A Grain of Salt for the WMT Manual Evaluation

Ondřej Bojar, Miloš Ercegovčević
Martin Popel, and Omar F. Zaidan
{bojar,popel}@ufal.mff.cuni.cz
ercegovcevic@hotmail.com
ozaidan@cs.jhu.edu

Sat July 30, 2011
Outline

• Two Interpretations of Ranking.
• Annotator Agreement.
• Rewarding Ties?
• Head-to-Head Comparisons.
• Reference Translations.
• Concluding Suggestions.
Manual Ranking of MT Outputs

Source: Záchranáři kromě dívky vylovili z vody několik těl a trosek letounu.
Reference: Apart from the girl rescuers also pulled a few bodies and some debris from the sea.

| Translation                                                                 | Rank (1=Best, 5=Worst, ties are OK) |
|----------------------------------------------------------------------------|-------------------------------------|
| Rescue workers in addition to the girls from water picked up a few of the carcases and the ruin of the plane. | 1 (Best) 2 3 4 5 (Worst)            |
| Rescue workers in addition to the girls picked out of the water several bodies and the ruins of the aeroplane. | 1 (Best) 2 3 4 5 (Worst)            |
| Then, in addition to the girls up from the water for a few bodies and wreckage of the plane. | 1 (Best) 2 3 4 5 (Worst)            |
| Rescue except girls fished out of the water several bodies and the wreckage of the plane. | 1 (Best) 2 3 4 5 (Worst)            |
| Apart from the girl rescuers also pulled a few bodies and some debris from the sea. | 1 (Best) 2 3 4 5 (Worst)            |
Interpreting Manual Ranks

better

A • • • • • •
B • • • • • •
C • • • • • •
D • • • • • •
Interpreting Manual Ranks

better

A • • • • • • • • •
B • • • • • • • • •
C • • • • • • • • •
D • • • • • • • • •

"block"
Interpreting Manual Ranks

better

A • • • • •
B • • • • •
C • • • • •
D • • • • •

A • • • • •
B • • • • •
C • • • • •
E • • • • •
Interpreting Manual Ranks

Who Wins WMT?
Interpreting Manual Ranks

[Systems] are ranked based on how frequently they were judged to be better than or equal to any other system.
Interpreting Manual Ranks

"≥ All in Block"

A: 1/2  
B: 0/2  
C: 0/2  
D: 0/1  
E: 1/1  

Sat July 30, 2011  A Grain of Salt for WMT Evaluation
Interpreting Manual Ranks

"≥ All in Block"

A: 1/2
B: 0/2
C: 0/2
D: 0/1
E: 1/1
Interpreting Manual Ranks

Simulated Pairwise

"≥ All in Block"

A: 1/2
B: 0/2
C: 0/2
D: 0/1
E: 1/1
Interpreting Manual Ranks

Simulated Pairwise

A > B
A > C
A > D
B = C
B > D
C > D

"≥ All in Block"

A: 1/2
B: 0/2
C: 0/2
D: 0/1
E: 1/1

Sat July 30, 2011 A Grain of Salt for WMT Evaluation
Interpreting Manual Ranks

Simulated Pairwise

A > B, A < B
A > C, A < C
A > D
B = C
B < E
C < E

"≥ All in Block"

A: 1/2
B: 0/2
C: 0/2
D: 0/1
E: 1/1

Sat July 30, 2011 A Grain of Salt for WMT Evaluation
Interpreting Manual Ranks

better

Simulated Pairwise

A:B  A>B  A<B
A>C  A>C
A>D
B=C  B=C
B>D
C>D

"≥ All in Block"

A: 1/2
B: 0/2
C: 0/2
D: 0/1
E: 1/1

Sat July 30, 2011
Interpreting Manual Ranks

Simulated Pairwise

```
A > B  A < B
A > C  A < C
A > D
B = C  B = C
B > D
C > D
```

"≥ Others"

```
A: 3/6
B: 4/6
C: 4/6
D: 0/3
E: 3/3
```

"≥ All in Block"

```
A: 1/2
B: 0/2
C: 0/2
D: 0/1
E: 1/1
```

Sat July 30, 2011

A Grain of Salt for WMT Evaluation
Interpreting Manual Ranks

Simulated Pairwise

- A > B
- A > C
- A > D
- B = C
- B > D
- C > D

"≥ Others"

- A: 3/6
- B: 4/6
- C: 4/6
- D: 0/3
- E: 3/3

"≥ All in Block"

- A: 1/2
- B: 0/2
- C: 0/2
- D: 0/1
- E: 1/1

Sat July 30, 2011 A Grain of Salt for WMT Evaluation
Interpreting Manual Ranks

Simulated Pairwise

"≥ Others"

"≥ All in Block"

A: 3/6
B: 4/6
C: 4/6
D: 0/3
E: 3/3

A: 1/2
B: 0/2
C: 0/2
D: 0/1
E: 1/1

A > B
A > C
A > D
B = C
B > D
C > D
A ≤ E
B ≤ E
C ≤ E
Interpreting Manual Ranks

Simulated Pairwise | "≥ Others" | "≥ All in Block"
---|---|---
A > B | A: 3/6 | A: 1/2
A > C | B: 4/6 | B: 0/2
A > D | C: 4/6 | C: 0/2
B = C | D: 0/3 | D: 0/1
B > D | E: 3/3 | E: 1/1
C > D | | |

A < E
B < E
C < E
Interpreting Manual Ranks

Simulated Pairwise

| Simulated Pairwise | "≥ Others" | "≥ All in Block" |
|--------------------|------------|------------------|
| A > B              | A < B      | A: 3/6           |
| A > C              | A < C      | B: 4/6           |
| A > D              |             |                  |
| B = C              | B = C      |                  |
| B > D              |             |                  |
| C > D              |             |                  |

AB C
better
A>B A<C A>D B=C B>D C>D
A<EB<E C<E

Simulated Pairwise

| Simulated Pairwise | "≥ Others" | "≥ All in Block" |
|--------------------|------------|------------------|
| A = B              | A < B      | A: 1/2           |
| B = C              | A < C      | A: 1/2           |
| B < E              | B < B      |                  |
| C < B              |             |                  |
| C < E              |             |                  |

A: 3/6 A: 1/2
A: 4/6 B: 0/2
“≥ All in Block” Similar

Czech-English
English-Czech
English-French
English-German
English-Spanish
French-English
German-English
Spanish-English

\[ a \times x + b \]

Sat July 30, 2011 A Grain of Salt for WMT Evaluation
Speculation

“≥ All in Block” ≈ “Best vs. Rest”

Moving from 5-way ranking to 2-way classification:

- Should be easier.
- Could have higher agreement.

Agreement for “≥ all in block”:

- No data in WMT10 evaluations.
- Some in WMT11 (analysis pending).
- WMT12 could sample even more to examine that.
Simulated Pairwise

\[ P(A) = \frac{\# \text{agree}}{\# \text{comparisons}} \]

To account for agreement by chance

\[ \kappa = P(A) - P(E) \]

\[ \frac{1}{1 - P(E)} \]

Different defs. of \( P(E) \).

(\( \text{WMT<11} \) happened to pick an overly optimistic one.)
Annotator Agreement

Simulated Pairwise

A > B  A < B
A > C  A < C
A > D
B = C  B = C
B > D
C > D

A < E
B < E
C < E
Annotator Agreement

\[ P(A) = \frac{\# \text{ agree}}{\# \text{ comparisons}} \]

To account for agreement by chance \( P(E) \):

\[ \kappa = \frac{P(A) - P(E)}{1 - P(E)} \]

- Different defs. of \( P(E) \).
  (Little absolute dif. in \( \kappa \).)
  WMT<11 happened to pick an overly optimistic one.
Agreement Results

“WMT kappa”
(Bennett et al., 1954)  \( P(E) = \frac{1}{3} \)

(Scott, 1955)  \( P(E) \) empirical

| “\(\geq\) Others” | \(S\) | \(\pi\) |
|-------------------|------|------|
| Inter incl. ref.  | 0.487 | 0.454 |
| excl. ref.        | 0.439 | 0.403 |
| Intra incl. ref.  | 0.633 | 0.609 |
| excl. ref.        | 0.601 | 0.575 |

- \(\geq 0.4\) is said to be moderate.
Lower for Longer Sentences

Intra. incl. ref.
Intra. excl. ref.
Inter. incl. ref.
Inter. excl. ref.

Moderate agreement

Kappa, $P(E) = 0.33$
Krippendorff (1980) suggests that Kappa, $P(E) = 0.33$, is a useful measure for assessing inter-rater reliability.
Lower for Longer Sentences

Kappa, $P(E) = 0.33$

Source length vs. Number of sentences

Sat July 30, 2011 A Grain of Salt for WMT Evaluation
## Rewarding Ties?

|               | “> Others”       | “> Others”       | “Ignore Ties”    |
|---------------|------------------|------------------|------------------|
| Favours       | “mainstream”     | “distinct”       | -                |
|               | \( \frac{\text{wins} + \text{ties}}{\text{wins} + \text{ties} + \text{losses}} \) | \( \frac{\text{wins}}{\text{wins} + \text{ties} + \text{losses}} \) | \( \frac{\text{wins}}{\text{wins} + \text{losses}} \) |

\[ \frac{24}{40} = \frac{6}{10} \]
### Rewarding Ties?

|                  | “$>$ Others” | “$>$ Others” | “Ignore Ties” |
|------------------|--------------|--------------|---------------|
| **Favours**      | wins + ties  | wins         | wins          |
|                  | wins + ties + losses | wins + ties + losses | wins + losses |
| **“mainstream”** | wins + ties  | wins         | -             |
| **“distinct”**   | wins         | wins         | -             |

Wanna cheat WMT?
# Rewarding Ties?

|                      | “$>\ Others$” | “$>\ Others$” | “Ignore Ties” |
|----------------------|---------------|---------------|---------------|
| **Favours**          | “mainstream”  | “distinct”    |               |
|                      | **wins+ties** | **wins**      | **wins**      |
|                      | **wins+ties+losses** | **wins+ties+losses** | **wins+ties+losses** |
|                      | **D**        | **M1**        | **M1**        |
|                      | **M1**       | **M2**        | **M2**        |
|                      | **M2**       | **M3**        | **M3**        |
|                      | **M3**       | **M4**        | **M4**        |
|                      | **6x**       | **4x**        |               |

![Diagram](image)

D 6 × 4 = 24/40

M1 10 × 3 + 4 = 34/40

Sat July 30, 2011
## Rewarding Ties?

| Favours | “> Others” | “> Others” | “Ignore Ties” |
|---------|-------------|-------------|---------------|
| “mainstream” | wins + ties / wins + ties + losses | wins / wins + ties + losses | wins / wins + losses |
| “distinct” | wins / wins + ties + losses | - | - |

**D**

- 6 \times 4 = 24/40
- 24/40
- 24 / 40 = 6/10

**M1**

- 10 \times 3 + 4 = 34/40
- 4/40
- 4/10

---

Sat July 30, 2011

A Grain of Salt for WMT Evaluation
WMT11 Results of English-Czech
Head-to-Head Comparisons

- WMT overview paper also reports head-to-head comparisons.
- Head-to-head not always in line with official \( \geq \) others.

|                  | \( \geq \) Others | H-to-H | \( \geq \) Others | H-to-H |
|------------------|-------------------|--------|-------------------|--------|
| CU-BOJAR         | 65.6              | 35.8   | 401               | 81     |
| CU-TECTO         | 60.1              | 45.7   | 392               |        |

Head-to-head is estimated:
- on much smaller dataset,
- different set of sentences.
Indistinguishable Systems

- Even a targeted pairwise comparison may not tell who is better.
- Six independent annotations of 63 sentences.

| Annotator | CU-BOJAR | CU-TECTO | fine | wrong | Σ |
|-----------|----------|----------|------|-------|---|
| A         | 24       | 23       | 5    | 11    | 63 |
| C         | 10       | 12       | 5    | 36    | 63 |
| D         | 32       | 20       | 2    | 9     | 63 |
| M         | 11       | 18       | 7    | 27    | 63 |
| O         | 23       | 18       | 4    | 18    | 63 |
| Z         | 25       | 27       | 2    | 9     | 63 |
| **Total** | **125**  | **118**  | **25** | **110** | **378** |

⇒ Different annotators focus on different errors.
“Being compared (more often) to the ref. disfavors my system.”

| Source   | Target | Correlation of Ref. vs. “≥ Others” |
|----------|--------|-----------------------------------|
| Spanish  | English| 0.341                             |
| English  | French | 0.164                             |
| French   | English| 0.098                             |
| German   | English| 0.088                             |
| Czech    | English| -0.041                            |
| English  | Czech  | -0.145                            |
| English  | Spanish| -0.411                            |
| English  | German | -0.433                            |
| Overall  |        | -0.107                            |

Overall no (neg.) correlation between the ref. and “≥ others”. 
Even the worst language pair is caused by just a few outliers.
Final Suggestions for WMT

Sample differently:

- Allow measuring agreement for “≥ all in block”.
- Sample reference fewer times.
  - It’s not harmful, but we can save the labor.
- Run a pilot study with fewer sentences in block.
  - Esp. if we’re not restricting sentence length.

Evaluate differently:

- Ignore ties.
- Use empirical $P(E)$, i.e. $\pi$ by Scott (1955).
Avoid Humans!

Subject and object swapped in reference translations:

SRC  FCC awarded a tunnel in Slovenia for 64 million
REF  FCC byl přidělen tunel ve Slovinsku za 64 milionů

Gloss FCC was awarded a tunnel in Slovenia for 64 million

Rankings by the same annotator:

| SRC              | REF              | Ranks |
|------------------|------------------|-------|
| It’s not completely ideal. | Není to úplně ideální. |       |
| PC-TRANS         | To není úplně ideální. | 2 5   |
| CU-BOJAR         | To není úplně ideální. | 5 4   |
E. M. Bennett, R. Alpert, and A. C. Goldstein. 1954. Communications through limited questioning. Public Opinion Quarterly, 18(3):303–308.

Klaus Krippendorff. 1980. Content Analysis: An Introduction to Its Methodology. Sage Publications, Beverly Hills, CA. Chapter 12.

William A. Scott. 1955. Reliability of content analysis: The case of nominal scale coding. Public Opinion Quarterly, 19(3):321–325.