Goals, Approach and Methodology

Content of
Linguistic
Annotation:
Standards and
Practices

New York University
November 7, 2009
Adam Meyers

Overview

- Standardization Goals
- Standards With the Greatest Impact on NLP
- How to Facilitate Standards Adoption
- Working Group Goals
- Goals of This Meeting

Conflicting Goals of Good Standardization

- Limit Alternative Representation
  - Burden of proof to create new analysis
    - Must show inadequacy of previous analyses
  - Require mapping rules
  - ISO Data Category Registry
- Do not stifle research
  - No unnecessary standardization
  - Minimal commitment to theoretical positions

More conflicting Goals: Descriptive Adequacy vs. Overhead

- Simple Standards
  - **Easy to implement, understand, agree with**
  - **Have Loopholes which cause disagreement**
- Complex Standards
  - **Harder to implement, understand, agree with**
  - **Remove Loopholes**
- Options
  - 2 Sets of Standards: simple and complex
  - Compromise:
    - Cover as much as possible
    - Leave some loopholes

Ultimate Goal of Standardization: Interoperability

- Annotation Mergable into One Framework
  - CONLL 2008/2009, GLARF, Ontonotes, Masc
  - Same tokens, phrases and/or anchors
  - The more shared assumptions, the better
- Combined features for machine learning
  - Best with shared assumptions about basic units
    - ACE, GALE and other large shared tasks
  - Ex: Are ARG0s of *attack* verbs likely to be coreferential with ARG1s of *prosecute* verbs?
- Creating larger body of training data
  - Manual POS tagsets, NE tags, SRL

Standardization of Basic Units is Crucial to Interoperability

- Characters
- Tokens
- Token Groupings: Constituents, Dependency Graph-lets, Chunks, etc.
- Sentences, Utterances, etc.
- Paragraphs, Turns, etc.
- Documents, Document Collections, Networks of Webpages, Genres, Epochs, ...

Tokenization

- Canonically: tokens = units separated by spaces
- Punctuation that divides words
  - *New York-based* → *New* + *York* + **-** + *based*
  - *U.S./Japan treaty* → *U.S.* + */* + *Japan* + *treaty*
- Beginning and End Word Punctuation
  - *"The beginning* → *"* + *The* + *beginning*
  - *ever after.* → *ever* + *after* + .
- Division without spaces or punctuation
  - *doesn't* → *does* + *n't*
  - cannot → can + not

Do Tokens Uniquely Partition the Input String into Substrings?

- *Shared periods*
  - *They had ducks, cows, etc.* →
    They + *had* + *ducks* + , + *cows* + , + *etc.* + .
- Contracted forms
  - *Can't* → *can* + *n't*
    - *can* + *'t*
    - *ca* + *n't*
    - *can* + *not*
  - *wanna* → *wan* + *na*
    - *want* + *na*
    - *want* + *to*

How Much String Regularization is Part of Tokenization?

- *Big Blue won't go topsy turvy* → *IBM + will + not + go + topsy turvy*
  - Aliasing: *Big Blue* → *IBM*
  - Contraction Regularization: *won't* → *will + not*
  - Recognition of multi word units: *topsy turvy*
- *They hav gone to the theatre* → *Pron + 3P + have + 3P + go + pastpart + to + the + theater*
  - Spelling correction: *hav* → *have*
  - Morphology: *gone* → *go + pastpart*
  - Spelling regularization: *theatre* → *theater*

Tokenization and the 800 Pound Gorilla

- Penn Treebank
  - www.cis.upenn.edu/~treebank/tokenizer.sed
  - New rules, e.g., hyphenation
- Tokenization Efforts Needs to Align with PTB
  - Adopt PTB rules or map to/from them
  - Coordination is Important as PTB tokenization is refined
- Are there other Gorillas for English, e.g., BNC?
  - crel.lancs.ac.uk/bnc2sampler/guide_c7.htm#m1b

Other “levels” related to Tokenization

- Penn Treebank POS, Morphology/Contraction
  - *gone/VBN*
  - *gonna → gon/VBG + na/TO*
- Aliasing is usually treated as part of anaphora
  - *IBM: Big Blue*
  - *NYC: Big Apple*
- Recognition of Multi-word-expressions
  - Non-trivial due to variation, modification, etc.

Identifying Larger Units

- Chunks: less constrained, phrase-like units
  - No commitment to internal structure, sentence partitioning or full theoretical framework
- Phrases
  - Theoretically grounded sets of tokens that completely partition sentence/utterance/string
  - No commitment to head assignment
- Dependency Graph-Lets
  - Rooted subgraphs of theoretically grounded dependency structure
  - Root = Head
  - Structure tends to be flatter than phrase structure
  - No commitment to consecutiveness of leaves

Phrase Structure Devotees are Likely to Agree On Dependency Heads

- Most Phrases are Assumed to Have Heads
- The ultimate head (head of head of head...) of a phrase is usually the same as the head of the corresponding dependency graph-let
- Choice of ultimate head is likely to be the same among competing phrase structure accounts
- Therefore, dependency heads might be a good basis for an easily sharable standard?
  - Head-based equivalency relations for phrases can be useful for evaluation purposes

Unfortunately, Not all Phrases Have (Uncontroversial) Heads

- Coordinate structures, named entities, range phrases, rate phrases, *the more the merrier*, etc.
- Different dependency theories have different strategies for identifying "heads" for such constructions
  - So dependency representations are usually rooted graphs, not forests
- This diversity of opinion is a barrier to standardization

Descriptively Adequate Account Generalizes Head to "Anchor"

- Most Anchors do one of the following:
  - Provide Lexical Features
    - For phrase or dependency graph-let
  - Provide a predictable path to
    - Anchors that provide lexical features
    - Coordination, support constructions, light verbs, copulas, etc.
  - Consist of multiple tokens
    - NEs, Dates, Certain Idioms
- May Require 2 Tiers: Surface and "Deep"
- Some Uglinesses
  - Null theory-internal anchors
  - Parenthetical material, false starts, etc.

Easy to Implement

- Headed Constituents
  - Same as descriptively adequate approach
- Heuristics Identify "Good Enough" Anchor
  - First or Last Constituent
    - Provides Link so Graph is Complete
  - Sometimes Provides Some Lexical Info
    - First Conjunct of Coordinate Structure
    - Note: Can be elaborated to account for lexical properties of all conjuncts

How can we overcome obstacles to the acceptance of standards?

- SIGANN Seal of Approval
  - Qualifying a Standard
  - Committee to Decide if Effort Follows Standards
  - Will this be meaningful?
- Peer Review
  - Papers, Conferences, Grant Proposals
  - Should standards compliance be a factor?
- Peer Pressure
  - A base of compliant efforts could get the ball rolling
  - Funding for small compliant annotation efforts?

Working Group Members

- Policy: Cieri, Joshi, Meyers, Palmer
- Scope: Calzari, Ide, Prasad, Pustejovsky, Wiebe
- Tokenization: Baker, Boguraev, Macleod, Mota, Xue
- Anchor: Flickinger, Fillmore, Hajic, Rambow, Sun, Uresova

Working Group Schedule

- Rooms (Alternative = 13th Floor Lounge)
  - Policy: Room 101
  - Scope: Room 512
  - Tokenization: Room 517
  - Anchor: Room 705
- 11:15 – 12:45
  - Create 20 minutes of Slides
  - Choose Leader to Present Them
- Presentations (20 minutes plus padding)
  - 1:45—2:08 Policy
  - 2:08—2:32 Scope
  - 2:32—2:56 Tokenization
  - 2:56—3:30 Anchor
- 4:15—4:45 Make 4 or 5 slides summarizing recommendations

Policy Working Group Goals

- What Process establishes a Standard?
  - Registry with ISO Data Registry?
  - Review and Recognition by SIGANN Committee?
  - What?
- How can/should enticements be used?
  - SIGANN seal of approval (or similar measure)
  - Peer Review
- How should standards apply to derivative tasks?
  - Ex: Does MT need to obey tokenization standards?
- How can/should "justified" exceptions be used to modify a standard?

Scope Working Group Goals

- Which Classes of Content Categories are Ready for Standardization? For each class answer:
  - Does this class effect interoperability?
  - Otherwise, what else could justify standardization?
- How should CLASP interact with Prior Standards?
  - Ex: ISO TC37 SC4 Language Resources Management
- To what extent should CLASP guidelines be anglocentric?
  - Should standards be written on a language by language basis?
  - Should standards be generalizable across languages

Tokenization Working Group

- What part of tokenization is absolutely **necessary**?
  - What part **should** be be standard?
- Does tokenization partition the set of characters in the input string (less white space)? Or are any characters changed, deleted or reused?
- How much regularization should tokenization include?
  - How much should be left to POS tagging, morphological analysis, NE tagging, etc.?
- Who are the 800 pound gorillas and should they be followed blindly?
- Ease of implementation
  - PTB's sed script
  - BNC's list of contracted forms

Anchor Working Group

- Anchors, in theory,
  - Carry lexical properties of graph-let
  - Or point to other anchors with this property
  - Useful for selection restrictions, etc.
- Multiple strategies – how many do we need?
  - Descriptively adequate, 2 types of anchors
  - Low-overhead heuristics, not always predictive
  - Compromises
- Completely connected graphs connecting all words
  - Convenient for systems that traverse such graphs
  - Requires idealization that ignore problem cases
  - Is it worth it? Are some compromises better than others?
- Possible implementation of guidelines: specifications and/or annotation effort?

What Should Come Out of this Workshop?

- A better understanding of the problem
  - What kind of content standards are needed?
  - How they should be implemented?
  - Who should pay attention to them and why?
- How specification development could be funded or otherwise supported?
  - Proposed annotation effort, e.g., anchors
  - Incentives for following standards
  - Accreditation processes for new standards
  - Accreditation procedures for compliance

Schedule

- 9:30-10:00 Breakfast
- 10:00-10:30 This talk
- 10:30-11:15 Nancy's talk
- 11:15-12:45 Working Groups
- 12:45-1:45 Catered Lunch (right outside)
- 1:45-3:15 Working Group Presentations
- 3:15-3:45 Group Discussion
- 3:45-4:15 Coffee Break
- 4:15-4:45 Preparation of Summary Slides
- 4:45-5:30 Summary Slide Presentations
- 5:30-6:00 Final Discussion