文献紹介:

Non-distributional Word Vector Representations

発表者: 長岡技術科学大学 竹野 峻輔

※ 発表資料中の図・式・表などは論文中のものから引用しています。

論文概要

- Faruqui, M., & Dyer, C. (2015). Non-distributional Word Vector Representations. ACL-2015, 464–469. http://doi.org/10.3115/v1/P15-2076

- Data-driven representation learning for words is a technique of central importance in NLP. While indisputably useful as a source of features in downstream tasks, such vectors tend to consist of uninter-pretable components whose relationship to the categories of traditional lexical seman-tic theories is tenuous at best. We present a method for constructing interpretable word vectors from hand-crafted linguis-tic resources like WordNet, FrameNet etc. These vectors are binary (i.e, contain only 0 and 1) and are 99.9% sparse. We analyze their performance on state-of-the-art eval-uation methods for distributional models of word vectors and find they are competi-tive to standard distributional approaches.

- 人手データ(辞書からWordnetなどのオントロジーまで)で作った

  バイナリ表現で state-of-the-art と いい勝負するができるよ！という話.

導入

- Word Embeddingの流行
  - コーパスから教師なしで学習できる にも関わらず 意味的関係がある程度自動で取れる(v(king) - v(man) + v (man) = v(queen))
  - 文脈が言葉の意味を定義する：分布仮説
  - GloVe, word2vec etc …
- では人の作ったデータは意味ないのか→当たり前だがそんなことはない
  - 人の作ったデータには色んな知見が詰まっている！ (ただ作るのが大変なだけ…)

理論

いろいろな辞書を

Lexicon	Vocabulary	Features
WordNet	10,794	92,117
Supersense	71,836	54
FrameNet	9,462	4,221
Emotion	6,468	10
Connotation	76,134	12
Color	14,182	12
Part of Speech	35,606	20
Syn. & Ant.	35,693	75,972
Union	119,257	172,418

Table 1: Sizes of vocabualry and features induced from different linguistic resources.

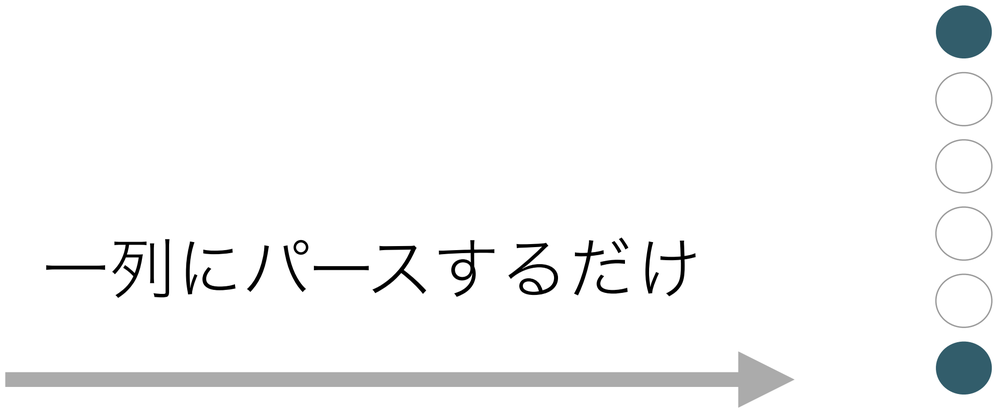
理論

いろいろな辞書を

Lexicon	Vocabulary	Features
WordNet	10,794	92,117
Supersense	71,836	54
FrameNet	9,462	4,221
Emotion	6,468	10
Connotation	76,134	12
Color	14,182	12
Part of Speech	35,606	20
Syn. & Ant.	35,693	75,972
Union	119,257	172,418

Table 1: Sizes of vocabualry and features induced from different linguistic resources.

一列にパースするだけ



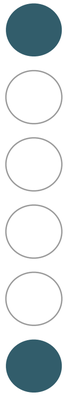
理論

いろいろな辞書を

Lexicon	Vocabulary	Features
WordNet	10,794	92,117
Supersense	71,836	54
FrameNet	9,462	4,221
Emotion	6,468	10
Connotation	76,134	12
Color	14,182	12
Part of Speech	35,606	20
Syn. & Ant.	35,693	75,972
Union	119,257	172,418

Table 1: Sizes of vocabualry and features induced from different linguistic resources.

一列にパースするだけ



学習用の入力に使う

理論

辞書のデータを0/1にパースするだけなので疎

平均 34 / 172,418(〜99.9%) × 119,257 (ave. 15 types)

- 疎行列用のデータ構造を使えば 内積の計算は効率的.

Word	POL.POS	COLOR.PINK	SS.NOUN.FEELING	PTB.VERB	ANTO.FAIR	···	CON.NOUN.POS
love	1	1	1	1	0		1
hate	0	0	1	1	0		0
ugly	0	0	0	0	1		0
beauty	1	1	0	0	0		1
refundable	0	0	0	0	0		1

Table 2: Some linguistic word vectors. 1 indicates presence and 0 indicates absence of a linguistic feature.

理論

Lexicon	Vocabulary	Features
WordNet	10,794	92,117
Supersense	71,836	54
FrameNet	9,462	4,221
Emotion	6,468	10
Connotation	76,134	12
Color	14,182	12
Part of Speech	35,606	20
Syn. & Ant.	35,693	75,972
Union	119,257	172,418

Table 1: Sizes of vocabualry and features induced from different linguistic resources.

WordNet: 解釈可能なsynset x POS,

上位語, 下位語, 同義語, 類義語, 反意語の0,1

Supersense :
(Ciaramita and Alutun 2006; Nastase, 2008; Tsvetkiv et al., 2014)
名詞・形容詞・動詞の意味分類

FrameNet:
語彙*品詞に対応するフレームを取り出すだけ

理論

Lexicon	Vocabulary	Features
WordNet	10,794	92,117
Supersense	71,836	54
FrameNet	9,462	4,221
Emotion	6,468	10
Connotation	76,134	12
Color	14,182	12
Part of Speech	35,606	20
Syn. & Ant.	35,693	75,972
Union	119,257	172,418

Table 1: Sizes of vocabualry and features induced from different linguistic resources.

Emotion & Sentiment:
(Mohammad and Turney 2013)
感情辞書(喜怒哀楽恐 etc … 8種),

感情極性(正・負)

Connotation: (Feng et al. 2013)
極性(負・正・中立) x 品詞.

暗示的な表現, 網羅性が高い

Color : 色辞典(uncertainly -> gray, blood -> 赤)

理論

Lexicon	Vocabulary	Features
WordNet	10,794	92,117
Supersense	71,836	54
FrameNet	9,462	4,221
Emotion	6,468	10
Connotation	76,134	12
Color	14,182	12
Part of Speech	35,606	20
Syn. & Ant.	35,693	75,972
Union	119,257	172,418

Table 1: Sizes of vocabualry and features induced from different linguistic resources.

Part of Speech Tags: PTBのPoS

Syn. & Ant. 類義語・反意語

(WordNetのものとは違って)

単語-単語の関係を表す

実験設定

L2正則化 対数エントロピー法で基本 学習.

入力 : skip-gram, GloVe, LSA　v.s. そのまま疎行列, SVD, そのまま+SVD

- Word Similarity (WS-353, RG-65, SimLex) :
  単語のpairを当てる or スペルマンの順位相関係数 で評価
- Sentiment Analysis (Socher et al. 2013) :
  木構造のpositive-negative判定タスク
- NP-Bracketing (Lazaridou et al. 2013)
  e.g. local (phone company), blood pressure machine

実験結果

- WordSimilarityタスクは そのまま の疎行列を使うだけでも かなりの強い
  ただ 300 billion 単語のSkip-gramもなかなかに強い.
  NP Bracketing に対して 疎行列が強い(Why? 品詞の情報？)

Vector	Length (D)	Params.	Corpus Size	WS-353	RG-65	SimLex	Senti	NP
Skip-Gram	300	$D \times N$	300 billion	65.6	72.8	43.6	**81.5**	80.1
Glove	300	$D \times N$	6 billion	60.5	76.6	36.9	77.7	77.9
LSA	300	$D \times N$	1 billion	**67.3**	77.0	49.6	81.1	79.7
Ling Sparse	172,418	–	–	44.6	**77.8**	56.6	79.4	**83.3**
Ling Dense	300	$D \times N$	–	45.4	67.0	**57.8**	75.4	76.2
Skip-Gram $\oplus$ Ling Sparse	172,718	–	–	67.1	80.5	55.5	82.4	82.8

Table 3: Performance of different type of word vectors on evaluation tasks reported by Spearman's correlation (first 3 columns) and Accuracy (last 2 columns). Bold shows the best performance for a task.

まとめ

- 分散表現の研究が盛んであるが，辞書的資源も捨てたもんじゃない
  - そこまで高度な辞書を使ってはいない：多くの言語で実現可能
  - ただの連結であるので, 組み合わせ方でより良い使い方ができるだろう
- 辞書を使ったベクトルの良いところ：人が解釈しやすい
- 分散表現はそういう動きがある、が 解釈しやすいと言えない