A World Leading SFI Research Centre

Science Foundation Ireland For what's next

Engaging Content
Engaging People

Translating Pro-Drop Languages with Reconstruction Models

Longyue Wang[♥] **Zhaopeng Tu**[♣] **Shuming Shi**[♣] **Tong Zhang**[♣] **Yvette Graham**[♥] **Qun Liu**[♥]

[♥]ADAPT Centre, Dublin City University [♣]Tencent AI Lab

Problem Statement

- **Pronouns** are frequently **omitted** in pro-drop languages (e.g., Chinese, Japanese) especially in **informal genres**.

Genres	Sents	ZH-Pro	EN-Pro	DP
Dialogue	2.15M	1.66M	2.26M	26.55%
Newswire	3.29M	2.27M	2.45M	7.35%

- It leads to significant **challenges** with respect to the production of complete **translations**.

Input	(它) 根本 没 那么 严重	Input	这块 面包 很 美味 ! 你 烤 的 (它) 吗 ?
Ref	**It** is not that bad	Ref	The bread is very tasty ! Did you bake **it** ?
SMT	Wasn 't that bad	SMT	This bread , delicious ! Did you bake ?
NMT	**It** 's not that bad	NMT	The bread is delicious ! Are you baked ?

Novelty of Work

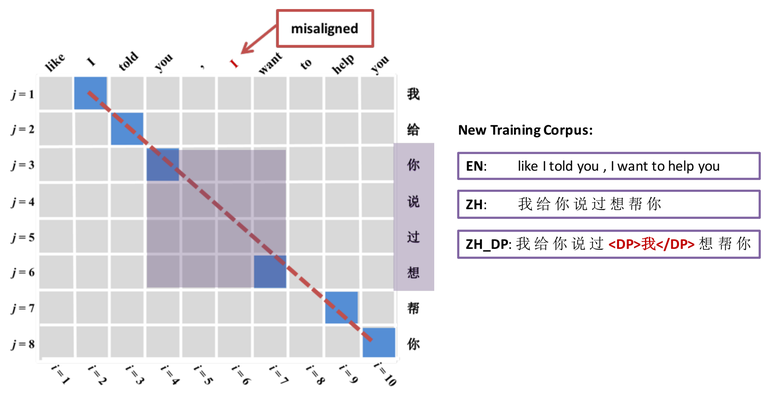
- We show that although NMT models advance SMT models on translating pro-drop languages, there is still **large room for improvement**;

System	Baseline	Oracle	△
SMT	30.16	35.26	+5.10
NMT	31.80	36.73	+4.93

- Little attention has been paid to the problem within NMT. We introduce a **reconstruction-based approach** (+ 3.28 BLEU);
- We release a large-scale **bilingual dialogue corpus** (2.2M Chinese–English sentence pairs).

DP Annotation and Generation

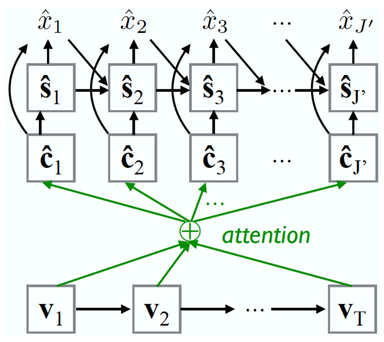
Annotation: as large parallel corpora are usually available, we automatically annotate DP using alignment information.



Generation: we apply RNN for DP position detection and MLP for DP word recovering.

Reconstructor

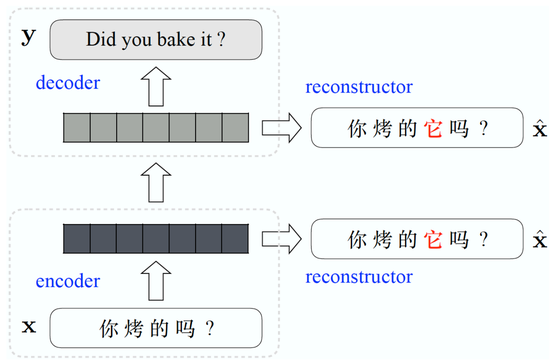
The reconstructor reads a sequence of **hidden states** and the labelled source sentence, and outputs a **reconstruction score**.



$$R(\hat{\mathbf{x}}|\mathbf{v}) = \prod_{j=1}^{J'} R(\hat{x}_j|\hat{x}_{<j}, \mathbf{v}) = \prod_{j=1}^{J'} g_r(\hat{x}_{j-1}, \hat{\mathbf{s}}_j, \hat{\mathbf{c}}_j)$$

Reconstructor-Augmented NMT

Two independent reconstructors with their own parameters, each of which reconstructs the labelled source sentence from the **encoder** and **decoder** hidden states.



$$J(\theta,\gamma,\psi) = \arg\max_{\theta,\gamma,\psi} \sum_{n=1}^{N} \Big\{ \underbrace{\log P(\mathbf{y}^n|\mathbf{x}^n;\theta)}_{likelihood} + \underbrace{\log R_{enc}(\hat{\mathbf{x}}^n|\mathbf{h}^n;\theta,\gamma)}_{enc\text{-}rec} + \underbrace{\log R_{dec}(\hat{\mathbf{x}}^n|\mathbf{s}^n;\theta,\psi)}_{dec\text{-}rec} \Big\}$$

Experiments

- **Data**

Data	|S|	|W|		|P|		|V|		|L|	
		Zh	En	Zh	En	Zh	En	Zh	En
Train	2.15M	12.1M	16.6M	1.66M	2.26M	151K	90.8K	5.63	7.71
Tune	1.09K	6.67K	9.25K	0.76K	1.03K	1.74K	1.35K	6.14	8.52
Test	1.15K	6.71K	9.49K	0.77K	0.96K	1.79K	1.39K	5.82	8.23

- **Main Results**

Model	#Params	Speed		BLEU	
		Training	Decoding	Test	△
Baseline	86.7M	1.60K	2.61	31.80	– / –
Baseline (+DPs)	86.7M	1.59K	2.63	32.67†	+0.87 / –
+ enc-rec	+39.7M	0.71K	2.63	33.67†‡	+1.87 / +1.00
+ dec-rec	+34.1M	0.84K	2.18	33.48†‡	+1.68 / +0.81
+ enc-rec + dec-rec	+73.8M	0.57K	2.16	**35.08**†‡	**+3.28 / +2.41**
Multi-source (Zoph and Knight 2016)	+20.7M	1.17K	1.27	32.81†	+1.01 / +0.14
Multi-layer (Wu et al. 2016)	+27.0M	0.53K	2.12	33.46†‡	+1.62 / +0.79
Enc-Dec-Rec (Tu et al. 2017)	+34.1M	0.87K	2.26	33.08†	+1.28 / +0.41

- **Effect of DP Generation Performance**

Model	Automatic	Manual	△
Baseline (+DPs)	32.67	36.73	+4.06
+ enc-rec	33.67	37.58	+3.91
+ dec-rec	33.48	37.23	+3.75
+ enc-rec + dec-rec	35.08	38.38	+3.30

- **Contribution Analysis**

Model	Test	△
Baseline	31.80	– / –
Baseline (+DPs)	32.67	+0.87 / –
+ enc-rec	33.67	+1.87 / +1.00
+ dec-rec	33.15	+1.35 / +0.48
+ enc-rec + dec-rec	**34.02**	**+2.22 / +1.35**

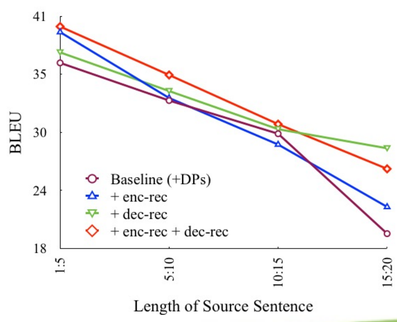
Reconstruction used in training only

- **Effect of Reconstruction**

Model	Test	△
Baseline	31.80	– / –
Baseline (+DPs)	32.67	+0.87 / –
+ enc-rec	33.21	+1.41 / +0.54
+ dec-rec	33.08	+1.28 / +0.41
+ enc-rec + dec-rec	**33.25**	**+1.45 / +0.58**

Replacing DP sentence with original one

- **Length Analysis**



- **Error Analysis**

Model	Error	Sub.	Obj.	Dum.	All
BASE	Total	112	41	45	198
+ ENC	Fixed	51	22	28	101
	New	25	8	4	37
+ DEC	Fixed	57	21	17	95
	New	19	10	6	36
+ ENC + DEC	Fixed	50	34	33	117
	New	11	14	7	32

Fixed Error	
Input	等 我 搬进 来 (**我**) 可以 买 一台 泡泡机 吗 ?
Ref.	When I move in, can I get a bubble machine?
NMT	When I move in *to* buy a bubble machine.
Our	When I move in, can **I** buy a bubble machine?
Non-Fixed Error	
Input	(**他**) 是 个 训练营 ?
Ref.	It is a camp?
NMT	*He* was a camp?
Our	*He*'s a camp?
Newly Introduced Error	
Input	(**我**) 要 把 这 戒指 还给 你
Ref.	I need to give this ring back to you.
NMT	**I**'m gonna give you the ring back.
Our	*To* give it back to you.

DUBLIN CITY UNIVERSITY

Tencent AI Lab

Fondúreacht Eolaíochta Éireann Dá bhfuil romhainn
Science Foundation Ireland For what's next

Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union

European Union
European Regional Development Fund

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.