# Overview for the Second Shared Task on Language Identification in Code-Switched Data

GIOVANNI MOLINA

# Content

Task Description

Data Sets

Survey of Shared Task Systems

Results

Conclusion

Future Work

# Task Description

The task consists of labeling each token/word in the input test data with one out of 8 labels:

| No. | Label | Stands For | Note |
|---|---|---|---|
| 1 | lang1 | English/MSA | English words only |
| 2 | lang2 | Spanish/EGY | Spanish words only |
| 3 | mixed | Mixed Language | Word is partially in both languages |
| 4 | NE | Named Entity | Names |
| 5 | ambiguous | Ambiguous words | Can't determine whether English or Spanish |
| 6 | FW | Foreign Word | Word is not English nor Spanish |
| 7 | UNK | Unknown | Unrecognizable word |
| 8 | other | Not words | Symbols, usernames, emoticons, … |

# Data Sets

Data collected from Twitter.

There were two language pairs:

- SPA-EN
- MSA-DA

Tweet Statistics

| 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) |

# SPA-EN

Train and Development from EMNLP 2014

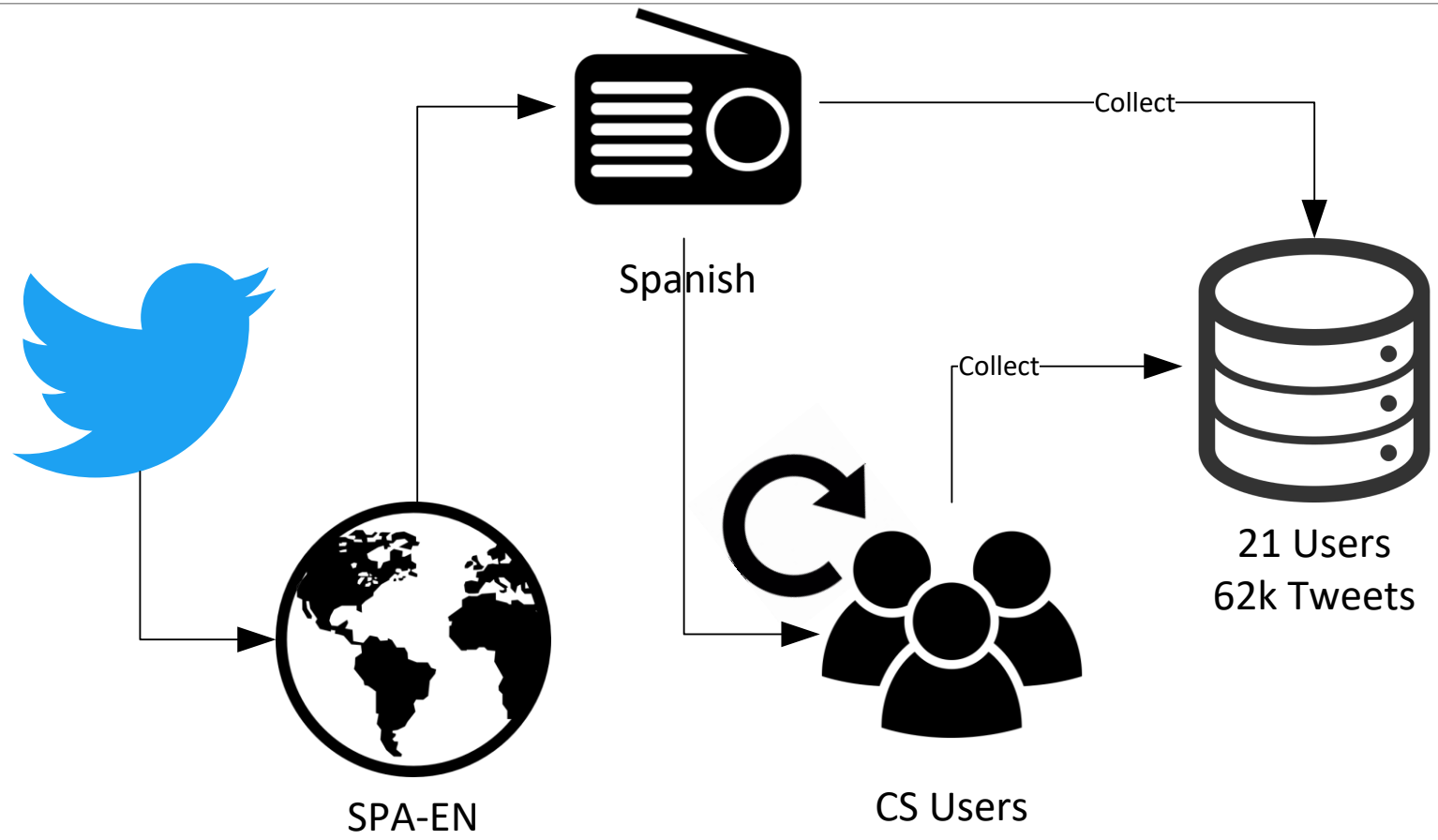
Quality Improvements

## Why?

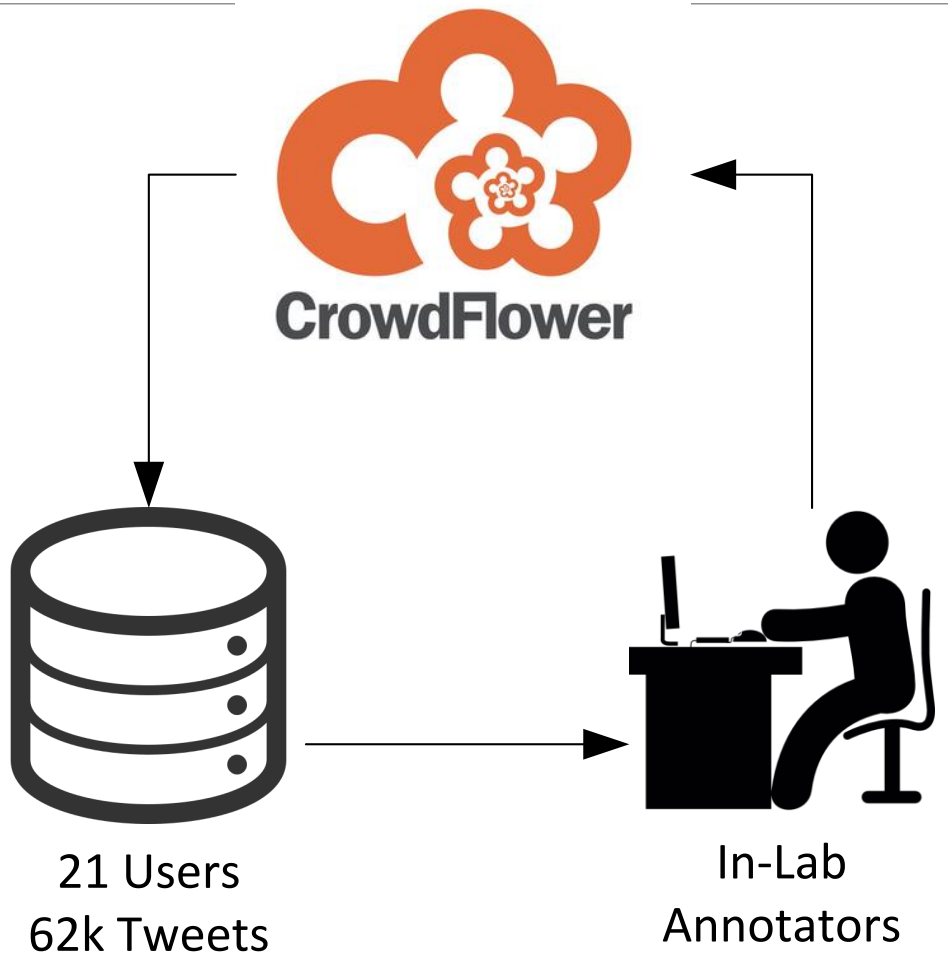
56.6 million Hispanics in the US, 17% of the population.

38.4 million Spanish speakers. 58% of these fully bilingual.

# SPA-EN Data Collection



# SPA-EN Data Annotation



# SPA-EN Test Data Statistics

| Monolingual Tweets | Code-Switched Tweets |
|---|---|
| 4,626 (43.2%) | 6,090 (56.8%) |
| **Label** | **Tokens** |
| ambiguous | 4 |
| lang1 | 16,944 |
| lang2 | 77,047 |
| mixed | 4 |
| ne | 2,092 |
| fw | 19 |
| other | 25,311 |
| unk | 25 |
| **Total** | **121,446** |

# MSA-DA

Train and Development from EMNLP 2014

Quality Improvements

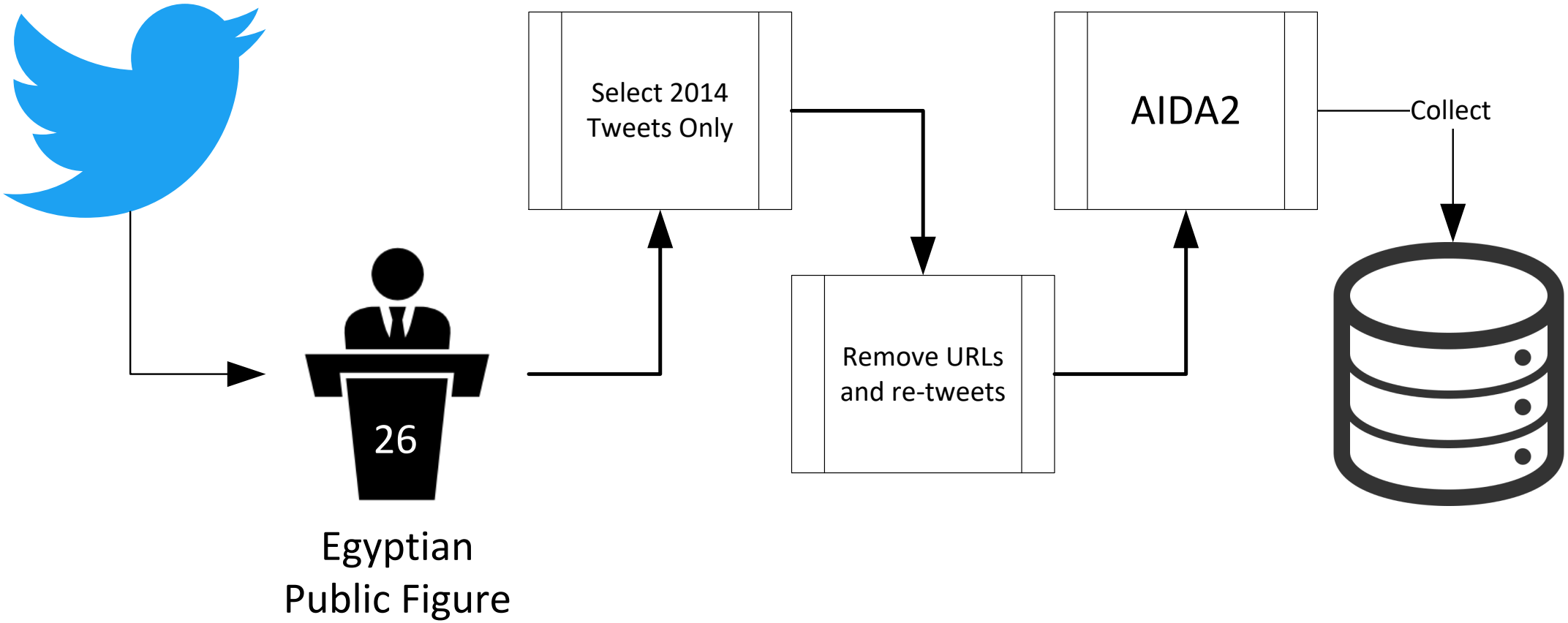
## Why?

Closely related.

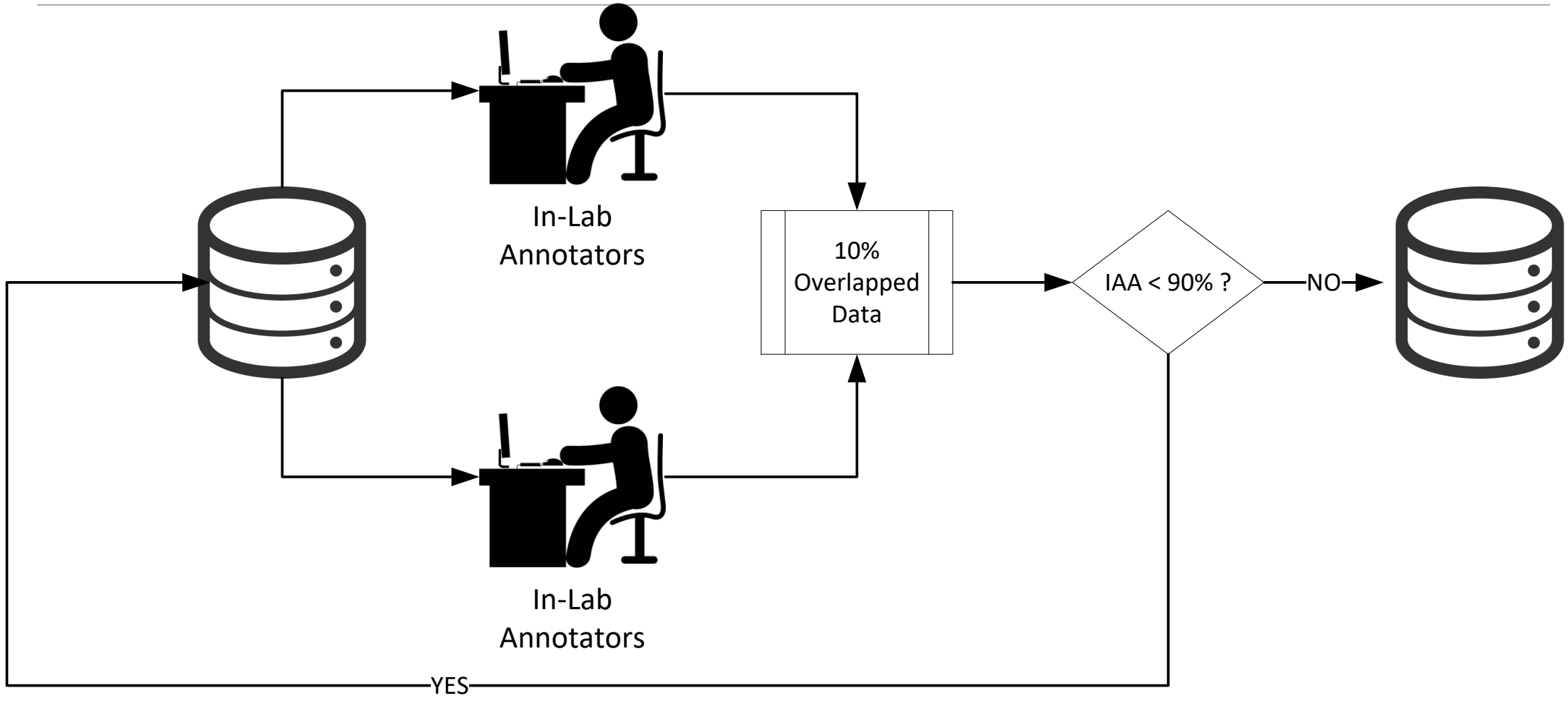
52 million native Egyptian speakers.

24 million L2 Egyptian speakers.

# MSA-DA Data Collection



# MSA-DA Data Annotation



# MSA-DA Test Data Statistics

| Monolingual Tweets | Code-Switched Tweets |
|---|---|
| 1,044 (83%) | 214 (17%) |
| **Label** | **Tokens** |
| ambiguous | 117 |
| lang1 | 5,804 |
| lang2 | 9,630 |
| mixed | 1 |
| ne | 2,363 |
| fw | 0 |
| other | 2,743 |
| unk | 0 |
| **Total** | **20,658** |

# Shared Task Submissions

9 participating teams.

9 system submissions from 9 teams for SPA-EN

5 system submissions from 4 teams for MSA-DA

# Survey of Shared Task Systems

| System | Traditional Machine Learning | Deep Learning | Rules | External Resources | LM | Case | Affixes | Context |
|---|---|---|---|---|---|---|---|---|
| (Al-Badrashiny and Diab, 2016) | CRF | - | - | SPLIT, Gigaword | ✓ | - | - | - |
| (Xia, 2016) | CRF | - | - | fastText | - | ✓ | ✓ | ±1 |
| (Jaech et al., 2016) | - | CNN, LSTM | - | - | - | - | - | - |
| (Shirvani et al., 2016) | Logistic Regres-sion | - | - | GNU Aspell, NER, POS tagger | - | - | ✓ | - |
| (Chanda et al., 2016) | - | - | ✓ | NER, Dictionaries | - | - | - | ±1 |
| (Samih et al., 2016) | CRF | LSTM | - | Gigaword, word2vec | - | ✓ | ✓ | ±1 |
| (Shrestha, 2016) | CRF | - | - | - | - | ✓ | ✓ | - |
| (Sikdar and Gambck, 2016) | CRF | - | - | Babelnet, Babelfy | - | ✓ | ✓ | ±2 |

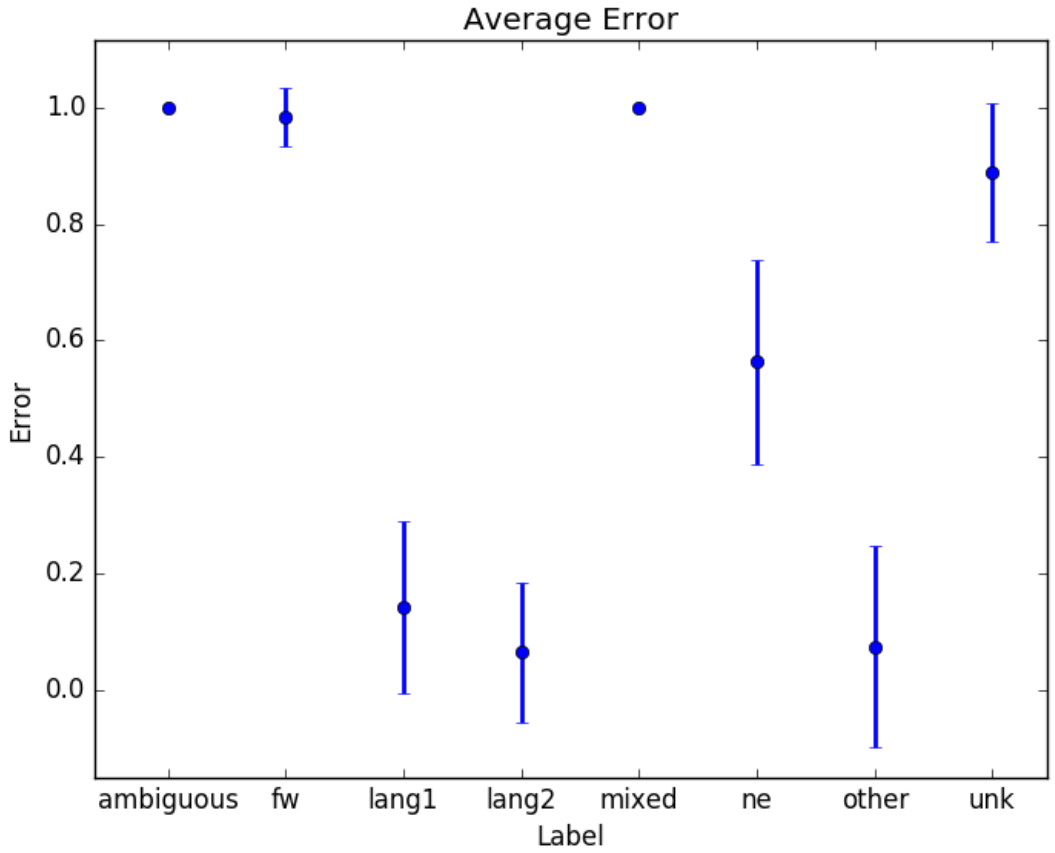
# Tweet Level Results (Top 3)

| Language Pair | System | Weighted F-1 |
|---|---|---|
| SPA-EN | Baseline | 0.607 |
| | Jaech et al. | 0.898 |
| | Samih et al. | 0.90 |
| | **Shirvani et al.** | **0.913** |
| MSA-DA | Baseline | 0.44 |
| | Jaech et al. | 0.73 |
| | Al-Badrishiny and Diab | 0.75 |
| | **Samih et al.** | **0.83** |

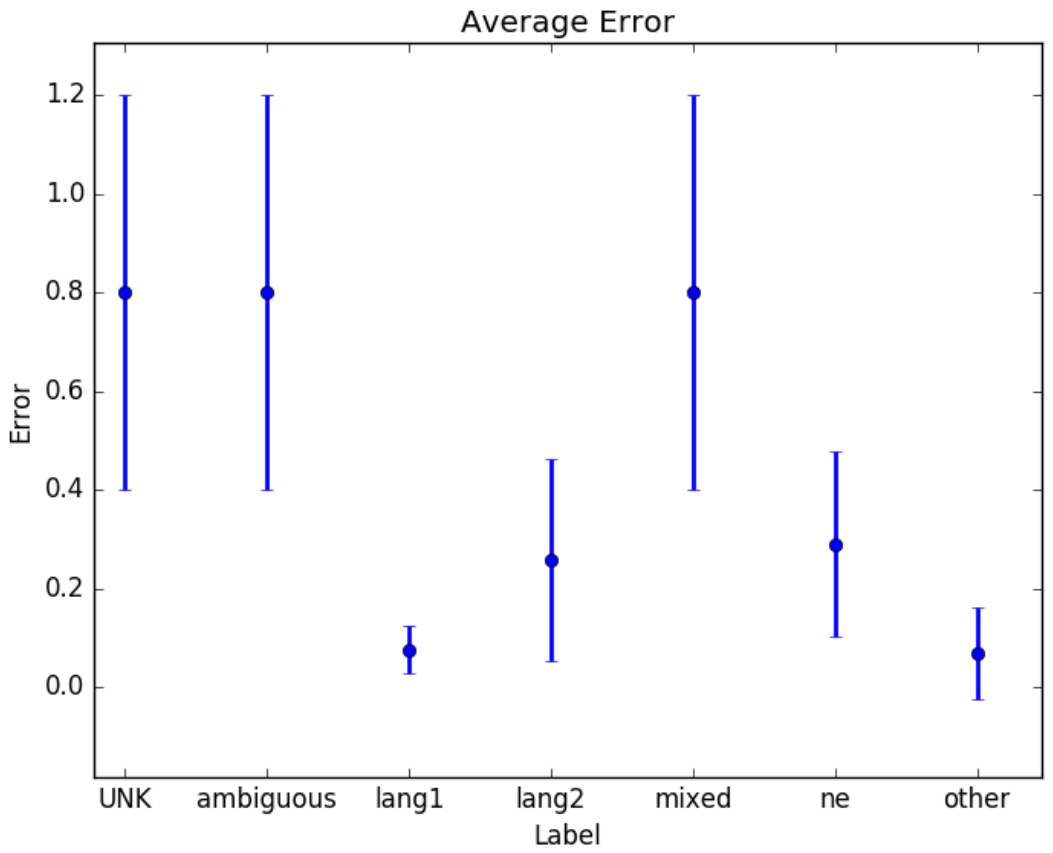
# Token Level Results (Top 3)

| Language Pair | System | Average F-1 |
|---|---|---|
| SPA-EN | Baseline | 0.828 |
| | Samih et al. | 0.968 |
| | IIIT Hyderabad | 0.969 |
| | **Shirvani et al.** | **0.973** |
| MSA-DA | Baseline | 0.463 |
| | Al-Badrishiny and Diab-2 | 0.828 |
| | Al-Badrishiny and Diab-1 | 0.851 |
| | **Samih et al.** | **0.876** |

# SPA-EN Average Errors



# MSA-DA Average Errors



# EMNLP 2014 vs EMNLP 2016

| Tweet Level Results | | |
|---|---|---|
| Language Pair | EMNLP 2014 F-1 Range | EMNLP 2016 F-1 Range |
| SPA-EN | 0.634 – 0.822 | 0.77 – 0.913 |
| MSA-DA | 0.196 – 0.417 | 0.66 – 0.83 |

| Token Level Results | | |
|---|---|---|
| Language Pair | EMNLP 2014 F-1 Range | EMNLP 2016 F-1 Range |
| SPA-EN | 0.704 – 0.940 | 0.603 – 0.973 |
| MSA-DA | 0.385 – 0.799 | 0.594 – 0.876 |

# Outstanding Challenges

SPA-EN:

*Haciendo unos* **looks** *Muy* **cool** *para el* **shoot** *del Nuevo* ***album*** *de* @username *con* #disquared #alexandermcqueen #johnvarvatos #nycshowrooms

**Translated**:

Making some very cool looks for the shoot of the new album by @username with #disquared #alexandermcqueen #johnvarvatos #nycshowrooms

MSA-DA:

@username **لازم اخذ رأيه** لكن رأى المفتى ورأيه استشارى **واخذ** باجماع هيئة المحكمة **ضرورى يكون** اى حكم بالاعدام

**Translated**:

@username Any death sentence **must be** confirmed unanimously by the members of the court and after **consulting** the mufti (i.e, an Islamic scholar who is an interpreter of Islamic law, entitled to issue fatwas). The mufti's opinion should be considered as a consultation but it is **important to consult him.**

# Conclusion

We had a very successful shared task!

Overall performance was higher than previous shared task.

Influence on system design from previous shared task.

New deep learning techniques.

There is great interest in the topic!

# Future Work

Another Shared Task!

POS Tagging

Additional Language Pairs

We want to hear your suggestions at the panel discussion!

# Thank You for participating!


National Science Foundation
WHERE DISCOVERIES BEGIN


## Special thanks:

Fahad AlGhamdi
Mona Diab
Salim El Awad
Pascale Fung
Mahmoud Ghoneim
Julia Hirschberg
Nicolas Rey-Villamizar
Raymund Riegl
Thamar Solorio
Victor Soto
Simon Tice

In-lab annotators
CrowdFlower contributors

# Questions?

Contact Information:

Giovanni Molina
gemolinaramos@uh.edu