The NiuTrans Machine Translation System for NTCIR-9 PatentMT

Tong Xiao, Qiang Li, Qi Lu, Hao Zhang, Haibo Ding, Shujie Yao
Xiaoming Xu, Xiaoxu Fei, Jingbo Zhu, Feiliang Ren, Huizhen Wang
Natural Language Processing Lab (www.nlplab.com), Northeastern University

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.

NiuTrans: An Open-Source Statistical Machine Translation System

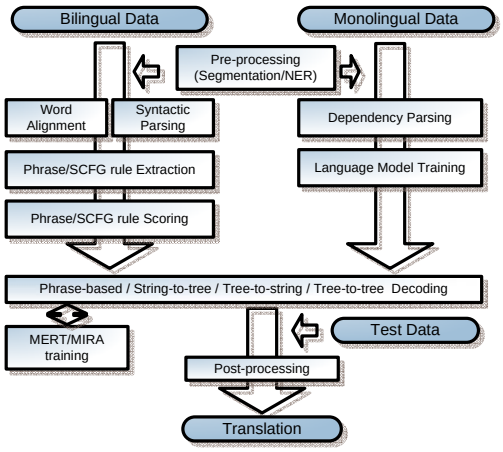
http://www.nlplab.com/NiuPlan/NiuTrans.html

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)
- Supports multiple SMT models a) Phrase-based model b) Hierarchical phrase-based model (coming soon) c) Syntax-based model (string-to-tree/tree-to-string/tree-to-tree) (coming soon)

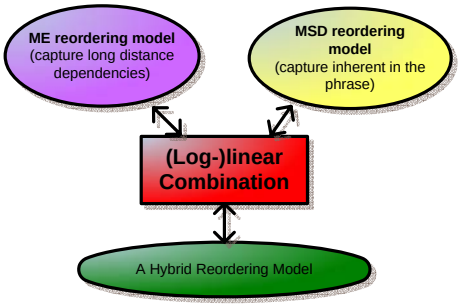
NiuTrans.Phrase (for NTCIR-9 PatentMT)

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



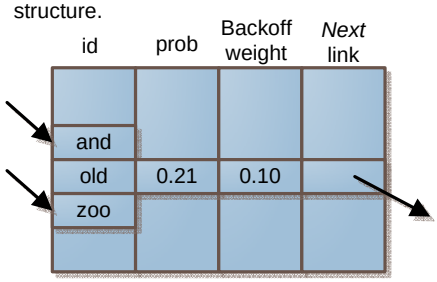
A Hybrid Reordering Model

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.



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

- Chinese and Japanese sentences were segmented using the NEUNLPLab Chinese segmentation system and the MeCab system, respectively.
- For 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.
- Bi-directional word alignments were performed on the bilingual sentences with GIZA++ & "grow-diag-final-both".
- To recover the case information, we used the recaser in Moses SMT toolkit which is based on heuristic rules and HMM models.

Main result

Table 1. Datasets used

Entry		Chinese-English C/E	Japanese-English J/E	Monolingual (English)
TRAINING	SENTENCES	1.0M	3.2M	282M
	WORDS	38M/43M	116M/110M	10882M
	VOCABULARY	300K/278K	184K/195K	1M
	ALIGNMENTS	36M	58M	N/A
DEVELOPMENT	SENTENCES	1500	2000	N/A
	WORDS	55K/60K	75K/70K	N/A
TEST	SENTENCES	2000	2000	N/A
	WORDS	55K/51K	74K/63K	N/A

Table 2. Results on NTCIR-9 PatentMT Evaluation Data

Entry	Chinese-English			Japanese-English		
	adequacy	accept	BLEU4	adequacy	accept	BLEU4
NiuTrans.Phrase	**3.51**	**0.543**	0.3229	2.37	0.416	0.2440
NiuTrans.Phrase + EBMT	N/A	N/A	**0.3273**	N/A	N/A	0.2488
Baseline 1 – Moses' hiero	3.29	0.476	0.3072	2.61	0.474	**0.2895**
Baseline 2 – Moses' phrasal	2.89	N/A	0.2932	2.42	0.447	0.2861
Baseline 3 – A rule-based system	2.27	N/A	0.1075	**3.53**	**0.674**	0.1885
Baseline 4 – Google's online translation	2.96	0.42	0.2569	2.27	0.417	0.1873

Using additional out-domain data

Table 3. Additional Open-domain Datasets

Entry	NIST news C/E	CWMT news C/E	Multi-domain dictionary
SENTENCES/ENTRIES	2.0M	3.1M	2.0M
WORDS	49M/55M	60M/65M	N/A
VOCABULARY	209K/135K	393K/374K	N/A
ALIGNMENTS	46M	55M	N/A

Table 4. Results of Using Additional Training Data

Entry	BLEU4 Dev	BLEU4 Test
Baseline (NTCIR-9 CE PatentMT)	**0.3311**	**0.3217**
+ NIST CE news	0.3257	0.3171
+ CWMT CE news	0.3279	0.3148
+ multi-domain bi-dict	0.3282	0.3172
+ all	0.3270	0.3165