Identified NER Errors + Some Examples

Onur Kara

May 1, 2020

NOTE: - This list is very much incomplete when compared to NER error analysis suite's additional error types/classes. There are only 30 or so types of errors listed here which are those I've taken note of over the past 3 years across a variety of general news and entertainment corpora (news outlets and clients). Thanks to this summers class of interns, the NER error analysis suite has upward of 60-80 errors classified and provides various modes of performing analysis on NER results to a great degree of depth/granularity if needed. Usages cases range from the analysis of errors resulting from some newly trained language model to domain specific analysis of the performance/accuracy of an existing 'good' model when used across arbitrary corpora. System can easily be extended to language agnostic.

Error Types

Disambiguation Errors
Mr. Smith, Joe Smith Joe Smith PhD
Galaxy S10, S10
Federal Bureau of Investigation, FBI

Noun-Entity Ambiguity
Occurs when entity is a homograph of a noun
i.e. plural common noun 'jobs' and surname 'Jobs'
So if Jobs is picked up and is 1st token in sentence and not in ground truth (not a fragment of actual ground entity) then it's of the following type
→ *Occurs when non-entity is start of sentence of start of internal quote*

Punctuation Related Errors
John B.B. Smith, Joe Smith Joe Smith PhD
B.B. King
L. Scott Fitz where L. could easily lead to sentence termination

Noun-Entity Concatenation
Again, start of sentence or with quote
e.g. end sentence. First Baptist Western Church lost its roof then the walls caved in.
Where "First" is incorrectly picked up

Fragments Numerical
e.g. end of sentence. Samsung Galaxy is now, what was the $\underbrace{\textit{iPhone}}_{\textit{easily fragmented '6 years'}}$ *6 years ago.*
e.g. start sentence. First Baptist Western Church lost its roof then the walls caved in.
Where "First" is incorrectly picked up

Company/Product Concatenation *e.g. The new Samsung Galaxy is great.*
not uncommon for some of the form "Samsung Galaxy" to be picked up rather than Samsung (company) and "Galaxy" product which would be the correct tagging

Start of Sentence Fragmentation
e.g. The Galaxy is great.

Conjunctive Adverb Concatenation
Similar to noun-entity concatenation but we take steps preprocessing to account for the majority of cases

Interior Entity Fragmentation
Capitalized tokens separated by preposition or conjunction incorrectly pick up individually rather than together

Interior Entity Concatenation
Capitalized tokens separated by preposition or conjunction but incorrectly concatenated Brazil and China

Numerical Token Fragmentation
e.g. Last June 3 6 Mafia released their new record..
Correct entity is 3 6 Mafia, however the context of Data messes up statistical model.
e.g. Andre 3000 years later was no longer a superstar.
Correct entity is Andre 3000

Numerical Token Concatenation
e.g. Joe Tomlin came in 1st, Elvin Hayes came in 2nd and Harold Smith 3rd.
Harold Smith is the correct entity.

Contraction Errors
Johnny's Pizzerea Fantastico is fantastic
Johnny's Pizzerea Fantastico is fantastic
see hand-written document for additional context and solutions

Location Split by Comma Fragment
e.g. was in Chicago, Illinois, a great place to live
correct entitty is Chicago, Illinois

Multi-Comma list with Complex Entities
e.g. The comma separated list Federation of the Best, Nation of Islam, Chicago, Wilmington, North Carolina, The Alliance for a Green and Dynamic Africa, and America is difficult.
This is extremely difficult and there are clearly various failure modes that exist in this single example

Multi-Comma Single Entity
e.g. The Office of Music, Entertainment and Cinema and Video
The Office of Music, Entertainment and Cinema and Video
The Office of Music, Entertainment and Cinema and Video
The Office of Music, Entertainment and Cinema and Video
The Office of Music, Entertainment and Cinematic and Video
where only the last one is correct and many other error fragments not listed above also exist

Title-Prefix Errors - Subset of Disambiguation Errors
e.g. John Jones
Captain John Jones
Capt John Jones
Mr. Jones

Title-Suffix Errors - Subset of Disambiguation Errors
e.g. John Jones
John Jones Republican
John Jones R

Sports Teams + Athletes
e.g. Athletico Madrid Arda Turan starts as winger.

Sports' Team-Nickname Disambiguation
e.g. Manchester United are on a roll. The Red Devils have won 9 of the last 10 games.
Manchester United are on a roll. Man. United have won 9 of the last 10 games.
Manchester United are on a roll. United have won 9 of the last 10 games.
Manchester United Red Devils are on a roll.

Athlete-Position Concatenation
e.g. Leo Messi F has scored at least one goal in his last 11 games.
Striker Leo Messi has scored at least one goal in his last 11 games.

Team-Score/Rank Concatenation
e.g. Barcelona 1st in the table are on a roll.
Barca won last time these two met, Barcelona 1 - Real Madrid 2 on matchday 2.

Mixed Casing Abbreviations Errors
e.g. Call of Duty: Modern Warfare 2
Call of Duty: Modern Warfare II
Call of Duty: MW2
Call of Duty: MWII
COD MW2
COD: MW2
Modern Warfare 2
CoD: Modern Warfare 2
etc...

Hyphen Induced Fragmentation
e.g. The Anglo-American Association

Hyphen Induced Concatenation
e.g. Tonights Bulls-Lakers game will be good.

Title-Colon Fragmentation
e.g. American Hero: A Story of Valor and Courage

List Colon-Entity Concatenation
e.g. Element 1: Cool Show is new to HBO
where only Cool Show should be considered.

List Colon-Entity Fragmentation
e.g. Element 1: Star of the Show is new to HBO
where only Star of the Show is the correct entity

Disease, Syndromes and other Domain Specific Entity Errors
e.g. The patient was diagnosed with Alzheimer's disease
The term to be picked up is Alzheimer's disease

Musician/Artist-Alias Disambiguation
e.g. Jay-Z
JayZ
Jayz
https://www.overleaf.com/project/5ef31e014ae9da0001bff938 Jay - Z
Jay Z
Jaÿ-Z
Jay-Hova
Hova
Hova da God
Jigga
S. Carter
Shawn C Carter
Shawn Carter
Shawn Corey Carter