

Specializing Evidence Extraction Using Transformation Based Learning

Elizabeth D. Liddy

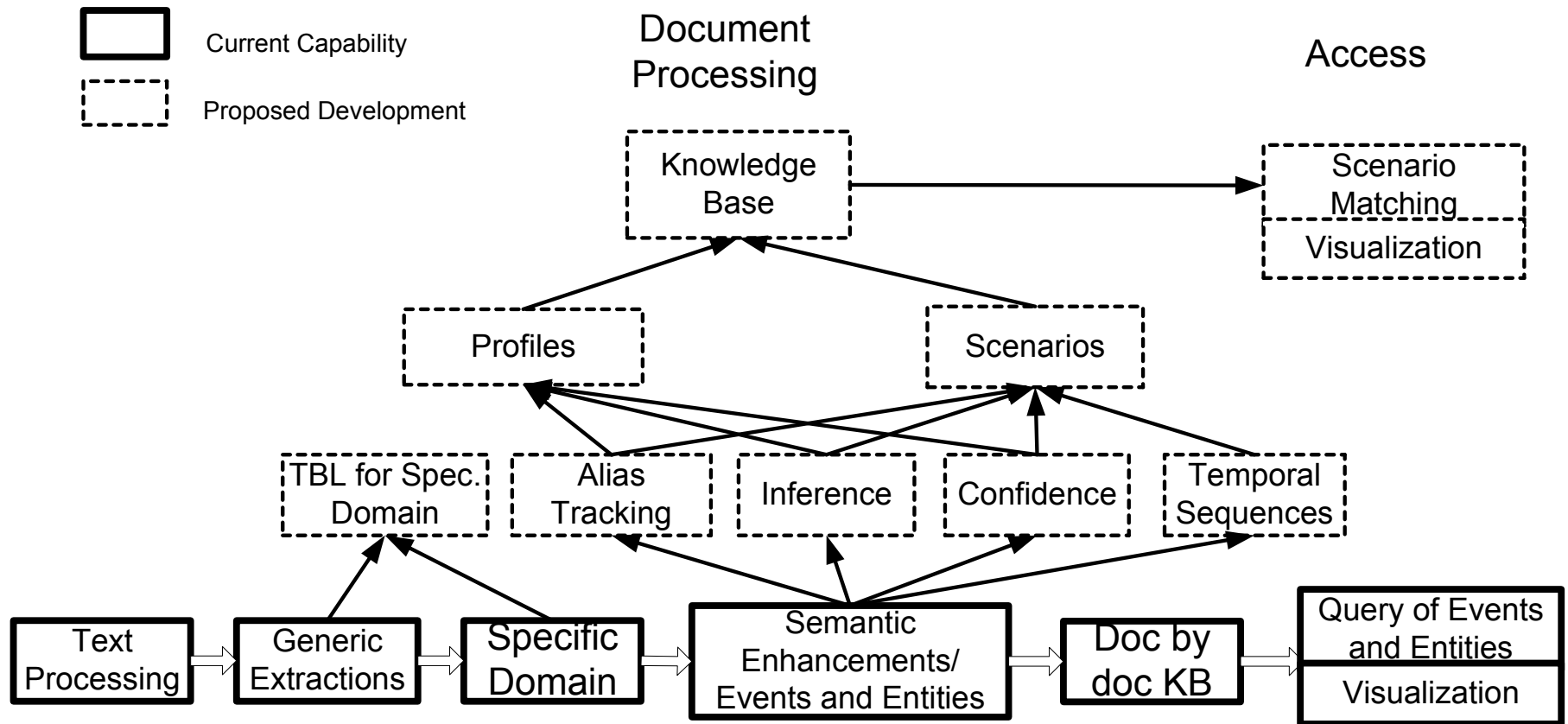
TIDES PI Meeting

July 24-26, 2002

CNLP

Center for Natural Language Processing

CNLP Evidence Extraction Capabilities



CNLP

Center for Natural Language Processing

Generic Event-based Extraction

- Shallow parsing rules analyze surface lexical & syntactic clues to map to semantic roles of verbs
 - Patterned after the case frames of Fillmore
 - First, parse rules for basic sentence structure
 - Generic roles, i.e. *agent, object, location, manner*
- Extract entities, relations, and events into frames
- Final stage is co-reference resolution

Relations Identified

- About 20 from Sowa's conceptual graph theory

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

- Plus, 30 new ones.....and more being added

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

In 1978-79 Kintex delivered supplies to anti-revolution movements in Angola and South Africa.

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

Specialization

- For applications in specific domains, events are “typed” and event roles specialized for the domain
 - Important to characterize precisely and distinctively
 - Extend category hierarchy and relations

| | | |
|---------------------------------|---|--------------------------------|
| event = <i>buy</i> | | event = <i>buy</i> |
| agent = <i>Kintex</i> | → | seller = <i>Kintex</i> |
| object = <i>supplies</i> | → | goods = <i>supplies</i> |

- Source of roles: FrameNet, VerbNet, PropBank

Transformation-Based Learning (TBL)

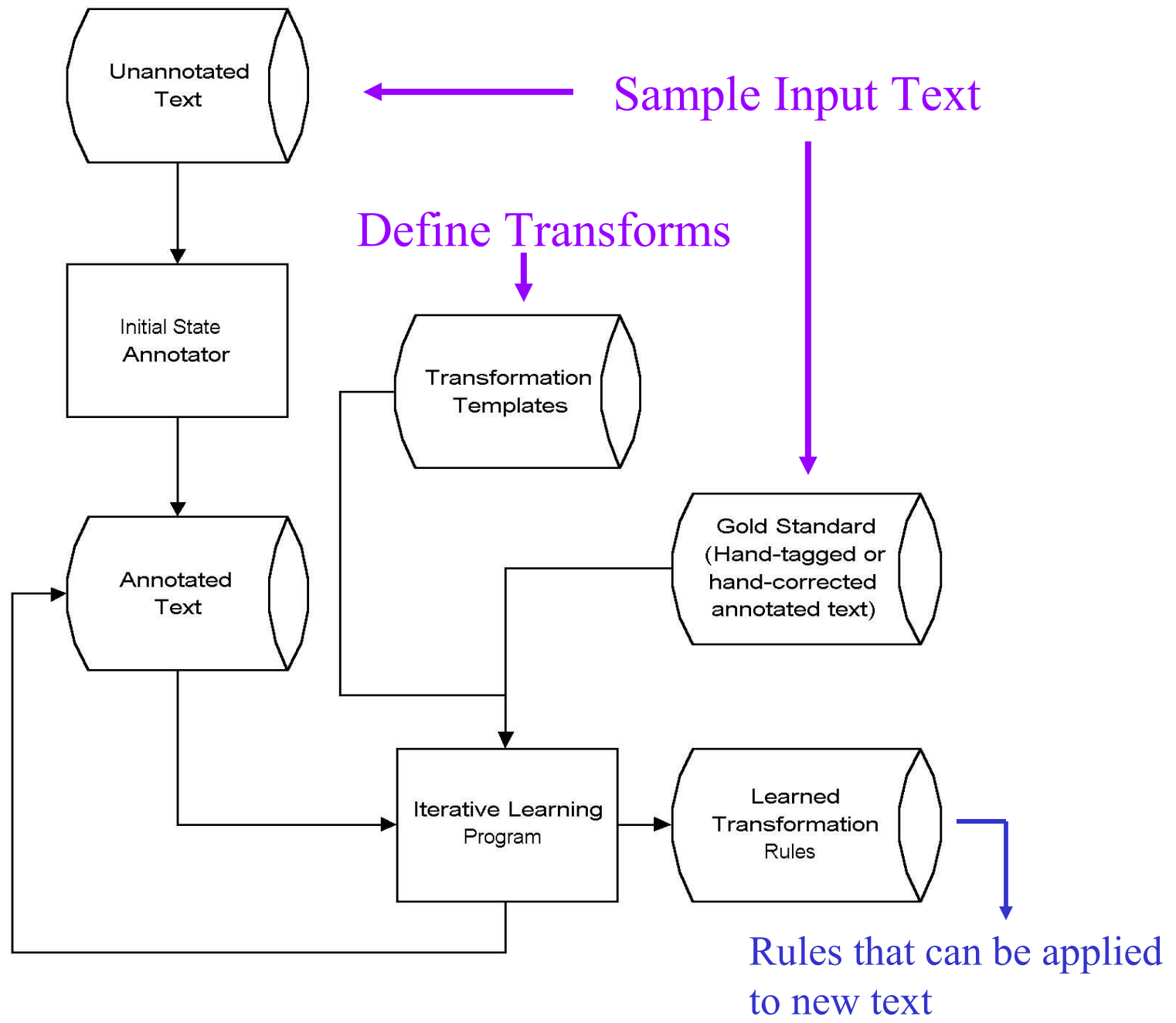
- A transformation-based, error-driven machine-learning approach to specialize to new domains
- Our two step process:
 1. System applies base annotation rules to produce **generic extractions** according to case frame roles
 2. System applies transformation rules to **specialize** the generic roles to a specific domain
- Iterative learning cycle continues, at each stage using the corpus which results from previously learned transformation as the basis for comparison

Transformations:

- Consist of 2 components:
 1. A triggering environment
 - The generic event, relations, values and types
 2. A re-write rule
 - Changes relation from generic to specialized
- Simple example of TBL rule format:
 - if [triggering environment] then [re-write rule]
 - if [event = *kill*, object = ?X, ?X.type = person] then [object-> *victim*]

Learning Algorithm

- At each iteration of learning, the transformation is found whose application results in the best score according to the objective function in use
- That transformation is added to the ordered transformation list
- The training corpus is updated by applying this transformation
- Learning of new transformation rules continues until no transformation can be found whose application results in an improvement in the annotated corpus



Developing Domain Model

- CNLP analysts examined Russian Contract Killing corpus to define domain model
 - Incorporated analysis from Veridian
 - Delimited event types of: **arrest, attempt to kill, deal, detain, disappear, investigation, kidnap, kill, payment, telephone conversation, theft, and warning**
 - Each event type may be lexicalized in text as one of several verbs or nominalizations, for example, type ‘kill’: ***assassination, commit murder, contract murder, kill, liquidate, and murder***
 - Each event has “roles” for most common attributes
 - E.g., arrest = frame-> ‘*authority*’, ‘*arrestee*’ & ‘*charge*’

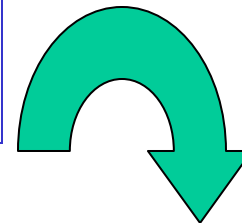
Training Process

- Goal is to learn domain specialization rules with minimal human annotation effort
 - Using a small number of exemplar sentences to bootstrap annotation
 - First training set chosen for main domain concepts
 - Identified by simple ‘event’ counts based on model
 - Bootstrapping rules applied to speed hand-annotation
 - Second training set chosen to elaborate other domain concepts and to “cover” domain by random choices
 - Hand annotation time for 35 documents, approximately 1600 sentences, was 8 hours

Example of Specialization

In March 1997, Petrosyn was arrested again by officers from the directorate, . . .

event = arrest
agent = officers of the directorate
object = Petrosyn
point-in-time = March 1997

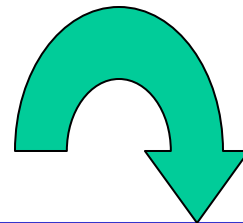


event = arrest
type = arrest
authority = officers of the directorate
arrestee = Petrosyn
point-in-time = March 1997

More Complex Specialization

Ogorodnikov returned fire, and the criminal was forced to flee.

event = return
agent = Ogorodnikov
object = fire



**Rule corrects object
when it is “fire”**

event = return fire
type = attempt to kill
perpetrator = Ogorodnikov
object = null (effect is to delete)

Results

- Not yet agreed-to metric for learning semantic roles
- Following the path of Gildea & Palmer (2002) on reporting predicate argument recognition
- For ‘model-based’ classes of verbs in RCK
 - Verbs and nominalizations of semantic classes of *arrest, attempt to kill, detain, investigation, kill, payment* with > 6 occurrences of that type
 - 70 documents (950 events requiring 2,275 role changes) using 80/20 split with 5-fold cross validation

Precision 88.93%

Recall 67.93%



Preliminary Error Analysis & Fixes

1. Incorrect generic extractions

- But, iterative process of specialization is giving valuable feedback to improve generic rules

2. Lack of type information on common nouns to provide in triggering environment

- Categorization of compositional noun phrases is under development

3. Some event phrases are infrequent in the corpus

- Larger, focused corpus will assist here

TBL Advantages

1. Iteratively *transforms* an initial automatic generic extraction into a more informative, specialized one.
2. Creates a relatively small number of linguistically motivated rules that are understandable by humans.
3. Has been successfully applied to other NLP tasks.
4. Has been shown more powerful than decision trees.
5. Combines statistical & symbolic approaches using transformation statistics to learn symbolic rules.
6. Requires an order of magnitude fewer decisions than estimating the parameters of statistical models.
7. Does not over-train.

Benefits of TBL for EE Specialization

- Rapid deployment of proven extraction capability to new domains.
- Refines the content of extractions to differentiate them in large bodies of processed text.
- Provides more informative, specialized extractions for utilization by Link Discovery & Pattern Learning and other down-stream tasks.