Ling573
NLP Systems and Applications
May 7, 2013

Local Context and SMT for Question Expansion

- "Statistical Machine Translation for Query Expansion in Answer Retrieval", Riezler et al, 2007
- Investigates data-driven approaches to query exp.

Local Context and SMT for Question Expansion

- “Statistical Machine Translation for Query Expansion in Answer Retrieval”, Riezler et al, 2007

- Investigates data-driven approaches to query exp.
  - Local context analysis (pseudo-rel. feedback)

Local Context and SMT for Question Expansion

- "Statistical Machine Translation for Query Expansion in Answer Retrieval", Riezler et al, 2007

- Investigates data-driven approaches to query exp.
  - Local context analysis (pseudo-rel. feedback)
  - Contrasts: Collection global measures
    - Terms identified by statistical machine translation
    - Terms identified by automatic paraphrasing

Motivation

- Fundamental challenge in QA (and IR)
  - Bridging the “lexical chasm”

Motivation

- Fundamental challenge in QA (and IR)
  - Bridging the "lexical chasm"
    - Divide between user's info need, author's lexical choice
    - Result of linguistic ambiguity
- Many approaches:
  - QA

Motivation

- Fundamental challenge in QA (and IR)
  - Bridging the “lexical chasm”
    - Divide between user’s info need, author’s lexical choice
    - Result of linguistic ambiguity
- Many approaches:
  - QA
    - Question reformulation, syntactic rewriting
    - Ontology-based expansion
    - MT-based reranking
  - IR:

Motivation

- Fundamental challenge in QA (and IR)
  - Bridging the "lexical chasm"
    - Divide between user's info need, author's lexical choice
    - Result of linguistic ambiguity
- Many approaches:
  - QA
    - Question reformulation, syntactic rewriting
    - Ontology-based expansion
    - MT-based reranking
  - IR: query expansion with pseudo-relevance feedback

Task & Approach

- Goal:
  - Answer retrieval from FAQ pages
    - IR problem: matching queries to docs of Q-A pairs
    - QA problem: finding answers in restricted document set

Task & Approach

- Goal:
  - Answer retrieval from FAQ pages
    - IR problem: matching queries to docs of Q-A pairs
    - QA problem: finding answers in restricted document set
- Approach:
  - Bridge lexical gap with statistical machine translation
  - Perform query expansion
    - Expansion terms identified via phrase-based MT

Challenge

- Varied results for query expansion
- IR:
  - External resources:

Challenge

- Varied results for query expansion
- IR:
  - External resources:
    - Inconclusive (Vorhees), some gain (Fang)

Challenge

- Varied results for query expansion
- IR:
  - External resources:
    - Inconclusive (Vorhees), some gain (Fang)
  - Local, global collection statistics: Substantial gains
- QA:

Challenge

- Varied results for query expansion
- IR:
  - External resources:
    - Inconclusive (Vorhees), some gain (Fang)
  - Local, global collection statistics: Substantial gains
- QA:
  - External resources: mixed results
  - Local, global collection stats: sizable gains

Creating the FAQ Corpus

- Prior FAQ collections limited in scope, quality

Creating the FAQ Corpus

- Prior FAQ collections limited in scope, quality
  - Web search and scraping ‘FAQ’ in title/url
  - Search in proprietary collections

Creating the FAQ Corpus

- Prior FAQ collections limited in scope, quality
  - Web search and scraping ‘FAQ’ in title/url
  - Search in proprietary collections
  - 1-2.8M Q-A pairs
    - Inspection shows poor quality
- Extracted from 4B page corpus

Creating the FAQ Corpus

- Prior FAQ collections limited in scope, quality
  - Web search and scraping 'FAQ' in title/url
  - Search in proprietary collections
  - 1-2.8M Q-A pairs
    - Inspection shows poor quality
- Extracted from 4B page corpus (they're Google)
  - Precision-oriented extraction

Creating the FAQ Corpus

- Prior FAQ collections limited in scope, quality
  - Web search and scraping 'FAQ' in title/url
  - Search in proprietary collections
  - 1-2.8M Q-A pairs
    - Inspection shows poor quality
- Extracted from 4B page corpus (they're Google)
  - Precision-oriented extraction
    - Search for 'faq', Train FAQ page classifier ➔ ~800K pages

Creating the FAQ Corpus

- Prior FAQ collections limited in scope, quality
  - Web search and scraping 'FAQ' in title/url
  - Search in proprietary collections
  - 1-2.8M Q-A pairs
    - Inspection shows poor quality
- Extracted from 4B page corpus (they're Google)
  - Precision-oriented extraction
    - Search for 'faq', Train FAQ page classifier ➔ ~800K pages
    - Q-A pairs: trained labeler: features?

Creating the FAQ Corpus

- Prior FAQ collections limited in scope, quality
  - Web search and scraping 'FAQ' in title/url
  - Search in proprietary collections
  - 1-2.8M Q-A pairs
    - Inspection shows poor quality
- Extracted from 4B page corpus (they're Google)
  - Precision-oriented extraction
    - Search for 'faq', Train FAQ page classifier ➔ ~800K pages
    - Q-A pairs: trained labeler: features?
      - punctuation, HTML tags (<p>,..), markers (Q:), lexical (what,how)
      - ➔ 10M pairs (98% precision)

Machine Translation Model

- SMT query expansion:
  - Builds on alignments from SMT models

Machine Translation Model

- SMT query expansion:
  - Builds on alignments from SMT models
- Basic noisy channel machine translation model:
  - e: English; f: French $\arg\max_e p(e \mid f) = \arg\max_e p(f \mid e)p(e)$

Machine Translation Model

- SMT query expansion:
  - Builds on alignments from SMT models
- Basic noisy channel machine translation model:
  - e: English; f: French $\arg\max_e p(e \mid f) = \arg\max_e p(f \mid e) p(e)$
  - p(e): 'language model'; p(f|e): translation model

Machine Translation Model

- SMT query expansion:
  - Builds on alignments from SMT models
- Basic noisy channel machine translation model:
  - e: English; f: French $\arg\max_e p(e \mid f) = \arg\max_e p(f \mid e)p(e)$
  - p(e): 'language model'; p(f|e): translation model
    - Calculated from relative frequencies of phrases
      - Phrases: larger blocks of aligned words

Machine Translation Model

- SMT query expansion:
  - Builds on alignments from SMT models
- Basic noisy channel machine translation model:
  - e: English; f: French $\arg\max_e p(e \mid f) = \arg\max_e p(f \mid e) p(e)$
  - p(e): 'language model'; p(f|e): translation model
    - Calculated from relative frequencies of phrases
      - Phrases: larger blocks of aligned words
  - Sequence of phrases:

Machine Translation Model

- SMT query expansion:
  - Builds on alignments from SMT models
- Basic noisy channel machine translation model:
  - e: English; f: French $\arg\max_e p(e \mid f) = \arg\max_e p(f \mid e) p(e)$
  - p(e): 'language model'; p(f|e): translation model
    - Calculated from relative frequencies of phrases
      - Phrases: larger blocks of aligned words
  - Sequence of phrases:

$$p(f_i^I \mid e_i^I) = \prod_{i=1}^{I} p(f_i \mid e_i)$$

Question-Answer Translation

- View Q-A pairs from FAQ as translation pairs
  - Q as translation of A (and vice versa)

Question-Answer Translation

- View Q-A pairs from FAQ as translation pairs
  - Q as translation of A (and vice versa)
- Goal:
  - Learn alignments b/t question words & synonymous answer words

Question-Answer Translation

- View Q-A pairs from FAQ as translation pairs
  - Q as translation of A (and vice versa)
- Goal:
  - Learn alignments b/t question words & synonymous answer words
    - Not interested in fluency, ignore that part of MT model
- Issues: Differences from typical MT

Question-Answer Translation

- View Q-A pairs from FAQ as translation pairs
  - Q as translation of A (and vice versa)
- Goal:
  - Learn alignments b/t question words & synonymous answer words
    - Not interested in fluency, ignore that part of MT model
- Issues: Differences from typical MT
  - Length differences

Question-Answer Translation

- View Q-A pairs from FAQ as translation pairs
  - Q as translation of A (and vice versa)
- Goal:
  - Learn alignments b/t question words & synonymous answer words
    - Not interested in fluency, ignore that part of MT model
- Issues: Differences from typical MT
  - Length differences → Modify null alignment weights
  - Less important words

Question-Answer Translation

- View Q-A pairs from FAQ as translation pairs
  - Q as translation of A (and vice versa)
- Goal:
  - Learn alignments b/t question words & synonymous answer words
    - Not interested in fluency, ignore that part of MT model
- Issues: Differences from typical MT
  - Length differences ➔ Modify null alignment weights
  - Less important words ➔ Use intersection of bidirectional alignments

Example

- Q: “How to live with cat allergies”
- Add expansion terms
  - Translations not seen in original query

(how, how) (to, to) (live, live) (with, with) (cat, **pet**) (allergies, allergies)
(how, how) (to, to) (live, live) (with, with) (cat, cat) (allergies, **allergy**)
(how, how) (to, to) (live, live) (with, with) (cat, cat) (allergies, **food**)
(how, how) (to, to) (live, live) (with, with) (cat, **cats**) (allergies, allergies)

SMT-based Paraphrasing

- Key approach intuition:
  - Identify paraphrases by translating to and from a 'pivot' language

SMT-based Paraphrasing

- Key approach intuition:
  - Identify paraphrases by translating to and from a 'pivot' language
  - Paraphrase rewrites yield phrasal 'synonyms'
    - E.g. translate E -> C -> E: find E phrases aligned to C

SMT-based Paraphrasing

- Key approach intuition:
  - Identify paraphrases by translating to and from a 'pivot' language
  - Paraphrase rewrites yield phrasal 'synonyms'
    - E.g. translate E -> C -> E: find E phrases aligned to C
- Given paraphrase pair (trg, syn): pick best pivot
-

SMT-based Paraphrasing

- Key approach intuition:
  - Identify paraphrases by translating to and from a 'pivot' language
  - Paraphrase rewrites yield phrasal 'synonyms'
    - E.g. translate E -> C -> E: find E phrases aligned to C
- Given paraphrase pair (trg, syn): pick best pivot
- $$p(syn \mid trg) = \max_{src} p(src \mid trg) p(syn \mid src)$$
  $$p(trg \mid syn) = \max_{src} p(src \mid syn) p(trg \mid src)$$

SMT-based Paraphrasing

- Features employed:
  - Phrase translation probabilities, lexical translation probabilities, reordering score, # words, # phrases, LM

SMT-based Paraphrasing

- Features employed:
  - Phrase translation probabilities, lexical translation probabilities, reordering score, # words, # phrases, LM
- Trained on NIST multiple Chinese-English translations
-

SMT-based Paraphrasing

- Features employed:
  - Phrase translation probabilities, lexical translation probabilities, reordering score, # words, # phrases, LM
- Trained on NIST multiple Chinese-English translations
- $$p(syn_1^I \mid trg_1^I) = \left(\prod_{i=1}^{I} p_\phi(syn_i \mid trg_i)^{\lambda_\phi} \right.$$
  $$\times p_{\phi'}(trg_i \mid syn_i)^{\lambda_{\phi'}} \times p_w(syn_i \mid trg_i)^{\lambda_w}$$
  $$\left. \times p_{w'}(trg_i \mid syn_i)^{\lambda_{w'}} \times p_d(syn_i, trg_i)^{\lambda_d} \right)$$
  $$\times l_w(syn_1^I)^{\lambda_l} \times c_\phi(syn_1^I)^{\lambda_c} \times p_{LM}(syn_1^I)^{\lambda_{LM}}$$

Example

- Q: “How to live with cat allergies”
- Expansion approach:
  - Add new terms from n-best paraphrases

(how, how) (to live, to live) (with cat, with cat) (allergies, **allergy**)
(how, **ways**) (to live, to live) (with cat, with cat) (allergies, allergies)
(how, how) (to live with, to live with) (cat, **feline**) (allergies, allergies)
(how to, how to) (live, **living**) (with cat, with cat) (allergies, allergies)
(how to, how to) (live, **life**) (with cat, with cat) (allergies, allergies)
(how, **way**) (to live, to live) (with cat, with cat) (allergies, allergies)
(how, how) (to live, to live) (with cat, with cat) (allergies, **allergens**)
(how, how) (to live, to live) (with cat, with cat) (allergies, **allergen**)

Retrieval Model

- Weighted linear combination of vector similarity vals
  - Computed between query and fields of Q-A pair

Retrieval Model

- Weighted linear combination of vector similarity vals
  - Computed between query and fields of Q-A pair
- 8 Q-A pair fields:
  - 1) Full FAQ text; 2) Question text; 3) answer text;
  - 4) title text; 5) 1-4 without stopwords

Retrieval Model

- Weighted linear combination of vector similarity vals
  - Computed between query and fields of Q-A pair
- 8 Q-A pair fields:
  - 1) Full FAQ text; 2) Question text; 3) answer text;
  - 4) title text; 5) 1-4 without stopwords
  - Highest weights:

Retrieval Model

- Weighted linear combination of vector similarity vals
  - Computed between query and fields of Q-A pair
- 8 Q-A pair fields:
  - 1) Full FAQ text; 2) Question text; 3) answer text;
  - 4) title text; 5) 1-4 without stopwords
  - Highest weights: Raw Q text;
    - Then stopped full text, stopped Q text
    - Then stopped A text, stopped title text
  - No phrase matching or stemming

Query Expansion

- SMT Term selection:
  - New terms from 50-best paraphrases
    - 7.8 terms added

Query Expansion

- SMT Term selection:
  - New terms from 50-best paraphrases
    - 7.8 terms added
  - New terms from 20-best translations
    - 3.1 terms added
  - Why?

Query Expansion

- SMT Term selection:
  - New terms from 50-best paraphrases
    - 7.8 terms added
  - New terms from 20-best translations
    - 3.1 terms added
    - Why? - paraphrasing more constrained, less noisy
- Weighting: Paraphrase: same; Trans: higher A text

Query Expansion

- SMT Term selection:
  - New terms from 50-best paraphrases
    - 7.8 terms added
  - New terms from 20-best translations
    - 3.1 terms added
    - Why? - paraphrasing more constrained, less noisy
- Weighting: Paraphrase: same; Trans: higher A text
- Local expansion (Xu and Croft)
  - top 20 docs, terms weighted by tfidf of answers

Query Expansion

- SMT Term selection:
  - New terms from 50-best paraphrases
    - 7.8 terms added
  - New terms from 20-best translations
    - 3.1 terms added
    - Why? - paraphrasing more constrained, less noisy
- Weighting: Paraphrase: same; Trans: higher A text
- Local expansion (Xu and Croft)
  - top 20 docs, terms weighted by tfidf of answers
    - Use answer preference weighting for retrieval
    - 9.25 terms added

Experiments

- Test queries from MetaCrawler query logs
  - 60 well-formed NL questions

Experiments

- Test queries from MetaCrawler query logs
  - 60 well-formed NL questions
- Issue: Systems fail on 1/3 of questions
  - No relevant answers retrieved
    - E.g. "how do you make a cornhusk doll?", "what does 8x certification mean", etc
  - Serious recall problem in QA DB

Experiments

- Test queries from MetaCrawler query logs
  - 60 well-formed NL questions
- Issue: Systems fail on 1/3 of questions
  - No relevant answers retrieved
    - E.g. "how do you make a cornhusk doll?", "what does 8x certification mean", etc
  - Serious recall problem in QA DB
- Retrieve 20 results:
  - Compute evaluation measures @10, 20

Experiments

- Test queries from MetaCrawler query logs
  - 60 well-formed NL questions
- Issue: Systems fail on 1/3 of questions
  - No relevant answers retrieved
    - E.g. "how do you make a cornhusk doll?", "what does 8x certification mean", etc
  - Serious recall problem in QA DB
- Retrieve 20 results:
  - Compute evaluation measures @10, 20

Evaluation

- Manually label top 20 answers by 2 judges

Evaluation

- Manually label top 20 answers by 2 judges
- Quality rating: 3 point scale
  - adequate (2): Includes the answer
  - material (1): Some relevant information, no exact ans
  - unsatisfactory (0): No relevant info

Evaluation

- Manually label top 20 answers by 2 judges
- Quality rating: 3 point scale
  - adequate (2): Includes the answer
  - material (1): Some relevant information, no exact ans
  - unsatisfactory (0): No relevant info
- Compute 'Success$_{type}$ @ n'
  - Type: 2,1,0 above
  - n: # of documents returned
- Why not MRR?

Evaluation

- Manually label top 20 answers by 2 judges
- Quality rating: 3 point scale
  - adequate (2): Includes the answer
  - material (1): Some relevant information, no exact ans
  - unsatisfactory (0): No relevant info
- Compute 'Success$_{type}$ @ n'
  - Type: 2,1,0 above
  - n: # of documents returned
- Why not MRR? - Reduce sensitivity to high rank
  - Reward recall improvement

Evaluation

- Manually label top 20 answers by 2 judges
- Quality rating: 3 point scale
  - adequate (2): Includes the answer
  - material (1): Some relevant information, no exact ans
  - unsatisfactory (0): No relevant info
- Compute 'Success$_{type}$ @ n'
  - Type: 2,1,0 above
  - n: # of documents returned
- Why not MRR? - Reduce sensitivity to high rank
  - Reward recall improvement
  - MRR rewards systems with answers in top 1, but poorly on everything else

Results

	S_2@10	S_2@20	$S_{1,2}$@10	$S_{1,2}$@20
baseline *tfidf*	27	35	58	65
local expansion	30 (+ 11.1)	40 (+ 14.2)	57 (- 1)	63 (- 3)
SMT-based expansion	38 (+ 40.7)	43 (+ 22.8)	58	65

Discussion

- Compare baseline retrieval to:
  - Local RF expansion
  - Combined translation, paraphrase based expansion

Discussion

- Compare baseline retrieval to:
  - Local RF expansion
  - Combined translation, paraphrase based expansion

- Both forms of query expansion improve baseline
  - 2 @10: Local: +11%; SMT: +40.7%
  - 2,1 (easier task): little change

Example Expansions

	how to live with cat allergies
-	allergens allergic infections filter plasmacluster rhinitis introduction effective replacement
+	allergy cats pet food
+	way allergens life allergy feline ways living allergen
	how to design model rockets
-	models represented orientation drawings analysis element environment different structure
+	models rocket
+	missiles missile rocket grenades arrow designing prototype models ways paradigm
	what is dna hybridization
-	instructions individual blueprint characteristics chromosomes deoxyribonucleic information genetic molecule
+	slides clone cdna sitting sequences
+	hibridization hybrids hybridation anything hibridacion hybridising adn hybridisation nothing
	how to enhance competitiveness of indian industries
+	resources production quality processing established investment development facilities institut
+	increase industry
+	promote raise improve increase industry strengthen
	how to induce labour
-	experience induction practice imagination concentration information consciousness different
-	relaxation
-	birth industrial induced induces
	way workers inducing employment ways labor working child work job action unions

Observations

- Expansion improves for rigorous criteria
  - Better for SMT than local RF

- Why?

Observations

- Expansion improves for rigorous criteria
  - Better for SMT than local RF

- Why?
  - Both can introduce some good terms
  - Local RF introduces more irrelevant terms
  - SMT more constrained
  - Challenge: Balance introducing info vs noise

Comparing Question Reformulations

- "Exact Phrases in Information Retrieval for Question Answering", Stoyanchev et al, 2008
- Investigates
  - Role of 'exact phrases' in retrieval for QA
  - Optimal query construction through document retrieval
    - From Web or AQUAINT collection
  - Impact of query specificity on passage retrieval

Motivation

- Retrieval bottleneck in Question-Answering
  - Retrieval provides source for answer extraction
  - If retrieval fails to return answer-contained documents, downstream answer processing is guaranteed to fail
  - Focus on recall in information retrieval phase
  - Consistent relationship b/t quality of IR and of QA

Motivation

- Retrieval bottleneck in Question-Answering
  - Retrieval provides source for answer extraction
  - If retrieval fails to return answer-contained documents, downstream answer processing is guaranteed to fail
  - Focus on recall in information retrieval phase
  - Consistent relationship b/t quality of IR and of QA
- Main factor in retrieval: query
  - Approaches vary from simple to complex processing or expansion with external resources

Approach

- Focus on use of ‘exact phrases’ from a question

Approach

- Focus on use of ‘exact phrases’ from a question
- Analyze impact of diff’t linguistic components of Q
  - Relate to answer candidate sentences

Approach

- Focus on use of 'exact phrases' from a question
- Analyze impact of diff't linguistic components of Q
  - Relate to answer candidate sentences
- Evaluate query construction for Web, Trec retrieval
  - Optimize query construction

Approach

- Focus on use of 'exact phrases' from a question
- Analyze impact of diff't linguistic components of Q
  - Relate to answer candidate sentences
- Evaluate query construction for Web, Trec retrieval
  - Optimize query construction
- Evaluate query construction for sentence retrieval
  - Analyze specificity

Data Sources & Resources

- Documents:
  - TREC QA AQUAINT corpus
  - Web

Data Sources & Resources

- Documents:
  - TREC QA AQUAINT corpus
  - Web
- Questions:
  - TREC2006, non-empty questions

Data Sources & Resources

- Documents:
  - TREC QA AQUAINT corpus
  - Web
- Questions:
  - TREC2006, non-empty questions
- Gold standard:
  - NIST-provided relevant docs, answer key: 3.5 docs/Q

Data Sources & Resources

- Documents:
  - TREC QA AQUAINT corpus
  - Web
- Questions:
  - TREC2006, non-empty questions
- Gold standard:
  - NIST-provided relevant docs, answer key: 3.5 docs/Q
- Resources:
  - IR: Lucene; NLTK, Lingpipe: phrase, NE annotation
    - Also hand-corrected

Query Processing Approach

- Exploit less-resource intensive methods
  - Chunking, NER
  - Applied only to questions, candidate sentences
    - Not applied to full collection

Query Processing Approach

- Exploit less-resource intensive methods
  - Chunking, NER
  - Applied only to questions, candidate sentences
    - Not applied to full collection → can use on Web
- Exact phrase motivation:
  - Phrases can improve retrieval
    - "In what year did the movie win academy awards?"

Query Processing Approach

- Exploit less-resource intensive methods
  - Chunking, NER
  - Applied only to questions, candidate sentences
    - Not applied to full collection → can use on Web
- Exact phrase motivation:
  - Phrases can improve retrieval
    - "In what year did the movie win **academy awards**?"
    - Phrase:

Query Processing Approach

- Exploit less-resource intensive methods
  - Chunking, NER
  - Applied only to questions, candidate sentences
    - Not applied to full collection → can use on Web
- Exact phrase motivation:
  - Phrases can improve retrieval
    - "In what year did the movie win **academy awards**?"
    - Phrase: Can rank documents higher
    - Disjunct:

Query Processing Approach

- Exploit less-resource intensive methods
  - Chunking, NER
  - Applied only to questions, candidate sentences
    - Not applied to full collection → can use on Web
- Exact phrase motivation:
  - Phrases can improve retrieval
    - "In what year did the movie win **academy awards**?"
    - Phrase: Can rank documents higher
    - Disjunct: Can dilute pool

Query Processing

- NER on Question and target
  - target: 1991 eruption on Mt. Pinatubo vs Nirvana

Query Processing

- NER on Question and target
  - target: 1991 eruption on **Mt. Pinatubo** vs **Nirvana**

Query Processing

- NER on Question and target
  - target: 1991 eruption on **Mt. Pinatubo** vs **Nirvana**
  - Uses LingPipe: ORG, LOC, PER
- Phrases (NLTK)
  - NP, VP, PP

Query Processing

- NER on Question and target
  - target: 1991 eruption on **Mt. Pinatubo** vs **Nirvana**
  - Uses LingPipe: ORG, LOC, PER
- Phrases (NLTK)
  - NP, VP, PP
- Converted Q-phrases:
  - Heuristic paraphrases on question as declarative
    - E.g. Who was|is NOUN|PRONOUN VBD ➔ NOUN|PRONOUN was|is VBD

Query Processing

- NER on Question and target
  - target: 1991 eruption on **Mt. Pinatubo** vs **Nirvana**
  - Uses LingPipe: ORG, LOC, PER
- Phrases (NLTK)
  - NP, VP, PP
- Converted Q-phrases:
  - Heuristic paraphrases on question as declarative
    - E.g. Who was|is NOUN|PRONOUN VBD → NOUN|PRONOUN was|is VBD
  - q-phrase: expected form of simple answer
    - E.g. When was Mozart born? → Mozart was born
  - How likely are we to see a q-phrase?

Query Processing

- NER on Question and target
  - target: 1991 eruption on **Mt. Pinatubo** vs **Nirvana**
  - Uses LingPipe: ORG, LOC, PER
- Phrases (NLTK)
  - NP, VP, PP
- Converted Q-phrases:
  - Heuristic paraphrases on question as declarative
    - E.g. Who was|is NOUN|PRONOUN VBD ➔ NOUN|PRONOUN was|is VBD
  - q-phrase: expected form of simple answer
    - E.g. When was Mozart born? ➔ Mozart was born
  - How likely are we to see a q-phrase? Unlikely
  - How likely is it to be right if we do see it?

Query Processing

- NER on Question and target
  - target: 1991 eruption on **Mt. Pinatubo** vs **Nirvana**
  - Uses LingPipe: ORG, LOC, PER
- Phrases (NLTK)
  - NP, VP, PP
- Converted Q-phrases:
  - Heuristic paraphrases on question as declarative
    - E.g. Who was|is NOUN|PRONOUN VBD → NOUN|PRONOUN was|is VBD
  - q-phrase: expected form of simple answer
    - E.g. When was Mozart born? → Mozart was born
  - How likely are we to see a q-phrase? Unlikely
  - How likely is it to be right if we do see it? Very

Comparing Query Forms

- Baseline: words from question and target

Comparing Query Forms

- Baseline: words from question and target
- Experimental:
  - Words, quoted exact phrases, quoted names entities
  - Backoff: Lucene: weight based on type

Comparing Query Forms

- Baseline: words from question and target
- Experimental:
  - Words, quoted exact phrases, quoted names entities
  - Backoff: Lucene: weight based on type
  - Backoff: Web: 1) converted q-phrases;
    - 2) phrases; 3) w/o phrases -- until 20 retrieved
    - Combined with target in all cases

Comparing Query Forms

- Baseline: words from question and target
- Experimental:
  - Words, quoted exact phrases, quoted names entities
  - Backoff: Lucene: weight based on type
  - Backoff: Web: 1) converted q-phrases;
    - 2) phrases; 3) w/o phrases -- until 20 retrieved
    - Combined with target in all cases
- Max 20 documents: expensive downstream process
  - Sentences split, ranked

Query Components

Target Question	United Nations What was the number of member nations of the U.N. in 2000?
Named Entity	U.N., United Nations
Phrases	“member nations of the U.N.”
Converted Q-phrase	“member nations of the U.N. in 2000”
Baseline Query	was the number of member nations of the U.N. in 2000 United Nations
Lucene Query with phrases and NE	was the number of member nations of the U.N. in 2000 “United Nations”, ”member nations of the u.n.”
Cascaded web query	
query1	“member nations of the U.N. in 2000” AND (United Nations)
query2	”member nations of the u.n.” AND (United Nations)
query3	(number of member nations of the U.N. in 2000) AND (United Nations)
query4	(United Nations)

Query Components in Supporting Sentences

	sent w/ answer		all sentences		precision
	num	proportion	num	proportion	
Named Entity	907	0.320	4873	0.122	.18
Phrases	350	0.123	1072	0.027	.34
Verbs	396	0.140	1399	0.035	.28
Q-Phrases	11	0.004	15	0.00038	.73
Words	2573	0.907	29576	0.745	.086
Total Sentences	2836		39688		

Highest precision:

Query Components in Supporting Sentences

	sent w/ answer		all sentences		precision
	num	proportion	num	proportion	
Named Entity	907	0.320	4873	0.122	.18
Phrases	350	0.123	1072	0.027	.34
Verbs	396	0.140	1399	0.035	.28
Q-Phrases	11	0.004	15	0.00038	.73
Words	2573	0.907	29576	0.745	.086
Total Sentences	2836		39688		

Highest precision: Converted q-phrase, then phrase,..

Query Components in Supporting Sentences

	sent w/ answer		all sentences		precision
	num	proportion	num	proportion	
Named Entity	907	0.320	4873	0.122	.18
Phrases	350	0.123	1072	0.027	.34
Verbs	396	0.140	1399	0.035	.28
Q-Phrases	11	0.004	15	0.00038	.73
Words	2573	0.907	29576	0.745	.086
Total Sentences	2836		39688		

Highest precision: Converted q-phrase, then phrase,..
Words likely to appear, but don't discriminate

Results

- Document and sentence retrieval
- Metrics:
  - Document retrieval:
    - Average recall
    - MRR
    - Overall document recall: % of questions w/>=1 correct doc
  - Sentence retrieval
    - Sentence MRR
    - Overall sentence recall
    - Average prec of first sentence
    - # correct in top 10, top 50

Results

- Document and sentence retrieval
- Metrics:
  - Document retrieval:
    - Average recall
    - MRR
    - Overall document recall: % of questions w/>=1 correct doc

Results

	avg doc recall	avg doc MRR	overall doc recall	avg sent MRR	overall sent recall	avg corr sent in top 1	avg corr sent in top 10	avg corr sent in top 50
IR with Lucene on AQUAINT dataset								
baseline (words disjunction from target and question)	0.530	0.631	0.756	0.314	0.627	0.223	1.202	3.464
baseline + auto phrases	0.514	0.617	0.741	0.332	0.653	0.236	1.269	3.759
words + auto NEs & phrases	0.501	0.604	0.736	0.316	0.653	0.220	1.228	3.705
baseline + manual phrases	0.506	0.621	0.738	0.291	0.609	0.199	1.231	3.378
words + manual NEs & phrases	0.510	0.625	0.738	0.294	0.609	0.202	1.244	3.368
IR with Yahoo API on WEB								
baseline words disjunction	-	-	-	0.183	0.570	0.101	0.821	2.316
cascaded using auto phrases	-	-	-	0.220	0.604	0.140	0.956	2.725
cascaded using manual phrases	-	-	-	0.241	0.614	0.155	1.065	3.016

Discussion

- Document retrieval:
  - About half of the correct docs are retrieved, rank 1-2
- Sentence retrieval:
  - Lower, correct sentence ~ rank 3

Discussion

- Document retrieval:
  - About half of the correct docs are retrieved, rank 1-2
- Sentence retrieval:
  - Lower, correct sentence ~ rank 3
- Little difference for exact phrases in AQUAINT

Discussion

- Document retrieval:
  - About half of the correct docs are retrieved, rank 1-2
- Sentence retrieval:
  - Lower, correct sentence ~ rank 3
- Little difference for exact phrases in AQUAINT
- Web:
  - Retrieval improved by exact phrases
    - Manual more than auto (20-30%) relative
  - Precision affected by tagging errors