Finding the needle in the haystack

Frustrations and lessons learned from collecting and annotating data

Thamar Solorio
University of Houston

This Talk

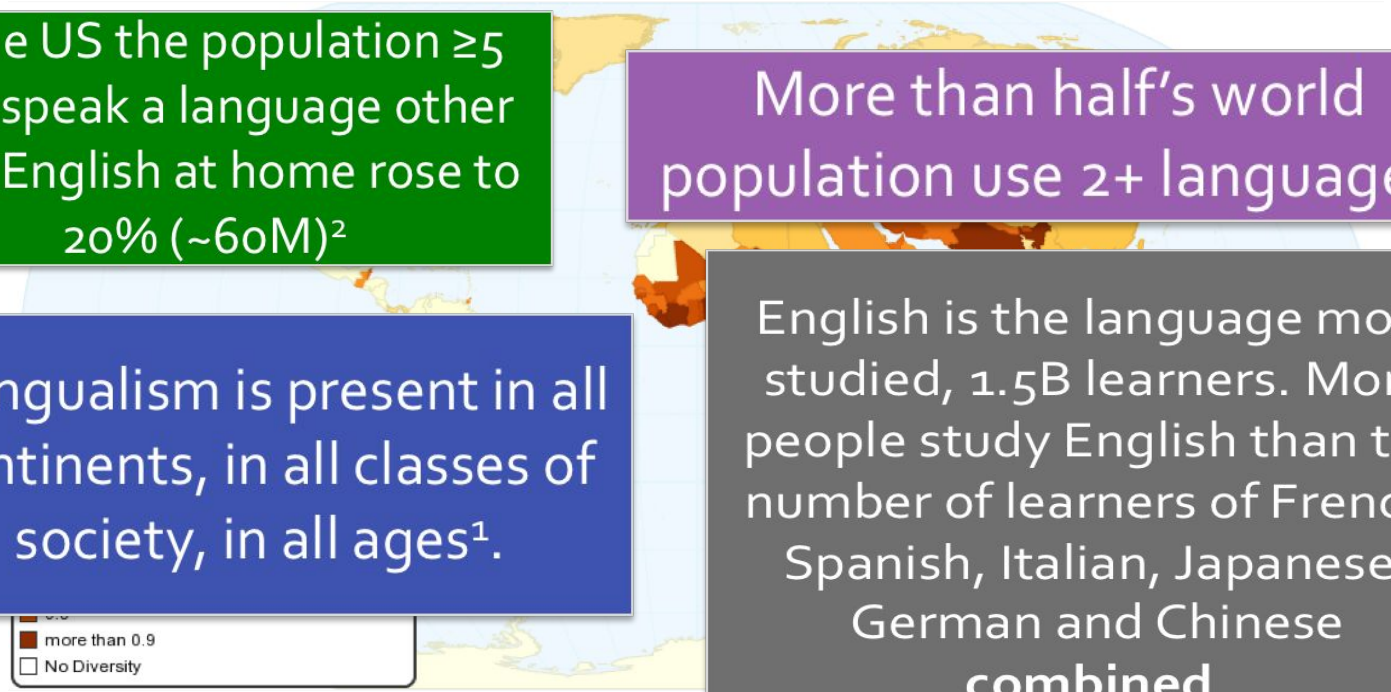
- Not all data are created equal
- Often we care about linguistic patterns and not about specific topics

Our Data Collection and Annotation Projects

- Mixed-Language Data
  - Spanish-English, MSA-Arabic Dialects, Nepali-English
  - Challenges with finding data
  - Our approach
- Extremely negative data
  - Definition of extremely negative
  - Challenges collecting and annotating data
  - Our approach

Task 1:
Mixed-Language Data

Languages in Contact



In the US the population ≥5 who speak a language other then English at home rose to 20% (~60M)[2]

More than half's world population use 2+ languages[1]

Bilingualism is present in all continents, in all classes of society, in all ages[1].

English is the language most studied, 1.5B learners. More people study English than the number of learners of French, Spanish, Italian, Japanese, German and Chinese **combined**.

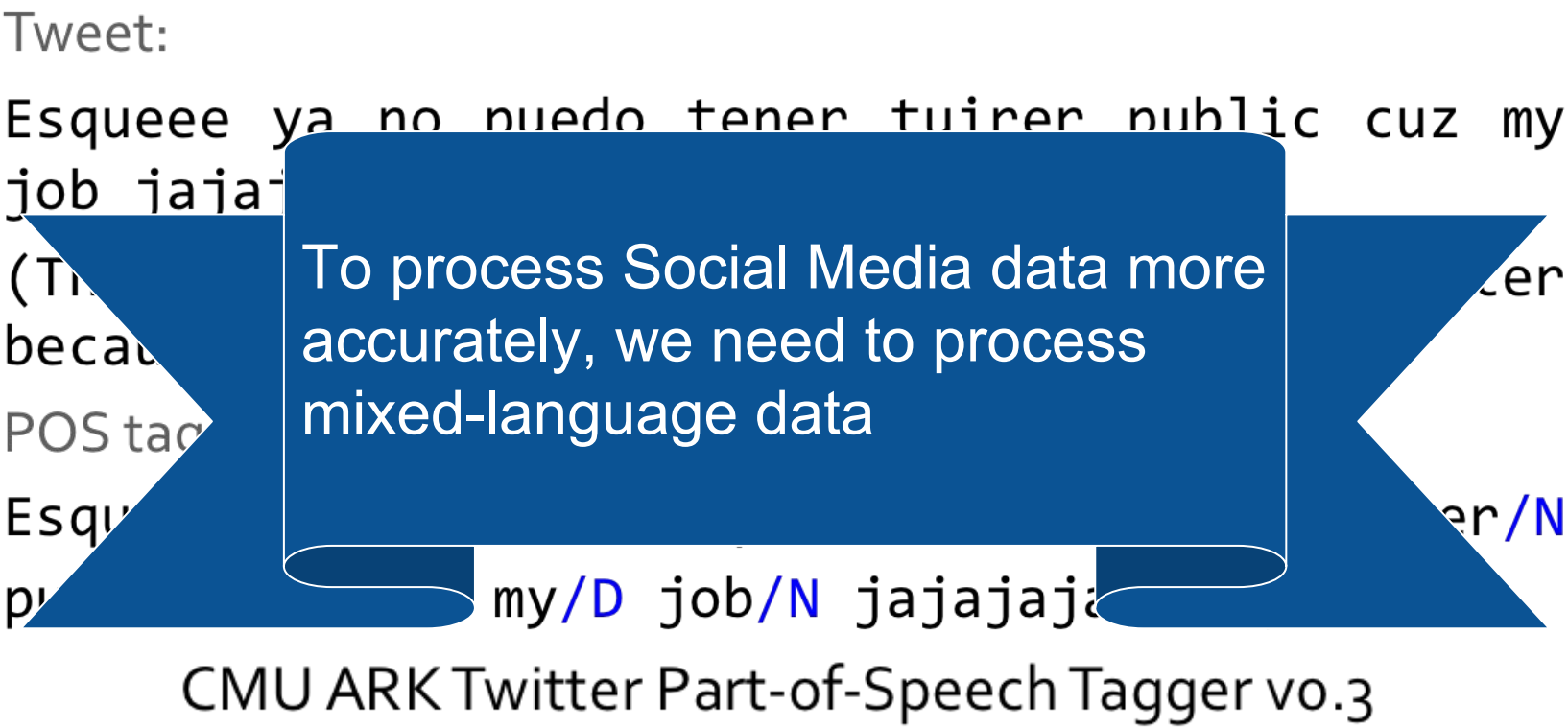
Social Media data can look like this

Tweet:

Esqueee ya no puedo tener tuirer public cuz my job jajajaja :(

(The thing is I can’t have a public tweeter because of my job hahahaha :()

Social Media data can look like this



To process Social Media data more accurately, we need to process mixed-language data

Our Goals

1. To design new NLP tools capable of dealing with more than one language in the input (enabling technology)
2. Motivate more research in this direction (shared tasks and workshops)

Our Goals

1. To design new NLP tools capable of dealing with more than one language in the input (enabling technology)
2. Motivate more research in this direction (shared tasks and workshops)

- Both of this require annotated data
- The most challenging language combination was Spanish-English (Spa-Eng).
- We focused on Social Media from Twitter and Facebook

Challenges in Finding Spa-Eng Data

- The political history of Spanish in the USA made it hard to find data
- We can’t just look for #code-switching
- Using only keywords will result in artificially small vocabulary

Our Solution to Collect Spa-Eng Data in Twitter

1. Search for common English words in Spanish tweets
2. Search for common Spanish words in English tweets
3. In-lab annotators labeled this data
4. Ranked users by the amount of mixed language data
5. Crawl more data from top code-switching users
6. Annotate this data with CrowdFlower (now Figure8)

Data Collected and Annotated

First Shared task on Language Identification

Language-pair	Training	Test
MAN-EN	1000	313
MSA-DA	5,838	2332, 1,777
NEP-EN	9,993	3,018 (2,874)
SPA-EN	11,400	3,060 (1,626)

T. Solorio, E. Blair, S.Maharjan, S. Bethard, M. Diab, M. Ghoneim, A. Hawwari, F. AlGhamdi, J. Hirschberg, A. Chang and P. Fung. *Overview for the First Shared Task on Language Identification in Code-Switched Data.* Proceedings from the First Workshop on Computational Approaches to Code-Switching, held at EMNLP 2014.

Data Collected and Annotated

Second shared task on Language Identification

Language Pair	Training	Development	Test
MSA-DA	8,862	1,117	1,262 (1,258)
SPA-ENG	8,733	1,857	18,237 (10,716)

G. Molina, F. AlGhamdi, M. Ghoneim, A.Hawwari, N. Rey-Villamizar, M. Diab, T. Solorio. *Overview for the Second Shared Task on Language Identification in Code-Switched Data.* Proceedings from the Second Workshop on Computational Approaches to Code-Switching, held at EMNLP 2016.

Data Collected and Annotated

Third shared task: Named Entity Recognition

Classes	ENG-SPA			MSA-EGY		
	Train	Dev	Test	Train	Dev	Test
Person	6,226	95	1,888	8,897	1,113	777
Location	4,323	16	803	4,500	474	332
Organization	1,381	10	307	2,596	263	179
Group	1,024	5	153	2,646	303	139
Title	1,980	50	542	2,057	258	18
Product	1,885	21	481	795	81	54
Event	557	6	99	902	121	81
Time	786	9	197	578	79	28
Other	382	7	62	122	19	2
NE Tokens	18,544	219	4,532	23,093	2,711	1,610
O Tokens	614,013	9,364	178,479	181,229	20,031	19,804
Tweets	50,757	832	15,634	?	?	?

G. Aguilar et al., *Overview for the CALCS Shared Task 2018: Named Entity Recognition on Code-Switched Data, (coming soon)*

Task 2:
Nasty Data


Bully @Bully
UR SO UGLY
6m
bully
You're worthless
Bully
2 seconds ago
NO ONE CARES!!
You're a LOSER.
You're so PATHETIC
You SUCK
fakebook
WORDS SCAR,
RUMORS DESTROY,
BULLIES KILL.
Bully @bully-now
Ew. Don't talk to me.
Expand
bully
Can't believe you even show up to school. Nobody wants you there.
bully
I HATE YOU
Nobody likes you.
1
Kristina Webb
Bully said:
Just give up now.
Bully
2 seconds ago
EW. such an attention seeker
Like · Comment · Share
Bully @Bully
You are STUPID
bully
WEIRDO
Bully @Bully 3s
You are so FAKE
Expand
You Tube
Bully
Stop trying to be cool, its not working.
1 second ago
Reply

Data Collection

- Start with low-hanging fruit: bad words
- Data source is ask.fm
  - Anyone can ask any question to any other user
  - We collect the question and the answer
  - Semi-anonymous social media platform
  - Very popular among teens and pre-teens
- We collected 586K question-answer pairs from 1,954 random users

Challenges

- Bad wor
- Even se
  respons
- We crav

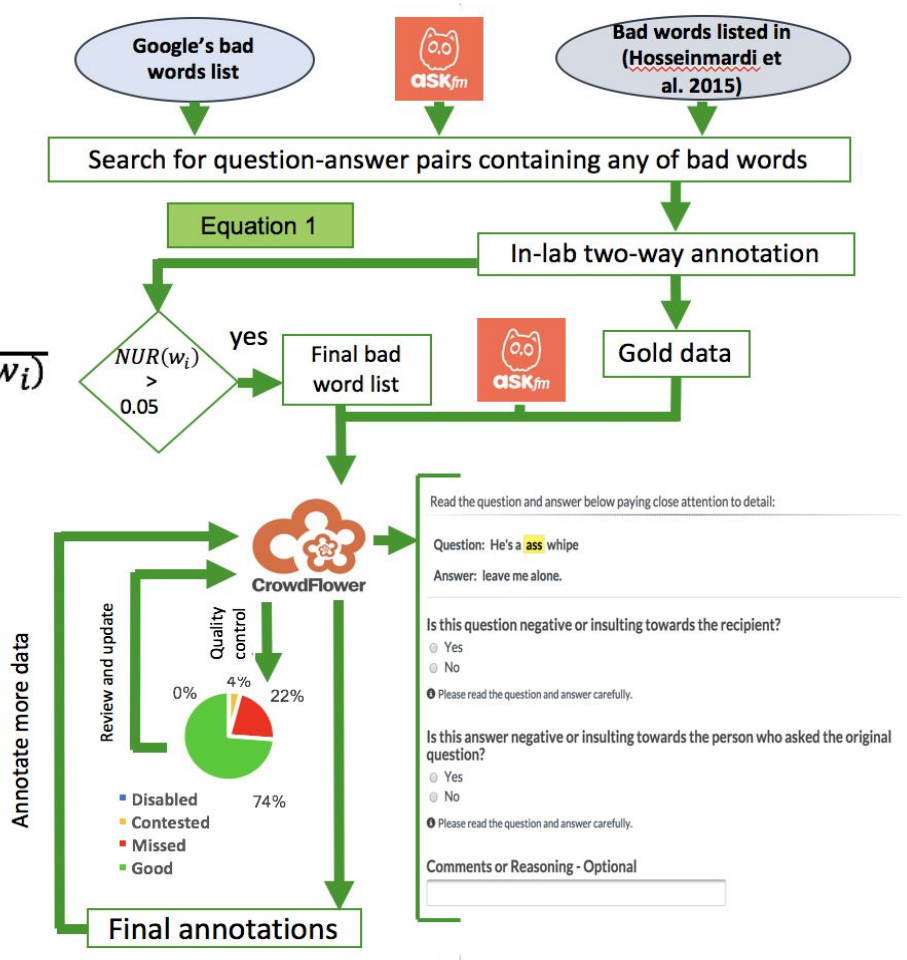
e contexts
r to receive a positive
of it was not relevant!

Annotation Process

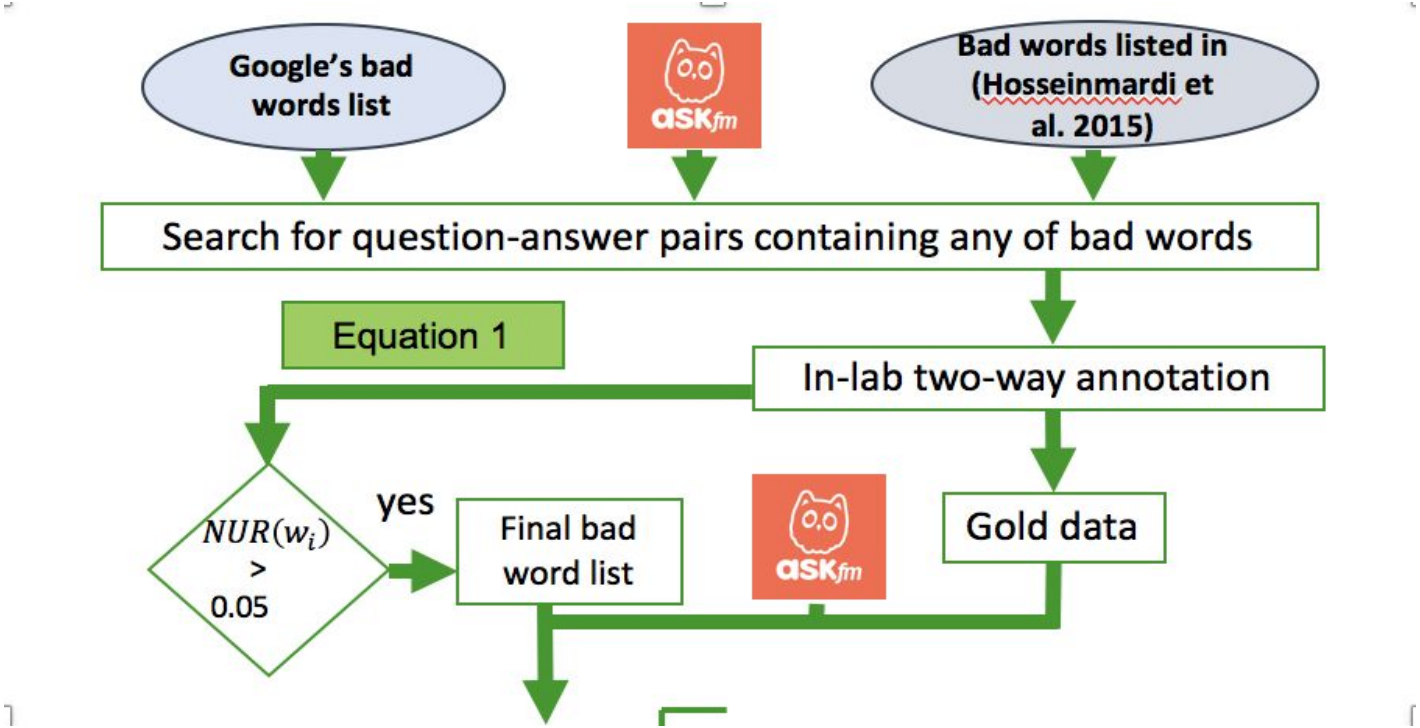
- Negative Use Rate of bad words:

$$NUR(w_i) = \frac{Count(Invective, w_i)}{Count(Invective, w_i) + Count(Neutral, w_i)}$$

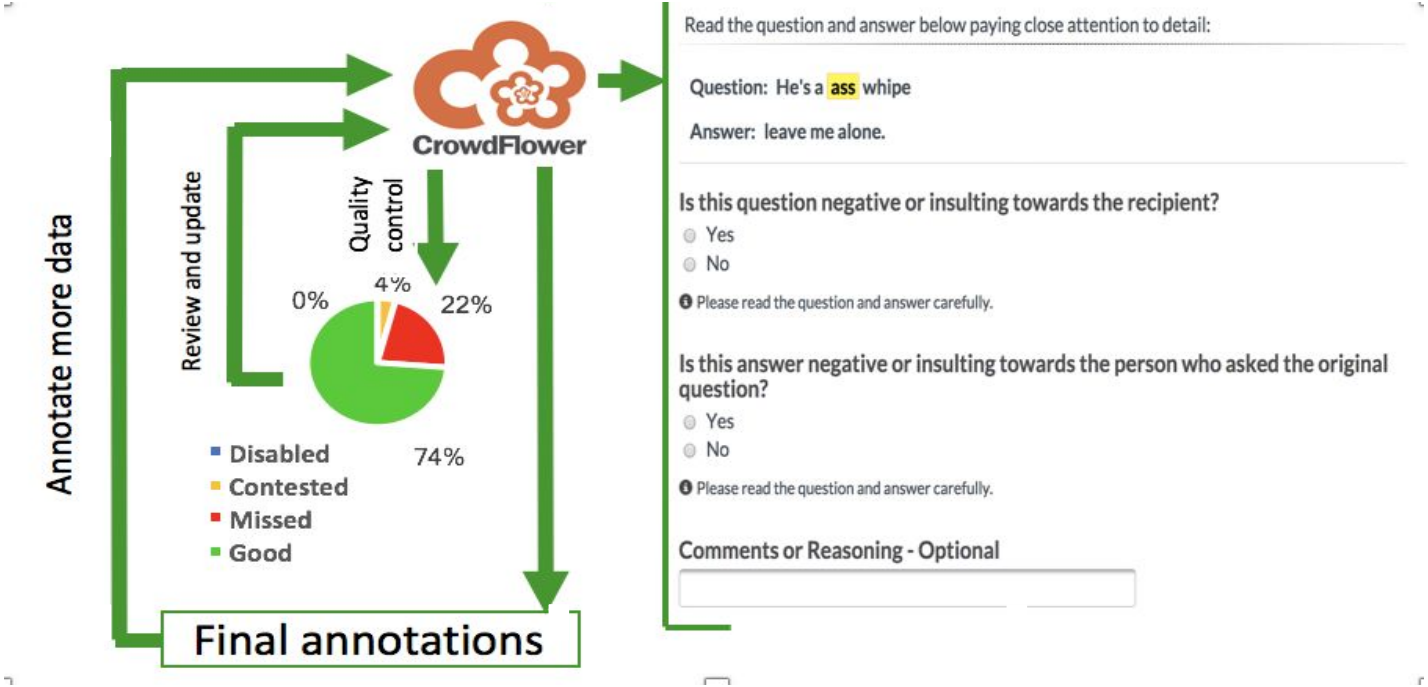
- The inter annotation kappa score is 0.453 which is fair agreement for this task.



Annotation Process



Annotation Process



Example

Q1: you s*ck big c*cks don't you? they've told me. hit me up b**ch.
Q2: I honestly don't give a f*ck if you let me call you names or not, I can do what I want, hoe. you ain't shit to me.
Q3: you shit.
Q4: b**ch shut up, you s*ck d*ck. reason why guys cheat on your asss.
Q5: You shit? Umm o: stay in school kids.
Q6: Yeah that's what frustrates me :/
Q7: I know, thanks skank.
Q8: are you really that stupid? when someone cheats on you, you're with them idiot. mario, and jacob. go suck black c*ck now.
Q9: who hasn't cheated on your ass, anything is possible.
Q10: you're awfully stupid.

Final Remarks

- We do have available lots of data
  - But finding what we're looking for can be a challenge
- We need to find a way to make annotated Twitter data easily shareable

Thanks to Collaborators and Funding Agencies

- Gustavo Aguilar
- Fahad AlGhamadi
- Steve Bethard
- Mona Diab
- Raquel Diaz
- Pascale Fung
- Mahmoud Ghoneim
- Abdelati Hawwari
- Giovanni Molina
- Nicolas Rey Villamizar
- Niloofar Samghabadi
- Victor Soto
- Alan Sprague
- Julia Hirschberg
- Suraj Maharjan

Thank you!

Find our corpora @ http://ritual.uh.edu

Labels in our data

Label	Description
lang1	Language 1
lang2	Language 2
mixed	Words with mixed morphemes
NE	Named Entities
ambiguous	Context can't help assign lang1 or lang2
other	Punctuation marks, emoticons, numbers
unknown	Unrecognizable token
FW	Foreign word (not lang1 or lang2)