Definition of Patent Machine Translation task at NTCIR-10

NTCIR-10 Patent Translation Task Organizers

June 22, 2012

1. Outline

There are three subtasks for the Patent Machine Translation task (PatentMT): Chinese to English (CE), Japanese to English (JE), and English to Japanese (EJ). Participants choose the subtasks that they would like to participate in. The training data and test data will be provided to participants by the National Institute of Informatics (NII) or The Hong Kong Institute of Education (HKIED). Participants will translate the test data using their machine translation systems and submit the translations to the PatentMT organizers. The PatentMT organizers will evaluate the submissions and return the evaluation results to the participants.

2. Training Data

The training data for JE and EJ subtasks will be distributed by NII. The training data for CE subtasks will be distributed by HKIED. Please read the *readme* files in the distributed data for a detailed description.

For the JE and EJ subtasks, the training data is the same as that for the NTCIR-8 patent translation task and the NTCIR-9 patent machine translation task.

For the CE subtask, the training data is the same as that for the NTCIR-9 patent machine translation task.

3. Test Data

There are several types of test data, and there are four types of evaluations performed for the NTCIR-10 PatentMT:

Intrinsic Evaluation (IE)	The quality of translated sentences will be evaluated using new test sets.
Patent Examination Evaluation (PEE)	New: The usefulness of machine translation for patent examination will be evaluated.
Chronological Evaluation (ChE)	New: A comparison between NTCIR-10 and 9 to measure progress over time, using the NTCIR-9 test sets.
Multilingual Evaluation (ME)	New: A comparison of CE and JE translations using the same English references to see the source language dependency.

- An Intrinsic Evaluation will be conducted for all subtasks.
- A Patent Examination Evaluation will be conducted for the CE and JE subtasks.
- A Chronological Evaluation will be conducted for all subtasks.
- A Multilingual Evaluation will be conducted for the CE and JE subtasks.

NTCIR-9 test data is being used as the test data for the NTCIR-10 Chronological Evaluation. NTCIR-9 test data is also being used as the test data for the NTCIR-10 Multilingual Evaluation for the JE subtask.

The NTCIR-9 test and reference data will be provided when the training and the development data is provided. However, the NTCIR-9 test and reference data must NOT be used for development and training purpose because the NTCIR-9 test data is used for Chronological Evaluation at NTCIR-10. The NTCIR-9 test and reference data can be used for writing your papers.

Please read the *readme* files in the distributed data for more details.

Each test data file is formatted as follows.

TEST-SENTENCE-ID sentence

Examples for Chinese test sentences:

19991015-1 当存在较多的机油时，闪点温度将升高。
20060712-2 本发明总的来说涉及图像传感器，更为具体地说，涉及具有浮动栅极设备的可调彩色图像传感器结构。
20070207-3 图1示出了根据本发明的一个实施例的改进的传感设备。
...

Examples for Japanese test sentences:

19990824-1 一方、可動プレート321 にはケーブル324 が接続されている。
19990730-2 次に、コントロールトランジスタの構造につき説明する。
19990730-3 先ず、選択トランジスタの構造につき説明する。
...

Examples for English test sentences:

20000202-025445-1 A water cooling jacket (not shown) is disposed outside the nipple 208.
20001011-310458-2 Next, operation of the first embodiment will be explained.
20010201-026030-3 The bit b4 of the status register 10 is a writing status bit.
...

The organizers will also provide context data for the intrinsic evaluation test data. The context data is comprised of patent documents that include the test sentences. Participants can use the context data to translate the test sentences. Please read the *readme* files in the distributed data for a detailed description of the context data.

4. XML entity and character normalization

The English sentences from USPTO and the Japanese sentences from JPO include XML entities. At NTCIR-10, the organizers will convert the XML entities into the corresponding original UTF-8 characters in reference sentences of the JE and EJ subtasks before the use.

The organizers will also convert some UTF-8 characters into the corresponding ASCII characters in the reference English sentences for normalization purposes. The organizers will also convert ASCII characters into the corresponding UTF-8 characters in reference Japanese sentences.

The conversion programs will be provided to the participants when the training data is provided. Please read the *readme* files in the distributed data for a detailed description of the programs.

The organizers recommend converting any XML entities in the monolingual data, the training data for JE and EJ, the development data for JE and EJ, and NTCIR-9 test data for JE and EJ into the original characters, then normalizing the characters using the conversion programs before the use.

NTCIR-10 test data for the intrinsic evaluation will be converted before release.

5. Policy for Resource Usage and Declaration

Participants may not use any information from patent data published in 2006-2007 (except for the development data) when they train and develop their MT systems, because the test data for this subtask will be selected from patents filed in this time period. However, participants will be allowed to use other external data for their MT systems, as long as they make a declaration of the resources they used for each submission.

The resource information will be used to classify the submissions into categories for further analysis. (Please see the section "Submission Format" below for how to format the submitted results.) The resource information will be displayed by the RESOURCE_BILINGUAL, RESOURCE_MONOLINGUAL, RESOURCE_EXTERNAL, and CONTEXT tags. If you have any questions regarding the usage of resources, please consult the organizers for further clarification.

NII have released the NTCIR-8 PATMT Japanese-English training data to the public for research use. HKIED released the NTCIR-9 PatentMT Chinese-English training data to the NTCIR-9 CE subtask participants. Participants are allowed to use the data before the 2012.7.1 release date.

6. Translation

Participants will translate the test sentences from the source language to the target languages using their MT systems. Case recovery and de-tokenization must be done on the translation results.

For intrinsic evaluation, each participating group is allowed to submit more than one translation result for a single subtask. At least one result must be produced using only the parallel corpus for training both translation and language models whenever they use

the corpus-based MT method for intrinsic evaluation. For the remaining results, participants are allowed to use additional information. For example, participants may use the monolingual English or Japanese patents from 1993-2005 to train their language models for the Chinese to English, Japanese to English and English to Japanese subtasks.

For other types of evaluation (Patent Examination Evaluation, Chronological Evaluation, and Multilingual Evaluation), each participant is requested to submit only one run for each evaluation type.

7. Submission

For Intrinsic Evaluation, each participant is allowed to submit as many translated results ("runs") as they want. However, the submitted runs should be prioritized by the group. The runs with smaller numbers have higher priority.

For other types of evaluation (Patent Examination Evaluation, Chronological Evaluation, and Multilingual Evaluation), each participant is requested to submit only one run for each evaluation type, given a priority number of 1.

Each run should be saved to a single file. The run files must be encoded in UTF-8. All run files should be combined into a single tar.gz file and submitted to ntc10adm-patentmt AT khn DOT nict DOT go DOT jp. Here, "AT" and "DOT" denote "@" and ".", respectively. Please do not submit plain text files because the character information may potentially be corrupted during the transmission.

8. Submission File Name

Each submission file should have a name conforming with the following format:

X-S-T-N.txt

X : System identifier that is the same as the group ID (e.g., NTC)

T : Subtask of evaluation:

- ze: CE subtask
- je: JE subtask
- ej: EJ subtask

T : Type of evaluation:

- int: intrinsic evaluation (for CE, JE, and EJ)
- exa: patent examination evaluation (for CE and JE)

- chr: chronological evaluation (for CE, JE, and EJ)
- mul: multilingual evaluation (for CE)
- chrmul: chronological and multilingual evaluation (for JE)

N : Priority of run (1, 2, 3, ...) for each evaluation type

For example, if the group “NTC” submits two files for the CE intrinsic evaluation, one file for the CE patent examination evaluation, one file for the CE chronological evaluation, one file for the multilingual evaluation, the names of the run files should be “NTC-ze-int-1.txt”, “NTC-ze-int-2.txt”, “NTC-ze-exa-1.txt”, “NTC-ze-chr-1.txt”, and “NTC-ze-mul-1.txt”.

The type name for JE is combined for chronological evaluation and multilingual evaluation because the test data for the chronological evaluation and multilingual evaluation are the same.

9. Submission Format

In this section, we describe the file format for submissions of translated results by MT systems. The translation submission files are organized with the following tags.

<TS-TEST-DATA> Each file has a single “TS-TEST-DATA” tag with an ID. The ID is the filename of this submitted file without its suffix.

<TASK> The type of evaluation. The type is “Intrinsic”, “Chronological”, “Multilingual”, “Chronological/Multilingual” (for JE), or “Patent Examination”.

<DIRECTION> The direction of translation. “CE” if Chinese into English, “JE” if Japanese into English, or “EJ” if English into Japanese.

<SYSTEM-ID> System identifier that is the same as the group ID.

<PRIORITY> Priority of the run.

<TYPE> Rough type of the system: “SMT”= statistical MT, “EBMT” = example based MT, “RBMT” = rule-based MT or “HYBRID” = hybrid MT system.

<RESOURCE_BILINGUAL>: “YES”, if you used the bilingual training data provided by the organizers; “NO”, if you did not use the bilingual training data provided by the organizers.

<RESOURCE_MONOLINGUAL>: “YES”, if you used the monolingual training data provided by the organizers; “NO”, if you did not use monolingual training data provided by the organizers.

<RESOURCE_EXTERNAL>: “YES”, if you used external knowledge other than data provided by the organizers or the system uses rule-based system, otherwise “NO”.
<CONTEXT> “YES” if the system uses context information, otherwise “NO”
<OFFLINE-TIME> Approximate time for offline training (e.g., 3 hours, 2 weeks, and N/A)
<ONLINE-TIME> Approximate time for online translation of the test sentences of the task (e.g., 10 minutes)
<MACHINE-SPEC> A brief description of computers used for the task (e.g., specification of a server and the number of nodes in a PC cluster)
<SYSTEM-DESCRIPTION> A brief description of the features of the system. If you submit multiple runs in the same subtask for intrinsic evaluation, please include the differences between runs.
<RESULTS> Translated sentences should be put into this field with a sentence in each line. Each sentence should be preceded by the TEST-SNTENCE-ID assigned to the original corresponding test sentence. The TEST-SNTENCE-ID and the translated sentence should be separated by white spaces.

Submission files should be encoded in UTF-8 format. (Although the NTCIR-9 test file in Japanese are in EUC. NTCIR-10 test file will be in UTF-8.)

Example files for the evaluations are as follows.

An example of run file for the intrinsic evaluation

```
<TS-TEST-DATA ID="NTC-ze-int-1">
<TASK>Intrinsic</TASK>
<DIRECTION>JE</DIRECTION>
<SYSTEM-ID>NTC</SYSTEM-ID>
<PRIORITY>1</PRIORITY>
<TYPE>SMT</TYPE>
<RESOURCE_BILINGUAL>YES</RESOURCE_BILINGUAL>
<RESOURCE_MONOLINGUAL>NO</RESOURCE_MONOLINGUAL>
<RESOURCE_EXTERNAL>NO</RESOURCE_EXTERNAL>
<CONTEXT>YES</CONTEXT>
<OFFLINE-TIME>3 weeks</OFFLINE-TIME>
<ONLINE-TIME>2 days</ONLINE-TIME>
<MACHINE-SPEC>Xeon 3GHz dual CPU, 4GB memory</MACHINE-SPEC>
<SYSTEM-DESCRIPTION>
Modified version of the Moses phrase-based MT system.
The language model takes advantage of context information using the trigger model.
The other runs use the different language models.
</SYSTEM-DESCRIPTION>
<RESULTS>
19990824-236884-1 On the other hand, a cable 324 ...
19990730-217401-2 Next, structure of the control ...
19990730-217401-3 First, structure of the selection ...
</RESULTS>
</TS-TEST-DATA>
```