The HIT-LTRC Machine Translation System for IWSLT 2012

Xiaoning Zhu, Yiming Cui, Conghui Zhu,
Tiejun Zhao and Hailong Cao

Harbin Institute of Technology

Outline

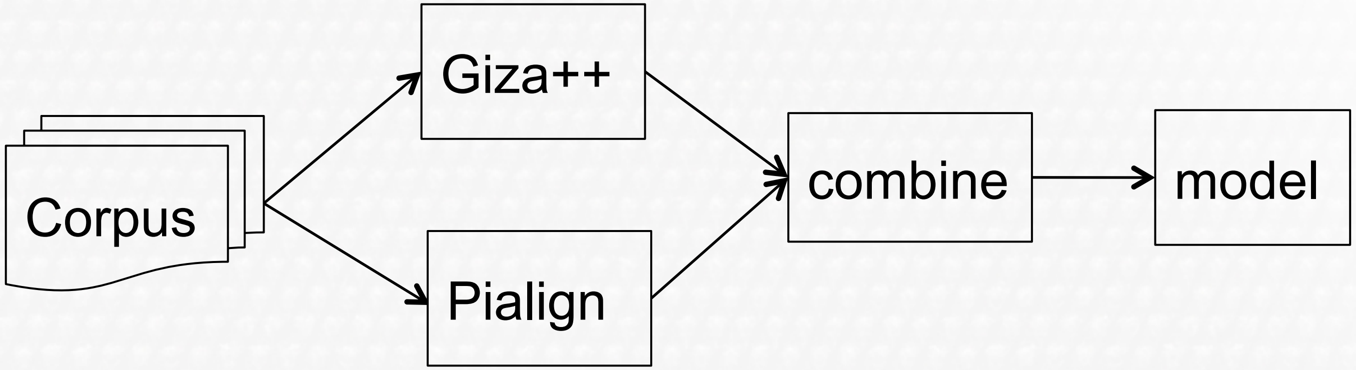
- Introduction
- System summary
- Pialign
- Experiments
- Conclusion and future work

Introduction

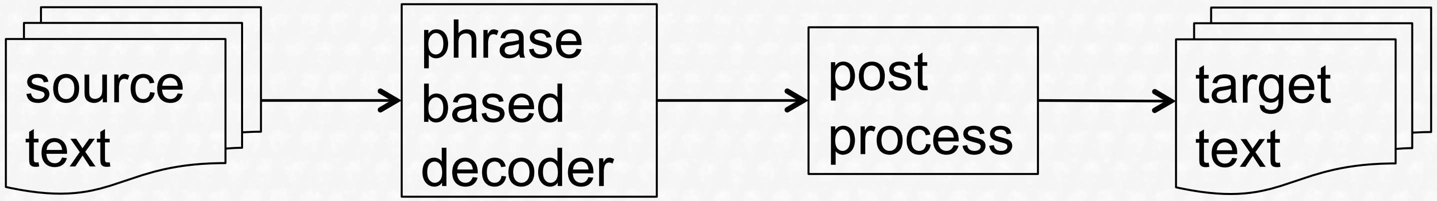
- Olympic shared task
- Phrase-based model
- Phrase table analysis
- Phrase table combination
  - Pialign
  - Giza++

System summary

- Training



- Decoding



System summary

- Tools
  - Moses decoder
  - Giza++ for phrase extraction
  - Pialign for phrase extraction
  - SRILM for language model training
  - Mert for tuning

System summary

- Feature sets
  - Bidirectional translation probabilities
  - Bidirectional lexical translation probabilities
  - MSD-reordering model
  - Distortion model
  - Language model
  - Word penalty
  - Phrase penalty

Pialign

- Phrases of multiple granularities directly modeled
    - + No mismatch between alignment goal and final goal
    - + Completely probabilistic model, no heuristics
    - + Competitive accuracy, smaller phrase table
- Uses a hierarchical model for Inversion Transduction Grammars (ITG)
- Uses Bayesian non-parametric Pitman Yor process

Parameter Tuning of Pialign

- Samps (Sampling frequency)
  - Small: cannot correctly reflect the translation knowledge
  - Big: will produce a sampling bias
  - Finally this value is set to 20 empirically

Sampling times	1	20	80
Phrase table scale	382,137	1,413,367	2,005,941

Experiments

- Corpus
  - HIT_train
  - HIT_dev
  - BTEC_train
  - BTEC_dev

Name	Corpus	#
Corpus 1	BTEC_train+HIT_train	72575
Corpus 2	Corpus 1 + BTEC_dev	75552
Corpus 3	Corpus 2 + HIT_dev	77609

Experiments

- Comparison of Giza++ and Pialign

Corpus	align	total	common	different
1	Giza++	1182913	409443	773470
	Pialign	1385520		976077
2	Giza++	1208128	418788	789340
	Pialign	1413367		994579
3	Giza++	1236688	428377	808306
	Pialign	1445577		1017200

Experiments

- Covering of test set
  - $c = \frac{\text{\# of phrases both in test set and in phrase table}}{\text{\# of phrases in test set}}$

Corpus	align	Chinese	English
1	Giza++	21.7%	36.0%
	Pialign	23.6%	38.3%
2	Giza++	21.7%	36.1%
	Pialign	23.8%	38.7%
3	Giza++	21.9%	36.6%
	Pialign	23.9%	38.9

Experiments

- Translation result with giza++ and pialign
  - After we tuned the parameters with HIT_dev, the result became worse.This may be caused by the mismatch between HIT_dev and HIT_train

Corpus	align	Before tuning	After tuning
1	Giza++	20.76	19.97
	Pialign	20.80	19.70
2	Giza++	20.62	18.40
	Pialign	21.20	19.66
3	Giza++	20.51	15.52
	Pialign	20.54	15.10

Experiments

- Phrase table combination
  - Linear Interpolation

Interpolate parameter	BLEU%
0.4	20.69
0.5	20.78
0.6	20.62

Conclusion and future work

- Tuning may not useful when the dev set not match the training set.
- Pialign can get a better result with a little phrase table