

Finding Questions to Your Answers

Ozgur Yilmazel, Grant Ingersoll, Elizabeth Liddy
Center for Natural Language Processing
Syracuse University
Syracuse, New York, USA

April 15, 2007

Typical vs. Atypical Scenario

- Typical - automated QA systems find answers to clients' new questions from a corpus of reports, websites, or newsfeeds
 - Types of questions
 - Factoid
 - List
 - Definition
 - How / Why
- Atypical – Matches new reports / postings / analyses as they are produced to pre-existing questions
 - Types of questions
 - Ongoing interests vs. spur of the moment information needs

Organizational Settings

- Knowledge-intensive organizations
 - Intelligence Community
 - Market analysis companies
- Rely on a supply chain of information
 - Connects Consumers of information with the Producers
 - Begins with RFIs / INs from the consumers that are typically quite general in nature
 - Eventually result in large, dynamic collections of reports being produced

Supply Chain of Information

- RFI's / INs
 - Complex – consisting of sets of related questions – not a single question
 - Can be thought of as 'containers' of 'Question Sets'
 - Groups of logically related questions

Scenario Development – Information Needs

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<InformationNeed>
  <Descriptor>ABC-2006-103</Descriptor>
  <Title>Global Warming</Title>
  <Background> Up to date information is needed to continually assess
the state, effects and impact of global warming. </Background>
  <QuestionSet>
    <identifier>A</identifier>
    <title>Metrics on Global Warming</title>
    <question id="1">What metrics, using measures such as
temperature, have been collected to study Global Warming?
</question>
    <question id="2">What evidence supports the increase of
Global Warming?
</question>
    <question id="3">What organizations are researching and
publishing papers which include metrics and data on Global
Warming?
</question>
    <question id="4">Which countries and organizations are
funding research on global warming?
</question>
    <question id="5">How much is being invested to research
Global Warming?
</question>
  </QuestionSet>
  <QuestionSet>
    <identifier>B</identifier>
    <title>Photographic and Imagery on Global Warming
</title>
    <question id="1"> Are there photographs or satellite images
available that provide evidence of Global Warming?
</question>
    <question id="2"> What organizations are using or providing
imagery on Global Warming?
</question>
    <question id="3"> What countries or organizations are involved
in photographing coastlines in North and South America?
</question>
  </QuestionSet>
  <QuestionSet>
    <identifier>C</identifier>
    <title>Impact of Global Warming</title>
    <question id="1">What evidence, if any, supports the
connection between Global Warming and the global or regional
economy?
</question>
    <question id="2">
What evidence, if any, supports the connection between Global
Warming and the fishing industry?
</question>
  </QuestionSet>
</InformationNeed>
```

Scenario Development: Sample Questions

<question>

- 1. What metrics, using measures such as temperature, have been collected to study Global Warming?**
- 2. What evidence supports the increase of Global Warming?**
- 3. What organizations are researching and publishing papers which include metrics and data on Global Warming?**
- 4. How much is being invested to research Global Warming?**

</question>

Producer Side

- High-paid experts in a specific topic / industry
 - Able to anticipate what the questions are / might be
 - Organization must ensure their insights are made available to every Consumer who might benefit
 - Consumers may not yet realize need
- Tasks to satisfy current & future INs
 - Research
 - Analysis
 - Reporting
- End products are large, dynamic reports
 - Can provide answers to multiple INs of multiple Consumers
 - May be produced asynchronously from INs, so key task is to ensure that Consumers with a standing IN for this information receive it

Problem / Motivation

- Intelligence Community or Market Analysis desire the capability to easily
 - Retrieve answers to questions
 - Determine who is asking what questions
 - Match intelligence products / reports with Information Needs

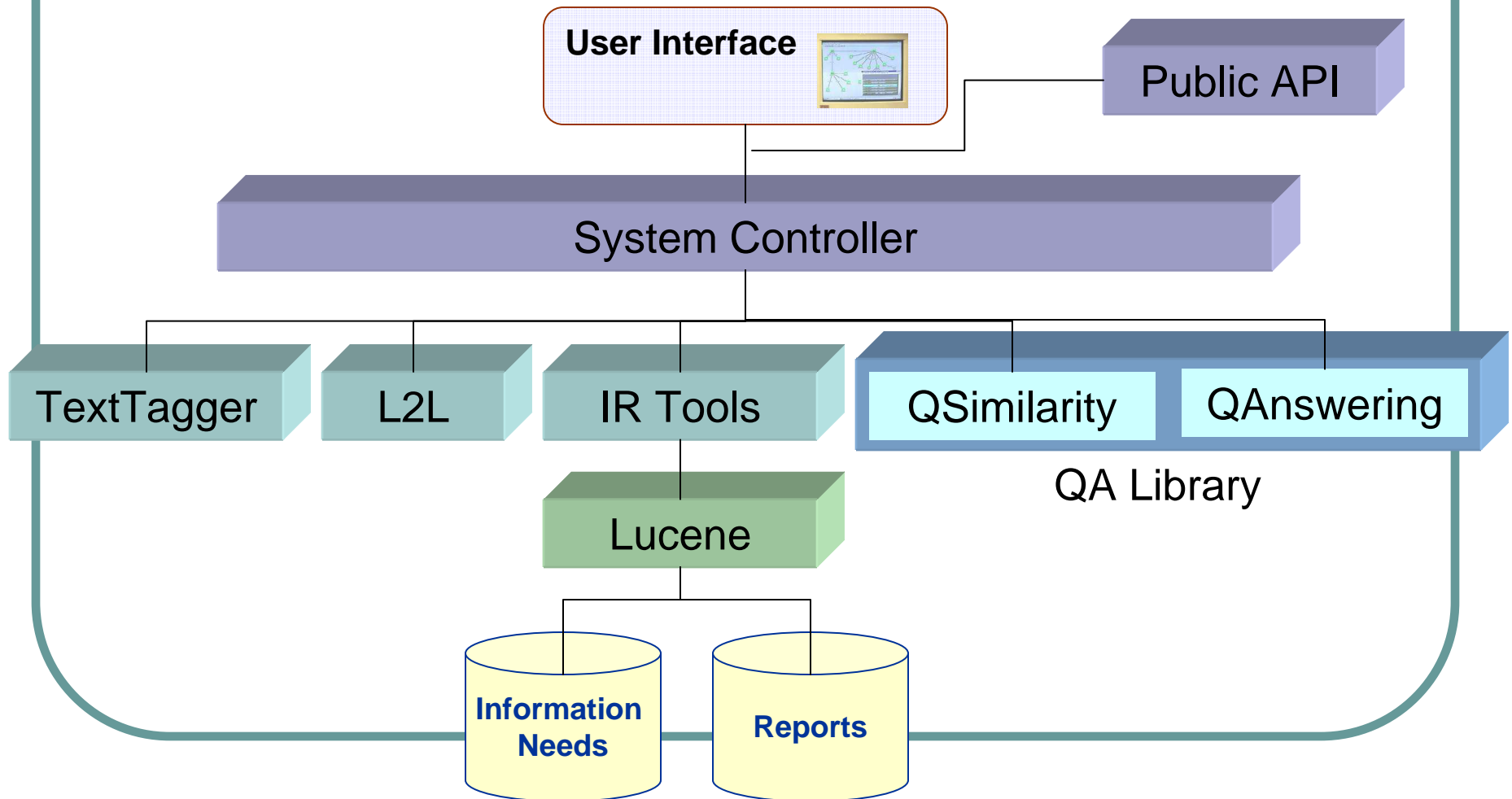
Goals

- **Build Prototype System to:**
 - Conduct traditional search and Question Answering over documents and questions
 - Identify and rank question sets that are answered by new documents
 - Identify similar questions across question sets to help facilitate collaboration among interested parties

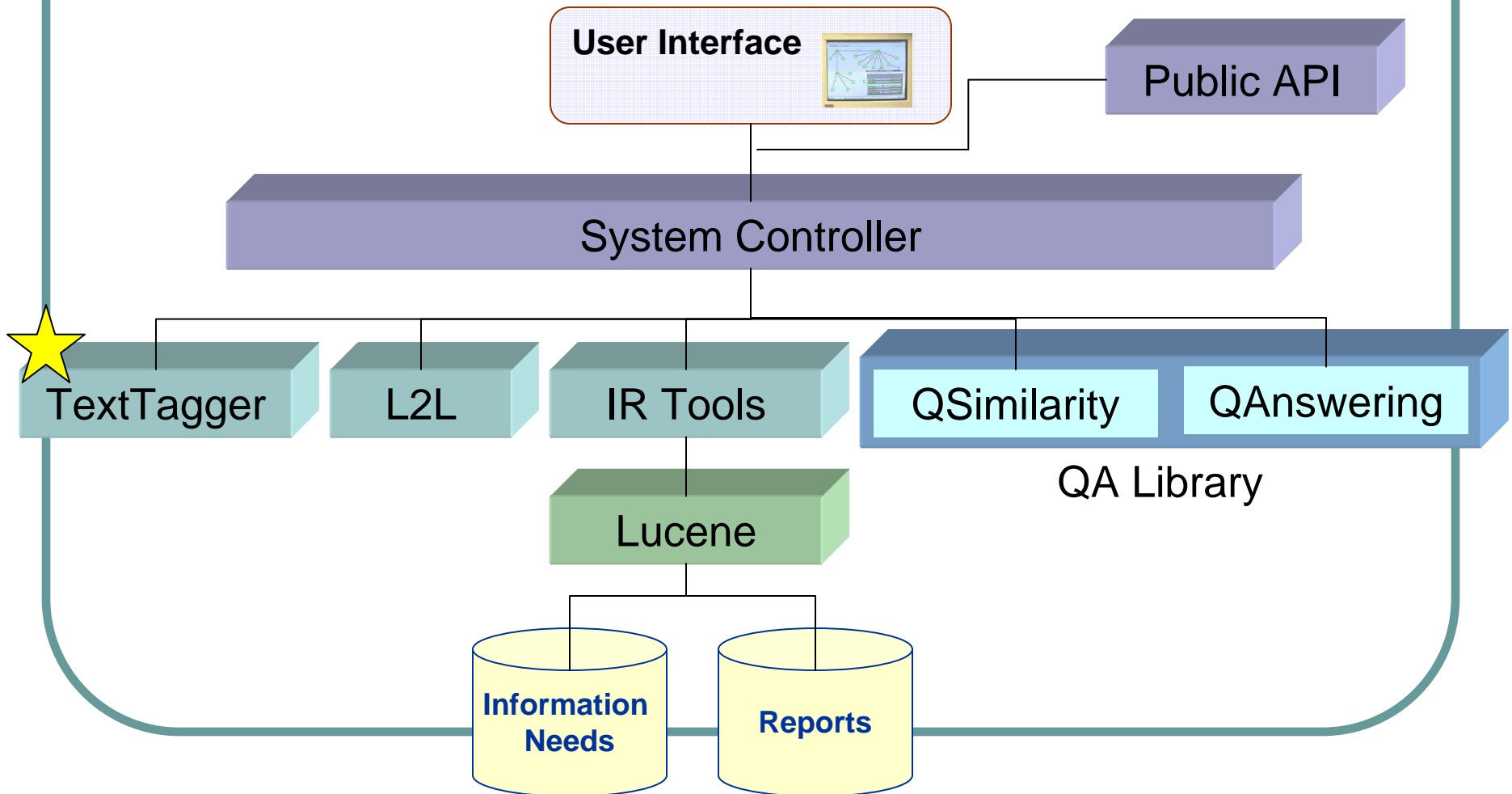
System Design

- Web-based application provides:
 - Traditional search of INs and QA on Reports
 - Match Incoming Reports (answers) to questions that can be answered by report
 - Browse Reports, Information Needs
 - Display contextual information about where questions are answered in document

Q&A System Design



Q&A System Design



TextTagger

- Rule based information extraction system developed at the Center for Natural Language Processing
- Analyzes unstructured text for lexical, syntactic and semantic information
- Uses a sequence of steps called phases

TextTagger Phases

- Tokenization
- Sentence Detection
- Part-of-Speech Tagging
- Stemming
- Non-compositional Identification
- Phrase Bracketing
 - Temporal Concepts
 - Numeric Concepts
 - Named Entity Phrases
 - Common Nouns
- Entity Categorization
- Event and Relation Extraction



TextTagger Output

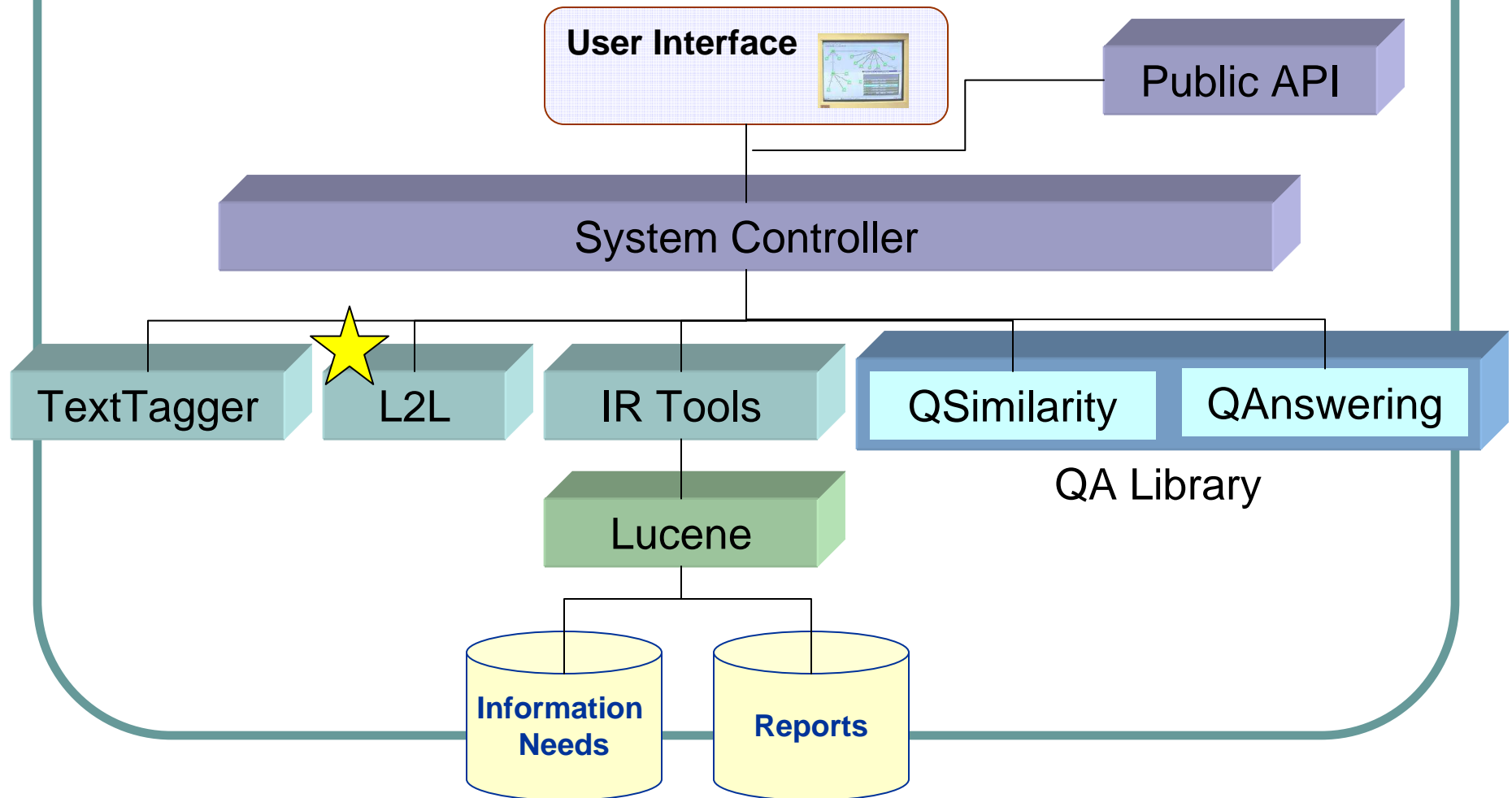
Central Capital Corp said it planned a three-for-two split of its common and class A subordinate voting shares, subject to shareholder approval at the April 23 annual meeting

```
<S> <NP cat="co" sf="Central|NP Capital|NP  
Corp|NP"> Central|NP Capital|NP Corp|NP  
</NP> say|VBD it|PRP plan|VBD a|DT
```

In addition to individual words

- part-of-speech
- entity categories
- co-references

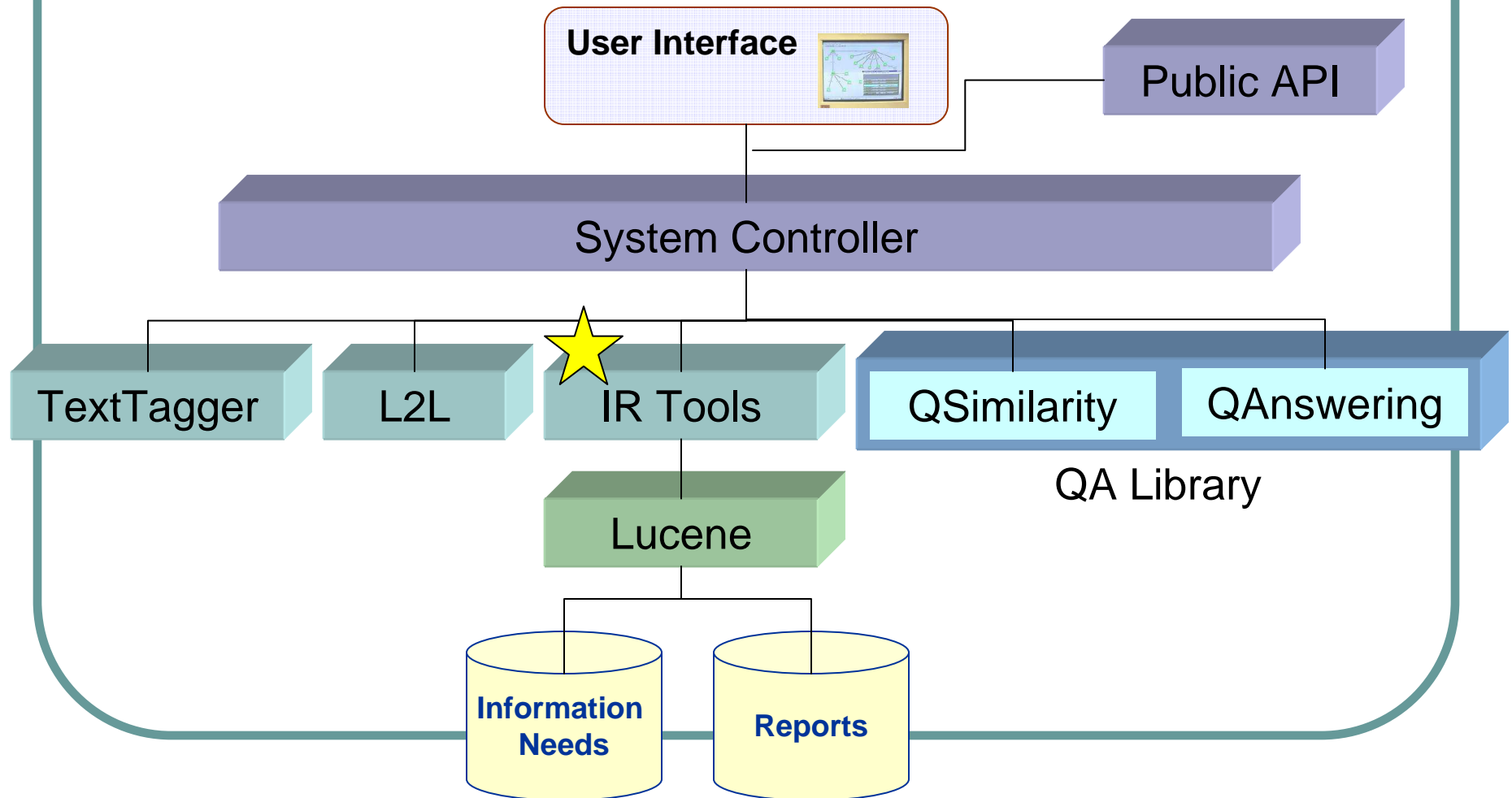
Q&A System Design



Language-to-Logic

- Identifies important features of a natural language question
 - Type of the answer expected
 - Important keywords and their synonyms
 - Focus of the question
 - Relative keyword importance (weighting)
 - Lexical clues for finding answers
 - Spelling variations

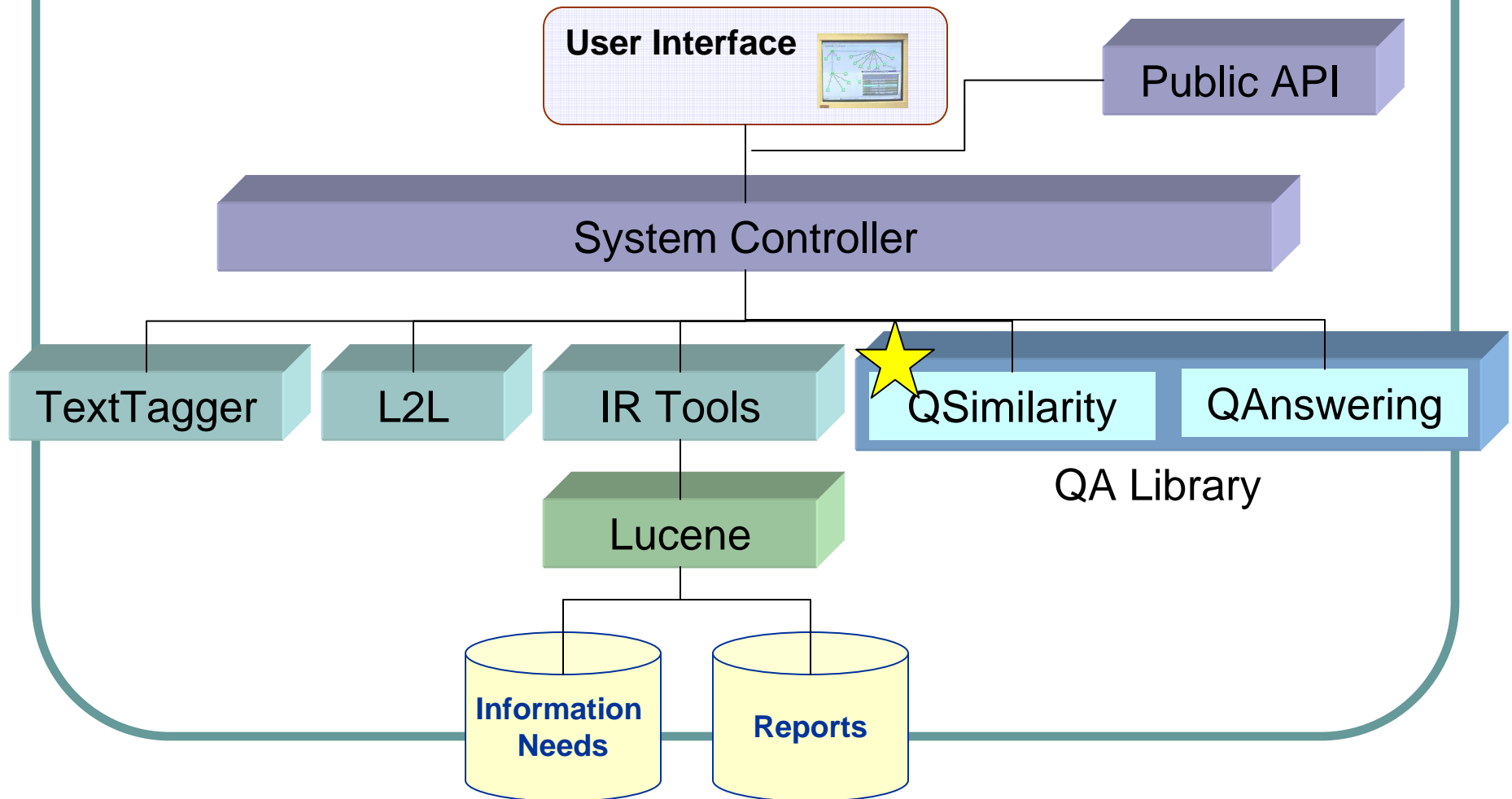
Q&A System Design



IRTools

- Generic information retrieval library
 - It can use various retrieval engines
 - Lucene, Google, MSN, others
 - Provides extensions to support TextTagger output
 - Matching for
 - L2L output
 - TextTagger output
 - Keyword queries

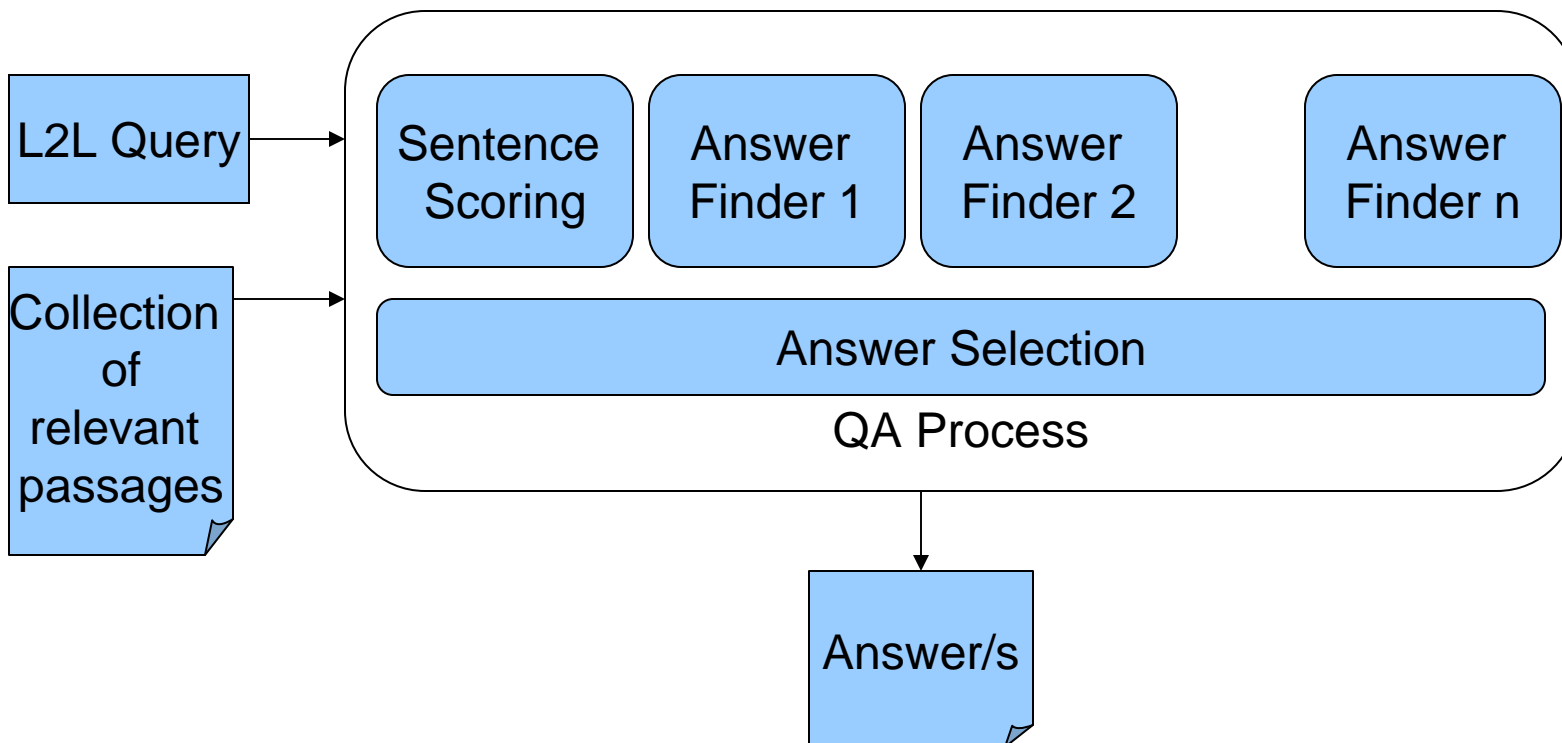
Q&A System Design



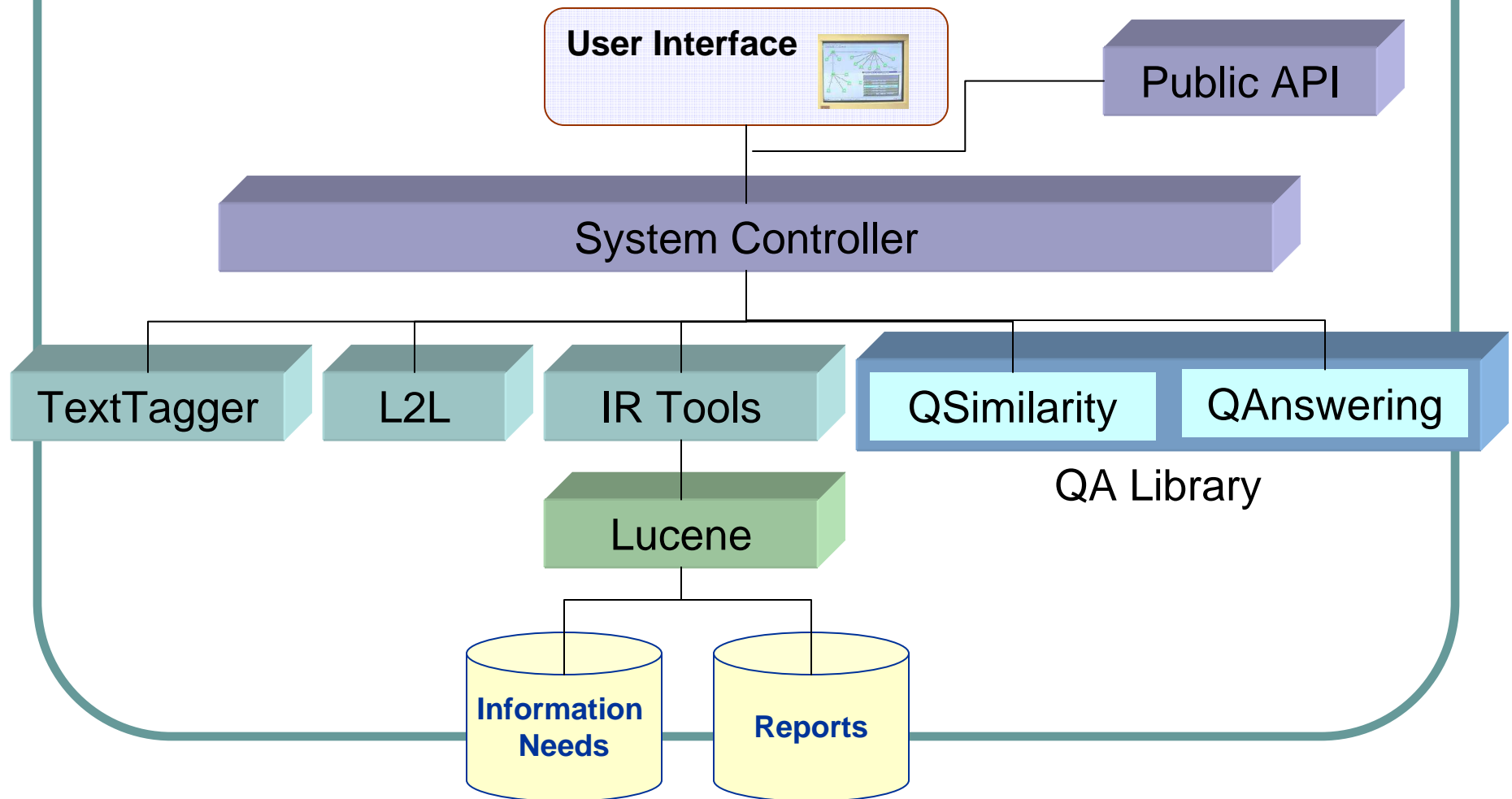
QA Library - I

- Multiple answer finding approaches
- PnP architecture – easy to add new Answer Finders
 - Current Answer Finders include:
 - Keyword based
 - Sentence based
 - Extractions and co-reference
 - Multi Sentence

QA Library - II



Q&A System Design



Indexing Reports

- Process Documents through TextTagger
 - Use rule set tailored to Report sublanguage
 - Rules developed by language analysts
- Tokens, phrases, and important extractions are indexed using IR Tools
 - Remove stopwords
 - Use stems from TextTagger
- Store the report for display

Indexing INs

- Questions are processed through L2L using rule set designed for appropriate domain
- Capture the hierarchical nature of INs for use during matches
- Identify keywords, phrases, important terms and index using IR Tools
 - Add synonyms, spelling variations, stem keywords

Hierarchical Indexing

- Three Options
 1. Index whole IN as a single document and use position info to match
 2. Each piece of IN is a separate field
 3. Index each question as a separate document and reconstruct IN hierarchy
- We chose #3
 - Higher priority given to fine-grain matching
 - Easy to reconstruct hierarchy by storing each IN as a field on the Lucene document

Matching

- **Standard Retrieval**

- Questions processed by L2L and converted into Lucene queries based on keywords by IR Tools
- Basis for other matching approaches

QA Matching for Reports

- Queries processed by L2L and used to identify candidate reports for QA
- QA library processes candidates and scores answers

Matching Reports to INs

- More difficult due to large vocabulary in reports compared to INs
- Process report with TextTagger to identify important tokens, phrases, extractions
- Pick representative content as basis for query
 - Area of ongoing research
- Search using query against INs index
- Use QA library to score found questions against given report

Reports to INs Issue

- Report length makes developing query difficult
 - How to pick the right terms to represent a report?
 - Ongoing research area
 - Strategy:
 - Remove stopwords
 - Use key phrases, tokens as identified by TextTagger
 - Investigate summarization algorithms

INs to INs

- Find similar information needs to expand question set
- Use L2L to process questions
- Utilize Query Similarity library to go beyond simple keyword matching
- Pluggable Interface allows for easily trying new approaches

Query Similarity

- Score query pairs between 0 and 1
- Keyword Approaches:
 - Keywords Overlap
 - Relaxed Keywords in Common allows some missing keywords
 - Synonyms
 - Test to see if a keyword is a synonym of other
 - Edit Distance
 - Account for spelling variations
 - Nominalization
 - See if one keyword is a nominal of the other

Query Similarity (cont.)

- Non-keyword approaches:
 - Answer Type Match
 - Are the two queries interested in the same category of answer
- Combine two or more of the approaches and weight them according to how important each piece is to identify final result

Conclusions & Future Work

- Approach has been validated with customers supporting the IC
- Discussions with financial services companies suggest need in such large, knowledge intensive organizations
- Other domains have same need
 - Systematic Reviews in Medicine
 - Query-based reports that comprehensively examine medical literature
 - Identify, evaluate, synthesize evidence-based studies
 - Formulate best approach for a particular diagnosis
 - Take up to 2 years to write
 - Need continuous updates
 - Q-to-A can provide this continuous updating process

TextTagger Phases

Text:

Sandoz Corp's Northrup King Co said it bought Stauffer Seeds, a unit of Stauffer Chemical Co. Terms were not disclosed.

Tokenization – identifies the basic units of text

Tokenized:

Sandoz | Corp | 's | Northrup | King | Co | said | it | bought
Stauffer | Seeds | , | a | unit | of | Stauffer | Chemical | Co | . | Terms |
were | not | disclosed | . |

Sentence Detection – identify sentence boundaries

Sentence Detection:

<S> Sandoz | Corp | 's | Northrup | King | Co | said | it | bought |
Stauffer | Seeds | , | a | unit | of | Stauffer | Chemical | Co | . | </S>
<S> Terms | were | not | disclosed | . | </S>

TextTagger Phases

Part-of-Speech Tagging – identifies the syntactic category of a word in a sentence

<S>Sandoz|NP Corp|NP 's|POS Northrup|NP King|NP Co|NP said|VBD it|PRP bought|VBD Stauffer|NP Seeds|NP ,|, a|DT unit|NN of|IN Stauffer|NP Chemical|NP Co|NP .|. </S>

<S>Terms|NNS were|VBD not|RB disclosed|VBN .|. </S>

Stemming (Lemmatization) – identifies the root form of words

<S>Sandoz|NP Corp|NP 's|POS Northrup|NP King|NP Co|NP say|VBD it|PRP buy|VBD Stauffer|NP Seeds|NP ,|, a|DT unit|NN of|IN Stauffer|NP Chemical|NP Co|NP .|. </S>

<S>Terms|NNS be|VBD not|RB disclose|VBN .|. </S>

TextTagger Phases

Non-compositional – identifies phrases with two or more words which the meaning of the phrase is different from the combination of meaning of the individual words

hot dog → hot_dog

real estate → real_estate

Phrase Bracketing

Temporal Concepts

(“April 5th”, “last week”, “100 years ago”)

Numeric Concepts

(“\$50”, “80 mph”, “lower 30s”, “half a pound”)

Named Entity Phrases

(“Ozgur Yilmazel”, “NASA”, “Syracuse University”)

Common Nouns

(“main buyer”, “five policemen”, “gross margin”)

TextTagger Phases

Entity categorization – assign semantic categories to phrases

Sandoz Corp → **<NP cat="company">** Sandoz Corp **</NP>**

President Kennedy → **<NP cat="person">** President Kennedy **</NP>**



TextTagger Phases

Event and Relation extraction – identifies attributes of entities and their relations

Text: Ozgur is a student at Syracuse University. He arrived in Syracuse in 1997.

1. namedentity = Ozgur
category = person
isa = student = entity3
2. namedentity = Syracuse University
category= education unit
associated = student = entity3
3. entity=student
category=person

4. namedentity = Syracuse
category = city
5. entity=1997
category=year
6. event=arrive
destination= Syracuse = entity4
agent = he = entity7
7. entity=he
coref=Ozgur=entity1

