Definition of Patent Machine Translation task at NTCIR-9

NTCIR-9 Patent Translation Task Organizers
April 21, 2011

1. Outline

There are three subtasks for the Patent Machine Translation task (PatentMT): Chinese to English (C-E), Japanese to English (J-E), and English to Japanese (E-J). Participants choose the subtasks that they would like to participate in. The training data and test data will be provided to participants by NII. Participants translate the test data using their machine translation systems and submit the translations to the PatentMT organizers. The PatentMT organizers will evaluate the submitted translations and return the evaluation results to the participants.

2. Training Data

The training data will be distributed by NII. Please read the *readme* files in the distributed data for the detailed description.

For the J-E and E-J subtasks, the training data is the same as that for the NTCIR-8 patent translation task.

For the C-E subtask, the training data will be newly built by Hong Kong Institute of Education and ChiLin Star Corp, and distributed by NII.

3. Test Data

The organizers will provide three sets of test sentences. One set (file) contains Chinese sentences, one set (file) contains Japanese sentences and the other set (file) contains English sentences. The filenames of these three files are:

- ntc9-patentmt-fmlrun-z.txt
- ntc9-patentmt-fmlrun-j.txt
- ntc9-patentmt-fmlrun-e.txt

The format of these three files is as follows.

TEST-SNT-ID sentence

Examples for Chinese test sentences are:

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　　1示出了根据本发明的一个实施例的改进的　感设备。
...

Examples for Japanese test sentences are:

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 are:

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 test data. The context data is 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. Policy for Resource Usage and Declaration

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

NTCIR-8 PATMT training data will be released to the public for research use by NII before 2011/01/05. Participants are allowed to use the data before the 2011/01/05 date that NTCIR-9 releases the training data.

The resource information will be used to classify the submissions into categories for further analysis. To be more precise, the categories are defined by the following three questions:

(1) did you use the provided data for the submission? (yes or no)

(2) did you use external knowledge other than the provided data for the submission? (yes or no)

(3) did you use context information around a target sentence for the submission? (yes or no).

Please see the section “Submission Format” below to know how to format the submitted results. If you have any question regarding the usage of resources, please consult the organizers for further clarification.

5. Translation

Participants 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.

Although 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 files in

ntcir9-patentmt-train-zh-en.tgz (for C-E subtask)
ntcir8_patmt_ntc8-patmt-train.tgz (for J-E and E-J subtasks)

for training both translation and language models whenever they use the corpus-based MT method. 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 Chinese to English, Japanese to English and English to Japanese subtasks.

The translated sentences will be evaluated through human evaluation and automatic evaluation. The primary evaluation is human evaluation.

6. Submission

Each participant is allowed to submit as many translated results (”runs”) as they want. However, the submitted runs should be prioritized by the group, because the runs with higher priority would be used for manual evaluations and for analysis purposes, though all submitted runs will be evaluated by automatic evaluation measure (BLEU and/or NIST). Priority no. should be assigned through all the submissions of each participant, and the runs with smaller numbers have higher priority. We would organize the manual evaluation for as many runs as our budget allows.

Each run should be saved in a single file. The run files must be encoded in UTF-8. Otherwise, the organizers will automatically convert them into UTF-8 for evaluation. All run files should be combined into a single tar.gz file and should be submitted to ntc9adm-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.

7. Submission File Name

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

X-T-N.txt

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

T : Type of evaluation:

・ ze: C-E intrinsic evaluation

・ je: J-E intrinsic evaluation

・ ej: E-J intrinsic evaluation

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

For example, if the group “NTC” submits two files for the J-E intrinsic evaluation, one file for the E-J evaluation, the names of the run files should be “NTC-je-1.txt”, “NTC-je-2.txt”, “NTC-ej-1.txt”.

8. Submission Format

In this section, we describe the file format for submissions of translated results by MT systems. The 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”
<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> “YES” If you used training data provided by the organizers. “NO” if you did not use training data provided by the organizers.
<EXTERNAL> “YES” if you use external knowledge other than data provided by the organizers or the system type is rule-based or hybrid, 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 features of the system. If you submit multiple runs in the same direction of translation, please include the difference between runs.
<RESULTS> Translated sentences should be put into this field with a sentence in each line. Each sentence should be preceded by TEST-SNT-ID assigned to the original corresponding test sentence. TEST-SNT-ID and the translated sentence are separated by white spaces.

Submission files should be encoded in UTF-8 format. (Although the input file in Japanese are in EUC.)

Example files for the evaluations are as follows.

An example of run file for the intrinsic evaluation

```
<TS-TEST-DATA ID="NTC-je1">
<TASK>Intrinsic</TASK>
<DIRECTION>JE</DIRECTION>
<SYSTEM-ID>NTC</SYSTEM-ID>
<PRIORITY>1</PRIORITY>
<TYPE>SMT</TYPE>
<RESOURCE>YES</RESOURCE>
<EXTERNAL>NO</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.
</SYSTME-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>
```