I²R's Machine Translation System for IWSLT 2010

Xiangyu Duan, Rafael E. Banchs , Jun Lang, Deyi Xiong, Aiti Aw , Min Zhang and Haizhou Li
Institute for Infocomm Research, Singapore
{xduan, rembanchs, jlang, dyxiong, aaiti, mzhang, hli}@i2r.a-star.edu.sg

1. Two Layer Machine Translation Framework: individual systems + system combination

Individual Systems

Phrasal Translation System

- **Lavender**: in-house phrase-based decoder
- **Moses**: off the shelf phrase-based decoder

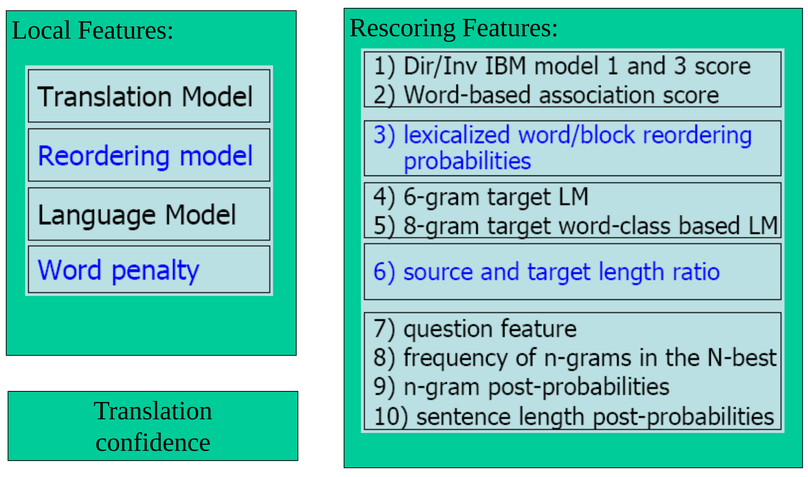
Syntax-based Translation System: Max-entropy-based BTG Translation System

- **Bound**: boundary words based reordering model
- **Lar**: boundary words based reordering + linguistically annotated reordering model
- **UniBrack**: boundary words based reordering + syntax-driven bracketing model

System Combination

Rescoring

Concatenate all system's outputs and rescore them



Confusion Network Based System Combination

Backbone selection
- Backbone determines the word order of final translation

Hypothesis alignment
- build word alignment between backbone and each hypothesis

Confusion network construction
- build confusion network from hypothesis alignment

Confusion network decoding
- find the best translation path through confusion network

2. Experimental Results

2.1 English-to-Chinese

Different Chinese Word Segmentation

	3-gram	4-gram	5-gram
Original	37.89	38.13	38.31
ICTCLAS	35.46	35.95	37.04
NUS tool	37.50	35.64	36.81
Character	35.81	37.39	**38.38**

- Character-base segmentation on Chinese side was adopted.

Data Pre-processing and Parameter Tunning

System	BLEU
Baseline	38.38
+ Training corpus enhanced	40.88
+ Heading and trailing blanks eliminated	44.34
+ English hyphens removed	44.76
+ Decoding parameters adjusted	**45.53**

- Training corpus enhancement: add Chinese-to-English dev data
- Heading and trailing blanks: caused by character segmentation tool
- Decoding parameters: Stack size, table size, and MBR

Individual Systems and System Combination (Rescoring)

System	ASR	CRR
Moses	38.62	45.55
Tranyu(Bound)	37.39	44.01
Lavender	37.62	45.45
Combination	**38.75**	**45.98**

2.2 Chinese-to-English

Different Chinese Word Segmentation

		BLEU	NIST
word segmentation	Original	**0.4603**	**7.4042**
	ICTCLAS	0.3956	6.5231
	NUS tool	0.4038	6.5812
ORI + number translation		0.4587	7.3248

- Original segmentation on Chinese side was adopted.

Word Alignment Combination

		BLEU	NIST
GIZA++	baseline	0.4603	7.2618
	combination	**0.4749**	7.3573
Berkeley	baseline	0.4608	7.1290
	combination	0.4717	**7.4001**

- GIZA++ word alignment combination was adopted.

Rescoring on Each System and Confusion Network Based System Combination

	Rescoring				Confusion Network
	Moses	Tranyu: Bound	Tranyu: UniBrack	Tranyu:LAR	
Before	0.4749	0.4719	0.4726	0.4685	-
After	0.4899	**0.4954**	0.4794	0.4845	**0.5054**

- Rescoring improves each system's performances.
- Confusion Network based system combination improves performance over best single system.