2015年8月3日 ERATO感謝祭Season II@国立情報学研究所

Learning word representations from relational graphs (AAAI'15) and Embedding semantic relations into word representations (IJCAI'15)

前原 貴憲（静岡大学）

元となる論文

- Danushka Bollegala, Takanori Maehara, Yuichi Yoshida and Ken-Ichi Kawarabayashi (2015): "**Learning word representations from relational graphs**", in Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI'15), January 25th–29th, 2015, Austin, Texas, United States, pp. 2146–2152.

- Danushka Bollegala, Takanori Maehara and Ken-ichi Kawarabayashi (2015): "**Embedding semantic relations into word representations**", in Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI'15), Buenos Aires, Argentina, July 25th–31th, pp. 1222–1228.

単語のベクトル表現

自然言語処理・単語のベクトル表現

自然言語処理：自然言語（英語, 日本語,...）を計算機で扱う
（例：機械翻訳・文章要約・音声認識，...）

多くの機械学習アルゴリズムはベクトルを入力とする
$\Rightarrow$ 単語をベクトルで表す手段が必要

- bag-of-words**表現**
- n-gram**表現**
- **行列分解型の表現**
- **共起数え上げ型の表現**
- **予測モデル型の表現**
- ...

bag-of-words表現

i番目の単語 $\mapsto$ i番目の単位ベクトル

- **高次元・疎ベクトル**
- **キーワードが効くタスクに有効**
  例：SPAM**メール分類** (discount, for sale, we regret, ...)

n-gram 表現

単語 $\mapsto$ 単語に含まれる n-gram 全体

例： apple $\mapsto$ {ap,pp,pl,le} $(n = 2)$

- **高次元・疎ベクトル**
- **字面が似た単語を同一視したい場合に有効**
  例：専門用語の分類・意味解釈（xxxholic $=$ xxx 依存症）

分布仮説 (distributional hypothesis)

bag-of-words/n-gram **表現は単語だけから定まる表現**

分布仮説 [Harris 1954; Firth 1957]：

似た意味の単語は似た文脈で出現する

$\Rightarrow$ **文脈データからベクトル表現を構成** (data-driven)

- **文脈の情報は収集可能**
  · · · 機械学習的・言語に対する知見をあまり必要としない
- **データが多いほど高品質**
  · · · 最近のビッグデータの流行とマッチ

行列分解型の表現

「文脈 $=$ どの文章に出現するか」

単語wが文章dに出現する確率$P(w|d)$を

$$P(w|d) = \sum_h \underline{P(w|h)} P(h|d)$$

などと分解（非負行列分解・特異値分解）

$$\text{単語} w \mapsto P(w|h)\ (h\text{次元ベクトル})$$

- **低次元・密ベクトル**
- **文書・単語分類系のタスクでよく使われる**
  例：ニュース記事の多クラス分類

共起数え上げに基づく表現

「文脈 = 周辺に共起する単語」

単語 $\mapsto$ それと共起する単語集合

例： “I have bread and butter for breakfast”

$\Rightarrow$ butter = [I, have, bread, and, for, breakfast]

- **高次元・疎ベクトル**
- SVD **などで次元圧縮することも多い**
- **汎用的・多くのタスクで高精度**

予測モデル型の表現

「文脈 $=$ 周辺に共起する単語」

アイデア：文章中の単語を隠してその単語を当てる

単語を周辺の単語のベクトルで予測するモデル
例： $p(i|v(w_{i\pm1}))$ **(ロジスティック回帰)**

このモデルがデータと最もあうベクトルを計算

- **低次元・密ベクトル**
- **現在活発に研究されている（モデルの自由度が高い）**
- **汎用的・多くのタスクで高精度**

予測モデル型流行の主要因：差による意味表現

$$\boldsymbol{v}(\text{king}) - \boldsymbol{v}(\text{man}) \approx \boldsymbol{v}(\text{queen}) - \boldsymbol{v}(\text{woman})$$

$\Rightarrow$ 連想ゲームが解ける！

応用が広いので，多くの人が食いついた

- tokyo - japan $\approx$ italy - paris (capital)
- france - french $\approx$ mexico - spanish (language)
- use - used $\approx$ go - went (past)
- car - cars $\approx$ city - cities (plural)
- ...

本研究のモチベーション
「単純共起」から「文脈共起」へ

単純共起から文脈共起へ

従来手法：共起は全て同等に扱う

- **Ostrich** is a large **bird**
- **Sparrow** is a small **bird**
- **Cat** chases a **bird**

⇒ 「どういう形で共起したか」の情報量は大きい

共起の形を陽に取り入れることで精度向上させたい

- **同じデータ量でより良い品質**
- **少ないデータ量で同等の品質**

提案手法の概要

- Learning word representations from relational graphs (AAAI'15)
  従来手法：単語単語共起（行列）
  提案手法：単語単語文脈共起（テンソル）

- Embedding semantic relations into word representations (IJCAI'15)
  従来手法：単語単語共起
  提案手法：単語単語共起 s.t. **文脈共起制約**

Learning word representations from relational graphs (AAAI'15)

共起グラフ

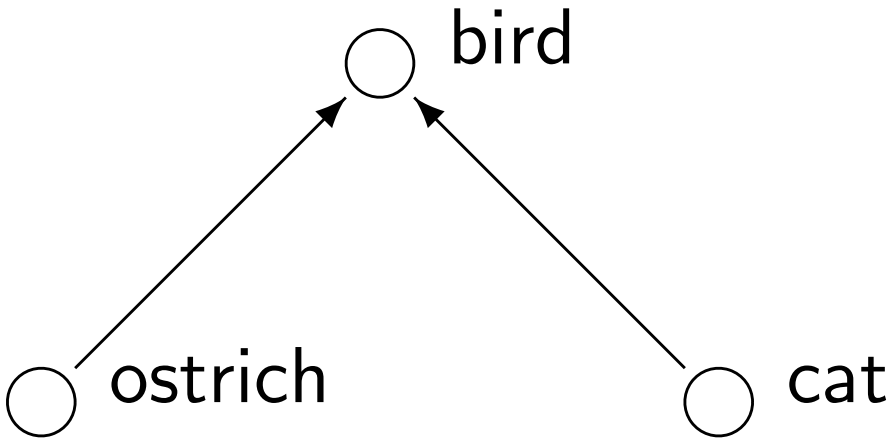
$G = (V, E)$ with $w : E \to \mathbb{R}$

where

V：単語全体

E：共起関係を表現する枝

$w(e)$：共起関係eの強さ



提案手法：関係グラフ

$G = (V, E)$ with $w : E \to \mathbb{R}$, $l : E \to L$

where

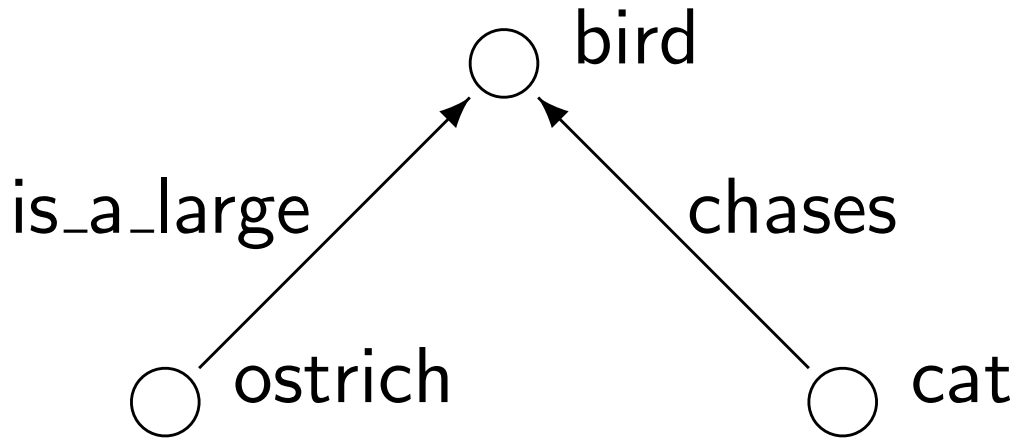
V：単語全体

E：共起関係を表現する枝

L：共起のラベル（例：is_a_large**）**

$w(e)$：共起関係eの強さ

$l(e)$：共起関係eのラベル



関係グラフ上の予測モデル

「単語iと単語jがラベルlで共起する強さ」を予測

- **単語$i \mapsto$ ベクトル$x(i)$**
- **ラベル$l \mapsto$ 半正定値$G(l)$**
- **i, jがlで共起する強さ**

$$x(i)^\top G(l) x(j)$$

ベクトル表現を求める最適化問題

$$\min_{x,G} \sum_{(i,j,w,l)\in E} (w - x(i)^\top G(l) x(j))^2$$

最適化アルゴリズム：交互最適化

$$\min_{x,G} \sum_{(i,j,w,l)\in E} (w - x(i)^\top G(l) x(j))^2$$

収束するまで以下を反復：

(1) $G(l)$ を固定して $x(i)$ を最適化

(2) $x(i)$ を固定して $G(l)$ を最適化

※ $G(l)$ は対角行列に限っても同等精度

各ステップは確率勾配法 (AdaGrad) の１ステップで実装

非凸最適化問題に対する勾配法なので初期値依存する

実験結果：初期値として既存手法の結果を使うと
初期値で得られる結果を consistent に上回る

Embedding semantic relations into word representations (IJCAI'15)

表現空間上に線型構造を仮定

ベクトル表現の差は意味をあらわす

$$\boldsymbol{v}(\text{king}) - \boldsymbol{v}(\text{man}) \approx \boldsymbol{v}(\text{queen}) - \boldsymbol{v}(\text{woman})$$

関係共起は意味に対応する

Ostrich is a large **bird**

$$\Rightarrow \text{is\_a\_large} \approx \boldsymbol{v}(\text{ostrich}) - \boldsymbol{v}(\text{bird})$$

「差が意味をあらわす」ことを直接扱うために
関係のベクトル表現 $p \mapsto \boldsymbol{v}(p)$ を中間変数として導入

$R(p)$: pで共起する単語全体（関係の意味の外延的定義）
$\boldsymbol{R}(\text{is\_a\_large}) \ni (\text{ostrich}, \text{bird})$

提案モデル

関係の表現は，共起する単語対の表現の差

$$v(p) = \frac{1}{|R(p)|} \sum_{(i,j) \in R(p)} w(i,j,p)(v(i) - v(j))$$

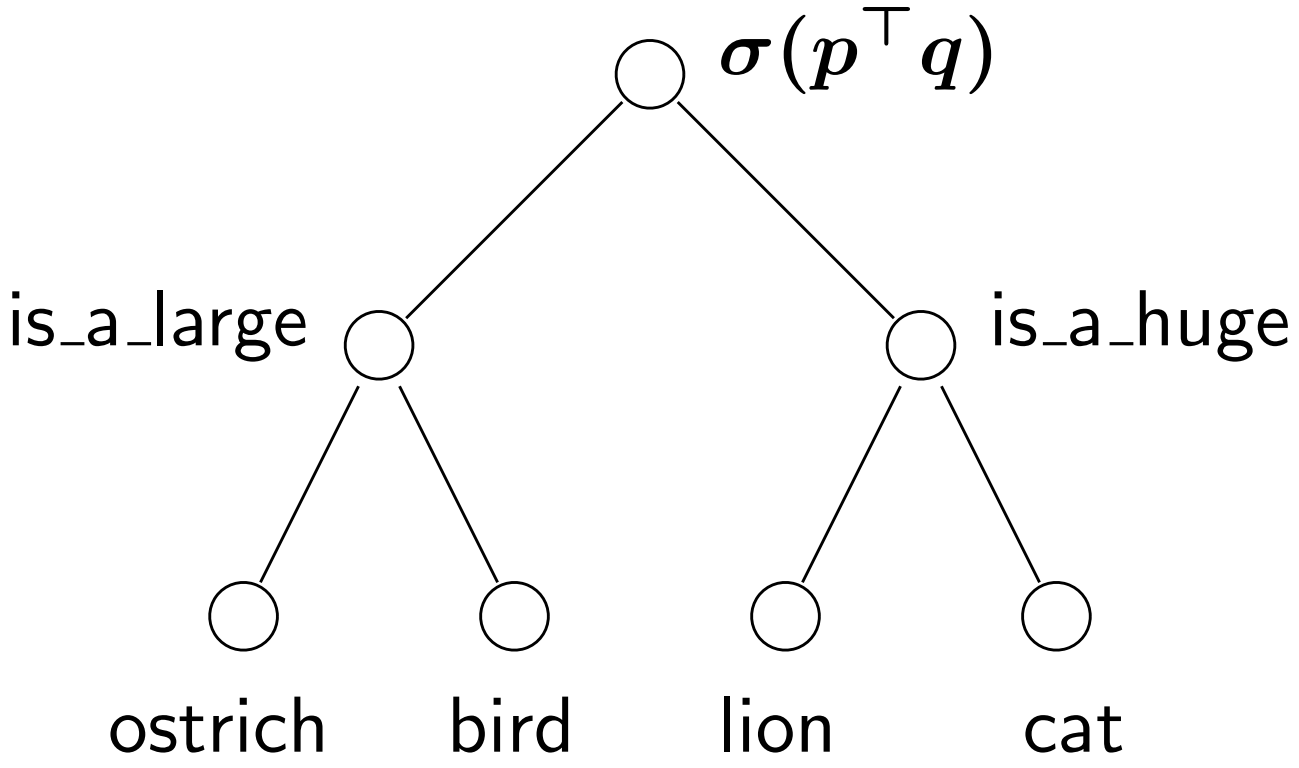
最適化する関数

$$L(v) = \sum_{(p,q) \in D} (s(p,q) - \sigma(v(p)^\top v(q)))^2$$

where

- s：**関係** (p,q) **の類似度** $(+1/-1)$
- $\sigma = \tanh$（**非線形変換**）

提案モデルのイメージ図



$$\Rightarrow \quad \text{is\_a\_large} = \text{ostrich - bird}$$
$$\approx \text{is\_a\_huge} = \text{lion - cat}$$

最適化アルゴリズム：交互最適化

収束するまで以下を反復：

(1) $v(i)$ を固定して $v(p)$ を計算（平均するだけ）

(2) $v(p)$ を固定して $v(i)$ を最適化

ステップ (2) は確率勾配法 (AdaGrad) の１ステップで実装
勾配が連鎖律から解析的にわかるので，高速に動く

非凸最適化問題に対する勾配法なので初期値依存する

実験結果：初期値として既存手法の結果を使うと
初期値で得られる結果を consistent に上回る

最適化屋として面白いと思うところ

非凸最適化問題の初期値依存性を積極的に使う

普通の考え方：

初期値ごとに違う結果が出るからつらい

ここでの考え方：

初期値の性質を継承した局所最適解が出てうれしい

最適化問題の大域最適解がNLP的に一番良いとは限らない

⇒ 良い解を局所最適化して性質付加する，という発想

まとめ

文脈を考慮した単語のベクトル表現（提案手法）

- **妥当そうな非凸最適化問題を設定**
  - **テンソル分解** (AAAI'15)
  - **線型構造の過程** (IJCAI'15)
- **文脈考慮しない表現を初期値にして局所最適化**
- **実験上，既存手法を** consistent **に上回る**