Study of Translation Edit Rate with Targeted Human Annotation

Matthew Snover (UMD)
Bonnie Dorr (UMD)
Richard Schwartz (BBN)
Linnea Micciulla (BBN)
John Makhoul (BBN)

Outline

- Motivations
- Definition of Translation Edit Rate (TER)
- Human-Targeted TER (HTER)
- Comparisons with BLEU and METEOR
- Correlations with Human Judgments

Motivations

- Subjective human judgments of performance have been the gold standard of MT evaluation metrics
- However…
  - Human Judgments are coarse grained
  - Meaning and fluency judgments tend to be conflated
  - Poor interannotator agreement at the segment level
- We want a more objective and repeatable human measure of fluency and meaning
  - We want a measure of the amount of work needed to fix a translation to make it both fluent and correct
  - Count the number of edits for a human to fix the translation

What is (H)TER?

- **Translation Edit Rate (TER): Number of edits needed to change a system output so that it exactly matches a given reference**
  - **MT research has become increasingly phrased-based, and we want a notion of edits that captures that**
  - **Allow movement of phrases using shifts**

- **Human-targeted TER (HTER): Minimal number of edits needed to change a system output so that it is fluent and has correct meaning**
  - **Infinite number of references could be used to find the one-best reference to count minimum number of edits**
  - **We have normally have 4 references at most though**
  - **Generate new targeted reference that is very close to system output**
  - **Measure TER between targeted reference and system output**

Formula of Translation Edit Rate (TER)

- With more than one reference:
  - TER = <# of edits> / <avg # of reference words>
  - TER is calculated against best (closest) reference
- Edits include insertions, deletions, substitutions and shifts
  - All edits count as 1 edit
  - Shift moves a sequence of words within the hypothesis
  - Shift of any sequence of words (any distance) is only 1 edit
- Capitalization and punctuation errors are included

Why Use Shifts?

- WER too harsh when output is distorted from reference
- With WER, no credit is given to the system when it generates the right string in the wrong place

Why Use Shifts?

```
REF:           saudi arabia denied this week
  information published in the american new york
  times

HYP: this week the   saudis denied
  information published in the          new york
  times
```

- WER too harsh when output is distorted from reference
- With WER, no credit is given to the system when it generates the right string in the wrong place

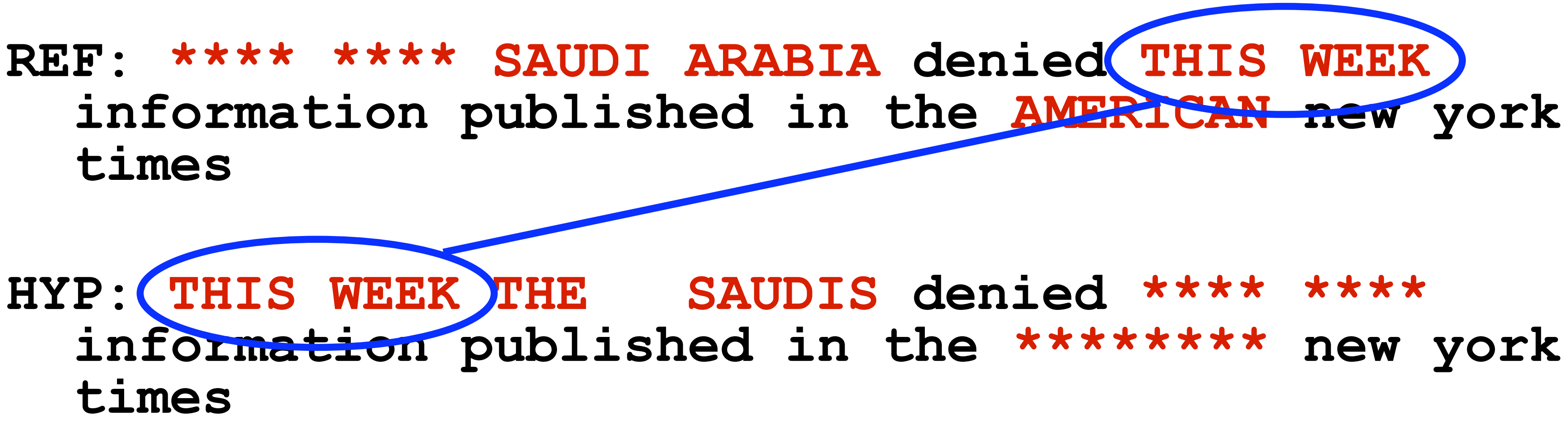
Why Use Shifts?

```
REF: **** **** SAUDI ARABIA denied THIS WEEK
   information published in the AMERICAN new york
   times

HYP: THIS WEEK THE   SAUDIS denied **** ****
   information published in the ******** new york
   times
```

- **WER too harsh when output is distorted from reference**
- **With WER, no credit is given to the system when it generates the right string in the wrong place**

Why Use Shifts?



```
REF: **** **** SAUDI ARABIA denied THIS WEEK
   information published in the AMERICAN new york
   times

HYP: THIS WEEK THE    SAUDIS denied **** ****
   information published in the ******** new york
   times
```

- WER too harsh when output is distorted from reference
- With WER, no credit is given to the system when it generates the right string in the wrong place

- TER shifts reflect the editing action of moving the string from one location to another

Example

```
REF:           saudi arabia denied  this week
  information published in the american new york
  times

HYP: this week the   saudis denied
  information published in the          new york
  times
```

Example

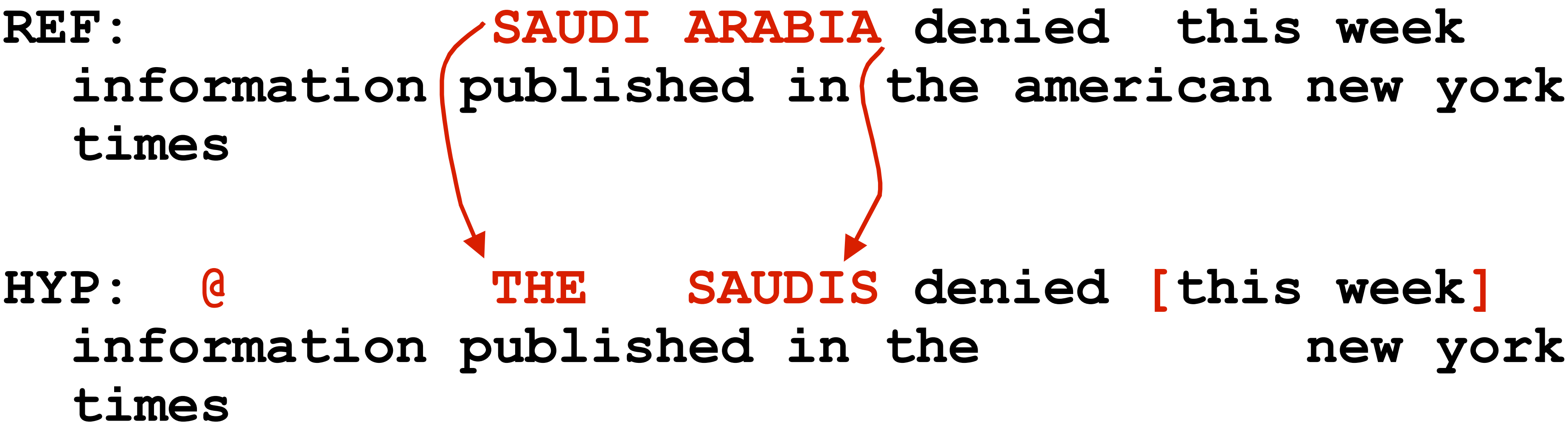
```
REF:              saudi arabia denied  this week
   information published in the american new york
   times

HYP:   @          the    saudis denied [this week]
   information published in the          new york
   times
```

Edits:

- **Shift “this week” to after “denied”**

Example



```
REF:           SAUDI ARABIA denied  this week
  information published in the american new york
  times

HYP:  @        THE   SAUDIS denied [this week]
  information published in the          new york
  times
```

Edits:

- **Shift "`this week`" to after "`denied`"**
- **Substitute "`Saudi Arabia`" for "`the Saudis`"**

Example

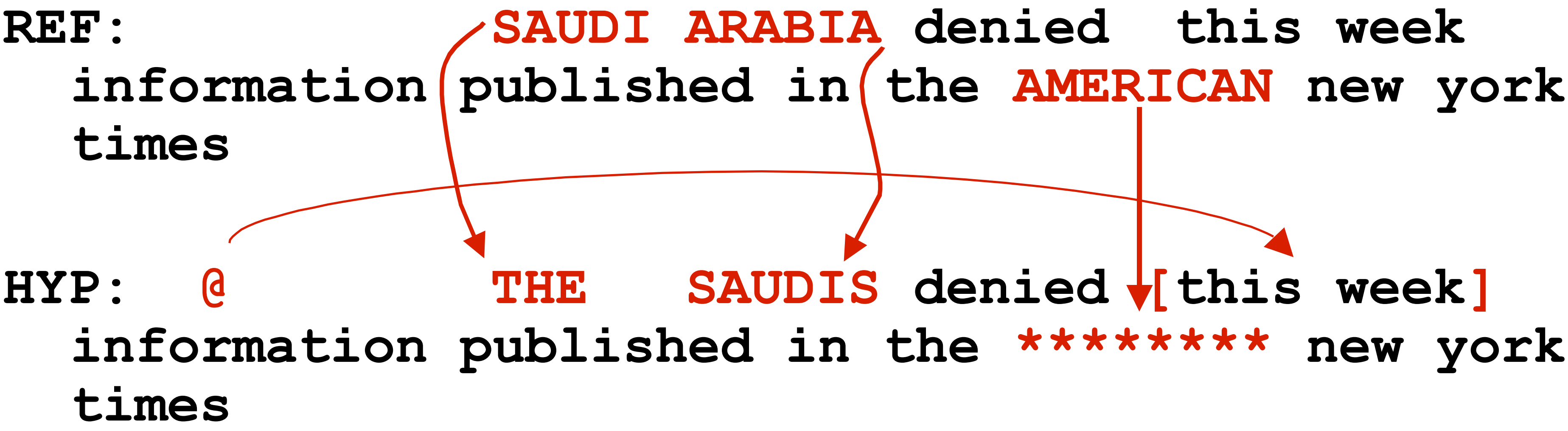
```
REF:            SAUDI ARABIA denied  this week
   information published in the AMERICAN new york
   times

HYP:  @         THE   SAUDIS denied ↓[this week]
   information published in the ********* new york
   times
```

Edits:

- Shift “this week” to after “denied”
- Substitute “Saudi Arabia” for “the Saudis”
- Insert “American”

Example



```
REF:          SAUDI ARABIA denied  this week
   information published in the AMERICAN new york
   times

HYP:   @          THE    SAUDIS denied [this week]
   information published in the ******** new york
   times
```

Edits:

- Shift “this week” to after “denied”
- Substitute “Saudi Arabia” for “the Saudis”
- Insert “American”

- 1 Shift, 2 Substitutions, 1 Insertion
  - 4 Edits (TER = 4/13 = 31%)

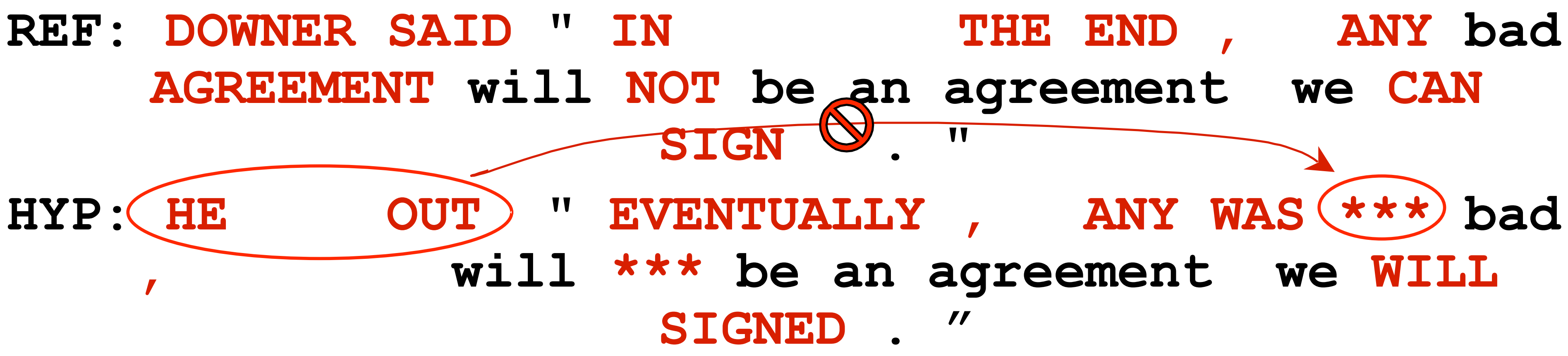
Calculation of Number of Edits

- **Optimal sequence of edits (with shifts) is very expensive to find**
- **Use a greedy search to select the set of shifts**
  - **At each step, calculate min-edit (Levenshtein) distance (number of insertions, deletions, substitutions) using dynamic programming**
  - **Choose shift that most reduces min-edit distance**
  - **Repeat until no shift remains that reduces min-edit distance**
- **After all shifting is complete, the number of edits is the number of shifts plus the remaining edit distance**

Shift Constraints

```
REF: DOWNER SAID " IN           THE END ,    ANY bad
    AGREEMENT will NOT be an agreement  we CAN
                 SIGN    . "
HYP: HE      OUT   " EVENTUALLY ,    ANY WAS *** bad
    ,         will *** be an agreement  we WILL
                 SIGNED . ″
```

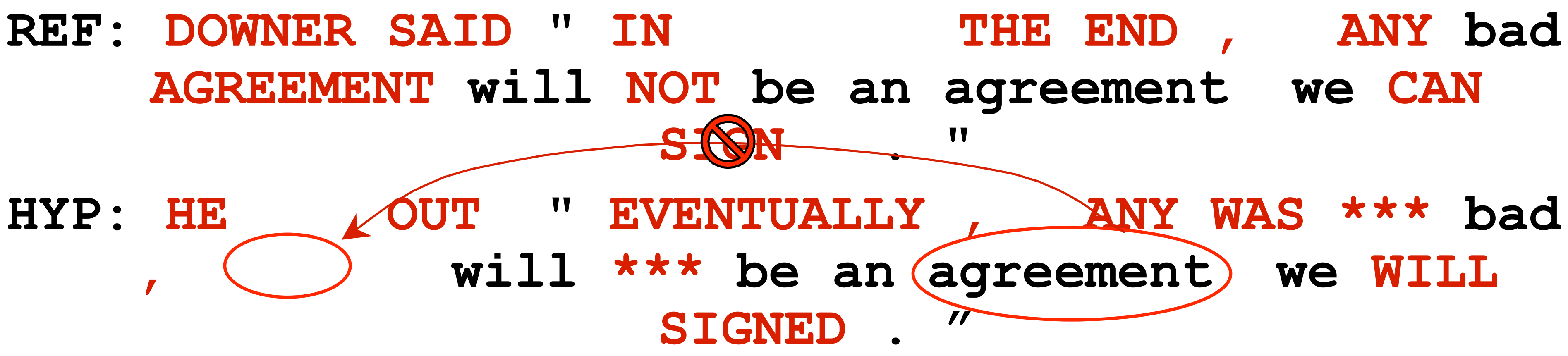
Shift Constraints



```
REF: DOWNER SAID " IN             THE END ,   ANY bad
     AGREEMENT will NOT be an agreement  we CAN
                  SIGN      . "
HYP: HE      OUT  " EVENTUALLY ,   ANY WAS *** bad
     ,             will *** be an agreement  we WILL
                  SIGNED . "
```

- **Shifted words must match the reference words in the destination position exactly**

Shift Constraints

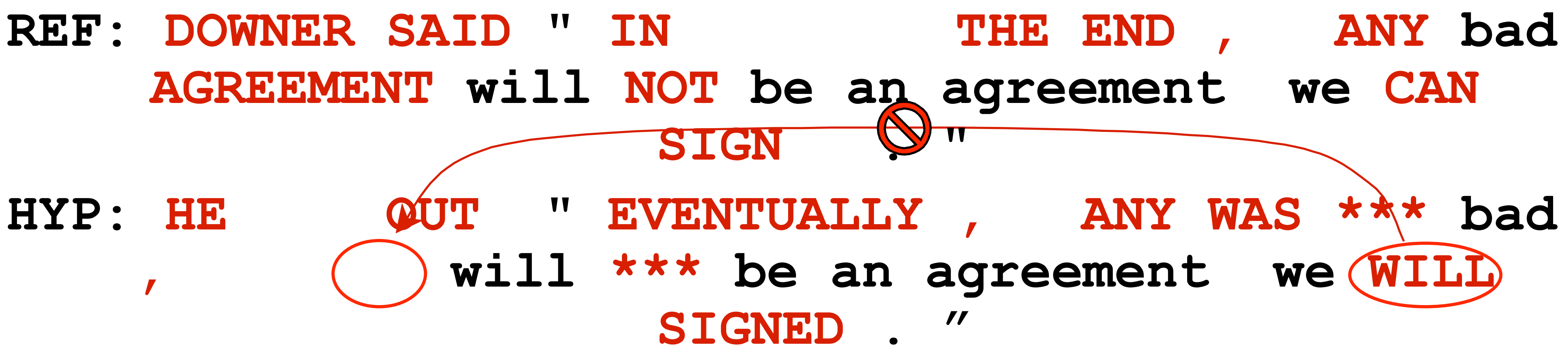


```
REF: DOWNER SAID " IN            THE END ,    ANY bad
     AGREEMENT will NOT be an agreement  we CAN
                    SIGN      .  "

HYP: HE     OUT   " EVENTUALLY ,    ANY WAS *** bad
     ,             will *** be an agreement  we WILL
                    SIGNED  .  "
```

- **Shifted words must match the reference words in the destination position exactly**
- **The word sequence of the hypothesis in the original position and the corresponding reference words must not match**

Shift Constraints

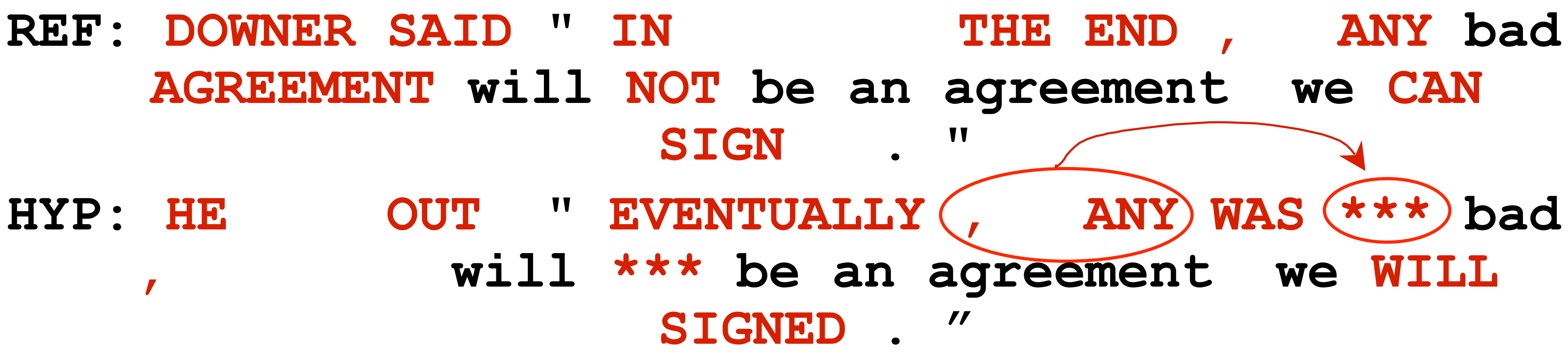


REF: DOWNER SAID " IN THE END , ANY bad AGREEMENT will NOT be an agreement we CAN SIGN . "

HYP: HE OUT " EVENTUALLY , ANY WAS *** bad , will *** be an agreement we WILL SIGNED . "

- Shifted words must match the reference words in the destination position exactly
- The word sequence of the hypothesis in the original position and the corresponding reference words must not match
- The word sequence of the reference that corresponds to the destination position must be misaligned before the shift

Shift Constraints



```
REF:  DOWNER SAID " IN            THE END ,     ANY bad
     AGREEMENT will NOT be an agreement  we CAN
                     SIGN    . "
HYP:  HE      OUT  " EVENTUALLY ,    ANY WAS *** bad
     ,          will *** be an agreement  we WILL
                     SIGNED . "
```

- **Shifted words must match the reference words in the destination position exactly**
- **The word sequence of the hypothesis in the original position and the corresponding reference words must not match**
- **The word sequence of the reference that corresponds to the destination position must be misaligned before the shift**

HTER: Human-targeted TER

- Procedure to create *targeted references*
  - Start with an automatic system output (hypothesis) and one or more human references.
  - Fluent speaker of English creates a new reference translation targeted for this system output by editing the hypothesis until it is fluent and has the same meaning as the reference(s)
  - Targeted references not required to be elegant English
- Compute minimum TER including new reference

Post-Editing Tool

- Post-Editing tool displays all references and hypothesis
- Tool shows where hypothesis differs from best reference
- Tool shows current TER for ‘reference in progress’
- Requires average 3-7 minutes per sentence to annotate
  - Time was relatively consistent over 4 annotators
  - Time could be reduced by a better post-editing tool

Post-Editing Tool

- Post-Editing tool displays all references and hypothesis
- Tool shows where hypothesis differs from best reference
- Tool shows current TER for 'reference in progress'
- Requires average 3-7 minutes per sentence to annotate
  - Time was relatively consistent over 4 annotators
  - Time could be reduced by a better post-editing tool

- Example:

Ref1: The expert, who asked not to be identified, added, "This depends on the conditions of the bodies."

Ref2: The experts who asked to remain unnamed said, "the matter is related to the state of the bodies."

Post-Editing Tool

- Post-Editing tool displays all references and hypothesis
- Tool shows where hypothesis differs from best reference
- Tool shows current TER for ‘reference in progress’
- Requires average 3-7 minutes per sentence to annotate
  - Time was relatively consistent over 4 annotators
  - Time could be reduced by a better post-editing tool

- Example:

Ref1: The expert, who asked not to be identified, added, "This depends on the conditions of the bodies."

Ref2: The experts who asked to remain unnamed said, "the matter is related to the state of the bodies."

Hyp: The expert who requested anonymity said that "the situation of the matter is linked to the dead bodies".

Post-Editing Tool

- Post-Editing tool displays all references and hypothesis
- Tool shows where hypothesis differs from best reference
- Tool shows current TER for ‘reference in progress’
- Requires average 3-7 minutes per sentence to annotate
  - Time was relatively consistent over 4 annotators
  - Time could be reduced by a better post-editing tool

- Example:

Ref1: The expert, who asked not to be identified, added, "This depends on the conditions of the bodies."

Ref2: The experts who asked to remain unnamed said, "the matter is related to the state of the bodies."

Hyp: The expert who requested anonymity said that "the situation of the matter is linked to the dead bodies".

Targ: The expert who requested anonymity said that "the matter is linked to the condition of the dead bodies".

Post-Editing Tool

- Post-Editing tool displays all references and hypothesis
- Tool shows where hypothesis differs from best reference
- Tool shows current TER for ‘reference in progress’
- Requires average 3-7 minutes per sentence to annotate
  - Time was relatively consistent over 4 annotators
  - Time could be reduced by a better post-editing tool

- Example:

Ref1: The expert, who asked not to be identified, added, "This depends on the conditions of the bodies."

Ref2: The experts who asked to remain unnamed said, "the matter is related to the state of the bodies."

Hyp: The expert who requested anonymity said that "the situation of the matter is linked to the dead bodies".

Targ: The expert who requested anonymity said that "the matter is linked to the condition of the dead bodies".

Post-Editing Instructions

- Three Requirements For Creating Targeted References
  1. Meaning in references must be preserved
  2. The targeted reference must be easily understood by a native speaker of English
  3. The Targeted Reference must be as close to the System Output as possible without violating 1 and 2.

- Grammaticality must be preserved
  - Acceptable: The two are leaving this evening
  - Not Acceptable: The two is leaving this evening
- Alternate Spellings (British or US or contractions) are allowed
- Meaning of targeted reference must be equivalent to at least one of the references

Targeted Reference Examples

- Four Palestinians were killed yesterday by Israeli army bullets during a military operation carried out by the Israeli army in the old town of Nablus .
- I tell you truthfully that reality is difficult the load is heavy and the territory is vibrant and gyrating .
- Iranian radio points to lifting 11 people alive from the debris in Bam

Experimental Design

- Two systems from MTEval 2004 Arabic
  - 100 randomly chosen sentences
  - Each system output was previously judged for fluency and adequacy by two human judges at NIST
  - S1 is one of the worst systems; S2 is one of the best
- Four annotators corrected system output
  - Two annotators for each sentence from each system
  - Annotators were undergraduates employed by BBN for annotation
- We ensured that the new targeted references were sufficiently accurate and fluent
  - Other annotators checked (and corrected) all targeted references for fluency and meaning
  - Second pass changed 0.63 words per sentence

Results (Average of S1 and S2)

	Ins	Del	Sub	Shift	TER
TER (4 UnTarg Ref)	4.6	12.0	25.8	7.2	49.6
HTER (1 Targ Ref)	3.0	8.2	8.9	4.9	33.5

- Insertion of Hypothesis Words (missing in reference)
- Deletion of Reference Words (missing in hypothesis)

Results (Average of S1 and S2)

	Ins	Del	Sub	Shift	TER
TER (4 UnTarg Ref)	4.6	12.0	25.8	7.2	49.6
HTER (1 Targ Ref)	3.0	8.2	8.9	4.9	33.5

- Insertion of Hypothesis Words (missing in reference)
- Deletion of Reference Words (missing in hypothesis)
- TER reduced by 33% using targeted references

Results (Average of S1 and S2)

	Ins	Del	Sub	Shift	TER
TER (4 UnTarg Ref)	4.6	12.0	25.8	7.2	49.6
HTER (1 Targ Ref)	3.0	8.2	8.9	4.9	33.5

- Insertion of Hypothesis Words (missing in reference)
- Deletion of Reference Words (missing in hypothesis)
- TER reduced by 33% using targeted references
  - 33% of edits using untargeted references are due to small sample of references

Results (Average of S1 and S2)

	Ins	Del	Sub	Shift	TER
TER (4 UnTarg Ref)	4.6	12.0	25.8	7.2	49.6
HTER (1 Targ Ref)	3.0	8.2	8.9	4.9	33.5

- Insertion of Hypothesis Words (missing in reference)
- Deletion of Reference Words (missing in hypothesis)
- TER reduced by 33% using targeted references
  - 33% of edits using untargeted references are due to small sample of references
  - Substitutions reduced by largest factor

Results (Average of S1 and S2)

	Ins	Del	Sub	Shift	TER
TER (4 UnTarg Ref)	4.6	12.0	25.8	7.2	49.6
HTER (1 Targ Ref)	3.0	8.2	8.9	4.9	33.5

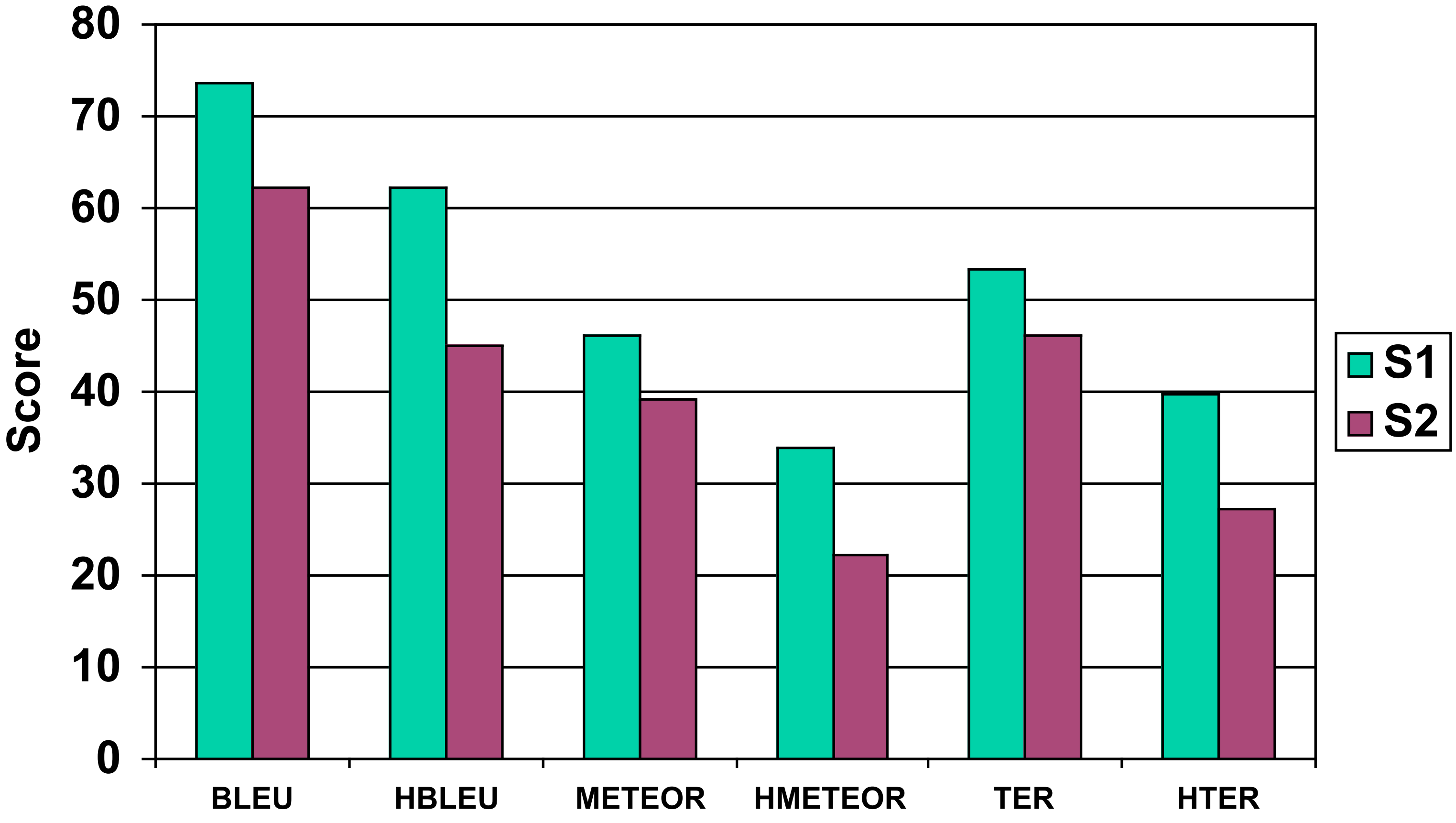
- Insertion of Hypothesis Words (missing in reference)
- Deletion of Reference Words (missing in hypothesis)
- TER reduced by 33% using targeted references
  - 33% of errors using untargeted references are due to small sample of references
  - Substitutions reduced by largest factor
- Majority of edits are substitutions and deletions

BLEU and METEOR

- BLEU (Papineni et al. 2002)
  - Counts number of n-grams (size 1-4) of the system output that match in the reference set
  - Contributed to recent improvements in MT
- METEOR (Banerjee and Lavie 2005)
  - Counts number of exact word matches between system output and reference
  - Unmatched words are stemmed, and then matched
  - Additional penalties for reordering words

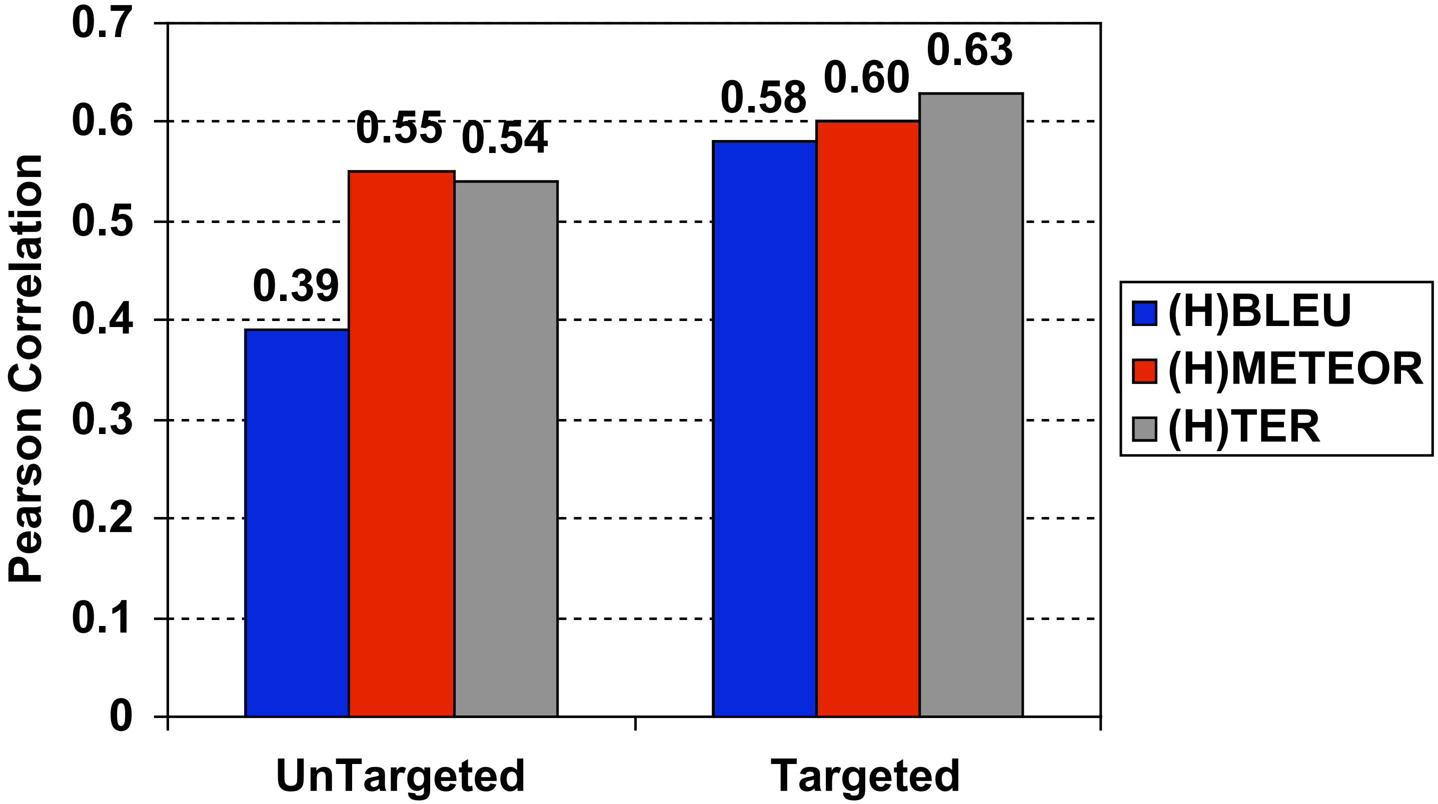
- To compare with error measures
  - 1.0 - BLEU and 1.0 - METEOR used in this talk
- HBLEU and HMETEOR
  - BLEU and METEOR when using human-targeted references

System Scores



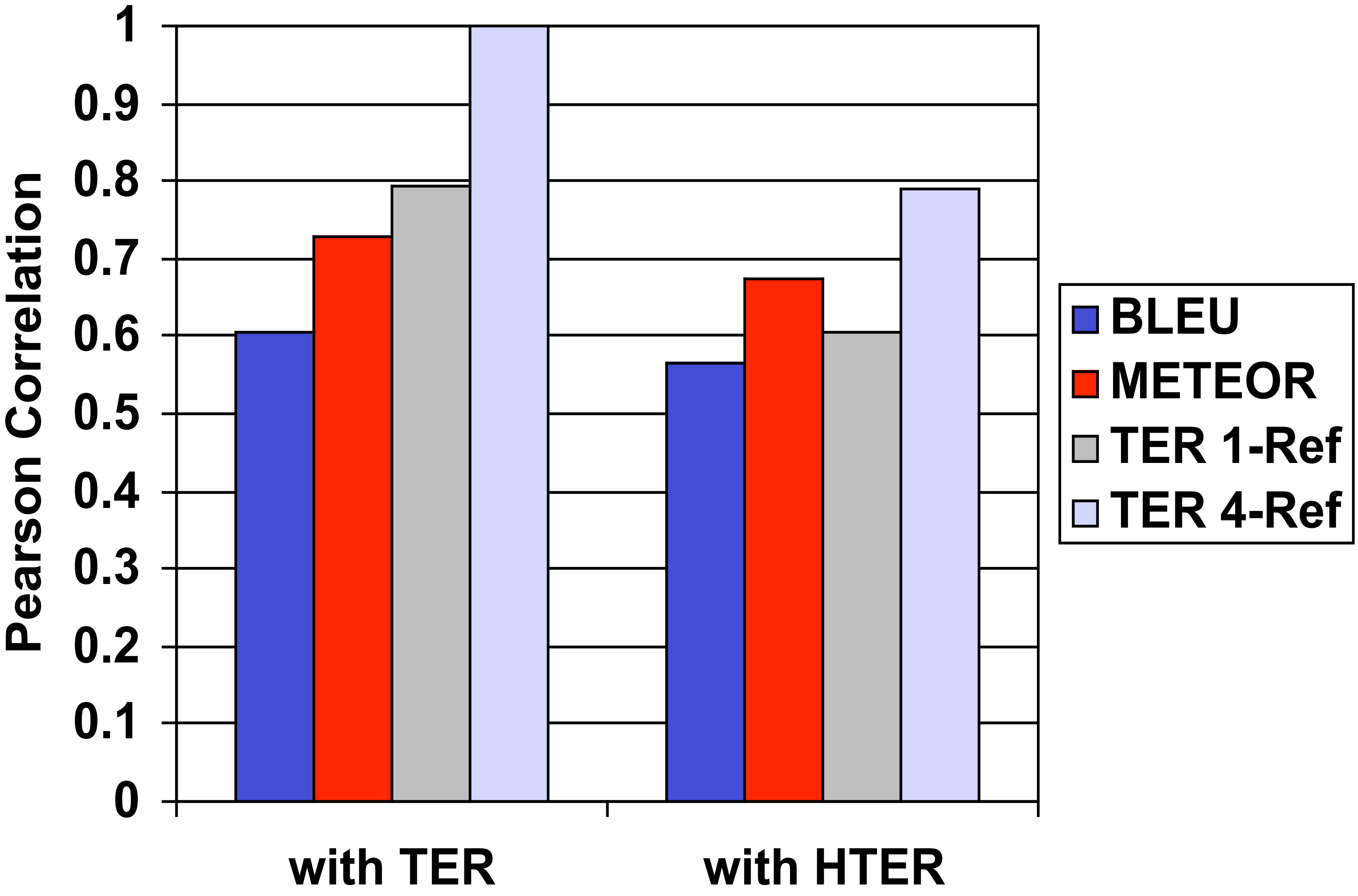
- 1.0 - BLEU and 1.0 - METEOR shown
- Low scores are better

Correlation with Human Judgments



- Segment Level Correlations (200 data points)
- Targeted correlations are the average of 2 correlations (2 targ refs)
- HTER correlates best with human judgments
- Targeted references increase correlation for evaluation metrics
- METEOR correlates better than TER
- HTER correlates better than HMETEOR

Correlation Between (H)TER / BLEU / Meteor



UNIVERSITY OF MARYLAND

BBN TECHNOLOGIES

Correlations between Human Judges

- **Each human judgment is the average of fluency and adequacy judgments**

	TER	HTER	BLEU	HBLEU	MET.	HMET.	HJ-1	HJ-2
HJ-1	0.46	0.51	0.34	0.46	0.50	0.51	1.00	0.48
HJ-2	0.47	0.58	0.33	0.53	0.45	0.53	0.48	1.00

Correlations between Human Judges

- Each human judgment is the average of fluency and adequacy judgments

	TER	HTER	BLEU	HBLEU	MET.	HMET.	HJ-1	HJ-2
HJ-1	0.46	0.51	0.34	0.46	0.50	0.51	1.00	0.48
HJ-2	0.47	0.58	0.33	0.53	0.45	0.53	0.48	1.00

- Subjective human judgments are noisy
  - Exhibit lower correlation than might be expected

Correlations between Human Judges

- Each human judgment is the average of fluency and adequacy judgments

	TER	HTER	BLEU	HBLEU	MET.	HMET.	HJ-1	HJ-2
HJ-1	0.46	0.51	0.34	0.46	0.50	0.51	1.00	0.48
HJ-2	0.47	0.58	0.33	0.53	0.45	0.53	0.48	1.00

- Subjective human judgments are noisy
  - Exhibit lower correlation than might be expected
- HTER correlates a little better with a single human judgment than another human judgment does
  - Rather than having judges give subjective scores, they should create targeted references

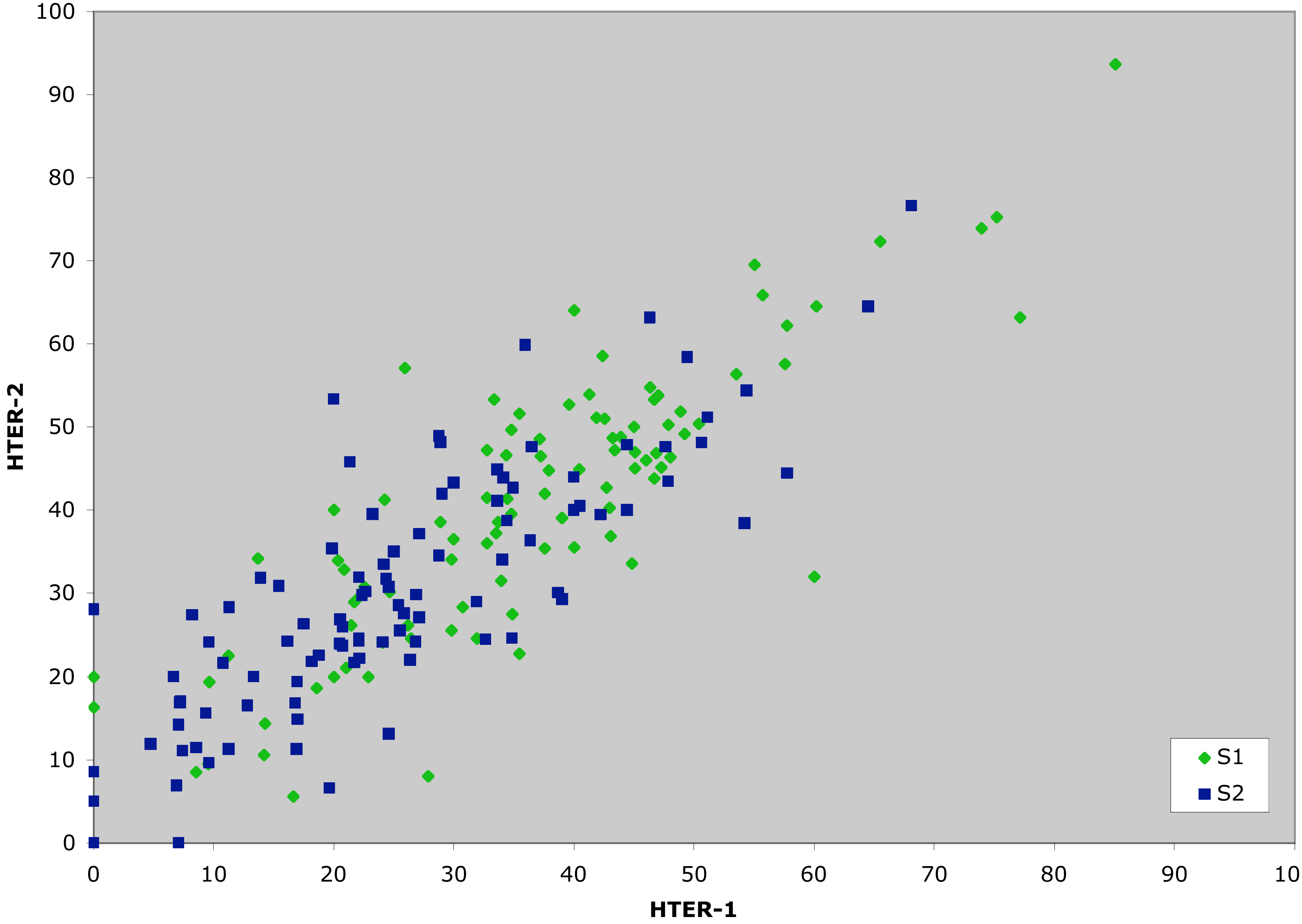
Correlations between Human Judges

- Each human judgment is the average of fluency and adequacy judgments

	TER	HTER	BLEU	HBLEU	MET.	HMET.	HJ-1	HJ-2
HJ-1	0.46	0.51	0.34	0.46	0.50	0.51	1.00	0.48
HJ-2	0.47	0.58	0.33	0.53	0.45	0.53	0.48	1.00

- Subjective human judgments are noisy
  - Exhibit lower correlation than might be expected
- HTER correlates a little better with a single human judgment than another human judgment does
  - Rather than having judges give subjective scores, they should create targeted references
- TER correlates with single human judgment about as well as another human judgment

Correlation Between HTER Post-Editors



r = 0.85

Examining MT Errors with HTER

- Subjective human judgments aren’t useful for diagnosing MT errors
- HTER indicates portion of output that is incorrect

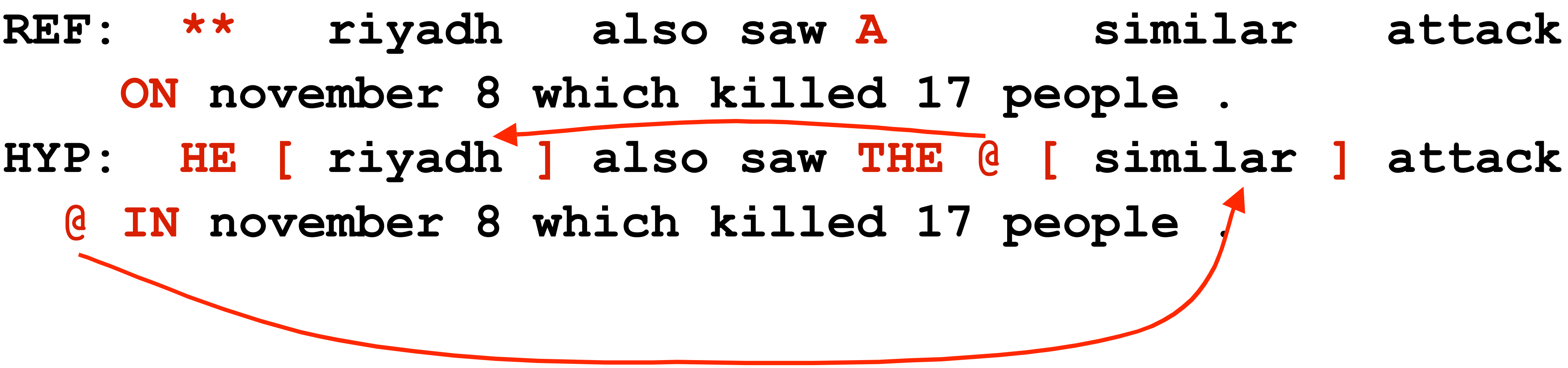
Hypothesis: he also saw the riyadh attack similar in november 8 which killed 17 people .

```
REF:       riyadh   also saw a       similar   attack
    on november 8 which killed 17 people .
HYP:  he   riyadh   also saw the     similar   attack
    in november 8 which killed 17 people .
```

Examining MT Errors with HTER

- Subjective human judgments aren’t useful for diagnosing MT errors
- HTER indicates portion of output that is incorrect

Hypothesis: he also saw the riyadh attack similar in november 8 which killed 17 people .



```
REF:   **   riyadh   also saw A        similar   attack
     ON november 8 which killed 17 people .
HYP:   HE [ riyadh ] also saw THE @ [ similar ] attack
   @ IN november 8 which killed 17 people .
```

Conclusions

- Targeted References decreases TER by 33%
  - In all subsequent studies TER reduction is ~50%
- HTER has high correlation with human judgments
  - But is very expensive
  - Targeted references not readily reusable
- HTER makes fine distinctions among correct, near correct, bad translations
  - Correct translations have HTER = 0
  - Bad translations have high HTER
  - May be substitute for Subjective Human Judgments
- HTER is easy to explain to people outside of MT community:
  - Amount of work to correct the translations

Future Work and Impact

- Compute HTER and Human Judgment correlations at the system level, rather than segment level
  - Caveat: HTER expensive to generate for many systems
- Better post-editing tool
  - Suggests edits to the annotator
- Investigate non-uniform weights for (H)TER
- HTER currently used in GALE Evaluation
- TER computation code available at http://www.cs.umd.edu/~snover/tercom

Questions