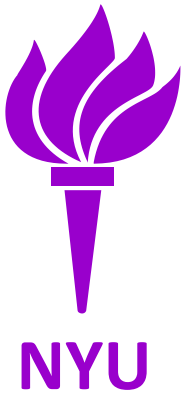


ICE:

a GUI for training extraction engines

CSCI-GA.2590

Ralph Grishman

Objectives

- Rapidly train extraction engines for new domains
- Use linguistic analysis to guide training
  - distributional analysis to build entity classes
  - bootstrapping to identify patterns for relations
- Interact with users in their own terms
  - using phrases, not formal representations
- Guide user
  - require judgments, not lots of examples from user
  - allow experienced users to direct process

Entity Classes

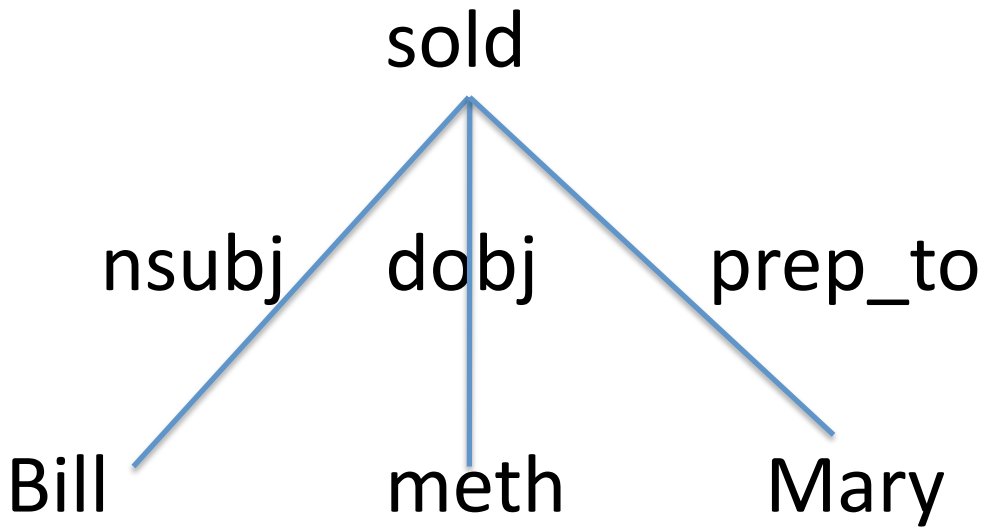
Entity classes represented by explicit sets

(Common sets – people, organizations, … – are predefined)

Relations

- Relations defined by sets of Lexicalized Dependency Paths (LDPs)
- Each LDP consists of
  - types of relation arguments
  - path in dependency tree, including
    - labels on dependency arcs
    - lemmatized forms of words

LDP example



LDP1: PERSON—nsubj^{-1}:sell:dobj—DRUG

LDP2: PERSON—nsubj^{-1}:sell:prep_to—PERSON

Process

- Read and analyze corpus
- Rank terms
    - both single and multi-word
- Create entity sets
- Find and rank labeled paths in corpus connecting pairs of entity mentions
- Build relations
    - get seed
    - bootstrap to find paraphrases

Using ICE

- We would like to try ICE out on several new domains
- To get started, we would set up an instance of ICE for each project *interested in using it*
- *We need*
  - *corpus with minimal mark-up*
  - *initial type dictionary*
  - *one or two relations with a couple of examples of each*