Some Extensions of Neural Machine Translation for Auto-formalization of Mathematics

Qingxiang Wang, Cezary Kaliszyk, Josef Urban

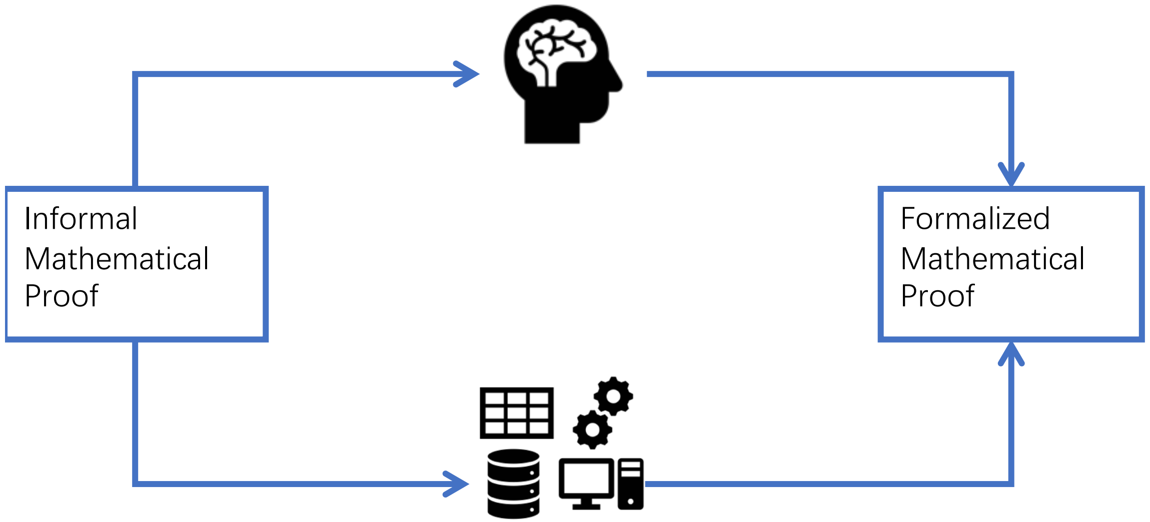
AITP 2019 – Obergurgl, Austria

April 11, 2019

Overview

- Auto-Formalization with Deep Learning
- Universal Approximation
- Supervised NMT (Luong et al.)
- Unsupervised NMT (Lample et al.)
- NMT with Type Elaboration
- Summary

Auto-Formalization with Deep Learning



Universal Approximation

Theorem 2. *Let σ be any continuous sigmoidal function. Then finite sums of the form*

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j)$$

are dense in $C(I_n)$. In other words, given any $f \in C(I_n)$ and $\varepsilon > 0$, there is a sum, $G(x)$, of the above form, for which

$$|G(x) - f(x)| < \varepsilon \qquad \text{for all} \quad x \in I_n.$$

G. Cybenko 89 - Approximation by Superpositions of a Sigmoidal Function

Supervised NMT (Luong et al.)

- Default: two-layer LSTM with attention.
- Lots of configurable hyper-parameters:

  (Attention, Layers, Unit Size, Unit Type, Residual, Encoding, Optimizers, etc)

- Formal abstracts of *Formalized mathematics*, which are generated latex from Mizar (v8.0.01_5.6.1169)
- 1,056,478 pairs of Latex– Mizar sentences in 90:10.

```
theorem P0:
for L being right_zeroed non empty addLoopStr,
    S,T being Subset of L
st 0.L in T holds S c= S + T
```

Formalized Mathematics ↓ ↑ Seq2Seq

Now we state the propositions:

(1) Let us consider a right zeroed, non empty additive loop structure L, and subsets S, T of L. If $0_L \in T$, then $S \subseteq S + T$.

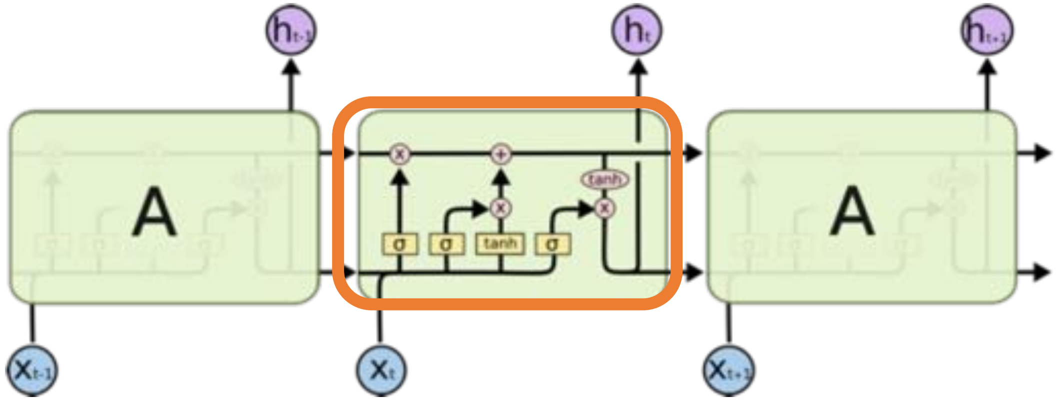
Supervised NMT (Luong et al.)

Latex	`If $ X \mathrel { = } { \rm the ~ } { { { \rm carrier } ~ { \rm of } ~ { \rm } } } { A _ { 9 } } $ and $ X $ is plane , then $ { A _ { 9 } } $ is an affine plane .`
Mizar	`X = the carrier of AS & X is being_plane implies AS is AffinPlane ;`

Latex	`If $ { s _ { 9 } } $ is convergent and $ { s _ { 8 } } $ is a subsequence of $ { s _ { 9 } } $ , then $ { s _ { 8 } } $ is convergent .`
Mizar	`seq is convergent & seq1 is subsequence of seq implies seq1 is convergent ;`

Supervised NMT (Luong et al.)

- Memory-cell unit types

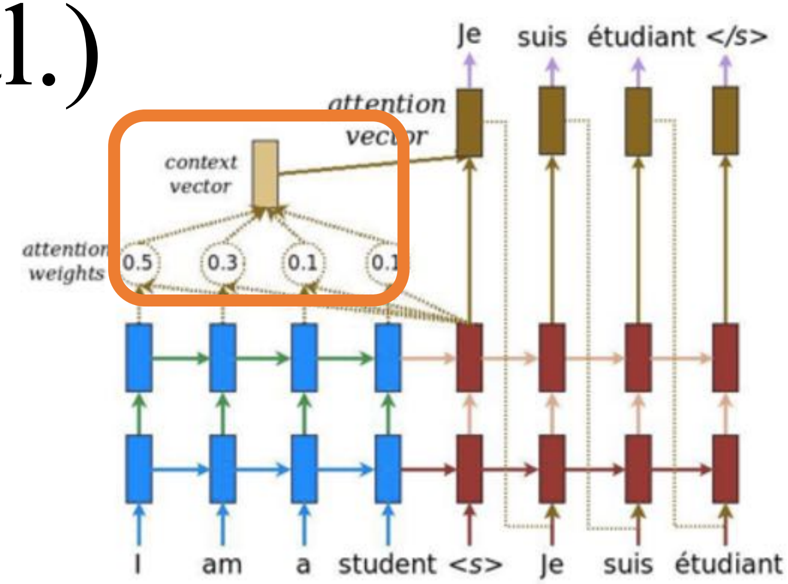


Parameter	Final Test Perplexity	Final Test BLEU	Identical Statements (%)	Identical No-overlap (%)
LSTM	**3.06**	**41.1**	**40121 (38.12%)**	**6458 (13.43%)**
GRU	3.39	34.7	37758 (35.88%)	5566 (11.57%)
Layer-norm LSTM	11.35	0.4	11200 (10.64%)	1 (0%)

Table 5. Evaluation on type of memory cell (attention not enabled)

Supervised NMT (Luong et al.)

- Attention

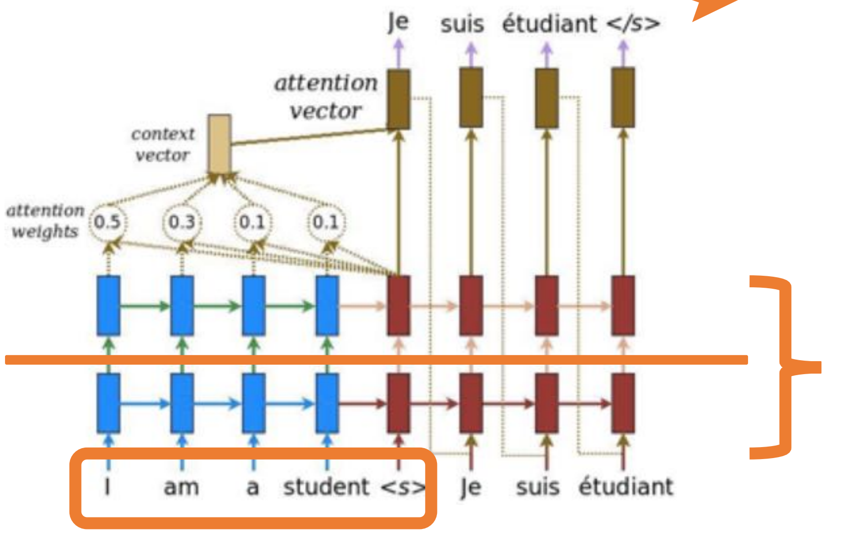


Parameter	Final Test Perplexity	Final Test BLEU	Identical Statements (%)	Identical No-overlap (%)
No Attention	3.06	41.1	40121 (38.12%)	6458 (13.43%)
Bahdanau	3	40.9	44218 (42.01%)	8440 (17.55%)
Normed Bahdanau	1.92	63.5	60192 (57.19%)	18057 (37.54%)
Luong	**1.89**	64.8	60151 (57.15%)	18013 (37.45%)
Scaled Luong	2.13	**65**	**60703 (57.68%)**	**18105 (37.64%)**

Table 6. Evaluation on type of attention mechanism (LSTM cell)

Supervised NMT

- Residuals, layers, etc.

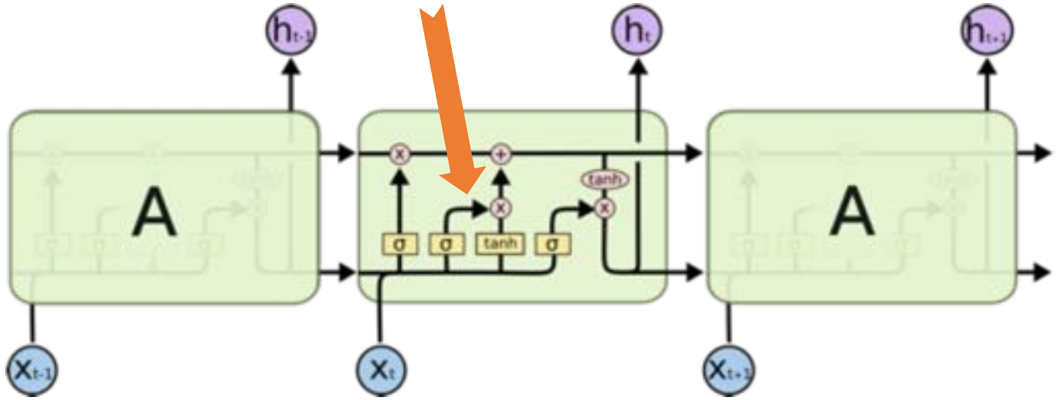


Parameter	Final Test Perplexity	Final Test BLEU	Identical Statements (%)	Identical No-overlap (%)
2-Layer	3.06	41.1	40121 (38.12%)	6458 (13.43%)
3-Layer	2.10	64.2	57413 (54.55%	16318 (33.92%)
4-Layer	2.39	45.2	49548 (47.08%)	11939 (24.82%)
5-Layer	5.92	12.8	29207 (27.75%)	2698 (5.61%)
6-Layer	4.96	20.5	29361 (27.9%)	2872 (5.97%)
2-Layer Residual	1.92	54.2	57843 (54.96%)	16511 (34.32%)
3-Layer Residual	1.94	62.6	59204 (56.25%)	17396 (36.16%)
4-Layer Residual	1.85	56.1	59773 (56.79%)	17626 (36.64%)
5-Layer Residual	2.01	63.1	59259 (56.30%)	17327 (36.02%)
6-Layer Residual	NaN	0	0 (0%)	0 (0%)
2-Layer Adam	1.78	56.6	61524 (58.46%)	18635 (38.74%)
3-Layer Adam	1.91	60.8	59005 (56.06%)	17213 (35.78%)
4-Layer Adam	1.99	51.8	57479 (54.61%)	16288 (33.86%)
5-Layer Adam	2.16	54.3	54670 (51.94%)	14769 (30.70%)
6-Layer Adam	2.82	37.4	46555 (44.23%)	10196 (21.20%)
2-Layer Adam Res.	1.75	56.1	63242 (60.09%)	19716 (40.97%)
3-Layer Adam Res.	1.70	55.4	64512 (61.30%)	20534 (42.69%)
4-Layer Adam Res.	1.68	57.8	64399 (61.19%)	20353 (42.31%)
5-Layer Adam Res.	1.65	64.3	64722 (61.50%)	20627 (42.88%)
6-Layer Adam Res.	1.66	59.7	65143 (61.90%)	20854 (43.35%)
2-Layer Bidirectional	2.39	**69.5**	63075 (59.93%)	19553 (40.65%)
4-Layer Bidirectional	6.03	63.4	58603 (55.68%)	17222 (35.80%)
6-Layer Bidirectional	2	56.3	57896 (55.01%)	16817 (34.96%)
2-Layer Adam Bi.	1.84	56.9	64918 (61.68%)	20830 (43.30%)
4-Layer Adam Bi.	1.94	58.4	64054 (60.86%)	20310 (42.22%)
6-Layer Adam Bi.	2.15	55.4	60616 (57.59%)	18196 (37.83%)
2-Layer Bi. Res.	2.38	24.1	47531 (45.16%)	11282 (23.45%)
4-Layer Bi. Res.	NaN	0	0 (0%)	0 (0%)
6-Layer Bi. Res.	NaN	0	0 (0%)	0 (0%)
2-Layer Adam Bi. Res.	1.67	62.2	65944 (62.66%)	21342 (44.37%)
4-Layer Adam Bi. Res.	**1.62**	66.5	65992 (62.70%)	21366 (44.42%)
6-Layer Adam Bi. Res.	1.63	58.3	**66237 (62.93%)**	**21404 (44.50%)**

Table 7. Evaluation on various hyperparameters w.r.t. layers

Supervised NMT (Luong et al.)

- Unit dimension in cell



Parameter	Final Test Perplexity	Final Test BLEU	Identical Statements (%)	Identical No-overlap (%)	Training Time (hrs.)
128 Units	3.06	41.1	40121 (38.12%)	6458 (13.43%)	1
256 Units	1.59	64.2	63433 (60.27%)	19685 (40.92%)	3
512 Units	1.6	**67.9**	66361 (63.05%)	21506 (44.71%)	5
1024 Units	**1.51**	61.6	**69179 (65.73%)**	**22978 (47.77%)**	11
2048 Units	2.02	60	59637 (56.66%)	16284 (33.85%)	31

Table 8. Evaluation on number of units

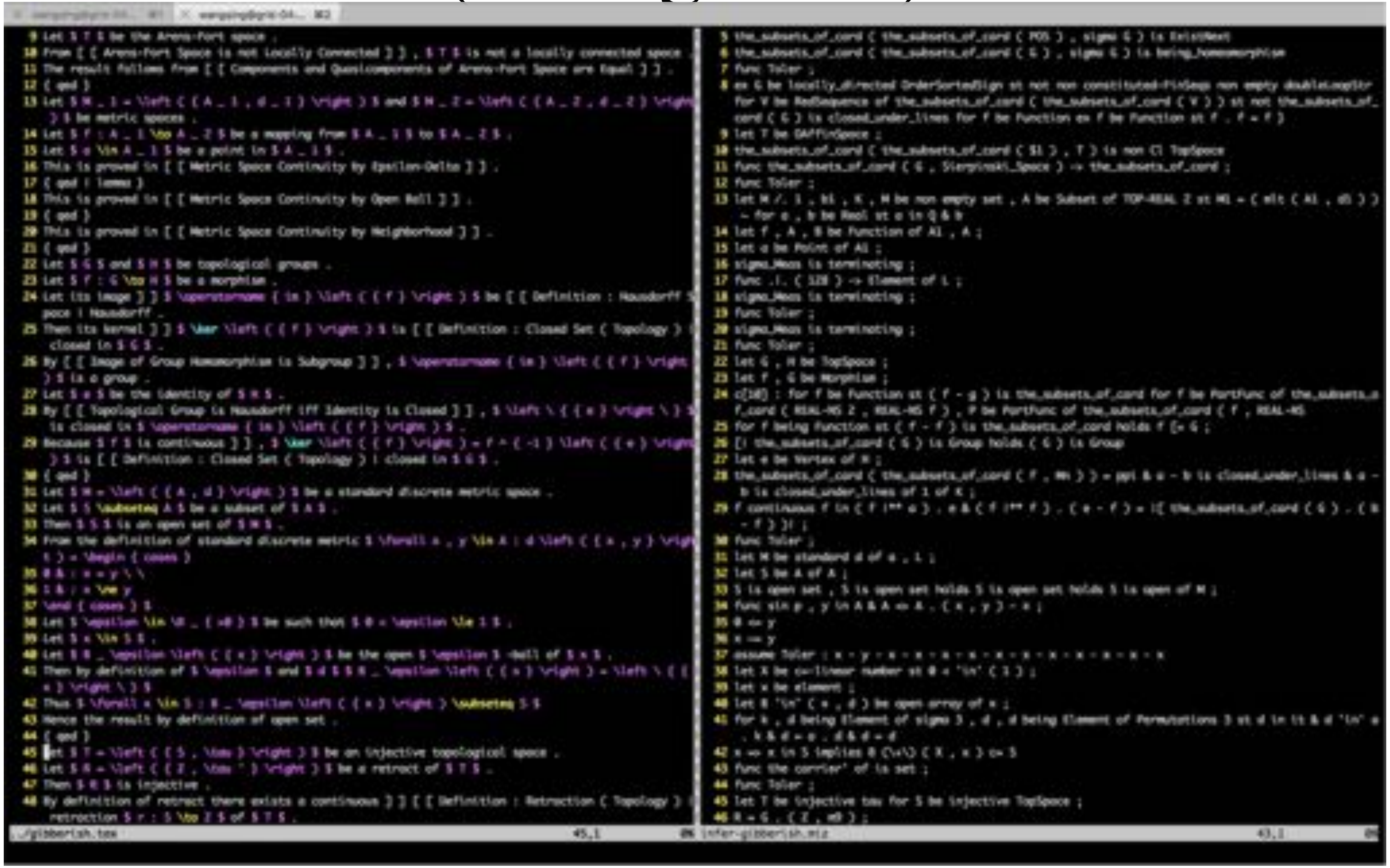
Supervised NMT (Luong et al.)

	Identical Statements	0	≤ 1	≤ 2	≤ 3
Best Model - 1024 Units	69179 (total)	65.73%	74.58%	86.07%	88.73%
	22978 (no-overlap)	47.77%	59.91%	70.26%	74.33%
Top-5 Greedy Cover - 1024 Units - 4-Layer Bi. Res. - 512 Units - 6-Layer Adam Bi. Res. - 2048 Units	78411 (total)	74.50%	82.07%	87.27%	89.06%
	28708 (no-overlap)	59.68%	70.85%	78.84%	81.76%
Top-10 Greedy Cover - 1024 Units - 4-Layer Bi. Res. - 512 Units - 6-Layer Adam Bi. Res. - 2048 Units - 2-Layer Adam Bi. Res. - 256 Units - 5-Layer Adam Res. - 6-Layer Adam Res. - 2-Layer Bi. Res.	80922 (total)	76.89%	83.91%	88.60%	90.24%
	30426 (no-overlap)	63.25%	73.74%	81.07%	83.68%
Union of All 39 Models	83321 (total)	79.17%	85.57%	89.73%	91.25%
	32083 (no-overlap)	66.70%	76.39%	82.88%	85.30%

Table 9. Coverage w.r.t. a set of models and edit distances

- But generates gibberish when we tried arbitrary LaTeX statements on the trained model...☹

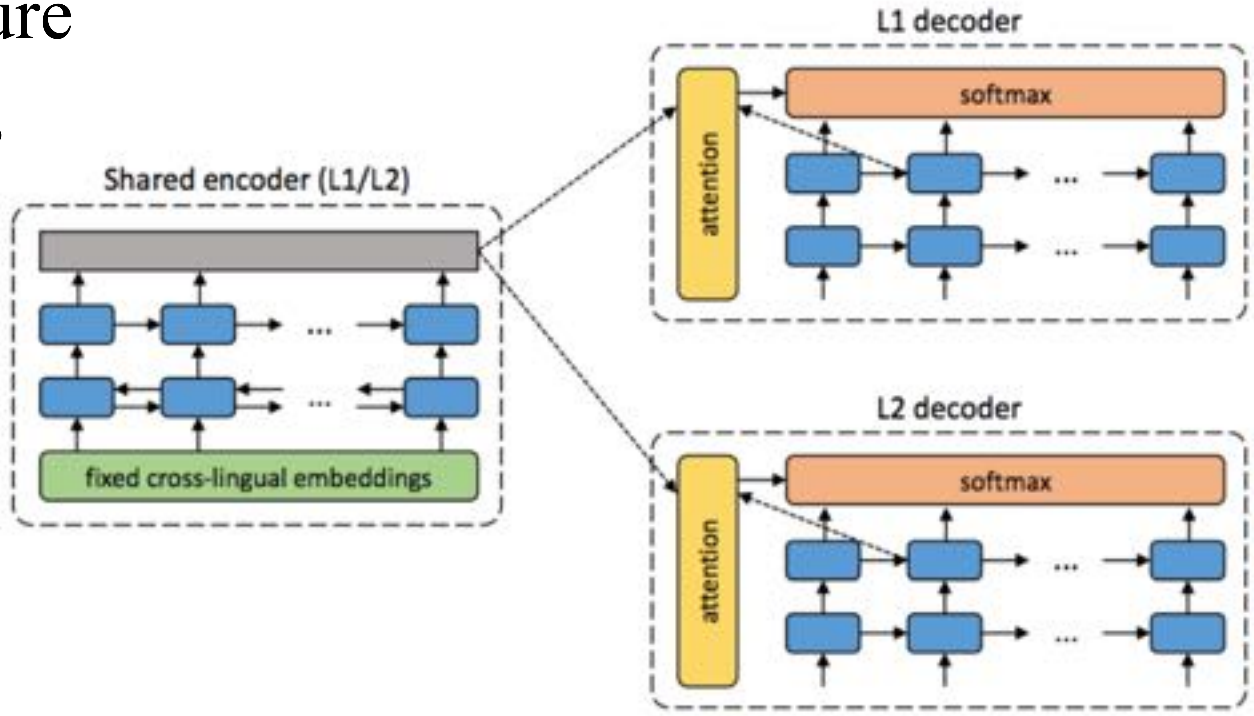
Supervised NMT (Luong et al.)



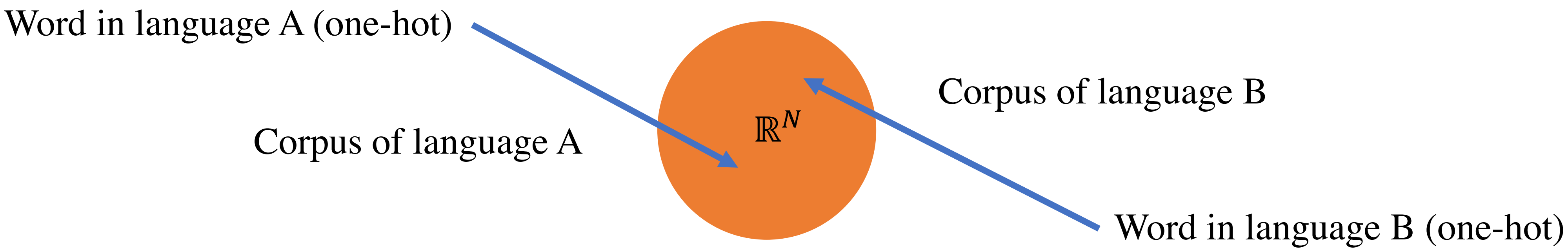
- Demo

Unsupervised NMT (Lample et al.)

- Two monolingual corpora instead of one parallel corpora (ProofWiki - Mizar)
- Shared-encoder NMT architecture
- Fixed cross-lingual embeddings
  - Word2Vec
  - BPE (Byte Pair Encoding)
- Denoising and backtranslation



Unsupervised NMT (Lample et al.)



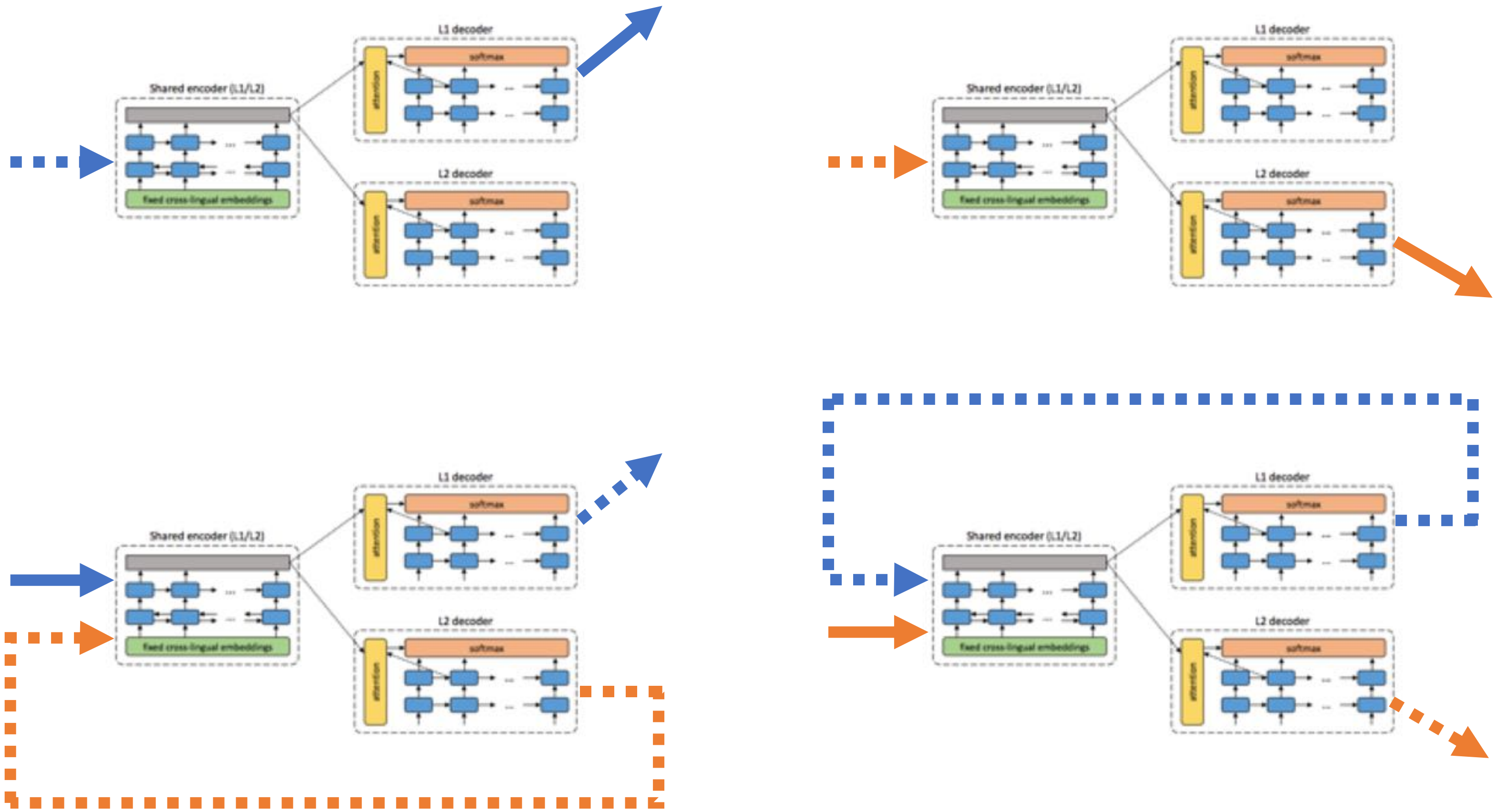
Word2Vec

3 BPE iterations on a corpus with the word "Lower"

BPE

{"L", "o", "w", "e", "r"} → {"L", "o", "w", "er"} → {"L", "ow", "er"} → {"Low", "er"}

Unsupervised NMT (Lample et al.)

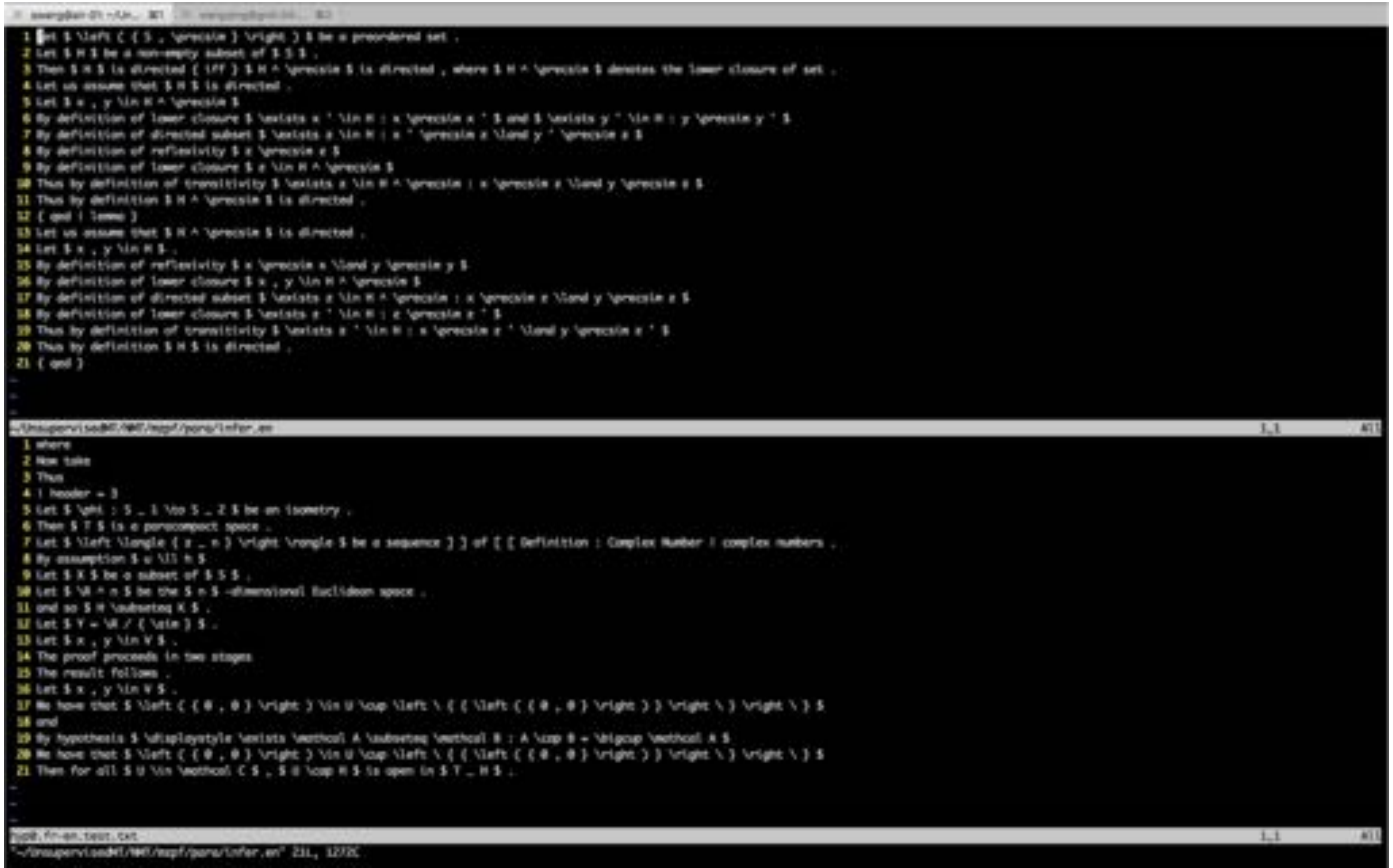


Denoising

Back Translation

- Generating gibberish on our data…☹

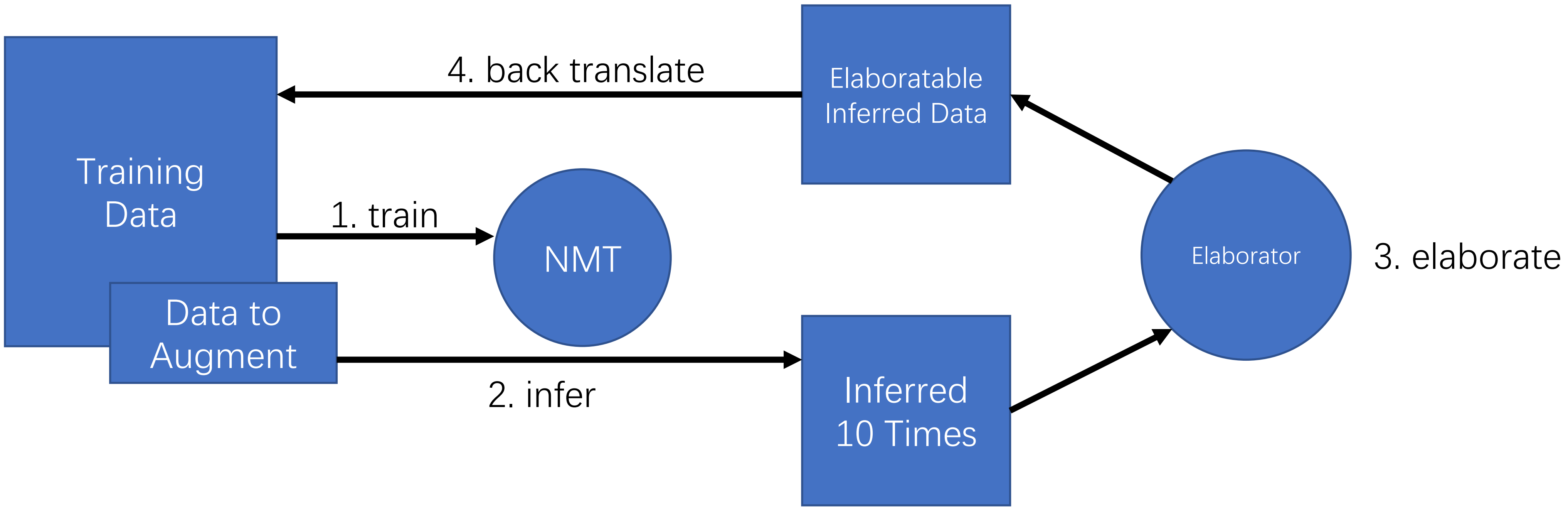
Unsupervised NMT (Lample et al.)



- Demo

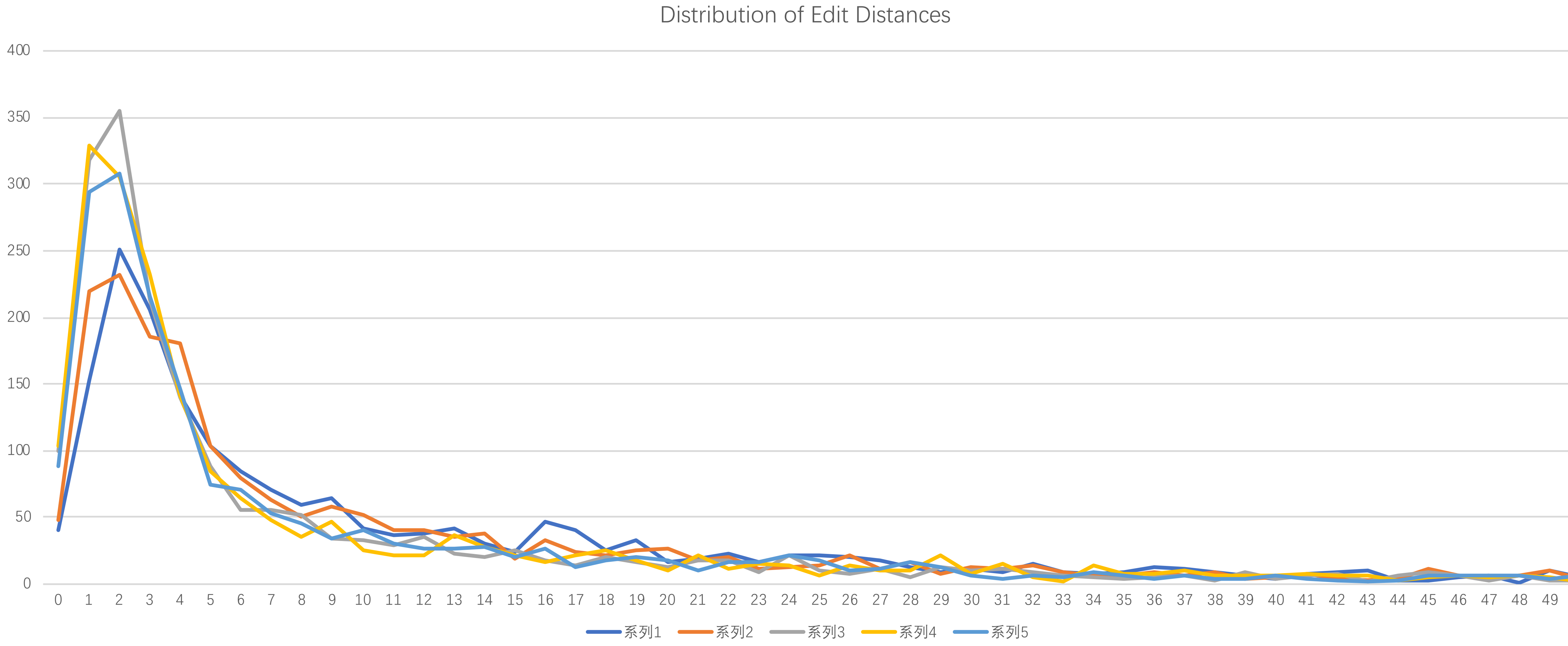
NMT with Type Elaboration

- Still Luong's NMT, but with Mizar -> TPTP (prefix format) as data.
- Augment our data through type elaboration and iterative training.



- Performance stabilizes after a few iterations…☹

NMT with Type Elaboration



Summary

- For auto-formalization, we hit a wall with NMT techniques with limited data.
- Focus on obtaining high-quality data.
- This is still a direction worth going as manual translation is too costly.

Thanks

All historical orientation is only living when we learn to see what is ultimately essential is due to our own interpreting in the free rethinking by which we gain detachment from all erudition.

Martin Heidegger – The Metaphysical Foundations of Logic