Задача машинного перевода для низкоресурсных языков

К. Ф. Сафин
Научный руководитель: к.ф.-м.н. Ю.В.Чехович

Московский физико-технический институт

6 декабря 2018 г.

Цель исследования

Цель

- Построить алгоритм обучения модели машинного перевода для низкоресурсных языков.
- Низкоресурсные языки — языки, для которых недостаточно данных, необходимых для обучения систем машинного перевода на достаточном уровне качества.

Цель исследования

Актуальность

- Самостоятельная задача построения переводных систем для редких пар языков.
- Вспомогательная задача в решении более общих задач — обнаружения заимствований текстов, сопоставление данных на разных языках и т.п.

Литература

G. Lample, M. Ott, A. Conneau, L. Denoyer, M. Ranzato

Phrase-Based & Neural Unsupervised Machine Translation.

arXiv, 2018.

G. Lample, A. Conneau, L. Denoyer, M. Ranzato

Unsupervised Machine Translation Using Monolingual Corpora Only.

ICLR 2018.

M. Artetxe, G. Labaka, E. Agirre, K. Cho

Unsupervised Neural Machine Translation.

ICLR 2018.

R. Sennrich, B. Haddow, A. Birch

Improving Neural Machine Translation Models with Monolingual Data.

ACL 2016.

D. Erhan, Y. Bengio, A. Courville, P. Manzagol, P. Vincent

Why Does Unsupervised Pre-training Help Deep Learning?.

Journal of Machine Learning Research, 2010.

Обшая задача машинного перевода

Дано:

$$\mathbf{X} - \textit{source language},$$

$$\mathbf{Y} - \textit{target language}.$$

Построить модель перевода m:

$$y = m(x), x \in \mathbf{X}, y \in \mathbf{Y}$$

Качество перевода:
Популярная (но не точная) мера качества BLEU:
доля совпавших n-gram истинного и машинного перевода.

Существующие модели

SMT (Statistical Machine Translation)
Вероятностный подход:

$$y_{trans} = \arg\max_y \mathbb{P}(y|x) = \arg\max_y \mathbb{P}(x|y)\mathbb{P}(y),$$

$\mathbb{P}(x|y)$ — translation table,
$\mathbb{P}(y)$ — language model.
Особенности:

- Быстрое обучение.
- Компоненты модели обучаются отдельно.
- Требуется большая выборка параллельных предложений для обучения.
- Плохая согласованность слов.

Существующие модели

NMT (Neural Machine Translation)
Векторизация предложений:

$$\mathbf{X} \xrightarrow[\text{encode}]{\mathbf{f}} \mathbf{H} \xrightarrow[\text{decode}]{\mathbf{g}} \mathbf{X}.$$

Особенности:

- Долгая настройка.
- Требуется большая выборка параллельных предложений для обучения.
- Более согласованное построение предложений.

Исследуемые подходы

- Двуязычные векторные представления слов.
- Искуственная генерация параллельных предожений.
- Backtranslation метод обучения.
- Перевод со вспомогательным (pivot) языком.
- Совмещение различных существующих подходов.

Базовые эксперименты

Train: 10k параллельных предложений корпуса WMT'17.
Test: 1k предложений.
Качество unsupervised подходов: 30-40 BLEU

	Word-by-word translation	seq2seq	moses (SMT)
BLEU	2.1	6,4	8,7

Спасибо за внимание!