Think out of the box!

Conceptions and Misconceptions in NLP Evaluation

Ido Dagan
Bar-Ilan University

Bar-Ilan University

B I U
N L P

DSI Understanding the world through Data Bar-Ilan University
DATA SCIENCE INSTITUTE

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?

UNIVERSITY of WASHINGTON

intel AI

Tradition of Excellence

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

UWNLP

B I U
N L P

Controlled Crowdsourcing for High-Quality QA-SRL Annotation

- Expert annotators are required.

PropBank

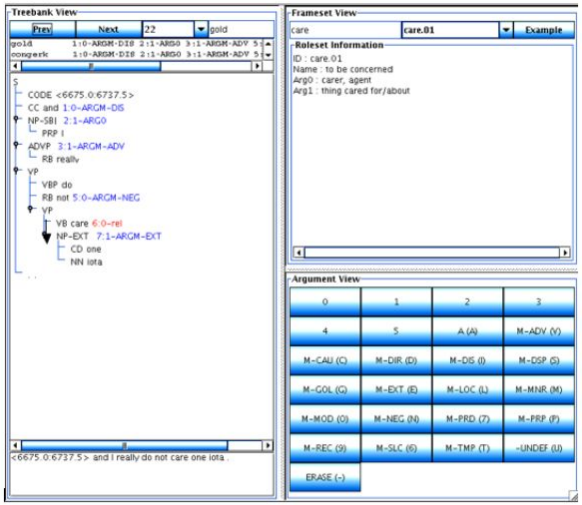


Figure 1.4: Find and annotate any sisters to the `rel`

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

He et. al. (2015) Fitzgerald et. al. (2018)

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?

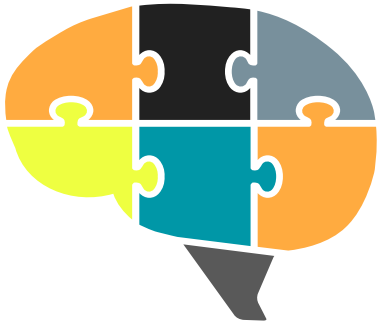
When was something proposed?

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.

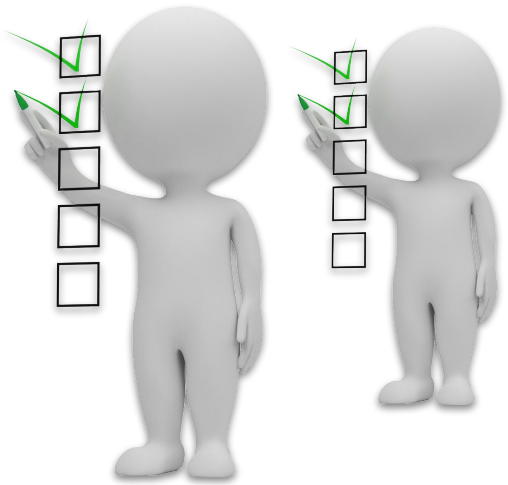
2. Some training and sensitivity is needed, particularly for annotating implicit roles.

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

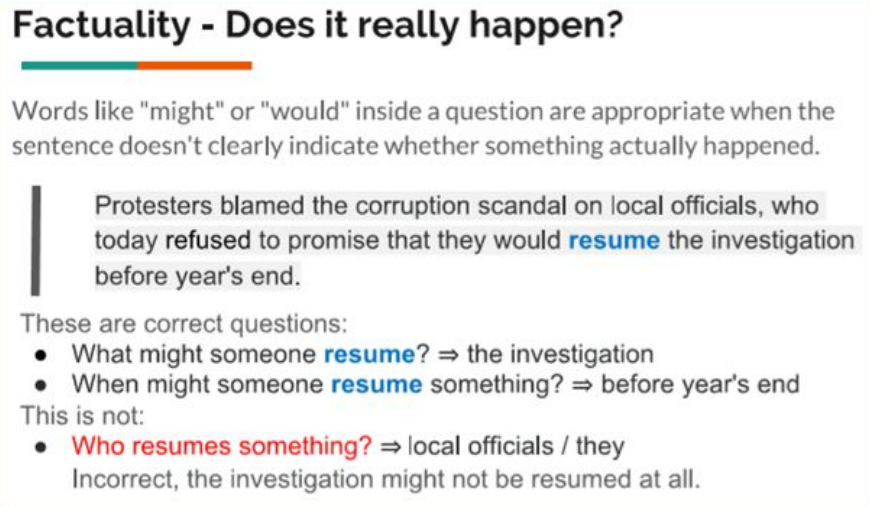
Selecting workers that do initially well on our task, on an expert-annotated sample, released as “trap task”.

Controlling for Annotator Skill and Qualification

Training 2

1. Providing, short, yet extensive guidelines

2. Conduct a short annotation round for practice



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''.				
PropBank	ARG0	QA-SRL	Who has proved something?	Thomas
	ARG1		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''.				
NomBank	ARG0	**QANom**	**Who has proved something?**	Thomas
	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

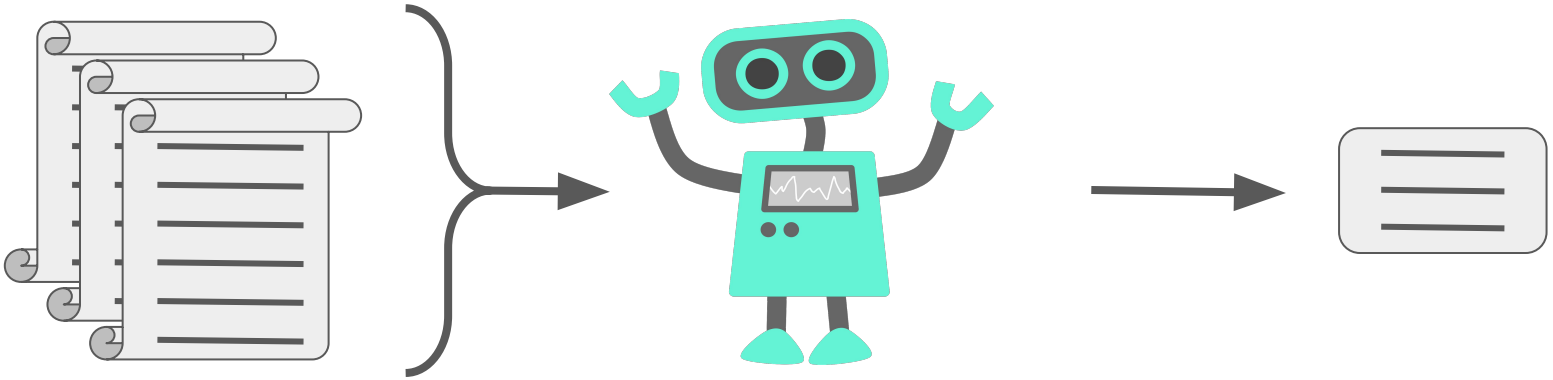
Ori Shapira, David Gabay,
Hadar Ronen, Judit Bar-Ilan, Yael Amsterdamer,
Ani Nenkova, Ido Dagan

EMNLP 2018

Texts to summarize

Summarization system

Summary

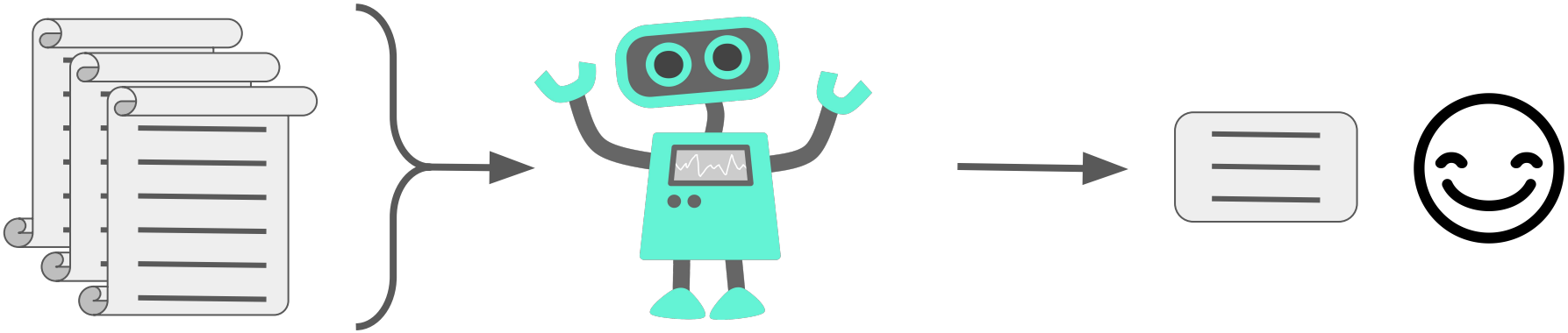


Summarization in a nutshell (multi-document setting)

Text to summarize

Summarization system

Summary

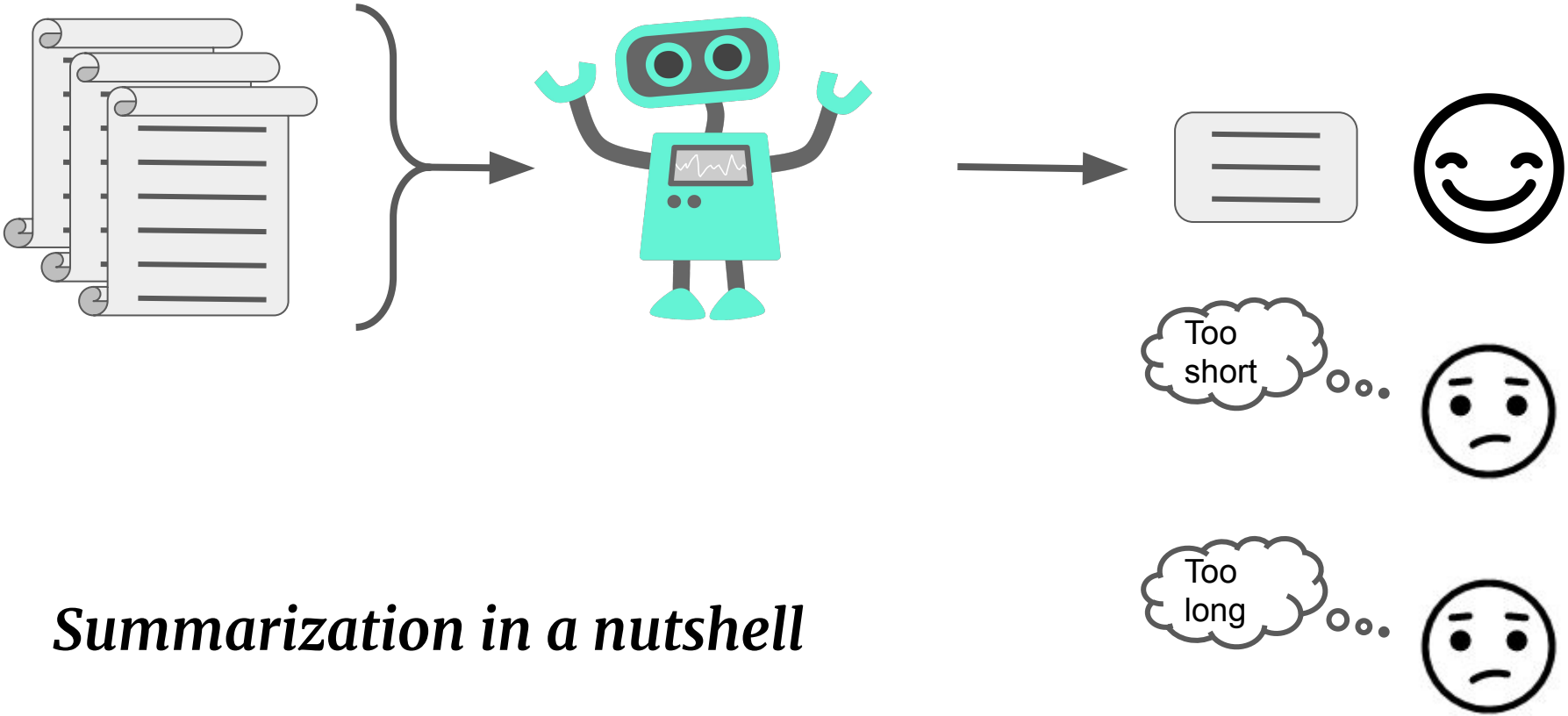


Summarization in a nutshell

Text to summarize

Summarization system

Summary

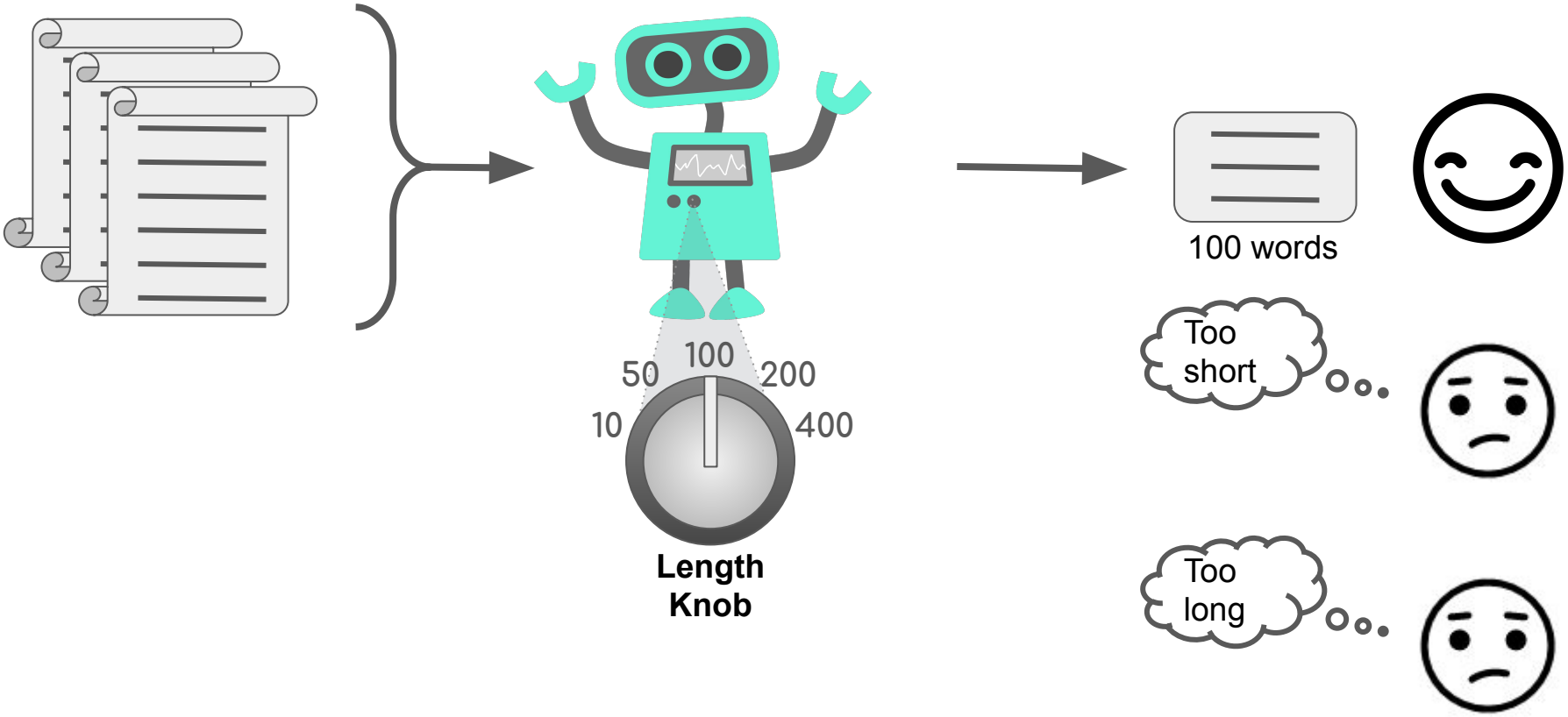


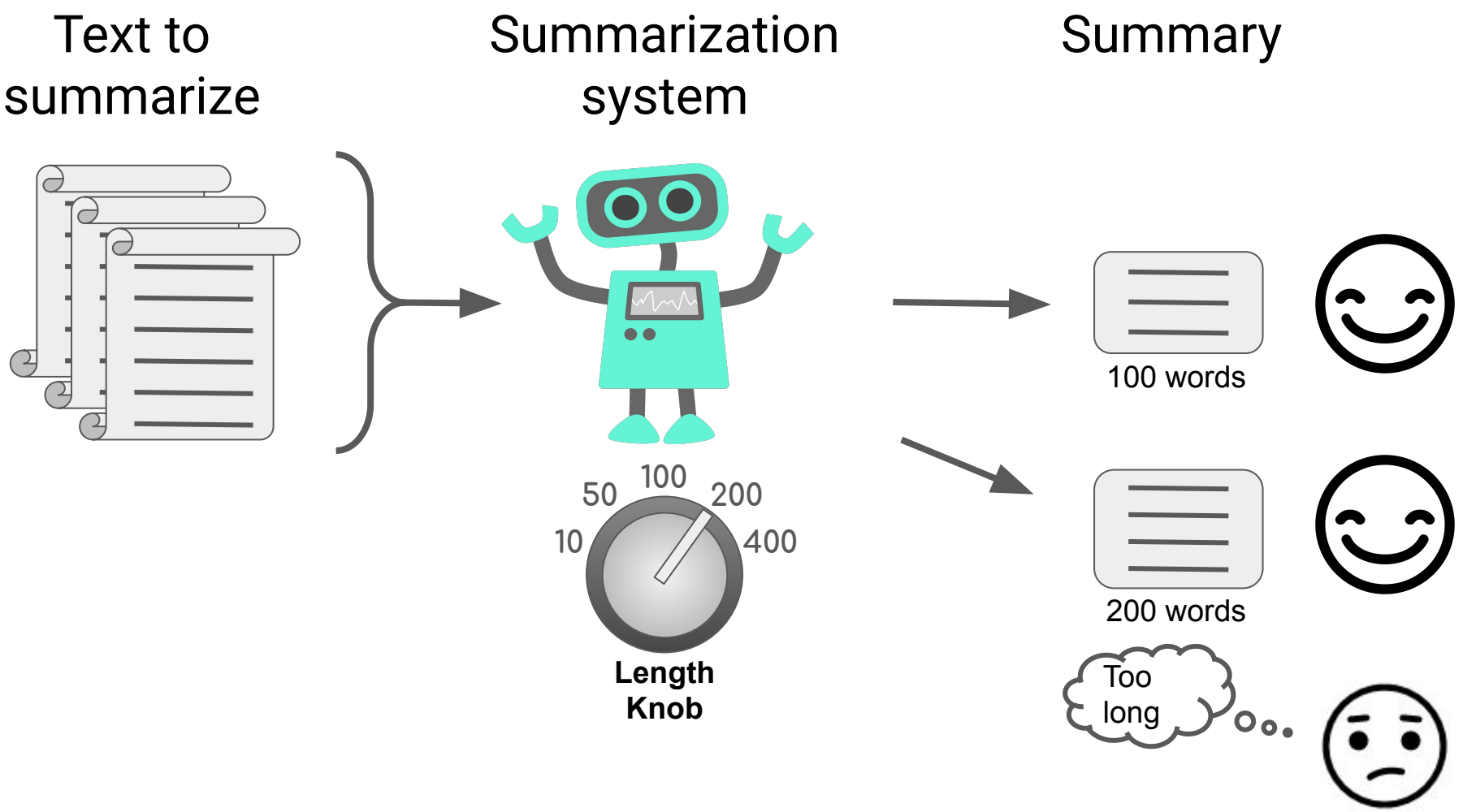
Summarization in a nutshell

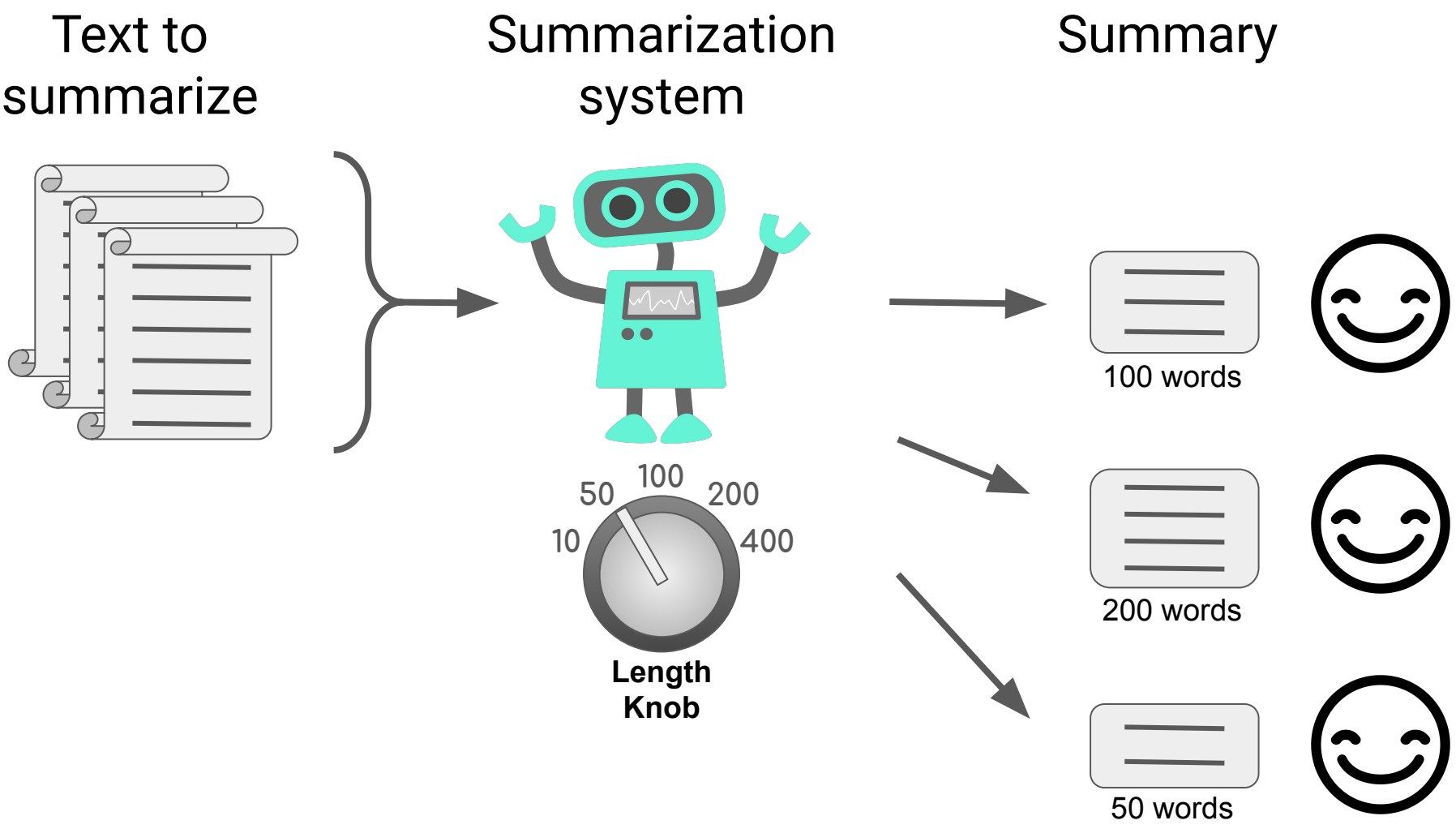
Text to summarize

Summarization system

Summary




Text to summarize
Summarization system
Summary
10
50
100
200
400
Length Knob
100 words
200 words
Too long


Text to summarize
Summarization system
Summary
100 words
200 words
50 words
10
50
100
200
400
Length Knob

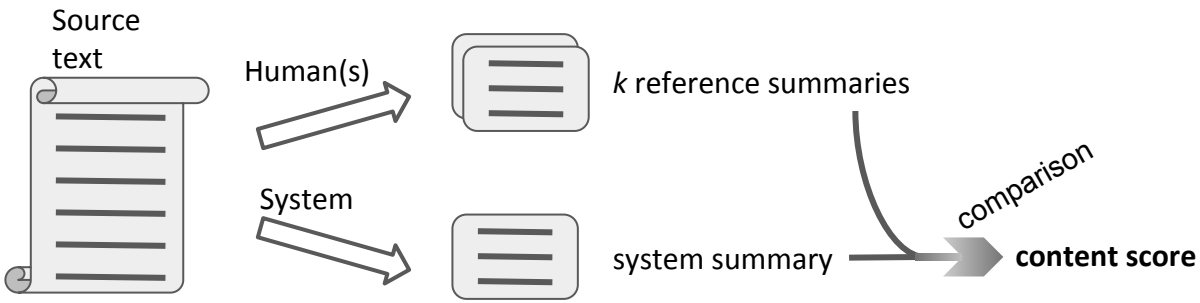
Multi-Length Summary Evaluation

- 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



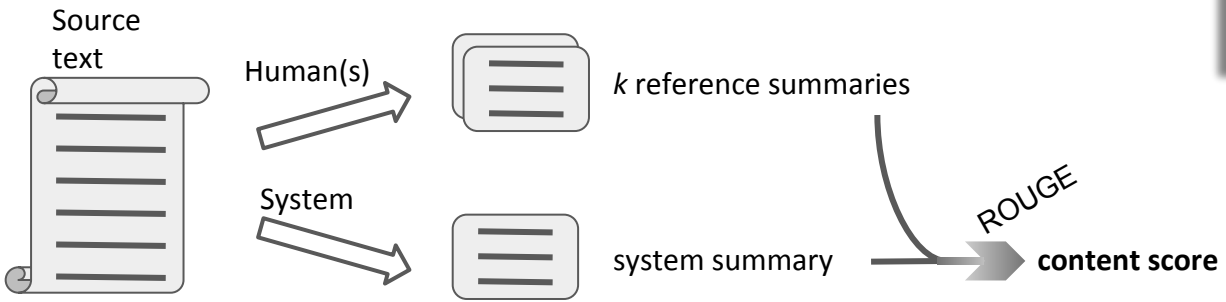
Automatic Evaluation

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

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

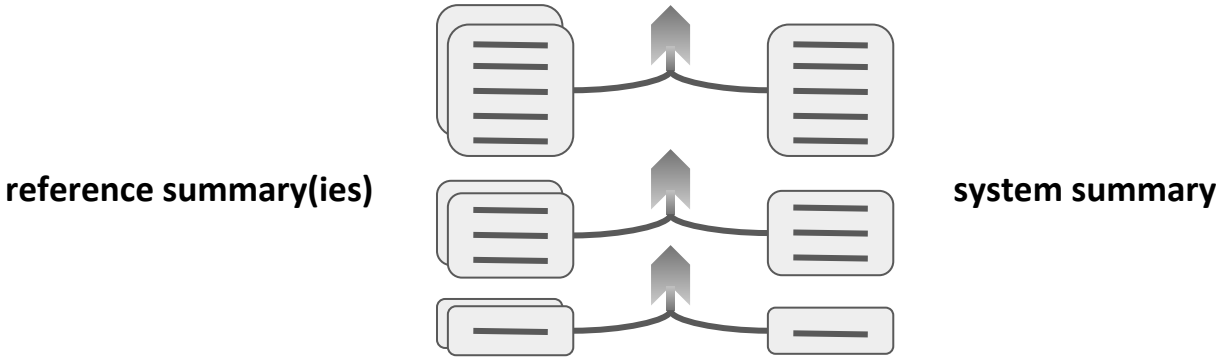
C.Y. Lin, 2004



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 miniscule 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 miniscule system ranking fluctuation between lengths

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

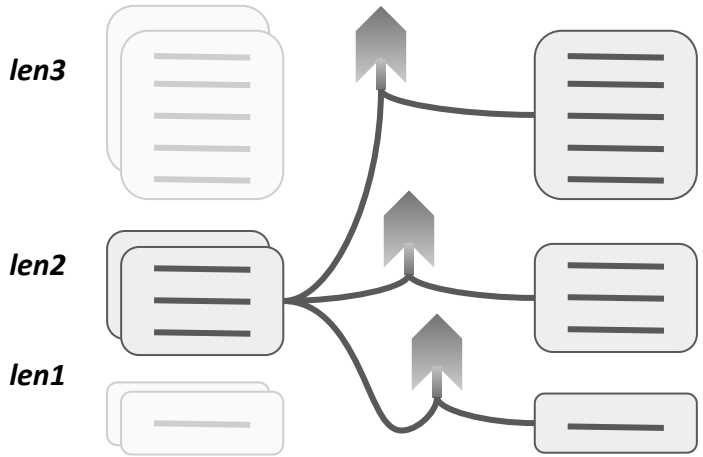
- More concretely - can we use a single reference summary length to evaluate system summaries of varying lengths?

  ⇒ *Need to assess validity of such method!*

Proposal: *Single* Reference Summary Length

To evaluate *multiple* system summary lengths!

reference summary(ies)



system summary

Standard vs. Proposed Evaluation

Assess automatic evaluation protocol by correlation to manual evaluation:

$$\textbf{\textit{Is}}\ \mathrm{Corr}_{\mathrm{Proposed}}\ \textbf{\textit{at least as good as}}\ \mathrm{Corr}_{\mathrm{Standard}}\ \textbf{\textit{?}}$$

Standard vs. Proposed Evaluation

Assess automatic evaluation protocol by correlation to manual evaluation:

$$\textbf{\textit{Is}}\ \mathrm{Corr}_{\mathrm{Proposed}}\ \textbf{\textit{at least as good as}}\ \mathrm{Corr}_{\mathrm{Standard}}\ \textbf{\textit{?}}$$

Data for analysis available from DUC 2001/2002 with:

- reference and system summaries
- manual summary scores

Results

Main observation:

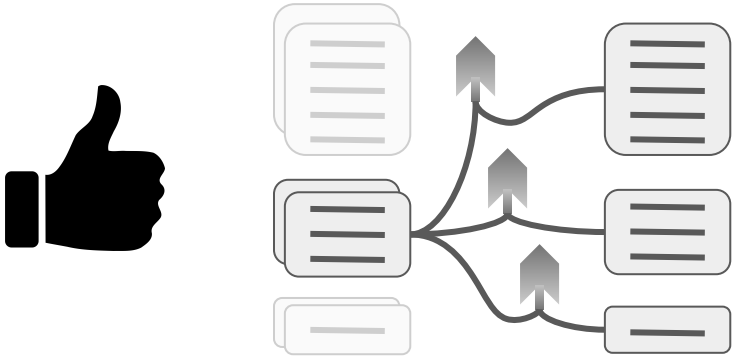
- A ***single*** set of reference summaries of length **50**, is as effective as ***separate*** sets of reference summaries per length
- ***Longer*** reference summary sets are overall less effective

Reference Set	System Summary Length: 50 3refs	50 1ref	100 3refs	100 1ref	200 3refs	200 1ref	400 3refs	400 1ref	Avg. across lengths 3refs	Avg. across lengths 1ref
Standard	0.72	0.65	0.88	0.85	0.9	0.86	0.95	0.94	0.86	0.83
Only50	0	0	+0.02	0	+0.01	+0.04	+0.01	+0.02	**+0.010**	**+0.015**
Only100	-0.01	+0.04	0	0	+0.01	-0.01	+0.02	0	+0.005	+0.008
Only200	-0.09	-0.09	-0.06	-0.08	0	0	+0.01	-0.01	-0.035	-0.0045
Only400	-0.06	+0.02	-0.09	-0.09	-0.01	+0.03	0	0	-0.040	-0.010

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

One Reference Length-
for Multiple System Summary Lengths

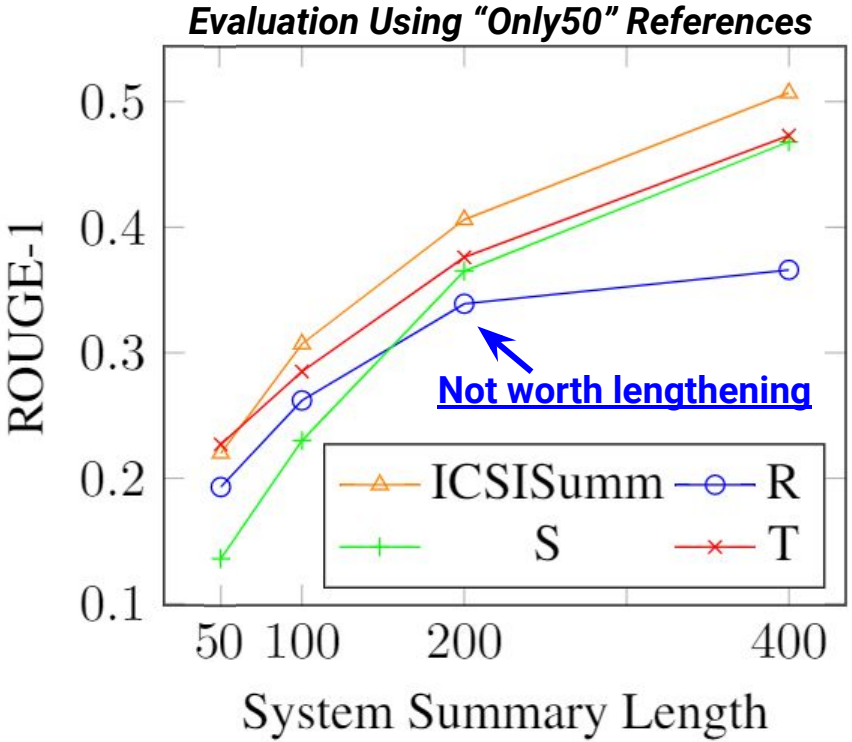
- Our finding: on DUC 2001 and 2002, correlation of automatic (ROUGE) assessment to human assessment using a *single set* of short reference summaries is as good as using a *different set* for each length



Additional Benefit:
Incremental Gain of Lengthening Summaries

A single reference set enables:

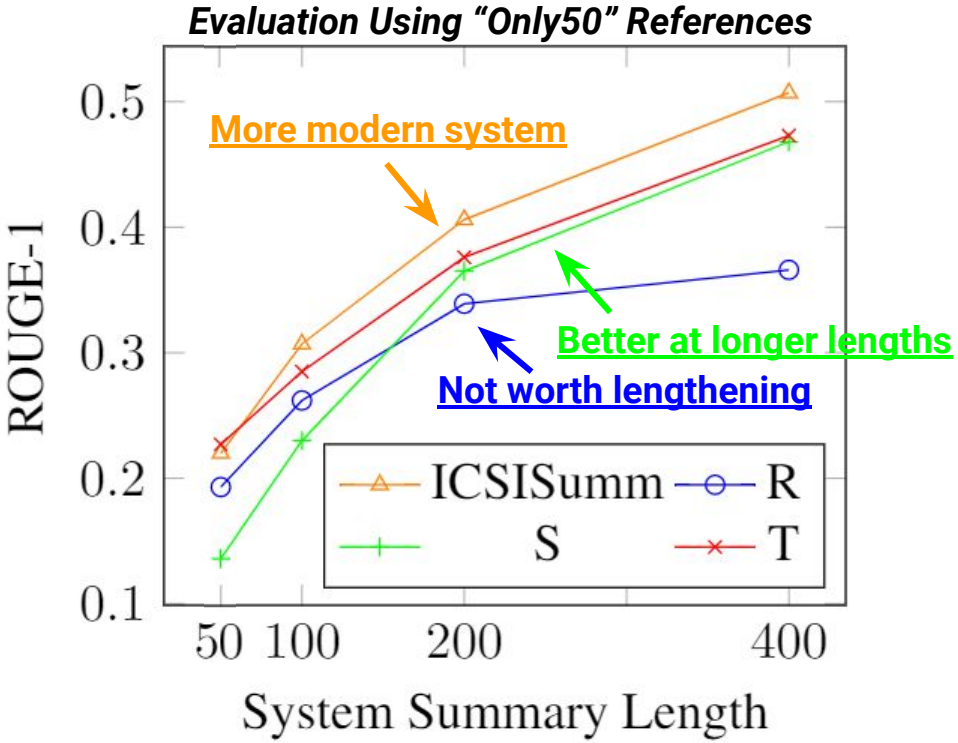
- Observing content gain as a system's summaries lengthen



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

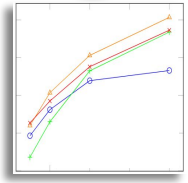
- Recent large-scale datasets include a single reference summary per input (document or document set)
- Summary varies across inputs
  - Vs. fixed length in early benchmarks
- ROUGE F1 - introduced to bias system summary content scope to equal that of the summary
  - Again - no “length knob”
- Our protocol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



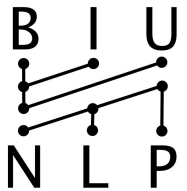
→ Varying-length summarization systems can be easily compared

Conception:

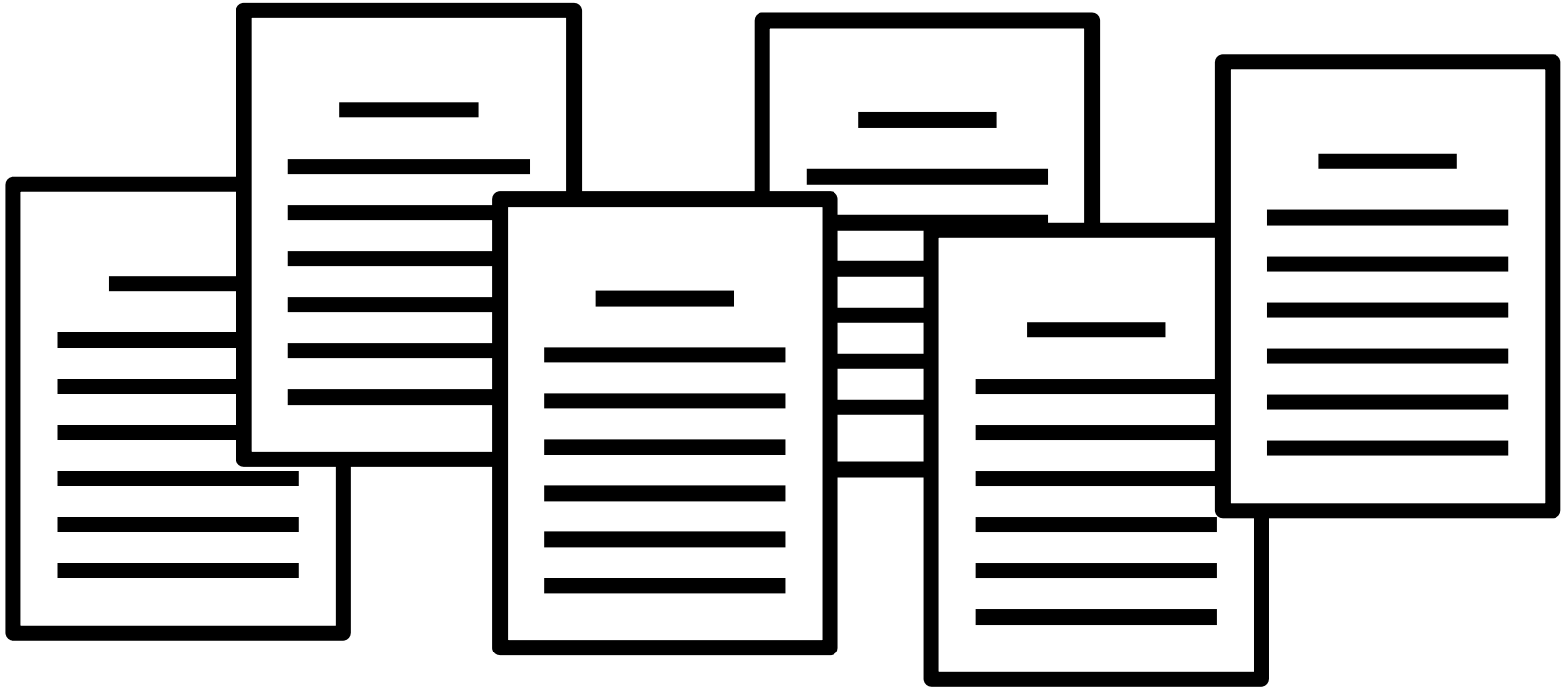
NLP eval is (mostly) about “static” tasks

Extending Multi-document Summarization Evaluation to the Interactive Setting

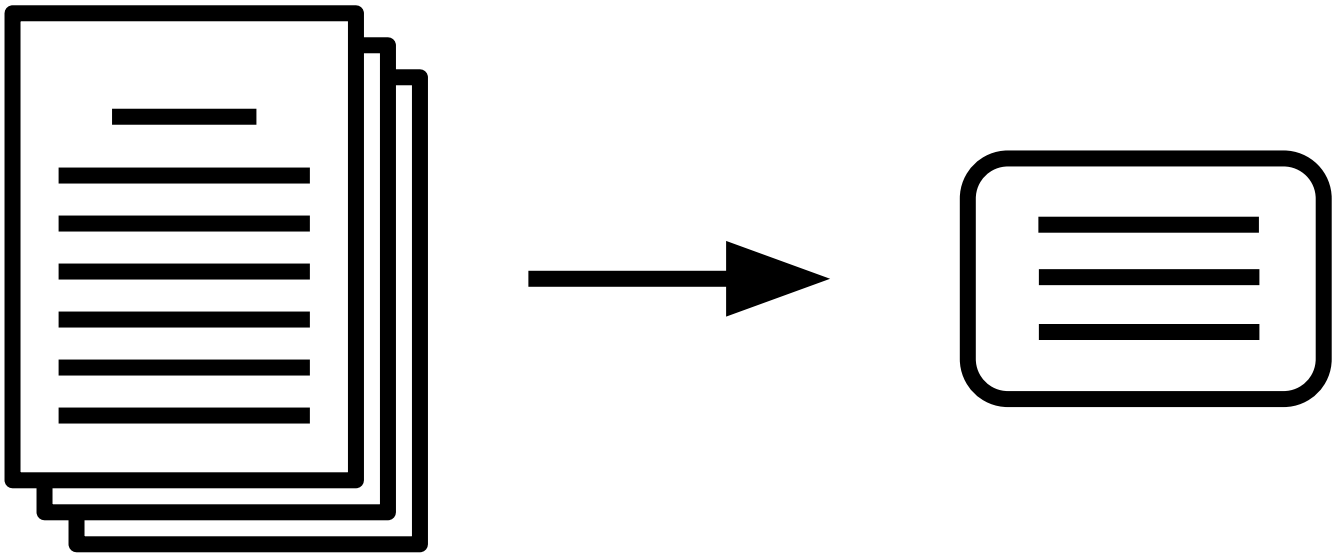
Lead researcher: Ori Shapira (ongoing)
With Ram Pasunuru, Mohit Bansal



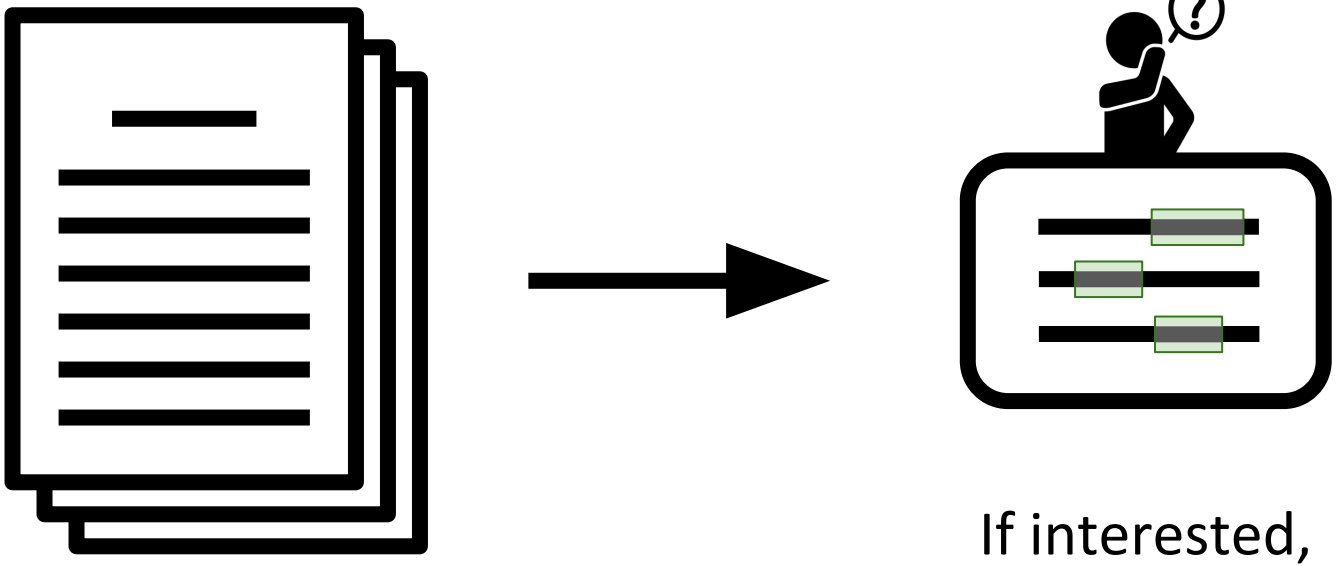
Multiple texts are hard to comprehend!



Is summarization enough?

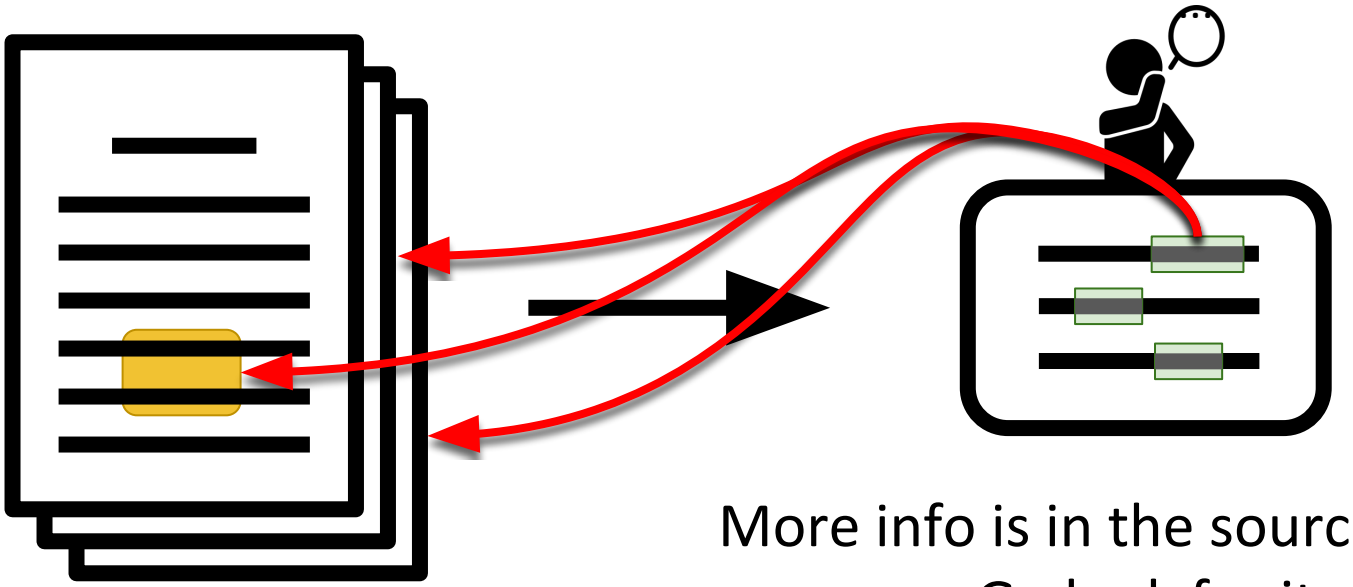


Is summarization enough?



If interested,
user wants to
explore more info!

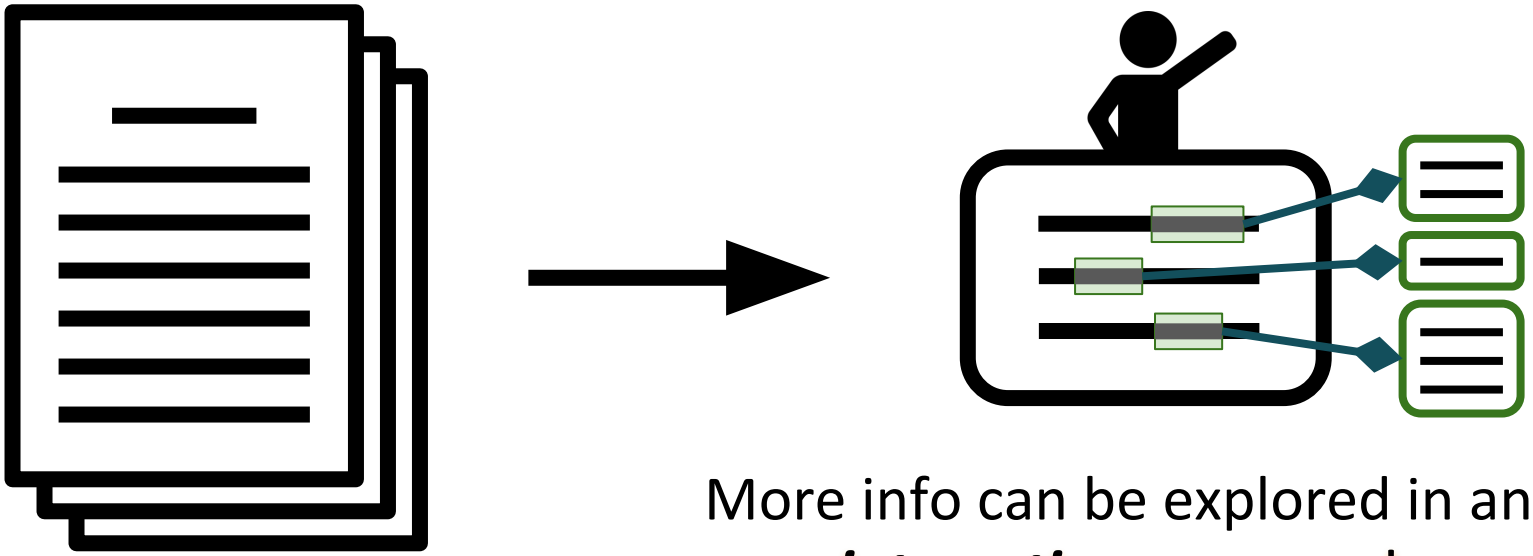
Summarization **Gap**



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

😕

Interactive Summarization



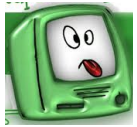
More info can be explored in an ***interactive*** summary!

Demo - prototype system

Interactive NLP

- Not much research conducted

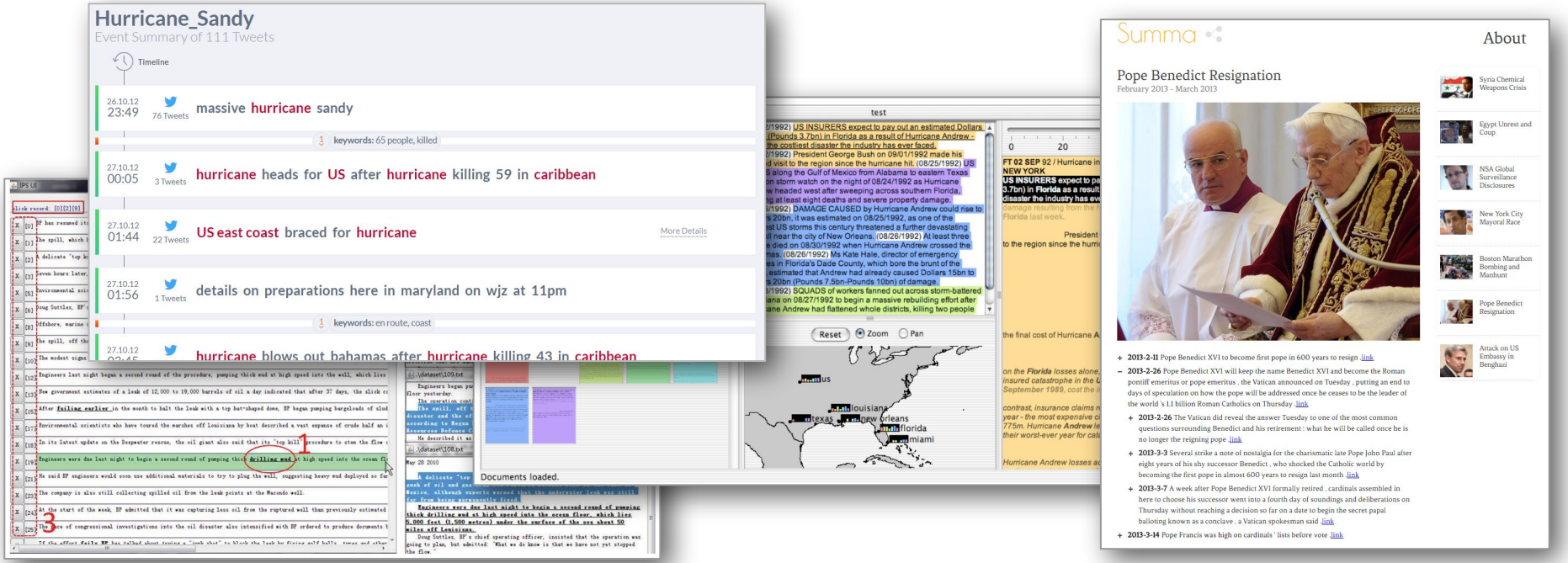
  - Very difficult to evaluate
  - Involves humans… eek!
- But very important



  - 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)



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

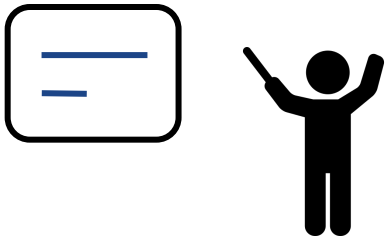
Interaction Setup Abstraction

- 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

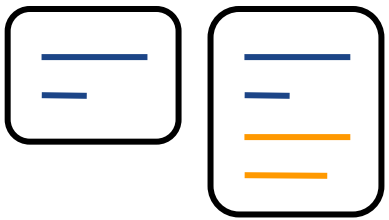
Interaction Setup Abstraction

- 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



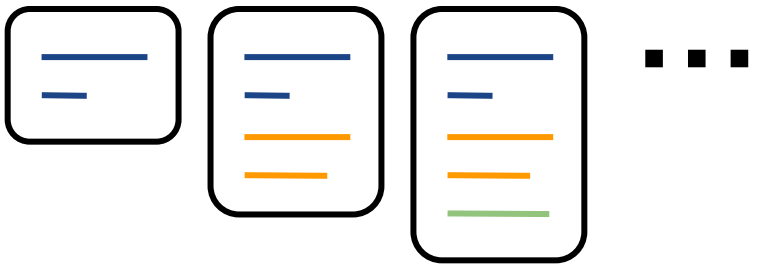
Second snapshot

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

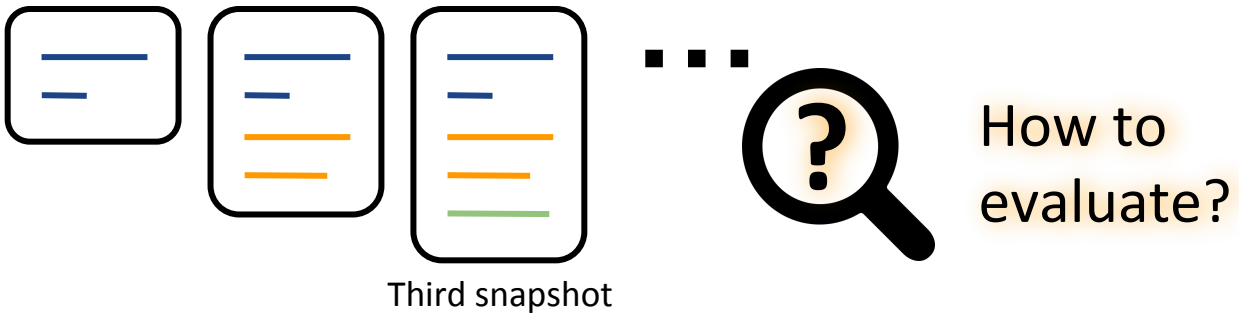
- After each interaction, new text is presented
- Interactive summarization = an *incrementally growing* summary
- Each interaction yields a snapshot summary of the information presented so far



Third 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

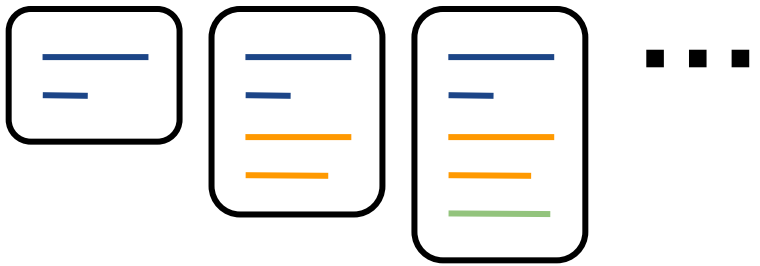


Our approach: Adapt static evaluation methods

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

How is an interactive summary then evaluated?

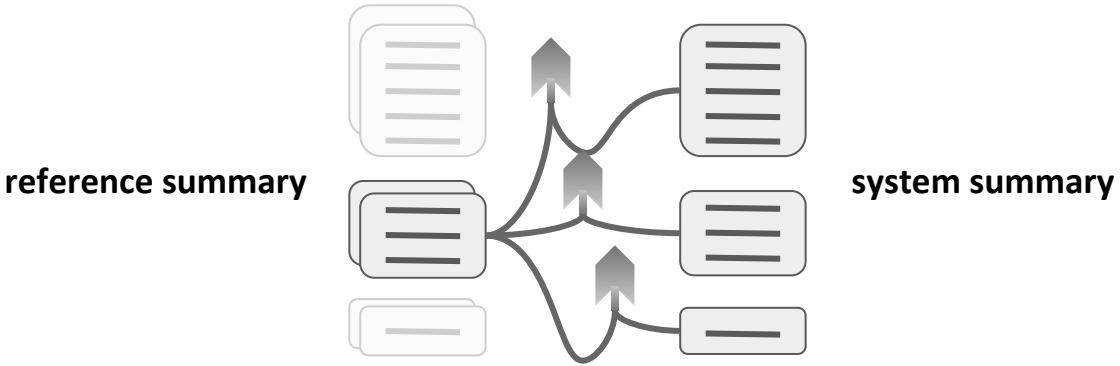
- In a user session, each snapshot can be evaluated as a static summary, with a recall-based measure



Evaluating the “Snapshots”

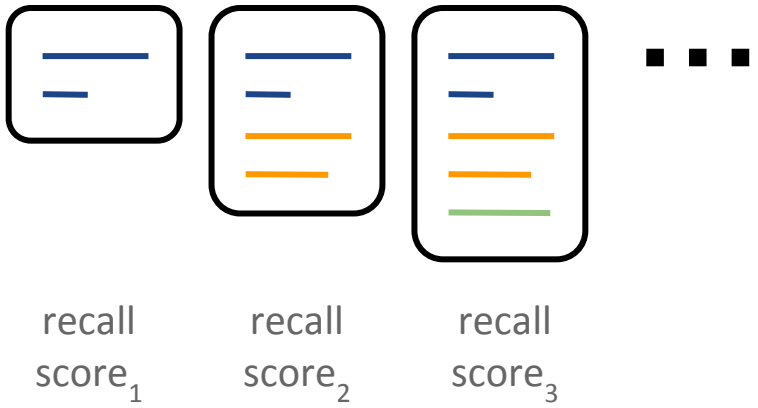
Based on our previous **single reference summary** length result:

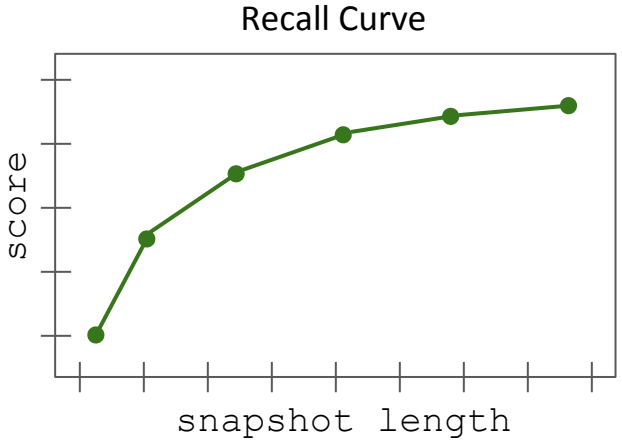
→ We can use a standard MDS dataset



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

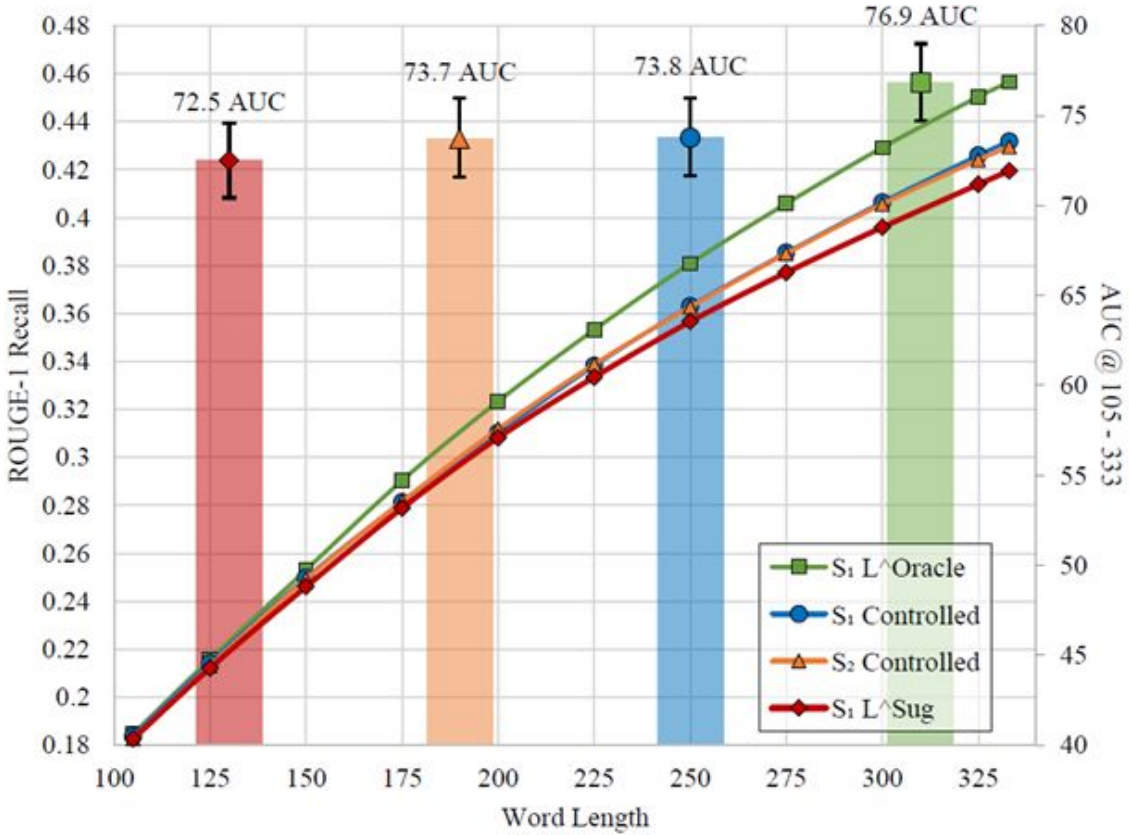




- *Score increases as text accumulates*
- *AUC can provide an overall score*

Real Interactive Evaluation Example

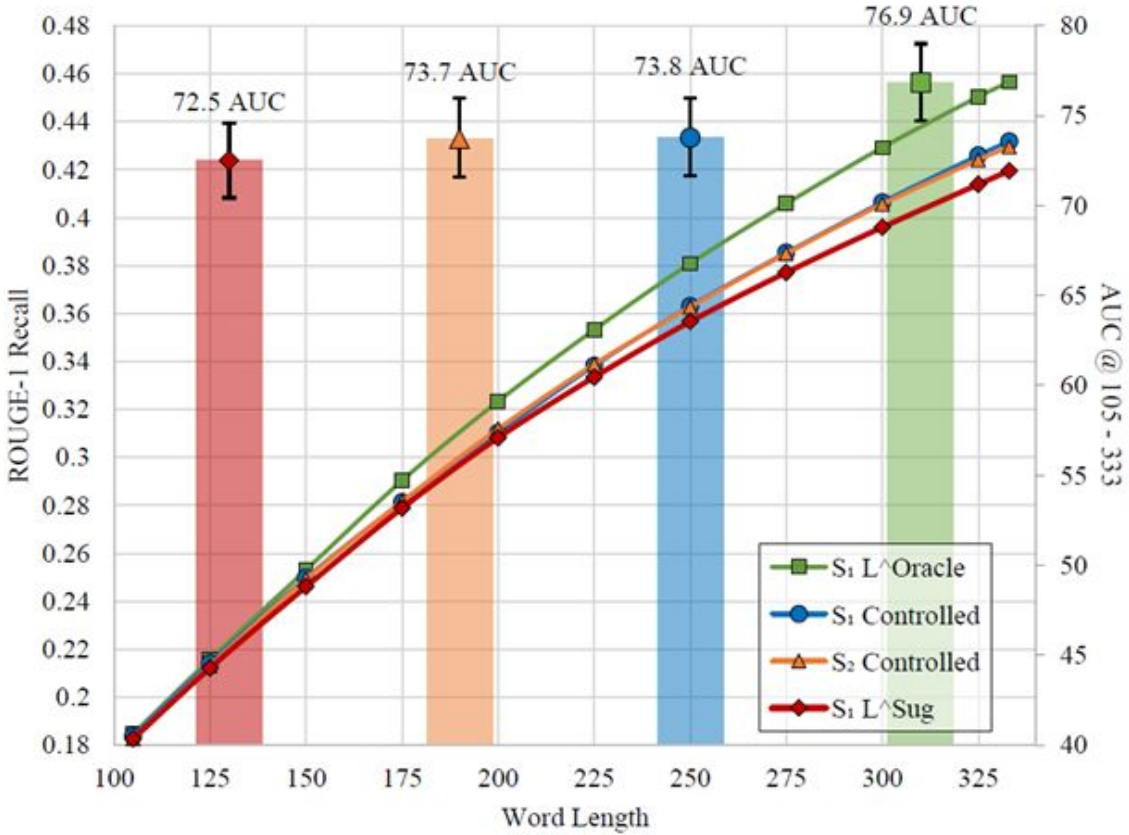
- Two baseline systems, crowdsourced sessions
- Upper bound
  - “Oracle” queries
- Lower bound
  - No human involvement



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 moneyyy!! $$$”
  - 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%
% ΔAUC from lower bound	+1.8%	−1.4%

More engaged

More thought in queries

Better than fully automatic baseline

Conclusions: Interactive Summarization

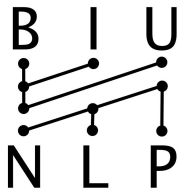
- Interactive summarization addresses the *real* user need
  - *Exploring* multiple-texts - closing the summarization gap
- A viable research task - via proposed evaluation framework
- Opens up plenty to research: modeling, interaction modes

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

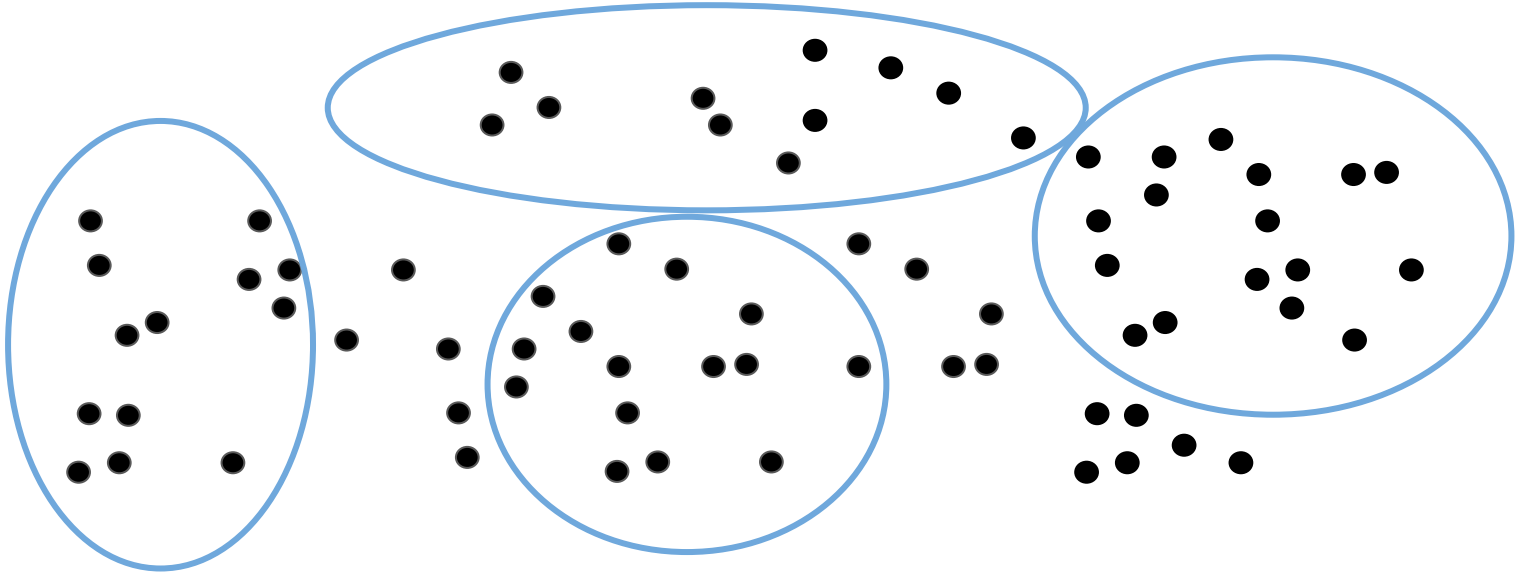
Relation Extraction

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

*Downstream tasks leverage coreference **links** to bridge information across mentions*

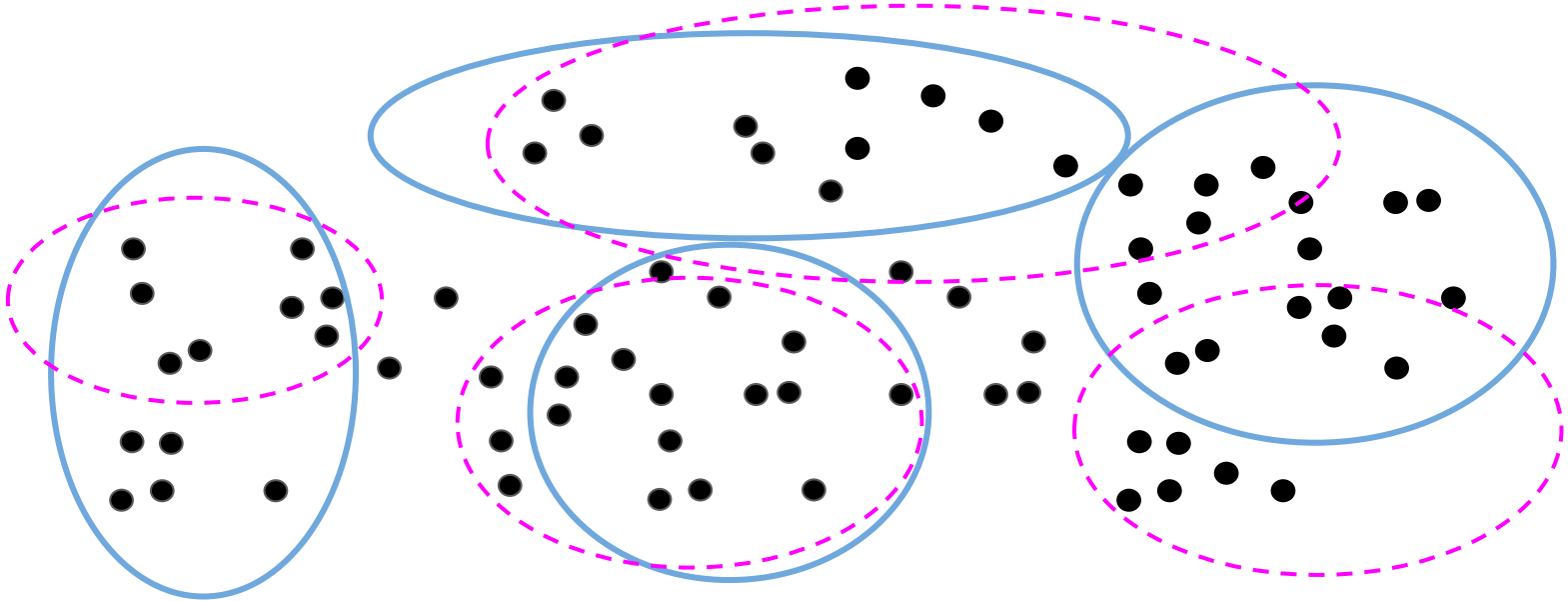
Evaluating Coreference Resolution as a Clustering Task

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	Hybrid	Mention-based
MUC *(Vilain et al., 1995)*	**LEA** *(Moosavi et al., 2016)*	**B^3** *(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

discover

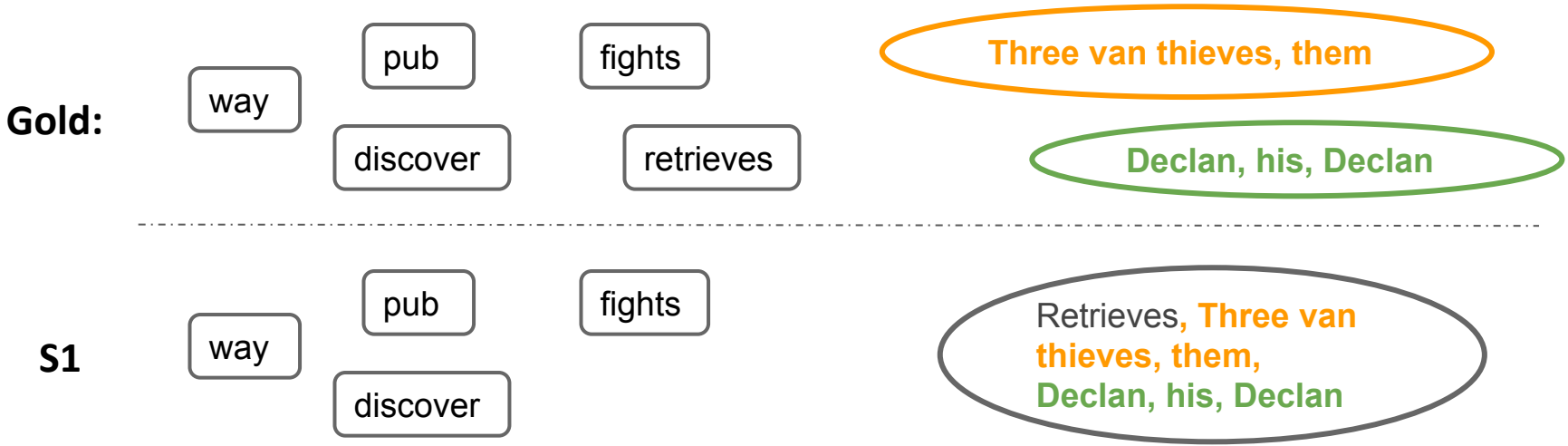
fights

retrieves

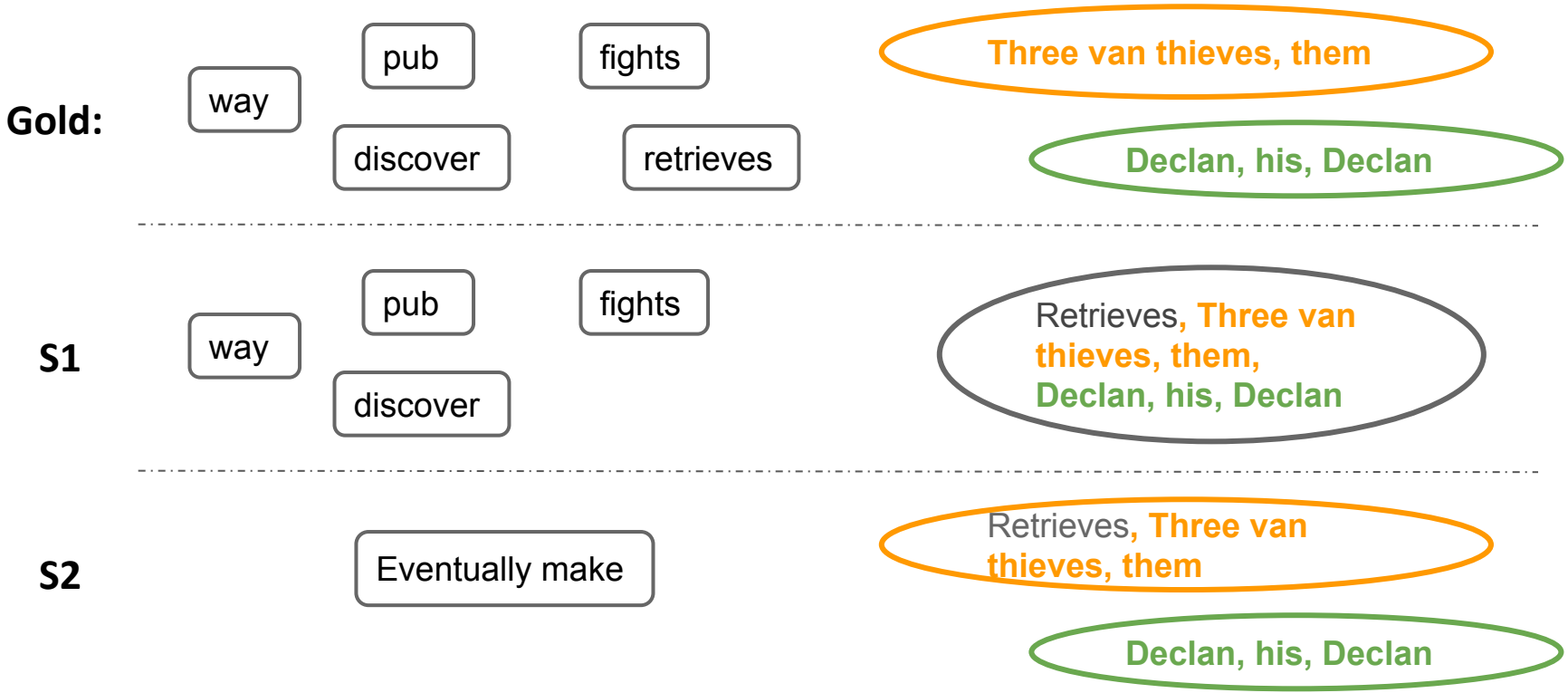
Three van thieves, them

Declan, his, Declan

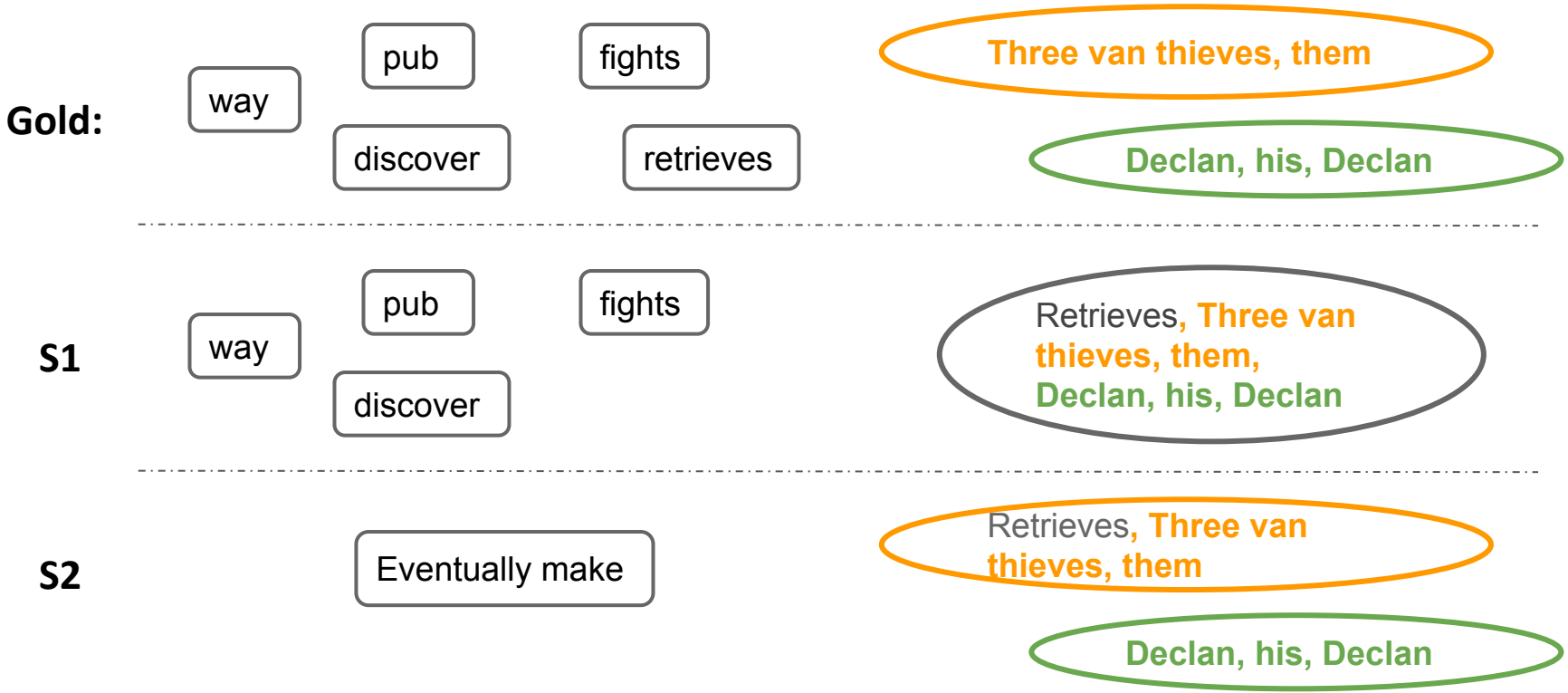
Singleton Effect



Singleton Effect



Singleton Effect



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
(Barhom et al, ACL 2019)	79.5
Our current system	81.0

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 ↓
Our current system	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 ↓
Our current system	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