

Домашнее задание 2. Машинный перевод

Deadline: 25.04.2019

Домашнее задание посвящено задаче машинного перевода. Мы рассмотрим несколько сценариев использования стандартной архитектуры seq2seq на основе рекуррентных нейронных сетей. Домашнее задание частично основывается на материалах [курса Johns Hopkins University](#).

Данные – мультязычный параллельный корпус: [доступен по ссылке](#). Будем работать с частью корпуса Multilingual Parallel Corpus.

Для оценки ваших решений используйте метрику BLEU, реализованную, в том числе, в NLTK (`nltk.translate.bleu_score`).

Задание 1 (6 баллов) Реализуйте стандартную архитектуру МТ [1, 2]:

- RNN-энкодер
- RNN-декодер
- механизм внимания

Протестируйте эту архитектуру на паре языков из мультязычного корпуса. Рекомендуем выбирать дистантные (неродственные) языки и переводить на знакомый вам язык.

Разбиение на обучающее и тестовое множество проведите любым образом, который кажется вам разумным. Попробуйте дать не только формальную, но субъективную оценку результатам.

Задание 2 (макс. 7 баллов) Реализуйте следующие идеи развития модели:

- beam-search при декодировании [1] (2 балла)
- в дополнение к эмбедингам слов – символьное представление входных слов или BPE в энкодере [3] (1 балл)
- другие варианты механизма внимания (аддитивные или мультипликативные варианты механизма внимания, скалярное произведение) [4] (2 балла)
- извлечение именованных сущностей и перевод именованных сущностей по словарю (например, топонимы можно переводить по дереву категорий Википедии) (1 балл)
- разные принципы формирования мини-батчей: по длине предложения на исходном языке, по длине предложения на целевом языке и др. [5] (1 балл)

Снова попытайтесь дать не только формальную, но субъективную оценку результатам – какая модификация большего всего влияет на качество результатов? Почему?

Задание 3 (4 балла)

Теперь будем переводить с нескольких языков одновременно на один целевой язык. Реализуйте архитектуру, в которой три энкодера и один декодер – т.н. мультиэнкодер. В этой части задания вы столкнетесь с проблемой неполноты данных: часть предложений на каких-то языках будет отсутствовать. Эту проблему можно решить двумя способами: не работать с неполными данными или использовать эвристику, предложенную в работе [6] – заменить предложения на специальную метку NULL.

Для этого эксперимента вам понадобится выбрать два дополнительных к предыдущим заданиям языка. Снова попытайтесь дать не только формальную, но субъективную оценку результатам – как использование дополнительных языков повлияло на качество перевода?

Задание 4 (бонус, до 3 баллов)

1. Решите любое задание с использованием альтернативных seq2seq архитектур (например, Transformer).
2. Попробуйте использовать методы ускорения обучения [7].
3. Визуализируйте и проанализируйте карты внимания.

Рекомендуемое чтение

1. *Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le* “Sequence to sequence learning with neural networks.” In Advances in neural information processing systems, pp. 3104-3112. 2014
2. *Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio* “Neural machine translation by jointly learning to align and translate.”. 2014. arXiv preprint arXiv:1409.0473.
3. *Ling, Wang, Isabel Trancoso, Chris Dyer, and Alan W. Black* “Character-based neural machine translation.”. 2015. arXiv preprint arXiv:1511.04586.
4. *Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning* “Effective approaches to attention-based neural machine translation.”. 2015. arXiv preprint arXiv:1508.04025.
5. *Morishita, Makoto, Yusuke Oda, Graham Neubig, Koichiro Yoshino, Katsuhito Sudoh, and Satoshi Nakamura* “An empirical study of mini-batch creation strategies for neural machine translation.”. 2017. arXiv preprint arXiv:1706.05765.

6. *Nishimura, Yuta, Katsuhito Sudoh, Graham Neubig, and Satoshi Nakamura* “Multi-Source Neural Machine Translation with Missing Data.”. 2018. arXiv preprint arXiv:1806.02525.
7. *Ott, Myle, Sergey Edunov, David Grangier, and Michael Auli* “Scaling Neural Machine Translation.”. 2018. arXiv preprint arXiv:1806.00187.