Multilingual Annotation

Nianwen Xue
7/8/2011
LSA Summer Institute
Course objective

• Design and annotate a small corpus through case studies of well-established annotation projects in multiple languages, mainly English and Chinese
Course schedule

- Class 1: overview
- Class 2: Multilingual Treebanking
- Class 3: Multilingual Propbanking
- Class 4: Multilingual Discourse relations
- Class 5: Multilingual Temporal Annotation
- Class 6: Annotation of subjective language
- Class 7: Multilayer annotation
- Class 8: project presentation
Project schedule

- Week 1: Come up with a few project ideas, and write them up in a few slides.
- Week 2: Write the specifications/guidelines
- Week 3: pilot annotation/guidelines revision
- Week 4: evaluation, presentation, writeup
Today’s topics

• Why annotating?
• What to annotate?
• Life cycle of an annotation project
• Annotation evaluation
• Annotation infrastructure/tools
Annotating for NLP

- Machine learning is the dominant approach in NLP
- Annotated data is fuel to machine learning-based systems
- Annotation defines the system output, and formulates problems that need to be solved
- Almost every federally funded program has an annotation component: MUC, ACE, TIDES, GALE, BOLT
Annotating for linguistics

- Allows linguists to target certain word types or syntactic constructions to study
- Query tools have been developed to extract certain constructions
Context dependency of annotation

- Dictionaries, thesauri, and ontologies
  - WordNet, LDCE
  - Can serve as specifications for annotation
- Annotations
  - Penn (English, Chinese, Arabic) Treebank
  - Propbank
- Purpose of annotation:
  - Resolve ambiguity
What is there to annotate?

- Word tokens
- Part of speech
- Word sense
- Time expressions
- Sentiment/opinions
- Named entities
- Relations
- Coreference
- Many many more
At one point, Zuckerberg -- forever the geek -- pulled up a chart about Facebook's projected growth, which he introduced as a "log-normalized graph." The Internet's collective eyes rolled back.

"Logarithmic graphs? Jesus Christ."

So what went wrong?

Essentially, Zuckerberg seems to have tried -- and failed -- to imitate the Undisputed king of the tech presser: Apple's Steve Jobs, who is known for leaving even the most skeptical of tech bloggers in a "hypnotic daze" after he finishes a talk.
At one point, Zuckerberg -- forever the geek -- pulled up a chart about Facebook's projected growth, which he introduced as a "log-normalized graph."

The Internet's collective eyes rolled back.

"Logarithmic graphs? Jesus Christ."

So what went wrong?

Essentially, Zuckerberg seems to have tried -- and failed -- to imitate the Undisputed king of the tech presser: Apple's Steve Jobs, who is known for leaving even the most skeptical of tech bloggers in a "hypnotic daze" after he finishes a talk.
Tokenization (Chinese)

日文 章鱼 怎么说？
Japanese octopus how say
“How to say octopus in Japanese?”

日本 article fish how say
“???”
Part-of-speech (POS) tagging

At one point, Zuckerberg -- forever the geek -- pulled up a chart about Facebook 's projected growth, which he introduced as a "log-normalized graph."
The/DT Internet/NR 's/POS collective/JJ eyes/NNS rolled/VBD back/RP.

"/" Logarithmic/JJ graphs/NNS ?/. Jesus/NR Christ/NR ./. "/""

So what went wrong?

Essentially, Zuckerberg seems to have tried -- and failed -- to imitate the Undisputed king of the tech presser: Apple 's Steve Jobs, who is known for leaving even the most skeptical of tech bloggers in a "hypnotic daze" after he finishes a talk.

Penn Treebank Tagset: http://www.cis.upenn.edu/~treebank/
POS-tagging (Chinese)

他背叛了我。
He betray ASP I.
“He betrayed me.”
“He betrayed me.”

他的背叛伤害了我。
His DE betrayal hurt ASP me.
“his betrayal hurt me.”
“His betrayal hurt I”
At one point, Zuckerberg -- forever the geek -- pulled up a chart about Facebook 's projected growth, which he introduced as a "log-normalized graph." The/DT Internet/NR 's/POS collective/JJ eyes/NNS rolled/VBD back/RP. "/") Logarithmic/JJ graphs/NNS ?/. Jesus/NR Christ/NR ./. ")"

So what went wrong?

Essentially, Zuckerberg seems to have tried -- and failed -- to imitate the Undisputed king of the tech presser: Apple 's Steve Jobs, who is known for leaving even the most skeptical of tech bloggers in a "hypnotic daze"

Penn Treebank Tagset: http://www.cis.upenn.edu/~treebank/
At one point, Zuckerberg -- forever the geek -- pulled up a chart about Facebook 's projected growth, which he introduced as a "log-normalized graph."
The Internet's collective eyes rolled back.
"/" Logarithmic graphs?/. Jesus Christ. "/" 

So what went wrong?

Essentially, Zuckerberg seems to have tried -- and failed -- to imitate the Undisputed king of the tech presser: Apple 's Steve Jobs, who is known for leaving even the most skeptical of tech bloggers in a "hypnotic daze" after he finishes a talk.

Prague Dependency Treebank: http://ufal.mff.cuni.cz/pdt2.0/
I saw her duck.

how about...

- I saw her duck with a telescope.
- I saw her duck with a telescope in the garden...
At one point, Zuckerberg -- forever the geek -- pulled up a chart about Facebook's projected growth, which he introduced as a "log-normalized graph."
The Internet's collective eyes rolled back.
"Logarithmic graphs? Jesus Christ."

So what went wrong?

Essentially, Zuckerberg seems to have tried -- and failed -- to imitate the Undisputed king of the tech presser: Apple's Steve Jobs, who is known for leaving even the most skeptical of tech bloggers in a "hypnotic daze."

Propbank: http://verbs.colorado.edu/propbank/framesets-english/introduce-v.html
At one point, Zuckerberg -- forever the geek -- pulled up a chart about Facebook 's projected growth, which he introduced as a "log-normalized graph."

The Internet's collective eyes rolled back.

"Logarithmic graphs? Jesus Christ."

So what went wrong?

Essentially, Zuckerberg seems to have tried -- and failed -- to imitate the Undisputed king of the tech presser: Apple's Steve Jobs, who is known for leaving even the most skeptical of tech bloggers in a "hypnotic daze" after he finishes a talk.

Propbank: http://verbs.colorado.edu/propbank/framesets-english/introduce-v.html
Semantic roles

Agent

↓

Cheney shot the wild goose

Target

↓

The wild goose was shot by Cheney

Cheney’s shooting of the wild goose
At one point, Zuckerberg -- forever the geek -- pulled up a chart about Facebook's projected growth, which he introduced as a "log-normalized graph."
The Internet's collective eyes rolled back.

So what went wrong?

Essentially, Zuckerberg seems to have tried -- and failed -- to imitate the Undisputed king of the tech presser: Apple's Steve Jobs, who is known for leaving even the most skeptical of tech bloggers in a "hypnotic daze" after he finishes a talk.

Word sense grouping: http://verbs.colorado.edu/html_groupings/introduce-v.html
Word sense (Chinese)

这张纸很白。
This CL paper very white.
“This piece of paper is very white.”
“This paper is white.”

白看一场电影
free watch one CL movie
“watch a movie for free”
“watch a free movie”
“White watch a movie.”
Discourse relations (PDTB)

[Arg1 At one point, Zuckerberg -- forever the geek -- pulled up a chart about Facebook 's projected growth, which he introduced as a "log-normalized graph."]
[conn Then] [Arg2 The Internet 's collective eyes rolled back.]

"Logarithmic graphs? Jesus Christ."

TEMPORAL-Precedence
Or
CONTINGENCY-Result?

Essentially, Zuckerberg seems to have tried -- and failed -- to imitate the Undisputed king of the tech presser: Apple 's Steve Jobs, who is known for leaving even the most skeptical of tech bloggers in a "hypnotic daze" after he finishes a talk.

http://www.seas.upenn.edu/~pdtb
At one point, Zuckerberg -- forever the geek -- pulled up a chart about Facebook 's projected growth, which he introduced as a "log-normalized graph." The Internet 's collective eyes rolled back.

"Logarithmic graphs? Jesus Christ."

So what went wrong?

Essentially, Zuckerberg seems to have tried -- and failed -- to imitate the Undisputed king of the tech presser: Apple 's Steve Jobs, who is known for leaving even the most skeptical of tech bloggers in a "hypnotic daze" after he finishes a talk.
At one point, Zuckerberg -- forever the geek -- pulled up a chart about Facebook's projected growth, which he introduced as a "log-normalized graph." The Internet's collective eyes rolled back.

"Logarithmic graphs? Jesus Christ."

So what went wrong?

Essentially, Zuckerberg seems to have tried -- and failed -- to imitate the undisputed king of the tech presser: Apple's Steve Jobs, who is known for leaving even the most skeptical of tech bloggers in a "hypnotic daze" after he finishes a talk.
At one point, Zuckerberg -- forever the geek -- pulled up a chart about Facebook's projected growth, which he introduced as a "log-normalized graph." The Internet's collective eyes rolled back.

"Logarithmic graphs? Jesus Christ."

So what went wrong?

Essentially, Zuckerberg seems to have tried -- and failed -- to imitate the Undisputed king of the tech presser: Apple's Steve Jobs, who is known for leaving even the most skeptical of tech bloggers in a "hypnotic daze" after he finishes a talk.

ACE guidelines: http://projects.ldc.upenn.edu/ace/annotation/
Named entity recognition

• Named entity is an object of interest such as person, organization, and location
• Named entity recognition involves recognizing word sequences as named entities, and classifying them

Dr. Roberto Feliz and Dr. Hiba Georges were quickly jolted from the most modern of medical care in Boston, Massachusetts, to the most rudimentary of care when they flew to Haiti last week to work at a hospital housed in two tents run by the University of Miami.

PERSON: Dr. Roberto Feliz, Dr. Hiba Georges
LOCATION: Boston Massachusetts, Haiti
ORGANIZATION: the University of Miami
What makes named entity recognition difficult?

• Ambiguity
  – The same name string can refer to different entities of the same type
    • George Bush: George H. Bush, George W. Bush
  – The same name string can refer to different entities of different types
    • Ford: the person or the company
  – Web name disambiguation campaign:
    • http://nlp.uned.es/weps

• Variations of the same entity type
• Attempt to create/find an exhaustive list of names is hopeless
John Kennedy

John Kennedy, Composer
Home | About | Music | Schedule | Writings | Press | Contact © 2009, John Kennedy, all rights reserved. Photos by Sara Stathas unless noted otherwise.
www.johnkennedymusic.com/ - Cached - Similar –

Home - Louisiana Department of the Treasury - John Neely Kennedy ...
Department of the Treasury. Search for over $200 million in unclaimed property.
www.treasury.state.la.us/ - Cached - Similar –

John Kennedy Magic - World Class Magic Tricks for the Serious ...
At John Kennedy Magic we take pride in designing and manufacturing world class magic tricks for the serious hobbyist and professional.
www.johnkennedymagic.com/ - Cached - Similar –

John F. Kennedy : Biography
John Fitzgerald Kennedy, the son of Joseph Patrick Kennedy and Rose Fitzgerald, was born in Brookline, Massachusetts, on 29th May, 1917. ...
www.spartacus.schoolnet.co.uk/USAkennedyJ.htm - Cached - Similar -

John Kennedy Dealerships | New Ford, Mazda, Lincoln, Mercury, and ...
Conshohocken, Feasterville, Pottstown, Phoenixville & Plymouth Meeting, PA New, John Kennedy Dealerships sells and services Ford, Mazda, Lincoln, Mercury, ...
www.kennedyauto.com/ - Cached - Similar -
Names in the blogsphere

- TonkaManOR said...
- Michelle Obama Says:
- HiTek LoLife
- DarwinSurvivor
- aRosebyNEothername
- a random John
- missing_piece
- Spirit fingers
- Pony_express
- DahlIELama
- Ice cream for breakfast
At one point, Zuckerberg -- forever the geek -- pulled up a chart about Facebook 's projected growth, which he introduced as a "log-normalized graph." The Internet 's collective eyes rolled back.

"Logarithmic graphs? Jesus Christ."

So what went wrong?

Essentially, Zuckerberg seems to have tried -- and failed -- to imitate the Undisputed king of the tech presser: Apple 's Steve Jobs, who is known for leaving even the most skeptical of tech bloggers in a "hypnotic daze" after he finishes a talk.

ACE guidelines: http://projects.ldc.upenn.edu/ace/annotation/
At one point, Zuckerberg -- forever the geek -- pulled up a chart about Facebook 's projected growth, which he introduced as a "log-normalized graph."
The Internet 's collective eyes rolled back.

"Logarithmic graphs? Jesus Christ."

So what went wrong?

Essentially, Zuckerberg seems to have tried -- and failed -- to imitate the Undisputed king of the tech presser: Apple 's Steve Jobs, who is known for leaving even the most skeptical of tech bloggers in a "hypnotic daze" after he finishes a talk.

ACE guidelines: http://projects.ldc.upenn.edu/ace/annotation/
ACE Relation Types (2008)

- Physical
  - Located, Near

- Part-Whole
  - Geographical, Subsidiary, Artifact

- Personal-Social
  - Business, Family, Lasting-Personal

- ORG-Affiliation
  - Employment, Ownership, Founder, Student-Alum, Sports-Affiliation, Investor-Shareholder, Membership

- Gen-Affiliation
  - Citizen-Resident-Religion-Ethnicity, Org-Location-Origin

- Agent-Artifact
  - User-Owner-Inventor-Manufacturer
Event and Temporal Analysis

- Extraction of events
- Extraction and normalization of time expressions
- Associating events with time expressions
- Determining the temporal location of events
- Temporal ordering of events
An example

• Apple Inc. will host a much-anticipated press event January 27 in San Francisco, California. Invitations went out Monday.
  – Events:
    • press conference, inviting
  – Time expressions:
    • January 27 [2010-01-27], Monday [2010-01-18]
  – Associating between events and time expression:
    • Press conference ~ January 27, Inviting ~ Monday
  – Event ordering:
    • Press conference is AFTER invitation
Annotation of events, times and their relations

- **Event extraction**
  - Detection: which text spans are anchors of events?
  - Classification: what types of events are there?

- **Times**
  - Detection: identifying the text spans of time expressions
  - Normalization of these time expressions

- **Relations between them**
  - Which time expressions are linked to which events?
  - How events are temporally ordered?
Event detection

- Find and classify all the events in a text.
  - Most verbs introduce events/states
    - But not all (*take a bath*)
  - Nominalizations often introduce events
    - *Collision, destruction, the running*...
CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.
CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York
IE in the Biomedical domain

- **Amiodarone weakly inhibited CYP2C9, CYP2D6, and CYP3A4-mediated activities with $Ki$ values of 45.1--271.6 $\mu$M**

- **Named entities:**
  - Amiodarone, CYP2C9, CYP2D6, CYP3A4

- **Relations:**
  - amiodarone inhibits CYP2C9 with $Ki=45.1--271.6$
  - amiodarone inhibits CYP2D6 with $Ki=45.1--271.6$
  - amiodarone inhibits CYP3A4 with $Ki=45.1--271.6$
Diseases and treatments

They felt that there was obstruction and possible obstruction to airflow generating some obstructive apnea, as well as partial obstructive events.

In 1980 she had quadruple coronary artery bypass graft surgery by Dr. Elks at Feargunwake Otacaa Community Hospital and did well until 1988 when she had exertional angina and a positive stress test and found that three or four grafts were occluded.
Sorting annotation by text unit

- Tokens/words: tokenization, POS tagging,
- Phrases: named entities, sentiment, semantic roles, multi-word expressions
- Clauses: syntax, semantic roles, discourse relations
- Sentences: syntax, discourse relations
- Paragraphs: high-level discourse relations
- Documents: topics, sentiment
- Invisible elements: empty categories
Can annotation be theory-neutral?

- atheoretical
- theoretical

- Named entities
- Coreference
- Relations
- Word sense
- sentiment

- POS-tagging
- syntax
- Semantic roles
- Discourse relations

Many schools of thought
Level of difficulty

Easy

- Tokenization
- POS tagging
- Named entities
- ...

Syntax
- Semantic roles
- Discourse relations
- Relations
- Coreference
- ...

Quantifier scope
- Negation
- Focus
- ...

hard
What has been annotated?

• Syntactic structures
• Semantics
• Discourse
• Subjective language
Sample from LREC 2010

- Word alignment, Persian pos-tagging, Italian discourse, disordered speech, concession, causation in French, near identities for coreference, named entity for Dutch, super-sense tagging in Italian, word sense in Wikipedia,
Life cycle of an annotation project

• Have an idea
• Write specifications/guidelines
  – Specification is the conceptual framework
  – Guidelines include specifications and instructions for annotators
• Develop/select an annotation tool
• Annotate
  – Multiple annotators
• Evaluate and validate
Annotation lifecycle

- Guideline design
- Pilot annotation
- Annotation tool/infrastructure
- Productive annotation adjudication
- Evaluation (ITA)
- Validation
Evaluation metrics

• Straight accuracy

• Precision/recall/F-score

• Kappa:
  \[ k = \frac{A - E}{1 - E} \]
The “Art” of annotation

• Find the “sweet spot” between
  – The depth of representation
  – Inter-Annotator Agreement
  – Productivity
Annotation tool: MAE and MAI

• A light-weight annotation tool
• Java-based, works on Mac, Windows and Linux systems
• Tested on English and Chinese
• Developed by Amber Stubbs at Brandeis
• Available for Download here: http://pages.cs.brandeis.edu/~astubbs/index.html

• Can be found here: http://pages.cs.brandeis.edu/~xuen/