

# Representing Textual Content in a Generic Extraction Model

**Nancy McCracken**

**Center for Natural Language Processing  
School of Information Studies  
Syracuse University**

**March 27, 2002**

**AAAI Workshop**

**Acquiring and Using Linguistic and World Knowledge  
for Information Access**

# Natural Language Processing

- Goal
  - a range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for knowledge intensive applications
- Current Status
  - a document processing system using NLP to annotate text, using a generic model of entities and events
  - information extraction applications and visualization
  - open domain Question & Answering in progress

# Goals of our Document Processing

- Process documents quickly
  - 10,000 documents overnight
  - real-time processing of small pieces of text such as queries
- Extract
  - the **existence** and **attributes**
  - of **concrete objects** and **actions**
- Compromise
  - between the need to process documents quickly
  - and as complete a document semantics as possible

# NLP Document Processing

- Document processing through several levels of natural language processing is achieved through phases of a cascading finite-state parser.
- In 1978-79 Kintex delivered supplies to anti-revolution movements in Angola and South Africa.
- POS tagging with a Brill tagger
- <S> in|IN 1978-79|CD Kintex|NP delivered|VBN supplies|NN to|TO anti-revolution|JJ movements|NNS in|IN Angola|NP and|CC South|NP Africa|NP .|. </S>

# Additional Phases of Document Processing

- Bracket numeric phrases (NC), proper noun phrases (NP) and complex nominals (CN)
- `<S> in|IN <NC> 1978-79|CD </NC> <NP> Kintex|NP </NP> deliver|VBN supplies|NN to|TO <CN> anti-revolution|JJ movement|NNS </CN> in|IN <NP> Angola|NP </NP> and|CC <NP> South|NP Africa|NP </NP> .|. </S>`
- Categorize proper noun phrases from syntactic clues and some resources
- `<S> in|IN <NC cat="time"> 1978-79|CD </NC> <NP cat="co"> Kintex|NP </NP> deliver|VBN supplies|NN to|TO <CN> anti-revolution|JJ movement|NNS </CN> in|IN <NP cat="cntry"> Angola|NP </NP> and|CC <NP cat="cntry"> South|NP Africa|NP </NP> .|. </S>`

# Category Hierarchy

- General hierarchy of human knowledge, but more elaborated in concrete areas of human activity
- 17 main categories – about 150 in all – example showing level of detail in one category:

- 1 People
  - 1.1 Titles / Positions
  - 1.2 Groups
    - 1.2.1 Organizations
      - 1.2.1.1 Government Orgs
        - 1.2.1.1.1 Courts
        - 1.2.1.1.2 Lawmaking groups
        - 1.2.1.1.3 Military
      - 1.2.1.2 Terrorist Groups
      - 1.2.1.3 Military divisions
      - 1.2.1.20 Organizational subdivisions
    - 1.2.2 Companies
    - 1.2.3 Religion
    - 1.2.4 Sports Teams
  - 2 Thought, Communication and Communication Channels
  - 3 Buildings & Structures
  - 4 Substances, Materials, Objects, and Equipment

# Generic Event Model

- Events are the dynamic processes describing interactions among entities – typically verbs
- Shallow parsing rules match surface syntax of sentences and map it to semantic roles of verbs.
  - Patterned after the case frames of Fillmore.
  - Parse rules for basic sentence structure
- Called generic because the roles are left very general
  - typically generic roles agent, object, location, manner
- Last stages of document processing extract entities and events into frames
- (Final stage is coreference resolution within the document.)

# Relations of Entities and Roles of Events

- About 20 from Sowa's conceptual graphs

agent	duration	material
cause	experiencer	method
characteristic	instrument	part-of
content	location	point-in-time
destination	manner	source . . .

- About 30 new ones

acronym	affiliation	age
alias	amount	area
associated	body-part	charge
condition	cost	date
date-of-birth	date-of-death	distance
frequency	geographic-affiliation	isa . . .

# Events and Roles Represented as Frames

id = 0  
name = Kintex  
type = company

id = 1  
name = Angola  
type = country  
Episode 0  
site-of = anti-revolution  
movement

id = 2  
name = South Africa  
type = country

id = 3  
entity = movement  
Episode 0  
characteristic = anti-revolution  
location = Angola = 1  
location = South Africa = 2

id = 4  
**event = deliver**  
Episode 0  
**object = supplies**  
**agent = Kintex = 0**  
**recipient = anti-revolution movement**  
**point-in-time = 1978-79**

# Entities and Relations

- Additional rules recognize phrases that modify noun phrases, such as appositives. Some prepositional phrases are also handled here as modifiers of nouns. Emphasis has been on people, places and organizations.

id = 3

**entity = movement**

Episode 0

**characteristic = anti-  
revolution**

**location = Angola**

**location = South Africa**

id = 5

**name = Palawan**

**type = island**

Episode 0

**isa\* = western Philippine  
province**

\* Note the use of “isa” for descriptive phrases, not category hierarchy

# Knowledge Representation

- Extracted information is represented in a frame logic
  - Frames representation in the standard existential conjunctive interpretation in logic
  - Note that our frames are only partially associated with the category hierarchy - not a strongly typed logic
  - Some context is represented: time referents, document source, speaker
  - Related Work
    - Narrative Knowledge Representation Language (NKRL)
      - European NOMOS project – (references: G.P. Zarri)
    - UNO – a natural language logic
      - frames with quantified noun phrases, negation and context
      - set and interval semantics (references: Lucja Iwanska)

# Inference in Frame Logic

- Goal of inference is to do approximate matching between one representation of knowledge (frames) and another
- Inference proof rule uses abductive inference
- Basic Inference: One frame entails another frame if
  - the event or entity values entail
  - each attribute of the first frame is present in the second
  - and each attribute has values that entail (call this value matching)
- Abductive Inference: Use the basic entailment rule except that each entailment has a weighted proof. The attribute values may entail at a probability less than 1, including 0 if an attribute is missing, and these are combined to give a final probability for the result.
  - (Reference: Hobbs et al)

# Axioms

- We use axioms to relate different NL representations of the same knowledge.
  - Example:  
event=?X && agent=?Y && object=?Z  
<= entity=?Y && isa=(nominal(?X,event,agent)) of ?Z
- Our axioms represent the interplay of equivalent or implications of language structure and knowledge representation.
  - For complete text understanding, one needs common sense knowledge, but that is beyond the scope of our work!

# Value Matching

- Several forms of value matching in the process of inference. Note that values are NL phrases.
  - Geographic regional containment from a geographic database
    - entity = ?X && location=China  
    <= entity = ?X && location=Beijing
  - Synonyms from WordNet
  - Category containment from the category hierarchy
  - Adjective/adverb containment
- Uses “world knowledge” sources
- Note that some rules require POS tagging to examine phrases at a finer grain.

# Information Applications: Information Extraction for Specific Domains

- For business competitive intelligence or national security intelligence
- Process documents every day and extract information of interest on particular topics
  - Visualization
  - Search for Patterns of Interest

# Map to Specific Domain

- For applications in a specific domain, events can be “typed” and event roles mapped to more specific terminology appropriate to the application
  - Important for subject matter experts of visualize and search in their own domain vocabulary
- Extend category hierarchy and relations

**event = buy**

**agent = Kintex**

**object = supplies**

**event = buy**

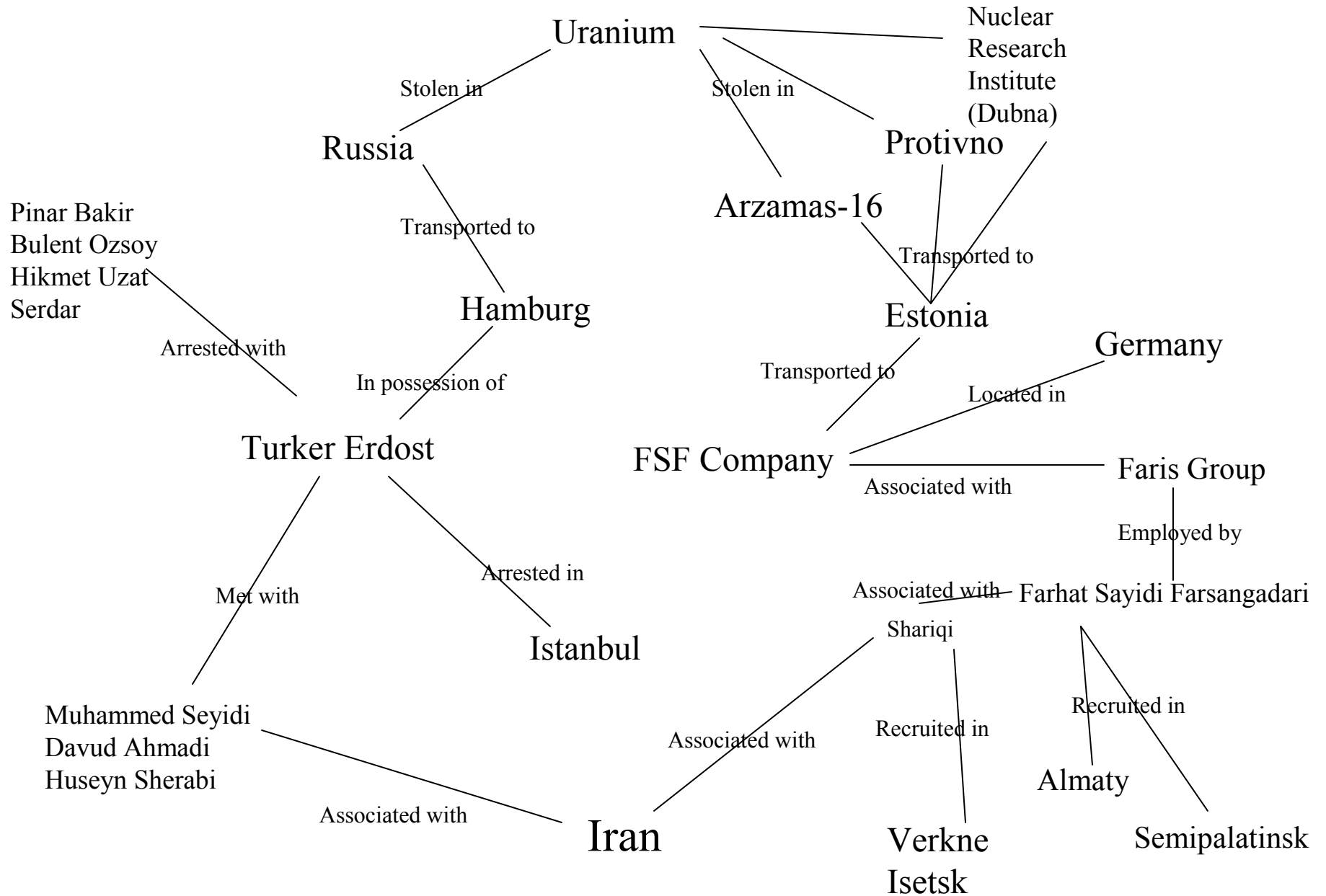
**seller = Kintex**

**goods = supplies**

- Sources of verb roles: FrameNet, VerbNet, Proposition Bank



# Analyst's View of 2 NUCLEAR SMUGGLING Incidents



# Information Applications: Open Domain Question & Answering

- Example Tasks
  - TREC Q&A track (concrete questions)
  - NLP question interface to university course resources
- For large document collections, 2 stage retrieval and processing
  - Retrieval to locate relevant documents
    - L2L (Language to Logic) for standard search engine
  - Document processing on relevant documents to extract information
  - Question processing to produce goal frame representation of desired answer
  - Inference engine to match goal frame to extractions from documents

# Work in Progress

- Better parsing of sentences
  - Identification and relation between clauses
  - Prepositional attachment
- Multi-document resolution of names
- Construction and use of time sequence information

# Challenge Question 1

- Adjusting model to new domains:
  - Obtain specific domain vocabulary for categories and relations
  - Use machine learning (Transformation Based Learning) to learn rules to map generic extractions to specific domain extractions
    - Linguistic analysts mark documents for training
- Adjusting NLP document processing to new text genre
  - Requires work by linguistic analysts to adjust shallow parsing rules for different syntactic patterns

# Challenge Question 2

- Inspect, assess and hand-edit knowledge in the model
  - Inspection of knowledge by visualization and question-answering.
    - We are also working on more extended patterns of events
  - Assessment and hand-editing not currently done. We expect in future to allow domain experts to annotate information from documents as to reliability and correctness.

# Challenge Question 3

- Performance on a large scale trial - efficiency
  - Document processing
    - Time to process 100,000 TREC documents: 24 hours on a 4 processor PC server.
  - Inference
    - As yet unknown
    - Performance is reason to choose frames over First Order Logic – far fewer inference steps, but each step is bigger

# Challenge Question 3

- Performance on a large scale trial – accuracy
  - Evaluation of information extraction
    - Hand evaluation by linguistic analysts
      - Proper name extraction  
precision: 92% recall: 94%
      - Proper name categorization  
precision: 97% recall: 71%
      - Event and entity extraction  
precision: 93% recall: 61%
      - Event attributes  
precision: 69% recall: 48%
      - Entity attributes  
precision: 87% recall: 44%

# Challenge Question 4

- Additional Knowledge or Theory
  - Looking forward to improvements in WordNet or similar word resources
    - Synonymy, hypernymy (superclass), hyponymy (subclass), glosses
  - Theory of confidence levels: assign confidence levels according to document source, how to use inference across documents with varying confidence levels, without resorting to probabilities
  - Theory of context that includes user information and previous queries