Обучение машинного перевода без параллельных текстов

Саттаров Тагир

Московский физико-технический институт

Курс: Численные методы обучения по прецедентам (практика, В.В. Стрижов)/Группа 674, весна 2019

Цель исследования

Цель работы

Улучшить качество перевода на низкоресурсных парах языков

Проблема

При переводе некоторых пар языков необходимо учитывать морфолгическое строение слов, что существенно усложняет задачу машинного перевода

Решение

Прдлагается построить общее скрытое пространство для обоих языков. Предлагается использовать seq2seq модель кодировщика-декодировщика и дискриминатора

Литература

- G Lample, A. Conneau, L. Denoyer, M. Ranzato. Unsupervised Machine Translation Using Monolingual Corpora Only. Conference paper at ICLR 2018
- A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jegou. Word translation without parallel data. ICLR 2018
- A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, L-P Morency. Multi-Attention Recurrent Network for Human Communication Comprehension. AAAI-18

Постановка задачи

Дано

Заданы $\mathfrak{D}_{src}$ и $\mathfrak{D}_{tgt}$, где $\mathfrak{D}_{src}$ – выборка предложений на английском языке, а $\mathfrak{D}_{tgt}$ – на французском. Также задана $\mathfrak{D}_{valid}$ – предложения на анлийском, переведенные на французский для проверки качества перевода.

Используются модели

Дискриминатора $\mathbf{d}$ – по вектору из скрытого пространства $\mathbf{Z}$ предсказывает исходный язык

Кодировщика $\mathbf{f}$ – переводит предложение из одного из языков в скрытое пространство

Декодировщика $\mathbf{g}$ – восстанавливает предложение по вектору из скрытого пространства

Общая схема решения

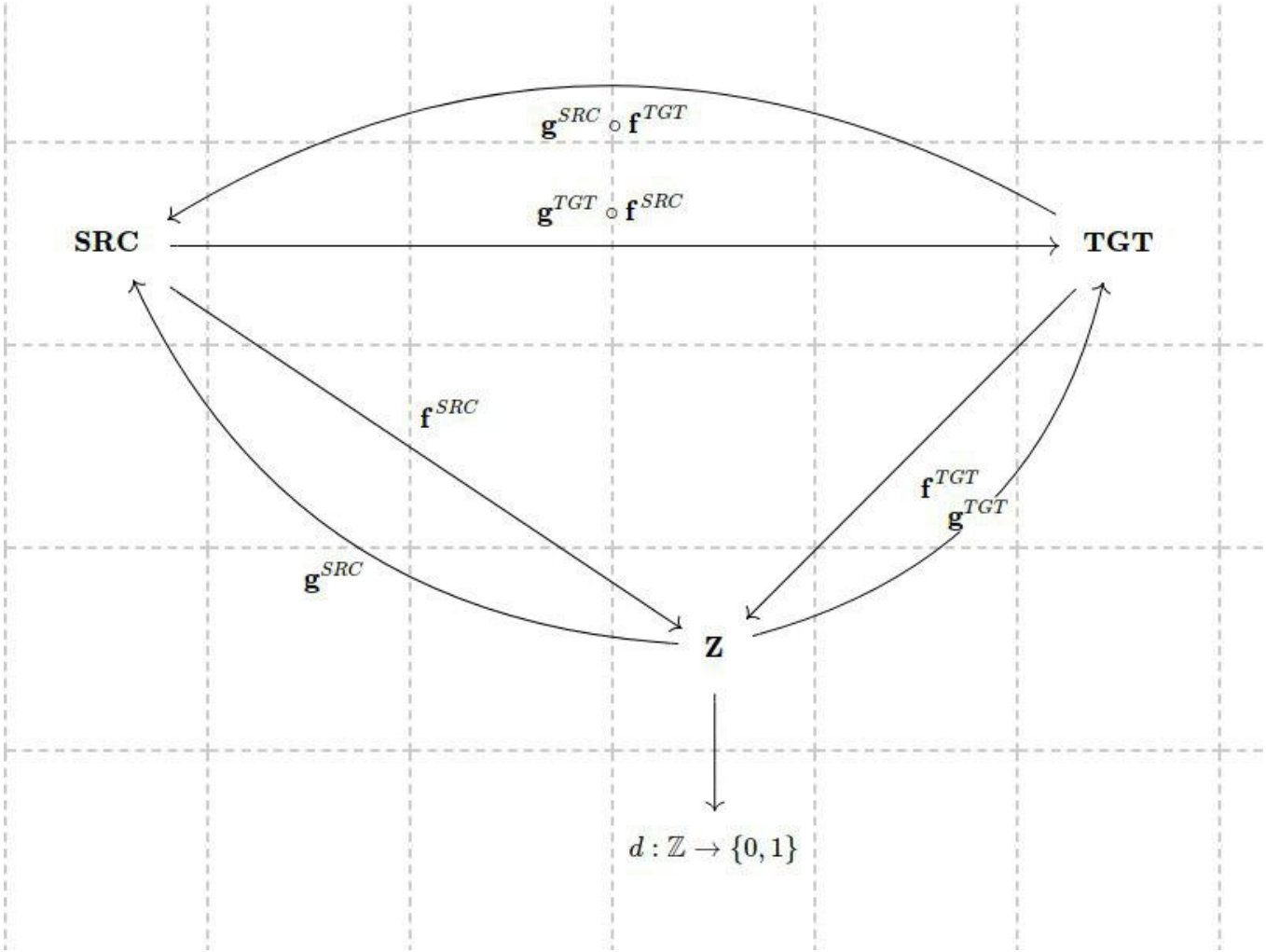


Рис.: Auto-encode

Требования к кодировщику и декодировщику

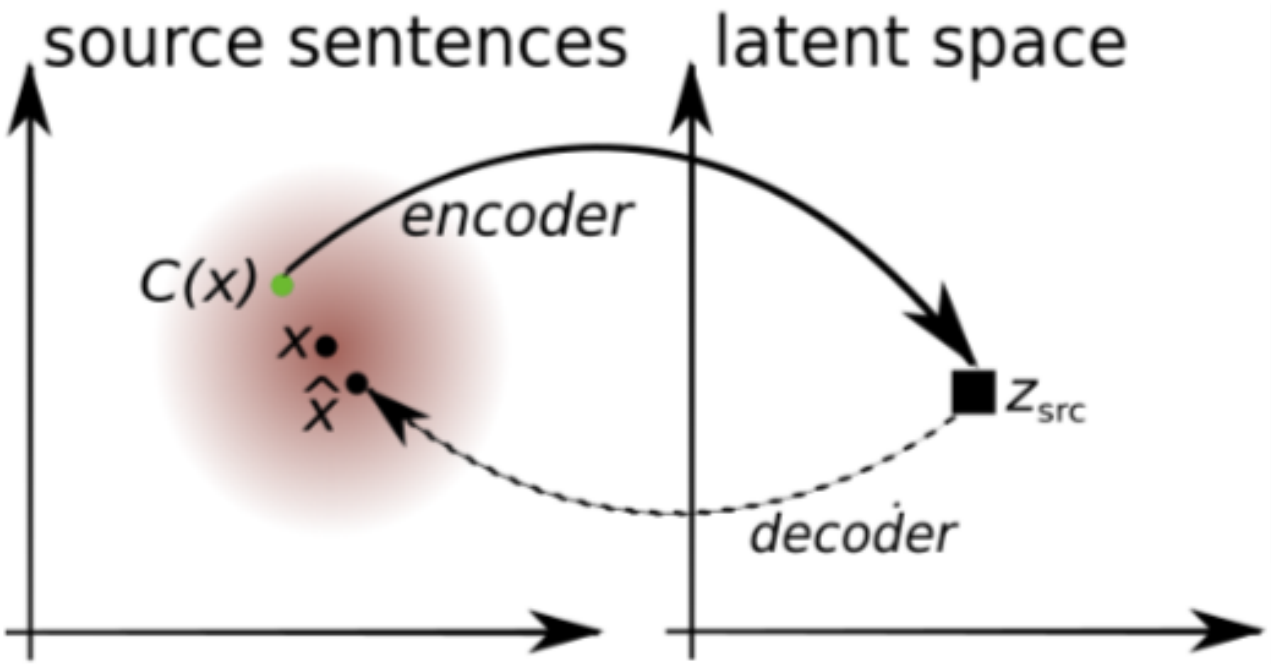


Рис.: Auto-encode

Требования к кодировщику и декодировщику

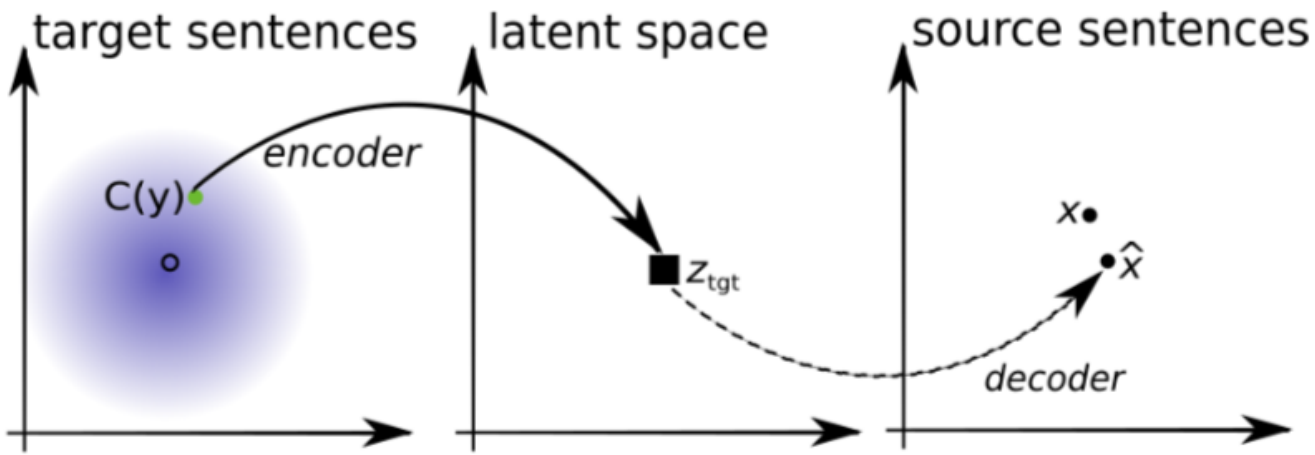


Рис.: Cross-domain

Критерий качества перевода

example

X_{auto}: am wearing older person a blue sitting on jacket a bench eating cream ice . cone
Y_{auto}: am older person wearing a blue jacket sitting on a bench eating an ice cream cone .
X_{cross}: deux hommes dans une sont bureaux wrestling deux comme autres regarder hommes .
Y_{cross}: two men in an office are wrestling as two other men watch on .

Критерий качества перевода

Дано

Один из наиболее распространенных критериев качества – BLEU(bilingual evaluation understudy).

Опредение

N-грамма – последовательность из N подряд идущих слов.

Метод

Рассматривается несколько вариантов профессианального перевода и смотрится количество N-грамм, лежащих и в машинном переводе, и в данной выборке.

Критерий качества перевода

Cand 2: Mary did not give a smack to a green witch.

Ref 1: Mary did not slap the green witch.
Ref 2: Mary did not smack the green witch.
Ref 3: Mary did not hit a green sorceress.

Cand 2 точность 2-грамм: 4/9

Критерий качества перевода

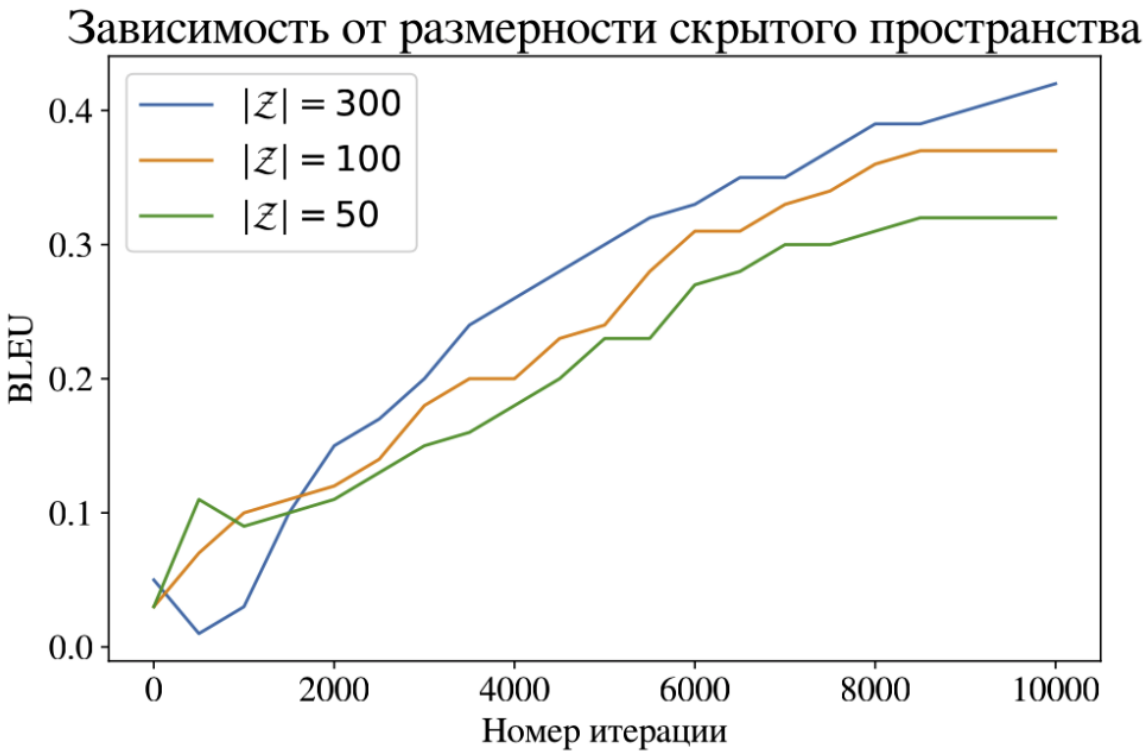
Cand 2: Mary did not give a smack to a green witch.

Ref 1: Mary did not slap the green witch.
Ref 2: Mary did not smack the green witch.
Ref 3: Mary did not hit a green sorceress.

Для каждой N-граммы счетчик не должен превышать максимального количества этой n-граммы в любом предложении

Cand 2 точность 1-грамм: 7/10

Вычислительный эксперимент



Заключение

- Применили модель для пары en->fr
- В будущем планируется использовать модель с четремя attention-ми и, возможно, применить Tensor2Tensor модель вместо seq2seq