Evaluating Neural Machine Translation in English-Japanese Task

Zhongyuan Zhu

Overview (Abstract)

We evaluated Neural Machine Translation (NMT) models in English-Japanese translation task. Various network architectures with different recurrent units are tested. Additionally, we examine the effect of using pre-reordered data for the training. Our experiments show that even simple NMT models can produce better translations compared with all SMT baselines. For NMT models, recovering unknown words is another key to obtaining good translations. We describe a simple workaround to find missing translations with a back-off system. Surprisingly, performing pre-reordering on the training data hurts the model performance. We provide a qualitative analysis demonstrates a specific error pattern in NMT translations which omits partial information and thus fail to preserve the complete meaning.

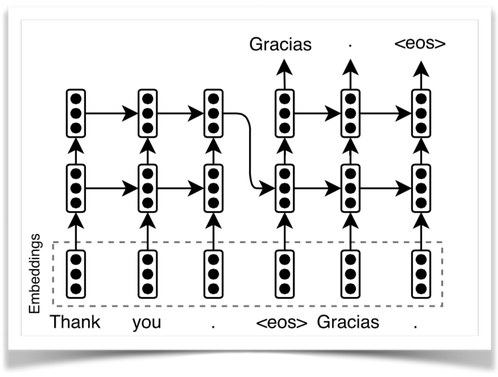
Evaluation results in English-Japanese task

	BLEU	RIBES	HUMAN
BASELINE T2S SMT	33.44	0.758	30.00
Ensemble of 2 LSTM Search	33.38	0.800	-
+ UNK replacing (submitted system 1)	**34.19**	**0.802**	**43.50**
+ System combination	35.97	0.807	-
+ 3 pre-reordered ensembles (submitted system 2)	**36.21**	**0.809**	**53.75**

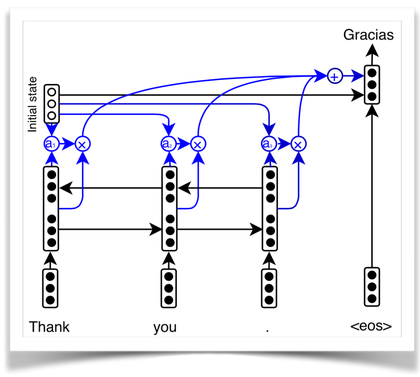
(JPO adequacy evaluation result of system 2: 3.81, best competitor: 4.04)

Experimental details

- A comparison of two network architectures

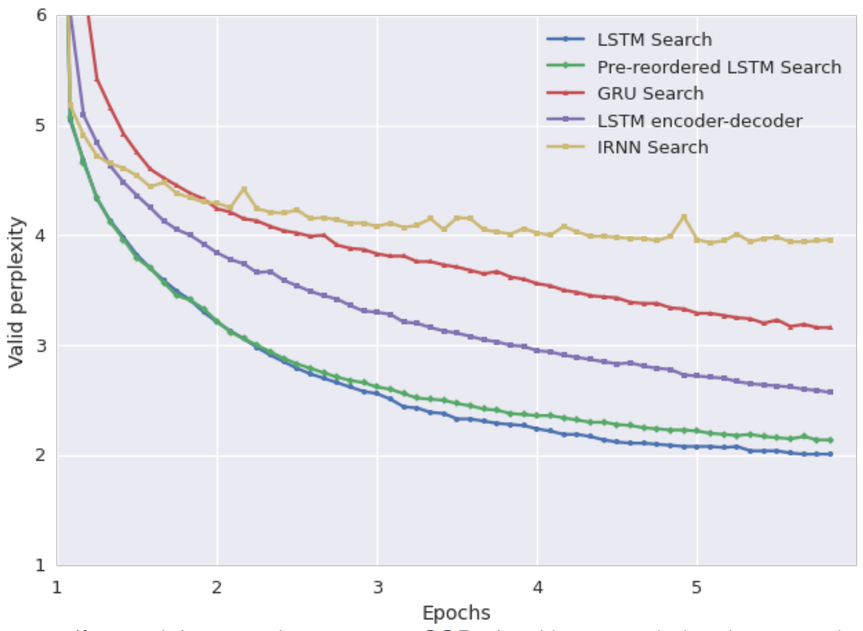


multi-layer encoder-decoder model



soft-attention model

- Visualization of the training process for different models



LSTM Search: soft-attention model with LSTM units
Pre-reordered LSTM Search: soft-attention model with LSTM units trained on pre-reordered data
GRU Search: soft-attention model with GRU units
LSTM encoder-decoder: 4-layer encoder-decoder model with LSTM units
IRNN Search: soft-attention model with IRNN units

(for model comparison, we use SGD algorithm to optimize the network, details are presented in the paper)

- Problem of unknown words

Replacing unknown words in the target side with "$unkpos_i$" (Luong et al., 2015) works well with soft-attention models trained on pre-reordered data. However, for models trained on data of natural order, other sophisticated solutions are required.

A simple workaround is to find the missing word in the translation result of a baseline system. As for the same target word, they usually share similar context even in different translations.

Findings

- Soft-attention models outperforms multi-layer encoder-decoder models

The evaluation of valid perplexity shows that soft-attention models outperforms simple encoder-decoder models with a substantial margin. This matches our expectation as the alignment between English and Japanese are far more complicated than English-French pair.

- Training models on pre-reordered data hurts the performance

Both the perplexity on valid data and automatic evaluation scores show that training soft-attention LSTM models on pre-reordered data degrades the performance.

	BLEU	RIBES
Single LSTM Search	32.19	0.797
Pre-reordered LSTM Search	30.97	0.779

- NMT models tend to make grammatically valid but incomplete translations

Input	this paper discusses some systematic uncertainties including casimir force , false force due to electric force , and various factors for irregular uncertainties due to patch field and detector noise .
NMT result	ここでは，Ｃａｓｉｍｉｒ力を考慮したいくつかの系統的不確実性について論じた。
Reference	Ｃａｓｉｍｉｒ力や電気力による偽の力，パッチ場や検出器雑音による不規則な不確かさの種々の要因を含め，幾つかの系統的不確かさを論じた。

Retrospection

We conducted a detailed qualitative analysis on a held-out development dataset. The existence of unknown words are found to drastically degrade the quality of translations. Even the missing word can be posteriorly recovered, some of the translations are still unnatural. In our experiments, we set vocabulary size to 80k and 40k for the input and output layer respectively. Increasing these numbers will significantly slow down the training. Overcoming this problem is expected to be the key of obtaining qualitative translations for NMT models.

Weblio Inc.