文書
の
ベクトル表現

テキスト処理を行うには

- 処理の対象である
  - 単語
  - 文
  - 文書

をうまく表現する必要がある

単語を表現するには

- 第一案: 類義語やオントロジーを用いて単語間の関係で表現する。欠点は
  - （近年大分と自動化されたとはいえ）手で作成
  - 文脈依存部分の表現が困難
- 第二案: 単語の共起を用いる. 欠点は
  - 共起を表すのに大量のメモリが必要
  - データも必要。文脈が考慮できない
- 第三案: one-hot-vector. 欠点は
  - 新語に対応できない
  - 単語間の関係が記述できない

第4案: 低次元ベクトル

- これが、今まで説明してきた SVD, LSI/LSA, pLSI/pLSA, LDA である.
- 単語を他との関係（それが出現する文書との関係）を用いて表現する
  - 「互いに」であることに注意！
  - 単語 vs. 文書 だけではない

補足: John Rupert Firth

"You shall know a word by the company it keeps" -1957

- 英国の言語学者
- （表層的に言えば）単語の意味は、共起する単語からわかる

recursive です

From Wikipedia

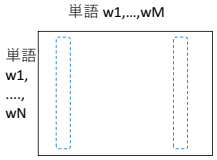
SVD 等の問題点

- 計算コスト
- 単語の意味が表現されているか？
  - LDAに至ってはずいぶんと良くなったが、あまりよくはない
- （作成後）新たな単語が来ると困る

単語の文脈を用いる word2vec

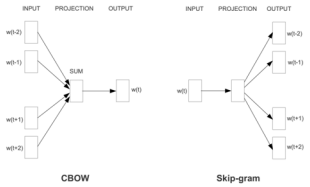
- アイデア: ある単語の文脈（前後の単語集合）を予測させよう。
  - その予測器が、単語の表現となろう

単語・単語行列でよいかな？



いや、予測器を作ろう

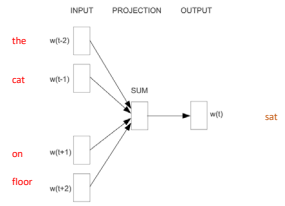
- 2 個のNNモデル:
  - CBOW (Continuous Bag of Word): 「窓」中の単語で、真ん中の単語を予測する
  - Skip-gram: 単語で、その周囲の単語を予測する

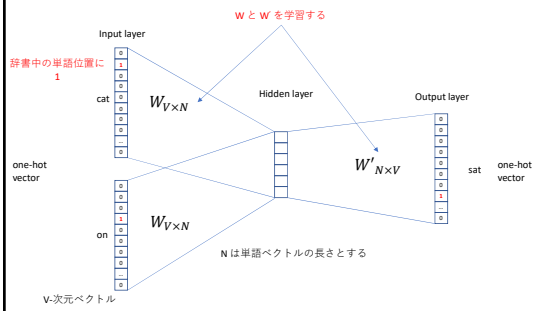


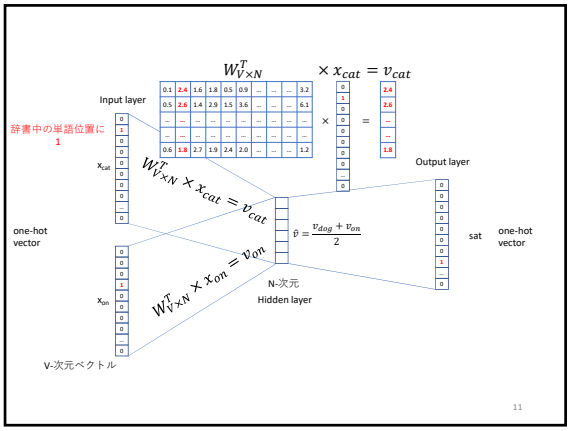
Mikolov et al. (2013)
Efficient Estimation of Word Representations in Vector Space
https://arxiv.org/pdf/1301.3781v3.pdf

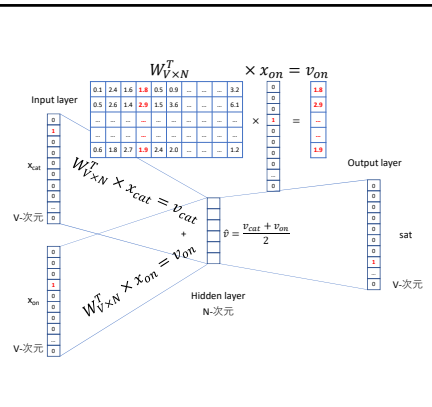
word2vec - CBOW

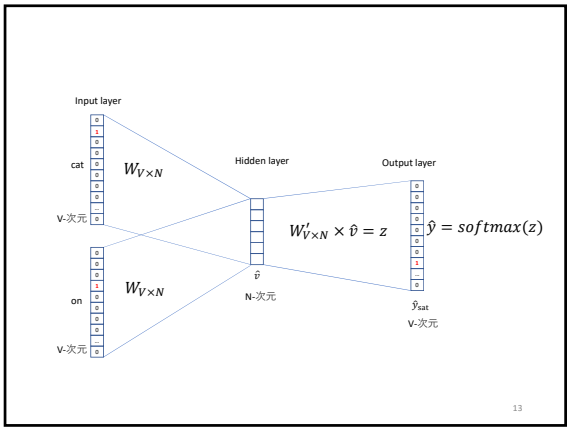
- 例: “The cat sat on floor”
  - 窓サイズ = 2



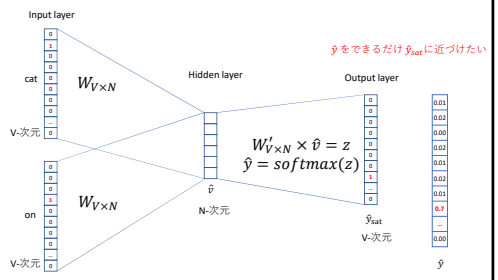




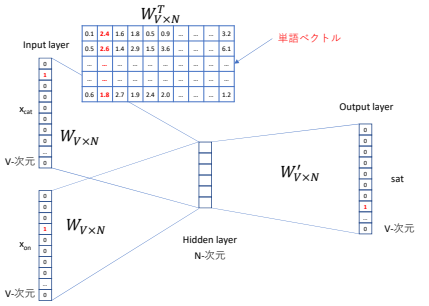




13



14



W と W′ のどちらも単語ベクトルとすることができる．
平均をとってもよい．

15

非常に面白い性質がある

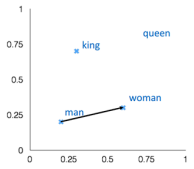
Mikolov et al. (2014) が示した線形関係

a:b :: c:?

$$d = \arg\max_x \frac{(w_b - w_a + w_c)^T w_x}{||w_b - w_a + w_c||}$$

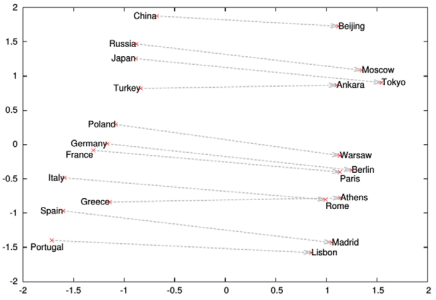
man:woman :: king:?

+	king	[0.30 0.70]
-	man	[0.20 0.20]
+	woman	[0.60 0.30]
	queen	[0.70 0.80]



16

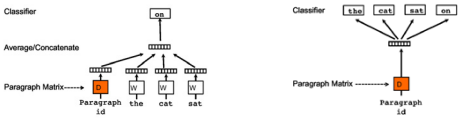
単語間の同一関係のマップ



17

文や文書はどう表現するか

- 段落(paragraph) vector (Quoc Le & Mikolov, 2014)
  - word2vec の単純な拡張
  - 同様に2モデルあり



18

応用

- 検索
- 感情分析（Sentiment analysis）
- 分類

19