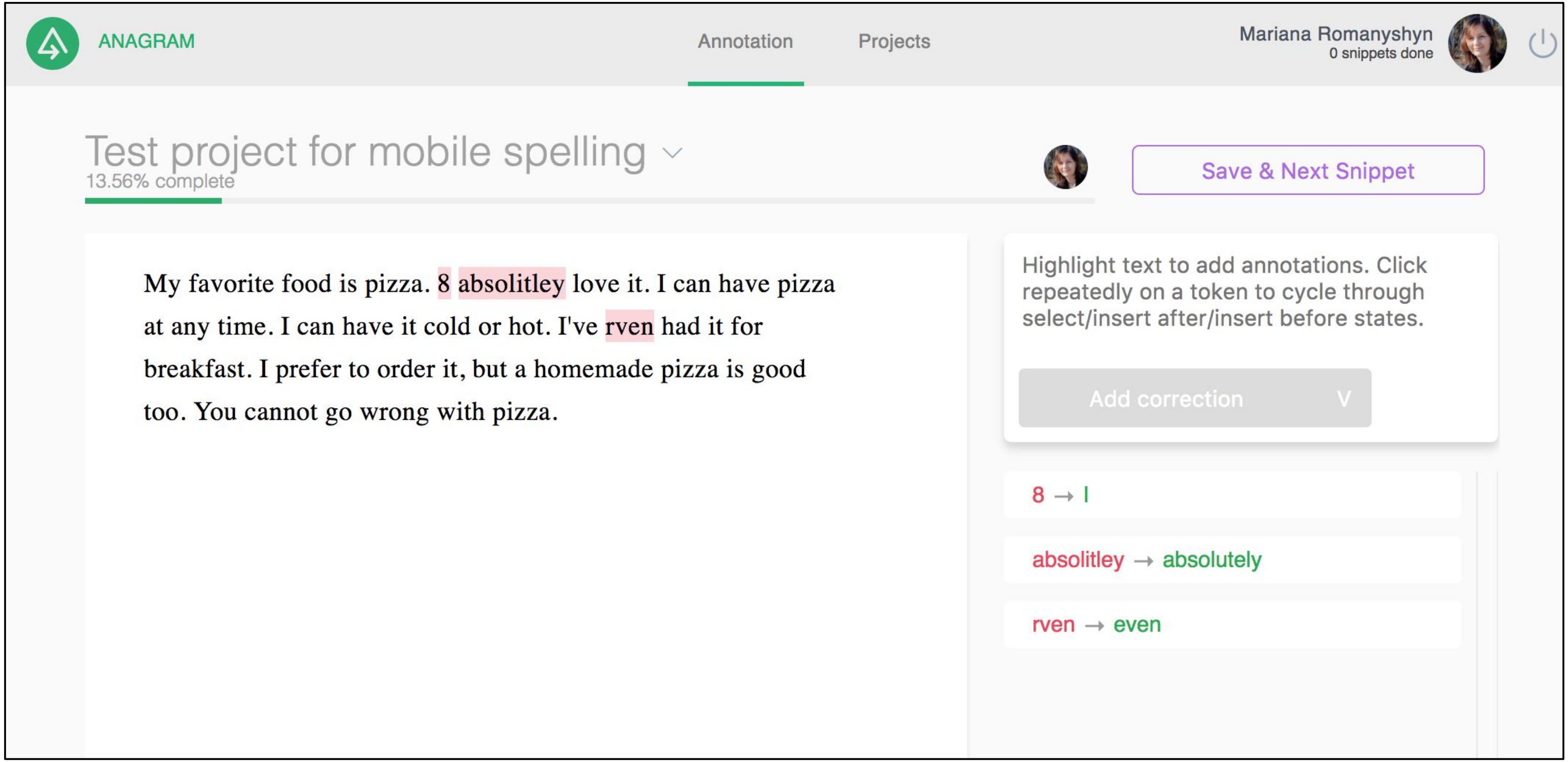
My favourite place is Guruvayur temple. I love Guruvayurappan and i feel relaxed there.

I live in mumbai, maharashtra, india. In mumbai there are many spots where we can enjoy...

Annotation

- Who: *expert linguists*
- Data: *SMS + Twitter + AMT project*
- Tool: *Anagram*

Annotation Tool



Test project for mobile spelling

13.56% complete

My favorite food is pizza. 8 absolitley love it. I can have pizza at any time. I can have it cold or hot. I've rven had it for breakfast. I prefer to order it, but a homemade pizza is good too. You cannot go wrong with pizza.

Highlight text to add annotations. Click repeatedly on a token to cycle through select/insert after/insert before states.

8 → I

absolitley → absolutely

rven → even

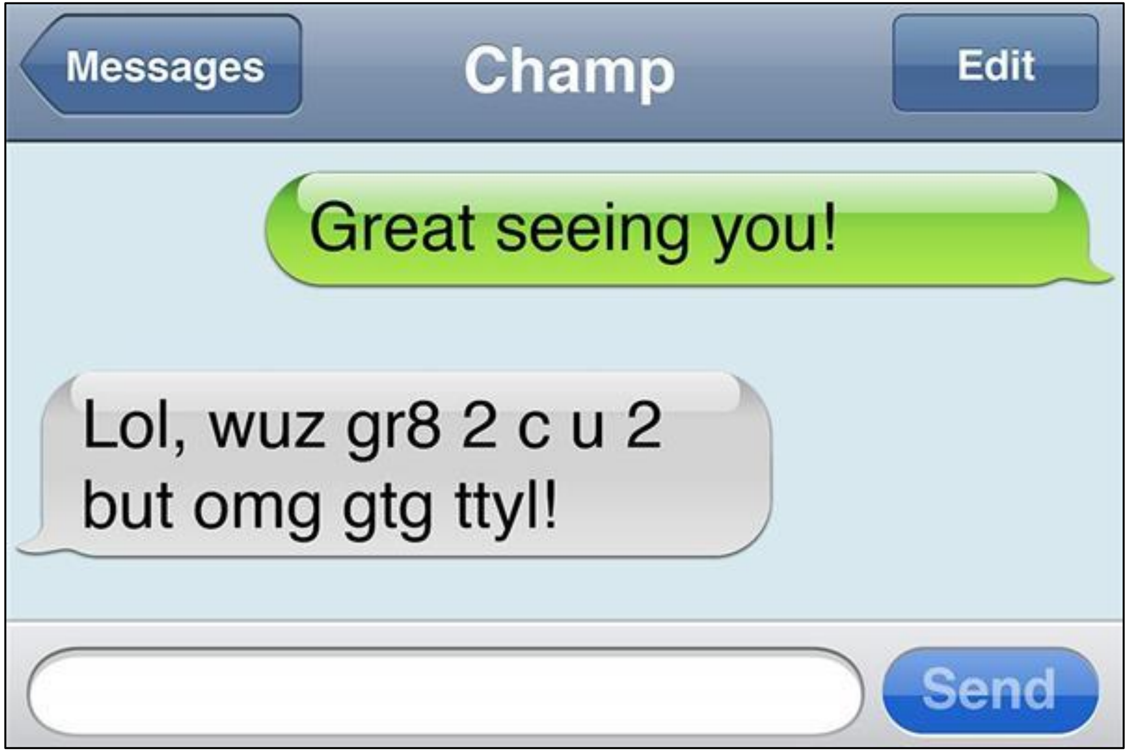
The main issue

cud u tell ppl im gona b a bit l8 cos 2 buses hav gon past cos they were full & im still waitin 4 1. Pete x

The main issue

cud u tell ppl im gona b a bit l8 cos 2 buses hav gon past cos they were full & im still waitin 4 1. Pete x

The main issue



Great seeing you!

Lol, wuz gr8 2 c u 2 but omg gtg ttyl!

Annotation Process

- Guidelines
- Training
- Calibration
- Annotation
- Disagreement resolution

Data Annotation: learnings

Learnings

1. Guidelines
    a. simple, short, non-contradicting
    b. a fall-back option
    c. as many examples as possible

Learnings

1. Guidelines
    1. simple, short, non-contradicting
    2. a fall-back option
    3. as many examples as possible
2. Quality control
    1. qualification tests / training stage
    2. annotators with specific qualifications
    3. cross-annotation
    4. automatic dismissal of the work

Learnings

3. Automatically annotated data saves time…

(and teach the annotators)

Learnings

3. Automatically annotated data saves time…

   (and teach the annotators)

4. Saving time and money
   a. extract 100% agreement from crowdsourcing
   b. use experts to reannotate the rest

Learnings

3. Automatically annotated data saves time…

   (and teach the annotators)

4. Saving time and money
   a. extract 100% agreement from crowdsourcing
   b. use experts to reannotate the rest

5. Pay quickly and be responsive to emails

Learnings

3. Automatically annotated data saves time…

   (and teach the annotators)

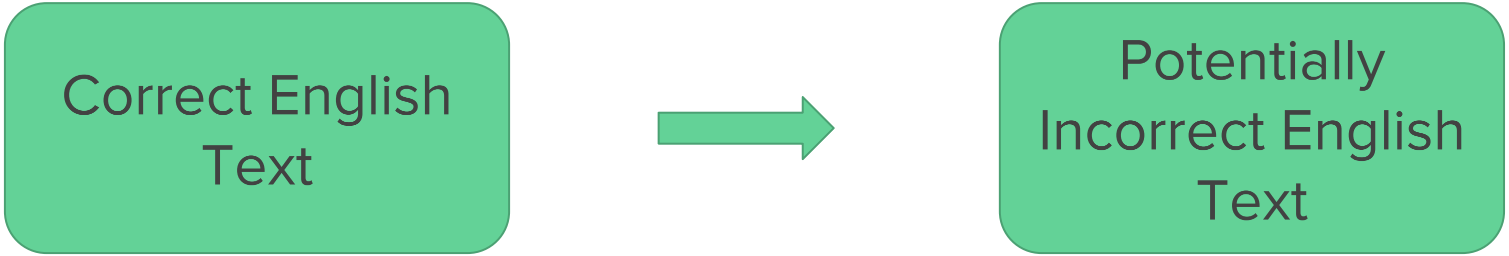
4. Saving time and money
   1. extract 100% agreement from crowdsourcing
   2. use experts to reannotate the rest
5. Pay quickly and be responsive to emails
6. Gamification

Learnings

3. Automatically annotated data saves time…
   (and teach the annotators)
4. Saving time and money
   a. extract 100% agreement from crowdsourcing
   b. use experts to reannotate the rest
5. Pay quickly and be responsive to emails
6. Gamification
7. Annotation bias

Data Generation

The Idea



Correct English Text

Potentially Incorrect English Text

Use case: collocation correction

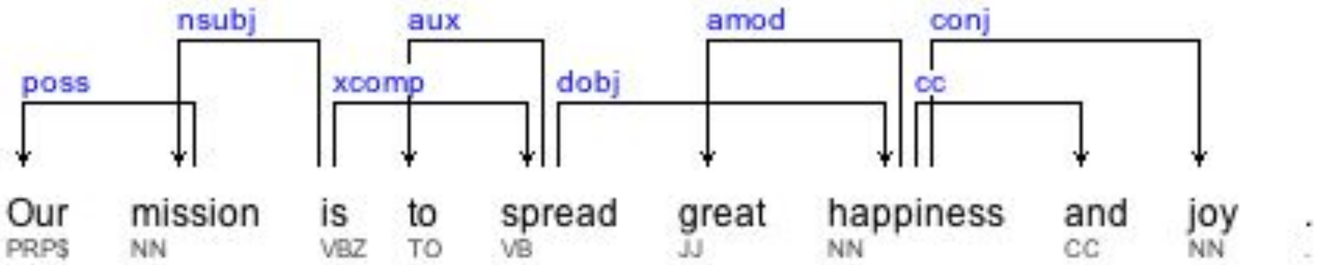
Our mission is to spread big happiness and joy.

Use case: collocation correction

Our mission is to spread {big => great} happiness and joy.

Use case: collocation correction

- Extract collocations from good texts



Use case: collocation correction

- Extract collocations from good texts
- Get synonyms from a thesaurus

Our mission is to spread great happiness and joy.

large

hefty

massive

big

considerable

...

Use case: collocation correction

- Extract collocations from good texts
- Get synonyms from a thesaurus
- Filters:
    - is the replacement a good collocation?
    - is the combination frequent?

Use case: collocation correction

- Extract collocations from good texts
- Get synonyms from a thesaurus
- Filters
- Replace the good adjective with a synonym

Our mission is to spread large happiness and joy.
Our mission is to spread hefty happiness and joy.
Our mission is to spread massive happiness and joy.
Our mission is to spread big happiness and joy.
Our mission is to spread considerable happiness and joy.

Results

- True positives
    - *I thought you did a {full => comprehensive} research...*
    - *...the most {beautiful => good-looking} men in the world.*
- Problems
    - Not all confusions are synonymous:
        - *{crowded => heavy} traffic*
    - Rare combinations can be treated as a mistake
        - *{Subversive=>Underground} lines characterize...*

Data Licensing

Data Owners

- Universities/companies/individuals

- Issues:
  - data owners have no idea about cost and/or license
  - legislation is different in different countries
  - be ready to spend about 3 months
  - and sometimes...

JOEY DOESN'T SHARE FOOD!
ROFLBOT

Sometimes you win, sometimes you learn

- The Story of a Missing Licence from Creators
- The Story of a Lost Electronic Copy
- The Story of a Never-Ending Divorce
- The Story of a Grumpy Data Owner

Conclusions

- Available data is scarce
- Data can be annotated, generated, and licenced effectively
- Good data == great effort
- Sometimes you win, sometimes you learn