A Report On Yisong's Internship 2017 Summer

1.Introduction.

In this summer, we(NUS WING Group) collaborate with Intelllex(a legal search Start-up), to refine current legal search.
Our motivation is to enable lawyer to search by fact, not only by keyword.
Facts are like the *Parties* show up in cases(e.g., Jason Fred/ Macrohard Company), and the relation between parties(e.g., Non-Competitive Contract).
So we set workflow like this:

Named Entity Recognition -> Relation Extraction

In June and July, I did some exploration work.

- Use Spacy(a NLP tool) to extract Named Entities(we will denote by NE hereafter).
- Create an inverted dictionary for NEs and compound nouns.
- Do some statistics on it, find terms with high frequency.
- Try K-means on NE set, happy to find some cluster do make sense.
- Try LDA on NE set, also happy to find some interesting topics. Thanks Yanchuan for teaching me many math.

When it turned to August, we found the result we obtain didn't make much sense. It is due to Spacy's NER's performance is not good enough, then the data we feed K-means and LDA is not convincing enough.

I later tried to post-processing Spacy's result, but found this task quie subtle, because Spacy's NER is not rule-based, it use context to predict.

The annotated data was not ready at that time, therefore I can't use CRF to build a NER model, then I decide to implement a rule-based NER myself from scratch.

2. The Rule-Based NER.

(1) Overview

We basically support all NE types in Spacy's NER.

TYPE	DESCRIPTION
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FACILITY	Buildings, airports, highways, bridges, etc.

ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LANGUAGE	Any named language.

(Spacy’s NE types)

We firstly emphasize on PERSON, ORG and LOC types, and compare result with Spacy’s result(will discuss it later)

(2)Implementation of the Rule-Based NER

The most powerful tool I use is Regular Expression.

(i) We first extract entities that has very fixed form, like Dates, Laws, Plaintiff, Defendant, and LOC

-Date

```
(((?<!\d)((\d{2}|\d{4})\s|[0-9]\s){0,1}(January|February|March|June|July|August|September|October|November|December)(\s(\d{2}|\d{4}))*))
((\d{2}|\d{4})\s|[0-9]\s){1}May|April(\s(\d{2}|\d{4}))*|((\d{2}|\d{4})\s|[0-9]\s){0,1}May|April(\s[0-9]{4})
```

We extract patterns like this:

1st Sept 2017

-Laws

```
(([tT]he|[A-Z][A-Za-z-&’]+)\s)(([A-Z][A-Za-z-&’]*|of|the|and|for)(\s)*)*((\(([A-Za-z-&]|\s|,)*?\))(\s)){0,1}(([A-Z][A-Za-z-&’]*|of|the|and|for)\s)*([A-Z][a-z]*)
And
re.search(r'(?<![A-Za-z])(bill(s)*|ordinance(s)*|act(s)*|rule(s)*|constitution(s)*|statute(s)*|regulation(s)*|charter(s)*|code(s)*)(?![A-Za-z])', text.lower())
```

We extract patterns like this:

Company (Amendment) Bill,

Copyright Act
the Securities Industry Act and the Securities Industry Regulations
Parliament (Privileges , Immunities and Powers) Act

-Plaintiff & Defendant

```
((?<![A-Za-z])((?!((?<![A-Za-z])(LR|Should|Lastly|Firstly|Does|Do|Did|Can|Like|Facts|Consequ
ently|Because|Accordingly|Later|Before|Whether|From|Now|So|Under|Then|In|Neither|Bot
h|If|Re|Although|One|When|While|Since|As|After|SLR|Even|And|But|That|What)(?![A-Za-z]))
)([tT]he|([A-Z][a-z-]*|[A-Z][a-z]+-[A-Za-z]+)('s)*(')*)(\s)*)((([A-Z][A-Za-z]*|[A-Za-z]+-[A-Za-z]+)(
's|-)*(')*|&|of|-|bin|de|the|binti|and|for)(\s)*)*(\(((?!(\s)*(consd|folld|distd|refd))([A-Za-z-&]|\
s)*(?<!(consd|folld|distd|refd)))\)\s)*((([A-Z][a-z-]*|[A-Za-z]+-[A-Za-z]+)('s|-)*(')*|&|of|-|and|th
e|bin|de|binti|for)(\s)*)*(([A-Z][A-Za-z]*)('s)*(')*|No\s\d+|and
(an)*other(s)*)(?![A-Za-z])|[A-Z][A-Za-z]*)(?=\sv\s)
```

The Trick is that, Plaintiff and Defendant are always surrounded by a ‘ **v** ’

We extract patterns like this:

Times Publishing Bhd and others v Sivadas
Television Broadcasts Ltd and others v Golden Line Video & Marketing Pte Ltd

-Location

```
(\d+(st|nd|rd|th)*|[A-Z][A-Za-z]+|#|-|\s|,|(\(.{1,5}\)))*
And
re.search(r'(?<![A-Za-z\d])(road|rd|street|avenue|ave|blk|block|floor|condominium|condo)(
?![A-Za-z\d])
```

We found that, LOC(or address) contains patterns like No 7, Blk 343, and keywords list above.
We extract patterns like this:

Block 44 Lorong 5 Toa Payoh Unit 01 - 205
1st Floor , Mandarin Theatre , 535 , Kallang Bahru , Singapore 1233

(ii) We then use a very tolerable Regular Expression to mark all possible terms/phrase as target, and use a score system to predict which NE type it belongs.

To find the target, we have this Regular Expression:

```
(?<![A-Za-z])((?!((?<![A-Za-z])(Have|Had|Should|Could|Lastly|Firstly|Does|Do|Did|Can|Like|F
acts|Consequently|Subsequently|Because|Accordingly|Later|Before|Whether|From|Now|So|
Under|Then|In|Neither|Both|If|Re|Although|One|When|While|Since|As|After|SLR|Even|And|
But|That|What)(?![A-Za-z])))([tT]he|([A-Z][A-Za-z-]*|[A-Z][a-z]+-[A-Za-z]+)('s)*(')*)(\s)*)((([A-Z
][A-Za-z]*|[A-Za-z]+-[A-Za-z]+)('s|-)*(')*|&|of|the|-|bin|de|binti|and|for)(\s)*)*((\(([A-Za-z]|'|-|\
s)*?\))(\s)){0,1}((([A-Z][a-z-]*|[A-Za-z]+-[A-Za-z]+)('s|-)*(')*|&|of|-|and|bin|de|binti|for)(\s)*)*((
?!((?<![A-Za-z])(Rep|I|O|Should|Lastly|Firstly|Does|Do|Did|Can|Like|Facts|Consequently|Bec
ause|Accordingly|Later|Before|Whether|From|Now|So|Under|Then|In|Neither|Both|If|Re|Alt
```

hough|One|When|While|Since|As|After|SLR|Even|And|But|That|What)(?![A-Za-z])))([A-Z][A-Za-z]*)('s)*(')*)(?![A-Za-z])

In human language, we basically extract all continuous words start with Capitalized Letter.

We extract patterns like this:

Hong Kong -- GPE
Tan Tee Jim -- PERSON
Keh Kee Guan & Co -- ORG
Lorong Buangkok -- LOC

We then use a scoring system to find the confidence the NE suits for each type, and assign it with the type having the highest confidence.
In the scoring system, the we have 2 types of rules: keywords inside the target span, and keywords in the context.
The keywords inside the target span is very reliable, while not so reliable in the context. The window size is very hard to set. When the window size is small, some keywords in context will escape, when the window size is large, we have many false positive.

--Confidence for PERSON

-Keyword:
We have prefix like, Mr, Miss, Mrs, Dr, Prof
We have title like, Sgt, Insp, chairman, speaker…
We have a chinese name set, following such regular expression:

(a|b|c|d|e|f|g|h|i|j|k|l|m|n|o|p|q|r|s|t|w|y|ch|sh|ng|zh)(ingh|am|oo|wee|au|oong|ang|ek|hee|eong|eung|eng|ie|heng|eo|an|ai|ao|ap|ay|ak|eck|ee|ei|en|eok|eong|eow|ew|ey|him|i|ia|ian|iao|iang|iew|im|in|iok|ing|ng|nn|o|oh|ow|uen|ock|ok|ong|oo|oon|oor|siao|su|u|ua|un|uo|uan|ung|ui|uk|oi|um|wee|wei)|ng|ah|sye)

We also have a western name set, using census data from U.S. government.

-Shape of Word:
Stephen King J
A V Winslow
Single Cap word can contribute to confidence.

-Context:
We have context like:

- Death of *
- * quote/say/speak

--Confidence for ORG

-Keywords:
Pte, Ltd, Sdn, Bhd, Company, co, corp…
Bank, School, Insititutes…

-Context: We found a very useful context:

A Malaysian Company, Pt Bigbang, has a debt...
Actually Company and Pt Bigbang is coreference.

--Confidence for LOC
We also use context to predict more LOC types.
In left context, key phrase like,

go/drive/come/travel to

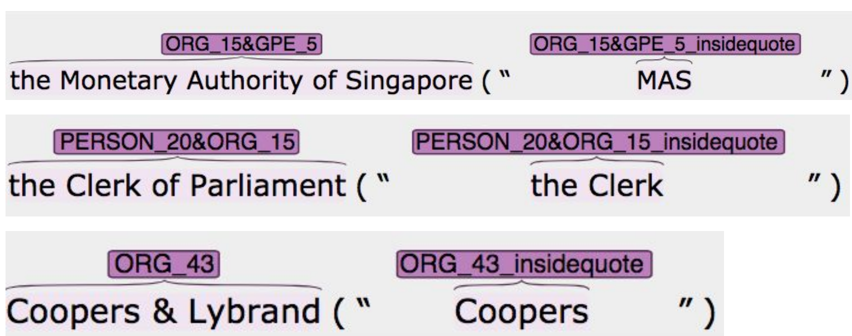
It does extract many hidden LOC entities out!

**Comment: we haven't found a good number of context, actually it is quite CRF, I'm waiting for the company side to feed me more annotated data, so that I can find more reliable contexts.*

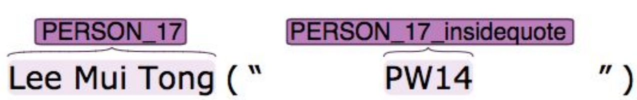
(iii) We further implement the detection for simple coreference, and annotate it the type when such coreference shows up again.
In our rule, they are what shows up in parenthesis or quote mark.

The coreference here: (1) abbreviation (2)legal terms

(1)abbreviation



(2)legal terms



(PW here means, Prosecuting Witness)

3. Evaluating the Rule-Based NER

* Yanchuan has given me several ‘golden’ annotated data, but I found many mistakes in it, like the NE is not annotated inside a case type.
*I will update this part later, after Yanchuan give me a ‘golden’ data with fewer mistakes.

(i) result on one ‘golden’ document with fewer mistakes, sg_cases_2883, the result goes:

	Precision	Recall	F1 Score
Yisong	0.95	0.80	0.87
Spacy	0.61	0.21	0.31

(For Person Type)

	Precision	Recall	F1 Score
Yisong	0.58	0.71	0.64
Spacy	0.38	0.5	0.43

(For ORG Type)

	Precision	Recall	F1 Score
Yisong	1	0.5	0.66
Spacy	0	0	0

(For LOC Type)

	Precision	Recall	F1 Score
Yisong	0.89	0.72	**0.80**
Spacy	0.48	0.20	**0.28**

(In Aggregate)

(ii) result on randomly picked ‘golden’ documents, some docs contain mistakes.(sg_cases_2883 sg_cases_407 sg_cases_47 sg_cases_80 sg_cases_102 sg_cases_147 sg_cases_177 sg_cases_192 sg_cases_203 sg_cases_240 sg_cases_52 sg_cases_556 sg_cases_66)

	Precision	Recall	F1 Score
Yisong	0.54	0.72	0.61
Spacy	0.49	0.35	0.41

(For Person Type)

	Precision	Recall	F1 Score
Yisong	0.24	0.74	0.61
Spacy	0.25	0.54	0.40

(For ORG Type)

	Precision	Recall	F1 Score
Yisong	0.47	0.74	0.57
Spacy	0.17	0.019	0.034

(For LOC Type)

	Precision	Recall	F1 Score
Yisong	0.373	0.725	**0.50**
Spacy	0.340	0.384	**0.36**

(In Aggregate)

Result Analysis:
Yisong’s performance is lower than in the previous case, due to low precision in ORG, I observe the golden documents, many ORG haven’t been annotated.

4. Future Plan

I plan to continue collaborating with Intelllex in long distance(I will be at Beijing), so that I can complete the plan we made at the very first beginning.

A basic timeline is like:
Sept: Use annotated data to build a **CRF-Based NER**, compare result with this Rule-Based NER.
Oct. - Dec.: After the entity recognition is done, finish **relation extraction**.

5. Acknowledgement

Thank you Prof Min, for giving me this opportunity to explore NLP, thanks for the meetings and your helpful advice. After hands-on NLP this summer, I still like it! (*good news, isn't it?*)

Thank you Dr Yanchuan, for coding tips, fixing my bugs, leading me to read few paper, deriving the math, and the drinks!

Thank you Dr-to-be Kishaloy, for always being there at the lab, answering my naive questions and offering quick advice!

Thank you.

6. References

Gibbs sampling in the generative model of Latent Dirichlet Allocation, Tom Griffiths
Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons, A McCallum, W Li