This paper describes the NiuTrans system submitted to the NTCIR-9 Patent Machine Translation task by the Natural Language Processing Lab at Northeastern University. Our submissions were generated using the phrase-based translation system implemented under the NiuTrans project. To fit the patent translation task, our system is improved in several ways.

**Reordering:** Unlike traditional approaches, we did not resort to a single reordering model, but instead used a hybrid approach that makes use of multiple reordering models.

**Large-scale n-gram LM:** We developed a simple and fast language model for n-gram scoring on very large patent data, and trained a 5-gram language model using all English data (57 GB raw text) provided within the task.

**SMT and EBMT:** We enhanced our SMT system using a simple EBMT system.

In addition to the phrase structure design, we also prune the model using both vocabulary filtering and r-gram filtering.

Chinese and Japanese sentences were segmented using the NEUNLPLab Chinese segmentation system and the MeCab system, respectively.

For the Chinese-English MT track, all number/date/time entities were generalized to be unique symbols. These entities were then translated using an additional rule-based translation engine when we decoded test sentences.

All sentence pairs with unreasonable target-length/source-length ratios (＞0.2 or ＜5.0) were filtered out to weaken the influence of noisy data.

To recover the case information, we used the rescaper in Moses SMT toolkit which is based on heuristic rules and HMM models.

NiuTrans: An Open-Source Statistical Machine Translation System

http://www.nlplab.com/NiuPlan/NiuTrans.html

**Features**

- Written in C++. So it is fast.
- Multi-thread supported
- Easy-to-use APIs for feature engineering
- Competitive performance for Chinese-foreign translation tasks
- A compact but efficient n-gram language model is embedded. It does not need external support from other softwares (such as SRILM).

**NiuTrans.Phrase (for NTCIR-9 PatentMT)**

- Based on Bracketing Transduction Grammar
- Two reordering models: ME and MSD
- CKY-style decoder with cube pruning and beam pruning.

**A Hybrid Reordering Model**

We developed a very simple approach to combine multiple reordering approaches modeled in different views: all reordering models (features) were jointly used during decoding.

**Large-scale n-gram LM**

Our LM builder is basically a "sorted" Trie structure.

- id
- prob
- Backoff
- weight
- Next link

In addition to the data structure design, we also prune the model using both vocabulary filtering and n-gram filtering.

**Combination of SMT and EBMT**

- In addition to the NiuTrans SMT system, we developed a simple EBMT system. Given a test sentence, it first scans the training corpus and finds the most "similar" samples using the Longest Common Subsequences (LCS) algorithm.
- Then it generates the translation output by only deleting unexpected target words.
- We used the "one-beat-all" strategy for final translation selection: if the EBMT output is trusted enough, we selected its result as the final output; otherwise, we chose the SMT output.

**Results**

- Using additional out-domain data:

<table>
<thead>
<tr>
<th>Entries</th>
<th>Chinese-English</th>
<th>Japanese-English</th>
<th>ChinaMet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Trans Phrases &amp; SMT models</td>
<td>26.04</td>
<td>31.95</td>
<td>26.04</td>
</tr>
<tr>
<td>EBMT</td>
<td>26.25</td>
<td>32.03</td>
<td>26.25</td>
</tr>
<tr>
<td>Total</td>
<td>26.25</td>
<td>32.03</td>
<td>26.25</td>
</tr>
</tbody>
</table>