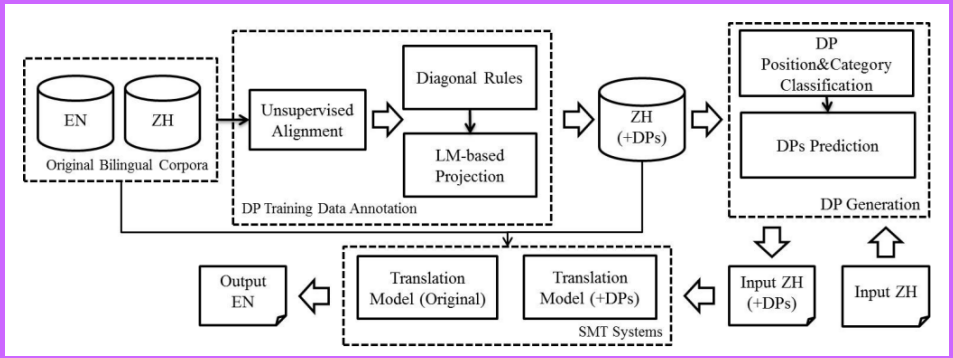
DROPPED PRONOUN GENERATION FOR DIALOGUE MACHINE TRANSLATION

Longyue Wang* Xiaojun Zhang* Zhaopeng Tu** Hang Li** Qun Liu*
*ADAPT Centre, School of Computing, Dublin City University, Ireland
** Noah's Ark Lab, Huawei Technologies, China

Question:

In the dialogue translation from Chinese into English, pronouns as anaphors are frequently dropped in the source language. These dropped pronouns (DPs) may not be the obstacles for human's understanding for people can easily recall the missing pronouns from the context. However, this does not hold in statistical machine translation (SMT) scenario since most missing pronouns fail to be translated appropriately, even with the state-of-the-art translation models.

Solution:



1. Given a parallel corpus, we use the source side as the training data for DP generation, which is automatically annotated with DPs by projecting aligned pronouns at the target side. We believe the first work on recovering DPs by using parallel corpus.
2. With the DP-inserted corpus, we propose a two-phase DP generation approach. First, we employ a sequence labeling model based on recurrent networks (RNNs) to predict the positions and the categories of DPs in the source sentences. Second, we use another *n*-gram language model trained on a larger data to select the best DP candidate.
3. Finally, we apply the DP generator to dialogue machine translation by: 1) training an additional translation model on the DP-inserted corpus; 2) completing the input sentences with the pronouns generated from the DP generation model, before feeding them to the translation system.

1(a) 这块 蛋糕 很美味 。 谁 烤的 （它） ？
1(b) This cake is very tasty . Who did bake **it** ?
2(a) （我） 不知道。 （你） 喜欢 （它） 吗 ？
2(b) **I** don't know . Do **you** like **it** ?

DP	Data	P	R	F1
Position	Dev	0.91	0.83	0.86
+Category	Test	0.87	0.79	0.83
+Pronoun	Dev	0.80	0.77	0.78
	Test	0.76	0.72	0.74

Systems	Dev Set	Test set
Baseline	19.57	17.66
+ DP-ins. TM	20.46	18.32
+ DP-gen. Input	21.59	19.34
Oracle	24.20	20.98

This research is supported by ADAPT Centre for Digital Content Technology funded under the SFI Research Centres Programme (Grant 13/RC/2106) and co-funded under the European Regional Development Fund, and also supported by the DCU-Huawei Joint Project (Grant 201504032 and Grant YB2015090061).