# 89688: Statistical Machine Translation

# Lecture 2: Evaluation

March 2020

Roee Aharoni
Computer Science Department
Bar Ilan University

Based in part on slides from Edinburgh University's MT class

"More has been written about machine translation evaluation than about machine translation itself."

“More has been written about machine translation evaluation than about machine translation itself.”

"More has been written about machine translation evaluation than about machine translation itself."

**Yorick Wilks**

# Why evaluate?

# Why evaluate?

- Rank competing systems

**English→German**

| Ave. | Ave. z | System |
|---|---|---|
| 90.3 | 0.347 | Facebook-FAIR |
| 93.0 | 0.311 | Microsoft-WMT19-sent-doc |
| 92.6 | 0.296 | Microsoft-WMT19-doc-level |
| 90.3 | 0.240 | HUMAN |
| 87.6 | 0.214 | MSRA-MADL |
| 88.7 | 0.213 | UCAM |
| 89.6 | 0.208 | NEU |
| 87.5 | 0.189 | MLLP-UPV |
| 87.5 | 0.130 | eTranslation |
| 86.8 | 0.119 | dfki-nmt |
| 84.2 | 0.094 | online-B |
| 86.6 | 0.094 | Microsoft-WMT19-sent-level |
| 87.3 | 0.081 | JHU |
| 84.4 | 0.077 | Helsinki-NLP |

# Why evaluate?

- Rank competing systems

**NAACL 2006 WORKSHOP ON STATISTICAL MACHINE TRANSLATION**

**June 8 and 9, 2006**

**English→German**

| Ave. | Ave. z | System |
|---|---|---|
| 90.3 | 0.347 | Facebook-FAIR |
| 93.0 | 0.311 | Microsoft-WMT19-sent-doc |
| 92.6 | 0.296 | Microsoft-WMT19-doc-level |
| 90.3 | 0.240 | HUMAN |
| 87.6 | 0.214 | MSRA-MADL |
| 88.7 | 0.213 | UCAM |
| 89.6 | 0.208 | NEU |
| 87.5 | 0.189 | MLLP-UPV |
| 87.5 | 0.130 | eTranslation |
| 86.8 | 0.119 | dfki-nmt |
| 84.2 | 0.094 | online-B |
| 86.6 | 0.094 | Microsoft-WMT19-sent-level |
| 87.3 | 0.081 | JHU |
| 84.4 | 0.077 | Helsinki-NLP |

# Why evaluate?

- Rank competing systems
- Make incremental improvements

**NAACL 2006 WORKSHOP ON STATISTICAL MACHINE TRANSLATION**

**June 8 and 9, 2006**

**English→German**

| Ave. | Ave. z | System |
|---|---|---|
| 90.3 | 0.347 | Facebook-FAIR |
| 93.0 | 0.311 | Microsoft-WMT19-sent-doc |
| 92.6 | 0.296 | Microsoft-WMT19-doc-level |
| 90.3 | 0.240 | HUMAN |
| 87.6 | 0.214 | MSRA-MADL |
| 88.7 | 0.213 | UCAM |
| 89.6 | 0.208 | NEU |
| 87.5 | 0.189 | MLLP-UPV |
| 87.5 | 0.130 | eTranslation |
| 86.8 | 0.119 | dfki-nmt |
| 84.2 | 0.094 | online-B |
| 86.6 | 0.094 | Microsoft-WMT19-sent-level |
| 87.3 | 0.081 | JHU |
| 84.4 | 0.077 | Helsinki-NLP |

# Why evaluate?

- Rank competing systems
- Make incremental improvements
  - More data?

**NAACL 2006 WORKSHOP ON STATISTICAL MACHINE TRANSLATION**

**June 8 and 9, 2006**

**English→German**

| Ave. | Ave. z | System |
|---|---|---|
| 90.3 | 0.347 | Facebook-FAIR |
| 93.0 | 0.311 | Microsoft-WMT19-sent-doc |
| 92.6 | 0.296 | Microsoft-WMT19-doc-level |
| 90.3 | 0.240 | HUMAN |
| 87.6 | 0.214 | MSRA-MADL |
| 88.7 | 0.213 | UCAM |
| 89.6 | 0.208 | NEU |
| 87.5 | 0.189 | MLLP-UPV |
| 87.5 | 0.130 | eTranslation |
| 86.8 | 0.119 | dfki-nmt |
| 84.2 | 0.094 | online-B |
| 86.6 | 0.094 | Microsoft-WMT19-sent-level |
| 87.3 | 0.081 | JHU |
| 84.4 | 0.077 | Helsinki-NLP |

# Why evaluate?

- Rank competing systems
- Make incremental improvements
  - More data?
  - Different preprocessing?

**NAACL 2006 WORKSHOP ON STATISTICAL MACHINE TRANSLATION**

**June 8 and 9, 2006**

**English→German**

| Ave. | Ave. z | System |
|---|---|---|
| 90.3 | 0.347 | Facebook-FAIR |
| 93.0 | 0.311 | Microsoft-WMT19-sent-doc |
| 92.6 | 0.296 | Microsoft-WMT19-doc-level |
| 90.3 | 0.240 | HUMAN |
| 87.6 | 0.214 | MSRA-MADL |
| 88.7 | 0.213 | UCAM |
| 89.6 | 0.208 | NEU |
| 87.5 | 0.189 | MLLP-UPV |
| 87.5 | 0.130 | eTranslation |
| 86.8 | 0.119 | dfki-nmt |
| 84.2 | 0.094 | online-B |
| 86.6 | 0.094 | Microsoft-WMT19-sent-level |
| 87.3 | 0.081 | JHU |
| 84.4 | 0.077 | Helsinki-NLP |

# Why evaluate?

- Rank competing systems
- Make incremental improvements
  - More data?
  - Different preprocessing?
  - Different hyperparameters?

**NAACL 2006 WORKSHOP ON STATISTICAL MACHINE TRANSLATION**

**June 8 and 9, 2006**

**English→German**

| Ave. | Ave. z | System |
|---|---|---|
| 90.3 | 0.347 | Facebook-FAIR |
| 93.0 | 0.311 | Microsoft-WMT19-sent-doc |
| 92.6 | 0.296 | Microsoft-WMT19-doc-level |
| 90.3 | 0.240 | HUMAN |
| 87.6 | 0.214 | MSRA-MADL |
| 88.7 | 0.213 | UCAM |
| 89.6 | 0.208 | NEU |
| 87.5 | 0.189 | MLLP-UPV |
| 87.5 | 0.130 | eTranslation |
| 86.8 | 0.119 | dfki-nmt |
| 84.2 | 0.094 | online-B |
| 86.6 | 0.094 | Microsoft-WMT19-sent-level |
| 87.3 | 0.081 | JHU |
| 84.4 | 0.077 | Helsinki-NLP |

# Why evaluate?

- Rank competing systems
- Make incremental improvements
  - More data?
  - Different preprocessing?
  - Different hyperparameters?
- Evaluate new ideas

**NAACL 2006 WORKSHOP ON STATISTICAL MACHINE TRANSLATION**

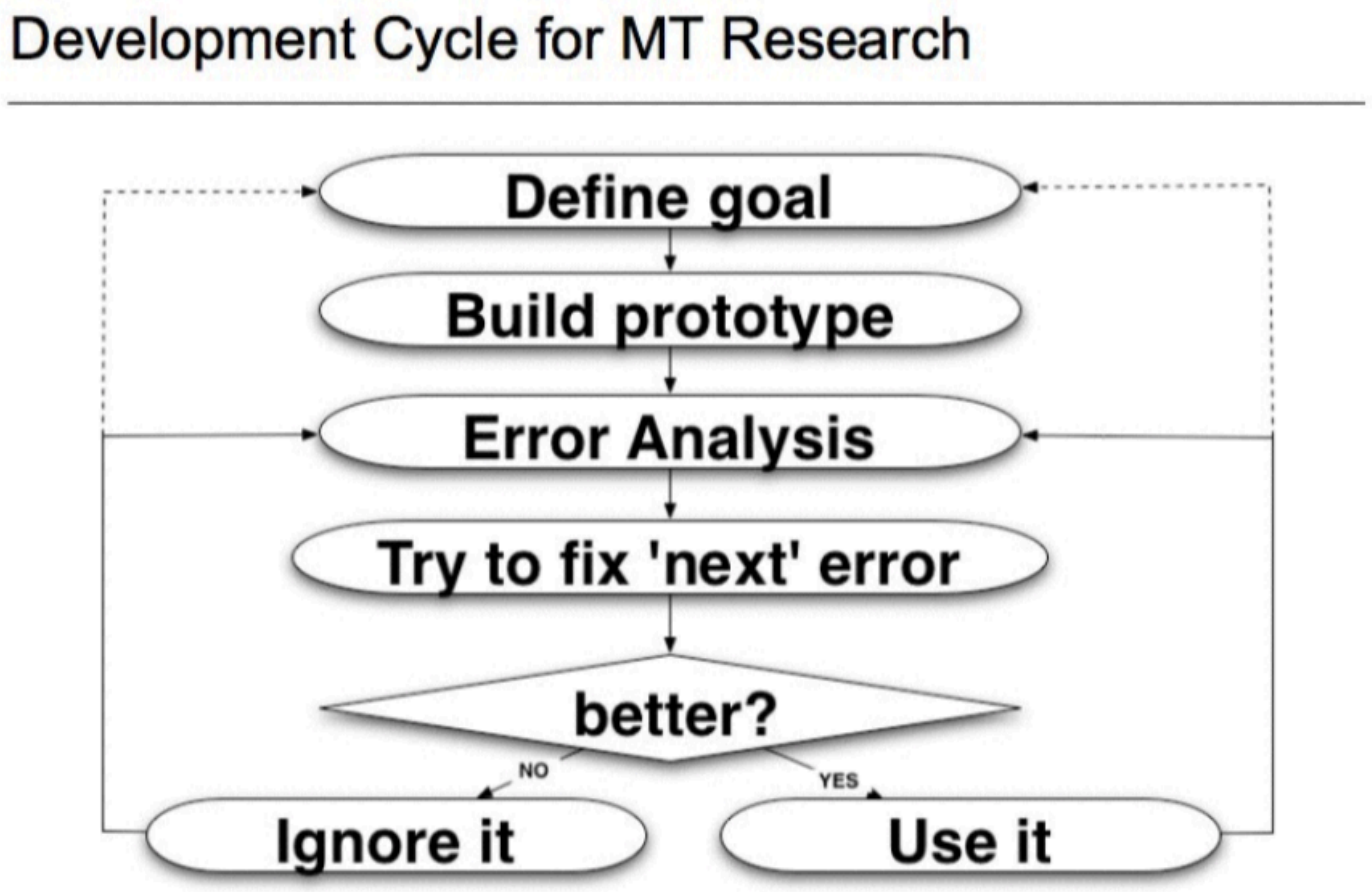
**June 8 and 9, 2006**

**English→German**

| Ave. | Ave. z | System |
|---|---|---|
| 90.3 | 0.347 | Facebook-FAIR |
| 93.0 | 0.311 | Microsoft-WMT19-sent-doc |
| 92.6 | 0.296 | Microsoft-WMT19-doc-level |
| 90.3 | 0.240 | HUMAN |
| 87.6 | 0.214 | MSRA-MADL |
| 88.7 | 0.213 | UCAM |
| 89.6 | 0.208 | NEU |
| 87.5 | 0.189 | MLLP-UPV |
| 87.5 | 0.130 | eTranslation |
| 86.8 | 0.119 | dfki-nmt |
| 84.2 | 0.094 | online-B |
| 86.6 | 0.094 | Microsoft-WMT19-sent-level |
| 87.3 | 0.081 | JHU |
| 84.4 | 0.077 | Helsinki-NLP |

# Why evaluate?

# Why evaluate?




**Evaluation enables progress**

# What is a good translation?

# What is a good translation?

# What is a good translation?

- Transitions from one language to another

# What is a good translation?

- Transitions from one language to another
- Preserves the meaning

# What is a good translation?

- Transitions from one language to another
- Preserves the meaning
- Fluent?

# What is a good translation?

- Transitions from one language to another
- Preserves the meaning
- Fluent?
- Preserves style?

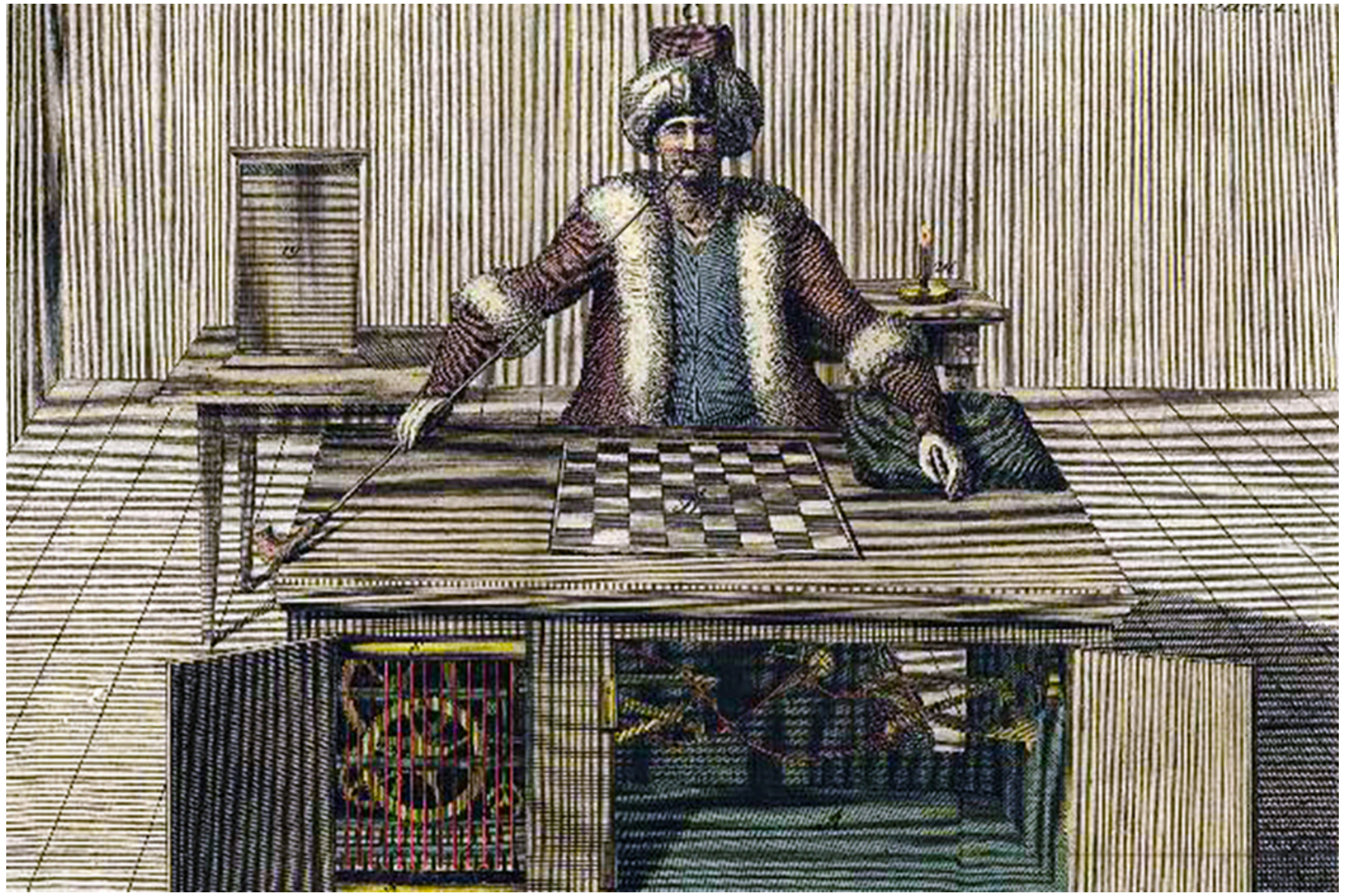
# What is a good translation?

- Transitions from one language to another
- Preserves the meaning
- Fluent?
- Preserves style?
- What else?

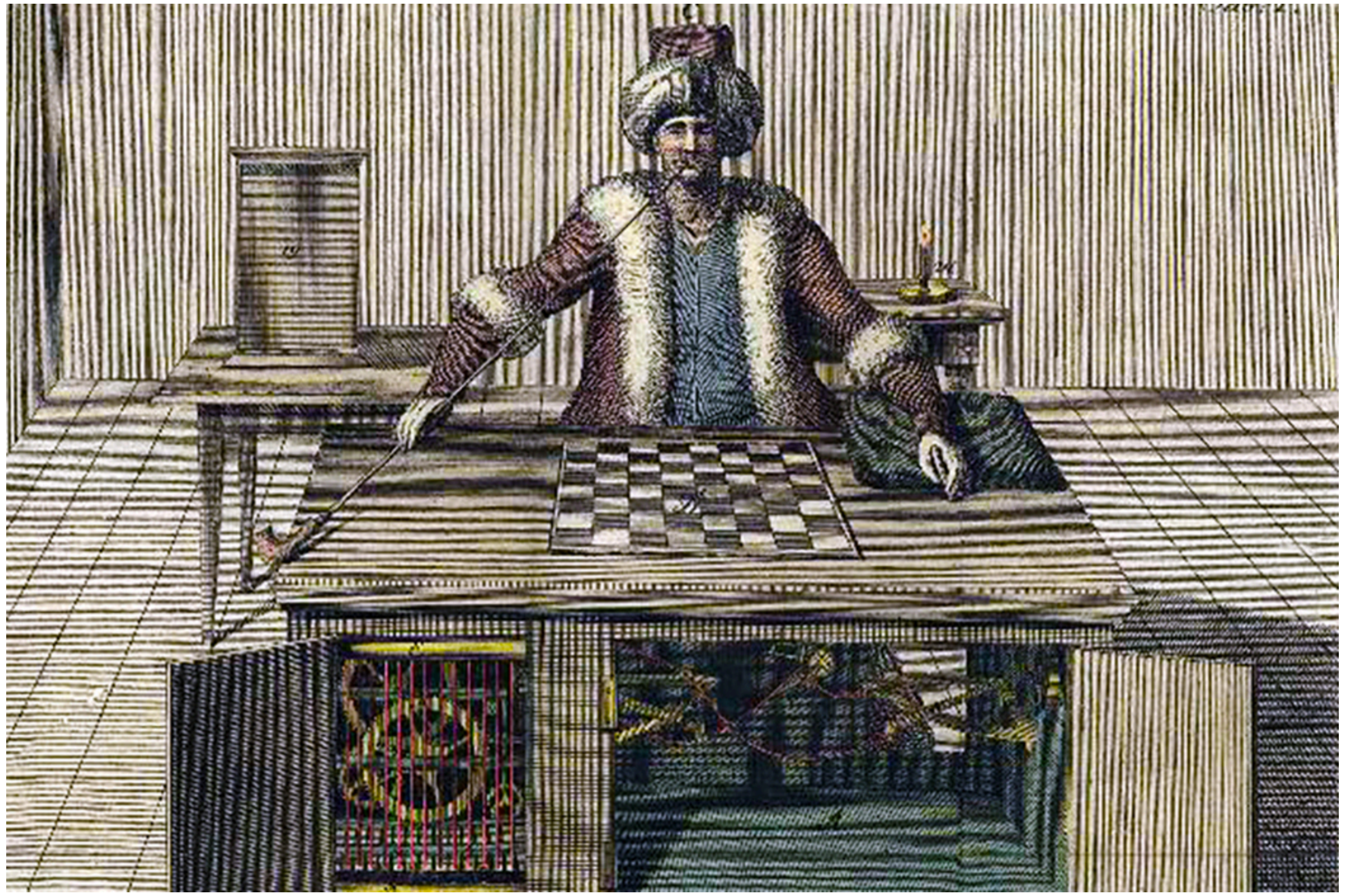
# How can we *measure* this?

# How can we *measure* this?

# How can we *measure* this?



# How can we *measure* this?



# How can we *measure* this?

# How can we *measure* this?

| | Human | Automatic |
|---|---|---|

# How can we *measure* this?

| | Human | Automatic |
|---|---|---|
| **Accurate** | Yes | Sometimes… |

# How can we *measure* this?

| | Human | Automatic |
|---|---|---|
| **Accurate** | Yes | Sometimes… |
| **Speed** | Slow | Fast |

# How can we *measure* this?

| | Human | Automatic |
|---|---|---|
| **Accurate** | Yes | Sometimes… |
| **Speed** | Slow | Fast |
| **Price** | Expensive | Cheap |

# How can we *measure* this?

| | Human | Automatic |
|---|---|---|
| **Accurate** | Yes | Sometimes… |
| **Speed** | Slow | Fast |
| **Price** | Expensive | Cheap |
| **Subjectivity** | Subjective | Objective |

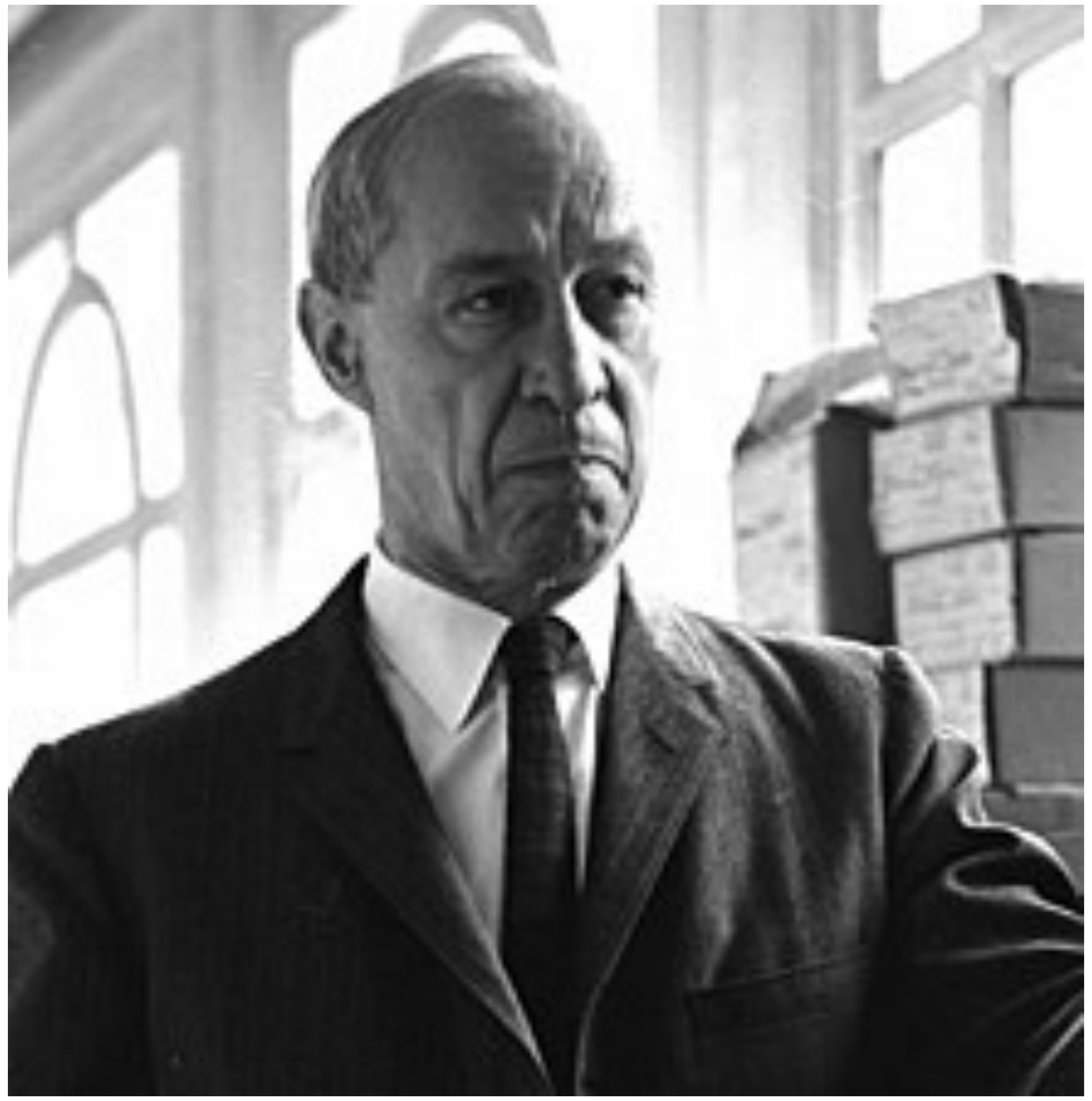
# How can we *measure* this?

| | Human | Automatic |
|---|---|---|
| **Accurate** | Yes | Sometimes… |
| **Speed** | Slow | Fast |
| **Price** | Expensive | Cheap |
| **Subjectivity** | Subjective | Objective |
| **Reproducible** | No | Yes |

# Human Evaluation Methods

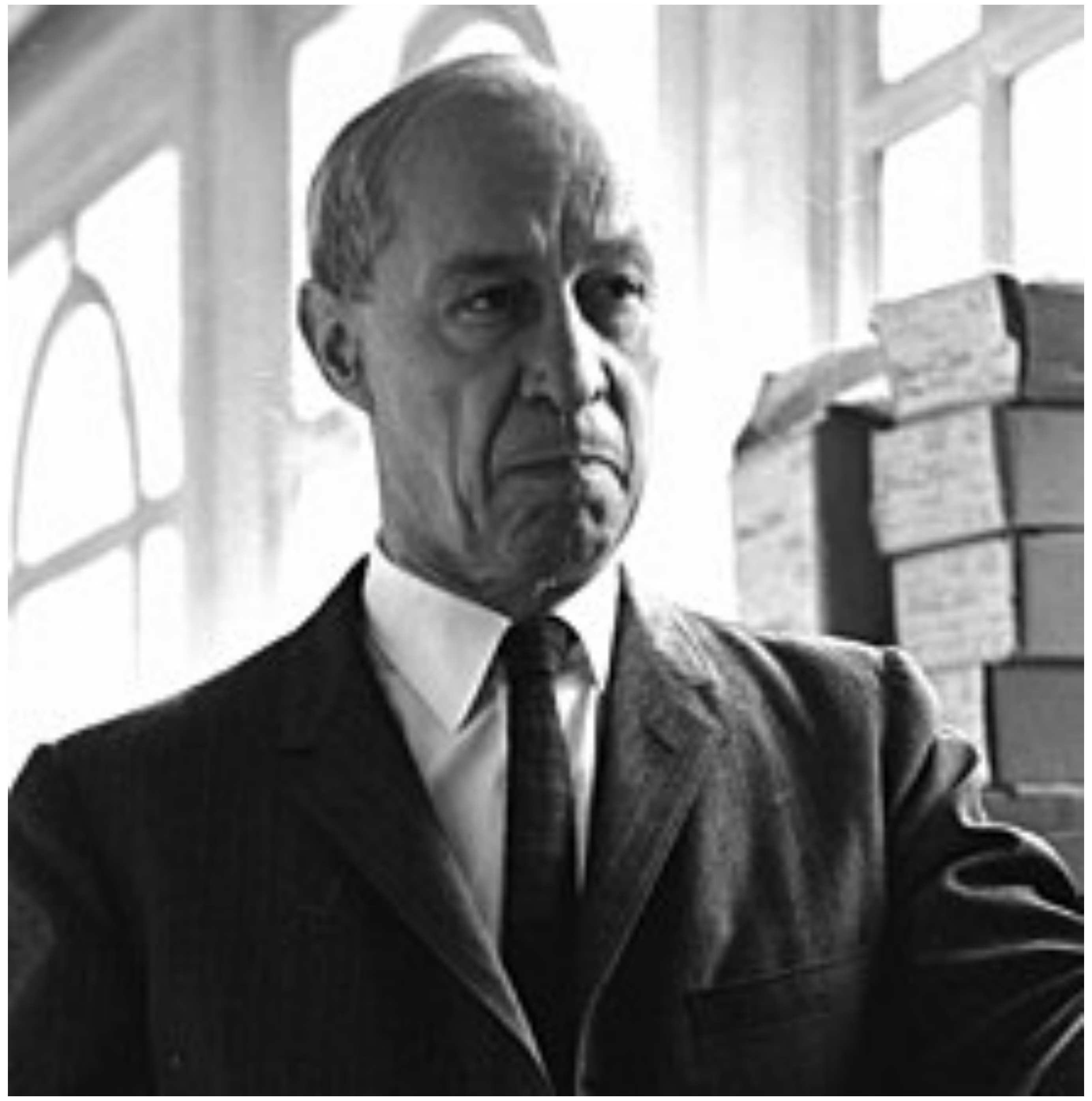
# The Likert Scale

# The Likert Scale

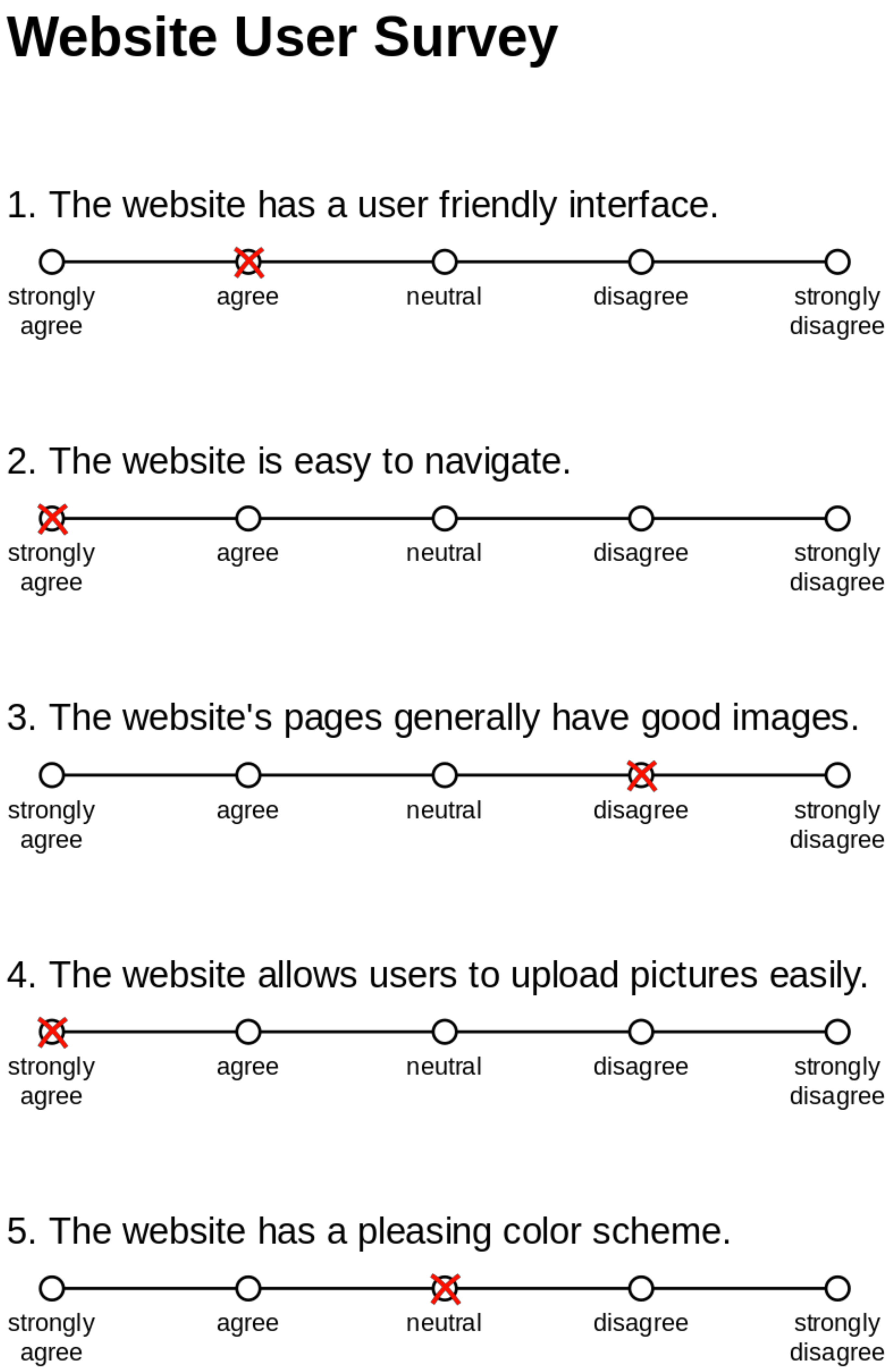


Rensis Likert

# The Likert Scale

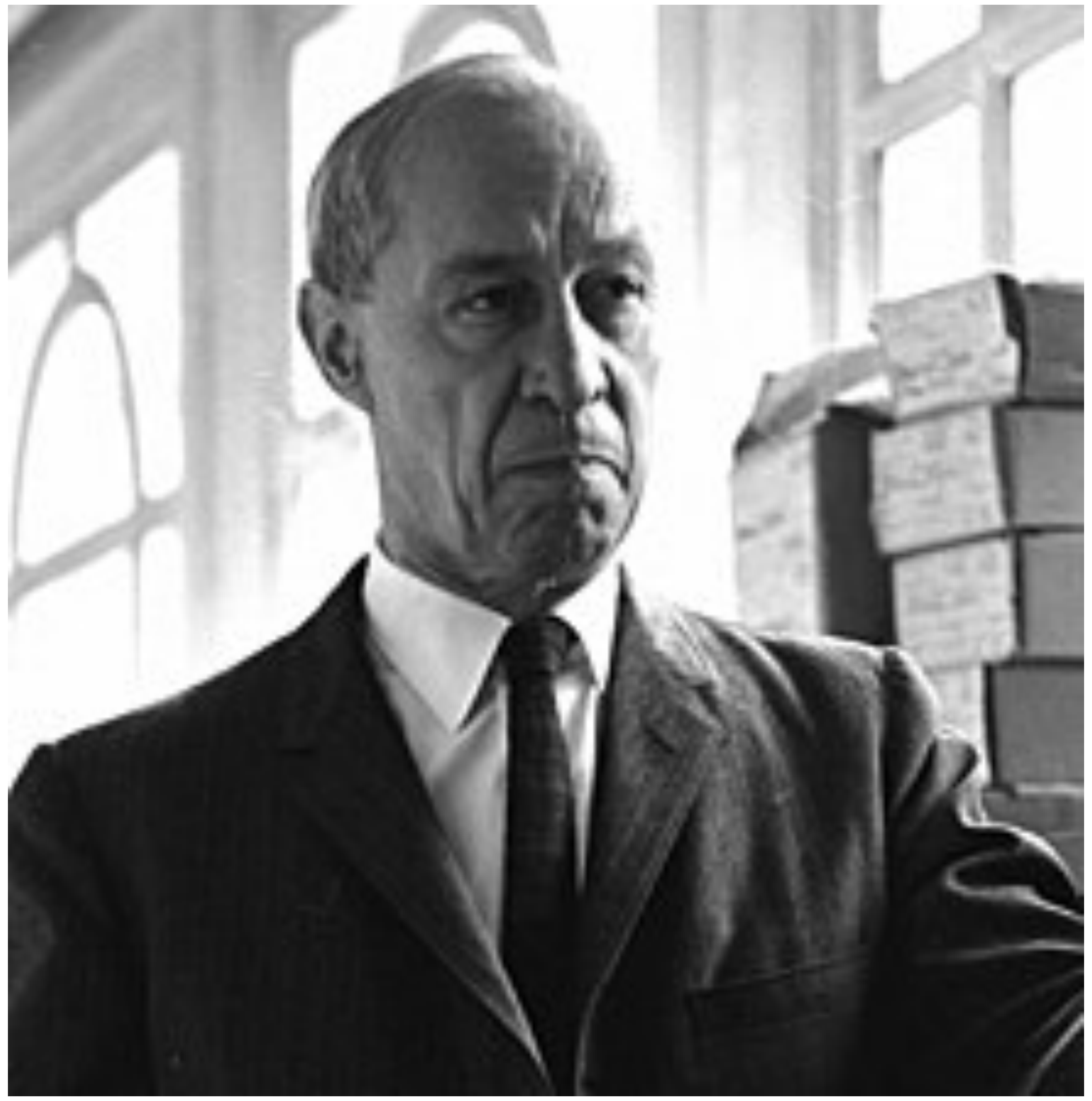


Rensis Likert

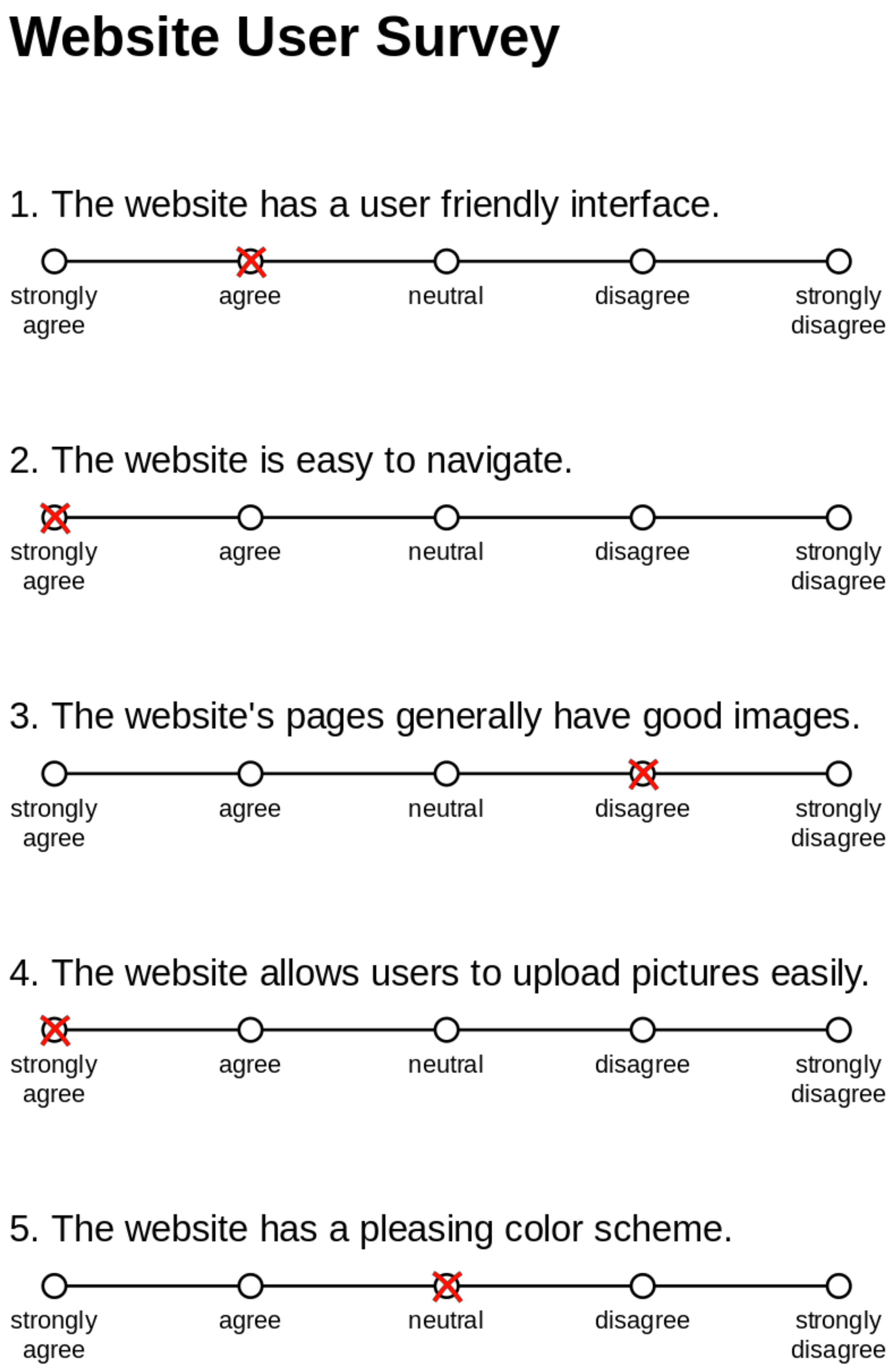


# The Likert Scale

- WMT 06’ - WMT 07’

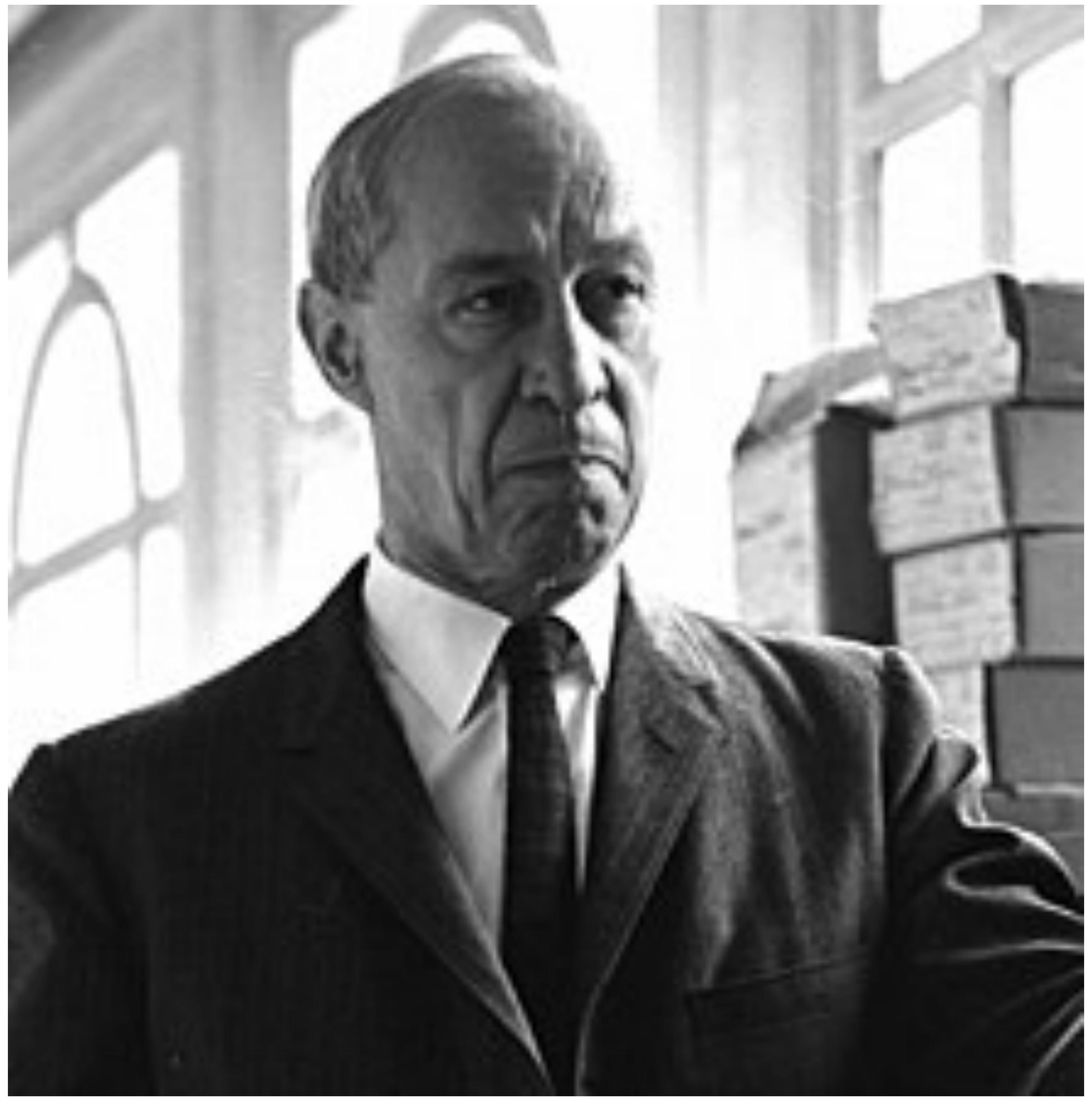


Rensis Likert

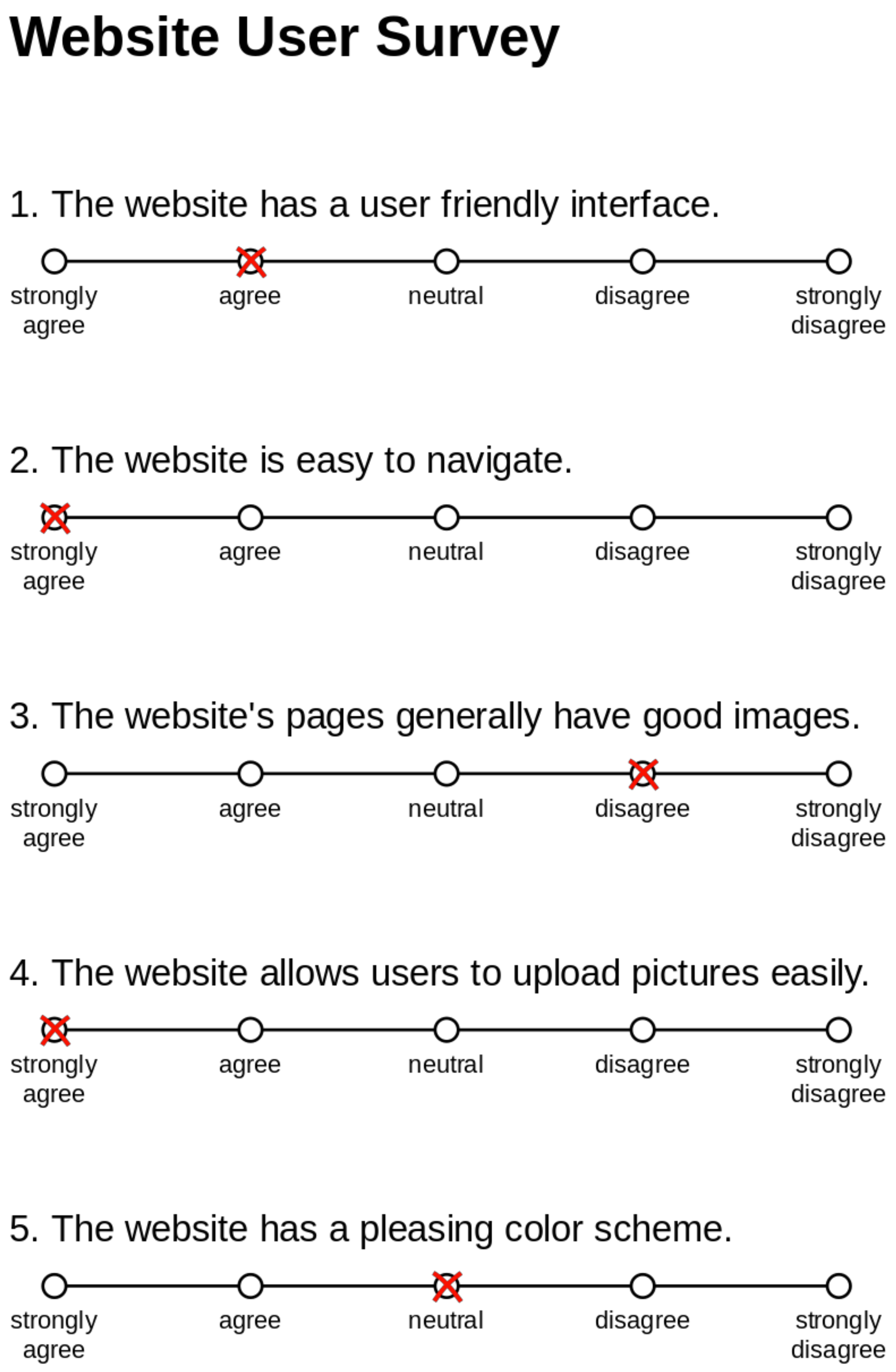


# The Likert Scale

- WMT 06’ - WMT 07’
- Rank based on:

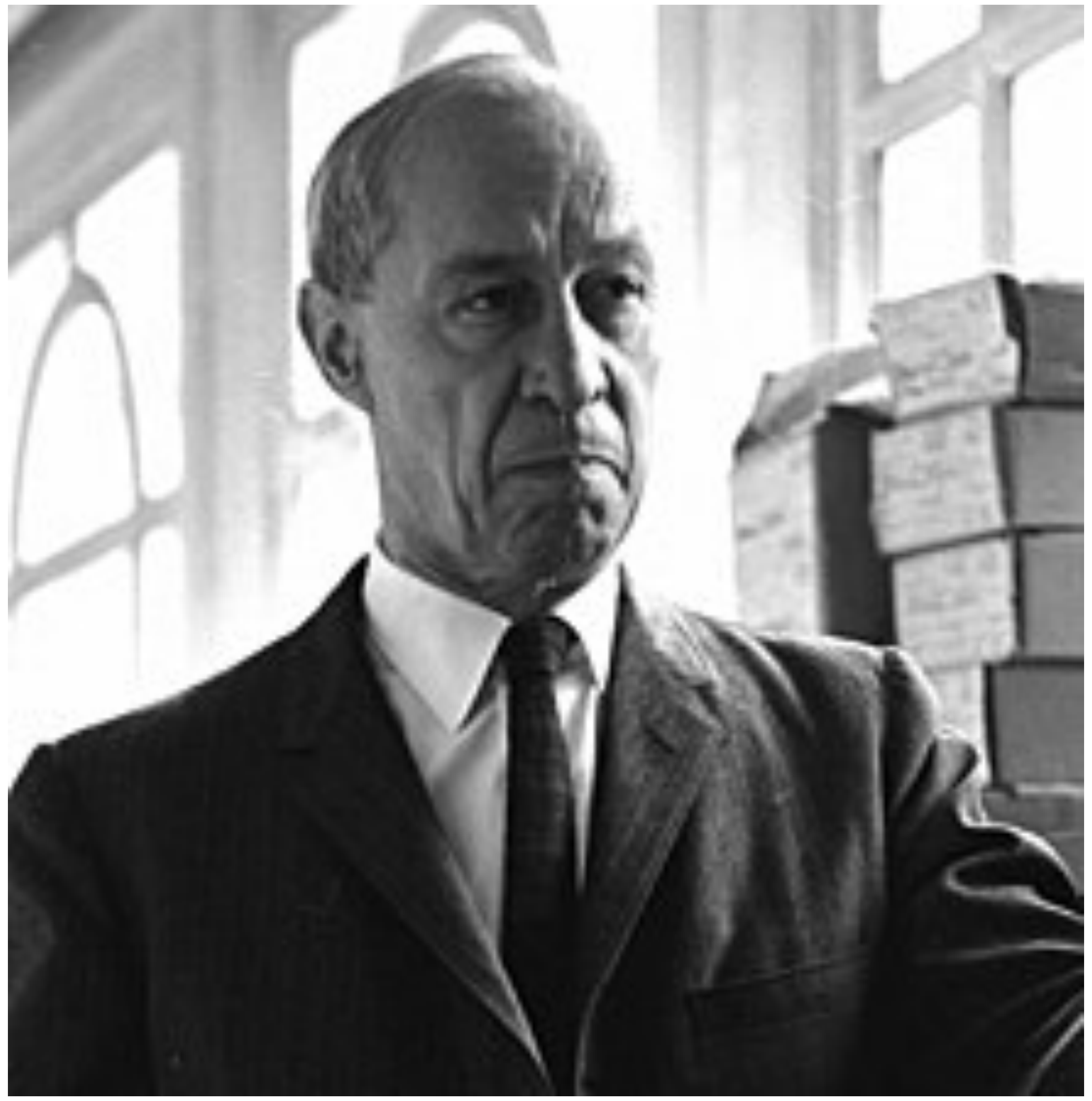


Rensis Likert

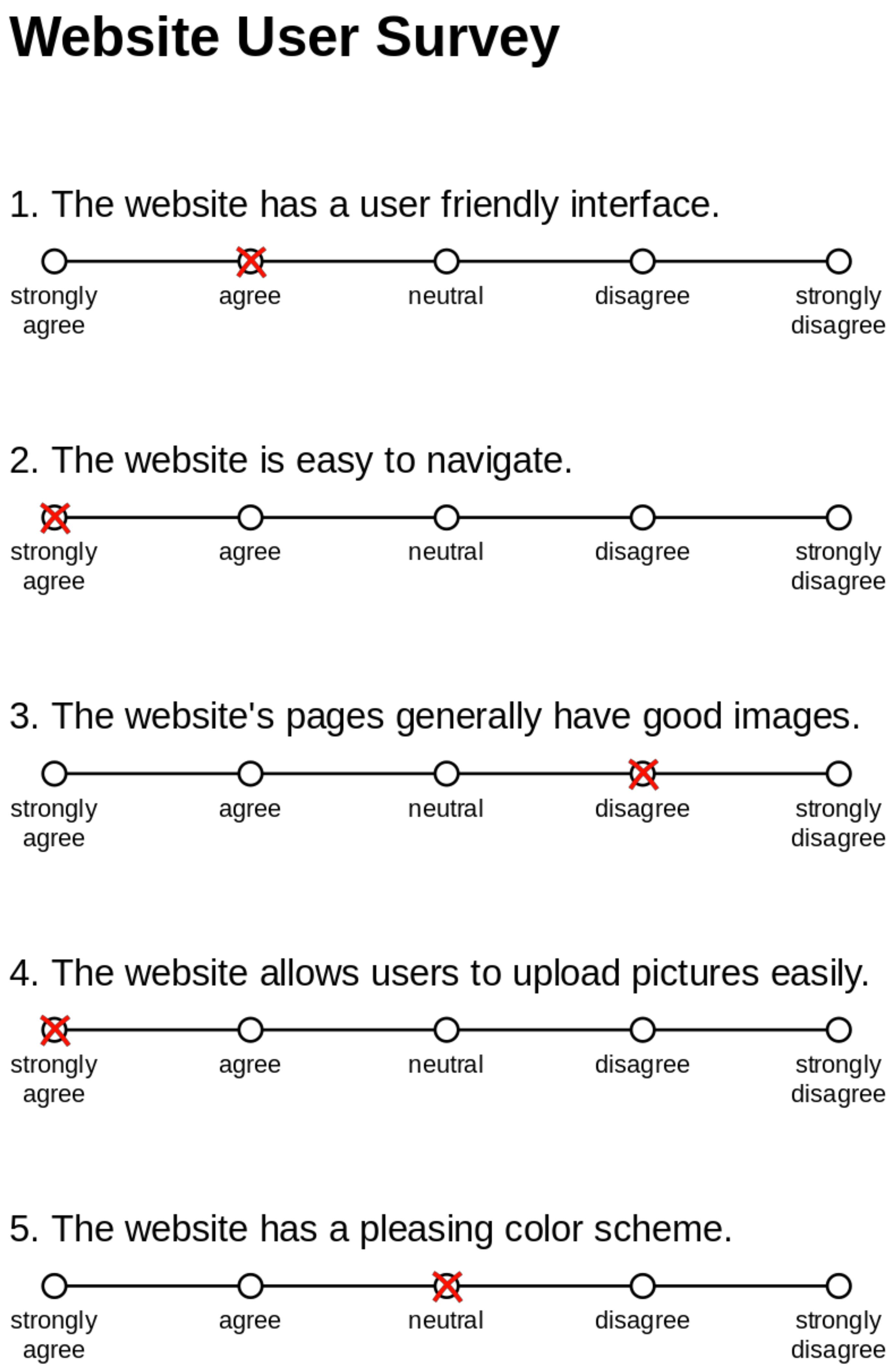


# The Likert Scale

- WMT 06’ - WMT 07’
- Rank based on:
    - **Adequacy** (“how much of the meaning expressed in the reference is also expressed in a hypothesis”?)

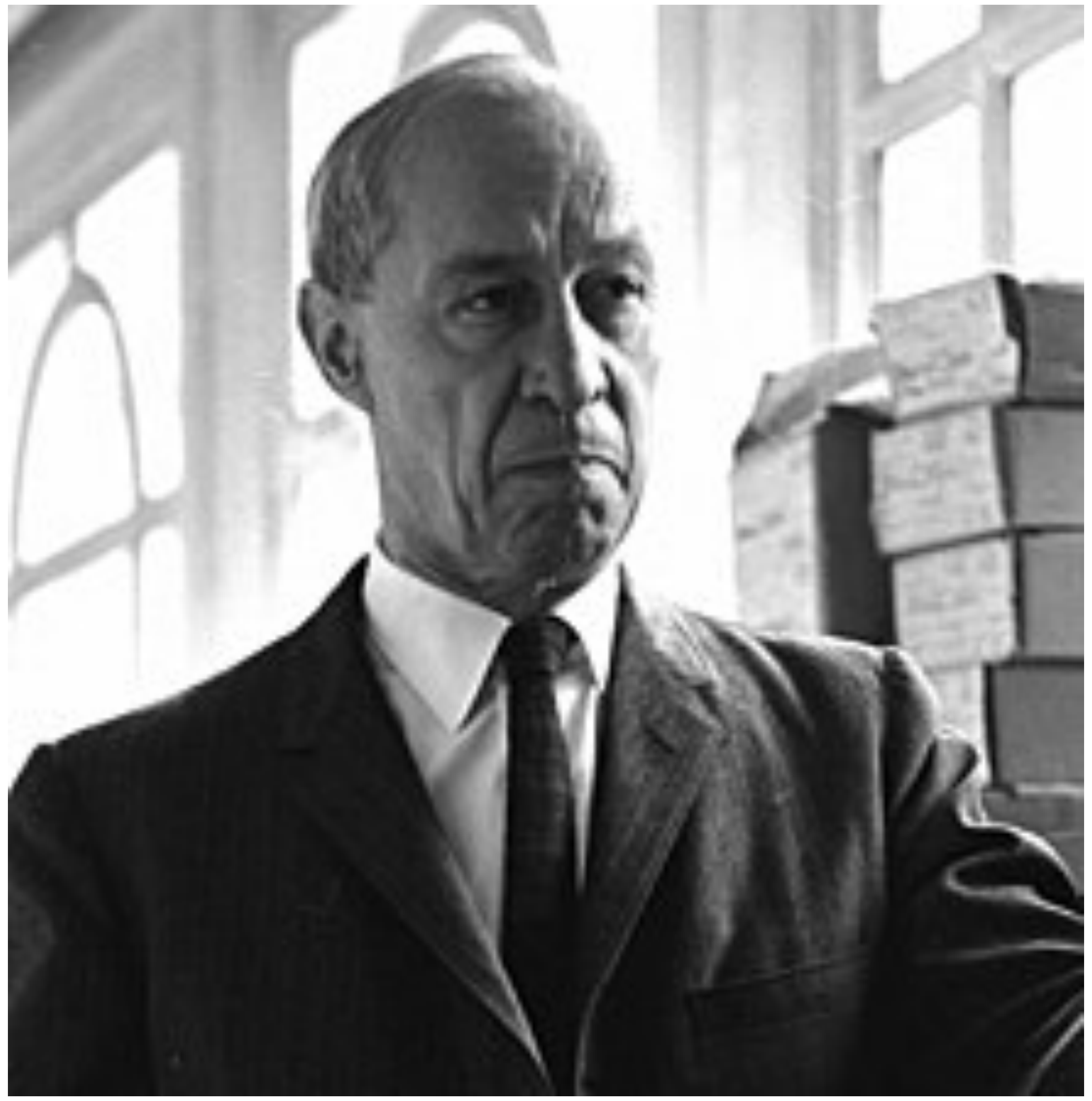


Rensis Likert

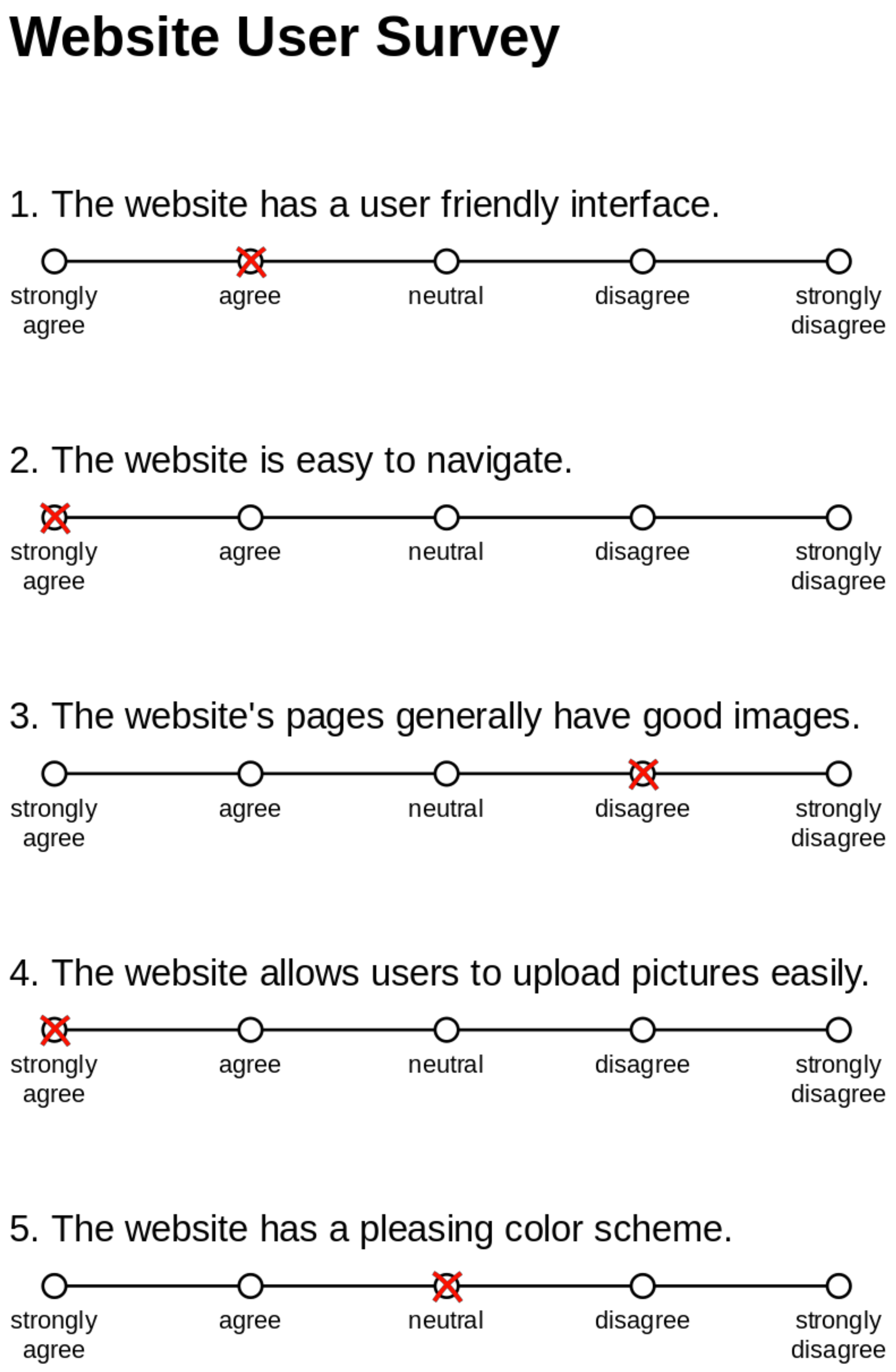


# The Likert Scale

- WMT 06’ - WMT 07’
- Rank based on:
  - **Adequacy** (“how much of the meaning expressed in the reference is also expressed in a hypothesis”?)
  - **Fluency** (“how fluent the translation is?”)

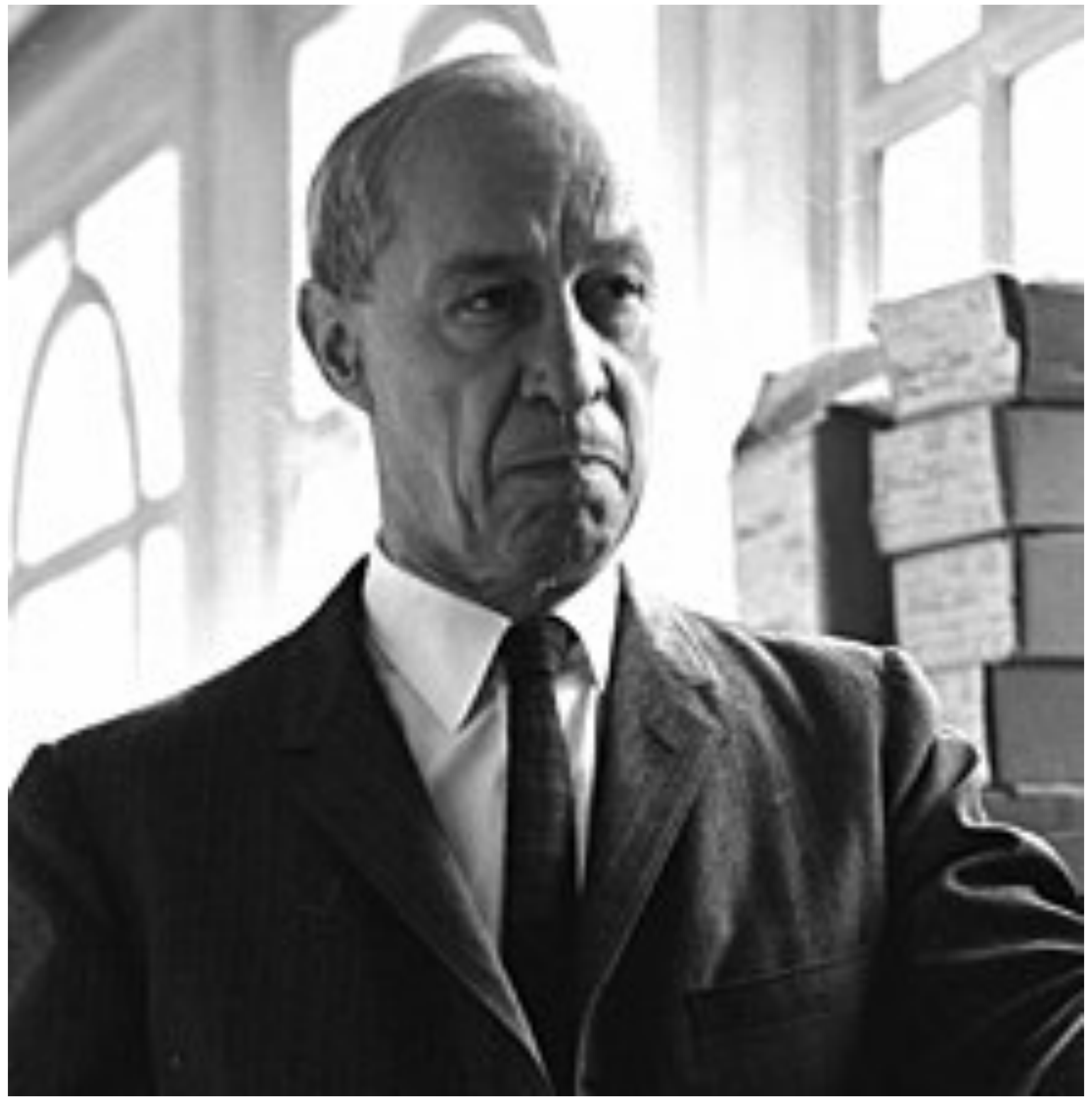


Rensis Likert

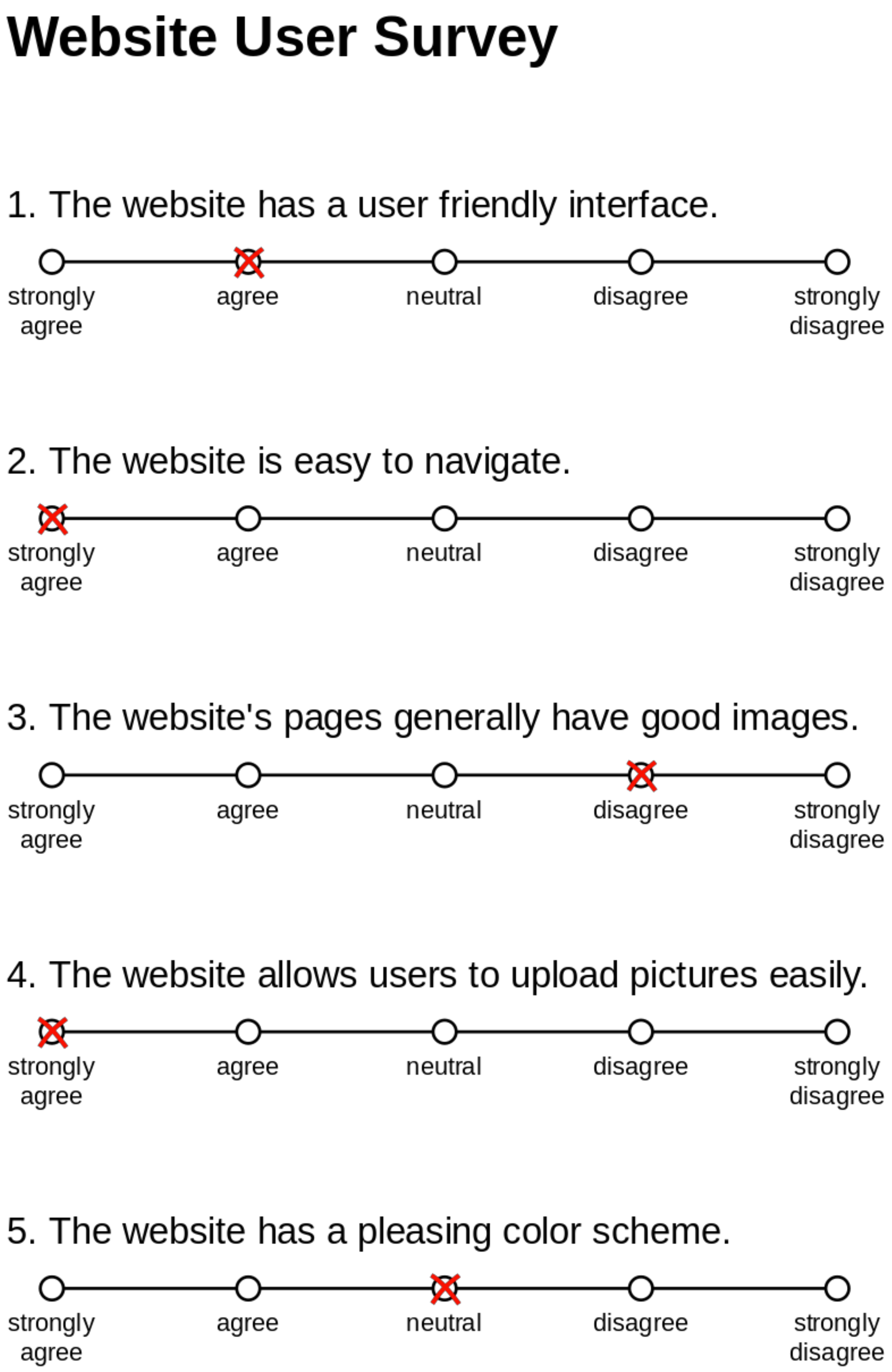


# The Likert Scale

- WMT 06’ - WMT 07’
- Rank based on:
  - **Adequacy** (“how much of the meaning expressed in the reference is also expressed in a hypothesis”?)
  - **Fluency** (“how fluent the translation is?”)
- Pros? Cons?

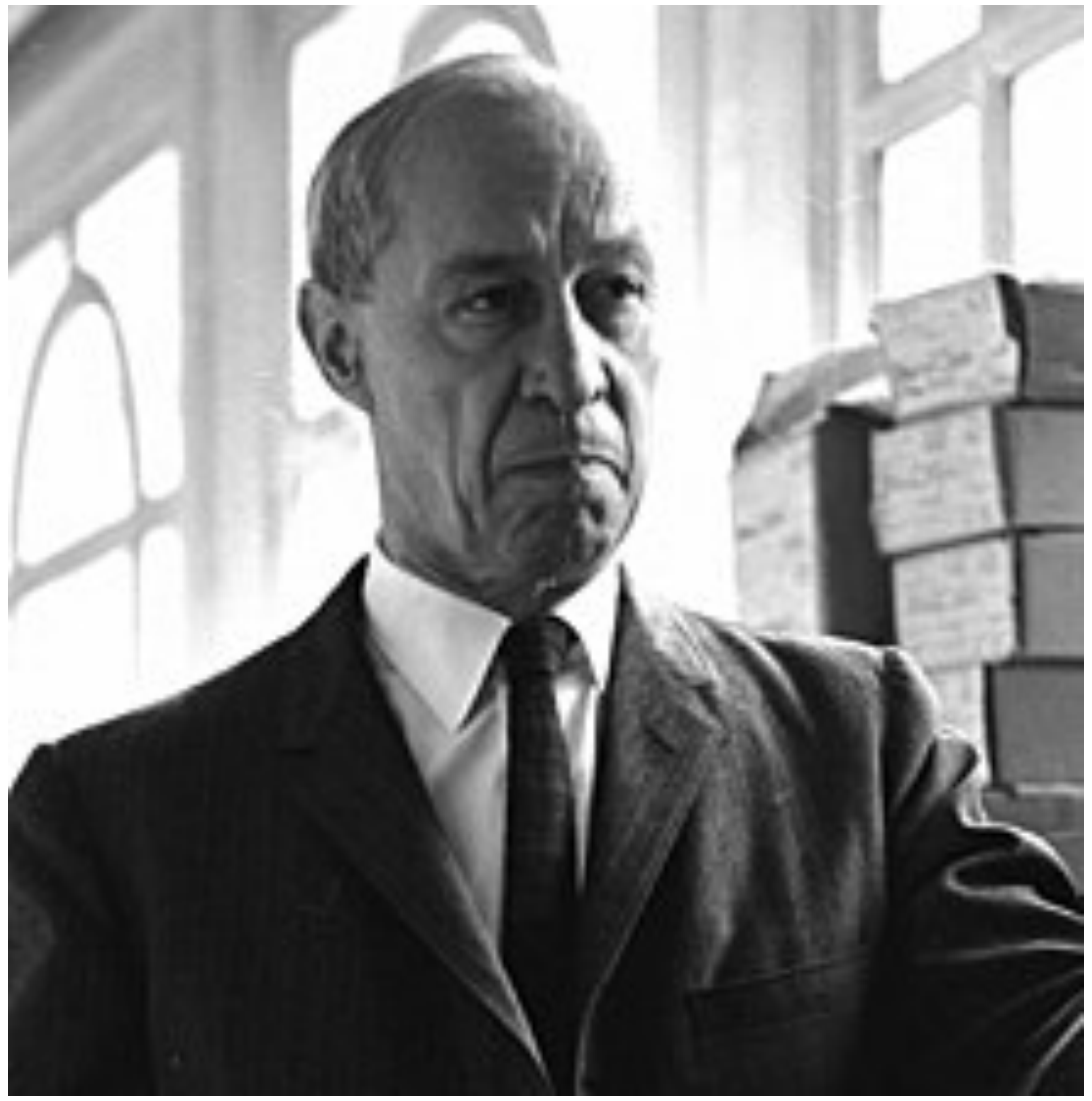


Rensis Likert

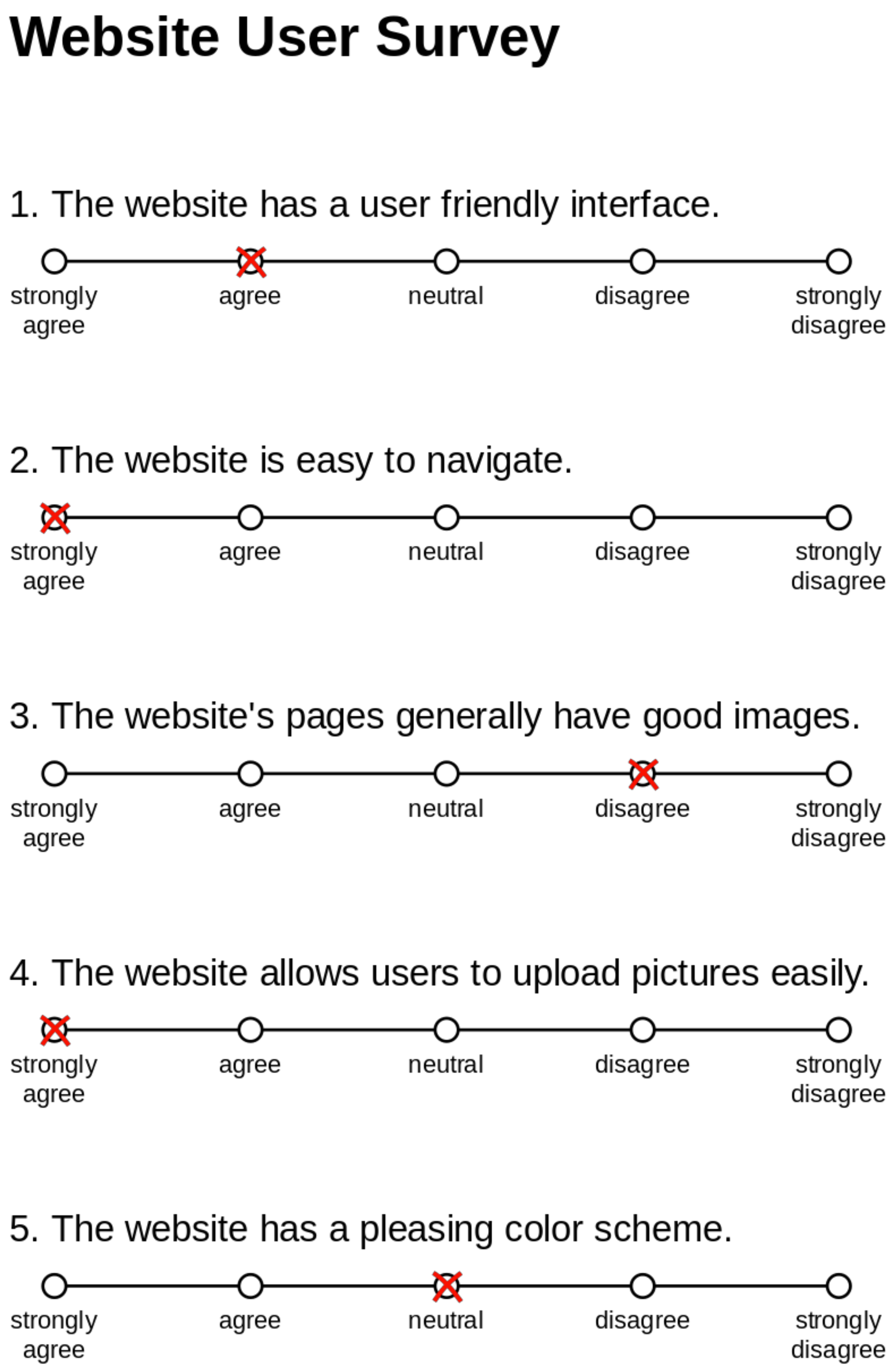


# The Likert Scale

- WMT 06’ - WMT 07’
- Rank based on:
  - **Adequacy** (“how much of the meaning expressed in the reference is also expressed in a hypothesis”?)
  - **Fluency** (“how fluent the translation is?”)
- Pros? Cons?

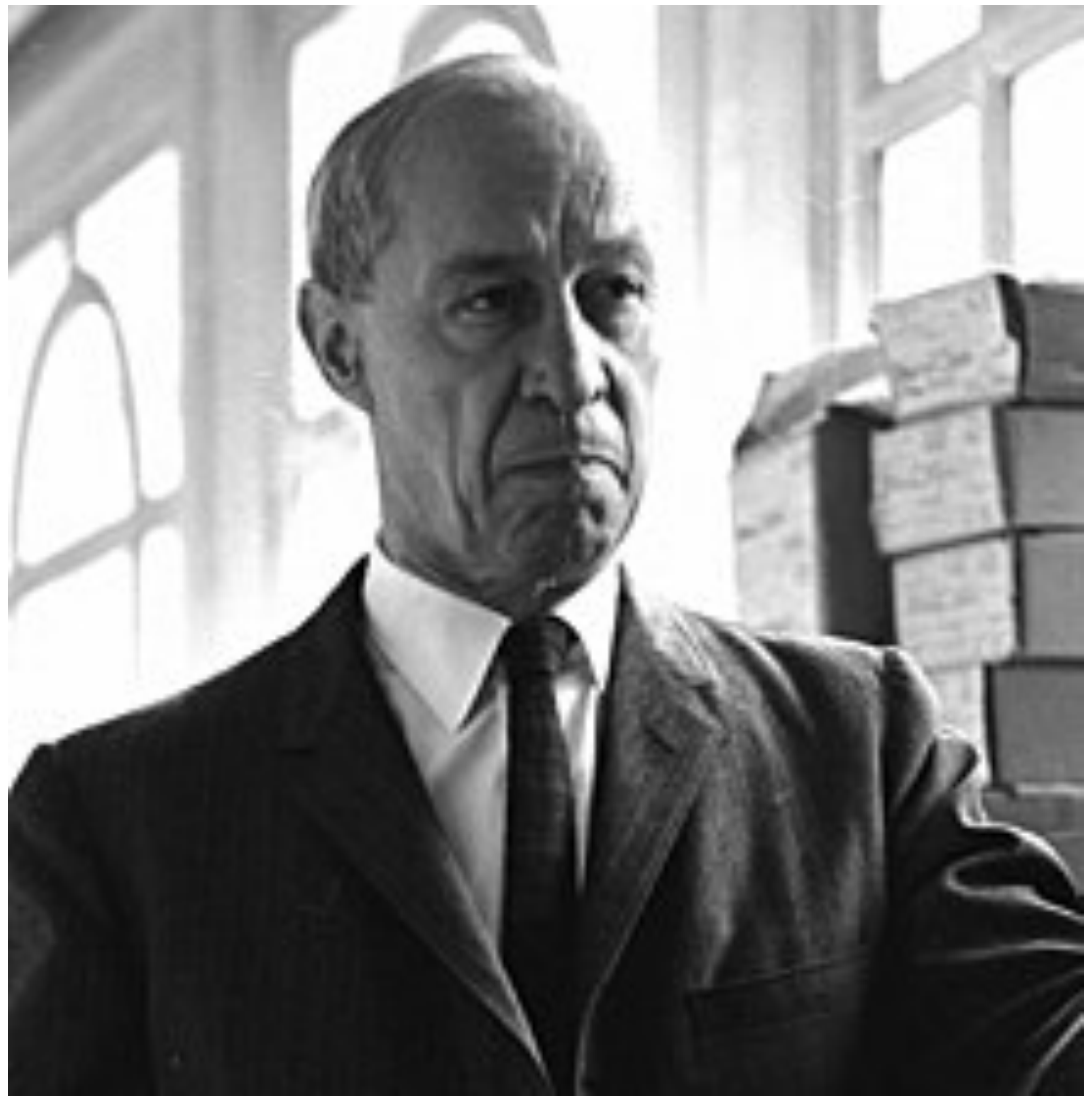


Rensis Likert

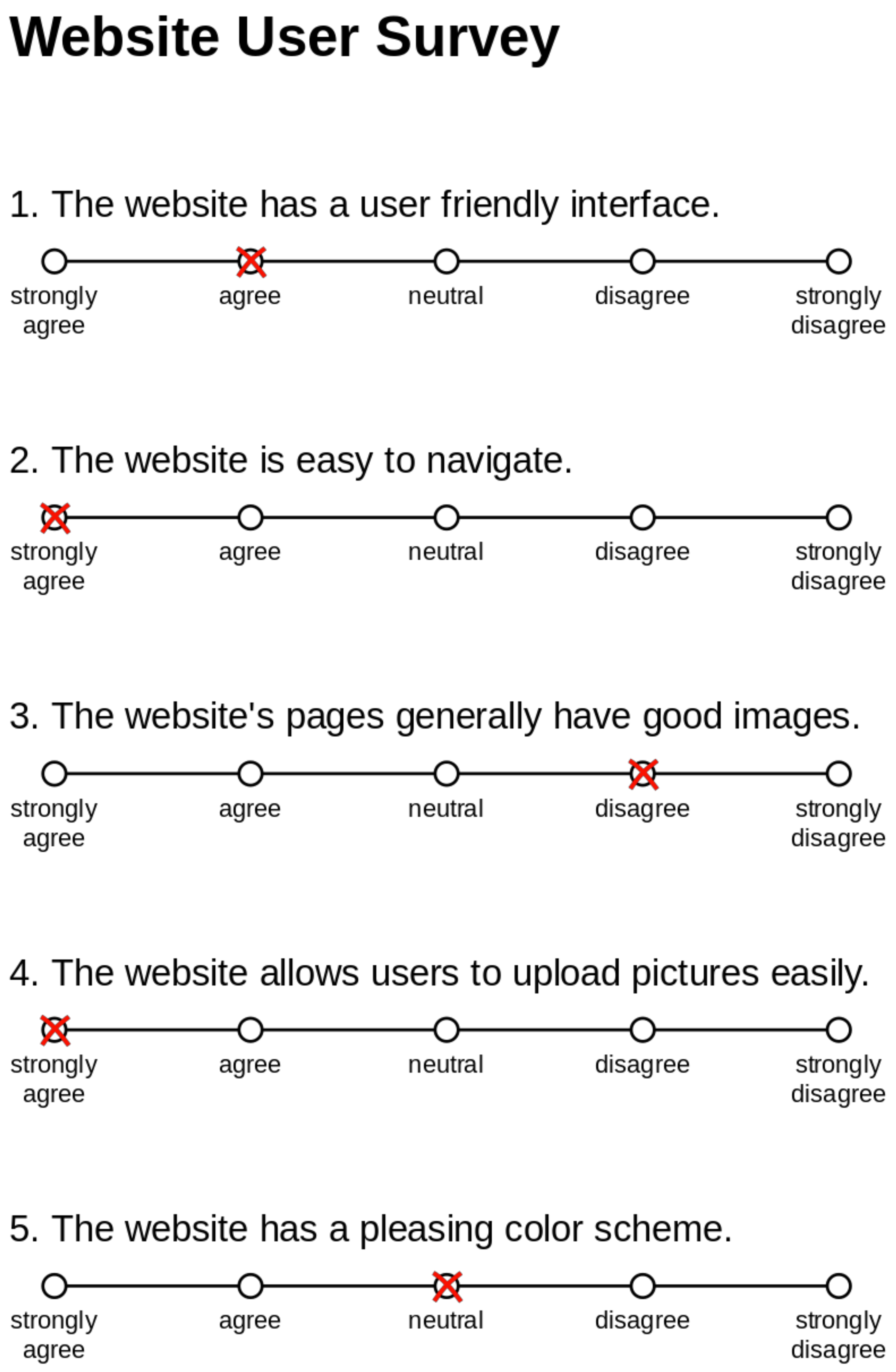


# The Likert Scale

- WMT 06’ - WMT 07’
- Rank based on:
  - **Adequacy** (“how much of the meaning expressed in the reference is also expressed in a hypothesis”?)
  - **Fluency** (“how fluent the translation is?”)
- Pros? Cons?



Rensis Likert



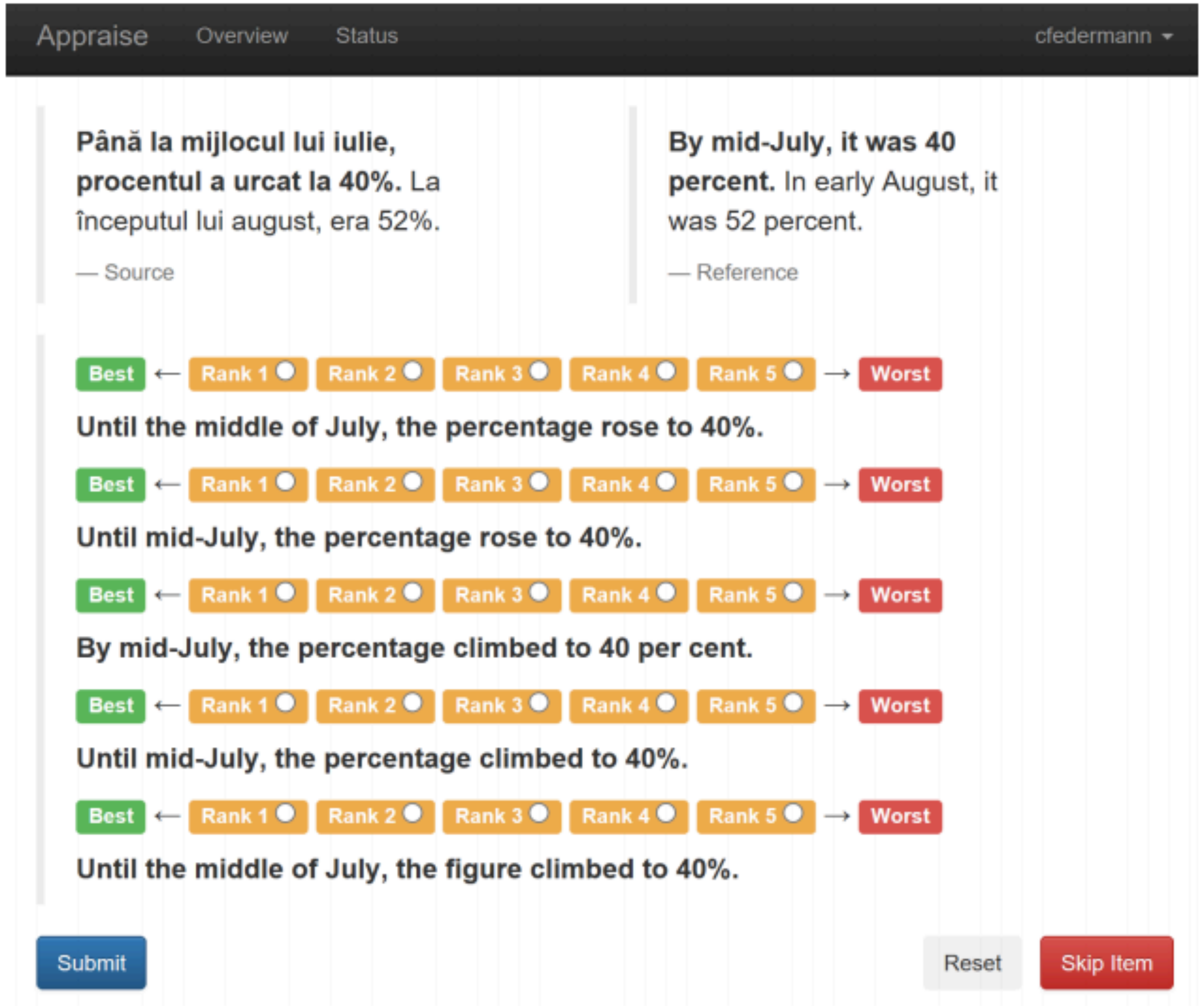
# Relative Ranking

# Relative Ranking

- WMT 07’-WMT 16’

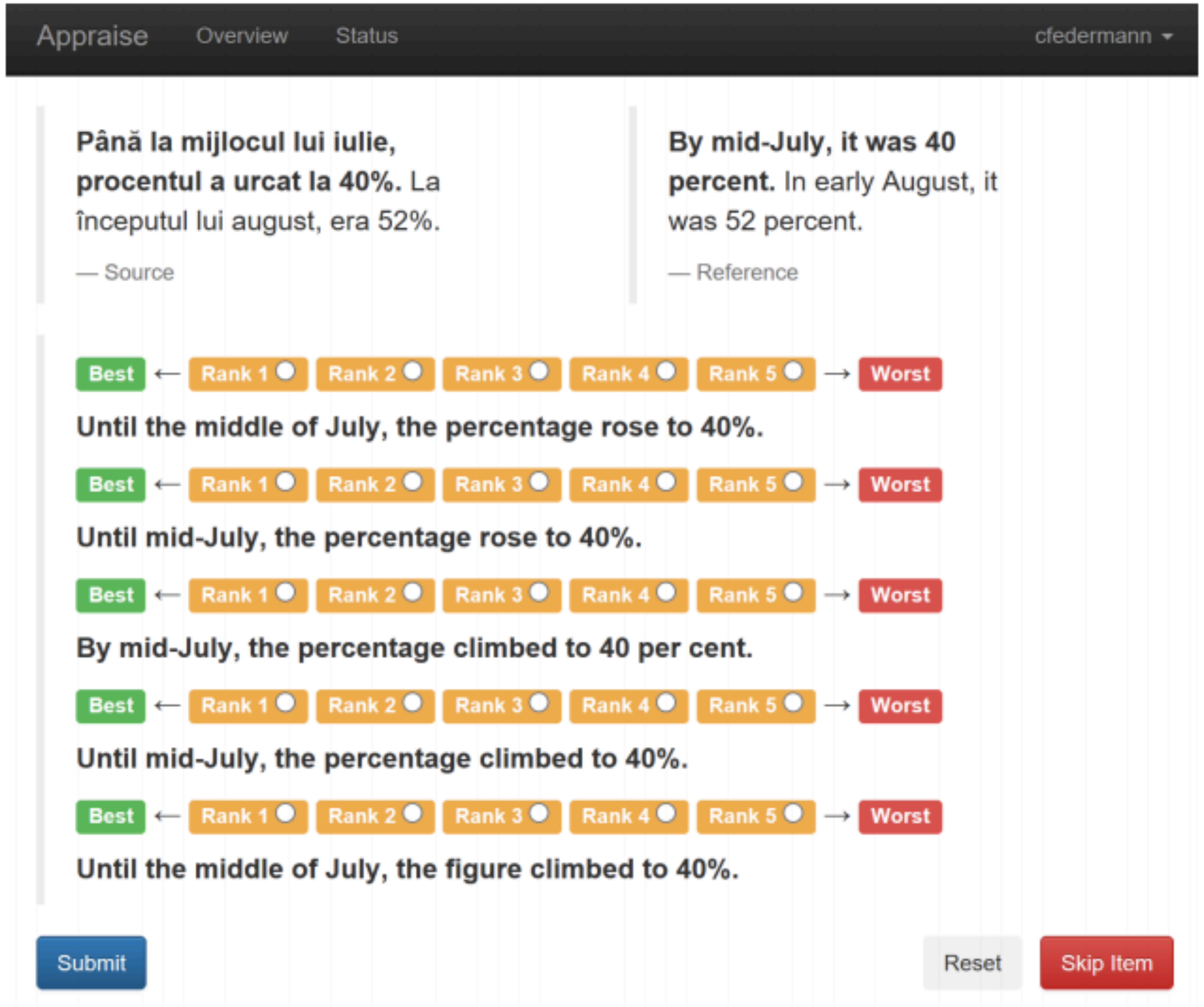
# Relative Ranking

- WMT 07'-WMT 16'
- Each rater ranks 5 systems



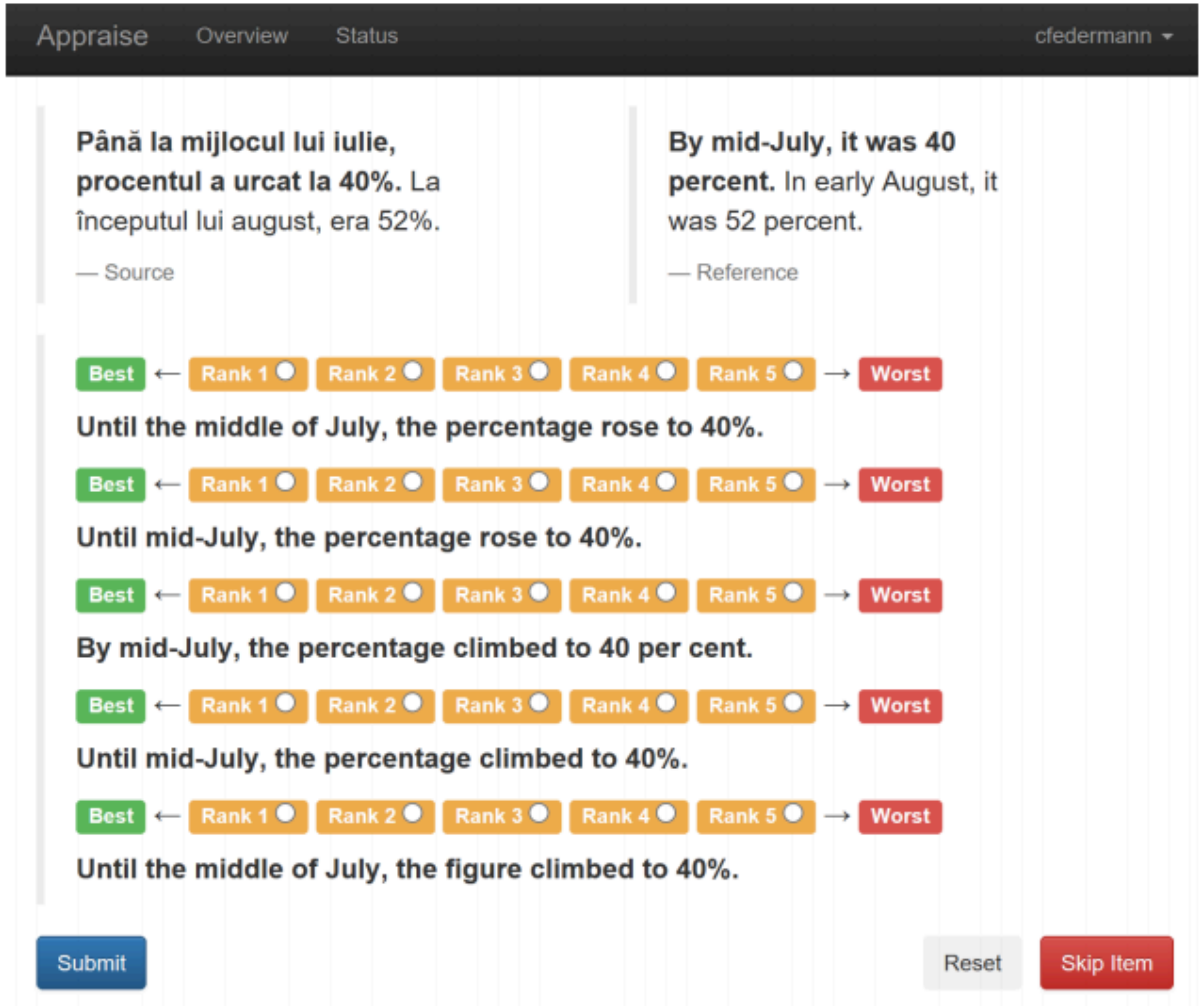
# Relative Ranking

- WMT 07'-WMT 16'
- Each rater ranks 5 systems
- Produces pairwise rankings



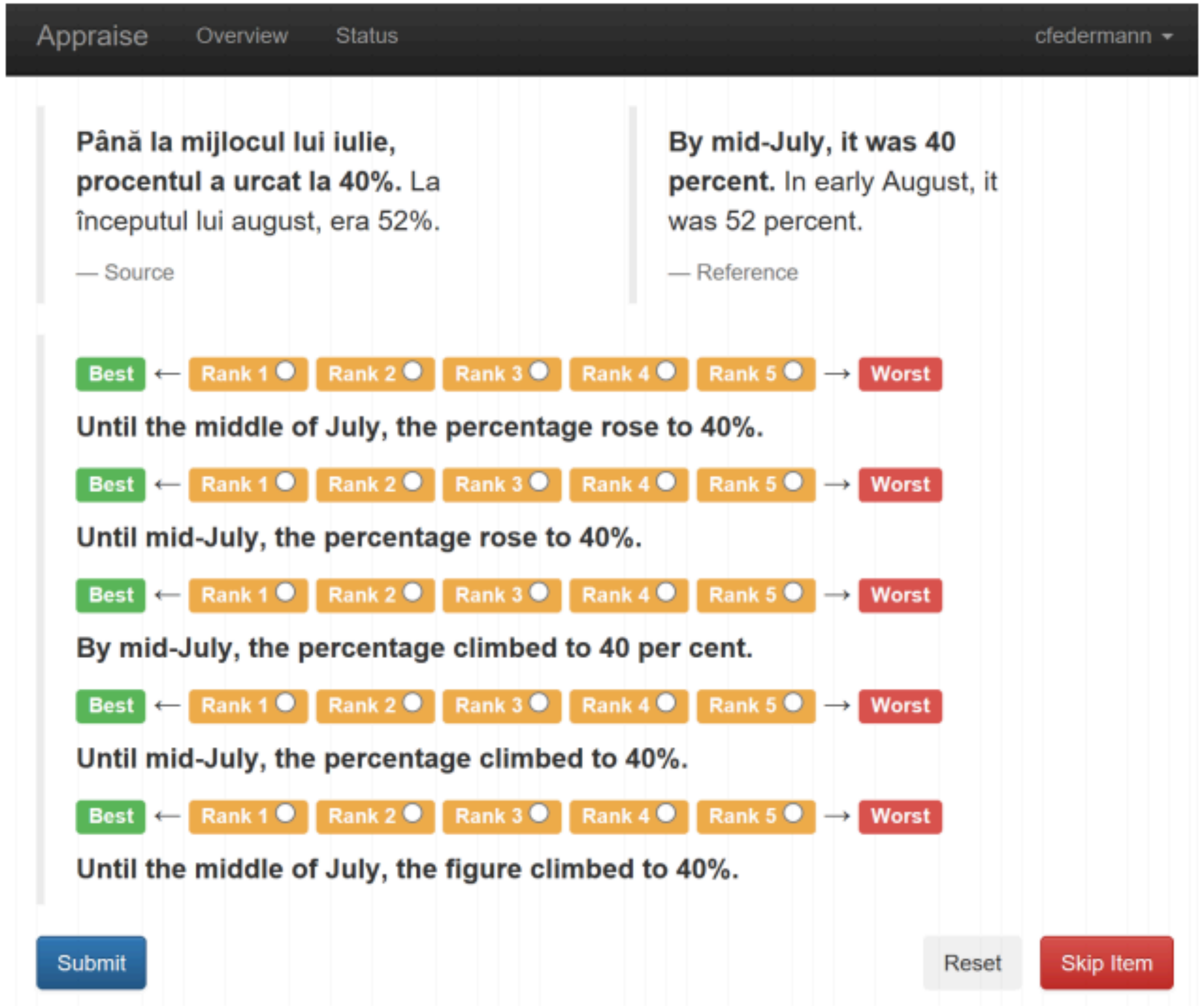
# Relative Ranking

- WMT 07'-WMT 16'
- Each rater ranks 5 systems
- Produces pairwise rankings
- Feed to the TrueSkill (or other) algorithm to obtain final rankings



# Relative Ranking

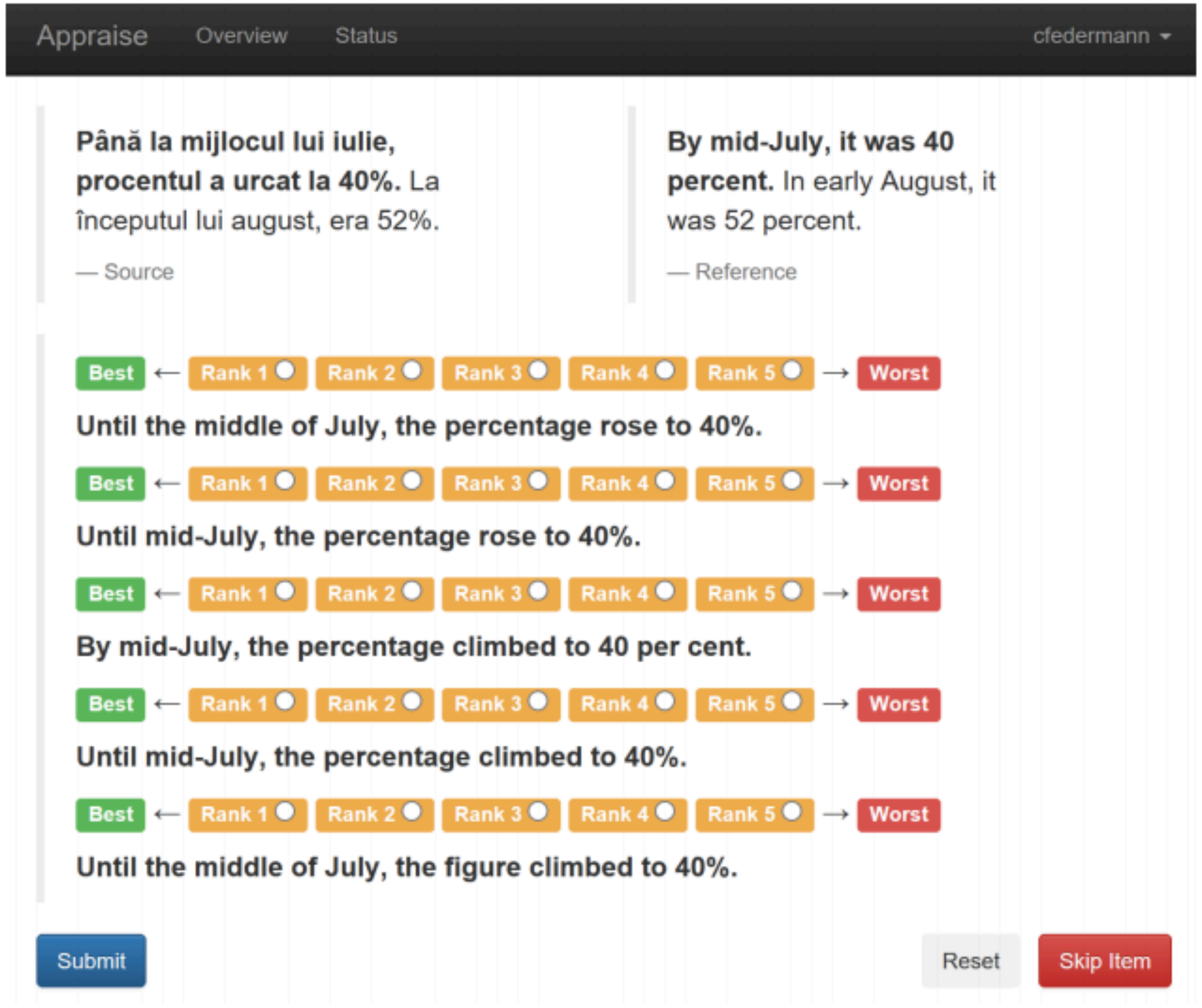
- WMT 07'-WMT 16'
- Each rater ranks 5 systems
- Produces pairwise rankings
- Feed to the TrueSkill (or other) algorithm to obtain final rankings
- Pros? Cons?



# Relative Ranking

- WMT 07’-WMT 16’
- Each rater ranks 5 systems
- Produces pairwise rankings
- Feed to the TrueSkill (or other) algorithm to obtain final rankings
- Pros? Cons?

XBOX LIVE



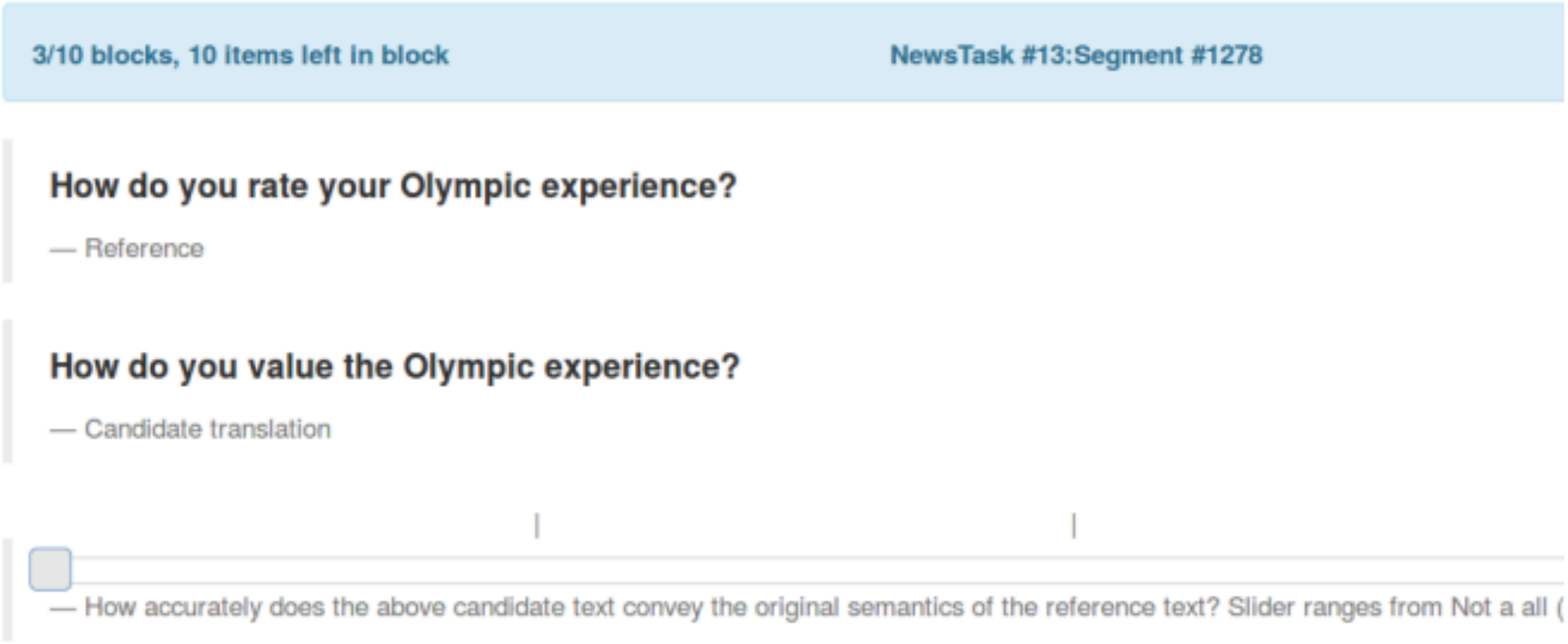
# Direct Assessment: Monolingual

# Direct Assessment: Monolingual

- WMT 16’-WMT 19’

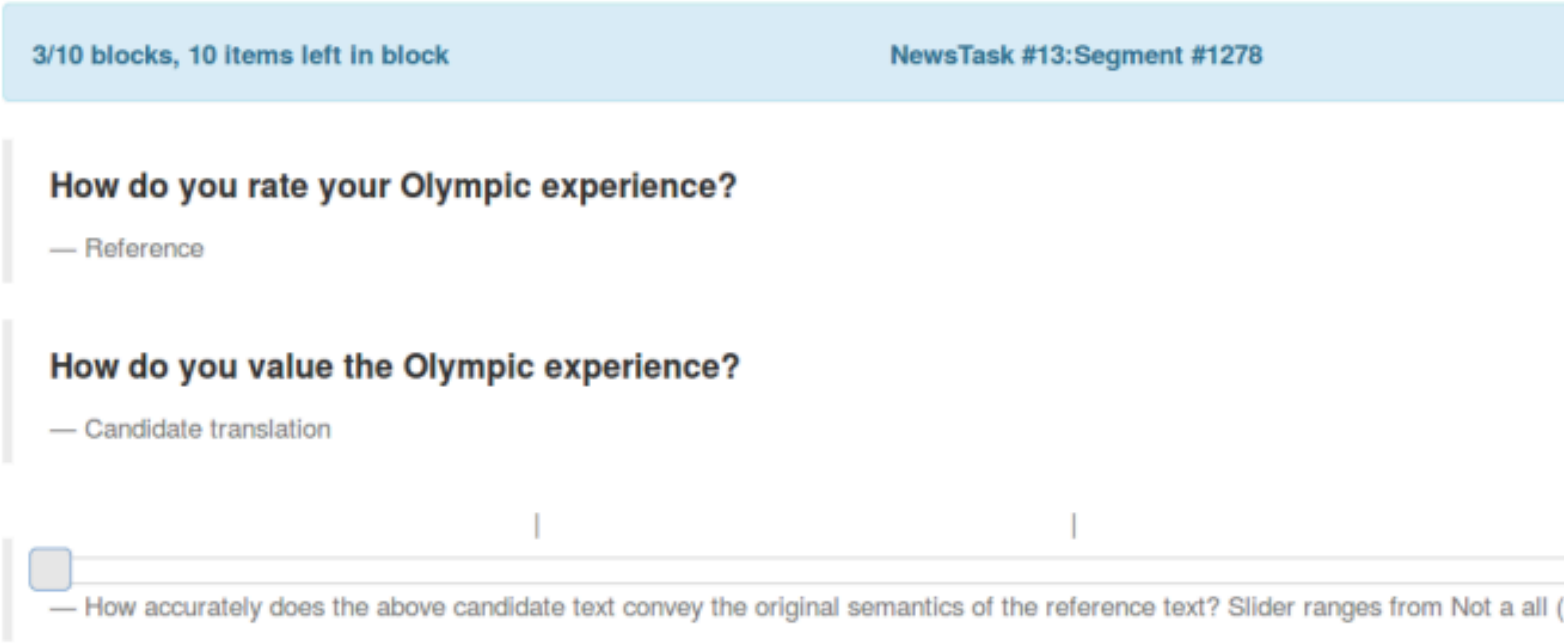
# Direct Assessment: Monolingual

- WMT 16'-WMT 19'



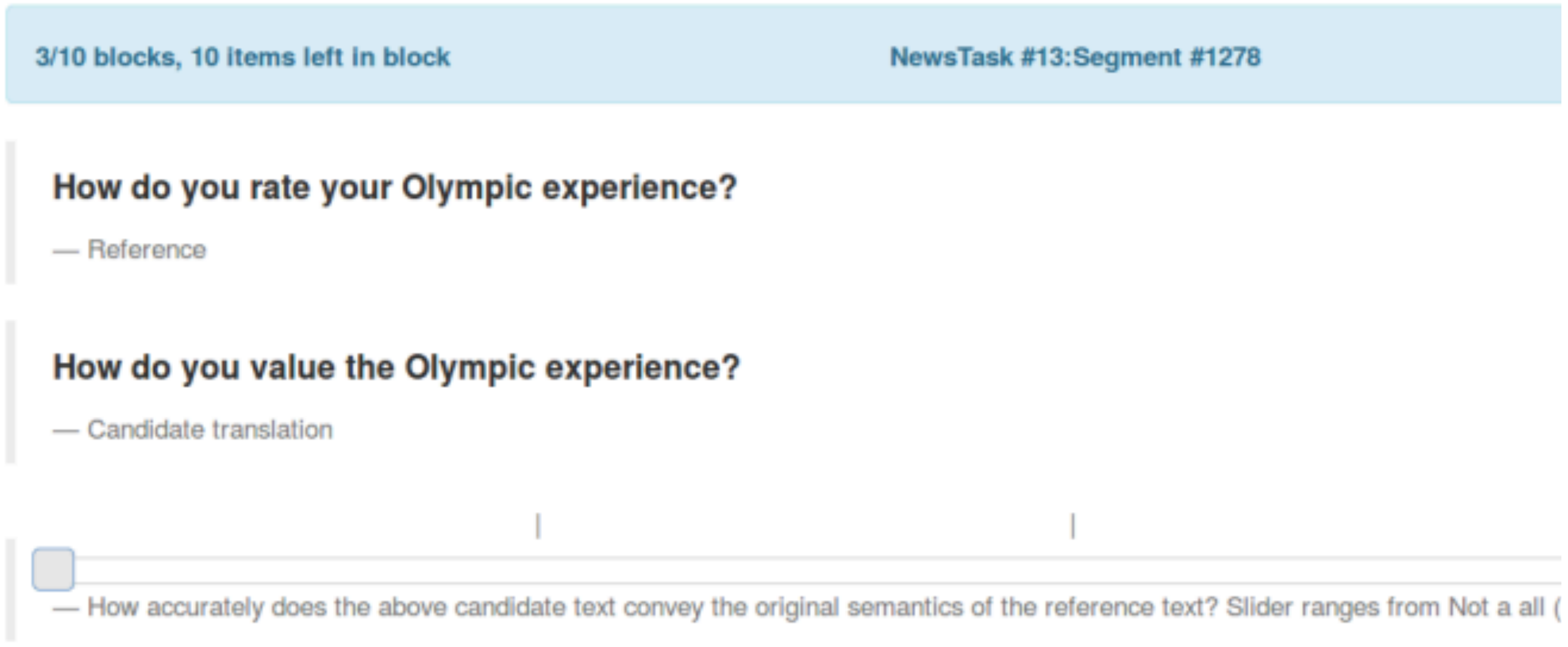
# Direct Assessment: Monolingual

- WMT 16’-WMT 19’
- Scores are standardised according to each individual worker’s overall mean and standard deviation



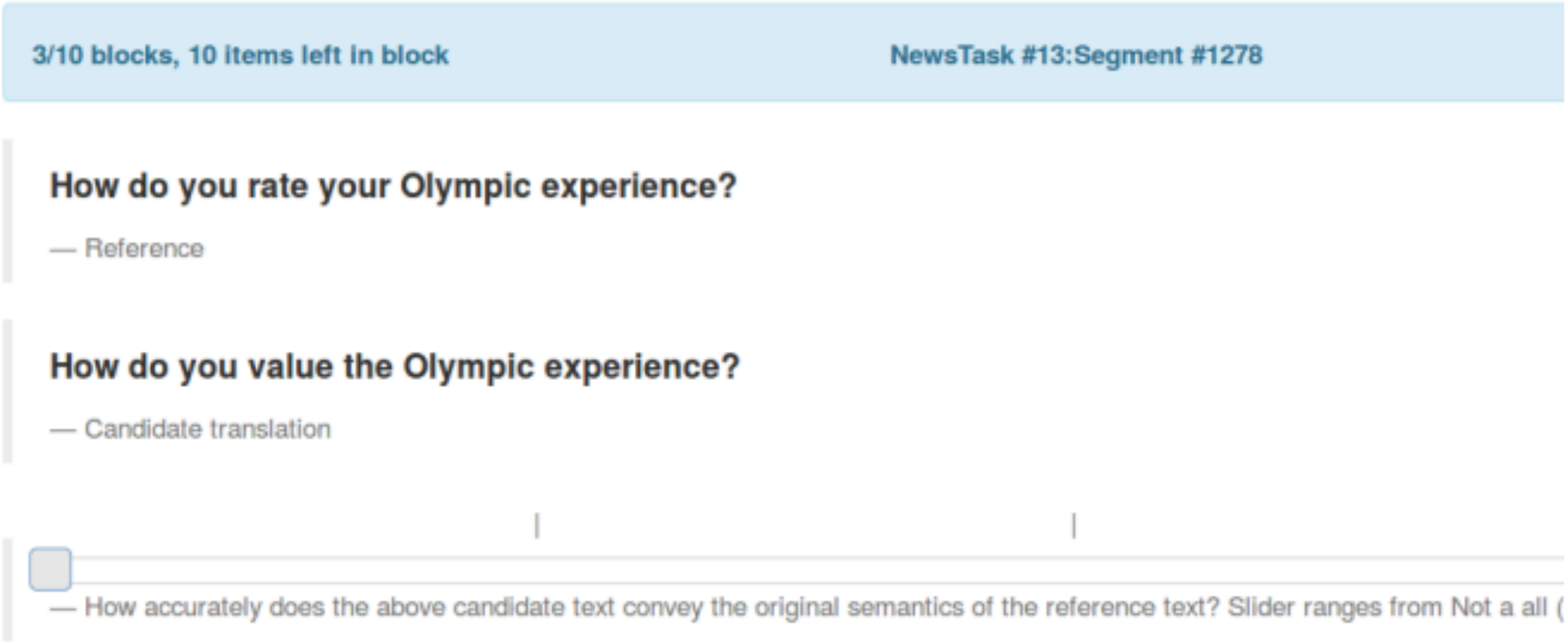
# Direct Assessment: Monolingual

- WMT 16'-WMT 19'
- Scores are standardised according to each individual worker's overall mean and standard deviation
- the overall score of a system is the mean (standardised) score of its translations



# Direct Assessment: Monolingual

- WMT 16’-WMT 19’
- Scores are standardised according to each individual worker’s overall mean and standard deviation
- the overall score of a system is the mean (standardised) score of its translations
- Adequacy is main, fluency used to break ties



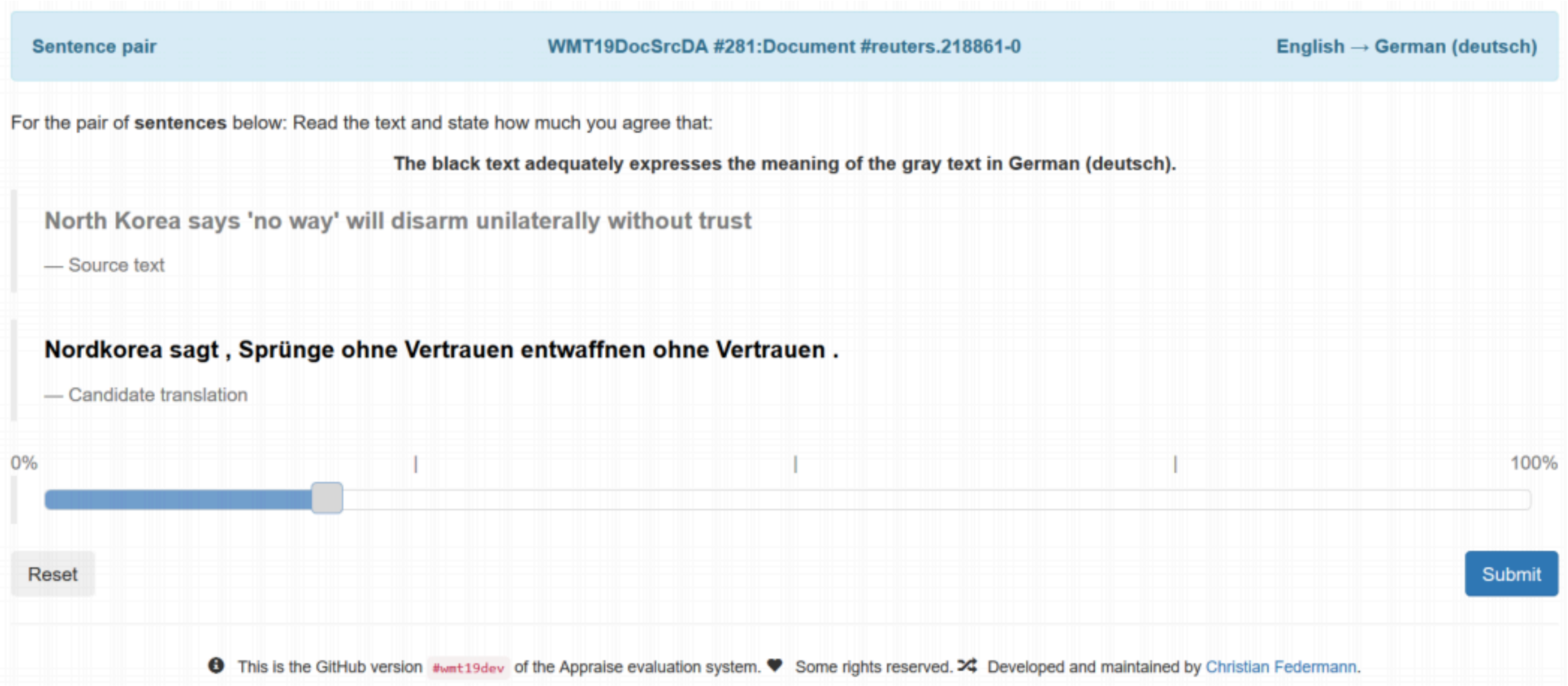
# Direct Assessment: Bilingual

# Direct Assessment: Bilingual

- WMT 18’-WMT 19’

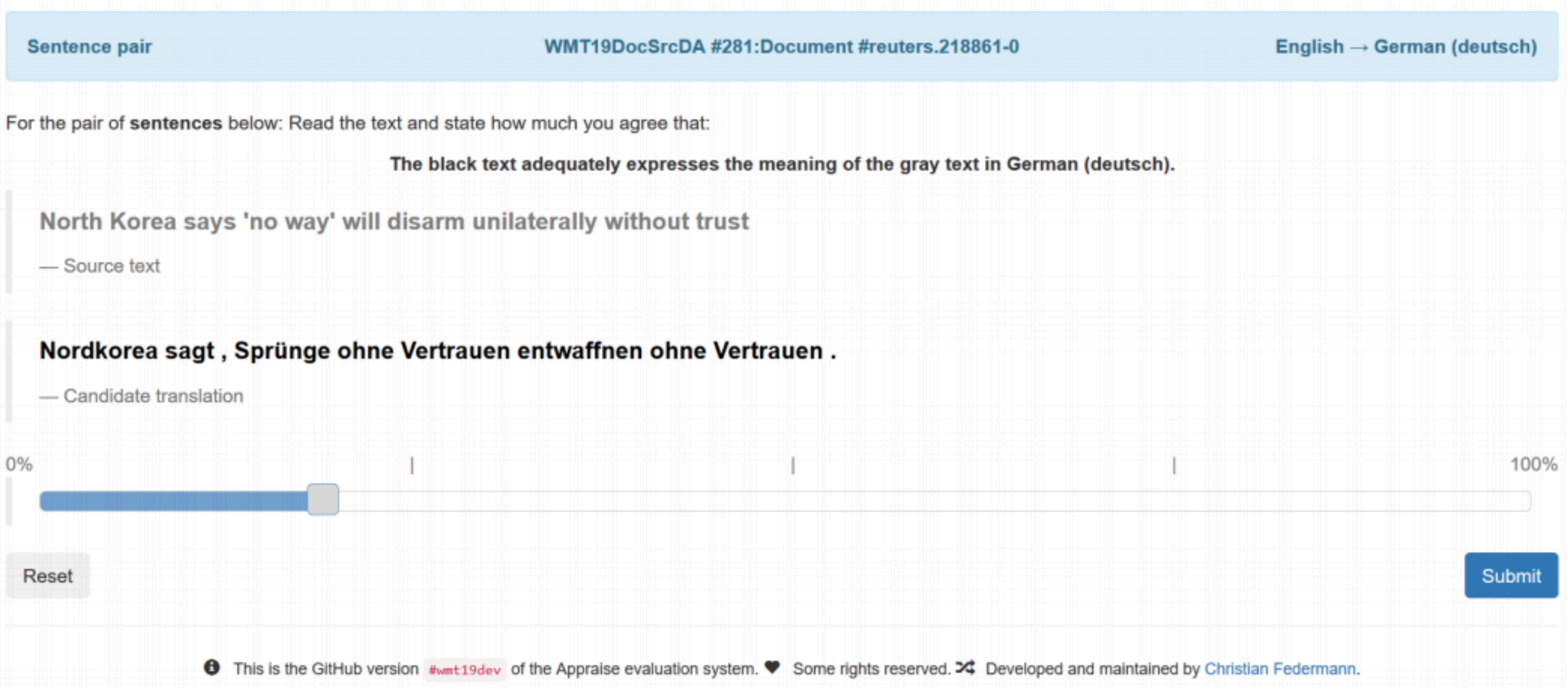
# Direct Assessment: Bilingual

- WMT 18'-WMT 19'
- Use a source sentence instead of a reference sentence ("source-based")



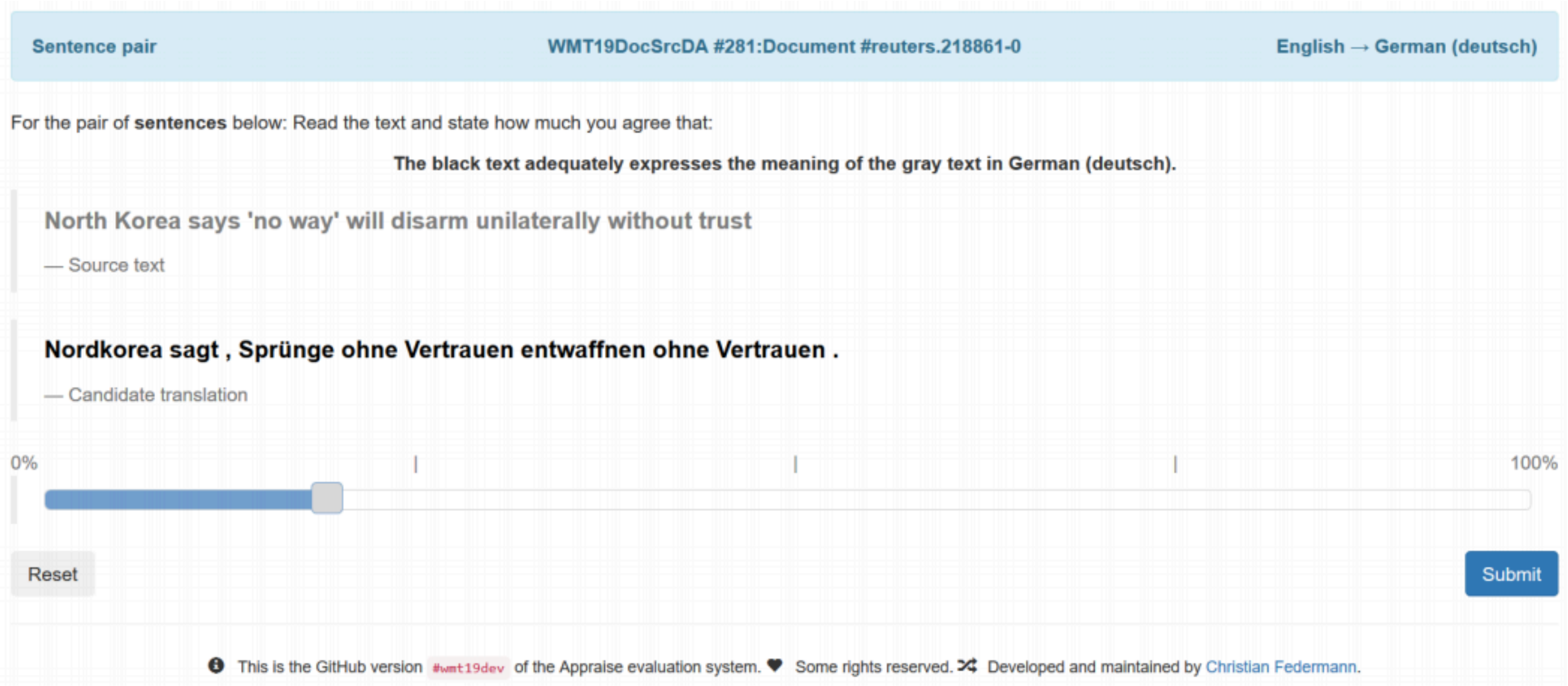
# Direct Assessment: Bilingual

- WMT 18'-WMT 19'
- Use a source sentence instead of a reference sentence ("source-based")
- Main motivation: enables to measure "human performance"



# Direct Assessment: Bilingual

- WMT 18’-WMT 19’
- Use a source sentence instead of a reference sentence (“source-based”)
- Main motivation: enables to measure “human performance”
- **Are we there yet?**



# Human Parity in MT

# Human Parity in MT

- Several papers previously claimed some some sort of human-parity

# Human Parity in MT

- Several papers previously claimed some some sort of human-parity

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi
`yonghui,schuster,zhifengc,qvl,mnorouzi@google.com`

Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey,
Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser,
Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens,
George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa,
Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean

# Human Parity in MT

- Several papers previously claimed some some sort of human-parity

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi
yonghui,schuster,zhifengc,qvl,mnorouzi@google.com

Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean

Achieving Human Parity on Automatic Chinese to English News Translation

Hany Hassan*, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou

Microsoft AI & Research

# Human Parity in MT

- Several papers previously claimed some some sort of human-parity
- WMT 18’ also presented such result for English - Czech

Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi
yonghui,schuster,zhifengc,qvl,mnorouzi@google.com
Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean

Achieving Human Parity on Automatic Chinese to English News Translation

Hany Hassan*, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou

**English→Czech**

| | Ave. % | Ave. z | System |
|---|---|---|---|
| 1 | 84.4 | 0.667 | CUNI-TRANSFORMER |
| 2 | 79.8 | 0.521 | UEDIN |
| | 78.6 | 0.483 | NEWSTEST2018-REF |
| 4 | 68.1 | 0.128 | ONLINE-B |
| 5 | 59.4 | −0.178 | ONLINE-A |
| 6 | 54.1 | −0.354 | ONLINE-G |

# Human Parity in MT

- Several papers previously claimed some some sort of human-parity
- WMT 18' also presented such result for English - Czech
- Possible caveats:

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi
yonghui,schuster,zhifengc,qvl,mnorouzi@google.com
Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean

Achieving Human Parity on Automatic Chinese to English News Translation

Hany Hassan,* Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou

**English→Czech**

| | Ave. % | Ave. z | System |
|---|---|---|---|
| 1 | 84.4 | 0.667 | CUNI-TRANSFORMER |
| 2 | 79.8 | 0.521 | UEDIN |
| | 78.6 | 0.483 | NEWSTEST2018-REF |
| 4 | 68.1 | 0.128 | ONLINE-B |
| 5 | 59.4 | −0.178 | ONLINE-A |
| 6 | 54.1 | −0.354 | ONLINE-G |

# Human Parity in MT

- Several papers previously claimed some some sort of human-parity
- WMT 18’ also presented such result for English - Czech
- Possible caveats:
  - Bad references

Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi
yonghui,schuster,zhifengc,qvl,mnorouzi@google.com
Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean

Achieving Human Parity on Automatic Chinese to English News Translation

Hany Hassan*, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou

**English→Czech**

| | Ave. % | Ave. z | System |
|---|---|---|---|
| 1 | 84.4 | 0.667 | CUNI-TRANSFORMER |
| 2 | 79.8 | 0.521 | UEDIN |
| | 78.6 | 0.483 | NEWSTEST2018-REF |
| 4 | 68.1 | 0.128 | ONLINE-B |
| 5 | 59.4 | −0.178 | ONLINE-A |
| 6 | 54.1 | −0.354 | ONLINE-G |

# Human Parity in MT

- Several papers previously claimed some some sort of human-parity
- WMT 18' also presented such result for English - Czech
- Possible caveats:
  - Bad references
  - Incompetent raters

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi
yonghui,schuster,zhifengc,qvl,mnorouzi@google.com

Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean

Achieving Human Parity on Automatic Chinese to English News Translation

Hany Hassan*, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou

**English→Czech**

| | Ave. % | Ave. z | System |
|---|---|---|---|
| 1 | 84.4 | 0.667 | CUNI-TRANSFORMER |
| 2 | 79.8 | 0.521 | UEDIN |
| | 78.6 | 0.483 | NEWSTEST2018-REF |
| 4 | 68.1 | 0.128 | ONLINE-B |
| 5 | 59.4 | −0.178 | ONLINE-A |
| 6 | 54.1 | −0.354 | ONLINE-G |

# Human Parity in MT

- Several papers previously claimed some some sort of human-parity
- WMT 18' also presented such result for English - Czech
- Possible caveats:
  - Bad references
  - Incompetent raters
  - Small sample size

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi
yonghui,schuster,zhifengc,qvl,mnorouzi@google.com

Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean

Achieving Human Parity on Automatic Chinese to English News Translation

Hany Hassan*, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou

**English→Czech**

| | Ave. % | Ave. z | System |
|---|---|---|---|
| 1 | 84.4 | 0.667 | CUNI-TRANSFORMER |
| 2 | 79.8 | 0.521 | UEDIN |
| | 78.6 | 0.483 | NEWSTEST2018-REF |
| 4 | 68.1 | 0.128 | ONLINE-B |
| 5 | 59.4 | −0.178 | ONLINE-A |
| 6 | 54.1 | −0.354 | ONLINE-G |

# Human Parity in MT

- Several papers previously claimed some some sort of human-parity
- WMT 18' also presented such result for English - Czech
- Possible caveats:
    - Bad references
    - Incompetent raters
    - Small sample size
    - Sentence-level evaluation

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi
yonghui,schuster,zhifengc,qvl,mnorouzi@google.com
Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean

Achieving Human Parity on Automatic Chinese to English News Translation

Hany Hassan*, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou

**English→Czech**

| | Ave. % | Ave. z | System |
|---|---|---|---|
| 1 | 84.4 | 0.667 | CUNI-TRANSFORMER |
| 2 | 79.8 | 0.521 | UEDIN |
| | 78.6 | 0.483 | NEWSTEST2018-REF |
| 4 | 68.1 | 0.128 | ONLINE-B |
| 5 | 59.4 | −0.178 | ONLINE-A |
| 6 | 54.1 | −0.354 | ONLINE-G |

# Not so fast…

# Not so fast...

- Several recent studies show how human parity was yet to be achieved:

# Not so fast…

- Several recent studies show how human parity was yet to be achieved:

**Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation**

**Antonio Toral**
Center for Language and Cognition
University of Groningen
The Netherlands
a.toral.ruiz@rug.nl

**Sheila Castilho** **Ke Hu** **Andy Way**
ADAPT Centre
Dublin City University
Ireland
firstname.secondname@adaptcentre.ie

# Not so fast…

- Several recent studies show how human parity was yet to be achieved:

**Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation**

**Antonio Toral**
Center for Language and Cognition
University of Groningen
The Netherlands
a.toral.ruiz@rug.nl

**Sheila Castilho** **Ke Hu** **Andy Way**
ADAPT Centre
Dublin City University
Ireland
firstname.secondname@adaptcentre.ie

**Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation**

**Samuel Läubli**[1] **Rico Sennrich**[1,2] **Martin Volk**[1]

# Not so fast…

- Several recent studies show how human parity was yet to be achieved:
  - The “Super-Human” sentence-level systems are inferior to humans when evaluated in document-level

**Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation**

**Antonio Toral**
Center for Language and Cognition
University of Groningen
The Netherlands
a.toral.ruiz@rug.nl

**Sheila Castilho** **Ke Hu** **Andy Way**
ADAPT Centre
Dublin City University
Ireland
firstname.secondname@adaptcentre.ie

**Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation**

**Samuel Läubli**[1] **Rico Sennrich**[1,2] **Martin Volk**[1]

# Not so fast…

- Several recent studies show how human parity was yet to be achieved:
  - The “Super-Human” sentence-level systems are inferior to humans when evaluated in document-level
  - The translation direction when producing the references is crucial (“Translationese”)

**Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation**

**Antonio Toral**
Center for Language and Cognition
University of Groningen
The Netherlands
a.toral.ruiz@rug.nl

**Sheila Castilho** **Ke Hu** **Andy Way**
ADAPT Centre
Dublin City University
Ireland
firstname.secondname@adaptcentre.ie

**Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation**

**Samuel Läubli**[1] **Rico Sennrich**[1,2] **Martin Volk**[1]

# Not so fast…

- Several recent studies show how human parity was yet to be achieved:
  - The “Super-Human” sentence-level systems are inferior to humans when evaluated in document-level
  - The translation direction when producing the references is crucial (“Translationese”)
  - The proficiency of the raters is crucial

**Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation**

**Antonio Toral**
Center for Language and Cognition
University of Groningen
The Netherlands
a.toral.ruiz@rug.nl

**Sheila Castilho** **Ke Hu** **Andy Way**
ADAPT Centre
Dublin City University
Ireland
firstname.secondname@adaptcentre.ie

**Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation**

**Samuel Läubli**[1] **Rico Sennrich**[1,2] **Martin Volk**[1]

# But is there hope?

# But is there hope?

- WMT 19’ included source-based sentence-level direct assessment with document context

# But is there hope?

- WMT 19' included source-based sentence-level direct assessment with document context
- For 3 out of 9 (De-En, En-De, En-Ru) language pairs, MT systems were tied or better than the reference translations

**German→English**

| Ave. | Ave. z | System |
|---|---|---|
| 81.6 | 0.146 | Facebook-FAIR |
| 81.5 | 0.136 | RWTH-Aachen |
| 79.0 | 0.136 | MSRA-MADL |
| 79.9 | 0.121 | online-B |
| 79.0 | 0.086 | JHU |
| 80.1 | 0.067 | MLLP-UPV |
| 79.0 | 0.066 | dfki-nmt |
| 78.0 | 0.066 | UCAM |
| 76.6 | 0.050 | online-A |
| 78.4 | 0.039 | NEU |
| 79.0 | 0.027 | HUMAN |
| 77.4 | 0.011 | uedin |
| 77.9 | 0.009 | online-Y |
| 74.8 | 0.006 | TartuNLP-c |
| 72.9 | −0.051 | online-G |
| 71.8 | −0.128 | PROMT-NMT |
| 69.7 | −0.192 | online-X |

# But is there hope?

- WMT 19’ included source-based sentence-level direct assessment with document context
- For 3 out of 9 (De-En, En-De, En-Ru) language pairs, MT systems were tied or better than the reference translations
  - **For sentence-level MT in high-resource settings, we can see some signs for human parity!**

**German→English**

| Ave. | Ave. z | System |
|---|---|---|
| 81.6 | 0.146 | Facebook-FAIR |
| 81.5 | 0.136 | RWTH-Aachen |
| 79.0 | 0.136 | MSRA-MADL |
| 79.9 | 0.121 | online-B |
| 79.0 | 0.086 | JHU |
| 80.1 | 0.067 | MLLP-UPV |
| 79.0 | 0.066 | dfki-nmt |
| 78.0 | 0.066 | UCAM |
| 76.6 | 0.050 | online-A |
| 78.4 | 0.039 | NEU |
| 79.0 | 0.027 | HUMAN |
| 77.4 | 0.011 | uedin |
| 77.9 | 0.009 | online-Y |
| 74.8 | 0.006 | TartuNLP-c |
| 72.9 | −0.051 | online-G |
| 71.8 | −0.128 | PROMT-NMT |
| 69.7 | −0.192 | online-X |

# But is there hope?

- WMT 19’ included source-based sentence-level direct assessment with document context
- For 3 out of 9 (De-En, En-De, En-Ru) language pairs, MT systems were tied or better than the reference translations
  - **For sentence-level MT in high-resource settings, we can see some signs for human parity!**
- However, for document level evaluation, the human translations are still significantly better, unlike in 2018

**German→English**

| Ave. | Ave. z | System |
|---|---|---|
| 81.6 | 0.146 | Facebook-FAIR |
| 81.5 | 0.136 | RWTH-Aachen |
| 79.0 | 0.136 | MSRA-MADL |
| 79.9 | 0.121 | online-B |
| 79.0 | 0.086 | JHU |
| 80.1 | 0.067 | MLLP-UPV |
| 79.0 | 0.066 | dfki-nmt |
| 78.0 | 0.066 | UCAM |
| 76.6 | 0.050 | online-A |
| 78.4 | 0.039 | NEU |
| 79.0 | 0.027 | HUMAN |
| 77.4 | 0.011 | uedin |
| 77.9 | 0.009 | online-Y |
| 74.8 | 0.006 | TartuNLP-c |
| 72.9 | −0.051 | online-G |
| 71.8 | −0.128 | PROMT-NMT |
| 69.7 | −0.192 | online-X |

**DR+DC**

| Ave. | Ave. z | System |
|---|---|---|
| 84.0 | 0.915 | HUMAN |
| 76.4 | 0.537 | CUNI-Transformer-T2T-2019 |
| 76.7 | 0.528 | CUNI-Transformer-T2T-2018 |
| 73.7 | 0.474 | CUNI-DocTransformer-T2T |
| 69.7 | 0.299 | CUNI-DocTransformer-Marian |
| 70.0 | 0.234 | uedin |
| 60.0 | −0.098 | TartuNLP-c |
| 59.9 | −0.169 | online-Y |
| 57.3 | −0.314 | online-B |
| 54.7 | −0.368 | online-G |
| 47.7 | −0.619 | online-A |
| 47.4 | −0.763 | online-X |

# Automatic Evaluation Methods

# Is it a good translation?

“奋进”号因机械手故障推迟到升空

Launch of “Endeavour” delayed by robotic arm problems.

“Progress” postponed because of mechanical hand into the sky.

# Is it a good translation?

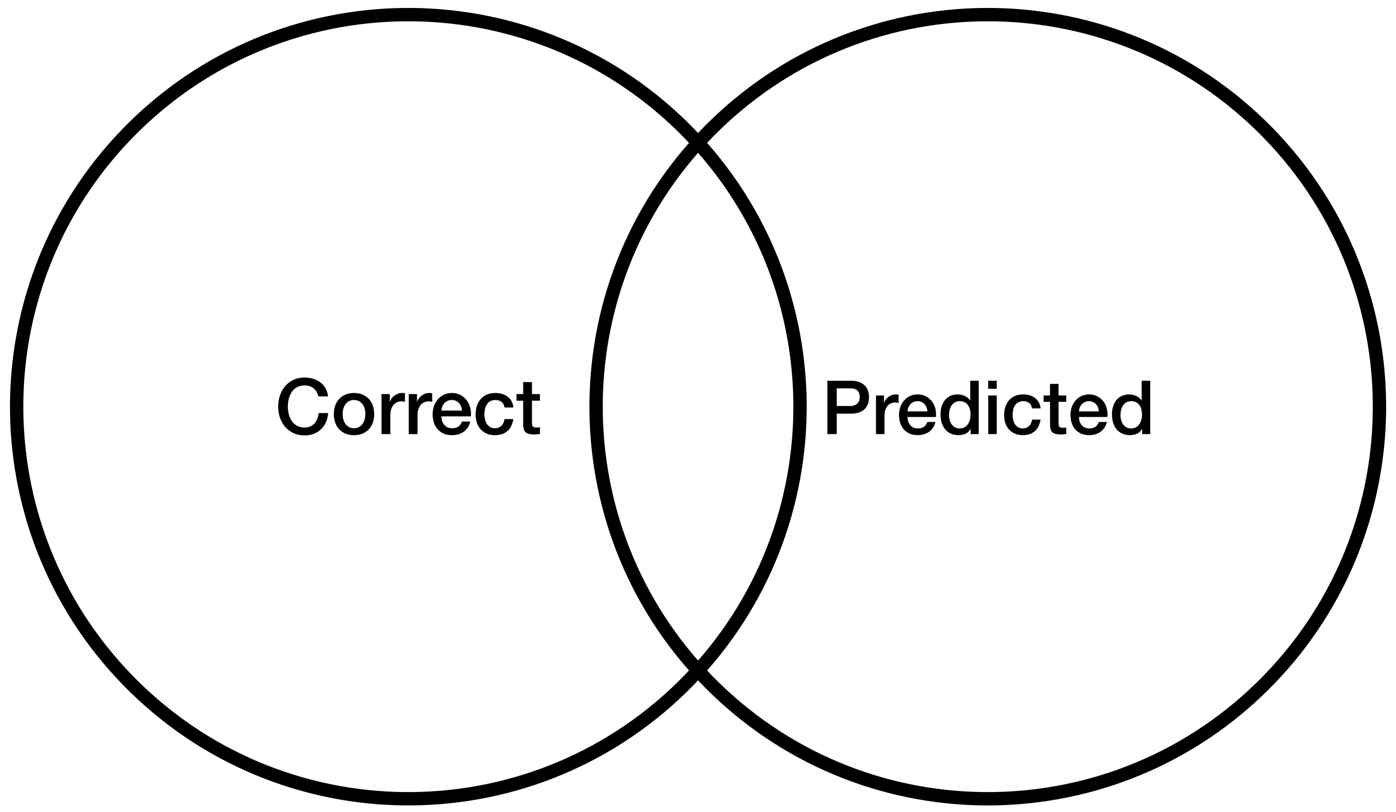
美国愿和北韩谈判但拒绝再付出报酬

US willing to negotiate with North Korea but not to pay more compensation.

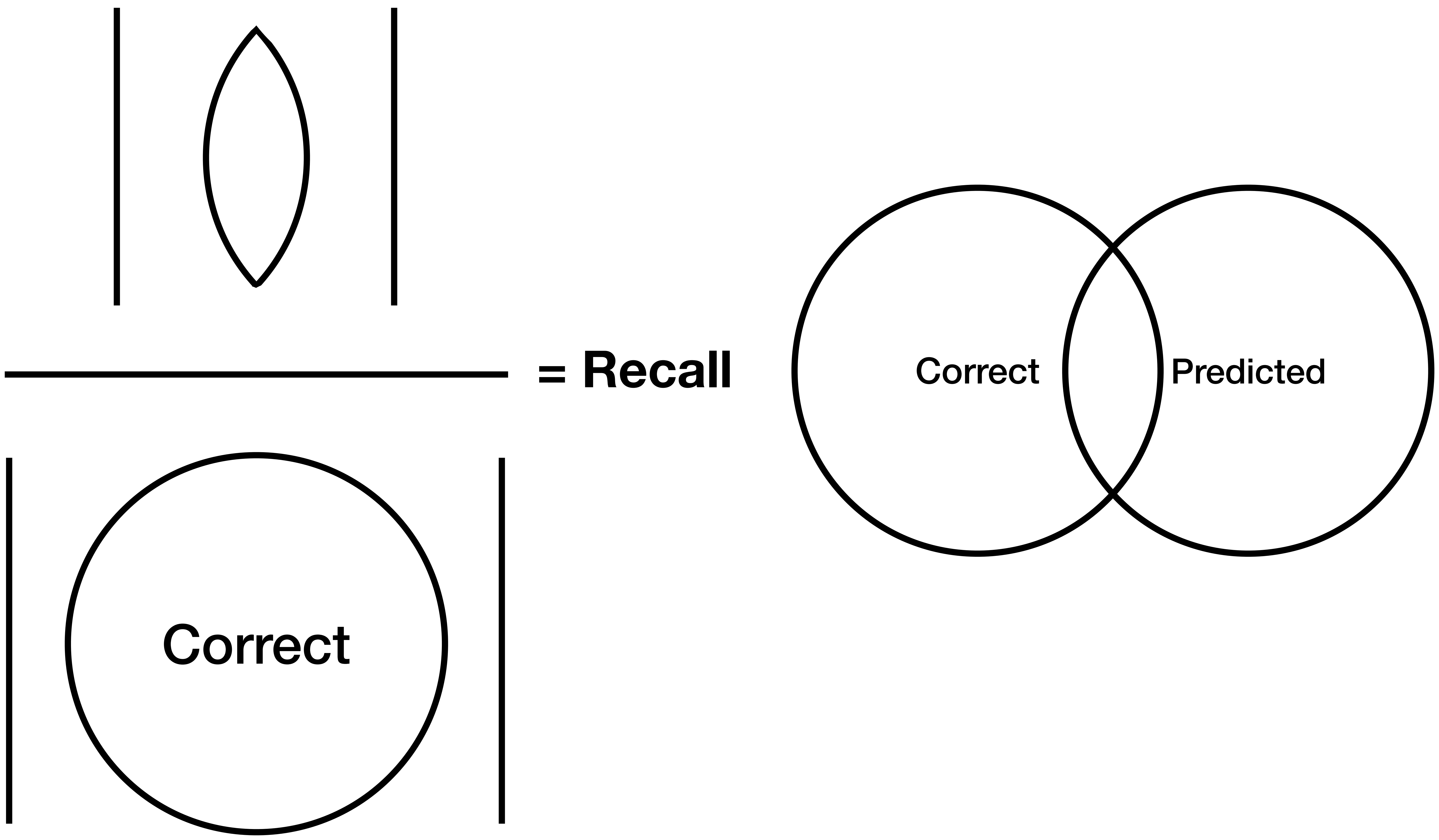
The United States is willing to hold talks with North Korea but refused to pay remuneration.

# Refresh - Precision/Recall

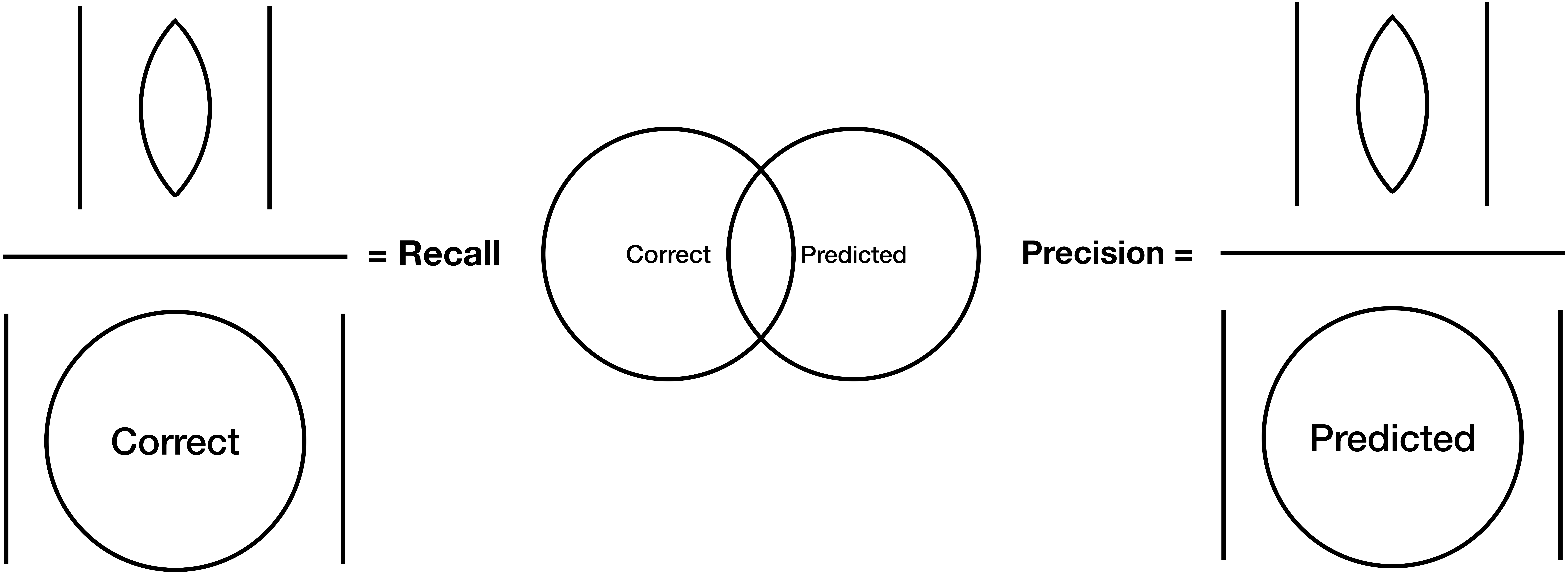
# Refresh - Precision/Recall



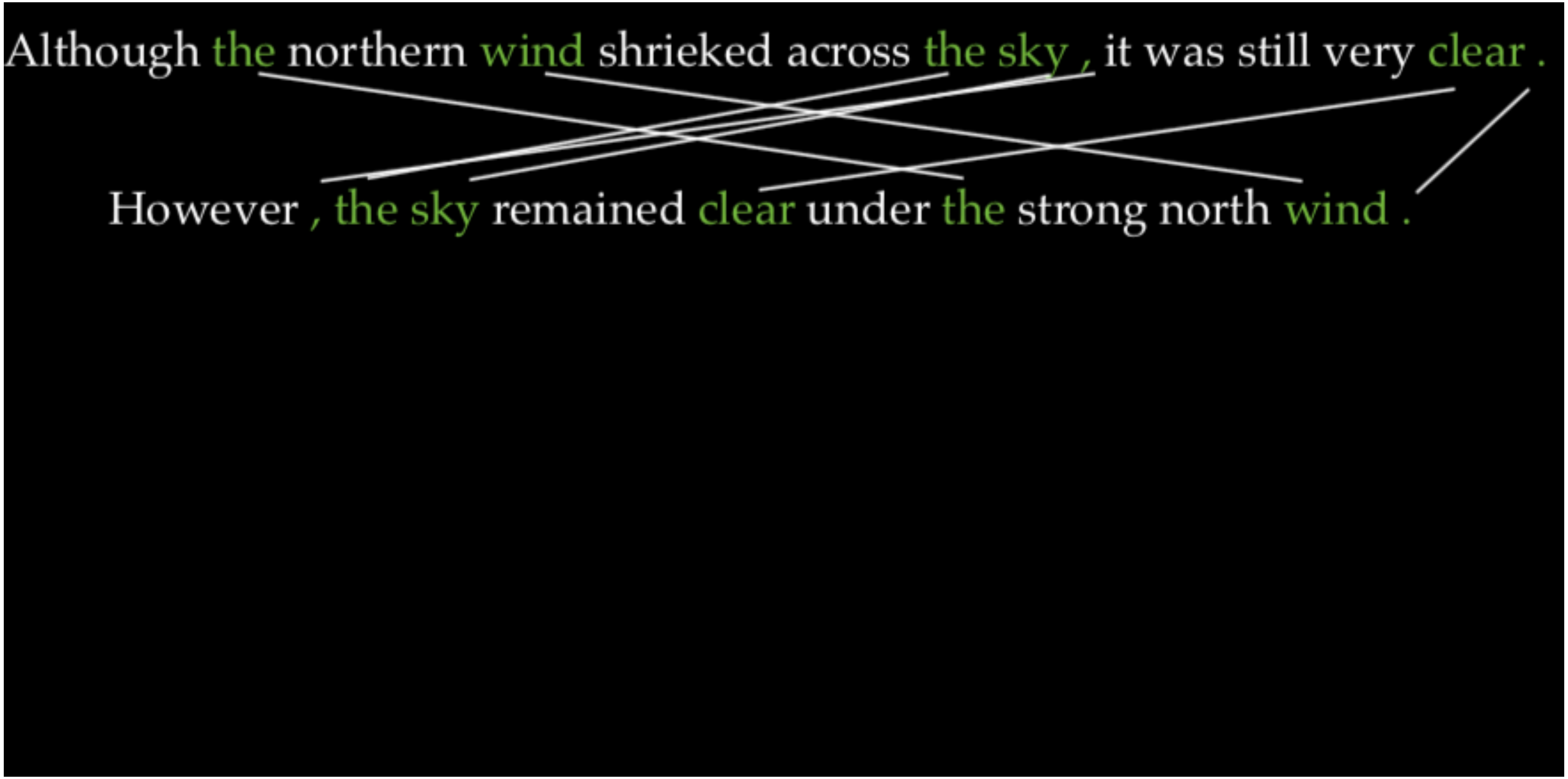
# Refresh - Precision/Recall



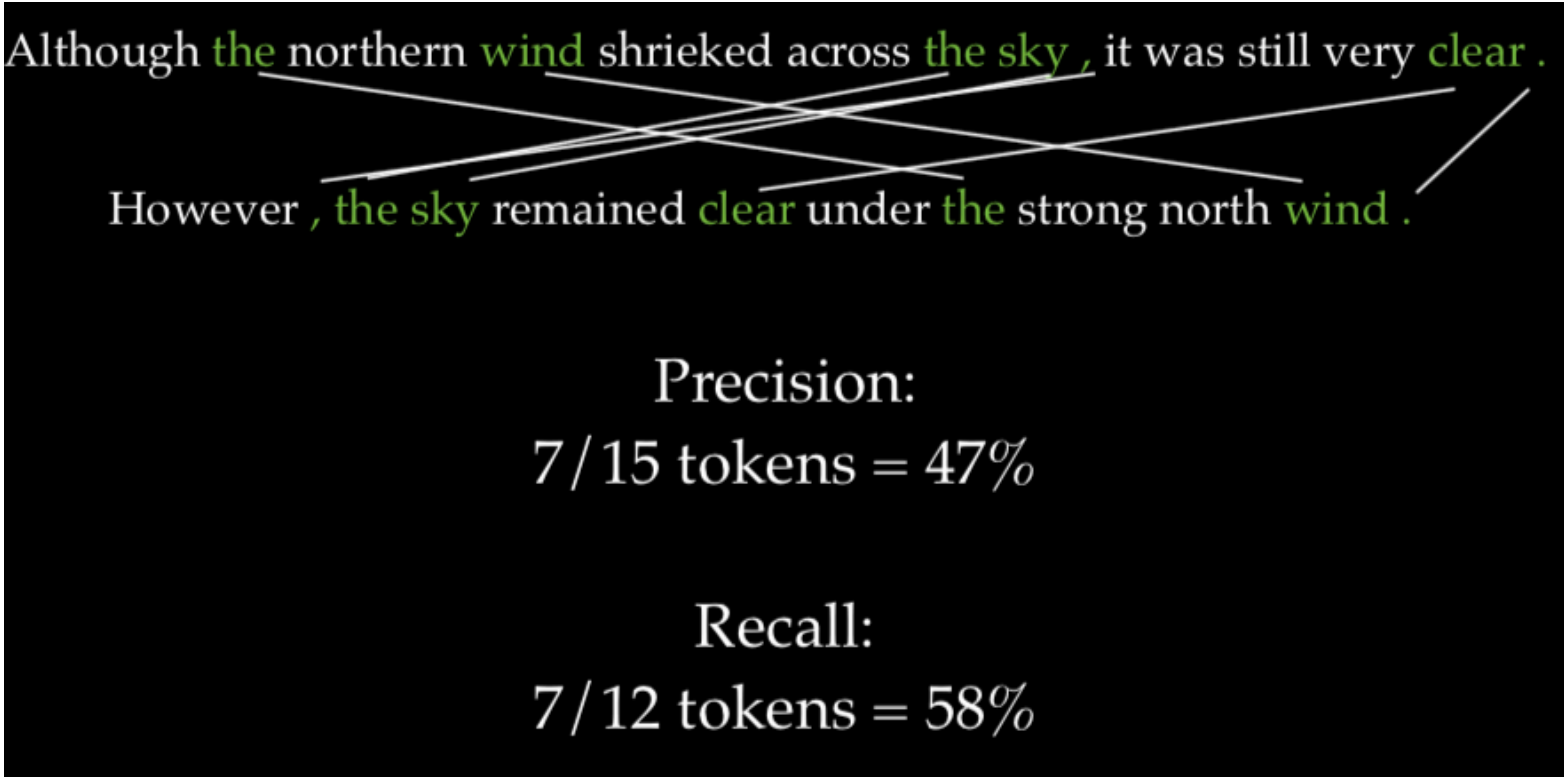
# Refresh - Precision/Recall



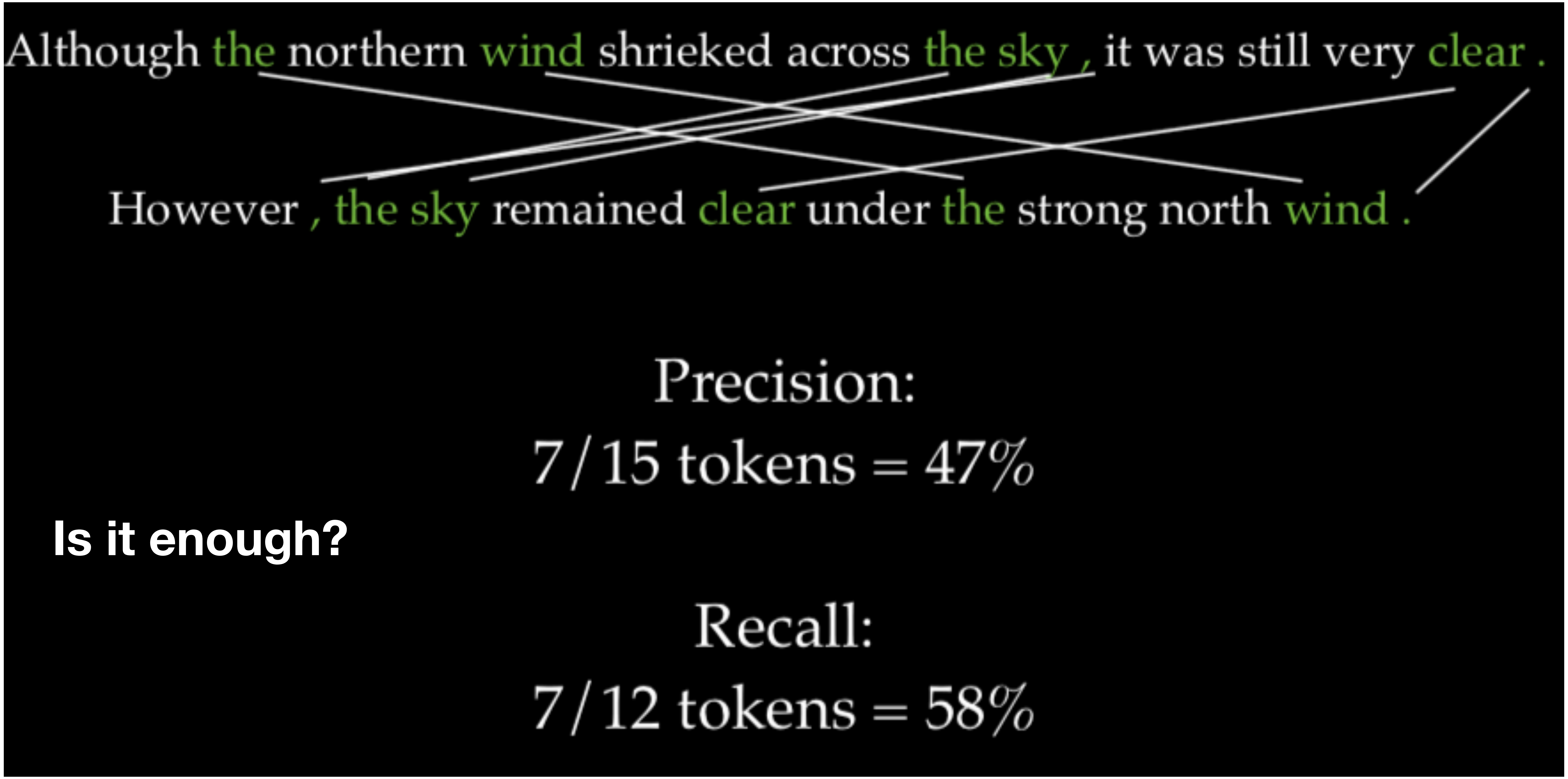
# Example - Precision and Recall



# Example - Precision and Recall



# Example - Precision and Recall



# Multiple References

Although the northern wind shrieked across the sky , it was still very clear .

However , the sky remained clear under the strong north wind .

Although a north wind was howling , the sky remained clear and blue .

The sky was still crystal clear , though the north wind was howling .

Despite the strong northerly winds , the sky remains very clear .

# Multiple References

Precision: 11/15 tokens

Although the northern wind shrieked across the sky , it was still very clear .

However , the sky remained clear under the strong north wind .

Although a north wind was howling , the sky remained clear and blue .

The sky was still crystal clear , though the north wind was howling .

Despite the strong northerly winds , the sky remains very clear .

# Multiple References

Precision: 11/15 tokens

Although the northern wind shrieked across the sky , it was still very clear .

**Is it enough?**

However , the sky remained clear under the strong north wind .

Although a north wind was howling , the sky remained clear and blue .

The sky was still crystal clear , though the north wind was howling .

Despite the strong northerly winds , the sky remains very clear .

# Capturing word order

sky very northern shrieked clear wind Although across the the , still was it .

However , the sky remained clear under the strong north wind .

Although a north wind was howling , the sky remained clear and blue .

The sky was still crystal clear , though the north wind was howling .

Despite the strong northerly winds , the sky remains very clear .

# Capturing word order

Precision: 11 / 15 tokens

sky very northern shrieked clear wind Although across the the , still was it .

However , the sky remained clear under the strong north wind .

Although a north wind was howling , the sky remained clear and blue .

The sky was still crystal clear , though the north wind was howling .

Despite the strong northerly winds , the sky remains very clear .

# Capturing word order

Precision: 11/15 tokens

sky very northern shrieked clear wind Although across the the , still was it .

**How can we fix this?**

However , the sky remained clear under the strong north wind .

Although a north wind was howling , the sky remained clear and blue .

The sky was still crystal clear , though the north wind was howling .

Despite the strong northerly winds , the sky remains very clear .

# N-gram Precision

Although the northern wind shrieked across the sky , it was still very clear .

However , the sky remained clear under the strong north wind .

Although a north wind was howling , the sky remained clear and blue .

The sky was still crystal clear , though the north wind was howling .

Despite the strong northerly winds , the sky remains very clear .

# N-gram Precision

Precision: 11/15 tokens
4/14 bigrams
1/13 trigrams

Although the northern wind shrieked across the sky , it was still very clear .

However , the sky remained clear under the strong north wind .

Although a north wind was howling , the sky remained clear and blue .

The sky was still crystal clear , though the north wind was howling .

Despite the strong northerly winds , the sky remains very clear .

# N-gram Precision

Precision: 11/15 tokens
0/14 bigrams
0/13 trigrams

sky very northern shrieked clear wind Although across the the , still was it .

However , the sky remained clear under the strong north wind .

Although a north wind was howling , the sky remained clear and blue .

The sky was still crystal clear , though the north wind was howling .

Despite the strong northerly winds , the sky remains very clear .

# Weakness of precision - low coverage

Precision: 3/1 tokens

2/2 bigrams

1/1 trigrams

very clear .

However , the sky remained clear under the strong north wind .

Although a north wind was howling , the sky remained clear and blue .

The sky was still crystal clear , though the north wind was howling .

Despite the strong northerly winds , the sky remains very clear .

# Weakness of precision - repetitions

Precision: 11/15 tokens

4/14 bigrams

1/13 trigrams

a north . the was and was the the the though the , the sky

However , the sky remained clear under the strong north wind .

Although a north wind was howling , the sky remained clear and blue .

The sky was still crystal clear , though the north wind was howling .

Despite the strong northerly winds , the sky remains very clear .

# BLEU

# BLEU

- “**B**i**l**ingual **E**valuation **U**nderstudy”

**BLEU: a Method for Automatic Evaluation of Machine Translation**

**Kishore Papineni, Salim Roukos, Todd Ward,** and **Wei-Jing Zhu**
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
{papineni,roukos,toddward,weijing}@us.ibm.com

# BLEU

BLEU: a Method for Automatic Evaluation of Machine Translation

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
{papineni,roukos,toddward,weijing}@us.ibm.com

- “**Bi**l**i**ngual **E**valuation **U**nderstudy”
- Published in 2002

# BLEU

BLEU: a Method for Automatic Evaluation of Machine Translation

**Kishore Papineni, Salim Roukos, Todd Ward,** and **Wei-Jing Zhu**
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
{papineni,roukos,toddward,weijing}@us.ibm.com

- "**Bi**lingual **E**valuation **U**nderstudy"
- Published in 2002
- 10852 citations, as of 3/2020

# BLEU

## BLEU: a Method for Automatic Evaluation of Machine Translation

**Kishore Papineni, Salim Roukos, Todd Ward,** and **Wei-Jing Zhu**
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
{papineni,roukos,toddward,weijing}@us.ibm.com

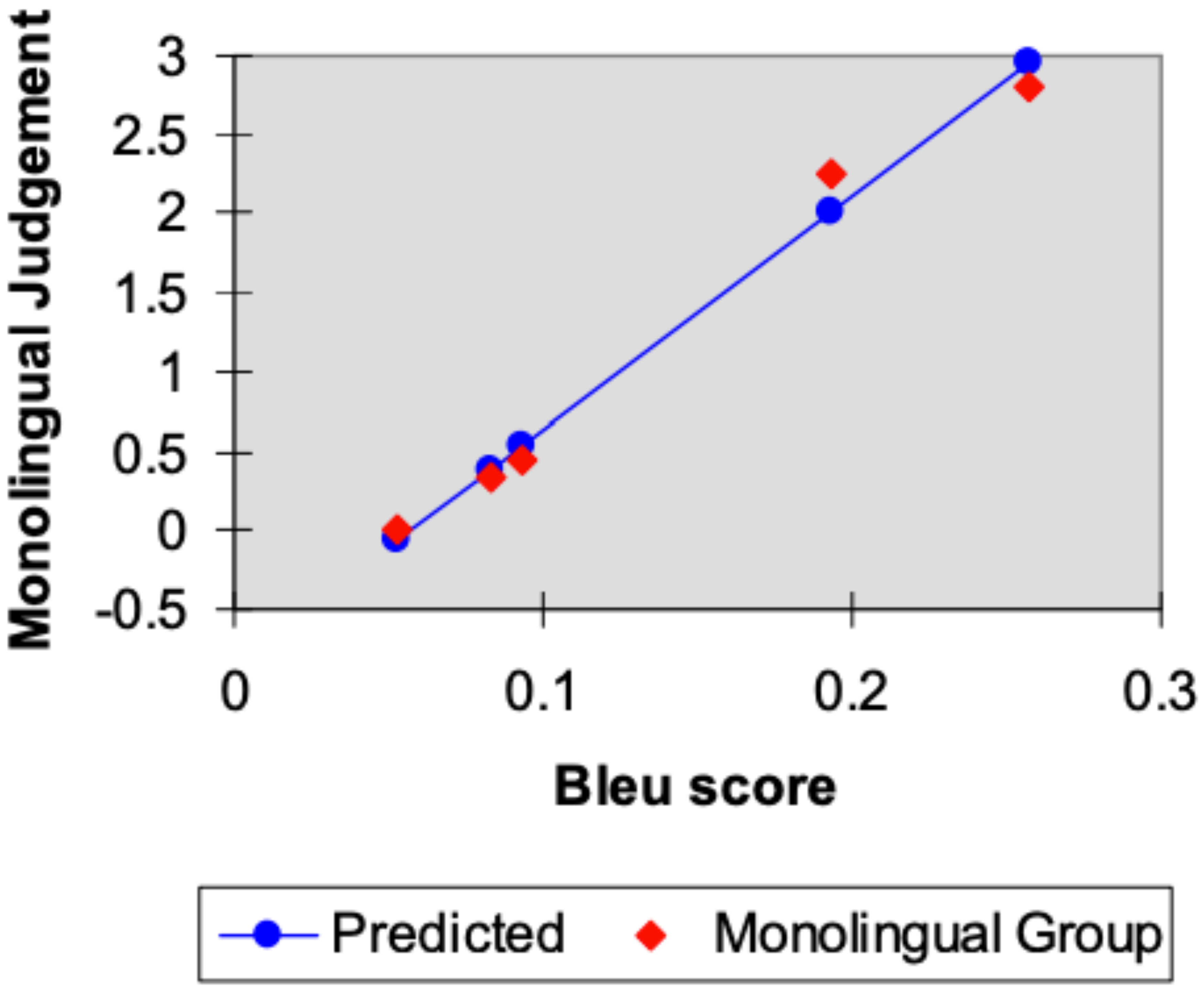
- “**Bi**l**i**ngual **E**valuation **U**nderstudy”
- Published in 2002
- 10852 citations, as of 3/2020
- Simple, reproducible, fast

# BLEU

**BLEU: a Method for Automatic Evaluation of Machine Translation**

**Kishore Papineni, Salim Roukos, Todd Ward,** and **Wei-Jing Zhu**
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
{papineni,roukos,toddward,weijing}@us.ibm.com

- "**Bi**lingual **E**valuation **U**nderstudy"
- Published in 2002
- 10852 citations, as of 3/2020
- Simple, reproducible, fast
- Correlated well with human evaluation



# BLEU - How it works?

# BLEU - How it works?

(clipped) precision for each n-gram size (usually 1-4)

$$p_n = \frac{\sum\limits_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum\limits_{n\text{-}gram' \in C'} Count(n\text{-}gram')}$$

# BLEU - How it works?

(clipped) precision for each n-gram size (usually 1-4)

$$p_n = \frac{\sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}$$

Brevity Penalty - punish if candidate is too short

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

# BLEU - How it works?

(clipped) precision for each n-gram size (usually 1-4)

$$p_n = \frac{\sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}$$

Brevity Penalty - punish if candidate is too short

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

BLEU score

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \qquad w_n = 1/N$$

# BLEU - Discussion

# BLEU - Discussion

- Can we compare BLEU scores across different systems?

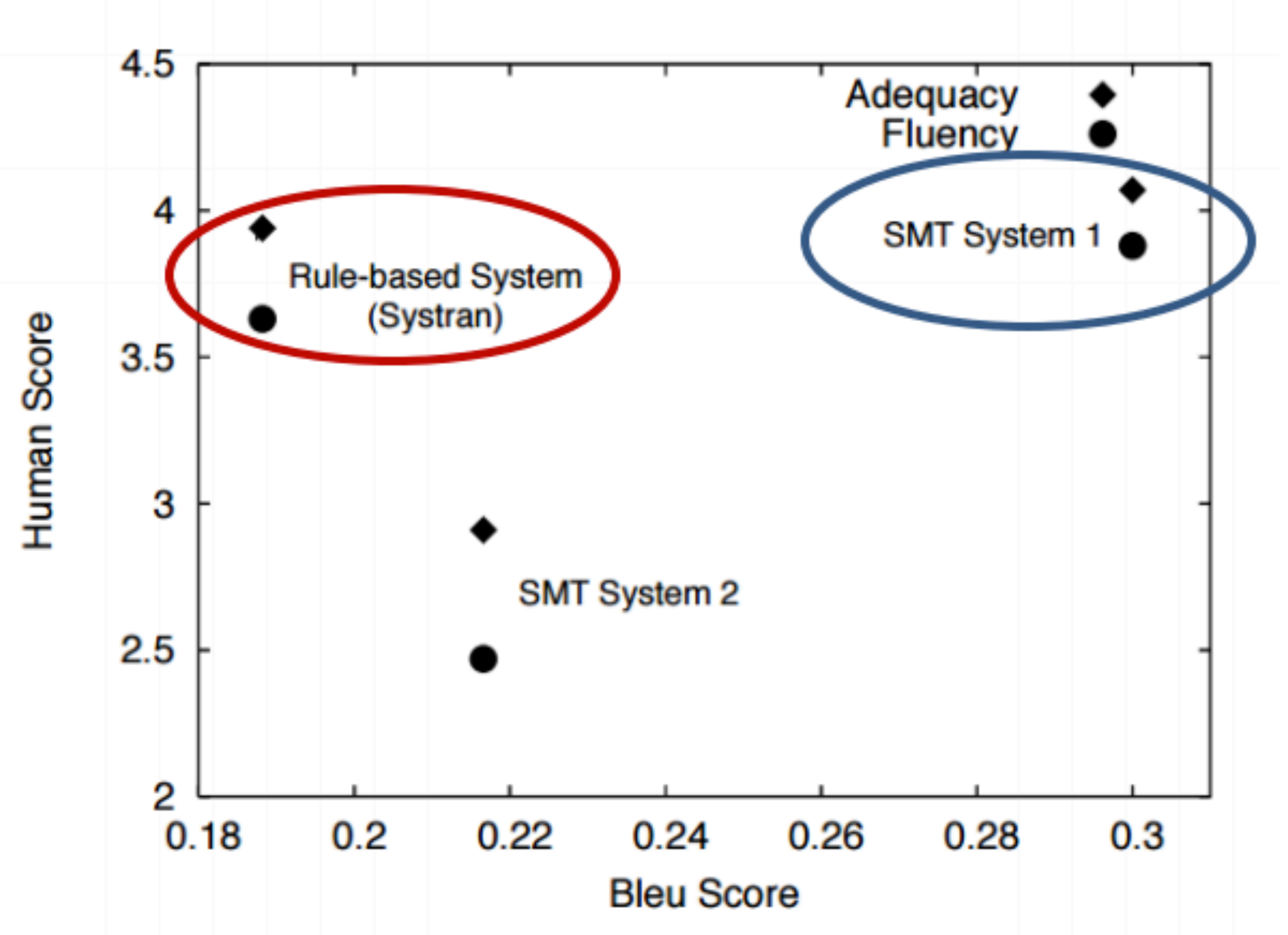
# BLEU - Discussion

- Can we compare BLEU scores across different systems?
- Can we compare BLEU scores across different languages?

# BLEU - Discussion

- Can we compare BLEU scores across different systems?
- Can we compare BLEU scores across different languages?
- Can we compare BLEU scores across different datasets?

# Issues with BLEU



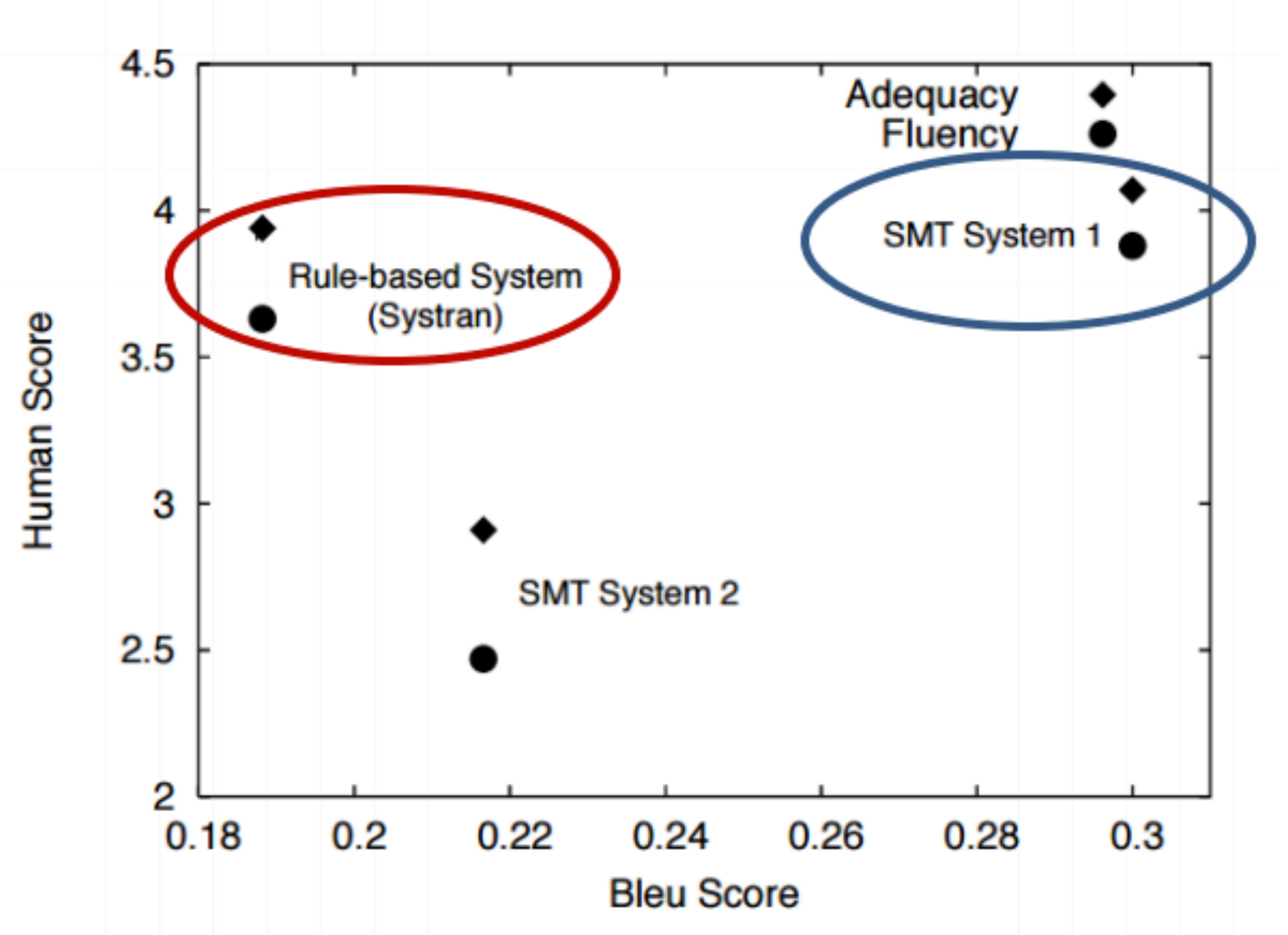

**Callison-Burch et al. 2006**

**A Call for Clarity in Reporting BLEU Scores**

**Matt Post**
Amazon Research
Berlin, Germany

# Issues with BLEU

- BLEU is not always correlated with human judgements



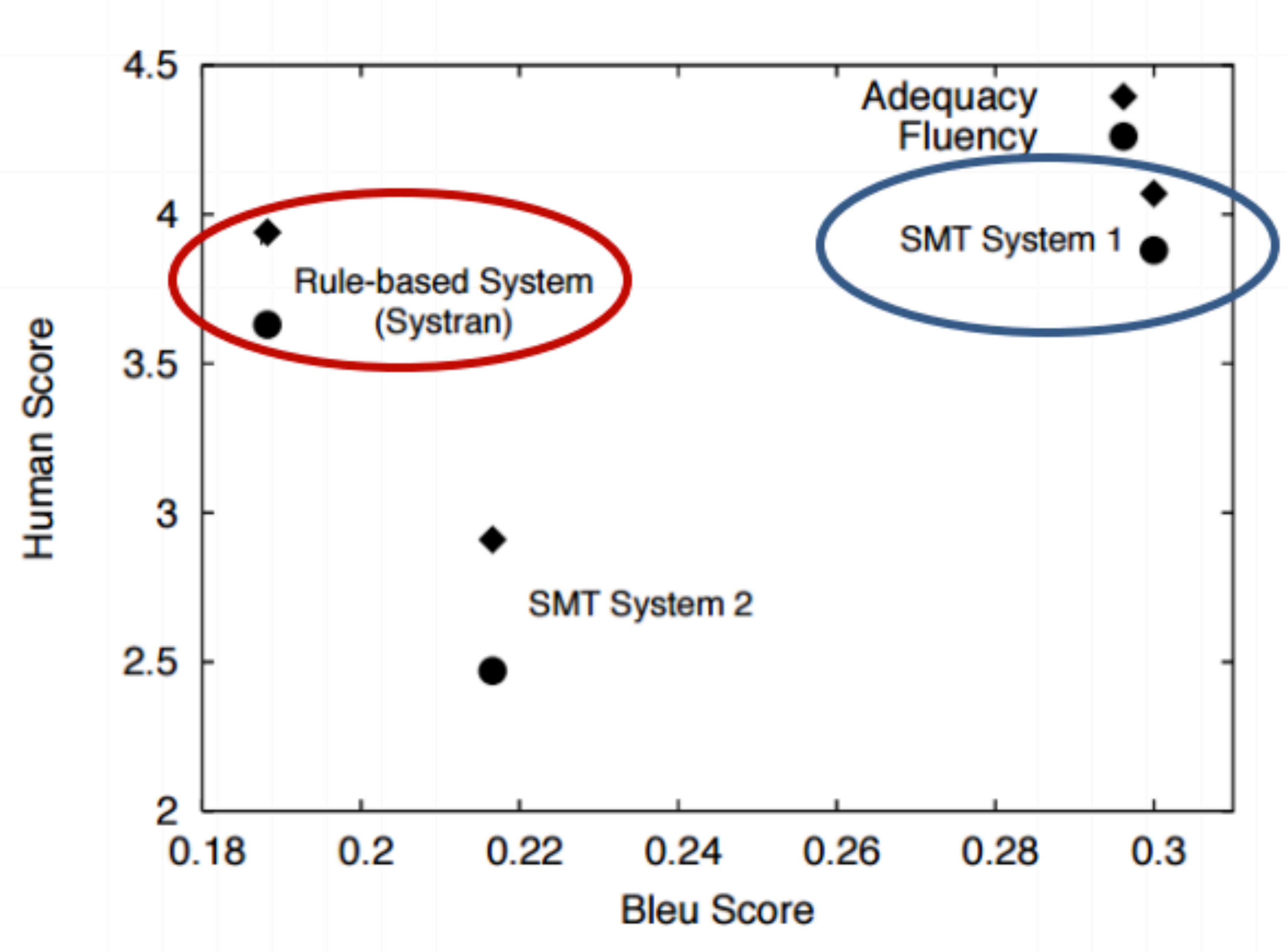

**Callison-Burch et al. 2006**

**A Call for Clarity in Reporting BLEU Scores**

**Matt Post**
Amazon Research
Berlin, Germany

# Issues with BLEU

- BLEU is not always correlated with human judgements
  - Most current works do not use multiple references



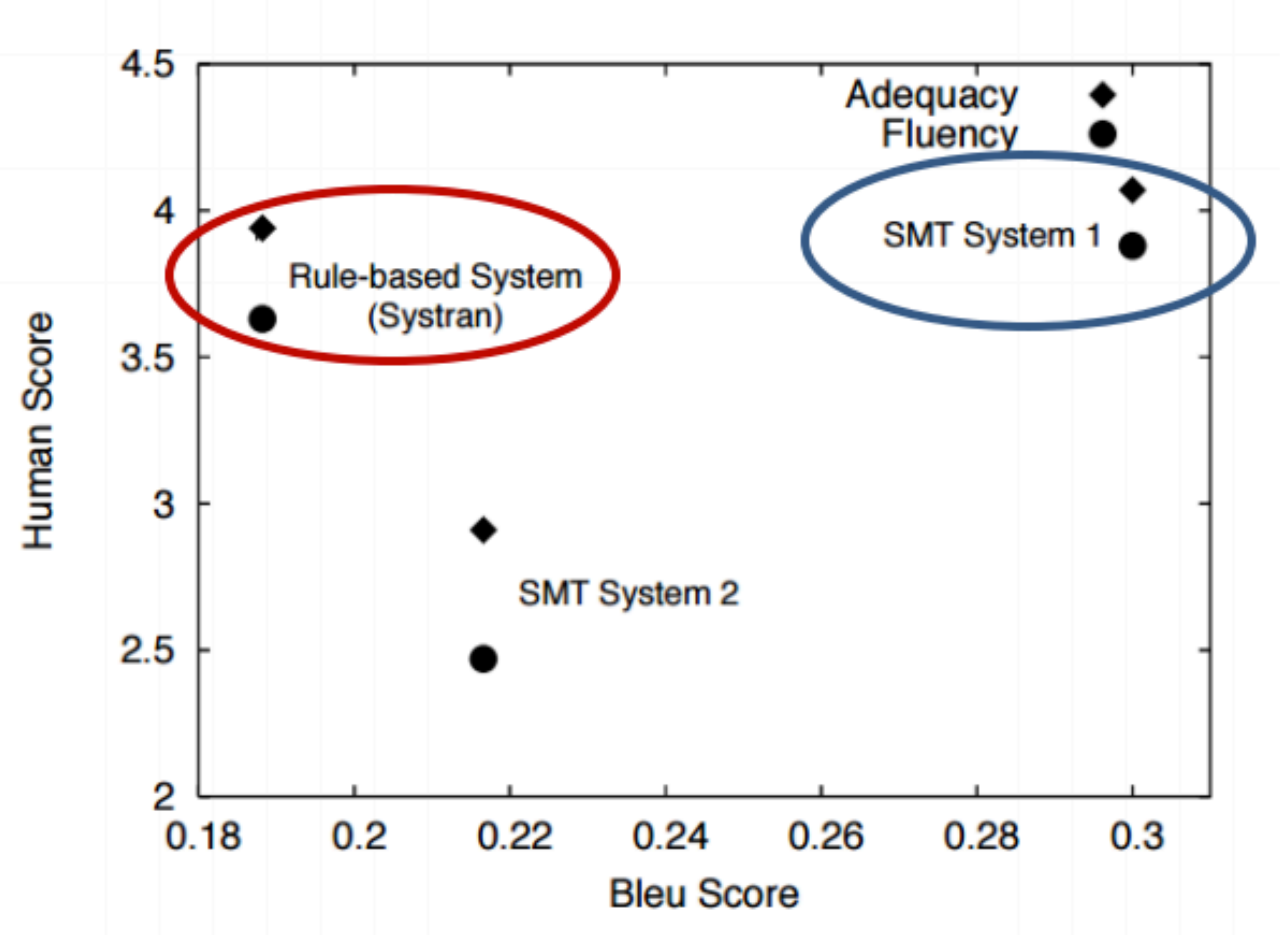

Callison-Burch et al. 2006

**A Call for Clarity in Reporting BLEU Scores**

**Matt Post**
Amazon Research
Berlin, Germany

# Issues with BLEU

- BLEU is not always correlated with human judgements
  - Most current works do not use multiple references
- Differences between implementations



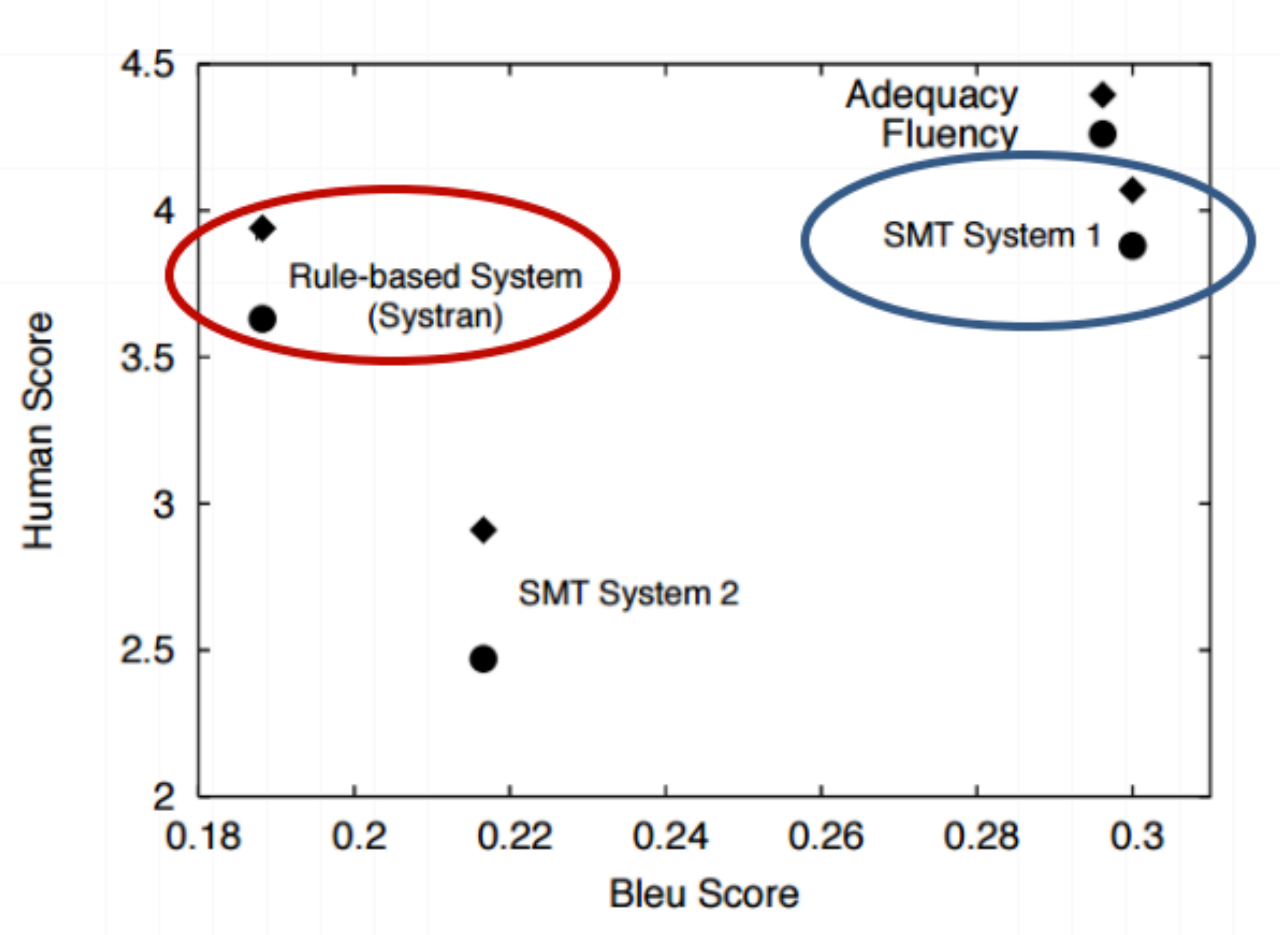

**Callison-Burch et al. 2006**

**A Call for Clarity in Reporting BLEU Scores**

**Matt Post**
Amazon Research
Berlin, Germany

# Issues with BLEU

- BLEU is not always correlated with human judgements
  - Most current works do not use multiple references
- Differences between implementations
  - Tokenisation, normalisation



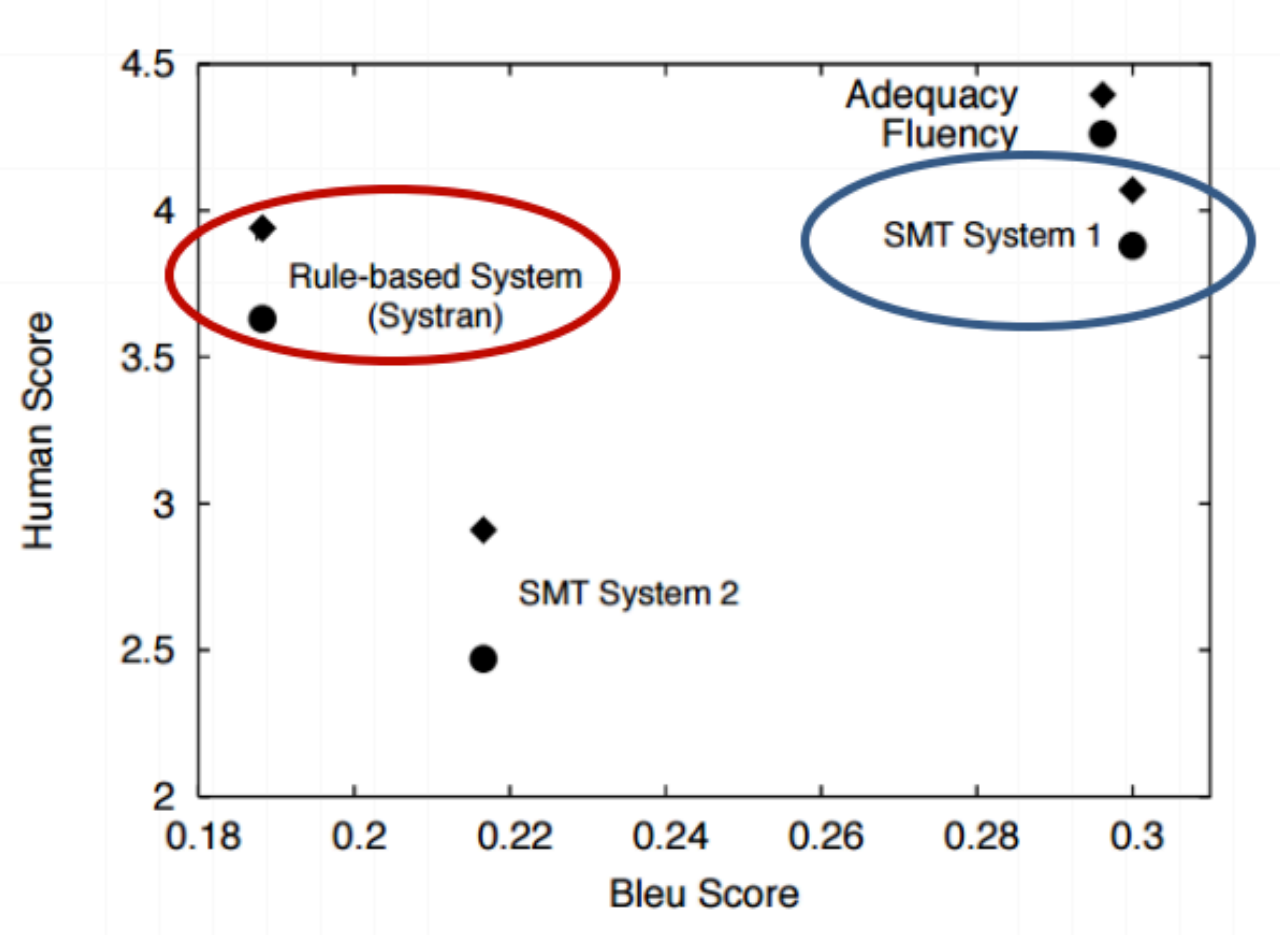

**Callison-Burch et al. 2006**

**A Call for Clarity in Reporting BLEU Scores**

**Matt Post**
Amazon Research
Berlin, Germany

# Issues with BLEU

- BLEU is not always correlated with human judgements
  - Most current works do not use multiple references
- Differences between implementations
  - Tokenisation, normalisation
  - Use SacreBLEU!




**Callison-Burch et al. 2006**

**A Call for Clarity in Reporting BLEU Scores**

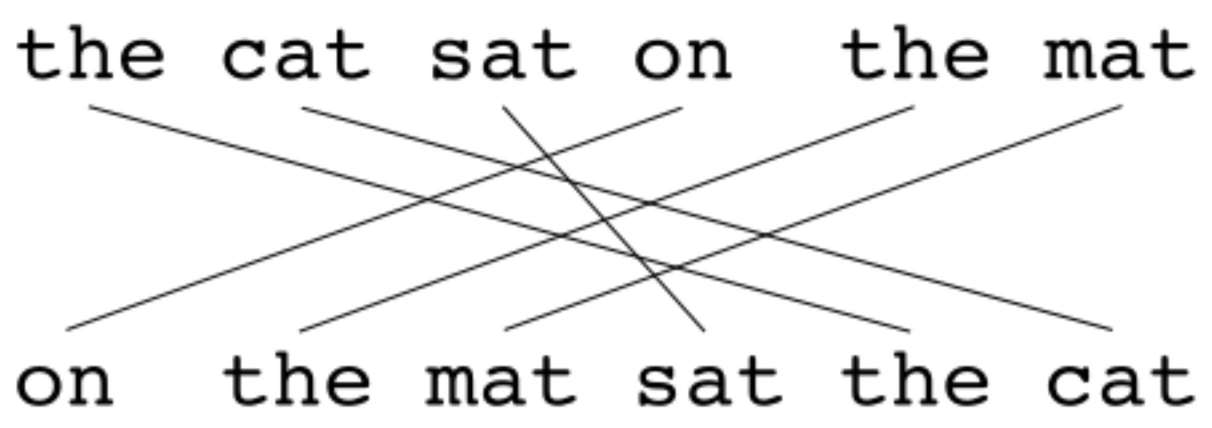
**Matt Post**
Amazon Research
Berlin, Germany

# METEOR

# METEOR

- “Metric for Evaluation of Translation with Explicit ORdering”

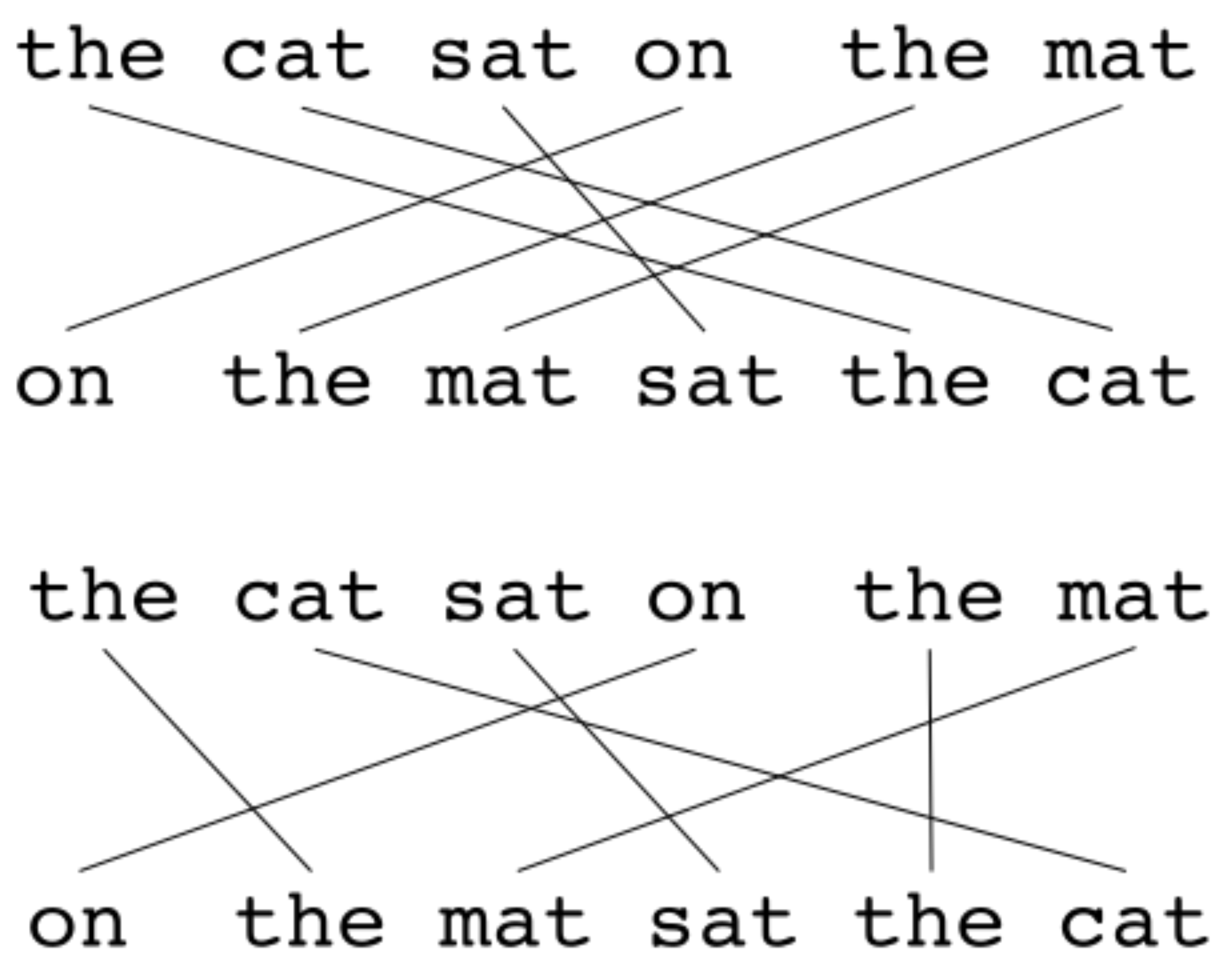
# METEOR




- “Metric for Evaluation of Translation with Explicit ORdering”
- Produces an alignment with iterative matching

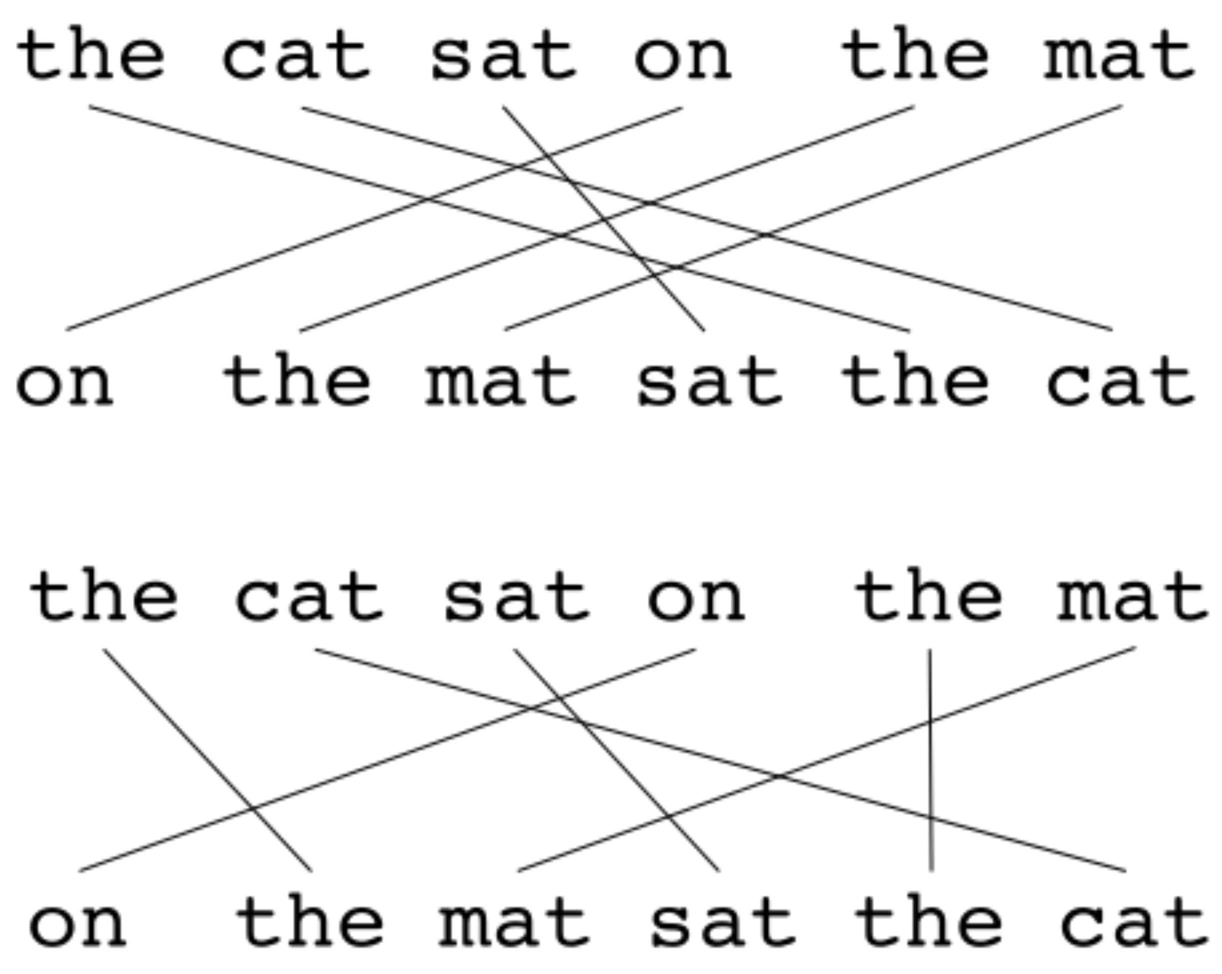
# METEOR

- "Metric for Evaluation of Translation with Explicit ORdering"
- Produces an alignment with iterative matching
  - Preferably with fewest crosses



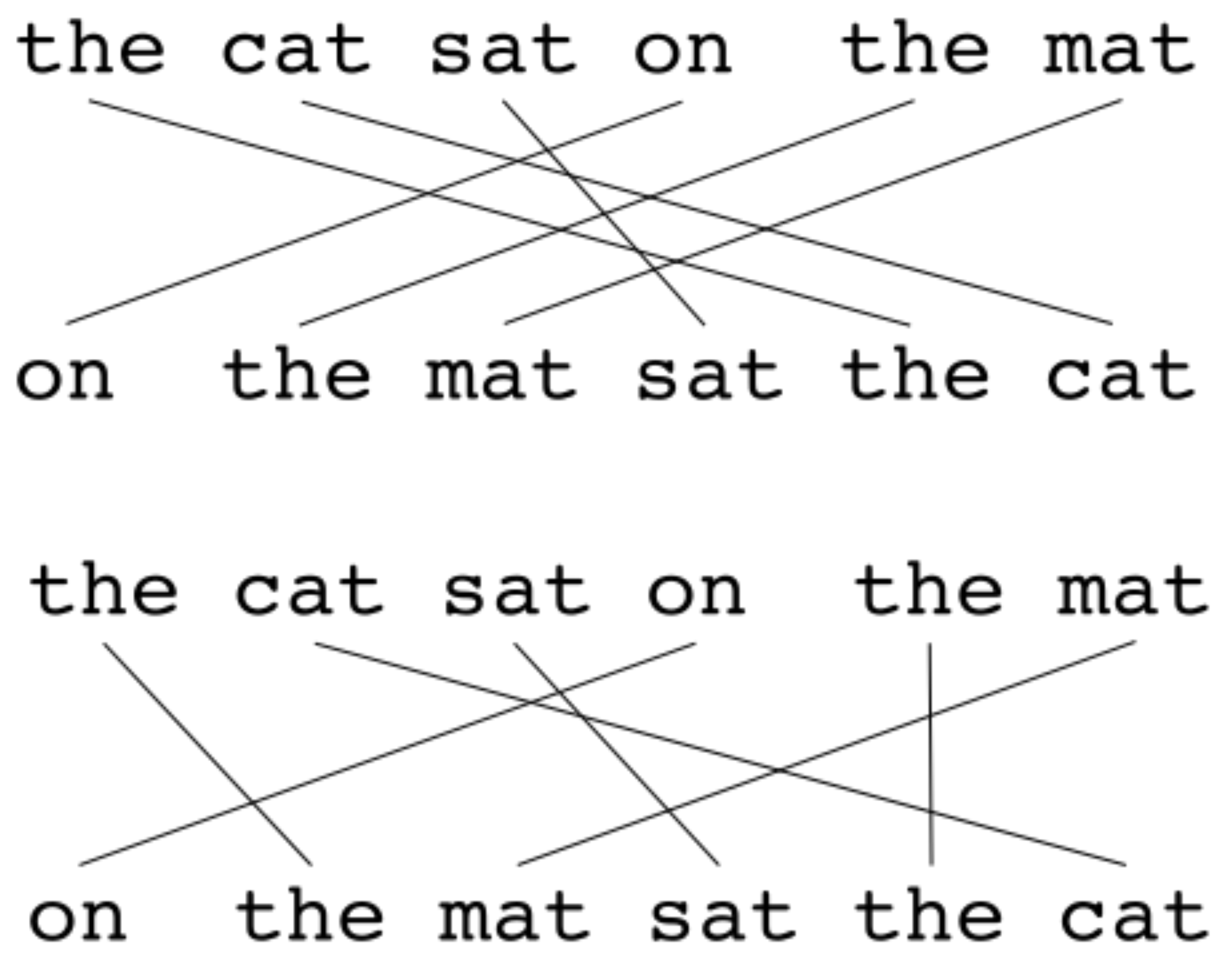
# METEOR

- “Metric for Evaluation of Translation with Explicit ORdering”
- Produces an alignment with iterative matching
  - Preferably with fewest crosses
- Computes unigram precision and recall



# METEOR

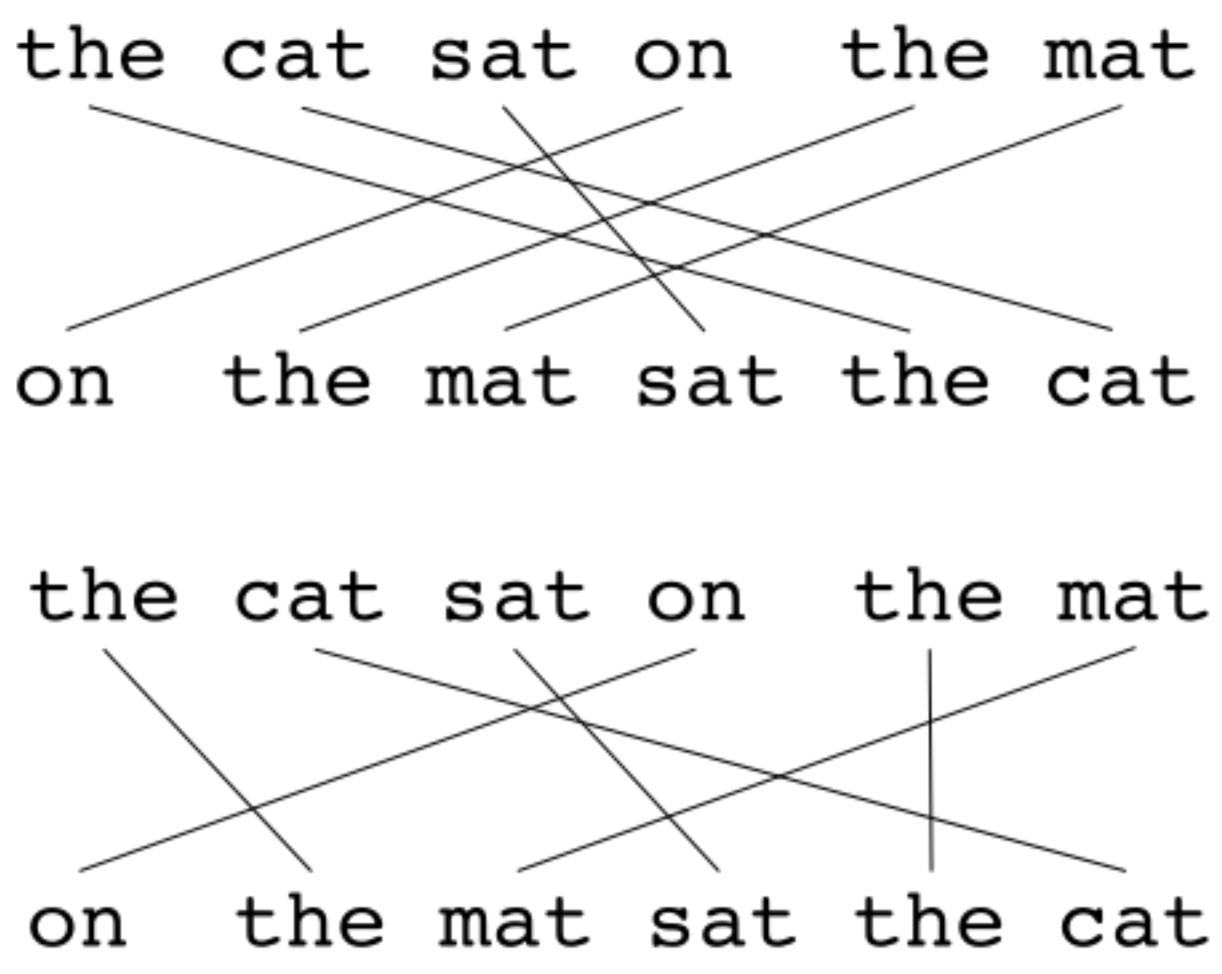
- “Metric for Evaluation of Translation with Explicit ORdering”
- Produces an alignment with iterative matching
  - Preferably with fewest crosses
- Computes unigram precision and recall
- Computes their harmonic mean, with more weight for recall




$$F_{mean} = \frac{10PR}{R + 9P}$$

# METEOR

- “Metric for Evaluation of Translation with Explicit ORdering”
- Produces an alignment with iterative matching
    - Preferably with fewest crosses
- Computes unigram precision and recall
- Computes their harmonic mean, with more weight for recall
- Penalise alignments with non-consecutive chunks

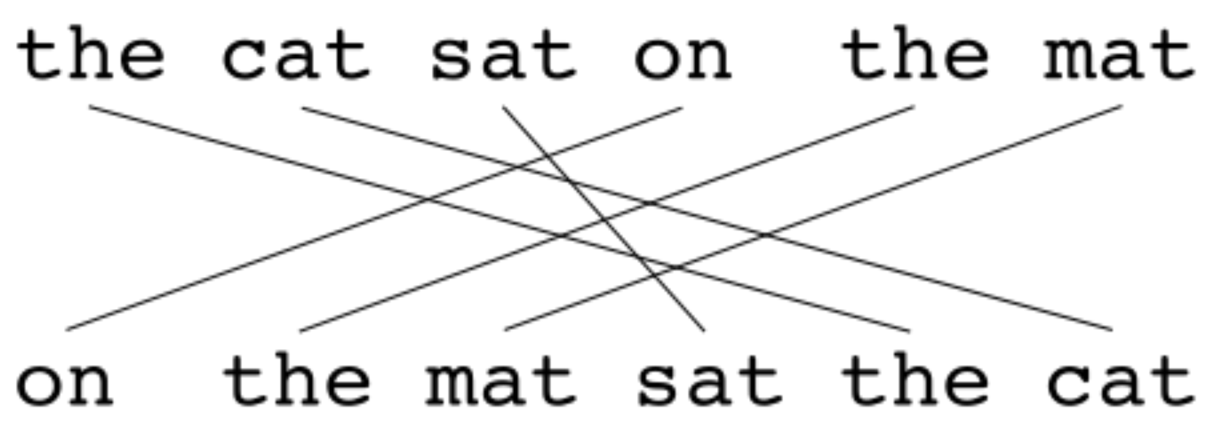


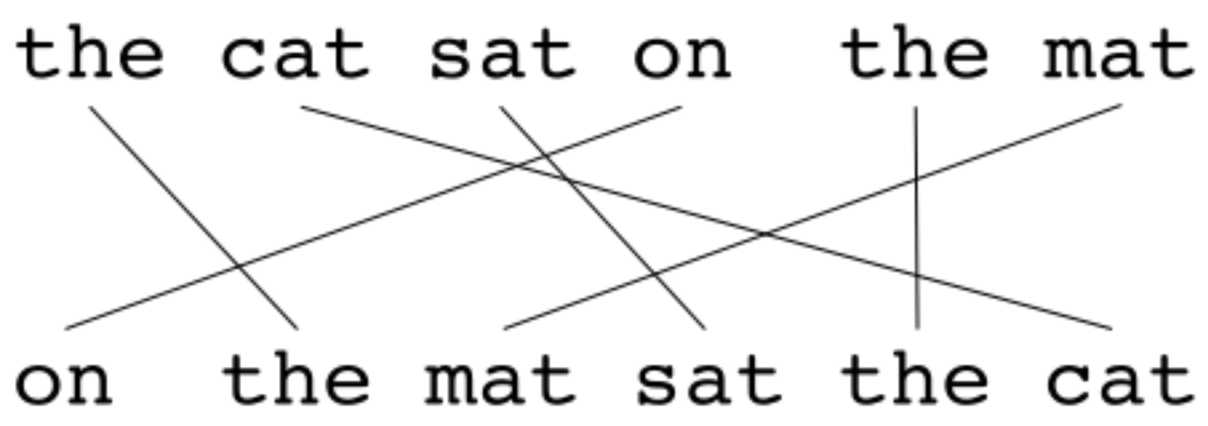

$$F_{mean} = \frac{10PR}{R + 9P}$$

$$p = 0.5\left(\frac{c}{u_m}\right)^3$$

# METEOR

- "Metric for Evaluation of Translation with Explicit ORdering"
- Produces an alignment with iterative matching
  - Preferably with fewest crosses
- Computes unigram precision and recall
- Computes their harmonic mean, with more weight for recall
- Penalise alignments with non-consecutive chunks
- Final score: $M = F_{mean}(1 - p)$



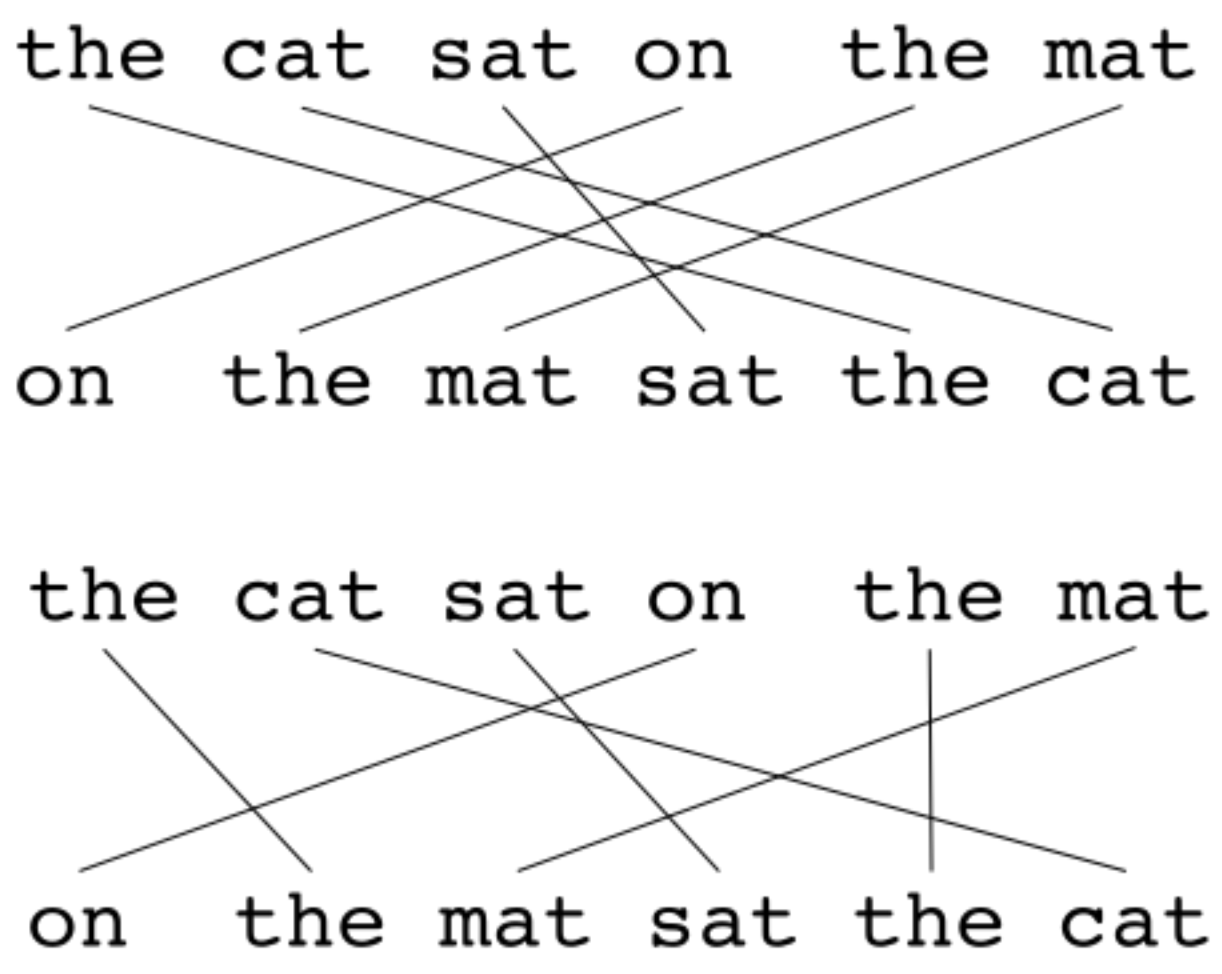




$$F_{mean} = \frac{10PR}{R + 9P}$$

$$p = 0.5\left(\frac{c}{u_m}\right)^3$$

# METEOR

- “Metric for Evaluation of Translation with Explicit ORdering”
- Produces an alignment with iterative matching
  - Preferably with fewest crosses
- Computes unigram precision and recall
- Computes their harmonic mean, with more weight for recall
- Penalise alignments with non-consecutive chunks
- Final score: $M = F_{mean}(1-p)$
- Also uses stemming (“goods”-“good”), synonyms (“well”-“good”)




$$F_{mean} = \frac{10PR}{R+9P}$$

$$p = 0.5\left(\frac{c}{u_m}\right)^3$$

# Contrastive Evaluation

# Contrastive Evaluation

- One problem with BLEU/METEOR - not specific:

# Contrastive Evaluation

- One problem with BLEU/METEOR - not specific:
  - What can we learn from X>Y? Why X>Y?

# Contrastive Evaluation

- One problem with BLEU/METEOR - not specific:
  - What can we learn from X>Y? Why X>Y?
- An alternative - measure a specific **linguistic phenomena**

# Contrastive Evaluation

- One problem with BLEU/METEOR - not specific:
  - What can we learn from X>Y? Why X>Y?
- An alternative - measure a specific **linguistic phenomena**

**How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs**

**Rico Sennrich**
School of Informatics, University of Edinburgh
{rico.sennrich}@ed.ac.uk

# Contrastive Evaluation

- One problem with BLEU/METEOR - not specific:
  - What can we learn from X>Y? Why X>Y?
- An alternative - measure a specific **linguistic phenomena**

**How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs**

**Rico Sennrich**
School of Informatics, University of Edinburgh
{rico.sennrich}@ed.ac.uk

**Transliteration:**
Aumann - אומן
Yonat - יונת

# Contrastive Evaluation

- One problem with BLEU/METEOR - not specific:
  - What can we learn from X>Y? Why X>Y?
- An alternative - measure a specific **linguistic phenomena**

**How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs**

Rico Sennrich
School of Informatics, University of Edinburgh
{rico.sennrich}@ed.ac.uk

**Transliteration:**
Aumann - אומן
Yonat - יונת

**Subject-verb agreement:**
She - הלכה, אכלה, ישבה
He - הלך, אכל, ישב

# Contrastive Evaluation

- One problem with BLEU/METEOR - not specific:
  - What can we learn from X>Y? Why X>Y?
- An alternative - measure a specific **linguistic phenomena**

**How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs**

**Rico Sennrich**
School of Informatics, University of Edinburgh
{rico.sennrich}@ed.ac.uk

**Transliteration:**
Aumann - אומן
Yonat - יונת

**Subject-verb agreement:**
She - הלכה, אכלה, ישבה
He - הלך, אכל, ישב

**Polarity/Negation:**
Uncertain - לא בטוח

# Contrastive Evaluation

- One problem with BLEU/METEOR - not specific:
  - What can we learn from X>Y? Why X>Y?
- An alternative - measure a specific **linguistic phenomena**

**How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs**

**Rico Sennrich**
School of Informatics, University of Edinburgh
{rico.sennrich}@ed.ac.uk

**Transliteration:**
Aumann - אומן
Yonat - יונת

**Subject-verb agreement:**
She - הלכה, אכלה, ישבה
He - הלך, אכל, ישב

**Polarity/Negation:**
Uncertain - לא בטוח

**Ambiguous words:**

| Example containing **ambiguous word** | Correct translations | Incorrect translations |
|---|---|---|
| It occurred to me that my **watch** might be broken. | Armbanduhr, Uhr | *Wache* |
| I hope you didn't get distracted during your **watch**. | *Wache* | Armbanduhr, Uhr |

# Contrastive Evaluation

- One problem with BLEU/METEOR - not specific:
  - What can we learn from X>Y? Why X>Y?
- An alternative - measure a specific **linguistic phenomena**
- General Idea - for each source sentence:

**How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs**

Rico Sennrich
School of Informatics, University of Edinburgh
{rico.sennrich}@ed.ac.uk

**Transliteration:**
Aumann - אומן
Yonat - יונת

**Subject-verb agreement:**
She - הלכה, אכלה, ישבה
He - הלך, אכל, ישב

**Polarity/Negation:**
Uncertain - לא בטוח

**Ambiguous words:**

| Example containing **ambiguous word** | Correct translations | Incorrect translations |
|---|---|---|
| It occurred to me that my **watch** might be broken. | Armbanduhr, Uhr | *Wache* |
| I hope you didn't get distracted during your **watch**. | *Wache* | Armbanduhr, Uhr |

# Contrastive Evaluation

- One problem with BLEU/METEOR - not specific:
  - What can we learn from X>Y? Why X>Y?
- An alternative - measure a specific **linguistic phenomena**
- General Idea - for each source sentence:
  - Create two translation options - one correct, one wrong

**How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs**

**Rico Sennrich**
School of Informatics, University of Edinburgh
{rico.sennrich}@ed.ac.uk

**Transliteration:**
Aumann - אומן
Yonat - יונת

**Subject-verb agreement:**
She - הלכה, אכלה, ישבה
He - הלך, אכל, ישב

**Polarity/Negation:**
Uncertain - לא בטוח

**Ambiguous words:**

| Example containing **ambiguous word** | Correct translations | Incorrect translations |
|---|---|---|
| It occurred to me that my **watch** might be broken. | Armbanduhr, Uhr | *Wache* |
| I hope you didn't get distracted during your **watch**. | *Wache* | Armbanduhr, Uhr |

# Contrastive Evaluation

- One problem with BLEU/METEOR - not specific:
  - What can we learn from X>Y? Why X>Y?
- An alternative - measure a specific **linguistic phenomena**
- General Idea - for each source sentence:
  - Create two translation options - one correct, one wrong
  - Score each option with the MT system

**How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs**

**Rico Sennrich**
School of Informatics, University of Edinburgh
{rico.sennrich}@ed.ac.uk

**Transliteration:**
Aumann - אומן
Yonat - יונת

**Subject-verb agreement:**
She - הלכה, אכלה, ישבה
He - הלך, אכל, ישב

**Polarity/Negation:**
Uncertain - לא בטוח

**Ambiguous words:**

| Example containing **ambiguous word** | Correct translations | Incorrect translations |
|---|---|---|
| It occurred to me that my **watch** might be broken. | Armbanduhr, Uhr | *Wache* |
| I hope you didn't get distracted during your **watch**. | *Wache* | Armbanduhr, Uhr |

# Contrastive Evaluation

- One problem with BLEU/METEOR - not specific:
  - What can we learn from X>Y? Why X>Y?
- An alternative - measure a specific **linguistic phenomena**
- General Idea - for each source sentence:
  - Create two translation options - one correct, one wrong
  - Score each option with the MT system
  - Count how many times the system preferred the correct option

**How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs**

**Rico Sennrich**
School of Informatics, University of Edinburgh
{rico.sennrich}@ed.ac.uk

**Transliteration:**
Aumann - אומן
Yonat - יונת

**Subject-verb agreement:**
She - הלכה, אכלה, ישבה
He - הלך, אכל, ישב

**Polarity/Negation:**
Uncertain - לא בטוח

**Ambiguous words:**

| Example containing **ambiguous word** | Correct translations | Incorrect translations |
|---|---|---|
| It occurred to me that my **watch** might be broken. | Armbanduhr, Uhr | *Wache* |
| I hope you didn't get distracted during your **watch**. | *Wache* | Armbanduhr, Uhr |

# Contrastive Evaluation

- One problem with BLEU/METEOR - not specific:
  - What can we learn from X>Y? Why X>Y?
- An alternative - measure a specific **linguistic phenomena**
- General Idea - for each source sentence:
  - Create two translation options - one correct, one wrong
  - Score each option with the MT system
  - Count how many times the system preferred the correct option
- Problem?

**How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs**

Rico Sennrich
School of Informatics, University of Edinburgh
{rico.sennrich}@ed.ac.uk

**Transliteration:**
Aumann - אומן
Yonat - יונת

**Subject-verb agreement:**
She - הלכה, אכלה, ישבה
He - הלך, אכל, ישב

**Polarity/Negation:**
Uncertain - לא בטוח

**Ambiguous words:**

| Example containing **ambiguous word** | Correct translations | Incorrect translations |
|---|---|---|
| It occurred to me that my **watch** might be broken. | Armbanduhr, Uhr | *Wache* |
| I hope you didn't get distracted during your **watch**. | *Wache* | Armbanduhr, Uhr |

# Summary

# Summary

- Evaluation is important!

# Summary

- Evaluation is important!
- Human evaluation is best, but: expensive, slow, subjective

# Summary

- Evaluation is important!
- Human evaluation is best, but: expensive, slow, subjective
- Automatic evaluation is cheap, fast and objective

# Summary

- Evaluation is important!
- Human evaluation is best, but: expensive, slow, subjective
- Automatic evaluation is cheap, fast and objective
    - BLEU is not perfect, but very popular

# Summary

- Evaluation is important!
- Human evaluation is best, but: expensive, slow, subjective
- Automatic evaluation is cheap, fast and objective
    - BLEU is not perfect, but very popular
    - Contrastive evaluation is informative

# Summary

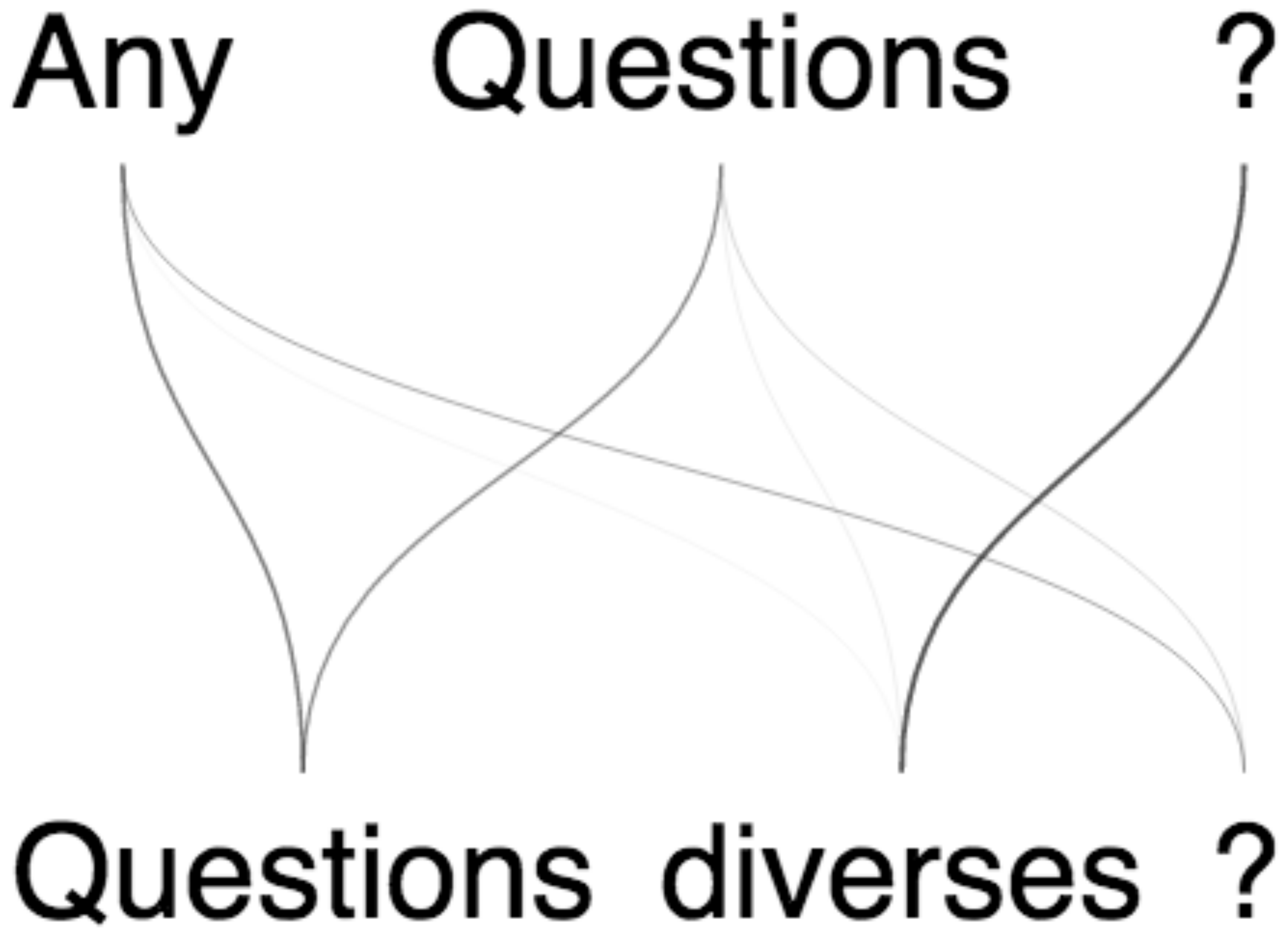
- Evaluation is important!
- Human evaluation is best, but: expensive, slow, subjective
- Automatic evaluation is cheap, fast and objective
    - BLEU is not perfect, but very popular
    - Contrastive evaluation is informative
- Human parity is here?

# Summary

- Evaluation is important!
- Human evaluation is best, but: expensive, slow, subjective
- Automatic evaluation is cheap, fast and objective
  - BLEU is not perfect, but very popular
  - Contrastive evaluation is informative
- Human parity is here?
  - Only in the sentence level, for high resource languages

# Summary

- Evaluation is important!
- Human evaluation is best, but: expensive, slow, subjective
- Automatic evaluation is cheap, fast and objective
  - BLEU is not perfect, but very popular
  - Contrastive evaluation is informative
- Human parity is here?
  - Only in the sentence level, for high resource languages
  - More work is needed!


Any Questions ?
Questions diverses ?