A Proposal on Evaluation Measures for RTE

Richard Bergmair

University of Cambridge Computer Laboratory
Natural Language Information Processing

ACL/IJCNLP 2009 Workshop on Applied Textual Inference,
Aug-6 2009

UNIVERSITY OF
CAMBRIDGE

Problems with the Current Methodology

- distribution neither balanced nor representative;
  so accuracy figures biased.
- notion of confidence-ranking misleading;
  accuracy & thresholding contradicts average precision.
- ENTAILMENT/CONTRADICTION symmetric;
  average precision doesn't reflect that.

Problems with the Current Methodology

- distribution neither balanced nor representative;
  so accuracy figures biased.
- notion of confidence-ranking misleading;
  accuracy & thresholding contradicts average precision.
- ENTAILMENT/CONTRADICTION symmetric;
  average precision doesn't reflect that.

Problems with the Current Methodology

- distribution neither balanced nor representative;
  so accuracy figures biased.
- notion of confidence-ranking misleading;
  accuracy & thresholding contradicts average precision.
- ENTAILMENT/CONTRADICTION symmetric;
  average precision doesn't reflect that.

Proposal for a New Methodology

- best: **mutual information**.
- average precision completely unsuitable!
  confidence-weighted score preferable, but there are still drawbacks;
- report baselines, be aware of bias with accuracies;
  perhaps use an artificially balanced subset for 3-way task.

Outline

Outline

RTE Data

N candidate entailments:

$$\mathrm{X} = \{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{N}\}.$$

gold standard:

$$\mathrm{G} : \mathrm{X} \mapsto \{\boxplus, \Diamond, \boxminus\}.$$

system output:

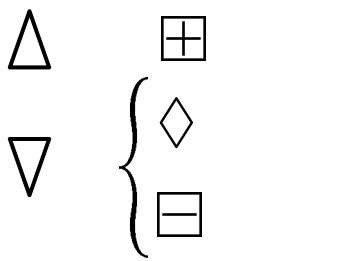
$$(\mathrm{L}, \succ) \text{ where}$$

$$\mathrm{L} : \mathrm{X} \mapsto \{\boxplus, \Diamond, \boxminus\},$$

strict total order $\succ$ on X.

RTE Data

class labels:



Logical Structure of Candidate Entailments

$$(x_{42}) \quad \frac{\begin{array}{ll} & \text{Socrates is a man and every man is mortal.} \quad (\varphi) \\ \therefore & \text{Socrates is mortal.} \quad (\psi) \end{array}}{}$$

Logical Structure of Candidate Entailments

$$(x_{42}) \quad \begin{array}{ll} & \text{Socrates is a man and every man is mortal.} \quad (\varphi) \\ \hline \therefore & \text{Socrates is mortal.} \quad (\psi) \end{array}$$

$$\Box(\varphi \rightarrow \psi)$$

$$G(x_{42}) = \boxplus \qquad \neg G(\neg x_{42}) = \boxminus$$

Logical Structure of Candidate Entailments

$$(x_{42}) \quad \begin{array}{ll} & \text{Socrates is a man and every man is mortal.} \quad (\varphi) \\ \hline \therefore & \text{Socrates is mortal.} \quad (\psi) \end{array}$$

$$\Box(\varphi \rightarrow \psi)$$

$$G(x_{42}) = \boxplus \qquad \neg G(\neg x_{42}) = \boxminus$$

Logical Structure of Candidate Entailments

$(\neg x_{42})$ Socrates is a man and every man is mortal. (φ)

$\therefore$ Socrates is **not** mortal. $(\neg\psi)$

$$\Box(\varphi \rightarrow \neg\psi)$$

$$G(x_{42}) = \boxplus \qquad \neg G(\neg x_{42}) = \boxminus$$

Logical Structure of Candidate Entailments

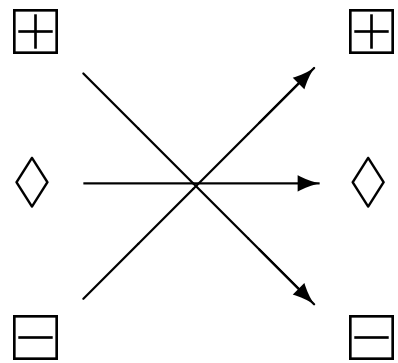
(x_{43}) Socrates is a man and every man is mortal. (φ)

$\therefore$ Socrates is (not) wise. $(\psi', \neg\psi')$

$$\Diamond(\varphi \to \psi) \qquad \Diamond(\varphi \to \neg\psi)$$

$$G(x_{43}) = \Diamond \qquad \neg G(\neg x_{42}) = \Diamond$$

Logical Structure of Candidate Entailments



Outline

Accuracy

$$\mathbb{A}_3(L;G) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\Big([L(x_i)]_3 = [G(x_i)]_3\Big),$$

$$\mathbb{A}_2(L;G) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\Big([L(x_i)]_2 = [G(x_i)]_2\Big),$$

$$\mathbb{A}_3(L;G) = \mathbb{A}_3(\neg L;\neg G)$$
$$\mathbb{A}_2(L;G) = \mathbb{A}_2(\neg L;\neg G)$$

Accuracy

$$\mathbb{A}_3(L;G) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\Big([L(x_i)]_3 = [G(x_i)]_3\Big),$$

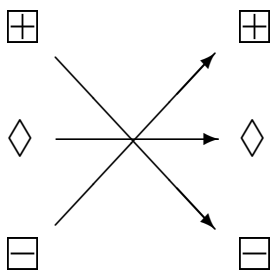
$$\mathbb{A}_2(L;G) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\Big([L(x_i)]_2 = [G(x_i)]_2\Big),$$

$$\mathbb{A}_3(L;G) = \mathbb{A}_3(\neg L;\neg G)$$
$$\mathbb{A}_2(L;G) = \mathbb{A}_2(\neg L;\neg G)$$

Accuracy & Logical Symmetry

$$\mathbb{A}_3\big(\mathrm{L};\mathrm{G}\big) = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \mathbb{1}\Big([\mathrm{L}(\mathrm{x}_i)]_3 = [\mathrm{G}(\mathrm{x}_i)]_3\Big),$$

$$\mathbb{A}_2\big(\mathrm{L};\mathrm{G}\big) = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \mathbb{1}\Big([\mathrm{L}(\mathrm{x}_i)]_2 = [\mathrm{G}(\mathrm{x}_i)]_2\Big),$$



$$\mathbb{A}_3\big(\mathrm{L};\mathrm{G}\big) = \mathbb{A}_3\big(\neg\mathrm{L};\neg\mathrm{G}\big)$$
$$\mathbb{A}_2\big(\mathrm{L};\mathrm{G}\big) = \mathbb{A}_2\big(\neg\mathrm{L};\neg\mathrm{G}\big)$$

Conditioned Accuracy

$$\mathbb{A}'_3(\mathrm{L};\mathrm{G},g) = \frac{\sum_{i=1}^{\mathrm{N}} \mathbb{1}\left([\mathrm{L}(\mathrm{x}_i)]_3 = [\mathrm{G}(\mathrm{x}_i)]_3 = g\right)}{\sum_{i=1}^{\mathrm{N}} \mathbb{1}\left([\mathrm{G}(\mathrm{x}_i)]_3 = g\right)},$$

$$\mathbb{A}'_2(\mathrm{L};\mathrm{G},g) = \frac{\sum_{i=1}^{\mathrm{N}} \mathbb{1}\left([\mathrm{L}(\mathrm{x}_i)]_2 = [\mathrm{G}(\mathrm{x}_i)]_2 = g\right)}{\sum_{i=1}^{\mathrm{N}} \mathbb{1}\left([\mathrm{G}(\mathrm{x}_i)]_2 = g\right)}.$$

$$\mathbb{A}'_2(\mathrm{L};\mathrm{G},\bigtriangleup) \rightsquigarrow \text{recall},$$

$$\mathbb{A}'_2(\mathrm{G};\mathrm{L},\bigtriangleup) \rightsquigarrow \text{precision}.$$

Bias

labels:	RTE-4	RTE-3 PILOT
ENTAILMENT ($\boxplus$)	50%	51%
UNKNOWN ($\lozenge$)	35%	40%
CONTRADICTION ($\boxminus$)	15%	9%

$\mathbb{A}_3(\mathrm{L}^{\boxplus};\mathrm{G}) = .500$ outperforms **1/3** of all RTE4 participants and **2/3** of all RTE3 PILOT participants!

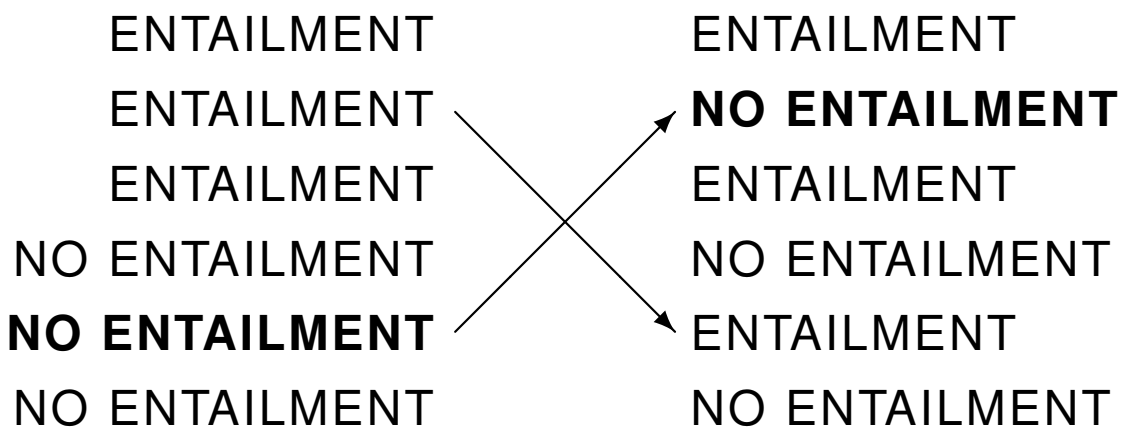
$\mathbb{A}_3(\mathrm{L}^{*};\mathrm{G}) = .394, \quad \mathbb{A}_3(\mathrm{L}^{\lozenge};\mathrm{G}) = .350, \quad \mathbb{A}_3(\mathrm{L}^{\boxminus};\mathrm{G}) = .150$

Outline

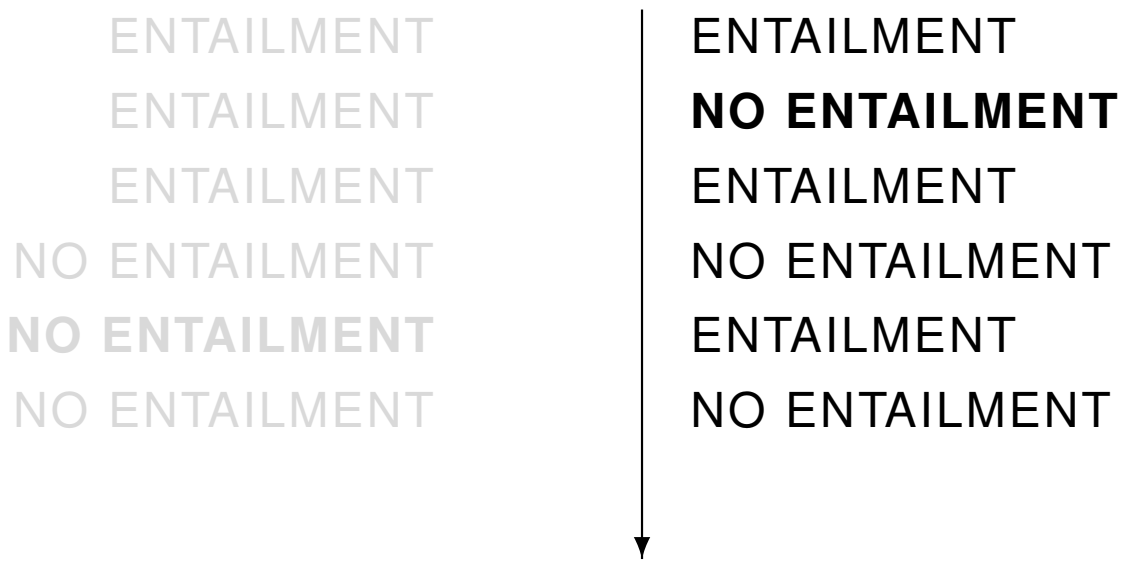
Average Precision

ENTAILMENT	ENTAILMENT
ENTAILMENT	NO ENTAILMENT
ENTAILMENT	ENTAILMENT
NO ENTAILMENT	NO ENTAILMENT
NO ENTAILMENT	ENTAILMENT
NO ENTAILMENT	NO ENTAILMENT

Average Precision



Confidence-Weighted Score



Average Precision vs. Confidence

- 2/3 of all RTE-4 participants who submitted confidence-ranked three-way labellings submitted confidence rankings, instead of $\mathbb{AP}$-style rankings.
- Stanford1: 44% $\rightarrow$ 62% after ranking down negative instances.
- confusing terminology!
- more generally: accuracy & average precision have contradictory preferences for rankings.

Average Precision vs. Confidence

- 2/3 of all RTE-4 participants who submitted confidence-ranked three-way labellings submitted confidence rankings, instead of $\mathbb{AP}$-style rankings.
- Stanford1: 44% $\rightarrow$ 62% after ranking down negative instances.
- confusing terminology!
- more generally: accuracy & average precision have contradictory preferences for rankings.

Average Precision vs. Confidence

- 2/3 of all RTE-4 participants who submitted confidence-ranked three-way labellings submitted confidence rankings, instead of $\mathbb{AP}$-style rankings.
- Stanford1: 44% $\rightarrow$ 62% after ranking down negative instances.
- confusing terminology!
- more generally: accuracy & average precision have contradictory preferences for rankings.

Average Precision vs. Confidence

- 2/3 of all RTE-4 participants who submitted confidence-ranked three-way labellings submitted confidence rankings, instead of $\mathbb{AP}$-style rankings.
- Stanford1: 44% $\rightarrow$ 62% after ranking down negative instances.
- confusing terminology!
- more generally: accuracy & average precision have contradictory preferences for rankings.

Average Precision & Logical Symmetry

id		system	gold
	1	ENTAILMENT	UNKNOWN
223	1	ENTAILMENT	UNKNOWN
	2	ENTAILMENT	ENTAILMENT
4	2	ENTAILMENT	CONTRADICTION
	3	UNKNOWN	UNKNOWN
313	3	UNKNOWN	UNKNOWN
	4	CONTRADICTION	CONTRADICTION
534	4	CONTRADICTION	ENTAILMENT
	5	CONTRADICTION	ENTAILMENT
415	5	CONTRADICTION	CONTRADICTION

Average Precision & Logical Symmetry

id		system	gold
	1	ENTAILMENT	UNKNOWN
223	1	~~ENTAILMENT~~	~~UNKNOWN~~
	2	ENTAILMENT	ENTAILMENT
4	2	~~ENTAILMENT~~	~~CONTRADICTION~~
	3	UNKNOWN	UNKNOWN
313	3	~~UNKNOWN~~	~~UNKNOWN~~
	4	CONTRADICTION	CONTRADICTION
534	4	~~CONTRADICTION~~	~~ENTAILMENT~~
	5	CONTRADICTION	ENTAILMENT
415	5	~~CONTRADICTION~~	~~CONTRADICTION~~

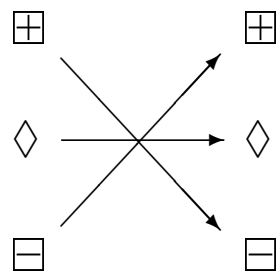
Average Precision & Logical Symmetry

id		system	gold
	1	CONTRADICTION	UNKNOWN
223	1	~~ENTAILMENT~~	~~UNKNOWN~~
	2	CONTRADICTION	ENTAILMENT
4	2	~~ENTAILMENT~~	~~CONTRADICTION~~
	3	UNKNOWN	UNKNOWN
313	3	~~UNKNOWN~~	~~UNKNOWN~~
	4	ENTAILMENT	CONTRADICTION
534	4	~~CONTRADICTION~~	~~ENTAILMENT~~
	5	ENTAILMENT	ENTAILMENT
415	5	~~CONTRADICTION~~	~~CONTRADICTION~~

Average Precision & Logical Symmetry

id		system	gold
	5	CONTRADICTION	UNKNOWN
223	1	~~ENTAILMENT~~	~~UNKNOWN~~
	4	CONTRADICTION	ENTAILMENT
4	2	~~ENTAILMENT~~	~~CONTRADICTION~~
	3	UNKNOWN	UNKNOWN
313	3	~~UNKNOWN~~	~~UNKNOWN~~
	2	ENTAILMENT	CONTRADICTION
534	4	~~CONTRADICTION~~	~~ENTAILMENT~~
	1	ENTAILMENT	ENTAILMENT
415	5	~~CONTRADICTION~~	~~CONTRADICTION~~

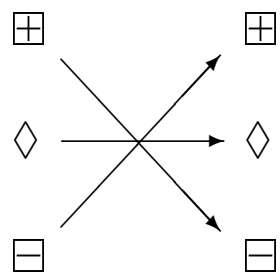
Average Precision & Logical Symmetry



$$\mathbb{A}_3(\mathrm{L};\mathrm{G}) = \mathbb{A}_3(\neg\mathrm{L};\neg\mathrm{G})$$
$$\mathbb{A}_2(\mathrm{L};\mathrm{G}) = \mathbb{A}_2(\neg\mathrm{L};\neg\mathrm{G})$$

$$\mathbb{AP}(\mathrm{G};\succ) \neq \mathbb{AP}(\neg\mathrm{G};\succ')$$

Average Precision & Logical Symmetry



$$\mathbb{A}_3\big(\mathrm{L};\mathrm{G}\big) = \mathbb{A}_3\big(\neg\mathrm{L};\neg\mathrm{G}\big)$$
$$\mathbb{A}_2\big(\mathrm{L};\mathrm{G}\big) = \mathbb{A}_2\big(\neg\mathrm{L};\neg\mathrm{G}\big)$$

$$\mathbb{AP}\big(\mathrm{G};\succ\big) \neq \mathbb{AP}\big(\neg\mathrm{G};\succ'\big)$$

Outline

Mutual Information: Defintion

marginals

20	25	5	$\mathbb{P}(\mathbf{G} = \boxplus) = .5$
9	18	9	$\mathbb{P}(\mathbf{G} = \Diamond) = .36$
1	7	6	$\mathbb{P}(\mathbf{G} = \boxminus) = .14$
$\mathbb{P}(\mathbf{L} = \boxplus) = .3$	$\mathbb{P}(\mathbf{L} = \Diamond) = .5$	$\mathbb{P}(\mathbf{L} = \boxminus) = .2$	$\mathbf{N} = 100$
$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxplus) = 1.0746$	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \Diamond) = 1.4277$	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxminus) = 1.5395$	

$$\mathbb{I}(\mathbf{G};\mathbf{L}) = \mathbb{H}(\mathbf{G}) - \mathbb{H}(\mathbf{G}|\mathbf{L})$$
$$= 1.4277 - 1.3441 = 0.0834$$

Mutual Information: Defintion

marginals

20	25	5	$\mathbb{P}(\mathbf{G} = \boxplus) = .5$
9	18	9	$\mathbb{P}(\mathbf{G} = \Diamond) = .36$
1	7	6	$\mathbb{P}(\mathbf{G} = \boxminus) = .14$
$\mathbb{P}(\mathbf{L} = \boxplus) = .3$	$\mathbb{P}(\mathbf{L} = \Diamond) = .5$	$\mathbb{P}(\mathbf{L} = \boxminus) = .2$	$\mathbf{N} = 100$
$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxplus) = 1.0746$	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \Diamond) = 1.4277$	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxminus) = 1.5395$	

$$\mathbb{I}(\mathbf{G};\mathbf{L}) = \mathbb{H}(\mathbf{G}) - \mathbb{H}(\mathbf{G}|\mathbf{L})$$
$$= 1.4277 - 1.3441 = 0.0834$$

Mutual Information: Defintion

marginals

20	25	5	$\mathbb{P}(\mathbf{G} = \boxplus) = .5$
9	18	9	$\mathbb{P}(\mathbf{G} = \Diamond) = .36$
1	7	6	$\mathbb{P}(\mathbf{G} = \boxminus) = .14$
$\mathbb{P}(\mathbf{L} = \boxplus) = .3$	$\mathbb{P}(\mathbf{L} = \Diamond) = .5$	$\mathbb{P}(\mathbf{L} = \boxminus) = .2$	$\mathbf{N} = 100$
$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxplus) = 1.0746$	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \Diamond) = 1.4277$	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxminus) = 1.5395$	

$$\mathbb{I}(\mathbf{G};\mathbf{L}) = \mathbb{H}(\mathbf{G}) - \mathbb{H}(\mathbf{G}|\mathbf{L})$$
$$= 1.4277 - 1.3441 = 0.0834$$

Mutual Information: Defintion

prior entropy

20	25	5	$\mathbb{P}(\mathbf{G} = \boxplus) = .5$
9	18	9	$\mathbb{P}(\mathbf{G} = \Diamond) = .36$
1	7	6	$\mathbb{P}(\mathbf{G} = \boxminus) = .14$
$\mathbb{P}(\mathbf{L} = \boxplus) = .3$	$\mathbb{P}(\mathbf{L} = \Diamond) = .5$	$\mathbb{P}(\mathbf{L} = \boxminus) = .2$	$\mathbb{H}(\mathbf{G}) = 1.4277$
$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxplus) = 1.0746$	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \Diamond) = 1.4277$	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxminus) = 1.5395$	$\mathbb{H}(\mathbf{G}\|\mathbf{L}) = 1.3441$

$$\mathbb{I}(\mathbf{G};\mathbf{L}) = \mathbb{H}(\mathbf{G}) - \mathbb{H}(\mathbf{G}|\mathbf{L})$$
$$= 1.4277 - 1.3441 = 0.0834$$

Mutual Information: Defintion

prior entropy

20	25	5	$\mathbb{P}(\mathbf{G} = \boxplus) = .5$
9	18	9	$\mathbb{P}(\mathbf{G} = \Diamond) = .36$
1	7	6	$\mathbb{P}(\mathbf{G} = \boxminus) = .14$
$\mathbb{P}(\mathbf{L} = \boxplus) = .3$	$\mathbb{P}(\mathbf{L} = \Diamond) = .5$	$\mathbb{P}(\mathbf{L} = \boxminus) = .2$	$\mathbb{H}(\mathbf{G}) = 1.4277$
$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxplus) = 1.0746$	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \Diamond) = 1.4277$	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxminus) = 1.5395$	$\mathbb{H}(\mathbf{G}\|\mathbf{L}) = 1.3441$

$$\mathbb{I}(\mathbf{G};\mathbf{L}) = \mathbb{H}(\mathbf{G}) - \mathbb{H}(\mathbf{G}|\mathbf{L})$$
$$= 1.4277 - 1.3441 = 0.0834$$

Mutual Information: Defintion

entropy after specific decisions

20	25	5	$\mathbb{P}(\mathbf{G} = \boxplus) = .5$
9	18	9	$\mathbb{P}(\mathbf{G} = \Diamond) = .36$
1	7	6	$\mathbb{P}(\mathbf{G} = \boxminus) = .14$
$\mathbb{P}(\mathbf{L} = \boxplus) = .3$	$\mathbb{P}(\mathbf{L} = \Diamond) = .5$	$\mathbb{P}(\mathbf{L} = \boxminus) = .2$	$\mathbb{H}(\mathbf{G}) = 1.4277$
$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxplus) = 1.0746$	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \Diamond) = 1.4277$	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxminus) = 1.5395$	$\mathbb{H}(\mathbf{G}\|\mathbf{L}) = 1.3441$

$$\mathbb{I}(\mathbf{G};\mathbf{L}) = \mathbb{H}(\mathbf{G}) - \mathbb{H}(\mathbf{G}|\mathbf{L})$$
$$= 1.4277 - 1.3441 = 0.0834$$

Mutual Information: Defintion

entropy after specific decisions

20	25	5	$\mathbb{P}(\mathbf{G} = \boxplus) = .5$
9	18	9	$\mathbb{P}(\mathbf{G} = \Diamond) = .36$
1	7	6	$\mathbb{P}(\mathbf{G} = \boxminus) = .14$
$\mathbb{P}(\mathbf{L} = \boxplus) = .3$	$\mathbb{P}(\mathbf{L} = \Diamond) = .5$	$\mathbb{P}(\mathbf{L} = \boxminus) = .2$	$\mathbb{H}(\mathbf{G}) = 1.4277$
$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxplus) = 1.0746$	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \Diamond) = 1.4277$	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxminus) = 1.5395$	$\mathbb{H}(\mathbf{G}\|\mathbf{L}) = 1.3441$

$$\mathbb{I}(\mathbf{G};\mathbf{L}) = \mathbb{H}(\mathbf{G}) - \mathbb{H}(\mathbf{G}|\mathbf{L})$$
$$= 1.4277 - 1.3441 = 0.0834$$

Mutual Information: Defintion

relative entropy

20	25	5	$\mathbb{P}(\mathbf{G} = \boxplus) = .5$
9	18	9	$\mathbb{P}(\mathbf{G} = \Diamond) = .36$
1	7	6	$\mathbb{P}(\mathbf{G} = \boxminus) = .14$
$\mathbb{P}(\mathbf{L} = \boxplus) = .3$	$\mathbb{P}(\mathbf{L} = \Diamond) = .5$	$\mathbb{P}(\mathbf{L} = \boxminus) = .2$	$\mathbb{H}(\mathbf{G}) = 1.4277$
$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxplus) = 1.0746$	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \Diamond) = 1.4277$	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxminus) = 1.5395$	$\mathbb{H}(\mathbf{G}\|\mathbf{L}) = 1.3441$

$$\mathbb{I}(\mathbf{G};\mathbf{L}) = \mathbb{H}(\mathbf{G}) - \mathbb{H}(\mathbf{G}|\mathbf{L})$$
$$= 1.4277 - 1.3441 = 0.0834$$

Mutual Information: Defintion

mutual information

20	25	5	$\mathbb{P}(\mathbf{G} = \boxplus) = .5$
9	18	9	$\mathbb{P}(\mathbf{G} = \Diamond) = .36$
1	7	6	$\mathbb{P}(\mathbf{G} = \boxminus) = .14$
$\mathbb{P}(\mathbf{L} = \boxplus) = .3$	$\mathbb{P}(\mathbf{L} = \Diamond) = .5$	$\mathbb{P}(\mathbf{L} = \boxminus) = .2$	$\mathbb{H}(\mathbf{G}) = 1.4277$
$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxplus) = 1.0746$	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \Diamond) = 1.4277$	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxminus) = 1.5395$	$\mathbb{H}(\mathbf{G}\|\mathbf{L}) = 1.3441$

$$\begin{aligned}\mathbb{I}(\mathbf{G};\mathbf{L}) &= \mathbb{H}(\mathbf{G}) - \mathbb{H}(\mathbf{G}|\mathbf{L}) \\ &= 1.4277 - 1.3441 = 0.0834\end{aligned}$$

Mutual Information: No Bias!

.

20 (20)	25 (25)	5 (5)	$\mathbb{P}(\mathbf{G} = \boxplus) = .5$
9 (9)	18 (18)	9 (9)	$\mathbb{P}(\mathbf{G} = \Diamond) = .36$
1 (1)	7 (7)	6 (6)	$\mathbb{P}(\mathbf{G} = \boxminus) = .14$
$\mathbb{P}(\mathbf{L} = \boxplus) = .3$ (.3)	$\mathbb{P}(\mathbf{L} = \Diamond) = .5$ (.5)	$\mathbb{P}(\mathbf{L} = \boxminus) = .2$ (.2)	$\mathbb{H}(\mathbf{G}) = 1.4277$
$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxplus) = 1.0746$ (1.0746)	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \Diamond) = 1.4277$ (1.4277)	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxminus) = 1.5395$ (1.5395)	$\mathbb{H}(\mathbf{G}\|\mathbf{L}) = 1.3441$ (1.3441)

$$\begin{aligned}\mathbb{I}(\mathbf{G};\mathbf{L}) &= \mathbb{H}(\mathbf{G}) - \mathbb{H}(\mathbf{G}|\mathbf{L}) \\ &= 1.4277 - 1.3441 = 0.0834\end{aligned}$$

Mutual Information: No Bias!

constant choice

50 (20)	0 (25)	0 (5)	$\mathbb{P}(\mathbf{G} = \boxplus) = .5$
36 (9)	0 (18)	0 (9)	$\mathbb{P}(\mathbf{G} = \Diamond) = .36$
14 (1)	0 (7)	0 (6)	$\mathbb{P}(\mathbf{G} = \boxminus) = .14$
$\mathbb{P}(\mathbf{L} = \boxplus) = 1$ (.3)	$\mathbb{P}(\mathbf{L} = \Diamond) = .0$ (.5)	$\mathbb{P}(\mathbf{L} = \boxminus) = .0$ (.2)	$\mathbb{H}(\mathbf{G}) = 1.4277$
$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxplus) = 1.4277$ (1.0746)	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \Diamond)$? (1.4277)	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxminus)$? (1.5395)	$\mathbb{H}(\mathbf{G}\|\mathbf{L}) = 1.4277$ (1.3441)

$$\begin{aligned}\mathbb{I}(\mathbf{G};\mathbf{L}) &= \mathbb{H}(\mathbf{G}) - \mathbb{H}(\mathbf{G}|\mathbf{L}) \\ &= 1.4277 - 1.4277 = 0.0\end{aligned}$$

Mutual Information: No Bias!

constant choice

0 (20)	50 (25)	0 (5)	$\mathbb{P}(\mathbf{G} = \boxplus) = .5$
0 (9)	36 (18)	0 (9)	$\mathbb{P}(\mathbf{G} = \Diamond) = .36$
0 (1)	14 (7)	0 (6)	$\mathbb{P}(\mathbf{G} = \boxminus) = .14$
$\mathbb{P}(\mathbf{L} = \boxplus) = .0$ (.3)	$\mathbb{P}(\mathbf{L} = \Diamond) = 1$ (.5)	$\mathbb{P}(\mathbf{L} = \boxminus) = .0$ (.2)	$\mathbb{H}(\mathbf{G}) = 1.4277$
$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxplus)$? (1.0746)	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \Diamond) = 1.4277$ (1.4277)	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxminus)$? (1.5395)	$\mathbb{H}(\mathbf{G}\|\mathbf{L}) = 1.4277$ (1.3441)

$$\begin{aligned}\mathbb{I}(\mathbf{G};\mathbf{L}) &= \mathbb{H}(\mathbf{G}) - \mathbb{H}(\mathbf{G}|\mathbf{L}) \\ &= 1.4277 - 1.4277 = 0.0\end{aligned}$$

Mutual Information: No Bias!

constant choice

0 (20)	0 (25)	50 (5)	$\mathbb{P}(\mathbf{G} = \boxplus) = .5$
0 (9)	0 (18)	35 (9)	$\mathbb{P}(\mathbf{G} = \Diamond) = .36$
0 (1)	0 (7)	14 (6)	$\mathbb{P}(\mathbf{G} = \boxminus) = .14$
$\mathbb{P}(\mathbf{L} = \boxplus) = .0$ (.3)	$\mathbb{P}(\mathbf{L} = \Diamond) = .0$ (.5)	$\mathbb{P}(\mathbf{L} = \boxminus) = 1$ (.2)	$\mathbb{H}(\mathbf{G}) = 1.4277$
$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxplus)$? (1.0746)	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \Diamond)$? (1.4277)	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxminus) = 1.4277$ (1.5395)	$\mathbb{H}(\mathbf{G}\|\mathbf{L}) = 1.4277$ (1.3441)

$$\begin{aligned}\mathbb{I}(\mathbf{G};\mathbf{L}) &= \mathbb{H}(\mathbf{G}) - \mathbb{H}(\mathbf{G}|\mathbf{L}) \\ &= 1.4277 - 1.4277 = 0.0\end{aligned}$$

Mutual Information: No Bias!

random choice

25 (20)	18 (25)	7 (5)	$\mathbb{P}(\mathbf{G} = \boxplus) = .5$
18 (9)	13 (18)	5 (9)	$\mathbb{P}(\mathbf{G} = \Diamond) = .36$
7 (1)	5 (7)	2 (6)	$\mathbb{P}(\mathbf{G} = \boxminus) = .14$
$\mathbb{P}(\mathbf{L} = \boxplus) = .5$ (.3)	$\mathbb{P}(\mathbf{L} = \Diamond) = .36$ (.5)	$\mathbb{P}(\mathbf{L} = \boxminus) = .14$ (.2)	$\mathbb{H}(\mathbf{G}) = 1.4277$
$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxplus) = 1.4277$ (1.0746)	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \Diamond) = 1.4277$ (1.4277)	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxminus) = 1.4277$ (1.5395)	$\mathbb{H}(\mathbf{G}\|\mathbf{L}) = 1.4277$ (1.3441)

$$\begin{aligned}\mathbb{I}(\mathbf{G};\mathbf{L}) &= \mathbb{H}(\mathbf{G}) - \mathbb{H}(\mathbf{G}|\mathbf{L}) \\ &= 1.4277 - 1.4277 = 0.0\end{aligned}$$

Mutual Information: No Bias!

.

(20)	(25)	(5)	$\mathbb{P}(\mathbf{G} = \boxplus) = .5$
(9)	(18)	(9)	$\mathbb{P}(\mathbf{G} = \Diamond) = .36$
(1)	(7)	(6)	$\mathbb{P}(\mathbf{G} = \boxminus) = .14$
$\mathbb{P}(\mathbf{L} = \boxplus)$ (.3)	$\mathbb{P}(\mathbf{L} = \Diamond)$ (.5)	$\mathbb{P}(\mathbf{L} = \boxminus)$ (.2)	$\mathbb{H}(\mathbf{G}) = 1.4277$
$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxplus)$ (1.0746)	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \Diamond)$ (1.4277)	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxminus)$ (1.5395)	$\mathbb{H}(\mathbf{G}\|\mathbf{L})$ (1.3441)

$$\begin{aligned}\mathbb{I}(\mathbf{G};\mathbf{L}) &= \mathbb{H}(\mathbf{G}) - \mathbb{H}(\mathbf{G}|\mathbf{L}) \\ &= 1.4277 - 1.3441 = 0.0834\end{aligned}$$

Mutual Information: Degradation Problem

degradation

20 (20)	25 (25)	5 (5)	$\mathbb{P}(\mathbf{G} = \boxplus) = .5$
9 (9)	18 (18)	9 (9)	$\mathbb{P}(\mathbf{G} = \Diamond) = .36$
1 (1)	7 (7)	6 (6)	$\mathbb{P}(\mathbf{G} = \boxminus) = .14$
$\mathbb{P}(\mathbf{L} = \boxplus) = .3$ (.3)	$\mathbb{P}(\mathbf{L} = \Diamond) = .5$ (.5)	$\mathbb{P}(\mathbf{L} = \boxminus) = .2$ (.2)	$\mathbb{A}_3(\mathrm{L};\mathrm{G}) = .44$ (.44)
$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxplus) = 1.0746$ (1.0746)	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \Diamond) = 1.4277$ (1.4277)	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxminus) = 1.5395$ (1.5395)	$\mathbb{H}(\mathbf{G}\|\mathbf{L}) = 1.3441$ (1.3441)

$$\mathbb{I}(\mathbf{G};\mathbf{L}) = \mathbb{H}(\mathbf{G}) - \mathbb{H}(\mathbf{G}|\mathbf{L})$$
$$= 1.4277 - 1.3441 = 0.0834 \quad (0.0834)$$

Mutual Information: Degradation Problem

degradation

45 (20)	0 (25)	5 (5)	$\mathbb{P}(\mathbf{G} = \boxplus) = .5$
27 (9)	0 (18)	9 (9)	$\mathbb{P}(\mathbf{G} = \Diamond) = .36$
8 (1)	0 (7)	6 (6)	$\mathbb{P}(\mathbf{G} = \boxminus) = .14$
$\mathbb{P}(\mathbf{L} = \boxplus) = .8$ (.3)	$\mathbb{P}(\mathbf{L} = \Diamond) = .0$ (.5)	$\mathbb{P}(\mathbf{L} = \boxminus) = .2$ (.2)	$\mathbb{A}_3(\mathrm{L};\mathrm{G}) = .51$ (.44)
$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxplus) = 1.3280$ (1.0746)	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \Diamond)$? (1.4277)	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxminus) = 1.5395$ (1.5395)	$\mathbb{H}(\mathbf{G}\|\mathbf{L}) = 1.3703$ (1.3441)

$$\mathbb{I}(\mathbf{G};\mathbf{L}) = \mathbb{H}(\mathbf{G}) - \mathbb{H}(\mathbf{G}|\mathbf{L})$$
$$= 1.4277 - 1.3703 = 0.0262 \quad (0.0834)$$

Mutual Information: Degradation Problem

degradation

45 (20)	0 (25)	5 (5)	$\mathbb{P}(\mathbf{G} = \boxplus) = .5$
27 (9)	0 (18)	9 (9)	$\mathbb{P}(\mathbf{G} = \Diamond) = .36$
8 (1)	0 (7)	6 (6)	$\mathbb{P}(\mathbf{G} = \boxminus) = .14$
$\mathbb{P}(\mathbf{L} = \boxplus) = .8$ (.3)	$\mathbb{P}(\mathbf{L} = \Diamond) = .0$ (.5)	$\mathbb{P}(\mathbf{L} = \boxminus) = .2$ (.2)	$\mathbb{A}_3(\mathrm{L};\mathrm{G}) = .51$ (.44)
$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxplus) = 1.3280$ (1.0746)	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \Diamond)$? (1.4277)	$\mathbb{H}(\mathbf{G}\|\mathbf{L} = \boxminus) = 1.5395$ (1.5395)	$\mathbb{H}(\mathbf{G}\|\mathbf{L}) = 1.3703$ (1.3441)

$$\mathbb{I}(\mathbf{G};\mathbf{L}) = \mathbb{H}(\mathbf{G}) - \mathbb{H}(\mathbf{G}|\mathbf{L})$$
$$= 1.4277 - 1.3703 = 0.0262 \ \ (0.0834)$$

Mutual Information vs. Accuracy: Degradation

In response to degradation:

- MI: .0834 $\rightarrow$ .0262
- Acc: .44 $\rightarrow$ .51

Outline

Final Recommendations

- in addition to accuracy, report MI,
- use MI for ranking,
- optimize systems for MI,
- drop average precision! use confidence-ranked MI and/or bring back CWS
- be aware of bias and baseline scores when looking at accuracies.

Final Recommendations

- in addition to accuracy, report MI,
- use MI for ranking,
- optimize systems for MI,
- drop average precision! use confidence-ranked MI and/or bring back CWS
- be aware of bias and baseline scores when looking at accuracies.

Final Recommendations

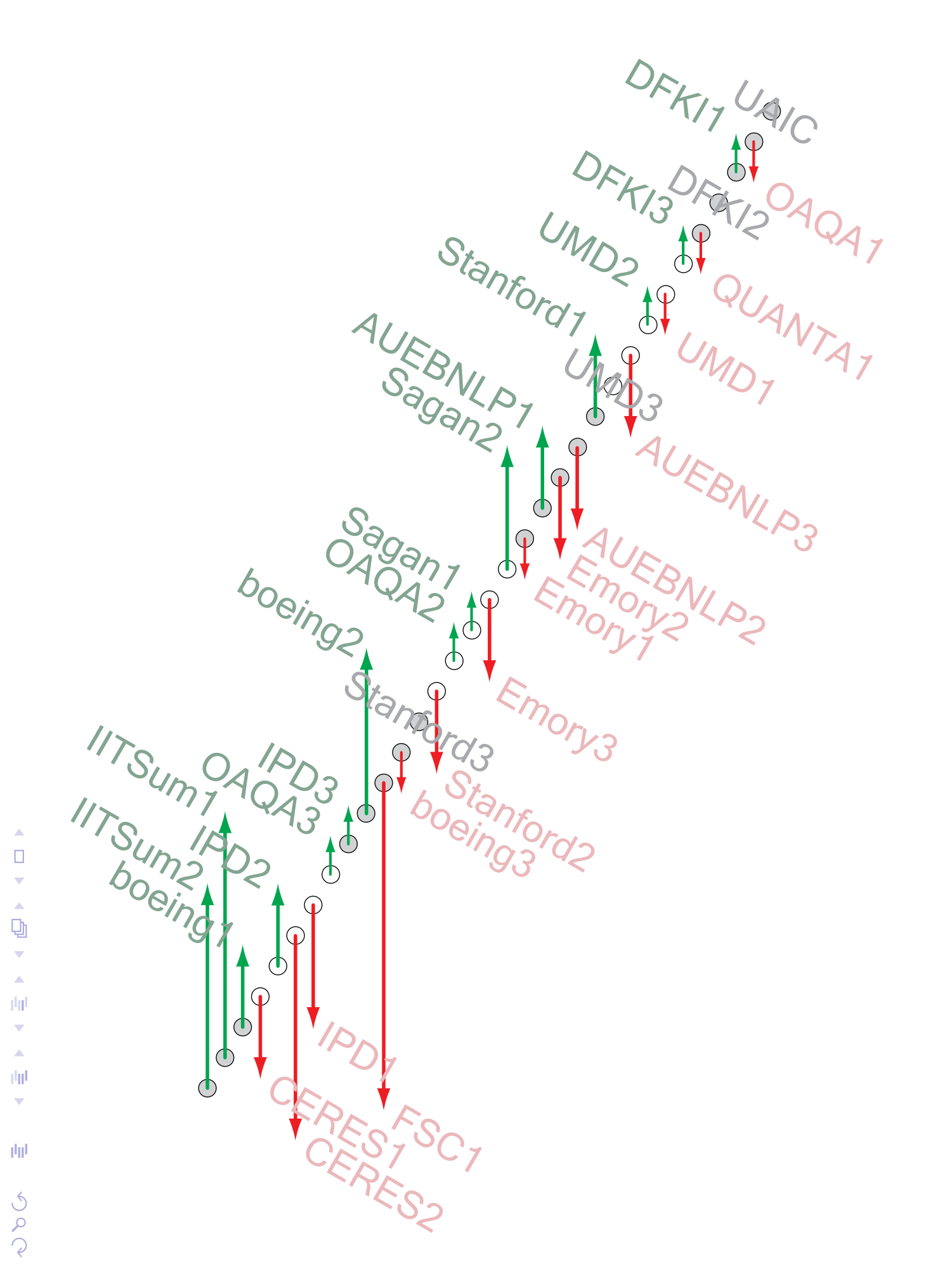
- in addition to accuracy, report MI,
- use MI for ranking,
- optimize systems for MI,
- drop average precision! use confidence-ranked MI and/or bring back CWS
- be aware of bias and baseline scores when looking at accuracies.

Final Recommendations

- in addition to accuracy, report MI,
- use MI for ranking,
- optimize systems for MI,
- drop average precision! use confidence-ranked MI and/or bring back CWS
- be aware of bias and baseline scores when looking at accuracies.

Final Recommendations

- in addition to accuracy, report MI,
- use MI for ranking,
- optimize systems for MI,
- drop average precision! use confidence-ranked MI and/or bring back CWS
- be aware of bias and baseline scores when looking at accuracies.


DFKI1
UAIC
DFKI3
DFKI2
OAQA1
UMD2
QUANTA1
Stanford1
UMD1
AUEBNLP1
UMD3
Sagan2
AUEBNLP3
AUEBNLP2
Sagan1
Emory2
OAQA2
Emory1
boeing2
Emory3
Stanford3
IPD3
Stanford2
IITSum1
OAQA3
boeing3
IPD2
IITSum2
boeing1
IPD1
FSC1
CERES1
CERES2