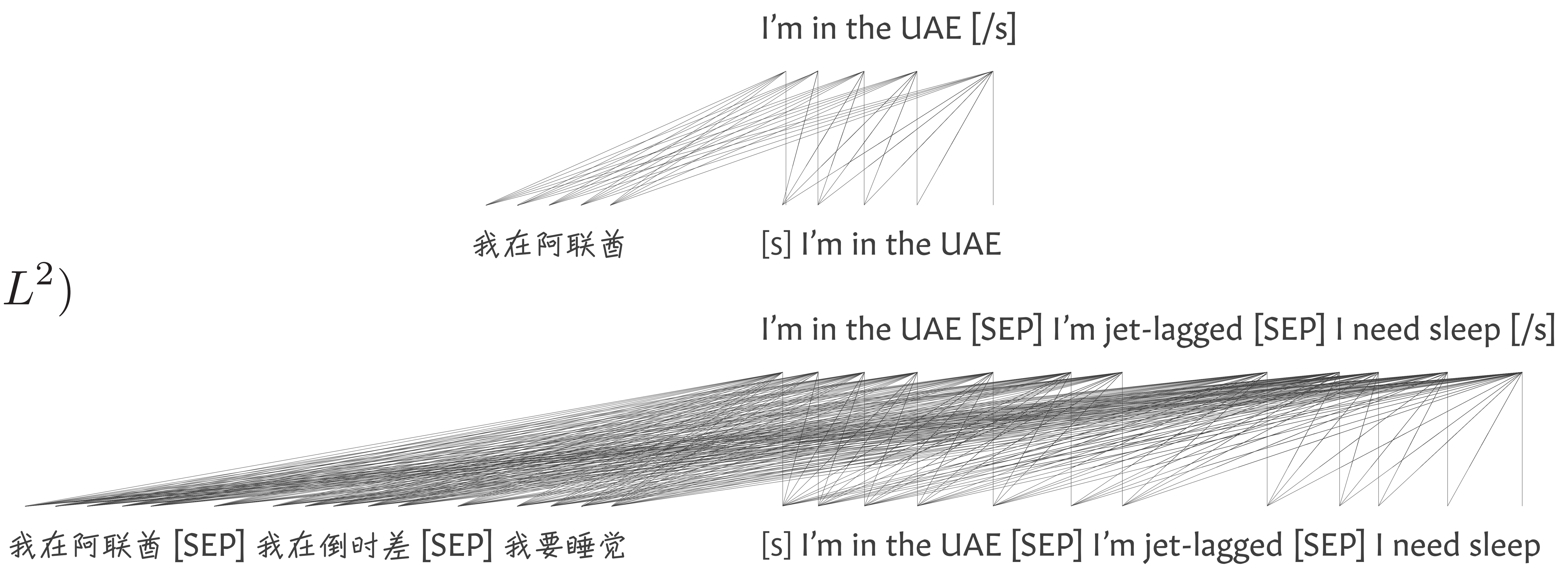
Zhaofeng Wu, Hao Peng, Nikolaos Pappas, Noah Smith
zfw@csail.mit.edu

Modeling Context With Linear Attention for Scalable Document-Level Translation

Replacing transformer attention to a linear approximation leads to substantial speedup on doc-level translation with similar/better quality

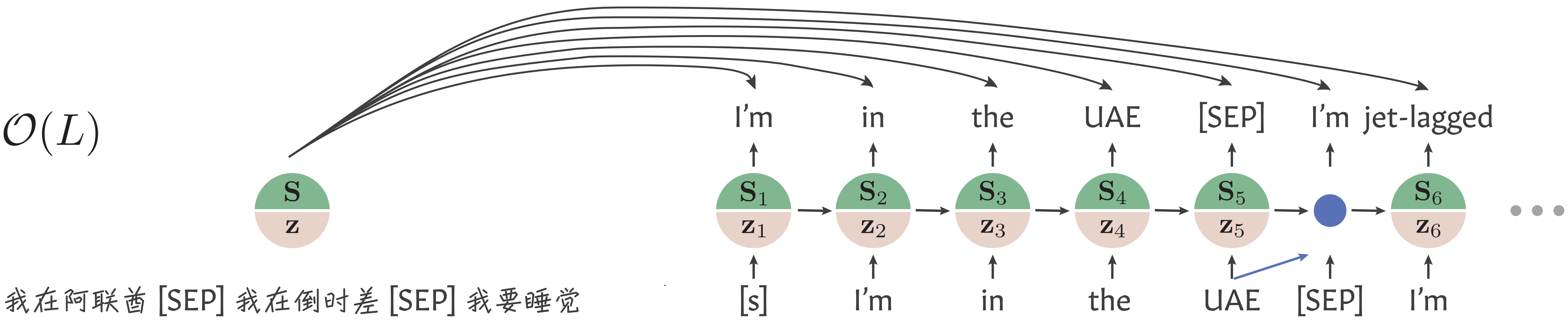
DOC-LEVEL TRANSLATION

$\mathcal{O}(L^2)$



RANDOM FEATURE ATTENTION

$\mathcal{O}(L)$



$$\sum_{i=1}^{N} \frac{\exp(\mathbf{q}_t\cdot\mathbf{k}_i)}{\sum_{j=1}^{N} \exp(\mathbf{q}_t\cdot\mathbf{k}_j)}\mathbf{v}_i^\top = \mathrm{attn}(\mathbf{q}_t,\{\mathbf{k}_i\},\{\mathbf{v}_i\}) \approx \mathrm{RFA}(\mathbf{q}_t,\{\mathbf{k}_i\},\{\mathbf{v}_i\}) = \frac{\phi(\mathbf{q}_t)\cdot\mathbf{S}_t}{\phi(\mathbf{q}_t)\cdot\mathbf{z}_t}$$

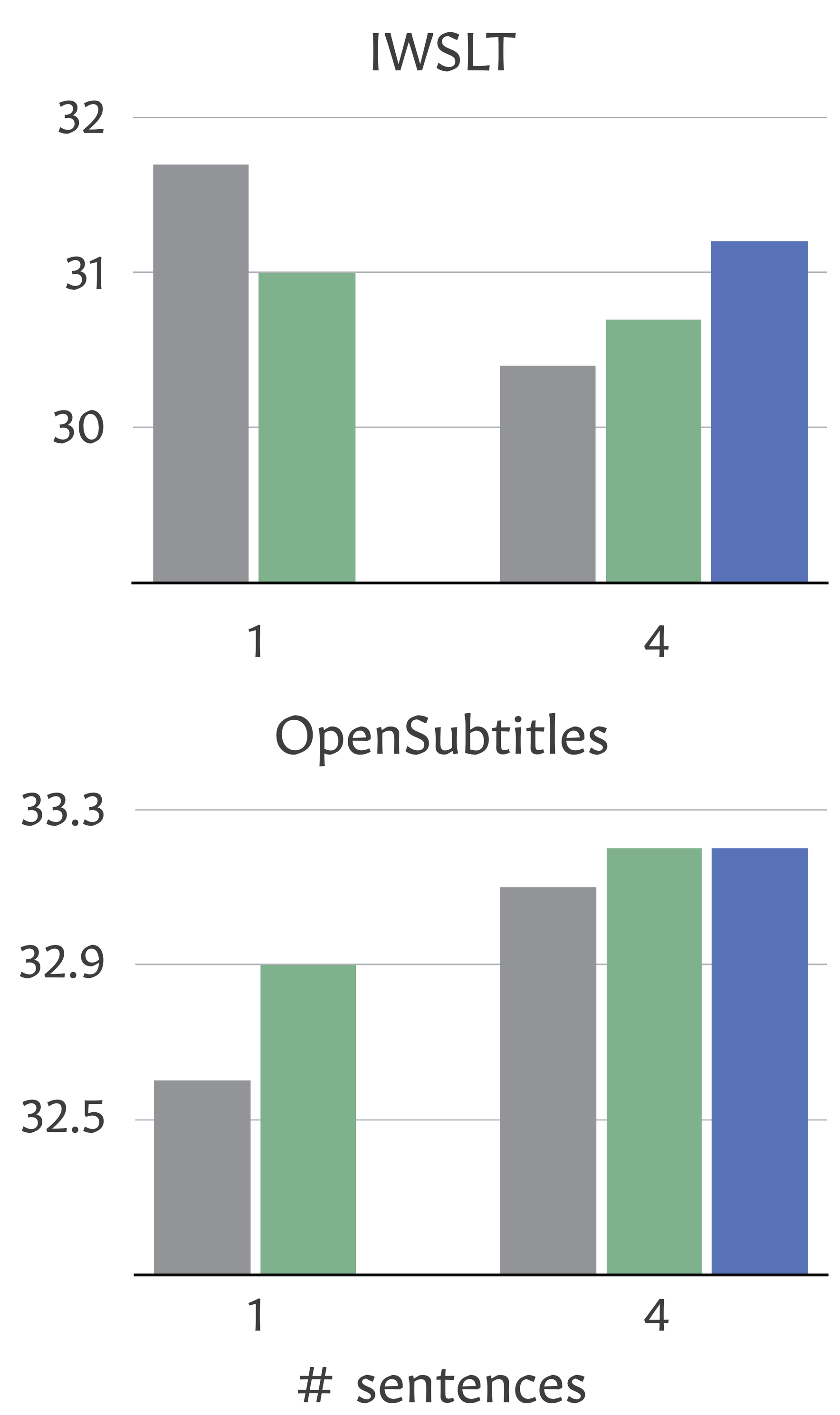
$$\mathbf{S} = \sum_{i=1}^{N}\phi(\mathbf{k}_i)\mathbf{v}_i^\top \qquad \mathbf{z} = \sum_{i=1}^{N}\phi(\mathbf{k}_i)$$

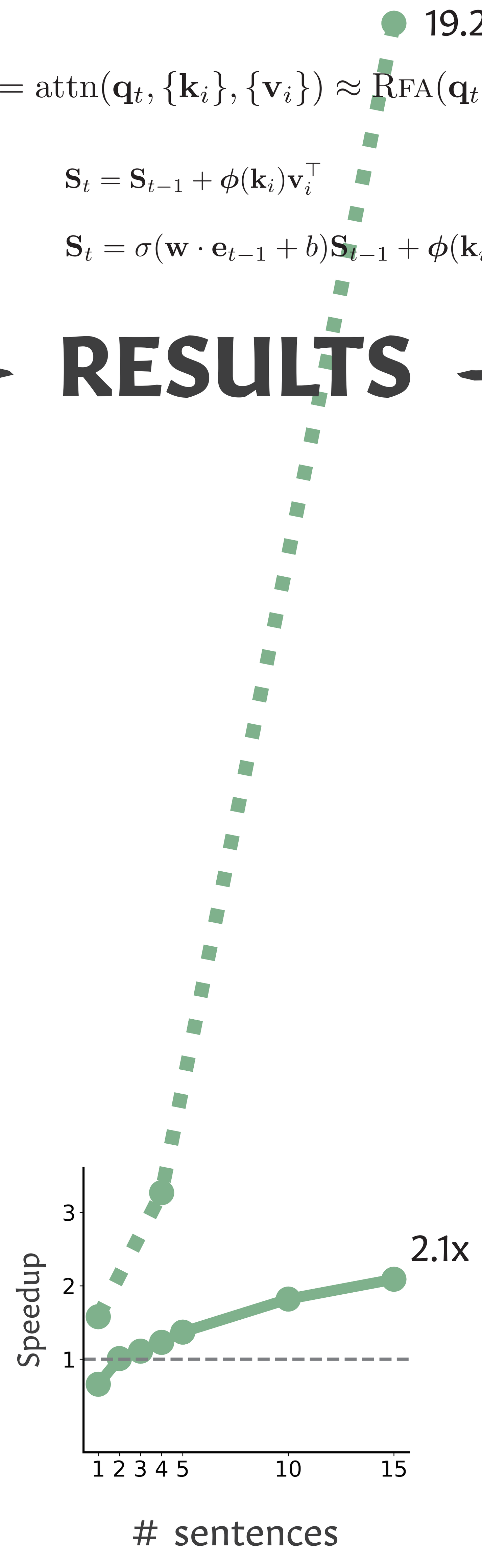
$\mathbf{S}_t = \mathbf{S}_{t-1} + \phi(\mathbf{k}_i)\mathbf{v}_i^\top \qquad \mathbf{z}_t = \mathbf{z}_{t-1} + \phi(\mathbf{k}_i)$ — Peng et al. (2021)

$\mathbf{S}_t = \sigma(\mathbf{w}\cdot\mathbf{e}_{t-1}+b)\mathbf{S}_{t-1} + \phi(\mathbf{k}_i)\mathbf{v}_i^\top \qquad \mathbf{z}_t = \sigma(\mathbf{w}\cdot\mathbf{e}_{t-1}+b)\mathbf{z}_{t-1} + \phi(\mathbf{k}_i)$ — + sentential gate

RESULTS

BLEU



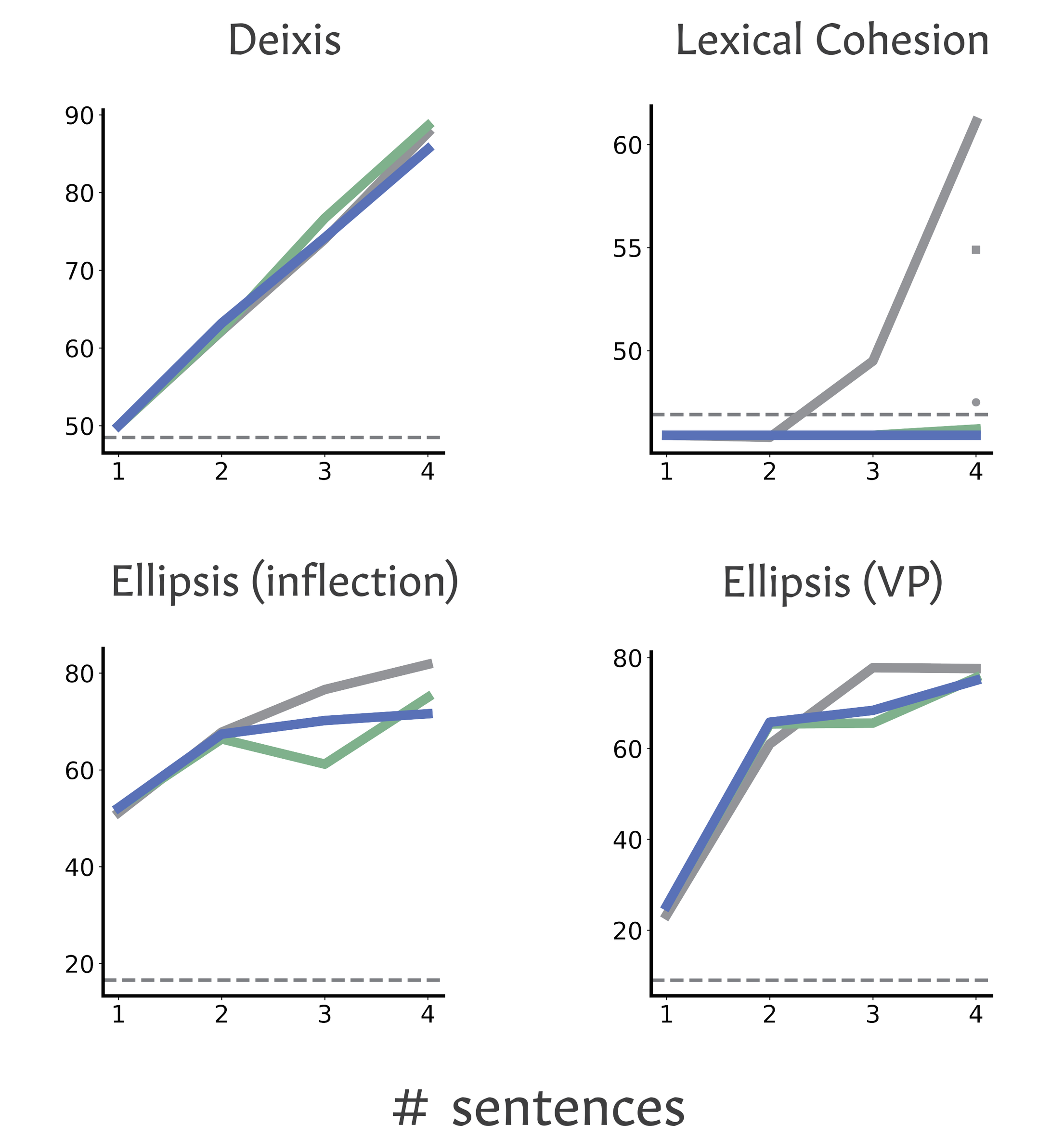


Consistency

I have received your gift [SEP] Thank you

我 收到 你 的 礼物 了 [SEP] 谢谢 您

I receive you GEN gift PFV Thank you-FOR



Ma, Edunov, Auli. A comparison of approaches to document-level machine translation. 2021.
Peng, Pappas, Yogatama, et al. Random feature attention. 2021.
Voita, Sennrich, Titov. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. 2019.

paper