Roadmap

- Motivation:
  - Retrieval gaps

- Query Formulation:
  - Question Series
  - Query reformulation:
    - AskMSR patterns
    - MULDER parse-based formulation
  - Classic query expansion
    - Semantic resources
    - Pseudo-relevance feedback
Retrieval Gaps

- Goal:
  - Based on question,
  - Retrieve documents/passages that best capture answer
Retrieval Gaps

- **Goal:**
  - Based on question,
  - Retrieve documents/passages that best capture answer

- **Problem:**
  - Mismatches in lexical choice, sentence structure
Retrieval Gaps

- **Goal:**
  - Based on question,
  - Retrieve documents/passages that best capture answer

- **Problem:**
  - Mismatches in lexical choice, sentence structure
    - Q: How tall is Mt. Everest?
Retrieval Gaps

- **Goal:**
  - Based on question,
  - Retrieve documents/passages that best capture answer

- **Problem:**
  - Mismatches in lexical choice, sentence structure
    - Q: How tall is Mt. Everest?
    - A: The height of Everest is...
Retrieval Gaps

- **Goal:**
  - Based on question,
  - Retrieve documents/passages that best capture answer

- **Problem:**
  - Mismatches in lexical choice, sentence structure
    - Q: How tall is Mt. Everest?
    - A: The height of Everest is...
    - Q: When did the first American president take office?
    - A: George Washington was inaugurated in...
Query Formulation

- Goals:
Query Formulation

• Goals:
  • Overcome lexical gaps & structural differences
  • To enhance basic retrieval matching
  • To improve target sentence identification

• Issues & Approaches:
Query Formulation

• Goals:
  • Overcome lexical gaps & structural differences
  • To enhance basic retrieval matching
  • To improve target sentence identification

• Issues & Approaches:
  • Differences in word forms:
Query Formulation

Goals:
- Overcome lexical gaps & structural differences
- To enhance basic retrieval matching
- To improve target sentence identification

Issues & Approaches:
- Differences in word forms:
  - Morphological analysis
- Differences in lexical choice:
Query Formulation

- Goals:
  - Overcome lexical gaps & structural differences
  - To enhance basic retrieval matching
  - To improve target sentence identification

- Issues & Approaches:
  - Differences in word forms:
    - Morphological analysis
  - Differences in lexical choice:
    - Query expansion
  - Differences in structure
Query Formulation

- Convert question suitable form for IR
- Strategy depends on document collection
  - Web (or similar large collection):
    - ‘stop structure’ removal:
      - Delete function words, q-words, even low content verbs
  - Corporate sites (or similar smaller collection):
    - Query expansion
      - Can’t count on document diversity to recover word variation
      - Add morphological variants, WordNet as thesaurus
      - Reformulate as declarative: rule-based
        - Where is X located -> X is located in
Question Series

- TREC 2003-...
- Target: PERS, ORG, ...
- Assessors create series of questions about target
  - Intended to model interactive Q/A, but often stilted
  - Introduces pronouns, anaphora
Question Series

- TREC 2003-...
- Target: PERS, ORG, ...
- Assessors create series of questions about target
  - Intended to model interactive Q/A, but often stilted
  - Introduces pronouns, anaphora
Handling Question Series

- Given target and series, how deal with reference?
Handling Question Series

- Given target and series, how deal with reference?
- Shallowest approach:
  - Concatenation:
    - Add the ‘target’ to the question
Handling Question Series

- Given target and series, how deal with reference?

- Shallowest approach:
  - Concatenation:
    - Add the ‘target’ to the question

- Shallow approach:
  - Replacement:
    - Replace all pronouns with target
Handling Question Series

- Given target and series, how deal with reference?

- Shallowest approach:
  - Concatenation:
    - Add the ‘target’ to the question

- Shallow approach:
  - Replacement:
    - Replace all pronouns with target

- Least shallow approach:
  - Heuristic reference resolution
Question Series Results

- No clear winning strategy
Question Series Results

- No clear winning strategy
- All largely about the target
  - So no big win for anaphora resolution
  - If using bag-of-words features in search, works fine
Question Series Results

- No clear winning strategy
  - All largely about the target
    - So no big win for anaphora resolution
    - If using bag-of-words features in search, works fine

- ‘Replacement’ strategy can be problematic
  - E.g. Target=Nirvana:
    - What is their biggest hit?
Question Series Results

- No clear winning strategy
  - All largely about the target
    - So no big win for anaphora resolution
    - If using bag-of-words features in search, works fine

- ‘Replacement’ strategy can be problematic
  - E.g. Target=Nirvana:
    - What is their biggest hit?
    - When was the band formed?
Question Series Results

- No clear winning strategy
  - All largely about the target
    - So no big win for anaphora resolution
    - If using bag-of-words features in search, works fine

- ‘Replacement’ strategy can be problematic
  - E.g. Target=Nirvana:
    - What is their biggest hit?
    - When was the band formed?
      - Wouldn’t replace ‘the band’
Question Series Results

- No clear winning strategy
  - All largely about the target
    - So no big win for anaphora resolution
    - If using bag-of-words features in search, works fine

- ‘Replacement’ strategy can be problematic
  - E.g. Target=Nirvana:
    - What is their biggest hit?
    - When was the band formed?
      - Wouldn’t replace ‘the band’

- Most teams concatenate
AskMSR

- Shallow Processing for QA
  - (Dumais et al 2002, Lin2007)

Diagram:
1. Question: Where is the Louvre Museum located?
2. Rewrite Query: “+the Louvre Museum +is located”
3. <Search Engine>
4. Collect Summaries, Mine N-grams
5. N-Best Answers
6. Tile N-Grams
7. Filter N-Grams
Intuition

- Redundancy is useful!
  - If similar strings appear in many candidate answers, likely to be solution
    - Even if can’t find obvious answer strings
Intuition

- Redundancy is useful!
  - If similar strings appear in many candidate answers, likely to be solution
    - Even if can’t find obvious answer strings

- Q: How many times did Bjorn Borg win Wimbledon?
  - Bjorn Borg blah blah blah Wimbledon blah 5 blah
  - Wimbledon blah blah blah Bjorn Borg blah 37 blah.
  - blah Bjorn Borg blah blah 5 blah blah Wimbledon
  - 5 blah blah Wimbledon blah blah Bjorn Borg.
Intuition

- Redundancy is useful!
  - If similar strings appear in many candidate answers, likely to be solution
    - Even if can’t find obvious answer strings

- Q: How many times did Bjorn Borg win Wimbledon?
  - Bjorn Borg blah blah blah Wimbledon blah 5 blah
  - Wimbledon blah blah blah Bjorn Borg blah 37 blah.
  - blah Bjorn Borg blah blah 5 blah blah Wimbledon
  - 5 blah blah Wimbledon blah blah Bjorn Borg.
- Probably 5
Query Reformulation

- Identify question type:
  - E.g. Who, When, Where,...

- Create question-type specific rewrite rules:
Query Reformulation

- Identify question type:
  - E.g. Who, When, Where,...

- Create question-type specific rewrite rules:
  - Hypothesis: Wording of question similar to answer
    - For ‘where’ queries, move ‘is’ to all possible positions
      - Where is the Louvre Museum located? =>
        - Is the Louvre Museum located
        - The is Louvre Museum located
        - The Louvre Museum is located, .etc.
Query Reformulation

- Identify question type:
  - E.g. Who, When, Where,…

- Create question-type specific rewrite rules:
  - Hypothesis: Wording of question similar to answer
    - For ‘where’ queries, move ‘is’ to all possible positions
      - Where is the Louvre Museum located? =>
        - Is the Louvre Museum located
        - The is Louvre Museum located
        - The Louvre Museum is located, .etc.

- Create type-specific answer type (Person, Date, Loc)
Query Form Generation

- 3 query forms:
  - Initial baseline query
Query Form Generation

- 3 query forms:
  - Initial baseline query
  - Exact reformulation: weighted 5 times higher
    - Attempts to anticipate location of answer
Query Form Generation

- 3 query forms:
  - Initial baseline query
  - Exact reformulation: weighted 5 times higher
    - Attempts to anticipate location of answer
    - Extract using surface patterns
      - “When was the telephone invented?”
Query Form Generation

- 3 query forms:
  - Initial baseline query
  - Exact reformulation: weighted 5 times higher
    - Attempts to anticipate location of answer
    - Extract using surface patterns
      - “When was the telephone invented?”
      - “the telephone was invented ?x”
Query Form Generation

3 query forms:
- Initial baseline query
- Exact reformulation: weighted 5 times higher
  - Attempts to anticipate location of answer
  - Extract using surface patterns
    - “When was the telephone invented?”
    - “the telephone was invented ?x”
- Generated by ~12 pattern matching rules on terms, POS
  - E.g. wh-word did A verb B -
Query Form Generation

• 3 query forms:
  • Initial baseline query
  • Exact reformulation: weighted 5 times higher
    • Attempts to anticipate location of answer
    • Extract using surface patterns
      • “When was the telephone invented?”
      • “the telephone was invented ?x”
  • Generated by ~12 pattern matching rules on terms, POS
    • E.g. wh-word did A verb B -> A verb+ed B ?x (general)
    • Where is A? -

Query Form Generation

- 3 query forms:
  - Initial baseline query
  - Exact reformulation: weighted 5 times higher
  - Attempts to anticipate location of answer
  - Extract using surface patterns
    - “When was the telephone invented?”
    - “the telephone was invented ?x”
  - Generated by ~12 pattern matching rules on terms, POS
    - E.g. wh-word did A verb B -> A verb+ed B ?x (general)
    - Where is A? -> A is located in ?x (specific)
- Inexact reformulation: bag-of-words
## Query Reformulation

### Examples

**What year did Alaska become a state?**

<table>
<thead>
<tr>
<th>baseline</th>
<th>What year did Alaska become a state</th>
</tr>
</thead>
<tbody>
<tr>
<td>inexact</td>
<td>Alaska became a state</td>
</tr>
<tr>
<td>exact</td>
<td>Alaska became a state ?x</td>
</tr>
</tbody>
</table>

**Who was the first person to run the mile in less than four minutes?**

<table>
<thead>
<tr>
<th>baseline</th>
<th>Who was the first person to run the mile in less than four minutes?</th>
</tr>
</thead>
<tbody>
<tr>
<td>inexact</td>
<td>the first person to run the mile in less than four minutes</td>
</tr>
<tr>
<td>exact</td>
<td>the first person to run the mile in less than four minutes was ?x</td>
</tr>
<tr>
<td>exact</td>
<td>?x was the first person to run the mile in less than four minutes</td>
</tr>
</tbody>
</table>
Deeper Processing for Query Formulation

- MULDER (Kwok, Etzioni, & Weld)
- Converts question to multiple search queries
  - Forms which match target
  - Vary specificity of query
    - Most general bag of keywords
    - Most specific partial/full phrases
Deeper Processing for Query Formulation

- MULDER (Kwok, Etzioni, & Weld)
- Converts question to multiple search queries
  - Forms which match target
  - Vary specificity of query
    - Most general bag of keywords
    - Most specific partial/full phrases
- Employs full parsing augmented with morphology
Syntax for Query Formulation

- Parse-based transformations:
  - Applies transformational grammar rules to questions
Syntax for Query Formulation

- Parse-based transformations:
  - Applies transformational grammar rules to questions
  - Example rules:
    - Subject-auxiliary movement:
      - Q: Who was the first American in space?
Syntax for Query Formulation

- Parse-based transformations:
  - Applies transformational grammar rules to questions
  - Example rules:
    - Subject-auxiliary movement:
      - Q: Who was the first American in space?
      - Alt: was the first American...; the first American in space was
    - Subject-verb movement:
      - Who shot JFK?
Syntax for Query Formulation

- Parse-based transformations:
  - Applies transformational grammar rules to questions
  - Example rules:
    - Subject-auxiliary movement:
      - Q: Who was the first American in space?
      - Alt: was the first American...; the first American in space was
    - Subject-verb movement:
      - Who shot JFK? => shot JFK
    - Etc

- Morphology based transformation:
  - Verb-conversion: do-aux+v-inf
Syntax for Query Formulation

- **Parse-based transformations:**
  - Applies transformational grammar rules to questions
  - Example rules:
    - **Subject-auxiliary movement:**
      - Q: Who was the first American in space?
      - Alt: was the first American...; the first American in space was
    - **Subject-verb movement:**
      - Who shot JFK? => shot JFK
    - Etc

- **Morphology based transformation:**
  - Verb-conversion: do-aux+v-inf => conjugated verb
Machine Learning Approaches

- Diverse approaches:
  - Assume annotated query logs, annotated question sets, matched query/snippet pairs
Machine Learning Approaches

• Diverse approaches:
  • Assume annotated query logs, annotated question sets, matched query/snippet pairs
  • Learn question paraphrases (MSRA)
    • Improve QA by setting question sites
    • Improve search by generating alternate question forms
Machine Learning Approaches

- Diverse approaches:
  - Assume annotated query logs, annotated question sets, matched query/snippet pairs
  - Learn question paraphrases (MSRA)
    - Improve QA by setting question sites
    - Improve search by generating alternate question forms
  - Question reformulation as machine translation
    - Given question logs, click-through snippets
      - Train machine learning model to transform Q -> A
Query Expansion

- Basic idea:
  - Improve matching by adding words with similar meaning/similar topic to query
Query Expansion

- Basic idea:
  - Improve matching by adding words with similar meaning/similar topic to query

- Alternative strategies:
  - Use fixed lexical resource
    - E.g. WordNet
Query Expansion

- Basic idea:
  - Improve matching by adding words with similar meaning/similar topic to query

- Alternative strategies:
  - Use fixed lexical resource
    - E.g. WordNet
  - Use information from document collection
    - Pseudo-relevance feedback
WordNet Based Expansion

- In Information Retrieval settings, mixed history
- Helped, hurt, or no effect
- With long queries & long documents, no/bad effect
WordNet Based Expansion

- In Information Retrieval settings, mixed history
  - Helped, hurt, or no effect
  - With long queries & long documents, no/bad effect

- Some recent positive results on short queries
  - E.g. Fang 2008
  - Contrasts different WordNet, Thesaurus similarity
  - Add semantically similar terms to query
    - Additional weight factor based on similarity score
Similarity Measures

- Definition similarity: $S_{\text{def}}(t_1, t_2)$
  - Word overlap between glosses of all synsets
  - Divided by total numbers of words in all synsets glosses
Similarity Measures

- Definition similarity: $S_{\text{def}}(t_1, t_2)$
  - Word overlap between glosses of all synsets
    - Divided by total numbers of words in all synsets glosses

- Relation similarity:
  - Get value if terms are:
    - Synonyms, hypernyms, hyponyms, holonyms, or meronyms
Definition similarity: $S_{\text{def}}(t_1, t_2)$
- Word overlap between glosses of all synsets
  - Divided by total numbers of words in all synsets glosses

Relation similarity:
- Get value if terms are:
  - Synonyms, hypernyms, hyponyms, holonyms, or meronyms

Term similarity score from Lin’s thesaurus
Results

- Definition similarity yields significant improvements
  - Allows matching across POS
  - More fine-grained weighting than binary relations
Deliverable #2 Discussion

- More training data available
- Test data released
- Requirements
- Deliverable Reports