Toward a ‘Science’ of Annotation

Eduard Hovy

Information Sciences Institute
University of Southern California
hovy@isi.edu
NLP Group at ISI

• **Goals:**
  – **Practical**: Build technology to help people communicate more naturally with computers, and better with one another
  – **Theoretical**: Understand foundations of language and semantics

• **Research methodology:**
  – Integrate **human theorizing** (model building) with **statistical processing** (machine learning) — find optimal balance
  – Evaluate frequently

• **Research group:**
  – One of largest university-based NLP groups in North America (in operation since late 1970s) — approx. 35 people
  – Funding from DARPA, NSF, ARDA, DoD, etc. (= $5M/year)
  – Active interaction with Computational Linguistics worldwide (Best Paper awards; leading roles in prof. societies; service on roadmap committees; etc.)
  – Significant focus on education (2–5 PhD students/year; 2 semester courses per year in CS; MS in CL program joint with Linguistics)
Information Extraction

- **Information Extraction task**: Identify fragments in texts that express important/useful info; extract; store in database (= IE is a kind of annotation)

- **Why do this?**
  - Create a single coherent database of just the info you care about
  - Cover a large number of sources — many more than you can possibly get by hand
  - Useful for study, reference, and teaching
  - Useful for data mining: finding trends/patterns over time

- **Who uses this today?**
  - Military / Intelligence community
  - Business community
  - Government
  - Biomedical, PoliSci, and other research communities
Example: IE of complex notions

• Context: build psychological models of people in focus areas — help Army avoid mistakes in action

• Automatically identify and extract people’s *Attitudes*, *Stimuli*, and *Actions*

• Problem: what are these concepts?
  – Domain experts disagree
  – For Attitude, Opinions are somehow important, but are not everything — also relevant are Goals and Beliefs

• Approach:
  – First extract all simpler pieces (entities, events, goals, beliefs, opinions, etc.)
  – Analyze internal structure of each piece
  – Then try to combine somehow, to identify useful sentences
Example: Data flow & modules

Preprocessing
• remove junk (html, etc.)
• insert parts of speech

Entity extraction
• find people, locations, dates, etc.
• COTS: IdentiFinder or Thingfinder

Event extraction
• find actions, events, etc.
• ISI extractor

Opinion, goal, belief extraction
• find additional kinds of info
• ISI system

Attitude construction
• combine extracted info to identify attitudes
• ISI system

Action construction
• combine extracted info to identify actions
• ISI system

Stimulus construction
• combine extracted info to identify stimuli
• ISI system

Parsing and analysis
• parse sentence structure and insert frames
• Charniak or MINIPAR parser and ISI frames

Event construction
• group events and associated info
• ISI system

Formatting
• standardize rep
• normalize strength scores

Tested on tens of thousands of articles, running nightly. Deployed at MITRE for operational use by Army
Example: Cascading complexity

- **Type**: Attitudes
- **Technology**: Combination of basic factors
  - Goals+Opinions+other things → Attitudes
  - Attitude includes 12 classes:
    - MOTIVATION
    - SUPERLATIVE
    - BELIEF
    - GOAL
    - OPINION
    - RELIGION
    - EXTREMUM
    - REPORT
    - GPE
    - DATE
    - TIME
    - LOCATION
  - Combination function: \( X = \sum_i \alpha_i f_i \) — work in progress to determine optimal coefficients \( \alpha_i \)
  - Annotation tests show much higher agreement among domain specialists than for basic factors alone

<table>
<thead>
<tr>
<th>Agree / disagree</th>
<th>Yes(%)</th>
<th>No(%)</th>
<th>System</th>
<th># sents at level of agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>0%</td>
<td>high</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>87.50%</td>
<td>12.50%</td>
<td>high</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>100%</td>
<td>none</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>20</td>
<td>high</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>25</td>
<td>high</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>80</td>
<td>medium</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>75</td>
<td>medium</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>60</td>
<td>medium</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>40</td>
<td>high</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>40</td>
<td>medium</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

- Each factor has an indicator engine
- Each engine returns a fragment of text (plus usually a score)
- To find the ‘attitude strength’ of a sentence, we combine the various scores using their *relative strengths*
- This we do by correlating human (SME) judgments
Annotation for IE

As the items to extract become more complex, IE definition phase becomes harder: move from pre-specified (hard-coded) rules to automated learning… …and this requires annotation…

• What is the role of annotation?
• How to define the IE, and how to determine IE acceptability?
• When IE is not acceptable, what can one do?
  – Improve IE algorithms
  – Fix or extend training data
  – Redefine extraction model
  – Refine domain theory
… etc. …
Generic IE methodology

- **Design phase** (domain experts):
  - Decide on the information types desired — based on domain theory and model
  - Obtain and prepare document corpus (domain + CS experts)
  - Annotate test documents to determine task feasibility for humans — often this requires further theoretical (model) adjustments or growth

- **Learning phase**: building the system (domain + CS experts):
  - Domain expert: annotate documents
  - CS person: create IE model (identify likely indicator cues; build cue recognizer functions, using words/phrase patterns/ling. info… as features)
  - CS: Deploy IE learning algorithms (CRF…)
  - Both: evaluate performance on unseen data and assign reliability scores

- **Is the result adequate?** If not, fix something; else go on

- **Application phase**: running the system (domain experts):
  - Obtain more documents
  - Run system (on input, apply all cues for the desired type; collect and merge results using merging functions; save output in database)
  - Reconcile inconsistencies and enjoy the results (or extend your theory!)
Two reasons to annotate

• **Traditional goal**: Fundamental belief that domain semantics is useful:
  – for reasoning in / studying the domain,
  – to help improve NLP.

• **Methodologies**: Transform pure text into interpreted/extracted/marked-up text
  – Old methodology: manually-built rules for transformations
  – New methodology: machine learning of transformations
    1. Have humans manually annotate texts with transformation info
    2. Train computers on the corpus to do the same job

• **Additional goal**: Use annotation as mechanism to test aspects of the theory of domain semantics empirically — actual theory formation as well
In NLP: Are we entering a new era of corpus building?

• The ‘statistics revolution’ in speech and NL processing is now complete:
  – Most people see speech and NL processing as a notation rewrite problem:
    • Speech → text, Italian → Chinese, sentence → parse tree → case frame, long text → short text…
    – Everyone uses machine learning to learn the rewriting ‘rules’
    – Everyone agrees creating rewriting rules by hand is infeasible for most transformations — the phenomena are too complex

• Results:
  – A new hunger for annotated corpora
  – A new class of researcher: the Annotation Expert

• BUT: How rigorous is Annotation as a ‘science’?
NLP at increasing depths

Direct: simple replacement

Small changes: demorphing, etc.

Adding info: POS tags, etc.

Medium changes: syntax

Adding more: semantic features

Shallow semantics: frames

Deep semantics: ?

Do interesting processing: filter, match parts, etc.

Small changes: demorphing, etc.

Direct: simple replacement
Shallow and deep semantics

• She sold him the book / He bought the book from her
  \( (X1 :\text{act Sell} :\text{agent She} :\text{patient (X1a :\text{type Book}) :\text{recip He})} \)
  \( (X2a :\text{act Transfer} :\text{agent She} :\text{patient (X2c :\text{type Book}) :\text{recip He}}) \)
  \( (X2b :\text{act Transfer} :\text{agent He} :\text{patient (X2d :\text{type Money}) :\text{recip She})} \)

• He has a headache / He gets a headache
  \( (X3a :\text{prop Headache} :\text{patient He}) \)

• Though it’s not perfect, democracy is the best system
  \( (X4 :\text{type Contrast} :\text{arg1 (X4a ...?...)} :\text{arg2 (X4b ...?...))} \)

- Which symbols?
- Which roles?
- How define states and state changes?
- How handle relations?
- How handle negation?
- How handle comparatives?
<table>
<thead>
<tr>
<th>Some phenomena to annotate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Somewhat easier</strong></td>
</tr>
<tr>
<td>Bracketing (scope) of predications</td>
</tr>
<tr>
<td>Word sense selection (incl. copula)</td>
</tr>
<tr>
<td>NP structure: genitves, modifiers…</td>
</tr>
<tr>
<td>Concepts: ontology definition</td>
</tr>
<tr>
<td>Concept structure (incl. frames and thematic roles)</td>
</tr>
<tr>
<td>Coreference (entities and events)</td>
</tr>
<tr>
<td>Pronoun classification (ref, bound, event, generic, other)</td>
</tr>
<tr>
<td>Identification of events</td>
</tr>
<tr>
<td>Temporal relations (incl. discourse and aspect)</td>
</tr>
<tr>
<td>Manner relations</td>
</tr>
<tr>
<td>Spatial relations</td>
</tr>
<tr>
<td>Direct quotation and reported speech</td>
</tr>
<tr>
<td><strong>More difficult</strong></td>
</tr>
<tr>
<td>Quantifier phrases and numerical expressions</td>
</tr>
<tr>
<td>Comparatives</td>
</tr>
<tr>
<td>Coordination</td>
</tr>
<tr>
<td>Information structure (theme/rheme)</td>
</tr>
<tr>
<td>Focus</td>
</tr>
<tr>
<td>Discourse structure</td>
</tr>
<tr>
<td>Other adverbials (epistemic modals, evidentials)</td>
</tr>
<tr>
<td>Identification of propositions (modality)</td>
</tr>
<tr>
<td>Opinions and subjectivity</td>
</tr>
<tr>
<td>Pragmatics/speech acts</td>
</tr>
<tr>
<td>Polarity/negation</td>
</tr>
<tr>
<td>Presuppositions</td>
</tr>
<tr>
<td>Metaphors</td>
</tr>
</tbody>
</table>
Annotation project desiderata

- Annotation must be:
  - Fast… to produce enough material
  - Consistent… enough to support learning
  - Deep… enough to be interesting

- Thus, need:
  - Simple procedure and good interface
  - Several people for cross-checking
  - Careful attention to the source theory!

- Example: Can this be done for semantics???
Annotation project desiderata

• Annotation must be:
  – **Fast**… to produce enough material
  – **Consistent**… enough to support learning
  – **Deep**… enough to be interesting

• Thus, need:
  – Simple **procedure** and good **interface**
  – Several people for **cross-checking**
  – Careful attention to the source **theory**!
Annotation as a science

• Increased need for corpora and for annotation raises new questions:
  – What kinds/aspects of ‘domain semantics’ to annotate?
    …it’s hardly an uncontroversial notion…
  – Which corpora? How much?
  – Which computational tools to apply once annotation is ‘complete’? When is it complete?
  – How to manage the whole process?

• Results:
  – A new hunger for annotated corpora
  – A new class of researcher: the Annotation Expert

• Need to systematize annotation process — BUT: How rigorous is Annotation as a ‘science’?
Talk overview

1. Introduction: A new role for annotation?
2. Example: Semantic annotation in OntoNotes
3. Toward a science of annotation: 7 questions
4. Conclusion
Semantic annotation projects

- Goal: corpus of pairs (sentence + semantic rep)
- Process: humans add information to sentences (and their parses)
- Recent projects:
  - OntoNotes (Weischedel et al. 05–)
  - PropBank (Palmer et al. 03–)
  - Framenet (Fillmore et al. 04)
  - Penn Treebank (Marcus et al. 99)
  - Interlingua Annotation (Dorr et al. 04)
  - I-CAB, Greek… banks
  - TIGER/SALSA Bank (Pinkal et al. 04–)
  - Prague Dependency Treebank (Hajic et al. 02–)
  - NomBank (Myers et al. 03–)
Other recent annotation projects

• US:
  – Time-ML (Pustejovsky et al.)
  – MPQA: subjectivity / ‘opinion’ (Wiebe et al.)

• EU:
  – Several annotation projects

• Japan:
  – Two ministries (MIC & METI) planning next 8 years’ NLP research — annotation important role
  – MIC theme: Universal communication (knowledge construction and multimedia integration, input and output)
The founder of Pakistan’s nuclear department, Abdul Qadeer Khan, has admitted he transferred nuclear technology to Iran, Libya, and North Korea. (slide credit to M. Marcus and R. Weischedel, 2004)
Example of result

Mrs. Hills said many of the 25 countries that she placed under varying degrees of scrutiny have made "genuine progress" on this touchy issue.

In various formats...

**Propositions**
- *predicate*: say
- *pb sense*: 01
- *on sense*: 1

**ARG0**: Mrs. Hills [10]
**ARG1**: many of the 25 countries that she placed under varying degrees of scrutiny have made "genuine progress" on this touchy issue

**Predicate**: make
- *pb sense*: 03
- *on sense*: None

**ARG0**: many of the 25 countries that she placed under varying degrees of scrutiny
**ARG1**: "genuine progress" on this touchy issue
Project structure & parts

- Syntactic structure
- Predicate/argument structure
- Disambiguated nouns and verbs
- Coreference links
- Ontology
- Decoders

(Slide by M. Marcus, R. Weischedel, et al.)
OntoNotes 1

- OntoNotes colleagues:
  - Sentence structure: U of Pennsylvania: Mitch Marcus et al.
  - Verb meanings: U of Colorado: Martha Palmer, Ann Houston, et al., plus about 20 annotators
  - Noun meanings: ISI: Robin Belvin, Ann Houston, Bonnie Glover Stalls, Rahul Bhagat, Mani Alagappan, Andrew Philpot, Jingbo Zhu, plus about 35 annotators
  - Coreference links: BBN: Ralph Weischedel, Lance Ramshaw, Sameer Pradhan, et al., plus 5 annotators

- Goal: In 4 years, annotate corpora of 1 mill words of English, Chinese, and Arabic text:
  - Manually provide semantic symbols for nouns, verbs, adjs, advs
  - Manually connect sentence structure in verb and noun frames
  - Manually link anaphoric references
  - Manually construct supporting ontology of senses
OntoNotes 2

• History:
  – PropBank (2002–): verb annotation procedure developed
  – OntoNotes Feasibility Study (2004): Test corpus built, with coref annotation
  – Project started October 2005 (English); Chinese added 2006; Arabic in 2007
  – Possible to continue until 2009, funding permitting
• Potential for the near future: semantics ‘bank’
  – May energize lots of research on semantic analysis, reps, etc.
  – May enable semantics-based IR, QA, MT, etc.
OntoNotes antecedents

Treebank

WordNet

FrameNet

PropBank + frames

NomBank + frames

PropBank2

VerbNet

Chinese

Chinese

Chinese

Chinese

Arabic

Sense tags, Coreference and Ontology links

OntoNotes

(Palmer et al.)
Even so: Many words untouched!

Results of automated annotation using system trained on OntoNotes corpus:

The Bush administration (WN-Poly ON-Poly) had heralded (WN-Poly False) the Gaza pullout (WN-Poly False) as a big step (WN-Poly ON-Mono) on the road (WN-Poly ON-Mono) map (WN-Poly False) to a separate Palestinian state (WN-Poly ON-Poly) that Bush hopes (WN-Poly ON-Mono) to see (WN-Poly ON-Poly) by the time (WN-Poly False) he leaves (WN-Poly False) office (WN-Poly False) but a Netanyahu victory (WN-Mono False) would steer (WN-Poly False) Israel away from such moves (WN-Poly ON-Poly).

The Israeli generals (WN-Poly ON-Mono) said (WN-Poly ON-Poly) that if the situation (WN-Poly ON-Mono) did not improve (WN-Poly ON-Mono) by Sunday Israel would impose (WN-Poly ON-Mono) "more restrictive and thorough security (WN-Poly False) measures (WN-Poly False)" at other Gaza crossing (WN-Poly ON-Mono) points (WN-Poly ON-Poly) that Israel controls (WN-Poly ON-Poly), according (WN-Poly False) to notes (WN-Poly False) of the meeting (WN-Poly False) obtained (WN-Poly ON-Mono) by the New York Times.
Three major subtasks

• How do you go from

    The founder of Pakistan’s nuclear department, Abdul Qadeer Khan, has admitted he transferred nuclear technology to Iran, Libya, and North Korea

to

    P1: :type Person3 :name “Abdul Qadeer Khan”
P2: :type Person3 :gender male
P3: :type Know-How4
P4: :type Nation2 :name “Iran”
P5: :type Nation2 :name “Libya”
P6: :type Nation2 :name “N. Korea”
X0: :act Admit1 :speaker P1 :saying X2
coref P1 P2

• Tasks:
  1. Create word senses for words (and insert into Omega ontology, as concepts)
  2. Annotate sentences with the senses
  3. Annotate sentences for co-reference

instances
semantic symbols
frame structure
coref links
OntoNotes annotation procedure

• **Sense creation** process goes by word:
  – Expert creates meaning options (shallow semantic senses) for verbs, nouns, [adjs, advs] … follows PropBank process (Palmer et al.)
  – Expert creates definitions, examples, differentiating features
  – (Ontology insertion: At same time, expert groups equivalent senses from different words and organizes/refines Omega ontology content and structure … process being developed at ISI)

• **Sense annotation** process goes by word, across docs:
  – Process developed in PropBank
  – Annotators manually…
    • See each sentence in corpus containing the current word (noun, verb, [adjective, adverb]) to annotate
    • Select appropriate senses (= ontology concepts) for each one
    • Connect frame structure (for each verb and relational noun)

• **Coref annotation** process goes by doc:
  – Annotators connect co-references within each doc
Ensuring trustworthiness/stability

• Problematic issues:
  1. What sense are there? Are the senses stable/good/clear?
  2. Is the sense annotation trustworthy?
  3. What things should corefer?
  4. Is the coref annotation trustworthy?

• Approach (from PropBank): “the 90% solution”:
  – Sense granularity and stability: Test with annotators to ensure agreement at 90%+ on real text
  – If not, then redefine and re-do until 90% agreement reached
  – Coref stability: only annotate the types of aspects/phenomena for which 90%+ agreement can be achieved
Sense annotation procedure

• Sense creator first creates senses for a word

• Loop 1:
  – Manager selects next nouns from sensed list and assigns annotators
  – Programmer randomly selects 50 sentences and creates initial Task File
  – Annotators (at least 2) do the first 50
  – Manager checks their performance:
    • 90%+ agreement + few or no NoneOfAbove — send on to Loop 2
    • Else — Adjudicator and Manager identify reasons, send back to Sense creator to fix senses and defs

• Loop 2:
  – Annotators (at least 2) annotate all the remaining sentences
  – Manager checks their performance:
    • 90%+ agreement + few or no NoneOfAbove — send to Adjudicator to fix the rest
    • Else — Adjudicator annotates differences
    • If Adj agrees with one Annotator 90%+, then ignore other Annotator’s work (assume a bad day for the other); else Adj agrees with both about equally often, then assume bad senses and send the problematic ones back to Sense creator
Pre-project test: Can it be done?

- Annotation process and tools developed and tested in PropBank (Palmer et al.; U Colorado)
- Typical results (10 words of each type, 100 sentences each):

<table>
<thead>
<tr>
<th></th>
<th>Round1</th>
<th>Round2</th>
<th>Round 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>verbs</td>
<td>.76 → .86 → .91</td>
<td>4.5 → 5.2 → 3.8</td>
<td>30 → 25 → 25</td>
</tr>
<tr>
<td>nouns</td>
<td>.71 → .85 → .95</td>
<td>7.3 → 5.1 → 3.3</td>
<td>28 → 20 → 15</td>
</tr>
<tr>
<td>adjs</td>
<td>.87 → – → .90</td>
<td>2.8 → – → 5.5</td>
<td>24 → – → 18</td>
</tr>
</tbody>
</table>

(by comparison: agreement using WordNet senses is 70%)
Before we start: Word statistics

Number of word tokens/types in 1000-word corpus
(95% confidence intervals on 85213 trials)

<table>
<thead>
<tr>
<th>1000-word corpus</th>
<th>tokens</th>
<th>types</th>
</tr>
</thead>
<tbody>
<tr>
<td>verbs</td>
<td>125.3</td>
<td>87.3</td>
</tr>
<tr>
<td>nouns</td>
<td>446.6</td>
<td>288.7</td>
</tr>
<tr>
<td>adjectives</td>
<td>103.2</td>
<td>80.6</td>
</tr>
</tbody>
</table>

Nouns: 57.2% of tokens
Monosemous nouns (but not names etc.):
14.6% of tokens
= 25.6% of nouns

Polysemy

<table>
<thead>
<tr>
<th>250K WSJ</th>
<th>verbs</th>
<th>nouns</th>
</tr>
</thead>
<tbody>
<tr>
<td>total</td>
<td>2341</td>
<td>5421</td>
</tr>
<tr>
<td>1 WN sense</td>
<td>428</td>
<td>(18%)</td>
</tr>
<tr>
<td>2 or 3 senses</td>
<td>966</td>
<td>(41%)</td>
</tr>
<tr>
<td>4+ senses</td>
<td>947</td>
<td>(40%)</td>
</tr>
</tbody>
</table>
Before we start:
Noun coverage, various corpora

Coverage in WSJ, Brown corpora of most frequent $N$ nouns

<table>
<thead>
<tr>
<th>Nouns</th>
<th>Tokens (total 205442)</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>76420</td>
<td>37%</td>
</tr>
<tr>
<td>500</td>
<td>140453</td>
<td>68%</td>
</tr>
<tr>
<td>1000</td>
<td>167715</td>
<td>82%</td>
</tr>
<tr>
<td>1500</td>
<td>181412</td>
<td>88%</td>
</tr>
<tr>
<td>2000</td>
<td>189641</td>
<td>92%</td>
</tr>
</tbody>
</table>

Coverage of corpus of most frequent $N$ polysemous-2 nouns (WSJ+Brown)
Compound noun groups

- Problem: N-N compounds ("kitchen knife", "party animal", etc.)
  - Do not want to annotate each noun independently (a party animal is neither a party nor an animal)
- Solution: automatically find multiple-noun tuples (pairs, triples, etc.) with high co-occurrence
  - Pantel at ISI used pointwise mutual information algorithm to identify high-reliability tuples (up to 4-grams)
  - Found 35,700 tuples
- Linking into Omega:
  - Automatically generated Omega superconcepts
  - Quasi-random check of 40 pairs showed about 72.5% accuracy
  - 1951 of the tuples cannot be attached into Omega because the head noun does not exist (e.g. proper nouns)
  - File at http://www.isi.edu/~pantel/wninte.txt

<table>
<thead>
<tr>
<th>wsj/00/wsj_0003.mrg</th>
<th>asbestos fiber</th>
</tr>
</thead>
<tbody>
<tr>
<td>wsj/00/wsj_0003.mrg</td>
<td>protection agency</td>
</tr>
<tr>
<td>wsj/00/wsj_0007.mrg</td>
<td>engineering industry</td>
</tr>
<tr>
<td>wsj/00/wsj_0008.mrg</td>
<td>government debt</td>
</tr>
<tr>
<td>wsj/00/wsj_0008.mrg</td>
<td>borrowing authority</td>
</tr>
<tr>
<td>wsj/00/wsj_0009.mrg</td>
<td>marketing arm</td>
</tr>
<tr>
<td>wsj/00/wsj_0009.mrg</td>
<td>auto maker</td>
</tr>
<tr>
<td>wsj/00/wsj_0009.mrg</td>
<td>marketing manager</td>
</tr>
<tr>
<td>wsj/00/wsj_0009.mrg</td>
<td>marketing executive</td>
</tr>
<tr>
<td>wsj/00/wsj_0010.mrg</td>
<td>boca raton</td>
</tr>
<tr>
<td>wsj/00/wsj_0099.mrg</td>
<td>air force contract</td>
</tr>
<tr>
<td>wsj/00/wsj_0099.mrg</td>
<td>intelligence data</td>
</tr>
</tbody>
</table>
# Nouns to be handled

## Monosemous
- 1253 trading
- 1117 investor
- 867 firm
- 585 tax
- 581 trader
- 567 chairman
- 566 income
- 462 asset
- 420 spokesman
- 338 customer
- 336 transaction
- 335 employee
- 324 shareholder
- 309 consumer
- 292 ad
- ... ... 1 academe
- 1 abstention
- 1 absorber

## Polyseamous
- 4178 year
- 3095 market
- 1796 sale
- 1467 month
- 1308 business
- 1253 trading
- 1211 rate
- 1141 time
- 1140 president
- 1117 investor
- 1053 day
- 1025 government
- 1012 quarter
- 974 bank
- 944 group
- ... ... 1 industrialization
- 1 globalization
- 1 diving

## Polyseamous unsensed
- 4076 company
- 3196 share
- 2393 stock
- 1874 price
- 1149 bond
- 1122 week
- 998 analyst
- 996 cent
- 919 interest
- 867 firm
- 774 product
- 759 earnings
- 749 industry
- 696 executive
- 667 money
- ... ... 1 aberration
- 1 abandonment
- 1 abacus
Annotation framework

• Data management:
  – Defined a data flow pathway that minimizes amount of human involvement, and produces status summary files (avg speed, avg agreement with others, # words done, total time, etc.)
  – Several interacting modules:
    • STAMP (built at UPenn, Palmer et al.): annotation
    • Server (ISI): store everything, with backup, version
    • Sense Creation interface (ISI): define senses
    • Sense Pooling interface (ISI): group together senses into ontology
    • Master Project Handler (ISI): annotators reserve word to annotate
    • Annotation Status interface (ISI): up-to-the-minute status
    • Statistics bookkeeper (ISI): individual annotator work

  (Re-)partition senses; (re-)create definitions and tests (1 person)
  Test: Annotate 50 sentences (2 people)
  >90% agreement? yes no
  Annotate all sentences with this word (2 people)
  >90% agreement? yes no
  Analyze disagreement
  Adjudicate the disagreements (adjudicator)
  Sense problem
  Annotator problem
  All sentences with this word annotated

USC INFORMATION SCIENCES INSTITUTE
The rest went to investors from France and Hong Kong. Earlier this year, Japanese investors snapped up a similar, $570 million mortgage-backed securities mutual fund. That fund was put together by Blackstone Group, a New York investment firm. The latest two funds were assembled jointly by Goldman Sachs & Co. of the U.S. and Japan’s Daiwa Securities Co.

The new, seven-year funds -- one offering a fixed-rate return and the other with a floating-rate return linked to the London interbank offered rate -- offer two key advantages to Japanese investors:

**STAMP annotation interface**
- Built for PropBank (Palme; UPenn)
- Target word
- Sentence
- Word sense choices (no mouse!)
Master Project Handler

Annotator ‘grabs’ word. Annotator name and date recorded (2 people per word)

When done, clicks here; system checks. When both are done, status is updated, agreement computed, and Manager is alerted

If Manager is happy, he clicks Commit; word is removed & stored for Database

Else he clicks Resense. Sensor and Adjudicator are alerted, and Sensor starts resensing. When done, she resubmits the word to the server, & it reappears here

This part visible to Admin people only

<table>
<thead>
<tr>
<th>Noun</th>
<th># of instances</th>
<th># of senses</th>
<th>Lock</th>
<th>Done</th>
<th>Agreement</th>
<th>Commit</th>
<th>Resense</th>
</tr>
</thead>
<tbody>
<tr>
<td>accident-n</td>
<td>22</td>
<td>2</td>
<td>Lock</td>
<td>Done</td>
<td>Lock: test(08-14-2006)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>accordance-n</td>
<td>2</td>
<td>2</td>
<td>Lock</td>
<td>Done</td>
<td>Lock: test(08-12-2006)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>activity-n</td>
<td>245</td>
<td>3</td>
<td>Lock</td>
<td>Done</td>
<td><em>Resensed</em>: slkvel, mcorne</td>
<td></td>
<td></td>
</tr>
<tr>
<td>advantage-n</td>
<td>76</td>
<td>2</td>
<td>Lock</td>
<td>Done</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>advertising-n</td>
<td>138</td>
<td>3</td>
<td>Lock</td>
<td>Done</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>agriculture-n</td>
<td>11</td>
<td>4</td>
<td>Lock</td>
<td>Done</td>
<td>Lock: test(08-12-2006)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>aid-n</td>
<td>101</td>
<td>3</td>
<td>Lock</td>
<td>Done</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aim-n</td>
<td>20</td>
<td>4</td>
<td>Lock</td>
<td>Done</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>air-n</td>
<td>89</td>
<td>7</td>
<td>Lock</td>
<td>Done</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>allocation-n</td>
<td>11</td>
<td>3</td>
<td>Lock</td>
<td>Done</td>
<td>Lock: test(08-12-2006)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ambassador-n</td>
<td>7</td>
<td>2</td>
<td>Lock</td>
<td>Done</td>
<td>Lock: test(08-12-2006)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>appraisal-n</td>
<td>7</td>
<td>2</td>
<td>Lock</td>
<td>Done</td>
<td>Lock: test(08-13-2006)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>arbitration-n</td>
<td>5</td>
<td>2</td>
<td>Lock</td>
<td>Done</td>
<td>Lock: test(08-13-2006)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>arm-n</td>
<td>53</td>
<td>8</td>
<td>Lock</td>
<td>Done</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Status page

Dynamically updated

http://arjuna.isi.edu:8000/Ontobank/AnnotationStats.html

Current status: # nouns annotated, # adjudicated; agreement levels, etc.

Agreement histogram

Individual noun stats: annotators, agreement, # sentences, # senses

Confusion matrix for results
### Agreement Analysis

Sometimes, one annotator is bad.

Sometimes, the senses are bad.

Sometimes, the word is just hard.

<table>
<thead>
<tr>
<th>noun</th>
<th>total annotated</th>
<th>number adjudicated</th>
<th>%adj</th>
<th>A1-A2 agr</th>
<th>A1-A2 agr%</th>
<th>A1-Adj agr%</th>
<th>A2-Adj agr%</th>
<th>Col G+H</th>
<th>What to do</th>
</tr>
</thead>
<tbody>
<tr>
<td>term</td>
<td>349</td>
<td>64</td>
<td>18.3</td>
<td>285</td>
<td>81.7</td>
<td>87.5</td>
<td>10.9</td>
<td>98.4</td>
<td>A2 bad</td>
</tr>
<tr>
<td>amount</td>
<td>310</td>
<td>78</td>
<td>25.2</td>
<td>232</td>
<td>74.8</td>
<td>91.0</td>
<td>8.9</td>
<td>99.9</td>
<td>A2 bad</td>
</tr>
<tr>
<td>return</td>
<td>281</td>
<td>52</td>
<td>18.5</td>
<td>229</td>
<td>81.5</td>
<td>13.4</td>
<td>84.6</td>
<td>98.0</td>
<td></td>
</tr>
<tr>
<td>payment</td>
<td>270</td>
<td>73</td>
<td>27.0</td>
<td>197</td>
<td>73.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>control</td>
<td>262</td>
<td>161</td>
<td>61.5</td>
<td>102</td>
<td>38.9</td>
<td>26.1</td>
<td>71.4</td>
<td>97.5</td>
<td></td>
</tr>
<tr>
<td>activity</td>
<td>245</td>
<td>140</td>
<td>57.1</td>
<td>108</td>
<td>44.1</td>
<td>10.7</td>
<td>91.4</td>
<td>102.1</td>
<td>A1 bad</td>
</tr>
<tr>
<td>building</td>
<td>231</td>
<td>38</td>
<td>16.5</td>
<td>193</td>
<td>83.5</td>
<td>36.8</td>
<td>63.2</td>
<td>100.0</td>
<td>A2 bad</td>
</tr>
<tr>
<td>average</td>
<td>220</td>
<td>16</td>
<td>7.3</td>
<td>191</td>
<td>86.8</td>
<td>100.0</td>
<td>0.0</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>place</td>
<td>205</td>
<td>137</td>
<td>66.8</td>
<td>68</td>
<td>33.2</td>
<td>65.7</td>
<td>26.3</td>
<td>92.0</td>
<td></td>
</tr>
<tr>
<td>support</td>
<td>198</td>
<td>27</td>
<td>13.6</td>
<td>171</td>
<td>86.4</td>
<td>25.9</td>
<td>74.1</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>department</td>
<td>145</td>
<td>0</td>
<td>0.0</td>
<td>145</td>
<td>100.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>marketing</td>
<td>167</td>
<td>85</td>
<td>50.9</td>
<td>83</td>
<td>49.7</td>
<td>60.0</td>
<td>40.0</td>
<td>100.0</td>
<td>split</td>
</tr>
<tr>
<td>game</td>
<td>163</td>
<td>60</td>
<td>36.8</td>
<td>125</td>
<td>76.7</td>
<td>86.7</td>
<td>60.0</td>
<td>146.7</td>
<td></td>
</tr>
<tr>
<td>import</td>
<td>157</td>
<td>104</td>
<td>66.2</td>
<td>59</td>
<td>37.6</td>
<td>76.0</td>
<td>29.8</td>
<td>105.8</td>
<td></td>
</tr>
<tr>
<td>competition</td>
<td>152</td>
<td>97</td>
<td>63.8</td>
<td>5</td>
<td>3.3</td>
<td>42.2</td>
<td>57.7</td>
<td>99.9</td>
<td>split</td>
</tr>
<tr>
<td>situation</td>
<td>143</td>
<td>49</td>
<td>34.3</td>
<td>76</td>
<td>53.1</td>
<td>65.3</td>
<td>42.9</td>
<td>108.2</td>
<td></td>
</tr>
<tr>
<td>material</td>
<td>129</td>
<td>30</td>
<td>23.3</td>
<td>99</td>
<td>76.7</td>
<td>10.0</td>
<td>90.0</td>
<td>100.0</td>
<td>A1 bad</td>
</tr>
<tr>
<td>form</td>
<td>131</td>
<td>31</td>
<td>23.7</td>
<td>100</td>
<td>76.3</td>
<td>58.1</td>
<td>38.7</td>
<td>96.8</td>
<td>split</td>
</tr>
<tr>
<td>trend</td>
<td>113</td>
<td>28</td>
<td>24.8</td>
<td>86</td>
<td>76.1</td>
<td>17.9</td>
<td>85.7</td>
<td>103.6</td>
<td></td>
</tr>
<tr>
<td>protection</td>
<td>111</td>
<td>41</td>
<td>36.9</td>
<td>70</td>
<td>63.1</td>
<td>22.0</td>
<td>78.0</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>date</td>
<td>102</td>
<td>84</td>
<td>82.4</td>
<td>18</td>
<td>17.6</td>
<td>23.8</td>
<td>72.6</td>
<td>96.4</td>
<td></td>
</tr>
<tr>
<td>requirement</td>
<td>95</td>
<td>86</td>
<td>90.5</td>
<td>9</td>
<td>9.5</td>
<td>95.4</td>
<td>3.5</td>
<td>98.9</td>
<td>A2 bad</td>
</tr>
<tr>
<td>saving</td>
<td>89</td>
<td>59</td>
<td>66.3</td>
<td>29</td>
<td>32.6</td>
<td>96.6</td>
<td>3.4</td>
<td>100.0</td>
<td>A2 bad</td>
</tr>
<tr>
<td>structure</td>
<td>87</td>
<td>19</td>
<td>21.8</td>
<td>68</td>
<td>78.2</td>
<td>100.0</td>
<td>0.0</td>
<td>100.0</td>
<td>A2 bad</td>
</tr>
<tr>
<td>recovery</td>
<td>75</td>
<td>17</td>
<td>22.7</td>
<td>58</td>
<td>77.3</td>
<td>76.5</td>
<td>23.5</td>
<td>100.0</td>
<td>A2 bad</td>
</tr>
<tr>
<td>traffic</td>
<td>57</td>
<td>16</td>
<td>28.1</td>
<td>42</td>
<td>73.7</td>
<td>81.2</td>
<td>6.2</td>
<td>87.4</td>
<td>A2 bad</td>
</tr>
<tr>
<td>challenge</td>
<td>54</td>
<td>26</td>
<td>48.1</td>
<td>34</td>
<td>63.0</td>
<td>73.0</td>
<td>50.0</td>
<td>123.0</td>
<td></td>
</tr>
<tr>
<td>location</td>
<td>54</td>
<td>17</td>
<td>31.5</td>
<td>37</td>
<td>68.5</td>
<td>88.2</td>
<td>11.8</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>merchant</td>
<td>51</td>
<td>34</td>
<td>66.7</td>
<td>17</td>
<td>33.3</td>
<td>0.0</td>
<td>100.0</td>
<td>100.0</td>
<td>A1 bad</td>
</tr>
<tr>
<td>beginning</td>
<td>50</td>
<td>25</td>
<td>50.0</td>
<td>26</td>
<td>52.0</td>
<td>60.0</td>
<td>44.0</td>
<td>104.0</td>
<td>split</td>
</tr>
</tbody>
</table>
## Annotation rates: English

### English

<table>
<thead>
<tr>
<th></th>
<th>#types = 9190</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>avg at 3/15</td>
</tr>
<tr>
<td></td>
<td>3/15 - 4/15</td>
</tr>
<tr>
<td></td>
<td>4/15 - 5/15</td>
</tr>
<tr>
<td></td>
<td>5/15 - 6/28</td>
</tr>
<tr>
<td></td>
<td>6/28 - 8/15</td>
</tr>
<tr>
<td></td>
<td>8/15 - 9/25</td>
</tr>
<tr>
<td></td>
<td>9/25 - 12/10</td>
</tr>
<tr>
<td></td>
<td>12/10 - 2/10</td>
</tr>
<tr>
<td></td>
<td>2/15 - 3/20</td>
</tr>
<tr>
<td>sensed</td>
<td>136 145 249 315 370 500 630 731 754</td>
</tr>
<tr>
<td>hours sensing</td>
<td>9 104 66 55 130 130 101 23</td>
</tr>
<tr>
<td>d-annot types</td>
<td>138 149 217 272 359 415 465 540 570</td>
</tr>
<tr>
<td>(words)</td>
<td>17.5 18.9 24.3 31.3 43.3 44.7 46.4 47.6 48.6</td>
</tr>
<tr>
<td>d-annot types (% of corpus)</td>
<td>1.4 5.4 7 12 1.4 1.7 1.2 1</td>
</tr>
<tr>
<td>hours annotating</td>
<td>353.9 115.1 69.7 106.4 197 56.8 111.2 165.7 352.9</td>
</tr>
</tbody>
</table>

Rate varies widely: due to re-sensing?
**Annotator work record**

Most recent week, each person:
- Total time
- Avg rate
- % of time working at acceptable rate (3/min)
- # sentences at acceptable rate

Full history of each person, weekly
English noun annotation stats

- Annotators at ISI:
  - About 9 regular annotators for English, 6 for Chinese
  - All trained on “bank”
  - Weekly telecons for discussion

- Status (October 06):
  - Avg. agreement: 91%
  - Residual disagreements adjudicated by linguist
  - Slow start, but speeding up

<table>
<thead>
<tr>
<th>2006</th>
<th>Date</th>
<th>Sensed</th>
<th>Sensing target</th>
<th>Double-annotated</th>
<th>Double-ann coverage</th>
<th>Adjudicated</th>
<th>Final target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>0</td>
<td>200</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>150</td>
<td></td>
</tr>
<tr>
<td>Feb</td>
<td>59</td>
<td>250</td>
<td>102</td>
<td>13</td>
<td>200</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mar</td>
<td>136</td>
<td>375</td>
<td>138</td>
<td>15.5%</td>
<td>44</td>
<td>275</td>
<td></td>
</tr>
<tr>
<td>Apr</td>
<td>145</td>
<td>450</td>
<td>149</td>
<td>18.9%</td>
<td>44</td>
<td>350</td>
<td></td>
</tr>
<tr>
<td>May</td>
<td>5/15</td>
<td>249</td>
<td>525</td>
<td>217</td>
<td>24.3%</td>
<td>51</td>
<td>450</td>
</tr>
<tr>
<td>Jun</td>
<td>6/28</td>
<td>315</td>
<td>625</td>
<td>272</td>
<td>31.3%</td>
<td>128</td>
<td>550</td>
</tr>
<tr>
<td>Jul</td>
<td></td>
<td>720</td>
<td></td>
<td></td>
<td></td>
<td>650</td>
<td></td>
</tr>
<tr>
<td>Aug</td>
<td>8/15</td>
<td>~370</td>
<td>850</td>
<td>359</td>
<td>43.3%</td>
<td>149</td>
<td>800</td>
</tr>
<tr>
<td>Sep</td>
<td>9/25</td>
<td>500</td>
<td>980</td>
<td>415</td>
<td>44.7%</td>
<td>214</td>
<td>950</td>
</tr>
<tr>
<td>Oct</td>
<td>10/30</td>
<td>568</td>
<td>1100</td>
<td>451</td>
<td>46.1%</td>
<td>221</td>
<td>1100</td>
</tr>
<tr>
<td>Nov</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dec</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Some preliminary statistics

- For 465 most frequent nouns in WSJ annotated:
  - total senses = about 2080
  - average number of senses per word = 4.47
  - 60.8% of nouns have 2–4 senses
    - “head” has largest number of senses: 32 senses
  - 78.9% of the polysemous nouns in WSJ need only one sense (predominant sense) (!)
  - 93.3% instances are covered by topmost 2 senses
  - 497 senses (23.9%) do not occur at all (!)
  - 254 nouns (54.6%) have at least one unseen sense (!)
  - Nouns, sorted by entropy of tags
- So: WSJ part of OntoNotes is an *unbalanced corpus* — we need another as well
  - It is very difficult to use such a skewed corpus for identifying infrequent or unseen senses

<table>
<thead>
<tr>
<th>Noun</th>
<th>Senses</th>
<th>Instances</th>
<th>Agreement</th>
<th>Entropy</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>play-n</td>
<td>13</td>
<td>41</td>
<td>0.490000</td>
<td>2.003143</td>
<td>0.239 0.217 0.130 0.109 0.087 0.087 0.065 0.043 0.021739 0.00</td>
</tr>
<tr>
<td>control-n</td>
<td>8</td>
<td>262</td>
<td>0.820000</td>
<td>1.788795</td>
<td>0.416 0.168 0.119 0.097 0.065 0.058 0.026 0.023 0.016 0.013</td>
</tr>
<tr>
<td>defense-n</td>
<td>8</td>
<td>149</td>
<td>1.000000</td>
<td>1.776716</td>
<td>0.261 0.248 0.164 0.128 0.106 0.058 0.035 0.00 0.00 0.00</td>
</tr>
<tr>
<td>bar-n</td>
<td>16</td>
<td>27</td>
<td>1.000000</td>
<td>1.767424</td>
<td>0.310 0.241 0.138 0.103 0.069 0.069 0.069 0.00 0.00 0.00</td>
</tr>
<tr>
<td>life-n</td>
<td>9</td>
<td>272</td>
<td>0.620000</td>
<td>1.763602</td>
<td>0.333 0.187 0.184 0.075 0.066 0.043 0.026 0.020 0.014 0.006</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>arbitration-n</td>
<td>2</td>
<td>5</td>
<td>0.800000</td>
<td>0.500402</td>
<td>0.80 0.20 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00</td>
</tr>
<tr>
<td>condition-n</td>
<td>4</td>
<td>171</td>
<td>0.930000</td>
<td>0.500555</td>
<td>0.817 0.178 0.005 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00</td>
</tr>
<tr>
<td>value-n</td>
<td>5</td>
<td>403</td>
<td>0.950000</td>
<td>0.497184</td>
<td>0.849 0.123 0.028 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00</td>
</tr>
<tr>
<td>technology-n</td>
<td>2</td>
<td>203</td>
<td>0.820000</td>
<td>0.494763</td>
<td>0.871 0.072 0.049 0.008 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00</td>
</tr>
<tr>
<td>policy-n</td>
<td>2</td>
<td>382</td>
<td>1.000000</td>
<td>0.493565</td>
<td>0.805 0.195 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>accordance-n</td>
<td>2</td>
<td>2</td>
<td>1.000000</td>
<td>0.000000</td>
<td>1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00</td>
</tr>
<tr>
<td>...same all the way down</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Current status (month 23)

- **Text annotated:**
  - Newswire text:
    - 300K words of Wall Street Journal: 800+ verbs and 500+ nouns; verb arg structure; coref links
  - Primarily broadcast news:
    - Broadcast news: 200K English and 300K Chinese (same number of nouns and verbs; all corefs)
    - Starting with 100K Arabic newswire data
  - Next year:
    - Broadcast conversations (200K words of English, 150K Chinese; all corefs)
  - Later:
    - Weblogs, newsgroups, etc.
- **Ontology:** Terms now being converted to concepts and taxonomized/inserted into overall structure
Database: Unified relational rep

(Slide by Sameer Pradhan, BBN)
Example: DB representation of syntax

- Treebank tokens (stored in the Token table) provide the common base
- The Tree table stores the recursive tree nodes, each with its span
- Subsidiary tables define the sets of function tags, phase types, etc.
Talk overview

1. Introduction: A new role for annotation?
2. Example: Semantic annotation in OntoNotes
3. Toward a science of annotation: 7 questions
4. Conclusion
The generic annotation pipeline

Theory 1
(Domain)

Theory 2
(Linguistics)

Theory 3
(Another field)

Model-building for IE / annotation

Annotation

IE engine: training and application

Evaluation and verification

90%?

Feedback

Corpus

1

2

3
Annotation: The 7 core questions

1. **Preparation**
   - Choosing the corpus — which corpus? What are the political and social ramifications?
   - How to achieve balance, representativeness, and timeliness? What does it even mean?

2. ‘**Instantiating’ the theory**
   - Creating the annotation choices — how to remain faithful to the theory?
   - Writing the manual: this is non-trivial
   - Testing for stability

3. **Interface design**
   - Building the interfaces. How to ensure speed and avoid bias?

4. **The annotators**
   - Choosing the annotators — what background? How many?
   - How to avoid overtraining? And undertraining? How to even know?

5. **Annotation procedure**
   - How to design the exact procedure? How to avoid biasing annotators?
   - Reconciliation and adjudication processes among annotators

6. **Validation**
   - Measuring inter-annotator agreement — which measures?
   - What feedback to step 2? What if the theory (or its instantiation) ‘adjusts’?

7. **Delivery**
   - Wrapping the result — in what form?
   - Licensing, maintenance, and distribution
Q1. Prep: Choosing the corpus

- Choose carefully—the future will build on your work!
  - (When to re-use something?—Today, we’re stuck with WSJ…)

- **Technical issues:** *Balance, representativeness, and timeliness*
  - When is a corpus representative? —“stock” in WSJ is *never* the soup base
    - Methodology of ‘principled’ corpus construction for representativeness
      (even BNC process rather ad hoc)
    - How to balance genre, era, domain…See (Kilgarriff and Grefenstette, CL 2003)
      - Effect of (expected) usage of corpus
  - **Experts:** corpus linguists or domain specialists

- **Social, political, funding issues:**
  - How do you ensure agreement / complementarity with others? Should you bother?
  - How do you choose which phenomena to annotate? Need high payoff…
  - How do you convince funders to invest in the effort?
Q1. Prep: What’s available

• Corpus collections are worth their weight in gold
  – Unencumbered by copyright
  – Available to whole community — standardized results for comparison

• Raw and processed text and speech:
  – Linguistic Data Consortium (LDC), UPenn: www.ldc.upenn.edu/
Q2: Instantiating the theory

- What to annotate? How ‘deeply’ to instantiate theory?
  - Design rep scheme / formalism very carefully — simple and transparent
  - ? Depends on theory — but also (yes? how much?) on corpus and annotators
  - Do tests first, to determine what is annotatable in practice

- Experts must create:
  - Annotation categories
  - Annotator instruction (coding) manual
  - Experts to build the manual: theoreticians? Or exactly NOT the theoreticians?

- Both must be tested! — Don’t ‘freeze’ the manual too soon
  - Experts annotate a sample set; measure agreements
  - Annotators keep annotating a sample set until stability is achieved

- Likely problems:
  - Categories not exhaustive over phenomena
  - Categories badly defined / unclear (intrinsic ambiguity, or relying on bg knowl?)

- Measuring stability — measures of agreement: (Teachman 1989)
  - Precision (correctness) = $P_i = \frac{\#\text{correct}}{N}$
  - Entropy (ambiguity, regardless of correctness) = $-\sum_i P_i \cdot \ln P_i$ (unambig $\to 0$)
  - Odds Ratio (distinguishability of categories) = $\frac{f_{xx}f_{yy}}{f_{xy}f_{yx}}$ (indistinguishable $\to 0$)
Q2: Theory and model

• ‘Neutering’ the theory: when the theory is controversial, or you cannot obtain stability — you may still be able to annotate, using a more neutral set of terms
  – E.g., from Case Roles (Agent, Patient, Instrument) to PropBank’s roles (arg0, arg1, argM) — user chooses desired role labels and maps PropBank roles to them

• What does this say about the theory, however?
Q3: The interface

- How to design adequate interfaces?
  - Maximize speed!
  - Create very simple tasks—but how simple? Boredom factor, but simple task means less to annotate before you have enough
  - Don’t use the mouse
  - Customize the interface for each annotation project?
  - Don’t bias annotators (avoid priming!)
    - Beware of order of choice options
    - Beware of presentation of choices
    - Is it ok to present together a whole series of choices with expected identical annotation? — annotate *en bloc*?
  - Check agreements and hard cases in-line?
    - Do you show the annotator how ‘well’ he/she is doing? Why not?
- Experts: Psych experimenters; Gallup Poll question creators
- Experts: interface design specialists
Q3. Interface: What’s available

- Interfaces/annotation tools:
  - ATLAS.TI: annotation toolkit (www.atlasti.com/)
  - Ad hoc annotation interfaces and tools from the NLP community

- Annotation standards:
  - Various XML and other notations
  - Standard backoff and other alternatives
  - Romary and Ide (2007): ISO annotation notation standards committee (ISO TC37 SC4 WG1)
    - Criteria: Expressive adequacy, media independence, semantic adequacy, incrementality for new info in layers, separability of layers, uniformity of style, openness to theories, extensibility to new ideas, human readability, computational processability, internal consistency
Q4: Annotators

• How to choose annotators?
  – Annotator backgrounds — should they be experts, or precisely not?
  – Biases, preferences, etc.
    – Experts: Psych experimenters

• How much to train the annotators?
  – Undertrain: Instructions are too vague or insufficient. Result: annotators create their own ‘patterns of thought’ and diverge from the gold standard, each in their own particular way (Bayerl 2006)
    • How to determine?: Use Odds Ratio to measure pairwise distinguishability of categories
    • Then collapse indistinguishable categories, recompute scores, and (?) reformulate theory — is this ok?
    • Basic choice: EITHER ‘fit’ the annotation to the annotators — is this ok? OR train annotators more — is this ok?
  – Overtrain: Instructions are so exhaustive that there is no room for thought or interpretation (annotators follow a ‘table lookup’ procedure)
    • How to determine: is task simply easy, or are annotators overtrained?
    • What’s really wrong with overtraining? No predictive power…

• Who should train the annotators?
  – Is it ok for the interface builder, or the learning system builder? — not: they have an agenda
Q5.1: Annotation procedure

• How to manage the annotation process?
  – When annotating multiple variables, annotate each variable separately, across whole corpus — speedup and local expertise … but lose context
  – The problem of ‘annotation drift’: shuffling and redoing items
  – Annotator attention and tiredness; rotating annotators
  – Complex management framework, interfaces, etc.

• The Wiebe ‘85% clear cases’ rule
  – Ask the annotators also to mark their certainty
  – There should be a lot of agreement at high certainty — the clear cases

• Reconciliation
  – Allow annotators to discuss problematic cases, then continue — can greatly improve agreement but at the cost of drift / overtraining

• Backing off: In cases of disagreement, what do you do?
  – (1) make option granularity coarser; (2) allow multiple options; (3) increase context supporting annotation; (4) annotate only major / easy cases

• Adjudication
  – Have an expert (or more annotators) decide in cases of residual disagreement — but how much disagreement can be tolerated before just redoing the annotation?

• Experts: …?
Q5.2: Annotation procedure

- Overall approach — Shulman’s rule: do the easy annotations first, so you’ve seen the data when you get to the harder cases
- The Rosé hypothesis: for up to 50% incorrect instances, it pays to show the annotator possibly buggy annotations and have them correct them (compared to having them annotate anew)

- **Active learning**: In-line process to dynamically find problematic cases for immediate tagging (more rapidly get to the ‘end point’), and/or to pre-annotate (help the annotator under the Rosé hypothesis)
  - Benefit: speedup; danger: misleading annotators
Q6.1: Validating annotations

- Evaluating individual pieces of information:
  - What to evaluate:
    - Individual agreement scores between creators
    - Overall agreement averages?
  - What measure(s) to use:
    - Simple agreement is biased by chance agreement — however, this may be fine, if all you care about is a system that mirrors human behavior
    - Kappa is better for testing inter-annotator agreement. But it is not sufficient — cannot handle multiple correct choices, and works only pairwise
    - Krippendorff’s alpha, Kappa variations…; see (Bortz 05; 6th ed; in German)
  - Tolerances:
    - When is the agreement no longer good enough? — why the 90% rule? (Marcus’s rule: if humans get N%, systems will achieve (N-10)%)
  - The problem of asymmetrical/unbalanced corpora
    - When you get high agreement but low Kappa — does it matter? An unbalanced corpus makes choice easy but Kappa low. Are you primarily interested in annotation qua annotation, or in doing the task?
- Experts: Psych experimenters and Corpus Analysis statisticians
Q6.2: Validating someone’s corpus

• But also, evaluate aspects of ‘metadata’:
  – **Theory and model**:
    • What is the underlying/foundational theory?
    • Is there a model of the theory for the annotation? What is it?
    • How well does the corpus reflect the model? And the theory? Where were simplifications made? Why? How?
  – **Creation**:
    • What was the procedure of creation? How was it tested and debugged?
    • Who created the corpus? How many people? What training did they have, and require? How were they trained?
    • Overall agreement scores between creators
    • Reconciliation/adjudication/purification procedure and experts
  – **Result**:
    • Is the result enough? What does ‘enough’ mean? (Sufficiency: when the machine learning system shows no increase in accuracy despite more training data)
    • Is the result consistent (enough)?
    • Is it correct? (can be correct in various ways!)
    • How was it used?
Q7: Delivery

• It’s not just about annotation…
  How do you make sure others use the corpus?

• Technical issues:
  – Licensing
  – Distribution
  – Support/maintenance (over years?)
  – Incorporating new annotations/updates: layering
  – Experts: Data managers
Talk overview

1. Introduction: A new role for annotation?
2. Example: Semantic annotation in OntoNotes
3. Toward a science of annotation: 7 questions
4. Conclusion
Writing a paper in the new style

• How to write a paper about an annotation project (and make sure it will get accepted at LREC, ACL, etc.)?

• Recipe:
  – Problem: phenomena addressed
  – Theory
    • Relevant theories and prior work
    • Our theory and its terms, notation, and formalism
  – The corpus
    • Corpus selection
    • Annotation design, tools, and work
  – Agreements achieved, and speed, size, etc.
  – Conclusion
    • Distribution, use, etc.
    • Future work

Current equiv
  problem
  past work
  training
  algorithm
  evaluation
  distribution
Some current technology and work

• Wide variety of NLP / machine learning technology available to learn to mimic human annotations:
  – Simple phrasal patterns (regular expressions)
  – Automated phrasal pattern learning algorithms
  – Markov Models and Conditional Random Fields

• Kinds of information typically used for learning experiments in NLP community:
  – Parts of speech — solved problem for many languages
  – Named Entities (people, places, organizations, dates, amounts, etc.) — e.g., BBN’s IdentiFinder
  – Syntactic structure — somewhat solved for some languages
  – Word senses and argument structure (lexico-semantics)
  – Opinions (both good/bad judgments and true/false beliefs)
  – Coreference links (pronouns and other anaphora)
  – Discourse structure
  – Various other semantic phenomena — more experimental
In conclusion…

Annotation is **both**:

- A mechanism for providing new training material for machines
- A mechanism for theory formation and validation — in addition to domain specialists, annotation can involve linguists, philosophers of language, etc. in a new paradigm
It’s not only NOT the most boring thing the world…
…it’s actually becoming COOL (obviously, since we are here now, in this workshop)

Thank you!