Learning to Rank Answers on Large Online QA Collections

Mihai Surdeanu[†,⋆], Massimiliano Ciaramita[⋆], Hugo Zaragoza[⋆]

[†]Barcelona Media Innovation Center, [⋆]Yahoo! Research Barcelona

June 16th, 2008

What is Question Answering?

Answer *natural language* questions with small fragments of *text*.

"What is the capital of Ohio?" $\rightarrow$ "Columbus"

"What is ACL?" $\rightarrow$ "Austin City Limits Music Festival: annual end of summer event held in Austin, Texas."

Motivation

- Most effort concentrated on factoid and definitional Question Answering (QA), e.g., TREC, CLEF evaluations.
- Little research and virtually no data available for non-factoid QA, such as manner or reason questions.
- Recent years have seen an explosion of user-generated content such as community-driven question-answering (Yahoo! Answers).
  - Advantages: large, open-domain, multilingual.
  - Disadvantages: high variance of quality.

Examples

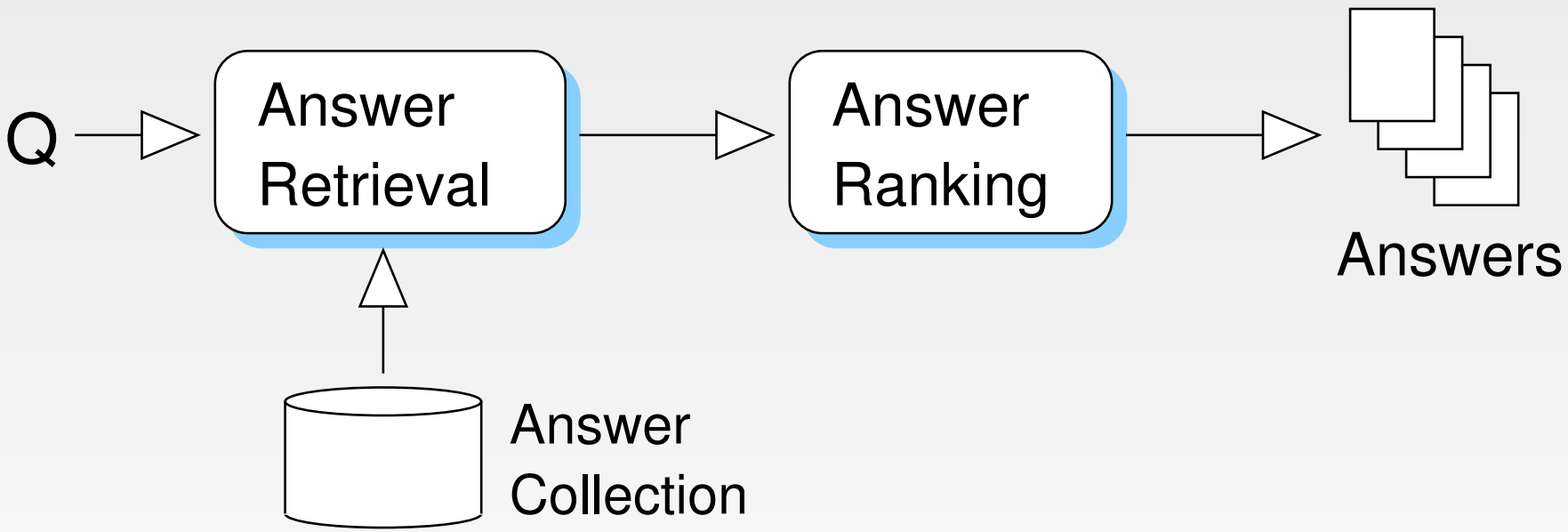
High Quality	*Q*: How do you quiet a squeaky door? *A*: Spray WD-40 directly onto the hinges of the door. Open and close the door several times. Remove hinges if the door still squeaks. Remove any rust, dirt or loose paint. Apply WD-40 to removed hinges. Put the hinges back, open and close door several times again.
High Quality	*Q*: How does a helicopter fly? *A*: A helicopter gets its power from rotors or blades. So as the rotors turn, air flows more quickly over the tops of the blades than it does below. This creates enough lift for flight.
Low Quality	*Q*: How to extract html tags from an html documents with c++? *A*: very carefully

Goal

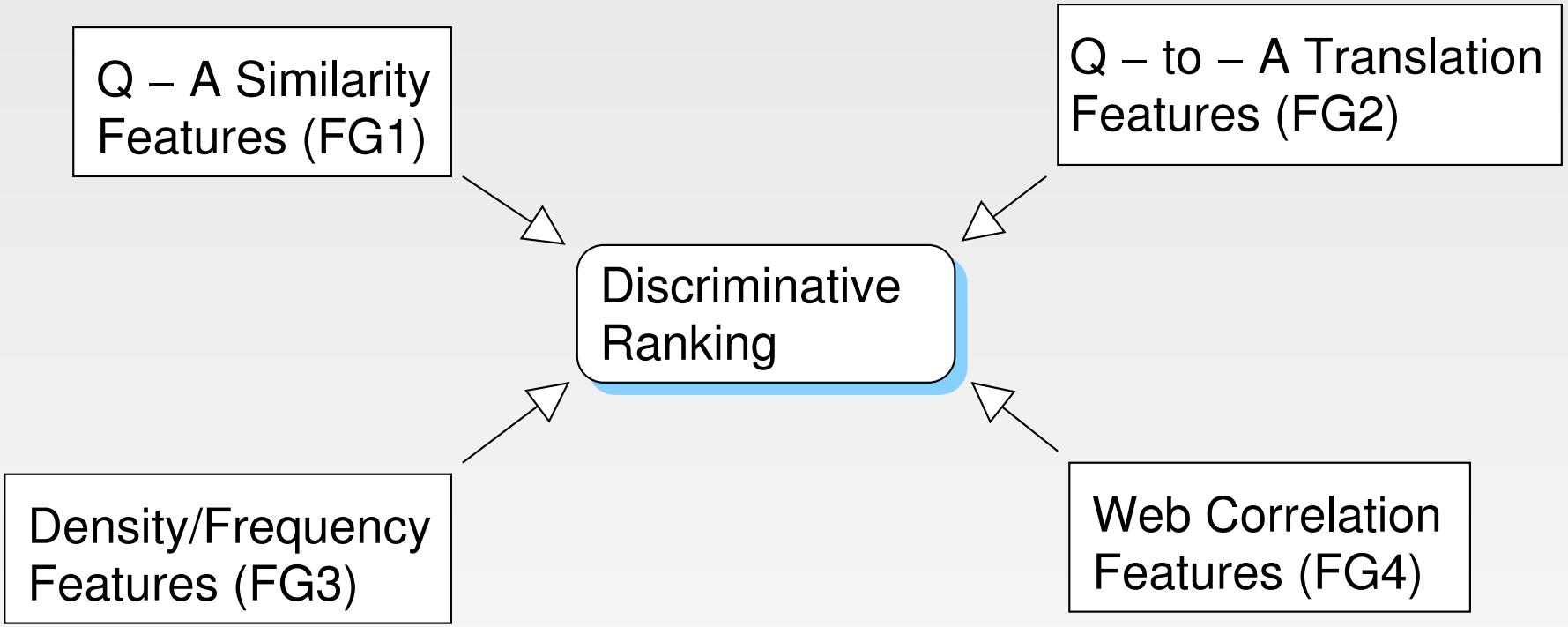
- Is it possible to learn an answer ranking model for complex questions from such noisy data?
- Which features/models are most useful in this scenario?

Outline

Approach: System Architecture



Approach: Learning Framework



- Classifier: ranking Perceptron of Shen and Joshi (2005)
- $+$ samples: Answers marked or voted as best in Y! Answers.
- $-$ samples: All other answers retrieved by IR.

Features (1/2)

FG1 Similarity Features

- The best answer is the one most similar to Q.
- BM25 and $tf \cdot idf$ between Q and A.

FG2 Translation Features

- A - source language, Q - target language.
- The best answer is the one most likely to translate to Q.
- $P(Q|A)$ computed using IBM Model 1.

Features (2/2)

FG3 Density and Frequency Features

- Same word sequence - Q terms recognized in the same order in A.
- Answer span - largest distance between two Q terms in A.
- Same sentence match - number of Q terms matched in a single sentence in A.
- Overall match - number of Q terms matched in A.
- Informativeness - number of NN, VB, JJ in A that are not found in Q.

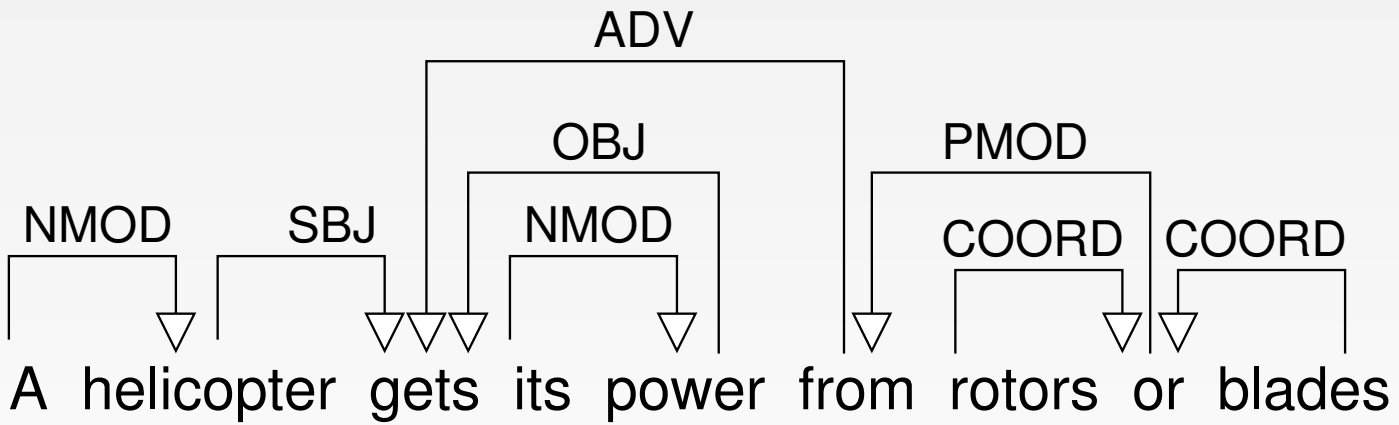
FG4 Web Correlation Features

- The best answer is the one most correlated with Q on the Web.
- Web correlation - CCP using search engine hits.
- Query-log correlation - PMI and χ^2 between (Q, A) words and a large query log.

Representation of Content: Structures

Features computed for several representations of content:

- *Words (W)* - the text is seen as a bag of words.
- *N-grams (N)* - the text is represented as a bag of *n*-grams.
- *Dependencies(D)* - the text is represented as a bag of syntactic dependencies.



Structure Parameters

- Degree of lexicalization:
  - Fully lexicalized structures, e.g., “helicopter” $\xrightarrow{\text{SBJ}}$ “get”.
  - Lexical elements replaced with coarse WordNet super senses (WNSS), e.g., `n.artifact` $\xrightarrow{\text{SBJ}}$ `v.possession`.
- Labels of relations: dependency relations can be labeled or unlabeled, e.g., “helicopter” $\xrightarrow{\text{SBJ}}$ “get” vs. “helicopter” $\rightarrow$ “get”.
- Structure size: controls the maximum number of elements in *n*-grams or dependency chains.

Outline

The Corpus

- Corpus build from a Nov. 2007 sample of Yahoo! Answers. Users ask questions and answer other users' questions. Best answers chosen by the asker or voted by participants.
- How to obtain:
  - Distributed through Yahoo!'s Webscope program.
  - Contact Kim Capps-Tanaka at research-data-requests@yahoo-inc.com
  - Ask for "Yahoo! Answers Manner Questions, version 1.0"
- 142,627 (Q, best A) pairs.
  - We index all As in this set as the collection **C**.
  - Partitioning of questions: 60% training, 20% development, 20% testing.

Evaluation Measures

- We evaluate results using two measures:
  1. Precision at rank 1 (P@1) - percentage of questions with correct answer on first position.
  2. Mean Reciprocal Rank (MRR) - average of question scores; score of a question is $1/k$, where k is position of correct answer.
- We are interested in the ranker's performance: we evaluate on the questions where the correct answer is retrieved from **C** in top N by Answer Retrieval.

Overall Results

	N = 15 (29.04% cov.)	N = 25 (32.81% cov.)	N = 50 (38.09% cov.)
Baseline (BM25)	41.48%	36.74%	31.66%
Ranking	**49.59%**$\pm$0.03	**43.98%**$\pm$0.09	**37.99%**$\pm$0.01
Relative Improvement	+19.55%	+19.70%	+19.99%

P@1 for the test partition for various values of N

	N = 15 (29.04% cov.)	N = 25 (32.81% cov.)	N = 50 (38.09% cov.)
Baseline (BM25)	56.12	50.31	43.74
Ranking	**63.84**$\pm$0.01	**57.76**$\pm$0.07	**50.72**$\pm$0.01
Relative Improvement	+13.75%	+14.80%	+15.95%

MRR for the test partition for various values of N

Model Selection Process

Iter.	Feature Set	MRR	P@1
0	BM25(W)	56.06	41.12%
1	+ translation(N_{WN})	61.13	46.24%
2	+ frequency/density(D)	62.50	48.34%
3	+ translation(W)	63.00	49.08%
4	+ query-log correlation	63.50	49.63%
5	+ frequency/density(W)	63.71	49.84%
6	+ query-log correlation	63.87	50.09%
7	+ frequency/density(W)	63.99	50.23%
8	+ translation(N)	64.03	50.30%
9	+ similarity(W)	64.08	50.42%
10	+ frequency/density(W)	64.10	50.42%
11	+ frequency/density(W)	64.18	50.36%
12	+ similarity(N)	64.22	50.36%
13	+ frequency/density(W)	64.33	50.54%
14	+ query-log correlation	64.46	50.66%
15	+ frequency/density(W)	64.55	50.78%
16	+ query-log correlation	64.60	50.88%
17	+ frequency/density(W)	64.65	50.91%

Contribution of the Various Content Representations

	Individual representations					Combined representations			
	W	N	N_{WN}	D	D_{WN}	W +N	W +N +N_{WN}	W +N +N_{WN} +D	W +N +N_{WN} +D +D_{WN}
FG1	0	**+1.06**	-2.01	+0.84	-1.75	+1.06	+1.06	+1.06	+1.06
FG2	+4.95	+4.73	+5.06	+4.63	+4.66	+5.80	+6.01	**+6.36**	+6.36
FG3	+2.24	+2.33	+2.39	+2.27	+2.41	+3.56	+3.56	**+3.62**	+3.62

MRR improvements on the development set (N = 15)

The NL analysis provides *complementary* information to the bag-of-word models!

Contribution of the Various Content Representations

	Individual representations					Combined representations			
	W	N	N_{WN}	D	D_{WN}	W +N	W +N +N_{WN}	W +N +N_{WN} +D	W +N +N_{WN} +D +D_{WN}
FG1	0	**+1.06**	-2.01	+0.84	-1.75	+1.06	+1.06	+1.06	+1.06
FG2	+4.95	+4.73	+5.06	+4.63	+4.66	+5.80	+6.01	**+6.36**	+6.36
FG3	+2.24	+2.33	+2.39	+2.27	+2.41	+3.56	+3.56	**+3.62**	+3.62

MRR improvements on the development set (N = 15)

The NL analysis provides *complementary* information to the bag-of-word models!

Conclusions

- Answer ranking engine built using a community-generated question-answer collection.
- Combination is key for improvement:
  - Combined several models: translation, similarity, frequency/density, web correlation.
  - Combined several representations of content: bag of words, n-grams, dependencies, word senses.
- NL analysis yields a small yet remarkable improvement, considering the scale of the evaluation.

How do you respond to a question when you don't know the answer?

It is simple, "I do not know the answer". Doing otherwise will leave you appearing awkward. It is good to be genuine in whatever you do.

I would give a stupid, yet humorous answer and hope for the best.

Just say... I don't know... follow it up with a blank stare.

Just play around the outside of the question avoiding the main question and hopefully someone won't ask you any questions.

YAHOO! ANSWERS

Thank you!