NIST 2015 Machine Translation Evaluation Official Results

Date of Release: December 23, 2015

Version 1

The 2015 Machine Translation Evaluation (OpenMT15) was the 10th evaluation conducted by NIST as part of an ongoing series of evaluations to support machine translation (MT) research and help advance the state-of-the-art in MT technology. The evaluation was implemented as described in the OpenMT15 evaluation plan.

Disclaimer

These results are not to be construed, or represented as endorsements of any participant's system or commercial product, or as official findings on the part of NIST or the U.S. Government. Note that the results submitted by developers of commercial MT products were generally from research systems, not commercially available products. Since OpenMT was an evaluation of research algorithms, the OpenMT test design required local implementation by each participant. As such, participants were only required to submit their translation system output to NIST for uniform scoring and analysis. The systems themselves were not independently evaluated by NIST.

Certain commercial equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the equipment, instruments, software or materials are necessarily the best available for the purpose.

There is ongoing discussion within the MT research community regarding the most informative metrics for machine translation. The design and implementation of these metrics are themselves very much part of the research. At the present time, there is no single metric that has been deemed to be completely indicative of all aspects of system performance.

The data, protocols, and metrics employed in this evaluation were chosen to support MT research and should not be construed as indicating how well these systems would perform in applications. While changes in the data domain, or changes in the amount of data used to build a system, can greatly influence system performance, changing the task protocols could indicate different performance strengths and weaknesses for these same systems.

Because of the above reasons, this should not be interpreted as a product testing exercise and the results should not be used to make conclusions regarding which commercial products are best for a particular application.

History

- December 23, 2015 v1: Official release

Evaluation Tasks

OpenMT15 task required a system automatically translated data from a source language into a target language. The source languages were Arabic (Egyptian) and Chinese (Mandarin), and the target language was English, resulting in two language pairs:

- Arabic-to-English
- Chinese-to-English

Input Tracks

OpenMT15 had two input tracks:

- audio* - the input to be translated was audio speech consisting of telephone conversations (audio-to-text translation)
- text - the input to be translated was SMS and chat messages as well as human transcripts of telephone conversations (text-to-text translation)

* This track was new to the OpenMT series which previously only had text-to-text translation.

Evaluation Conditions

OpenMT15 had two evaluation conditions:

- constrained - the system used only LDC provided data for training and development
- unconstrained - the system used additional data for training and development

Unlike previous OpenMT's, OpenMT15 did not place a constraint on the training data epoch because the evaluation data (telephone conversations and sms/chat messages) were not publicly available.

Source Data

The table below indicates the size of the source data. In addition to the source, sentence-unit segmentation was also given.

Language Pair	Genre	Number of Files	Number of Source Words
Arabic-to-English	CTS	48	27.5K
	SMS/Chat	198	24.5K
Chinese-to-English	CTS	53	26.3K
	SMS/Chat	162	21.9K

Reference Data

The evaluation data had one gold standard reference instead of four like previous OpenMT's. However, 5K of each language pair and genre combination had HyTER networks.

Performance Metrics

Several automated MT metrics were used to assess the MT output and are listed below.

- BLEU4 - mteval-v13a implementation of
  Papineni, K., Roukos, S., Ward, T. and Zhu, W. J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Compu-tatiol Linguistics, pages 311-318.
- NIST - mteval-v13a implementation of
  Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In Proceedings of the Second International Conference on Human Lan-guage Technology Research, pages 138-145.
- METEOR - meteor-1.3 implementation of
  Banerjee, S. and Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, pages 65-72.
- TERCOM - tercom.7.25 implementation of
  Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In Proceedings of Association for Machine Translation in the Americas, pages 223-231.
- HyTER - hyter-v1.0 implementation of
  Dreyer, M. and Marcu, D. (2012). HyTER: Meaning-Equivalent Semantics for Translation Evaluation. In Proceedings of NAACL 2012 Seventh Workshop on Statistical Machine Translation, pages 162-171.

Data Normalization

The reference and hypothesis were normalized prior to scoring for BLEU-4, NIST, METEOR, and TERCOM metrics. The normalizations performed were:

- remove word-level and phrase-level translation alternatives
- expand unambiguous contractions using only text patterns rather than meaning or context
- remove some punctuations and repeated punctuations
- normalize XML entities
- convert the data to lower case
- tokenize the data

Participants

The table below lists the sites and the tasks in which they participated in OpenMT15.

Team xID	Affiliation	Arabic-to-English				Chinese-to-English			
		Audio		Text		Audio		Text	
		cn	un	cn	un	cn	un	cn	un
bbn	Raytheon BBN Technologies		x	x			x	x	
bit2	Beijing Institute of Technology							x	
camt	Center for Applied Machine Translation #				x				
i2r	Institute for Infocomm Research							x	
ict	Institute of Computing Technology Chinese Academy of Sciences & Dublin City University							x	
jhu	Johns Hopkins University							x	
li2	University of Le Mans & University of Montreal							x	
naver	Naver Labs							x	
ncu	National Central University							x	
nju	Nanjing University							x	
pos	Pohang University of Science & Technology							x	
qcn	Qatar Computing Research Institute, Columbia University, & New York University Abu Dhabi			x					
tubitak	Tubitak & Bilgem	x		x					
ustc	University of Science and Technology of China			x				x	
uva	University of Amsterdam			x	x			x	x
	Beijing Institute of Technology	dropout							
	Carnegie Mellon University, Heidelberg University, & Oxford University								
	Carnegie Mellon University Qatar								
	Saarland University								
	SAS Research and Development Beijing Co., Ltd.								
	SRI International								
	TALP Research Center								

	University of Maryland

Site did not fulfill its obligations to the evaluation.

Evaluation Results

The tables below list the results by language pair, input track, and evaluation condition combinations. If a cell is blank, it means that the team did not participate in that condition.

Constrained Arabic-to-English Primary Systems

Team ID	Audio Input Track				Text Input Track											
	CTS (speech)				CTS (transcripts)				Chat				SMS			
	BLEU4	NIST	METEOR	100-TER	BLEU4	NIST	METEOR	100-TER	BLEU4	NIST	METEOR	100-TER	BLEU4	NIST	METEOR	10 TI
bbn					0.2778	4.9583	0.3075	47.5555	0.3456	3.8002	0.3171	51.4478	0.3251	3.7931	0.3243	52.3
qcn					0.2413	4.7276	0.2877	43.0901	0.3262	3.7056	0.3199	49.0496	0.2998	3.6964	0.3131	48.4
tubitak	0.0810	2.1932	0.1419	17.6352	0.2275	4.5776	0.2793	40.9260	0.3011	3.6286	0.3093	46.5684	0.2767	3.5887	0.3050	46.4
ustc					0.2364	4.5924	0.2934	37.1136	0.2307	3.0094	0.2575	36.2309	0.1879	2.7029	0.2418	33.9
uva					0.2051	4.1912	0.2658	40.8163	0.2207	2.8206	0.2368	38.5017	0.1814	2.7066	0.2323	36.7

Unconstrained Arabic-to-English Primary Systems

Team ID	Audio Input Track				Text Input Track											
	CTS (speech)				CTS (transcripts)				Chat				SMS			
	BLEU4	NIST	METEOR	100-TER	BLEU4	NIST	METEOR	100-TER	BLEU4	NIST	METEOR	100-TER	BLEU4	NIST	METEOR	100 TE
bbn	0.1508	2.7628	0.2073	32.8393												
camt					0.0948	2.9126	0.1868	26.3735	0.1119	2.0817	0.1738	26.1387	0.1046	2.1736	0.1783	27.84
uva					0.2095	4.3362	0.2698	40.7752	0.2205	2.9165	0.2416	37.9295	0.1899	2.8578	0.2384	37.08

Constrained Chinese-to-English Primary Systems

Team ID	Audio Input Track				Text Input Track											
	CTS (speech)				CTS (transcripts)				Chat				SMS			
	BLEU4	NIST	METEOR	100-TER	BLEU4	NIST	METEOR	100-TER	BLEU4	NIST	METEOR	100-TER	BLEU4	NIST	METEOR	100-TER
bbn					0.2012	4.0121	0.2748	40.9489	0.2039	3.4274	0.2615	39.3318	0.2137	2.9059	0.2799	41.427
bit2					0.1180	3.4834	0.2238	30.3630	0.1642	3.2287	0.2447	30.6611	0.1812	2.7919	0.2660	34.256
i2r					0.1789	4.0408	0.2629	36.9574	0.1861	3.3833	0.2536	33.8344	0.1909	2.8498	0.2677	35.749
ict					0.1670	3.9211	0.2588	35.1344	0.1603	3.0244	0.2340	32.4972	0.1567	2.5127	0.2413	33.375
jhu					0.1039	2.7237	0.1988	29.8344	0.0972	1.9619	0.1884	27.2982	0.1144	1.7972	0.2048	29.633
li2					0.1665	3.7731	0.2533	36.6175	0.1556	2.9641	0.2248	34.7634	0.1497	2.3435	0.2248	34.483
naver					0.1359	3.6537	0.2389	30.0262	0.1268	2.8839	0.2180	27.4276	0.1411	2.5127	0.2375	29.328
ncu					0.1504	3.7366	0.2457	34.0931	0.1409	2.8976	0.2253	30.5303	0.1530	2.4600	0.2350	31.647
nju					0.1556	3.6818	0.2437	33.7627	0.1523	3.0689	0.2336	30.5270	0.1718	2.6382	0.2520	33.144
postech					0.0896	2.8109	0.2196	3.5789	0.0886	2.3752	0.1962	6.3979	0.1075	2.0863	0.2128	9.2216
ustc					0.1896	4.1040	0.2667	37.3797	0.1915	3.4610	0.2632	34.3244	0.1960	2.8983	0.2754	36.176
uva					0.1506	3.7835	0.2458	32.6718	0.1462	2.9862	0.2285	31.1098	0.1562	2.5754	0.2458	32.245

Unconstrained Chinese-to-English Primary Systems

Team ID	Audio Input Track				Text Input Track											
	CTS (speech)				CTS (transcripts)				Chat				SMS			
	BLEU4	NIST	METEOR	100-TER	BLEU4	NIST	METEOR	100-TER	BLEU4	NIST	METEOR	100-TER	BLEU4	NIST	METEOR	100 TE

bbn	0.1063	2.1552	0.1787	26.0932												
uva					0.1521	3.7476	0.2483	34.7148	0.1424	2.8659	0.2278	31.0159	0.1582	2.4752	0.2434	33.7

Constrained Arabic-to-English Primary Systems (HyTER Subset)

Team ID	Audio Input Track					Text Input Track										
	CTS (speech)					CTS (transcripts)					Chat					
	BLEU4	NIST	METEOR	100-TER	100-HYTER	BLEU4	NIST	METEOR	100-TER	100-HYTER	BLEU4	NIST	METEOR	100-TER	100-HYTER	BL
bbn						0.2979	4.6988	0.3135	50.5448	53.8170	0.3149	3.7587	0.3136	49.5178	56.3367	0.3
qcn						0.2690	4.5941	0.2967	46.8770	51.8060	0.2790	3.6292	0.3011	46.0113	54.7887	0.3
tubitak	0.0971	1.9492	0.1461	18.6623	25.2320	0.2550	4.4357	0.2881	44.2166	50.6250	0.2769	3.5183	0.2938	43.1534	49.3277	0.2
ustc						0.2750	4.6252	0.3045	41.9232	49.0120	0.2401	3.1926	0.2637	37.6101	38.8437	0.1
uva						0.2235	3.9297	0.2744	43.3393	46.7860	0.2085	2.8322	0.2367	38.0970	41.1243	0.1

Unconstrained Arabic-to-English Primary Systems (HyTER Subset)

Team ID	Audio Input Track					Text Input Track										
	CTS (speech)					CTS (transcripts)					Chat					
	BLEU4	NIST	METEOR	100-TER	100-HYTER	BLEU4	NIST	METEOR	100-TER	100-HYTER	BLEU4	NIST	METEOR	100-TER	100-HYTER	BLE
bbn	0.1550	2.3549	0.2070	33.4717	39.3950											
camt						0.0947	2.4155	0.1777	26.3148	25.8760	0.1033	2.0755	0.1687	25.5194	27.0647	0.10
uva						0.2280	4.0887	0.2777	43.7112	47.0780	0.2025	2.8917	0.2377	36.2620	39.9270	0.19

Constrained Chinese-to-English Primary Systems (HyTER Subset)

Team ID	Audio Input Track					Text Input Track										
	CTS (speech)					CTS (transcripts)					Chat					
	BLEU4	NIST	METEOR	100-TER	100-HYTER	BLEU4	NIST	METEOR	100-TER	100-HYTER	BLEU4	NIST	METEOR	100-TER	100-HYTER	BLEU
bbn						0.2180	3.9297	0.2773	42.0581	42.1333	0.2025	3.3338	0.2599	38.9363	43.1959	0.2170
bit2						0.1369	3.4494	0.2215	32.2561	28.2289	0.1628	3.0619	0.2355	29.2985	40.1471	0.1924
i2r						0.2085	3.9984	0.2659	38.9569	42.2233	0.1945	3.2721	0.2488	33.6918	44.2900	0.1924
ict						0.1878	3.8839	0.2621	35.9354	36.7222	0.1730	2.9979	0.2366	34.4103	42.3588	0.1791
jhu						0.1250	2.7488	0.2022	31.4307	34.9444	0.1085	1.9625	0.1834	27.7614	35.0318	0.1208
li2						0.1807	3.6478	0.2556	37.3479	42.2556	0.1667	2.9320	0.2219	35.1968	45.1776	0.1545
naver						0.1587	3.6027	0.2371	31.7666	35.5056	0.1248	2.6821	0.2066	26.2067	32.0347	0.1491
ncu						0.1766	3.6734	0.2470	36.1349	33.3289	0.1396	2.7129	0.2094	29.4871	32.1059	0.1520
nju						0.1912	3.7935	0.2471	35.6251	38.8844	0.1575	3.0191	0.2300	30.9022	39.9541	0.1684
postech						0.1038	2.7613	0.2184	3.7760	3.5600	0.0838	2.2577	0.1852	6.6562	19.0012	0.1005
ustc						0.2211	4.1317	0.2710	39.5690	43.5567	0.2013	3.3562	0.2561	34.8654	44.8959	0.2023
uva						0.1692	3.7427	0.2501	34.3502	31.1678	0.1459	2.8672	0.2178	30.2538	31.4718	0.1683

Unconstrained Chinese-to-English Primary Systems (HyTER Subset)

Team ID	Audio Input Track					Text Input Track										
	CTS (speech)					CTS (transcripts)					Chat					
	BLEU4	NIST	METEOR	100-TER	100-HYTER	BLEU4	NIST	METEOR	100-TER	100-HYTER	BLEU4	NIST	METEOR	100-TER	100-HYTER	BLE
bbn	0.1176	1.9972	0.1806	26.4351	28.8378											
uva						0.1710	3.6777	0.2483	36.0964	35.0611	0.1445	2.7428	0.2213	30.4765	32.5676	0.16