Scalable Syntax-Aware Language Models Using Knowledge Distillation Kuncoro et al. [2019]

馬目 華奈

お茶の水女子大学 戸次研究室

最先端 NLP
September 28, 2019
※図表は論文より引用

背景・概要

- 構造を理解している LM（RNNG）は系列の LM よりも（文法タスクにおいて）成功している
- 構造を学習する LM は計算が複雑でスケールさせるのが難しい
- 系列モデルが大量の教師データを使える場合に、構造のバイアスは必要か？

この論文の内容

- 小さいデータセットから統語情報を学習した LM を LSTM に移転させる Knowledge Distillation(KD) を導入
- 提案手法（DSA-LSTM）は、多くのタスクで SOTA

問題点

- RNNG: 構造（文法）を捉える LM
- LSTM の学習の 10 倍遅い（GPU の恩恵を受けない）
- 大量のコーパスから RNNG LM を構築するのは非実用的

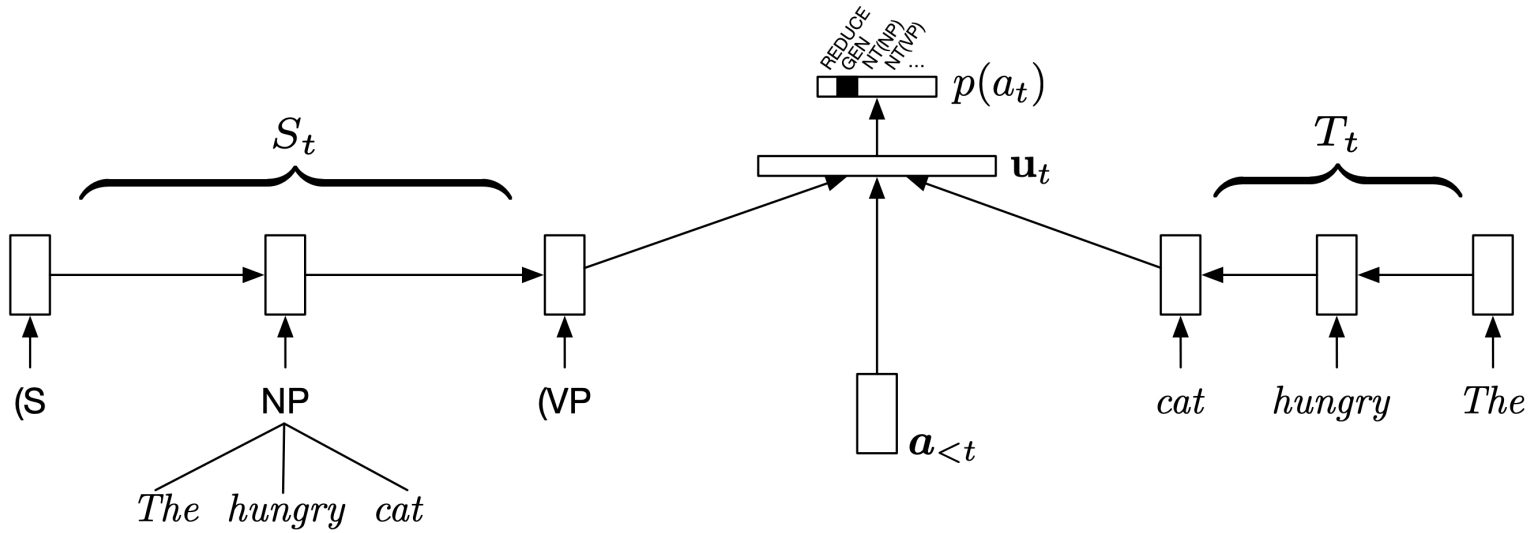
解決方法

- RNNG はスケーリングが難しいため、代わりに小さなデータセットでトレーニングされた RNNG の予測を教師モデルとして使用する
- RNNG を構文構造のガイド役とし LSTM は教師データ全体で学習（生徒モデル）
- このモデルを distilled syntax aware LSTM (DSA-LSTM) と呼ぶ ← 提案手法

RNNG (Dyer et al. [2016])

Recurrent Neural Network Grammar

- 句構造解析と文生成の同時学習モデルであり
- shift reduce 操作による遷移型モデルによって記述
- 生成は、shift の操作を Gen 操作に代替



Knowledge Distillation

- RNNG の $t(x)$ と LSTM の $q_\theta(x)$ のカルバック・ライブラー (KL) 情報量を最小化にするパラメータを探す

$$\begin{aligned}
\hat{\theta}_{KD} &= \arg_\theta \min D_{KL}(t(x)||q_\theta(x)) && (1)\\
&= \arg_\theta \min -\sum_{x\in\Sigma^*} t(x)\log q_\theta(x) && (2)\\
&= \arg_\theta \min -\mathbb{E}_{x\sim t(x)}\log q_\theta(x) && (3)
\end{aligned}$$

- 分散が大きいので KL を local word-level で最小化

$$\mathbb{E}_{x\sim t(x)}\log q_\theta(x) \approx$$
$$\mathbb{E}_{x^*\sim p^*(x)}\sum_{j=1}^{|x^*|} D_{KL}(t(w|x^*_{<j})||q_\theta(w|x^*_{<j}))$$

- 学習データにある文に確率 $\frac{1}{|D|}$ それ以外の文に 0

$$\hat{\theta}_{KD} \approx \arg_\theta \min -\frac{1}{|D|}\sum_{x^*\in D} l_{KD}(x^*_{<j};\theta)$$

$$l_{KD}(x^*;\theta)\sum_{j=1}^{|x^*|}\sum_{w\in\Sigma} t(w|x^*_{<j})\log q_\theta(w|x^*_{<j})$$

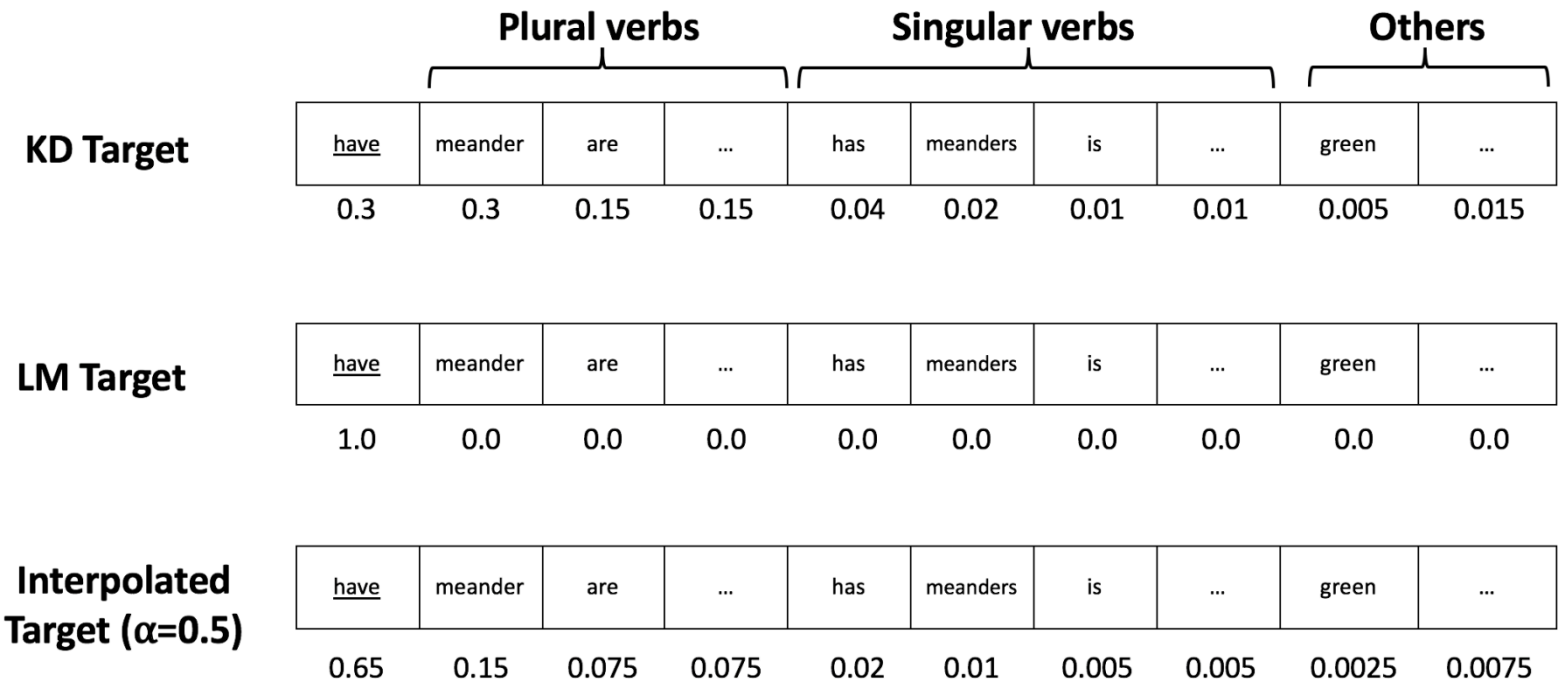
- 実際には Berkeley parser を使って RNNG を学習

$$t(w|x^*_{<j}) \approx t(w|x^*_{<j}, \hat{y}^{berk}_{<j}(x^*))$$

Interpolation

$$\hat{\theta}_{\alpha-int} = \arg_{\theta}\min -\frac{1}{|D|}\sum_{x^* \in D}$$

$$[\alpha l_{KD}(x^*_{<j};\theta) + (1-\alpha)\sum_{j=1}^{|x^*|}\log q_{\theta}(x^*_j|x^*_{<j})]$$



Parts of the river valley $(\alpha = 0.5)$

DSA-LSTM の特徴

- RNNG は文構造についての知識は持つが、学習データが限られるため意味的な知識は限られる
  本アプローチにより文構造と意味の両方について強力な言語モデルを構築できる
- 損失関数のみが従来の LSTM と異なるため、従来の LSTM と同様はるかに高速

実験

	Small Training Set			Full Training Set				
	Small LSTM†	**S-DSA-LSTM†**	**RNNG†**	**Full LSTM**	**BA-LSTM**	**DSA-LSTM**	**BERT**	**Humans**
Gulordava et al. (2018) **test ppl.**	94.54	93.95	**92.30**	**53.73**	54.64	56.74	N/A	N/A
SUBJECT-VERB AGREEMENT								
Simple	0.89	0.96	**0.99**	**1.00**	**1.00**	**1.00**	1.00	0.96
In a sentential complement	0.89	**0.98**	0.93	0.97	**0.98**	**0.98**	0.83	0.93
Short VP coordination	0.90	0.88	**0.96**	0.96	0.95	**0.99**	0.89	0.94
Long VP coordination	0.78	0.74	**0.94**	**0.82**	0.80	0.80	0.98	0.82
Across a prepositional phrase	0.83	0.88	**0.95**	0.89	0.89	**0.91**	0.85	0.85
Across a subject relative clause	0.81	0.87	**0.95**	0.87	0.87	**0.90**	0.84	0.88
Across an object relative clause	0.54	0.69	**0.95**	0.77	0.81	**0.84**	0.89	0.85
Across an object relative clause (no *that*)	0.55	0.61	**0.93**	0.70	0.74	**0.77**	0.86	0.82
In an object relative clause	0.79	0.87	**0.96**	0.90	0.91	**0.92**	0.95	0.78
In an object relative clause (no *that*)	0.72	0.88	**0.96**	0.86	0.83	**0.92**	0.79	0.79
Average of subject-verb agreement	0.77	0.84	**0.95**	0.87	0.88	**0.90**	0.89	0.86
REFLEXIVE ANAPHORA								
Simple	**0.93**	0.90	0.83	0.91	**0.92**	0.91	0.94	0.96
In a sentential complement	0.77	**0.78**	0.46	0.81	0.81	**0.82**	0.89	0.91
Across a relative clause	0.63	0.67	**0.82**	0.64	0.64	**0.67**	0.80	0.87
Average of reflexive anaphora	**0.78**	**0.78**	0.70	0.79	0.79	**0.80**	0.88	0.91
NEGATIVE POLARITY ITEMS								
Simple	0.93	0.84	0.28	0.96	**0.98**	0.94	N/A	0.98
Across a relative clause	0.82	0.73	0.78	0.75	0.70	**0.91**	N/A	0.81
Average of negative polarity items	**0.88**	0.79	0.53	0.86	0.84	**0.92**	N/A	0.90
Average of all constructions	0.79	0.82	**0.85**	0.85	0.86	**0.89**	N/A	0.88

結果のまとめ

評価で DSA-LSTM は以下３つより複数タスクで精度が向上

- ハイパーパラメーターを調整することにより、先行研究より優れた LSTM
- 構造バイアスを活用するが、スケーラビリティに欠ける教師モデルの RNNG
- KD から学習するが、構造バイアスはない born-again ネットワーク（Furlanello et al. [2018]）

まとめ

Qestion

系列モデルが大量の教師データを使える場合に、構造のバイアスは必要か？

Answer

構造バイアスは、大量学習ができても、なお大事

- 増え続ける教師データの恩恵を受けることができるモデルでも、構造バイアスは、構文能力の向上に関連

参考文献 I

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 199–209, San Diego, California, 2016. Association for Computational Linguistics.

Tommaso Furlanello, Zachary Chase Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born-again neural networks. In *In Proc. of ICML*, 2018.

Adhiguna Kuncoro, Chris Dyer, Laura Rimell, Stephen Clark, and Phil Blunsom. Scalable syntax-aware language models using knowledge distillation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3472–3484. Association for Computational Linguistics, 2019.