Natural Language Processing

Named Entity Recognition

Hutchatai Chanlekha

คณะวิศวกรรมศาสตร์

มหาวิทยาลัยเกษตรศาสตร์

Department of Computer Engineering, Kasetsart University

Named Entity Recognition

- Extracting entities of interest
  - In information extraction, one of the roles of NER is to filter candidate slot fillers
  - Introduced in MUC
    - PERSON, LOCATION, ORGANITION, DATE/TIME, MONEY/PERCENT
  - Extended to cover various area
    - Biomedical: protein, gene
    - Agricultural: plant, animal
    - Etc.
- Task
  - Identify position and boundary of NE
  - Recognize class of NE

Example

<ORG>The Royal Embassy of Saudi Arabia</ORG> issued a statement <DATE>Saturday</DATE> saying <PERSON>Masood</PERSON> visited <LOC>Saudi Arabia</LOC> from <DATE>November 2005</DATE> to <DATE>November 2006</DATE>

<PERSON>Sam Schwartz</PERSON> retired as executive vice president of famous hot dog manufacturer, <ORG>Hupplewhite Inc.</ORG>
He will be succeeded by <PERSON>Harry Himmelfarb</PERSON>.

<PERSON>พ.ต.อ. วรพงษ์ ภวเวส</PERSON> ผกก.สน.<LOC>วังทองหลาง</LOC>
เดินทางเข้าตรวจสอบบ้านร้างแห่งหนึ่งใน<LOC>ซอย ลาดพร้าว 62</LOC>

Named Entity Recognition

- Approaches for NER
  - Dictionary-based (Dictionary Matching)
    - Dictionary is one of an important knowledge sources for NER
    - Problem: Ambiguity when name is the same as common word, or as other NE class
  - Rule-based (Rule or Pattern Matching)
    - Suit for
      - NE with predictable pattern
      - structure or semi-structure documents
    - Problem: Not all occurrences of NE can be captured by rules
  - Statistical-Based
    - Machine learning approach
    - Need training corpus
  - Hybrid approach

Rule-based NER

- Various types of rules
  - Regular Expression
  - Heuristic rules
  - Automata
  - Other symbolic patterns
- Information used in the rules
  - Trigger words, such as person title, organization title, etc.
  - Name dictionary (i.e. gazetteers)
  - Linguistic information, such as POS, shallow semantic, etc.
  - Word feature, such as capital letter, character type, etc.
  - Prefix, suffix, infix
  - Name coreference

Example of rule-based system

- A Rule-based Named Entity Recognition System for Speech Input
  - http://mi.eng.cam.ac.uk/reports/svr-ftp/auto-pdf/kim_icslp2000.pdf
- RENAR: A Rule-Based Arabic Named Entity Recognition System
  - https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/RENAR.pdf
- LingPipe
  - http://alias-i.com/lingpipe/demos/tutorial/ne/read-me.html
- Rule-based Named Entity Recognition in Urdu
  - http://www.aclweb.org/old_anthology/W/W10/W10-2419.pdf
- Malay Named Entity Recognition Based on Rule-Based Approach
  - http://www.ijmlc.org/papers/428-LC038.pdf

Machine Learning for NER

GENERAL APPROACH

What is machine learning?

- Learn to improve automatically with experience
  - Learn from experience (e.g. examples, environments, etc.) to improve its performance in a certain task
  - Technically, it means class of programs that improve through experience.
- Application on Machine learning
  - Data mining: using historical data to improve decisions
    - Medical records -> medical knowledge
    - Weather information -> weather forecast
  - Software applications that can’t be programmed by hand
    - Autonomous driving
    - Speech recognition
    - NLP application
  - Self customizing programs
    - Newsreader that learns user interests

Why machine learning?

ML became possible …

- Progress in algorithms and theory
- Large number of online data
- High performance computing is available

Why using ML …

- Too difficult to program by hand
  - Such as too many features/attributes, relations between features/attributes are too complicated, etc.
- Portability → less human-expert time and effort
- Models are developed from data
  - Avoid bias from human developer, performance doesn't depend on human expertise

What is the learning problem?

- Learning = Improving with experience at some task
  - Improve over task ***T***
  - With respect to the performance measure, ***P***
  - Based on experience, ***E***

Examples of learning problems in NLP

- Part-of-speech tagging
  - Task T: recognizing part-of-speech of each word in a sentence
  - Performance measure P
    - percent of words with correctly recognized POS, etc.
  - Training experience E
    - POS-tagged sentences, etc.
- Sentiment analysis:
  - Task T: recognizing sentiment orientation of a clause
  - Performance measure P
    - percent of clauses that are correctly recognized sentiment, etc.
  - Training experience E
    - Set of clauses annotated with sentiment orientation, set of clauses
    - Set of clauses whose sentiment words were identified and annotated with orientation
    - etc.

Design Steps

- Choose training experience
- Choose the Target Function
  - Mapping from Data to Target Value
  - What will be your data and what will be your target value
- Choose Representation for Target Function
  - Features/attributes used for learning
  - Representation for target function
- Choose Learning Algorithm

STEP1: Choose the Training Experience

- Choose training experience
  - Type of training examples can have a significant impact on success or failure of the learner
    - Direct VS Indirect
    - Teacher or Not
  - The degree to which the learner controls the training examples
    - Training experience provided by a random process outside learner's control
    - Learners may pose various types of queries to an expert teacher
    - Learners collect training examples by autonomously exploring its environment
  - How well it represents the distribution of examples
    - Learning is most reliable when the training examples follow a distribution similar to that of future test examples
    - Quality of training experiences

STEP2: Choose the Target Function

- What type of knowledge will be learned?
- How this will be used by the program?
  - Direct VS Indirect
- Example:
  - Fn: token → Named entity class
  - Fn: sentence → sequence of part-of-speech
  - Fn: token (in the sentence) → token’s part-of-speech
  - Fn: sequence of characters → lattice representing word boundary

STEP 3: Choose Representation for Target Function

- Tradeoff in selecting choice of representation
  - Very expressive representation to allow representing as close an approximation as possible to the ideal target function V.
  - But… the more expressive the representation, **the more training data the program will require**.
- Example:
  - Collection of rules
  - Neural network
  - Decision hyperplane
  - Probability model

Step4: Choose Learning Algorithm

- Various machine learning techniques used in NLP
  - Naïve Bayes
  - Decision tree
  - Rule or pattern learning
  - Hidden Markov Model
  - Maximum Entropy
  - Conditional Random Field
  - Support Vector Machine
  - Neural Network
  - k-NN, k-mean, RBF, etc.
  - Etc.
- Many learning techniques need parameters setting

Problem Design

- Problem design
  - ML used in NLP are usually used to solve classification problem
    - Predict class of each example
  - Example of formulating NLP problems as classification problems
    - POS tagging → Classify token into its POS
    - IR → Classify document into its class
    - NE → Classify token (or seq. of tokens) into NE categories (or non-NE)
    - IE → Classify relation between pair of relevant entities
      - Another approach is to use rule learning algorithms, which is different from this setting.
    - Semantic tagging → Classify token into its sense

For NER

The Royal Embassy of Saudi Arabia issued a statement Saturday saying Khalid Masood visited Saudi Arabia from November 2005 to November 2006

พ.ต.อ. วรพงษ์ ภวเวส ผกก.สน. วังทองหลาง เดินทางเข้าตรวจสอบบ้านร้างแห่งหนึ่งในซอยลาดพร้าว 62

พ.ต.อ.| |วรพงษ์| |ภวเวส| |ผกก.| |สน.| |วังทองหลาง| |เดินทาง|เข้า|ตรวจสอบ|บ้าน|ร้าง|แห่ง|หนึ่ง|ใน|ซอย|ลาดพร้าว| |62|

- What is the target function?
  - Mapping from what to what?

Example

- Example of Named Entity Recognition as classification problem

<ORG>The Royal Embassy of Saudi Arabia</ORG> issued a statement <DATE>Saturday</DATE> saying <PERSON>Khalid Masood</PERSON> visited <LOC>Saudi Arabia</LOC> from <DATE>November 2005</DATE> to <DATE>November 2006</DATE>

- ○ Merge between NE position/boundary identification and NE categorization
- ○ Example: suppose N = {Person, Org, Loc}
  - Scheme I: *N*, *N*_start, *N*_cont, *N*_end, Other
  - Scheme II: *N*, *N*_in, *N*_out, Other

พ.ต.อ./p_start| /p_cont|วรพงษ์/p_cont| /p_cont|ภววเวสp_end| |ผกก.| |สน.|
|วังทองหลาง/loc| |เดินทาง|เข้า|ตรวจสอบ|บ้าน|ร้าง|แห่ง|หนึ่ง|ใน|ซอย/loc_start|
ลาดพร้าว/loc_cont| /loc_cont|62/loc_end|

Feature design for general ML problem

- Feature or Attribute
  - Describe data/example
    - called *feature vector*
    - [*Somchai*, *Mr.*, *said*, T, F, T, F, F]
    - characteristic observation
  - Fix number of features
    - Feature types must be predefined
    - Detailed enough to help the model in the decision task
  - Feature values
    - Not fix
      - Real value, string, …
    - Fix, predefined values
      - {a, b, c}, {True, False}, …

Feature Design (cont.)

- Select features or attributes describing training experience (training data)
  - Generally, features should be enough for training system to use for making decision
  - In NLP:
    - Internal token
    - Local context
    - Global context
    - External knowledge

What should be features for NER problem?

Statistical-based (ML) NER

- Features
  - Lexicon feature
    - token *w*
  - *N*-ary feature
    - previous *n* words, next *n* words, bigram, trigram, …
  - Word feature
    - Capital letter, number, contain special characters
  - Dictionary feature
    - Contain or is a word that appears in dictionary/word list
- Other possible features
  - Character
  - POS, Semantic
  - Prefix/suffix
  - Section (e.g. headline, body, preamble, etc.)
  - External systems
  - Etc.

Feature Encoding

- Feature encoding
  - depends on learning technique/tool
    - numeric (0,1,2,3,1.53, …)
      - For example: SVM
    - Any (0, 1, “person”, “unknown”, …)
      - For example: CRF
    - binary value, generally 0 or 1
      - For example: Maximum Entropy
    - First order logic (predicate and argument)
      - For example: FOIL
    - etc.

Discussion of feature encoding

- Suppose we have a real-valued attribute
  - How to preprocess this attribute to be used in ML technique that take predefined, discrete value as an input?
- Suppose we have an attribute with predefined, discrete value {small, medium, big}
  - How to preprocess this attribute to be used in ML technique that …
    - Take real-valued data as an input?
    - Take binary-valued data as an input?
- Suppose we have an attribute with predefined, discrete value {eat, Mr., said, can, has}
  - How to preprocess this attribute to be used in ML technique that …
    - Take real-valued data as an input?
    - Take binary-valued data as an input?

ML for NER

- Process
  - Corpus preprocessing
    - Zoning, Removing irrelevant parts of a document, etc.
    - Tokenization
    - Linguistic preprocessing, such as POS tagging, lemmatization, etc.
  - Feature extraction
    - Internal features
      - Features derived from the text, such as word feature, lexicon feature, contextual information, other appearances, etc.
    - External features
      - Features derived from external sources, such as dictionary, other NER systems, etc.
  - Generating feature vector for each token in the corpus
  - Select ML technique and train the model with the training data

Example: Feature Extraction

- Republican Sen. Marco Rubio of Florida was asked on Tuesday

w_{-2} w_{-1} w_0 w_{+1} w_{+2}

w_0	w_{-1}	w_{-2}	w_{+1}	w_{+2}	$InDict_1$ (w_0)	$InDict_1$ (w_1)	$InDict_2$ (w_{-1})	$InDict_3$ (w_0)	Cap (w_0)	Num (w_0)	Answer
Marco	Sen.	Republican	Rubio	of	1	1	1	0	1	0	Per_start

Dict1 (Name)
Anna
Bill
Mark
Rubio
Marco
…

Dict2 (Person title)
Mr.
Mrs.
Miss
Ms.
Dr.
Sen.

Dict3 (common word)
List of common words (i.e. words in general dictionary)