Tweeki: Linking Named Entities on Twitter to a Knowledge Graph

Bahareh Harandizadeh, Sameer Singh

https://ucinlp.github.io/tweeki/

Introduction and Motivation

Existing entity linking systems for Twitter:

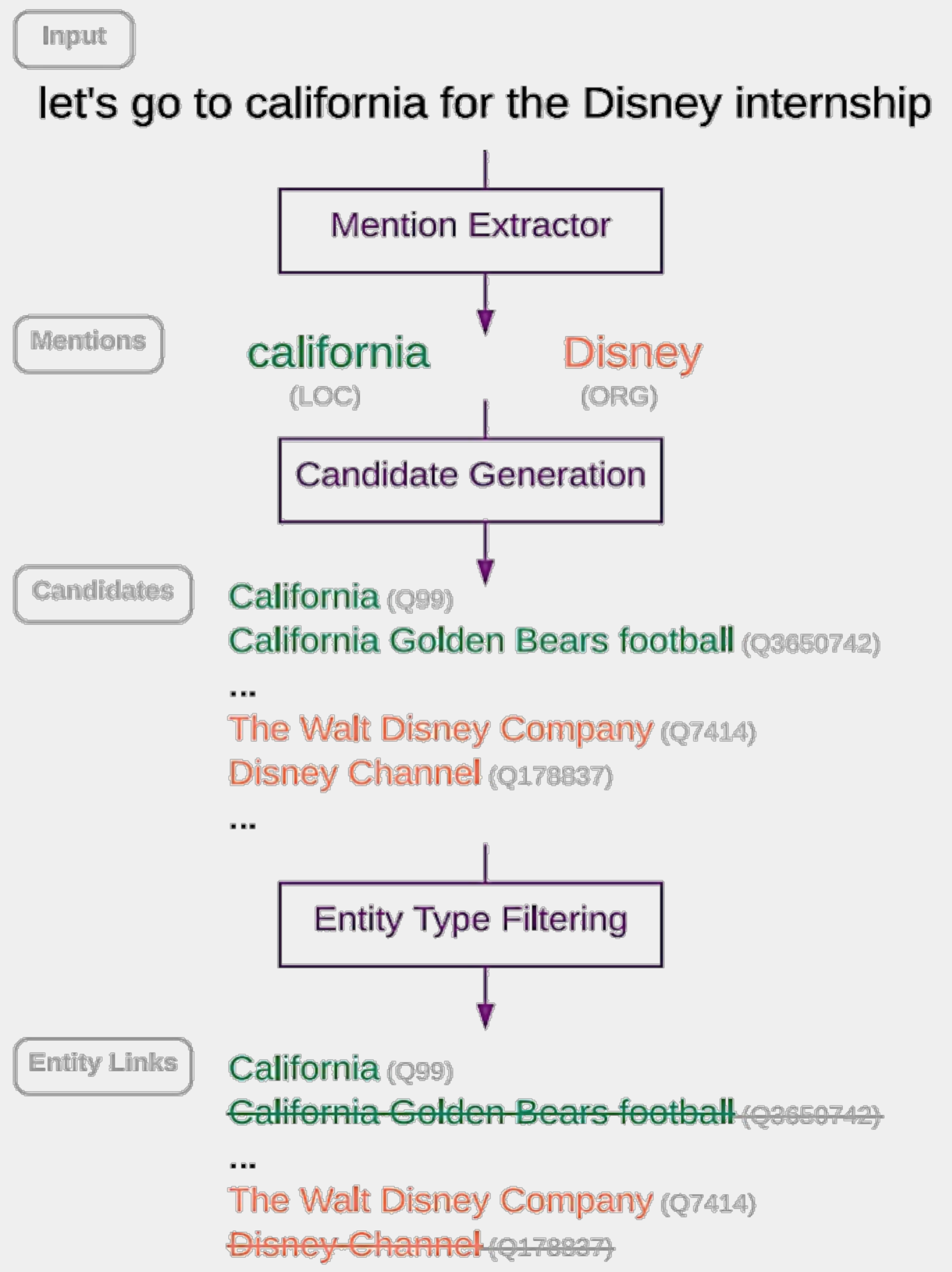
- Many of them are supervised (hence brittle)
- Unsupervised ones are heavily-engineered
  - Difficult to update and maintain
- Lack of linked datasets to use for social media analysis

We introduce:

1. **Tweeki:** unsupervised, modular entity linker for Twitter
2. **TweekiData:** large, automatically-annotated corpus of Tweets linked to entities in WikiData
3. **TweekiGold:** a gold dataset for entity linking evaluation.

Tweeki Pipeline

Tweeki has three major components:



Input

let's go to california for the Disney internship

Mention Extractor

Mentions

california (LOC)

Disney (ORG)

Candidate Generation

Candidates

California (Q99)
California Golden Bears football (Q3650742)
...
The Walt Disney Company (Q7414)
Disney Channel (Q178837)
...

Entity Type Filtering

Entity Links

California (Q99)
~~California Golden Bears football (Q3650742)~~
...
The Walt Disney Company (Q7414)
~~Disney Channel (Q178837)~~

Tweeki Datasets

TweekiData
Automated linking on a large corpus

TweekiGold
Manually annotate a subset for mentions and entity links

	TweekiGold	TweekiData
# tweets	500	5M
# tokens/tweet	16.31	14.41
# mentions (toks)	8,155	8,010,253
# mentions (spans)	958	5,038,870
# links	852	1,954,229
# uniq entities	638	273,685

Component: Mention Extraction

- Mentions are same as named entity recognition (NER)
- We use mentions of type PER, LOC, ORG, MISC.
- To find the most useful NER Tagger, we compare AllenNLP, Spacy, and StanfordNLP (Token-wise ACC):

	Spacy		Stanford		AllenNLP	
	P	R	P	R	P	R
TweekiGold	43.8	57.9	71.8	65.2	81.1	80.9
BTC-A	10.8	42.4	41.1	56.5	48.1	66.6
BTC-H	7.3	14.3	40.6	19.9	74.2	54.2
Average	20.6	38.2	51.2	47.2	**67.8**	**67.2**

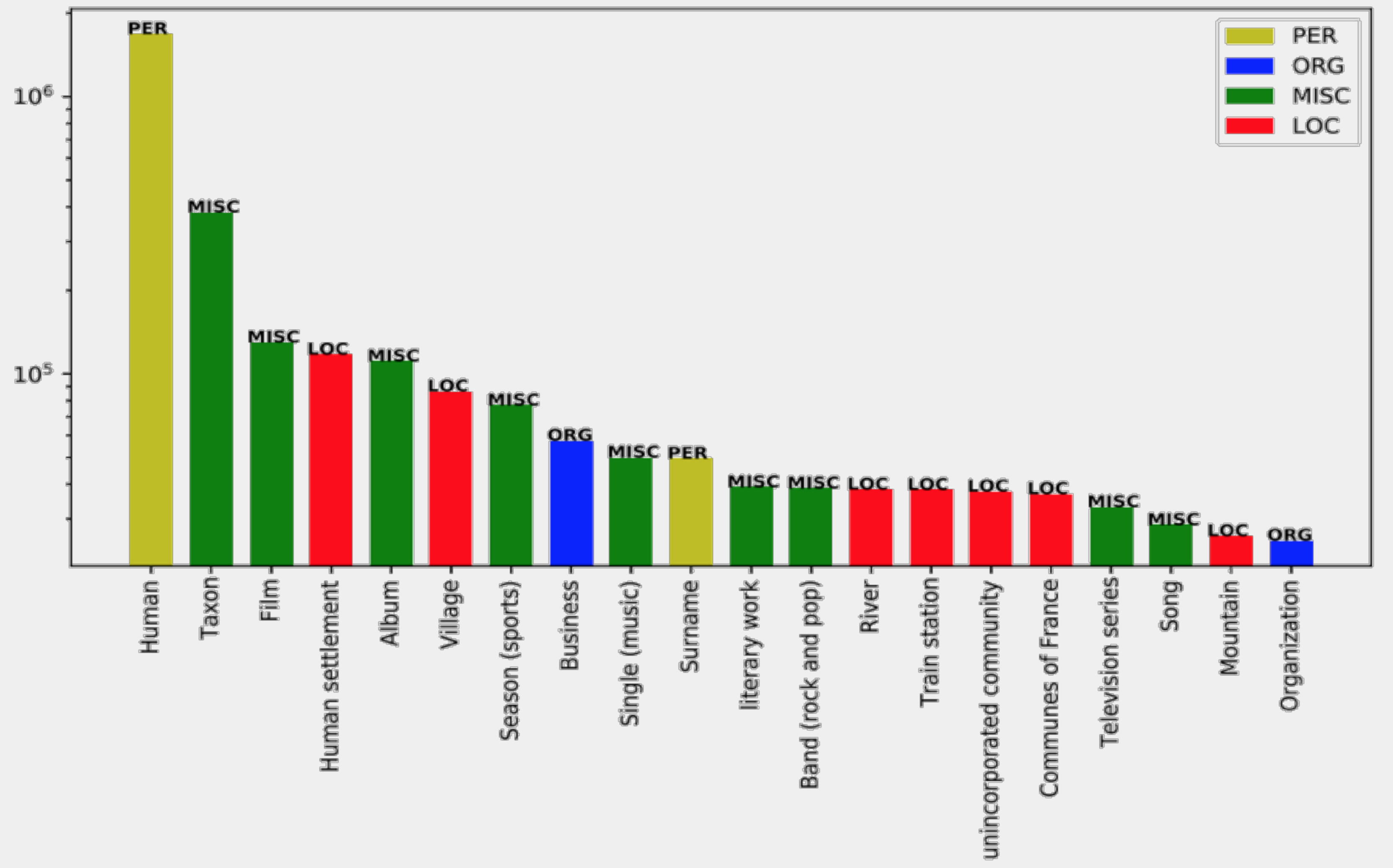
Component: Candidate Generation

- For each mention, get prior probability over entities using existing links between Wikipedia and WikiData entities
- To check the effectiveness of the candidate generation, we tested linking coverage of TweekiData by Types:

Type	#mentions	#entities	Coverage
PER	2.1m	550k	25%
LOC	1.8m	950k	52%
ORG	550k	200k	35%
MISC	490k	200k	40%

Component: Type Filtering

- Identify types for Entities (and filter mention by them)
- We use ***InstanceOf*** relation from WikiData to get the entity types (manually categorize ones that occur more than 100 times into PER, LOC, ORG, and MISC)



- Only consider entity types that match mention types

For each mention, predicted entity link is the remaining entity with highest prior probability

Results

Analyze the performance of Tweeki linker on TweekiGold, NEEL2016 and Derczynski linking datasets

	NEEL2016			Derczynski			TweekiGold		
	P	R	F1	P	R	F1	P	R	F1
TagMe	19.1	**30.0**	24.1	18.2	**50.1**	26.3	38.1	56.1	45.0
Babelfy	8.08	10.6	9.06	9.0	41.1	15.2	17.1	47.2	25.1
AIDA	-	-	-	-	-	-	53.2	32.1	38.5
End-to-End	**87.9**	13.1	22.8	**57.05**	29.2	**39.0**	**79.1**	35.2	49.4
OpenTapioca	11.0	19.1	14.8	9.1	36.0	14.0	20.2	50.4	29.1
Tweeki	58.0	15.2	**24.8**	41.1	34.2	37.1	69.0	**61.0**	**65.0**