An Introduction to NUS's Participation to the Definitional Question Answering Task at TREC 2003

Hang Cui
School of Computing, NUS

Hang Cui
School of Computing, NUS

NUS
National University of Singapore

Outline

- Introduction to TREC 2003 Definitional QA Task.

- Technologies in Top Definitional QA Systems.

- Description of Our System.

- Ongoing Work – Soft Patterns.

Outline

- Introduction to TREC 2003 Definitional QA Task.

- Technologies in Top Definitional QA Systems.

- Description of Our System.

- Ongoing Work – Soft Patterns.

TREC QA Task 2003

- Main Task (500)
  - Factoid (413)
  - List (37)
  - Definition (50)
- Passage Task
  - Factoid (same questions as factoid questions in Main task)

QA Main Task

- **One response for each question**
- **Factoid**
  - A single [answer-string, docid] pair or NIL
  - Exact answers needed (same as 2002)
- **List**
  - An unordered, non-empty set of [answer-string, docid] pairs
  - No NIL answers
  - Distinct answers
  - Exact answers needed
- **Definition (Who is … and What is …)**
  - An unordered, non-empty set of [answer-string, docid] pairs
  - No NIL answers
  - Distinct answers

Data Set

- Document Set
  - AQUAINT disk set (1,033,461 documents)
    - New York Times (1998 – 2000)
    - Associated Press (1998 – 2000)
    - Xinhua News Agency (1996 – 2000)
  - XML format – doc_id, doc_type, doc_time, header, body.
- Question Set
  - Sources: MSN Search logs and AOL logs.

Main Task Scoring

- **Final Score = 0.5*factoid_score+0.25*list_score+0.25*definition_score**

- **Factoid_score = # questions answered correctly / total # factoid questions**

- **List_score = mean (F_list)**
  - F_list = (2*IP*IR)/(IP+IR)
  - IR = # instances judged correct & distinct / # final answers
  - IP = # instances judged correct & distinct / # instances returned

- **Definition_score = mean (F_def)**
  - F_def = (26*NP*NR) / (25*NP+NR)
  - NR = # essential nuggets returned / # essential answer nuggets
  - NP = 1,   if length < allowance,
        1-[(length-allowance)/length] ,   otherwise
  - allowance = 100*(#essential+acceptable nuggets returned)
  - length = # non-white-space characters in strings returned

Example of Definition Questions - 1

- Q 1. Who is Gunter Blobel?

- Answer:

Dr. Gunter Blobel won the 1999 Nobel prize for Medicine. In his research, he discovered that proteins carry chemical signals that act as zip codes, helping them find their correct "address" within the cells. Dr. Blobel's discoveries have shed light on some hereditary diseases, such as cystic fibrosis, hyperoxaluria (kidney stones at an early age). This research also laid the basis for using biotechnology to produce drugs like insulin and growth hormone.

Dr. Blobel has conducted his research at Rockefeller University in New York City since 1967. He is a cellular and microbiologist.

Dr.Blobel was born in 1936 in Waltersdorf, Silesia, Germany, which is not part of Poland. He witnessed the bombing of Dresden as a young boy. In the early 1950's he escaped East Germany through Berlin. He earned his medical degree at the University of Tubingen. He earned his PhD at the University of Wisconsin in 1967. In 1980, Dr. Blobel became a U.S. citizen.

Example of Definition Questions - 2

- Q2. What is Goth?

- Answer:

  Goth, from the word Gothic, is a subculture of youths which originated in England in the mid- to late 1970's and spread to the U.S. in the early 1980's. Those who call themselves "Goths" dress in black. They often wear leather armor, cloaks, capes, or long black coats. They will often dye their hair jet-black and adorn themselves with powdered faces, garish black eyeliner, black lipstick, and black nail polish. Goths feel alienated from the mainstream, indulge in dark fantasy, and often think about death. The subculture is associated with the style of music known as Goth or industrial, also known as "death rock". Their music is characterized by minor chords, with lyrics expressing a downbeat worldview, futility, frustration and despair.

Outline

- Introduction to TREC 2003 Definitional QA Task.

- Technologies in Top Definitional QA Systems.

- Description of Our System.

- Ongoing Work – Soft Patterns.

Definitional QA Task at TREC 2003

- 54 runs from 25 participating groups.
- Questions – 50 in total
  - 30 for people (*e.g.* Aaron Copland)
  - 10 for organizations (*e.g.* ETA)
  - 10 for other terms (*e.g.* TB and Quasars)
- Assumption for readers and answers:
  - Average readers who are native speakers.
  - Intend to learn more about the term.
- Manually assessed.

Evaluation Metrics

- F Measure - Nugget Recall (NR) and Nugget Precision (NP).
  - NR is 5 times as important as NP.

- Nuggets given by assessors.
  - *e.g.* Qid 1901: Who is Aaron Copland?

1901 1	vital	American composer
1901 2	vital	musical achievements ballets symphonies
1901 3	vital	born Brooklyn NY 1900
1901 4	okay	son of Jewish immigrant
1901 5	okay	American communist
1901 6	okay	civil rights advocate
1901 8	vital	established home for composers
1901 9	okay	won Oscar for "the Heiress"

Overview of Top Ranked Systems

- Top 5 ranked systems – BBN, NUS, ISI, LCC and Columbia.

- Technologies shared in these systems
  - Corpus statistics.
  - External knowledge.
  - Definition patterns.

- Using patterns is the most distinct characteristic of definitional QA from other retrieval tasks.

Statistical Method

- Statistical ranking – centroid (profile)
  - Assumption – definition sentences are relevant sentences to the target.
  - Centroid words
    - Co-occurrences with the target, or
    - From external knowledge.
  - Ranking
    - Similarity with the centroid, or
    - Boost the weights of centroid words.

External Knowledge

- Two categories
  - General knowledge.
  - Specific knowledge.

- General knowledge
  - General search engines, *e.g.* Google.
  - Online glossaries: WordNet, Merriam-Webster (www.m-w.com).
  - Online encyclopedia: Encarta, Wikipedia (www.wikipedia.com)

- Specific knowledge – *e.g.* www.biographies.com

- No restriction to use any knowledge base, but risky.
  - Well structured web sites or glossaries cannot cover very recent terms, like SARS and Clay Aiken.

Definition Patterns

- Basic syntactic components
  - Appositives, *e.g. Gunter Blobel, a cellular biologist, won 1999 Nobel ….*
  - Copulas, *e.g. Blobel is a professor at Rockefeller University.*
- Predicates (relations)
  - *XXX was born ……*
- Other lexicon-syntactic patterns
  - *XXX can be used to ……*
- All participating systems used hand-crafted patterns.
  - ISI learns part of surface patterns from question-answer pairs (ACL 2002, 2003).

Outline

- Introduction to TREC 2003 Definitional QA Task.

- Technologies in Top Definitional QA Systems.

- **Description of Our System.**

- Ongoing Work – Soft Patterns.

Overview of Our System

- We envision the definitional QA problem as
  - Definition sentences selection (IR) + summarization.

- Pre-processing
  - Document retrieval and sentence splitting.
  - Sentence filtering (a window surrounding the target).
  - Anaphora resolution.

The Pipeline System

Hang Cui
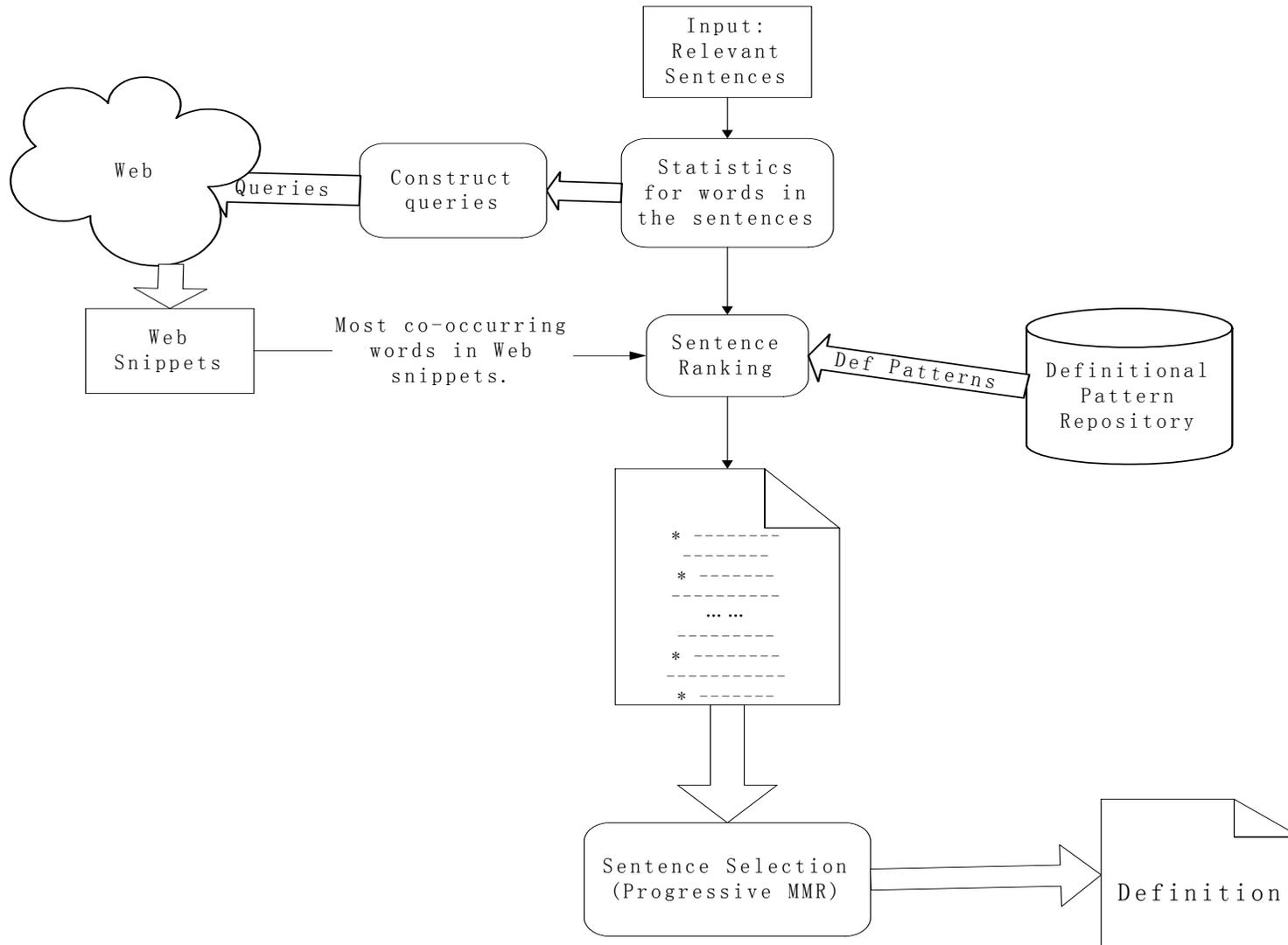
Query Construction

- For each question, up to 100 queries constructed to do web search (Google).
  - Each query contains up to 10 words selected by their sentence frequency with the target.
  - Up to 100 snippets per query retrieved.
  - Weights of those words also appearing in snippets are boosted.

Sentence Ranking

- 3 sources of evidence
  - Corpus statistics
    - Sentence frequency (co-occurrences with the target).
    - Global IDF (from web statistics).
  - Web search
    - Boosting the weights of the words in snippets.
  - Pattern matching
    - Augmenting the sentence weight by the previous two steps if the sentence matches one pattern.

Definition Pattern Matching

- Manually constructed patterns in regular expressions (around 30 in expanded format).
  - *<Sch_Target> , (a|an|the)*
  - *<Sch_Target> (is|are) (called|referred to|…)**
  - *" (.+?) " by <Sch_Target>*
- 11% improvement over statistical method.

Sentence Selection

- Redundancy removing – summarization.
- A variation of MMR (Maximum Marginal Relevance by Carbonell 1998).

  1. All sentences are ordered in descending order by weights.
  2. Add the first sentence to the summary.
  3. Examine the following sentences.

     If Weight(stc)- avg_sim(stc) < Weight(next_stc)

          Continue

     else

          Add stc to summary.

  4. Go to Step 3 till the length limit of the target summary is satisfied.

Outline

- Introduction to TREC 2003 Definitional QA Task.

- Technologies in Top Definitional QA Systems.

- Description of Our System.

- Ongoing Work – Soft Patterns.

The Problem

- Breaking news can provide timely definitions to those recently popular terms.

- Definition patterns are more obscure and more flexible than those in IE task.

  - *e.g. <Sch_Term> (also known as XXX).*

- Two problems in state-of-art systems

  - Coverage - hand-crafted rules cannot cover most of definition patterns.

  - Pattern flexibility - hard matching rules cannot satisfy the flexibility of definition sentences in news.

Soft Matching Patterns

- No generalized rules.
  - Keep instances in each slot.
  - Instance based learning.
- Preprocessing
  - Parsing and chunking.
  - Substitution for centroid words.
- Matching in a probabilistic framework.

1) Definition sentences (bold terms are search terms)

……galaxies, **quasars**, the brightest lights in distant universe ……

…… according to **Nostradamus**, a 16th century French apothecary ……

…… severance packages, known as **golden parachutes**, included ……

A **battery** is a cell which can provide electricity.

……

2) Reduced pattern instances (capitalized tags are chunks and syntactic classes):

NN , <Search_Term> , DT$ NN

according to <Search_Term> , DT$ NNP

known as <Search_Term> , VB

<Search_Term> BE$ DT$

……

3) Soft patterns based on the instances:

…… $\langle Slot_{-2} \rangle$	$\langle Slot_{-1} \rangle$	$\langle Search\_Term \rangle$	$\langle Slot_1 \rangle$	$\langle Slot_2 \rangle$ ……
NN 0.12	, 0.11		, 0.40	DT$ 0.2
according 0.03	to 0.03		BE$ 0.2	VB 0.1
known 0.09	as 0.20			

Probabilistic Matching

- Test sentences are pre-processed the same way as done to training sentences.

- Window size
  - Set to 2 in our experiments.

- Probabilistic matching degree
  - Individual slots' probabilities: Naïve Bayes.
  - Sequential evidence: bi-gram model.
    - Weight right sequence higher.

Unsupervised Labeling Definition Sentences

- To obtain labeled definition sentences
  - Hand labeled.
  - Unsupervised fashion.

- Unsupervised labeling
  - Pseudo-relevance feedback (PRF) from statistical ranking.
    - Group PRF – for 50 questions.
    - 33% top ranked sentences are definition sentences.

Initial Evaluation Results

- Soft patterns outperforms both hand-crafted rules and machine learned generalized rules.

- Outperforms the baseline (statistical method) by 27.20%.

- Outperforms hand-crafted rules (our TREC system) by 14.06% (with unsupervised learning).

- Outperforms generalized rules by a rule induction system (GRID) by 8.36% (with unsupervised learning).

- Seems hard rules are appropriate for IE task but not for definitional QA.

Conclusions

- Introduced definitional QA task at TREC 2003.

- Reviewed top systems at TREC.

- Our system – corpus statistics, web evidence and patterns.

- Extension of our TREC work – soft patterns.

- Future work – summary based QA.

Thanks!

☺