

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 reports, websites, or newsfeeds
 - Provide the specific information, not simply list of URLs with all / some of the keywords input by the user + and links to pages containing undifferentiated types of information
 - Rather, QA recognizes specific aspect of topic asked about
 - Provides responses that address the specific question from a single document or across multiple documents

Typical vs. Atypical Scenario

- **Typical** - Automated Question-Answering systems find answers to clients' new questions from reports, websites, or newsfeeds
 - Provide the specific information, not simply list of URLs with all / some of the keywords input by the user + and links to pages containing undifferentiated types of information
 - Rather, QA recognizes specific aspect of topic asked about
 - Provides responses that address the specific question from a single document or across multiple documents
- **Atypical** – Inverse QA
 - Matches new reports / postings / analyses dynamically as they are produced to pre-existing questions
 - Answer-providing documents become the questions

Organizational Settings

- Knowledge-intensive organizations
 - Intelligence Community
 - Market analysts
 - Competitive intelligence
- Model is 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>RBC-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 >

<Background - Up to date information to continually assess the state, effects, impacts of 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 on a specific topic or industry
 - Provide answer-specific responses / reports
 - Able to anticipate what the questions are / might be
 - Consumers may not yet realize need
- SME tasks to satisfy current & future INs
 - Research
 - Analysis
 - Reporting
- End-products are lengthy, wide-ranging reports
 - Can provide answers to multiple INs or more granular Questions of multiple Consumers
 - May be produced asynchronously from INs, so organization must ensure that every Consumer who might benefit from this information receive it

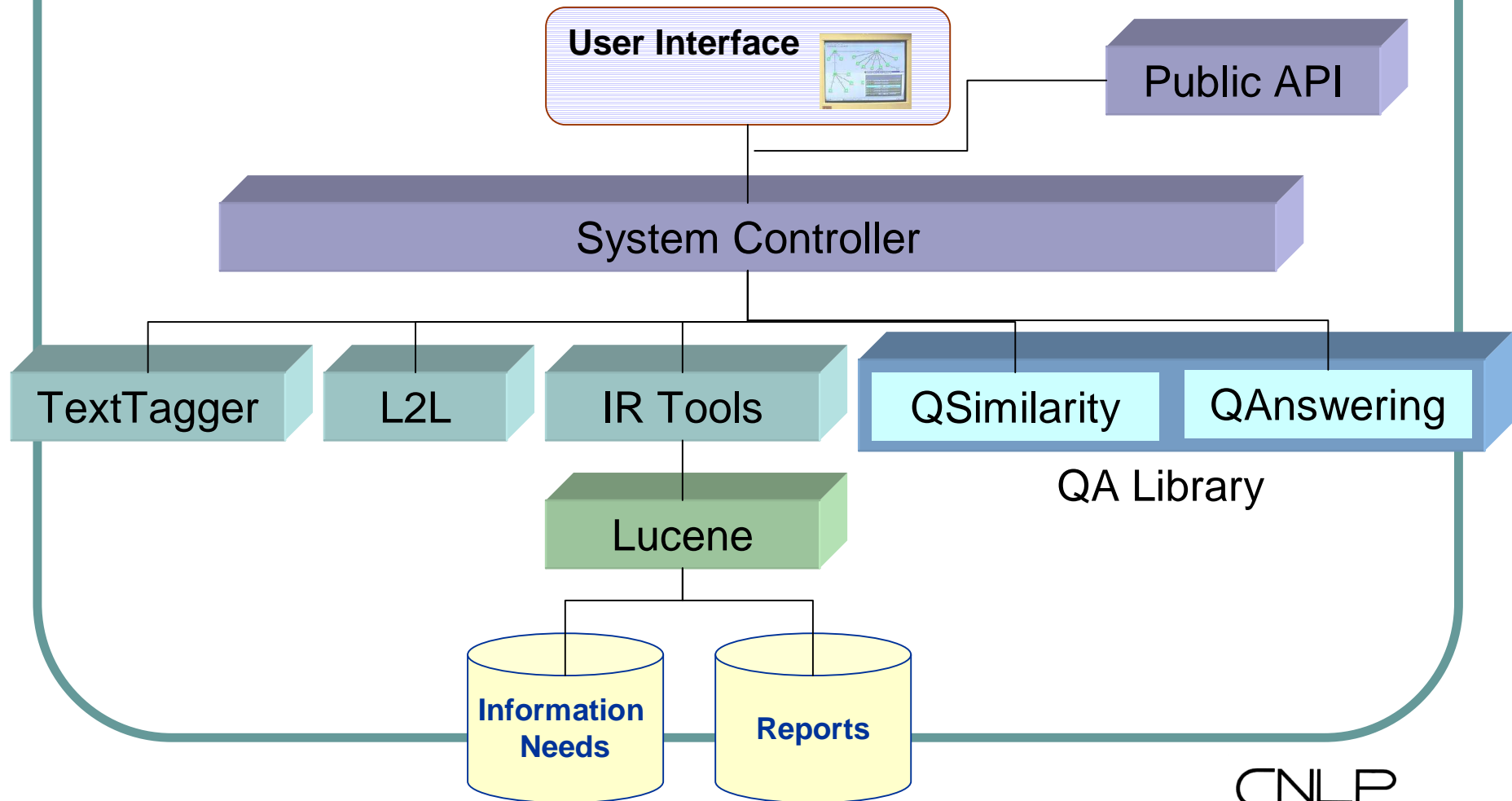
Organizational Perspective

- 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

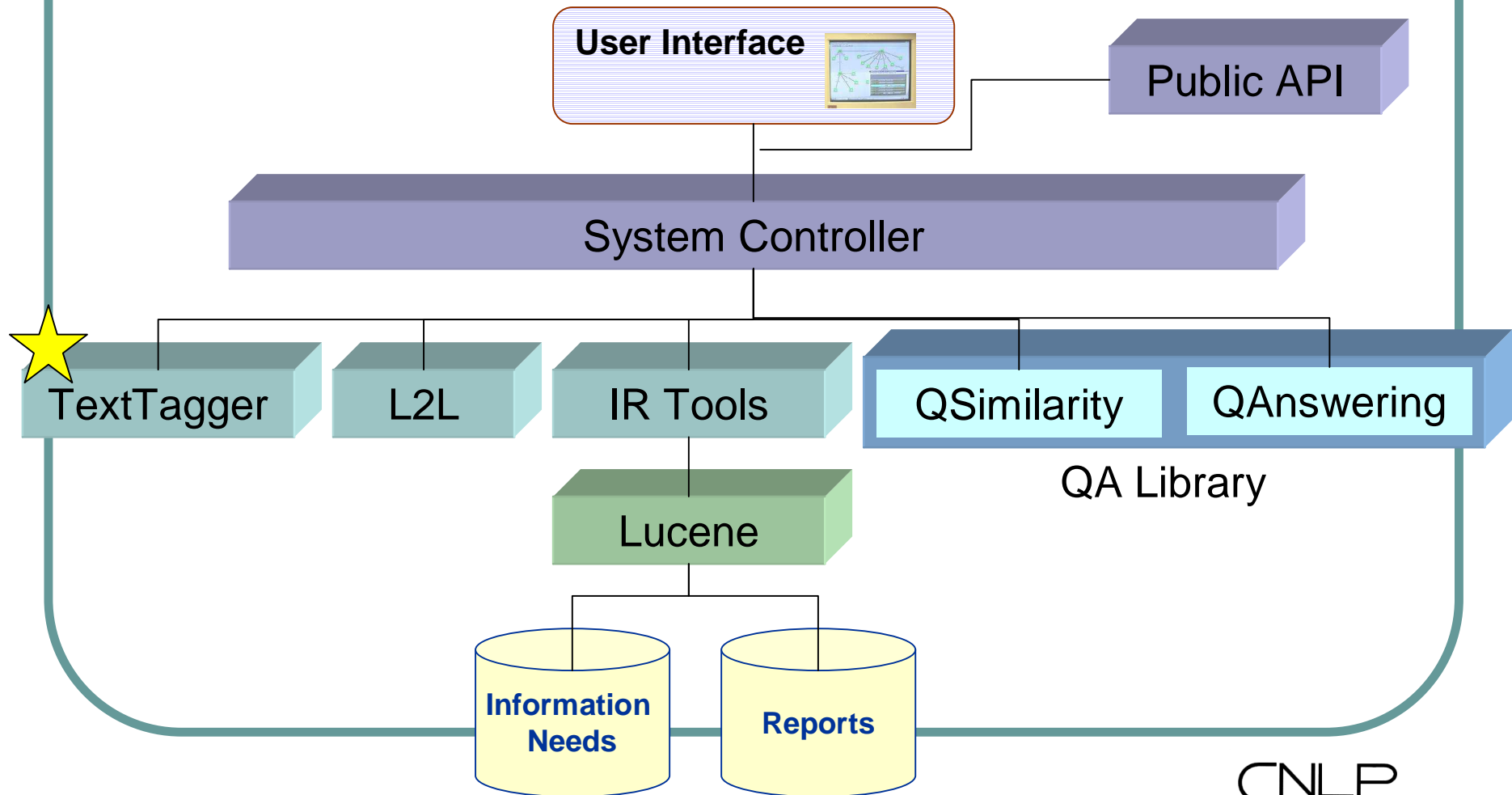
Our Response

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

QA System Design



QA System Design



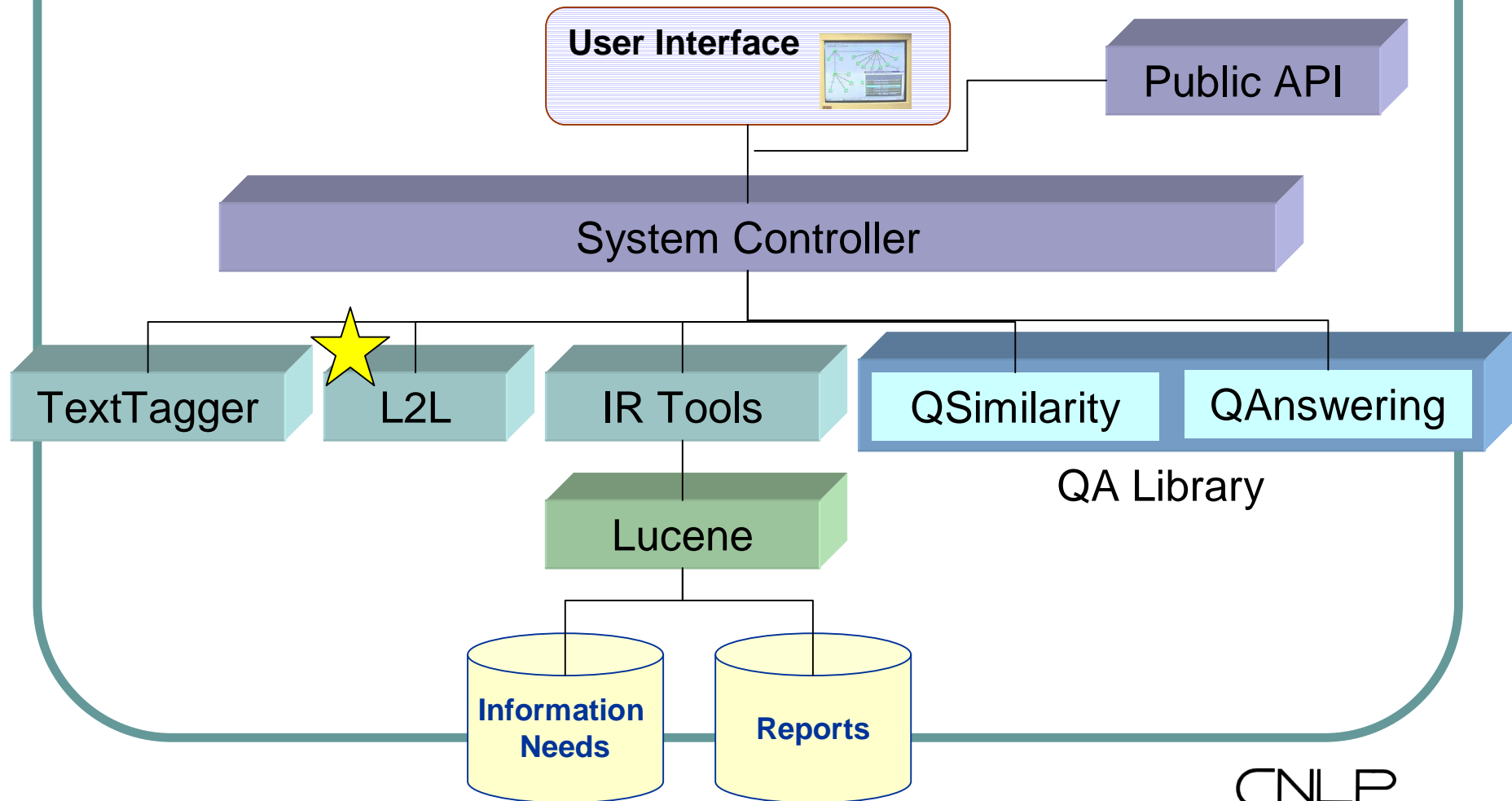
TextTagger

- Well-vetted 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
- Produces a rich representation of content using a sequence of configurable 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
- Reference Resolution

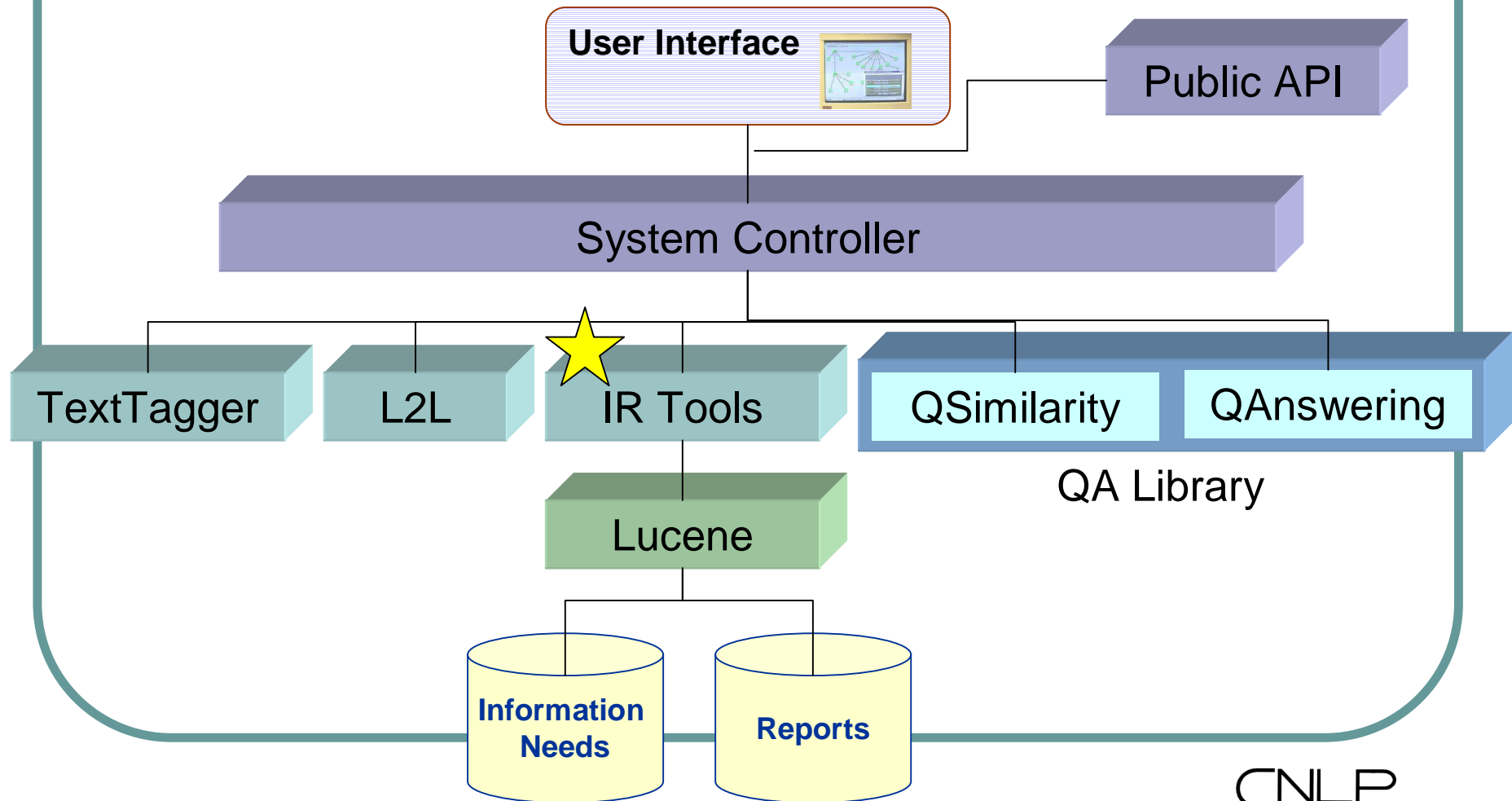
QA System Design



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
- Uses natural characteristics of language to produce a logical representation
 - Conjunctions, prepositions, modifications, syntax
 - Essential for QA, because not going for just topic match

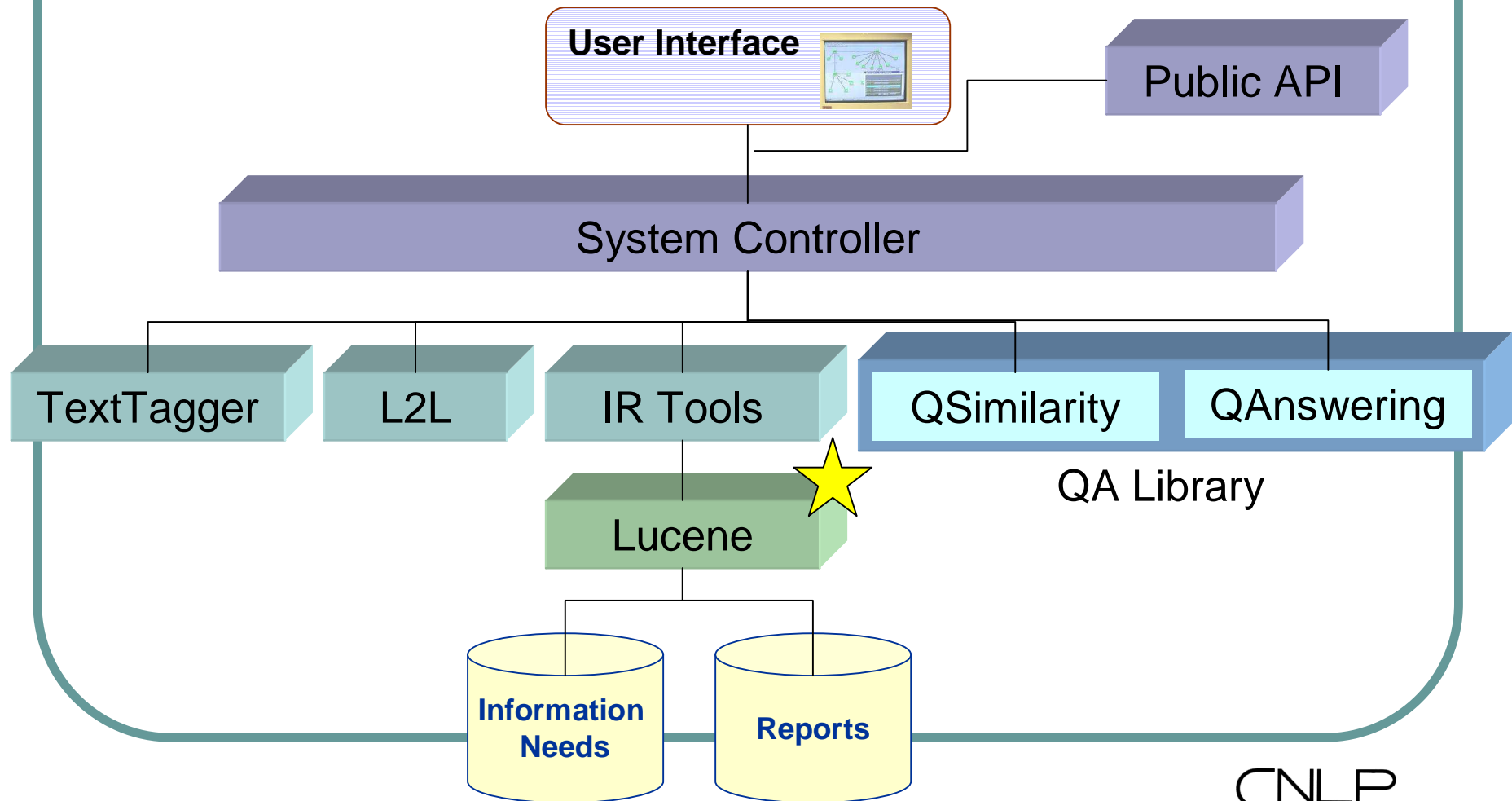
QA System Design



IR Tools

- Generic Information Retrieval library
 - Can use own search engine or various retrieval engines
 - Lucene, Google, MSN, Solr, others
 - Working with range of corporate customers who've chosen open-source search engines, but need both software and theory-based assistance
- Plus NLP extensions to support matching based on
 - L2L output
 - TextTagger output
 - Enriching index with NLP output to improve results from open source engines

QA System Design



Indexing INs

- Individual Questions / Question Sets / INs are processed through L2L
- Captures the hierarchical nature of INs for use during matches
- Identifies keywords, phrases, extractions and indexes them using our IR Tool specialization of Lucene
 - Adds synonyms & spelling variations
 - Tunes weighting parameters

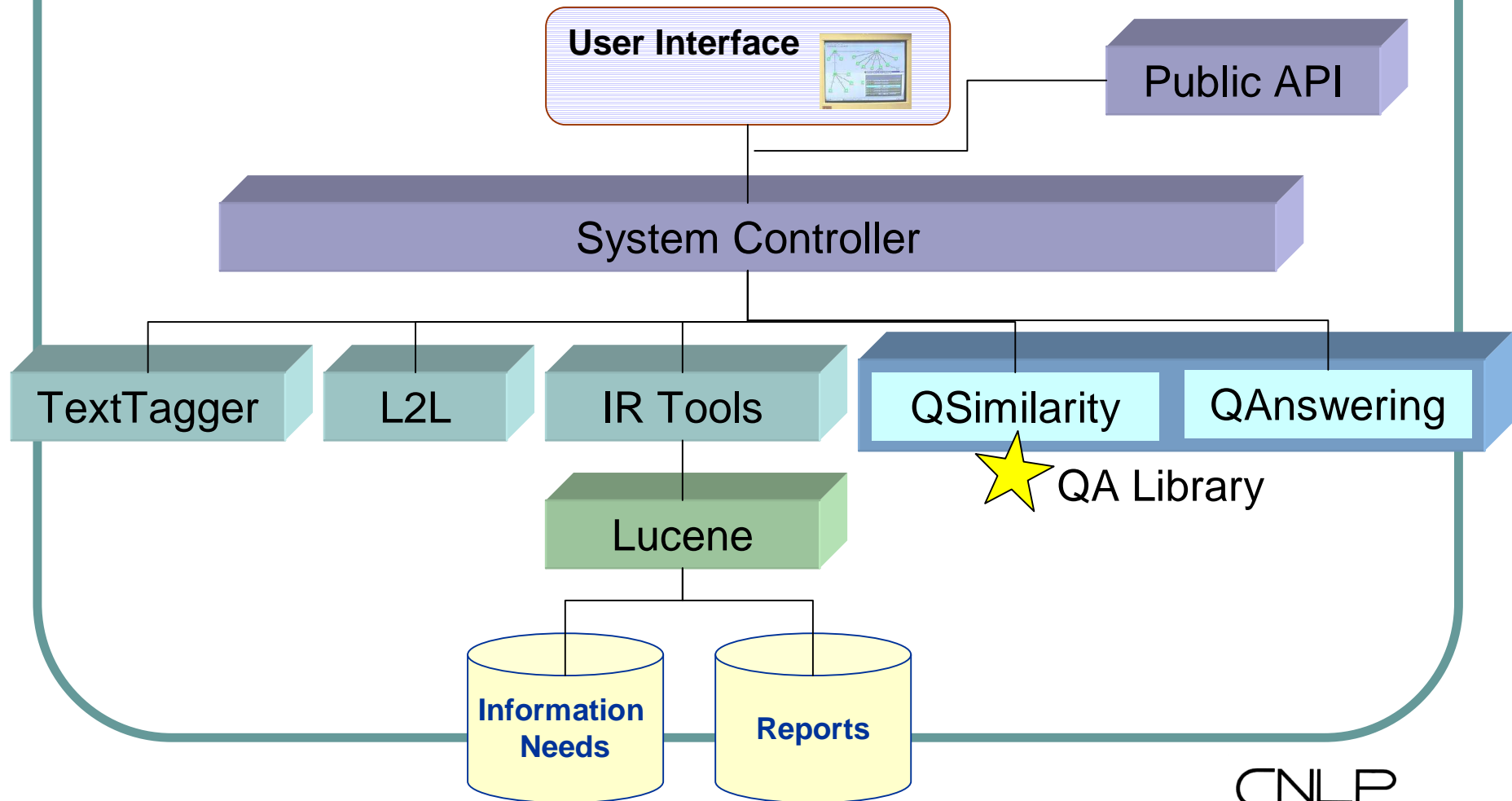
Hierarchical Indexing of INs

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

Indexing Reports

- Initially processed through TextTagger
 - Can use our tools to quickly tailor algorithms to genre / domain Sublanguage
- Terms, phrases, and important extractions are indexed using IR Tools
 - Using output from TextTagger
 - Reports will serve as both answers and 'questions'
- Store the initial report for later display

Q&A System Design



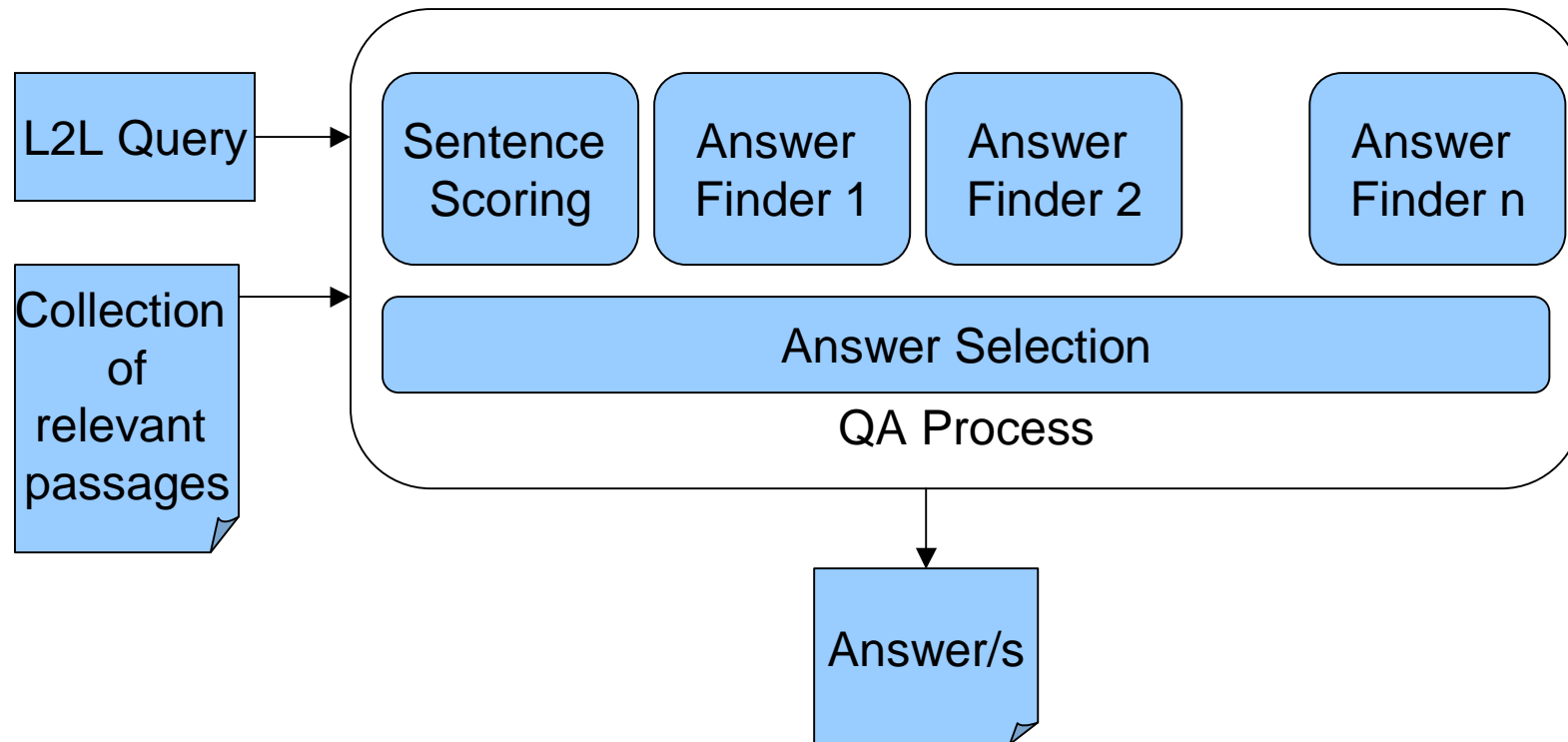
QA Matching for Standard Queries

- 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 next stage of more refined matching
- QA library processes candidate sections from reports and scores answers
 - Ranks them if multiple answers
 - Cut-off criterion utilized to prevent poorly responsive answers from being shown

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



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, entities, events, relations
- Pick representative content as basis for query
 - Alternatives available based on discourse linguistic theory
- Search using report-based query against INs' index
- Use QA library to score potentially relevant Questions & INs against given report

Matching Reports-to-INs Issue

- Report length makes developing query / queries based on them a difficult task
 - Which sections to represent?
 - Level of detail of indexing?
 - How to pick the right terms to represent them?
- Strategy:
 - Morphological, lexical, syntactic, semantic processing
 - Enables automatic content-based metadata generation by TextTagger for indexing
 - Precise enough to enable QA to INs or Questions

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 inserting new approaches
 - Social Network Analysis
 - Clustering

Question – to – Question Similarity

- Score query pairs between 0 and 1
- Simple Keyword Overlap
 - Relaxing Keywords in Common allows some missing keywords to be present
- Synonyms
 - Tests if a keyword in 1 query is a synonym of keyword in other
- Edit Distance
 - Accounts for spelling variations
- Nominalization
 - Checks if one keyword is nominalization of verb in the other
- Answer Type Match
 - Are the two queries interested in the same category of answer?
- 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 these 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