Learning to Rank Answers on Large Online QA Collections

Mihai Surdeanu, Massimiliano Ciaramita, Hugo Zaragoza

Barcelona Media Innovation Center, Yahoo! Research Barcelona

May 6, 2008

What is Question Answering?

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

- "What is the capital of Spain?" $\longrightarrow$ "Madrid"

Social QA

- “What is the difference between Quakers and Mennonites?” $\rightarrow$ “Both Mennonites (or Anabaptists) and Quakers (Friends Society) are peace faiths born out of a movement of groups who objected to basic tenets of the Church of England....”
- “What do you do with a 15month old girl who holds her breath till she gets what she wants?” $\rightarrow$ “Ignore her she can’t hurt herself she’ll start breathing if she passes out.”
- “Is it illegal to sleep in the back of my car while drunk?”$\rightarrow$ “don’t have the keys on your person or in the ignition... stick them in the console or glove compartment you should be good to go”

Social QA

- “Do journalism and politics mix?” →
  “Perhaps the major issue here is defining public interest and the form that the presentation thereof takes. People’s opinions on what defines relevant news varies, and objective presentation, like beauty, is in the eye of the beholder. Hyperbole is maligned by some, praised by others. It’s less a matter of what is objective than it is of who can determine what is objective.
  Journalists have the duty to seek and present the truth, so inherently they must question authority, but their questioning must be based on contrary information, or to inquiry about verification. Challenging authority should never be about egos. Illegitimate questions come at the expense of integrity.”

Social QA

- “What can we do to make poverty history?”→

“I admire Bono for his work on this issue. I think that the ONE campaign, the debt reduction campaign and the work with the G-8 is a very impressive step in the right direction.
However, I think that you have to attack the problem from the top down and the bottom up.
If debt reduction and foreign aid happen to the degree that has been promised the momentum is lost unless the people in Africa are able to capitalize on the opportunity.
It is hard to make a living, provide for your family, or start a new business when you spend your time and energy looking for safe drinking water or your next meal. We in the West take for granted our infrastructure that allows us to focus more of our energy on work and our ability to get ahead, provide for our families, etc.
How much time did you spend today looking for water or standing in line for water or food?
My answer is that the next step must be to establish a framework for what basic infrastructure must be present in each community (ie: access to food, water, health facilities, transportation, roads, education, daycare) to allow the people in those communities to work effectively and productively.
Working from the ground up I would suggest that the ONE campaign expand to coordinate the partnering of a sponsoring community/city/town with a similar sized one in Africa. But, unlike sister communities around the world that are strictly symbolic, the communities that adopt a "sister" in Africa must work with them to build their community to the minimum infrastructure framework that is established by the campaign. It could even be a competition in the sense that the people of a city in Norway could be competing with the people of a city in Canada or the USA to meet and exceed the framework standards set out in a given time frame.
Unlike many charities that ask for money, wouldn’t it be rewarding to know that your city is working to provide a better life and future for one in Africa. You would know where your money is going, where work is being done, and who is doing it. We all have a certain skill set and in your community there is a group of people who could make a real difference to ONE city/town/community in Africa.
We all tend to sit back and expect our politicians and leaders to fix the problems of the world and then complain that things don’t seem to be getting much better.

The people need to take control and be part of the solution.”

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; “i dunno”, “www.DailyMakeover.com”, “What?”, “LOL”, etc.

Examples

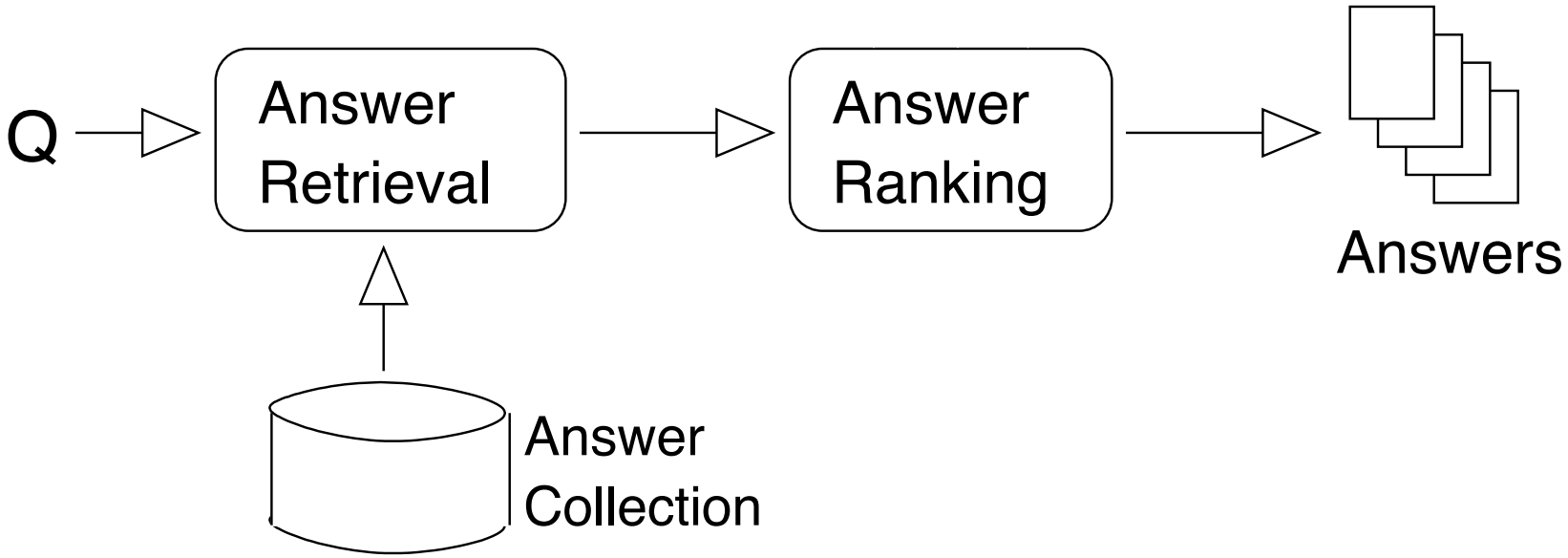
High Quality	*Q*: How do you quiet a squeaky door? *A*: Spray WD-40 directly onto the hinges of the door. Open and close the door several times. Remove hinges if the door still squeaks. Remove any rust, dirt or loose paint. Apply WD-40 to removed hinges. Put the hinges back, open and close door several 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

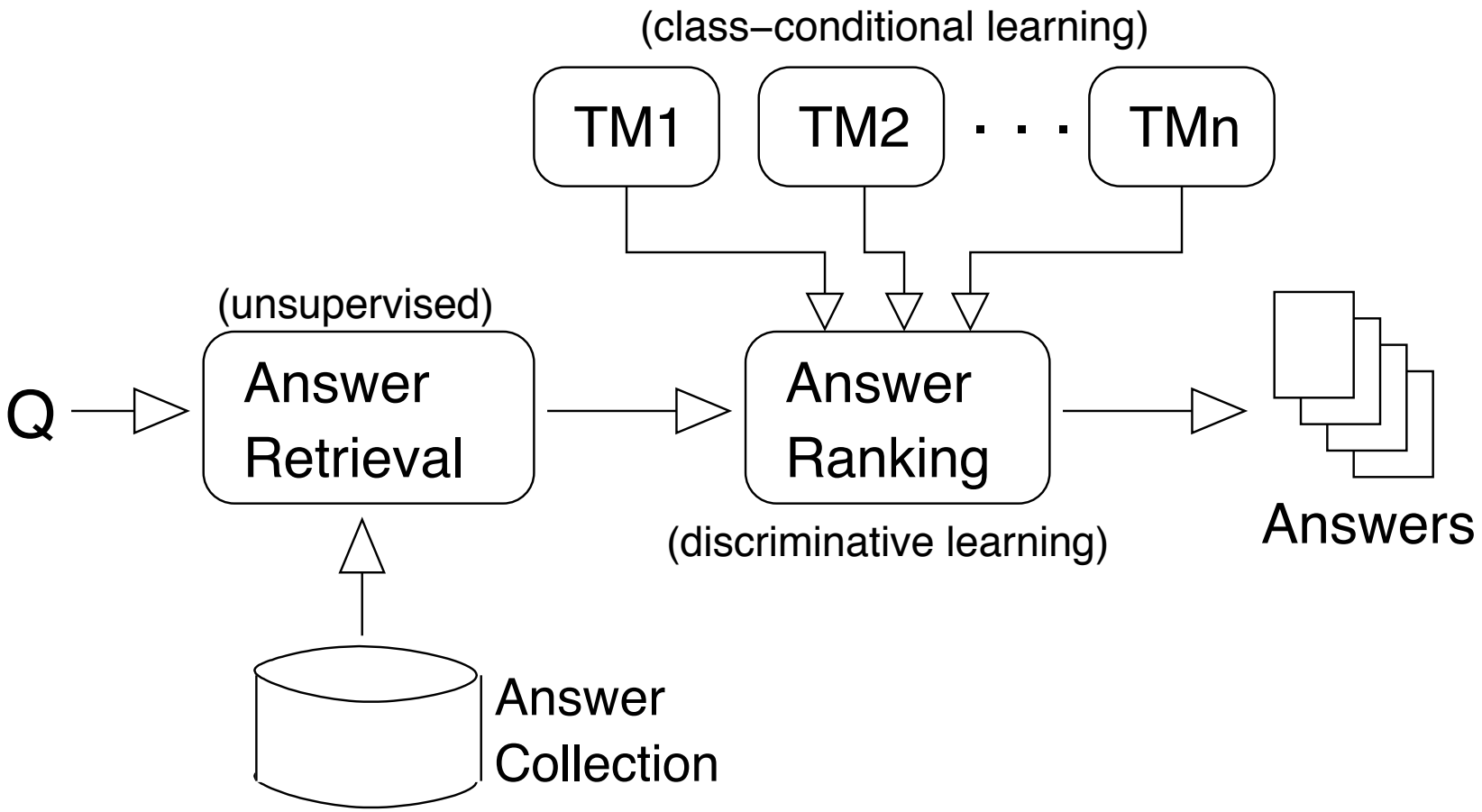
- Is it possible to learn an answer ranking model for complex questions from such noisy data?
- Which features/models are most useful in this scenario?

Outline

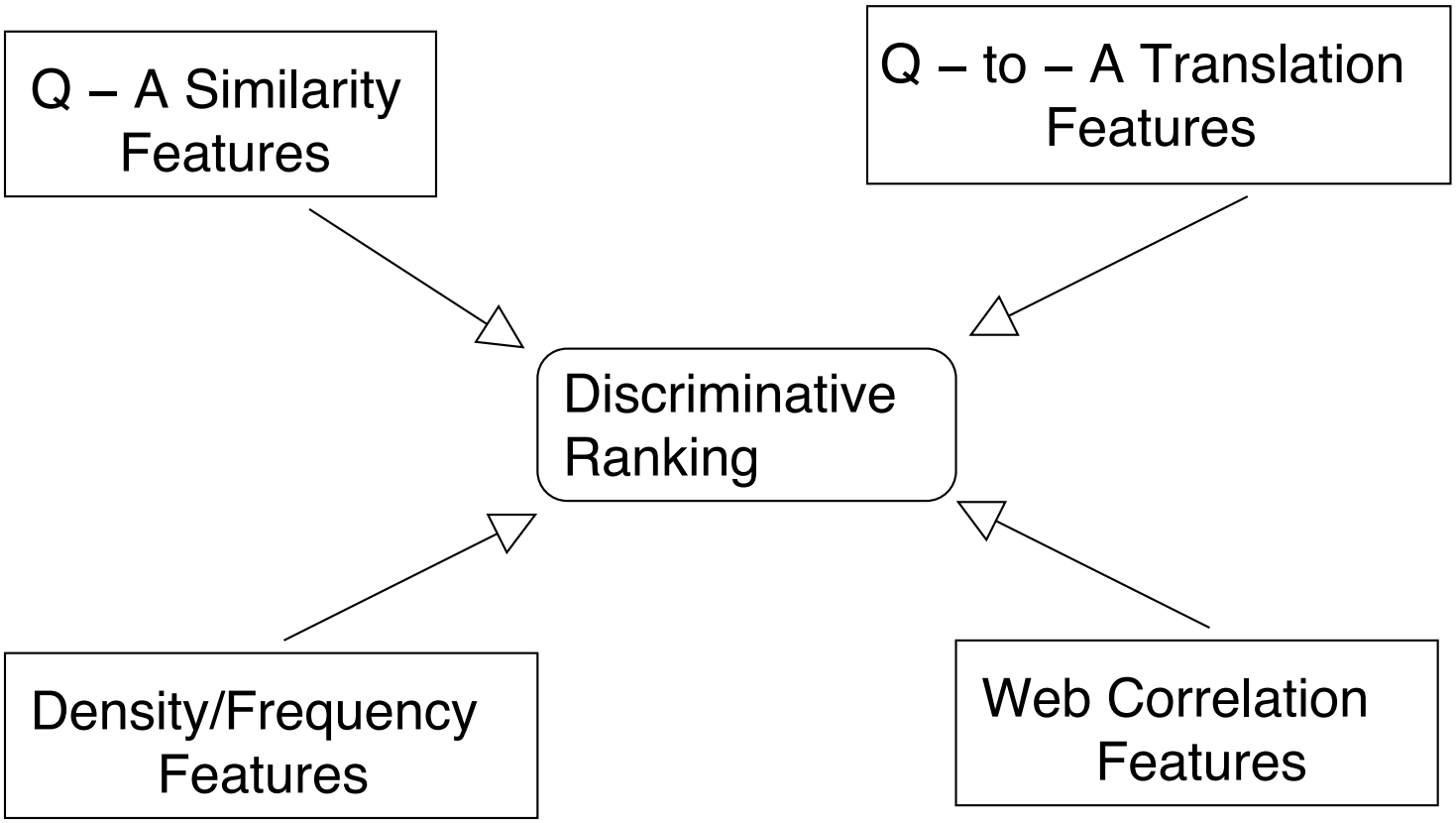
Approach: System Architecture



Approach: System Architecture



Approach: Learning Framework



- Positive samples: Answers marked as best in Yahoo! Answers.
- Negative samples: All other answers retrieved by IR.

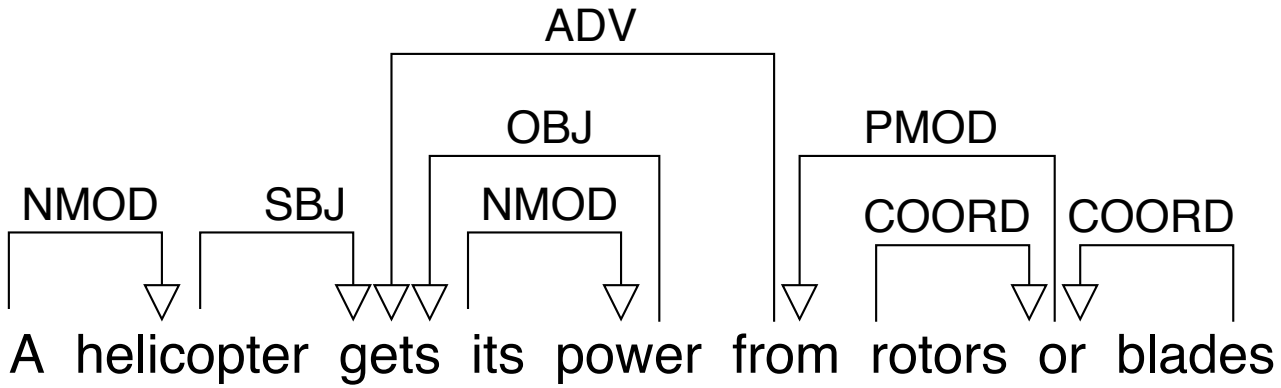
Features

1. FG1: Similarity Features
   - BM25 and $tf \cdot idf$ between Q and A.
2. FG2: Translation Features
   - $P(Q|A)$ given by IBM Model 1.
3. 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.
4. FG4: Web Correlation Features
   - 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

To investigate the contribution of NLP, we replicate most features for five different 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.



Representation of Content: 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`.
  - Lexical elements replaced with WSJ NE tags, e.g., `VEHICLE` $\xrightarrow{\text{SBJ}}$ “get”.
- 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.
- Focused on manner ("how to") questions. Corpus built using 2 filtering steps:
  1. Kept only questions that match the regular expression: `how (to|do|did|does|can|would|could|should)` and have an answer selected as best either by the asker or by the participants in the thread.
     - 364,419 (Q, best A) pairs.
  2. Removed the questions and answers of obvious low quality.
     - Heuristic: Both Q and A must have at least 4 words, out of which at least 1 noun and 1 verb.
     - 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.

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) - 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

P@1	N = 10 (26.25% cov.)	N = 15 (29.04% cov.)	N = 25 (32.81% cov.)	N = 50 (38.09% cov.)
IR	45.94%	41.48%	36.74%	31.66%
Ranking	**53.48%**±0.01	**49.65%**±0.03	**43.52%**±0.09	**37.51%**±0.09
Relative Improvement	+16.41%	+19.69%	+18.45%	+18.47%

MRR	N = 10 (26.25% cov.)	N = 15 (29.04% cov.)	N = 25 (32.81% cov.)	N = 50 (38.09% cov.)
IR	61.33	56.12	50.31	43.74
Ranking	**67.77**±0.09	**63.85**±0.01	**56.90**±0.07	**49.81**±0.08
Relative Improvement	+10.50%	+13.77%	+13.09%	+13.87%

Model Selection Process (1/2)

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%
18	+ similarity(N) $\times$ freq/dens(W)	64.67	50.88%
19	+ freq/dens(W) $\times$ translation(N_{WN})	64.76	51.04%
20	+ freq/dens(D_g) $\times$ query-log correlation	64.81	51.10%

Model Selection Process (2/2)

- Feature combination is key for improvement: 60% of improvement due to translation features, 20% due to frequency/density features, the rest caused by query-log-correlation features.
- The first two features chosen use NL analysis: NL structures complement well bag-of-word representations.
- Web-hit correlation not useful here because queries are too long. After query relaxation most meaning is lost.

Contribution of NL Analysis

	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

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

Contribution of NL Analysis

	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

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

Conclusions

- Answer ranking engine built using a community-generated question-answer collection:
  - Large-scale experimentation with various models/features.
  - Potential application: retrieval from social media.
  - Potential application: open-domain QA on the Web.
- 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, NEs.
- NL analysis yields a small, yet statistically-significant improvement. OK considering that:
  - We use off-the-shelf NLP processors.
  - We evaluate on a large corpus with noisy and subjective information.

Conclusions

- Answer ranking engine built using a community-generated question-answer collection:
  - Large-scale experimentation with various models/features.
  - Potential application: retrieval from social media.
  - Potential application: open-domain QA on the Web.
- 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, NEs.
- NL analysis yields a small, yet statistically-significant improvement. OK considering that:
  - We use off-the-shelf NLP processors.
  - We evaluate on a large corpus with noisy and subjective information.

Conclusions

- Answer ranking engine built using a community-generated question-answer collection:
  - Large-scale experimentation with various models/features.
  - Potential application: retrieval from social media.
  - Potential application: open-domain QA on the Web.
- 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, NEs.
- NL analysis yields a small, yet statistically-significant improvement. OK considering that:
  - We use off-the-shelf NLP processors.
  - We evaluate on a large corpus with noisy and subjective information.

Thank you!