A Tree-based Decoder for Neural Machine Translation

Xinyi Wang, Hieu Pham, Pengcheng Yin, Graham Neubig

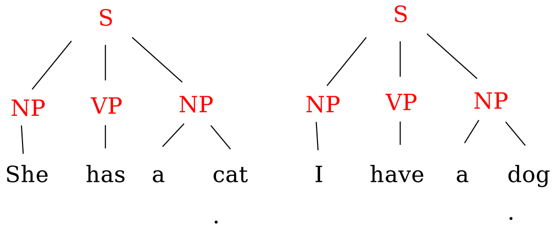
Carnegie Mellon University Language Technologies Institute

November 4, 2018

Tree Structures for Language

Language Technologies Institute

- Tree structures: captures inherent hierarchical structure of language
- Hypothesis: Improve generalization for low-resource data



Previous works on Syntactic MT

- Standard sequence decoder w/ multi-task objective
  - CCG interleaving [Nadejde et al., 2017]
    N **Jane** (S[dcl])/NP **had** NP/N **a** N **cat** . **.**
  - Linearized tree [Aharoni and Goldberg, 2017]
    (ROOT (S (NP **Jane**)NP (VP **had** (NP **a cat**)NP)VP **.**)S)ROOT
  - Sequence decoder + RNNG multi-task [Eriguchi et al., 2017]
- Restricted to specific type of syntactic structures
  - Dependency tree [Wu et al., 2017]

TrDec

- Natural integration of tree topology into decoding

While there is open non-terminal :

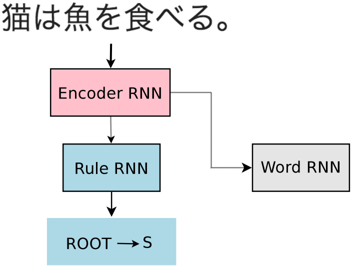
Rule RNN generates the next rule

If a preterminal is generated

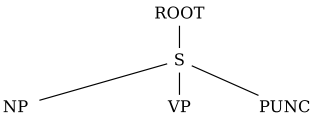
Word RNN generates words until eop

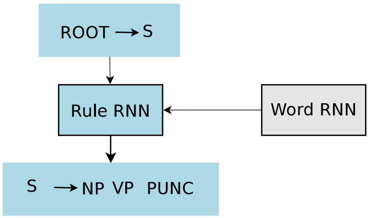
- Can flexibly work with any type of tree structure and compare tree topologies

TrDec: decoding process

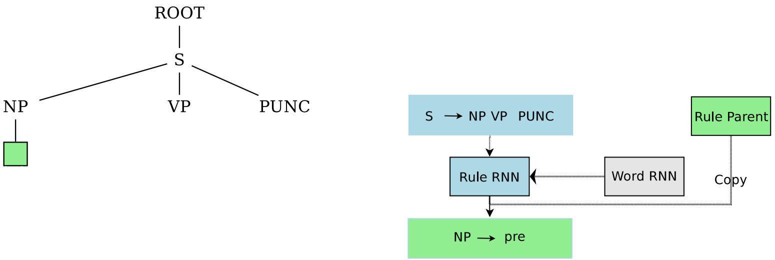


TrDec: decoding process

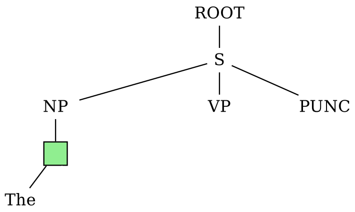


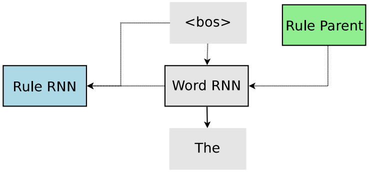


TrDec: decoding process

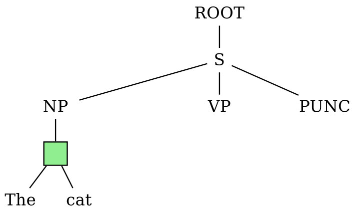


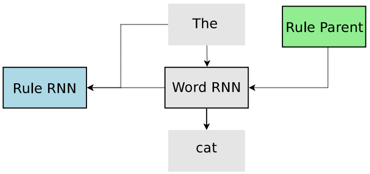
TrDec: decoding process



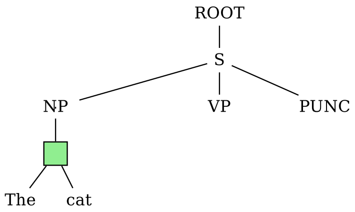


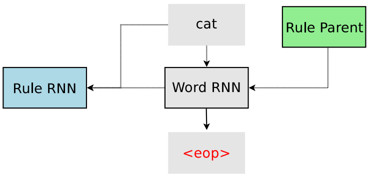
TrDec: decoding process



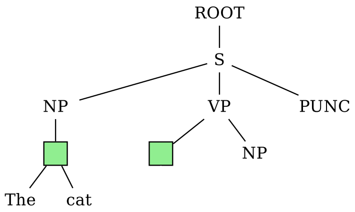


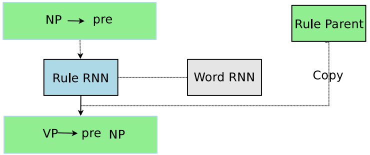
TrDec: decoding process



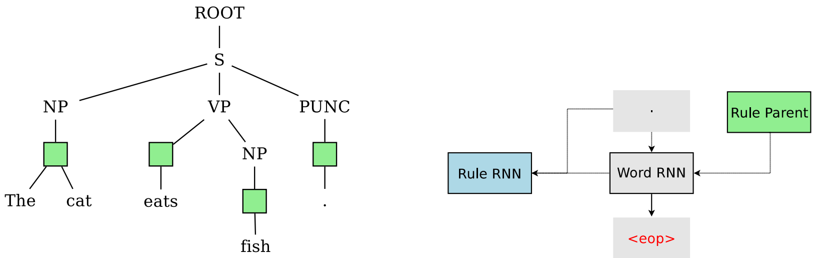


TrDec: decoding process



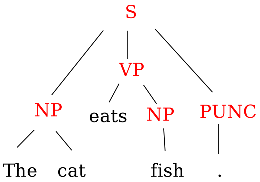


TrDec: decoding process

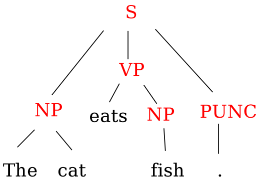


Tree Structures: Syntactic

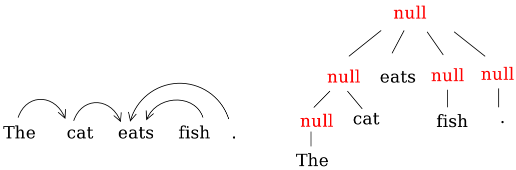
- Constituency (TrDec-con)



- Constituency (TrDec-con)



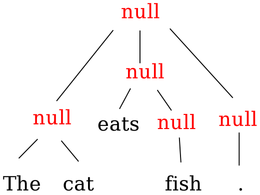
- Dependency (TrDec-dep)



Tree Structures: Ablated Syntactic Labels

- Does label information help?

- Does label information help?
- Unlabeled constituency (TrDec-con-null)



Tree Structures: Non-syntactic

- Does syntactic information help?

Tree Structures: Non-syntactic

- Does syntactic information help?
- Two types of balanced binary trees (TrDec-binary)

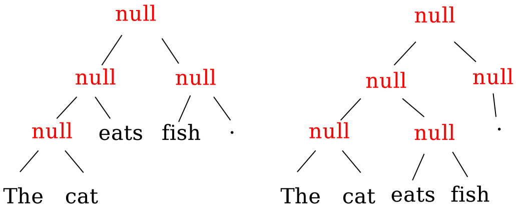


Figure: *Left*: built from top down; *Right*: built from bottom up

Experiments

- Data
  - or-en: small
  - de-en: medium
  - ja-en: medium
- Baselines
  - seq2seq with attention [Bahdanau et al., 2015]
  - CCG interleaving (CCG) [Nadejde et al., 2017]
  - NULL interleaving (CCG-null)
  - Linearized constituency tree (LIN) [Aharoni and Goldberg, 2017]

Model	**ja-en**	**de-en**	**or-en** (mean ± std)
seq2seq	21.10	32.26	10.90 ± 0.57
TrDec-con	21.59	31.93	11.43 ± 0.58
TrDec-con-null	22.72	31.21	11.35 ± 0.55
TrDec-dep	21.41	31.23	8.40 ± 0.5
TrDec-binary	**23.14***	32.65	**13.10**** ± **0.61**

Table: BLEU score of the models

Language Technologies Institute

- Syntactic tags don't have large effect

Model	ja-en	de-en	or-en (mean $\pm$ std)
seq2seq	21.10	32.26	10.90 ± 0.57
TrDec-con	21.59	31.93	11.43 ± 0.58
TrDec-con-null	22.72	31.21	11.35 ± 0.55
TrDec-dep	21.41	31.23	8.40 ± 0.5
TrDec-binary	**23.14***	32.65	$\mathbf{13.10}^{**} \pm \mathbf{0.61}$

Table: BLEU score of the models

Results: TrDec

- Syntactic tags don't have large effect
- Balanced binary trees win

Model	**ja-en**	**de-en**	**or-en** (mean $\pm$ std)
seq2seq	21.10	32.26	10.90 $\pm$ 0.57
TrDec-con	21.59	31.93	11.43 $\pm$ 0.58
TrDec-con-null	22.72	31.21	11.35 $\pm$ 0.55
TrDec-dep	21.41	31.23	8.40 $\pm$ 0.5
TrDec-binary	**23.14***	32.65	**13.10**** $\pm$ **0.61**

Table: BLEU score of the models

Results: TrDec

- Syntactic tags don't have large effect
- Balanced binary trees win
- Constituency trees perform better than dependency trees

Model	**ja-en**	**de-en**	**or-en** (mean $\pm$ std)
seq2seq	21.10	32.26	10.90 ± 0.57
TrDec-con	21.59	31.93	11.43 ± 0.58
TrDec-con-null	22.72	31.21	11.35 ± 0.55
TrDec-dep	21.41	31.23	8.40 ± 0.5
TrDec-binary	$\mathbf{23.14}^{*}$	32.65	$\mathbf{13.10}^{**} \pm \mathbf{0.61}$

Table: BLEU score of the models

Results: Other Syntactic Decoders

Model	ja-en	de-en	or-en (mean $\pm$ std)
seq2seq	21.10	32.26	10.90 ± 0.57
CCG	22.44	32.84	12.55 ± 0.60
CCG-null	21.31	**33.10**	11.96 ± 0.57
LIN	21.55	31.79	12.66 ± 0.61
TrDec-binary	**23.14***	32.65	$\mathbf{13.10}^{**} \pm \mathbf{0.61}$

Table: BLEU score of the models

Results: Other Syntactic Decoders

Language Technologies Institute

- TrDec-binary outperforms the alternatives for two of the three datasets

Model	**ja-en**	**de-en**	**or-en** (mean $\pm$ std)
seq2seq	21.10	32.26	10.90 ± 0.57
CCG	22.44	32.84	12.55 ± 0.60
CCG-null	21.31	**33.10**	11.96 ± 0.57
LIN	21.55	31.79	12.66 ± 0.61
TrDec-binary	**23.14***	32.65	$\mathbf{13.10}^{**} \pm \mathbf{0.61}$

Table: BLEU score of the models

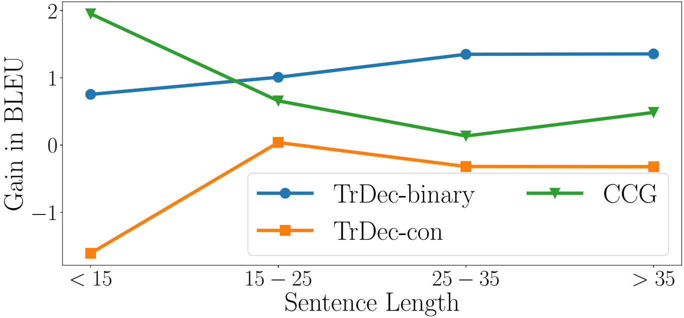
Results: Other Syntactic Decoders

- TrDec-binary outperforms the alternatives for two of the three datasets
- Alternatives in general outperform seq2seq

Model	**ja-en**	**de-en**	**or-en** (mean $\pm$ std)
seq2seq	21.10	32.26	10.90 $\pm$ 0.57
CCG	22.44	32.84	12.55 $\pm$ 0.60
CCG-null	21.31	**33.10**	11.96 $\pm$ 0.57
LIN	21.55	31.79	12.66 $\pm$ 0.61
TrDec-binary	**23.14***	32.65	**13.10**** $\pm$ **0.61**

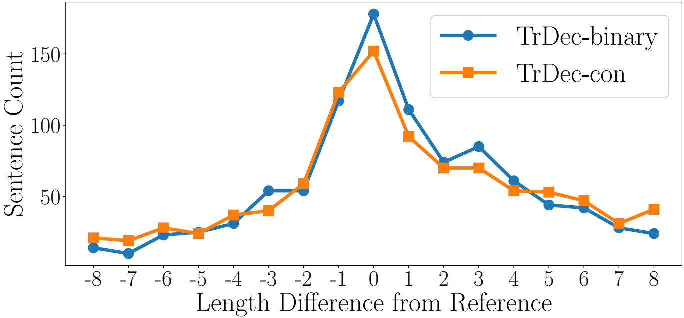
Table: BLEU score of the models

Why does TrDec outperform sequence decoders?



- Particular gain on longer sentences
  - Tree structures facilitate passing information over long distances?

Why syntactic trees don't work as well?



- Binary trees are better at modeling target length

Some thoughts on Syntactic NMT

- Structural bias is helpful

Some thoughts on Syntactic NMT

- Structural bias is helpful
- Identifying the right amount of bias is hard

Some thoughts on Syntactic NMT

- Structural bias is helpful
- Identifying the right amount of bias is hard
- Necessary to distinguish the gain from
  - syntactic information
  - modified model architecture

Some thoughts on Syntactic NMT

- Structural bias is helpful
- Identifying the right amount of bias is hard
- Necessary to distinguish the gain from
  - syntactic information
  - modified model architecture

Code: `https://github.com/cindyxinyiwang/TrDec_pytorch`

Thanks a lot for listening! Questions?

References

- Roee Aharoni and Yoav Goldberg (2017) Towards string-to-tree neural machine translation. In ACL.
- Nadejde et al. (2017) Predicting Target Language CCG Supertags Improves Neural Machine Translation. In WMT.
- Eriguchi et al. (2017) Learning to Parse and Translate Improves Neural Machine Translation. In ACL.
- Wu et al. (2017) Sequence-to-dependency neural machine translation. In ACL.
- Bahdanau et al. (2015) Neural Machine Translation by Jointly Learning to Align and Translate. In ICLR.

TrDec: decoding process