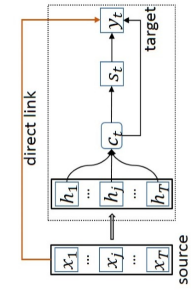




Shaohui Kuang<sup>1</sup> Junhui Li<sup>1</sup> António Branco<sup>2</sup> Weihua Luo<sup>3</sup> Deyi Xiong<sup>1</sup>  
<sup>1</sup>Soochow University, Suzhou, China <sup>2</sup>University of Lisbon, Lisboa, Portuga <sup>3</sup>Alibaba Group, Hangzhou, China  
 shaohuikuang@foxmail.com, {lijunhui, dxyxiong}@suda.edu.cn, antonio.branco@di.fc.ul.pt, weihua.luowh@alibaba-inc.com

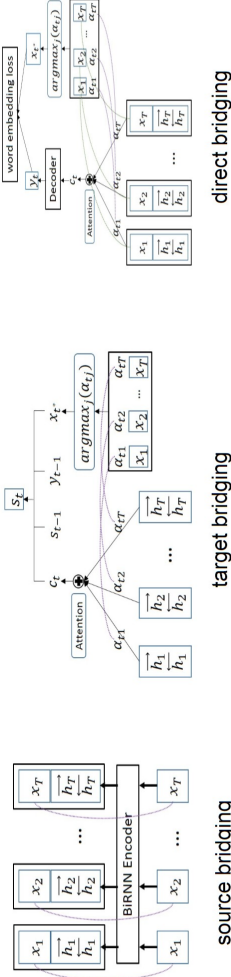
## Motivation

- In NMT, source and target word embeddings locate at the two ends of a long information processing procedure
- The associations and interactions between source and target words are loose in NMT



## Solution

- Focusing on modeling connections between source and target words
- Three methods: Source-side bridging model, Target-side bridging model, Direct bridging model



## Experiments

### Data

Chinese-English bilingual corpora (1.25M sentence pairs)

Model	NIST06	NIST02	NIST03	NIST04	NIST08	Avg
cdc (SMT)	34.00	35.81	34.70	37.15	25.28	33.23
RNNSearch*	35.92	37.88	36.21	38.83	26.30	34.81
Source bridging	36.79†	38.71†	37.24†	40.28†	27.40†	35.91
Target bridging	36.69	39.04†	37.63†	40.41†	27.98†	36.27
Direct bridging	36.97†	39.77†	38.02†	40.83†	27.83†	36.62
cdc (SMT)	58.29	59.65	59.28	58.12	61.54	59.64
RNNSearch*	59.56	57.79	59.25	57.88	64.22	59.78
Source bridging	58.13	56.25	57.33	56.32	62.13	58.01
Target bridging	58.01	56.27	57.76	56.33	62.12	58.12
Direct bridging	57.20	56.68	57.29	55.62	62.49	58.02

## Analysis of Word Alignment

### Better eos translation

System	Percentage (%)
RNNSearch*	49.82
Direct bridging	81.30

Percentage of target side eos translated from source side eos on the development set

### Better Word Translation

System	Target POS Tag		Source POS Tag	
	V	N	V	N
RNNSearch*	64.95	39.24	•	12.09
	7.31	33.37	33.40	•
Direct bridging	•	26.79	14.67	•
	66.29	39.71	•	10.94
Direct bridging	7.19	32.25	56.29	•
	•	26.12	15.22	•

Confusion matrix for translation by part-of-speech (POS)

### Better Word Alignment

System	SAER	AER
RNNSearch*	62.68	47.61
Direct bridging	59.72	44.71

A lower score indicates better alignment

Reference	two warring sides in sri lanka agreed to hold talks in geneva late this month
Baseline	sri lanka UNK to hold talks in geneva eos
Source	斯里兰卡 交成 双方 同意 本(this) 月(month) 下(下)周(week) 在 日内瓦 谈判 eos
Reference	the french athletes participating in special winter olympics returned to paris with honors
Baseline	the french athletes, who have participated in the disabled... have returned to paris. eos
Source	参加 残疾人 冬奥会(winter olympics) 的 法国 运动员 载誉(honors) 返回 巴黎 eos

Examples of NMT output with incorrect alignments of source and target words

## Related Work

- Attention Model: plays a crucial role in the alignment quality (Luong et al., 2015; Cohnet et al., 2016)
- Learning better Word Alignments: eg., from SMT model (Mi et al., 2016; Liu et al., 2016; Arthur et al., 2016)
- Cross-Lingual Embeddings: learn a linear mapping between vector spaces of languages (Mikolov et al., 2013; Gehring et al., 2017)

## Analysis of Over and Under Translation

### Over Translation

System	POS	ROU(%)
RNNSearch*	NN	8.63
	NR	12.92
	DT	4.01
	AD	5.28
Direct bridging	NN	7.56
	NR	10.88
	DT	2.37
	AD	4.49

### Under Translation

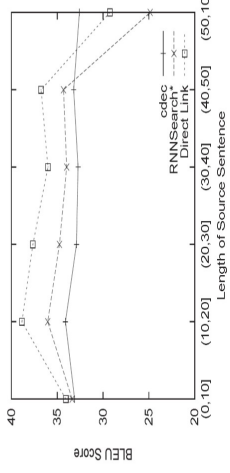
System	1-gram BLEU
cdc (SMT)	77.09
RNNSearch*	72.70
Direct bridging	74.22

Where  $|w|$  is the number of words in consideration,  $t(w_i)$  is the times of over translation for word  $w_i$

$$ROT = \frac{\sum_{w_i} t(w_i)}{|w|}$$

## Analysis of Long Sentence Translation

- During training, the maximum source sentence length limit is set to 50, thus making the learned models not ready to cope well with sentences over this maximum length limit
- For long input sentences, NMT systems tend to stop early in the generation of the translation



## Analysis of Learned Word Embeddings

- The closest target words obtained via the transformation matrix of our direct bridging method are very consistent with those obtained from the SMT lexical table
- Our improved model has a good capability of capturing the translation equivalence between source and target word embeddings

Src	Transformation	Lexical Table
是	is	is
和	and	and
及	and	and
将	will	will
会	will	will
发展	countries	countries
发展	development	development
经济	economic	economic
问题	question	question
人民	people	people