Social Media Machine Translation Toolkit (SMMTT)

Wang Ling
Carolin Haas
Chris Dyer
Adam Lopez
Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at [https://github.com/wlin12/SMMTT](https://github.com/wlin12/SMMTT)
  - Microblog Data
  - Tokenizer
  - Normalization
Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at [https://github.com/wlin12/SMMTT](https://github.com/wlin12/SMMTT)
  - Microblog Data
  - Tokenizer
  - Normalization

3 Million Microblog
Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at [https://github.com/wlin12/SMMTT](https://github.com/wlin12/SMMTT)
  - Microblog Data
  - Tokenizer
  - Normalization

![Diagram showing filtering from 3 million microblogs to 10,500 microblogs.](image)
Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at https://github.com/wlin12/SMMTT
  - Microblog Data
  - Tokenizer
  - Normalization

3 Million Microblog

Filtering

10500 Microblog

8000 Train
1250 Dev
1250 Test
Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at [https://github.com/wlin12/SMMTT](https://github.com/wlin12/SMMTT)
  - Microblog Data
  - Tokenizer
  - Normalization
Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at [https://github.com/wlin12/SMMTT](https://github.com/wlin12/SMMTT)
  - Microblog Data
  - Tokenizer
  - Normalization
Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at https://github.com/wlin12/SMMTT
  - Microblog Data
  - Tokenizer
  - Normalization

<table>
<thead>
<tr>
<th>Dataset</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weibo (8K)</td>
<td>14.33</td>
</tr>
</tbody>
</table>
Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at https://github.com/wlin12/SMMTT
  - Microblog Data
  - Tokenizer
  - Normalization

<table>
<thead>
<tr>
<th>Dataset</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weibo (8K)</td>
<td>14.33</td>
</tr>
<tr>
<td>FBIS (300K)</td>
<td>12.84</td>
</tr>
</tbody>
</table>
Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at [https://github.com/wlin12/SMMTT](https://github.com/wlin12/SMMTT)
  - Microblog Data
  - Tokenizer
  - Normalization

<table>
<thead>
<tr>
<th>Dataset</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weibo (8K)</td>
<td>14.33</td>
</tr>
<tr>
<td>FBIS (300K)</td>
<td>12.84</td>
</tr>
<tr>
<td>Weibo+FBIS (308K)</td>
<td>16.28</td>
</tr>
</tbody>
</table>
Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at https://github.com/wlin12/SMMTT
  - Microblog Data
  - Tokenizer
  - Normalization
Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at [https://github.com/wlin12/SMMTT](https://github.com/wlin12/SMMTT)
  - Microblog Data
  - Tokenizer
  - Normalization
Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at [https://github.com/wlin12/SMMTT](https://github.com/wlin12/SMMTT)
  - Microblog Data
  - Tokenizer
  - Normalization

<table>
<thead>
<tr>
<th>Original</th>
<th>Normalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>To Unknown Nick: haha. You pick up the food; I'll cook it <strong>brotha</strong>!</td>
<td>to unknown nick: haha. you pick up the food; <strong>i will</strong> cook it, <strong>brother</strong>!</td>
</tr>
</tbody>
</table>
Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at [https://github.com/wlin12/SMMTT](https://github.com/wlin12/SMMTT)
  - Microblog Data
  - Tokenizer
  - Normalization

<table>
<thead>
<tr>
<th>Original</th>
<th>Normalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>To Unknown Nick: haha. You pick up the food; <em>I'll</em> cook it <em>brotha</em>!</td>
<td>to unknown nick: haha. you pick up the food; <em>i will</em> cook it, <em>brother</em>!</td>
</tr>
<tr>
<td><strong>u guys are awesome</strong> ! cheers !</td>
<td><strong>you are the best</strong> ! cheers !</td>
</tr>
</tbody>
</table>
Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at https://github.com/wlin12/SMMTT
  - Microblog Data
  - Tokenizer
  - Normalization

<table>
<thead>
<tr>
<th>Original</th>
<th>Normalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>To Unknown Nick: haha. You pick up the food; I'll cook it brotha!</td>
<td>to unknown nick: haha. you pick up the food; i will cook it, brother !</td>
</tr>
<tr>
<td>u guys are awesome ! cheers !</td>
<td>you are the best ! cheers !</td>
</tr>
<tr>
<td>To Fuzzy Mannerz, ha! You're cute for sayin that</td>
<td>to vague mannerz, ha! you are cute for saying that</td>
</tr>
</tbody>
</table>
Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at [https://github.com/wlin12/SMMTT](https://github.com/wlin12/SMMTT)
  - Microblog Data
  - Tokenizer
  - Normalization
Thanks guys!!!!!