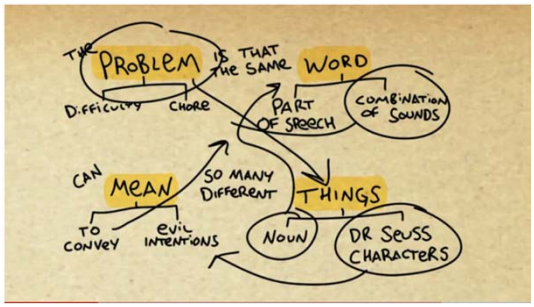
600.465 Connecting the dots - I (NLP in Practice)

Delip Rao
Click to edit Master subtitle style
delip@jhu.edu



What is "Text"?

Shall I compare thee to a summer's day?
Thou art more lovely and more temperate;
Rough winds do shake the darling buds of May,
And summer's lease hath all too short a date;
Sometime too hot the eye of heaven shines,
And often is his gold complexion dimm'd;
And every fair from fair sometime declines,
By chance or nature's changing course untrimm'd;
But thy eternal summer shall not fade,
Nor lose possession of that fair thou ow'st;
Nor shall Death brag thou wander'st in his shade,
When in eternal lines to time thou grow'st:
So long as men can breathe or eyes can see,
So long lives this, and this gives life to thee.

What is "Text"?

Shall I compare thee to a summer's day?

A relation-preserving isomorphism [edit]

If one object consists of a set X with a binary relation R and the other object consists of a set Y with a binary relation S then an isomorphism from X to Y is a bijective function $f : X \to Y$ such that

$$\mathrm{S}(f(u), f(v)) \iff \mathrm{R}(u, v)$$

S is reflexive, irreflexive, symmetric, antisymmetric, asymmetric, transitive, total, trichotomous, a partial order, total order, strict weak order, total preorder (weak order), an equivalence relation, or a relation with any other special properties, if and only if R is.

For example, R is an ordering $\leq$ and S an ordering $\sqsubseteq$, then an isomorphism from X to Y is a bijective function $f : X \to Y$ such that

$$f(u) \sqsubseteq f(v) \iff u \leq v.$$

Such an isomorphism is called an *order isomorphism* or (less commonly) an *isotone isomorphism*.

If $X = Y$ we have a relation-preserving automorphism.

An operation-preserving isomorphism [edit]

Suppose that on these sets X and Y, there are two binary operations $*$ and $\Diamond$ which happen to constitute the groups $(X, *)$ and $(Y, \Diamond)$. Note that the operators operate on elements from the domain and range, respectively, of the "one-to-one" and "onto" function f. There is an isomorphism from X to Y if the bijective function $f : X \to Y$ happens to produce results, that sets up a correspondence between the operator $*$ and the operator $\Diamond$.

$$f(u) \Diamond f(v) = f(u * v)$$

for all u, v in X.

What is "Text"?

"Real" World

- Tons of data on the web
- A lot of it is text
- In many languages
- In many genres

Language by itself is complex.
The Web further complicates language.

Click to edit Master subtitle style

11/30/11

NLP for fun and profit

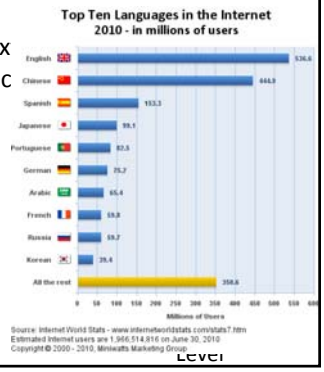
- Making NLP more accessible
  - Provide APIs for common NLP tasks

```
var text = document.get(…);
var entities = agent.markNE(text);
```

- **Big $$$$**
- Backend to intelligent processing of text

Desideratum: Multilinguality

- Except for feature ex
  be language agnostic

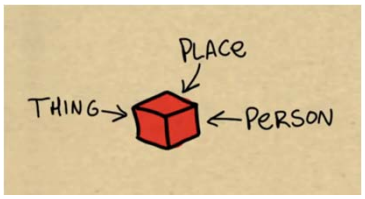


Level

In this lecture

- Understand how to solve and ace in NLP tasks
- Learn general methodology or approaches
- End-to-End development using an example task
- Overview of (un)common NLP tasks

Case study: Named Entity Recognition



Case study: Named Entity Recognition

- Demo: http://viewer.opencalais.com

- How do we build something like this?
- How do we find out well we are doing?
- How can we improve?

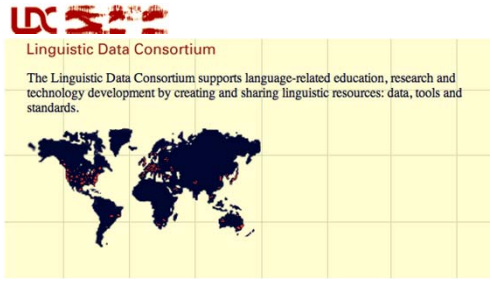
Click to edit Master subtitle style

11/30/11

Case study: Named Entity Recognition

- Collect data to learn from
  - Sentences with words marked as PER, ORG, LOC, NONE
- How do we get this data?

Pay the experts



Wisdom of the crowds

Getting the data: Annotation

- Time consuming
- Costs **$$$**
- Need for quality control
  - Inter-annotator aggreement
  - Kappa score (Kippendorf, 1980)
- Smarter ways to annotate
  - Get fewer annotations: Active Learning
  - Rationales (Zaidan, Eisner & Piatko, 2007)

Click to edit Master subtitle style

11/30/11

Our recipe …

- 1. Formalize some insights
- 2. Study the formalism mathematically
- 3. Develop & implement algorithms
- 4. Test on real data

NER: Designing features

Only
France
and
Great
Britain
backed
Fischler
's
proposal
.

Need to segment sentences

Tokenize the sentences

Preprocessing

- Not as trivial as you think
- Original text itself might be in an ugly HTML
- Cleaneval!

NER: Designing features

Only	IS_CAPITALIZED
France	IS_CAPITALIZED
and	
Great	IS_CAPITALIZED
Britain	IS_CAPITALIZED
backed	
Fischler	IS_CAPITALIZED
's	
proposal	
.	

NER: Designing features

Only	IS_CAPITALIZED	IS_SENT_START
France	IS_CAPITALIZED	
and		
Great	IS_CAPITALIZED	
Britain	IS_CAPITALIZED	
backed		
Fischler	IS_CAPITALIZED	
's		
proposal		

NER: Designing features

Only	IS_CAPITALIZED	IS_SENT_START
France	IS_CAPITALIZED	
and		
Great	IS_CAPITALIZED	
Britain	IS_CAPITALIZED	
backed		
Fischler	IS_CAPITALIZED	
's		
proposal		

NER: Designing features

Only	IS_CAPITALIZED	IS_SENT_START	
France	IS_CAPITALIZED		IN_LEXICON_LOC
and			
Great	IS_CAPITALIZED		
Britain	IS_CAPITALIZED		IN_LEXICON_LOC
backed			
Fischler	IS_CAPITALIZED		
's			
proposal			

NER: Designing features

These are extracted during preprocessing!

Only			START	
France	POS=			IN_LEXICON_LOC
and	POS=CC			
Great	POS=NNP	IS_CAPITALIZED		
Britain	POS=NNP	IS_CAPITALIZED		IN_LEXICON_LOC
backed	POS=VBD			
Fischler	POS=NNP	IS_CAPITALIZED		

NER: Designing features

Only	POS=RB	IS_CAPITALIZED	…	PREV_WORD=_NONE_
France	POS=NNP	IS_CAPITALIZED	…	PREV_WORD=only
and	POS=CC		…	PREV_WORD=france
Great	POS=NNP	IS_CAPITALIZED	…	PREV_WORD=and
Britain	POS=NNP	IS_CAPITALIZED	…	PREV_WORD=great
backed	POS=VBD		…	PREV_WORD=britain
Fischler	POS=NNP	IS_CAPITALIZED	…	PREV_WORD=backed

NER: Designing features

Only	POS=RB	IS_CAPITALIZED	…	PREV_WORD=_NONE_	…
France	POS=NNP	IS_CAPITALIZED	…	PREV_WORD=only	…
and	POS=CC		…	PREV_WORD=france	…
Great	POS=NNP	IS_CAPITALIZED	…	PREV_WORD=and	…
Britain	POS=NNP	IS_CAPITALIZED	…	PREV_WORD=great	…
backed	POS=VBD		…	PREV_WORD=britain	…
Fischler	POS=NNP	IS_CAPITALIZED	…	PREV_WORD=backed	…
's	POS=XX			PREV_WORD=fischler	

NER: Designing features

- Can you think of other features?

WORD
PREV_WORD
NEXT_WORD
PREV_BIGRAM
NEXT_BIGRAM
POS
PREV_POS
NEXT_POS
PREV_POS_BIGRAM
NEXT_POS_BIGRAM
IN_LEXICON_PER
IN_LEXICON_LOC
IN_LEXICON_ORG
IS_CAPITALIZED
HAS_DIGITS
IS_HYPHENATED
IS_ALLCAPS
FREQ_WORD
RARE_WORD
USEFUL_UNIGRAM_PER
USEFUL_BIGRAM_PER
USEFUL_UNIGRAM_LOC
USEFUL_BIGRAM_LOC
USEFUL_UNIGRAM_ORG
USEFUL_BIGRAM_ORG
USEFUL_SUFFIX_PER
USEFUL_SUFFIX_LOC
USEFUL_SUFFIX_ORG

Click to edit Master subtitle style

11/30/11

Click to edit Master subtitle style

11/30/11

NER: How else can we improve?

- Unlabeled data!

With more unlabeled data
instance 1: ... headquartered in (Washington State)^L ...
instance 2: ... (Mr. Washington)^P, the vice president of ...
instance 3: ... headquartered in (Kazakhstan) ...
instance 4: ... flew to (Kazakhstan) ...
instance 5: ... (Mr. Smith), a partner at Steptoe & Johnson ...
test: ... (Robert Jordan), a partner at ...
test: ... flew to (China) ...

example from Jerry Zhu

NER : Challenges

- Domain transfer
  WSJ → NYT
  WSJ → Blogs ??
  WSJ → Twitter ??!?
- Tough nut: Organizations
- Non textual data?

Entity Extraction is a Boring Solved Problem – or is it?
(Vilain, Su and Lubar, 2007)

NER: Related application

- Extracting real estate information from Craigslist Ads

Our oversized one, two and three bedroom apartment homes with floor plans featuring 1 and 2 baths offer space unlike any competition. Relax and enjoy the views from your own private balcony or patio, or feel free to entertain, with plenty of space in your large living room, dining area and eat-in kitchen. The lovely pool and sun deck make summer fun a splash. Our location makes commuting a breeze – Near MTA bus lines, the Metro station, major shopping areas, and for the little ones, an elementary school is right next door.

NER: Related Application

- BioNLP: Annotation of chemical entities

Type	Description	Example
CM	chemical compound	citric acid
RN	chemical reaction	1,3-dimethylation
CJ	chemical adjective	pyrazolic
ASE	enzyme	methylase
CPR	chemical prefix	1,3-

Corbet, Batchelor & Teufel, 2007

Shared Tasks: NLP in practice

- Shared Task
  - Everybody works on a (mostly) common dataset
  - Evaluation measures are defined
  - Participants get ranked on the evaluation measures
  - Advance the state of the art
  - Set benchmarks
- Tasks involve common hard problems or new interesting problems