Policy Working Group
CLASP Workshop
NYU

November 7, 2009

Adam Meyers, Chris Cieri, Aravind Joshi, Martha Palmer

Catching more flies with honey

- Standards can't be forced. *US view?*
- Standards emerge naturally from groups working on similar tasks
- Honey1: Market forces, Ex. quickly integrating new languages into a multilingual system
- Honey2: Community Practices to encourage standards
  - Peer review that expects new frameworks to be situated with respect to pre-existing work
  - Workshops that raise awareness
  - Groups working on similar tasks can be encouraged to see how their annotation schemes can be reconciled
  - Emphasis on documentation

What processes can establish standards?

- It is constructive for specific annotation types to define mappings between alternative representations
  - Facilitates development of “adapters”!
  - Ensures international compatibility
  - Can be encouraged by a community consensus that could be encoded in the ISO Registry

- SIGANN Committee involvement
  - Only in promoting awareness, not mandating standards for proposal or paper reviews, for example

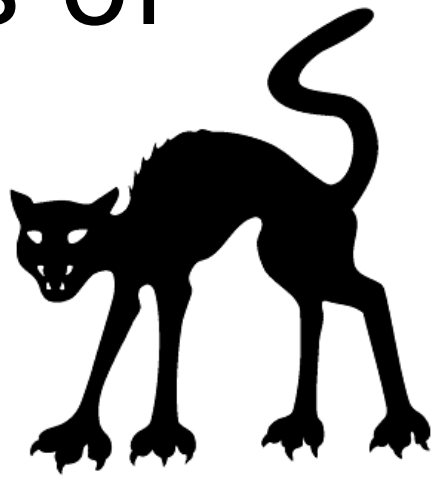
Ex. Semantic Roles

- FrameNet and PropBank (VerbNet) have quite different approaches
- These projects have communicated closely for quite awhile
  - Joint funding
  - Common ACL2004 tutorial (with Prague and Salsa)
  - Students working both projects, mapping between FrameNet and PropBank
  - Now a commitment to developing an ISO standard that will clarify similarities and differences and define a common ground which other SRL systems could easily map to

Ex. Discourse Treebank

- Pursuing annotations in English, Hindi, Turkish and Arabic simultaneously requires extensive discussions back and forth
- Facilitates the development of compatible annotation guidelines
- Highlights linguistic motivations for intrinsic differences

To publicize benefits of standards: Lessons learned



- Horror stories panel (w/ beer if possible)
  - The problems of merging different annotation layers too closely
    - The problems with PropBank using inflexible pointers to tree nodes
  - Lesson learned: Use standoff
    - Additional utility of standoff: Being able to view 2 different layers (or 2 alternatives to the same layer) simultaneously without having to actually merge the annotation, Xbank, GRAF, Panacea

Additional horror story

- The mind-boggling implications of different Chinese segmentations for later processing stages
  - CTB has one set of segmentation guidelines
  - GALE forced LDC to use slightly different guidelines for Chinese/English alignment
  - Result: none of the statistical Chinese parsers produced segmentations that were compatible with the alignments.
  - Segmentations need to be consistent for tokenization/postagging/parsing/word alignment/language models – new guidelines means redoing the entire pipeline