

論文紹介
潜在意味分析、Predication
algorithmとメタファ理解

中本 敬子
kenakamoto@nifty.com

1

今回発表の流れ

- 潜在意味分析の紹介(Landauer, Foltz, & Laham, 1998)
- Predication Algorithmの紹介(Kintsch, 1998)
- メタファ理解への応用の紹介(Kintsch, 2000)
- 感想

2

潜在意味分析(LSA)とは？

- 大規模なテキストコーパスから単語の意味を意味空間上のベクトルとして表現するための数学的・統計的技法
- 大規模なテキストコーパスを扱い、今までの(心理学の)感覚からすると高次元での情報圧縮を行い、様々な行動データの予測を行うのに成功したところに意義がある(と思う)。

3

基本コンセプト

- 類似した意味を持つ単語は、類似した文脈の中に現れる。
– 昨日の夕飯のデザートはリンゴとオレンジだった。
– 昨日の晩ご飯のデザートは桃とオレンジだった。
- あるテキストにおける単語の出現 / 非出現を調べ、そのパターンを分析すれば、単語の使用状況から、単語の意味を特定できるのではないか？

4

LSAの概要

- テキスト・データ(意味あるまとまりを持つ文あるいは文章)から、行列(単語の出現頻度: バスケット・データ)を作成。
- 値を適当に変換(対数変換とエントロピーを使う)。
- 特異値分解する。
- 上位幾つかの特異値だけを使って、行列を復元
- 語(テキスト)間の意味関係をベクトル間のコサインで表す。

5

入力データ

- 行を単語、列をテキスト要素を単語の頻度とする行列。通常、LSAでは、60000 × 30000等の巨大な行列になる。

```
computer use data files of some industrial firms
c1: Most studies involve the EPC computer application
c2: A survey of use patterns of computer control systems
c3: The EPC use nuclear management system
c4: Some studies focus engineering technology of EPC
c5: Related use pattern of system file transfer software
...
m1: The generation of random binary ordered rows
m2: The generation of random binary ordered rows
m3: Some users try to find out how to use the system
m4: Some users try to find out how to use the system
```

- A small example

$[X] =$

| | c1 | c2 | c3 | c4 | c5 | m1 | m2 | m3 | m4 |
|------------|----|----|----|----|----|----|----|----|----|
| computer | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| use | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| data | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| files | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| of | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| some | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| industrial | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| firms | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

6

特異値分解 (Singular Value Decomposition)

- 因子分析と類似した手法
- 長方形行列を3つの行列に分解: 1. 元の行列の列の要素を直行するベクトルの値で表現したもの、2. 元の行列の列の要素を同じように表現したもの、3) 重み付けの値が対角要素に入った対角行列。

7

特異値分解 - 例

$$[X] = [U][\Sigma][V^T]$$

$$[U] =$$

| | | | | | | | | | |
|------|------|------|------|------|------|------|------|------|------|
| 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |

$$[\Sigma] =$$

| | | | | | | | | | |
|------|------|------|------|------|------|------|------|------|------|
| 2.88 | 2.39 | 1.64 | 1.33 | 0.91 | 0.67 | 0.36 | 0.36 | 0.36 | 0.36 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

$$[V^T] =$$

| | | | | | | | | | |
|------|------|------|------|------|------|------|------|------|------|
| 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |

Figure 1. Example SVD of matrix in Figure 0.

8

次元の縮約

- 上位n次元のみを用いて、行列を復元する(通常は、300-500次元)。
- 次元数は経験的に決められている。

$$[X] =$$

| | u1 | u2 | u3 | u4 | u5 | u6 | u7 | u8 | u9 |
|----------|------|------|------|------|------|------|------|------|------|
| human | 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| robotics | 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| computer | 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| minor | 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| science | 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| language | 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| film | 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| EPS | 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| activity | 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| books | 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| graph | 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| domains | 0.18 | 0.25 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |

11

次元の縮約 - 効用

- 直接経験したことのない単語の共起関係でも、予測できる。
- 縮約することで、元データでは表せなかった類似性が表現できる。
 - 元データ $r(\text{human}, \text{user}) = -.38$ $r(\text{human}, \text{minors}) = -.29$
 - LSA $r(\text{human}, \text{user}) = .94$ $r(\text{human}, \text{minors}) = -.83$

10

LSAで出来ること

- 発達(語彙獲得)
 - 語彙テストにおける同義語判断の予測
 - 単語分類課題(boy, girl, horse, flower)の予測
- 言語理解(読み)
 - テキストの一貫性と理解の予測
 - 語彙判断におけるプライミング効果の予測
 - テキストの主観評定(小論文の成績等)の予測
 - 脳機能障害における様々な言語障害の予測
- クエリーとドキュメントの類似性判断

11

LSAの心理学的意義

- プラトン問題への一解答
- 言語コミュニケーション成立に対する説明



12

Predication Algorithm

- Construction-integration model (Kintsch, 1998)をLSA上で実現したアルゴリズム
- 語の多義性を解決する方法
 - My horse *ran*.
 - The machine *ran*.
 - The color *ran*.
 - Runの意味は、それぞれで異なる。

13

Predication Algorithm概要

- 文(語句) Predicate (Argument) 例) horse runs.
- Predicate Pのベクトル, Argument Aのベクトルを特定.
- Pの近傍ベクトルI(m個)抽出. 近傍はcosineで定義.
- P, A, Iをユニットとするネットワークを構成
 - P及びAからIへの活性リンク(Cosineに対応する重み)
 - I同士の間で抑制リンク(小さな負の重み)
 - AとPの間に活性リンク(cosineに対応する重み)
- ネットワークが安定するまで走らせる.
- Iの内, 最も強く活性化されたk個のベクトルを採用
- P, Aおよびk個のベクトルの合成(centroid)として, P(A)の意味ベクトルを計算

14

Predication Algorithm 図解

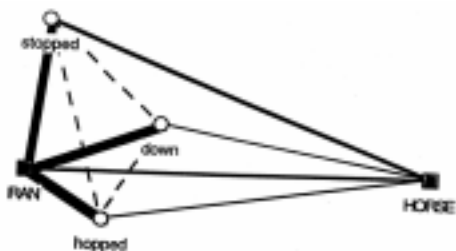


Fig. 1. A fragment of an integration network. Moves are the predicate ran and the argument down, and these items form the neighborhood of ran. Solid lines indicate positive connections among nodes, dashed lines indicate arbitrary connections.

15

Predication Algorithm 結果の例1

Table 1
The nearest neighborhood of ran, rank orders and activation values for horse and color

| neighborhood | nearest neighbor | activation value | nearest neighbor | activation value |
|--------------|------------------|------------------|------------------|------------------|
| off-ran | horse | 0.895 | color | 0.000 |
| ran | -0.28 | 0.895 | color | 0.220 |
| jumped | -0.17 | 0.699 | color | 0.000 |
| rolled | -0.09 | 0.609 | color | 0.000 |
| stopped | -0.28 | 0.500 | color | 0.000 |
| ran | -0.16 | 0.609 | color | 0.200 |
| downed | -0.16 | 0.609 | color | 0.180 |
| down | -0.17 | 0.609 | color | 0.000 |
| tried | -0.16 | 0.609 | color | 0.000 |
| tried | -0.16 | 0.609 | color | 0.000 |
| quitted | -0.14 | 0.609 | color | 0.000 |
| ran | -0.19 | 0.200 | color | 0.000 |
| recovered | -0.11 | 0.609 | color | 0.000 |
| threw | -0.11 | 0.609 | color | 0.200 |
| worked | -0.17 | 0.609 | color | 0.200 |
| with | -0.26 | 0.177 | color | 0.000 |
| color | -0.28 | 0.476 | color | 0.200 |
| color | -0.19 | 0.200 | color | 0.000 |
| rolled | -0.12 | 0.609 | color | 0.200 |
| down | -0.16 | 0.200 | color | 0.200 |
| hopped | -0.12 | 0.609 | color | 0.000 |

16

Predication Algorithm 結果の例2

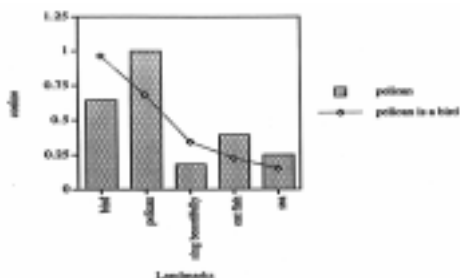


FIG. 2. Cosine between the vectors for police and police is a bird and five landmarks.

17

LSA + Predication Algorithmで説明できること

- メタファ理解
 - The student dropped the glass. (.76)
 - The glass was broken. (.66)
- 因果推論
 - The student dropped the glass. (.76)
 - The glass was broken. (.66)
- 類似性判断

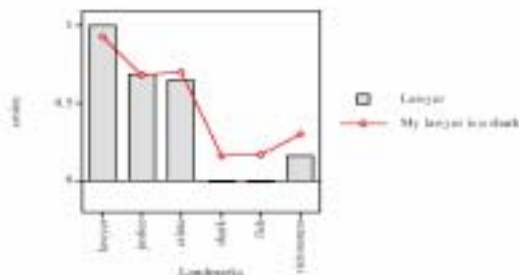
18

メタファへの応用

- Predication Algorithmを用いれば、メタファも字義通り文と同じように意味を計算することができる。
 - メタファと字義通り文に質的区別をおかないモデル(たとえば、類包含モデル; Glucksberg & Keysar, 1990)の見解と一致。
- ただし・・・
 - 字義通り文では、 $m(P\text{の近傍の数})=20$ 程度でよいのに対し、
 - メタファでは、 $m=500$ 等たくさん必要 = メタファの難しさ。

19

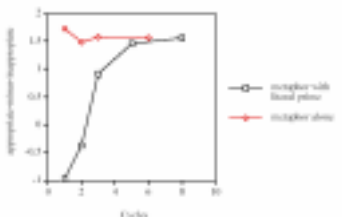
メタファへの適用例 My lawyer is a shark.



20

メタファ理解における 意味の活性化と抑制 (1)

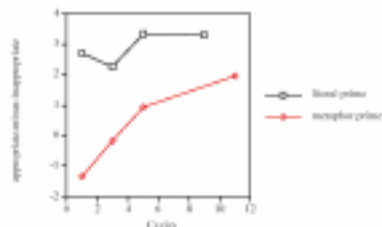
- 比喩に無関連な意味特徴が活性化されていると、比喩の読みは遅くなる (MacGlone & Manfredi, 2001).
 - Sharks can swim. My lawyer is a shark. Slow
 - My lawyer is a shark. Fast



21

メタファ理解における 意味の活性化と抑制 (2)

- 比喩文を呈示すると、その喩辞に関連した字義通り文の読みは遅くなる (Gernsbacher et al., 2001).
 - My lawyer is a shark. sharks are good swimmers. Slow
 - The hammerhead is a shark. sharks are good swimmers. Fast



22

Kintschモデルによる メタファの説明の特徴

- メタファと字義通り文の連続性を、明示的に説明している。
- TopicとVehicleの役割が、初期の処理過程から異なる。
- 定量的な予測を行っているため、検証がしやすい (Kintschらの言い方では、Computational Theory)。

23

感想 (1)

- LSAについて
 - ずいぶん、おおざっぱなやり方で作った意味の表現で色々な行動データが説明できるものだ。
 - 日本語で同様のことをやろうとすると大変だろうなあ・・・(分かち書き、表記違いの問題とか)
 - と書いていたら、すでに工学の分野では、日本語への応用がかなり進められているようです。
- Predicationについて
 - 今のところ、Predicate、Argumentの設定は人力でコーディングされている。モデルを詰めていくためには統語的知識を組み込んでいく必要があるのだろう。
 - 行動データの予測に失敗したときに、Predication Algorithmに問題があるのか、LSAに問題があるのか、直感的にしが決まらないらしい。

24

感想(2)

- メタファへの応用について
 - Glucksbergらの研究を後押しする初めての動くモデルか？
 - こういうモデルで説明されるメタファ(理解)と構造写像理論(Gentner, 1983等)で説明されるメタファ(理解)とは、やはり分けていく必要があるだろう。
 - メタファの種類分け(梶見先生との共同研究で着手)と下位プロセスの特定(ナカモトの野望。なかなか進みません)が、実験心理学者としての仕事か？

25

References

- Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*, **7**, 257-266.
- Kintsch, W. (2001). Predication. *Cognitive Science*, **25**, 173-202.
- Landauer, T. K., Foltz, P. W., & Lajam, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, **25**, 259-284.
- LSAと言語獲得は
 - Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, **104**, 211-240.
- 論文のダウンロードとLSAのオンラインデモは、
 - <http://lsa.colorado.edu/>

26