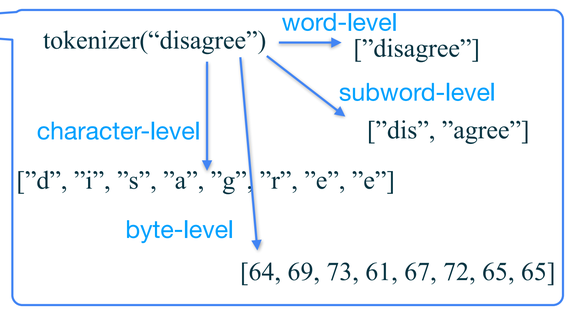
Q3-5 ByT5のAttentionを用いたトークン結合

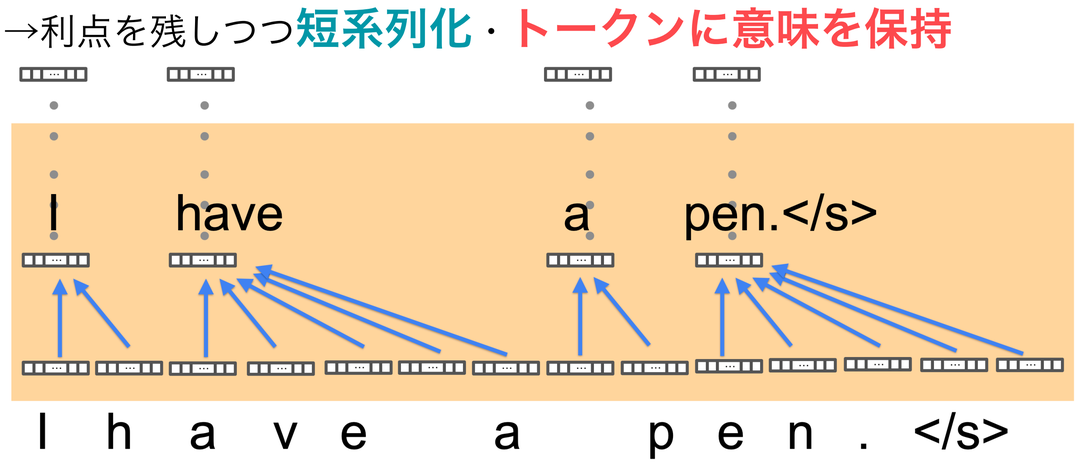
田中康紀　須藤克仁　中村哲
奈良先端科学技術大学院大学

概要

【ByT5 (byte-level T5)】
利点：語彙に非依存
欠点：長系列化
　　トークンと意味の不一致



【提案手法】
ByT5を文字レベル分割とみなし，Attentionを元にトークンを結合
→利点を残しつつ**短系列化**・**トークンに意味を保持**



【結果】
英語テキスト分類において
処理速度：ByT5に比べ1.78倍↑
分類精度：ランダムに結合する手法に比べ 1.8pt↑

実験

英語の感情2値分類タスクであるSST-2にfine-tuningを行った

実験結果1: 分類精度

モデル名	accuracy(%)
ByT5-base	**90.6**
ランダムに選択し結合(ベースライン)	83.5
Attentionから選択し結合(提案手法)	**85.3**
単語レベルで結合	84.5
Attentionから選択し抽出	82.1
6層目で結合	87.1
12層目で結合	89.6

実験設定
CPU
AMD EPYC 7742 64-Core
メモリ 1TB
GPU
NVIDIA A100-SXM 40GB
バッチサイズ12

実験結果2: 処理速度

モデル名	処理速度(it/s)	比率
ByT5-base(ベースライン)	4.88	1
提案手法	8.67	**1.78**
6層目で結合　※論文中記載誤り	7.12	1.47
12層目で結合	5.12	1.15

提案手法はCPU上でのテンソル操作が入る

提案手法

K,u,b,e,r,n,e,t,e,s, ,i,s, ,a,b,b,e,r,v,i,a,t,e,d, ,a,s, ,k,8,s,.,</s>

↓ 2層目の隠れベクトルを得る

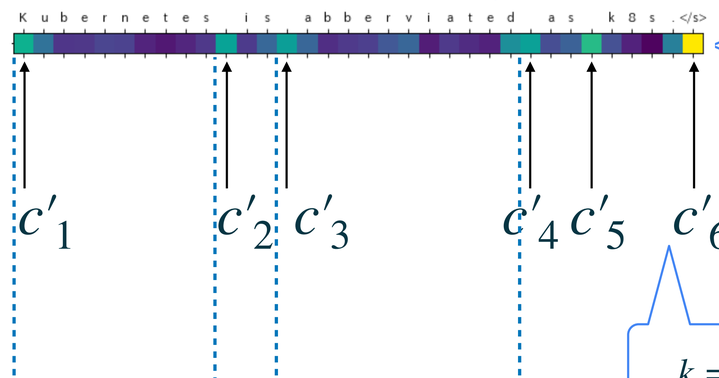
$\mathbf{v}_K \mathbf{v}_u \ \mathbf{v}_b \ \cdots \ \mathbf{v}_s \mathbf{v}_. \ \mathbf{v}_{</s>}$

↓ 2層目のAttention行列を参照

Significant Score

$$\mathrm{Sig}(c) = \frac{1}{L}\sum_{l}^{L} \bar{A}\,[c_l, c]$$

query方向に平均

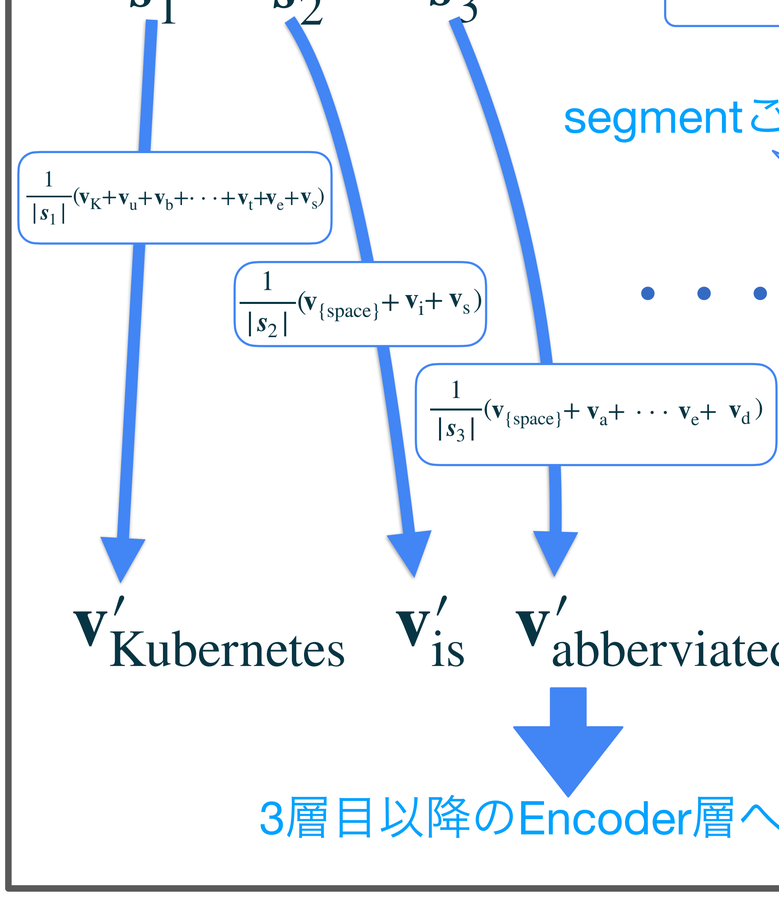


$\mathrm{Sig}(\boldsymbol{c}) = \left(\mathrm{Sig}\,(c_1), \ldots, \mathrm{Sig}\,(c_L)\right)$

上位k個数抽出

$k = \lfloor 34/5 \rfloor = 6$

$\boldsymbol{c}' = \mathrm{topk}(\mathrm{Sig}(\boldsymbol{c})) = \left(c'_1, \ldots, c'_k\right)$



segmentごとにベクトルを平均

$$\boldsymbol{v}'_i = \frac{1}{|\boldsymbol{s}_i|}\sum_{c \in \boldsymbol{s}_i} \boldsymbol{v}_c$$

3層目以降のEncoder層へ入力

考察

短系列化・分類精度

ランダムな手法 < 提案手法
→Attentionのまとまりは，意味のまとまとまりを持つ

ByT5-base　＞　提案手法
→周り文字の結合はAttentionの計算で十分
　ベクトル結合の際に，必要な情報が失われた

処理速度と分類精度はトレードオフの関係にある

トークンの獲得

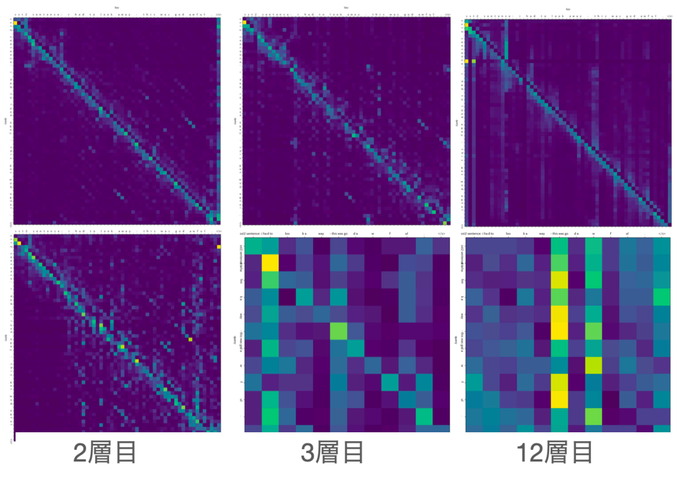
生成タスク・大規模なデータで学習の必要性
単語レベル　＜　提案手法
NNでは単語よりも良いまとまりが存在

入力文
sst2 sentence: it 's a charming and often affecting journey .
結合結果
sst2_sentence: it's_ / a_ / cha / r / ming_and_o / fte / n_ / affecting_ / jo / urne / y_ / . / <s>

入力文
sst2 sentence: unflinchingly bleak and desperate
結合結果
sst2_sentence:_ / unf / linc / hingly_ble / a / k_and_ / / des / pe / r / ate / <s>

Attention行列

類似：query方向の強い値の連続
差異：残差NWの位置による対角成分



validationデータ"sst2 sentence: i had to look away - this was god awful . "を入力