The *NiuTrans* 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 and Huizhen Wang

Natural Language Processing Lab
Northeastern University
http://www.nlplab.com

Outline

- Our NiuTrans System
- Improvements for Patent MT
- Evaluation Results
- Future Work

Our Submissions

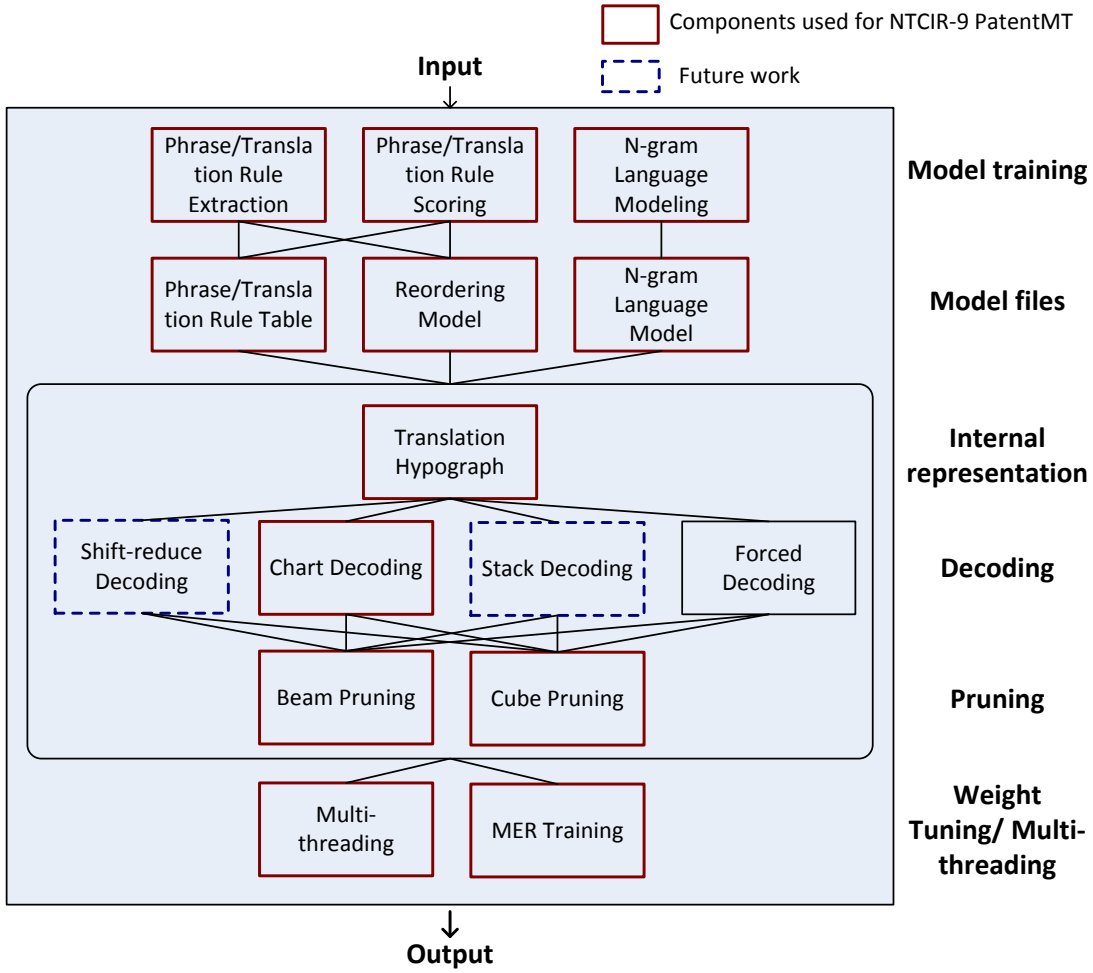
- Patent MT Tasks
  - Chinese-English (2 submissions)
  - Japanese-English (2 submissions)
  - *English-Japanese* (*no submission*)

NiuTrans

- NiuTrans is an open-source Statistical Machine Translation (SMT) system which is developed by our group - NLP Lab at Northeastern University.
  - Publicly released on this July, and so far shared by more than 200+ research groups all over the world.
- Features
  - Written in C++. It is fast, easy to install and use.
  - Multi-thread supported
  - Competitive performance for Chinese-Foreign translation
  - A compact but efficient n-gram language model is embedded.
  - Support multiple SMT models
    - Phrase-based model
    - Hierarchical phrase-based model
    - Syntax-based model (string-to-tree/tree-to-string/tree-to-tree) (coming soon)
- Available at http://www.**nlplab**.com

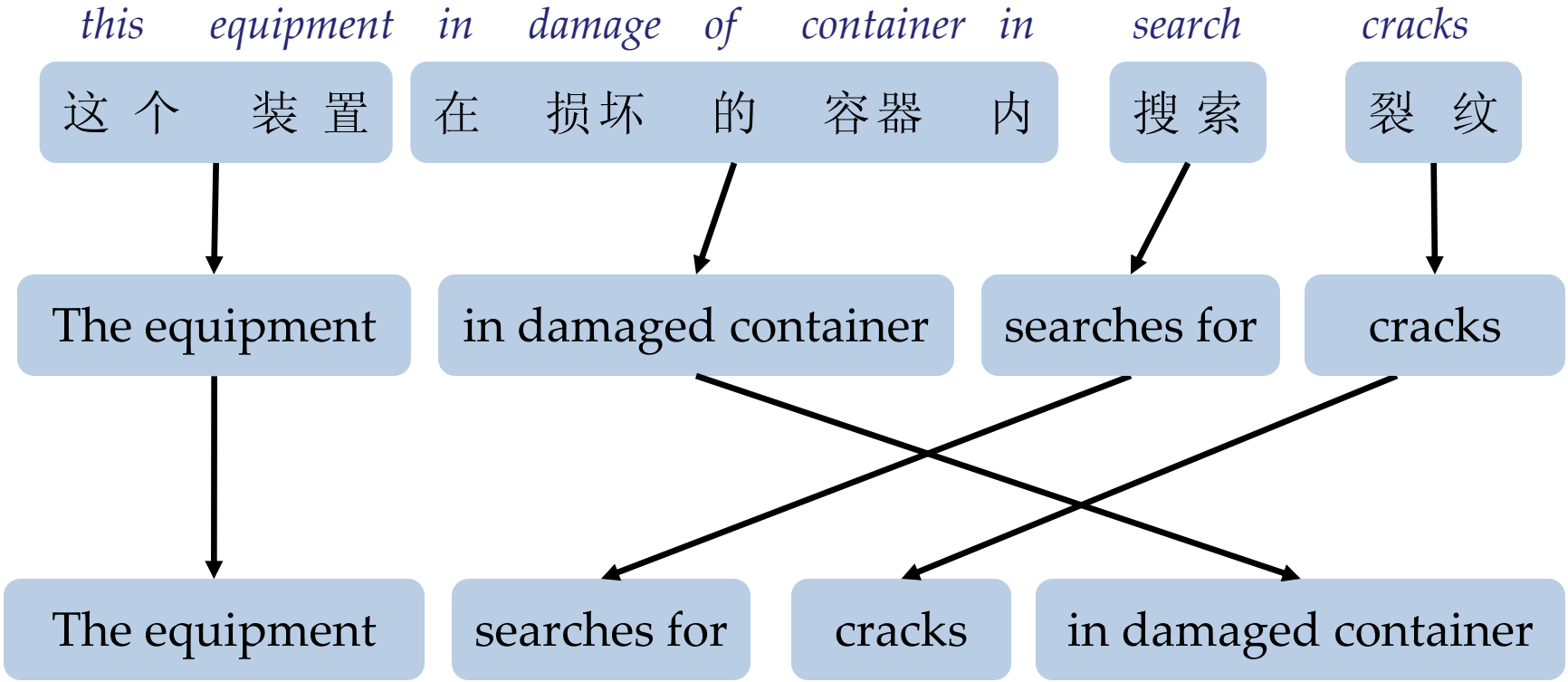
NiuTrans for NTCIR-9 PatentMT

- All of the SMT models are implemented in the same framework
  - We chose the phrase-based engine for NTCIR-9
- Note that we did not use the syntax-based engine in this task because current parsing accuracy is far from satisfactory on non-news domain technical documents.



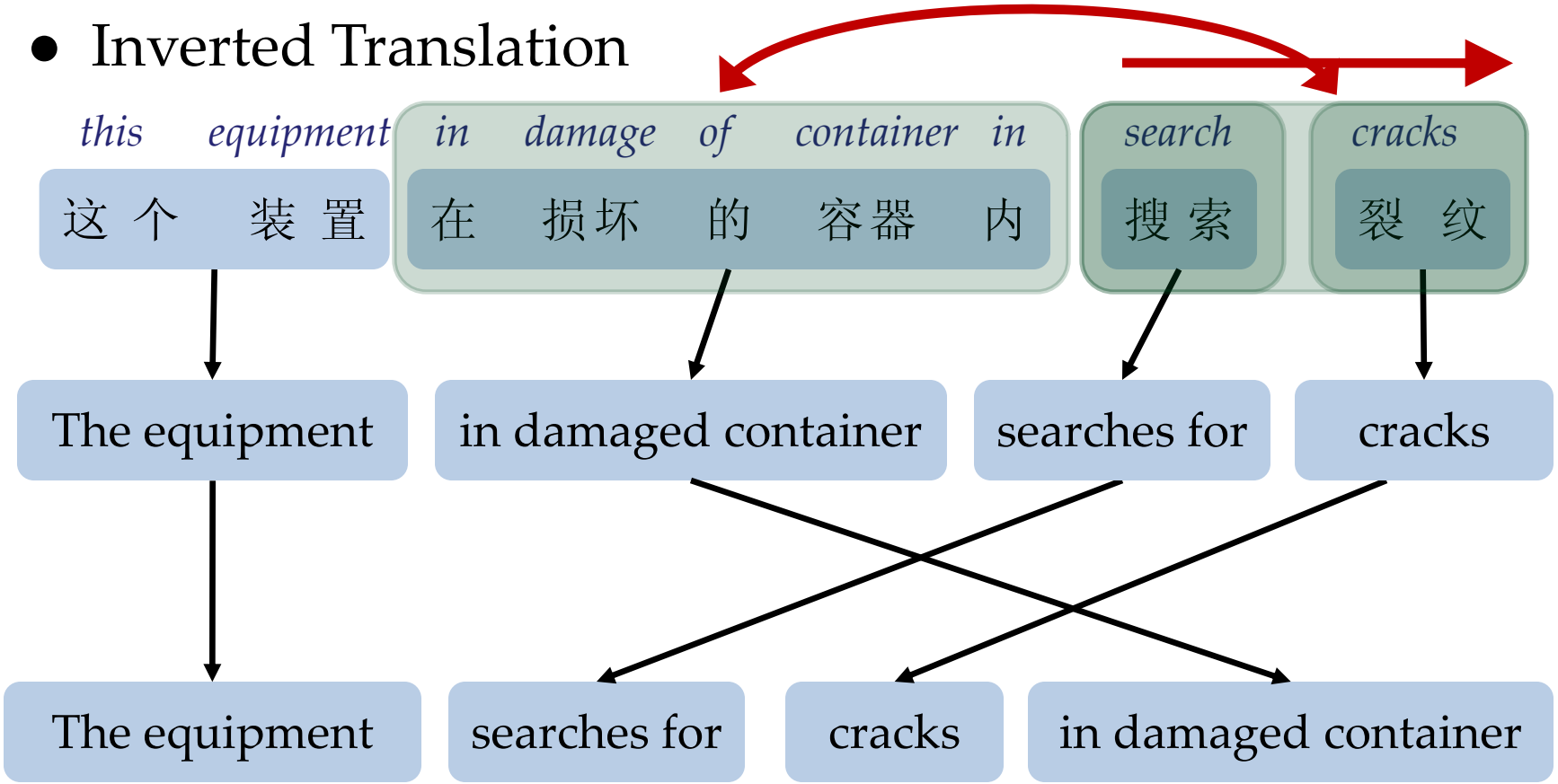
NiuTrans.Phrase

- NiuTrans.Phrase system follows the general framework of Inversion Transduction Grammar (ITG)

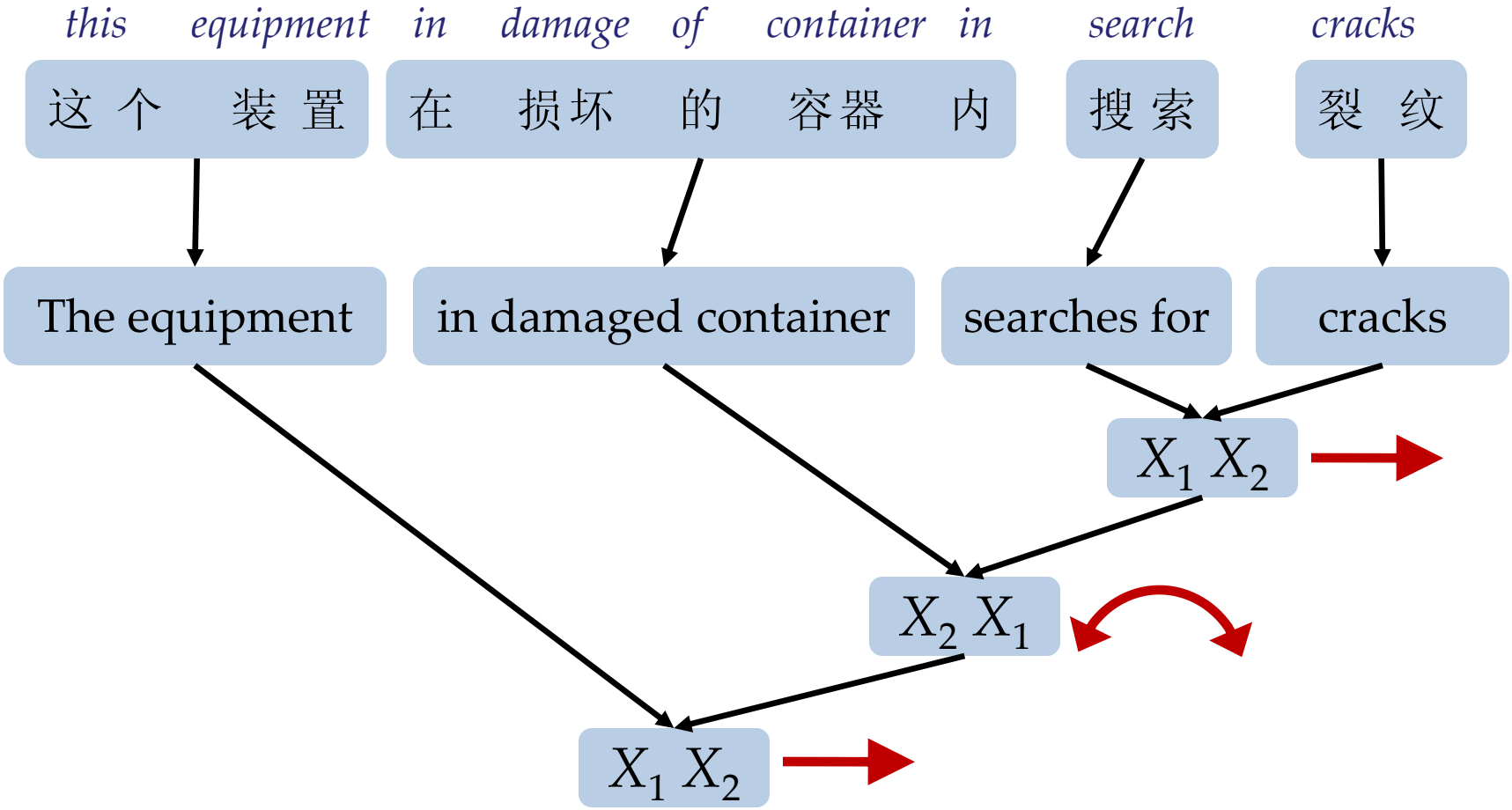


NiuTrans.Phrase

- Monotone Translation
- Inverted Translation



NiuTrans.Phrase



Improvements for Patent Translation

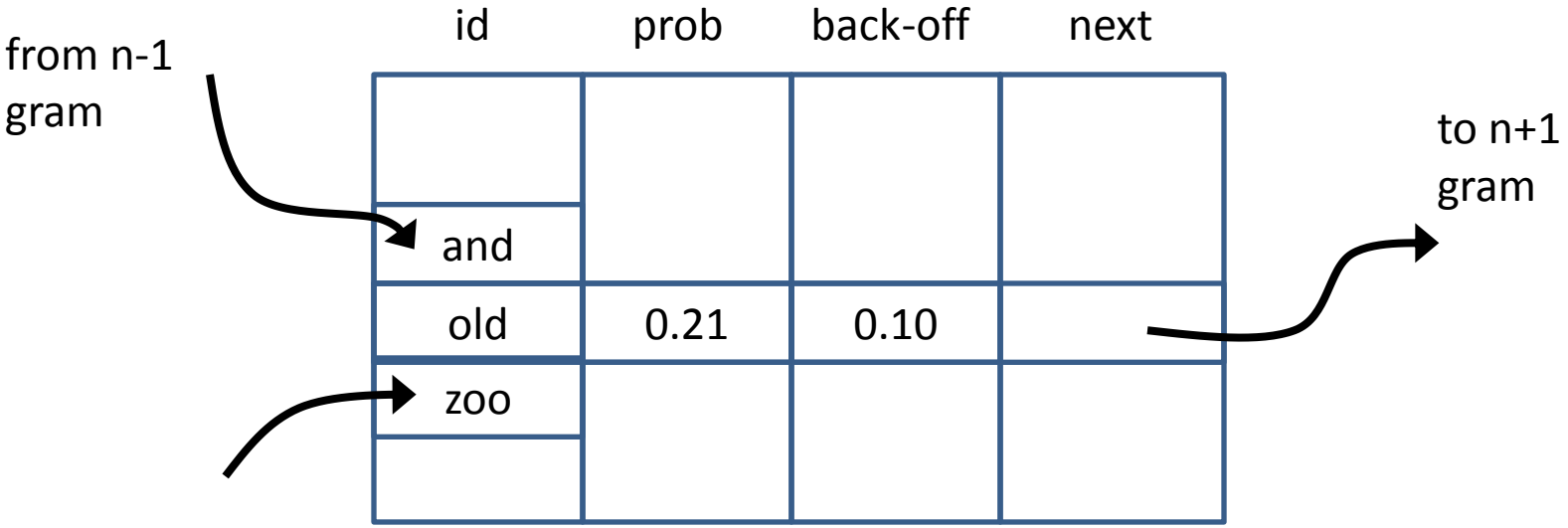
- Hybrid Reordering Model for Phrase-based SMT
- Large-scale N-gram Language Modeling
- Combining SMT and EBMT

A hybrid reordering model

- Many reordering models are available
  - competitive translation quality
  - different strengths and weaknesses
    - ME ordering: characterizes the movement of hierarchical structures by phrase boundary features
    - MSD ordering: powerful in local reordering that is inherent in the phrase translations
- It is natural to explore approaches that use or combine multiple reordering models
  - Our Solution: jointly use them during decoding

Large-Scale N-gram Language Modeling

- Our LM builder is basically a "sorted" Trie structure (Pauls and Klein, 2011)



- Pruning
  - Vocabulary filtering
  - N-gram filtering
- 57GB raw text → 6.1GB (5-gram) LM file (binary format)

Combining SMT and EBMT

- Combination is a desirable way to achieve higher translation accuracy than any individual approach does.
  - SMT: NiuTrans.Phrsae
  - EBMT: a naïve word-based EBMT system
    - Longest Common Subsequences
    - Delete unexpected target words
  - Select EBMT output only when very similar sentences are found
- Problem
  - Noisy data
  - Needs a better combination strategy

Features

- These improvements result in 17 features for our submitted (SMT) system

	Feature	Description	Weight (*ch-en*)	Weight (*jp-en*)
1	$\Pr(t \mid s)$	Phrase trans-probability	0.089	0.107
2	$\Pr_{lex}(t \mid s)$	Lexical weight	0.043	0.034
3	$\Pr(s \mid t)$	Inverted $\Pr(t \mid s)$	0.017	0.050
4	$\Pr_{lex}(s \mid t)$	Inverted $\Pr_{lex}(t \mid s)$	0.033	0.039
5	$\Pr_{\mathrm{LM5}}(t)$	5-gram language model	0.157	0.063
6	Length(t)	# of target words	0.095	0.154
7	Count(Phr)	# of phrases	0.111	0.104
8	WD	# of word deletions	-0.006	-0.018
9	Bi-Lex	# of bi-lex links	0.082	0.051
10	Count(low-freq)	# of low-frequency rules	-0.040	-0.031
11	$f_{BTG\text{-}ME}$	ME-based reordering feature	0.193	0.201
12	$f_{M\text{-}previous}$	M orientation (previous)	0.037	0.024
13	$f_{S\text{-}previous}$	S orientation (previous)	0.017	0.014
14	$f_{D\text{-}previous}$	D orientation (previous)	0.018	0.030
15	$f_{M\text{-}following}$	M orientation (following)	0.017	0.031
16	$f_{S\text{-}following}$	S orientation (following)	0.036	0.011
17	$f_{D\text{-}following}$	D orientation (following)	0.002	0.028

Evaluation (formal run)

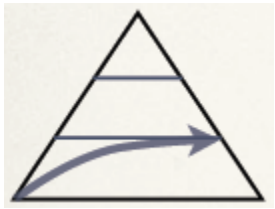
- The recourses we used were constrained to those provided for NTCIR-9 PatentMT
  - Chinese-English: 38M/43M words
  - Japanese-English: 116M/110M words
  - English (USPTO): 10,882M words
- Results

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

- NiuTrans.Phrase outperforms all five baselines on CE MT
- EBMT is useful in enhancing SMT output
- Maybe there is something wrong with the use of the open-source Japanese segmentation tool for our task.

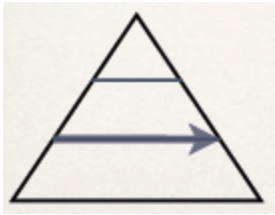
Future work:
NiuTrans will support more features

string-to-string



string-to-tree

tree-to-string



tree-to-tree

Current Version (v0.3.0)
- (Hierarchical) Phrase-based models

Standard Version (v1.0.0)
- Various syntax-based models
- Tree-parsing and parsing-based decoding

Coming soon!

Thank you!

Google *NiuTrans* to find it

Welcome