Meteor Universal: Language Specific Translation Evaluation for Any Target Language

Michael Denkowski  Alon Lavie
{mdenkows,alavie}@cs.cmu.edu

Meteor Universal
Support for any target language using only bitext used to build MT systems:
• Linguistic resources (paraphrases and function words) extracted from bitext.
• Universal parameter set learned by pooling data from all WMT languages.
• Significantly outperforms baseline metrics on unseen languages with no development data.

Meteor Scoring
Align MT hypothesis to reference translation using flexible matching:
Compute similarity score based on precision, recall, and fragmentation (measure of gaps and reordering)
Parameters tuned to maximize agreement with human judgments:
• $cs$: balance between precision and recall
• $\beta, \gamma$: shape and severity of fragmentation penalty
• $d$: balance between content and function words
• $w_{stem}, w_{synonym}, w_{paraphrase}$: weights of approximate matches

Universal Parameter Set
Parameter set learned using all WMT12 data (Callison-Burch et al., 2012):
• 100,000 binary rankings covering 8 language directions.
• Restrict scoring for all languages to exact and paraphrase matching.
Parameters encode human preferences that generalize across languages:
• Prefer recall over precision.
• Prefer word choice over word order.
• Prefer correct translations of content words over function words.
• Prefer exact matches over paraphrase matches, while still giving significant credit to paraphrases.

Linguistic Resources
Paraphrases extracted using phrase pivoting (Bannard and Callison-Burch, 2005) and filtered for precision (Denkowski and Lavie, 2011):

After a sharp drop in the morning ...
Después de la rápida caída de la mañana ...
... Una caída de volumen parecido se registró por última vez ...
... having registered a similarly-ranked fall the last time ...
Learning paraphrase ("drop", "fall") by pivoting through "caída"

Function words learned by relative frequency in monolingual text:
The weight of one of the world’s longest-running conflicts is resting on ...
All of this is designed to reinforce one point: the Gaza withdrawal belongs ...
For the source of the problem is neither the European Central Bank ...
So it is surprising that this choice is not at the center of the political ...
Learning function words "the", "of", "is", "this" by high frequency

Experiments
Compare to (language-specific) Meteor 1.3 on WMT13 data (Macháček and Bojar, 2013):

<table>
<thead>
<tr>
<th>Language</th>
<th>Meteor 1.3</th>
<th>Meteor Universal</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>0.214</td>
<td>0.206</td>
<td>0.124</td>
</tr>
<tr>
<td>Czech</td>
<td>0.092</td>
<td>0.085</td>
<td>0.044</td>
</tr>
<tr>
<td>German</td>
<td>0.163</td>
<td>0.157</td>
<td>0.097</td>
</tr>
<tr>
<td>Spanish</td>
<td>0.106</td>
<td>0.101</td>
<td>0.068</td>
</tr>
<tr>
<td>French</td>
<td>0.150</td>
<td>0.137</td>
<td>0.099</td>
</tr>
</tbody>
</table>

Sentence-level agreement with human judges on WMT13 data

Compare to BLEU on two unseen languages, Russian and Hindi:

<table>
<thead>
<tr>
<th>Language</th>
<th>Meteor Universal</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Russian (WMT13)</td>
<td>0.128</td>
<td>0.068</td>
</tr>
<tr>
<td>Hindi (WMT14)</td>
<td>0.264</td>
<td>0.227</td>
</tr>
</tbody>
</table>

Sentence-level agreement with human judges on data for unseen languages

Software
Meteor 1.5 (including Meteor Universal) is freely available under the GNU Lesser General Public License.
Meteor Universal tutorial:
http://www.cs.cmu.edu/~mdenkows/meteor-universal.html
Meteor download:
https://www.cs.cmu.edu/~alavie/METEOR/