Rule-based Method
-- Named Entity Recognition

Fang Li

Dept. of Computer Science & Technology

Contents

- **Regular Expression**

- **Problems on Entity Identification**
- **Features of Entity Identification**
- **Rule-based method**

Regular Expressions

- A formal language for specifying text strings:
- ✓ A kind of pattern
- How can we search for any of these?
  - woodchuck
  - woodchucks
  - Woodchuck
  - Woodchucks

Regular expressions

Metacharacters: Twelve characters

the backslash \,
the caret ^,
the dollar sign $,
the period or dot .,
the vertical bar or pipe symbol |,
the question mark ?,
the asterisk or star *,
the plus sign +,
the opening parenthesis (,
the closing parenthesis),
the opening square bracket [,
the opening curly brace {.

Shorthand for character class:

\d: a digit.
\w: a "word character" (alphanumeric characters plus underscore)
\s: matches a whitespace character (includes tabs and line breaks).

Regular Expressions: Disjunctions

- Letters inside square brackets []

Pattern	Matches
[wW]oodchuck	Woodchuck, woodchuck
[1234567890]	Any digit

Pattern	Matches	
[A-Z]	An upper case letter	Drenched Blossoms
[a-z]	A lower case letter	my beans were impatient
[0-9]	A single digit	Chapter 1: Down the Rabbit Hole

Regular Expressions: Negation in Disjunction

- Negations [^Ss]
  - **Carat after the opening square bracket** negates the character class.

Pattern	Matches	
[^A-Z]	Not an upper case letter	Oyfn pripetchik
[^Ss]	Neither 'S' nor 's'	I have no exquisite reason"
q[^e]	Not e	Match qu in **qu**estion, but not match ira**q**
a^b	The pattern a carat b	Look up a^b now

More Disjunction

- The pipe | for disjunction

Pattern	Matches
yours|mine	yours mine
a|b|c	= [abc]
(cat |dog) food	cat food dog food

- **Repetition**: Use **curly braces** to specify a specific amount of repetition. Examples:

\b[1-9][0-9]{3}\b match a number between 1000 and 9999.
\b[1-9][0-9]{2,4}\b matches a number between 100 and 99999.

Regular Expressions: ? * + .

Makes the preceding token optional

Pattern	Matches	
colo**u**?r	Optional previous char	color colour
o**o***h!	0 or more of previous char	oh! ooh! oooh! ooooh!
o+h!	1 or more of previous char	oh! ooh! oooh! ooooh!
ba**a**+		baa baaa baaaa baaaaa
beg.n	a single character, except line break characters	begin begun begun beg3n

Anchors: ^ $

Anchors do not match any characters. They **match a position**

Pattern	Matches
^[A-Z]	Palo Alto
^[^A-Za-z]	1 “Hello”
\.$	The end.
.$	The end? The end! (dot match any any character)
\b	matches at a word boundary.

Example

- Find all instances of the word "the" in a text.

the Misses capitalized examples

[tT]he Incorrectly returns other or theology

[^a-zA-Z][tT]he[^a-zA-Z]

How to describe the regular expression of an email address?

\b[A-Z0-9._%+-]+@[A-Z0-9.-]+\.[A-Z]{2,4}\b

^[A-Z0-9._%+-]+@[A-Z0-9.-]+\.[A-Z]{2,4}$

Summarization for regular expression

- Regular expressions play a surprisingly large role
  - the first model for any text processing text
- Used in machine learning classifiers
  - as features, very useful in capturing generalizations

Is it possible to write "*an expression*" to identify a Named entity?

Named Entity Recognition

``*TWA has not been a normal company,'' said Robert Peiser, chief financial officer.*

We can not substantiate the claims.

Entities include:

- **Named entities**: *TWA, Robert Peiser …*
- **pronoun entities**: *we,…*
- **nominal entities**: the *company, …*

Named entities are the most important one among the 3 categories which is the anchor point for IE.

Difficulties of NE recognition

- Potential set of NE is too large to include in dictionaries/Gazetteers.
- Names changing constantly.
- Names appear in many variant forms. E.g. John Smith, Mr Smith or John
- Subsequent occurrences of names might be abbreviated.
- Ambiguity of NE types. E.g. John Smith is a person name or a company name? depends on:
- Internal structure: Mr. John Smith
- Context: The new company, John Smith will make….
- Ambiguity: Washington is a person or a location?

Features of Named Entities

According to its position in the text:

- **Features** that occur in the information unit itself, such as the composition of letters and digits of an entity name.
- **Features** that close neighborhood or context window of the token string to be classified.
- **Features** that occur in the complete document or document collection.

Features of Named Entities

According to their types

- Lexical: variations concerning punctuation (USA versus U.S.A), capitalization (e.g., Citibank versus CITIBANK)
- Syntactic: The part-of-speech of a word.词性
- Semantic: refer to semantic classifications of single- or multi-word information units.
- Discourse features refer to features computed by using text fragments, larger than the sentence.

Typical lexical features in a named entity recognition
(candidate entity name i that occur in the context window of l words)

FEATURE	VALUE TYPE	VALUE
Short type	Boolean	True if *i* matches the short type *j*; False otherwise.
POS	Nominal	Part-of-speech tag of the syntactic head of *i*.
Context word	Boolean or real value between 0 and 1; Or nominal.	True if the context word *j* occurs in the context of *i*; False otherwise; If a real value is used, it indicates the weight of the context word *j*. Alternatively, the context word feature can be represented as one feature with nominal values.
POS left	Nominal	POS tag of a word that occurs to the left of *i*.
POS right	Nominal	POS tag of a word that occurs to the right of *i*.
Morphological prefixes/suffixes	Nominal	Prefix or suffix of *i*.

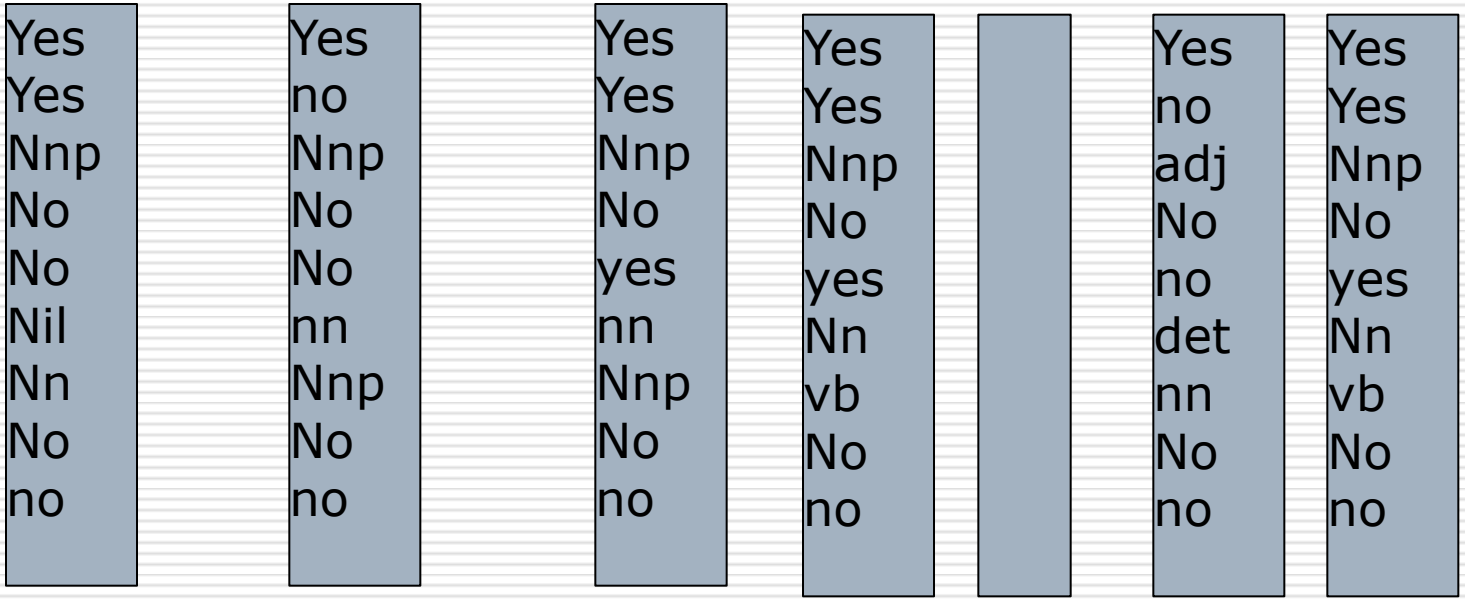
For example: person name identification

Microsoft spokesman John Smith is a popular man.

Features	Values	comments
Full string cap	True True	The first letter of each word is capitalized.
POS	nnp	Brown corpus: nn for singular common nouns,nns for plural common nouns, np for singular proper nouns.
Contain "Mr" | "Dr" Contain [CEO, CFO,spokesman,…] before	No Yes	Before and after the word
POS left	nn	
POS right	vb	
Morphological prefix or suffix	No no	Prefix: co (joint, with), pro(for, forward), re(again,back)

Each word is transferred as a feature vector.

- *Microsoft spokesman John Smith is a popular man.*



Features → Representation

- ***Feature selection:*** **use which features to identify person name.**
- ***Different methods have different features.***

- For rule-based method

to design Patterns or Rules

- For Machine Learning method

Features → features of models→ train

Features can have:

- ✓ Numeric values: discrete or real values
- ✓ Boolean value
- ✓ Nominal values: certain words
- ✓ Ordinal values: 0=small, 1=medium, 2=large
- ✓ Interval or ratio scaled values

Pattern vs. Rule

- Pattern (like regular expressions)

Price pattern(P):\b[1-9][0-9]{[0-9]}\.[1-9]{0-6}\b

Person name pattern (P1):姓+名

姓属于（姓名库），名：单字或双字

- Rule

If x match P then x is a **price**.

If x match P1 then x is a **person**.

Basic Steps for Named Entity Recognition

1. **Build linguistic patterns or rules to identify Entities or Relations**

"Dr.Yiming Yang was appointed as CEO of IBM at …"

"Smith was appointed as chairman of the account board. →

Pattern:

person **be appointed as** *post of company*

Basic Steps for Named Entity Recognition (cont.)

2. Apply rules or patterns to text and extraction

"Smith was appointed as Akim of Akmola region" →

Person: Smith

Post: Akim (head of local government)

Company: Akmola region ✗

Pattern Needs

- **general enough:** to have a broad applicability.
- **specific enough:** to be consistently reliable over a large number of texts.
- For example:

"*Person, post convinced company*" → ***too general***

"*company named person to post*" → ***too specific***

Difficulties to Collect the Patterns

- Different words:

  named, appointed, selected, chosen, promoted, …

- Different constructions:

  IBM named Fred president

  IBM announced the appointment of Fred as president

  Fred, who was named president by IBM

- Different names:

  George H. W. Bush, former President Bush, 41

Difficulties to Collect the Patterns (cont.)

- **Ambiguity**

Fred's appointment as professor vs.
Fred's 3 PM appointment with the dean

- **Complex structures**

For the Federal Election Commission, Bush picked Justice Department employee and former Fulton County, Ga., Republican chairman **Hans von Spakovsky** for one of three openings.

- **Reference**

George Garrick has served as president of Sony USA for 13 years. *The company* announced *his* retirement effective *next May*.

Who build patterns?

- Human experts
- Machine automatically learned from data.

Steps of Rule-based methods

- ☐ ***Use a lexicon to identify some named entities.***
- ☐ **Identify possible parts of names with lexical features**
- ☐ ***Write rules to recognize names***
  - Take advantage of capitalization
  - Take advantage of internal structure
  - ✓ Mumble Mumble City → probably a location
  - ✓ Mumble Mumble GmbH → probably a company
- ☐ ***Run over a corpus, find errors:***
  - ✓ General Electric is a company, not a general
  - ✓ Yesterday IBM Corp. announced …
- ☐ ***A large set of complex rules will be the result***

Use a lexicon to identify some named entities

- Advantages - Simple, fast, language independent, easy to retarget.
- Disadvantages - collection and maintenance of lists, cannot deal with name variants, cannot resolve ambiguity and include all.
- How to find a lexicon?

Using Gazetteer (Lexicon)

- **Online phone directories and yellow pages** for person and organization names
- U.S. census bureau: http://www.census.gov/genealogy/www/data/1990surnames/
- Locations lists:

  US GEOnet Names Server (GNS) data - 3.9 million locations with 5.37 million names

  http://earth-info.nga.mil/gns/html/
- The **World Gazetteer** provides a comprehensive set of population data and related statistics:

  http://www.world-gazetteer.com/
- http://www.fallingrain.com/world
- **Wikipedia , Linked data**

Write rules to recognize names

R1: if features then **person**

R2: if features then **location**

R3: if features then **organization**

Features like capitalization: (not enough)

- Full-string=U.S. → Location
- Full-string=I.B.M → organization

Lexical & Context Features

- **Set of spelling features**
  - Full-string=x (full-string=Maro Cooper)
  - Contains(x) (contains(Maco))
  - Allcap1 (IBM)
  - Allcap2 (N.Y.)
  - Nonalpha=x A.T.&T. nonalpha=.&.
- **Set of context features**
  - Context=x (context=president)
  - Context-type=x (prep or apposition)

Parsing-based Features

- **Has_Predicate**: from logical subject to verb

  e.g. *He said she would want him to join* →

  he: Has_Predicate(*say*), she: Has_Predicate(want), him: Has_Predicate(*join*)

- **Has_Amod**: from noun to its adjective modifier

  e.g. *He is a smart, handsome young man* → *man*: Has_AMod(*smart*)

- **Possess**: from the possessive noun-modifier to head noun

  e.g. *His son was elected as mayor of the city* → *his*: Possess(*son*), *city*: Possess(*mayor*)

Example: some rules for person

Possess(*wife*)→ PER
Possess(*brother*)→ PER
Possess(*daughter*) → PER
Possess(*bravery*) → PER
Possess(*father*) → PER

Has_Predicate(*divorce*) → PER
Has_Predicate(*remarry*) → PER

Some rules for Location

Possess(*concert_hall*) → LOC
Possess(*mayor*) → LOC

Has_AMod(*coastal*) → LOC
Has_AMod(*northern*) → LOC
Has_AMod(*eastern*) → LOC
Has_AMod(*northeastern*) → LOC

For example: Birthdate extraction

- ✓ George Washington was born in 1725.
- ✓ Washington was born on Feb. 12, 1725.
- ✓ Feb. 12 is Washington's birthday.
- ✓ Washington's birth date is Feb. 12, 1725.
- – George Washington was born in America.
- – Washington's standard was born by his troops in 1778.

Some Rules

- <Name> "was born" {"in"|"on"} <Date>
  => extraction (Name, Date)
- <Date> "is" <Name, possessive> "birthday"
  => extraction (Name, Date)
- <Name,possessive> "birth" "date" "is" <Date>
  => Extraction (Name, Date)

Pattern Models

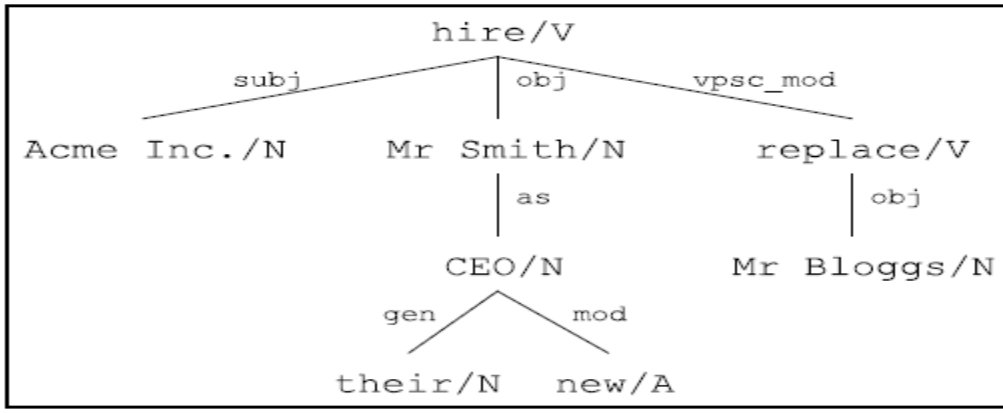
Predicate-Argument Model (SVO)

Chains: a path between a verb node and any other node in a dependency tree passing through zero or more intermediate nodes

Linked Chains: a pair of chains which share the same verb but no direct descendants.

Sub-tree: any subtree of a dependency tree can be used as an extraction pattern

Pattern Examples



SVO

[V/hire](subj[N/Acme Inc.]+obj[N/Mr Smith])
[V/replace](obj[N/Mr Bloggs])

Chains

[V/hire](subj[N/Acme Inc.])
[V/hire](obj[N/Mr Smith])
[V/hire](obj[N/Mr Smith](as[N/CEO]))
[V/hire](obj[N/Mr Smith](as[N/CEO](gen[N/their])))
[V/hire](obj[N/Mr Smith](as[N/CEO](mod[A/new]))))
[V/hire](vpsc_mod[V/replace])
[V/hire](vpsc_mod[V/replace](obj[N/Mr Bloggs]))
[V/replace](obj[N/Mr Bloggs])

Linked Chains

[V/hire](subj[N/Acme Inc.]+obj[N/Mr Smith])
[V/hire](subj[N/Acme Inc.]+obj[N/Mr Smith](as[N/CEO]))
[V/hire](obj[N/Mr Smith]+vpsc_mod[V/replace](obj[N/Mr Bloggs]))

Summarization

- ☐ *Regular Expression → pattern introduction*
- ☐ *How to identify named entities using rules?*
- ✓ *Find features: indicative, informative*
- ✓ *Build patterns : using many features.* ***Not too general, not too specific.***
- ✓ *Apply patterns (rules)*
- ☐ *Rule-based methods: build patterns by human beings.*

References:

- 沈嘉懿，李芳 “中文机构名称与简称的识别”中文信息学报 2007年11月
- 张小衡，王玲玲 “中文机构名称的识别与分析”中文信息学报
- 黄德根等 “基于SVM和CRF的双层模型中文机构名识别” 大连理工大学学报2010年9月

English Named Entity free Software

- Stanford NER (Java package, based on linear Chain Conditional Random Field)
- spaCy(https://spacy.io) implemented in Python
- Alias-i LingPipe (implemented in Java, supports both rule-based and supervised training method).
- Natural Language Toolkit (NLTK) (is a python NLP toolkit, based on Maximum Entropy Classifier).

Evaluation Corpus and Metrics

- Test Corpus: http://downloads.schwa.org/wikiner/wikigold.conll.txt

Entity Type	PER	LOC	ORG	MISC	Total
No.	931	1014	898	712	3555

- Evaluation
  - ✓ Exact matching
  - ✓ Partial Matching

Evaluation Results

		PER	LOC	ORG	OVERALL
Stanford	P	0.7195	0.7753	0.6992	0.7359
	R	0.8733	0.7416	0.4143	0.6813
	F	0.7890	0.7581	0.5203	0.7075
	PP	0.7496	0.8309	0.8083	0.7914
	PR	0.9098	0.7949	0.4788	0.7327
	PF	0.8220	0.8125	0.6014	0.7609
spaCy	P	0.7286	0.7321	0.3346	0.6110
	R	0.7325	0.6144	0.2873	0.5498
	F	0.7305	0.6681	0.3092	0.5788
	PP	0.7788	0.8085	0.5642	0.7240
	PR	0.7830	0.6785	0.4844	0.6514
	PF	0.7809	0.7378	0.5213	0.6858
LingPipe	P	0.4840	0.5067	0.2425	0.4026
	R	0.4211	0.4822	0.2806	0.3985
	F	0.4504	0.4941	0.2602	0.4005
	PP	0.6025	0.6052	0.4341	0.5412
	PR	0.5242	0.5759	0.5022	0.5357
	PF	0.5606	0.5902	0.4657	0.5384
NLTK	P	0.4802	0.4463	0.3115	0.4228
	R	0.7164	0.5493	0.3396	0.5378
	F	0.5750	0.4925	0.3249	0.4734
	PP	0.5587	0.4832	0.4883	0.5136
	PR	0.8335	0.5947	0.5323	0.6532
	PF	0.6690	0.5332	0.5094	0.5750

- Organization Recognition is a much hard task.
- Standford and Spicy show better performance than the other two's in this dataset.

Classroom Discussion

- How to identify named organization name ?

- 清远绿由环保科技有限公司主要从事固体废物无害化处置和资源化利用项目。在展台上，记者看到了工业污泥、陶瓷废渣处理加工制成的环保科技砖块，将废胶“变废为宝”制成的各类毛刷等。国家工信部是*中小企业*的*行政主管部门*，来自工信部的总工程师朱宏任介绍说，国家出台的这份文件既考虑解决了小型微型企业当前面临的生产经营困难，又注重引导*企业*增强内生动力，还提出了支持*企业*长期平稳健康发展的长效机制。说到*中小企业*创新的问题，兴业银行首席经济学家鲁政委认为，*中小企业*融资最难的时候已经过去了，恰恰相反，*中小企业*投融资难都得到了比较好的改善，所有*中小企业*最缺的是优质的客户，在技术、手段、平台上还需要更具体的、务实的创新。

Analysis the task (语法语义特性)
以机构特征词为中心语的定名词性短语

- 机构名称的组成：名称组成词 （前部判断）+公司特征词 （后缀判断）
- **公司特征词**是有限的，可以放在字典中，如国家机关名（部委），教育科研机构（大学），公共设施及场所（公园，体育馆），医疗机构，商业机构，社会组织，体育组织，体育组织，娱乐场所等。
- **名称组成词**包括：地名，人名（李宁体育公司），学科（电子科技），研究生产经营对象（五金工具批发市场/商店，软件研究所/公司），音译词（协和医院），创办，工作方式 （集团，股份）。

Analysis the task (组成规律)

机构名称::〈地名〉〈机构团体〉〈序数词〉〈人名〉〈专造名〉〈产品、对象〉〈功能/方式/等级〉〈学科/行业〉+〈机构特征词〉

For example:

长沙有色金属中等专科学校

香港第四广播电台

第一分校

Rule Deduction

- Org=[ModifierWord]*+[FeatureWord]
- FeatureWord=公司 | 大学 | 机构 | …
- POS (ModifierWord) = adj | np | nnp | nz | vn| …

1. [n|nz|…]{1,5}+ [Company|Corp.|Ltd.]
2. [ns|nz|…]{1,5}+[University|college|school]
3. [adj]+[Foundation|Agency]

Problems

- **边界识别**

"美国华盛顿大学"、"北京中央美术学院"

规则：如首词为地名，且后接有地名、人名、机构团体名或专造名，则该地名不能包括在高校名称中。

- **错误机构名称**

"美国女子大学已经由。。。""欧洲大学"

规则：修饰语不可以只含有国家名。

Rule-based method Implementation (实现方法)

1.找到第一个机构特征词;

2.根据相应规则往前逐个检查各词作为修饰词的合法性, 直到发现非法词;

3.如所接收的修饰词同机构特征词构成一个合法机构名称, 则分析记录该机构名称;

4.找下一个机构特征词, 如找到, 则跳至步骤2;

5.结束。

Other Solutions:

- Lexicon: keep famous company names
- Parsing-based Rules:

  Has_AMod(*advisory*) → ORG

  Has_AMod(*non-profit*) → ORG

  Possess(*ceo*)→ORG

  Possess(*operate loss*) → ORG

  Has_AMod(*multinational*) → ORG

  Has_AMod(*non-governmental*) → ORG
- Heuristic Clues:

- It is consecutive, not cross sentence boundary or any punctuation.
- It appears often.

Context Features

- Context word: {董事长|经理|发言人 }
- POS Left: {}
- POS right: {}

For example:

阿里巴巴董事长马云指出，互联网将会颠覆整个服务行业。

Open Questions Remains

- The boundary of Named Entities

上海交通大学**校友会**最近宣布一项新举措. √

上海交通有一项新举措 ×

- The abbreviations

上海**交**通**大**学→ 交大

华东**师**范**大**学→华师大

交通银**行**→交行

Submit report example

- Group member : XX,XX,XXX
- Aim: **organization name identification**
- Problem:
- Information collect:
- Method:
- Reason or assumption: