Name Entity Recognition in Tweets

Tugberk Goc, Kadir Akgul

University of Bilgi

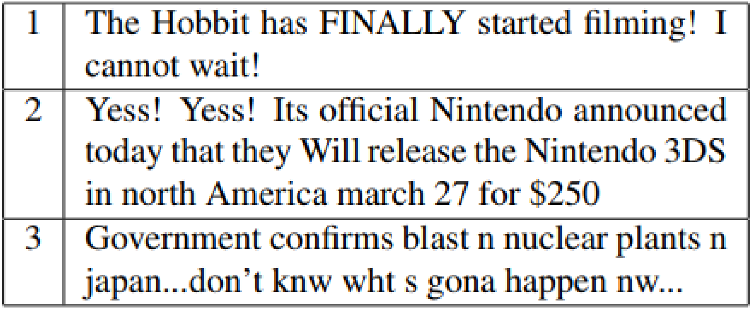
tugberkgoc@gmail.com, kadirakgul@gmail.com

May 17, 2019

Overview

1. Introduction
2. Shallow Syntax in tweets
    - Part of Speech Tagging
    - Shallow Parsing
    - Capitalization
3. Named Entity Recognition
    - Segmenting Name Entities
    - Classifying Named Entities
    - Classification Experiments
4. Conclusion

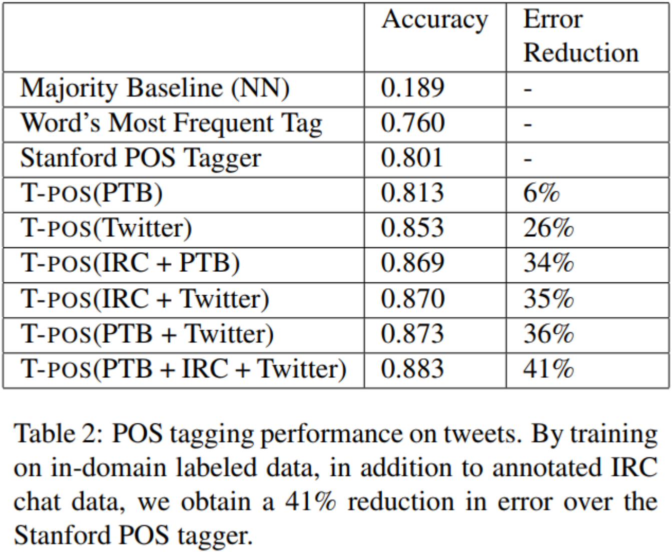
1 - Introduction



1	The Hobbit has FINALLY started filming! I cannot wait!
2	Yess! Yess! Its official Nintendo announced today that they Will release the Nintendo 3DS in north America march 27 for $250
3	Government confirms blast n nuclear plants n japan...don't knw wht s gona happen nw...

Table 1: Examples of noisy text in tweets.

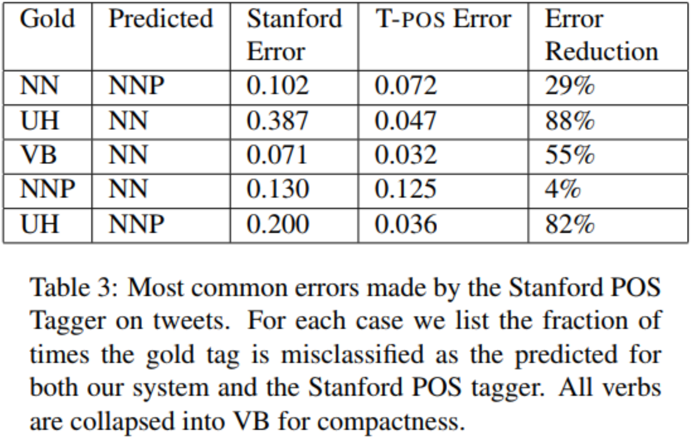
2 - Shallow Syntax in Tweets



	Accuracy	Error Reduction
Majority Baseline (NN)	0.189	-
Word's Most Frequent Tag	0.760	-
Stanford POS Tagger	0.801	-
T-POS(PTB)	0.813	6%
T-POS(Twitter)	0.853	26%
T-POS(IRC + PTB)	0.869	34%
T-POS(IRC + Twitter)	0.870	35%
T-POS(PTB + Twitter)	0.873	36%
T-POS(PTB + IRC + Twitter)	0.883	41%

Table 2: POS tagging performance on tweets. By training on in-domain labeled data, in addition to annotated IRC chat data, we obtain a 41% reduction in error over the Stanford POS tagger.

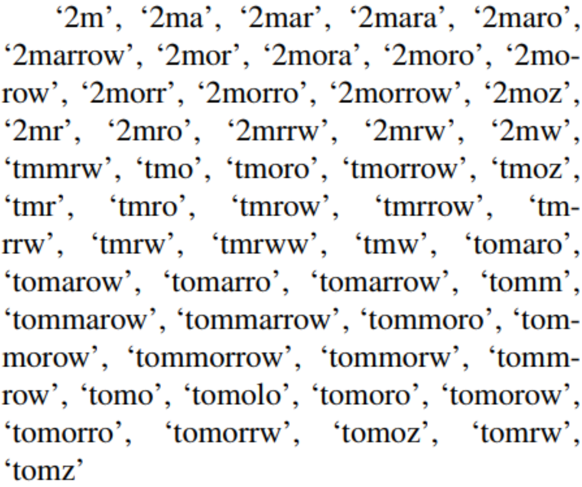
2.1 - Part of Speech Tagging



Gold	Predicted	Stanford Error	T-POS Error	Error Reduction
NN	NNP	0.102	0.072	29%
UH	NN	0.387	0.047	88%
VB	NN	0.071	0.032	55%
NNP	NN	0.130	0.125	4%
UH	NNP	0.200	0.036	82%

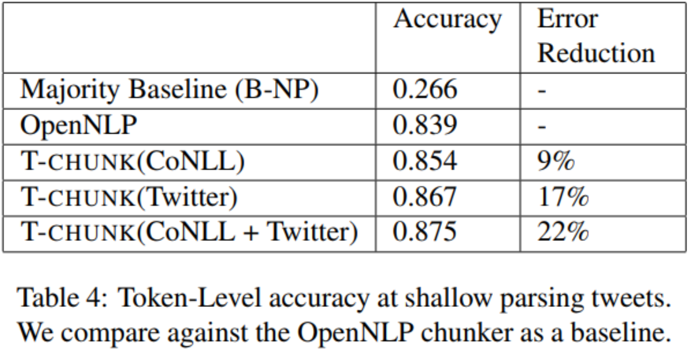
Table 3: Most common errors made by the Stanford POS Tagger on tweets. For each case we list the fraction of times the gold tag is misclassified as the predicted for both our system and the Stanford POS tagger. All verbs are collapsed into VB for compactness.

2.1 - Part of Speech Tagging



'2m', '2ma', '2mar', '2mara', '2maro', '2marrow', '2mor', '2mora', '2moro', '2morow', '2morr', '2morro', '2morrow', '2moz', '2mr', '2mro', '2mrrw', '2mrw', '2mw', 'tmmrw', 'tmo', 'tmoro', 'tmorrow', 'tmoz', 'tmr', 'tmro', 'tmrow', 'tmrrow', 'tmrrw', 'tmrw', 'tmrww', 'tmw', 'tomaro', 'tomarow', 'tomarro', 'tomarrow', 'tomm', 'tommarow', 'tommarrow', 'tommoro', 'tommorow', 'tommorrow', 'tommorw', 'tommrow', 'tomo', 'tomolo', 'tomoro', 'tomorow', 'tomorro', 'tomorrw', 'tomoz', 'tomrw', 'tomz'

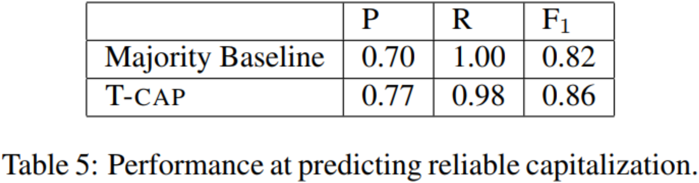
2.2 - Shallow Parsing



	Accuracy	Error Reduction
Majority Baseline (B-NP)	0.266	-
OpenNLP	0.839	-
T-CHUNK(CoNLL)	0.854	9%
T-CHUNK(Twitter)	0.867	17%
T-CHUNK(CoNLL + Twitter)	0.875	22%

Table 4: Token-Level accuracy at shallow parsing tweets. We compare against the OpenNLP chunker as a baseline.

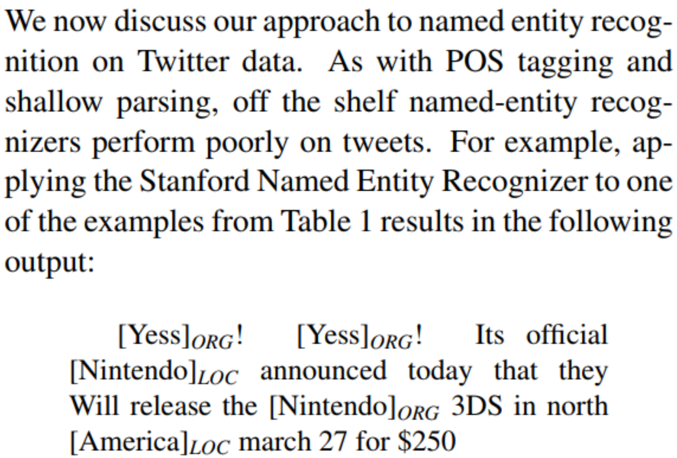
2.3 - Capitalization



	P	R	F_1
Majority Baseline	0.70	1.00	0.82
T-CAP	0.77	0.98	0.86

Table 5: Performance at predicting reliable capitalization.

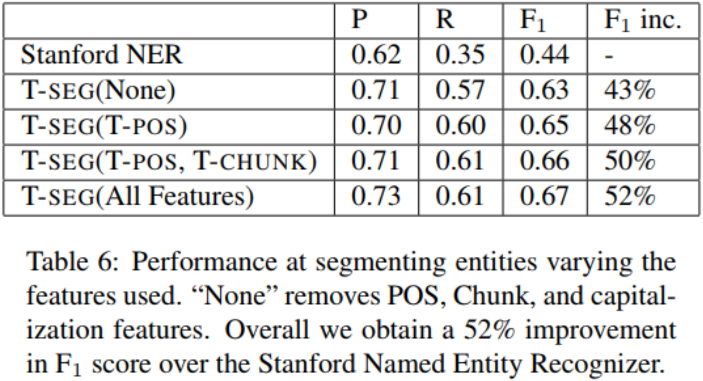
3 - Named Entity Recognition



We now discuss our approach to named entity recognition on Twitter data. As with POS tagging and shallow parsing, off the shelf named-entity recognizers perform poorly on tweets. For example, applying the Stanford Named Entity Recognizer to one of the examples from Table 1 results in the following output:

> [Yess]$_{ORG}$! [Yess]$_{ORG}$! Its official [Nintendo]$_{LOC}$ announced today that they Will release the [Nintendo]$_{ORG}$ 3DS in north [America]$_{LOC}$ march 27 for \$250

3.1 - Segmenting Name Entities



	P	R	F_1	F_1 inc.
Stanford NER	0.62	0.35	0.44	-
T-SEG(None)	0.71	0.57	0.63	43%
T-SEG(T-POS)	0.70	0.60	0.65	48%
T-SEG(T-POS, T-CHUNK)	0.71	0.61	0.66	50%
T-SEG(All Features)	0.73	0.61	0.67	52%

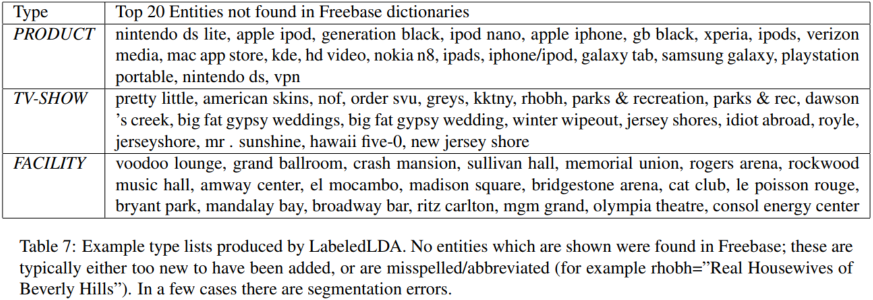
Table 6: Performance at segmenting entities varying the features used. "None" removes POS, Chunk, and capitalization features. Overall we obtain a 52% improvement in F_1 score over the Stanford Named Entity Recognizer.

3.2 - Classifying Named Entities

- KKTNY in 45min..........

Freebase Baseline: Although Freebase has very broad coverage, simply looking up entities and their types is inadequate for classifying named entities in context (0.38 F-score, 3.2.1). For example, according to Freebase, the mention China could refer to a country, a band, a person, or a film. This problem is very common: 35% of the entities in our data appear in more than one of our (mutually exclusive) Freebase dictionaries. Additionally, 30% of entities mentioned on Twitter do not appear in any Freebase dictionary, as they are either too new (for example a newly released videogame), or are misspelled or abbreviated (for example mbp is often used to refer to the mac book pro).

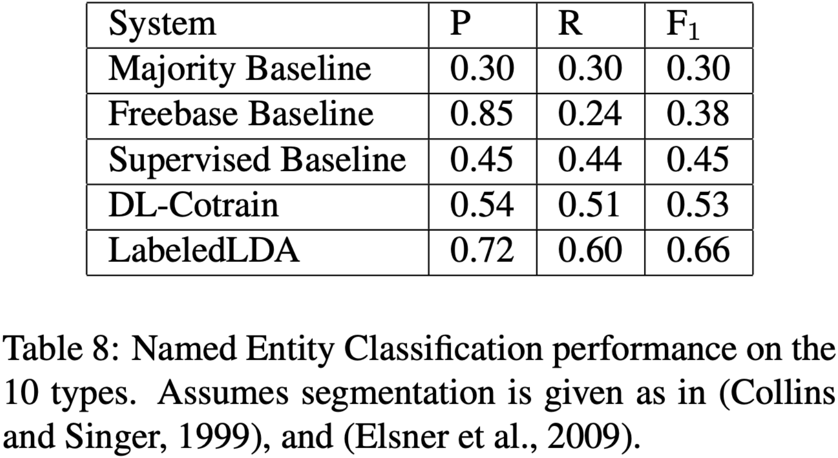
3.2 - Classifying Named Entities



Type	Top 20 Entities not found in Freebase dictionaries
PRODUCT	nintendo ds lite, apple ipod, generation black, ipod nano, apple iphone, gb black, xperia, ipods, verizon media, mac app store, kde, hd video, nokia n8, ipads, iphone/ipod, galaxy tab, samsung galaxy, playstation portable, nintendo ds, vpn
TV-SHOW	pretty little, american skins, nof, order svu, greys, kktny, rhobh, parks & recreation, parks & rec, dawson 's creek, big fat gypsy weddings, big fat gypsy wedding, winter wipeout, jersey shores, idiot abroad, royle, jerseyshore, mr . sunshine, hawaii five-0, new jersey shore
FACILITY	voodoo lounge, grand ballroom, crash mansion, sullivan hall, memorial union, rogers arena, rockwood music hall, amway center, el mocambo, madison square, bridgestone arena, cat club, le poisson rouge, bryant park, mandalay bay, broadway bar, ritz carlton, mgm grand, olympia theatre, consol energy center

Table 7: Example type lists produced by LabeledLDA. No entities which are shown were found in Freebase; these are typically either too new to have been added, or are misspelled/abbreviated (for example rhobh=”Real Housewives of Beverly Hills”). In a few cases there are segmentation errors.

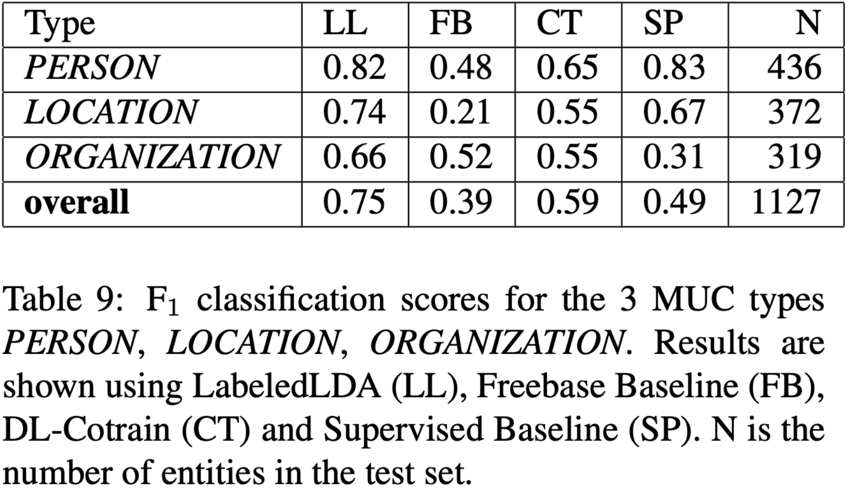
3.2.1 Classification Experiments



System	P	R	F_1
Majority Baseline	0.30	0.30	0.30
Freebase Baseline	0.85	0.24	0.38
Supervised Baseline	0.45	0.44	0.45
DL-Cotrain	0.54	0.51	0.53
LabeledLDA	0.72	0.60	0.66

Table 8: Named Entity Classification performance on the 10 types. Assumes segmentation is given as in (Collins and Singer, 1999), and (Elsner et al., 2009).

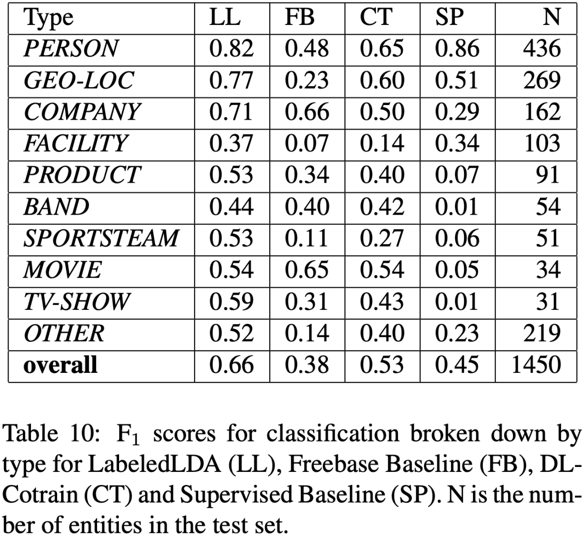
3.2.1 Classification Experiments



Type	LL	FB	CT	SP	N
PERSON	0.82	0.48	0.65	0.83	436
LOCATION	0.74	0.21	0.55	0.67	372
ORGANIZATION	0.66	0.52	0.55	0.31	319
overall	0.75	0.39	0.59	0.49	1127

Table 9: F_1 classification scores for the 3 MUC types *PERSON*, *LOCATION*, *ORGANIZATION*. Results are shown using LabeledLDA (LL), Freebase Baseline (FB), DL-Cotrain (CT) and Supervised Baseline (SP). N is the number of entities in the test set.

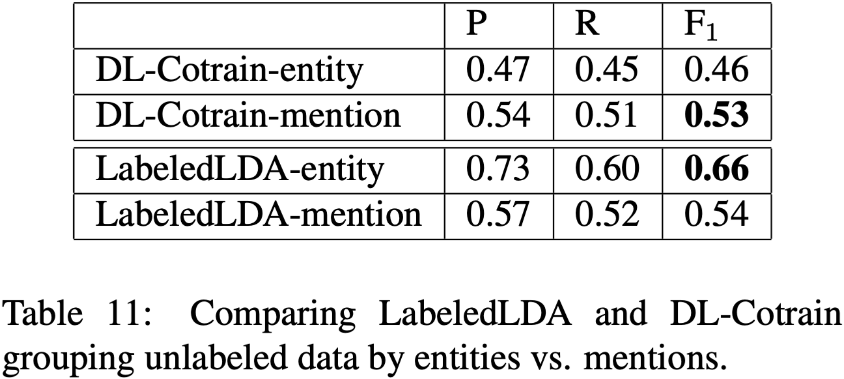
3.2.1 Classification Experiments



Type	LL	FB	CT	SP	N
PERSON	0.82	0.48	0.65	0.86	436
GEO-LOC	0.77	0.23	0.60	0.51	269
COMPANY	0.71	0.66	0.50	0.29	162
FACILITY	0.37	0.07	0.14	0.34	103
PRODUCT	0.53	0.34	0.40	0.07	91
BAND	0.44	0.40	0.42	0.01	54
SPORTSTEAM	0.53	0.11	0.27	0.06	51
MOVIE	0.54	0.65	0.54	0.05	34
TV-SHOW	0.59	0.31	0.43	0.01	31
OTHER	0.52	0.14	0.40	0.23	219
overall	0.66	0.38	0.53	0.45	1450

Table 10: F_1 scores for classification broken down by type for LabeledLDA (LL), Freebase Baseline (FB), DL-Cotrain (CT) and Supervised Baseline (SP). N is the number of entities in the test set.

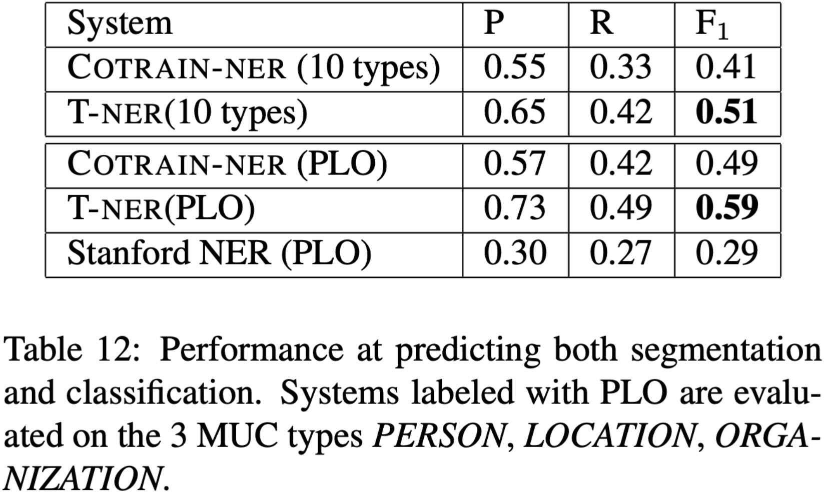
3.2.1 Classification Experiments



	P	R	F_1
DL-Cotrain-entity	0.47	0.45	0.46
DL-Cotrain-mention	0.54	0.51	**0.53**
LabeledLDA-entity	0.73	0.60	**0.66**
LabeledLDA-mention	0.57	0.52	0.54

Table 11: Comparing LabeledLDA and DL-Cotrain grouping unlabeled data by entities vs. mentions.

3.2.1 Classification Experiments



System	P	R	F_1
COTRAIN-NER (10 types)	0.55	0.33	0.41
T-NER(10 types)	0.65	0.42	**0.51**
COTRAIN-NER (PLO)	0.57	0.42	0.49
T-NER(PLO)	0.73	0.49	**0.59**
Stanford NER (PLO)	0.30	0.27	0.29

Table 12: Performance at predicting both segmentation and classification. Systems labeled with PLO are evaluated on the 3 MUC types *PERSON*, *LOCATION*, *ORGANIZATION*.

The End