単語の分散表現

導入:Questions

Q1,自然言語処理(言語解析)とは？

人間が普段扱う言語(日本語、英語等)を機械が理解できる言語で処理・解析すること

Q2,文書をそのまま処理できるの？

いきなりは無理。文書は文の構造体であるから、文と文の関係、文自体の構造、単語の接続関係、単語の意味、単語分割等ステップを踏まないと機械が扱えるデータにはならない。

品詞分解

吾輩は / 猫で / ある

吾輩、猫	名詞
は、で	助詞
ある	動詞

導入:Questions

Q3,何ができそう？
機械の処理向上や言語学の知識の蓄積により、研究は進んでいる。

例としてモデル性能評価指標
（GLUEデータセット）
右図のタスクについては今後
できそうなことがまとまっている。

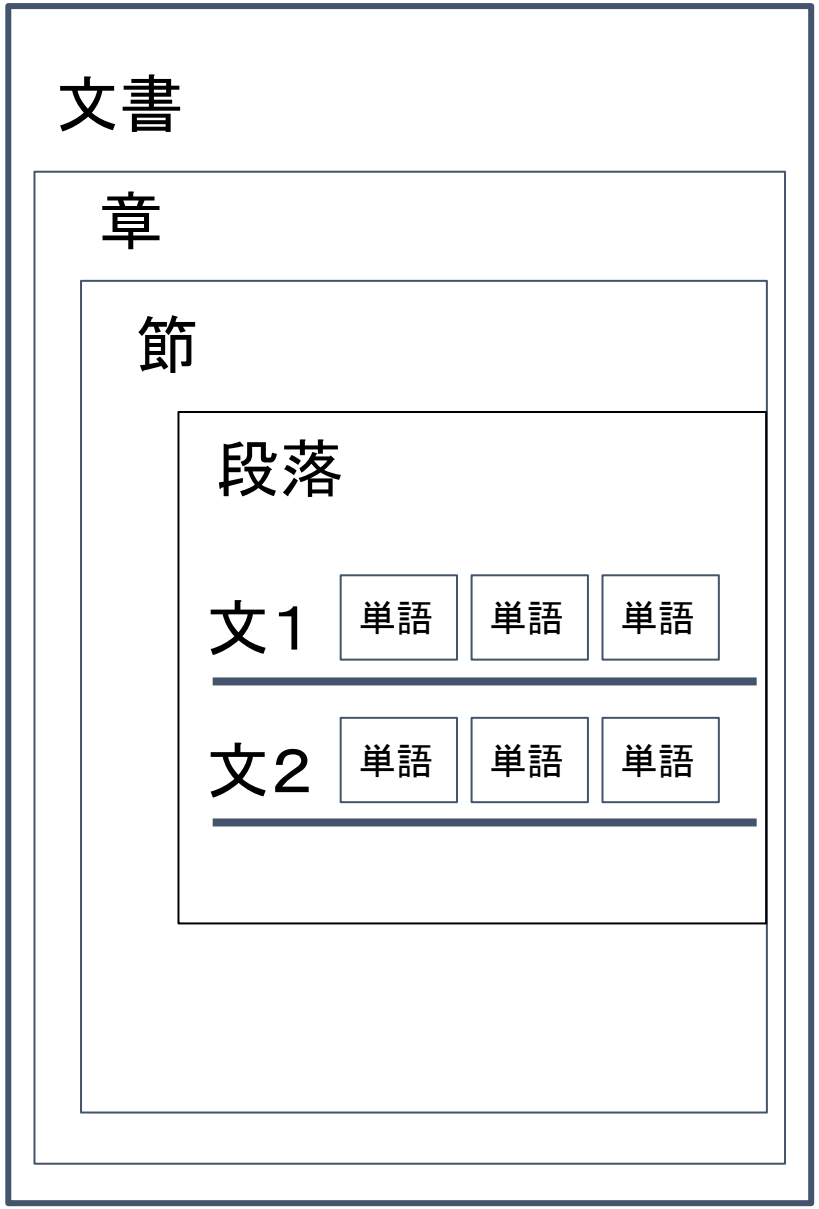
タスク	http://deeplearning.hatenablog.com/entry/menhera_chan	概要
GLUE		8種の言語理解タスク
1. MNLI		2入力文の含意/矛盾/中立を判定
2. QQP		2質問文が意味的に等価か判定
3. QNLI		SQuADの改変．陳述文が質問文の解答を含むか判定
4. SST-2		映画レビューの入力文のネガポジを判定
5. CoLA		入力文が言語的に正しいか判定
6. STS-B		ニュース見出しの2入力文の意味的類似性をスコア付け
7. MRPC		ニュース記事の2入力文の意味的等価性を判定
8. RTE		2入力文の含意を判定
SQuAD		質疑応答タスク．陳述文から質問文の解答を抽出
CoNLL		固有表現抽出タスク．単語に人物/組織/位置のタグ付け
SWAG		入力文に後続する文を4つの候補文から選択

導入:Questions

Q4,Q3のようなことをやるための第一歩は？
文書を解体して、その最小要素をまずは機械に理解してもらうこと。
文書>章>節>段落>文>単語

単語を機械が理解できる形式に変換
->ベクトル化（分散表現）、構造化

分散表現の例：
色（桔梗色）->色（RGB|R85 G85 B153）



このあとの流れ

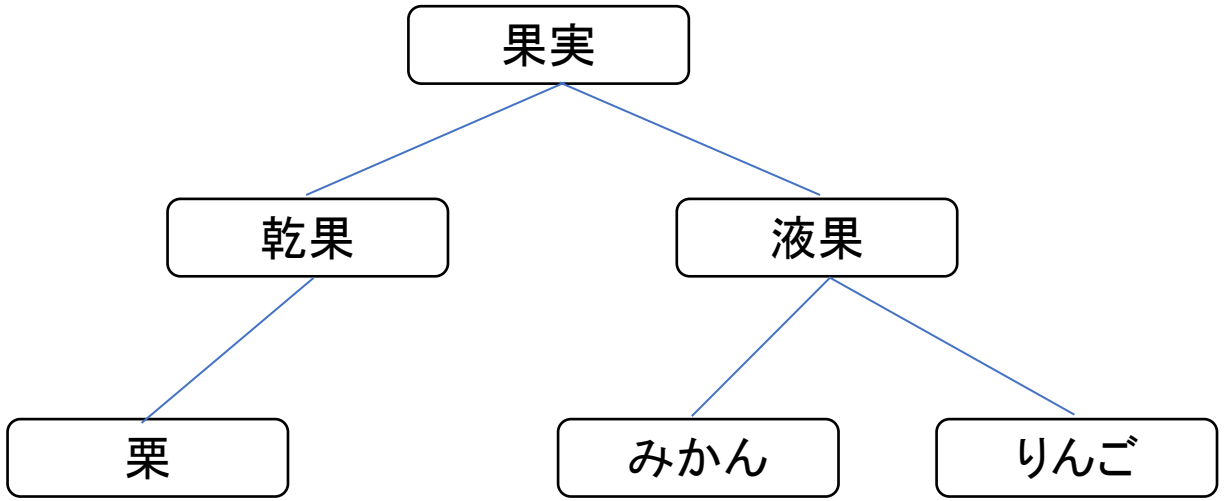
- 単語分散表現の手法について三種紹介
  - シソーラスベース
  - カウントベース
  - 推論ベース

- 単語の類似度判定基準の大別
- 分散表現の使いどころ

表現手法：構造化手法（シソーラスベース）

- 体系化された単語辞書を作成し、これをもとに単語の類語判定する
  - 利点：既存の言語構造の反映/類語検索が楽
  - 欠点：新語反映・そもそも作成が大変/ニュアンス（文脈依存な意味変化）を捉えられない

利点例:学術構造の反映



欠点例:類義でも些細な違いのもの

- 草　->植物の意味 or 笑いの表現
- 死んだ　->俺死んだわ～（失敗の意味）
  or 生命活動の停止の意味

表現手法：統計的手法（カウントベース）

- 分布仮説に基づき、単語をベクトル化、単語間の類似度を計算し類語判定する
  - 分布仮説：単語の意味は周囲の単語によって形成される
  - 利点：文章における単語の頻度から計算可能
  - 欠点：大規模な文章に対して計算コストが高い/頻度は文章に依存する

利点例：文脈から類語判定
I drink beer. We drink beer -> I guzzle beer. We guzzle beer
-> beer ≈ guzzle

欠点例：時間経過による意味の変化
すばらしい地震災害 ->否定的な意味
すばらしい人 ->肯定的な意味

分布仮説 You say goodbye and I say hello.

say は Youとgoodbye（コンテキスト）から
意味が形成される

表現手法：統計的手法（カウントベース）

・分布仮説に基づき、単語をベクトル化、単語間の類似度を計算し類語判定する

- 分布仮説：単語の意味は周囲の単語によって形成される
- 利点：文章における単語の頻度から計算可能
- 欠点：大規模な文章に対して計算コストが高い/頻度は文章に依存する

分布仮説に基づき、ベクトル化
Text="You say goodbye and I say hello."
=[you,say,goodbye,,and,i,say,hello,.]
->0,1で表現（OneHot表現）
->各単語のOneHot表現を行列にする

共起行列 ->

	you	say	goodbye	and	i	hello	.
you	0	1	0	0	0	0	0
say	1	0	1	0	1	1	0
goodbye	0	1	0	1	0	0	0
and	0	0	1	0	1	0	0
i	0	1	0	1	0	0	
hello	0	1	0	0	0	0	1
.	0	0	0	0	0	1	0

表現手法：統計的手法（カウントベース）

- 分布仮説に基づき、単語をベクトル化、単語間の類似度を計算し類語判定する
  - 分布仮説：単語の意味は周囲の単語によって形成される
  - 利点：文章における単語の頻度から計算可能
  - 欠点：大規模な文章に対して計算コストが高い/頻度は文章に依存する

共起行列を元に類似度判定
->コサイン類似度
->同じ方向を向いていたら1（類義）、反向を向いていたら-1（反義）

$$cos\_similar = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}||||\mathbf{y}||} \frac{x_1 y_1 + x_2 y_2 + \cdots x_n y_n}{\sqrt{x_1^2 + x_2^2 \cdots x_n^2}\sqrt{y_1^2 + y_2^2 \cdots y_n^2}}$$

表現手法：NeuralNetwork（推論ベース）

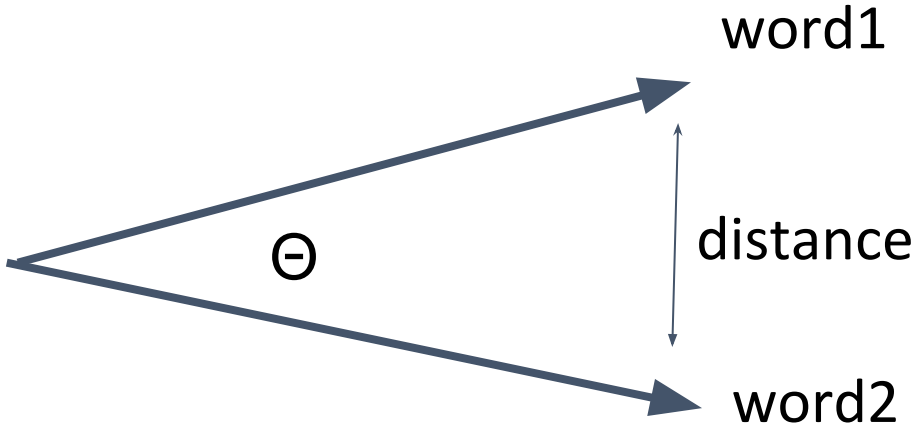
・単語をOneHot表現、文中の単語の出現確率をNNで推論、類語判定する
- 分布仮説：単語の意味は周囲の単語によって形成される
- 利点：大規模な文章に対応可能
- 欠点：ハイパーパラメータが多い（層数、ニューロン数、学習率とか）

Word2vec

・CBOWモデル:You ? goodbye and I say hello.
?に何が入るかをYouとgoodbye（コンテキスト）から推論

・Skip-gramモデル：? say ? and I say hello.
?に何が入るかをsay（コンテキスト）から推論

単語の類似度判定



類似度の判定->距離 or 方向を基準にする

シソーラス（距離）　->構造化したIDで距離測定
カウント（ベクトル）->単語共起行列の潜在意味解析
推論（ベクトル）　->単語ベクトルを入力としたNNで出現確率の推論

カウント：文の単語分布を捉える（統計情報の活用）が、単語の類推は弱い
推論：分布を捉えるのは弱いが（単語ベクトルを入力とする）が、単語の類推は強い
word2vecは類推問題をベクトルの加減算で解くためである。

両方に強いモデルを作る->Global Vectorモデル(GloVe)

単語の分散表現の使いどころ

- メールやツイートの感情分析
- 感情分析を元にして文書分類（アプリに不満を持つ意見を優先表示）
- 大規模コーパス(Wikipedia,GoogleNews)を元に転移学習

単語の分散表現の評価方法

評価指標：類似性、類推問題->単語類似度の評価セットを利用

Model：使用したモデル
Dim：作成した層数
Size：語彙数
Semantics：単語意味類推問題の正答率
->king:queen=man:women
Syntax：単語の形態情報を問う問題
->bad:worst=good:best

→モデルと語彙数の兼ね合いで精度が変化
→単語ベクトルの次元数は適度なサイズが良い

NER (Tjong Kim Sang and De Meulder,2003)
GloVeより抜粋

Model	Dim.	Size	Sem.	Syn.	Tot.
ivLBL	100	1.5B	55.9	50.1	53.2
HPCA	100	1.6B	4.2	16.4	10.8
GloVe	100	1.6B	67.5	54.3	60.3
SG	300	1B	61	61	61
CBOW	300	1.6B	16.1	52.6	36.1
vLBL	300	1.5B	54.2	64.8	60.0
ivLBL	300	1.5B	65.2	63.0	64.0
GloVe	300	1.6B	80.8	61.5	70.3
SVD	300	6B	6.3	8.1	7.3
SVD-S	300	6B	36.7	46.6	42.1
SVD-L	300	6B	56.6	63.0	60.1
CBOW†	300	6B	63.6	67.4	65.7
SG†	300	6B	73.0	66.0	69.1
GloVe	300	6B	77.4	67.0	71.7
CBOW	1000	6B	57.3	68.9	63.7
SG	1000	6B	66.1	65.1	65.6
SVD-L	300	42B	38.4	58.2	49.2
GloVe	300	42B	**81.9**	**69.3**	**75.0**

参考文献

・ゼロから作るDeepLearning2/斎藤康毅/オライリージャパン

・メンヘラちゃんと学ぶDeepLearning最新論文
http://deeplearning.hatenablog.com/entry/menhera_chan

・GloVe: Global Vectors forWord Representation/Jeffrey Pennington et.al
/Computer Science Department, Stanford University, Stanford, CA 94305