UIMA

From Extraction to Reasoning

Chris Welty
IBM Research

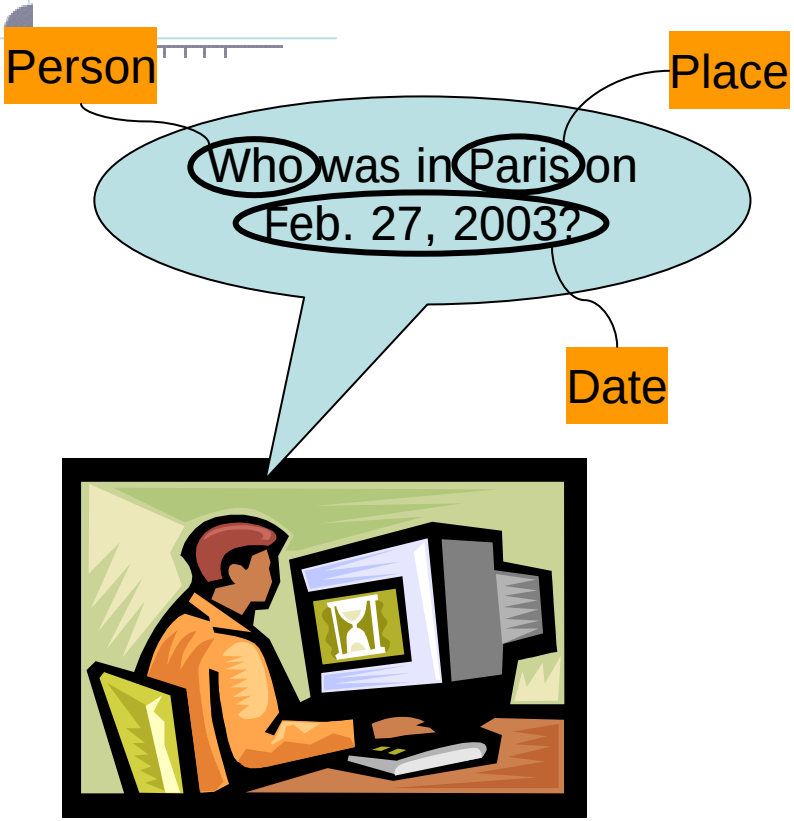
IBM

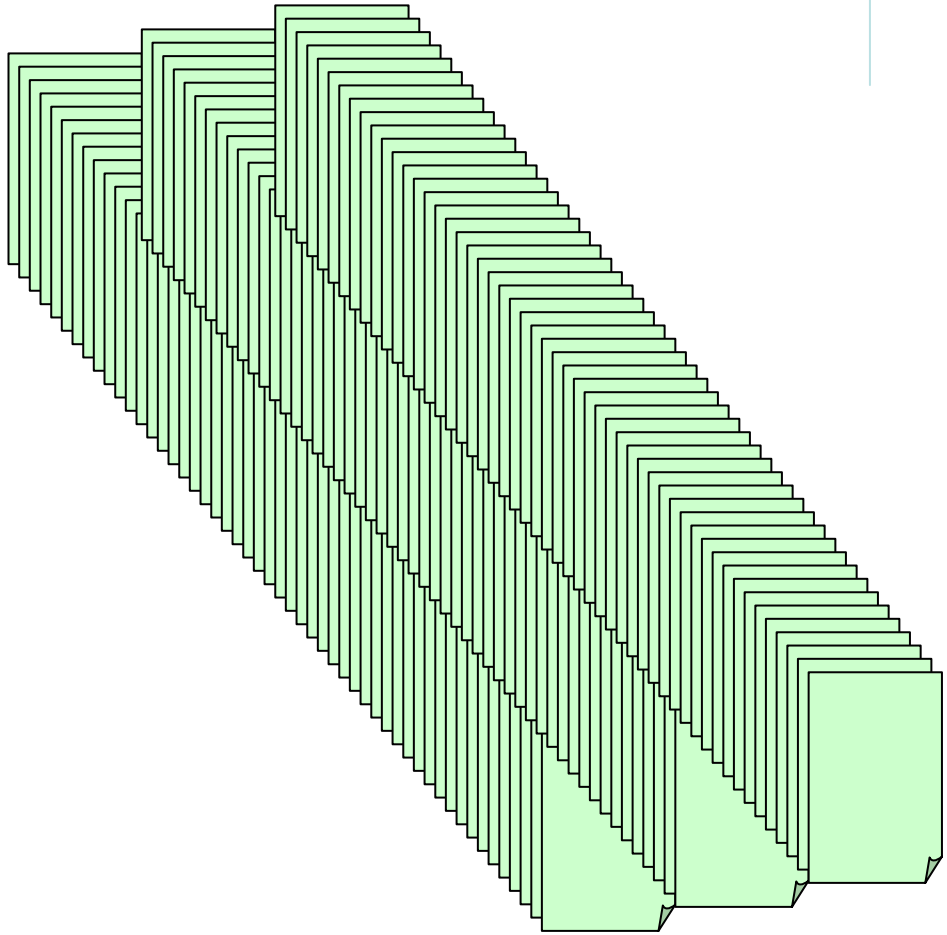
Acknowledgements

- IBM Unstructured Information Management Architecture (UIMA) Project

- David Ferrucci (IBM)
- Bill Murdock (IBM)
- Mary Neff (IBM)
- Richard Fikes (Stanford)
- Selene Makarios (Stanford)
- Rob McCool (Stanford)

Why extraction?





Why reasoning?

Former Iraqi Ambassador to Canada
Hisham Al Shawi, who defected to the UK in
August 1993, was a key figure in launching Iraq's nuclear program. From 1972 to
1974, Al Shawi was chairman of Iraq's Atomic Energy Commission and was Iraq's representative on the board of the IAEA. Al Shawi helped coordinate Iraq's
efforts to train Iraqi students and send them abroad to universities and international research facilities. However, there is no evidence that Al Shawi knew of efforts to use these trainees in a military nuclear programme.
During the week of Feb. 26, 2003, Al Shawi travelled to France and proceeded up the Seine by boat. He spent three days in the latin quarter, and returned on Mar. 1.
The purpose of his trip was not known.
Al Shawi and a fellow defector, former ambassador to Tunisia Hamid al Jabbouri, denied having any detailed knowledge of Iraq's nuclear program or its procurement network.

Why reasoning?

Former Iraqi Ambassador to Canada

Hisham Al Shawi, who defected to the UK in August 1993, was a key figure in launching Iraq's nuclear program. From 1972 to 1974, Al Shawi was chairman of Iraq's Atomic Energy Commission and was Iraq's representative on the board of the IAEA. Al Shawi helped coordinate Iraq's efforts to train Iraqi students and send them abroad to universities and international research facilities. However, there is no evidence that Al Shawi knew of efforts to use these trainees in a military nuclear programme.

During the week of Feb. 26, 2003, Al Shawi travelled to France and proceeded up the Seine by boat. He spent three days in the latin quarter, and returned on Mar. 1.

The purpose of his trip was not known.

Al Shawi and a fellow defector, former ambassador to Tunisia Hamid al Jabbouri, denied having any detailed knowledge of Iraq's nuclear program or its procurement network.

What is needed?

- Reasoning
  - Temporal containment
  - Spatial containment
- Knowledge
  - Geography
  - Basic Ontology
- Coreference
  - Multiple mentions of same *entities, relations*
  - Find time that something happened
- Extraction
  - Recognize people, places
  - Recognize times, relations

Former Iraqi Ambassador to Canada
Hisham Al Shawi, who defected to the UK in August 1993, was a key figure in launching Iraq's nuclear program. From 1972 to
1974, Al Shawi was chairman of Iraq's Atomic Energy Commission and was Iraq's representative on the board of the IAEA. Al Shawi helped coordinate Iraq's efforts to train Iraqi students and send them abroad to universities and international research facilities. However, there is no evidence that Al Shawi knew of efforts to use these trainees in a military nuclear programme.
During the week of Feb. 26, 2003, Al Shawi travelled to France and proceeded up the Seine by boat. He spent three days in the latin quarter, and returned on Mar. 1.
The purpose of his trip was not known.
Al Shawi and a fellow defector, former ambassador to Tunisia Hamid al Jabbouri, denied having any detailed knowledge of Iraq's nuclear program or its procurement network.

Real World Reasoning?

- The acquisition bottleneck
- Reasoning is *hard*
  - Complexity
    NP < Exp < Non-Elementary < Undecidable
  - More than 1,000,000 RDF triples
  - Can't hold results of 1000 documents in memory
- Reasoning is *sound*
  - Not tolerant of errors
    <Person>Bush International</Person>
  - Not just precision
    Iraqi press agency says <event>the war is ended</event>
- Reasoning can be *inscrutable*
  - If I have no friends then all my friends are doctors

Text Analysis

How much reasoning when

Bounded reasoning

Explanation

General Problem

- *Given*
  - an ontology in OWL
  - A background knowledge base in RDF
  - Inference procedures
  - A collection of existing IE components
- Use the results of IE to populate the KB
  - Map IE semantics to KB semantics
  - Map extracted entities to possibly existing KB instances

Starting Point

Simple Token and Sentence Annotator

Annotates tokens and sentences.

Ace Annotator

A statistical entity and relation annotator that performs within-document coreference resolution.

KS Relation Detector Aggregate

Aggregate that includes KS Relation Detector and TAE's that provide its inputs

XsgParsing Annotator

Performs deep parse using slot grammar parsers

PhraseFinder Annotator

Creates annotations for phrases to be passed to ESG. Phrase sources are WordNet and pre-annotated Resporator phrases.

EAnnotator

IBM EAnnotator - A statistical entity, relation, and event annotator.

Cross Annotator Coreference

Resolves correference disputes across annotators

KS Relation Annotator

Knowledge Structures Relation Annotator (pattern-based relation detector)

Simple Phone Call Relation Annotator

Finds mentions of phone calls in text and annotates them as MadePhoneCallTo relations.

Simple Phone Number Annotator

Finds phone numbers in text

Starting Point

UIMA

KS Relation Detector Aggregate

Aggregate that includes KS Relation Detector and TAE's that provide

Ace Annotator

tistical entity and
on annotator that
rms within-
ment corefer

GlossOnt

Finds taxonomic and other definitional relations from

Simple Token a
Sentence Anno

A
se

HoldsDuring

Extracts Relations between TimeEx3 entities and relations.

CrossDocumentCoreference

Merges coreference chains across documents.

Nominator

Finds prope
and other clues

tations for
passed to ESG.
ordNet

E

IBM EAnnotator - A
statistical entity,
relation, and event
ann

KDDAnnotator

A statistical entity
relation annotator
within-
coreference
.

TFSTOnBoard

Finds onBoard relations between people and vehicles.

TFSTVehicle

Finds vehicles.

entities.

TFSTAddr

Finds addresses and extracts the subplace relation.

Simp
Annc

Finds mentions of phone
calls in text and annotates
them as MadePhoneCallT
relations.

eference

TFSTTime

Extracts TimeE
entities.

nnotator

e Structures
nnotator (pattern-

JResporator

Extracts over 80 classes of entities.

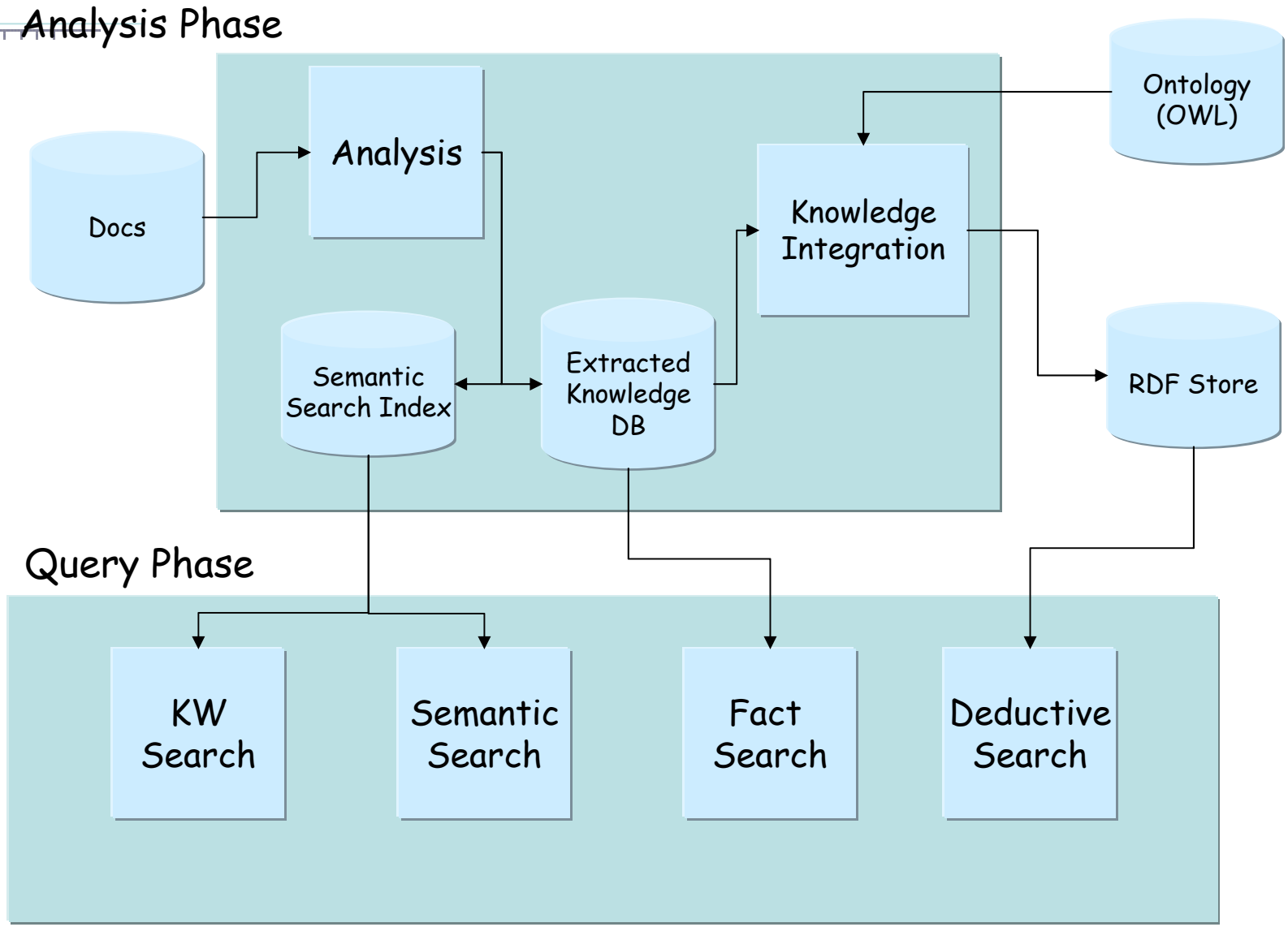
Finds a number of standard named entities.

Finds phone numbers in text

Issues in the combination

- Components have *overlapping* semantics
  - Common type system, but…
  - Different meanings, precision, recall
  - ACE, Resporator
- Multiple annotations on a single span
  - Disagree 20% of the time
- Multiple overlapping co-reference chains
  - U.S. subplace of Russia

Knowledge Integration

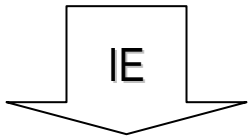


Semantic Integration Goals

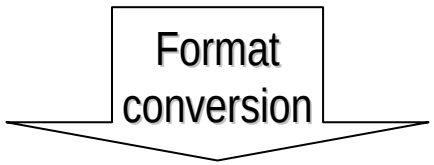
- Process results of IE into a form *suitable for reasoning*
  - i.e. by advanced reasoning components (time, space, etc.)
- Map linguistic structures into knowledge-base
  - Different ontologies - Moving targets
  - Different semantics - Declarative
- Explore utility of reasoning
  - Clean up the IE results using ontology semantics
  - Improve precision and recall
  - Propose candidate contexts
  - Experiment with different kinds of reasoning
- Scale the results along some dimension of “massive”
- Evaluate the quality of the results

Semantic Integration

“13 delegates from Turkey arrived today.”



“13 delegates from <country>Turkey</country> arrived today.”



<OWL:country rdf:ID=“Turkey” />

Easy!!!

IE ↔ KR

- Entities vs. Mentions
- Relations
- Precision
- Brittleness
- Explainability
- Recall
- Scale

IE ↔ KR

- Entities vs. Mentions
- Relations
- Precision
- Brittleness
- Explainability
- Recall
- Scale

IE ↔ KR

- Entities vs. Mentions
- Relations
- Precision
- Brittleness
- Explainability
- Recall
- Scale

IE ↔ KR

- Entities vs. Mentions
- Relations
- Precision
- Brittleness
- Explainability
- Recall
- Scale

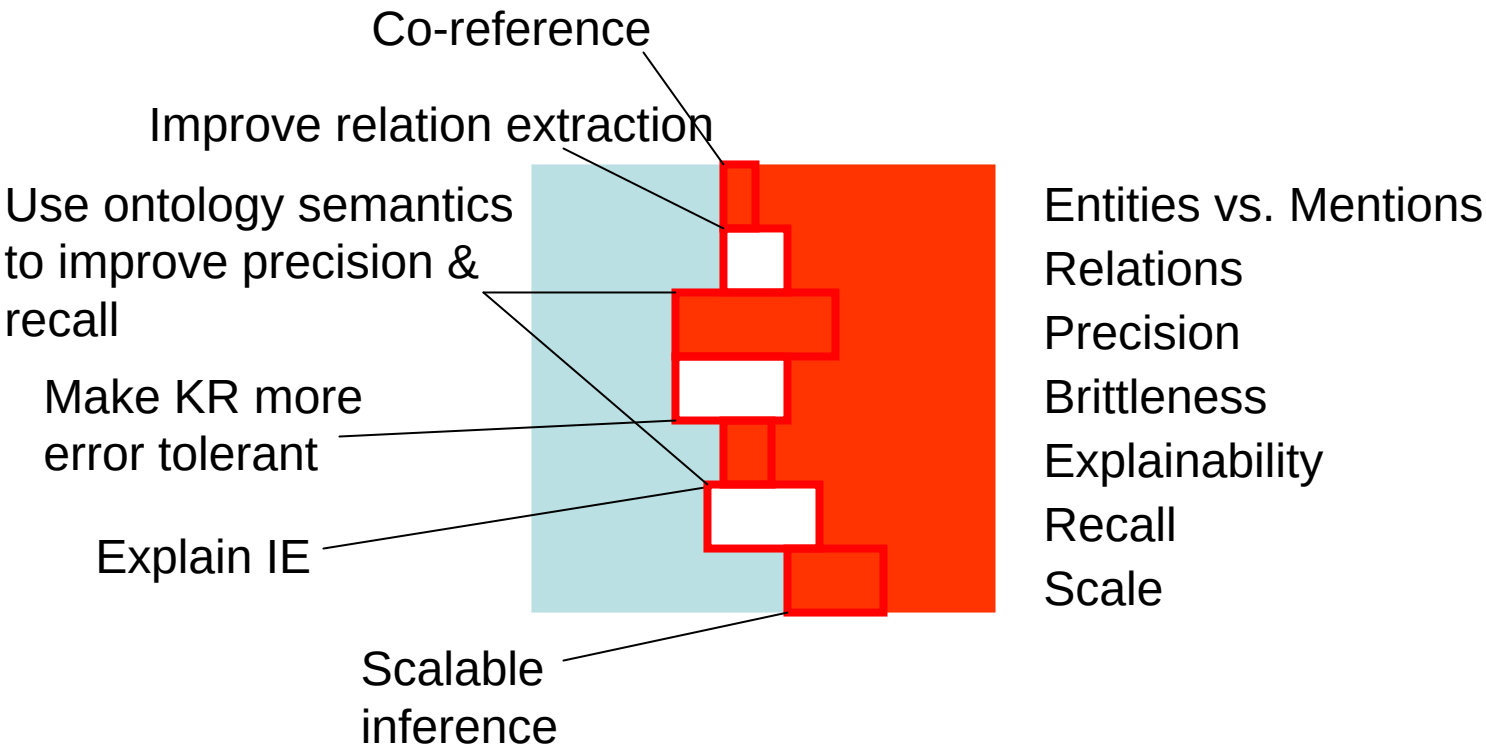
IE ↔ KR

- Entities vs. Mentions
- Relations
- Precision
- Brittleness
- Explainability
- Recall
- Scale

IE ↔ KR

- Entities vs. Mentions
- Relations
- Precision
- Brittleness
- Explainability
- Recall
- Scale

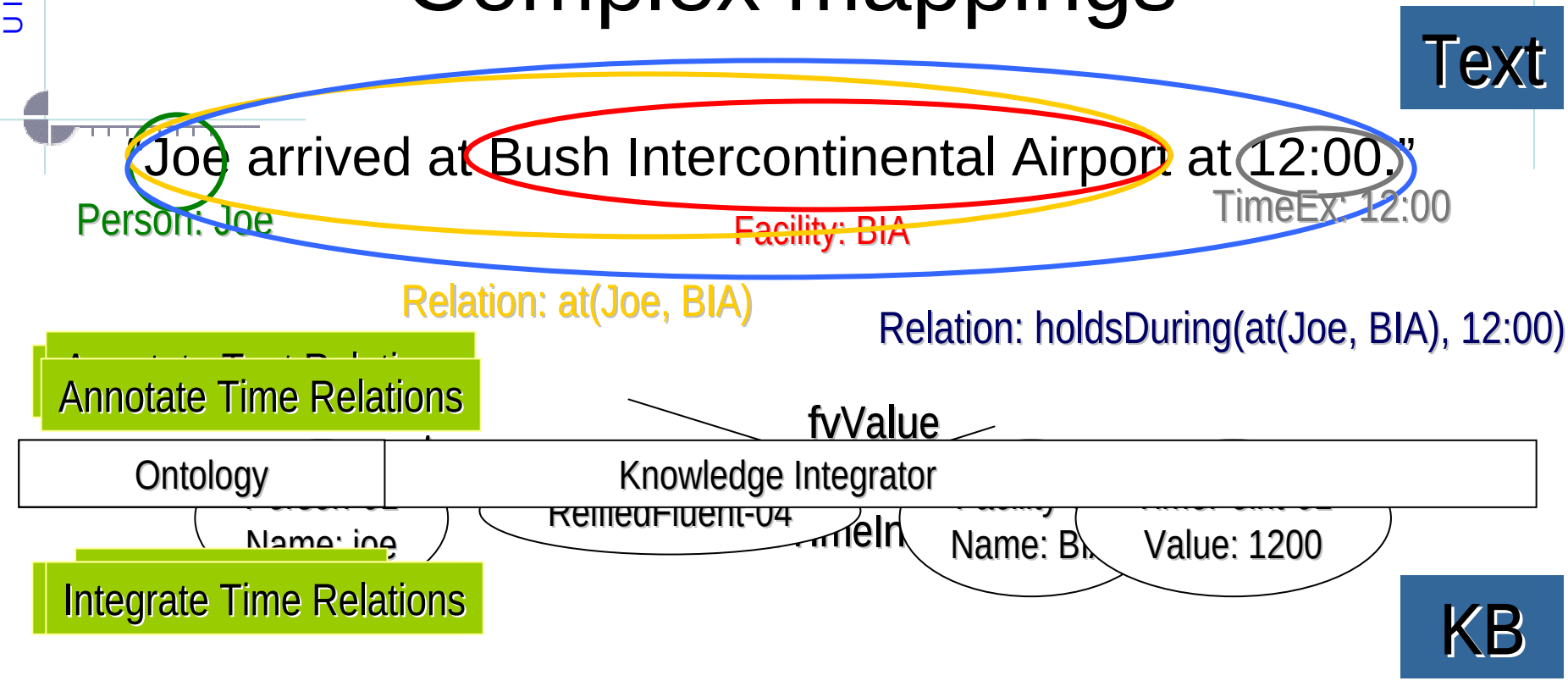
IE ↔ KR


Co-reference
Improve relation extraction
Use ontology semantics to improve precision & recall
Make KR more error tolerant
Explain IE
Scalable inference
Entities vs. Mentions
Relations
Precision
Brittleness
Explainability
Recall
Scale

Mapping from Extraction to Knowledge

- Type-Class Mapping
  - Person → Person
  - Country → Location & Political Entity
- Entity Mapping
  - Person(Abduhl Ramazi) → kani:person-101
- Relations
  - IE:At(Person,Place) → KANI:At
  - IE:At(Place,Place) → KANI:subPlace
- Complex Mappings
  - HoldsDuring(At(Person, Place), TimeInterval) →
    At(Person, *fv*) & fvValue(*fv*, Place) & fvTimeInterval(*fv*, TimeInterval)
  - Uses(BioTerrorism, Diseases) → …

Complex mappings



Broader Contributions

- Reuse&Adapt existing work
  - Chimaera, OntoMerge, Prompt
- Develop catalog of integration inferences
  - Extension to Prompt
  - Moving towards semi-automation for linguistic ontologies
- Mapping TimeML ↔ DAML-Time
- Expose deeper semantic requirements
- Improving P&R of IE

Catalog of Ontology Merging Operations

- Simple mappings
  - $Class_1 \rightarrow$ Existing $Class_2$
  - $Class_1 \rightarrow$ New $Class_2$
  - $Class_1$+$Class_2$ Subclass new $Class_2$
  - …
- Complex mappings
  - $Class_1 \rightarrow Set_2$ of Classes
  - Set_1 of Classes $\rightarrow Class_2$
  - …
- Language Expressivity
  - DL-expressible [Halevy] [Calvanese & DeGiacamo]
  - Function-free FOL expressible [McDermott & Smith]
  - FOL expressible [Chalupsky & MacGregor]
  - Non-FOL

Mapping TimeML to OWL-Time

- TimeML
  - Markup language with linguistic-based semantics
  - time expressions (Timex)
  - Events
  - Links (before, after, …)
- DAML-Time
  - Ontology-based specification of time points and intervals
  - Based on Allen calculus
  - No events
- High level correspondence [Pustejovsky&Hobbs]
- Generate complete OWL-Time RDF for TimeBank 1.1 corpus

Deeper Semantic Requirements

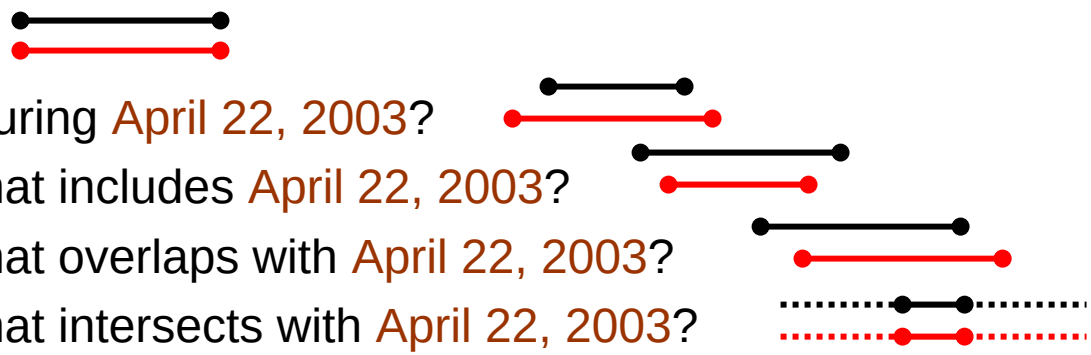
- IE focused on surface semantics
- Surface semantics appear obvious
  - Person(Chris)
  - onBoard(Chris, train)
- Requirement for reasoning exposes problems

"Chris was in his office on April 22, 2003."

Holds(in(Chris, office), *t1*)

What is *t1:*

- April 22, 2003?
- A time interval during April 22, 2003?
- A time interval that includes April 22, 2003?
- A time interval that overlaps with April 22, 2003?
- A time interval that intersects with April 22, 2003?



Using Semantics to boost precision/recall

- "A man was arrested, his name was given as Chris"

Co-reference chain

Relation: nameOf(his, Chris)

Cannot find a link from "his" to "Chris" – the relation is not lexical, it's semantic

During integration, the semantics of *name* relation are processed and Chris assigned as name

- "He arrived at Bush Intercontinental at 12:00"

Relation: at(he, Bush Intercontinental)

Type: Person, Facility

Entity extraction tags "Bush Intercontinental" with Person and Facility

Relation extraction finds at(person, Bush Intercontinental)

Semantics of *at* relation requires range be a facility or place

Semantics of Person and Facility are of disjoint classes

Person annotation thrown away

Knowledge for Improving Precision

- Range/Domain Constraints
  - At $\mapsto$ Person $\times$ Place
- Background Knowledge
  - "Kisumu is the AIDS capital of Kenya."
  - Capital of Kenya is Nairobi
- Logical Consistency
  - Vehicle in two places at the same time
- Temporal Consistency
  - A time point both *before* and *after* the same event

Knowledge for Improving Recall

- Semantic Relations
  - Has-alias
- Containment axioms
  - Boston in Massachusetts in US
  - April 23, 2003 in April, 2003, in 2003
- Classification axioms
  - Author of newspaper article a journalist
- Spatial axioms
  - Passenger on vehicle located where vehicle is located
- Temporal ordering axioms
  - Transitivity of *before, after*
  - Full Allen calculus
- Requires different notion of evaluation

Trials and Tribulations

- Reasoning helps information processing
  - Increase in recall through deductive inference
  - Increase in precision through constraint processing
- General-purpose reasoning algorithms are complex
  - OWL-lite is Co-NP
  - OWL-DL is EXPTIME
  - OWL, General First-order reasoning is undecidable
  - Allen Calculus undecidable
- In *practice* this means reasoning must be bounded
  - A *tradeoff* between scale and effectiveness of reasoning
- Research agenda – explore this tradeoff
  - Techniques for bounding data (e.g. partitioning)
  - Techniques for “hiding” data
  - Incrementally apply more sophisticated reasoning (staged reasoning)
  - Measure KB and Ontology *complexity* (what is massive?)

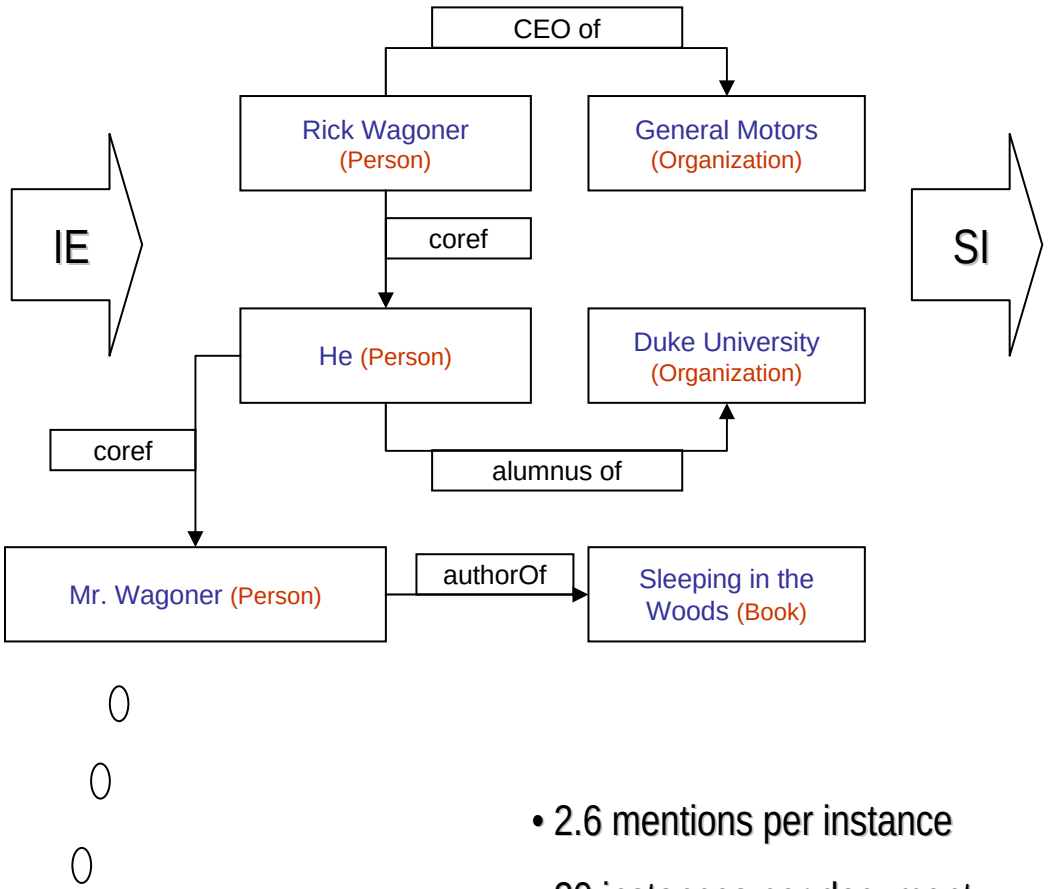
Hiding Data - Motivation

- IE generates a lot of data
- Need to reduce the burden on reasoning-based processes
  - RDF-based reasoners limited to ~500K nodes
- Can we reduce the amount of data?
- What if it is needed later?

Data Volume Analysis

… Rick Wagoner is CEO of General Motors. He is an alumnus of Duke University…

… Mr. Wagoner, the author of *Sleeping in the Woods*…



P1 type Person
P2 type Person
O1 type Org
O2 type Org
P1 ceoOf O1
P2 alumnusOf O2
P1 sameAs P2
P3 type Person
P3 sameAs P2
P3 authorOf O3
O3 type Book

Two triples for every co-reference

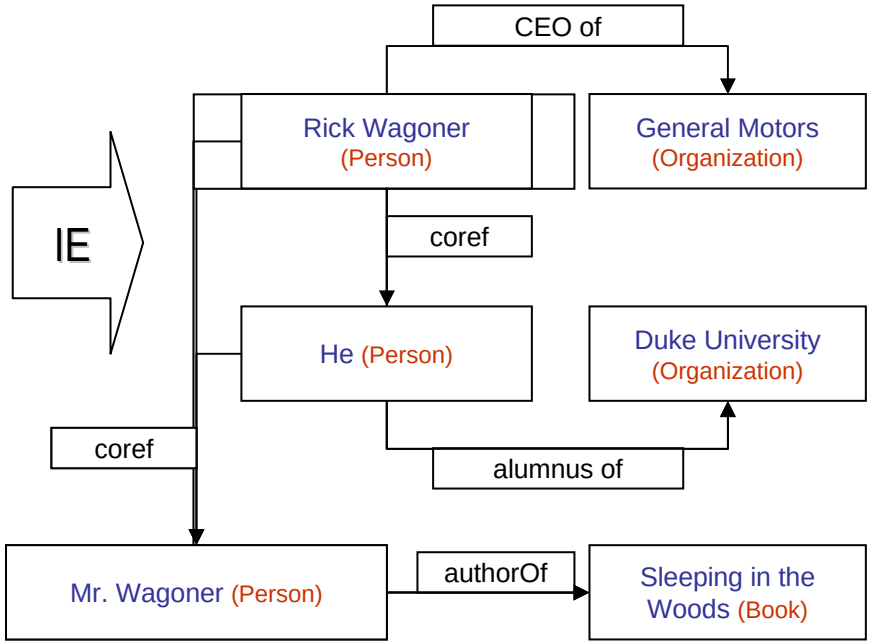
- 2.6 mentions per instance
- 20 instances per document
- 136 triples per document
- *64 triples per document for representing coreference.*

IBM

Hiding Data

… Rick Wagoner is CEO of General Motors. He is an alumnus of Duke University…

… Mr. Wagoner, the author of *Sleeping in the Woods*…



P1 type Person

O1 type Org
O2 type Org
P1 ceoOf O1
P1 alumnusOf O2

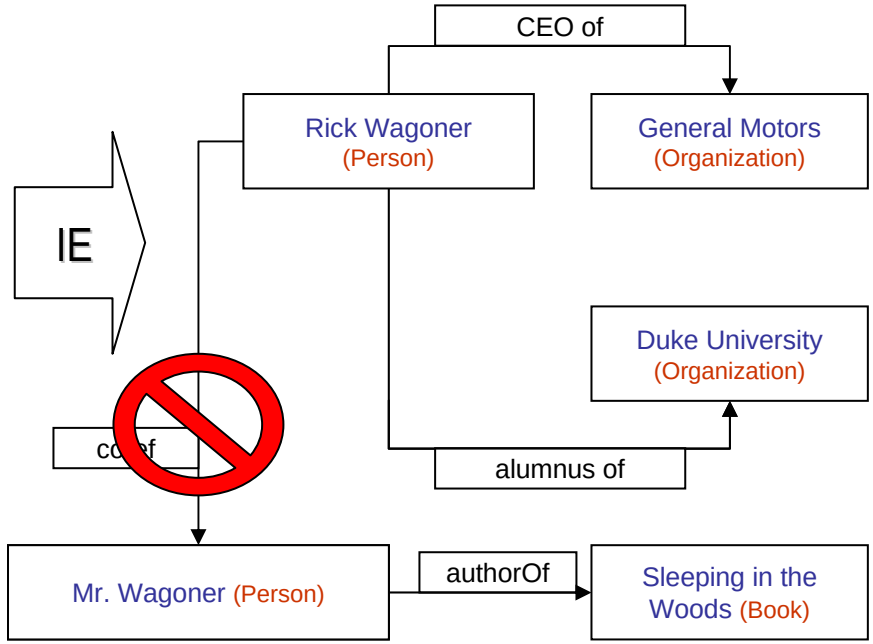
P1 authorOf O3
O3 type Book

- 47% reduction in the number of RDF triples
- 62% reduction in the number of RDF nodes
- But what if it's wrong…

Hiding Data

… Rick Wagoner is CEO of General Motors. He is an alumnus of Duke University…

… Mr. Wagoner, the author of *Sleeping in the Woods*…



SI

P1 type Person

O1 type Org
O2 type Org
P1 ceoOf O1
P1 alumnusOf O2

P3 type Person
P3 sameAs P1
P3 authorOf O3
O3 type Book

- 47% reduction in the number of RDF triples
- 62% reduction in the number of RDF nodes
- But what if it's wrong…
- "Hide" IE inferences in provenance
- Expose when problem arises

Key Challenges

- How much reasoning when
- Artifacts of established IE
  - Imprecision
  - Co-reference
  - Multiple annotations
- Confidence, Trustworthiness
- Maintaining provenance through mapping
- Limits of Automated Reasoning for integration
  - Limits of description logic reasoning (OWL/DL) [Lenzerini, 2002], [Halevy, 2001]
  - Limits of first order reasoning (FOL) [McDermott, et al, 2002]
- Axiomatic semantics of information extraction
  - Further clarify semantics [ACE, KDD]
  - Boost precision
- Adaptability
- Evaluation

How much reasoning when?

- Incremental value of increased processing
- Many dimensions of complexity
  - Representational
    - Worst-case complexity
    - Special-purpose reasoning
    - Optimizations
  - Instance level
    - Number of instances, relations
    - Connectedness
    - Precision
  - Ontology level
    - Number of classes, properties
    - Number of axioms, constraints