Domain Adaptation

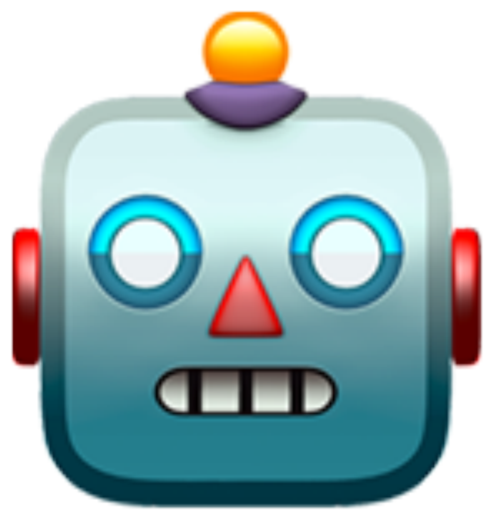
Neural Machine Translation

Yasmin Moslem, ADAPT Centre, Dublin City University

Motivation

Story

- You got a new Machine Translation project.
- Surprise: in-domain dataset is too small
- Generic MT model: Vice President vs. Deputy Chairperson
- What to do?!



Solution

Domain Adaptation

- Big generic/out-of-domain dataset (vocabulary & syntax)
- Small specialised/in-domain dataset (terminology & style)

Quiz

- Think for two domains or two languages; one has a lot of resources and one has limited resources.

Combining Training Data

Does it work?

- Hint: Just combining both the big generic corpus and the specialised one will ***not*** work. Why?

Combining Training Data

Does it work?

- Hint: Just combining both the big generic corpus and the specialised one will ***not*** work. Why?
    - Retraining for a long time
    - In-domain data is small and its effect is limited

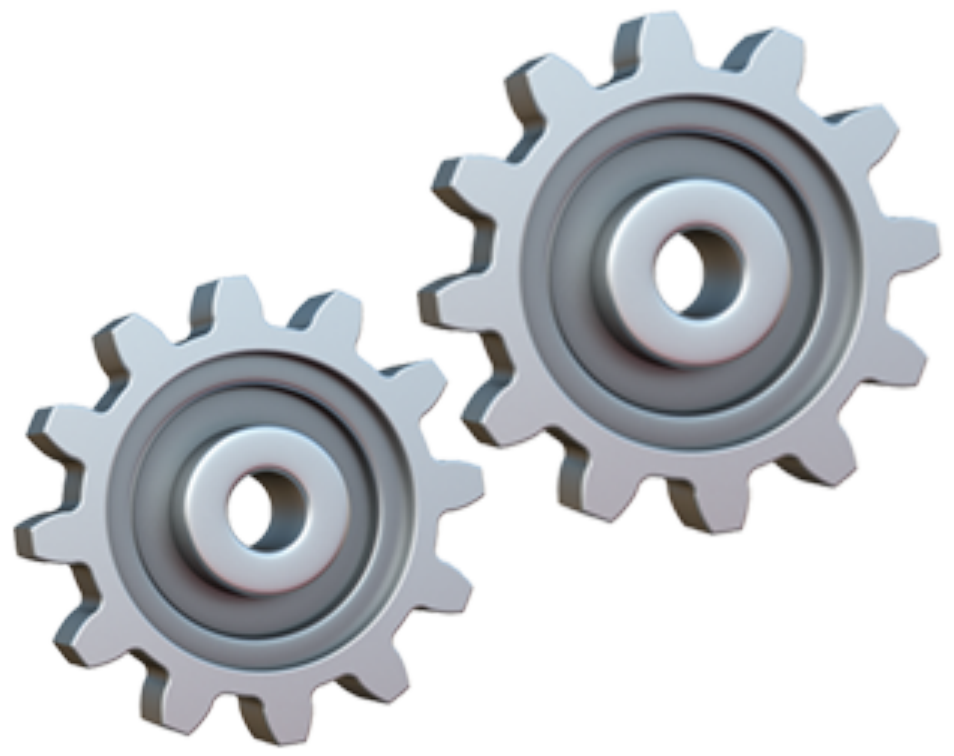
Combining Training Data

Does it work?

- Hint: Just combining both the big generic corpus and the specialised one will ***not*** work. Why?
  - Retraining for a long time
  - In-domain data is small and its effect is limited
- Suggestions to make it work?

Domain Adaptation Approaches
(that work!)

- Incremental Training / Fine-tuning
- Ensemble Decoding
- Data Weighting
- Using Monolingual Synthetic Data
- Iterative/Multilingual Transfer Learning
- On-the-fly Domain Adaptation



Incremental Training / Fine-tuning

Steps

1. Train a baseline model on the generic/out-of-domain corpus

2. Continue training the baseline model on the in-domain corpus (e.g. Luong & Manning, 2015)

Incremental Training / Fine-tuning

catastrophic forgetting

- Generic sentences are translated badly (e.g. unidiomatic structure or unknown words) by the fine-tuned model while they are translated better by the baseline model.
- Solution?
  - Avoid fine-tuning for too many steps/epochs.
  - In the fine-tuning step, add out-of-domain data to the in-domain data (e.g. (Chu et al., 2017)

Ensemble Decoding

- Ensemble Decoding: a method that allows using multiple models simultaneously during the translation time.
- Preprocessing requirements: include vocabulary of both corpora
- Ensemble the base model with the fine-tuned model (Freitag & Al-Onaizan, 2016)
- Better than using the fine-tuned model as it helps avoid “over-fitting”.

Data Weighting

- **Approach 1:** train one model on two corpora at the same time while giving a higher weight for the specialised corpus over the other generic corpus, or
- **Approach 2:** train the model on only one corpus that includes both generic segments and specialised segments, giving higher weights for specialised segments.

Using Monolingual Synthetic Data

Approach 1: Back Translation

- Improving Neural Machine Translation Models with Monolingual Data (Sennrich et al., 2016):

    1. Utilising monolingual *target sentences* after filling the source side with back-translation.

    2. Mixing the new synthetic data with parallel data for either training or fine-tuning.

Using Monolingual Synthetic Data

Approach 2: Pseudo In-Domain Data Selection

- Domain Adaptation via Pseudo In-Domain Data Selection (Chinea-Ríos et all., 2017):

  1. Selecting, from a large monolingual pool of sentences <u>in the source language</u>, those instances that are more related to a given [in-domain] test-set. (Chinea-Ríos et all., 2017 or Axelrod et al., 2011)

  2. Next, this selection is automatically translated and the generic baseline neural machine translation system is fine-tuned with this data.

Using Synthetic Data

- Better OOV Translation with Bilingual Terminology Mining (Huck et al., 2019):
    1. Translate the text and get OOVs.
    2. Translate them using bilingual word embeddings (BWEs), created with MUSE, and take the 5–best candidates.
    3. Using the 5 proposed target language words as queries to mine target-language sentences.
    4. Back-translate the sentences, forcing the back-translation of each of the five proposed target-language OOV-translation-candidates to be the original source-language OOV.
    5. Use this synthetic data to fine-tune the system; as a result, the translation of OOVs can be dramatically improved.

Iterative Transfer Learning

- UCAM Biomedical translation at WMT19: Transfer learning multi-domain ensembles (Saunders et al., 2019):
  - Iterative transfer learning, B→A→B:
    1. Fine-tune the in-domain model B with the generic model A.
    2. Fine-tune the generic model A with the in-domain model B.
  - Conditions:
    - Large in-domain data.
    - Similar to the generic data.

Multilingual Transfer Learning

- Exploiting Out-of-Domain Parallel Data through Multilingual Transfer Learning for Low-Resource Neural Machine Translation (Imankulova et al. 2019):

  1. Following the idea of multilingual zero-shot (Johnson et al., 2017) - M2M (multi-to-multi) model by adding an artificial token that specifies the target language to the beginning of each source sentence and shuffling the entire training data

  2. Multi-stage Fine-tuning: Ja↔Ru MT system: 1) train a multilingual NMT model on out-of-domain Ja↔En and Ru↔En data; 2) fine-tune it on in-domain Ja↔En and Ru↔En data; and 3) further fine-tune it on Ja↔Ru data.

  3. Using synthetic data via back-translation (with M2M) is useful and provides the best system.

On-the-fly Domain Adaptation

- Multi-Domain Neural Machine Translation through Unsupervised Adaptation (Farajian et al., 2017):

    1. Given an input sentence *q*, extract from the pool of parallel data the top (*source, target*) pairs in terms of similarity between the *source* and *q*.

    2. Use the retrieved pairs to fine-tune the model, which is then applied to translate *q*.

    3. Reset the adapted model to the original parameters, translate the next input sentence, and so on.

- Other approaches:

    - Domain Adaptive Inference for Neural Machine Translation (Saunders, et all., 2019)

    - Compact Personalized Models for Neural Machine Translation (Wuebker, et all., 2018)

Final Note: Full Words vs. Sub-words

- **Methods of sub-wording:** Byte Pair Encoding (BPE) & unigram language model
- **Popular Segmentation Strategy:** SentencePiece
- **Sub-wording can help in cases:**
  - Word variations in the same language, e.g. “translate vs. translation”
  - Compound words in the same language, e.g. “multi-tasking”. So now you model is not only able to translate “multi-tasking”, but any other phase that includes the word “multi”.
  - Shared words between languages
  - Common misspellings, like forgetting accents.

References

- Amittai Axelrod & Xiaodong He, JianfengGao. (2011). Domain Adaptation via Pseudo In-Domain Data Selection. EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, *Proceedings of the Conference*. 355-362.
- Mara Chinea-Ríos, Álvaro Peris, Francisco Casacuberta. (2017). Adapting Neural Machine Translation with Parallel Synthetic Data. *Proceedings of the Second Conference on Machine Translation* (pp. 138–147). Association for Computational Linguistics.
- Chenhui Chu, Raj Dabre, Sadao Kurohashi. (2017). An Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers) (pp. 385–391). Association for Computational Linguistics.
- Chenhui Chu, Rui Wang. (2018). A Survey of Domain Adaptation for Neural Machine Translation. *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1304–1319). Association for Computational Linguistics.
- M. Amin Farajian, Marco Turchi, Matteo Negri, Marcello Federico. (2017). Multi-Domain Neural Machine Translation through Unsupervised Adaptation. *Proceedings of the Conference on Machine Translation (WMT), Volume 1: Research Papers*, pages 127–137 Copenhagen, Denmark, September 7-11, 2017.
- Markus Freitag, Yaser Al-Onaizan. (2016). Fast Domain Adaptation for Neural Machine Translation.
- Matthias Huck, Viktor Hangya and Alexander Fraser. (2019). Better OOV Translation with Bilingual Terminology Mining. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5809–5815). Association for Computational Linguistics.
- Aizhan Imankulova, Raj Dabre, Atsushi Fujita, Kenji Imamura. (2019). Exploiting Out-of-Domain Parallel Data through Multilingual Transfer Learning for Low-Resource Neural Machine Translation. *Proceedings of Machine Translation Summit XVII Volume 1: Research Track* (pp. 128–139). European Association for Machine Translation.
- Minh-Thang Luong, Christopher D. Manning. (2015). Stanford Neural Machine Translation Systems for Spoken Language Domains. *Proceedings of the International Workshop on Spoken Language Translation*.
- Rico Sennrich, Barry Haddow, Alexandra Birch. (2016). Improving Neural Machine Translation Models with Monolingual Data. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 86–96). Association for Computational Linguistics.
- Danielle Saunders, Felix Stahlberg, Adria de Gispert, & Bill Byrne. (2019). Domain Adaptive Inference for Neural Machine Translation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 222–228). Association for Computational Linguistics.
- Danielle Saunders, Felix Stahlberg, Bill Byrne. (2019). UCAM Biomedical translation at WMT19: Transfer learning multi-domain ensembles. *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers*, Day 2) (pp. 169–174). Association for Computational Linguistics.
- Sander Tars, Mark Fishel. (2018). Multi-Domain Neural Machine Translation.
- Joern Wuebker, Patrick Simianer, John DeNero. (2018). Compact Personalized Models for Neural Machine Translation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 881–886). Association for Computational Linguistics.

Let’s Connect!

- LinkedIn: https://www.linkedin.com/in/yasmin-moslem/
- Website: https://machinetranslation.io/
- Email: yasmin@machinetranslation.io