NLP若手の会 第14回 シンポジウム (YANS2019)
at 北海道 ホテルエミシア札幌 2019/08/26 - 28

解釈性向上のための注意機構と損失勾配に対する関連損失の導入

北田 俊輔, 彌冨 仁

法政大学 理工学研究科 応用情報工学専攻

法政大学
HOSEI University

■ 背景 | 注意機構と損失勾配に対する関連損失

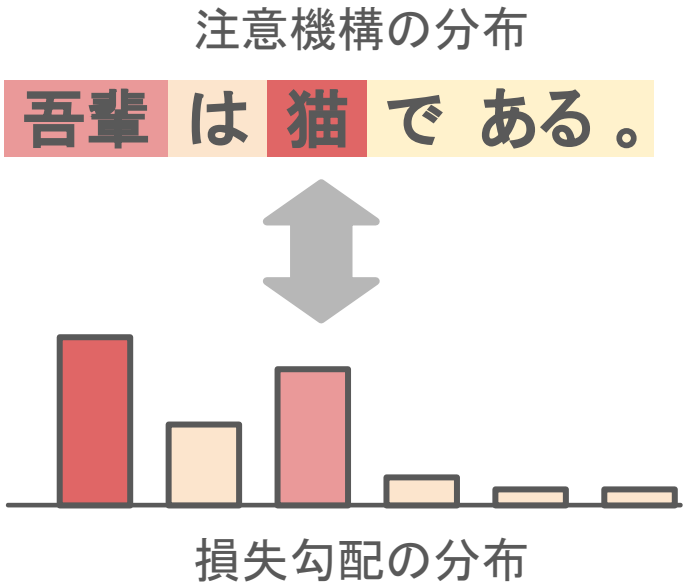
自然言語処理における注意機構

- 予測精度の向上に寄与
- 予測根拠の提示として利用

注意機構における単語の重要度
損失勾配を基にした単語の重要度

➜ 必ずし<u>も関連するとは限らない</u>

[Jain and Wallance, NAACL19]



注意機構の分布と損失勾配を基にした値の分布を
関連するように学習する**関連損失**を導入する

$$\mathcal{L} = \mathcal{L}\left(\mathbf{x}^{(n)}, y^{(n)}\right) + \lambda \mathcal{L}_r\left(\sigma\left(\frac{\partial y^{(n)}}{\partial \mathbf{x}^{(n)}}\mathbf{h}^{(n)}\right), \hat{\alpha}^{(n)}\right)$$

■ 評価実験 | 関連損失を導入したモデル

テキスト分類タスクを用いた比較

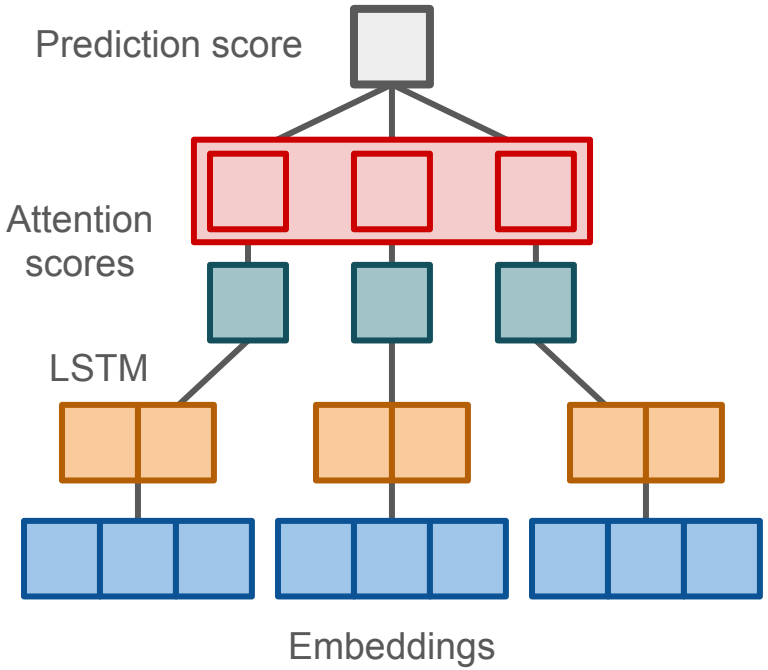
モデル

- 注意機構つき1層LSTM

評価データセット (2クラス)

2クラスになるよう予め前処理

- 20 Newsgroups
- Stanford Sentiment Treebank
- IMDB Movie Review Corpus



学習済みモデルを用いた分析

- 関連損失の有無による注意機構の可視化結果の差異
- 注意機構のスコアと損失勾配を基にしたスコアの関連度合い