University of Freiburg
Chair of Algorithms and Data Structures

Seminar Information Extraction

Lymba (QA System)

Fabian Schillinger
schillif@informatik.uni-freiburg.de
15.01.2014

TREC 2007

- Question anwering track
  - Blog data & Newswire documents
  - Factoid questions

    "How many calories are there in a Big Mac?"
  - List questions

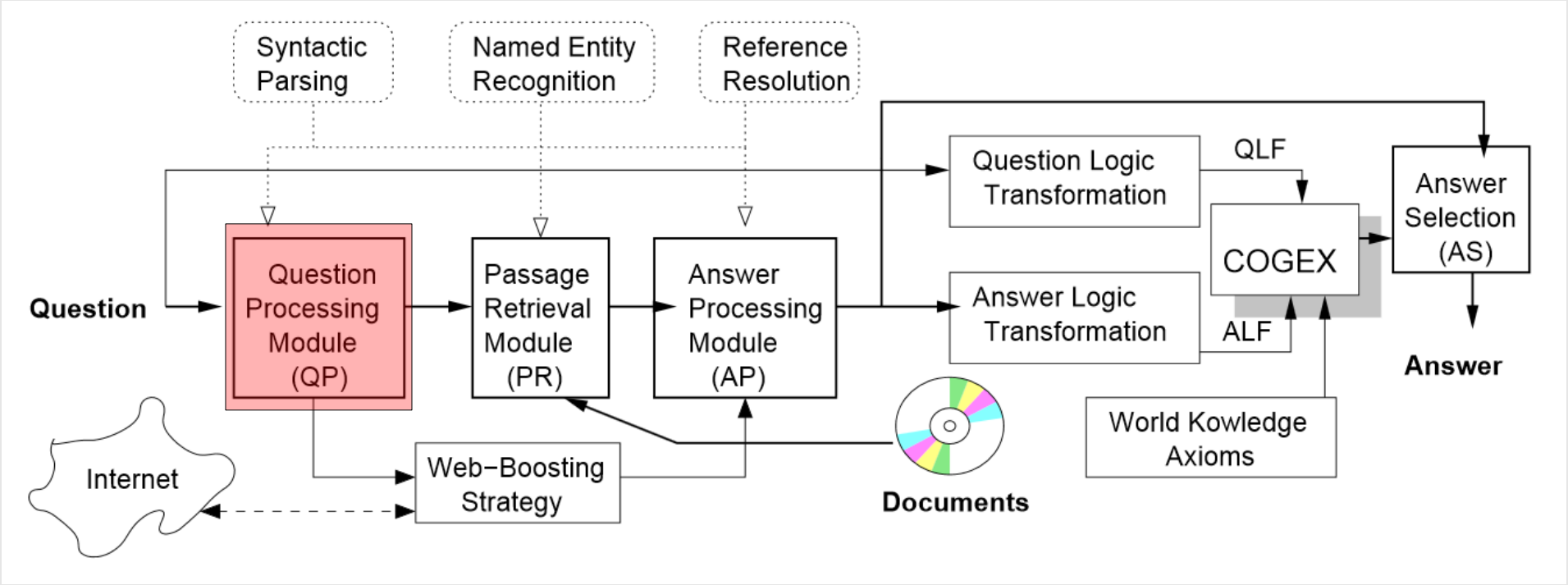
    "List the names of chewing gums."
  - "Other" questions

    interesting facts about some target

PowerAnswer 4 Overview

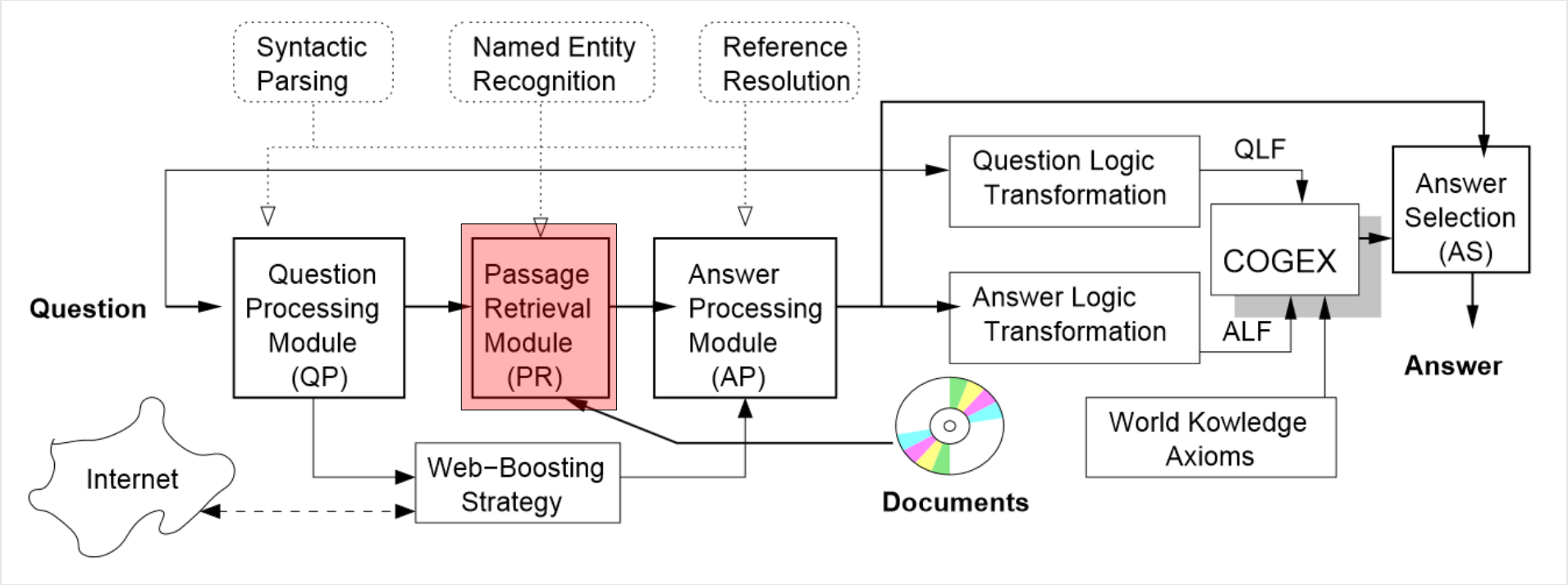
- Distributed strategy-based QA-System
- Strategy is collection of the components
  - Question Processing (QP)
  - Passage Retrieval (PR)
  - Answer Processing (AP)

PowerAnswer 4 Overview



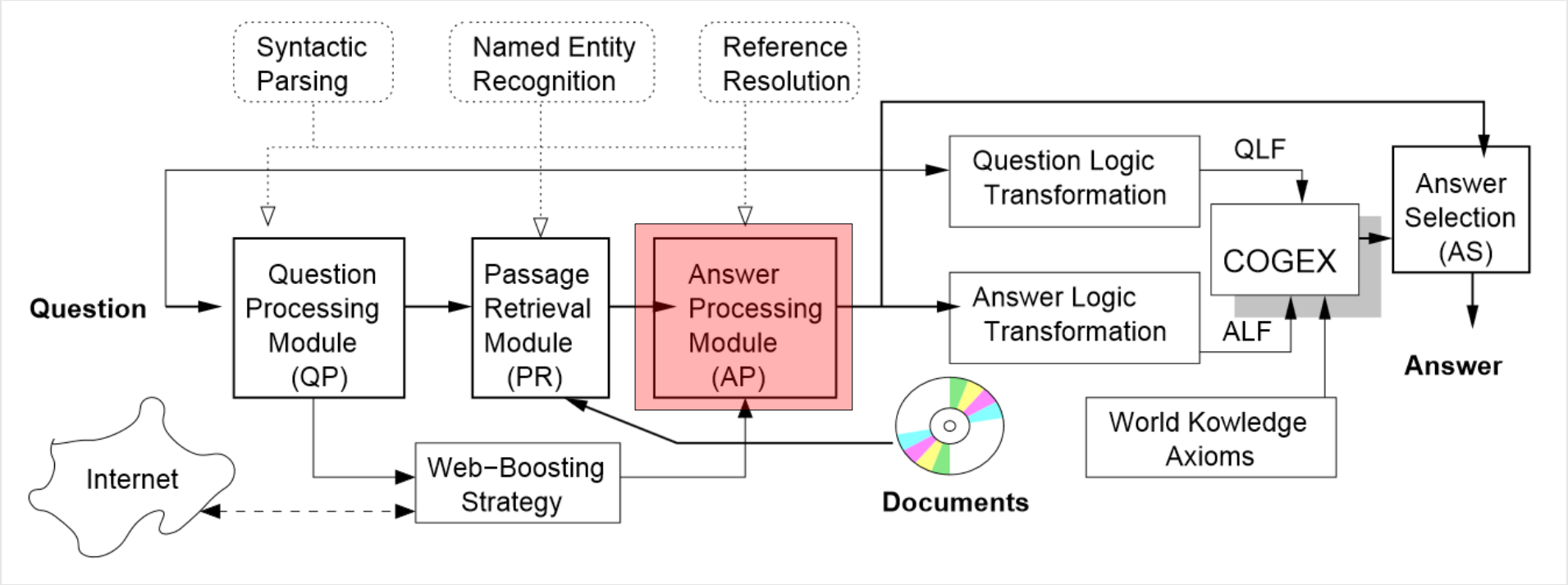
- determine temporal constraints
- detect the expected answer type
- select the keywords used in retrieving relevant passages
- decide which question class to use

PowerAnswer 4 Overview



- ranks passages that are retrieved by the IR system

PowerAnswer 4 Overview



- extracts and scores the candidate answers

Question Answering over Blog Data

- 175 GB of blog data
  - with surrounding HTML/XML
- parsed to identify unique content
- language detection to remove non-english text, spam and empty entries
- 13.1 GB of data (92.5% reduction)

Temporal Event Processing

- Concept Tagger Module
  - Detects events in question or candidate passage
  - Labels them

Event Class	Question
Occurrence marry	Who is he planning to marry?
State held	In what city were … held?
Aspectual begin	On what date did the court begin?

Temporal Event Processing

- Concept Tagger Module
  - Detects events in question or candidate passage
  - Labels them
  - Identifies temporal expressions
    - Absolute dates
    - Times
    - Durations

Expression	Question
Absolute Date 2004	What company acquired IMG in 2004?
Duration Three months	In three months following ...
Sets Each year	How many grants … each year?

Temporal Event Processing

- Concept Tagger Module
  - Uses set of rules working on full parse tree of text
  - All temporal expressions normalized

Q249.5: How many grants does the Fulbright Program award each year?

P: The program named after the former Senator J. William Fulbright awards approximately 4,500 new grants annually.

ROSE – a new NER System

- Uses 3-step process:
  - Pass text through pattern based grammar system
  - Pass grammar annotated data through ML system
  - Perform partial matching on the text

Answer Likelihood for Factoid Answer Selection

- Goal
  - Group questions from previous TRECs into classes
  - Build language model for each class on features extracted from Question and Answer

Answer Likelihood for Factoid Answer Selection

- Three methods
  - Generate REGEX-style paraphrases and group them together by paraphrase identifiers
  - Use hierarchical clustering based on
    - Expected answer type
    - Most relevant keywords
    - Named entity types
  - Group by answer type

Answer Likelihood for Factoid Answer Selection

- For questions in classes and correctly judged answers the following features were extracted
  - Stemmed keywords
  - Morphological alternations for keywords
  - Named entity tags

Answer Likelihood for Factoid Answer Selection

- Implemented in two stages:
  - During question processing
  - During answer processing
- Use score of answer likelihood to re-rank candidates
- Best observations with grouping the questions by answer type

List Questions

- Strategy
  - Try to maximize recall by returning as many answers as possible during passage retrieval using
    - Lexico-semantic alternations
    - Relaxing the query to include
      - Target keywords
      - Most relevant keywords from primary question text
  - But how to filter answers?

List Questions

- Strategy
  - Try to maximize recall by returning as many answers as possible during passage retrieval using
    - Lexico-semantic alternations
    - Relaxing the query to include
      - Target keywords
      - Most relevant keywords from primary question text
  - But how to filter answers?
    - Use lists from Wikipedia
    - Integrate COGEX

List Questions

- external data for specialized answer types
  - Bots for Amazon.com, imdb.com if question was in domain of books, songs or movies
  - Bot for Wikipedia
    - Google "I'm feeling lucky" to locate relevant articles

List Questions

- COGEX for list questions
  - Potential candidates from passage retrieval
  - Each candidate is hypothesized to be answer
  - COGEX checks if assertion is entailed by corresponding candidate answer passage
  - Only candidates with entailment score over some threshold are returned as valid answers

"Other" Questions

- The challenge is selection of interesting and novel nuggets from large corpus
  - Definition pattern matching module
  - List of over 200 positive and negative pre-computed patterns

    Extended by
  - Hierarchy of nugget patterns and automatically derived generic answer patterns

"Other" Questions

- Nugget hierarchy based on question classes from previous TREC question sets
  - 35 target classes
    - Animal, actor, musician, literature...
  - Each class is associated with a set of minimal information
    - Person pattern: full name, birth, death, place of birth, residence, occupation, etc.
    - Event: begin time, end time, duration, location, participants, etc.

"Other" Questions

- Example patterns:
  - _nationality _profession _var

    German chancellor Angela Merkel
  - _var (_nationality _profession

    Angela Merkel (Germany's chancellor
  - _nationality _JJ _profession _var

    Germany's first female chancellor Angela Merkel

Results

- Factoid answer selection

Run Tag	Submitter	Accuracy
LymbaPA07	Lymba Corporation	0.706
LCCFerret	Language Computer Corporation	0.494
lsv2007c	Saarland University	0.289

Results

- List questions

Run Tag	Submitter	F-Score
LymbaPA07	Lymba Corporation	0.479
LCCFerret	Language Computer Corporation	0.324
ILQUA1	State University of New York (SUNY) at Albany	0.147

Results

- "other" questions

Run Tag	Submitter	F-Score (beta=3)
FDUQAT16B	Fudan University	0.329
lsv2007c	Saarland University	0.299
QASCU2	Concordia University	0.281
LymbaPA07	Lymba Corporation	0.281
LCCFerret	Language Computer Corporation	0.261

Thank you for your attention.

Any questions?