NIST 2005 Machine Translation Evaluation Official Results

Date of Release: Mon, Aug 1, 2005
Version 3

Introduction

The NIST 2005 Machine Translation Evaluation (MT-05) was part of an ongoing series of evaluations of human language translation technology. NIST conducts these evaluations in order to support machine translation (MT) research and help advance the state-of-the-art in machine translation technology. These evaluations provide an important contribution to the direction of research efforts and the calibration of technical capabilities.

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 MT-05 was an evaluation of research algorithms, the MT-05 test design required local implementation by each participant. As such, participants were only required to submit their translation system output to NIST for scoring. The systems themselves were not evaluated.

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. For that reason, this should not be considered a product testing exercise.

Evaluation Tasks

The MT-05 evaluation consisted of two tasks. Each task required performing translation from a given source language into the target language. The source languages were Arabic and Chinese, and the target language was English.

Evaluation Conditions

MT research and development requires language data resources. System performance is strongly affected by the type and amount of resources used. Therefore, two different resource categories were defined as conditions of evaluation. The categories differed solely by the amount of data that was available for use in system training and development. The evaluation conditions were called "Large Data Track" and "Unlimited Data Track".

1. Large Data Track consisted of the data existing before December 1st, 2004 that was distributed by LDC.
2. Unlimited Data Track consisted of any publicly available data existing before December 1st, 2004.

Evaluation Data

Source Data

There were two source sets, one for each language under test. Each source set contained 100 articles (or documents). These articles were drawn from newswire documents published by the Agence France Presse and the Xinhua News Agency from December 1, 2004 to January 24, 2005. The source documents were encoded in UTF-8 for Arabic and GB for Chinese.

Reference Data

Each source set had four sets of high quality independently generated human translations. Each translation agency was required to have native speaker(s) of the source and target languages, working on their translations.

Performance Measurement

Machine translation quality was measured automatically using an N-gram co-occurrence statistic metric developed by IBM and referred to as BLEU. BLEU measures translation accuracy according to the N-grams or sequence of N-words that it shares with one or more high quality reference translations. Thus, the more co-occurrences the better the score. BLEU is an accuracy metric, ranging from "0" to "1" with "1" being the best possible score. A detailed description of BLEU can be found in the paper Papineni, Roukos, Ward, Zhu (2001). "*Bleu: a Method for Automatic Evaluation of Machine Translation*" (keyword = RC22176).

The main benefit of BLEU is that it is automatically generated when given a system translation and one or more reference translations, allowing quick, inexpensive, and repeatable evaluations that do not require human assessments. It has been found to generally rank systems in the same order as human assessments. BLEU, however, does not have the power to distinguish subtle differences in high quality translations.

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 protocols and metrics employed in this evaluation were created to support MT research and not application effectiveness.

A software utility implementing BLEU is provided on the NIST website as a downloadable tool for anyone who wants to use it to support his own research effort, independent of NIST evaluations.

Evaluation Participants

The table below lists the sites, and the tasks which they participated in, for this year's machine translation evaluation.

NIST ID	Site	Location	Arabic	
			Large	*Unlimited*
ARL	U.S. Army Research Laboratory	USA		x
ATR	Advanced Telecommunications Research Institute International Spoken Language Translation Research Laboratories	Japan		
EDINBURGH	University of Edinburgh	UK	x	
FSC	Fitchburg State College	USA	x	
GOOGLE	Google	USA	x	x
HIT#	Harbin Institute of Technology Machine Intelligence & Translation Laboratory	China		
IBM	IBM	USA	x	
ICT#	Chinese Academy of Sciences Institute of Computing Technology	China		
ISI	University of Southern California Information Sciences Institute	USA	x	
ITCIRST	ITC-IRST	Italy		
JHU-CU	Johns Hopkins University & University of Cambridge	USA, UK	x	
LINEARB	Linear B	UK		
MITRE	MITRE Corporation	USA	x	
NRC	National Research Council of Canada	Canada		

NTT	NTT Communication Science Laboratories	Japan		
RWTH	RWTH Aachen University	Germany		
SAAR	Saarland University	Germany		
SAKHR	Sakhr Software	USA		x
SYSTRAN	SYSTRAN Language Translation Technologies	USA	x	
UMD	University of Maryland	USA	x	

# Sites that did not fulfill their obligation of attending the follow-up workshop.

Evaluation Results

The tables below list the official results of the NIST 2005 Machine Translation Evaluation.

- Tables 1-3 list primary system results for the Arabic-to-English task
- Tables 4-5 list primary system results for the Chinese-to-English task

Notes:

- MITRE submitted two contrastive systems (for the Arabic-to-English task, Large Data track) before submitting their primary system.
- FSC submitted their contrastive system (for Arabic-to-English task, Large Data track) a few minutes before submitting their primary system
- The submission from Linear B involved a human in the translation process. This system is not directly comparable with others because the human has knowledge past the cut-off date for training December 1, 2004.

Release History

- Version 3: Public distribution version of the results
- Version 1: Initial release of results to evaluation participants

Primary Systems - Arabic

This section reviews the results for each site's primary submission for the Arabic-to-English translation task. There is a separate table for each data track (large and unlimited).

Arabic-to-English Task, *Large* Data Track

Table 1	
Site	**BLEU-4 Score**
GOOGLE	0.5131
ISI	0.4657
IBM	0.4646

UMD	0.4497
JHU-CU	0.4348
EDINBURGH	0.3970
SYSTRAN	0.1079
MITRE	0.0772
FSC	0.0037

Arabic-to-English Task, *Unlimited* Data Track

Table 2	
Site	**BLEU-4 Score**
GOOGLE	0.5137
SAKHR	0.3403
ARL	0.2257

Arabic-to-English Task, Data Track Undefined

Table 3	
LINEARB*	0.4300*

* Linear B was not a submission of a fully automatic MT system. Rather, it was a human-aided statistical MT system that used non-Arabic speakers to correct the English fluency by selecting optional English phrases from the system's lattices. Search engines were used to look up the spelling of proper names.

By having humans in the loop and using a search engine at the time of evaluation, the Linear B system had access to data past the training cut-off date (December 1, 2004). Therefore, this system did not belong to either the "Large Data Track" or the "Unlimited Data Track" and is not directly comparable to the other systems in either track.

Primary Systems - Chinese

This section reviews the results for each site's primary submission for the Chinese-to-English translation task. There is a separate table for each data track (large and unlimited).

Chinese-to-English Task, *Large* Data Track

Table 4	
Site	**BLEU-4 Score**
GOOGLE	0.3531

ISI	0.3073
UMD	0.3000
RWTH	0.2931
JHU-CU	0.2827
IBM	0.2571
EDINBURGH	0.2513
ITCIRST	0.2445
NRC	0.2323
NTT	0.2321
ATR	0.1822
SYSTRAN	0.1471
SAAR	0.1310
MITRE	0.0542

Chinese-to-English Task, *Unlimited* Data Track

Table 5	
Site	**BLEU-4 Score**
GOOGLE	0.3516
ICT	0.1293
HIT	0.0797