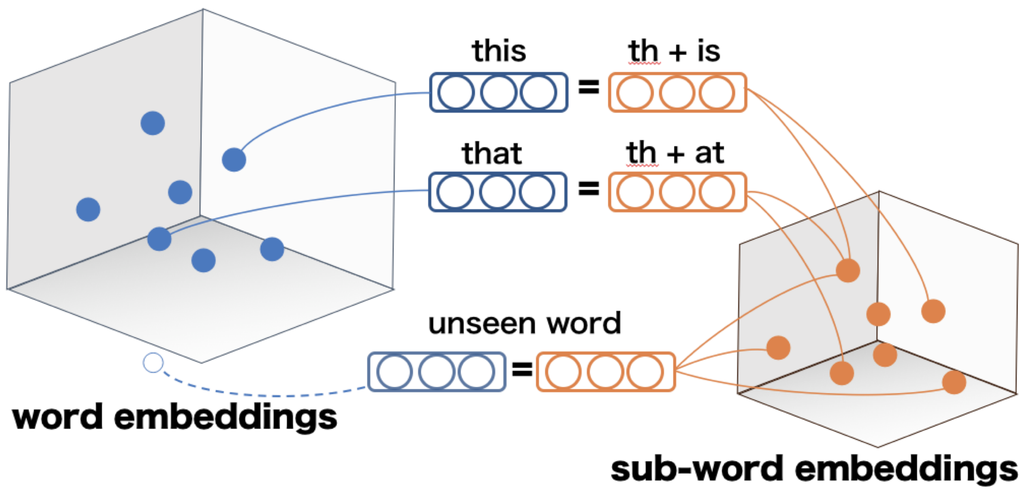
サブワードに基づく単語ベクトルの再構築

佐々木 翔大[1], 鈴木 潤[1,2] 乾 健太郎[1,2] ([1]東北大学, [2]理研 AIP)

概要

- 事前に学習された大規模な単語ベクトル (例 *GloVe, FastText vec.*)には2つの問題が存在
  ①**モデルサイズ**: 各単語に１つずつベクトルを付与→膨大なベクトル数
  ②**未知語**: 事前に決めたボキャブラリ内の単語しか対応できない
- サブワードベースのベクトルの構成によって単語ベクトル空間を表現する手法で、①モデルサイズを**1/10**に削減し、②未知語タスクにおいて既存手法を上回った



単語ベクトルの再構築

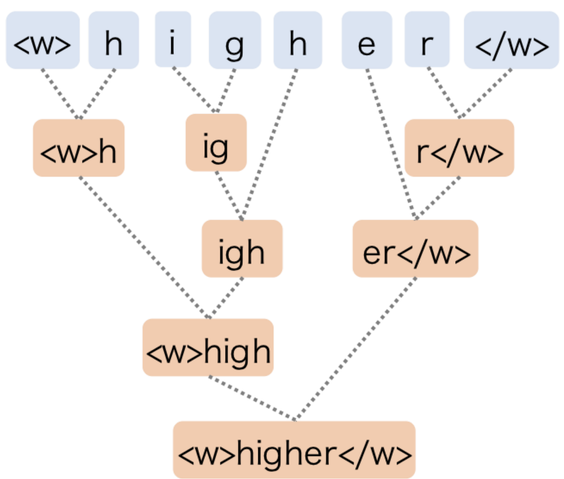
文字 N-gram モデル （既存手法と同等）

✓ 各文字N-gram にベクトルを割り当て、和をとる

unigram	<w>	h	i	g	h	e	r	</w>
bigram	<w>h	hi	ig	gh	he	er	r</w>	
trigram	<w>hi	hig	igh	ghe	her	er</w>		
⋮			⋮					
7-gram	<w>higher	higher</w>						
8-gram	<w>higher</w>							

BPE モデル （提案モデル）

✓ BPEの各マージルールにベクトルを割り当て、和をとる



BPE rule	
① r, </w>	→ r</w>
② <w>, h	→ <w>h
③ e, r</w>	→ er</w>
⋮ ⋮	⋮ ⋮
⑥ <w>h, igh	→ <w>high
⑦ <w>high, er</w>	→ <w>higher</w>

*ただしいずれの手法でも、コーパスでの頻度上位k個のN-gram / BPEマージルール までを用いる

- k個のベクトルを用いて元の単語ベクトル空間を表現
- kをできるだけ小さくしたい （モデルサイズの削減）

サブワードベクトルの学習: サブワードベクトルの和を元の単語ベクトルに近づける

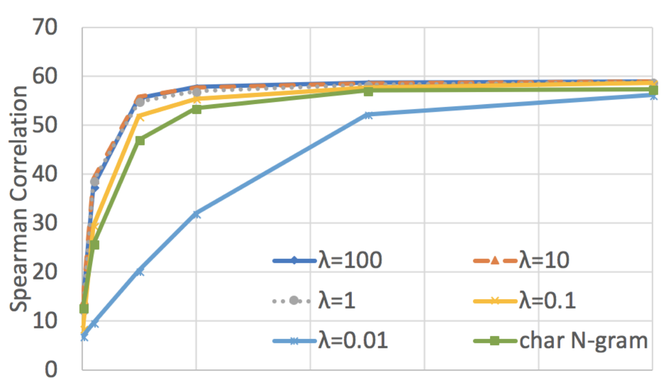
ロス関数: $$\sum_i \left(||A_i - w_i||_2^2 + \lambda \sum_{j \in \alpha(i,M)} \frac{1}{M} \left(A_i w_j - w_i w_j\right)^2 \right)$$

w_i: 単語iのベクトル（教師信号）, A_i: サブワードベクトルの和
$\alpha(i,M)$: 単語iと類似度の高い単語の集合, M: $\alpha(i,M)$の大きさ

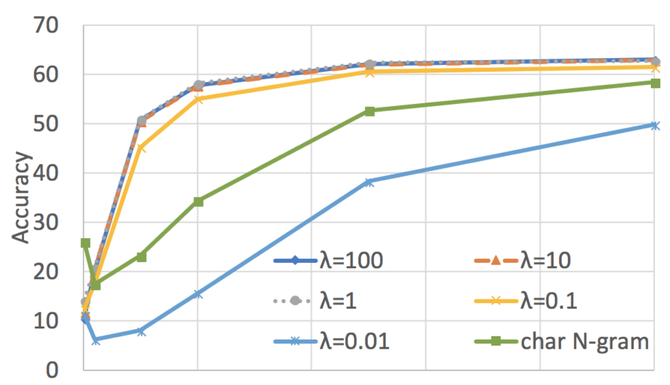
実験①：モデルサイズと性能の関係

＊実験設定
- 事前学習した単語ベクトル（教師信号）
  - Wikipediaコーパス上でSGNS (200d)
  - GloVe.42B.300d

＊タスク
- 単語類似度判定、単語アナロジー



単語類似度判定：モデルサイズ ― スピアマン順位相関係数



単語アナロジー：モデルサイズ ― 正解率

＊結果と考察
- 提案手法が既存手法（*FastText*相当）の性能を上回った
- 元の単語ベクトルの性能を維持しながら、**モデルサイズを 約 1/10 に削減することに成功**
- 提案手法は情報量の高いサブワードにベクトルを割り当て、効率的に単語ベクトル空間を表現していると考えられる

method	model size k	単語類似度判定 MEN	MC	MTurk	RARE	R&G	SCWS	Simlex	WSR	WSS	アナロジー GL	MSYN
SGNS-Wiki												
(Orig. word emb.)	434k	(71.0)	(68.9)	(71.6)	(42.1)	(73.6)	(65.0)	(32.4)	(53.5)	(72.9)	(66.7)	(53.9)
Reconst. N-gram	100k	65.4	70.7	65.5	33.5	56.6	61.4	28.6	47.6	65.3	32.8	38.6
Reconst. BPE	**100k**	**70.9**	**71.2**	**67.9**	**33.9**	**68.4**	**63.7**	**31.8**	**52.9**	**72.4**	**61.8**	**49.0**
GloVe												
(Orig. word emb.)	1917k	(74.3)	(64.6)	(77.7)	(38.4)	(81.7)	(54.0)	(37.4)	(60.0)	(69.6)	(75.0)	(71.6)
Reconst. N-gram	200k	50.4	28.9	56.4	**35.3**	36.3	43.5	18.5	46.6	45.7	29.2	43.9
Reconst. BPE	**200k**	**76.8**	**80.5**	**65.5**	27.4	**83.1**	**54.9**	**40.3**	**63.5**	**72.3**	**73.7**	**59.3**

実験②：未知語タスクにおける性能

＊実験設定
- 未知語を含む問題も評価
- lower bound:
  未知語にランダムベクトルを割り当てる

＊結果
- 各手法はlower boundの性能から改善が見られる
- **提案手法が文字N-gramの性能を上回った**

method	model size k	単語類似度判定 RARE	SCWS	WSR	WSS	アナロジー MSYN
Reconst. N-grams	100k	28.1	61.1	43.5	64.7	37.3
(lower bound)	100k	(22.2)	(60.3)	(38.6)	(63.1)	(36.6)
Reconst. BPE-rules	**100k**	**29.3**	**63.5**	**46.5**	**72.1**	**46.4**
(lower bound)	100k	(22.6)	(62.5)	(43.3)	(69.8)	(46.4)