Sparse and Constrained Attention for Neural Machine Translation

Chaitanya Malaviya[1], Pedro Ferreira[2], André F.T. Martins[2,3]

[1]Carnegie Mellon University, [2]Instituto Superior Técnico, [3]Unbabel

Adequacy in Neural Machine Translation

Repetitions

Source: und wir benutzen dieses wort mit solcher verachtung

Reference: and we say that word *with such contempt* .

Translation: and we use this word with such contempt contempt .

Dropped words

Source: Ein 28-jähriger Koch, der kürzlich nach Pittsburgh gezogen war, wurde diese Woche im Treppenhaus eines örtlichen Einkaufszentrums tot aufgefunden .

Reference: A 28-year-old chef who recently moved to Pittsburgh was found dead in the staircase *of a local shopping mall* this week .

Translation: A 28-year-old chef who recently moved to Pittsburgh was found dead in the staircase this week .

Previous Work

- Conditioning on coverage vectors to track attention history (Mi, 2016 ; Tu, 2016).

- Gating architectures and adaptive attention to control amount of source context (Tu, 2017; Li & Zhu, 2017).

- Reconstruction Loss (Tu, 2017).

- Coverage penalty during decoding (Wu, 2016).

Main Contributions

J'ai mangé le sandwich

1. Fertility-based Neural Machine Translation Model
   (Bounds on source attention weights)

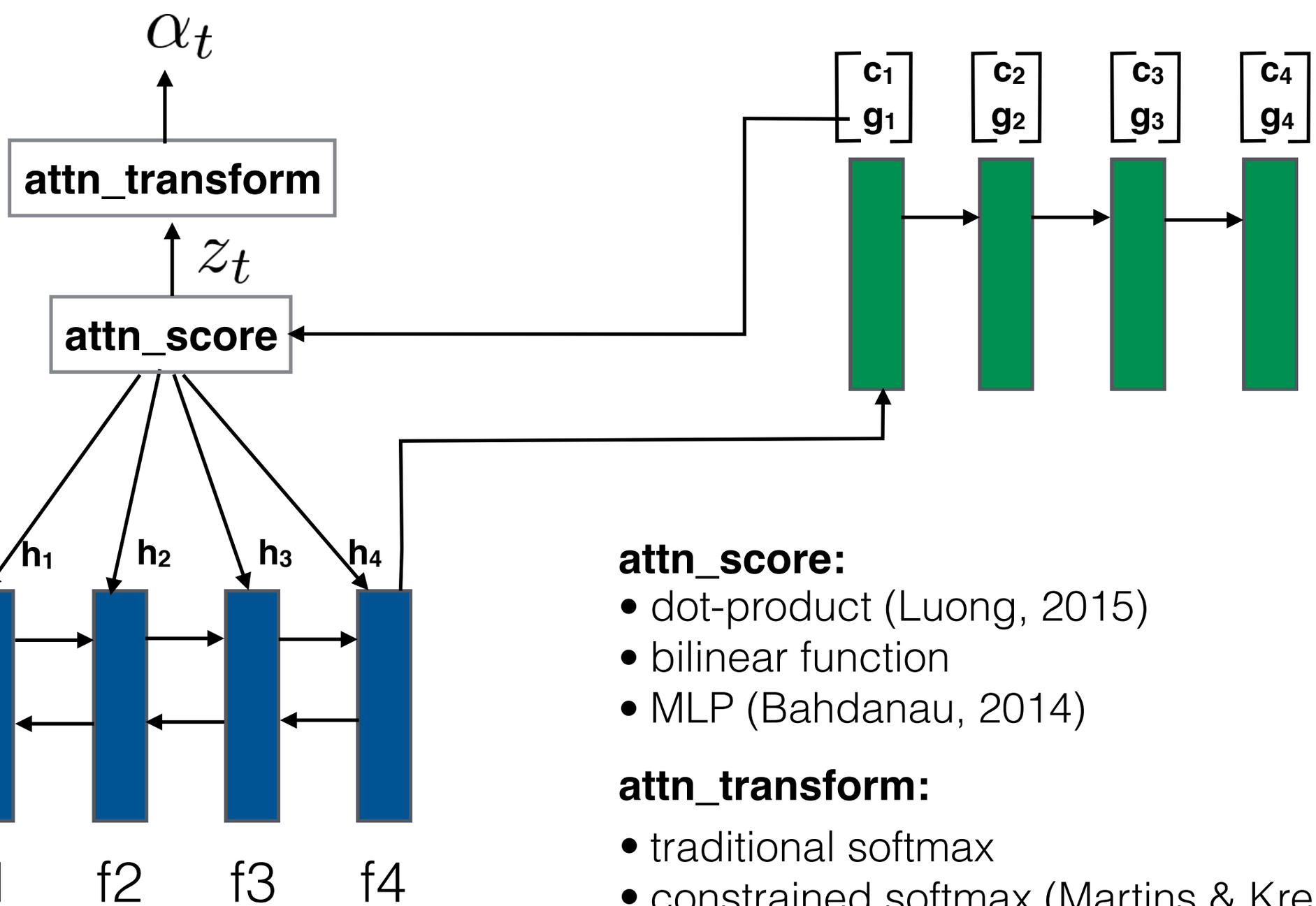
2. Novel attention transform function: *Constrained Sparsemax*
   (Enforces these bounds)

3. Evaluation Metrics: REP-Score and DROP-Score

NMT + Attention Architecture

$$p_t = \text{softmax}(W[g_t; c_t])$$

$$c_t = H\alpha_t$$



attn_score:

- dot-product (Luong, 2015)
- bilinear function
- MLP (Bahdanau, 2014)

attn_transform:

- traditional softmax
- constrained softmax (Martins & Kreutzer, 2017)
- sparsemax (Martins & Astudillo, 2016)
- **constrained sparsemax** (**this work**)

Attention Transform Functions

- Sparsemax: Euclidean projection of z provides sparse probability distributions.

$$\text{sparsemax}(\boldsymbol{z}) \quad := \quad \underset{\boldsymbol{\alpha} \in \Delta^J}{\arg\min} \, \|\boldsymbol{\alpha} - \boldsymbol{z}\|^2$$

- Constrained Softmax: Returns the distribution closest to softmax whose attention probabilities are bounded by upper bounds u.

$$\text{csoftmax}(\boldsymbol{z}; \boldsymbol{u}) \; := \; \underset{\boldsymbol{\alpha} \in \Delta^J}{\arg\min} \, \text{KL}(\boldsymbol{\alpha} \| \, \text{softmax}(\boldsymbol{z}))$$

$$\text{s.t.} \quad \boldsymbol{\alpha} \leq \boldsymbol{u}$$

Attention Transform Functions

- Sparsemax: Euclidean projection of z provides sparse probability distributions.

$$\text{sparsemax}(\boldsymbol{z}) \quad := \quad \underset{\boldsymbol{\alpha} \in \Delta^J}{\arg\min} \|\boldsymbol{\alpha} - \boldsymbol{z}\|^2$$

Sparse **and** Constrained?

- Constrained softmax: Returns the distribution closest to softmax whose attention probabilities are bounded by upper bounds u.

$$\text{csoftmax}(\boldsymbol{z}; \boldsymbol{u}) \quad := \quad \underset{\boldsymbol{\alpha} \in \Delta^J}{\arg\min} \, \text{KL}(\boldsymbol{\alpha} \| \, \text{softmax}(\boldsymbol{z}))$$

$$\text{s.t.} \quad \boldsymbol{\alpha} \leq \boldsymbol{u}$$

Constrained Sparsemax

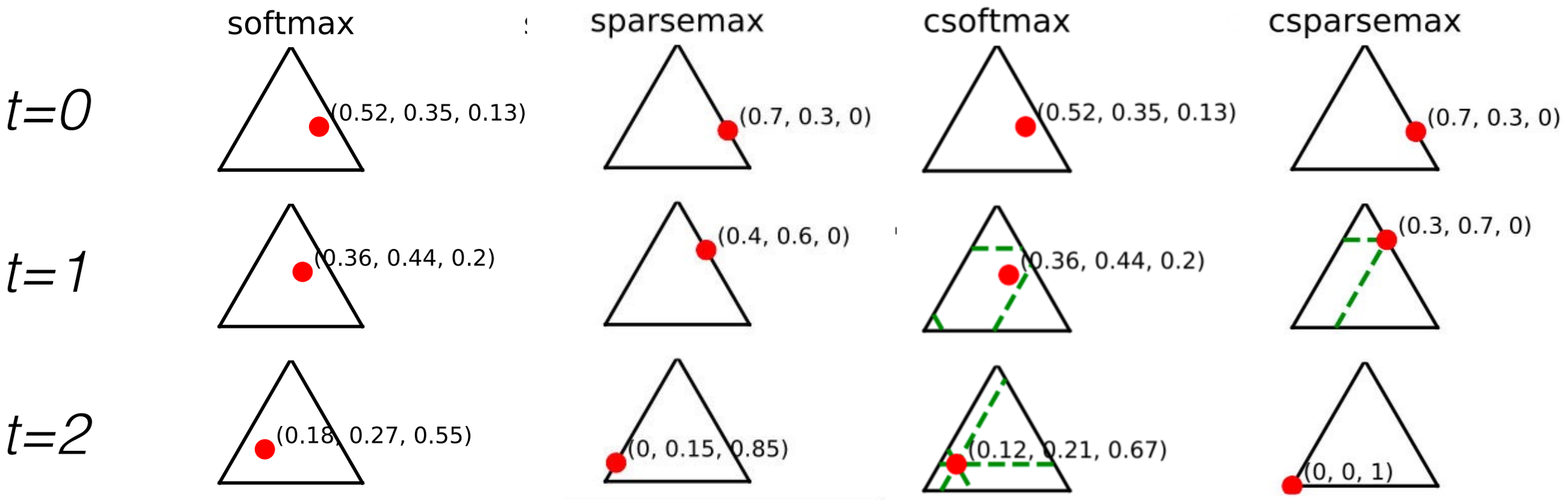
- Provides sparse ***and*** bounded probability distributions.

$$\text{csparsemax}(\boldsymbol{z}; \boldsymbol{u}) := \underset{\boldsymbol{\alpha} \in \Delta^J}{\arg\min} \|\boldsymbol{\alpha} - \boldsymbol{z}\|^2 \quad \text{s.t.} \quad \boldsymbol{\alpha} \leq \boldsymbol{u}.$$

- This transformation has two levels of sparsity:
  over time steps & over attended words at each step.

- Efficient linear and sublinear time algorithms for
  forward and backward propagation.

Visualization: Attention transform functions



- csparsemax provides sparse and constrained probabilities.

Fertility-based NMT Model

Fertility-based NMT

- Allocate fertilities f for each source word as attention budgets that exhaust over decoding.
- Fertility Predictor : Train biLSTM model supervised by fertilities from fast_align (IBM Model 2).

Fertility-based NMT

- Fertilities incorporated as:

$$\alpha_t = \text{csparsemax}(z_t, \underbrace{f - \beta_{t-1}}_{u_t})$$

$$\beta_{t-1} := \sum_{\tau=1}^{t-1} \alpha_\tau$$

- Exhaustion strategy to encourage more attention for words with larger credit remaining:

$$z'_t = z_t + cu_t$$

Experiments

Experiments

- Experiments performed on 3 language pairs:
  De-En (IWSLT 2014), Ro-En (Europarl), Ja-En (KFTT).

- Joint BPE with 32K merge operations.

- Default hyperparameter settings in OpenNMT-Py.

- Baselines: *Softmax*, + *CovPenalty* (Wu, 2016) and
  + *CovVector* (Tu, 2016)

Evaluation Metrics: REP-Score & DROP-Score

REP Score:

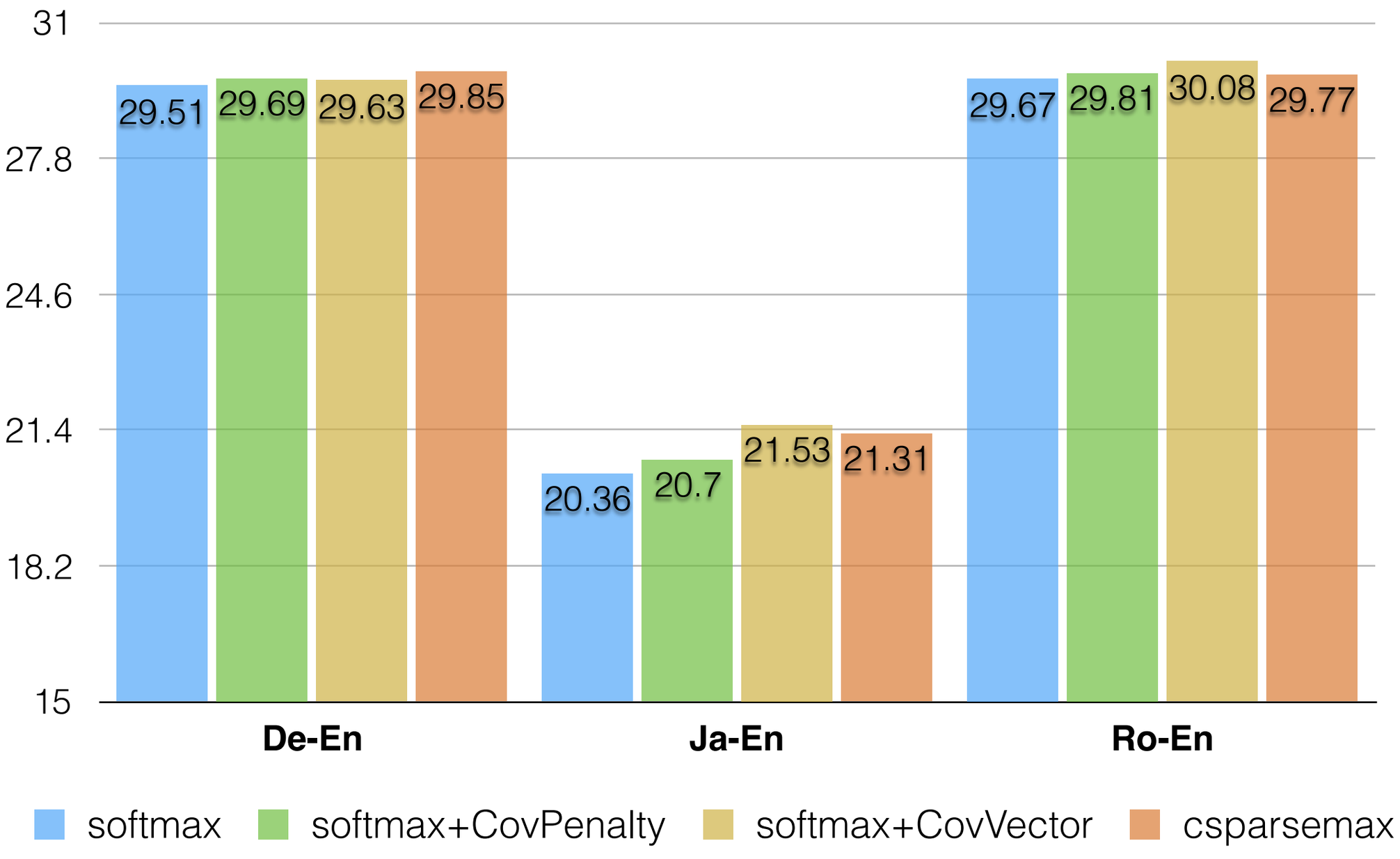
- Penalizes n-gram repetitions in predicted translations.
- Normalize by number of words in reference corpus.

DROP Score:

- Find word alignments from source to reference & source to predicted.
- % of source words aligned with some word in reference, but not with any word in predicted translation.

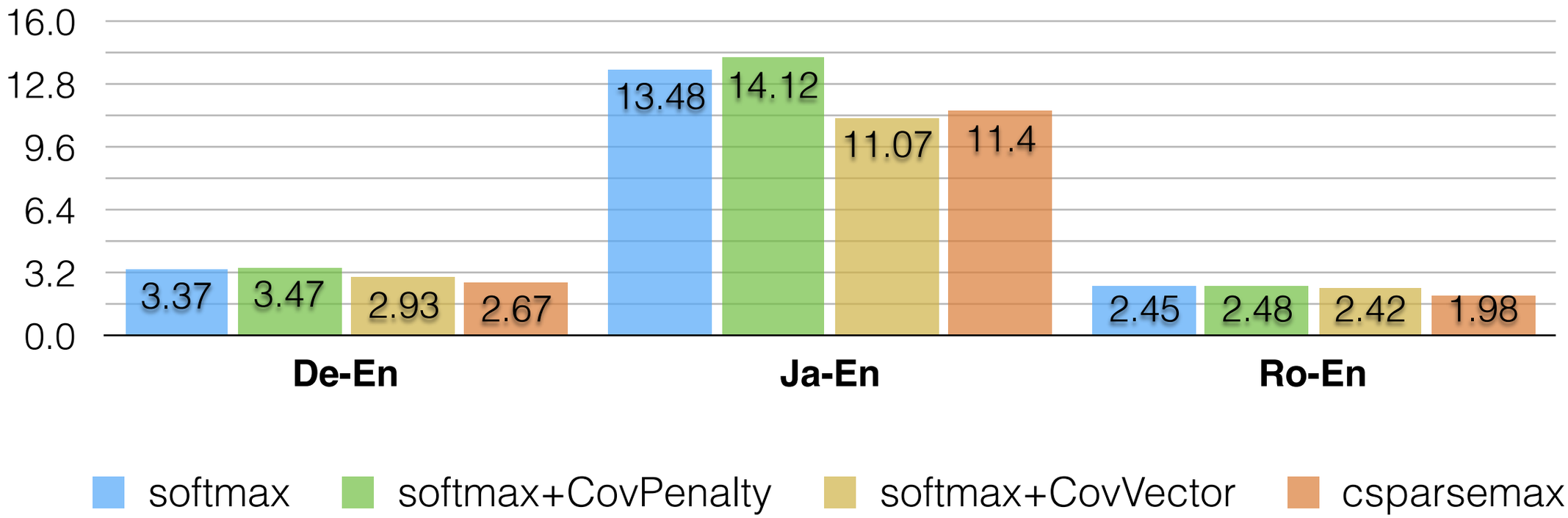
Results

BLEU Scores

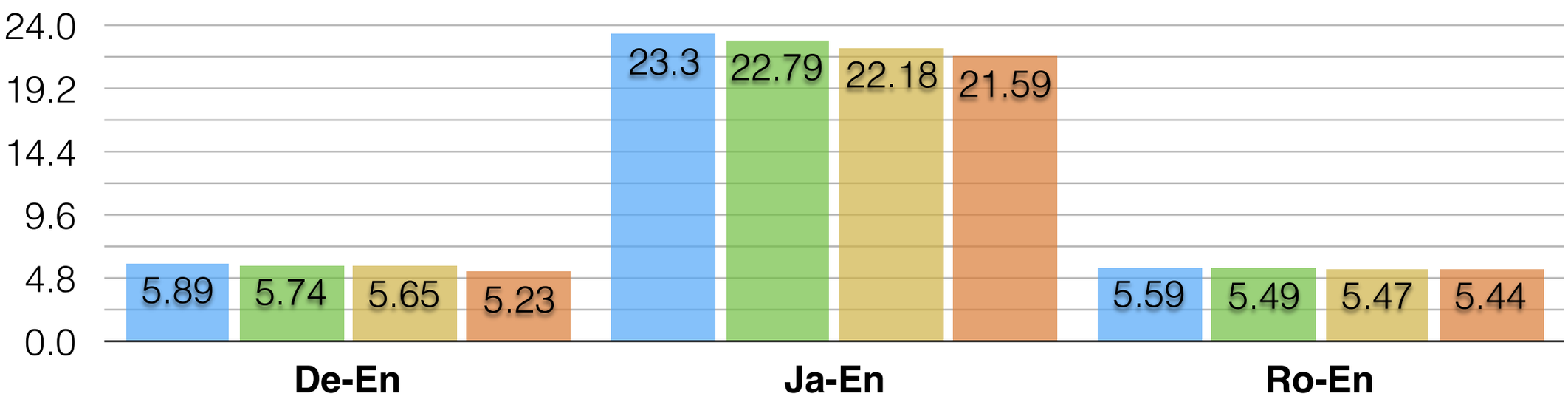


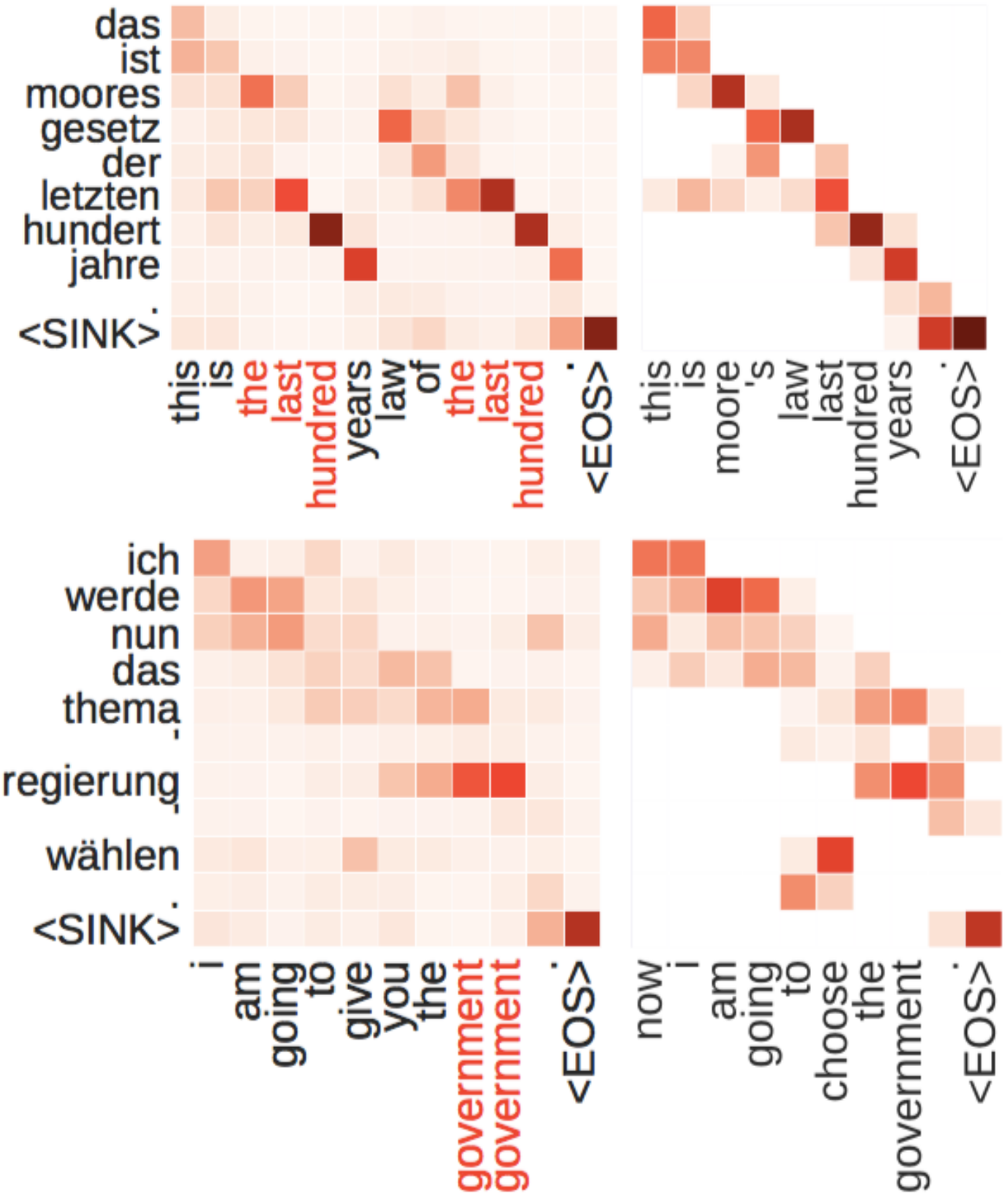
REP Scores

Lower is better!



DROP Scores





softmax ***csparsemax***

- csparsemax yields sparse set of alignments and avoids repetitions.

Examples of Translations

input	überlassen sie das ruhig uns .
reference	*leave that up to us* .
softmax	**give us a silence** .
csparsemax	leave it to us .

input	so ungefähr , sie wissen schon .
reference	*like that , you know* .
softmax	**so , you know , you know** .
csparsemax	like that , you know .

input	wir sehen das dazu , dass phosphor wirklich kritisch ist .
reference	we can see *that* phosphorus is really critical .
softmax	we see **that that** phosphorus is really critical .
csparsemax	we see that phosphorus is really critical .

More in the paper…

Thank You!

Code: www.github.com/Unbabel/sparse_constrained_attention

Questions?