Neural Machine Translation Improvements Using Back-Translation, and Applications in CNN-based Protein-Ligand Docking Evaluation

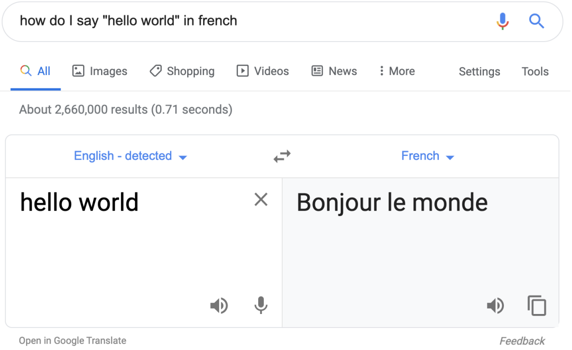
Nathan Wood, TECBIO 2020

July 15, 2020

Table of Contents

Machine Translation

- Computationally intensive means of bridging and interconverting language pairs



Types of Machine Translation [3]

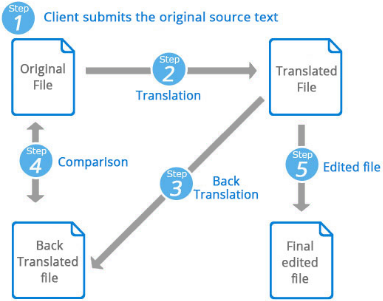
- Traditional Rules-Based Machine Translation:
    - Language organized into a corpus, or collections of known words or statements
    - Run through patterns, grammar, and lexicons
    - Reorganize against syntactic structures (ADJ-N-V-ADV)
- Statistical Machine Translation
    - No lexical or grammatical foreknowledge
    - References previous translations from database
- Neural Machine Translation
    - Data is processed into multiple layers iteratively using parallel processing and hardware acceleration
    - Algorithm use allows linguistic rules to be developed from previous models
    - Structured on encoding/decoding associated with human sensory processing

Essential Question

What if Language Data is Sparse?

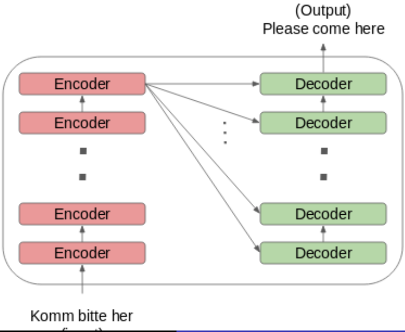
Back Translation

- Training a "Target to Source" pathway
- Monolingual language sets bridged together with synthetic data



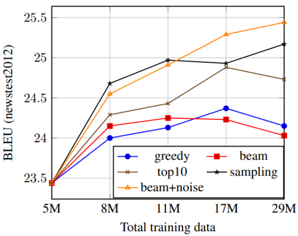
Edunov, Ott, Auli, and Grangier (2018) [1]

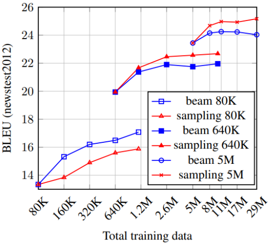
- WMT-2018 English-German Competitive Set w/o 250 word sentences or longer (226M sentences)
- Training Set: 52K sentence pairs
- 6 encode/decode blocks, 4096 feed-foward layers ("Big Transformer")



Edunov, et.al. Results (2018)

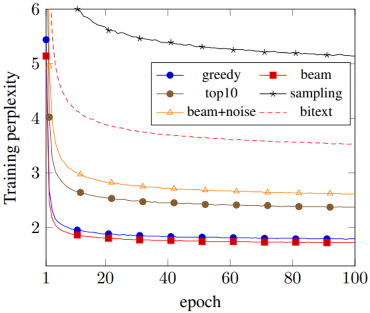
- BLEU(Bilingual Evaluation Understudy)
  - Individually translated segments are compared to a qualified reference
  - Individual scores are averaged against whole corpus
- Bitext- The alignment of relevant words and patterns
- Adding synthetic data (beam+noise and sampling) perform best





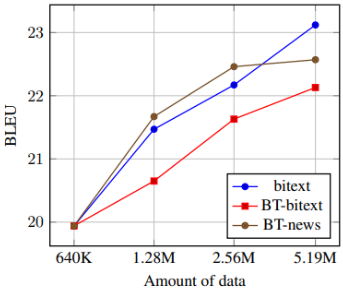
Edunov, et.al. Results (2018)

- Perplexity - how well a model predicts a sample
  - Greedy and Beam do not introduce synthetic data

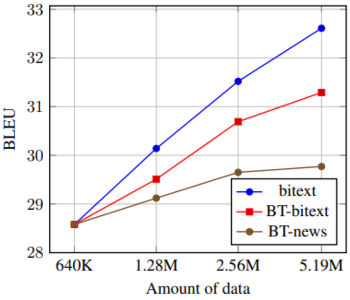


Edunov, et.al. Examining Synthetic Data

- subsample with: remaining data, aligned back-translated bitext, and raw-backtranslated data
- Back-Translated data performs almost as well as the bitext in comparison to raw newstest2012 set and a hybrid set



(a) newstest2012



(b) valid-mixed

Edunov, et.al Conclusions

- Back Translation is effective at augmenting datasets with respect to machine translation
- Synthetic data mostly increased performance
- Future work entails optimized selection of helpful synthetic data

Hoang, Haffari, Koehn, and Cohn

- Iterative Back Translation- feeding back translation back into the model
- Hypothesis: Better Back Translation = Better Synthetic Corpus = Better Translation Quality
- German-English and English-German Scenarios

Hoang, et.al Back Translation [2]

- Worst: 10k iterations
- Best: Convergence

German–English	Back	Final
no back-translation	-	29.6
10k iterations	10.6	29.6 (+0.0)
100k iterations	21.0	31.1 (+1.5)
convergence	23.7	32.5 (+2.9)

English–German	Back	Final
no back-translation	-	23.7
10k iterations	14.5	23.7 (+0.0)
100k iterations	26.2	25.2 (+1.5)
convergence	29.1	25.9 (+2.2)

Hoang, et.al Iterative "Re-Back" Back Translation

- Cycle through inputted back translations multiple times
- High and low resource conditions with shallow and deep architectures
  - Base - parallel data only, not yet back translated, shallow only
  - First - parallel and synthetic data, deep only
  - Final - shallow, deep, and 4-stage ensemble
  - Low Resources only- English-French, English-Farsi

Hoang, et.al High Resource Iterative BT

- Re-back back translation typically outperformed conventional back translation
- Out-performed best translation models of WMT 2017 Competition

German–English	Back*	Shallow	Deep	Ensemble
back-translation	23.7	32.5	35.0	35.6
re-back-translation	27.9	33.6	36.1	36.5
Best WMT 2017	-	-	-	35.1

English–German	Back*	Shallow	Deep	Ensemble
back-translation	29.1	25.9	28.3	28.8
re-back-translation	34.8	27.0	29.0	29.3
Best WMT 2017	-	-	-	28.3

Hoang, et.al Low Resource Iterative BT

- Moses- a Statistical Machine Translation model set, for comparison purposes
- Slight improvements over conventional back translation

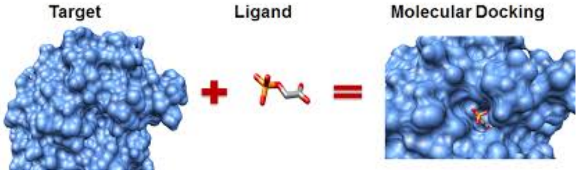
Setting	French–English		English–French		Farsi–English	English-Farsi
	100K	1M	100K	1M	100K	100K
NMT baseline	16.7	24.7	18.0	25.6	21.7	16.4
back-translation	22.1	27.8	21.5	27.0	22.1	16.7
back-translation iterative+1	22.5	-	22.7	-	22.7	17.1
back-translation iterative+2	22.6	-	22.6	-	22.6	17.2
back-translation (w/ Moses)	23.7	27.9	23.5	27.3	21.8	16.8

Hoang, et.al. Conclusions

- Both standard and iterative back-translation is quality dependent, and dependent on sampling
- Iterative back-translation clearly shows improved performance in comparison to standard back translation
- Future considerations will entail developing a unified end-to-end system

How is this Relevant?

- Quantum Mechanical Computations
  - Time and Resource Intensive
- Machine Learning Methods
  - Docking ligands (potential drugs) into a protein and evaluating the pose
  - Pose is evaluated using Convolutional Neural Networks once partitioned into 24Å grid spaces



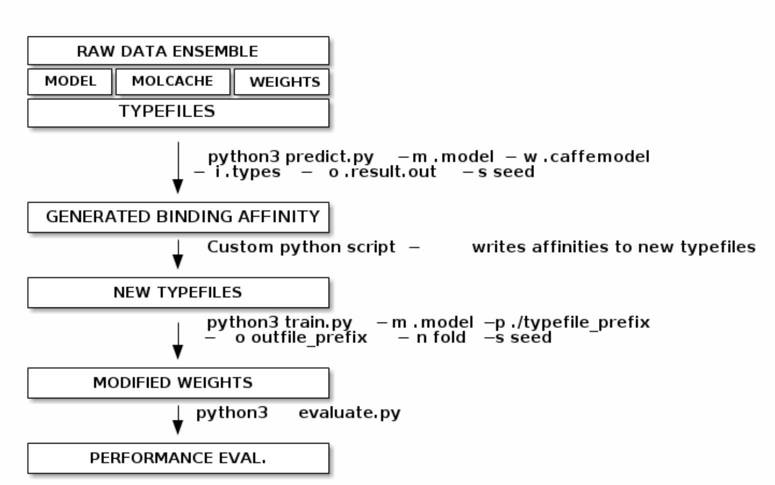
Essential Question

What if Data is Partially Unavailable?

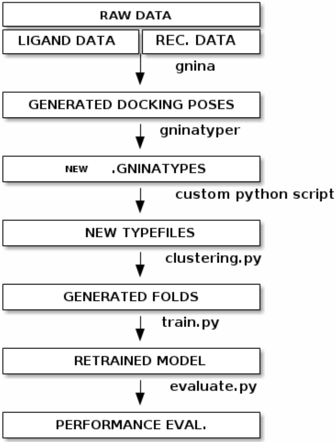
We Can Generate It!

- Project Proposal
  - Evaluate if adding computationally generated data improves model scoring performance
  - Generated Binding Affinities (CrossDock2020 Completeset) and Generated Protein-Ligand docking poses (BindingDB and PDBBind)

How This is Accomplished: Affinities



How This is Accomplished: Poses



References I

EDUNOV, S., OTT, M., AULI, M., AND GRANGIER, D.
Understanding back-translation at scale, 2018.

HOANG, V. C. D., KOEHN, P., HAFFARI, G., AND COHN, T.
Iterative back-translation for neural machine translation.
In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation* (Melbourne, Australia, July 2018), Association for Computational Linguistics, pp. 18–24.

References II

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J.

Google's neural machine translation system: Bridging the gap between human and machine translation.

CoRR abs/1609.08144 (2016).

Questions