

Evaluation Plan for GALE Go/No-Go Phase 3 / Phase 3.5 Translation Evaluations

1 INTRODUCTION

The goal of the Global Autonomous Language Exploitation program (GALE) is to develop and apply computer software technologies to absorb, analyze and interpret huge volumes of speech and text in multiple languages. NIST is tasked with evaluating the “translation” aspect of GALE¹.

Specifically, the GALE Translation (MT) evaluation will test machine translation of text and machine translation recorded speech. The test will include language data from both Arabic and Chinese, with system performance tallied separately for each language and separately for text and recorded speech sources.

GALE contractors will be the only participants in this evaluation, and the participants must meet specific Go/No-Go levels of performance.

The third formal GALE Go/No-Go Translation evaluation will be conducted in the summer (Phase 3) and winter (Phase 3.5) of 2008.

This document describes the evaluation protocols for GALE P3/P3.5 Go/No-Go Translation evaluations.

2 EVALUATION TASKS

There are two tasks being evaluated in GALE P3/P3.5 MT evaluations, namely *translation of text* and *translation of recorded speech*. These tasks are being evaluated for two source languages – Arabic and Chinese.

2.1 TRANSLATION OF TEXT

Translation of Text tests a system’s ability to translate foreign text data into understandable and accurate English text. The input for this task will be a variety of mostly unstructured source language documents taken from newswire publications and web texts. Systems must produce English text that completely captures the meaning conveyed by the source data, using easily understandable English.

2.2 TRANSLATION OF RECORDED SPEECH

Translation of Recorded Speech tests a system’s ability to transcribe foreign language audio into understandable and accurate English text. The input for this task will be a variety of audio broadcasts from the news domain and from call-in talk shows. System must produce English text that completely captures the meaning conveyed by the source data, using easily understandable English.

3 EVALUATION DATA

The input data for the GALE Translation evaluation will consist of Arabic and Chinese language data from a variety of audio and text sources.

The evaluation data set provided to test GALE MT, may not be used to update other system components being evaluated separately for the GALE P3/P3.5 Go/No-Go Evaluations.

The GALE P3/P3.5 Translation Evaluation data will come from (1) new data collected between June 1 – 30, 2007 and (2) sequestered data from the GALE P2.5 Translation Evaluation.

All data produced by the LDC that are outside of the GALE P3 Translation evaluation epoch (June 1 – 30, 2007) and do not overlap in content with the sequestered P2.5 data are allowed in training. Note that this means several LDC produced corpora² will be allowed in training for the GALE P3 Translation evaluation and will overlap in time with the GALE P3 Distillation evaluation epoch and the GALE P2/P2.5 evaluation epoch (November 1 – December 22, 2006).

All data produced by the sites that are outside of the GALE P3 Translation evaluation epoch (June 1 – 30, 2007) and outside of the GALE P2.5/P2 Translation evaluation epoch (November 1 – December 22, 2006) are allowed in training.

GALE teams may continue to collect their own data outside of the evaluation epochs (June 1-30, 2007; November 1 – December 22, 2006). Sites must take steps to remove data from this additional collection that cannot be unequivocally identified as having been created outside of the evaluation epochs mentioned above.

GALE teams must share all collected Arabic data by May 15th, 2008.

GALE teams must share all collected Chinese data by TBD.

For GALE P3/P3.5, systems will process more data than the final amount listed in sections 3.1 and 3.2. The additional data will facilitate the selection of the final test set for post-editing.

3.1 TEXT SOURCES

Text data will come from a variety of “Newswire” and “Web Text” sources. The sources will be very similar to last year’s DARPA GALE Machine Translation (MT) evaluation test material. The web text data will be drawn from web logs and discussion forums which will include data that is less well formulated text.

There will be approximately 20,000 words (15,000 new, 5000 sequestered from P2.5) for each of the text sources genres (*newswire* and *web text*). Document length may range from 500-1500 words (as measured by the English reference), but systems will only be evaluated on a pre-defined section of each document³. The pre-defined sections will be identified with the <GALE_P2> tags for the sequestered data and with the <GALE_P3> tags for the new data.

The system under test will not be given direct access to the categorization of each test document, that is, as to whether it is a newswire document or web text data. This must be determined

¹ BAE is tasked with evaluating the “distillation” aspect of GALE.

² Specifically the GALE P2 MT development set (Nov 2006) and the Gigaword v3 corpus (Jan 2005 – Dec 2006)

³ Data selection is described in the “GALE Evaluation Data Selection” document. See: <http://www.nist.gov/tests/gale/2008>

by automatic means. However, metadata will be provided along with the text data that may be used to assist the system in the categorization determination. Metadata information will be included in the test document and identified by SGML attribute tags. All information included inside each source text document is available information to the system under test. In addition, developers may wish to use filenames to assist in determining the categorization of the data. Standard filenames (as are used in the GALE training data) will be used for the GALE P3/P3.5 evaluations.

3.2 AUDIO SOURCES

Audio data will come from a variety of “Broadcast News” and “Broadcast Conversation or Talk Show” sources. The sources will be very similar to last year’s DARPA GALE MT recorded speech evaluation test material. Broadcast conversation sources will focused more on round-table discussions and call-ins that have a conversational style of speech.

There will be approximately 20,000 words (15,000 new, 5000 sequestered from P2.5) for each of the audio sources genres (*broadcast news* and *broadcast conversation*)⁴. Broadcast length may range from 20-60 minutes, but systems will only be evaluated on a pre-defined section of each broadcast, each divided equally between *broadcast news* and *broadcast conversation* sources. For both types of data, the test data will be excerpts of a broadcast.

The system under test will not be given direct access to the categorization of each test broadcast, that is, as to whether it is broadcast news or broadcast conversation data.

Note: Developers may wish to use filenames to assist in determining the categorization of the data. Note that the same file may contain both *broadcast news* and *broadcast conversation* data. Standard filenames (as are used in the GALE training data) will be used for the GALE P3/P3.5 evaluations.

4 DATA FORMATS

The test data formats for the text and audio sources will include segmentation marks, same as in P2.5. However, segmentation will not be provided in Phase 4 and beyond.

4.1 INPUT FORMATS

This section describes the formats of the source data files that will be distributed as evaluation test data for use in the GALE P3/P3.5 Translation evaluations.

4.1.1 Text Input

The source data for text input will be UTF-8 encoded files, with each file corresponding to a single document. The native text inside each document will be segmented.

Each document will contain a series of SGML tags that are used for document identification and document structure. There will be a beginning and ending identifier tags <GALE_P2> (for sequestered data) <GALE_P3> (for new data) that indicate the portion of the document to be processed for evaluation. Systems may use other portions of the document but are only required to output translations for the

identified section, including the beginning and ending identifier tags. Other miscellaneous tags (e.g., <h1>, <p>) might be present in the source file. Only the native source language that is between the identifier tags (i.e., <GALE_P2> or <GALE_P3>) and <seg> tags is to be translated. (*Web text data may have a <QUOTE> tag that contains native source text as an attribute. This data is not to be translated.*)

An example of a source file for text input:

```
<doc id="NYT-doc1">
<body>
<headline>
<seg id="1">ARABIC LANGUAGE TEXT</seg>
</headline>
<text>
<seg id="2">ARABIC LANGUAGE TEXT</seg>
<seg id="3">ARABIC LANGUAGE TEXT</seg>
<GALE_P3 id="S1">
<seg id="4">ARABIC LANGUAGE TEXT</seg>
<seg id="5">ARABIC LANGUAGE TEXT</seg>
</GALE_P3>
<seg id="6">ARABIC LANGUAGE TEXT</seg>
</text>
</body>
</doc>
```

Note that only segments 4 and 5 are to be translated for evaluation.

4.1.2 Audio Input

The source data for audio input will be a separate audio waveform for each broadcast. Each waveform will be distributed in 16-bit PCM format and will include a NIST SPHERE header.

There will be one Un-partitioned Evaluations Map (UEM) file which specifies the time regions within each audio recording to be evaluated.

The UEM file structure is as follows:

```
<F><SP><C><SP><BT><SP><ET><SP><SN>
```

where

<F> indicates the file id, consisting of the path, filename and extension of the waveform to be processed.

<SP> indicates a space (“ ”).

<C> indicates the waveform channel, which, for GALE is always set to 1.

<BT> indicates the beginning time of the segment measured in seconds from the beginning of the file which is time 0.

<ET> indicates the ending time of the segment measured in seconds from the beginning of the file which is time 0.

<SN> indicates the snippet ID.

4.2 OUTPUT FORMATS

This section describes the file formats that the systems evaluated for GALE P3 Translation, must produce.

Each GALE team is to submit system translations for *exactly* one system. *Contrastive* systems will not be evaluated. Translations must be submitted for the complete evaluation test set for each language.

System translations should have proper capitalization and all punctuation should be attached to the text.

⁴ The goal is to create a test set with approximately the same amount of text data as audio data (as measured by the English reference text data). It is estimated that Arabic broadcast news contains about ~5700 words per hour, and Chinese broadcast news contains ~11,000 words per hour (1.5 characters = 1 word). Rates for broadcast conversation data are similar.

4.2.1 MT Output from Text

The system translations of text sources must adhere to the following NIST MT data format.

System translations must begin with the same <doc> tag as is present in the source data. System translations must end with a closing </doc> tag.

System translations are only required to process the native source language that exists between the beginning and ending <GALE_P3> (or <GALE_P2>) tags. These section identifying tags must be included in the system output.

Although the original source text documents will contain segment information, the system translations are not required to output the same segmentation, but they must identify segments by placing segment tags around the translated text. Each segment tag must have an id attribute which sequentially identifies the segments. Each segment tag has a corresponding closing tag. An example of a MT text translation file might be (only what is in red is required):

```
<doc id="NYT-doc1">
<body>
<headline>
<seg id="1">TRANSLATED ENGLISH TEXT</seg>
</headline>
<text>
<seg id="2">TRANSLATED ENGLISH TEXT</seg>
<seg id="3">TRANSLATED ENGLISH TEXT</seg>
<GALE_P3 id="S1">
<seg id="4">TRANSLATED ENGLISH TEXT</seg>
<seg id="5">TRANSLATED ENGLISH TEXT</seg>
</GALE_P3>
<seg id="6">TRANSLATED ENGLISH TEXT</text>
</body>
</doc>
```

4.2.2 MT Output from Audio

The MT system translation output format from audio is exactly the same as for text, but **must** contain the optional segment attributes for time boundaries (start & end):

```
<audiofile fileid="CNN_HEADLINE-file1">
<seg id="1" start="1.25"
end="12.33">TRANSLATED ENGLISH TEXT</seg>
<seg id="2" start="20.95"
end="55.42">TRANSLATED ENGLISH TEXT</seg>
...
</audiofile>
```

5 REFERENCE DATA

System translation output will be evaluated (post-edited) by comparing the system output with a single gold-standard English reference translation. The Linguistic Data Consortium will be responsible for creating the GALE Translation reference data.

In cases where the original source language is ambiguous, the reference data will contain allowable alternatives for words or phrases.

6 DATA PREPROCESSING

As was the case for P2.5, segmentations will be provided for both text and audio sources for the GALE P3/P3.5 evaluation data. Teams are free to use the provided segmentation or to submit results with their own segmentation, in which case NIST will run an automatic alignment procedure with minimal checks for gross errors before post-editing.

6.1 EVALUATION METRIC

Like previous GALE evaluations, GALE P3/P3.5 will use an edit-distance metric to evaluate system translation quality. This will be accomplished by having one or more qualified human editor(s)⁵ make changes to the MT output so that the resulting edited-MT uses understandable English that contains exactly the same information as the reference data in as few edits as they can use. The editors will be given specific guidelines⁶ to follow while performing the edits.

6.2 POST EDITING PROCESS

NIST has developed an editing interface⁷ designed for the post editing task. An editor will view the contents of one complete document⁸ with the focus on a single sentence-like unit. The aligned reference and system translations will be displayed in two separate columns. Alternative words and phrases will be given to the editor in instances when the original source language data was ambiguous or if independent translators did not agree on the exact meaning.

The post editor will modify the segment under focus until they feel that the MT output completely captures the meaning conveyed in the reference data, and nothing more. They are instructed to make modifications using as few edits as possible. Although the editor will be looking at the aligned segments, they will be free to use context before and after the current line of focus. See the post editing guidelines for more details.

Each translated document, by each system, will be post-edited by 2 editors. Both edited documents will be reviewed in a second pass. There will be quality control measures in place to verify that the post editors are performing their job in an acceptable manner.

6.3 THE EDIT DISTANCE METRIC

NIST will compare the resulting edited-MT with the original MT and count the number of edits. Each edit is weighted equally. The number reported will be the ratio of the number of edits to the number of words in the gold standard reference data. In the case of alternative words and phrases, only the first choice listed will be counted as part of the reference.

This score will be automatically calculated using the BBN supplied evaluation script `tercom.0.7.2.jar`⁹. NIST will report the mean HTER scores over the first-pass and second-pass edited data. The final HTER score is found by taking the lowest HTER segment score when comparing the two edited versions.

NIST will distribute to the GALE community, all of the alignments and HTER scores for each document and for each

⁵ For GALE P3 the LDC will be responsible for hiring qualified post editors and for the implementation of the post editing process. LDC will make an effort to retain the same post editors from previous phases, but some new editors may be trained/employed.

⁶ The "Post Editing Guidelines For GALE Machine Translation Evaluation" maybe accessed via the NIST GALE website at: URL <https://www.nist.gov/speech/tests/gale/2007/doc/>

⁷ The JAVA based post editing interface maybe accessed via the NIST GALE website at: <https://www.nist.gov/speech/tests/gale/2008>

⁸ A document could be a complete newswire document, a broadcast news story, a section taken from a user group, or a section from a broadcast talk show.

⁹ The BBN supplied evaluation script is available via the NIST GALE website at: <https://www.nist.gov/speech/tests/gale/2008>

data genre. Scores will be distributed on the document and segment level.

7 SUBMITTING RESULTS TO NIST

FTP is the preferred method for submitting system translation files to NIST.

7.1 MACHINE TRANSLATION OUTPUT

NIST will score one set of machine translations from each participant using the above mentioned post-editing protocols. *Contrastive systems* will not be evaluated.

7.2 PACKAGING SYSTEM TRANSLATIONS

Create a directory that identifies the GALE team:

“Agile”, “Rosetta”, or “Nightingale”

Under your team directory create the following structure:

```
./Arabic/audio
./Arabic/text
./Chinese/audio
./Chinese/text
```

Place the system translations in their proper directory.

System translation files should have the same name as the input file but replace audio file “.sph” and text file “.sgm” extensions with “.system.sgm”

Tar and compress the directory and FTP it to NIST via anonymous ftp to [jaguar.ncsl.nist.gov/gale/incoming](ftp://jaguar.ncsl.nist.gov/gale/incoming)

Notify NIST of your submission by sending an email to gale_poc@nist.gov.

7.3 SYSTEM DESCRIPTIONS

A system description will NOT be required for the GALE evaluation.

8 SCHEDULE

Date	Event
Jun-01-2007	Begin collection of GALE evaluation data
Jun-30-2007	End collection of GALE evaluation data
P3	
Jun-30-2008	GALE P3 Arabic translation evaluation begins
Jul-31-2008	Translations of text and audio due at NIST
Aug-18-2008	Post-editing begins
Sep-29-2008	Post-editing ends
Sep-30-2008	Final scores to DARPA
P3.5	
Nov-17-2008	GALE Arabic and Chinese translation evaluation begins
Dec-08-2008	Translations of text due at NIST
Dec-15-2008	Translations of audio due at NIST

Dec-22-2008	Post-editing begins
Mar -01-2008	Post-editing ends
Mar-04-2008	Final scores to DARPA
TBD	GALE P3 PI meeting