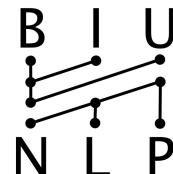


Think out of the box!

Conceptions and Misconceptions in NLP Evaluation

Ido Dagan
Bar-Ilan University



Rationale

- Sometimes - evaluation practices limit research evolution
 - ⇒ We should look out to expand our evaluation protocols, to encourage research expansion
- In other times, certain evaluation aspects might yield misleading results
 - ⇒ While following prior practices, we should always be on the watch

Outline

Four such recent experiences:

- Controlled crowdsourcing for challenging data collection
- Revisiting old “same length” assumption for reference and system summaries
- Extending summarization evaluation to the interactive setting
- Singleton annotations in coreference - consider or ignore?

Conception:
Crowdsourcing is for the crowd

Expert vs. Crowd Annotation

- NLP used to rely on high quality expert annotations
 - Complex guidelines
 - Very costly, hard to replicate
- Crowdsourcing became very attractive, but
 - Tasks should be simplified
 - Data is more noisy
- Can we get the best of both worlds?

Controlled Crowdsourcing for High-Quality QA-SRL Annotation



By: Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou,
Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer and Ido Dagan

ACL 2019



Controlled Crowdsourcing for High-Quality QA-SRL Annotation

- Expert annotators are required.



PropBank

Treebank View

Frameset View

Argument View

0	1	2	3
4	5	A (A)	M-ADV (V)
M-CAU (C)	M-DIR (D)	M-DIS (I)	M-DSP (S)
M-GOL (G)	M-EXT (E)	M-LOC (L)	M-MNR (M)
M-MOD (O)	M-NEG (N)	M-PRD (T)	M-PPR (F)
M-REC (S)	M-SLC (S)	M-TMP (T)	-UNDEF (U)
ERASE (-)			

Figure 1.4: Find and annotate any sisters to the rel



Controlled Crowdsourcing for High-Quality QA-SRL Annotation

- SNLI
- SQuAD
- DocRED





Controlled Crowdsourcing for High-Quality QA-SRL Annotation

- SNLI
- SQuAD
- DocRED



OFTEN, IN SUBTLE TASKS, CROWDSOURCING YIELDS
REDUCED QUALITY, OR AVOIDED ALTOGETHER.



Our experience with QA-SRL

- We discovered some shortcomings in the crowd-sourced dataset
- Which we overcame by proposing a – broadly applicable – methodology for ***Controlled Crowdsourcing***





QA-SRL:

Predicate Argument

Structure for Laymen



Semantic Role Labeling (SRL)



In 1950 Alan M. Turing published “Computing machinery and intelligence” in Mind, in which he proposed that machines could be tested for intelligence using questions and answers

ARG-0: Proposer

ARG-1: Proposition

ARGM-LOC: Locative

ARGM-TMP: Temporal



QA-SRL:

The promise for crowd annotation



In 1950 Alan M. Turing published “Computing machinery and intelligence” in Mind, in which he proposed that machines could be tested for intelligence using questions and answers

Who proposed something?

Where was something proposed?

What did someone propose?

When was something proposed?

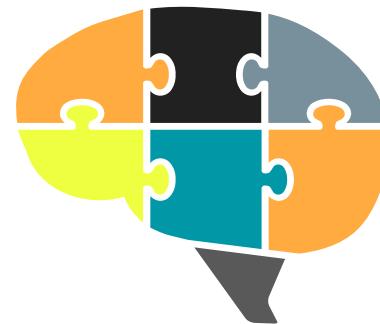
Semantics as layman Q&A

“Who did what to whom, where when and how?”

Annotated by laymen
in large-scale.



Uncover roles that are
naturally understood.





**Did QA-SRL crowdsourcing
yield sufficient quality?**



Were we satisfied with prior crowdsourcing?

Coverage is insufficient

-68% recall vs. expert

1. Annotators often miss some roles.
2. Some training and sensitivity is needed, particularly for annotating implicit roles.



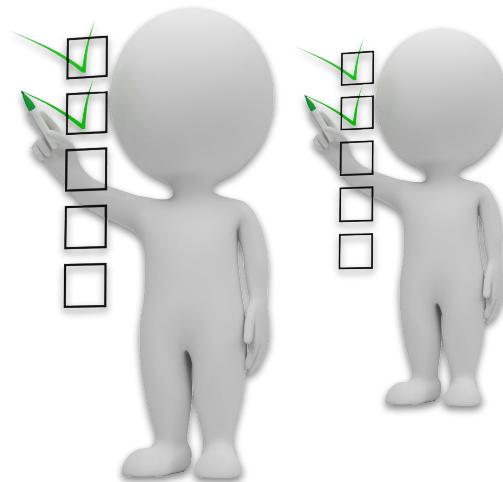


Were we satisfied with the results?

Authoritative source of truth is missing

1. No authoritative arbitration between answers is performed.

2. Again – arbitration requires training, and higher awareness





How to control for quality in crowdsourcing?

And still be cost-effective.

**IN QASRL AND IN OTHER
DEMANDING ANNOTATION TASKS.**



How to control for quality in crowdsourcing?

Provide training and carefully select your annotator pool

Use multiple trained annotators with human conflict resolution



How to control for quality in crowdsourcing?

Provide training and carefully select your annotator pool

Use multiple trained annotators with human conflict resolution



Controlling for Annotator Skill and Qualification

Screening

1

Selecting workers that do initially well on our task, on an expert-annotated sample, released as “trap task”.





Controlling for Annotator Skill and Qualification

Training

2

1. Providing, short, yet extensive guidelines
2. Conduct a short annotation round for practice

Factuality - Does it really happen?

Words like "might" or "would" inside a question are appropriate when the sentence doesn't clearly indicate whether something actually happened.

Protesters blamed the corruption scandal on local officials, who today refused to promise that they would **resume** the investigation before year's end.

These are correct questions:

- What might someone **resume**? ⇒ the investigation
- When might someone **resume** something? ⇒ before year's end

This is not:

- **Who resumes something?** ⇒ local officials / they
Incorrect, the investigation might not be resumed at all.



Controlling for Annotator Skill and Qualification

Feedback

3

Providing detailed personal feedback.



Evaluation for: XXXXX

Your score: 77.89% for a total of 15 tasks, with 2.40 questions per task

Overall comments: Good job!

Notice for the verb “have” you’ve included “might”, since the sentence describes something that might have to be done. You should also do so for the verb “prepare”.



Controlling for Annotator Skill and Qualification

Selection

4

Selecting best performing workers for the actual dataset annotation.





How to control for quality in crowdsourcing?

Provide training and carefully select your annotator pool

Use multiple trained annotators with human conflict resolution



Streamlining Annotation and Mitigating Conflicts



Multiple Generators

2 Workers per annotation task creating Q&A pairs

1. Who **proposed** something? ⇒ Alan M. Turing
2. How was something **proposed**? ⇒ Turing published



-
1. Who **proposed** ⇒ He, Turing
 2. When was something **proposed**? ⇒ In 1950





Streamlining Annotation and Mitigating Conflicts



Single Consolidator

Selected among best performing annotators.

Removing redundancies or erroneous questions (roles).

1. Who **proposed** something? ⇒ Alan M. Turing ✓
2. How was something **proposed**? ⇒ Turing published ✗

-
1. Who **proposed** ⇒ He, Turing redundant
 2. When was something **proposed**? ⇒ In 1950 ✓



Streamlining Annotation and Mitigating Conflicts



Single Consolidator

Creating a consolidated, non redundant set of roles (questions) and arguments (answers).

1. Who **proposed** something? ⇒ He, Alan M. Turing



2. When was something **proposed**? ⇒ In 1950





Evaluation and Agreement

Are we satisfied with the results now?



Inter Annotator Agreement

Annotator vs. Annotator

79 F1 points.

Average agreement of one trained worker vs. another worker

Pipeline vs. Pipeline

83 F1 points.

Average agreement between one set of annotators (*2 gen + 1 cons*) versus another set



Our Results vs. Previous Annotation

Labelled Argument Detection

Our process results in a significant increase in coverage **vs. an expert set** and a somewhat better precision.

	P	R	F1
Ours	88.0	95.5	91.6
Fitzgerald et. al.	83.1	67.8	74.7



Comparison with PropBank

- High recall on all roles (> 93%)
- High actual precision (> 90%)
- Detecting many **implicit roles** not covered by PropBank
- Our annotators are on-par with experts.

QANom (*Klein et. al., COLING 2020*)

QA-driven SRL of *deverbal nominalizations*

Use QA-SRL annotation format to attain a unified verbal+nominal SRL scheme

Thomas has proved in different ways that God exists, including an assertion dubbed "the Ontological argument".				
	ARGO		Who has proved something?	Thomas
PropBank	ARG1	QA-SRL	What has someone proved?	that God exists
	ARGM-MNR		How did someone prove something?	in different ways the Ontological argument
Thomas has provided different proofs for the existence of God, including an assertion dubbed "the Ontological argument".				
	ARGO	QANom	Who has proved something?	Thomas
NomBank	ARG1		What has someone proved?	the existence of God
	-		How did someone prove something?	the Ontological argument

QADiscourse (*Pyatkin et. al., EMNLP 2020*)

Represent informational Discourse Relations through QA pairs

They're risking their reputation and a charity's reputation is very precious.

Q: Despite what are they risking their reputation?

A: a charity's reputation is very precious

It was her government that started putting people on incapacity benefit rather than register as unemployed [...].

Q: Instead of what did her government put people on incapacity benefit?

A: rather than register as unemployed

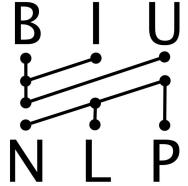
Controlled Crowdsourcing Takeouts

- Adapt expert annotation practices to crowdsourcing
- Substantially pushing the boundaries of crowdsourcing quality
 - High quality annotation in the first place, rather than struggling, often times hopelessly, with noisy annotations
- Adapted in our lab for many different tasks
 - Similar methodologies in some other labs, but usually not publicized

Misconception:

A system summary should correspond to the reference summary length

Evaluating *Multiple* System-Summary-Lengths: A Case Study

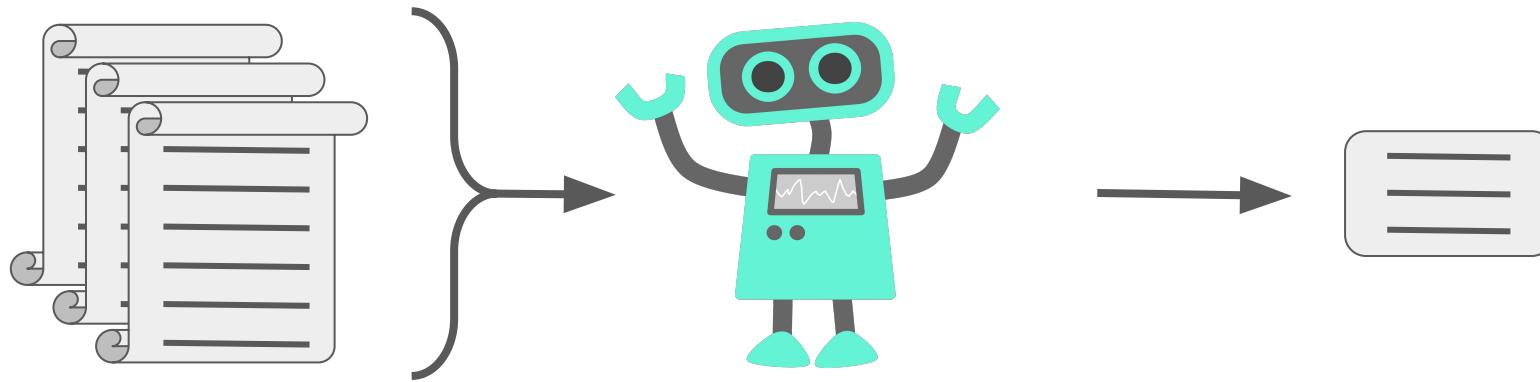


Ori Shapira, David Gabay,
Hadar Ronen, Judit Bar-Ilan, Yael Amsterdamer,
Ani Nenkova, Ido Dagan

Texts to
summarize

Summarization
system

Summary

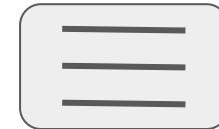
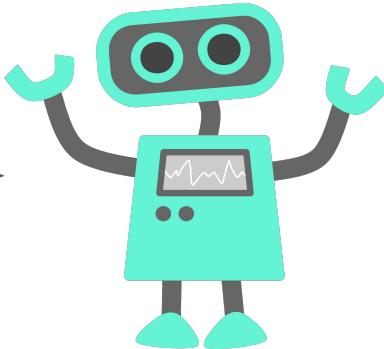
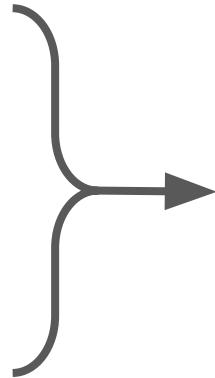
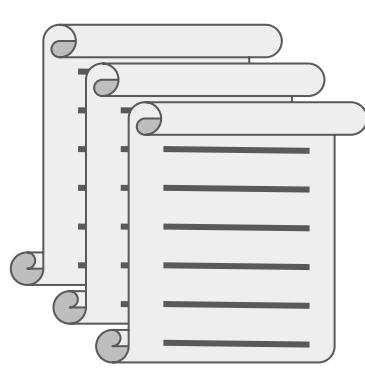


Summarization in a nutshell (multi-document setting)

Text to
summarize

Summarization
system

Summary

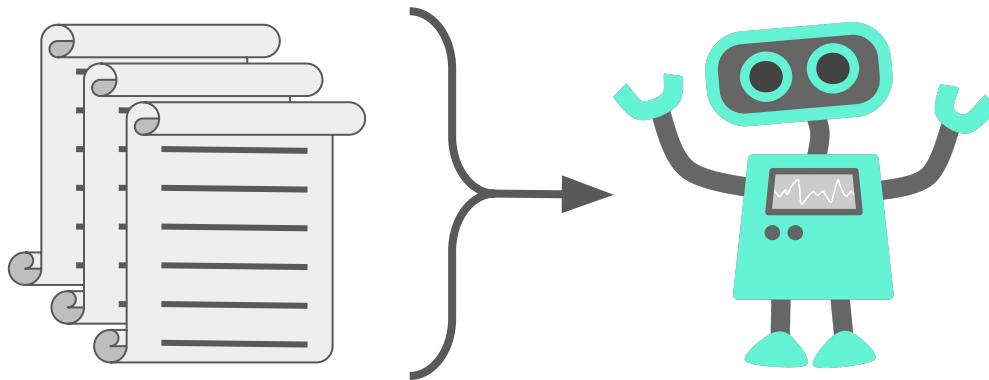


Summarization in a nutshell

Text to
summarize

Summarization
system

Summary

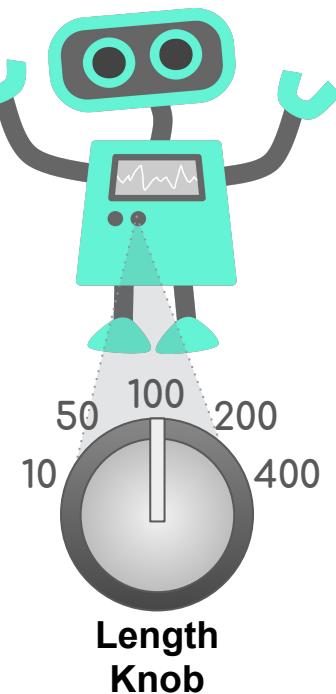
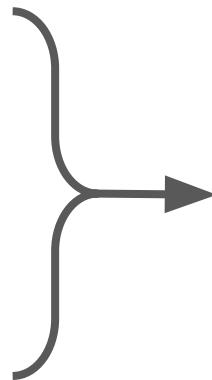
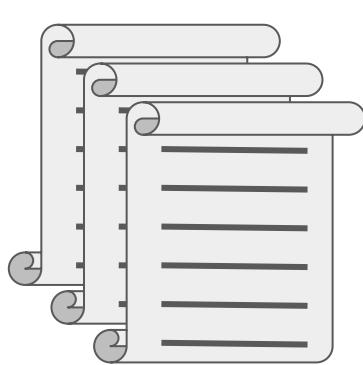


Summarization in a nutshell

Text to
summarize

Summarization
system

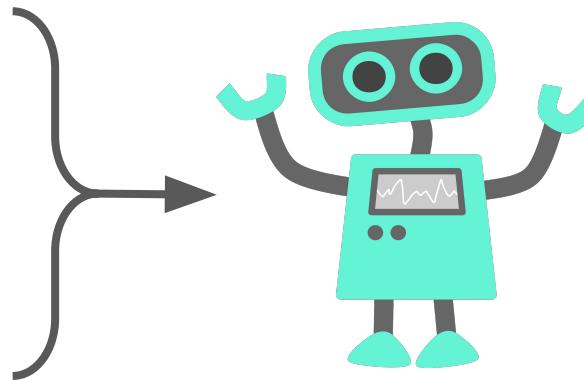
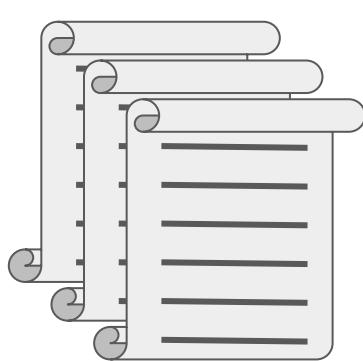
Summary



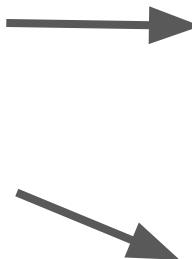
Text to
summarize

Summarization
system

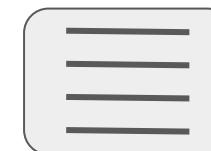
Summary



Length
Knob



100 words



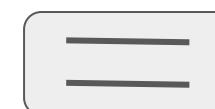
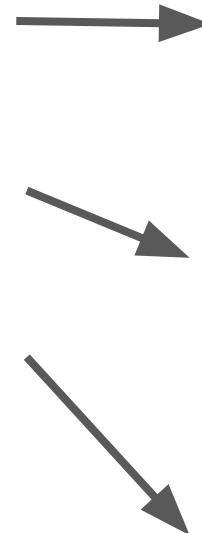
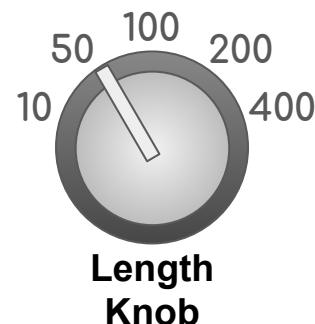
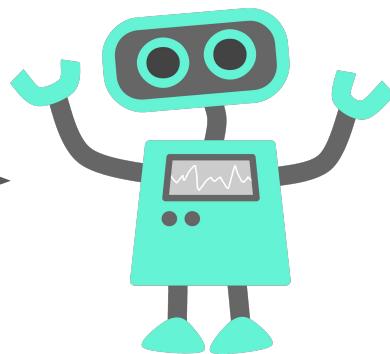
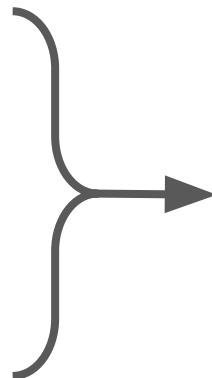
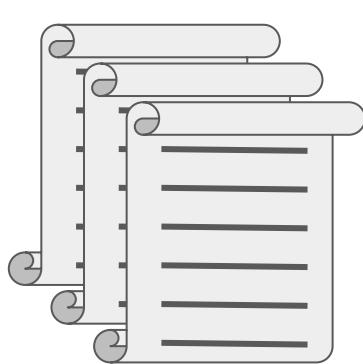
200 words



Text to
summarize

Summarization
system

Summary

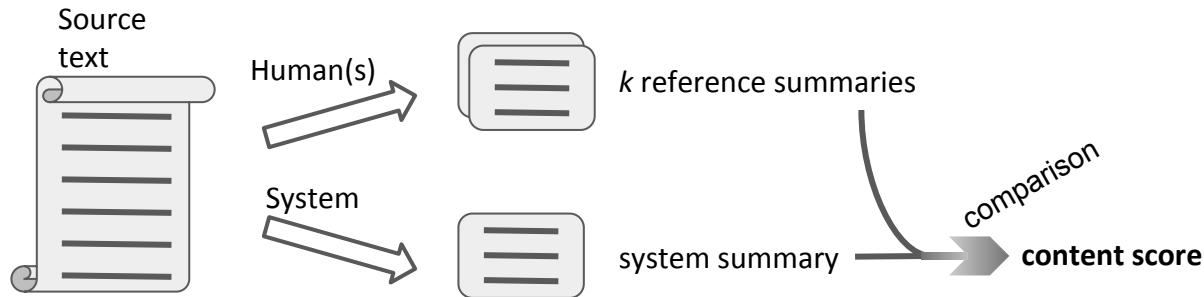


Multi-Length Summary Evaluation

- Early benchmarks asked systems to output summaries of multiple predefined lengths (DUC 2001 and 2002)
 - 50, 100, 200 and 400 word summaries in 2001
 - 10, 50, 100, 200 word summaries in 2002
 - Provided manual reference summaries for all lengths
 - Each system summary evaluated against reference summary of the same length
- Multi-length methodology abandoned due to high cost
- Since then, **hardly any evaluation, and research**, on multi-length summaries

How are Summaries Evaluated (for content)?

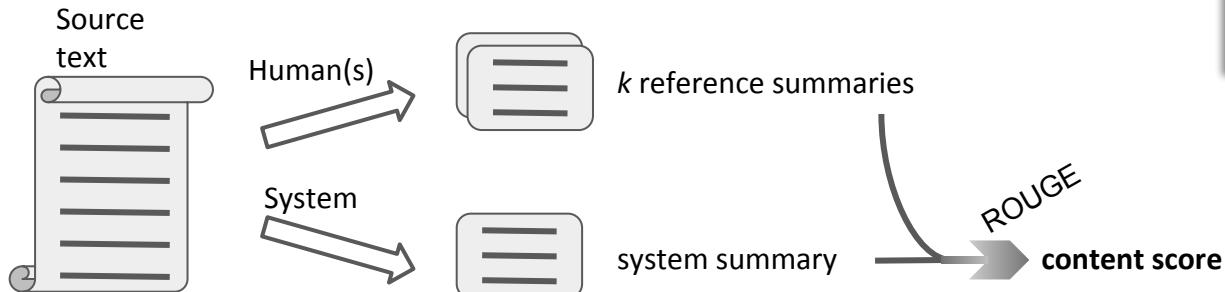
- Common evaluation: Compare a *system* summary to a *reference* summary
 - Manual: semantic content comparison by humans -- effective but expensive
 - Automatic: lexical (mostly) overlap -- shown/assumed to correlate decently to manual evaluation



Automatic Evaluation

ROUGE - most widely-accepted *automatic* summary-content evaluation metric

- Word sequence overlap between a system summary and one or more reference summaries
 - Originally - ROUGE Recall

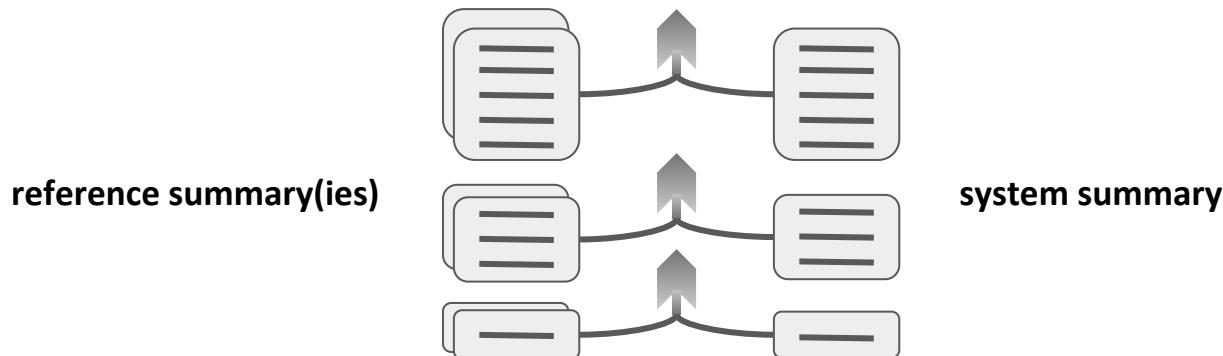


$$\text{ROUGE-N} = \frac{\sum_{S \in \{\text{Reference Summaries}\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{\text{Reference Summaries}\}} \sum_{gram_n \in S} Count(gram_n)}$$

C.Y. Lin, 2004

Common “Same-Length” Presupposition

- It is commonly assumed that a reference summary of a certain length can be used to evaluate only system summaries of a matching length
- Either
 - Same length by construction: prescribed reference and system summary length
 - System summary expected to exactly match reference content (CNN/Daily Mail) (Use ROUGE F1)



Multi-Length Evaluation Worthiness

- After DUC 2001 / 2002, benchmarks discontinued varying length evaluation:
 1. Costly (preparation and comparison)
 2. Seemingly minuscule system ranking fluctuation between lengths

Multi-Length Evaluation Worthiness

- After DUC 2001 / 2002, benchmarks discontinued varying length evaluation:
 1. Costly (preparation and comparison)
 2. **Seemingly miniscule system ranking fluctuation between lengths**

However: fluctuation analysis (DUC 2001):

- Spearman correlation between human rankings of systems at the 50-word and 400-word lengths: **0.61**
- System ranked **1st** at length 50, ranked **6th** at lengths 200 and 400 (of 14 systems)
- The larger the difference between a pair of summary lengths, the greater the fluctuation in system rankings

⇒ **We're missing out - desired to resume multi-length evaluation**

Multi-Length Evaluation Worthiness

- After DUC 2001 / 2002, benchmarks discontinued varying length evaluation:
 1. **Costly (preparation and comparison)**
 2. Seemingly minuscule system ranking fluctuation between lengths

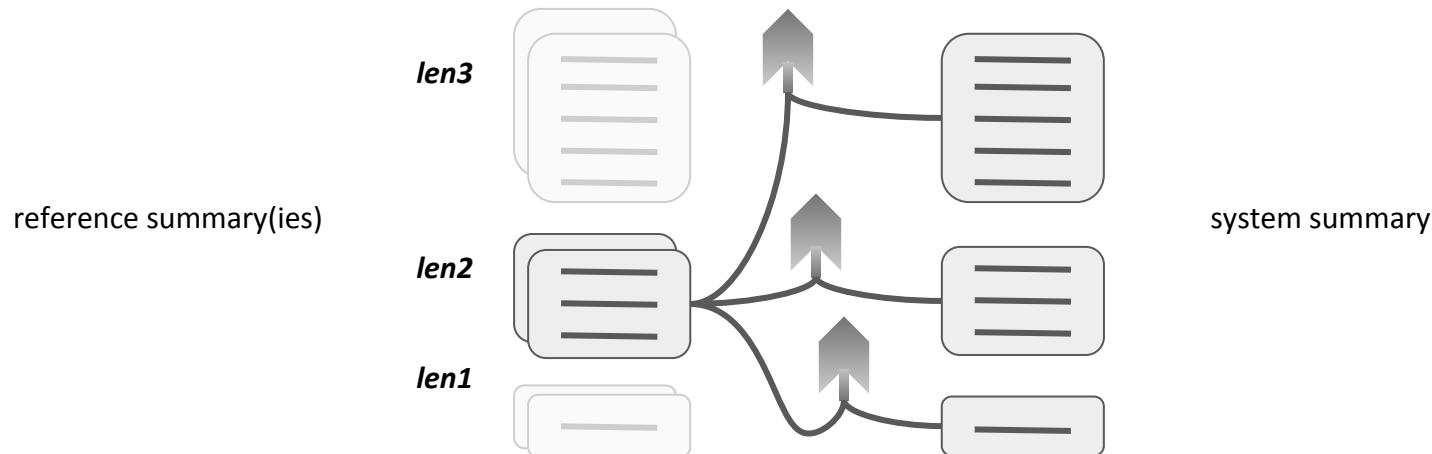
Our research question: Is the “same length” presupposition required?

- More concretely - can we use a single reference summary length to evaluate system summaries of varying lengths?

⇒ *Need to assess validity of such method!*

Proposal: *Single* Reference Summary Length

To evaluate *multiple* system summary lengths!



Standard vs. Proposed Evaluation

Assess automatic evaluation protocol by correlation to manual evaluation:

*Is $\text{Corr}_{\text{Proposed}}$ ***at least as good as*** $\text{Corr}_{\text{Standard}}$?*

Standard vs. Proposed Evaluation

Assess automatic evaluation protocol by correlation to manual evaluation:

Is $\text{Corr}_{\text{Proposed}}$ at least as good as $\text{Corr}_{\text{Standard}}$?

Data for analysis available from DUC 2001/2002 with:

- reference and system summaries
- manual summary scores

Results

Main observation:

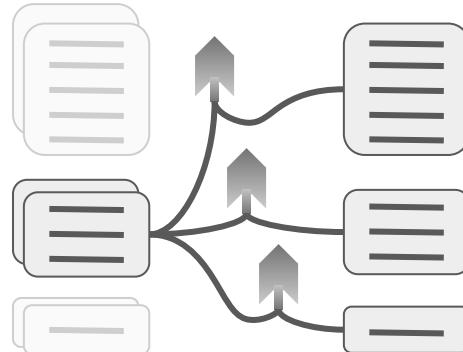
- A **single** set of reference summaries of length **50**, is as effective as **separate** sets of reference summaries per length
- **Longer** reference summary sets are overall less effective

Reference Set	Standard	System Summary Length										Avg. across lengths
		50		100		200		400		3refs		
3refs	1ref	3refs	1ref	3refs	1ref	3refs	1ref	3refs	1ref	3refs	1ref	3refs
Only50	0	0	+0.02	0	+0.01	+0.04	+0.01	+0.02	+0.02	+0.010	+0.015	0.86
Only100	-0.01	+0.04	0	0	+0.01	-0.01	+0.02	0	+0.02	+0.005	+0.008	-0.09
Only200	-0.09	-0.09	-0.06	-0.08	0	0	+0.01	-0.01	-0.01	-0.035	-0.0045	-0.09
Only400	-0.06	+0.02	-0.09	-0.09	-0.01	+0.03	0	0	-0.040	-0.010	-0.09	-0.09

Correlation of ROUGE-1 ranking to manual evaluation ranking on DUC 2001 dataset

One Reference Length- for Multiple System Summary Lengths

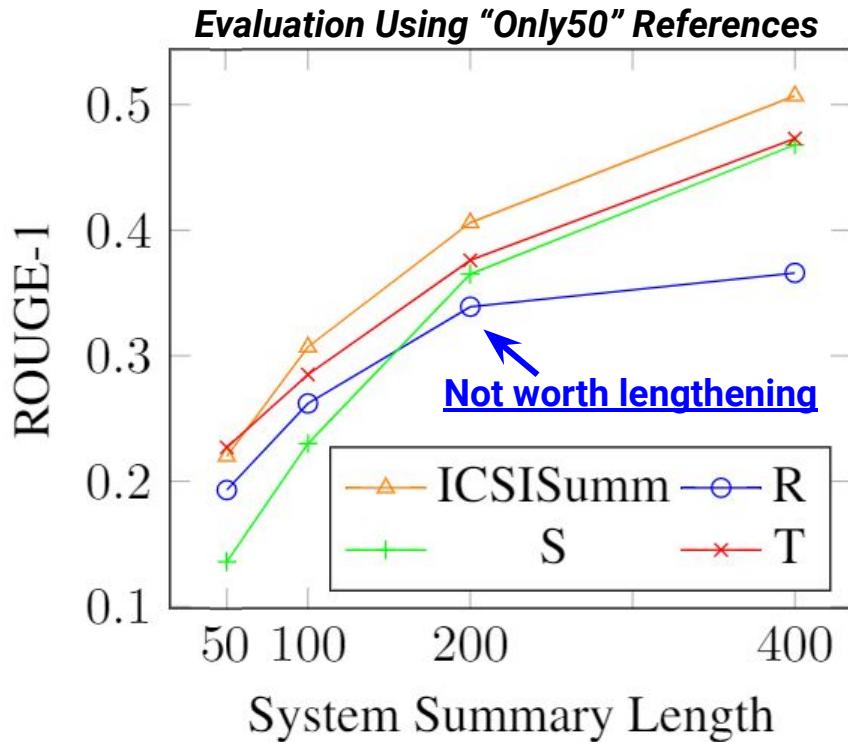
- Our finding: on DUC 2001 and 2002, correlation of automatic (ROUGE) assessment to human assessment using a *single set* of short reference summaries is as good as using a *different set* for each length



Additional Benefit: Incremental Gain of Lengthening Summaries

A single reference set enables:

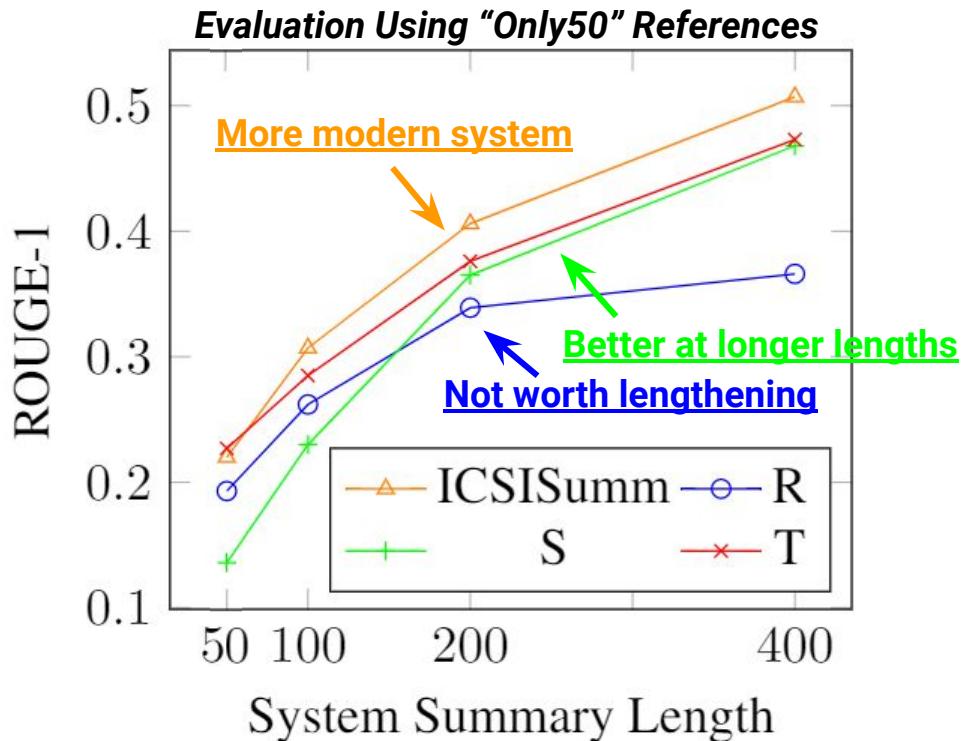
- Observing content gain as a system's summaries lengthen



Additional Benefit: Incremental Gain of Lengthening Summaries

A single reference set enables:

- Observing content gain as a system's summaries lengthen
- Easy comparison between systems at different lengths, and overall



A note on ROUGE F1

- Recent large-scale datasets include a single reference summary per input (document or document set)
- Summary varies across inputs
 - Vs. fixed length in early benchmarks
- ROUGE F1 - introduced to bias system summary content scope to equal that of the summary
 - Again - no “length knob”
- Our protocol is applicable for these datasets too!

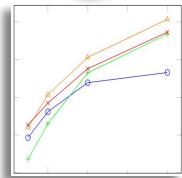
Conclusion - Variable-length Summaries

On the available data:

a set of rather short reference summaries of a single length can effectively evaluate multiple length system summaries



→ Evaluating systems with a length knob can now be cheaply reinstated



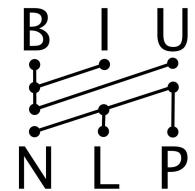
→ Varying-length summarization systems can be easily compared

Conception:

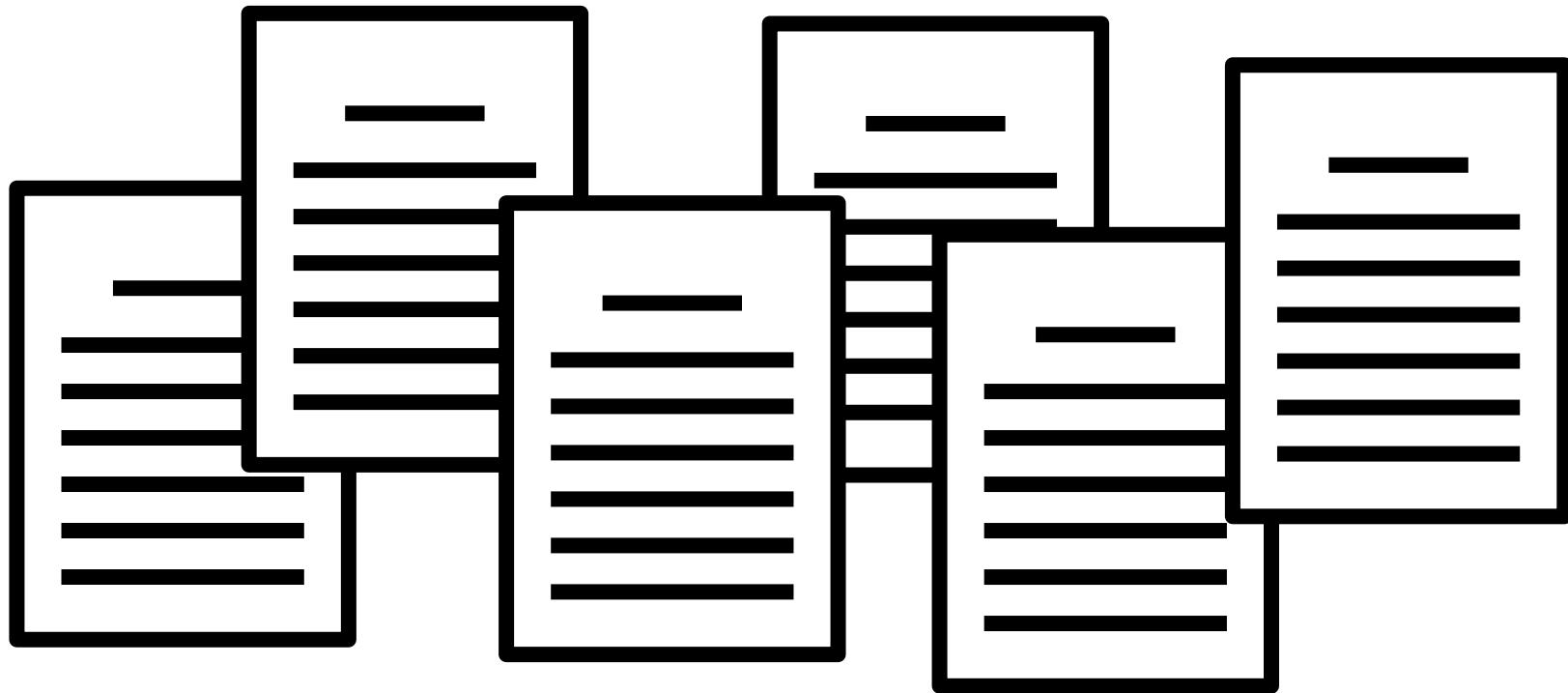
NLP eval is (mostly) about “static” tasks

Extending Multi-document Summarization Evaluation to the Interactive Setting

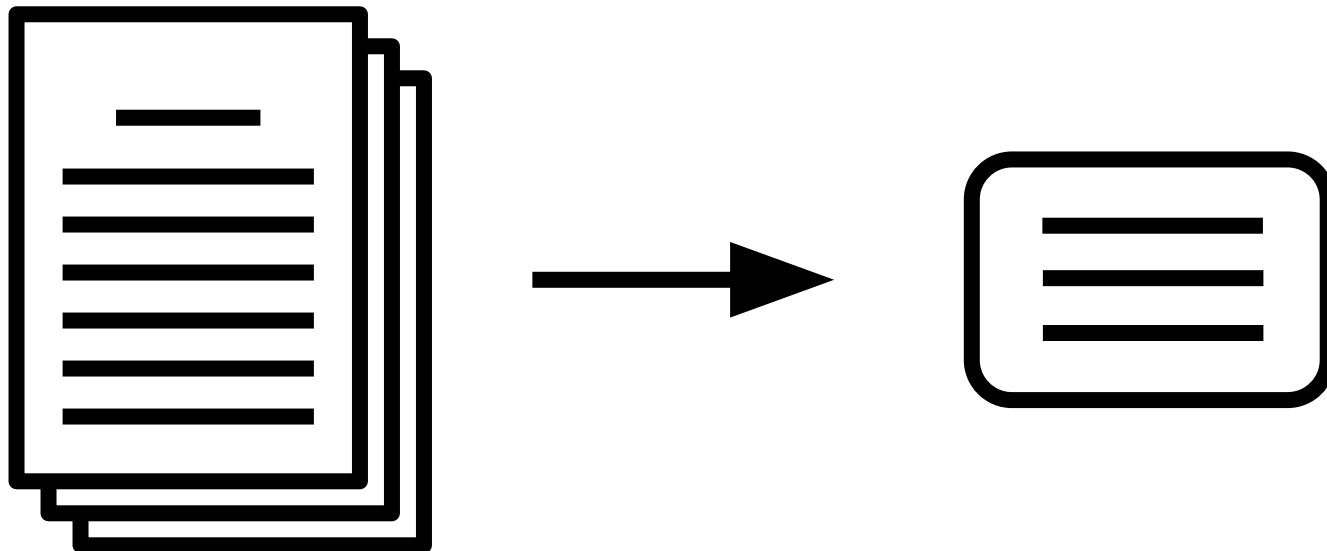
Lead researcher: Ori Shapira (ongoing)
With Ram Pasunuru, Mohit Bansal



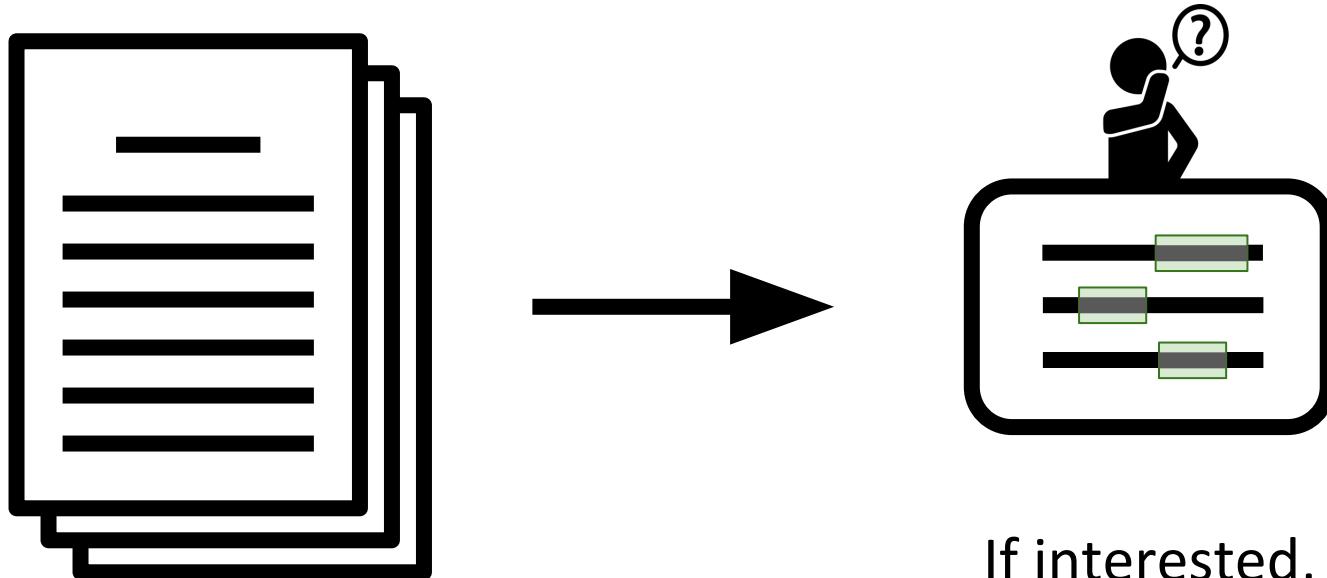
Multiple texts are hard to comprehend!



Is summarization enough?

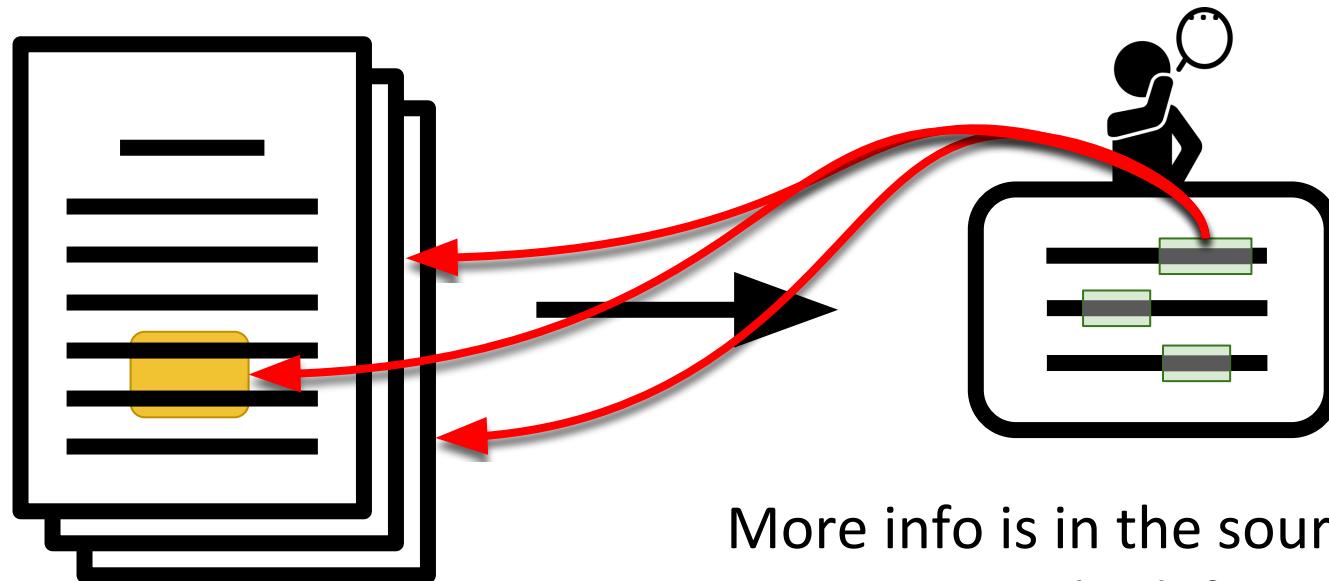


Is summarization enough?



If interested,
user wants to
explore more info!

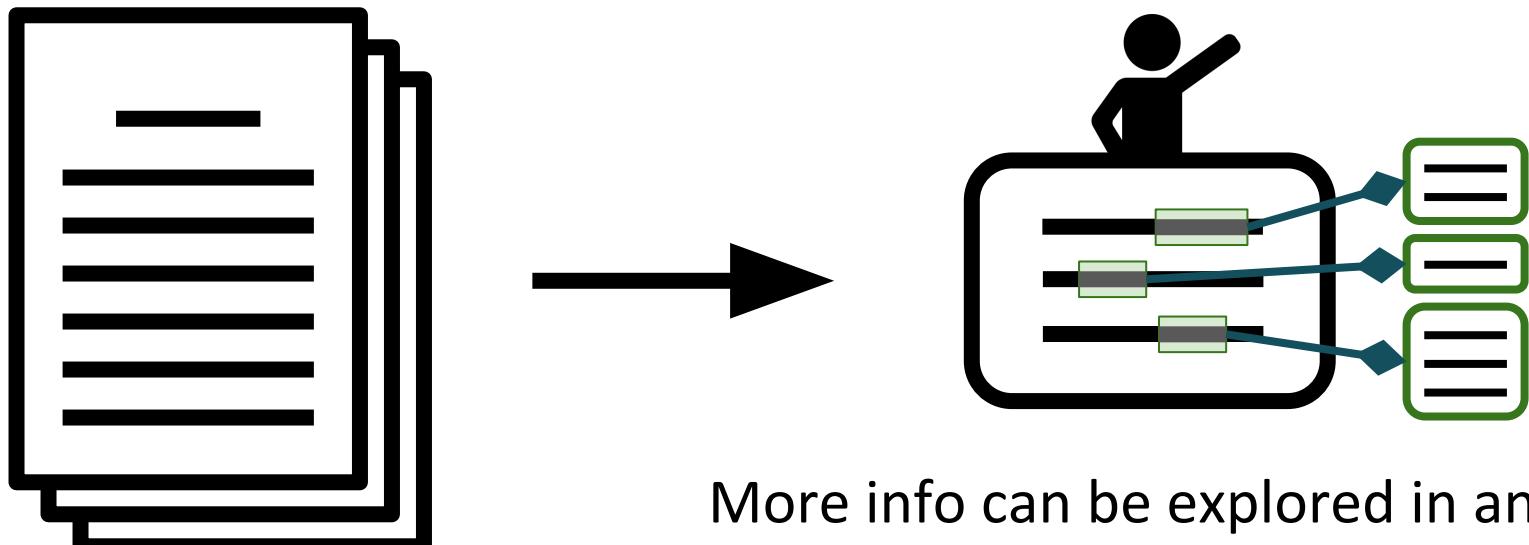
Summarization Gap



More info is in the source texts...
Go look for it



Interactive Summarization



More info can be explored in an
interactive summary!



Demo - prototype system

Summary of 25 articles on

NATIVE AMERICAN CHALLENGES

have worked to wean themselves from federal dependency as funding for Indian programs has waned.

About one-third of the 557 Indian tribes around the nation, including tribes in Connecticut, Minnesota and Wisconsin, now offer some form of gambling.

How useful is this for the journalist's generic overview of the topic? 4-Very useful ★★★★

high school

Only about 17 percent of high school graduates go to college, and most do not finish.

Aside from that, only 63 percent of Indians are high school graduates.

How much useful info does this add to the journalist's overview (regardless of how well it matched your query)? ★★★★

Tip: Highlight text to query

C +

You may move on to the questionnaire if you're done exploring. Done exploring →

1

2

3

4

5

6

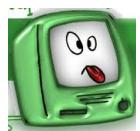
Interactive NLP

- Not much research conducted



- Very difficult to evaluate
- Involves humans... eek!

- But very important



- Human-computer cooperation addresses human dynamic needs
- Improves outputs, and reduces computer foolishness

Interactive Summarization

A few works -- all are evaluated differently and incompletely (if at all)

Hurricane Sandy

Event Summary of 111 Tweets

Timeline

26.10.12 23:49 massive hurricane sandy
76 Tweets | keywords: 65 people, killed

27.10.12 00:05 hurricane heads for US after hurricane killing 59 in caribbean
3 Tweets

27.10.12 01:44 US east coast braced for hurricane
22 Tweets | More Details

27.10.12 01:56 details on preparations here in maryland on wjz at 11pm
1 Tweets | keywords: en route, coast

27.10.12 03:46 hurricane blows out bahamas after hurricane killing 43 in caribbean

1 [click record: 0][2][9]

X [0] BP has resumed its cleanup operation.

X [1] the spill, which is believed to be leaking from a damaged wellhead.

X [2] a delicate "top kill" procedure to stop the flow of oil.

X [3] seven hours later, BP's chief operating officer, Doug Suttles, said the spill had been stopped.

X [5] Environmental activists have been critical of BP's handling of the spill.

X [6] Doug Suttles, BP's chief operating officer, has been critical of BP's handling of the spill.

X [8] Offshore, marine life has been impacted by the spill.

X [9] The spill, off the coast of the Gulf of Mexico, has been leaking since last month.

X [10] The modest signs of progress have been welcomed by environmentalists.

X [12] Engineers last night began a second round of the procedure, pumping thick mud at high speed into the well, which lies about 10 miles off the coast of Louisiana.

X [13] New government estimates of a leak of 12,000 to 19,000 barrels of oil a day indicated that after 37 days, the slick could grow to 100 square miles by the end of the month.

X [15] After failing earlier in the month to halt the leak with a top kill-hatched down, BP began pumping hundreds of thousands of gallons of mud into the well to stop the flow of oil.

X [17] Environmental scientists who have toured the marshes off Louisiana by boat described a vast expanse of crude half an inch thick.

X [18] In its latest update to the Deepwater Horizon, the oil giant also said that its "top kill" procedure to stem the flow of oil had been successful.

X [19] Engineers were due last night to begin a second round of pumping thick drilling mud at high speed into the ocean floor.

X [21] We said BP engineers would soon use additional materials to try to plug the well, suggesting heavy sand deployed as far as 10 miles offshore.

X [22] The company is also still collecting spilled oil from the leak points at the Macondo well.

X [24] At the start of the week, BP admitted that it was capturing less oil from the ruptured well than previously estimated.

X [25] This was the result of congressional investigations into the oil disaster also intensified with BP ordered to produce documents to prove the claim.

X [26] The start of the investigation came after a "leak shot" to block the leak had failed, twice and other attempts to fix the problem had failed.

[click record: 0][2][10]

Engineers began pumping thick mud into the well to stop the flow of oil. The operation cost \$10 million, and it is expected to take several days to complete according to BP's chief operating officer, Doug Suttles. He described it as a "last-ditch effort".

May 28 2010

Documents loaded.

The figure is a map of the southeastern United States and the surrounding Caribbean Sea. It shows the coastline from the Carolinas down to the Yucatan Peninsula. A thick, dark line represents the path of Hurricane Andrew. The hurricane formed in the Bahamas on August 26, 1992, moved westward across the Gulf of Mexico, and made landfall near Homestead, Florida, on August 24, 1992. The path then continued inland through central Florida and into the western Gulf of Mexico, ending near Galveston, Texas. Major cities labeled include Miami, Fort Lauderdale, West Palm Beach, Naples, Tampa, St. Petersburg, Clearwater, Orlando, Lakeland, Tallahassee, Pensacola, Mobile, and New Orleans.

test

U/1992 US INSURERS expect to pay out an estimated Dollars 15billion 3.7bn in Florida as a result of Hurricane Andrew. The costliest disaster the industry has ever faced.

7/1992 President George Bush on 08/01/1992 made his visit to the region since the hurricane hit. (08/25/1992) US Senate passes a bill on the night of 08/24/1992 as Hurricane Andrew headed west after sweeping across southern Florida, at least eight deaths and severe property damage.

8/2002 CATASTROPHES CAUSED by Hurricane Andrew could rise to \$200 billion by 2010, it was estimated on 08/26/1992. The costliest of the US storms this century threatened a further devastating hit near the city of New Orleans. (08/26/1992) At least three people were killed as Hurricane Andrew crossed the state of Florida on 08/26/1992. Ms Kate Hagan, director of emergency services in Florida's Dade County, which bore the brunt of the estimated that Andrew had already caused Dollars 15bn in insured catastrophe losses.

7/1992 SQUADS of workers fanned out across storm-battered Florida on 08/27/1992 to begin a massive rebuilding effort after Hurricane Andrew had flattened whole districts, killing two people

Reset Zoom Pan

PT 04 SEP 02 / Hurricane in NEW YORK

US INSURERS expect to pay 3.7bn in Florida as a result of the costliest disaster the industry has ever faced. The damage resulting from the Hurricane Andrew last week.

President Clinton

to the region since the hurricane

the final cost of Hurricane Andrew

on the Florida losses alone, insured catastrophe in the US in September 1999, cost the insurance industry

contrast, insurance claims in the year - the most expensive claim in the US in 1999, Hurricane Andrew, was their worst-ever year for catastrophe losses.

Hurricane Andrew losses alone,

Summa

About

Pope Benedict Resignation

February 2013 - March 2013

- + 2013-2-11 Pope Benedict XVI to become first pope in 600 years to resign [link](#)
- 2013-2-26 Pope Benedict XVI will keep the name Benedict XVI and become the Roman pontiff emeritus or pope emeritus , the Vatican announced on Tuesday , putting an end to days of speculation on how the pope will be addressed once he ceases to be the leader of the world's 1.1 billion Roman Catholics on Thursday. [link](#)

+ 2013-2-26 The Vatican did reveal the answer Tuesday to one of the most common questions surrounding Benedict and his retirement : what he will be called once he is no longer the reigning pope. [link](#)

+ 2013-3-7 Several strike a note of nostalgia for the charismatic late Pope John Paul after eight years of his shy successor Benedict , who shocked the Catholic world by becoming the first pope in almost 600 years to resign last month. [link](#)

+ 2013-3-7 A week after Pope Benedict XVI formally retired , cardinals assembled in Rome to choose his successor went into a fourth day of soundings and deliberations on Thursday without reaching a decision so far on a date to begin the secret papal balloting known as a conclave . A Vatican spokesman said [link](#)

+ 2013-3-14 Pope Francis was high on cardinals' lists before vote [link](#)

Syria Chemical Weapons Crisis

Egypt Unrest and Coup

NSA Global Surveillance Disclosures

New York City Mayoral Race

Boston Marathon Bombing and Manhunt

Pope Benedict Resignation

Attack on US Embassy in Benghazi

Evaluating Interactive Summarization - A Challenge!

Till now - evaluations not comparable:

- Inconsistent user studies
- Pairwise preference of proprietary systems
- ROUGE for end-of-session summary against improvised reference
- No standard consideration of session evolution

Evaluating Interactive Systems - A Challenge!

- Often requires
 - Organized shared tasks
 - Experts as users
- Subjective system use
- Noisy when employing crowd-workers

Contribution

Extending summarization evaluation to the interactive setting:

- Absolute, comparable measures
- Using *controlled* crowdsourcing

Interaction Setup Abstraction

- After each interaction, new text is presented
- Interactive summarization = an *incrementally growing* summary
- Each interaction yields a snapshot summary of the information presented so far

Interaction Setup Abstraction

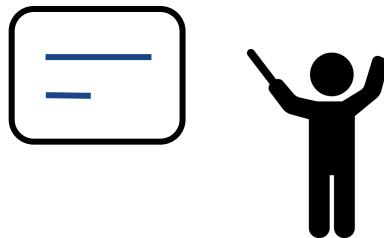
- After each interaction, new text is presented
- Interactive summarization = an *incrementally growing* summary
- Each interaction yields a snapshot summary of the information presented so far



Initial summary

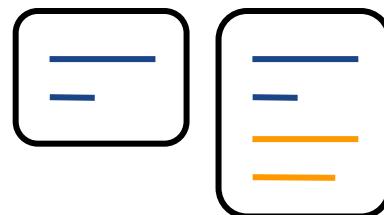
Interaction Setup Abstraction

- After each interaction, new text is presented
- Interactive summarization = an *incrementally growing* summary
- Each interaction yields a snapshot summary of the information presented so far



Interaction Setup Abstraction

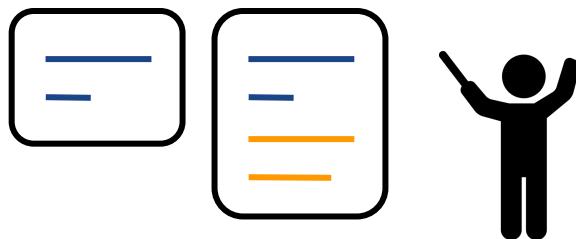
- After each interaction, new text is presented
- Interactive summarization = an *incrementally growing* summary
- Each interaction yields a snapshot summary of the information presented so far



Second snapshot

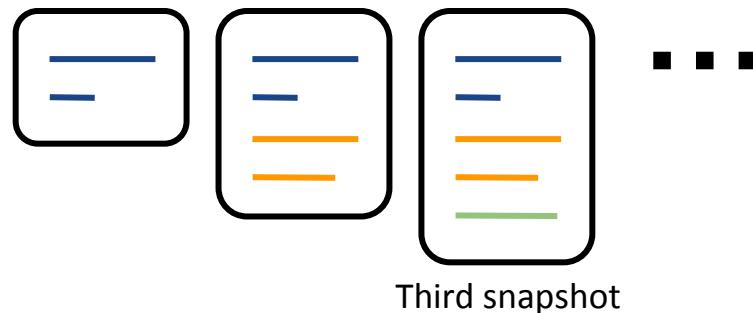
Interaction Setup Abstraction

- After each interaction, new text is presented
- Interactive summarization = an *incrementally growing* summary
- Each interaction yields a snapshot summary of the information presented so far



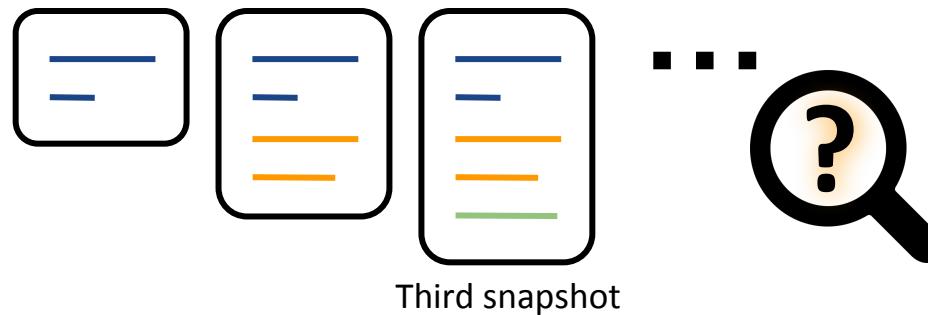
Interaction Setup Abstraction

- After each interaction, new text is presented
- Interactive summarization = an *incrementally growing* summary
- Each interaction yields a snapshot summary of the information presented so far



Interaction Setup Abstraction

- After each interaction, new text is presented
- Interactive summarization = an *incrementally growing* summary
- Each interaction yields a snapshot summary of the information presented so far



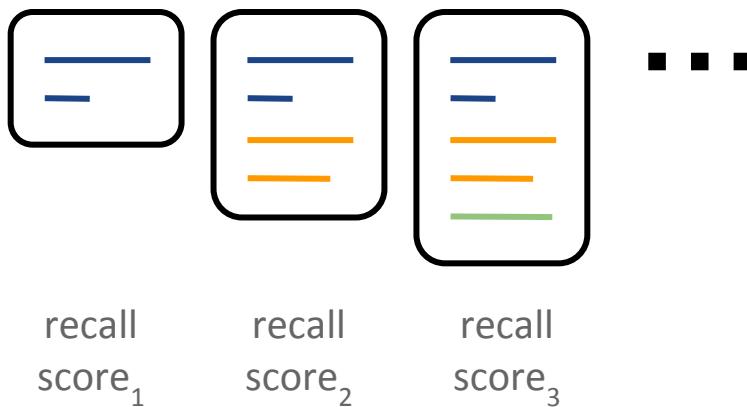
How to evaluate?

Our approach: Adapt static evaluation methods

- Both *Automatic* and *manual* summary content evaluation
- Applying ***recall-based*** measures
 - Particularly *ROUGE Recall* for automatic evaluation

How is an interactive summary then evaluated?

- In a user session, each snapshot can be evaluated as a static summary, with a recall-based measure

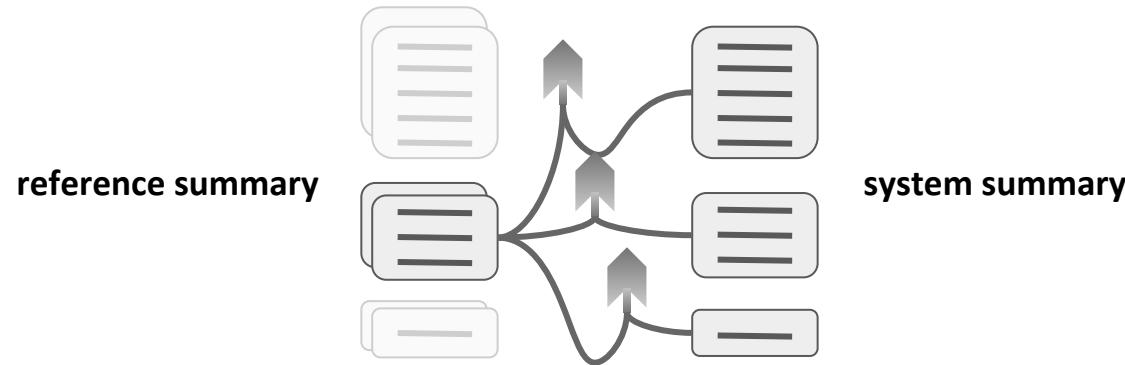




Evaluating the “Snapshots”

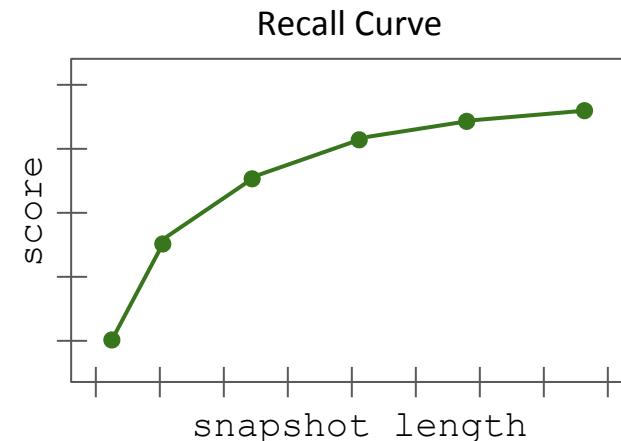
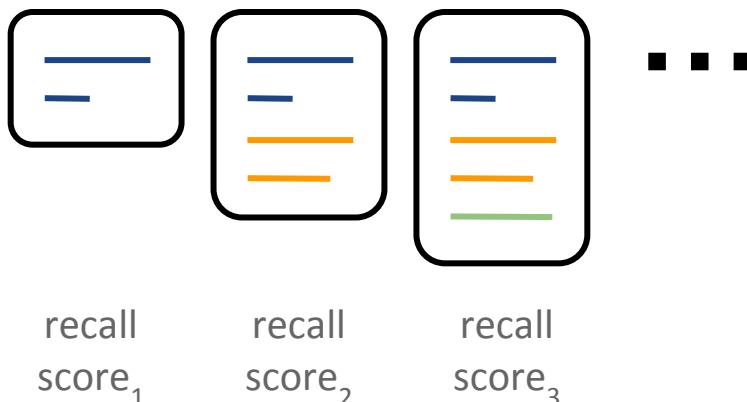
Based on our previous **single reference summary** length result:

→ We can use a standard MDS dataset



How is an interactive summary then evaluated?

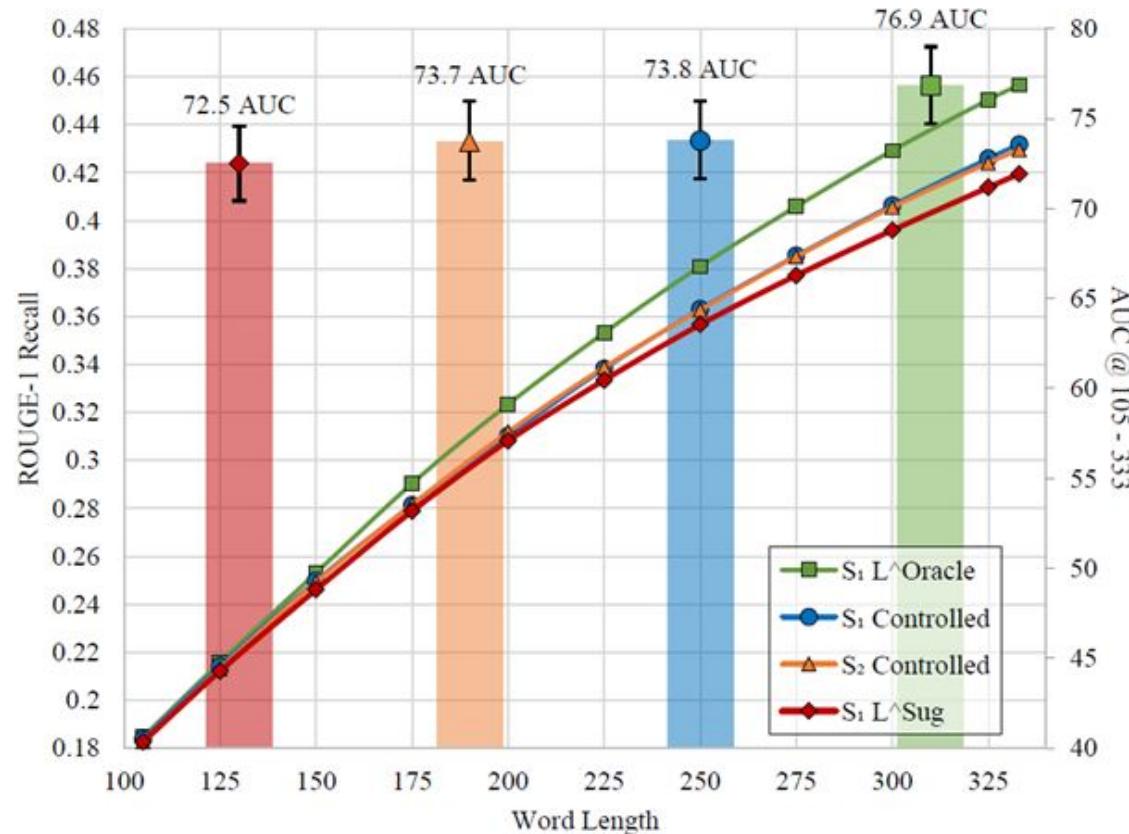
- In a user session, each snapshot can be evaluated as a static summary, with a recall-based measure



- Score increases as text accumulates*
- AUC can provide an overall score*

Real Interactive Evaluation Example

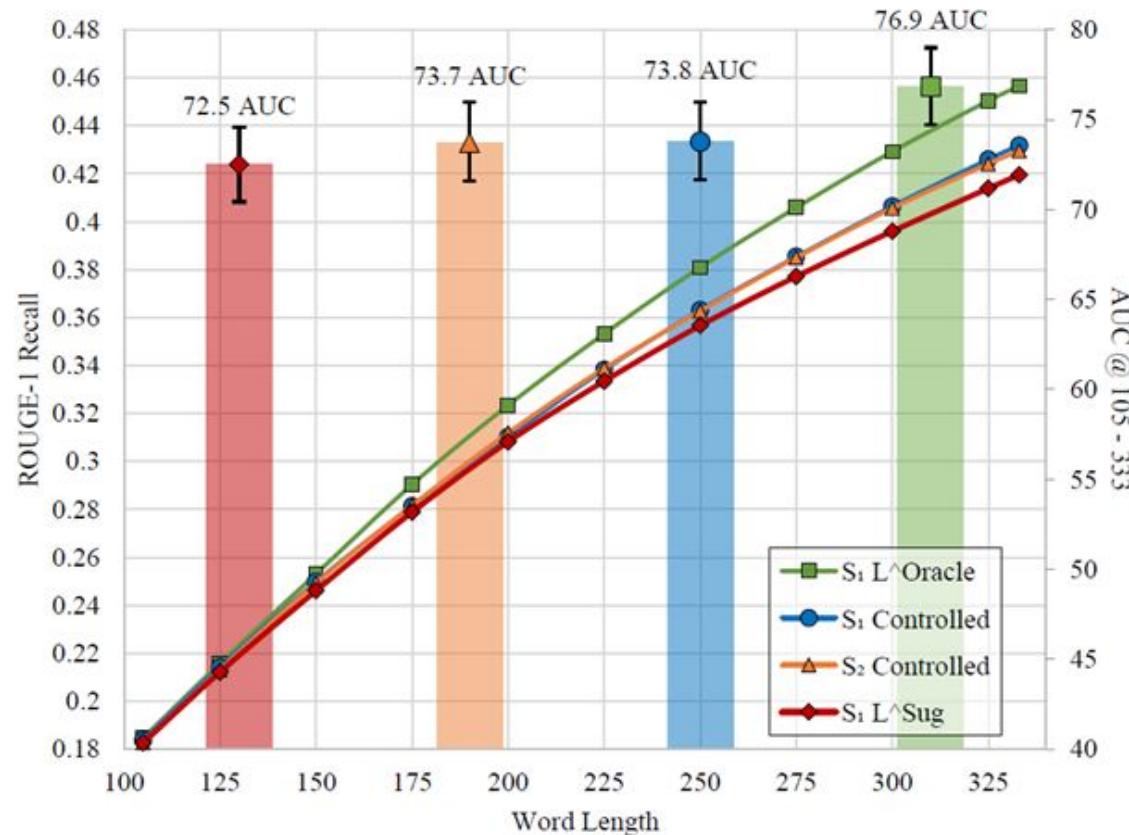
- Two baseline systems, crowdsourced sessions
- Upper bound
 - “Oracle” queries
- Lower bound
 - No human involvement



Real Interactive Evaluation Example

- Two baseline systems, crowdsourced sessions
- Upper bound
 - “Oracle” queries
- Lower bound
 - No human involvement

Human-computer synergy helps!



Automatic Evaluation - Advantages

- Scores are absolute, and comparable from one session/system to another
- Extends upon established *static* summarization practices
 - While utilizing **existing MDS datasets** reference summaries

Manual Evaluation

Leverage the session collection to also get human ratings [1-to-5 scale]

- Informativeness of initial summary
- How much useful info each response adds
- How well the system responded to the requests overall (at the end)
- UMUX-Lite questionnaire (Lewis et al., 2013) - perceived usability
- In our experiments:
 - Human evaluations found consistent with the automatic ones

Session Collection

True system quality reflected only in actual human sessions



Challenges of Crowdsourced Session Collection

- Prior work relied mostly on in-house user studies
 - Expensive, not scalable
- Our initial “standard” crowdsourcing:
 - Noisy - did not reflect realistic system use
 - Non-characteristic experimental behavior - “what would happen if...”
 - Insincere work - “I want easy money!! \$\$\$”
 - Incomparable user interests
 - No clear informational goal - unguided users show subjective interests

Proposal: Controlled Crowdsourcing for User Sessions

Purpose:

- Improve session quality (addressing above challenges)
- Filter out users with low engagement or lacking relevant skills
- Inspired by controlled crowdsourcing for QA-SRL annotation
 - (Roit et al., ACL 2019)

Controlled Crowdsourcing Stages

- Trap Task
 - Filter out insincere workers
 - Discover workers with ability/engagement for text exploration

Controlled Crowdsourcing Stages

- Trap Task
 - Filter out insincere workers
 - Discover workers with exploratory orientation (likely users)
- Practice Tasks
 - Familiarize user to the interface
 - Present objective “cover story”/goal for the session
 - E.g. “Produce an informative summary draft text which a journalist could use to best produce an overview of the topic”
 - Strongly emphasized throughout the practice sessions
 - Integrated wizard guidelines

Session Collection

- Session collection by workers who
 - Passed the trap task
 - Successfully complete two practice tasks

Controlled Crowdsourcing Better than “Wild”

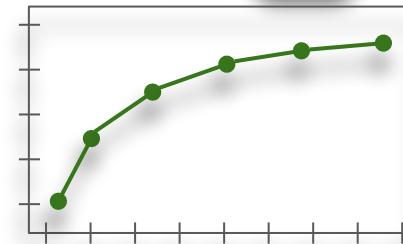
Controlled approach is better in various measurements:

Measure	Controlled	Wild	
# interactions	12.3	7.0	
Approx. explore time	250 sec.	170 sec.	
% suggested query	36.2%	62.7%	
% free-text query	25.3%	2.2%	
% Δ AUC from lower bound	+1.8%	-1.4%	

More engaged
More thought in queries
Better than fully automatic baseline

Conclusions: Interactive Summarization

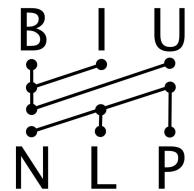
- Interactive summarization addresses the *real* user need
 - *Exploring* multiple-texts - closing the summarization gap
- A viable research task - via proposed evaluation framework
- Opens up plenty to research: modeling, interaction modes



Misconception:
Including singletons in (cross-document)
coreference evaluation

Streamlining Cross-Document Coreference Evaluation

Lead researcher: Arie Cattan (ongoing)
With: Gabriel Stanovsky, Mandar Joshi



Coreference Resolution

Anna and Declan eventually make their way on foot to a roadside pub, where they discover the three van thieves going through Anna's luggage. Declan fights them and retrieves the bag of his sister.

Coreference Resolution --- and its Downstream Use

Anna and **Declan** eventually make their way on foot to a roadside pub, where they discover the **three van thieves** going through **Anna**'s luggage. **Declan** fights **them** and retrieves the bag of **his sister**.

Question Answering (Quoref)

*Who does **Declan** get into a fight with?*

Three van thieves

Coreference Resolution --- and its Downstream Use

Anna and **Declan** eventually make their way on foot to a roadside pub, where they discover the **three van thieves** going through **Anna**'s luggage. **Declan** fights **them** and retrieves the bag of **his sister**.

Question Answering (Quoref)

*Who does **Declan** get into a fight with?*

Three van thieves

Relation Extraction

*(**sibling**, **Anna**, **Declan**)*

Coreference Resolution --- and its Downstream Use

Anna and **Declan** eventually make their way on foot to a roadside pub, where they discover the **three van thieves** going through **Anna**'s luggage. **Declan** fights **them** and retrieves the bag of **his sister**.

Question Answering (Quoref)

*Who does **Declan** get into a fight with?*

Three van thieves

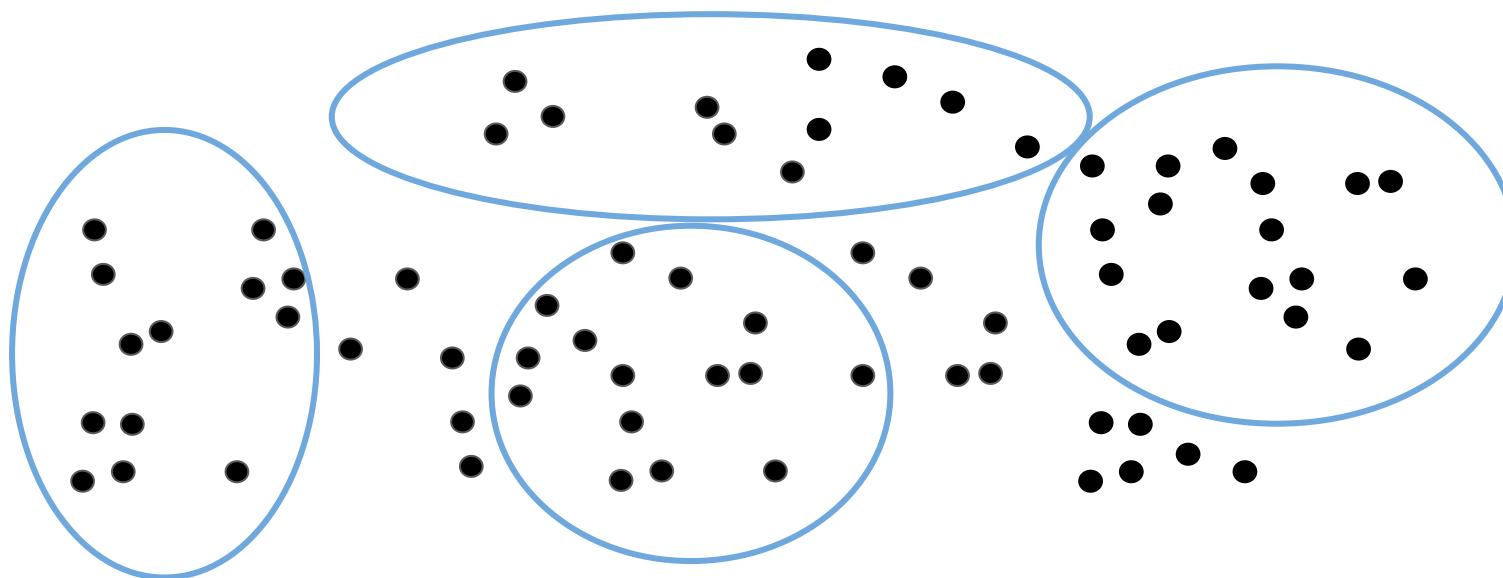
Relation Extraction

*(**sibling**, **Anna**, **Declan**)*

Downstream tasks leverage coreference links to bridge information across mentions

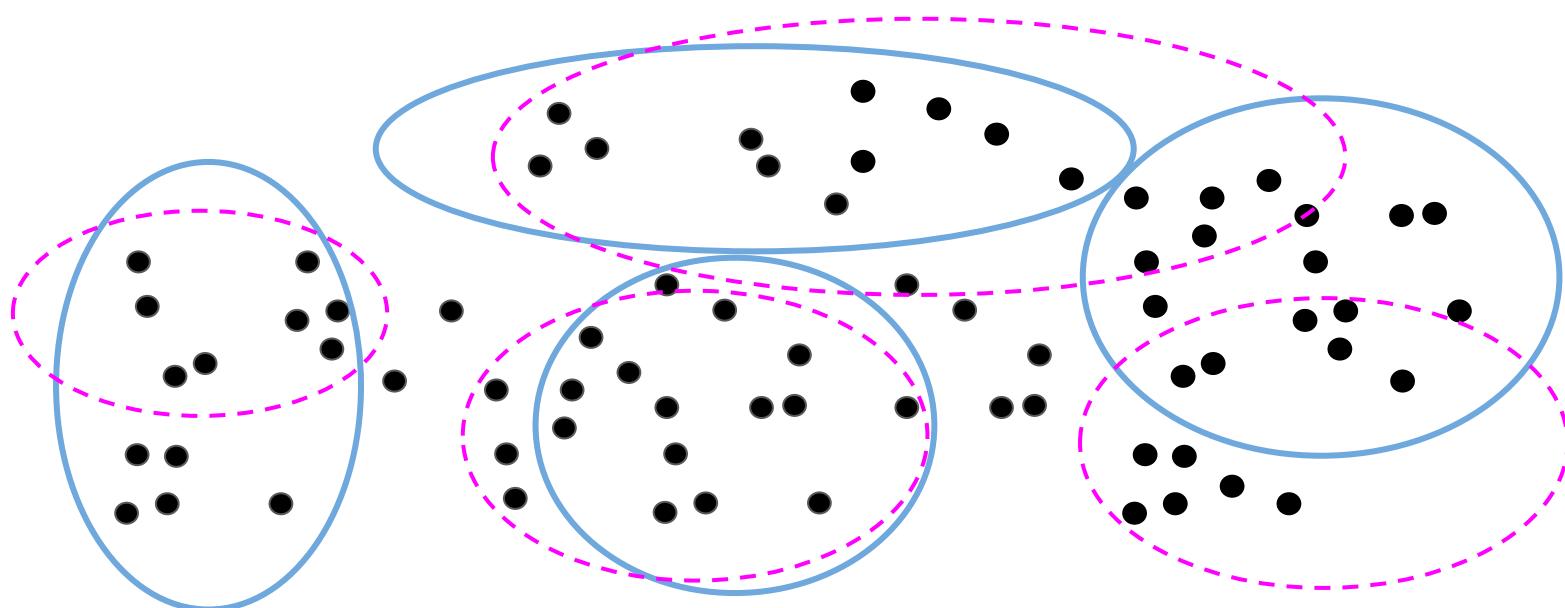
Evaluating Coreference Resolution as a Clustering Task

Measuring “overlap” between gold and predicted clusters:



Evaluating Coreference Resolution as a Clustering Task

Measuring “overlap” between gold and predicted clusters:



Evaluation Metrics

There are multiple evaluation metrics for coreference resolution, each one using recall, precision and F1.

Link-based

MUC (*Vilain et al., 1995*)

Hybrid

LEA (*Moosavi et al., 2016*)

Mention-based

B³ (*Bagga and Baldwin, 1998*)

CEAF (*Luo, 2005*)

Pivot Score: CoNLL F1: (MUC + B3 + CEAF) / 3

Singleton Annotation

- CoNLL-2012 (*Pradhan et al., 2012*) doesn't include singleton annotation
 - Might be considered as a limitation of the dataset?
- ECB+ (*Cybulska and Vossen., 2014*) - the standard dataset for *cross-document* coreference, and PreCo (*Chen et al., 2018*): include singletons
 - Singletons are typically included in evaluations over these datasets
 - They're abundant - ~50% in PreCo, make a major impact on evaluation

Singleton Annotation

- CoNLL-2012 (*Pradhan et al., 2012*) doesn't include singleton annotation
 - Might be considered as a limitation of the dataset
- ECB+ (*Cybulska and Vossen., 2014*) - the standard dataset for *cross-document* coreference, and PreCo (*Chen et al., 2018*): include singletons
 - Singletons are typically included in evaluations over these datasets
 - They're abundant - ~50% in PreCo, make a major impact on evaluation

With or without - what's right?

Singleton Effect

Gold:

way

pub

fights

discover

retrieves

Three van thieves, them

Declan, his, Declan

Singleton Effect

Gold:



Three van thieves, them

Declan, his, Declan

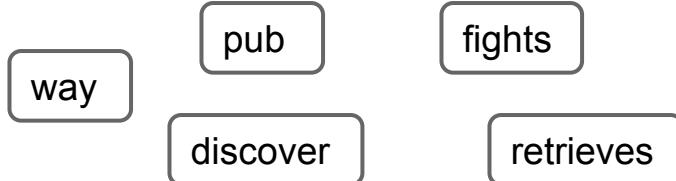
S1



Retrieves, Three van
thieves, them,
Declan, his, Declan

Singleton Effect

Gold:



Three van thieves, them

Declan, his, Declan

S1



Retrieves, Three van
thieves, them,
Declan, his, Declan

S2

Eventually make

Retrieves, Three van
thieves, them

Declan, his, Declan

Singleton Effect

Gold:



Three van thieves, them

Declan, his, Declan

S1



Retrieves, Three van
thieves, them,
Declan, his, Declan

S2

Eventually make

Retrieves, Three van
thieves, them

Declan, his, Declan

From a coreference resolution perspective, S2 is a better model!

Singleton Effect - Results

	System	MUC	B3	CEAF	LEA	CoNLL
With Singletons	S1	75.0	77.6	77.8	69.0	76.8
	S2	85.7	63.2	36.0	53.3	61.6

Singleton Effect - Results

	System	MUC	B3	CEAF	LEA	CoNLL
With Singletons	S1	75.0	77.6	77.8	69.0	76.8
	S2	85.7	63.2	36.0	53.3	61.6
Without Singletons	S1	75.0	53.1	44.4	42.1	57.5
	S2	85.7	83.9	90.0	80.0	86.5

Singleton Effect - Results

	System	MUC	B3	CEAF	LEA	CoNLL
With Singletons	S1	75.0	77.6	77.8	69.0	76.8
	S2	85.7	63.2	36.0	53.3	61.6
Without Singletons	S1	75.0	53.1	44.4	42.1	57.5
	S2	85.7	83.9	90.0	80.0	86.5



Even without singletons, S2 is still penalized because it links a singleton to a cluster.

Singleton Effect - Results

	System	MUC	B3	CEAF	LEA	CoNLL
With Singletons	S1	75.0	77.6	77.8	69.0	76.8
	S2	85.7	63.2	36.0	53.3	61.6
Without Singletons	S1	75.0	53.1	44.4	42.1	57.5
	S2	85.7	83.9	90.0	80.0	86.5

Our proposal:

Even when a dataset (e.g ECB+, PreCo) includes singletons, gold and predicted singletons should be ignored in the evaluation.

- They may still be useful for training/testing a *mention detection* model

SOTA Results on ECB+

- Standard cross-doc coreference benchmark

	With Singletons
<i>(Barhom et al, ACL 2019)</i>	79.5
<i>Our current system</i>	81.0

CoNLL F1 Scores of SOTA models on ECB+

SOTA Results on ECB+

- Standard cross-doc coreference benchmark

	With Singletons	Without singletons
<i>(Barhom et al, ACL 2019)</i>	79.5	67.6 ↓
<i>Our current system</i>	81.0	71.1 ↓

CoNLL F1 Scores of SOTA models on ECB+

SOTA Results on ECB+

- Standard cross-doc coreference benchmark

	With Singletons	Without singletons
(Barhom et al, ACL 2019)	79.5	67.6 ↓
<i>Our current system</i>	81.0	71.1 ↓

CoNLL F1 Scores of SOTA models on ECB+

Including singletons dramatically inflates SOTA results!

Takeaways Singletons

- Including singletons in the evaluation distorts the results and biases models towards mention detection
- Singletons artificially inflate SOTA results in cross-doc coref
- Singletons are not consistently annotated across datasets, but their annotation shouldn't change the task

Overall Conclusions - right out of the box...

- Don't take evaluation practices for granted (or anything...)
- Look carefully for flaws and limiting assumptions
- Look how to extend NLP research scope
 - For real-world needs, beyond comfort zone
 - While correspondingly extending evaluation protocols