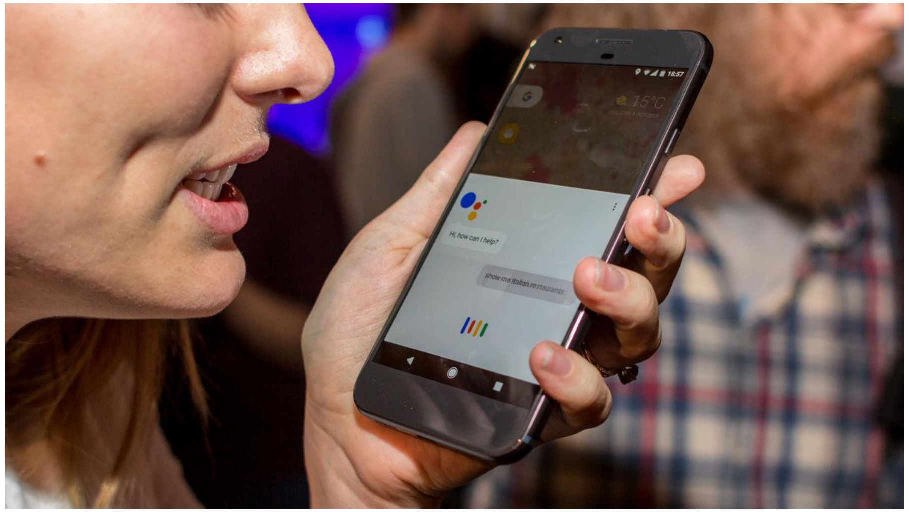
The Relevance of Answers

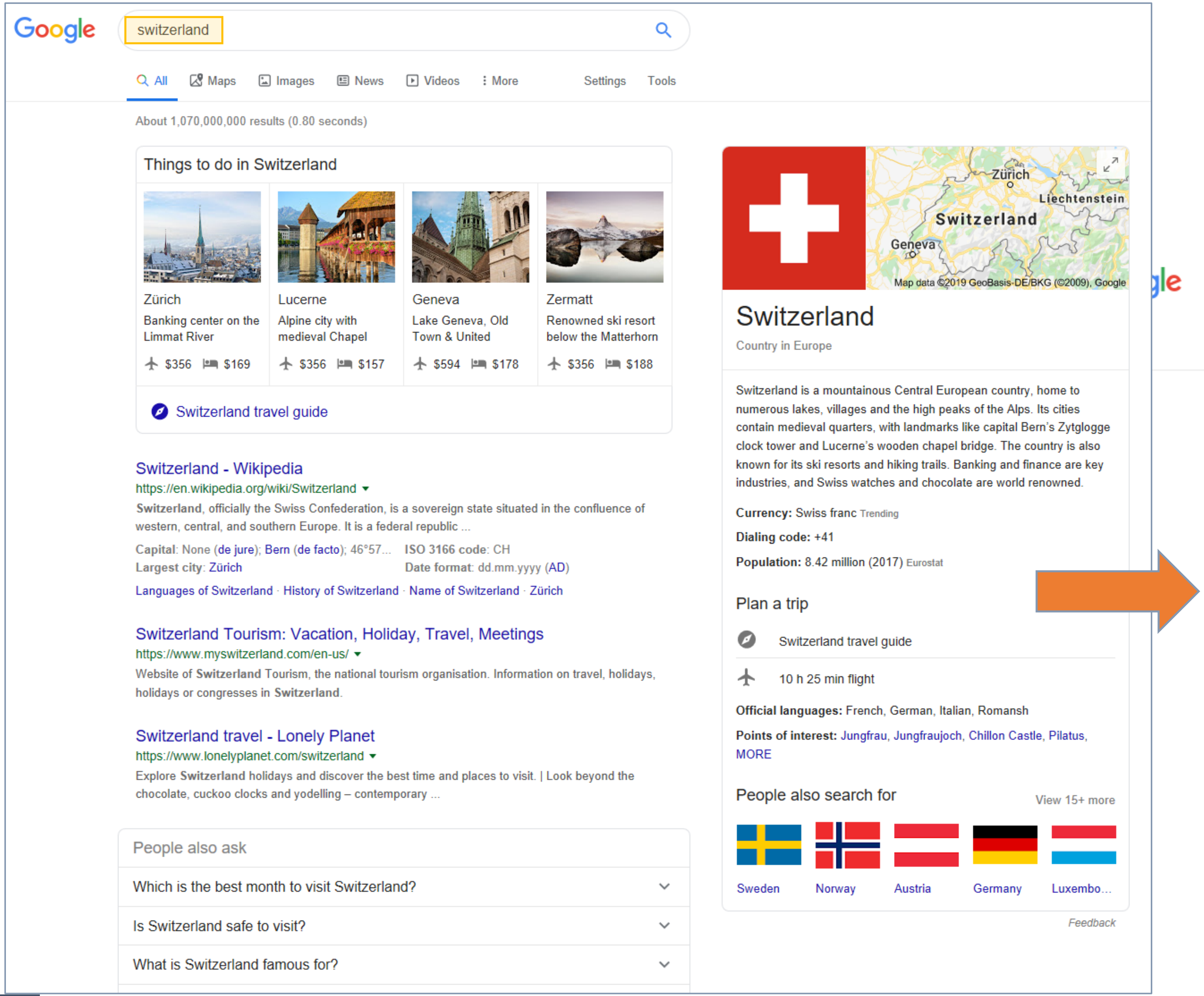
Bruce Croft
UMass Amherst and RMIT University

CLEF 2019

Google


Hi, how can I help?
show me italian restaurants

amazon



Google switzerland

All Maps Images News Videos More Settings Tools

About 1,070,000,000 results (0.80 seconds)

Things to do in Switzerland

Zürich	Lucerne	Geneva	Zermatt
Banking center on the Limmat River	Alpine city with medieval Chapel	Lake Geneva, Old Town & United	Renowned ski resort below the Matterhorn
$356 $169	$356 $157	$594 $178	$356 $188

Switzerland travel guide

Switzerland - Wikipedia
https://en.wikipedia.org/wiki/Switzerland ▾
Switzerland, officially the Swiss Confederation, is a sovereign state situated in the confluence of western, central, and southern Europe. It is a federal republic ...
Capital: None (de jure); Bern (de facto); 46°57... ISO 3166 code: CH
Largest city: Zürich Date format: dd.mm.yyyy (AD)
Languages of Switzerland · History of Switzerland · Name of Switzerland · Zürich

Switzerland Tourism: Vacation, Holiday, Travel, Meetings
https://www.myswitzerland.com/en-us/ ▾
Website of Switzerland Tourism, the national tourism organisation. Information on travel, holidays, holidays or congresses in Switzerland.

Switzerland travel - Lonely Planet
https://www.lonelyplanet.com/switzerland ▾
Explore Switzerland holidays and discover the best time and places to visit. | Look beyond the chocolate, cuckoo clocks and yodelling – contemporary ...

People also ask

Which is the best month to visit Switzerland?

Is Switzerland safe to visit?

What is Switzerland famous for?

Switzerland

Country in Europe

Switzerland is a mountainous Central European country, home to numerous lakes, villages and the high peaks of the Alps. Its cities contain medieval quarters, with landmarks like capital Bern's Zytglogge clock tower and Lucerne's wooden chapel bridge. The country is also known for its ski resorts and hiking trails. Banking and finance are key industries, and Swiss watches and chocolate are world renowned.

Currency: Swiss franc Trending

Dialing code: +41

Population: 8.42 million (2017) Eurostat

Plan a trip

Switzerland travel guide

10 h 25 min flight

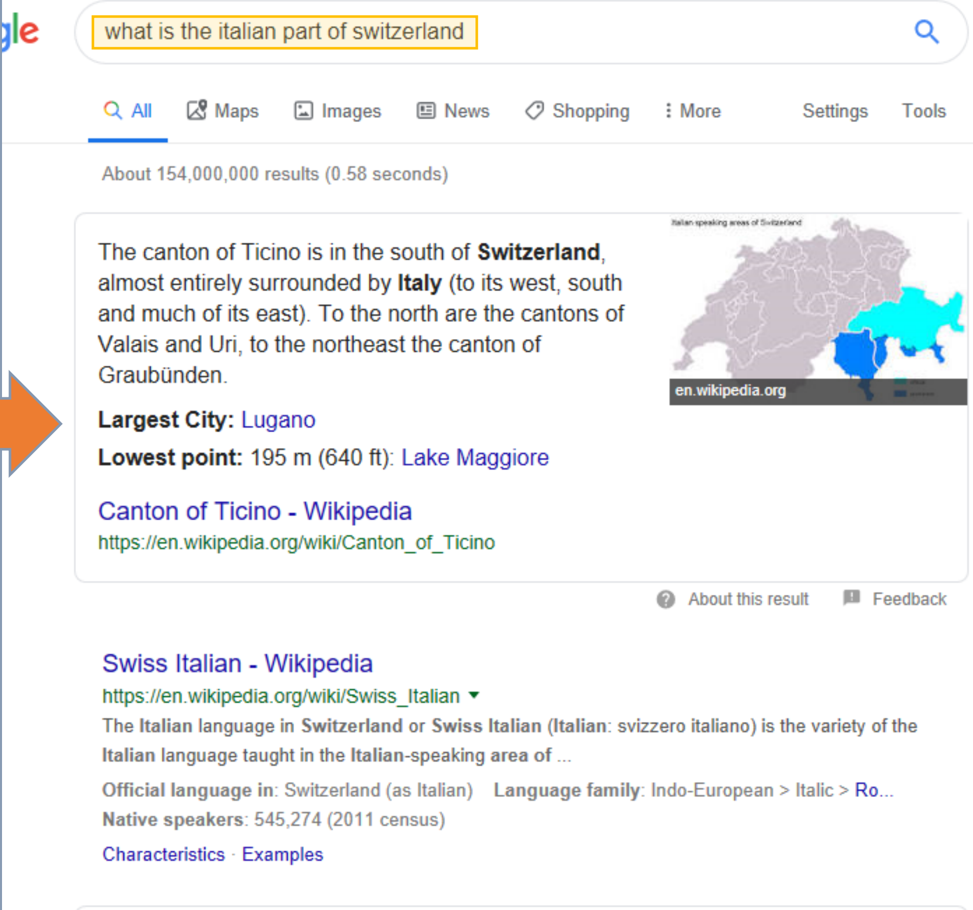
Official languages: French, German, Italian, Romansh

Points of interest: Jungfrau, Jungfraujoch, Chillon Castle, Pilatus, MORE

People also search for View 15+ more

Sweden Norway Austria Germany Luxembo...

Feedback



what is the italian part of switzerland

All Maps Images News Shopping More Settings Tools

About 154,000,000 results (0.58 seconds)

The canton of Ticino is in the south of **Switzerland**, almost entirely surrounded by **Italy** (to its west, south and much of its east). To the north are the cantons of Valais and Uri, to the northeast the canton of Graubünden.

Largest City: Lugano

Lowest point: 195 m (640 ft): Lake Maggiore

Canton of Ticino - Wikipedia
https://en.wikipedia.org/wiki/Canton_of_Ticino

About this result Feedback

Swiss Italian - Wikipedia
https://en.wikipedia.org/wiki/Swiss_Italian ▾
The Italian language in Switzerland or Swiss Italian (Italian: svizzero italiano) is the variety of the Italian language taught in the Italian-speaking area of ...
Official language in: Switzerland (as Italian) Language family: Indo-European > Italic > Ro...
Native speakers: 545,274 (2011 census)
Characteristics · Examples

This talk

- History of answer retrieval
- Answers vs. documents
- Relevance vs. correctness
- Ranking vs. interaction
- Tasks and test collections
- State-of-the-art
- Challenges

A Timeline

Time

Document Retrieval

Answer Passage Retrieval

Sentence Retrieval

Passages as Features

Snippet Retrieval

QA Factoid Retrieval

CQA or Non-Factoid QA

Conversational Answer Retrieval

Answer Passage Retrieval Revisited

Response Retrieval/Generation

Question Answering/Machine Comprehension

Complex Answer Retrieval
(Passages as Summaries)

Dimensions of Answer Retrieval

- Granularity
  - entity, sentence, passage, document, multi-document
- Extractive
  - answer extracted from text or retrieved from a collection of answers
- Generated
  - answer based on single existing text, composed from multiple existing texts, or created using text generation model
- Conversational
  - “one-off” or taking session history into account
- Personalized
  - generic or customized to user

Answer Passage Retrieval

- O’Connor (1975, 1977, 1980) pioneered work in sentence and passage retrieval, in scientific and legal domains
- “Answer passage”[1]
  - “answer-reporting”: passage from which an answer to a question can be inferred, perhaps using specialized knowledge
  - “answer-indicative”: passage from which it can be inferred that the document does contain an answer-reporting passage
  - Assumed full questions
    - e.g., “What is the evidence that lung adenocarcinoma can be chemically induced?”
  - Manually created “search-word lists” for each subject-matter word in the question

[1] *Answer-Passage Retrieval by Text Searching*, JASIS, 1980

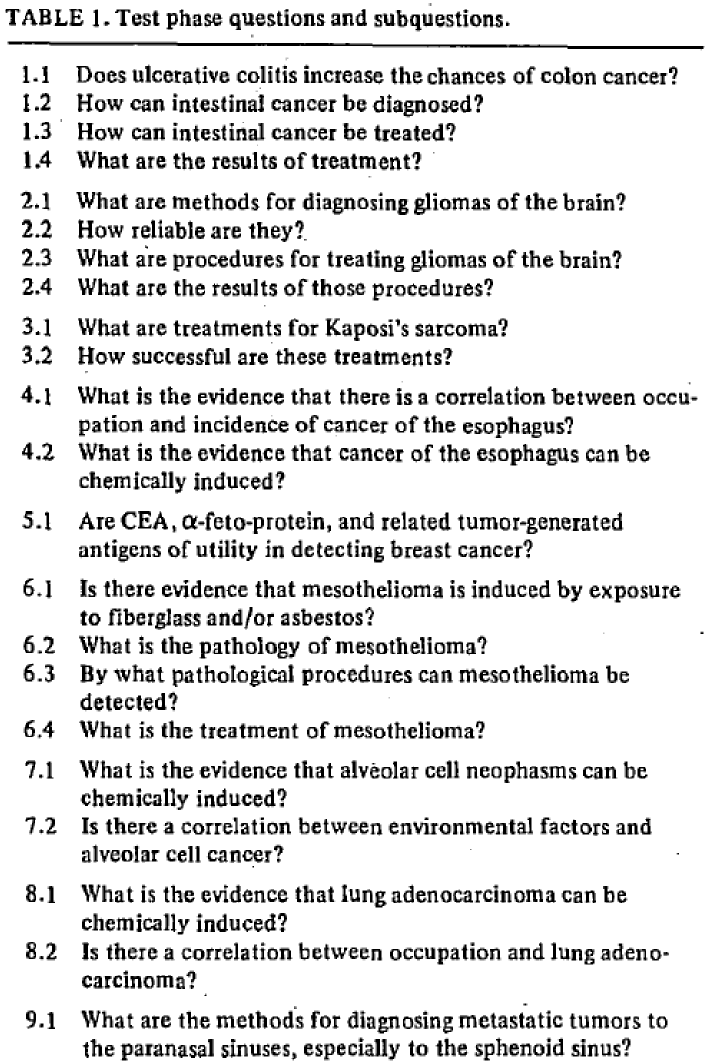


TABLE 1. Test phase questions and subquestions.

1.1 Does ulcerative colitis increase the chances of colon cancer?
1.2 How can intestinal cancer be diagnosed?
1.3 How can intestinal cancer be treated?
1.4 What are the results of treatment?

2.1 What are methods for diagnosing gliomas of the brain?
2.2 How reliable are they?
2.3 What are procedures for treating gliomas of the brain?
2.4 What are the results of those procedures?

3.1 What are treatments for Kaposi's sarcoma?
3.2 How successful are these treatments?

4.1 What is the evidence that there is a correlation between occupation and incidence of cancer of the esophagus?
4.2 What is the evidence that cancer of the esophagus can be chemically induced?

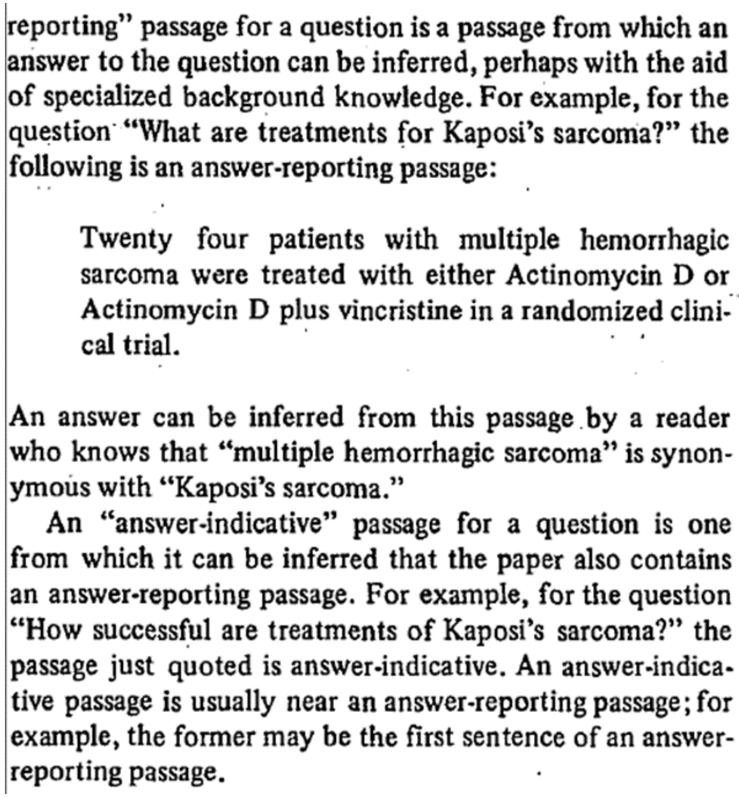
5.1 Are CEA, α-feto-protein, and related tumor-generated antigens of utility in detecting breast cancer?

6.1 Is there evidence that mesothelioma is induced by exposure to fiberglass and/or asbestos?
6.2 What is the pathology of mesothelioma?
6.3 By what pathological procedures can mesothelioma be detected?
6.4 What is the treatment of mesothelioma?

7.1 What is the evidence that alveolar cell neophasms can be chemically induced?
7.2 Is there a correlation between environmental factors and alveolar cell cancer?

8.1 What is the evidence that lung adenocarcinoma can be chemically induced?
8.2 Is there a correlation between occupation and lung adenocarcinoma?

9.1 What are the methods for diagnosing metastatic tumors to the paranasal sinuses, especially to the sphenoid sinus?

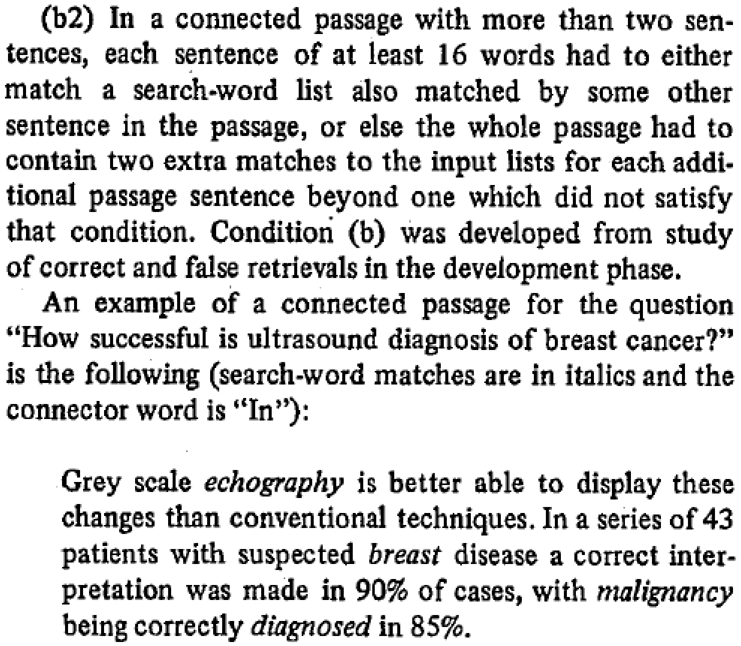


reporting" passage for a question is a passage from which an answer to the question can be inferred, perhaps with the aid of specialized background knowledge. For example, for the question "What are treatments for Kaposi's sarcoma?" the following is an answer-reporting passage:

> Twenty four patients with multiple hemorrhagic sarcoma were treated with either Actinomycin D or Actinomycin D plus vincristine in a randomized clinical trial.

An answer can be inferred from this passage by a reader who knows that "multiple hemorrhagic sarcoma" is synonymous with "Kaposi's sarcoma."

An "answer-indicative" passage for a question is one from which it can be inferred that the paper also contains an answer-reporting passage. For example, for the question "How successful are treatments of Kaposi's sarcoma?" the passage just quoted is answer-indicative. An answer-indicative passage is usually near an answer-reporting passage; for example, the former may be the first sentence of an answer-reporting passage.



(b2) In a connected passage with more than two sentences, each sentence of at least 16 words had to either match a search-word list also matched by some other sentence in the passage, or else the whole passage had to contain two extra matches to the input lists for each additional passage sentence beyond one which did not satisfy that condition. Condition (b) was developed from study of correct and false retrievals in the development phase.

An example of a connected passage for the question "How successful is ultrasound diagnosis of breast cancer?" is the following (search-word matches are in italics and the connector word is "In"):

> Grey scale *echography* is better able to display these changes than conventional techniques. In a series of 43 patients with suspected *breast* disease a correct interpretation was made in 90% of cases, with *malignancy* being correctly *diagnosed* in 85%.

Passage Retrieval

- Approach then shifted to using passages to improve document ranking effectiveness
    - Combining paragraph and document scores (Salton, Allan, and Buckley, 1993)
    - Combining topic segments from text tiling with document scores (Hearst and Plaunt, 1993)
    - HMMs for identifying relevant passages as part of document retrieval (Mittendorf and Schauble, 1994)
- Callan (1994) showed that fixed length “window” passages produced the best results for improving document ranking
    - Incorporated into Inquery, Indri, and Galago search engines

Passage Retrieval

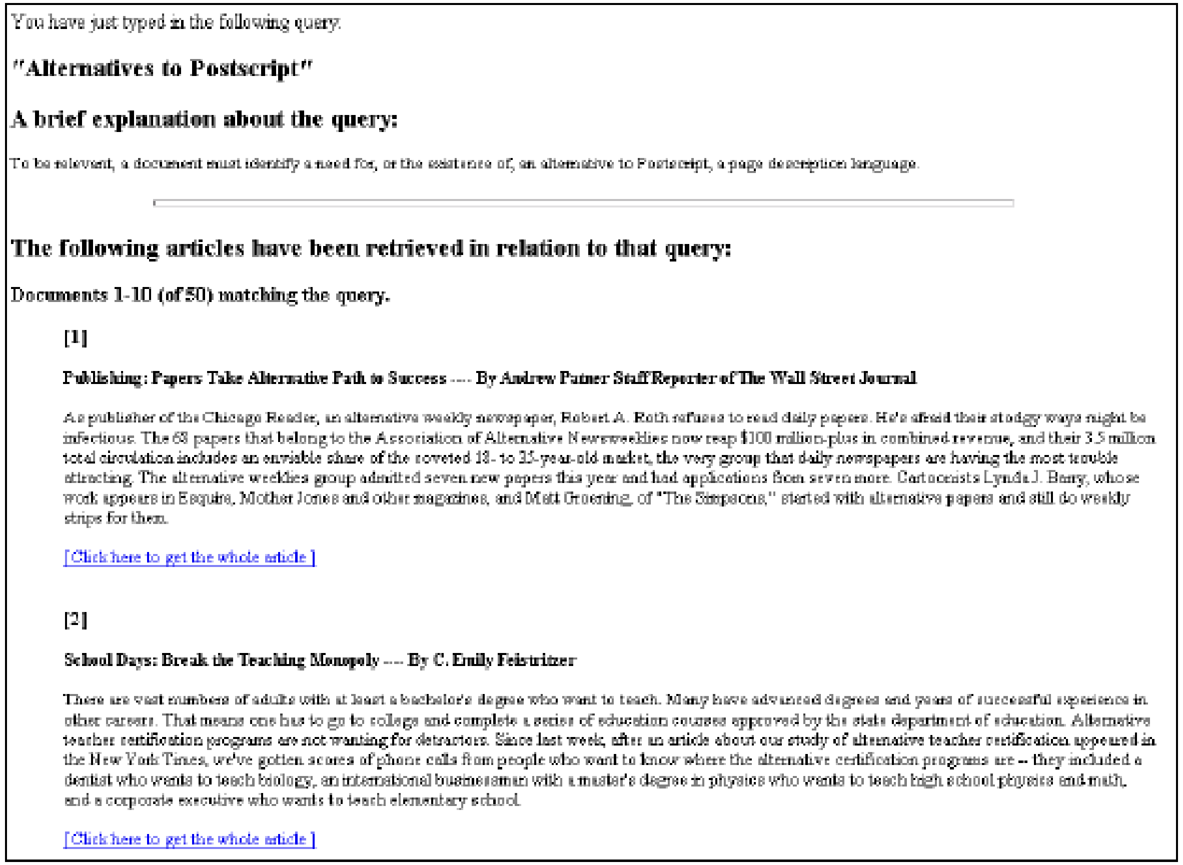
- Kaszkiel and Zobel (1997) tested “arbitrary” (variable width) passages, verified Callan’s results
- Liu and Croft (2002) described passage retrieval using language models
- Bendersky and Kurland (2010) showed best performance for this approach by varying document smoothing based on a homogeneity feature
- Lv and Zhai (2009) described the positional language model and applied it to passage retrieval

Sentence Retrieval

- Luhn (1958) ranked sentences by *significance* to create abstracts
  - Many summarization approaches based on identifying “best” sentences
- O’Connor (1975) retrieved “answer sentences”
- TREC Novelty track (2002) defined sentence retrieval tasks based on relevance and novelty assessments
- Murdock (2005) used a translation model for sentence retrieval and tested using TREC Novelty and QA data
- Balsubramanian (2007) compared variety of models for sentence retrieval
- Metzler and Kanungo (2008) used L2R models to rank sentences based on a range of features

Summaries and Snippets

Google patent, 2005.



You have just typed in the following query:

"Alternatives to Postscript"

A brief explanation about the query:

To be relevant, a document must identify a need for, or the existence of, an alternative to Postscript, a page description language.

The following articles have been retrieved in relation to that query:

Documents 1-10 (of 50) matching the query.

[1]

Publishing: Papers Take Alternative Path to Success ---- By Andrew Patner Staff Reporter of The Wall Street Journal

As publisher of the Chicago Reader, an alternative weekly newspaper, Robert A. Roth refuses to read daily papers. He's afraid their stodgy ways might be infectious. The 68 papers that belong to the Association of Alternative Newsweeklies now reap $100 million-plus in combined revenue, and their 3.5 million total circulation includes an enviable share of the coveted 18- to 35-year-old market, the very group that daily newspapers are having the most trouble attracting. The alternative weeklies group admitted seven new papers this year and had applications from seven more. Cartoonists Lynda J. Barry, whose work appears in Esquire, Mother Jones and other magazines, and Matt Groening, of "The Simpsons," started with alternative papers and still do weekly strips for them.

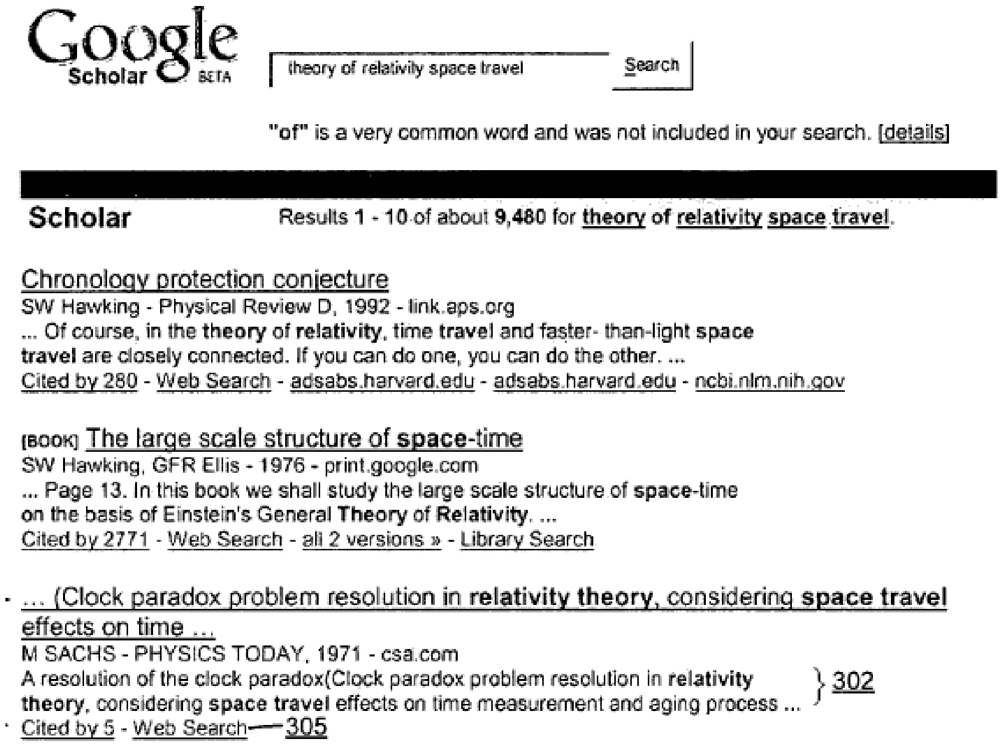
[Click here to get the whole article]

[2]

School Days: Break the Teaching Monopoly ---- By C. Emily Feistritzer

There are vast numbers of adults with at least a bachelor's degree who want to teach. Many have advanced degrees and years of successful experience in other careers. That means one has to go to college and complete a series of education courses approved by the state department of education. Alternative teacher certification programs are not wanting for detractors. Since last week, after an article about our study of alternative teacher certification appeared in the New York Times, we've gotten scores of phone calls from people who want to know where the alternative certification programs are -- they included a dentist who wants to teach biology, an international businessman with a master's degree in physics who wants to teach high school physics and math, and a corporate executive who wants to teach elementary school.

[Click here to get the whole article]



"of" is a very common word and was not included in your search. [details]

Scholar Results **1** - **10** of about **9,480** for **theory** of **relativity space travel**.

Chronology protection conjecture
SW Hawking - Physical Review D, 1992 - link.aps.org
... Of course, in the **theory** of **relativity**, time **travel** and faster- than-light **space travel** are closely connected. If you can do one, you can do the other. ...
Cited by 280 - Web Search - adsabs.harvard.edu - adsabs.harvard.edu - ncbi.nlm.nih.gov

[BOOK] The large scale structure of **space**-time
SW Hawking, GFR Ellis - 1976 - print.google.com
... Page 13. In this book we shall study the large scale structure of **space**-time on the basis of Einstein's General **Theory** of **Relativity**. ...
Cited by 2771 - Web Search - all 2 versions » - Library Search

... (Clock paradox problem resolution in **relativity theory**, considering **space travel** effects on time ...
M SACHS - PHYSICS TODAY, 1971 - csa.com
A resolution of the clock paradox(Clock paradox problem resolution in **relativity theory**, considering **space travel** effects on time measurement and aging process ...
Cited by 5 - Web Search

Tombros and Sanderson. 1998. Advantages of query biased summaries in information retrieval.

Answer Retrieval

- Factoid QA
  - Started in TREC QA track, 1999
  - Retrieving short answers (typically entities) for a limited set of (popular) questions
    - e.g., “where”, “who”, “when”
  - Questions often classified by entity type of answer
    - e.g., *time*, *money*, *person*, *place*, *quantity*
    - e.g., “Where was Roger Federer born?”
  - Relied on initial answer passage or sentence retrieval
  - Extractive QA from the web and knowledge bases

Answer Retrieval

- Community-based Question Answering (CQA)
    - e.g., Yahoo Answers
    - People answer other peoples’ questions
    - Generates huge archives of questions and answers
    - More general questions than factoid QA and answers often one or more paragraphs
    - FAQs and forums also provide large archives and questions and answers (e.g., Stack Overflow)
    - Non-factoid QA, not extractive

Answer Retrieval

- Berger et al (2000) used a translation model to retrieve answers written for FAQs
- Jeon (2005) retrieved similar questions to improve answer retrieval in CQA services
- Jeon (2006) used a feature-based model to predict answer quality for CQA
- Xue (2008) tested translation-based retrieval models for answer archives for CQA
- Surdeanu et al (2011) studied a range of features in a learning to rank framework used to rank potential answers for CQA

Conversational Answer Retrieval

(from SWIRL 2012)

- Open-domain, natural language text questions
- Answers extracted from the corpus (or corpora) being searched, and may be at different levels of granularity, depending on the question
  - Focus on passage-level answers
- Dialogue is about questions and answers, including history, with the aim of refining the understanding of questions and improving the quality of answers
- Evaluated as an open-domain IR task, in contrast to conversational chat or template-based conversation

Research Challenges for CAR

- Tasks
  - Breaking down the research required into manageable pieces
- Test Collections
  - Creating test collections that capture aspects of conversational retrieval for training and testing
- Evaluation
  - Creating (or agreeing on) measures that can be used for evaluating multi-turn, conversational interactions directed at addressing information needs

Answer Passage Retrieval

- Keikha et al (2014)
- CIIR answer passage collection (WebAP)
  - Based on TREC GOV2 web collection and “description” queries
    - e.g., “What evidence is there that aspirin may help prevent cancer?”
  - 82 queries selected as likely to have answers
  - Answer passages (av. 45 words) annotated manually
    - using relevant web pages in top 50 ranked by SDM
  - Over 8,000 passages annotated (av. 97 per query)
    - 43% “perfect”
    - 44% “excellent”
    - 10% “good”
    - 3% “fair”
  - Reasonable annotator agreement

GOV2 Queries

Which ones might have passage-level answers?

714	61	What restrictions are placed on older persons renewing their drivers' licenses in the U.S.?
715	80	What organizations (private or governmental) are developing drugs to combat schizophrenia?
716	66	Have any spammers been arrested or sued for sending unsolicited e-mail?
717	537	What states or localities offer programs for gifted and talented students?
718	617	What methods are used to control acid rain and its effects?
719	315	What kinds of harm do cruise ships do to sea life such as coral reefs, and what is the extent of the damage?
720	561	Find documents about Federal welfare reform legislation, regulation, and policy.
721	362	What applications are there for U.S. decennial census data, and how is it used?
722	203	In what ways does Iran support terrorism?
723	109	What is the U.S. government's definition of "executive privilege?"
724	30	What was the Iran Contra scandal and what were the consequences?
725	211	What would cause a lowered white blood cell count?

Q705. Identify any efforts, proposed or undertaken, by world governments to seek reduction of Iraq's foreign debt.

Document: GX019-35-14384668.html

Preview of Iraq Donors' Conference in Madrid, October 23-23, 2003

MR. DENIG: Good afternoon, and welcome to the Washington Foreign Press Center. We hope to be connected with London as well in a few minutes. We are very pleased today to have two experts to provide a preview for us of the Iraq donors conference in Madrid, which will be going on tomorrow and Friday. We have, first of all, Under Secretary of State for Economic, Business and Agricultural Affairs Al Larson, and we have Under Secretary of the Treasury for International Affairs John Taylor. Each one of them will have a brief opening statement to make, and then we'll be glad to take your questions.

…

UNDER SECRETARY TAYLOR: With respect to the assets, the -- I would actually begin by referring to the Security Council Resolution, Paragraph 24 here, which calls on the member states to remember their obligations to immediately cause the transfer of these funds, these funds that Saddam Hussein and his regime took out of the country, and to return it to the development fund for Iraq for the benefit of the Iraqi people. So that call is out there.

The United States has sent well over a billion dollars back of this money to the pay the Iraqi people. The Japanese have begun to do that. More effort needs to be done to return those assets that Saddam took out of the country and return them to the rightful owners in Iraq. We're working on that.

On the debt, there's a lot of progress that's being made on the effort to get a substantial reduction in the value of the debt. The debt is very high. We're getting more and more information about the size of it. The G8 governments, including Russia, agreed not to accept any payments on the debt, at least through the end of 2004, and in Dubai, the G7 governments agreed to resolve the debt issue by next year so that there can be a clear vision in front of the Iraqi people so they don't have the burden of this in front of them. But that is something that's ongoing, and the process is in place.

MR. DENIG: Let's got to Turkey up here, please.

Evaluating Answer Passage Retrieval

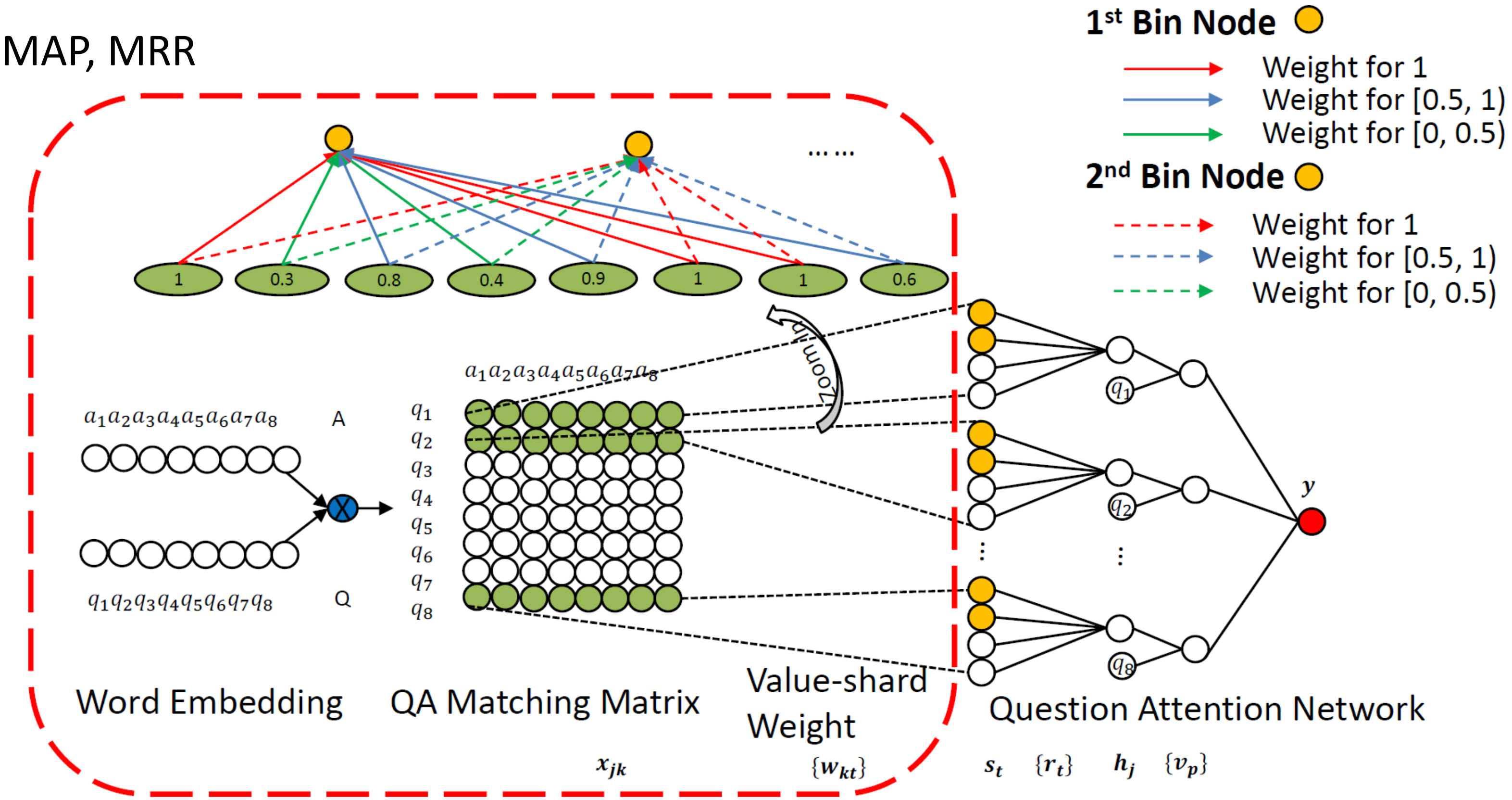
- TREC QA factoid retrieval relatively easy to evaluate
  - Answer sentence or not
  - P@N (Precision at rank N), MRR (Mean reciprocal rank), MAP (Mean Average Precision)
- CQA data usually produces low values
  - One “right” answer per question
- Defining evaluation metrics for passages has been a long-standing problem
  - Boundaries of passages are ill-defined – different models retrieve different passages
  - Manual annotation is very expensive – many more paragraphs than documents
  - Character-based overlap measures can be difficult to understand
  - Word overlap measures (e.g., Rouge) are indicative but indirect
  - Assessing relevance for short text fragments can be very vague

Baselines for Answer Passage Retrieval

- Comparing standard passage retrieval models for the task of answer passage retrieval
  - Query likelihood
  - Sequential dependence model
  - Bendersky-Kurland interpolation of passage and document scores
  - Positional language model with different kernel functions
  - Pseudo-relevance feedback
- Overlapped windows of 50 words used for passages
- Evaluated using character-level measures and Rouge overlap
- Bottom line: Term-based retrieval models are not good at finding answers

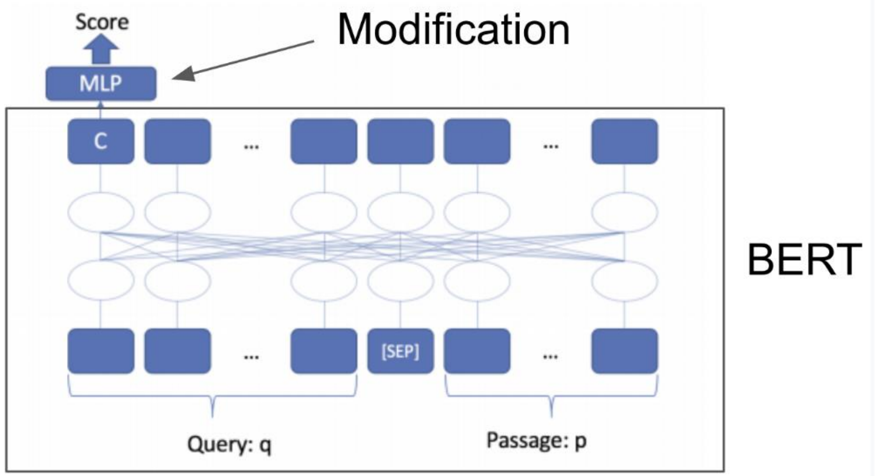
Answer Retrieval with Neural Models

- **Paper:** Yang, Ai, Guo, and Croft. 2016. aNMM: Ranking Short Answer Texts with Attention-Based Neural Matching Model.
- **Test Collection:** TREC QA, Yahoo CQA
- **Evaluation:** MAP, MRR
- **Model:**



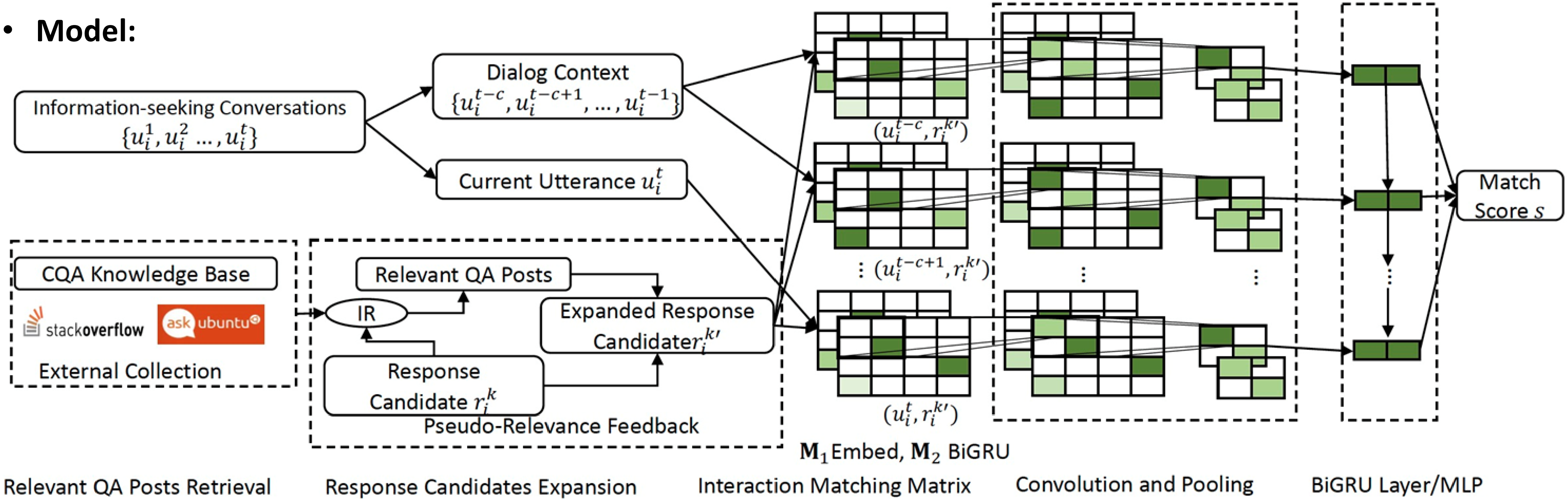
SOTA for Answer Retrieval

- BERT is the leading approach by far
  - 30-40% improvement across all metrics on MS MARCO, CQA collections
  - Continues trend of greater interaction/attention in Pre-BERT models
  - Difficult to specify the best configuration of BERT as the training method/fine tuning/performance distributions are often not provided in enough detail
  - Results on reading comprehension promising for extractive retrieval
  - P@1 is still only .25 on MARCO (.7 ON WikipassageQA)



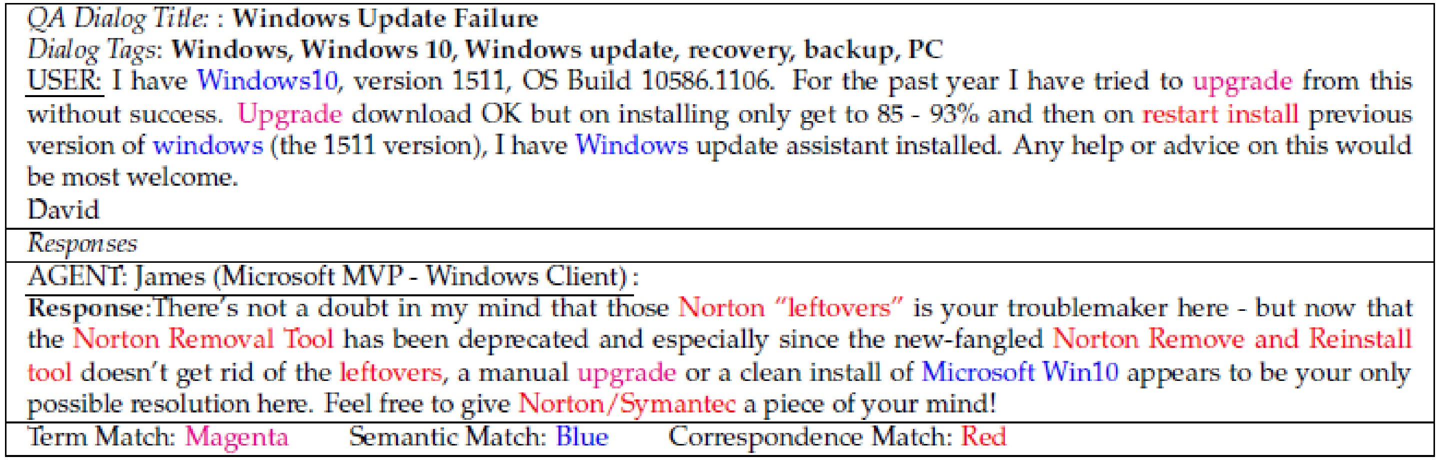
Response Retrieval

- **Paper:** Yang, Qiu, Qu, Guo, Zhang, Croft, Huang, and Chen. 2018. Response Ranking with Deep Matching Networks and External Knowledge in Information-seeking Conversation Systems.
- **Test Collection:** UDC, MSDialog, AliMe
- **Evaluation:** MAP, Recall@1, 2, 5
- **Model:**



Information-seeking Conversations $\{u_i^1, u_i^2 \dots, u_i^t\}$
Dialog Context $\{u_i^{t-c}, u_i^{t-c+1}, \dots, u_i^{t-1}\}$
Current Utterance u_i^t
CQA Knowledge Base
stackoverflow
ask ubuntu
External Collection
IR
Relevant QA Posts
Expanded Response Candidate$r_i^{k'}$
Response Candidate r_i^k
Pseudo-Relevance Feedback
$(u_i^{t-c}, r_i^{k'})$
$(u_i^{t-c+1}, r_i^{k'})$
$(u_i^t, r_i^{k'})$
$\mathbf{M}_1$ Embed, $\mathbf{M}_2$ BiGRU
Match Score s
Relevant QA Posts Retrieval
Response Candidates Expansion
Interaction Matching Matrix
Convolution and Pooling
BiGRU Layer/MLP

Response Retrieval

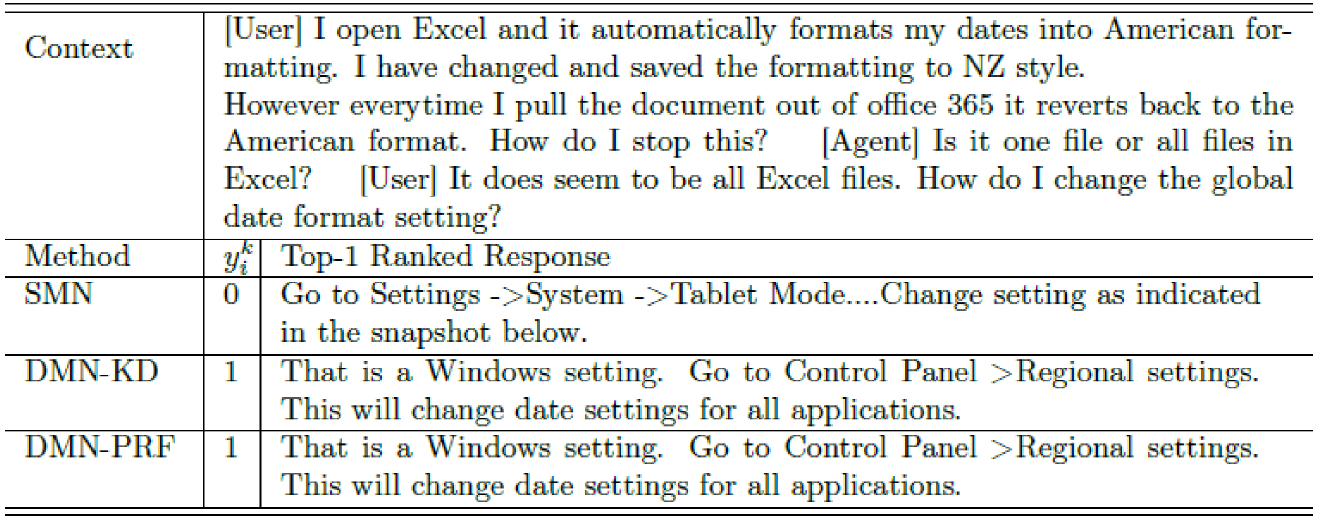


QA Dialog Title: : **Windows Update Failure**
Dialog Tags: **Windows, Windows 10, Windows update, recovery, backup, PC**
USER: I have Windows10, version 1511, OS Build 10586.1106. For the past year I have tried to upgrade from this without success. Upgrade download OK but on installing only get to 85 - 93% and then on restart install previous version of windows (the 1511 version), I have Windows update assistant installed. Any help or advice on this would be most welcome.
David

Responses

AGENT: James (Microsoft MVP - Windows Client) :
Response:There's not a doubt in my mind that those Norton "leftovers" is your troublemaker here - but now that the Norton Removal Tool has been deprecated and especially since the new-fangled Norton Remove and Reinstall tool doesn't get rid of the leftovers, a manual upgrade or a clean install of Microsoft Win10 appears to be your only possible resolution here. Feel free to give Norton/Symantec a piece of your mind!

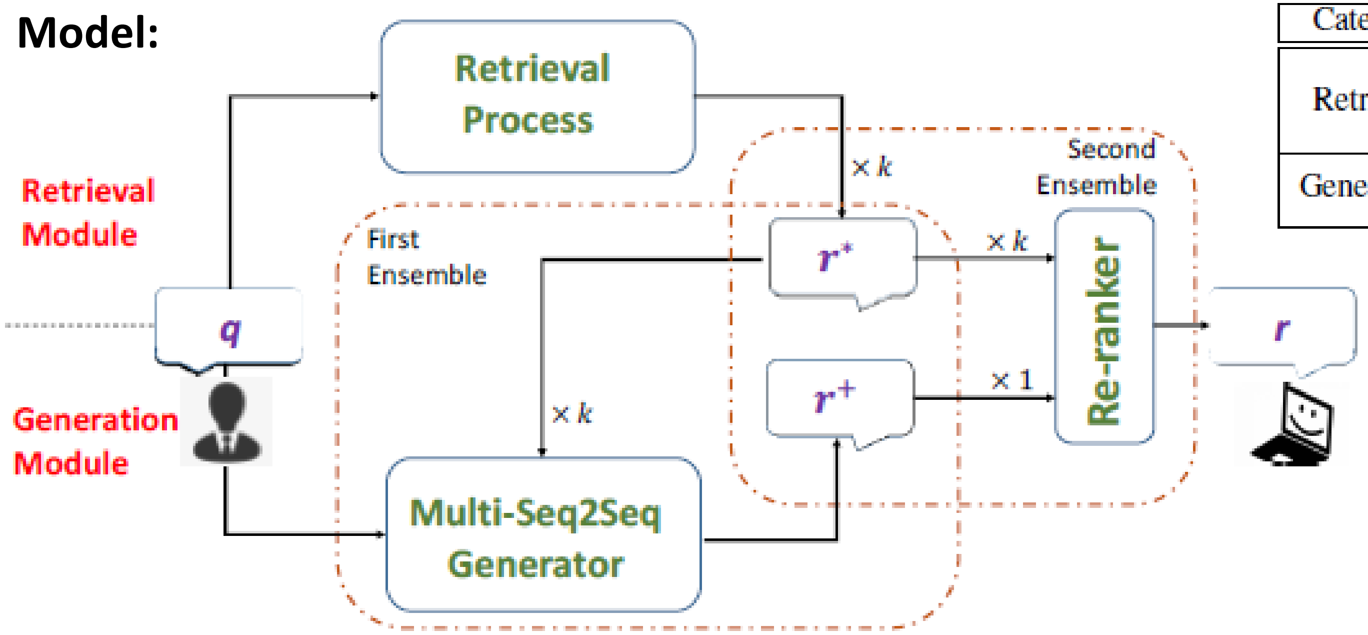
Term Match: Magenta Semantic Match: Blue Correspondence Match: Red

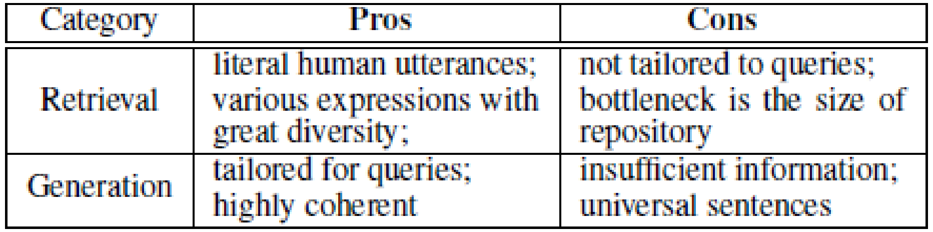


Context	[User] I open Excel and it automatically formats my dates into American formatting. I have changed and saved the formatting to NZ style. However everytime I pull the document out of office 365 it reverts back to the American format. How do I stop this? [Agent] Is it one file or all files in Excel? [User] It does seem to be all Excel files. How do I change the global date format setting?	
Method	y_i^k	Top-1 Ranked Response
SMN	0	Go to Settings ->System ->Tablet Mode....Change setting as indicated in the snapshot below.
DMN-KD	1	That is a Windows setting. Go to Control Panel >Regional settings. This will change date settings for all applications.
DMN-PRF	1	That is a Windows setting. Go to Control Panel >Regional settings. This will change date settings for all applications.

Hybrid Response Generation/Retrieval

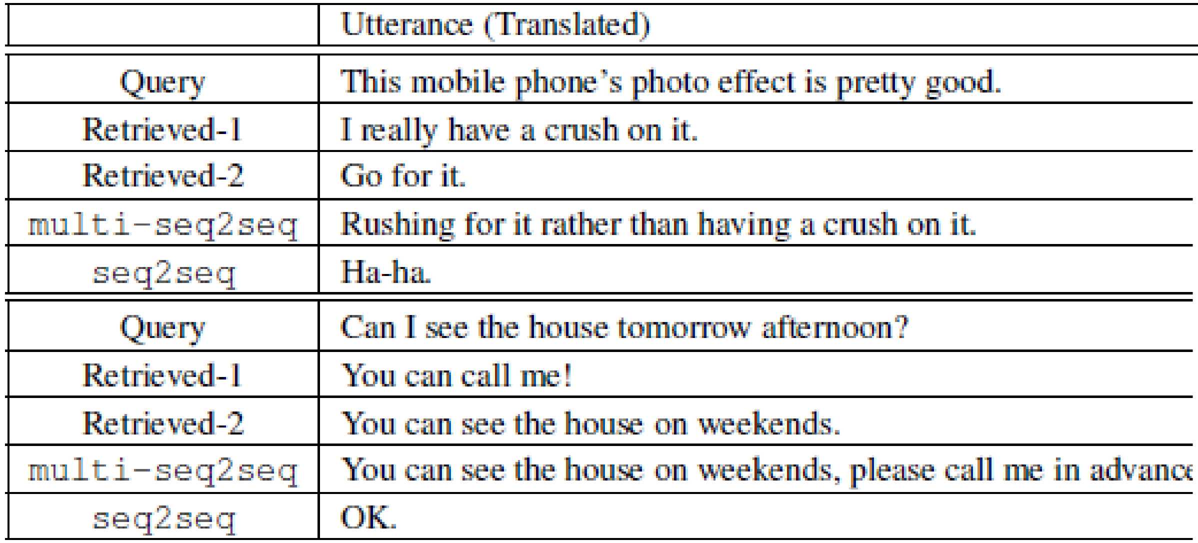
- **Paper:** Song, Li, Nie, Zhang, Zhao, and Yan. 2018. An Ensemble of Retrieval-Based and Generation-Based Human-Computer Conversation Systems.
- **Test Collection:** Wiebo, Tieba, Twitter/Foursquare (Ghazvininejad et al, A Knowledge-Grounded Neural Conversation Model. In AAAI ’18)
- **Evaluation:** Bleu, Rouge-L, human
- **Model:**



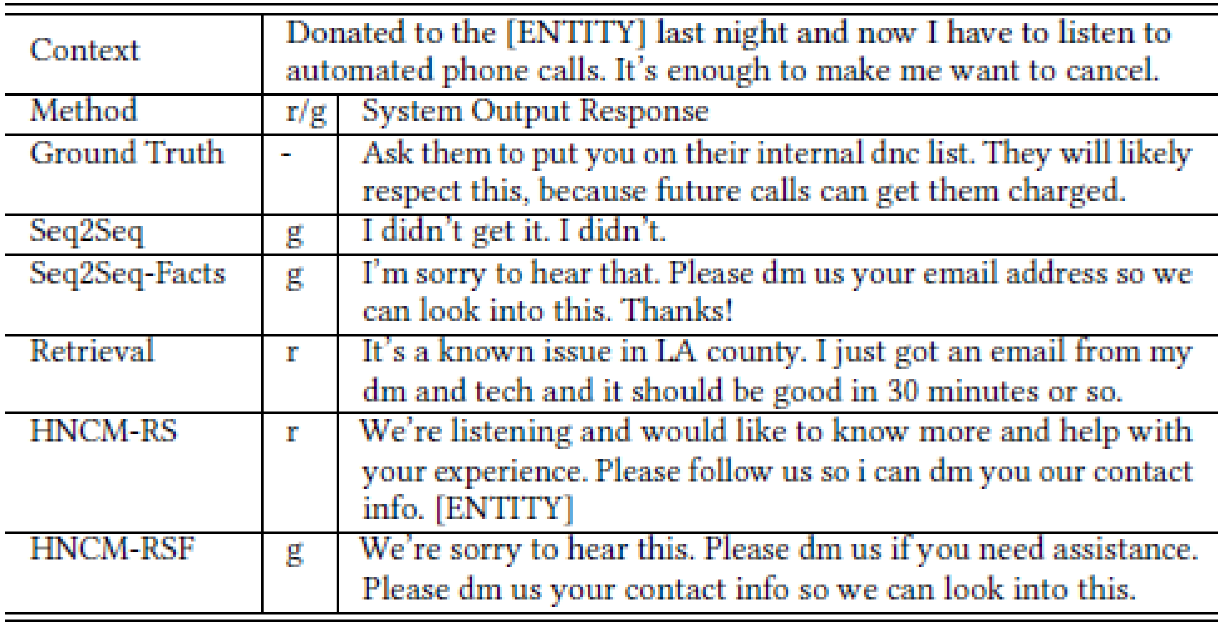


Category	Pros	Cons
Retrieval	literal human utterances; various expressions with great diversity;	not tailored to queries; bottleneck is the size of repository
Generation	tailored for queries; highly coherent	insufficient information; universal sentences

Hybrid Response Generation/Retrieval



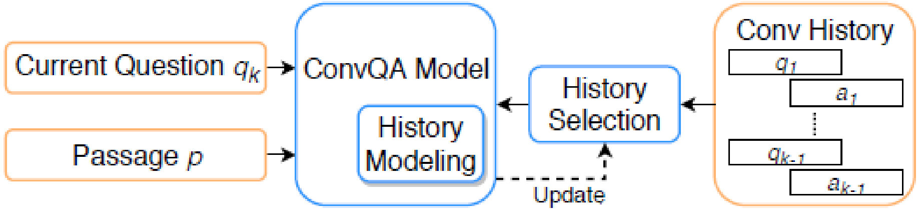
	Utterance (Translated)	**Selected by re-ranker**
Query	This mobile phone's photo effect is pretty good.	
Retrieved-1	I really have a crush on it.	
Retrieved-2	Go for it.	
`multi-seq2seq`	Rushing for it rather than having a crush on it.	
`seq2seq`	Ha-ha.	
Query	Can I see the house tomorrow afternoon?	
Retrieved-1	You can call me!	
Retrieved-2	You can see the house on weekends.	
`multi-seq2seq`	You can see the house on weekends, please call me in advanc	
`seq2seq`	OK.	

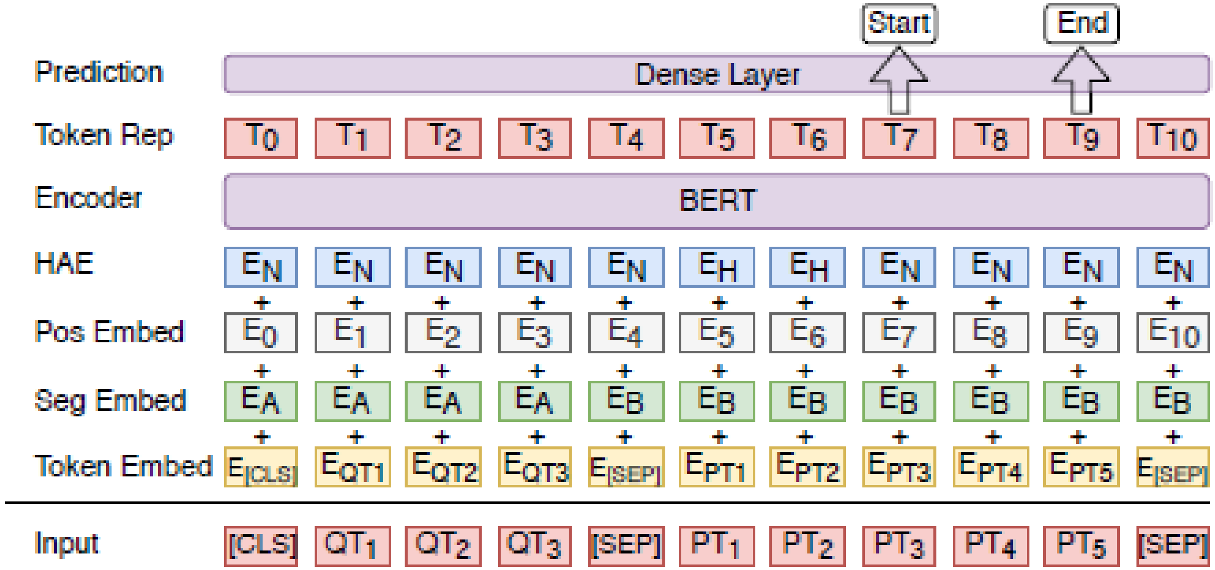


Context	Donated to the [ENTITY] last night and now I have to listen to automated phone calls. It's enough to make me want to cancel.	
Method	r/g	System Output Response
Ground Truth	-	Ask them to put you on their internal dnc list. They will likely respect this, because future calls can get them charged.
Seq2Seq	g	I didn't get it. I didn't.
Seq2Seq-Facts	g	I'm sorry to hear that. Please dm us your email address so we can look into this. Thanks!
Retrieval	r	It's a known issue in LA county. I just got an email from my dm and tech and it should be good in 30 minutes or so.
HNCM-RS	r	We're listening and would like to know more and help with your experience. Please follow us so i can dm you our contact info. [ENTITY]
HNCM-RSF	g	We're sorry to hear this. Please dm us if you need assistance. Please dm us your contact info so we can look into this.

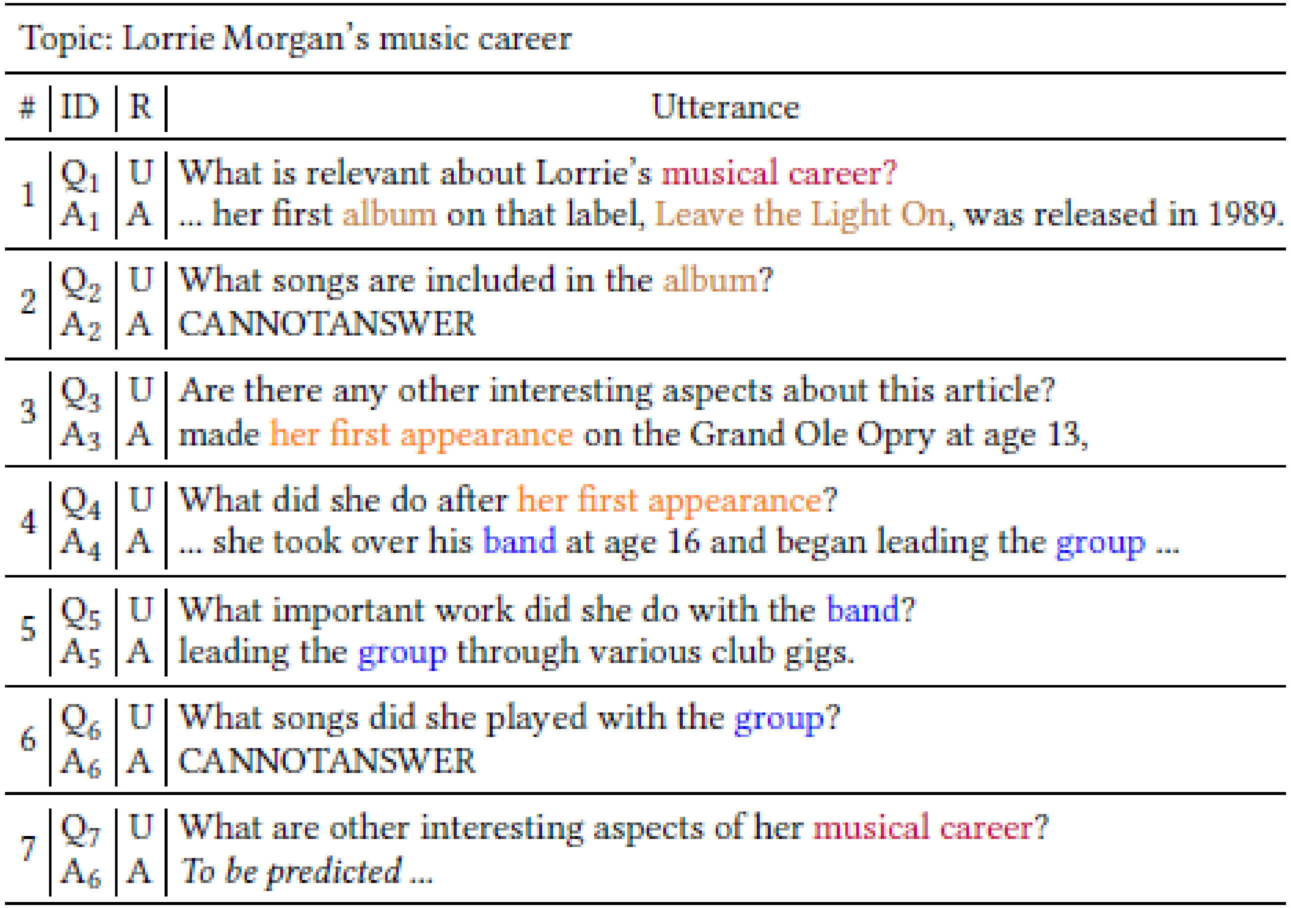
Conversational Question Answering

- **Paper:** Qu, Yang, Qiu, Croft, Zhang, and Iyer, 2019. BERT with History Answer Embedding for Conversational Question Answering.
- **Test Collection: QuAC dataset**
- **Evaluation: F1, HEQ-Q, HEQ-D**
- **Model:**




Start
End
Prediction
Dense Layer
Token Rep
T_0 T_1 T_2 T_3 T_4 T_5 T_6 T_7 T_8 T_9 T_{10}
Encoder
BERT
HAE
E_N E_N E_N E_N E_N E_H E_H E_N E_N E_N E_N
Pos Embed
E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}
Seg Embed
E_A E_A E_A E_A E_B E_B E_B E_B E_B E_B E_B
Token Embed
$E_{[CLS]}$ E_{QT1} E_{QT2} E_{QT3} $E_{[SEP]}$ E_{PT1} E_{PT2} E_{PT3} E_{PT4} E_{PT5} $E_{[SEP]}$
Input
[CLS] QT_1 QT_2 QT_3 [SEP] PT_1 PT_2 PT_3 PT_4 PT_5 [SEP]

Conversational Question Answering



Topic: Lorrie Morgan's music career

#	ID	R	Utterance
1	Q_1	U	What is relevant about Lorrie's musical career?
	A_1	A	... her first album on that label, Leave the Light On, was released in 1989.
2	Q_2	U	What songs are included in the album?
	A_2	A	CANNOTANSWER
3	Q_3	U	Are there any other interesting aspects about this article?
	A_3	A	made her first appearance on the Grand Ole Opry at age 13,
4	Q_4	U	What did she do after her first appearance?
	A_4	A	... she took over his band at age 16 and began leading the group ...
5	Q_5	U	What important work did she do with the band?
	A_5	A	leading the group through various club gigs.
6	Q_6	U	What songs did she played with the group?
	A_6	A	CANNOTANSWER
7	Q_7	U	What are other interesting aspects of her musical career?
	A_6	A	*To be predicted ...*

Time to take a step back and consider the big picture…

Answers or Documents?

- Documents contain answers to many possible questions
- SERPs present a range of answers to the likely underlying questions
- A list of documents is only a satisfactory answer to one type of information need
- Questions and answers are the natural communication tools for solving information needs
  - Document retrieval is only an intermediate step
- However, we know a lot about how people interact with lists of documents but very little about how they interact with potential answers

Answers or Documents?

- Answer passages are not just “little documents”
  - Text should have a strong relationship to the question
- Techniques developed for document retrieval may not be appropriate for answer retrieval
  - Ranked lists
  - Relevance feedback
  - Diversification
  - Evaluation

Relevance or Correctness?

- Relevance is at the core of most IR evaluation – but what is it?
  - Topical relevance, user relevance…
  - Still being debated
- PEGFB judgments for queries are difficult for users and require significant interpretation
- Correctness of an answer for a question well understood by crowdsourcing annotators
  - Definition of an answer?
- Disagreement is about the quality of an answer (and the text spans)
  - PEGFB generally makes more sense

Ranking or Interaction?

- Ranked lists of answers may not be an appropriate presentation
  - … snippets?
  - Answer confidence more important?
- Bandwidth may limit response to a single answer
- Interaction is a natural part of question and answer dialogue
  - e.g., clarification questions, feedback
- Identifying similar, redundant, alternative answers is similar to document diversification but requires more than term matching
- Negative feedback is particularly important for answers, but no guidance from previous work with documents

Tasks (or Challenges)

- Given a (non-factoid) question, find the best answers in a collection of answers
  - Ranking, P@1, determining confidence…
- Given a question, find the best answer passages in a collection of documents
  - Gold standard, overlap, relevant documents…
- Given a question and answer dialogue, find the best answer in a collection of answers or documents
  - Predict conversation response, partial history, session history…
- Given a question, determine the best answer across a range of granularities in a collection of documents
  - Single answers, sets of answers, summaries…

Tasks

- Given a set of retrieved answers, group them into categories
  - Redundant, similar, instances, alternatives…
- Given a top-ranked answer that is incorrect, rerank based on user feedback
  - Yes/no, word-based, entity-based, conversation response…

QA Test Collections

- TREC QA: 1.5K factoid questions with 60K paired potential answer sentences
- Yahoo L6 Webscope: 4.5M questions and associated answer passages from CQA service (Manner Questions subset: 150K "how" questions)
- WikiQA: 3K factoid questions with 30K answer sentences from associated Wiki page
- MS MARCO: 1M factoid questions from Bing log with 9M "companion" passages and 180K manually generated answers
- SQUAD: 100K manually generated questions with associated answers that are text spans in 530 Wikipedia articles
- WebAP: 8K text span answer passages (av. 45 words) from relevant documents for 80 TREC Gov2 questions
- Yahoo nfL6 subset: 85K non-factoid question and answer pairs
- WikiPassageQA: 4K non-factoid queries and answer passages created from 860 Wikipedia pages
- ANTIQUE: 2.5K questions from nfL6 with more complete relevance judgments

Conversation Test Collections

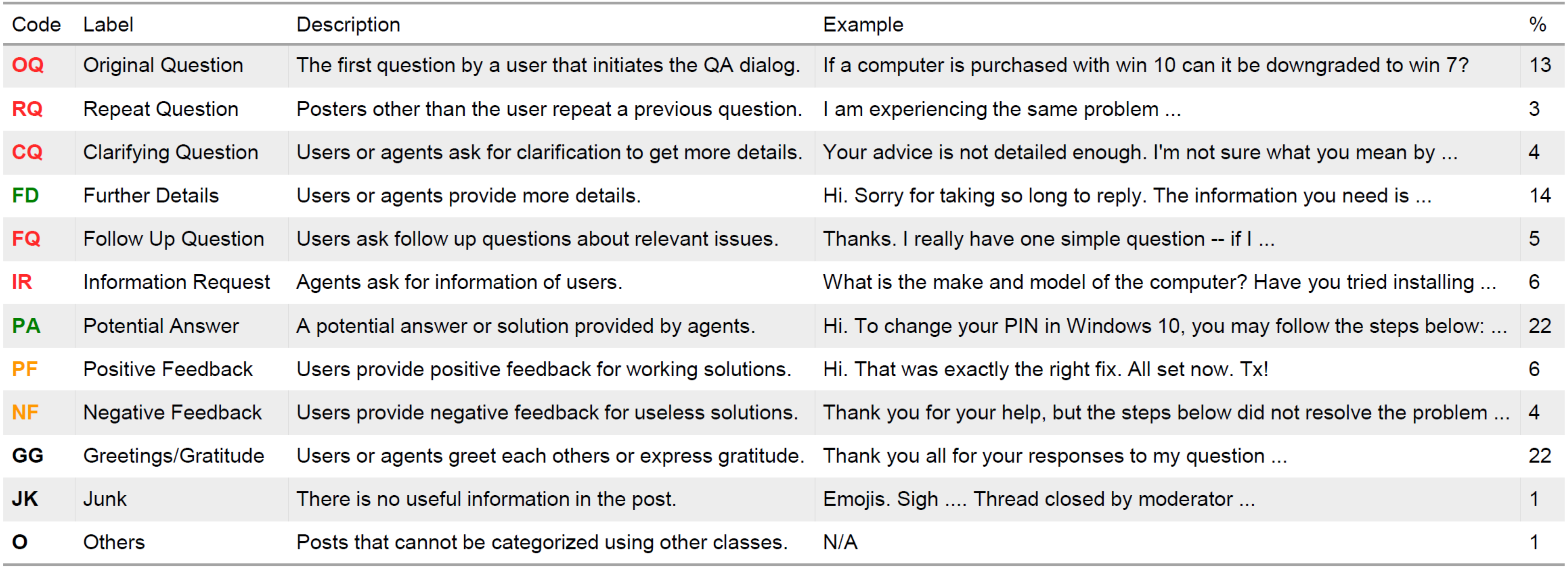
- Ubuntu (UDC): 1M conversations from technical support chat logs
- QuAC: 14K crowdsourced QA dialogs based on Wikipedia articles
- MSDialog: 35K conversations from MS technical support forum, 2K labelled with utterance intent
- AliMe: 63K context-response pairs from commercial online help chatbot (Chinese)
- Qulac: 10K crowdsourced clarifying question-answer pairs related to 200 TREC topics
- Amazon: Simulated product purchase conversations based on product facets
- MSMARCO Conversational Search: 45M user sessions containing 340K unique queries
- TREC CASt: New TREC track building on MSMARCO, others

Back

Crowdsourcing

- As IR researchers, we should be interested in more than results on leaderboards
- To study answers in more detail, crowdsourcing experiments are needed
- Examples:
  - Qu, Yang, Croft, Trippas, Zhang, and Qiu. 2018. Analyzing and Characterizing User Intent in Information-seeking Conversations.
  - Qu, Yang, Croft, Scholer, and Zhang. 2019. Answer Interaction in Non-factoid Question Answering Systems.

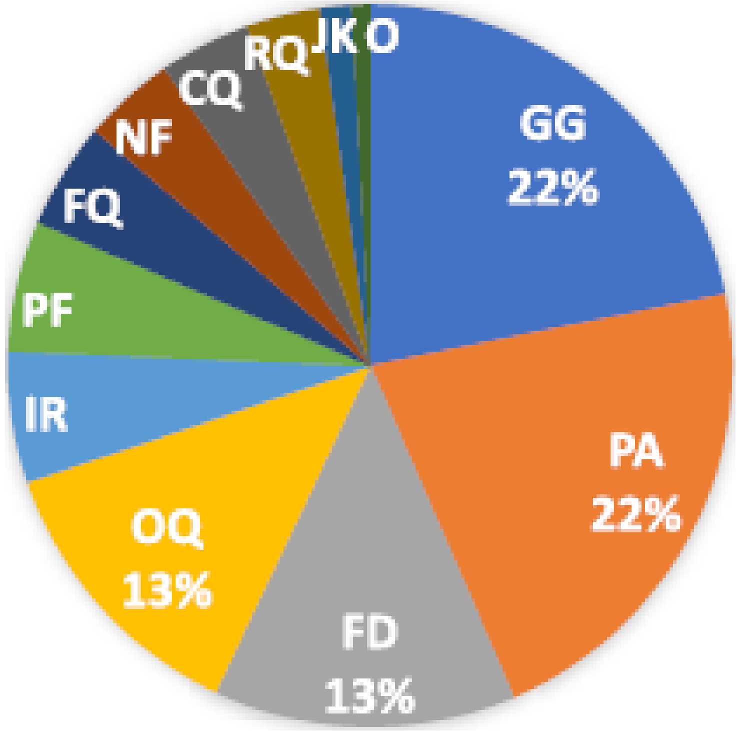
User Intent Taxonomy

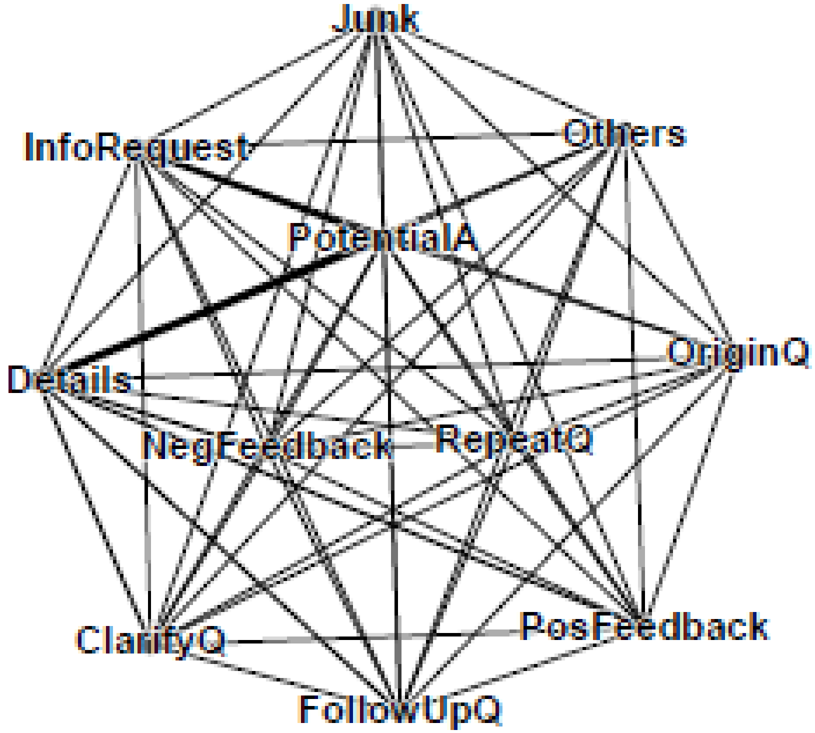


Code	Label	Description	Example	%
OQ	Original Question	The first question by a user that initiates the QA dialog.	If a computer is purchased with win 10 can it be downgraded to win 7?	13
RQ	Repeat Question	Posters other than the user repeat a previous question.	I am experiencing the same problem ...	3
CQ	Clarifying Question	Users or agents ask for clarification to get more details.	Your advice is not detailed enough. I'm not sure what you mean by ...	4
FD	Further Details	Users or agents provide more details.	Hi. Sorry for taking so long to reply. The information you need is ...	14
FQ	Follow Up Question	Users ask follow up questions about relevant issues.	Thanks. I really have one simple question -- if I ...	5
IR	Information Request	Agents ask for information of users.	What is the make and model of the computer? Have you tried installing ...	6
PA	Potential Answer	A potential answer or solution provided by agents.	Hi. To change your PIN in Windows 10, you may follow the steps below: ...	22
PF	Positive Feedback	Users provide positive feedback for working solutions.	Hi. That was exactly the right fix. All set now. Tx!	6
NF	Negative Feedback	Users provide negative feedback for useless solutions.	Thank you for your help, but the steps below did not resolve the problem ...	4
GG	Greetings/Gratitude	Users or agents greet each others or express gratitude.	Thank you all for your responses to my question ...	22
JK	Junk	There is no useful information in the post.	Emojis. Sigh .... Thread closed by moderator ...	1
O	Others	Posts that cannot be categorized using other classes.	N/A	1

Modeling Intent in Search Interactions

- **Test Collection: MSDialog, UDC**



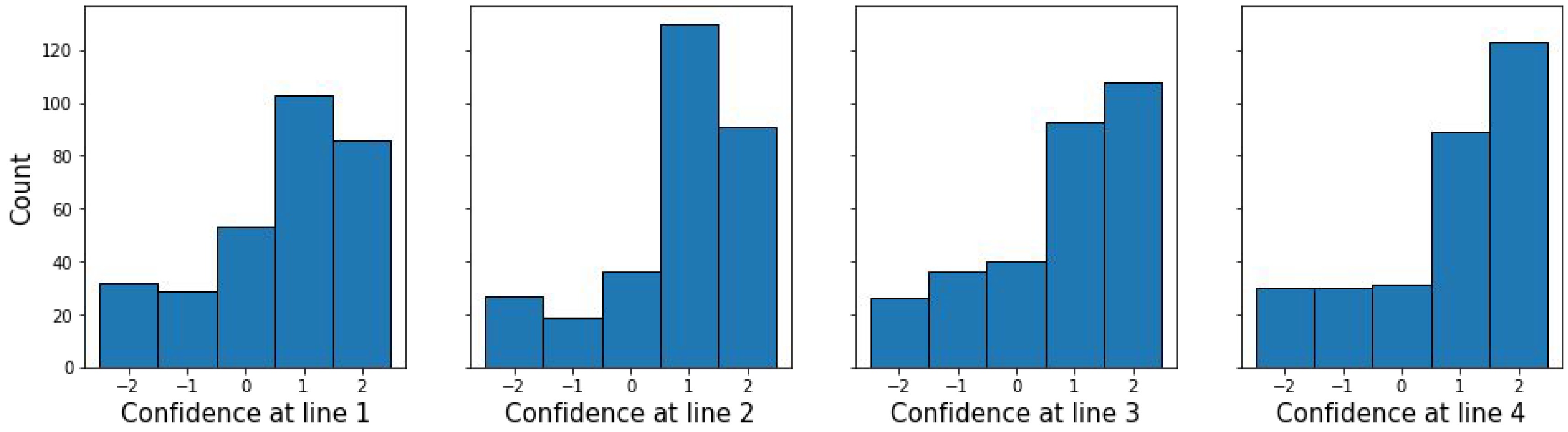


Answer Interaction Study

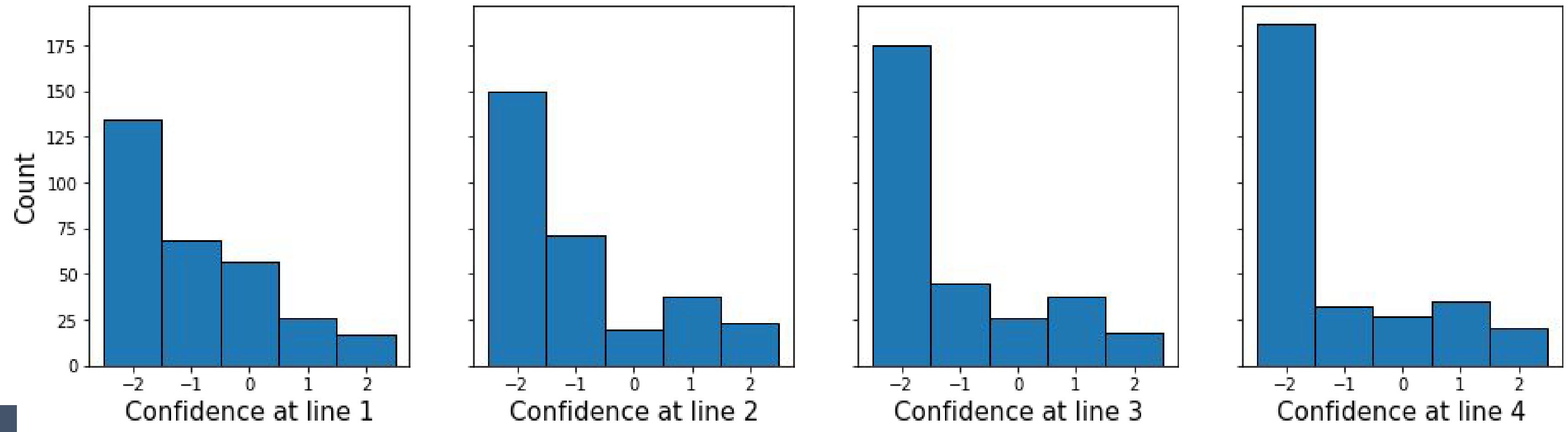
The **turkers are given a question and a short passage**. The 200 QA pairs are from nfl6, one good and one bad (but highly ranked) for each question.

- **“line by line”**: reveals the passage line by line. The turkers indicate their confidence level so far that this passage is a good answer.
- **“passage highlight”**: highlight important words/phrases (sentences are not encouraged) that helped them make their decision -- either positive or negative.
- **“passage highlight (with suggested words)”**: highlight important words/phrases with the presence of system suggested words. Turkers do not have to stick to the suggested words.

Distribution of confidence ratings



Distribution of confidence ratings from line 1 to line 4 for good answers



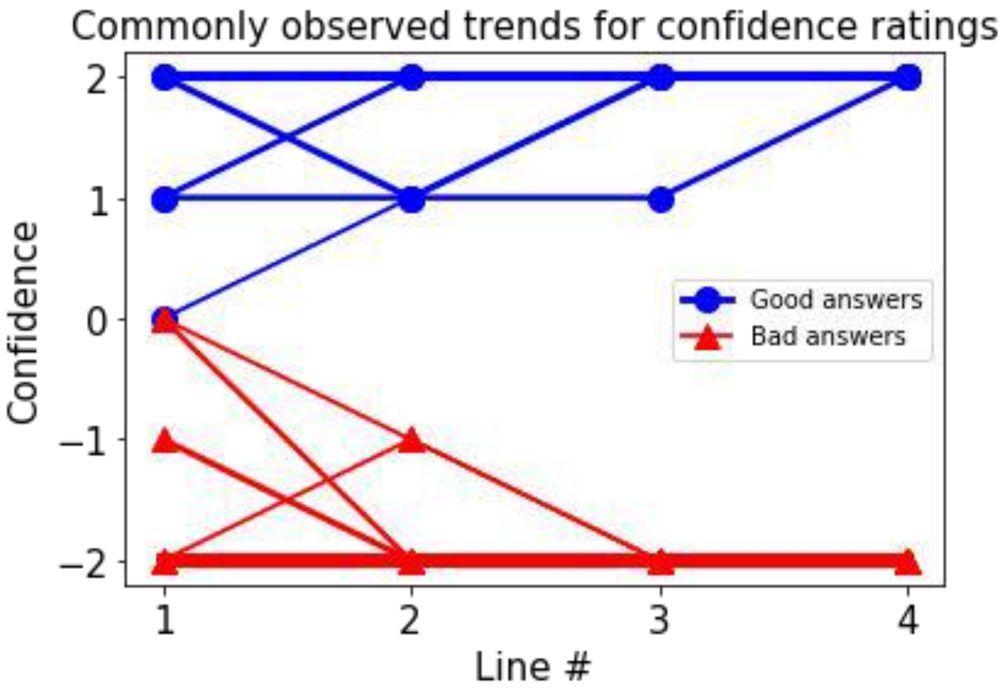
Distribution of confidence ratings from line 1 to line 4 for bad answers

Good answers: have a sense that the answers might be good at the beginning, but hesitate to make a confident rating until the latter half

Bad answers: can determine the answer quality from the very beginning.

Analysis of confidence ratings on a question level

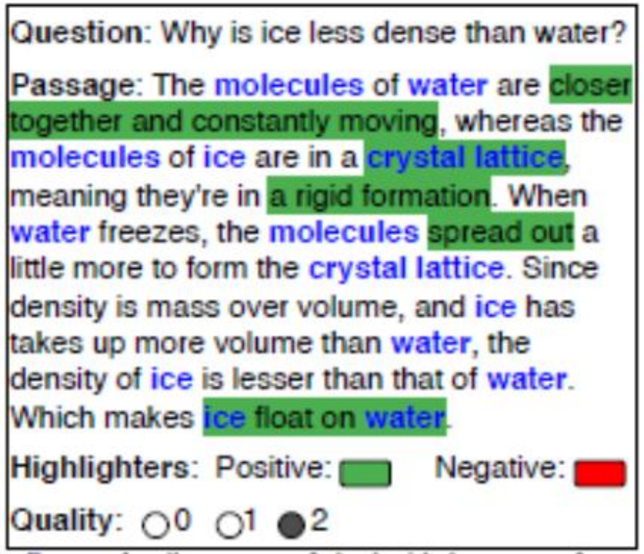
Answer type	Increase	Decrease	Constant	All positive	All negative
Good (out of 100)	24	3	22/17	51	4
Bad (out of 100)	10	19	44/40	8	54



People's initial impressions on answer quality are usually correct, and people become more and more confident on answer quality as they go through the answer.

The “passage highlight” setting

- The turker is given a question and an answer passage and is asked to **highlight positive and negative words** or phrases in the passage.
- At least one highlight for each answer needs to be made. In addition, the turkers are asked to give an overall answer quality.



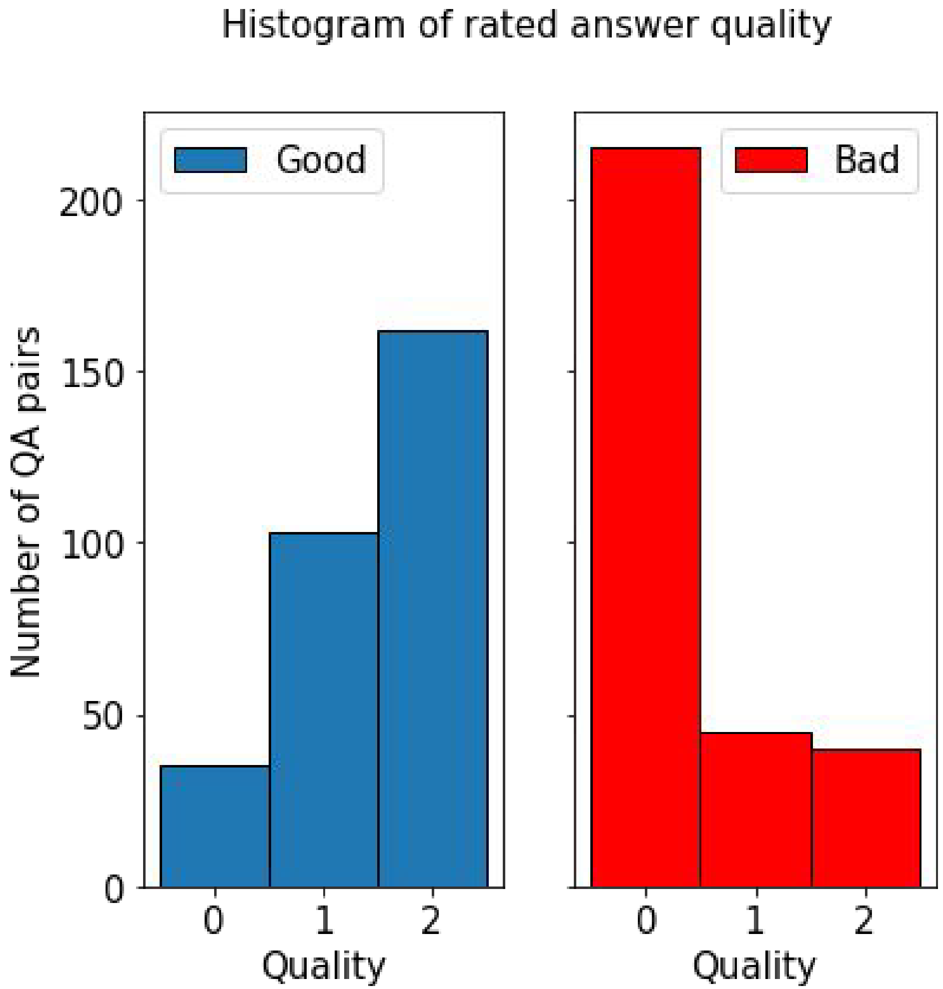
Question: Why is ice less dense than water?

Passage: The molecules of water are closer together and constantly moving, whereas the molecules of ice are in a crystal lattice, meaning they're in a rigid formation. When water freezes, the molecules spread out a little more to form the crystal lattice. Since density is mass over volume, and ice has takes up more volume than water, the density of ice is lesser than that of water. Which makes ice float on water.

Highlighters: Positive: 🟩 Negative: 🟥

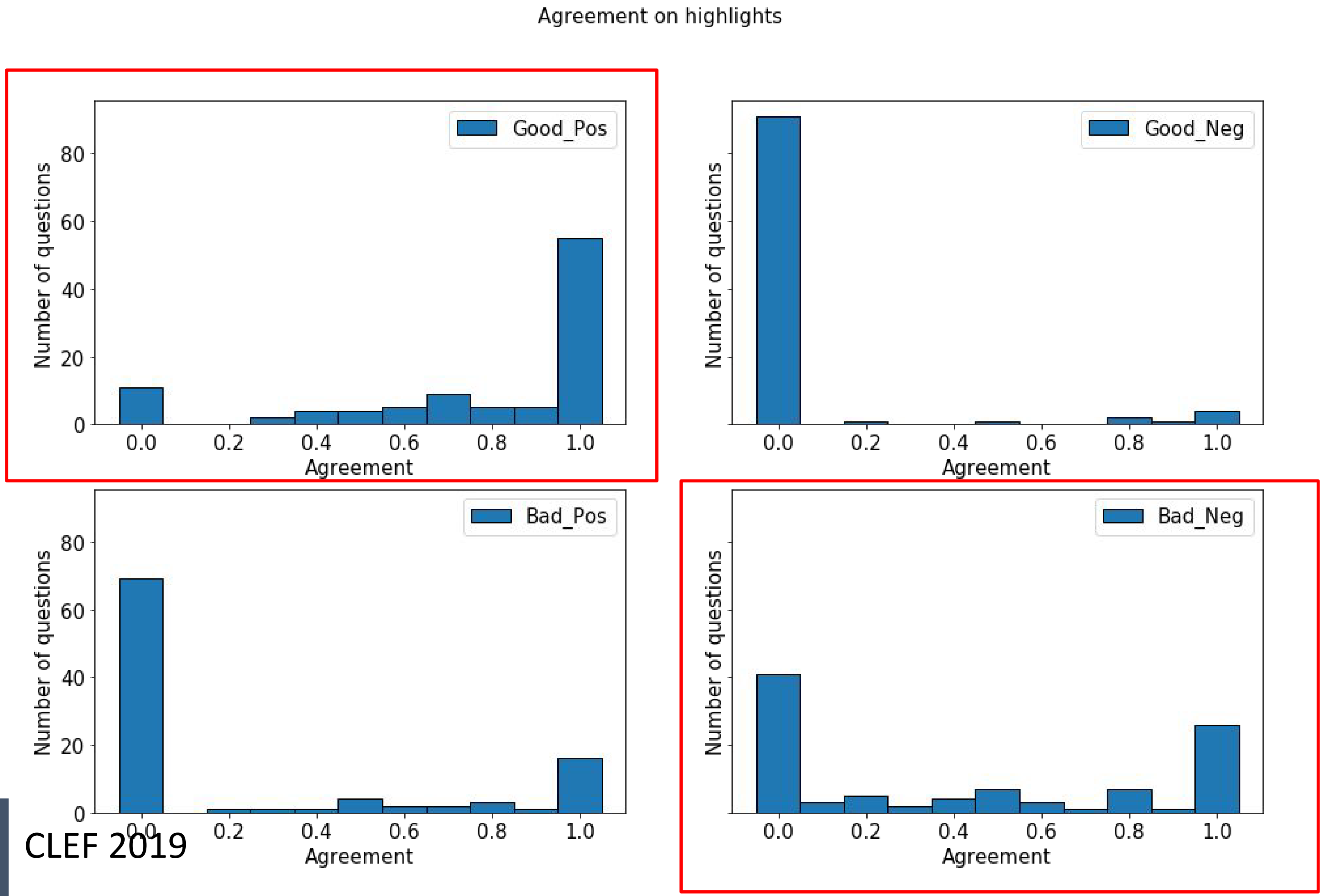
Quality: ○0 ○1 ●2

Distribution of rated answer quality



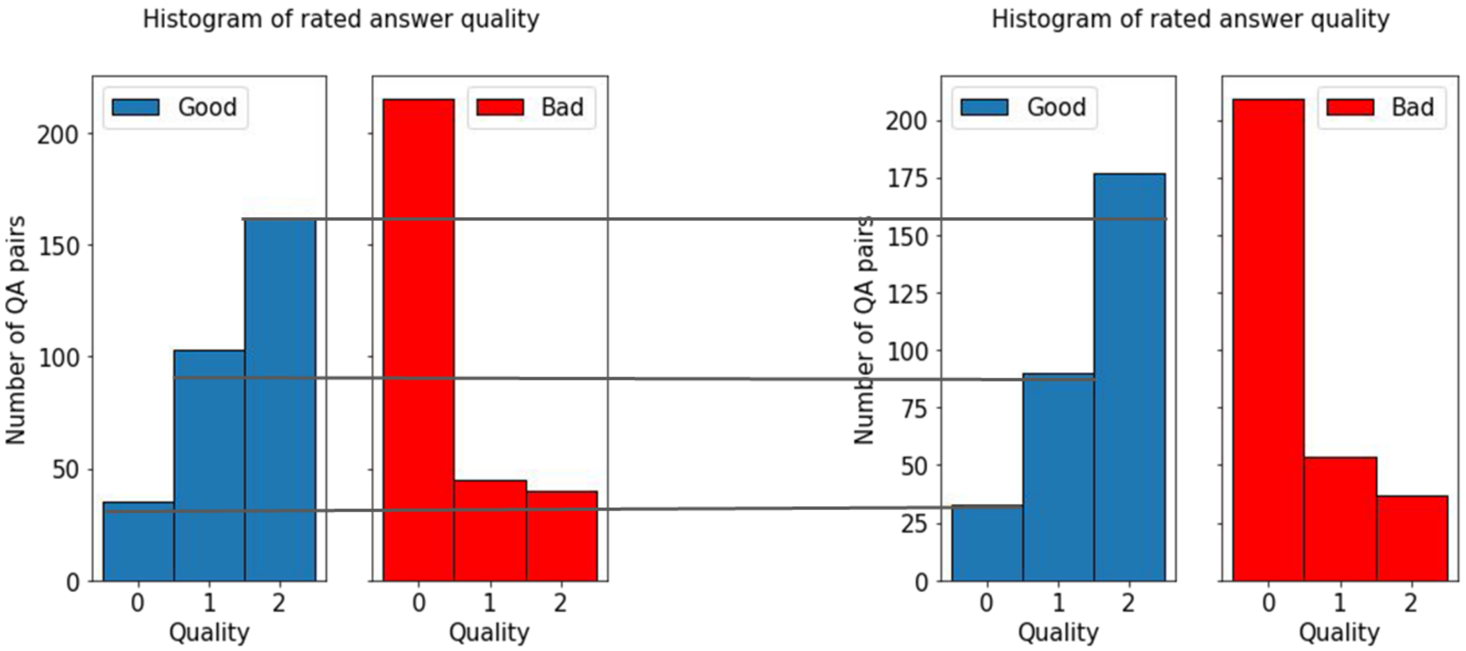
The two figures indicate that turkers are decisive in rating a bad answer. In addition, they can be harsh on the answer quality rating even if the passage is correct.

Quantifying agreement on highlights (Cont'd)



In general, the results indicate that people tend to get good agreement on what makes a good answer good. In contrast, when deciding what makes a bad answer bad, people tend to have more diverse opinions while still manage to achieve an agreement to some extent.

The “passage highlight (suggested words)” setting



“passage highlight”

“passage highlight (with suggested words)”

The suggested words could lead to more decisive evaluations on answer quality.

More Crowdsourcing…

- Using eyetracking to confirm/expand results
- Negative feedback experiments with different interaction modes
- Understanding passage boundaries in documents
- Formulating models of answers and testing them by comparing and categorizing

Summary

- Finding answers in response to questions is the key to progress in information retrieval
- We currently are only beginning to explore the research challenges in dealing with answers rather than documents
- New theories and tasks need to be developed
- New test collections and user studies need to be done
- Collaboration with NLP, ML, and HCI will become increasingly important

CLEF 2019

THANK YOU