-SOTA-
Document-level Neural Machine Translation

by Lorenzo Lupo

April 2020

Plan

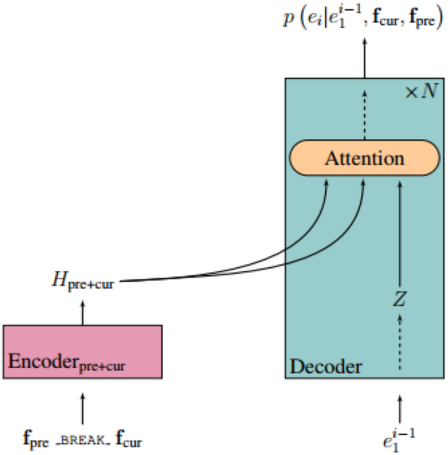
Plan

Overview

- gnee

Concatenation Approaches

Concatenation approaches to DLNMT consist in feeding a standard encoder-decoder architecture with a concatenation of sentences.



Concatenation Approaches

For instance:

- [Tiedemann and Scherrer, 2017] firstly introduced this approach proposing an RNN-based model that incorporate the preceding sentence by prepending it to the current one, separated by a <CONCAT> token. They propose two methods:
  - **2-TO-2**: the previous and the current sentences are translated together. The translation of the current sentence is then obtained by only retaining the tokens following the concatenation token.
  - **2-TO-1**: only the current sentence is translated.
- [Agrawal et al., 2018, Scherrer et al., 2019] investigated the concatenation approach with the Transformer as base model, extending the number of context sentences both on the source (s:-3,+1) and the target (t:-2) side.

Separate Encoding Approaches

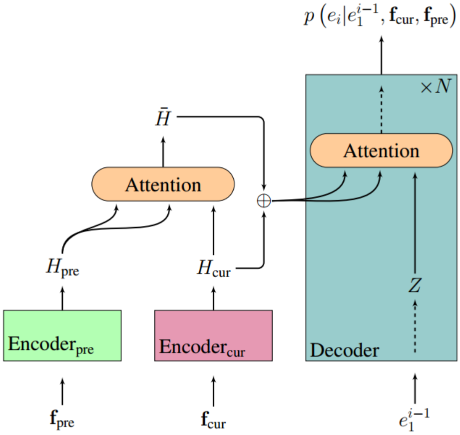
Separate encoding approaches to DLNMT consist in encoder-decoder models that encode the current and context sentences separately. This can be undertaken by:

- **Multiple encoders** working in parallel for the current and previous sentence. E.g. [Wang et al., 2017].
- **Multiple encoders with shared weights**. In this case, the parallel-working encoders not only have the same architecture, but also the same weights. E.g. [Voita et al., 2018].
- **Two-pass approaches**, in which the encoder makes a first sentence-level encoding pass of the source, and a second in which it encodes contextual information too. See Slide 13.

Separate Encoding Approaches

Once the encoding of the current and the context sentences has been carried out, they can be integrated in different ways:

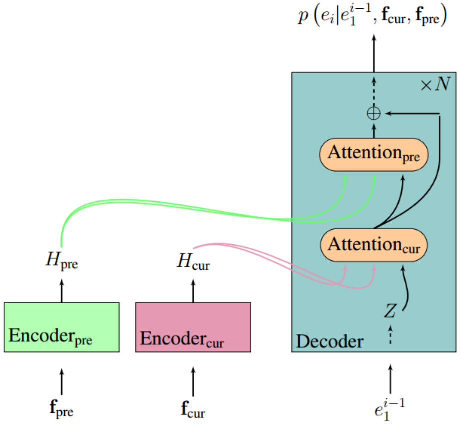
- **Outside** the decoder.
  - $(+)$ symbol represents a gate, a sum or a concatenation.



Separate Encoding Approaches

Once the encoding of the current and the context sentences has been carried out, they can be integrated in different ways:

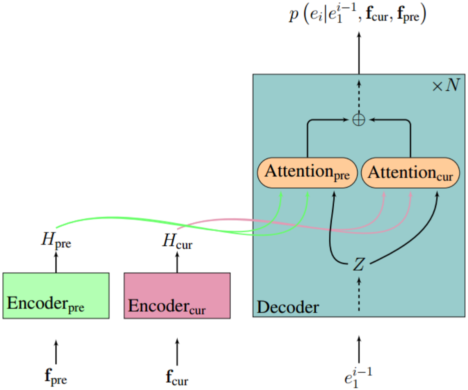
- **Outside** the decoder.
  - $(+)$ symbol represents a gate, a sum or a concatenation.
- **Inside** the decoder, **sequentially**.



Separate Encoding Approaches

Once the encoding of the current and the context sentences has been carried out, they can be integrated in different ways:

- **Outside** the decoder.
  - $(+)$ symbol represents a gate, a sum or a concatenation.
- **Inside** the decoder, **sequentially**.
- **Inside** the decoder, **in parallel**.



Separate Encoding Approaches

Architecture

The encoder-decoder architectures depicted above can be both RNN-based (until 2017) or Transfomer-based (after 2017), as for any approach to DLNMT. However, often some modifications are applied. For example:

- In the case of RNN-based architectures, integration inside the decoder can be undertaken without attention by simply concatenating context representations to the cell state of the deocdrr's RNN [Wang et al., 2017].
- beside contextual representation of words, the context encoder can also higher level representations such as sentence or document representations. This representations can also be attended by the decoder [Miculicich et al., 2018, Maruf et al., 2019] or added to the word-representations [Tan et al., 2019].
- Parallel integration inside the decoder can also happen within a single multi-head attention that takes as values and queries the concatenations of the current and context sentence representations [Voita et al., 2019b]

Separate Encoding Approaches

Including target-side context

Despite some have considered including past target-side context harmful because of the *error propagation* problem [Zhang et al., 2018], most recent works have showed it to be of utmost importance for making the most out of context. Past works have successfully included target-side context information in different ways:

- Translating past sentences (usually 1) along with the current one, and then discarding them, as in concatenation approaches [Bawden et al., 2018].
- By making the decoder attend the target-side hidden representations or embeddings of previously decoded sentences [Miculicich et al., 2018, Voita et al., 2019b, Maruf et al., 2019, Zheng et al., 2020].

Multi-Encoder

Reference	Context	Two-Pass Approach	Outside Integr.	Inside Integr.	Lang. Pair
[Wang et al., 2017]	s:-3		aut...	...aut	Zh→En
[Voita et al., 2018]	s:-1		yes		En→Ru
[Zhang et al., 2018]	s:-2		yes	sequential	Zh→En
[Miculicich et al., 2018]	s:-3; t:-3		yes		Zh/Es→En
[Maruf et al., 2019]	s:all; t:all	optional	yes		En→De
[Zheng et al., 2020]	s:all; t:all	yes	yes		Zh/En→En/De
[Jean et al., 2017]	s:-1			parallel	En→De/Fr
[Bawden et al., 2018]	s:-1; t:-1			parallel	En→Fr
[Fu et al., 2019]	s:all	yes		parallel	En/Zh→De/En
[Tan et al., 2019]	s:all	yes		parallel	Zh/De→En
[Voita et al., 2019b]	s:-3; t:-3	yes		parallel*	En→Ru

Caches

- [Maruf and Haffari, 2018] **Two-pass approach (not training! to explain...). All context (source and target). Fr/De/Et→En. Decoder without attention (integration in the RNN gates).** First pass for sentence representations (filling the memories), second pass for integrating current sentence representation with information stocked into memories (via coarse attention). extention: attention to target memory!
- [Zheng et al., 2018] borrow the decoder architecture from the TransformerXL [Dai et al., 2019]. This is very similar to the Transformer's decoder, but every time it translates a sentence it keeps its hidden representations into memory so that it can attend to them while decoding the following sentence. This mechanism creates a recurrence that enables past target-side information to be remembered.

Others

discourse-related information

- [Ohtani et al., 2019] concatenates multiple inputs (past and future), up to -7+3, Bi-LSTM, En-¿Jap. Add coreference information to the input and modify lstm to merge into its hidden state the antecedent or descendant's hidden states.
- [Stojanovski and Fraser, 2018] see survey
- [Rios Gonzales et al., 2017] see survey

Others

using monolingual corpora

- [Martnez Garcia et al., 2019] generate translation by integrating the scores output by Semantic Space Language Model and a NMT model, following the shallow fusion strategy [Gulcehre et al., 2015].
- [Sugiyama and Yoshinaga, 2019] augment dl parallel data by backtranslating BookCorpus (En-¿Ja). Then finds that the synthetic dl parallel corpus improves training of a DLNMT Transformer with concatenation of the inputs. Parallel corpus: IWSLT2017, Ja/Fr-¿En. Improvements are visible in both terms of BLUE but also discourse phenomena.
- [Jean and Cho, 2019] inputing context sentences with different strategies.
- [Li et al., 2019] pre-trained BERT for initializing the encoder
- [Voita et al., 2019a] Automatic Post Editing from sentence level to context level.

Others

dlnmt as learning problem

- [Jean and Cho, 2019]

Positional embedding schemas

1. positional encoding [Vaswani et al., 2017]
2. sentence distance embedding [Voita et al., 2019b]
3. inverse embedding and shit [Maruf et al., 2018]
4. segment embedding [Zheng et al., 2020]

Notes

1. two-step training what is. E.g. [Zhang et al., 2018, Miculicich et al., 2018]. Explain that DL training corpus is small! possible future direction...
2. using reference translations as target context during training have been proven useful but only to a certain extent. Better to mix reference with candidate outputs [Voita et al., 2019b]

Remarks and conclusions

Possible Future Research Directions

- build a large DL corpus for training systems;
- design models exploiting full context;
- design models performing good for single-sentence translation [Zheng et al., 2020]
- design post-processing models that are lightweight and can be trained on little data [Kim et al., 2019]. Nonetheless, beware that While this kind of approach is easy to deploy, the two-stage generation process may result in error accumulation.

Thank you for your attention!

References I

Agrawal, R. R., Turchi, M., and Negri, M. (2018).
Contextual Handling in Neural Machine Translation: Look Behind, Ahead and on Both Sides.
pages 11–20.
00007 Accepted: 2018-08-08T15:15:28Z.

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018).
Evaluating Discourse Phenomena in Neural Machine Translation.
In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.
00055.

References II

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., and Salakhutdinov, R. (2019).
Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context.
arXiv:1901.02860 [cs, stat].
00238 arXiv: 1901.02860.

Fu, H., Liu, C., and Sun, J. (2019).
Reference Network for Neural Machine Translation.
In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3002–3012, Florence, Italy. Association for Computational Linguistics.
00000.

References III

Gulcehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H.-C., Bougares, F., Schwenk, H., and Bengio, Y. (2015).
On Using Monolingual Corpora in Neural Machine Translation.
arXiv:1503.03535 [cs].
00207 arXiv: 1503.03535.

Jean, S. and Cho, K. (2019).
Context-Aware Learning for Neural Machine Translation.
arXiv:1903.04715 [cs].
00003 arXiv: 1903.04715.

Jean, S., Lauly, S., Firat, O., and Cho, K. (2017).
Does Neural Machine Translation Benefit from Larger Context?
arXiv:1704.05135 [cs, stat].
00038 arXiv: 1704.05135.

References IV

Kim, Y., Tran, D. T., and Ney, H. (2019).
When and Why is Document-level Context Useful in Neural Machine Translation?
arXiv:1910.00294 [cs].
00001 arXiv: 1910.00294.

Li, L., Jiang, X., and Liu, Q. (2019).
Pretrained Language Models for Document-Level Neural Machine Translation.
arXiv:1911.03110 [cs].
00001 arXiv: 1911.03110.

Martnez Garcia, E., Creus, C., and Espaa-Bonet, C. (2019).
Context-Aware Neural Machine Translation Decoding.
In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 13–23, Hong Kong, China. Association for Computational Linguistics.
00000.

References V

Maruf, S. and Haffari, G. (2018).
Document Context Neural Machine Translation with Memory Networks.
In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia. Association for Computational Linguistics.
00032.

Maruf, S., Martins, A. F. T., and Haffari, G. (2018).
Contextual Neural Model for Translating Bilingual Multi-Speaker Conversations.
In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 101–112, Brussels, Belgium. Association for Computational Linguistics.
00002.

References VI

Maruf, S., Martins, A. F. T., and Haffari, G. (2019).
Selective Attention for Context-aware Neural Machine Translation.
arXiv:1903.08788 [cs].
00012.

Miculicich, L., Ram, D., Pappas, N., and Henderson, J. (2018).
Document-Level Neural Machine Translation with Hierarchical Attention Networks.
arXiv:1809.01576 [cs].
00024 arXiv: 1809.01576.

Ohtani, T., Kamigaito, H., Nagata, M., and Okumura, M. (2019).
Context-aware Neural Machine Translation with Coreference Information.
In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 45–50, Hong Kong, China. Association for Computational Linguistics.
00000.

Rios Gonzales, A., Mascarell, L., and Sennrich, R. (2017).
Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings.
In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.
00030.

References VIII

Scherrer, Y., Tiedemann, J., and Loiciga, S. (2019).
Analysing concatenation approaches to document-level NMT in two different domains.
In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 51–61, Hong Kong, China. Association for Computational Linguistics.
00001.

Stojanovski, D. and Fraser, A. (2018).
Coreference and Coherence in Neural Machine Translation: A Study Using Oracle Experiments.
In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 49–60, Brussels, Belgium. Association for Computational Linguistics.
00003.

References IX

Sugiyama, A. and Yoshinaga, N. (2019).
Data augmentation using back-translation for context-aware neural machine translation.
In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44, Hong Kong, China. Association for Computational Linguistics.
00000.

Tan, X., Zhang, L., Xiong, D., and Zhou, G. (2019).
Hierarchical Modeling of Global Context for Document-Level Neural Machine Translation.
In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1576–1585, Hong Kong, China. Association for Computational Linguistics.

References X

00002.

Tiedemann, J. and Scherrer, Y. (2017).
Neural Machine Translation with Extended Context.
In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.
00038.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017).
Attention Is All You Need.
arXiv:1706.03762 [cs].
05728 arXiv: 1706.03762.

References XI

Voita, E., Sennrich, R., and Titov, I. (2019a).
Context-Aware Monolingual Repair for Neural Machine Translation.
arXiv:1909.01383 [cs].
00003 arXiv: 1909.01383.

Voita, E., Sennrich, R., and Titov, I. (2019b).
When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion.
In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.
00007.

References XII

Voita, E., Serdyukov, P., Sennrich, R., and Titov, I. (2018).
Context-Aware Neural Machine Translation Learns Anaphora Resolution.
In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.
00047.

Wang, L., Tu, Z., Way, A., and Liu, Q. (2017).
Exploiting Cross-Sentence Context for Neural Machine Translation.
In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics.
00045.

References XIII

Yamagishi, H. and Komachi, M. (2019).
Improving Context-aware Neural Machine Translation with Target-side Context.
arXiv:1909.00531 [cs].
00001 arXiv: 1909.00531.

Zhang, J., Luan, H., Sun, M., Zhai, F., Xu, J., Zhang, M., and Liu, Y. (2018).
Improving the Transformer Translation Model with Document-Level Context.
In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.
00028.

References XIV

Zheng, Z., Huang, S., Sun, Z., Weng, R., Dai, X.-Y., and Chen, J. (2018).
Learning to Discriminate Noises for Incorporating External Information in Neural Machine Translation.
arXiv:1810.10317 [cs].
00003 arXiv: 1810.10317.

Zheng, Z., Yue, X., Huang, S., Chen, J., and Birch, A. (2020).
Toward Making the Most of Context in Neural Machine Translation.
arXiv:2002.07982 [cs].
00000 arXiv: 2002.07982.

Markov Decision Processes

Reinforcement Learning

General class of algorithms that allow an agent to learn how to behave in a stochastic and possibly unknown environment by trial-and-error.

Markov Decision Process (MDP)

stochastic dynamical system specified by $< \mathbb{S}, \mathbb{A}, \mathcal{P}, \mathcal{R}, \gamma >$

1. $(\mathbb{S}, \mathcal{S})$ is a measurable state space
2. $(\mathbb{A}, \mathcal{A})$ is a measurable action space
3. $\mathcal{P} : \mathbb{S} \times \mathbb{A} \times \mathcal{S} \to \mathbb{R}$ is a Markov transition kernel
4. $\mathcal{R} : \mathbb{S} \times \mathbb{A} \to \mathbb{R}$ is a reward function
5. $0 < \gamma < 1$ is the discount factor.

Monte-Carlo Policy Gradient: Pseudocode

Input: Stochastic policy π_θ, Initial parameters θ_0, learning rate $\{\alpha_k\}$
Output: Approximation of the optimal policy $\pi_{\theta*} \approx \pi_*$

1: **repeat**
2: Sample M trajectories $h^{(m)} = \{(s_t^{(m)}, a_t^{(m)}, r_{t+1}^{(m)})\}_{t=0}^{T^{(m)}}$ under policy π_{θ_k}
3: Approximate policy gradient

$$\nabla_\theta J(\theta_k) \approx \frac{1}{M} \sum_{m=0}^{M} \sum_{u=0}^{T^{(m)}-1} \nabla_\theta \log \pi_{\theta_k} \left(s_u^{(m)}, a_u^{(m)} \right) \sum_{v \geqslant u}^{T^{(m)}-1} \gamma^{v-u} r_{v+1}^{(m)}$$

4: Update parameters using gradient ascent $\theta_{k+1} = \theta_k + \alpha_k \nabla_\theta J(\theta_k)$
5: $k \leftarrow k+1$
6: **until** converged

Episodic PGPE Algorithm: Pseudocode

Input: Controller F_θ, hyper-distribution p_ξ, initial guess ξ_0, learning rate $\{\alpha_k\}$
Output: Approximation of the optimal policy $F_{\xi*} \approx \pi_*$

repeat
  for $m = 1, \ldots, M$ **do**
    Sample controller parameters $\theta^{(m)} \sim p_{\xi_k}$
    Sample trajectory $h^{(m)} = \{(s_t^{(m)}, a_t^{(m)}, r_{t+1}^{(m)})\}_{t=0}^{T^{(m)}}$ under policy $F_{\theta^{(m)}}$
  end for
  Approximate policy gradient

$$\nabla_\xi J(\xi_k) \approx \frac{1}{M} \sum_{m=1}^{M} \nabla_\xi \log p_\xi \left(\theta^{(m)}\right) \left[G\left(h^{(m)}\right) - b\right]$$

  Update hyperparameters using gradient ascent $\xi_{k+1} = \xi_k + \alpha_k \nabla_\xi J(\xi_k)$
  $k \leftarrow k + 1$
until converged

Truncated Multiple Importance Sampling Estimator

Importance Sampling

Given a bounded function $f : \mathcal{Z} \to \mathbb{R}$, and a set of i.i.d. outcomes $z_1, \ldots, z_N$ sampled from Q, the importance sampling estimator of $\mu := \underset{z \sim P}{\mathbb{E}}[f(z)]$ is:

$$\widehat{\mu}_{\text{IS}} = \frac{1}{N} \sum_{i=1}^{N} f(z_i) w_{P/Q}(z_i), \tag{1}$$

which is an unbiased estimator, i.e., $\underset{z_i \overset{\text{iid}}{\sim} Q}{\mathbb{E}} [\widehat{\mu}_{IS}] = \mu$.

Truncated Estimator With Balance Heuristic

$$\check{\mu}_{\text{BH}} = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N_k} \min \left\{ M, \frac{p(z_{ik})}{\sum_{j=1}^{K} \frac{N_j}{N} q_j(z_{ik})} \right\} f(z_{ik}). \tag{2}$$

OPTIMIST2

Theorem

regretdiscretized Let $\mathcal{X}$ be a d-dimensional compact arm set with $\mathcal{X} \subseteq [-D, D]^d$. For any $\kappa \geqslant 2$, under Assumptions 1 and 2, OPTIMIST2 with confidence schedule $\delta_t = \dfrac{6\delta}{\pi^2 t^2 \left(1 + \lceil t^{1/\kappa} \rceil^d\right)}$ and discretization schedule $\tau_t = \lceil t^{\frac{1}{\kappa}} \rceil$ guarantees, with probability at least $1 - \delta$:

$$
\begin{aligned}
Regret(T) \leqslant \Delta_0 + C_1 T^{\left(1-\frac{1}{\kappa}\right)} d + C_2 T^{\frac{1}{1+\epsilon}} \\
\cdot \left[v_\epsilon \left((2 + {}^{d}\!/\!_{\kappa}) \log T + d \log 2 + \log \frac{\pi^2}{3\delta} \right) \right]^{\frac{\epsilon}{1+\epsilon}},
\end{aligned}
$$

where $C_1 = \dfrac{\kappa}{\kappa - 1} LD$, $C_2 = (1 + \epsilon) \left(2\sqrt{2} + \dfrac{5}{3} \right) \|f\|_\infty$, and Δ_0 is the instantaneous regret of the initial arm $\boldsymbol{x}_0$.