2015/08/29 第7回最先端NLP勉強会

Sparse Overcomplete Word Vector Representations

Manaal Faruqui Yulia Tsvetkov Dani Yogatama Chris Dyer Noah A. Smith

読む人 : 中路紘平
スマートニュース株式会社

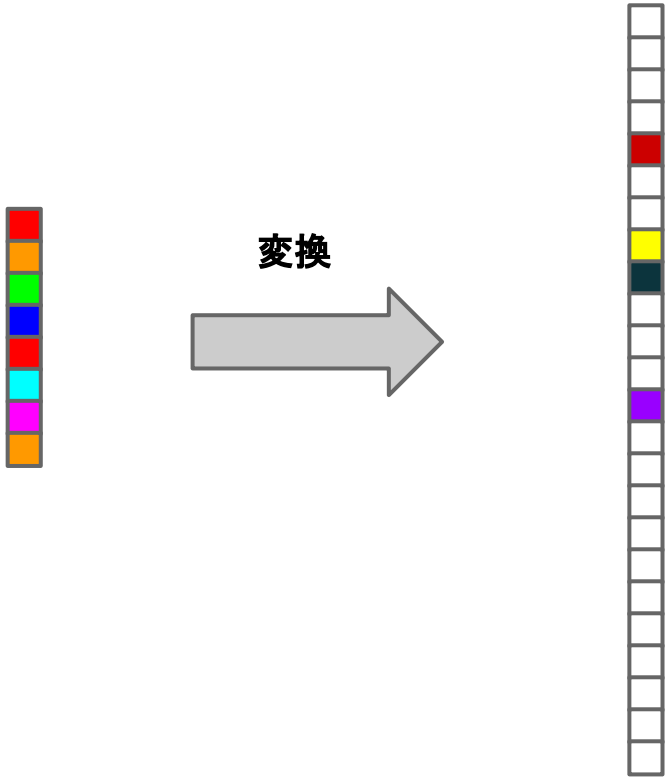
どんな人が書いたの？

- 第一著者：Manaal Faruqui@CMU
  - Noah Smithの研究室
  - 2015年だけで、ACL, NAACL, EMNLP に1stで通してる（ACLとNAACLは2本）
  - 最近はword vector representationが多い

本論文の概要

- word2vecなどで作られた**低次元密ベクトル**を、**高次元疎ベクトル**に変換する
- 既存タスクでの精度向上や、解釈しやすくなるなどのメリットが得られる

こういうイメージ



この論文を選んだ理由

- 高次元かつスパースな表現に魅力を感じた
  - 脳は明らかに高次元かつスパース
  - スパースならCPUで扱いやすい
- skip-gramの正則化は難しいが、本論文ではその目的を達成している

提案手法の概要

- 低次元ベクトルは既存手法で与えられる
- 目的関数を最適化することで高次元ベクトルを作り出す
  - 提案手法A：**高次元疎ベクトル**
  - 提案手法B：**高次元疎*バイナリ*ベクトル**

用字

X: 低次元密ベクトル, L×V行列

A: 高次元疎ベクトル, K×V行列

D: 変換(?)行列, L×K行列

x_i: Xのi列目の列ベクトル

提案手法A

- 目的関数は以下

$$\arg\min_{\mathbf{D},\mathbf{A}} \sum_{i=1}^{V} \|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2^2 + \lambda\|\mathbf{a}_i\|_1 + \tau\|\mathbf{D}\|_2^2$$

- 最適化はAdaGrad + RDA

提案手法B

- 目的関数は以下

$$\underset{\mathbf{D}\in\mathbb{R}_{\geq 0}^{L\times K},\mathbf{A}\in\mathbb{R}_{\geq 0}^{K\times V}}{\arg\min}\sum_{i=1}^{V}\|\mathbf{x}_i-\mathbf{D}\mathbf{a}_i\|_2^2+\lambda\|\mathbf{a}_i\|_1+\tau\|\mathbf{D}\|_2^2$$

- ただし、aの各要素は0 or 1
- 混合整数計画問題なのでそのまま解けない

提案手法Bの最適化

- 0,1の制約はとりあえず外す
- AdaGrad + RDAで最適化
  - 値が負になりそうな場合は0にクリップする
- v > 0の項はあとで全部1に丸める

実験概要

- 以下のタスクで性能を評価
  a. Word Similarity
  b. Sentiment Analysis
  c. Question Classification (TREC)
  d. 20 Newsgroup Dataset
  e. NP bracketing (NP)

※b,c,dでは単語ベクトルの平均を素性とした

次元数の設定

- K = {10L, 20L} で実験し、development setで性能が高い方を選んだ
  - 4種類中3種類が10Lになった
- % Sparseは91%～98%
  - 具体的な計算方法は不明

実験結果

Vectors		SimLex Corr.	Senti. Acc.	TREC Acc.	Sports Acc.	Comp. Acc.	Relig. Acc.	NP Acc.	Average
Glove	X	36.9	77.7	76.2	95.9	79.7	86.7	77.9	76.2
	A	38.9	**81.4**	**81.5**	**96.3**	**87.0**	**88.8**	**82.3**	**79.4**
	B	**39.7**	81.0	81.2	95.7	84.6	87.4	81.6	78.7
SG	X	**43.6**	81.5	77.8	97.1	80.2	85.9	80.1	78.0
	A	41.7	**82.7**	81.2	**98.2**	84.5	86.5	81.6	79.4
	B	42.8	81.6	**81.6**	95.2	**86.5**	**88.0**	**82.9**	**79.8**
GC	X	9.7	68.3	64.6	75.1	60.5	76.0	79.4	61.9
	A	12.0	73.3	77.6	77.0	68.3	**81.0**	**81.2**	67.2
	B	**18.7**	**73.6**	**79.2**	**79.7**	**70.5**	79.6	79.4	**68.6**
Multi	X	**28.7**	75.5	63.8	83.6	64.3	81.8	79.2	68.1
	A	28.1	**78.6**	79.2	93.9	78.2	84.5	81.1	74.8
	B	**28.7**	77.6	**82.0**	**94.7**	**81.4**	**85.6**	**81.9**	**75.9**

考察

- 多くのタスクで提案手法の性能がよい
  - 元のベクトルよりも性能がよい
- SimirarityとNP Bracketingでは、提案手法AとBはほぼ互角
  - これ以外の実験は個人的には信用しづらい…
  - 元の低次元ベクトルよりはほぼ確実によくなる

解釈性についての実験

実験概要：5つの単語を提示し、そこから仲間はずれの単語を人間に選んでもらう

- 提示する単語は以下のように決定する
  - 分散表現のi次元目の値が大きなものを4つ、値が小さなものを1つ

実験結果

Vectors	A1	A2	A3	Avg.	IAA	κ
X	61	53	56	57	70	0.40
A	71	70	72	71	77	0.45

- 元のベクトルよりも正解率が大幅に向上
- Inter Annotator Agreementも7ポイント向上

まとめ

- **低次元で密**な分散表現を**高次元で疎**な分散表現へと変換する方法を提案した
- 変換後の分散表現が良い性能を示すことをいくつかの実験で示した

感想・疑問

- 提案手法Aではデータ量はあまり減ってない
- 性能向上はそれほど大きくはないのでは
  - 大幅に上がってるのは、そもそも実験が適当だからでは？
- word intrusionの実験で結果が良くなっているというのは何を示唆しているのか？