

Shallow Parsing, Named Entities, and Machine Learning

Adam Meyers
New York University



Outline

- Shallow Parsing
- What is a Named Entity?
- Converting Shallow Parsing Tasks to Sequence Labeling Tasks (like POS tagging)
- Applying Machine Learning Packages to Sequence Labeling Tasks



Shallow or Partial Parsing

- Finding constituents in a sentence, but not parsing the whole sentence
- Shallow Parsing identifies “short phrases”:
 - Noun Group and Verb Group Chunking
 - NE tagging
 - Identifying Time Expressions
 - Identifying other phrases in text



What is a Named Entity?

- Definition 1: A single or multi-word expression that meets any of the following criteria:
 - is a proper noun phrase
 - *Adam L. Meyers, PhD.*
 - *Professor Meyers*
 - *New York University*
 - is a proper adjective phrase, e.g., *Latin American*
 - has external distribution of NP, but different internal structure
 - January 3, 2012
 - Five Hundred Thirty
 - waffles@cs.nyu.edu
- Definition 2: A class of words and multi-word expressions defined by specifications tuned to information extraction tasks (can conflict with 1 by including “normal” nouns)
 - <http://nlp.cs.nyu.edu/ene/> is a large NE hierarchy following definition 2.



Annotating Names in Sample Documents

- Sample Documents to be annotated with Mae with name.dtd
 - State of the Union addresses by Obama and Trump
 - Einstein's Theory of Relativity
- Named Entity Definitions
 - GPE = location with government or set of GPEs
 - PER = person or set of people
 - ORG = organization, club, society, etc. – set of people with (governing) structure
 - Other = Word sequence widely recognized as name (somewhat slippery definition)
- Attempts at annotation of samples illustrate:
 - Difficulty of applying these criteria
 - Suggests need for more detailed specifications
 - Specifications may be influenced by goal of research, e.g., is the “name” of a scientific theory a type of “name”?



What is a Proper Noun (Phrase)?

- Definition: A name of something that is (in English) capitalized even in non-initial position, typically representing a unique individual object. Proper nouns don't typically take determiners.
- What's unique?
 - Is *Adam Meyers* a proper NP even though there are more than one person with that name?
 - Are *Thursday* or *September 3* proper NPs even though there are more than one instance of these days?
 - What about car models such as the *Fiesta* which represent a type of objects rather than a specific object?
 - Color terms, e.g., *azure*, *salmon*, *peach*, ... identify unique types, just like car models, yet they are not technically proper nouns
- Capitalization can be inconsistent
 - fields of study (like *computer science*) are capitalized inconsistently
 - different languages use different capitalization conventions



Internal Structure of Person Names

- $NP \rightarrow \text{First_Name}$
- $NP \rightarrow (\text{TitleP})?(\text{First_Name})? (\text{Middle_Name}|\text{Initial})?\text{Last_Name} (\text{Post_Honorific})?$
- $\text{TitleP} \rightarrow (\text{Mod})^* \text{Title}$
- $\text{Mod} \rightarrow \textit{vice} \mid \textit{assistant} \mid \textit{assist.} \mid \textit{deputy}, \dots$
- $\text{Title} \rightarrow \textit{Mr.} \mid \textit{Ms.} \mid \textit{Mrs.} \mid \textit{Miss} \mid \textit{Master} \mid \textit{Dr.} \mid \textit{President}, \dots$
- $\text{First_Name} \rightarrow \textit{Adam} \mid \textit{Jenny} \mid \textit{Joshua} \mid \textit{Nurit} \mid \textit{Giancarlo} \mid \textit{Ralph} \mid \textit{Cristina} \mid \textit{Satoshi} \mid \textit{Heng} \mid \textit{Xiang} \mid \textit{Shasha} \mid \textit{Wei} \mid \textit{Ang} \mid \textit{Bonan} \mid \dots$
- $\text{Last_Name} \rightarrow \textit{Meyers} \mid \textit{Matuk} \mid \textit{Lee} \mid \textit{Grishman} \mid \textit{Mota} \mid \textit{Sekine} \mid \textit{Ji} \mid \textit{Li} \mid \textit{Liao} \mid \textit{Xu} \mid \textit{Min} \mid \dots$
- $\text{Post_Honorific} \rightarrow \textit{Esq.} \mid \textit{Jr.} \mid \textit{Sr.} \mid \textit{I} \mid \textit{II} \mid \textit{III} \mid \textit{PhD.} \mid \dots$
- Note: specifications vary about whether titles and Post_Honorifics are or are not part of the name (ACE excludes titles, but includes post-honorifics)



Structure of Organization/Location/... Names

- Many Different Structures Possible
 - *Advanced Micro Devices* (ORG, normal NP)
 - *Council of Indian Nations* (ORG, normal NP)
 - *Yucatan Penninsula* (LOC, normal NP)
 - *United States of America* (GPE, normal NP)
 - *Ford Motors, Inc.* (ORG, NP plus right modifier)
 - *Alcoholics Anonymous* (ORG, NP plus right modifier)
 - *Head, Heart, Hands, Health* (list of nouns)
 - *Alfac* (ORG, newly coined single word)
 - *Addis Abba* (GPE, two foreign words)
 - *Merrill Lynch* (ORG, Person name structure)
 - *Nobody Can Beat the Wiz* (ORG, normal S)
 - *Hi Ho* (SONG, idiom)
- Unambiguous (like fixed phrases)
 - Name of ORG: *Advanced Micro Devices* (Advanced modifies Devices)
 - *[Advanced biology] textbook* vs. *Advanced [biology textbook]*



Some Other Entities

- Numbers and Quantities
 - twenty five thousand, five hundred fifty eight
 - \$200 million
- Times and Dates (not always names)
 - January 3, 2011
 - Ten o'clock
 - 10:30
 - last Thursday
 - St. Valentine's Day
- Addresses (street, email, url, ...)
 - 1313 Mockingbird Lane, New York, NY 10003
 - hm1313@cs.nyu.edu
 - <http://nlp.cs.nyu.edu/people/meyers.html>



ACE Named Entities

- ACE Specifications online (name mentions only)
 - <https://www ldc.upenn.edu/sites/www ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf>
- GPE – location with a government
 - city, state, county, country
 - people, physical location, government
 - *US, New York City, Queens, Greenwich Village*
 - *The US attacked Sweden*
 - *The US likes quacamole*
 - *The US is in North America*
- Location – geographical location
 - lake, mountain, natural structure
 - *Hudson River, Mt. McKinley, the Grand Canyon*



ACE Named Entities 2

- Facility – man-made structure
 - bridge, street, building
 - *The Brooklyn Bridge, 12 Street, The Forbes building*
- Person – person or group of people
 - *Adam Meyers, The Smiths* (meaning a group of people with the last name *Smith*)
- Organization – group of people with structure
 - commercial, government, club, non-profit
 - *New York University, the Glee Club, the NYU Fencing Team, New York Police Department, Alphabet, Inc., Google, Inc., The Dungeons and Dragons Club*



The ACE Task

- 2000-2008 Government-sponsored shared tasks (or bake-offs)
- Full Entity task
 - Annotation of mentions
 - Names, common noun, pronoun phrases that fall into the semantic classes (ultimately a superset of previous slide)
 - Coreference
 - Entity = Sets of mentions that refer to the same thing
- Other tasks
 - Relations: between two entities
 - located, part-whole, family, employment, ...
 - Events: entities are arguments of predicates
 - Movement, attack, be_born, marry, die, business_merge, declare_bankruptcy, ...
- Languages: English, Chinese, Arabic, (plus limited Spanish)



Some Historical Notes

- Before ACE, NEs were introduced in 1995 as part of the MUC6 government task
 - <http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html>
- The ACE task and several other NE tasks extended MUC6 in various ways.
- Other NE tasks, both government and SIG sponsored:
 - CONLL 2002-2003: English, Dutch, German, Spanish
 - IREX 1998-1999: Japanese (co-chairs: Sekine at NYU and Isahara at CRL)
 - SIGHan 2006: Chinese
 - TAC/KBP 2009 – Present: English (NIST)
- **For NE for final projects**
 - **NYUClasses (Resources) or Linguistic Data Consortium**



BIO Tagging and HMM

- HW 3 used an HMM to identify POS tags
- Property of POS tagging:
 - Each token has exactly one Tag
- BIO Tags
 - Provide a way to analyze short phrases using one tag per token
 - Example with Noun Group Identification
 - *The*/B-NG *blue*/I-NG *book*/I-NG *is*/O *in*/O *the*/B-NG *box*/I-NG ./O
 - BGs from annotated sentence: “*The blue book*” & “*the box*”
 - B-X = Beginning X, I-X = Inside X, O = Outside of constituent (or other)
 - Tags can be specific to phrase type: B-Per, I-Per, B-GPE, I-GPE, ... etc.



More Examples of BIO Tags

- NE Annotation:
 - *Adam*/B-PER *Meyers*/I-PER *is*/O *at*/O *New*/B-ORG *York*/I-ORG *University*/I-ORG ./O
- Noun Group Chunking:
 - *He*/B-NG *teaches*/O *NLP*/B-NG *in*/O *the*/B-NG *Department*/I-NG *of*/O *Computer*/B-NG *Science*/I-NG ./O
- Time Expressions
 - *It*/O *was*/O *10*/B-TMP *o'clock*/I-TMP *on*/O *Saturday*/B-TMP.
- Typically, only 1 task is covered at a time



Partial Parsing as Sequence Labeling

- POS tagging is a sequence labeling task
 - Assigning labels to tokens in a sequence
- We used an HMM for POS tagging
 - HMM uses the following “features” to predict POS:
 - Previous POS
 - Current word
- BIO tags (like POS) are labels on individual words
- Can we use an HMM or something like it to predict BIO tags and therefore short phrases?



Features that May Predict Short Phrases

- Previous BIO Tag
- Word Related: Previous word (or Beg_Sentence), 2 words previous, Current Word, Following Word (or End Sentence), 2 words ahead, etc.
- Previous POS, POS of 2 words previous, Current POS, Next POS, etc.
- Captial or lowercase properties, last letter of current word, class in dictionary class, member of word list, etc.



Can we use lots of features in an HMM, like our POS tagger?

- Nymble: an HMM-style NE tagger
 - Replaces words with sets of features about orthography of word, e.g., *TwoDigitNum*, *ContainsDigitandAlpha*, *allCaps*, *firstWord*, *initCap*, ...
 - Different probabilities for beginning, ending and inside elements of NE differently
 - OOV model based on 20% sample from corpus
- Bikel, et. al. (1996). *Nymble: A High-performance Learning Name-finder*. in ANPL 1997



If Lots of Evidence, Do Machine Learning

- Suppose you want to combine lots of features together and take advantage of any correlation to predict outcomes
- Methods for doing this fall into the area called machine learning
- HMM (and Nymble's approach) are ways of doing machine learning, but now we will discuss machine learning more generally
- Algorithm Used for Homework 5: *Maximum Entropy*
- Supervised or Unsupervised
 - **Supervised: Methods in which statistical models are “trained” based on manually annotated text.**
 - We will focus on these.
 - Unsupervised: Methods in which statistical models are based on assumptions about un-annotated data
 - Semi-supervised: Methods that combine supervised and Unsupervised



High Level Description of Supervised ML

- Input = Data correctly annotated with observable set of features
 - Training Corpus
 - Test or Development Corpus
- Machine Learning Algorithms
 - Methods for combining evidence and making predictions
- Toolkits for Multiple Machine Learning Algorithms
 - JAVA
 - OpenNLP maxent package: <http://maxent.sourceforge.net/howto.html>
 - Default for HW6
 - WEKA: <http://www.cs.waikato.ac.nz/ml/weka/>
 - MALLET: <http://mallet.cs.umass.edu/>
 - Python
 - NLTK's classification package (Chapter 6)
 - Scikit-learn: <http://scikit-learn.org/stable/>



Making and Tuning ML Systems

- Experiment with Different ML Algorithms
 - Use the same set of features
 - Toolkits make switching easy
 - May help to understand differences between algorithms
 - Speed/complexity → limit size of training data
 - Assumptions about Feature Independence
 - Tweaking features, making new algorithms and making new more efficient versions of current ML algorithms
- Experiment with Different Sets of Features
 - Keep algorithm fixed
 - Vary numbers of features
 - Possible strategy: use as many features as possible
 - When these systems work, It cannot always be explained why
 - Possible strategy: use features that can be expected to make a prediction
- Possible to make an excellent ML system while treating algorithms as black boxes

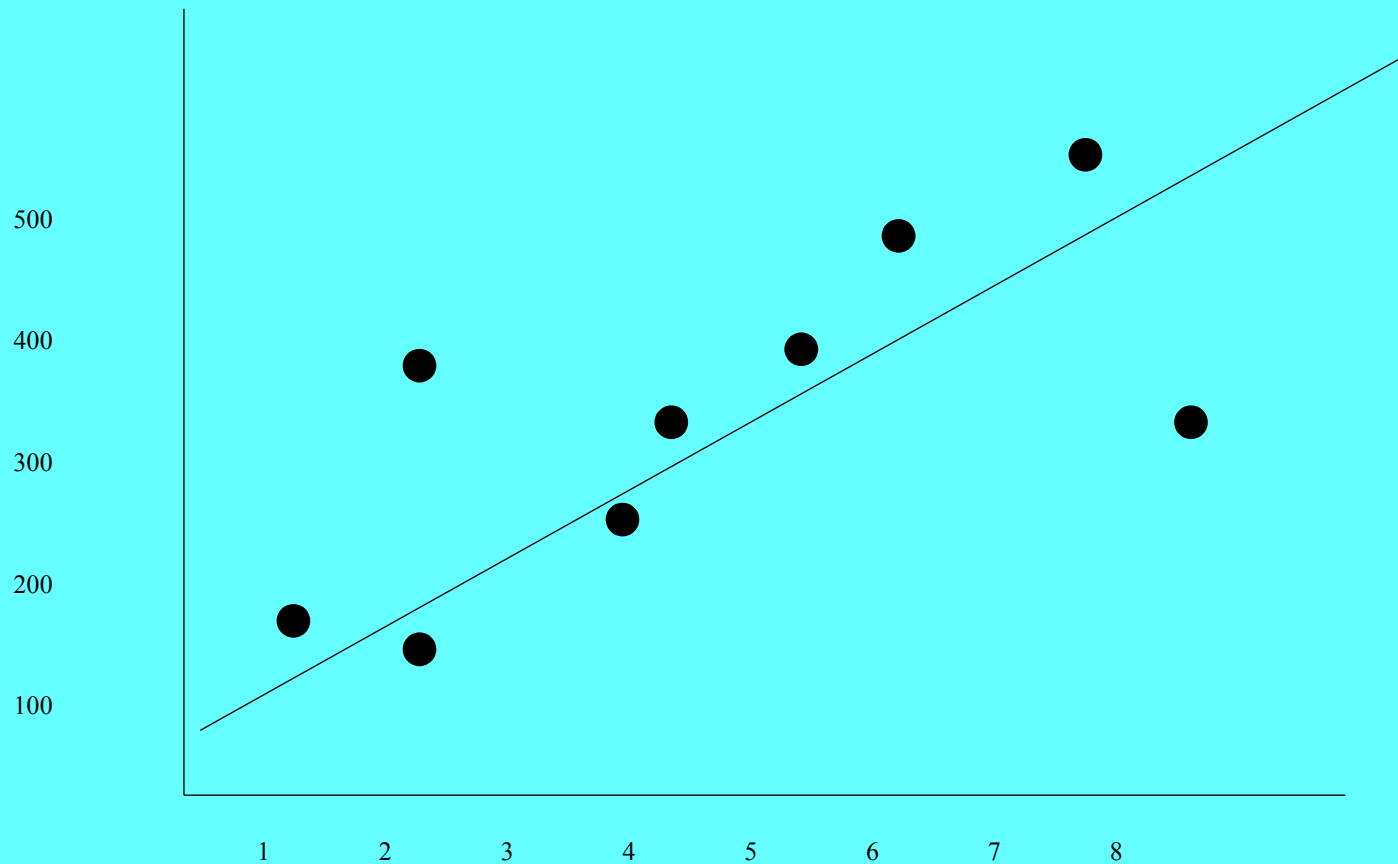


Regression Analysis

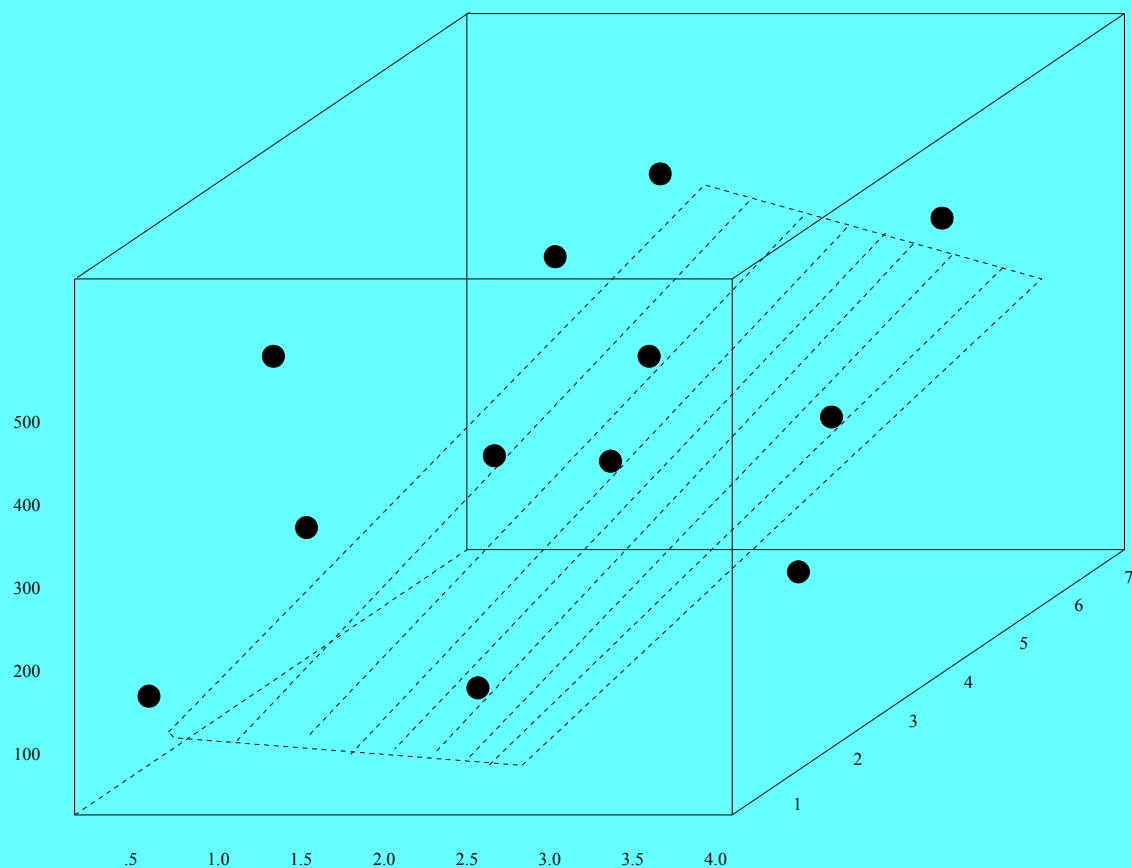
- Represent features as dimensions in a graph
- Approximate correlations using a figure with fewer dimensions
- 2 dimensions/features – approximate with a line (a 1 dimensional representation)
 - 1 feature “predicts” the other, e.g., height predicts shoe size
- 3 dimensions/features
 - approximate with a plane (2 dimensions) or
 - 2 features (height and age) predicts a 3rd feature (shoe size)
 - a line (1 dimension)
 - 1 feature (age) predicts 2 features (height and shoe size)
- Correlations Used to Predict Values in ML



Scatter Plot for 2 Features approximated by Regression Line:



Scatter Plot with 3 features approximated with a Regression plane



Machine Learning Algorithms

- Naive Bayes: Assumes that all features are independent of each other. Basically the probabilities of features in each category are multiplied together.
- Maximum Entropy: Combines features using weights that are adjusted via “smoothing”.
 - Normalizes result to a number between 0 and 1.
 - MEMM: Viterbi algorithm w/ Maximum entropy
- Other “traditional”: Support Vector Machines, Regression, Kernels, Conditional Random Fields, etc.
- Deep Learning: CNN, RNN, ...



Summary

- Named Entities: Classifications of names and sometimes other special noun phrases
- BIO Tags: Encoding Phrases as tags on tokens
- Supervised Machine Learning: Means of predicting a class in test data, given observed co-occurring features in training data



HW 5

- <https://cs.nyu.edu/courses/fall18/CSCI-UA.0480-006/priv/homework5.html>
- Due October 23, 2018

