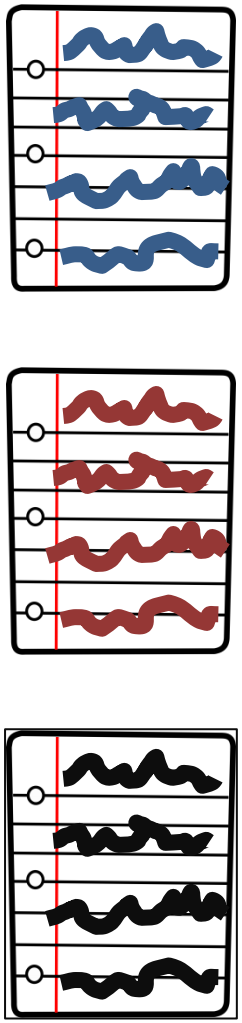
LQVSumm: A Corpus of **L**inguistic **Q**uality **V**iolations in Multi-Document **Summ**arization

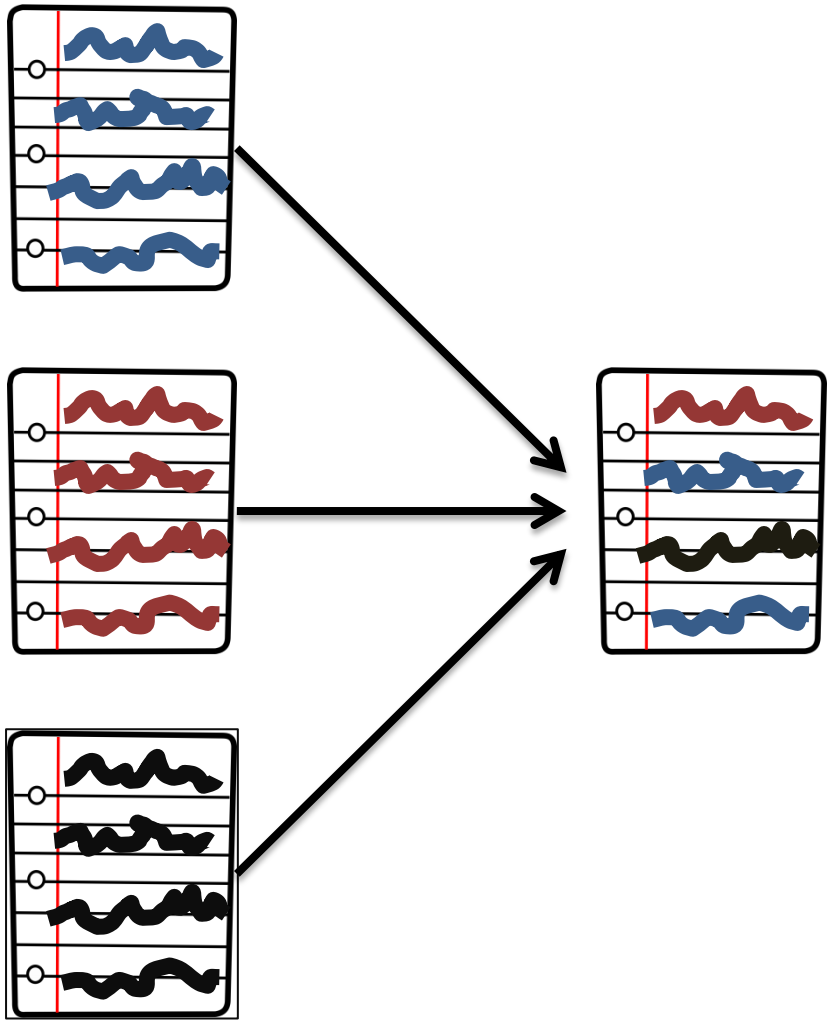
Annemarie Friedrich, Marina Valeeva and Alexis Palmer

COMPUTATIONAL LINGUISTICS & PHONETICS
SAARLAND UNIVERSITY, GERMANY

Extractive Multi-Document Summarization



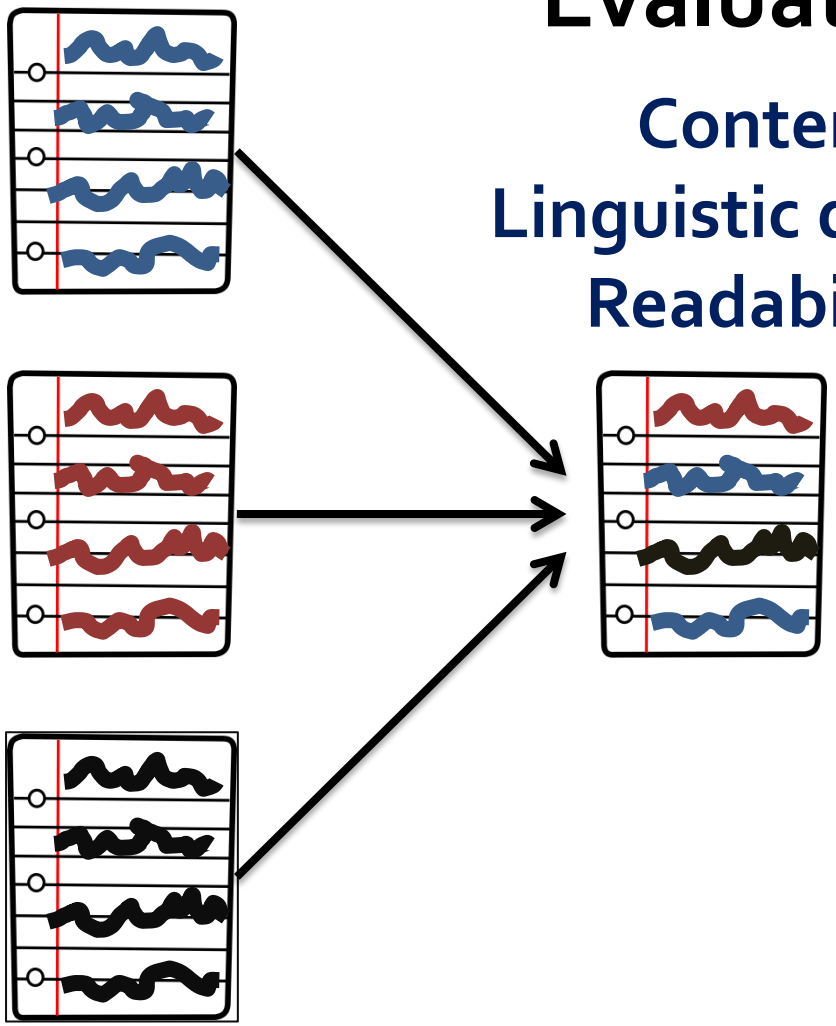
Extractive Multi-Document Summarization



Extractive Multi-Document Summarization

Evaluation

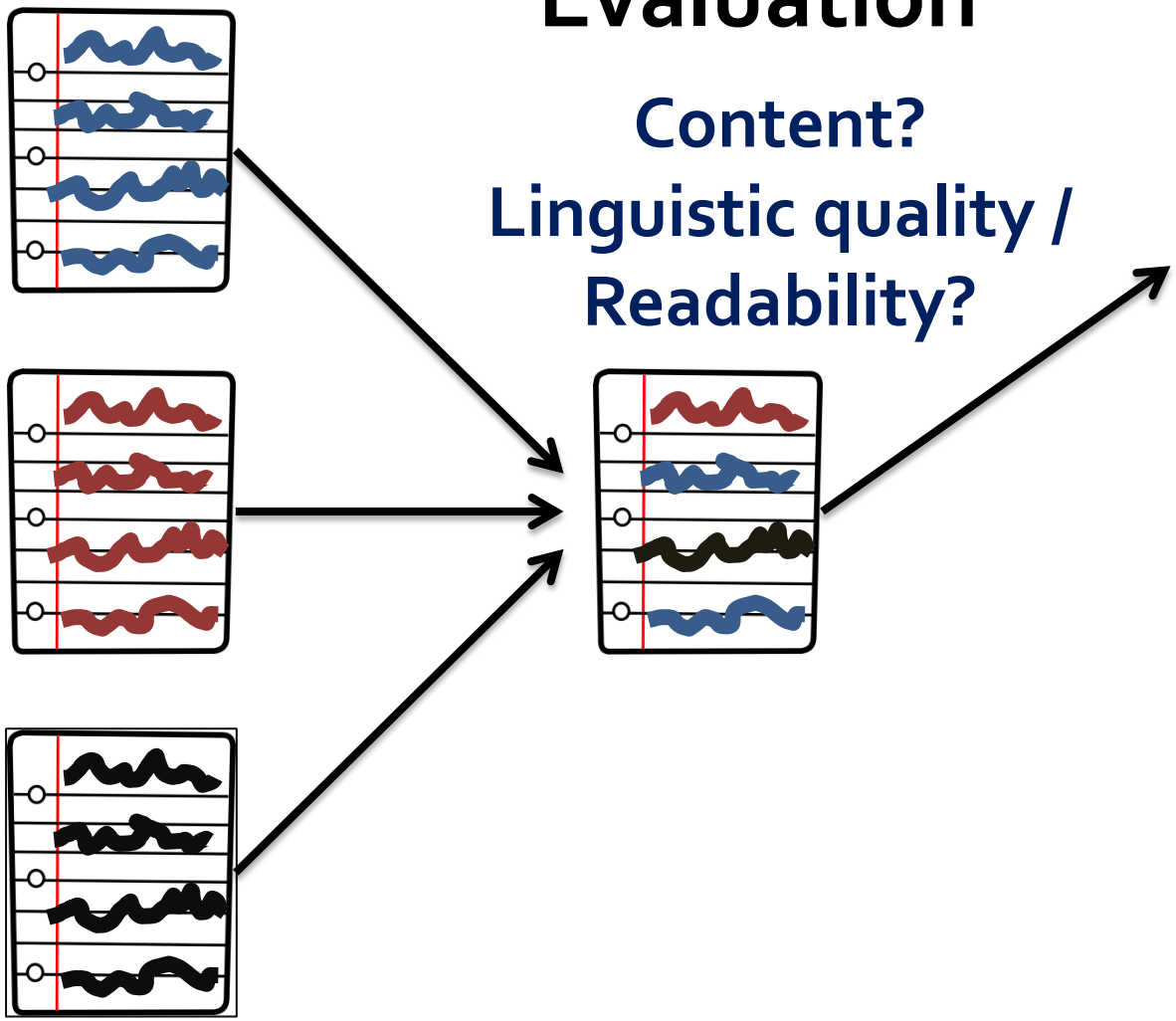
Content?
Linguistic quality / Readability?

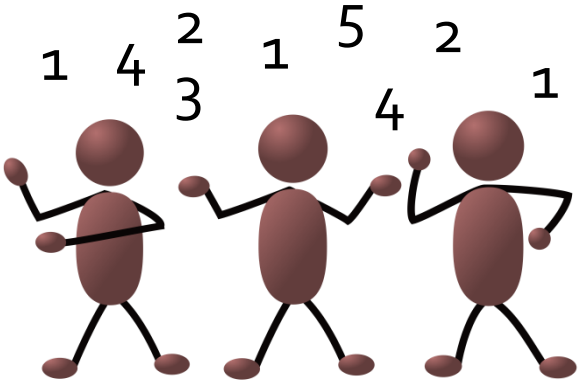


Extractive Multi-Document Summarization

Evaluation

Content?
Linguistic quality / Readability?

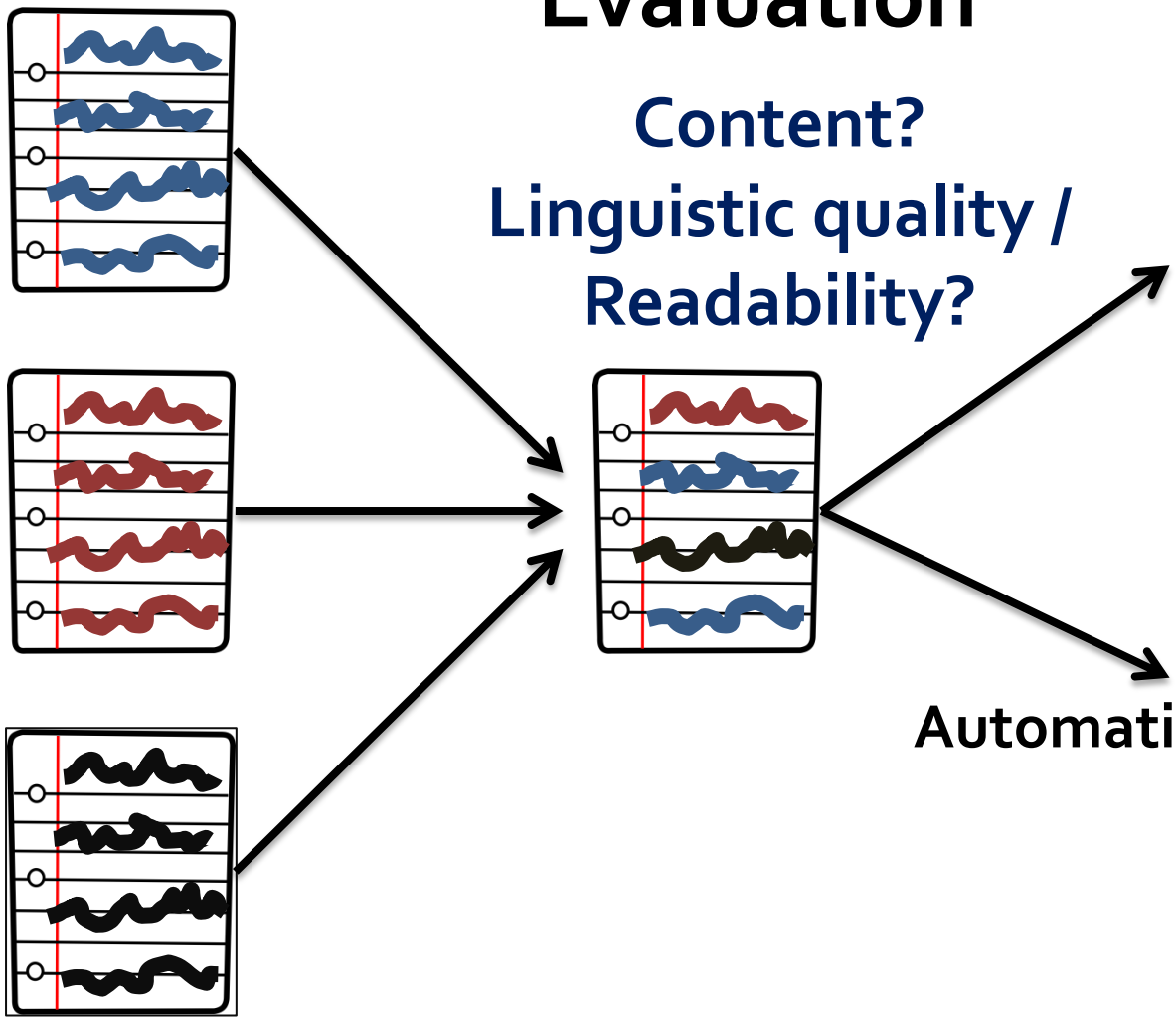


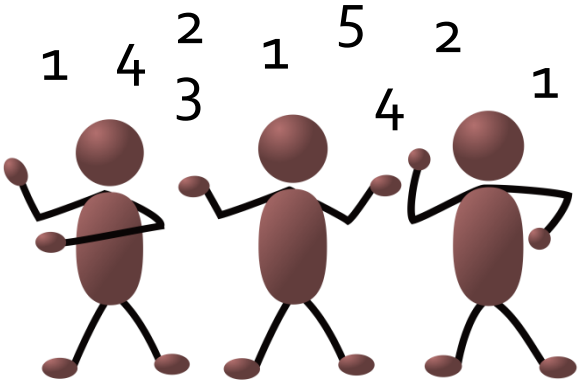


Extractive Multi-Document Summarization

Evaluation

Content?
Linguistic quality / Readability?



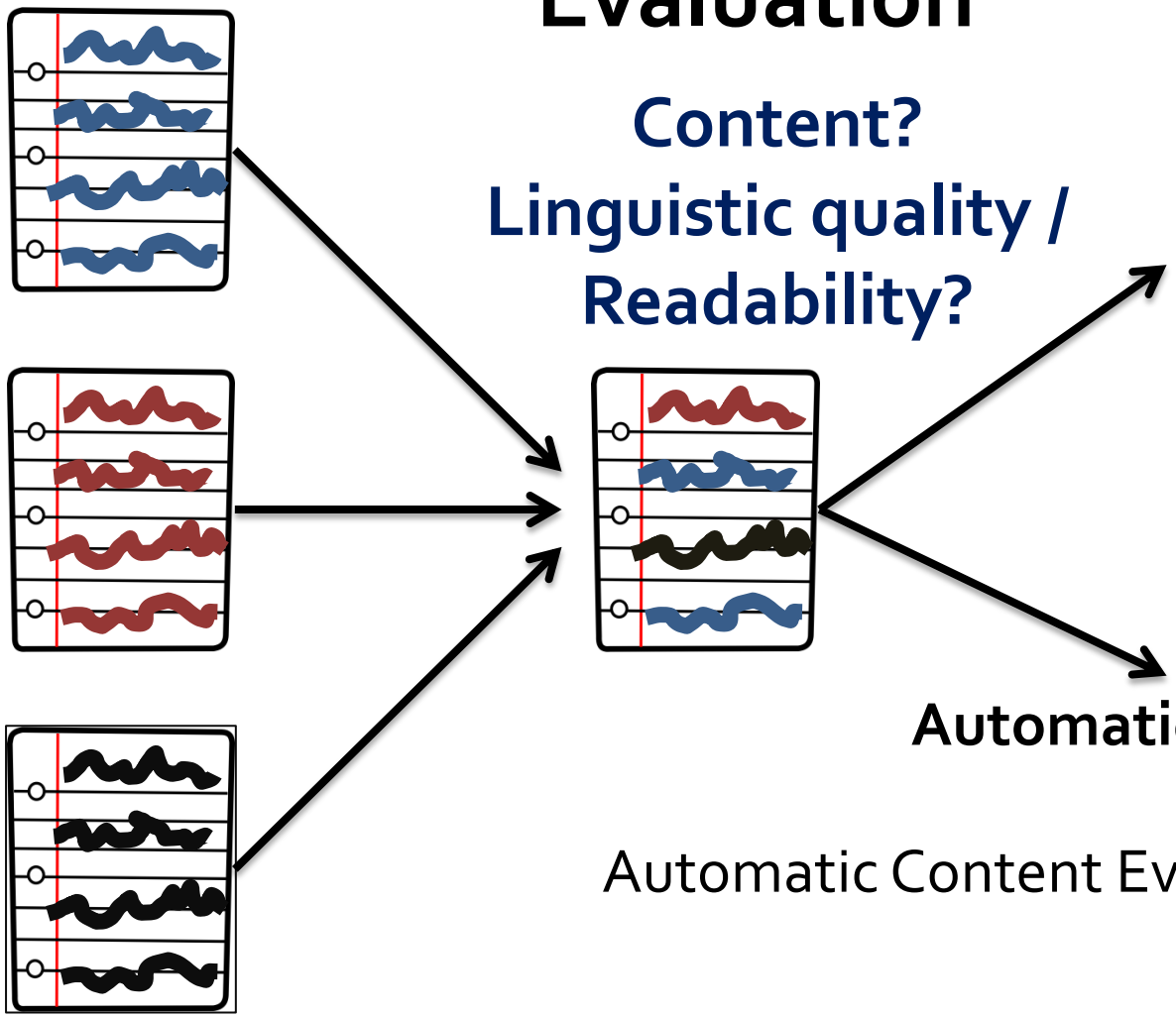


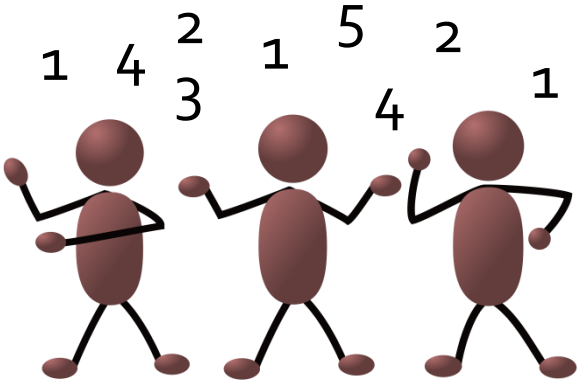
Automatic Evaluation Methods

Extractive Multi-Document Summarization

Evaluation

Content?
Linguistic quality / Readability?





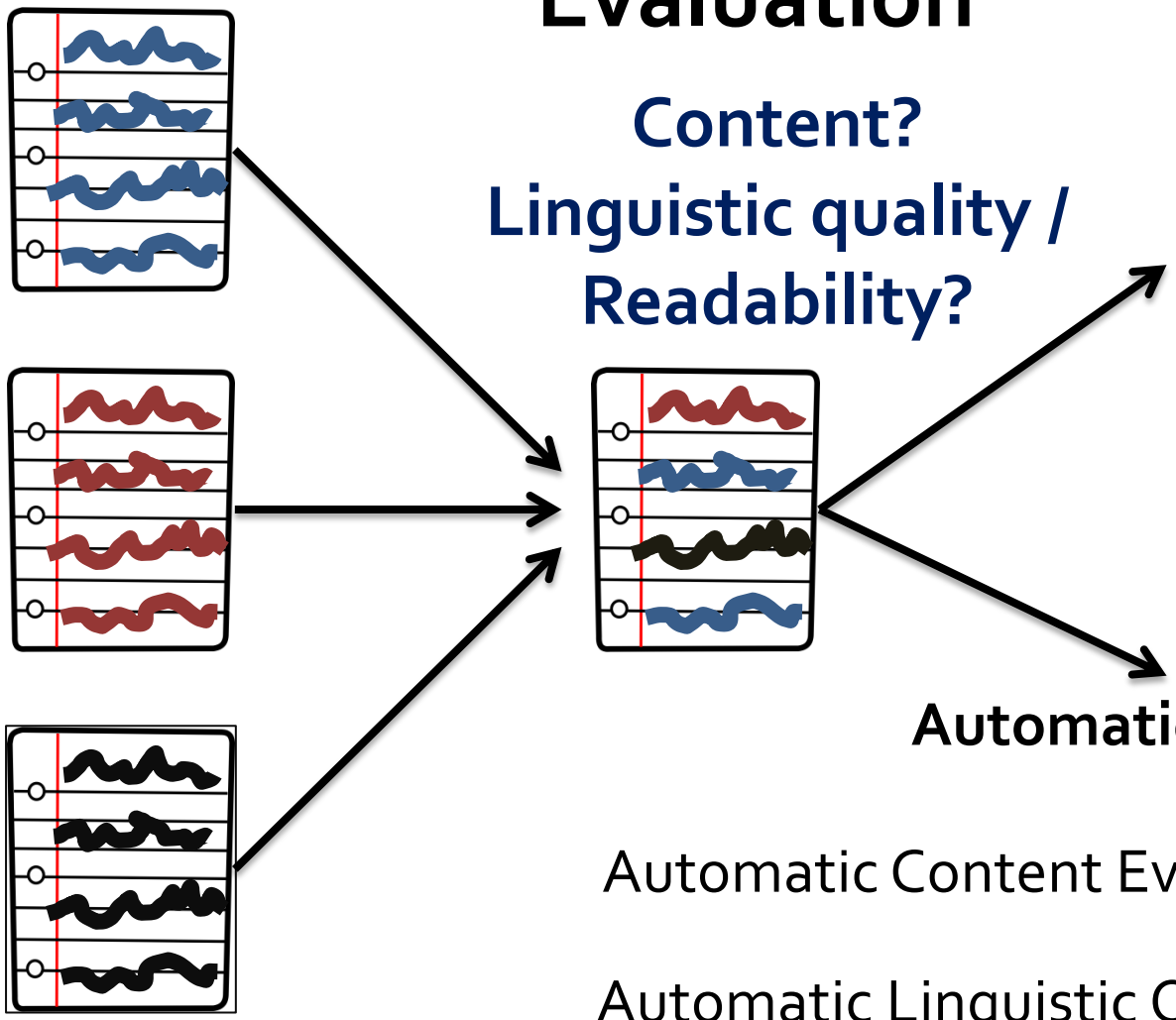
Automatic Evaluation Methods

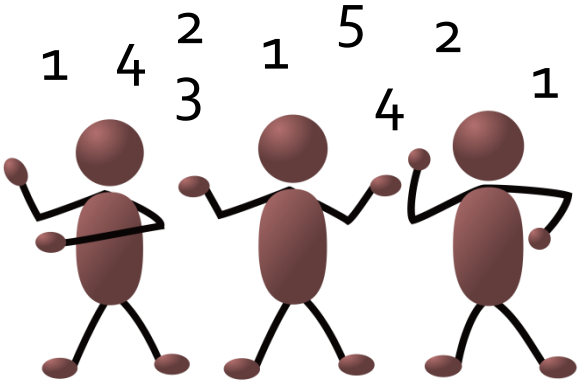
Automatic Content Evaluation ✓

Extractive Multi-Document Summarization

Evaluation

Content?
Linguistic quality / Readability?



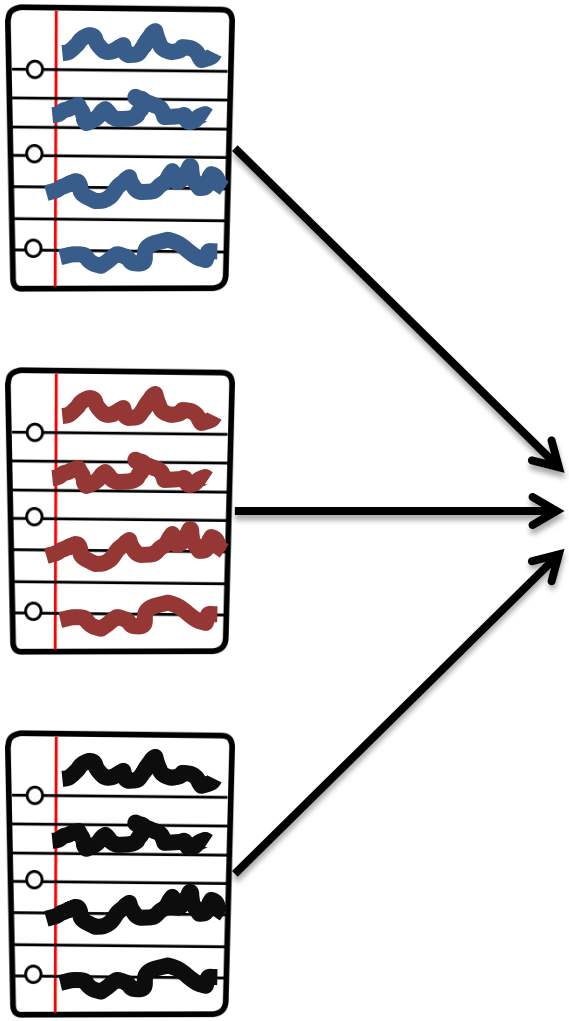


Automatic Evaluation Methods

Automatic Content Evaluation ✓

Automatic Linguistic Quality Evaluation ?

Violations of Linguistic Quality



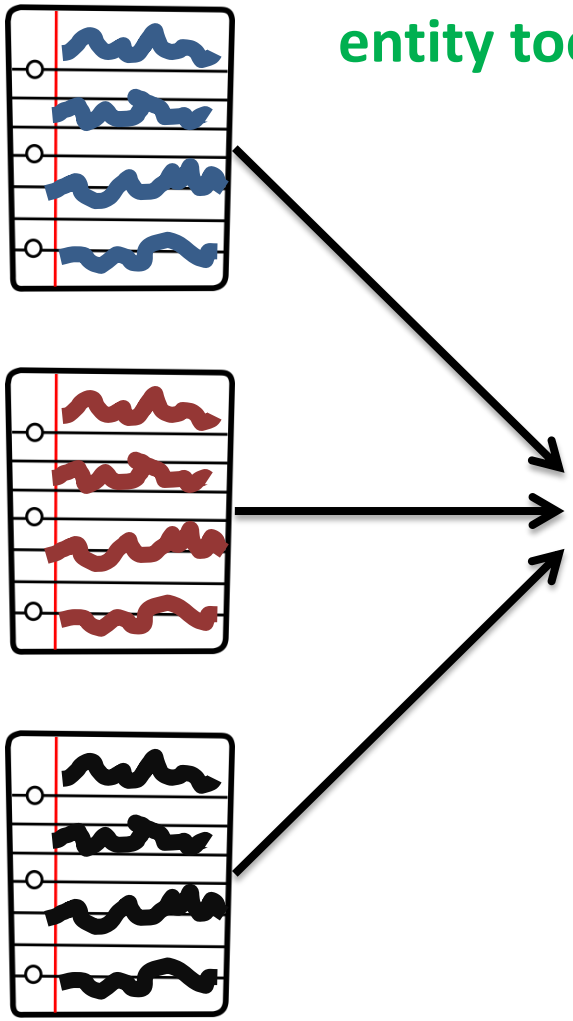
entity mentions: reference unclear

The suspect apparently called **her** from a cell phone shortly before the shooting began, saying he was "**acting out in revenge for something that happened 20 years ago**, Miller said. **The gunman, a local truck driver Charles Roberts**, was apparently **acting in "revenge" for an incident that happened to him 20 years ago**. Charles Carl Roberts IV may have planned to

Violations of Linguistic Quality

subsequent mention of entity too specific

entity mentions: reference unclear

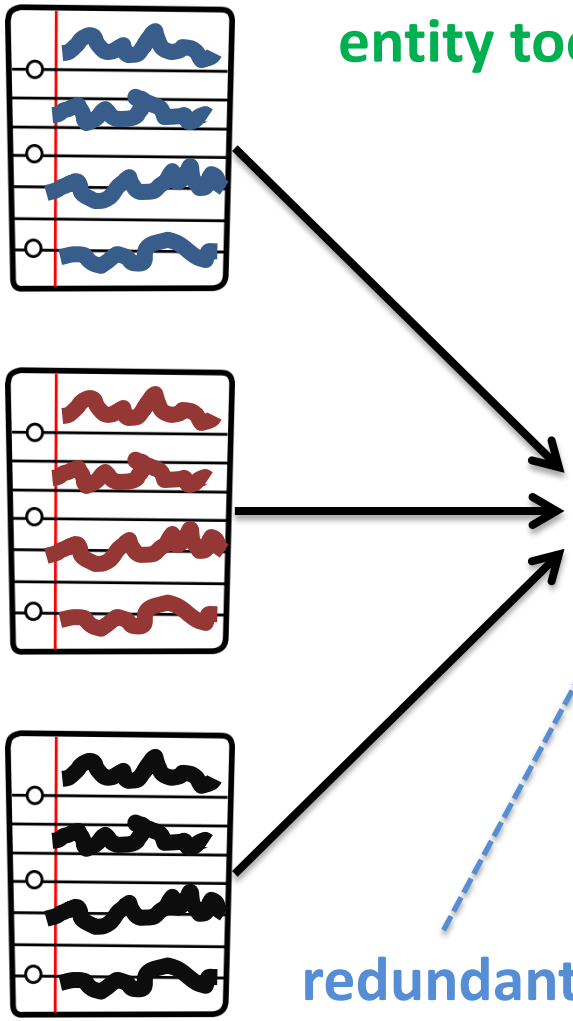


The suspect apparently called **her** from a cell phone shortly before the shooting began, saying he was "**acting out in revenge for something that happened 20 years ago**, Miller said. **The gunman, a local truck driver Charles Roberts**, was apparently **acting in "revenge" for an incident that happened to him 20 years ago**. Charles Carl Roberts IV may have planned to

Violations of Linguistic Quality

subsequent mention of entity too specific

entity mentions: reference unclear



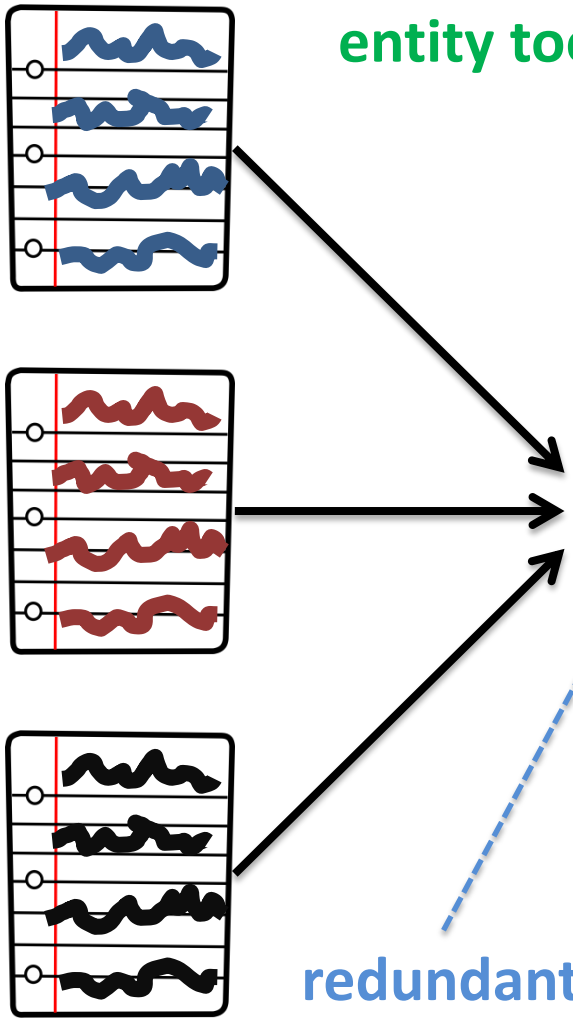
The suspect apparently called **her** from a cell phone shortly before the shooting began, saying he was "**acting out in revenge for something that happened 20 years ago**, Miller said. **The gunman, a local truck driver Charles Roberts**, was apparently **acting in "revenge" for an incident that happened to him 20 years ago**. Charles Carl Roberts IV may have planned to

redundant information

Violations of Linguistic Quality

subsequent mention of entity too specific

entity mentions: reference unclear

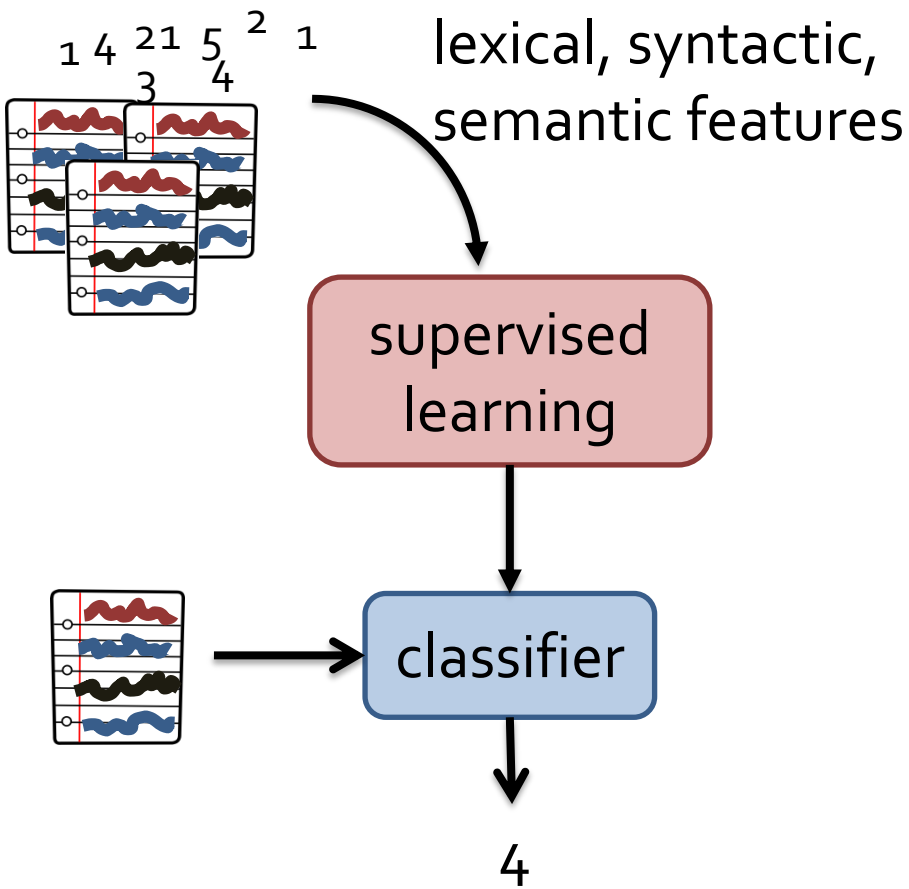


The suspect apparently called **her** from a cell phone shortly before the shooting began, saying he was "**acting out in revenge for something that happened 20 years ago**, Miller said. **The gunman, a local truck driver Charles Roberts**, was apparently **acting in "revenge" for an incident that happened to him 20 years ago**. Charles Carl Roberts IV may have planned to

redundant information

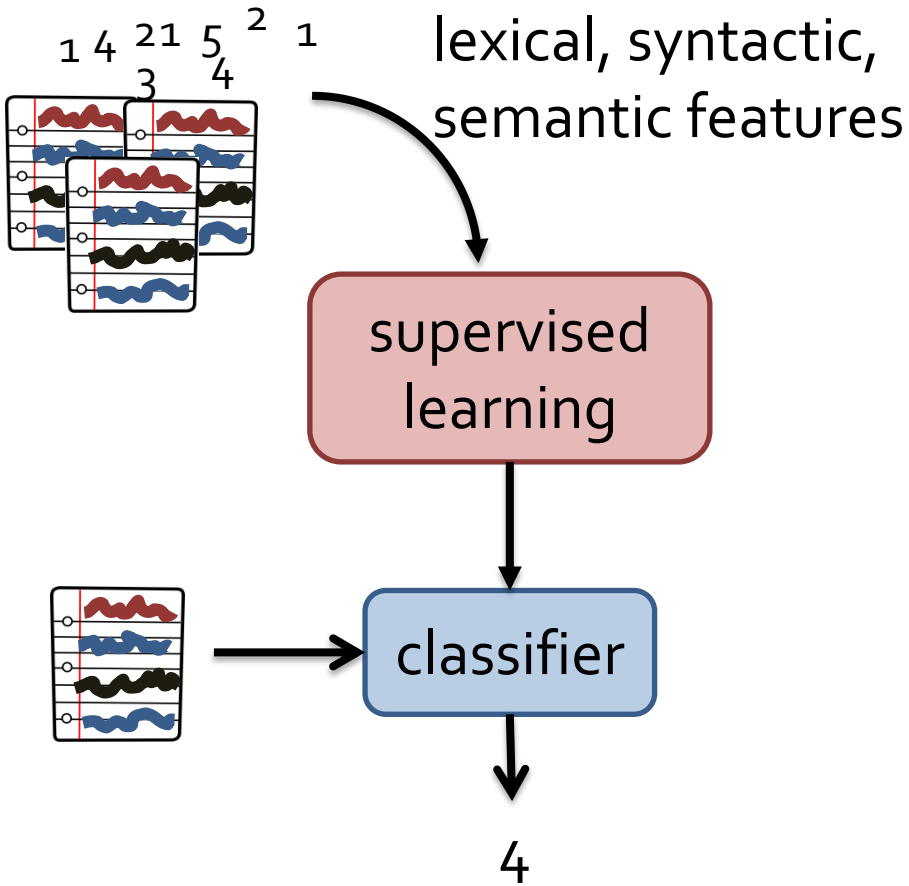
incomplete sentence

Automatic Evaluation of Linguistic Quality for Automatic Summarization



[Pitler et al., 2010; Conroy et al., 2011; Giannakopoulos and Karkaletsis, 2011; de Oliveira, 2011; Lin et al., 2012]

Automatic Evaluation of Linguistic Quality for Automatic Summarization



[Pitler et al., 2010; Conroy et al., 2011; Giannakopoulos and Karkaletsis, 2011; de Oliveira, 2011; Lin et al., 2012]

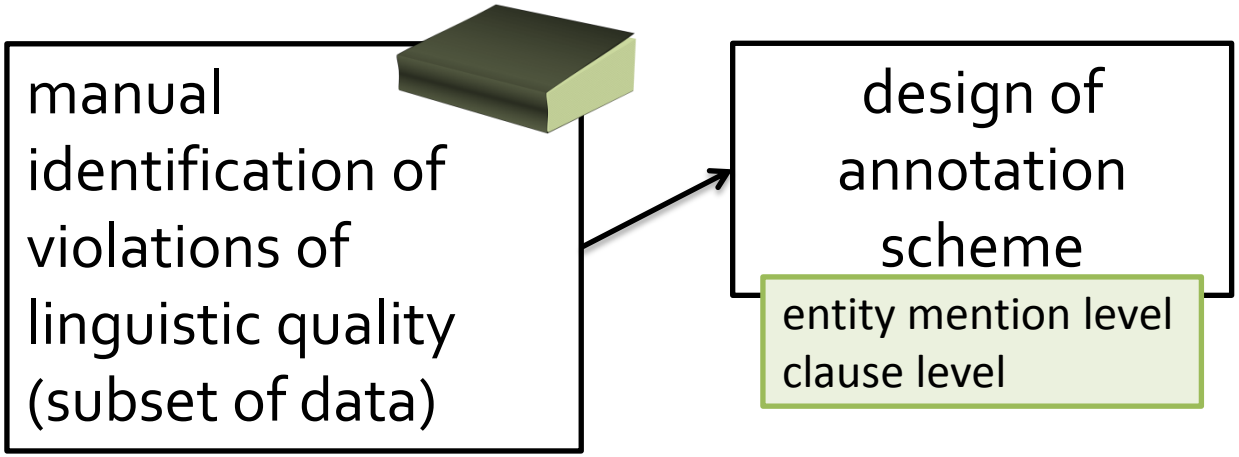
Revision-based approach

[Mani et al. 1999, Jing & McKeown 2000, Otterbacher et al. 2002]

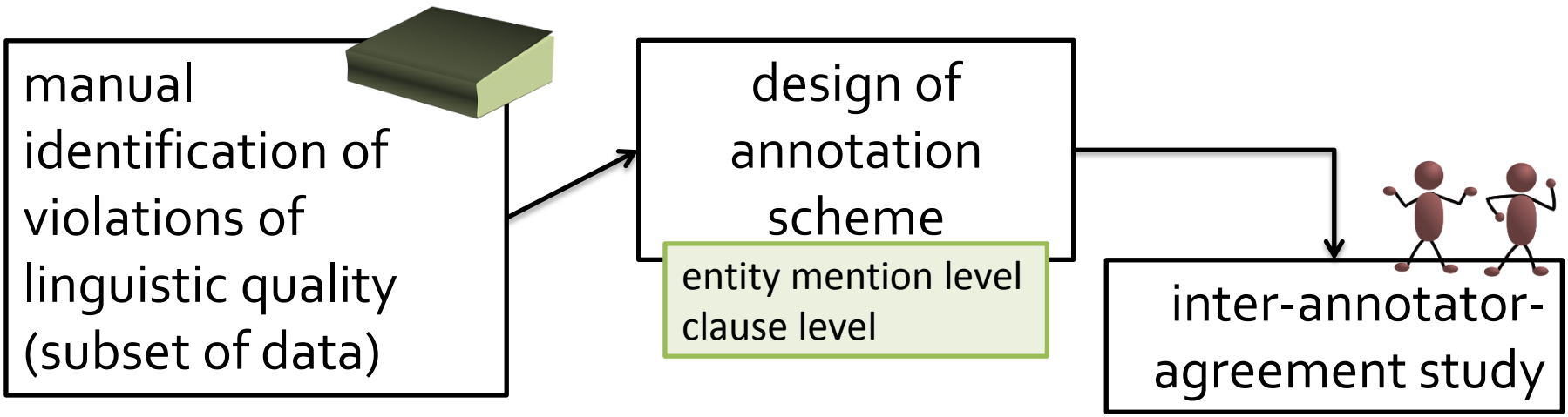
LQVSumm corpus

manual identification of violations of linguistic quality (subset of data)

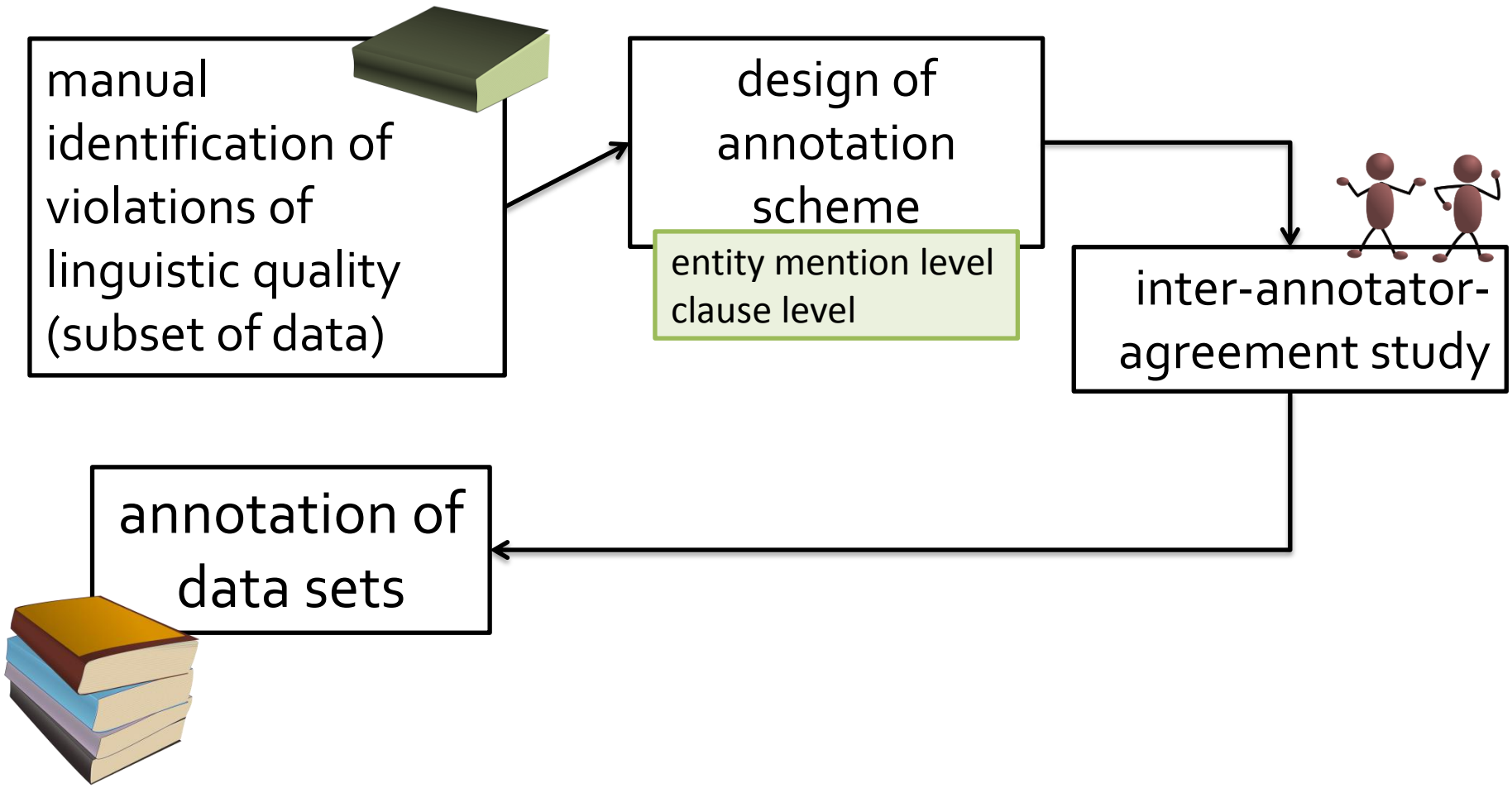
LQVSumm corpus



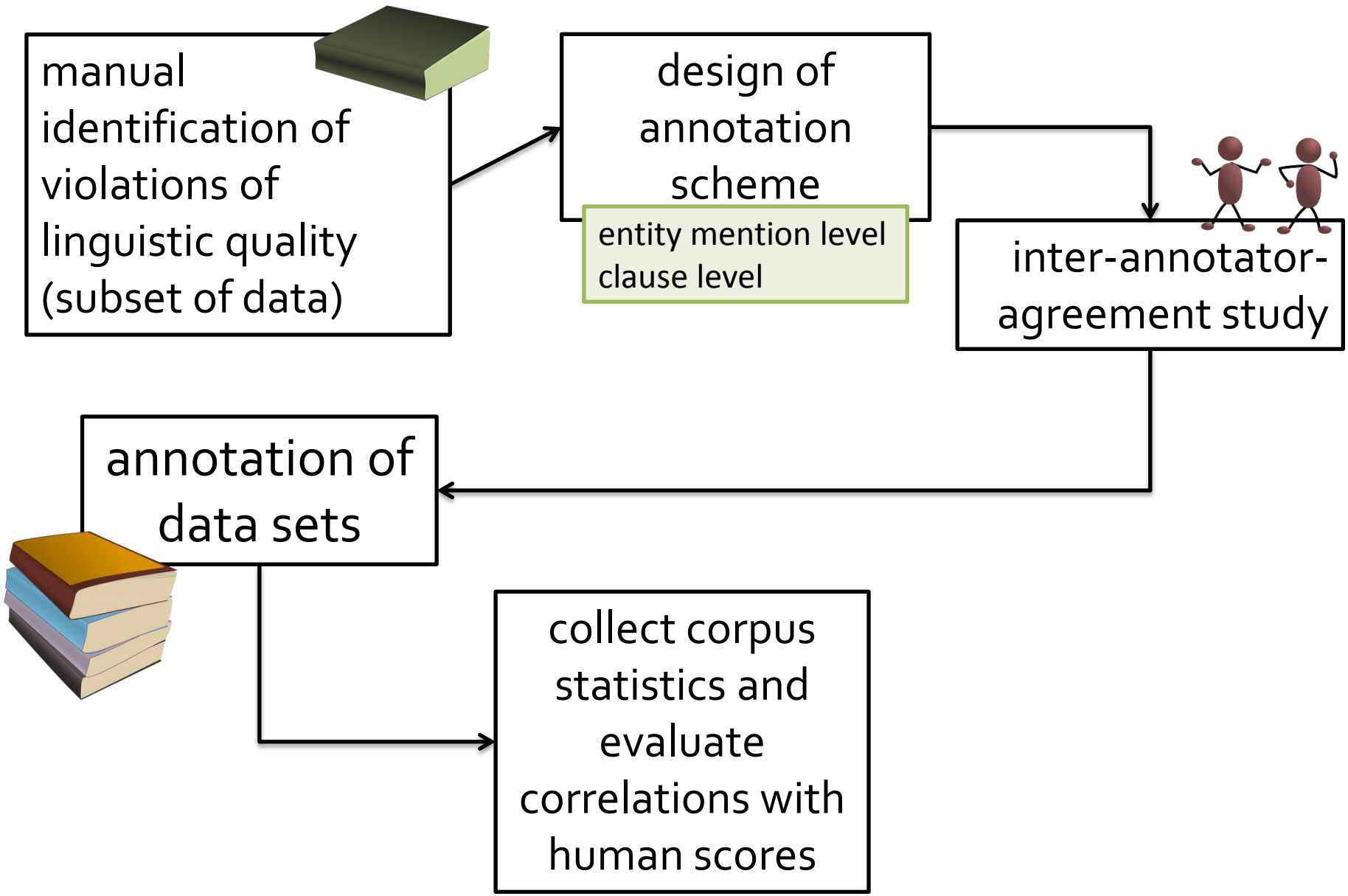
LQVSumm corpus



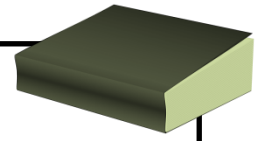
LQVSumm corpus



LQVSumm corpus



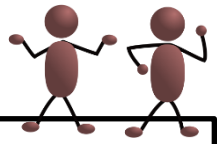
LQVSumm corpus



manual identification of violations of linguistic quality (subset of data)

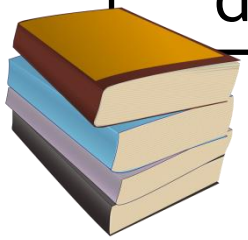
design of annotation scheme

- entity mention level
- clause level



inter-annotator-agreement study

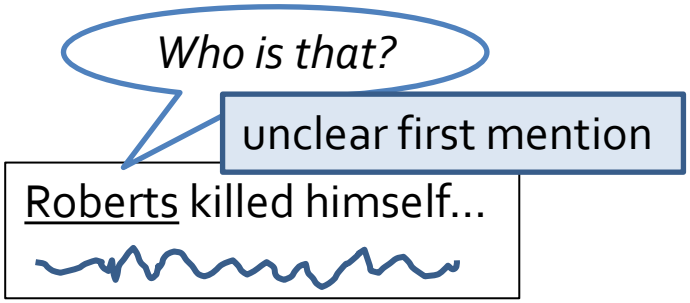
annotation of data sets



collect corpus statistics and evaluate correlations with human scores

FUTURE WORK: modeling: detection of violation types, evaluation tool

Annotation Scheme: **Entity Mention level**



Annotation Scheme: **Entity Mention level**

Who is that?

unclear first mention

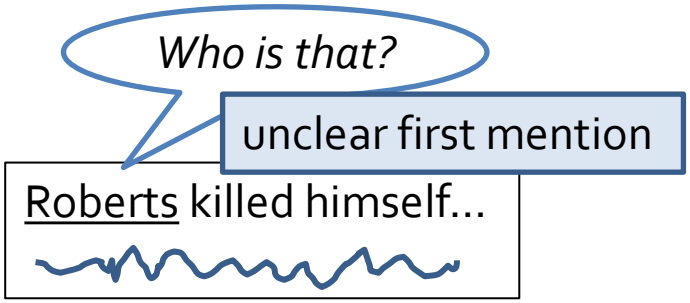
Roberts killed himself…

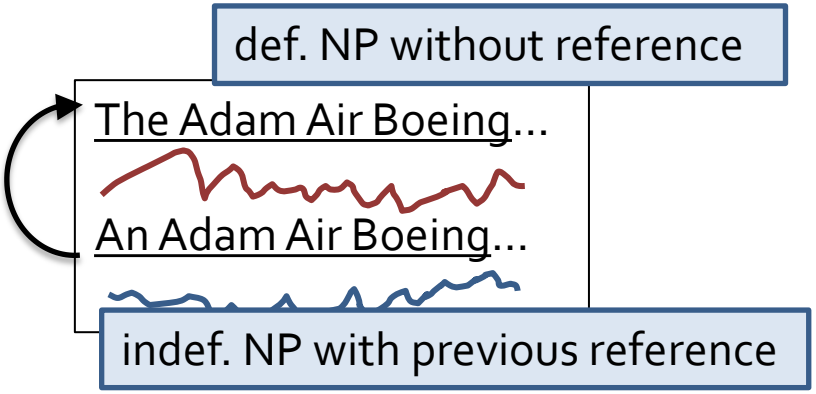
Taylor's attorney …

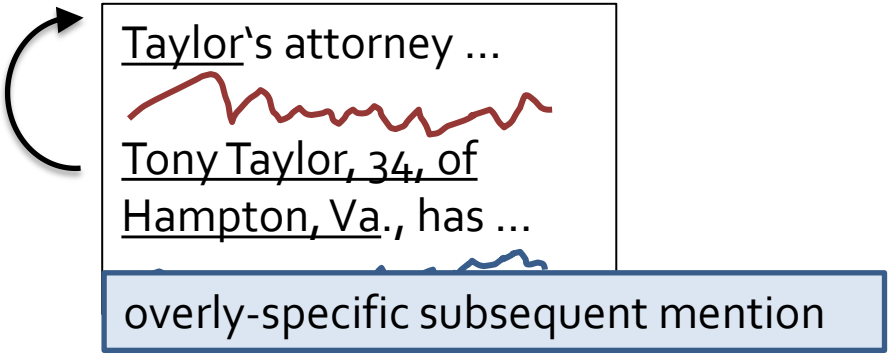
Tony Taylor, 34, of Hampton, Va., has …

overly-specific subsequent mention

Annotation Scheme: **Entity Mention level**







Annotation Scheme: **Entity Mention level**

Who is that?

unclear first mention

Roberts killed himself…

def. NP without reference

The Adam Air Boeing…

An Adam Air Boeing…

indef. NP with previous reference

Taylor's attorney …

Tony Taylor, 34, of Hampton, Va., has …

overly-specific subsequent mention

pronouns without antecedents

pronouns with misleading antecedents

unclear acronyms

Annotation Scheme: **Clause level**

(sentence, phrase, sequence of tokens)

ungrammaticality

incomplete sentence

Annotation Scheme: **Clause level**

(sentence, phrase, sequence of tokens)

ungrammaticality

incomplete sentence

dateline included

GEORGETOWN, Pennsylvania
2006-10-05 16:53:53 UTC

Annotation Scheme: **Clause level**

(sentence, phrase, sequence of tokens)

ungrammaticality

incomplete sentence

dateline included

GEORGETOWN, Pennsylvania
2006-10-05 16:53:53 UTC

no semantic relatedness between clauses

It is popularly known as the pink city.
He said there was no justification for such killings.

Annotation Scheme: **Clause level**

(sentence, phrase, sequence of tokens)

ungrammaticality

incomplete sentence

dateline included

GEORGETOWN, Pennsylvania
2006-10-05 16:53:53 UTC

redundant information

He was acting out in revenge for something that happened 20 years ago....

...was apparently acting in revenge for an incident that happened to him 20 years ago.

no semantic relatedness between clauses

It is popularly known as the pink city.
He said there was no justification for such killings.

Annotation Scheme: **Clause level**

(sentence, phrase, sequence of tokens)

ungrammaticality

incomplete sentence

dateline included

GEORGETOWN, Pennsylvania
2006-10-05 16:53:53 UTC

redundant information

He was acting out in revenge for something that happened 20 years ago....

...was apparently acting in revenge for an incident that happened to him 20 years ago.

no semantic relatedness between clauses

It is popularly known as the pink city.
He said there was no justification for such killings.

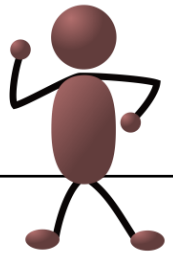
inappropriate use of discourse connective

LQVSumm: Annotated Data

	TAC
data source	1935 summaries, TAC 2011 (initial summaries), generated by 44 different extractive summarization systems
input to systems	sets of 10 news articles
Output	100-word summaries
summarization approaches	sentence selection + compression

LQVSumm: Annotated Data

	TAC
data source	1935 summaries, TAC 2011 (initial summaries), generated by 44 different <u>extractive</u> summarization systems
input to systems	sets of 10 news articles
Output	100-word summaries
summarization approaches	sentence selection + compression
manual scores for summaries	Readability (1-5), Pyramid (content), Responsiveness (1-5)



Inter-annotator agreement

- 100 randomly chosen summaries
- two annotators (A) and (B)
- annotations match if same type & overlapping span

Inter-annotator agreement

- 100 randomly chosen summaries
- two annotators (A) and (B)
- annotations match if same type & overlapping span

level	**Precision(B:A)**	**Recall(B:A)**	**F1**
entity mention	90.4	54.5	67.5
clause	84.1	83.3	83.6

Inter-annotator agreement

- 100 randomly chosen summaries
- two annotators (A) and (B)
- annotations match if same type & overlapping span

level	Precision(B:A)	Recall(B:A)	F1
entity mention	90.4	54.5	67.5
clause	84.1	83.3	83.6

A creates twice as many annotations,
B's annotations are a subset of A's

Inter-annotator agreement

- 100 randomly chosen summaries
- two annotators (A) and (B)
- annotations match if same type & overlapping span

level	Precision(B:A)	Recall(B:A)	F1
entity mention	90.4	54.5	67.5
clause	84.1	83.3	83.6

Agreement higher on clause level than on entity mention level

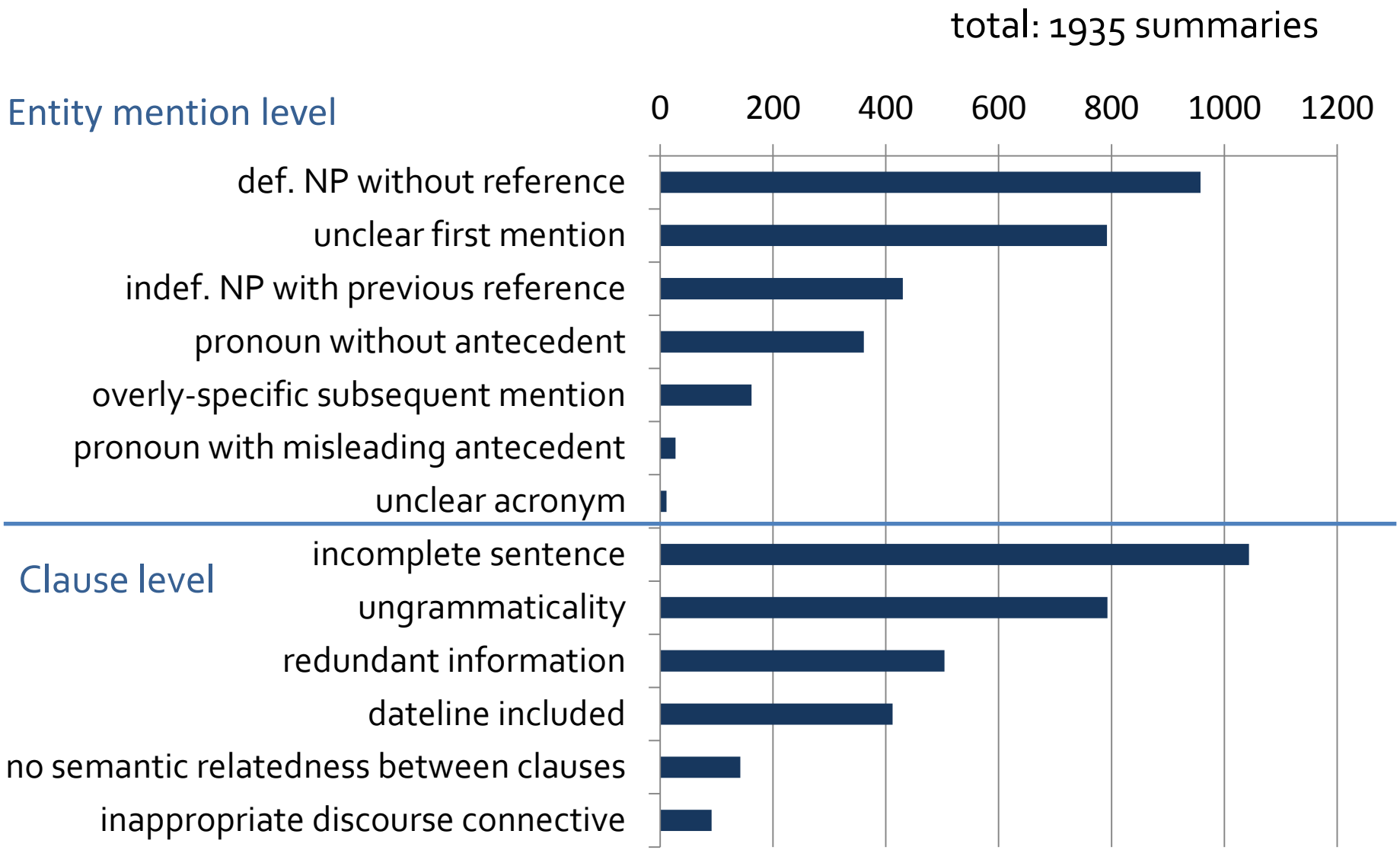
Inter-annotator agreement

- 100 randomly chosen summaries
- two annotators (A) and (B)
- annotations match if same type & overlapping span

level	**Precision(B:A)**	**Recall(B:A)**	**F1**
entity mention	90.4	54.5	67.5
clause	84.1	83.3	83.6

→ degree of subjectivity is manageable

Absolute Frequencies of LQVs by type



Ranking systems: average number of violations per summary

- compare rankings with TAC 2011 rankings
- draw conclusions about strengths/weaknesses of systems

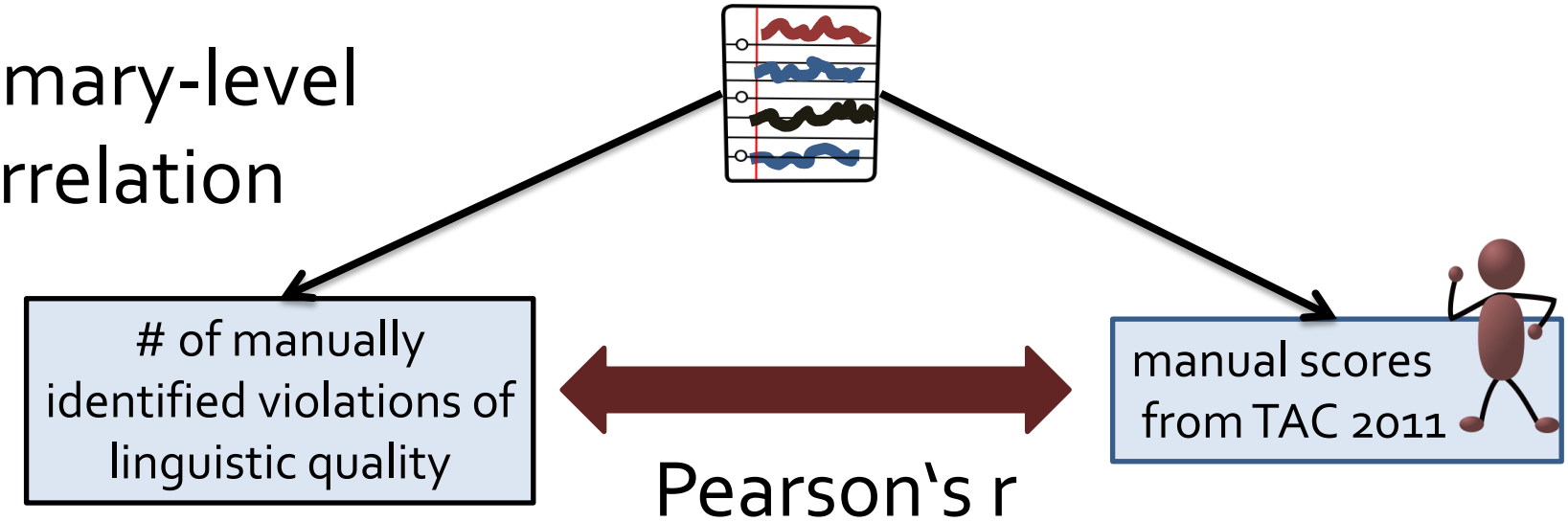
System	Entity mention level	Clause level	All LQV types
1 (baseline using first 100 words as summary)	0.34	1	1.34
21	0.84	0.45	1.3
…	…	…	…
7	1.14	4.63	5.77

Ranking systems: average number of violations per summary

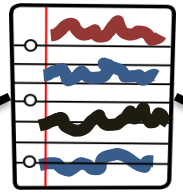
- compare rankings with TAC 2011 rankings
- draw conclusions about strengths/weaknesses of systems

System	Entity mention level	Clause level	All LQV types
1 (baseline using first 100 words as summary)	0.34	1	1.34
21	0.84	0.45	1.3
…	…	…	…
7	1.14	4.63	5.77
Best TAC system (differs for each column, TAC 2011)	(System 1) 0.34	(System 16) 0.23	(System 21) 1.30
Average of systems in TAC	1.42	1.54	2.96

Summary-level correlation



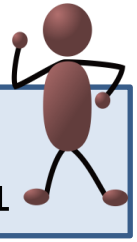
Summary-level correlation

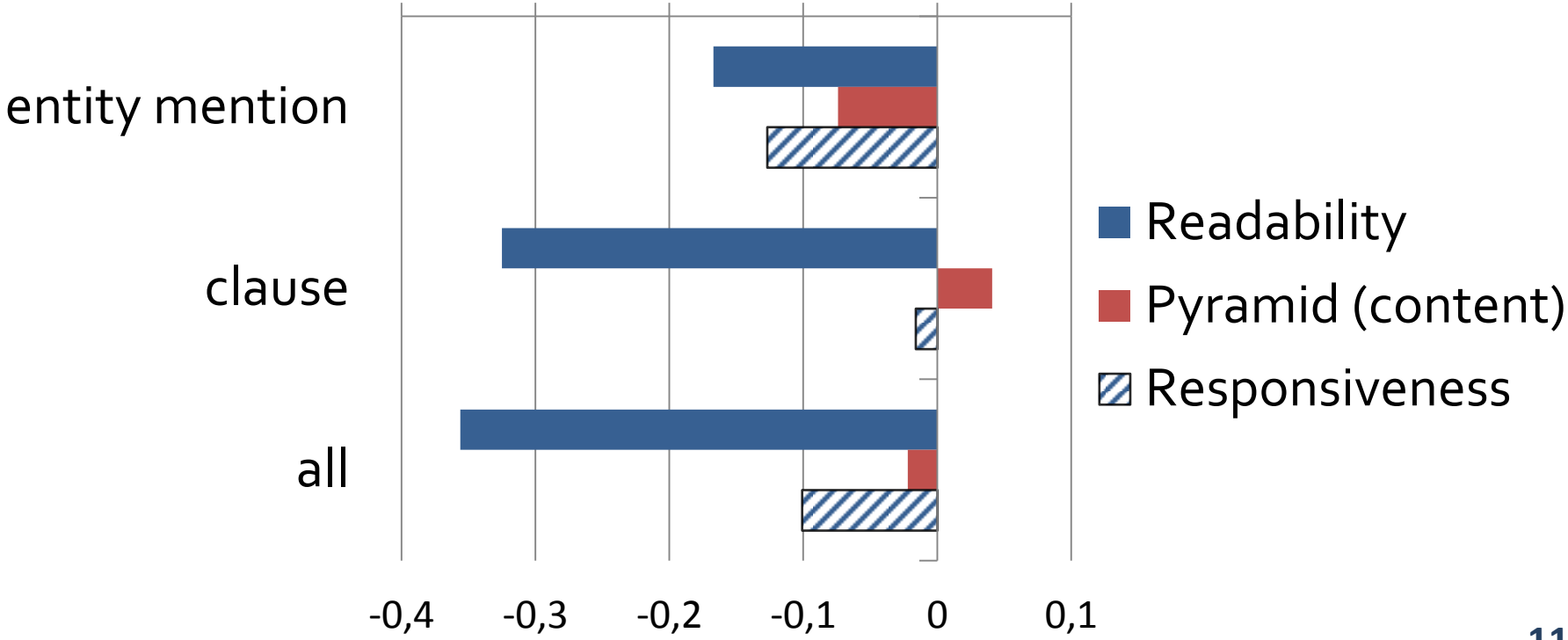


# of manually identified violations of linguistic quality

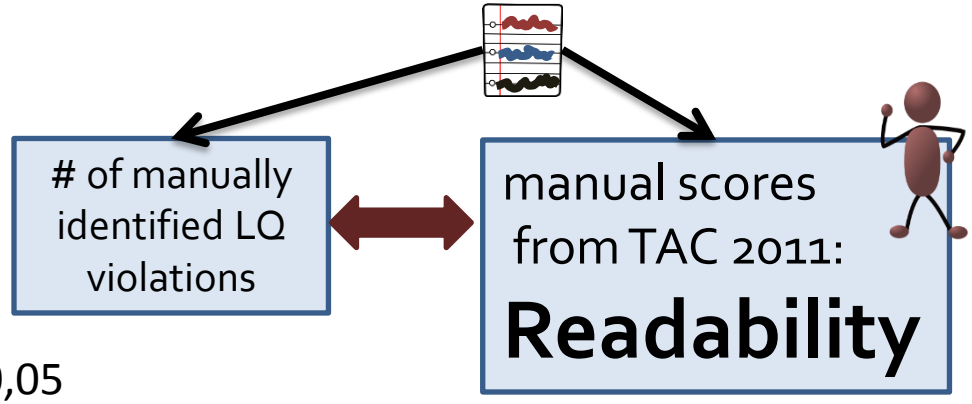
Pearson's r

manual scores from TAC 2011

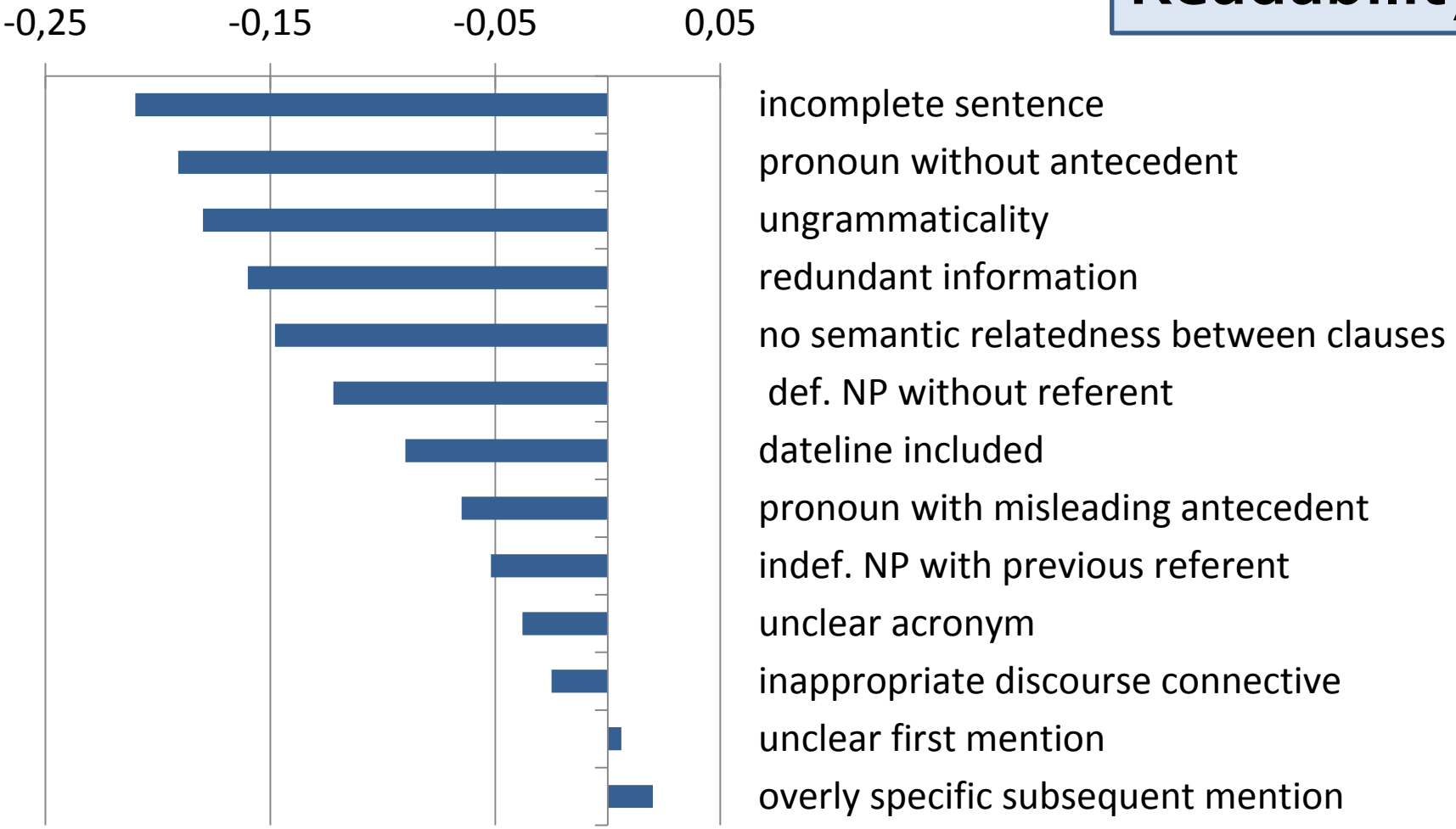




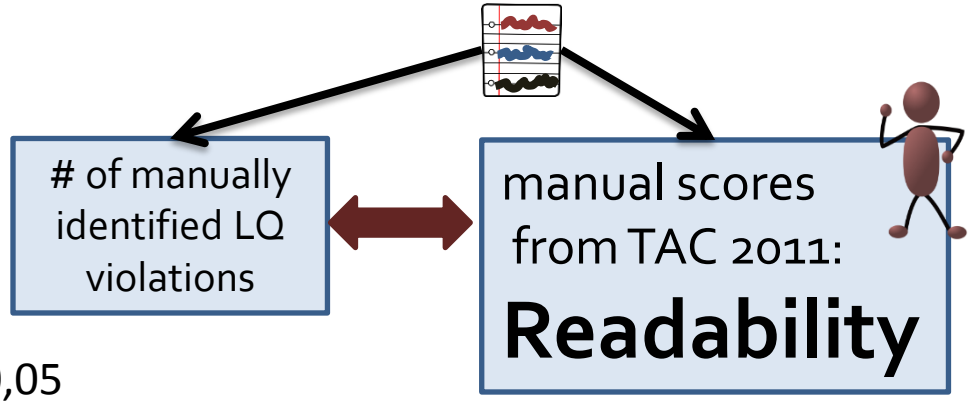
Summary-level correlation



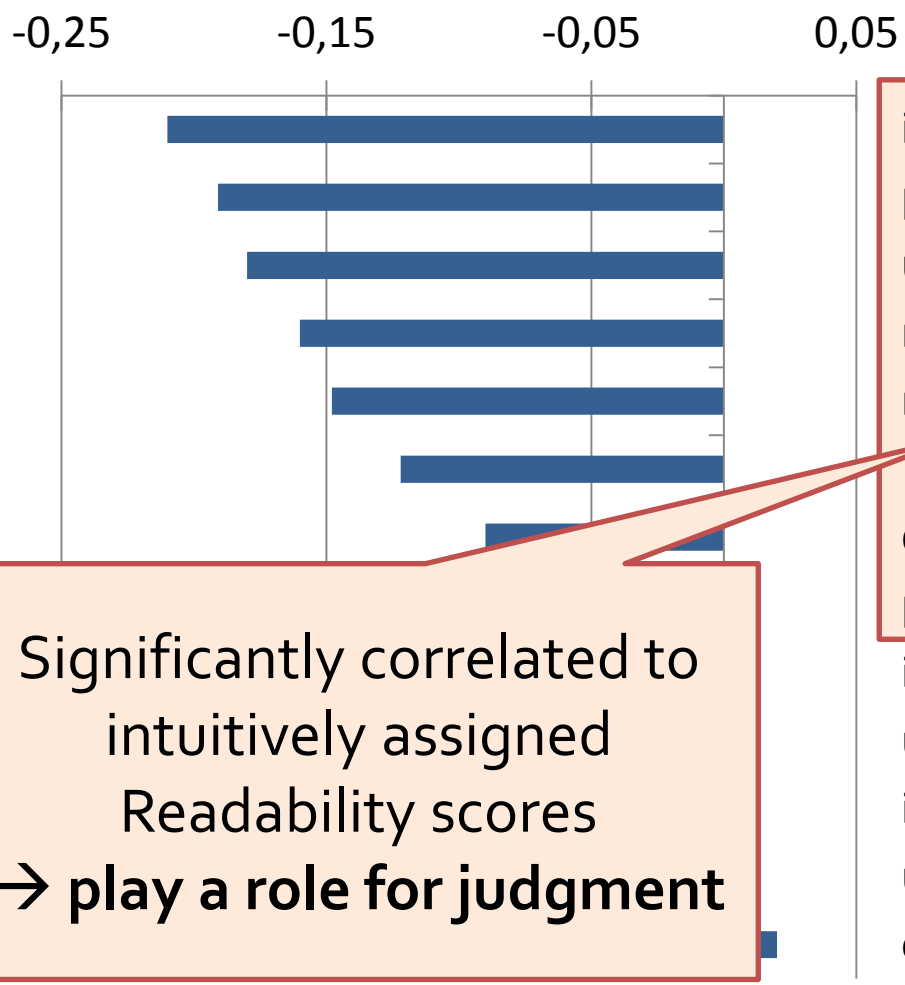
Pearsons's r



Summary-level correlation



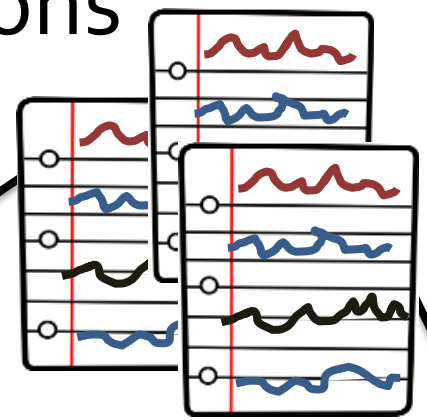
Pearsons's r



incomplete sentence
pronoun without antecedent
ungrammaticality
redundant information
no semantic relatedness between clauses
def. NP without referent
dateline included
pronoun with misleading antecedent
indef. NP with previous referent
unclear acronym
inappropriate discourse connective
unclear first mention
overly specific subsequent mention

Significantly correlated to intuitively assigned Readability scores
→ **play a role for judgment**

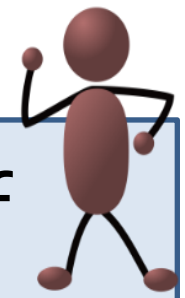
System-level correlations



All summaries created by one system

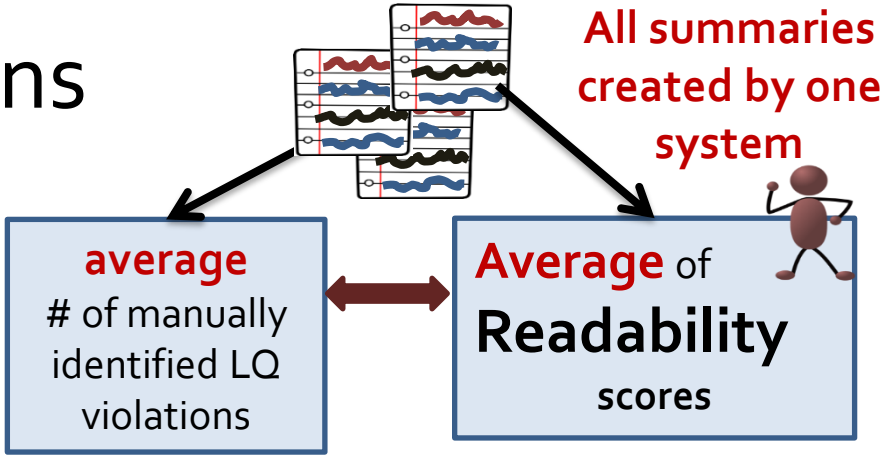
average
# of manually identified LQ violations

Average of **Readability scores**



System 21	1.30	3.75
System 2	1.74	3.34
System 7	5.77	2.09
…	…	…

System-level correlations

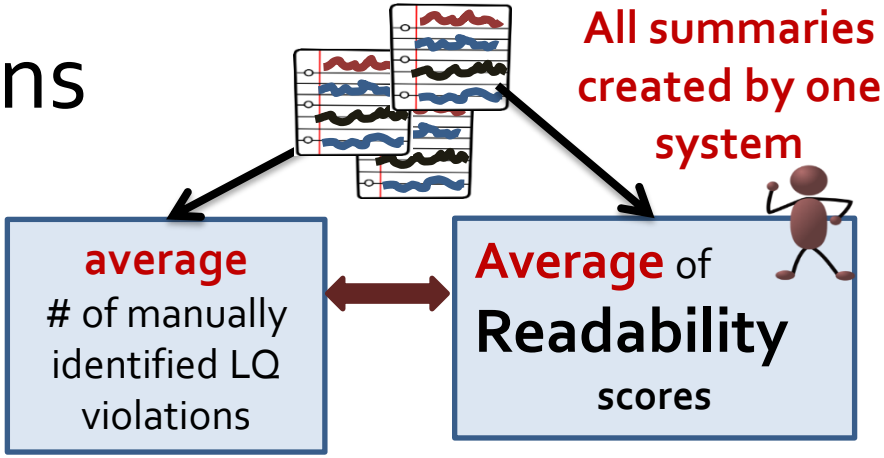


DICOMER: features from Penn Discourse TreeBank-style discourse parser

higher absolute correlation → better ranking

Method	***Ranking of***	**Pearson's r**	**Spearman's ρ**	**Kendall's τ**
DICOMER [Lin et al. 2012]	*all 50 systems*	**0.867**	0.712	0.535
LQVSumm $sum(violations)$	*44 systems*	-0.820	**-0.858**	**-0.713**

System-level correlations



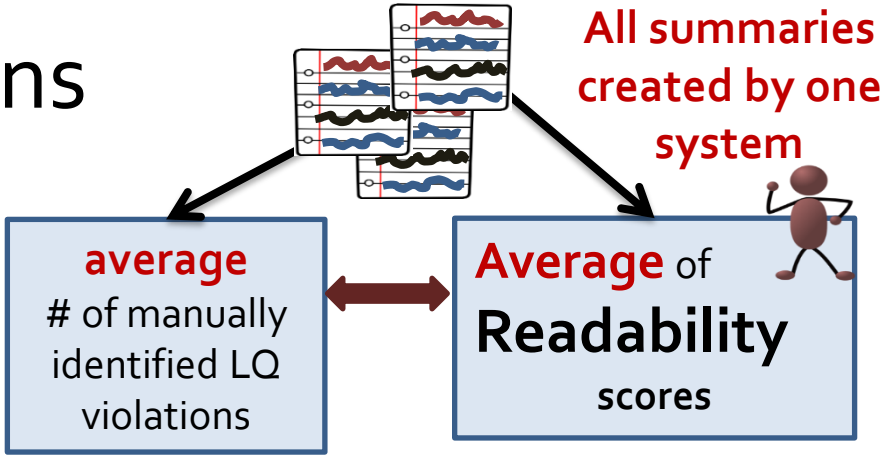
DICOMER: features from Penn Discourse TreeBank-style discourse parser

higher absolute correlation → better ranking

Method	*Ranking of*	Pearson's r	Spearman's ρ	Kendall's τ
DICOMER [Lin et al. 2012]	*all 50 systems*	**0.867**	0.712	0.535
LQVSumm $sum(violations)$	*44 systems*	-0.820	**-0.858**	**-0.713**

Pearson's r	Spearman's ρ, Kendall's τ
actual scores	ranking only

System-level correlations



DICOMER: features from Penn Discourse TreeBank-style discourse parser

higher absolute correlation → better ranking

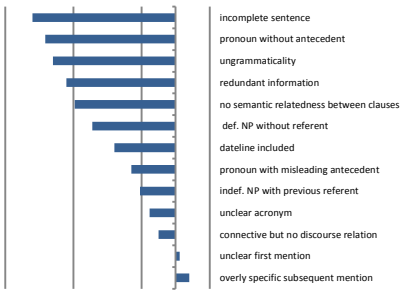
Method	*Ranking of*	Pearson's r	Spearman's ρ	Kendall's τ
DICOMER [Lin et al. 2012]	*all 50 systems*	**0.867**	0.712	0.535
LQVSumm $sum(violations)$	*44 systems*	-0.820	**-0.858**	**-0.713**

Pearson's r	Spearman's ρ, Kendall's τ
actual scores	ranking only
DICOMER is better (trained on TAC 2009 & TAC 2010)	**counting the number of violations works better than a supervised system.**

Conclusions

LQVSumm:
≈ 2000 summaries
marked with LQV types

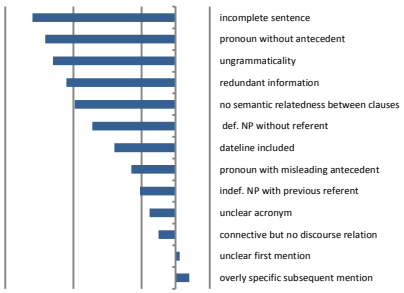
Conclusions



most types correlated to human judgments; others are infrequent

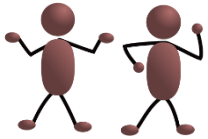
LQVSumm:
≈ 2000 summaries
marked with LQV types

Conclusions



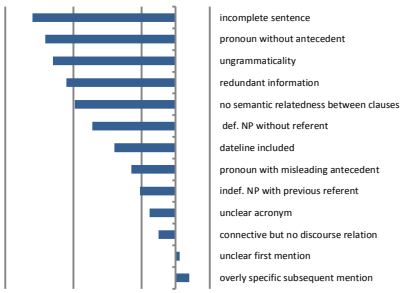
most types correlated to human judgments; others are infrequent

LQVSumm:
≈ 2000 summaries
marked with LQV types



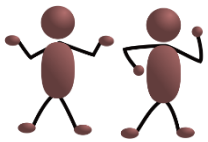
good inter-annotator agreement

Conclusions



most types correlated to human judgments; others are infrequent

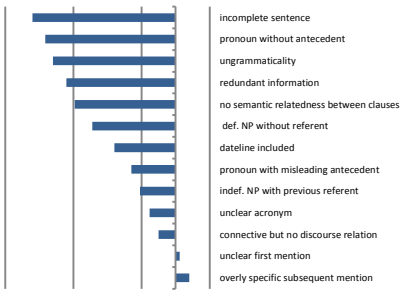
LQVSumm:
≈ 2000 summaries
marked with LQV types



good inter-annotator agreement

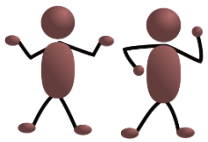
counts and marked instances of linguistic quality violations allow for:

Conclusions



most types correlated to human judgments; others are infrequent

LQVSumm:
≈ 2000 summaries
marked with LQV types

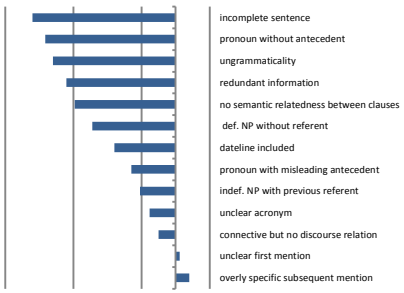


good inter-annotator agreement

counts and marked instances of linguistic quality violations allow for:

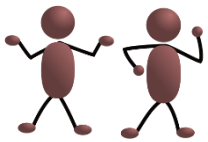
analyzing <u>what a particular system is good/bad at</u> (rather than just obtaining a numeric score)

Conclusions



most types correlated to human judgments; others are infrequent

LQVSumm:
≈ 2000 summaries
marked with LQV types



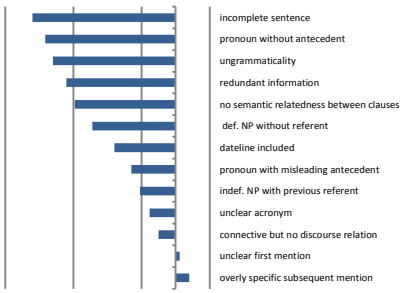
good inter-annotator agreement

counts and marked instances of linguistic quality violations allow for:

analyzing <u>what a particular system is good/bad at</u> (rather than just obtaining a numeric score)

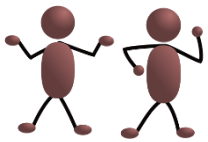
developing automatic methods to detect LQVs (future work)

Conclusions



most types correlated to human judgments; others are infrequent

LQVSumm:
≈ 2000 summaries
marked with LQV types



good inter-annotator agreement

counts and marked instances of linguistic quality violations allow for:

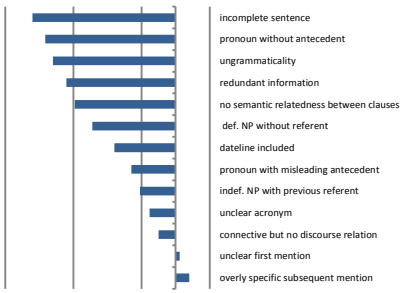
analyzing <u>what a particular system is good/bad at</u> (rather than just obtaining a numeric score)

developing automatic methods to detect LQVs (future work)

Available in stand-off format at: `www.coli.uni-saarland.de/~afried`

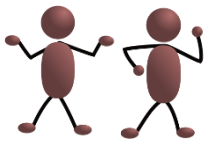
Conclusions

Thanks!



most types correlated to human judgments; others are infrequent

LQVSumm:
≈ 2000 summaries
marked with LQV types



good inter-annotator agreement

counts and marked instances of linguistic quality violations allow for:

analyzing <u>what a particular system is good/bad at</u> (rather than just obtaining a numeric score)

developing automatic methods to detect LQVs (future work)

Available in stand-off format at: `www.coli.uni-saarland.de/~afried`

Backup Slides

Annotation Scheme: Overview

entity mention level

- pronouns without antecedents
- indefinite NPs with a previous mention
- …

clause level

(sentence, phrase, sequence of tokens)

- ungrammatical sentences
- no semantic relatedness
- …

Performance of the G-Flow summarization system

- G-Flow system: Christensen et al. (NAACL 2013): Towards Coherent Multi-Document Summarization
- system incorporates coherence information into sentence extraction
- marked 50 summaries provided on the web site of the authors

System	Entity mention level	Clause level	All LQV types
Best TAC system (differs for each column, TAC 2011)	(System 1) 0.34	(System 16) 0.23	(System 21) 1.30
G-Flow (DUC 2004 data)	**0.30**	**0.20**	**0.50**

G-Flow succeeds in producing more coherent / readable summaries

inappropriate use of discourse connective