
Automatic Answer Typing for How-Questions

Christopher Pinchak & Shane Bergsma

University of Alberta

HLT-NAACL 2007

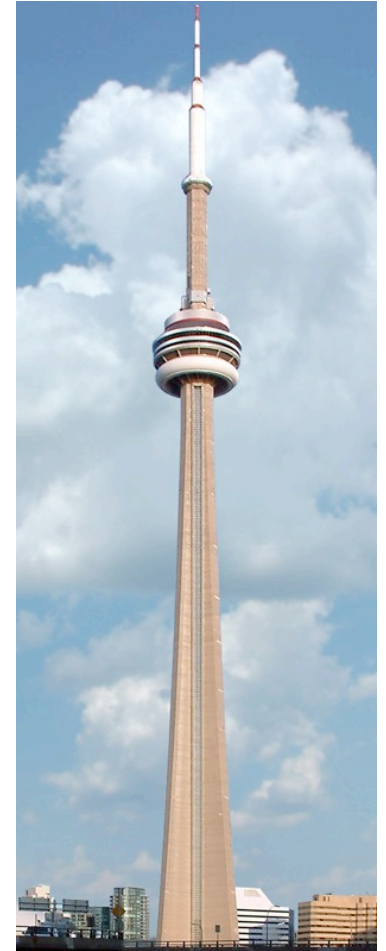


Outline

- Quantifiable How-questions
 - Definition
 - Relative importance
 - Dealing with types
- Our approach to typing
 - Units as types
 - Exploiting the adjective
- Evaluation of our approach

Quantifiable How-questions

- E.g., “How tall is the CN tower?”
- Presence of a how-adjective (“tall”)
- Excludes:
 - How do/does/did, How can/could, etc.
 - E.g., “How do I climb the CN tower?”
 - “How much” and “How many”
- Answers are in terms of units

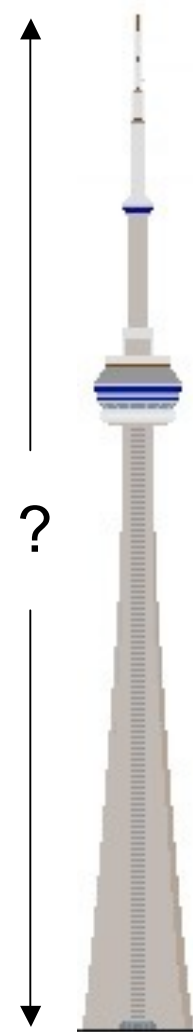


Importance

- How prevalent are quantifiable How-questions?
- AOL query set
 - 36 million queries
 - 11,000 quantifiable how-questions
 - ~1 in every 3,500 queries
- 152 unique adjectives in 11,000 questions
 - Not all are easily typed
 - e.g., “How Orwellian are we?”

Dealing with Types

- Want to find appropriate answers
 - How tall is the CN tower?
 - Blue, Bill Gates, 1,815 ft, Microsoft?
 - 1,815 ft, 200 pounds, 50 years, 5 miles?
- Popular approach is to use the NUM type ?
 - “The CN Tower, located in Toronto, Ontario, Canada, is the world's tallest freestanding structure on land, standing 553.33 meters (1,815 ft 5 in) tall... at 351 metres (1,150 ft) is the 360 Restaurant, which completes a full revolution once every 72 minutes.”



Our approach

- Goals
 - Finer-grained types than just NUM
 - Generate type information off-line from AOL questions
 - Capture types related to kinds of questions
- Key idea: Units form the basis of type information
 - All answers are expressed in terms of one or more units of measure
- Key idea: Exploiting the adjective
 - Units related to the adjective are appropriate
 - Unrelated units and other words are inappropriate
 - Adjective is shared by many different questions
 - “How tall is _____?”

Finding Related Units

- Tall → Height
 - WordNet *attributes* (height) and *derivationally-related terms* (tallness)
 - Good coverage on the quantifiable adjectives
- Use Google to discover units
 - Query pattern: “<*term*> is measured in _____”
 - E.g, “*height* is measured in _____” → feet, inches, ...

Expanding Units Lists

- Limited by Google query maximum and retrieval time (200 results per pattern)
- Pattern may not capture all suitable units
 - Slang *gigs* for *gigabytes*
- Use a database of automatically-generated similar words to expand the units lists
 - E.g., word *gigabytes* could add *GB*, *megabytes*, *kilobytes*, *KB*, *byte*, *GHz*, *gigs*, ...
 - Can pollute a units list, especially with low-similarity words

Filtering the Units List

- Not all units discovered are good
 - Come from parse errors or pattern mismatches
 - E.g., “*height* is measured in our study...”
 - Pollution from similar words lists
- Solution: use a value similar to IDF to discount words appearing on many adjectives’ units lists
 - Assumption is that units on all lists are likely noise
- Impose a threshold to eliminate low-scoring units

Example

- “How tall is the CN tower?”
 - *Tall* has four senses in WordNet:
 1. Attribute: *stature, height*
 2. Related term: *grandiloquence, magniloquence*
 3. Related term: *improbableness*
 4. No attribute or related term (e.g., “a tall order”)
 - “<*term*> is measured in” for <*term*> in {*stature, height, grandiloquence, magniloquence, improbableness*}
 - Unexpanded: Hand, inch, meter, feet, centimeter, metre, physical model, row, conjunction, point, m, ...
 - Expanded: inch, meter, hand, feet, **yard**, centimeter, metre, point, physical model, foot, row, model, ...

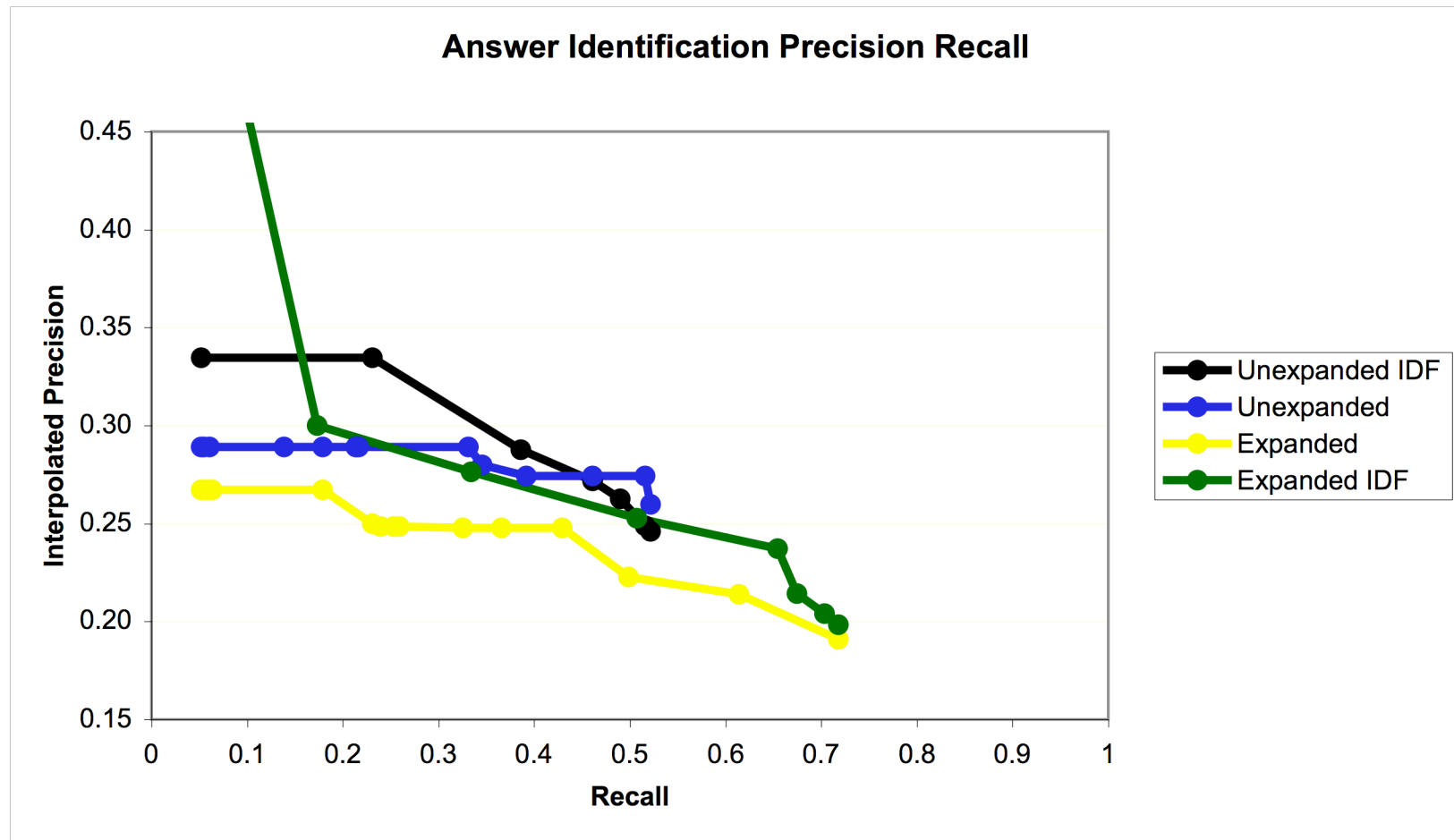
Evaluation

- Evaluation set
 - 86 quantifiable how-questions from TREC 2002-2005
- Evaluation goals
 - Observe the effects of varying threshold values
 - Compare expanded vs unexpanded units lists
 - Compare IDF modification with non-IDF
 - Compare with other approaches
 - Establish upper and lower bounds on performance

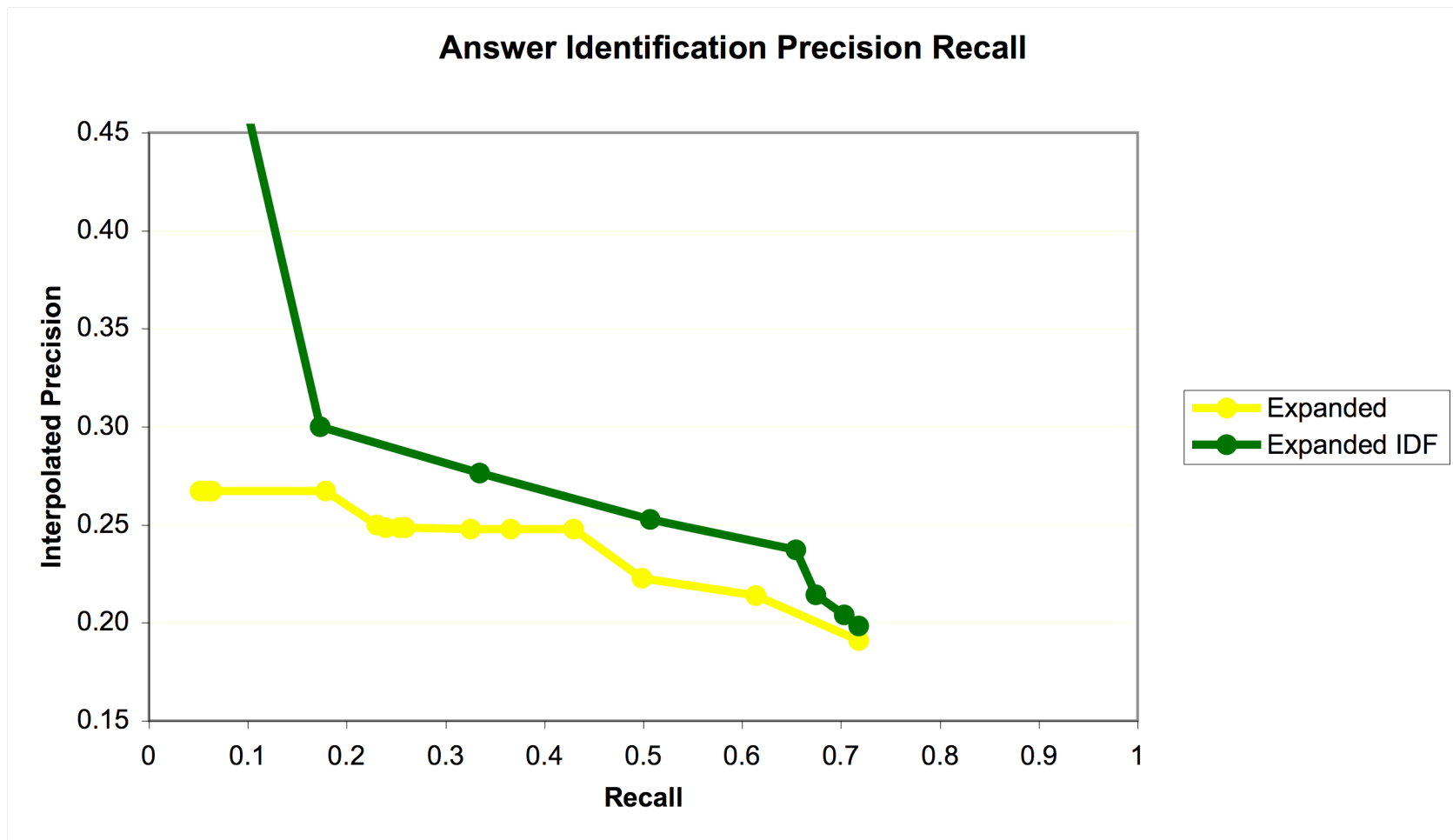
Evaluation Measure

- We will use Answer-Identification Precision Recall
 - Based on a set of correct answers from TREC
 - Propose as answers every numerical entity that matches a unit on our units list
 - Precision: number of proposed answers that are correct
 - Recall: number of correct answers are also proposed by our units list

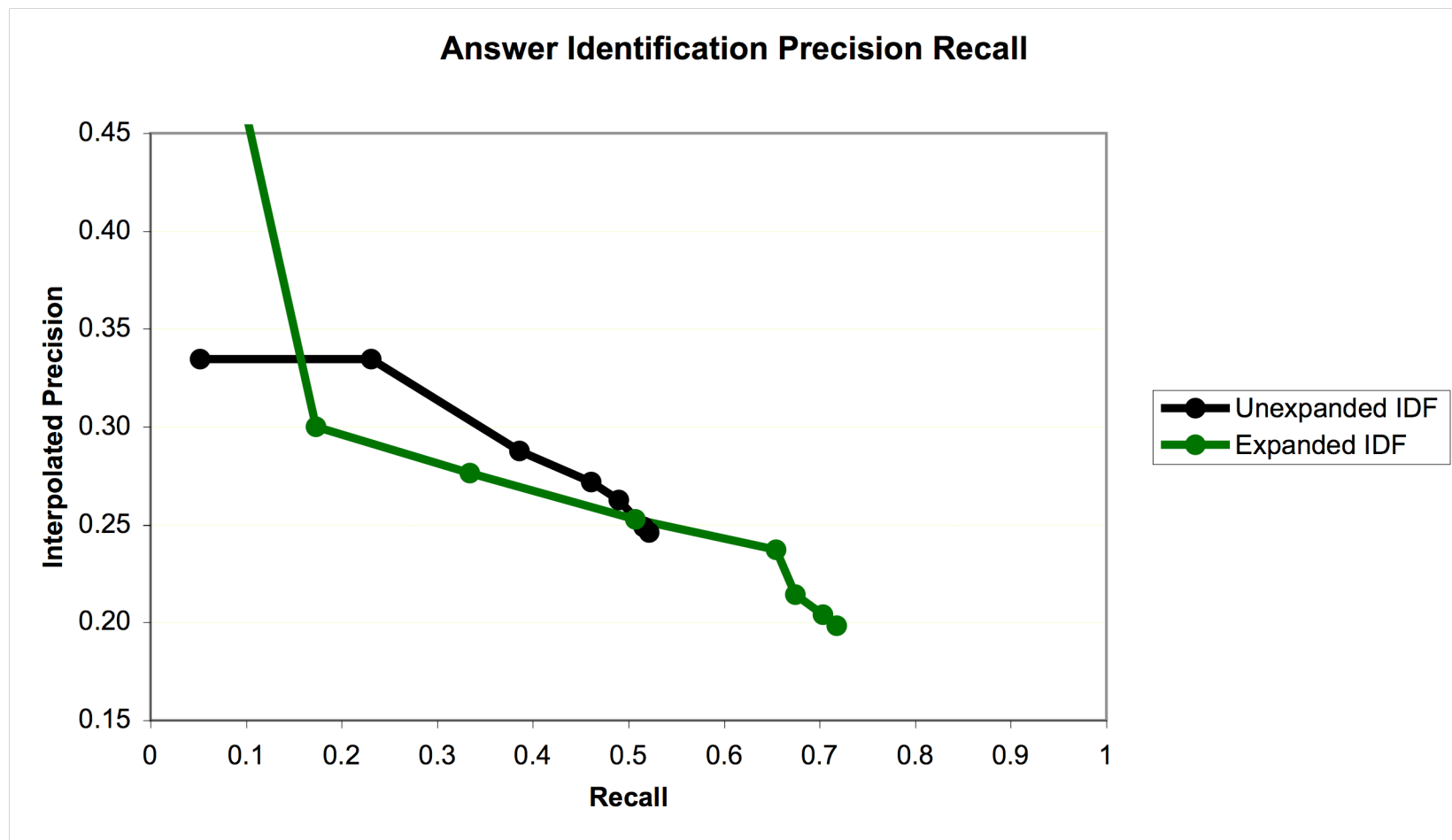
Effects of Varying Threshold



Effects of IDF



Effects of Expansion



Remarks

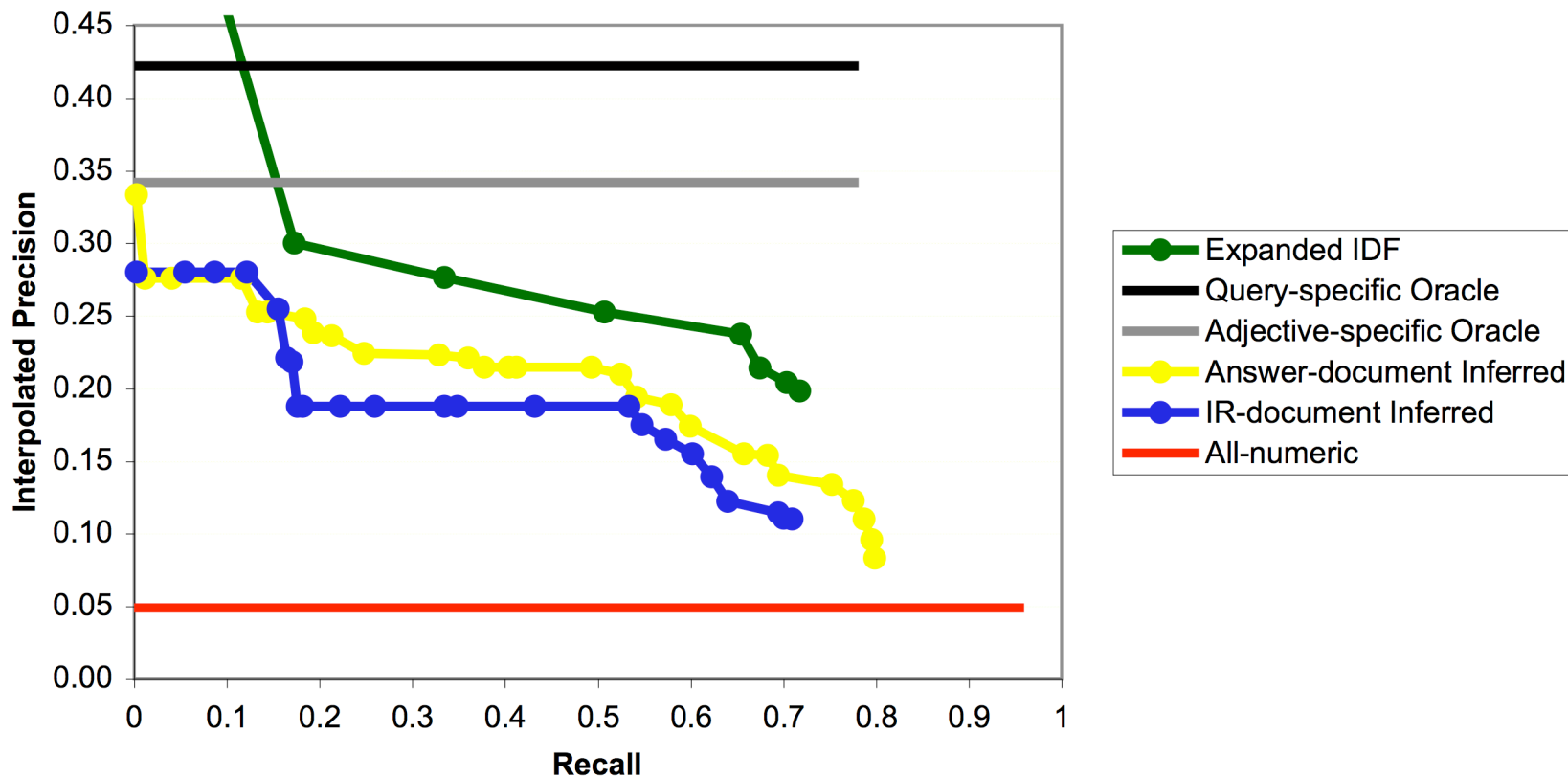
- Expanding units achieves higher possible recall with some loss in precision
 - Adds good undiscovered units
 - Pollutes the units lists with some inappropriate terms
- Would more Google results improve maximum recall?
 - Early experiments showed small improvements when going from 50 to 100 to 200 results
- IDF variant improves performance across the entire range of recall

Comparison Systems

- Query-specific Oracle
 - Units list for a question is the set of units found in correct TREC answers
- Adjective-specific Oracle
 - Units lists for questions sharing the same adjective are merged into one
- All-numeric
 - All numerical entities are placed on the units list
- IR-Document Inferred
 - Extract units from all numerical entities in documents returned by TREC's PRISE/Lucene, scored by frequency
- Answer-Document Inferred
 - Extract units from all numerical entities in documents containing an answer, scored by frequency

Comparison Systems

Answer Identification Precision Recall



Remarks

- Our approach performs worse than the oracles and better than the inferred
 - Oracles are upper bounds
 - Inferred are semi-realistic
 - All-numeric is a lower bound
- Oracles cannot achieve 100% recall due to unit-less answers
 - E.g., “How old was Duke Ellington when he died?” → “... died at age 75”

Conclusions and Future Work

- Quantifiable how-questions benefit from unit-based typing
 - Units are automatically discovered via WordNet and Google
 - Can pre-generate the list of units off-line
 - Performance is superior to fixed-category typing (i.e., All-numeric) and automatic inference of types from a document set (IR and Answer doc inferred)
- Incorporating the answer types into an IR approach could improve IR accuracy
 - E.g., Documents for “How tall is the CN tower?” should contain *meters, feet, etc.*
 - Less than 60% of top-10 IR docs contain a correct unit

Questions? Comments? Suggestions?

Support provided by:



Alberta Ingenuity Fund



Natural Sciences and Engineering
Research Council of Canada



Alberta Informatics Circle of
Research Excellence