Results of the WMT16 Metrics Shared Task

Ondřej Bojar
Yvette Graham
Amir Kamran
Miloš Stanojević

WMT16, Aug 11, 2016
Overview

- Summary of Metrics Task.
- Updates to Metric Task in 2016.
- Results
Metrics Task in a Nutshell
Metrics Task in a Nutshell

MT Outputs

Reference Translation
Metrics Task in a Nutshell

MT Outputs

Reference Translation

Metrics Developers
Metrics Task in a Nutshell

MT Outputs
Reference Translation

Metric Scores
Metric Scores
Metric Scores

Metrics Developers
Metrics Task in a Nutshell

MT Outputs

Reference Translation

Metrics Scores

Metrics Scores

Metrics Scores

Metrics Developers
Metrics Task in a Nutshell

Metrics Developers
Metrics Task in a Nutshell

MT Outputs
Reference Translation

Metrics Developers
Metric Scores

Human Judgements
Metrics’ Correlations
System- and Segment-Level Evaluation

- **System Level**
  - Participants compute one score for the whole test set, as translated by each of the systems

```
Econo
For example
The new in
The company m
From Friday's job
“The unification
Čermák, which
New common D
```

- **Segment Level**

```
0.387
```

```
Econo
For example
The new in
The company m
From Friday's job
“The unification
Čermák, which
New common D
```

```
0.211
0.583
0.286
0.387
0.354
0.221
0.438
0.144
```

```
0.354
0.221
0.438
0.144
```
System- and Segment-Level Evaluation

▶ System Level
  ▶ Participants compute one score for the whole test set, as translated by each of the systems

▶ Segment Level
  ▶ Participants compute one score for each sentence of each system’s translation
## Nine Years of Metrics Task

<table>
<thead>
<tr>
<th></th>
<th>'07</th>
<th>'08</th>
<th>'09</th>
<th>'10</th>
<th>'11</th>
<th>'12</th>
<th>'13</th>
<th>'14</th>
<th>'15</th>
<th>'16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participating Teams</td>
<td>-</td>
<td>6</td>
<td>8</td>
<td>14</td>
<td>9</td>
<td>8</td>
<td>12</td>
<td>12</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>Evaluated Metrics</td>
<td>11</td>
<td>16</td>
<td>38</td>
<td>26</td>
<td>21</td>
<td>12</td>
<td>16</td>
<td>23</td>
<td>46</td>
<td>16</td>
</tr>
<tr>
<td>Baseline Metrics</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**System-level**
- Spearman Rank Corr. • • • • • • • ○
- Pearson Corr. Coeff. ○ • • • •

**Segment-level**
- Ratio of Concordant Pairs •
- Kendall’s $\tau$ • • • • * * * *
- Pearson Corr. Coeff. •

* • main and ○ secondary score reported for the system-level evaluation.
* •, *, and * are slightly different variants regarding ties.

- Stable number of participating teams.
- A growing set of “baseline metrics”.
- Stable but gradually improving evaluation methods.
Updates to Metrics Task in 2016

- More Domains
  - News, IT, Medical.
- Two Golden Truths in News Task
  - Relative Ranking, Direct Assessment.
- Third golden truth in Medical Domain.
- Confidence for Sys-level Computed Differently.
  - Participants needed to score 10K systems.
- More languages (18 pairs):
  - Basque, Bulgarian, Czech, Dutch, Finnish, German, Polish, Portuguese, Romanian, Russian, Spanish, and Turkish
  - Paired with English in one or both directions.
## Metrics Task Madness

<table>
<thead>
<tr>
<th>Track</th>
<th>Test set</th>
<th>News Task</th>
<th>Tuning Task</th>
<th>IT Task</th>
<th>HimL</th>
<th>Year 1</th>
<th>Hybrid</th>
<th>cs</th>
<th>de</th>
<th>ro</th>
<th>fi</th>
<th>ru</th>
<th>tr</th>
<th>English into-English</th>
<th>bg</th>
<th>es</th>
<th>eu</th>
<th>nl</th>
<th>pl</th>
<th>pt</th>
</tr>
</thead>
<tbody>
<tr>
<td>RRsysNews</td>
<td>newstest2016</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>RRsysIT</td>
<td>it-test2016</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>DAsysNews</td>
<td>newstest2016</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>RRsegNews</td>
<td>newstest2016</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>DAsegNews</td>
<td>newstest2016</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>HUMEseg</td>
<td>himl2015</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>

“✓”: sets of underlying MT systems
“●”: language pairs covered in the evaluation
“•”: language pairs planned but abandoned
## Metrics Task Madness

<table>
<thead>
<tr>
<th>Track</th>
<th>Test set</th>
<th>News Task</th>
<th>Tuning Task</th>
<th>IT Task</th>
<th>HimL</th>
<th>Year 1</th>
<th>Hybrid</th>
<th>cs</th>
<th>de</th>
<th>ro</th>
<th>fi</th>
<th>ru</th>
<th>tr</th>
<th>English into-English</th>
<th>English into</th>
</tr>
</thead>
<tbody>
<tr>
<td>RRsysNews</td>
<td>newest2016</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RRsysIT</td>
<td>it-test2016</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DAsysNews</td>
<td>newest2016</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RRsegNews</td>
<td>newest2016</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DAsegNews</td>
<td>newest2016</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HUMEseg</td>
<td>himl2015</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

“✓”: sets of underlying MT systems  
“•”: language pairs covered in the evaluation  
“·”: language pairs planned but abandoned

For participants, this was cut down to the standard:

**Sys-level**  
0.387

and **seg-level**  
0.211, 0.583, 0.286, 0.387, 0.354, 0.221, 0.438, 0.144

Econo...
Metrics Task Domains

- **WMT16 News Task**
  - Systems and language pairs from the main translation task.
  - Truth: Primarily RR, DA into English and Russian.

- **WMT16 IT Task**
  - IT domain.
  - Only out of English.
  - Interesting target languages: (Czech, German,) Bulgarian, Spanish, Basque, Dutch, Portuguese.
  - Truth: Only RR

- **HimL Medical Texts**
  - Just one system per target language.
  - (So only seg-level evaluation.)
  - Truth: A new semantics-based metric.
Golden Truths

- **Relative Ranking (RR)**
  - 5-way relative comparison.
    - Interpreted as 10 pairwise comparison.
    - Identical outputs deduplicated.
    - Finally converted to a score using TrueSkill.

- **Direct Assessment (DA)**
  - Absolute adequacy judgement over individual sentences.
    - Judgements from each worker standardized.
    - Multiple judgements of a candidate averaged.
    - Finally averaged over all sentences of a system.
  - Fluency optionally to resolve ties.
  - Provided by Turkers (only English and Russian).
    - Planned but not done with Researchers.

- **HUME**
  - A composite score of manual judgements of meaning preservation.
  - Used only in the “medical” track.
Effects of DA vs. RR for Metrics Task

Benefits:
- More principled golden truth.
- Possibly more reliable, assuming enough judgements.

Negative aspects:
- Sampling for sys-level and seg-level is different.
- Perhaps impossible for seg-level out of English:
  - Too few Turker annotations.
  - Too few researchers. (Repeated judgements work as well.)

This year, only English and Russian news systems have DA judgements.
<table>
<thead>
<tr>
<th>Metric</th>
<th>Participant</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEER</td>
<td>ILLC – UvA (Stanojević and Sima’an, 2015)</td>
</tr>
<tr>
<td>Character</td>
<td>RWTH Aachen University (Wang et al., 2016)</td>
</tr>
<tr>
<td>chrF1,2,3</td>
<td>Humboldt University of Berlin (Popović, 2016)</td>
</tr>
<tr>
<td>wordF1,2,3</td>
<td>Humboldt University of Berlin (Popović, 2016)</td>
</tr>
<tr>
<td>DepCheck</td>
<td>Charles University, no corresponding paper</td>
</tr>
<tr>
<td>DPMFcomb-without-RED</td>
<td>Chinese Academy of Sciences and Dublin City University (Yu et al., 2015)</td>
</tr>
<tr>
<td>MPEDA</td>
<td>Jiangxi Normal University (Zhang et al., 2016)</td>
</tr>
<tr>
<td>UoW.ReVal</td>
<td>University of Wolverhampton (Gupta et al., 2015)</td>
</tr>
<tr>
<td>upf-cobalt</td>
<td>Universitat Pompeu Fabra (Fomicheva et al., 2016)</td>
</tr>
<tr>
<td>CobaltF</td>
<td>Universitat Pompeu Fabra (Fomicheva et al., 2016)</td>
</tr>
<tr>
<td>MetricsF</td>
<td>Universitat Pompeu Fabra (Fomicheva et al., 2016)</td>
</tr>
<tr>
<td>DTED</td>
<td>University of St Andrews, (McCaffery and Nederhof, 2016)</td>
</tr>
</tbody>
</table>
Standard Presentation of the Results

<table>
<thead>
<tr>
<th></th>
<th>cs-en</th>
<th>de-en</th>
<th>fi-en</th>
<th>ro-en</th>
<th>ru-en</th>
<th>tr-en</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RR</td>
<td>DA</td>
<td>RR</td>
<td>DA</td>
<td>RR</td>
<td>DA</td>
</tr>
<tr>
<td>Human Systems</td>
<td>6</td>
<td>6</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>MPEDA</td>
<td>.996</td>
<td>.993</td>
<td>.956</td>
<td>.937</td>
<td>.967</td>
<td>.976</td>
</tr>
<tr>
<td>UoW.ReVal</td>
<td>.993</td>
<td>.986</td>
<td>.949</td>
<td>.985</td>
<td>.958</td>
<td>.970</td>
</tr>
<tr>
<td>BEER</td>
<td>.996</td>
<td>.990</td>
<td>.949</td>
<td>.879</td>
<td>.964</td>
<td>.972</td>
</tr>
<tr>
<td>ChrF1</td>
<td>.993</td>
<td>.986</td>
<td>.934</td>
<td>.868</td>
<td>.974</td>
<td>.980</td>
</tr>
<tr>
<td>ChrF2</td>
<td>.992</td>
<td>.989</td>
<td>.952</td>
<td>.893</td>
<td>.957</td>
<td>.967</td>
</tr>
<tr>
<td>ChrF3</td>
<td>.991</td>
<td>.989</td>
<td>.958</td>
<td>.902</td>
<td>.946</td>
<td>.958</td>
</tr>
<tr>
<td>Character</td>
<td>.997</td>
<td>.995</td>
<td>.985</td>
<td>.929</td>
<td>.921</td>
<td>.927</td>
</tr>
<tr>
<td>mtevalNIST</td>
<td>.988</td>
<td>.978</td>
<td>.887</td>
<td>.801</td>
<td>.924</td>
<td>.929</td>
</tr>
<tr>
<td>mtevalBLEU</td>
<td>.992</td>
<td>.989</td>
<td>.905</td>
<td>.808</td>
<td>.858</td>
<td>.864</td>
</tr>
<tr>
<td>MosesCDER</td>
<td>.995</td>
<td>.988</td>
<td>.927</td>
<td>.827</td>
<td>.846</td>
<td>.860</td>
</tr>
<tr>
<td>MosesTER</td>
<td>.983</td>
<td>.969</td>
<td>.926</td>
<td>.834</td>
<td>.852</td>
<td>.846</td>
</tr>
<tr>
<td>WordF2</td>
<td>.991</td>
<td>.985</td>
<td>.897</td>
<td>.786</td>
<td>.790</td>
<td>.806</td>
</tr>
<tr>
<td>WordF3</td>
<td>.991</td>
<td>.985</td>
<td>.898</td>
<td>.787</td>
<td>.786</td>
<td>.803</td>
</tr>
<tr>
<td>WordF1</td>
<td>.992</td>
<td>.984</td>
<td>.894</td>
<td>.780</td>
<td>.796</td>
<td>.808</td>
</tr>
<tr>
<td>MosesPER</td>
<td>.981</td>
<td>.970</td>
<td>.843</td>
<td>.730</td>
<td>.770</td>
<td>.767</td>
</tr>
<tr>
<td>MosesBLEU</td>
<td>.991</td>
<td>.983</td>
<td>.880</td>
<td>.757</td>
<td>.752</td>
<td>.759</td>
</tr>
<tr>
<td>MosesWER</td>
<td>.982</td>
<td>.967</td>
<td>.926</td>
<td>.822</td>
<td>.773</td>
<td>.768</td>
</tr>
</tbody>
</table>

- Bold in RR indicates “official winners”.
- Some setups fairly non-discerning, here e.g. csen:
  - All but ChrF1, ChrF3, MtevalNIST and MosesPER tie at top.
## News RR Winners Across Languages

<table>
<thead>
<tr>
<th>Metric</th>
<th># Wins</th>
<th>Language Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEER</td>
<td>11</td>
<td>csen, encs, ende, enfi, enro, enru, entr, fien, roen, ruen, tren</td>
</tr>
<tr>
<td>UoW.ReVal</td>
<td>6</td>
<td>csen, deen, fien, roen, ruen, tren</td>
</tr>
<tr>
<td>chrF2</td>
<td>6</td>
<td>csen, encs, enro, fien, ruen, tren</td>
</tr>
<tr>
<td>chrF1</td>
<td>5</td>
<td>enro, fien, ruen, tren</td>
</tr>
<tr>
<td>chrF3</td>
<td>4</td>
<td>deen, enfi, entr, ruen</td>
</tr>
<tr>
<td>mosesCDER</td>
<td>4</td>
<td>csen, enfi, enru, entr</td>
</tr>
<tr>
<td>CharacTer</td>
<td>3</td>
<td>csen, deen, roen</td>
</tr>
<tr>
<td>mosesBLEU</td>
<td>3</td>
<td>csen, encs, enfi</td>
</tr>
<tr>
<td>mosesPER</td>
<td>3</td>
<td>enro, ruen, tren</td>
</tr>
<tr>
<td>mtevalBLEU</td>
<td>3</td>
<td>csen, encs, enro</td>
</tr>
<tr>
<td>wordF1</td>
<td>3</td>
<td>csen, encs, enro</td>
</tr>
<tr>
<td>wordF2</td>
<td>3</td>
<td>csen, encs, enro</td>
</tr>
<tr>
<td>mosesTER</td>
<td>2</td>
<td>csen, encs</td>
</tr>
<tr>
<td>mtevalNIST</td>
<td>2</td>
<td>encs, tren</td>
</tr>
<tr>
<td>wordF3</td>
<td>2</td>
<td>csen, entr</td>
</tr>
<tr>
<td>mosesWER</td>
<td>1</td>
<td>csen</td>
</tr>
</tbody>
</table>
Graphical Presentation of Significant Wins

- Williams (1959) test of significant improvement in Pearson correlation.
  - Green cell indicates that the metric in the row has significantly better correlation than the metric in the column.

So for Czech-English RR, we have:
Williams (1959) test of significant improvement in Pearson correlation.

- Green cell indicates that the metric in the row has significantly better correlation than the metric in the column.

So for Czech-English RR, we have:
Williams (1959) test of significant improvement in Pearson correlation.

- Green cell indicates that the metric in the row has significantly better correlation than the metric in the column.

So for Czech-English RR, we have:
Czech-English Direct Assessments

With just 6 systems, correlations do not differ reliably.
Czech-English Direct Assessments

With just 6 systems, correlations do not differ reliably.
6 standard and 6 tuning task systems indicate:

CharacTer, BEER, and mosesCDER not outperformed by anyone else.
6 standard and 6 tuning task systems indicate: **CharacTer**, **BEER** and **mosesCDER** not outperformed by anyone else.
Czech-English DA with Hybrids

10,000 synthetized systems allow to find almost total ordering.
Czech-English DA with Hybrids

10,000 synthetized systems allow to find almost total ordering.
“Hybrids” = Hybrid Super-Sampling

- 10,000 “new systems” constructed by mixing sentences.
- Puts extra burden on task participants.
  - Need to score 10k “system” outputs, full test set each.
  - 200MB–1.1GB bzipped input file per language pair.
- Allows to distinguish sys-level metrics much better.
- Applicable to both RR and DA.
  - Done with DA only now because RR human judgements of individual sentences have to be carried over to these 10k systems.

Winners according to DA hybrids:

<table>
<thead>
<tr>
<th>Metric</th>
<th># Wins</th>
<th>Language Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoW.ReVal</td>
<td>3</td>
<td>deen, roen, ruen</td>
</tr>
<tr>
<td>CharacTer</td>
<td>2</td>
<td>csen, enru</td>
</tr>
<tr>
<td>MPEDA</td>
<td>2</td>
<td>fien, tren</td>
</tr>
<tr>
<td>BEER</td>
<td>1</td>
<td>tren</td>
</tr>
</tbody>
</table>
Ex. English-Russian RR BLEU
Ex. English-Russian RR CharacTer
Ex. English-Finnish RR BLEU

HUMAN

MTEVALBLEU

12 13 14 15 16 17

-0.3 -0.1 0.0 0.2 0.4

ONLINE-A

ONLINE-B

ONLINE-G

ABUMATRAN-PBSMT

ABUMATRAN-NMT

ABUMATRAN-COMBO

UH-OPUS

JHU-PBMT

UH-FACTORED

AALTO

JHU-HLTCOE

UUT

JHU-PBMT

NYU-UMONTREAL
Ex. English-Finnish RR chrF3

HUMAN

chrF3

-0.3

-0.1

0.0

0.2

0.4

47.00 47.78 48.56 49.34 50.12 50.90 51.68

ONLINE-G

ABUMATRAN-NMT

ONLINE-B

ABUMATRAN-COMBO

UH-OPUS

ONLINE-A

NYU-UMONTREAL

ABUMATRAN-PBSMT

JHU-PBMT

UH-FACTORED

JHU-HLTCOE

AALTO

UUT
Sys-Level Metrics on IT Task

- To test metrics in domain-specific setting.
- Unfortunately, often too few participating systems.

```
<table>
<thead>
<tr>
<th>Human</th>
<th>en-bg</th>
<th>en-cs</th>
<th>en-de</th>
<th>en-es</th>
<th>en-eu</th>
<th>en-nl</th>
<th>en-pt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systems</td>
<td>RR</td>
<td>RR</td>
<td>RR</td>
<td>RR</td>
<td>RR</td>
<td>RR</td>
<td>RR</td>
</tr>
<tr>
<td>CHARACTer</td>
<td>1.000</td>
<td>0.901</td>
<td>0.930</td>
<td>0.963</td>
<td>1.000</td>
<td>0.927</td>
<td>0.976</td>
</tr>
<tr>
<td>chrF3</td>
<td>1.000</td>
<td>0.831</td>
<td>0.700</td>
<td>0.938</td>
<td>1.000</td>
<td>0.961</td>
<td>0.990</td>
</tr>
<tr>
<td>chrF2</td>
<td>1.000</td>
<td>0.837</td>
<td>0.672</td>
<td>0.933</td>
<td>1.000</td>
<td>0.959</td>
<td>0.986</td>
</tr>
<tr>
<td>BEER</td>
<td>1.000</td>
<td>0.744</td>
<td>0.621</td>
<td>0.931</td>
<td>1.000</td>
<td>0.983</td>
<td>0.989</td>
</tr>
<tr>
<td>chrF1</td>
<td>1.000</td>
<td>0.845</td>
<td>0.588</td>
<td>0.915</td>
<td>1.000</td>
<td>0.951</td>
<td>0.967</td>
</tr>
<tr>
<td>mtevalNIST</td>
<td>1.000</td>
<td>0.905</td>
<td>0.524</td>
<td>0.926</td>
<td>1.000</td>
<td>0.722</td>
<td>0.993</td>
</tr>
<tr>
<td>MPEDA</td>
<td>1.000</td>
<td>0.620</td>
<td>0.599</td>
<td>0.951</td>
<td>1.000</td>
<td>0.856</td>
<td>0.989</td>
</tr>
<tr>
<td>mosesTER</td>
<td>1.000</td>
<td>0.616</td>
<td>0.628</td>
<td>0.908</td>
<td>1.000</td>
<td>0.835</td>
<td>0.994</td>
</tr>
<tr>
<td>mtevalBLEU</td>
<td>1.000</td>
<td>0.750</td>
<td>0.621</td>
<td>0.976</td>
<td>1.000</td>
<td>0.596</td>
<td>0.997</td>
</tr>
<tr>
<td>mosesWER</td>
<td>1.000</td>
<td>0.009</td>
<td>0.656</td>
<td>0.916</td>
<td>1.000</td>
<td>0.903</td>
<td>0.991</td>
</tr>
<tr>
<td>mosesCDER</td>
<td>1.000</td>
<td>0.181</td>
<td>0.652</td>
<td>0.932</td>
<td>1.000</td>
<td>0.914</td>
<td>0.997</td>
</tr>
<tr>
<td>wordF1</td>
<td>1.000</td>
<td>0.240</td>
<td>0.644</td>
<td>0.959</td>
<td>1.000</td>
<td>0.911</td>
<td>0.997</td>
</tr>
<tr>
<td>wordF2</td>
<td>1.000</td>
<td>0.266</td>
<td>0.652</td>
<td>0.965</td>
<td>1.000</td>
<td>0.900</td>
<td>0.997</td>
</tr>
<tr>
<td>wordF3</td>
<td>1.000</td>
<td>0.274</td>
<td>0.655</td>
<td>0.966</td>
<td>1.000</td>
<td>0.897</td>
<td>0.996</td>
</tr>
<tr>
<td>mosesBLEU</td>
<td>1.000</td>
<td>0.296</td>
<td>0.650</td>
<td>0.974</td>
<td>1.000</td>
<td>0.886</td>
<td>0.992</td>
</tr>
<tr>
<td>mosesPER</td>
<td>1.000</td>
<td>0.307</td>
<td>0.548</td>
<td>0.911</td>
<td>1.000</td>
<td>0.938</td>
<td>0.998</td>
</tr>
</tbody>
</table>
```

... so only English-German tells us something.
**CharacTer** wins in both domains.
## Segment-Level News Task Evaluation

<table>
<thead>
<tr>
<th>Relative Ranking</th>
<th>Direct Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>⊖ Genuine comparisons</td>
<td>⊖ Only 1 candidate shown</td>
</tr>
<tr>
<td>⊖ 5-way comparison hard?</td>
<td>⊖ Principled Pearson</td>
</tr>
<tr>
<td>⊖ Non-standard Kendall’s $\tau$</td>
<td>⊖ Distinct sampling needed</td>
</tr>
<tr>
<td>⊖ Conf. estimation unclear</td>
<td></td>
</tr>
</tbody>
</table>

25 / 32
Segment-Level News Task Results

- DA and RR correlate at .85–.99 (.92 avg across langs).
- Top RR metric always among DA winners.
- RR Winners

<table>
<thead>
<tr>
<th>Metric</th>
<th># Wins</th>
<th>Language Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEER</td>
<td>4</td>
<td>encs, ende, enro, entr</td>
</tr>
<tr>
<td>DPMFcomb</td>
<td>3</td>
<td>csen, fien, ruen</td>
</tr>
<tr>
<td>metrics-f</td>
<td>3</td>
<td>deen, roen, tren</td>
</tr>
<tr>
<td>chrF2</td>
<td>1</td>
<td>enru</td>
</tr>
<tr>
<td>chrF3</td>
<td>1</td>
<td>enfi</td>
</tr>
</tbody>
</table>

- Williams’ test for DA reveals more top-performing metrics:
  - cobalt-f (deen, ruen), MPEDA (enru)
Ex. Russian-English DA Significance
Semantic Golden Truth (HUME)

HUME (Birch et al., 2016) uses two-stage annotation:
1. Semantic annotation (structure) of source.
2. Correctness assessment of corresponding parts of candidate.

- Final sentence-level score aggregated over source components.
HUME in Metrics Task

- A first probe.
- One test set:
  - Medical texts from Cochrane and NHS24
  - Translated by year 1 MT systems of the EU project HimL.
  - Source English annotated once.
  - Targets: Czech, German, Romanian, Polish
  - ~340 sentences

- Used only in segment-level evaluation.
Results of Semantic Evaluation

<table>
<thead>
<tr>
<th>Direction</th>
<th>en-cs</th>
<th>en-de</th>
<th>en-ro</th>
<th>en-pl</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>339</td>
<td>330</td>
<td>349</td>
<td>345</td>
</tr>
<tr>
<td>chrF3</td>
<td>.544</td>
<td>.480</td>
<td>.639</td>
<td>.413</td>
</tr>
<tr>
<td>chrF2</td>
<td>.537</td>
<td>.479</td>
<td>.634</td>
<td>.417</td>
</tr>
<tr>
<td>BEER</td>
<td>.516</td>
<td>.480</td>
<td>.620</td>
<td>.435</td>
</tr>
<tr>
<td>chrF1</td>
<td>.506</td>
<td>.467</td>
<td>.611</td>
<td>.427</td>
</tr>
<tr>
<td>MPEDA</td>
<td>.468</td>
<td>.478</td>
<td>.595</td>
<td>.425</td>
</tr>
<tr>
<td>wordF3</td>
<td>.413</td>
<td>.425</td>
<td>.587</td>
<td>.383</td>
</tr>
<tr>
<td>wordF2</td>
<td>.408</td>
<td>.424</td>
<td>.583</td>
<td>.383</td>
</tr>
<tr>
<td>wordF1</td>
<td>.392</td>
<td>.415</td>
<td>.569</td>
<td>.381</td>
</tr>
<tr>
<td>sentBLEU</td>
<td>.349</td>
<td>.377</td>
<td>.550</td>
<td>.328</td>
</tr>
</tbody>
</table>

- Bold again indicates metrics not significantly outperformed by any other (Williams, 1959).
- **chrF3** and other character-level metrics clearly win.
- **sentBLEU** by far the worst.
Summary

- The **2017 golden truth** will follow the main translation task.
- Whether DA or RR, we will use hybrids for sys-level.
- Domain-specific evaluation of metrics needs **enough systems** to participate (or plan seg-level evaluation).
- Top metrics consider again **character sequences** and are trained.
- Even “semantics” seems well captured by **character-level metrics**.
References


Lili Zhang, Zhou Wang, Wansu Xiang, Liang Wu, Zhiying Chen, Xinyi Tan, Mengshiao Huang,...