Test Report

Analyze The Organization Name Recognition Tool spaCy & Dataset Annotations

Michelle Ning

April 14, 2020

Introduction

In order to analyze spaCy performance and its application in the specific organization name recognition, two datasets are adopted. One is generic test dataset(OntoNotes.json) and the other is target domain dataset(EnronSentences.json).

1. Model performance analysis

1.1 Performance Metrics on the Generic Test Dataset

There are totally 5000 emails with annotations in OntoNotes.json file. Each email has been annotated with organization name (ORG) and other types of named entities, such as PERSON, LOCATION, DATE, etc. Here only the organization name annotation is analyzed.

Following are the results after all emails in OntoNotes.json are checked with spaCy:

- 1397 emails are annotated with correct organization name at correct text position;
- organization names of 190 emails are annotated, but the start or end annotation position are not exactly correct;
- 3341 emails don’t have any annotated organization name, which is the same as the results from spaCy;
- in 11 emails, spaCy finds organization names with false positive error;
- in 61 emails, spaCy missed real organization names (false negative error).

Confusion Matrix is used for the accuracy evaluation. The Confusion Matrix table is listed below.

	Predict Positive	Predict Negative
Positive	True Positive	False Negative
Negative	False Positive	True Negative

In our case,

- True Positive = 1397 + 190 = 1587
- True Negative = 3341
- False Positive = 11
- False Negative = 61
- Accuracy = (True Positive + True Negative)/ Total = (1587 + 3341) / 5000 = 0.99
- Recall = True Positive/(True Positive + False Negative) = 1587/(1587+61) = 0.96

1.2 Error Analysis and Performance on the Target Domain Dataset

There are 310 emails with labeled organization named in EnronSentences.json file. The pre-annotated names by machine are checked and edited by human. Each email annotation is checked by three annotators. Hence there are two labeled results for each email.

Compareing between the labeled two results, it is easy to find:

- 59 emails are re-annotated by annotators;
- for other 251 emails, annotators agreed with the machine recognition results.

1.2.1 Analysis spaCy performance based on annotations by human

It is better to know:

- counts_name_found: ORG names are correctly annotated by machine,
- counts_name_missed: ORG names are missed by machine and annotated by annotators,
- counts_name_wrong: ORG names are deleted from machine annotation by annotators.

After searching the whole file, these three name lists are listed here.

counts_name_found listed the names and their frequency:

{'Enron': 17, 'ENA': 9, 'ISDA': 5, 'Nicor': 4, 'Texas': 3, 'Cinergy': 3, 'FYI': 3, 'VAR': 3, 'CEC': 2, 'Sierra': 2, 'GISB': 2, 'Central': 2, 'HPL': 2, 'MILLS': 2, 'Yahoo': 2, 'The': 2, 'EOTT': 2, 'NDA': 2, 'Harvard': 2, 'CPUC': 2, 'Amazon.com': 2, 'Master': 2, 'ProCaribe': 2, 'Socal': 2, 'Waddington': 2, 'Containerisation': 2, 'Model': 2, 'PG&E': 2, 'Legal': 2, 'VMAC': 2, 'ISO': 2, 'Real': 2, 'CNN': 2, 'NNG': 1, 'Westheimer': 1, 'Gingerbread': 1, 'Oxley': 1, 'East': 1, 'ICAP': 1, 'Stinson/Vince': 1, 'Client': 1, 'Confirmation': 1, 'Ken': 1, 'Principal-Protected*': 1, 'Trust': 1, 'Nasdaq-100': 1, 'Ambac': 1, "Moody's": 1, 'AAA': 1, 'Standard': 1, 'Enerfax': 1, 'GMT-06:00': 1, 'Global': 1, 'Seller': 1, 'PEPL': 1, 'TE': 1, 'Submit': 1, 'Christi': 1, 'Dominion': 1, 'ESA': 1, 'RAC': 1, 'Eastern': 1, 'PPT': 1, 'Supervisor': 1, 'SHRM': 1, 'CRRA': 1, 'TransPecos': 1, 'FBI': 1, 'CommodityLogic': 1, 'ASAP': 1, 'Peoples': 1, 'EFF_DT': 1, 'Jesse': 1, 'Specialist': 1, 'Logistics': 1, 'Associates': 1, 'Morton': 1, 'GCP': 1, 'HSC': 1, 'Kinder': 1, 'Stan': 1, 'West': 1, 'CSA': 1, 'CIAC': 1, 'Oil-NG-Hedge-Spec': 1, 'Wal-Mart': 1, 'SoCal': 1, 'Corp.': 1, 'NDA-Credit2B.com': 1, 'DealBench': 1, 'LPG': 1, '~2900': 1, 'Bracewell': 1, 'Marketing': 1, 'GPG': 1, 'Level': 1, 'Craver': 1, 'Reserve': 1, 'Financial': 1, 'Hearing': 1, 'GTC': 1, 'Delphi': 1, 'Bobinchuck': 1, 'Louis': 1, 'Travelocity.com': 1, 'Border': 1, 'Credit': 1, 'ECT': 1, 'IROQ': 1, 'CI': 1, 'Kenneth.Wong@gm.com': 1, 'Reuters': 1, 'Shelley': 1, 'Japanese': 1, 'EES': 1, 'Direct': 1, 'Truluck': 1, 'Kelly': 1, 'Board': 1, 'AEP': 1, 'Tracy': 1, 'Gallup': 1, 'Agave': 1, 'Dennis': 1, 'CLEC': 1, 'Swidler': 1, 'Unit': 1, 'Zone': 1, 'Kimberly': 1, 'Katherine': 1, 'Canadian': 1, 'Cordes': 1, 'Paul,': 1, 'Topock': 1, 'Lotus': 1, 'Lavarato': 1, 'DETM': 1, 'Groups': 1, 'Independent': 1, 'Corry': 1, 'OATI': 1, 'Enron01': 1, 'CFTC': 1, 'Attached': 1, 'NNG/TW': 1, 'WordPerfect': 1, 'Rich': 1, 'Kansas': 1, 'RisktRac': 1, 'Howzabout': 1, 'Business': 1, 'Weekly': 1, 'Theresa': 1, 'Weil': 1, 'Baker': 1, 'LiveLink': 1, 'LV': 1, 'Akilesh': 1, '\\\\nahou-psecn01v': 1, 'Essie': 1, 'Leon': 1, 'xferring': 1, 'Buffet': 1, 'Fastow': 1, 'Megawatt-Daily': 1, 'EOL': 1, 'US': 1, 'Rob': 1, 'Legislature': 1, 'Tri': 1, 'New': 1, 'Transwestern': 1, 'TW': 1, 'Minnis,': 1, 'Burlington': 1, 'EOT': 1, 'Christie,': 1, 'Suzanne': 1, 'BoozAllen': 1, 'Antitrust': 1, 'Nymex': 1, 'EIM': 1, 'CGAS': 1, 'Rosalee': 1, 'Intergen': 1, 'EDI': 1, 'GMC': 1, 'Franklin': 1, 'Illinois': 1, 'Dunn': 1, 'AG': 1, 'Zimin': 1, 'LA': 1, 'Morrow': 1, 'BPA': 1, 'Annette': 1, 'Lone': 1, 'Plant': 1, 'Contact': 1, 'Maclaren': 1, 'Allen': 1, 'Meredith': 1, 'Sports': 1}

counts_name_missed listed the names and their frequency:

{'PEPL': 3, 'PEP': 3, '6/13/01': 3, 'Client': 3, 'GE': 2, 'FERC': 2, 'Supervisor': 2, 'Enron': 2, 'K#66940': 2, 'California': 2, 'UBS': 2, 'Master': 2, 'ROFR': 2, 'ISO': 2, 'The': 2, 'EnronOnline': 1, 'ISDA': 1, 'CO2': 1, 'Henry': 1, 'BNP/EML': 1, 'Hector': 1, 'LPG': 1, '2600MT': 1, 'LNG': 1, 'LOI': 1, "L/C's": 1, 'L/C': 1, 'AEP': 1, 'MGI': 1, 'FYI': 1, 'Weekly': 1, 'MAC': 1, 'NBSK': 1, 'CLEC': 1, 'CLE': 1, 'EC': 1, 'IV': 1, 'FX': 1, 'GPG': 1, 'NBPL': 1, 'GISB': 1, 'INGAA': 1, 'BPA': 1, 'Mobil': 1, 'Outages': 1, 'Twanda': 1, 'What t': 1, 'ENA': 1, 'Pau': 1}

counts_name_wrong listed the names and their frequency:

{'FERC': 2, 'ROFR': 2, 'CEC': 1, 'ABB': 1, 'GE': 1, 'ISDA': 1, 'PEPL': 1, 'BNP/EML': 1, 'Mariella': 1, 'LPG': 1, 'PEP': 1, 'Supervisor': 1, 'January': 1, 'DPC': 1, 'LNG': 1, 'K#66940': 1, 'Interstate': 1, 'California': 1, 'Central': 1, 'LOI': 1, "L/C's": 1, 'L/C': 1, 'Harry': 1, '6/13/01': 1, 'MGI': 1, 'FYI': 1, 'Master': 1, 'Weekly': 1, 'NBSK': 1, 'CLE': 1, 'Underwriting': 1, 'IV': 1, 'FX': 1, 'GPG': 1, 'NBPL': 1, 'Veroinca': 1, 'GISB': 1, 'GMT-06:00': 1, 'INGAA': 1, 'BPA': 1, 'Client': 1, 'Outages': 1, 'BB': 1, 'NGX': 1, 'ENA': 1, 'The': 1}

Here, if words in **counts_name_found** are considered as true positive cases, and words in **counts_name_missed** list are considered as false positive cases, then we could get recall of the whole file.

Through building two dictionaries based on these two lists, there are 203 names labeled by spaCy and 49 names missed by spaCy but annotated by human. Hence, based on annotators' results, for EnronSentences.json file, the recall of spaCy is:

$$\text{Recall} = \text{True Positive}/(\text{True Positive} + \text{False Negative}) = 230 / (203 + 49) = 0.81$$

1.2.2 Error Analysis

Analyze False Negative Cases

It is important to know what spaCy missed, hence, I focused on **counts_name_missed** list first. From **counts_name_missed** list, it is found:

(1) Some names in **counts_name_found** also appears in **counts_name_missed** list. They are:

- {'Master', 'PEPL', 'GISB', 'Client', 'CLEC', 'The', 'Weekly', 'FYI', 'Enron', 'LPG', 'GPG', 'AEP', 'Supervisor', 'BPA', 'ISDA', 'ENA', 'ISO'}

This means for some names, spaCy can annotated correctly in some emails, but it also can missed in some other cases. More detailed analysis over email content is needed to understand reasons.

(2) Some names are actual ORG name, but spaCy doesn't reorganized them, such as: 'GE'.

(3) Some words are not ORG name, but both spaCy and annotators labeled them as ORG, such as: 'Weekly', 'FYI', 'Supervisor'.

Analysis False Positive Cases

Then analyze names in **counts_name_wrong** list, which might help developers to remove some names from ORG dataset.

(1) spaCy annotated city name, date, time, etc. as ORG name,

- such as: 'California', 'January', 'GMT-06:00', 'Weekly'.

(2) spaCy annotated Capitalized words as ORG name,

- such as 'K#66940', 'Master', 'Supervisor', 'Underwriting', 'Outages', 'The'

2. *Analysis of Target Domain Dataset Annotations*

2.1 Analyze Annotations

Comparing the three lists, some observations are made below and require further clarifications or changes:

(1) Some names annotated by spaCy are consistent with those from annotators, however, sometimes they are deleted by annotators. Those are the words in both **counts_name_found** and **counts_name_wrong** list:

{'Master', 'PEPL', 'GMT-06:00', 'GISB', 'Client', 'The', 'Weekly', 'Central', 'FYI', 'LPG', 'GPG', 'Supervisor', 'CEC', 'ISDA', 'BPA', 'ENA'}

(2) Some names are added as ORG names by annotators, but sometimes they are removed from ORG name by annotators. Those are the words in both **counts_name_missed** and **counts_name_wrong** list:

{'Outages', 'K#66940', '6/13/01', 'CLEC', 'NBPL', 'Pau', 'GE', 'Twanda', 'CLE', 'Mobil', "L/C's", 'LOI', 'L/C', 'ENA', 'CO2', 'Weekly', 'Supervisor', 'BPA', 'ROFR', 'GISB', 'Hector', 'ISO', 'EnronOnline', 'The', 'MAC', 'MGI', 'INGAA', '2600MT', 'FERC', 'California', 'UBS', 'GPG', 'Client', 'BNP/EML', 'IV', 'LNG', 'Master', 'What t', 'PEPL', 'FYI', 'Enron', 'Henry', 'LPG', 'ISDA', 'AEP', 'NBSK', 'PEP', 'EC', 'FX'}

In addition, annotators might also make other mistakes. For example,

(1) some company names are removed from annotation by the annotator, like, 'ABB', which shouldn’t be because ABB is a true company name.

(2) Some non-ORG words are annotated incorrectly by the annotator, such as: '6/13/01', '2600MT', 'MAC'.

2.2 Suggestions for Annotation Guidelines

Based on the analysis of the target dataset, I have two suggestions for the annotation guidelines.

1. Give a ORG name list for annotators as a reference.

2. Give negative examples in annotation guidelines. May it could include:

- not all capitalized word is ORG
- city name, location name is not ORG
- people name is not ORG
- date, time is not ORG
- etc

2.3 Suggestions for Manual Annotation

Generally, manual annotation needs three annotators to annotate each sample. Based on EnronSentences.json file analysis results, some suggestions are listed here:

(1) It is better to give annotators original emails without providing any predicted annotation from spaCy's results, such that the annotators' decisions are not possibly affected by the spaCy's prediction. This approach ensures the annotators' labels are independent to the machine learning for accuracy validation purpose.

(2) For each example, it is better to show annotation results of all three annotators with their work ID, even they have the same results. The manual annotation quality may depend on the annotators skills.

- It is hard to make the same mistake by all three person.
- If they have different annotation, we can compare later and also can find out which annotator have high accuracy.

(3) If annotators prefer get the original emails with spaCy results, no matter they agree or disagree with these results, all three annotators should give their own annotation.

- Go through the whole json file, I observe that 17 annotators have worked in this file. It is better to know which three annotators worked on each example.
- There is an example in EnronSentences.json, and it is annotated by 9 annotators not 3.

(4) The structure of the file or format should be unified. And the suggested structure with original email, spaCy and human annotation is below:

```
{
   "text": "example email",
   "machine_entities": [
     [
       start_position1,
       end_position1,
       "ORG",
     ]
   ],
   "human_entities": {
     "work ID - 1": [
       [
         start_position1,
         end_position1,
         "ORG",
       ],
       [
         start_position2,
         end_position2,
         "ORG",
       ],
"work ID - 2": [
       [
```

```
        start_position2,
        end_position2,
        "ORG",
      ]
    ],
    "work ID - 3": [
      [
        start_position1,
        end_position1,
        "ORG",
      ],
      [
        start_position2,
        end_position2,
        "ORG",
      ],
      [
        start_position3,
        end_position3,
        "ORG",
      ]
    ],
  }
}
```

- Annotators just input their ORG results, no matter it the same as or different from machine results.
- If there is no ORG, then keep the input empty with work ID.