2008 NIST Metrics for Machine Translation (MetricsMATR08) Development Data

…
Authors: Mark Przybocki, Kay Peterson, Sébastien Bronsart
…

Introduction

The 2008 NIST Metrics for Machine Translation (MetricsMATR08) Development Data contains part of the data provided to MetricsMATR08 participants for metric development.

NIST MetricsMATR is a series of research challenge events for machine translation (MT) metrology, promoting the development of innovative, even revolutionary, machine translation (MT) metrics that correlate highly with human assessments of MT quality. Participants submit their metrics to NIST. NIST runs these metrics on held-back test data for which we have various human assessments as measures of MT quality, and calculates correlations between the automatic metric scores and the human assessments. This test data set remains unseen outside of NIST.

The development data in this release contains human reference translations, system translations, and human assessments of adequacy and preference for a subset of the data and systems of the NIST Open MT06 evaluation. The complete MT06 source data and reference translations are available from LDC's catalog as LDC2007E59 (MT06 Evaluation Source Text and Reference Translations).

The subset of MT06 evaluation data that forms this development set consists of 25 documents with a total of 249 segments. The source language is Arabic and the genre is newswire. For each segment, four human reference translations, in English, and system translations from eight different MT06 machine translation systems are included.

The software used for obtaining the human judgments included in this data set is the same as the software used for the NIST Open MT08 human assessments. It includes a description of the adequacy and preference assessment tasks and the instructions given to the judges. All segments assessed were judged by two independent judges. Adequacy judgments were performed for all segments of each document. Preference judgments were performed for the first four segments of each document such that full pair-wise comparisons between all eight MT systems were obtained. All judgments were performed against only one reference translation. The score represents an adjudicated score over the two individual judgments.

The MetricsMATR08 test data remains unseen to allow for repeated use as test data. The MetricsMATR08 development data set described here is reflective of the test data set only to a degree; the evaluation data set contains more varied data - from more genres, more source languages, more systems, and different evaluations - than this development data set. There are also more types of human assessments for the test data.

The official results of MetricsMATR08 on the test data for the metrics submitted to MetricsMATR08 are publicly available. NIST performed the same analyses on the MetricsMATR08 development data after the evaluation. These results are not publicly available, but will likely be available on request in the future by contacting mt_poc@nist.gov.

Sample

The sample below shows reference translations, system translations, and human assessments of adequacy and preference for one document from the development set. Reference and system

translations are in XML format. The README in the release contains a legend for the fields in the human assessment files.

Reference translation:
<refset setid="mt06_arabic_evlset_nist_part_v1" srclang="Arabic" trglang="English">
<DOC docid="AFP_ARB_20060206.0155" genre="newswire" sysid="reference03">
<p>
<seg id="1"> Roed Larsen: The Middle East: " Barrel of Gunpowder with a Burning Wick" </seg>
</p>
<p>
<seg id="2"> Oslo 2/06 (A.F.P) - The former U.N. envoy to the Middle East, Terje Roed Larsen, believes that the situation in this region has never been as dangerous as it is now, comparing the region to "a barrel of gunpowder with a burning wick." </seg>
</p>
<p>
<seg id="3"> In an interview with Norwegian State Broadcasting (NRK), Roed Larsen said, "The region now can be compared, in many aspects, to a barrel of gunpowder with a burning wick". </seg>
</p>
<p>
<seg id="4"> He added, "I believe the situation is more difficult, complex, and dangerous than it has been in many decades." </seg>
</p>
<p>
<seg id="5"> Among the current and latent crises in the Middle East, Roed Larsen mentioned the Arab-Israeli conflict, the Iranian nuclear issue, and the crisis between Lebanon and Syria. </seg>
</p>
<p>
<seg id="6"> He said that "All this brings us back to questions of fundamental values and opinions that put the situation at risk of getting out of control at any moment more than it has been in recent days." </seg>
</p>
<p>
<seg id="7"> Roed Larsen did not explicitly mention the violence that pervaded the demonstrations in the Islamic World in protest against the publication of cartoons about the Prophet Muhammad in the European newspapers. He just stressed the necessity of "dialogue" instead of "throwing stones and setting fires." </seg>
</p>
<p>
<seg id="8"> Demonstrators set the Danish and Norwegian embassies in Damascus on fire Saturday, and a building containing an office of the Danish Embassy in Beirut was also set on fire in response to the publication of the drawings in September in the Danish newspaper Jyllands-Posten, and their subsequent republication in a number of European media, foremost among them the Norwegian magazine Magazinet. </seg>
</p>
<p>
<seg id="9"> Roed Larsen is currently assigned to monitor the implementation of Security Council Resolution 1559 which calls for the disarming of the militias and armed groups in Lebanon, including, according to the international organization, the Shiite Lebanese Hezbollah, and Palestinian groups. </seg>
</p>
</DOC>
...
</refset>

Machine translation:
<tstset setid="mt06_arabic_evlset_nist_part_v1" srclang="Arabic" trglang="English">
<DOC docid="AFP_ARB_20060206.0155" genre="newswire" sysid="system04">
<seg id="1"> Rod Larsen : the Middle East is " gunpowder barrel by a burning wick "</seg>
<seg id="2"> Oslo 6 - 2 (AFP) - Terje Roed Larsen, the former UN envoy to the Middle East, considered that the situation in this region was never as serious as it is today comparing the region to "a gunpowder barrel with a burning wick " .</seg>
<seg id="3"> And Rod Larsen said a speaker of the general radio Al Nrougia " that I see Kaye " that " likening the region from several sides to gunpowder barrel to a burning wick now is possible " .</seg>
<seg id="4"> And he added " I believe that the situation is more difficult, complicated and serious than what it was since several decades " .</seg>
<seg id="5"> And between the existing or latent crises in the Middle East Rod Larsen mentioned the Arab Israeli conflict, the Iran's nuclear file and the crisis between Lebanon and Syria .</seg>
<seg id="6"> And he said that " all this takes us back to the issues of essential values and opinions that make the situation at any moment liable to go out of control more than what it was in the last days " .</seg>

<seg id="7"> And Rod Lars did not mention the frankness of acts of violence that interposed the demonstrations in the Islamic world in protest at the publication of caricatures about the prophet Mohamed in the European newspapers content with emphasizing on a necessity " the dialogue " a substitution " sticking the stones and the ignition of fires " .</seg>
<seg id="8"> And demonstrators have burned the Danish and Norwegian embassies buildings in Damascus Saturday while Ihqnbna includes an office for the Danish embassy in Beirut, in response to the publication of the drawings in September in the Danish newspaper of Jyllands-Posten and the repetition of their publication later in a number of the European media and in their lead was the Norwegian " Maghazint " magazine .</seg>
<seg id="9"> Larsen's arrival is expensive now the surveillance of the application of the International Security Council resolution is 1559 the ones that states the disarmament of militias and armed groups in Lebanon, including with the Lebanese the international organization's look of the Shiite Hezbollah and Palestinian groups .</seg>
</DOC>

...
</tstset>

Adequacy scores:
systemId,documentId,segmentId,judgeId,referenceId,score,scoreQuality,startTime,endTime
system04,AFP_ARB_20060206.0155,1,judgeAdj,reference03,6,yes,2007-12-07 14:37:10,2007-12-07 14:37:52
system04,AFP_ARB_20060206.0155,2,judgeAdj,reference03,7,yes,2007-12-07 14:42:44,2007-12-07 14:42:57
system04,AFP_ARB_20060206.0155,3,judgeAdj,reference03,3,no,2007-12-07 14:45:57,2007-12-07 14:46:38
system04,AFP_ARB_20060206.0155,4,judgeAdj,reference03,6,yes,2007-12-07 14:50:08,2007-12-07 14:51:04
system04,AFP_ARB_20060206.0155,5,judgeAdj,reference03,7,yes,2007-12-07 15:07:43,2007-12-07 15:07:54
system04,AFP_ARB_20060206.0155,6,judgeAdj,reference03,7,yes,2007-12-07 15:13:14,2007-12-07 15:13:24
system04,AFP_ARB_20060206.0155,7,judgeAdj,reference03,5,no,2007-12-07 15:16:00,2007-12-07 15:16:47
system04,AFP_ARB_20060206.0155,8,judgeAdj,reference03,7,no,2007-12-07 15:27:01,2007-12-07 15:28:30
system04,AFP_ARB_20060206.0155,9,judgeAdj,reference03,5,no,2007-12-06 14:42:47,2007-12-06 14:44:01

Preference scores:
system1Id,system2Id,documentId,segmentId,judgeId,referenceId,preferred,startTime,endTime
system04,system03,AFP_ARB_20060206.0155,1,judgeAdj,reference03,1,2008-03-12 13:43:47,2008-03-12 13:43:56
system06,system04,AFP_ARB_20060206.0155,1,judgeAdj,reference03,2,2008-03-12 13:43:57,2008-03-12 13:44:58
system04,system01,AFP_ARB_20060206.0155,1,judgeAdj,reference03,1,2008-03-12 13:44:59,2008-03-12 13:45:02
system07,system04,AFP_ARB_20060206.0155,1,judgeAdj,reference03,1,2008-03-12 13:46:29,2008-03-12 13:47:11
system05,system04,AFP_ARB_20060206.0155,1,judgeAdj,reference03,2,2008-03-12 13:47:25,2008-03-12 13:47:26
system04,system08,AFP_ARB_20060206.0155,1,judgeAdj,reference03,1,2008-03-12 13:48:52,2008-03-12 13:48:54
system02,system04,AFP_ARB_20060206.0155,1,judgeAdj,reference03,2,2008-03-12 13:48:54,2008-03-12 13:48:56
system02,system04,AFP_ARB_20060206.0155,2,judgeAdj,reference03,2,2008-03-12 13:50:28,2008-03-12 13:50:29
system04,system06,AFP_ARB_20060206.0155,2,judgeAdj,reference03,1,2008-03-12 13:50:38,2008-03-12 13:51:18
system04,system07,AFP_ARB_20060206.0155,2,judgeAdj,reference03,1,2008-03-12 13:51:18,2008-03-12 13:52:23
system04,system03,AFP_ARB_20060206.0155,2,judgeAdj,reference03,1,2008-03-12 13:52:34,2008-03-12 13:52:35
system04,system01,AFP_ARB_20060206.0155,2,judgeAdj,reference03,1,2008-03-12 13:52:35,2008-03-12 13:52:36
system04,system05,AFP_ARB_20060206.0155,2,judgeAdj,reference03,1,2008-03-12 13:53:16,2008-03-12 13:53:18
system04,system08,AFP_ARB_20060206.0155,2,judgeAdj,reference03,1,2008-03-12 13:53:24,2008-03-12 13:53:25
system02,system04,AFP_ARB_20060206.0155,3,judgeAdj,reference03,2,2008-03-12 13:55:11,2008-03-12 13:55:29
system04,system01,AFP_ARB_20060206.0155,3,judgeAdj,reference03,1,2008-03-12 13:55:31,2008-03-12 13:55:33
system05,system04,AFP_ARB_20060206.0155,3,judgeAdj,reference03,3,2008-03-12 13:56:19,2008-03-12 13:57:17
system08,system04,AFP_ARB_20060206.0155,3,judgeAdj,reference03,2,2008-03-12 13:57:35,2008-03-12 13:58:04
system06,system04,AFP_ARB_20060206.0155,3,judgeAdj,reference03,3,2008-03-12 13:59:00,2008-03-12 13:59:50
system03,system04,AFP_ARB_20060206.0155,3,judgeAdj,reference03,2,2008-03-12 14:00:18,2008-03-12 14:00:20
system04,system07,AFP_ARB_20060206.0155,3,judgeAdj,reference03,2,2008-03-12 14:00:21,2008-03-12 14:00:23
system05,system04,AFP_ARB_20060206.0155,4,judgeAdj,reference03,3,2008-03-12 14:00:23,2008-03-12 14:00:54
system02,system04,AFP_ARB_20060206.0155,4,judgeAdj,reference03,2,2008-03-12 14:00:58,2008-03-12 14:00:59
system04,system03,AFP_ARB_20060206.0155,4,judgeAdj,reference03,1,2008-03-12 14:01:44,2008-03-12 14:01:47
system01,system04,AFP_ARB_20060206.0155,4,judgeAdj,reference03,2,2008-03-12 14:02:21,2008-03-12 14:02:23
system04,system07,AFP_ARB_20060206.0155,4,judgeAdj,reference03,1,2008-03-12 14:03:48,2008-03-12 14:03:52
system04,system08,AFP_ARB_20060206.0155,4,judgeAdj,reference03,3,2008-03-12 14:06:02,2008-03-12 14:06:19
system04,system06,AFP_ARB_20060206.0155,4,judgeAdj,reference03,2,2008-03-12 14:06:39,2008-03-12 14:06:59

Content Copyright