

Pro-Active Question-Answering

Elizabeth D. Liddy

Center for Natural Language Processing
Syracuse University

April 23, 2007

Typical vs. Atypical Scenario

- **Typical** - Automated Question-Answering systems find answers to clients' new questions from a corpus of reports, websites, or newsfeeds
 - Types of questions
 - Factoid
 - List
 - Definition
 - How / Why

Typical vs. Atypical Scenario

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

Typical QA

- Well-known application that goes one step further than document retrieval
- Provides the specific information asked for in a natural language question
 - Not simply a listing of URLs that have all or some of the keywords input by the user -- and links to pages that contain undifferentiated types of information that the user must then search individually to find the information they need
- But rather, question-answering recognizes the specific aspect of the topic that is being asked about
 - Provides either short answers / answer-providing snippets that address the specific question
- QA is proving to be a highly desirable capability, with utility in the enterprise setting of closed-domain question-answering

Atypical QA

- The system tracks who has been asking what questions, and then matches recently acquired documents or newly produced reports to the standing information needs of users within an enterprise whose history suggest that the information in a new report would be useful.
- This capability is similar to what years ago was referred to as "Selective Dissemination of Information", but is now done with a great deal more precision due to richer NLP-based representations of both queries and new documents, and more sophisticated matching algorithms.

Organizational Settings

- Knowledge-intensive organizations
 - Intelligence Community
 - Market analysts
 - Competitive intelligence
- Rely on a supply chain of information
 - Connects Consumers of information with the Producers
 - Begins with Requests For Information / Information Needs from the Consumers that may start out as general in nature
 - Can be thought of as ‘containers’ of ‘Question Sets’
 - Groups of logically related, frequently complex questions
 - Fulfilled by Producer for single consumers, but could be of interest to many others

Information Needs Hierarchies

- < Information Need – 1 >
 - < Question Set – 1.A >
 - < Question – 1.A.1 >
 - < Question – 1.A.2 >
 - < Question – 1.A.3 >
 - < Question Set – 1.B >
 - < Question – 1.B.1 >
 - < Question – 1.B.2 >

- < Information Need – 2 >
 - < Question Set – 2.A >
 - < Question – 2.A.1 >
 - < Question – 2.A.2 >
 - < Question Set – 2.B >

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>
```

Information Needs Hierarchies

< Information Need – 1. Global Warming >

< Question Set – 1.A Metrics >

< Question – 1.A.1 What metrics, using measures such as temperature, have been collected to study Global Warming? >

< Question – 1.A.2 What evidence supports the increase of Global Warming? >

< Question – 1.A.3 What organizations are researching and publishing papers which include metric-based data on Global Warming? >

< Question Set – 1.B Photographic & Imagery

Producer Side

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

Organizational Perspective

- Intelligence Community & Market Analysis companies desire the capability to easily:
 - Match intelligence products / reports as they are produced to standing Information Needs
 - Retrieve answers to new questions
 - Detect similarity between questions
 - When a new question is asked, the collected responses to earlier, similar questions will be shared with the new inquirer
 - Determine similarity of interests of those asking questions
- In a single, integrated, easy to use QA System

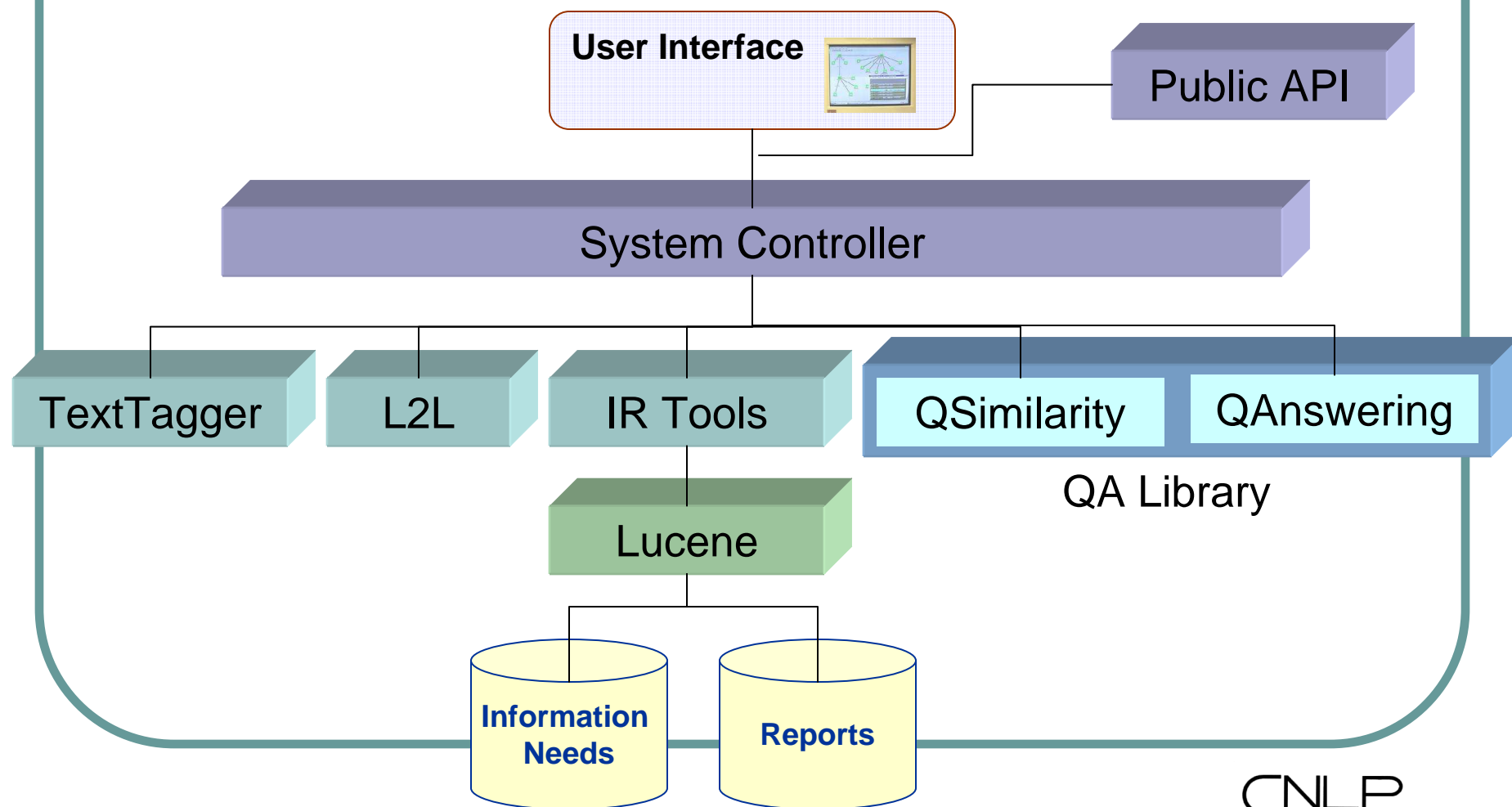
Our Goals

- Build Prototype System to:
 - Identify individual Questions / Question Sets that are answered by new documents
 - Either as produced or acquired
 - Conduct traditional search and Question-Answering over documents and questions

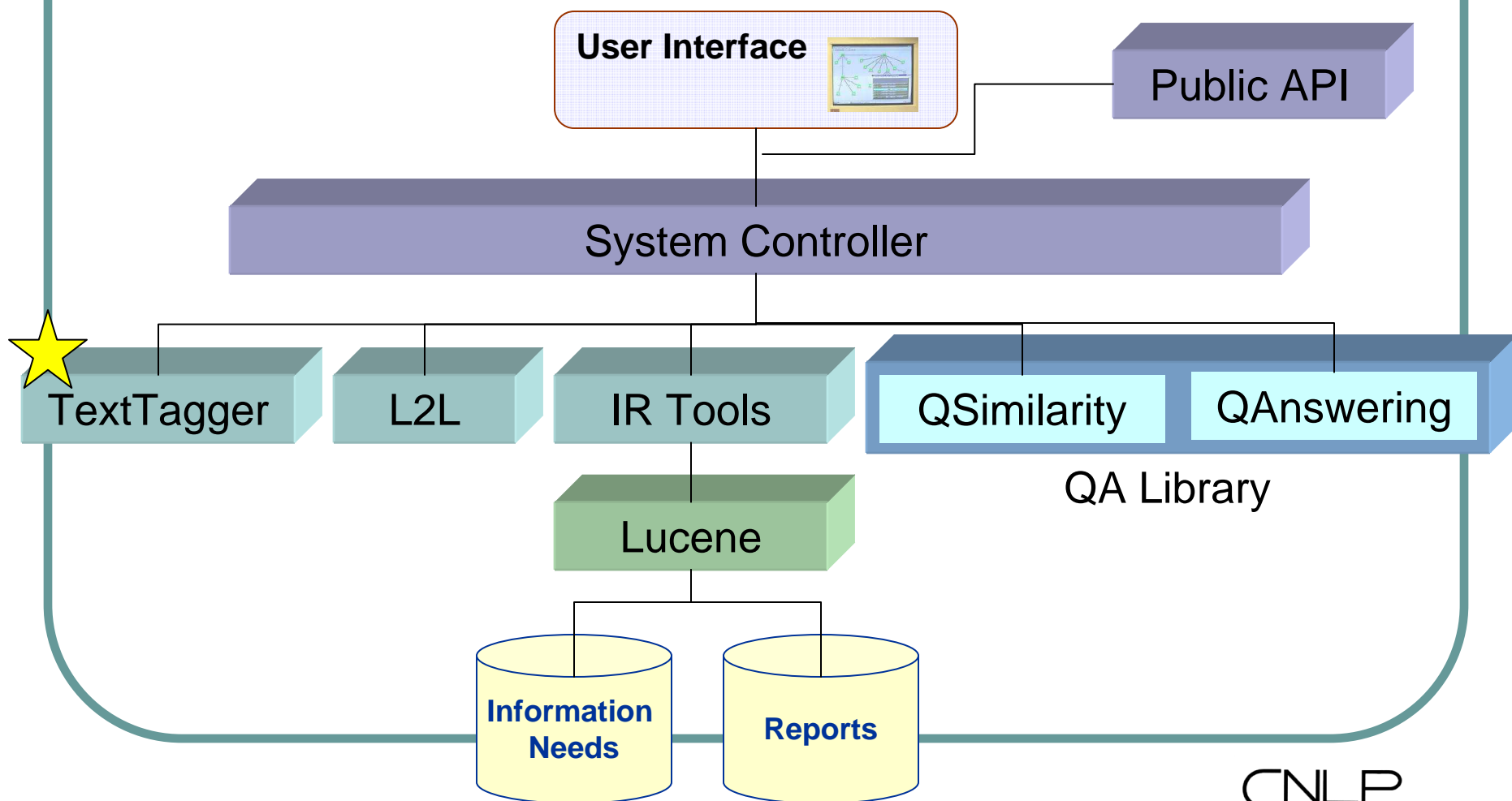
System Design

- Single web-based application provides:
 - Traditional search of new INs against existing reports
 - Matching of newly produced reports (answers) to existing INs or Questions that are answered by report
 - Ability to browse both reports & INs
 - Display of precise context where an IN or question is answered in report
 - Identify similar questions across Question Sets to help facilitate collaboration among interested parties

QA System Design



QA System Design



TextTagger

- NLP-based Information Extraction system developed at CNLP
- Analyzes unstructured text of any type at all the levels of language at which meaning is conveyed
 - Morphological
 - Lexical
 - Syntactic
 - Semantic
 - Discourse
- Uses a sequence of steps called phases

TextTagger Phases

- Tokenization
- Part-of-Speech Tagging
- Stemming
- Non-compositional Phrase Identification
- Phrase Bracketing
 - Named Entities
 - Temporal Concepts
 - Numeric Concepts
- Entity Categorization
- Event and Relation Extraction

TextTagger Output

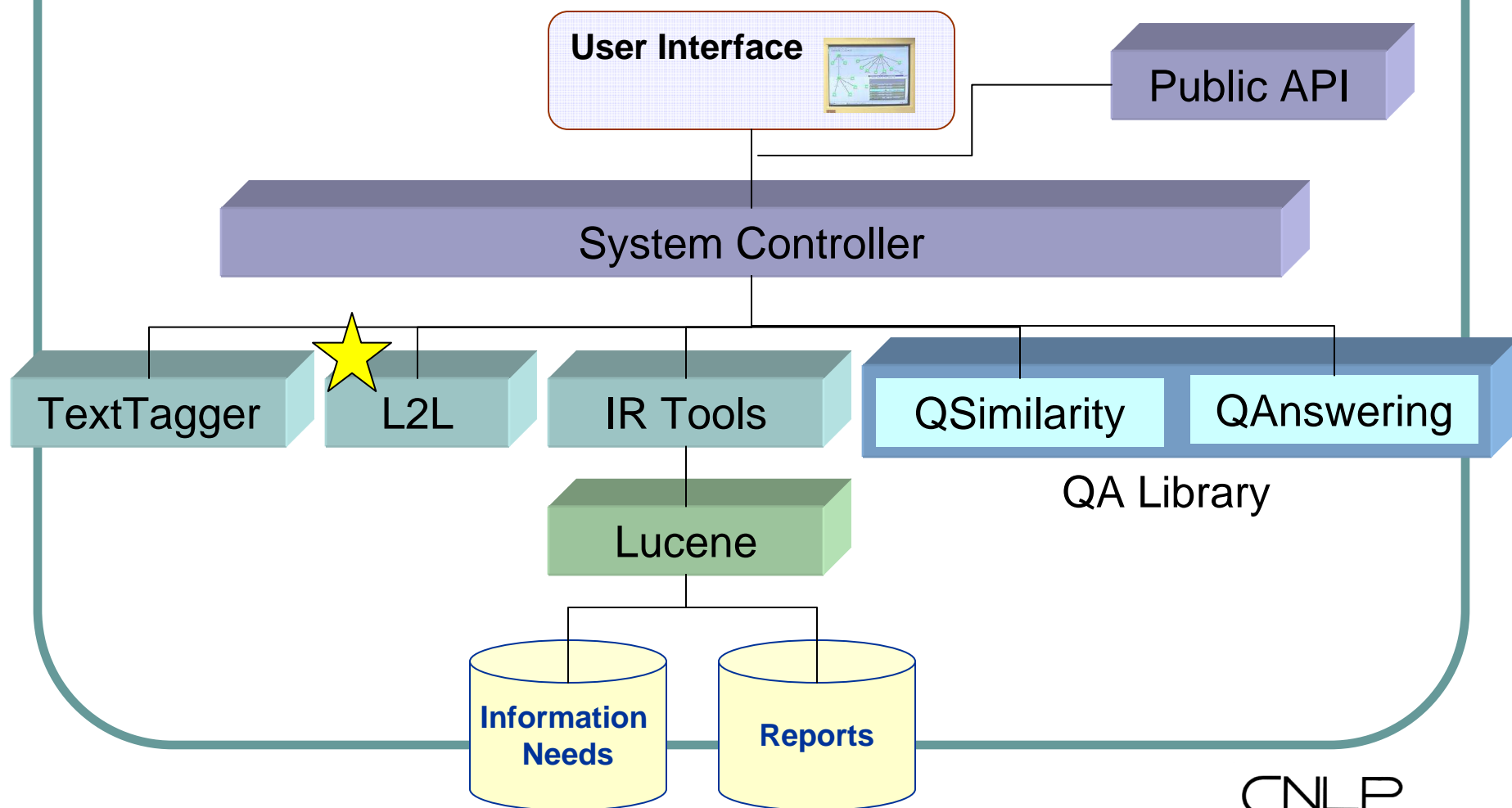
Central Capital Corp said it planned a three-for-two split of its common and class A subordinate voting shares, subject to shareholder approval at the April 23 annual meeting.

<S> <NP> cat="co" Central_Capital_Corp </NP> say|VBD it|PRP
plan|VBD a|DT

In addition to individual words'

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

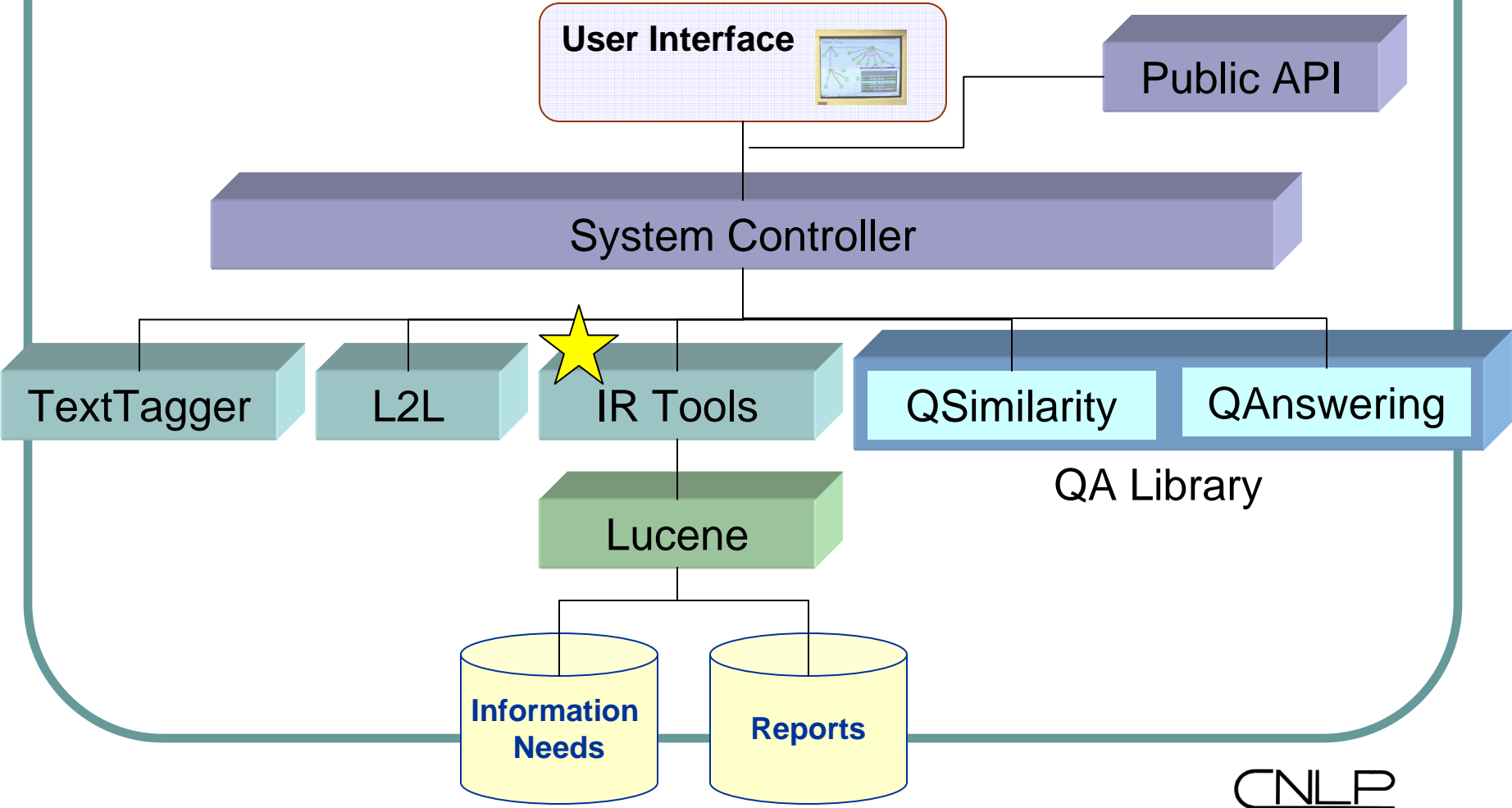
QA System Design



Language-to-Logic Query Analyzer

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

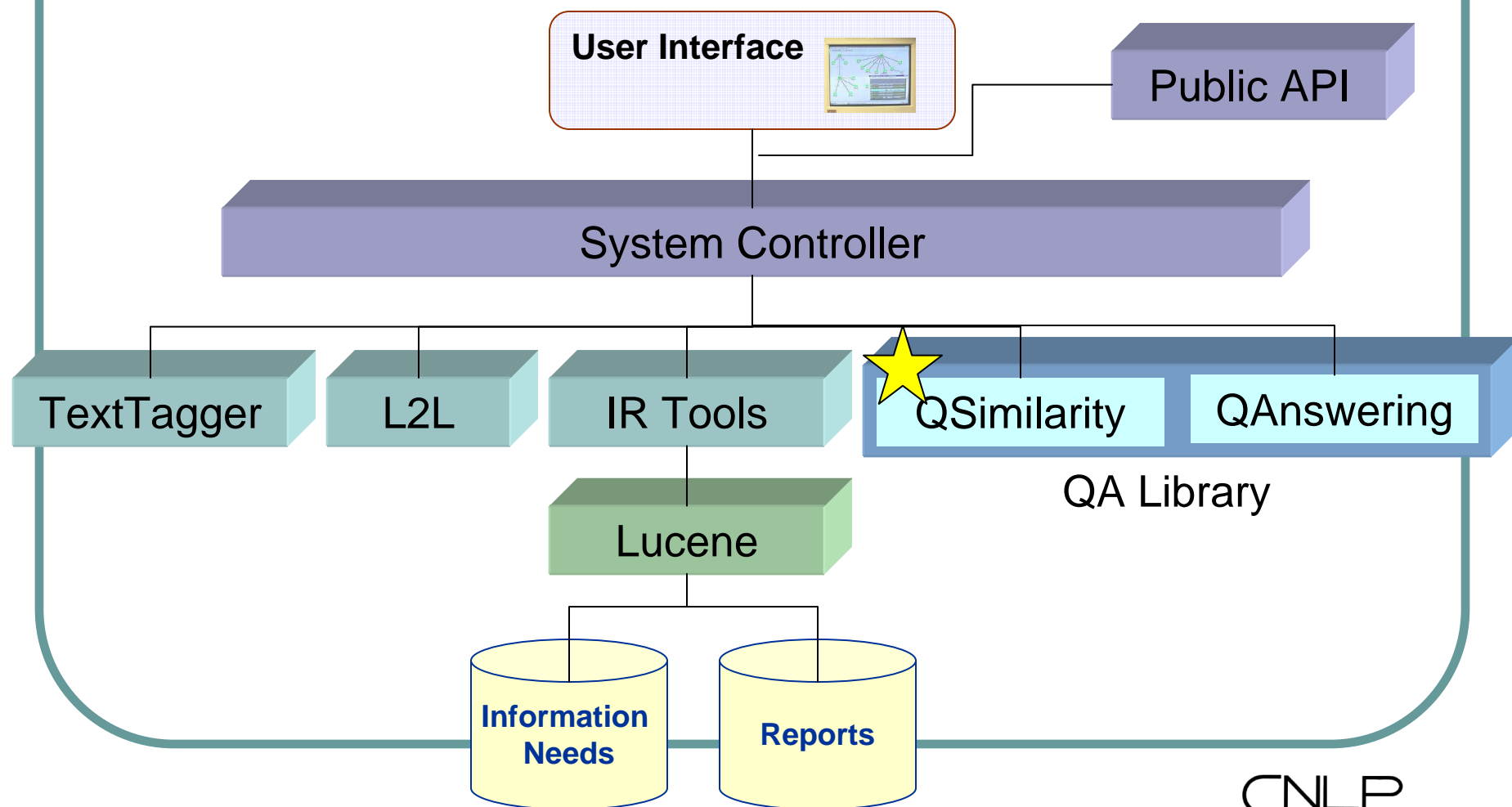
QA System Design



IR Tools

- Generic Information Retrieval library
 - Can use various retrieval engines
 - Lucene, Google, MSN, others
- Plus extensions to support matching for:
 - L2L output
 - TextTagger output
 - Keyword queries

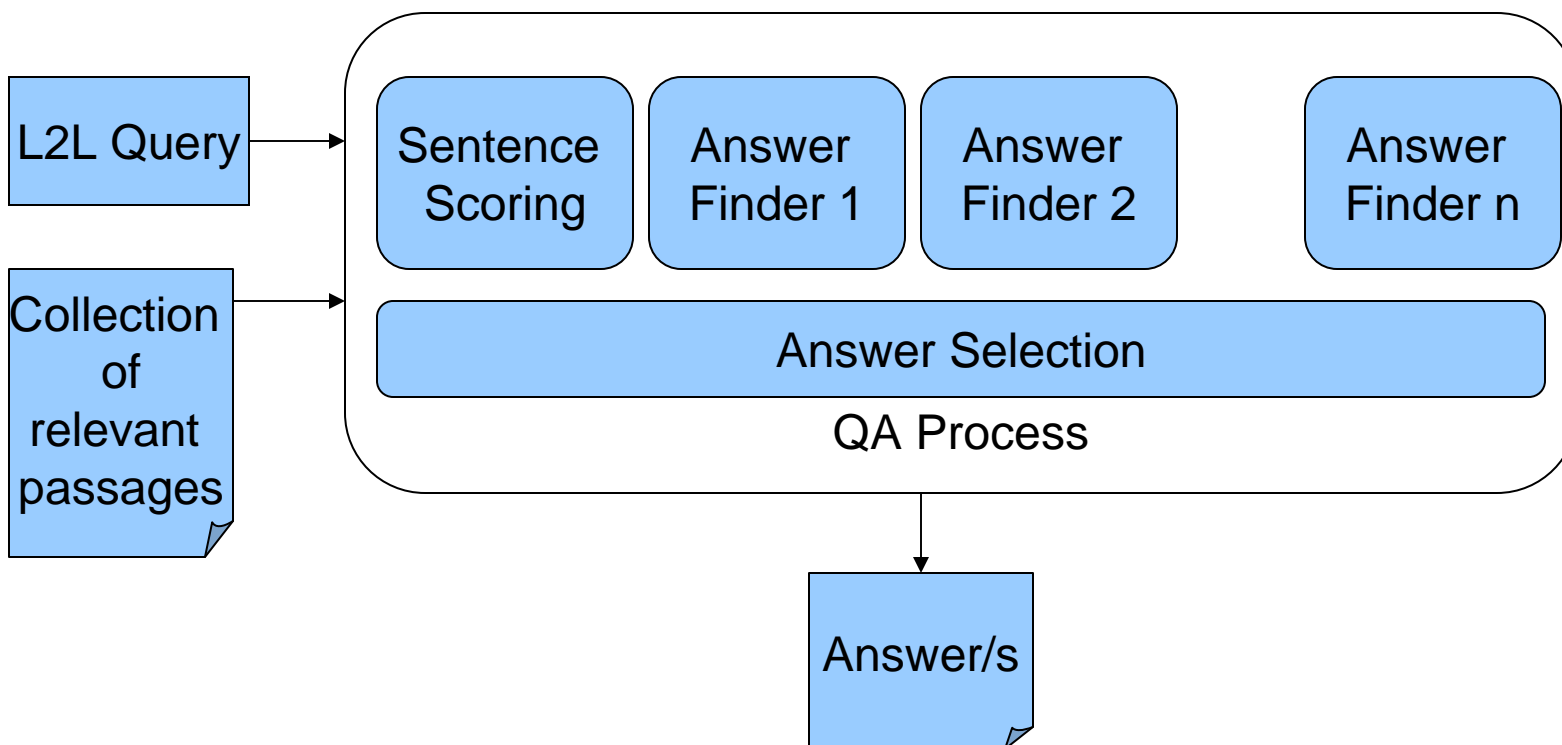
Q&A System Design



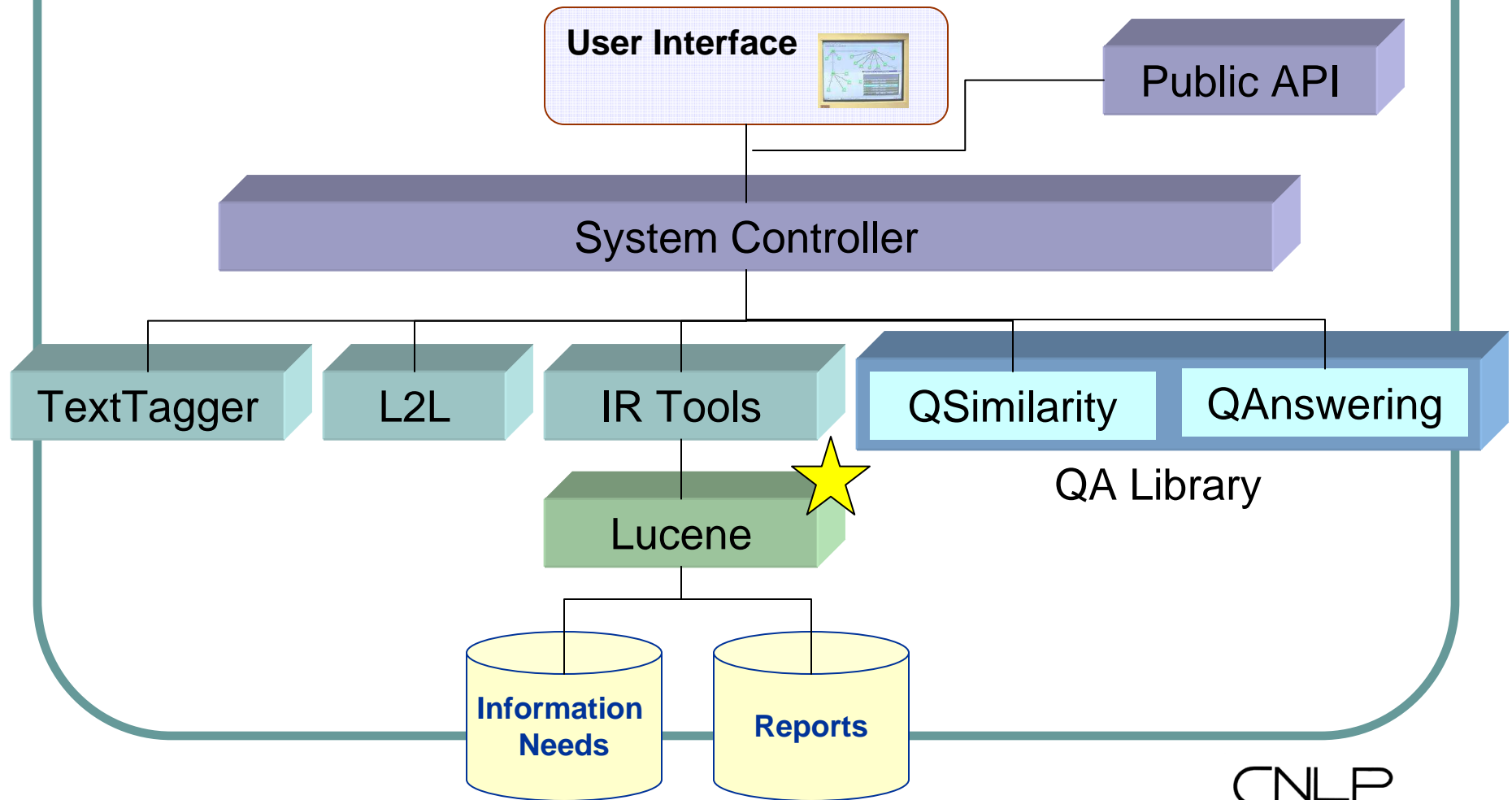
QA Library - I

- 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



QA System Design



Indexing Reports

- Initially processed through TextTagger
 - Use algorithms tailored to Report sublanguage
 - Developed by language analysts
- Terms, phrases, and important extractions are indexed using IR Tools
 - Using output from TextTagger
- Store the initial report for later display

Indexing INs

- Individual questions / sets / INs are processed through L2L using rule set designed for appropriate domain
- Captures the hierarchical nature of INs for use during matches
- Identifies keywords, phrases, important terms and indexes them using IR Tools
 - Adds synonyms & spelling variations

Hierarchical Indexing of INs

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

QA Matching for Standard QA

- Standard Retrieval
 - Questions processed by L2L and converted into Lucene queries based on output of IR Tools
 - Used to identify candidate reports for QA
 - Basis for other matching approaches
- QA library processes candidate sections from reports and scores answers

Matching New Reports to INs

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

Reports-to-INs Issue

- Report length makes developing query / queries based on it a difficult task
 - Which sections to represent?
 - Level of detail of indexing?
 - How to pick the right terms to represent it?
- Strategy:
 - Morphological, lexical, syntactic processing
 - Use key phrases, terms as identified by TextTagger
 - Investigating summarization algorithms
- Ongoing research area

INs to INs

- Goal is to 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 Approaches

- Score query pairs between 0 and 1
- Simple Keyword Overlap
 - Relaxing Keywords in Common allows some missing keywords to be present
- Synonyms
 - Tests to see if a keyword is a synonym of other
- Edit Distance
 - Accounts for spelling variations
- Nominalization
 - Checks if one keyword is nominalization of a verb in the other
- Answer Type Match
 - Are the two queries interested in the same category of answer?
- We combine several approaches and weight them according to how important each piece appears to be in order 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 on a particular disease / disorder
 - Identify, evaluate, synthesize evidence-based studies
 - Formulate best approach for a particular diagnosis
 - Take up to 2 years to write
 - Need continuous updates
 - Inverse QA 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

