Sampling and Filtering of Neural Machine Translation Distillation Data

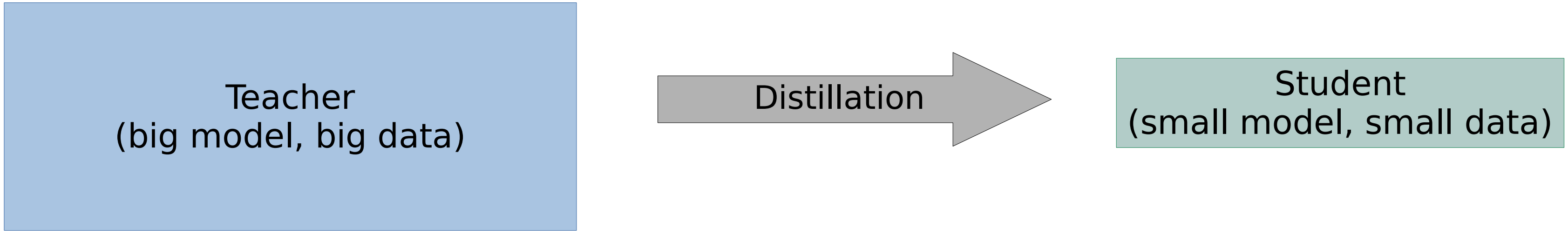
Vilém Zouhar

NAACL-SRW 2021

Motivation

Model {stealing,distillation}

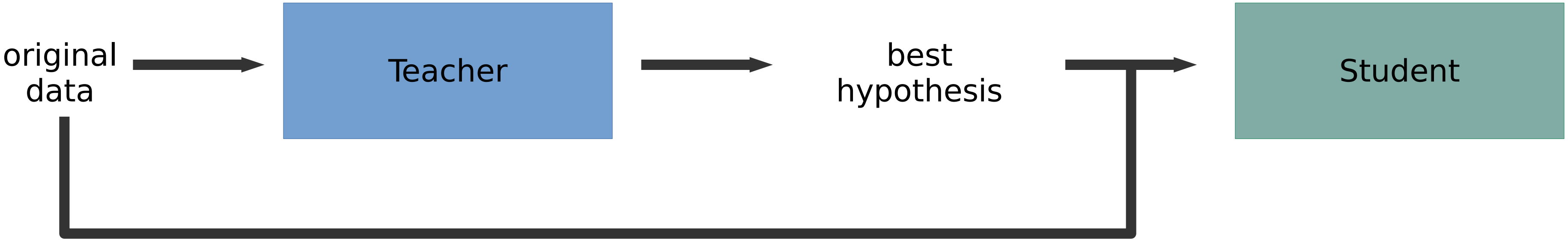
- Transferring knowledge from large model (teacher) to smaller one (student)
- Filtering improves performance (German et. al.)
  - Model inferences have no additional features that imply the quality



Related Work

Most prevalent

- Take all the best decoder hypotheses
- Combine them with original (not teacher train) data
- Output: two translations per one sentence → Twice as much data



Related Work

More sophisticated

- Filter sentences which are suspicious
  - Not recognized by language recognizer (German et. al.)
  - More than / less than k*length of the reference /source
- Filter based on reference quality, e.g. TER (Freitag et. al.)
  - Not main topic of their work

→ TER is very arbitary, how do other metrics behave?

→ What about other sampling methods?

→ What about the combination of them?

Idea

Distillation

- Query teacher on source sentences
- Get 12 translations for every sentence
- Treat them w.r.t. their quality
  - Filter worse ones
  - Oversample better ones
- Quality measured by
  - Decoder score
  - Reference based metrics (BLEU, ChrF, TER, SP)

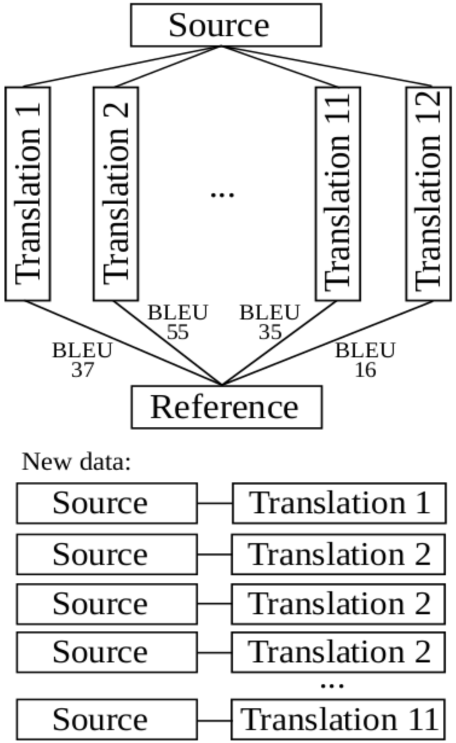


Figure 1. Scheme of an example of hypothesis sampling with BLEU metrics.

Experimental Setup

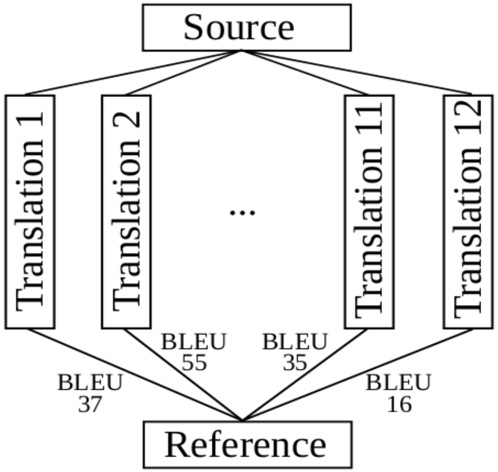
Process

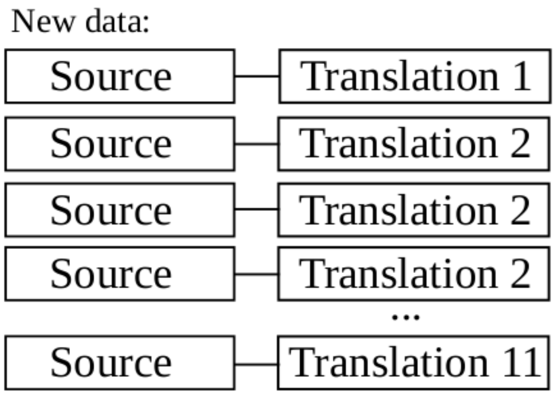
- Create new data set (data sampling)
  - Use different sampling methods
- Train student on the new data
  - CS→EN, EN→CS, EN→DE
- Evaluate and see the effect of the chosen method

Data Sampling

Notation

- $F_{metrics}^{k_1, k_2, k_3, \ldots}$
  - take top translation k_1 times, second one k_2 times, etc. …





second best $k_2 = 1$

top best $k_1 = 3$

.
.
.

Data Sampling

Notation

- $T^{n}_{metrics}$
  - Take *n* top translation hypotheses according to *metrics* (equals $F^{1,1,1,\ldots,1(n)}_{metrics}$)
- $G^{m}_{metrics}$
  - Take all sentence translations with *metrics*
  - At least *m*
- $Dedup[X]$
  - Deduplicate sentence pairs of *X*

Data Sampling

Notation

- Concatenation of sampling methods
- $T_{BLEU}^{2} + G_{score}^{-10}$
  - Join the top *2* sentences measured by *BLEU*
  - Add them to hypotheses with *decoder score* at least *-10*

Results – Combination

Dataset	CS→EN	EN→CS	EN→DE
T^1_{score} + Original	44.4	36.4	28.3
$Dedup[T^4_{\text{BLEU}} + T^4_{\text{score}}]$ + Original	43.7	35.3	29.1
$S^{4,3,2,1}_{\text{score}} + 2 \times$ Original	43.9	36.1	28.3
$S^{4,3,2,1}_{\text{BLEU}} + 2 \times$ Original	**45.5**	**37.3**	28.8
$S^{4,3,2,1}_{\text{BLEU}} + 4 \times$ Original	**45.5** $\star$	**37.4** $\star$	28.9
$T^4_{\text{score}} + T^{12}_{-}$	41.6	33.2	28.3
$T^4_{\text{BLEU}} + T^{12}_{-}$	42.6	33.9	28.7
$T^4_{\text{BLEU}} + T^4_{\text{score}}$	43.3	33.2	28.9
$Dedup[\sum T^2_{\text{metric}}]$	43.6	34.7	29.1
$Dedup[\sum T^2_{\text{metrics}}] + T^{12}_{-}$	40.8	32.0	27.2
$Dedup[T^4_{\text{BLEU}} + T^4_{\text{score}}] + T^1_{\text{BLEU}} + T^1_{\text{score}}$	43.5	34.7	29.2
$Dedup[T^4_{\text{BLEU}} + T^4_{\text{score}}] + Dedup[T^1_{\text{BLEU}} + T^1_{\text{score}}]$	42.6	34.9	**29.6** $\star$
$Dedup[T^4_{\text{BLEU}} + T^4_{\text{score}}]$	43.5	35.0	**29.3**

Table 6: BLEU scores for students trained on datasets made of combination of sampling methods. $\sum_{\text{metric}}$ sums over all used metrics (BLEU, ChrF, TER, SP, score).

Contributions

Meaning of results

- Smaller students
  - Improve results by 0.9 BLEU score
  - Not much, but orthogonal to other tricks

- Students with the same architecture like teacher
  - Improve results by 2.0 BLEU score

Conclusion

Although widely used:

- Taking only highest-scoring sentence & original → not the best results
- Combination of oversampled good hypotheses & original
- Choice of reference metric does not significantly influence results

Caveats:

- Evaluation on custom dataset, not standardized ones
  - Computation power limitations
- Oversampling vectors chosen at random
  - They are an independent variable as well
- Models very small

Future Work

Bigger models:

- Behavior on larger data & models
- Computationally expensive, but focus only on the best methods

Experiment with other ML domains:

- Does distillation oversampling work in other ML domains as well?

More oversampling schemes

- Sampling parameters chosen arbitrarily