JAECS 2018 年度 春季研究会

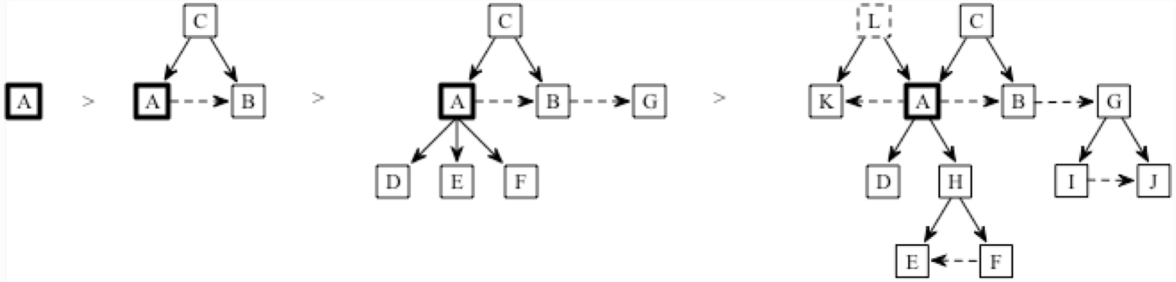
語の意味的粒度とコロケーションに関する試論
word2vecを用いた概念ネットワーク構築

長谷部 陽一郎
同志社大学

はじめに

本発表で扱う理論的な問い

- **コロケーション情報**に基づいて**語彙概念ネットワーク**のモデルを構築するにはどうしたらよいか
- 言語使用者の直感に合致する結果を得るには語の**意味的な粒度や階層性**を考慮する必要があるのではないか



Langacker (1990: 271)

本発表では上記の問いについて考えるべく進めている研究の**途中経過**を話します。

はじめに

技術的な課題

- 大規模なコーパス・データから共起情報を抽出し、**語どうしで比較できる形**にするにはどうしたらよいか
- 意味の粒度を考慮できる（内容的・サイズ的に）**適切なデータセット**を用意するにはどうしたらよいか

本発表で提案する方法

- **Wikipedia**のダンプデータのカテゴリー情報から意味の領域を適切に切り分けたサブコーパスを作成して、**word2vec**を用いた分析を実施

word2vecについて

特徴

- 2層のニューラルネットを用いた、**分散表現**（word embeddings）を作成するための手法
- 語の意味を共起情報に基づいて**ベクトル化**
- 語どうしで意味を「**計算**」できる

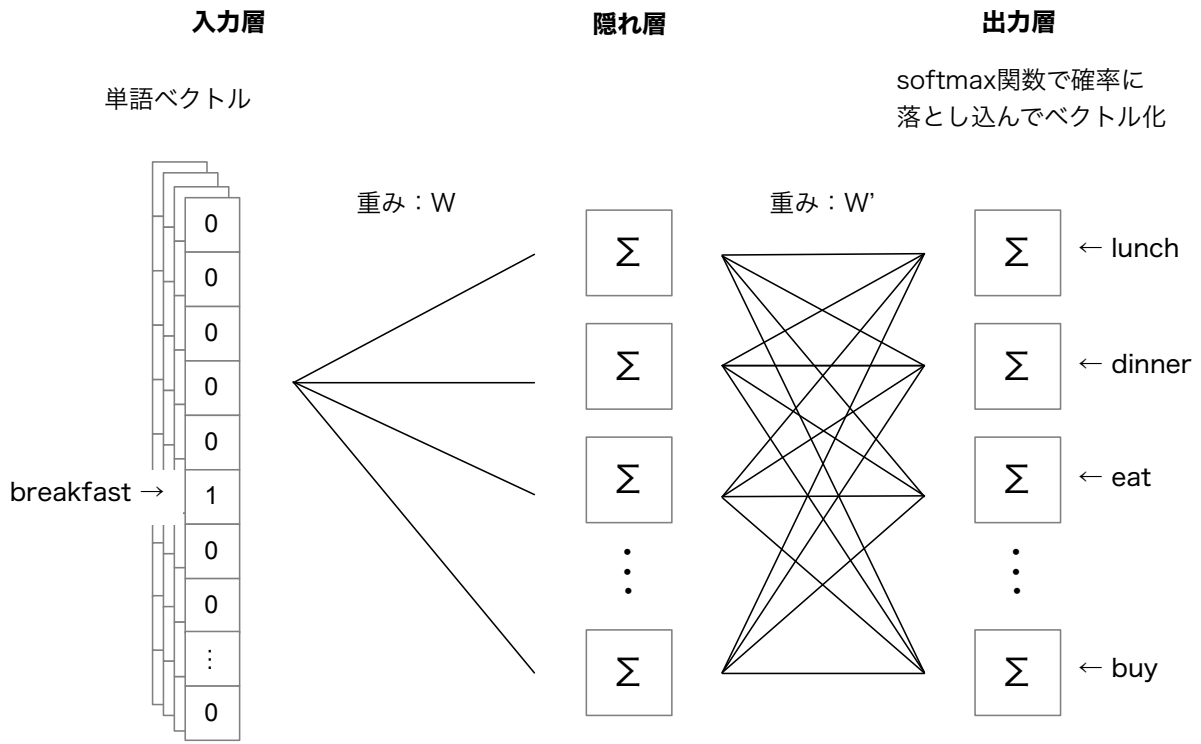
文献

- **Mikolov, Tomas, et al. 2013**. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781
- **西尾泰和. 2014**『word2vecによる自然言語処理』オライリー.

など

拙作wp2txtを用いて
日本語Wikipediaデータを抽出する
方法も載っています。

word2vecについて



McCormic (2016) を参考に作図

Wikipedia (enwiki) テキストの抽出

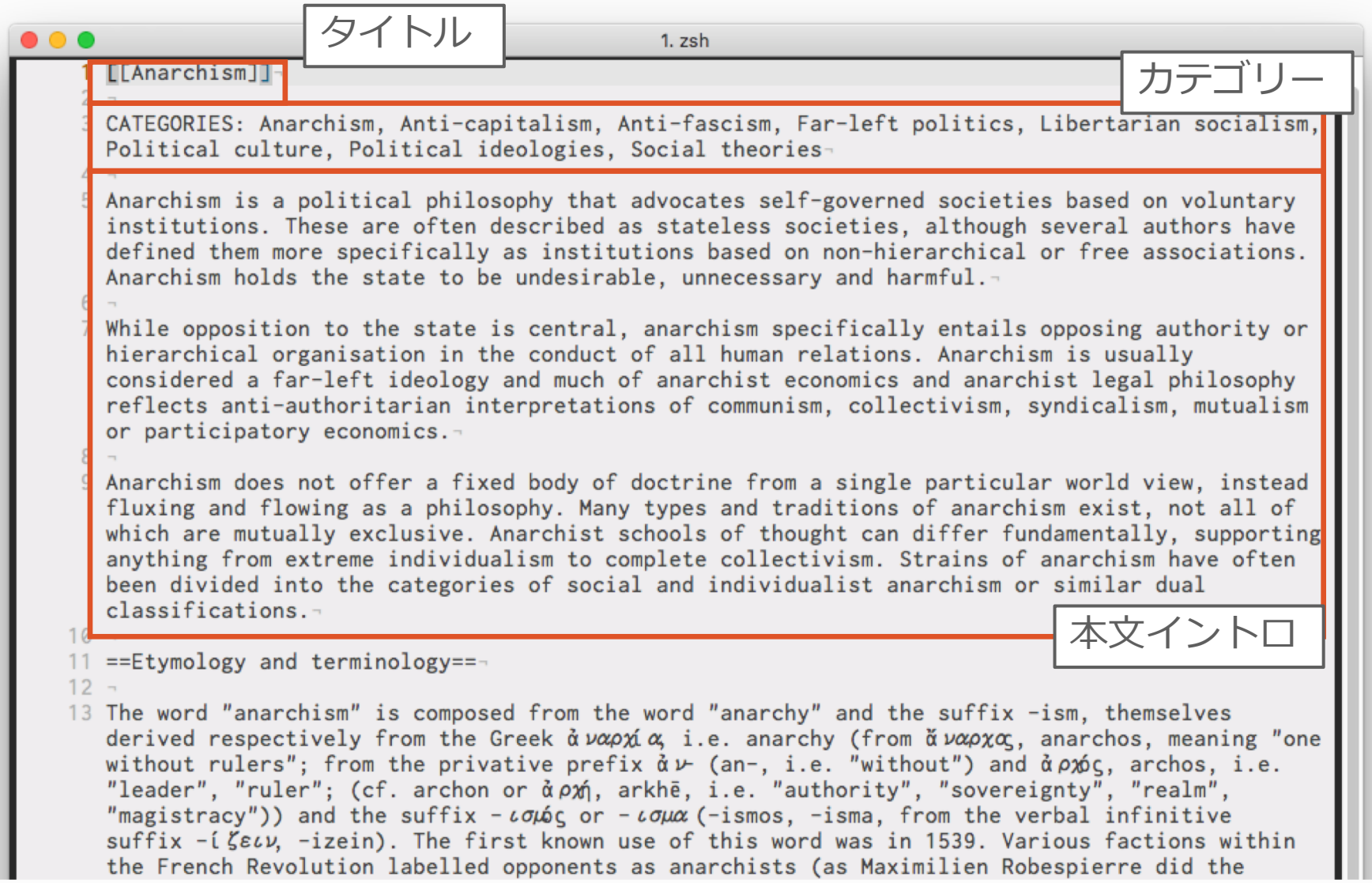
データの性質

- 公式サイトで定期的にスナップショットが配信されている
  2018年4月1日版・・・BZ圧縮ファイル約16GB
  https://en.wikipedia.org/wiki/Wikipedia:Database_download
- 約500万件の記事・約20億語・テキストファイル約14GB

ダンプデータからテキストデータを得るには・・・

- **wp2txt: Wikipedia dump file to text converter**
  https://github.com/yohasebe/wp2txt
- 上記ツールでは本文プレーンテキストの他、各記事の
  カテゴリー属性を取得することができる。（長谷部 2006）

Wikipedia (enwiki) テキストの抽出［結果］



Enwiki-Intro Corpusの作成

Wikipediaデータのどの部分をみるか・・・

- 各記事の**イントロダクション**のテキストを抽出
  → データ量の削減・不要データの排除のため
  → カテゴリーへの帰属度の高いデータを効率的に抽出し
  特定の事物や概念を「説明」するテキストのコーパスを得る

得られたコーパス

- **Enwiki-Intro Corpus**
  bz2圧縮ファイル約1.1GB・テキストファイル約3GB
  4,939,000記事・460,694,181語
- 公開の方法や時期は今のところ未定（圧縮データ or 構築ツール）

Enwiki-intro Corpusの作成

インターフェイスの作成

- 記事のカテゴリー、タイトル、本文（イントロ）をRDBに格納
- SQLでカテゴリー属性テキストの部分検索による絞り込み、
  領域を絞ったサブコーパスを作成できるように
- **カテゴリーと記事タイトルは「多対多」**の関係
  部分検索を工夫することでかなり柔軟に絞り込みが可能

```
1 select categories.title as ctitle, articles.id as aid, articles.title as atitle,
2 articles.introduction as intro from categories
3 inner join articles_categories on categories.id = articles_categories.category_id
4 inner join articles on articles.id = articles_categories.article_id
5 where categories.title ~* 'language' or categories.title ~* 'linguistics'
6
```

Load Query... Save Query... Cancel Execute Statement

ctitle	aid	atitle	intro
English-language idioms	397	Bootstrapping	In general, bootstrapping usually refers to a self-starting process that is…
C (programming language)	42	ANSI C	ANSI C, ISO C and Standard C refer to the successive standards for the C pro…
Programming language standards	42	ANSI C	ANSI C, ISO C and Standard C refer to the successive standards for the C pro…
English-language films	56	Army of Darkness	Army of Darkness is a 1992 American horror comedy film directed and co-writt…
English-language films	88	The Birth of a Nation	The Birth of a Nation is a 1915 American silent epic drama film directed and…
Member states of the Dutch Language Union	93	Belgium	Belgium ,België Nl-belgie.ogg; Belgique Fr-belgique.ogg; Belgien De-belgien.…
17th-century Latin-language writers	112	Baruch Spinoza	Baruch Spinoza was a Dutch philosopher of Sephardi/Portuguese origin. …
Regions of Europe with multiple official languages	201	Brussels	Brussels , officially the Brussels-Capital Region , is a region of Belgium c…
Agglutinative languages	209	Basque language	Basque is the language spoken in the Basque country. …
Basque language	209	Basque language	Basque is the language spoken in the Basque country. …
Languages of France	209	Basque language	Basque is the language spoken in the Basque country. …
Synthetic languages	209	Basque language	Basque is the language spoken in the Basque country. …
Subject-object-verb languages	209	Basque language	Basque is the language spoken in the Basque country. …

作成できるサブコーパスの例

芸術・メディア

- art または media を含むカテゴリー → 約40MB・6,556,683語

ビジネス

- business_を含むカテゴリー → 約31MB・4,980,842語

コンピュータ・ソフトウェア

- computer_ または softwareを含むカテゴリー → 約26MB・4,103,806語

動物・植物

- plantsまたはanimal_を含むカテゴリー → 約23.5MB・3,803,610語

政治

- politic_を含むカテゴリー → 約110MB・18,017,634語

社会

- social_またはsocietyを含むカテゴリー → 約34MB・ 5,530,901語

word2vecによる分析処理

Gensim (3.4.0) word2vec モジュール
以下の「例」では次の設定を使用

```
from gensim.models import Word2Vec
min_count = 5
size = 300
window = 5
model = Word2Vec(sentences,
                 min_count = min_count,
                 size = size,
                 window = window)
```

- 生起頻度 **5** 以上の語を対象とする
- 隠れ層の次元は **300**
- 近接語は左右各 **5** 語

word2vecの使用例 1

芸術・メディア サブコーパス

処理対象異なり語数：48,618

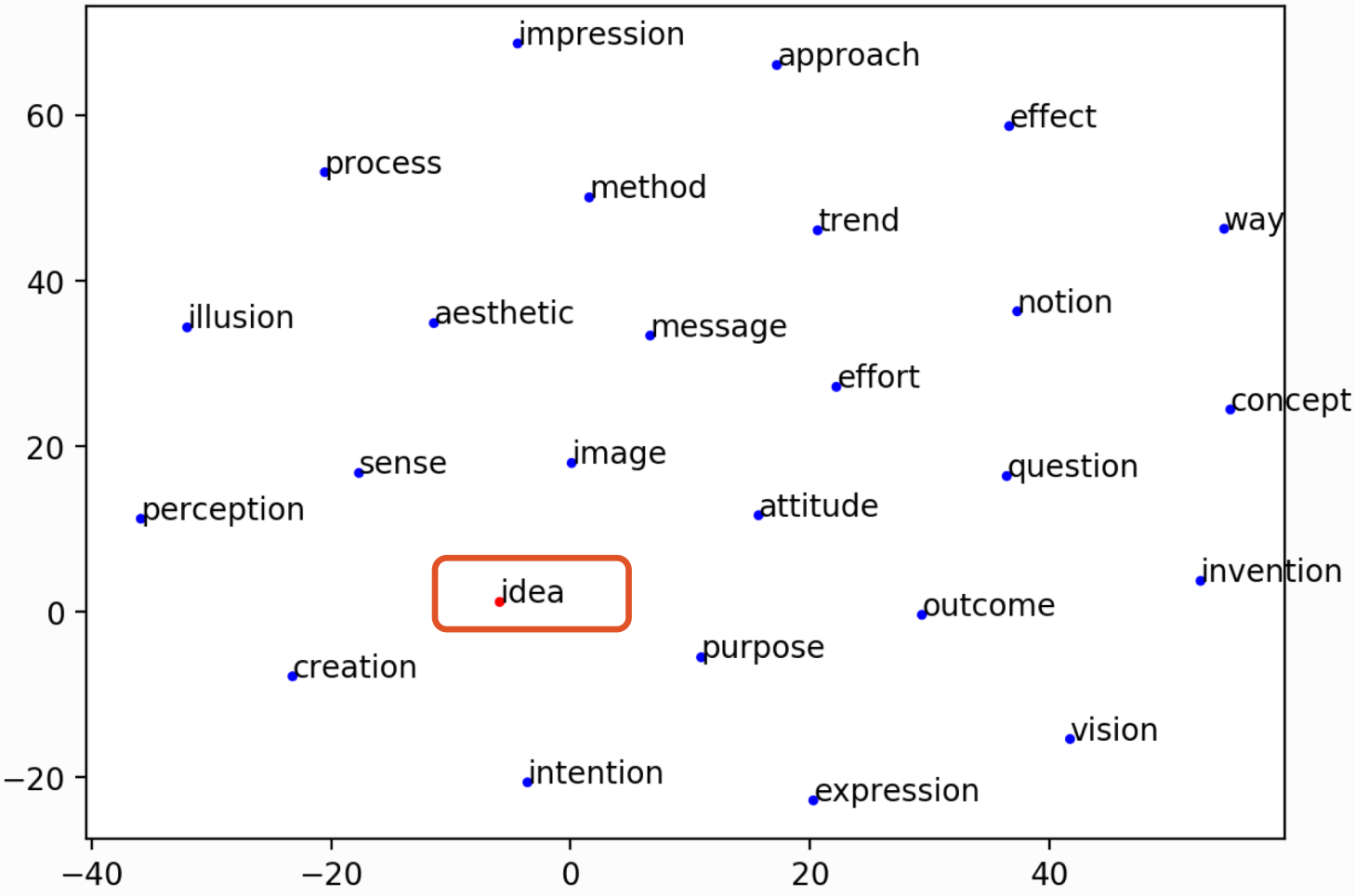
“idea” の意味的近接語

```
model.wv.similar_by_word ('idea', topn = 25)

[('concept', 0.7832673192024231),
 ('process', 0.6979352831840515),
 ('expression', 0.696306467056274),
 ('effect', 0.694710910320282),
 ('invention', 0.688177227973938),
 ('sense', 0.680100739002278),
 ... ]
```

t-SNE（t-Distributed Stochastic Neighbor Embedding）により
300次元を2次元に圧縮して可視化

word2vecの使用例1



芸術・メディアの領域において特徴的な語・概念のネットワークを
いわゆる放射状カテゴリーを思わせる形状で得ることができている

word2vecの使用例2

“accept” の意味的近接語のネットワークを異なる領域でモデル化して比較

教育 サブコーパス

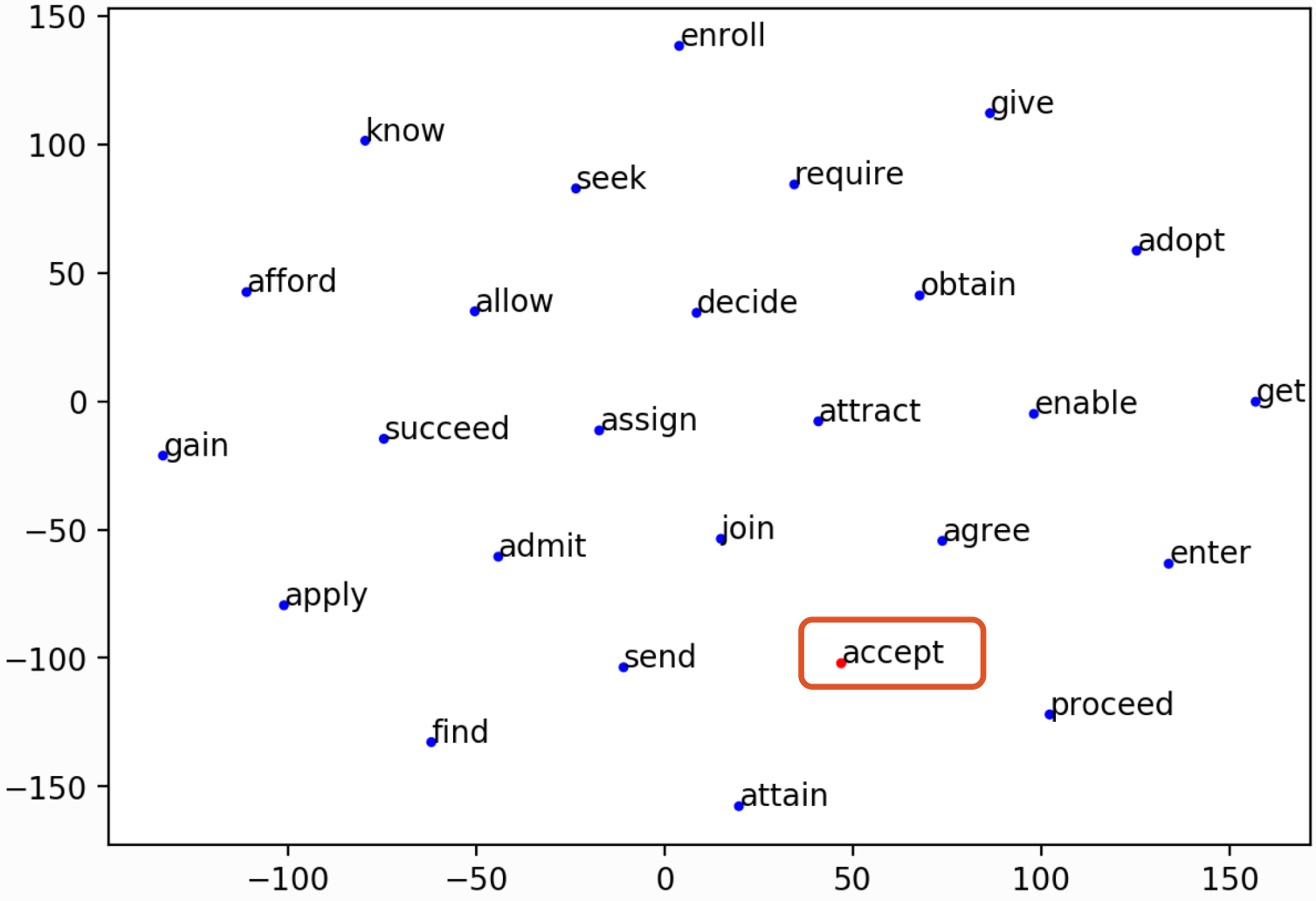
処理対象異なり語数： 41,269

政治 サブコーパス

処理対象異なり語数： 85,442

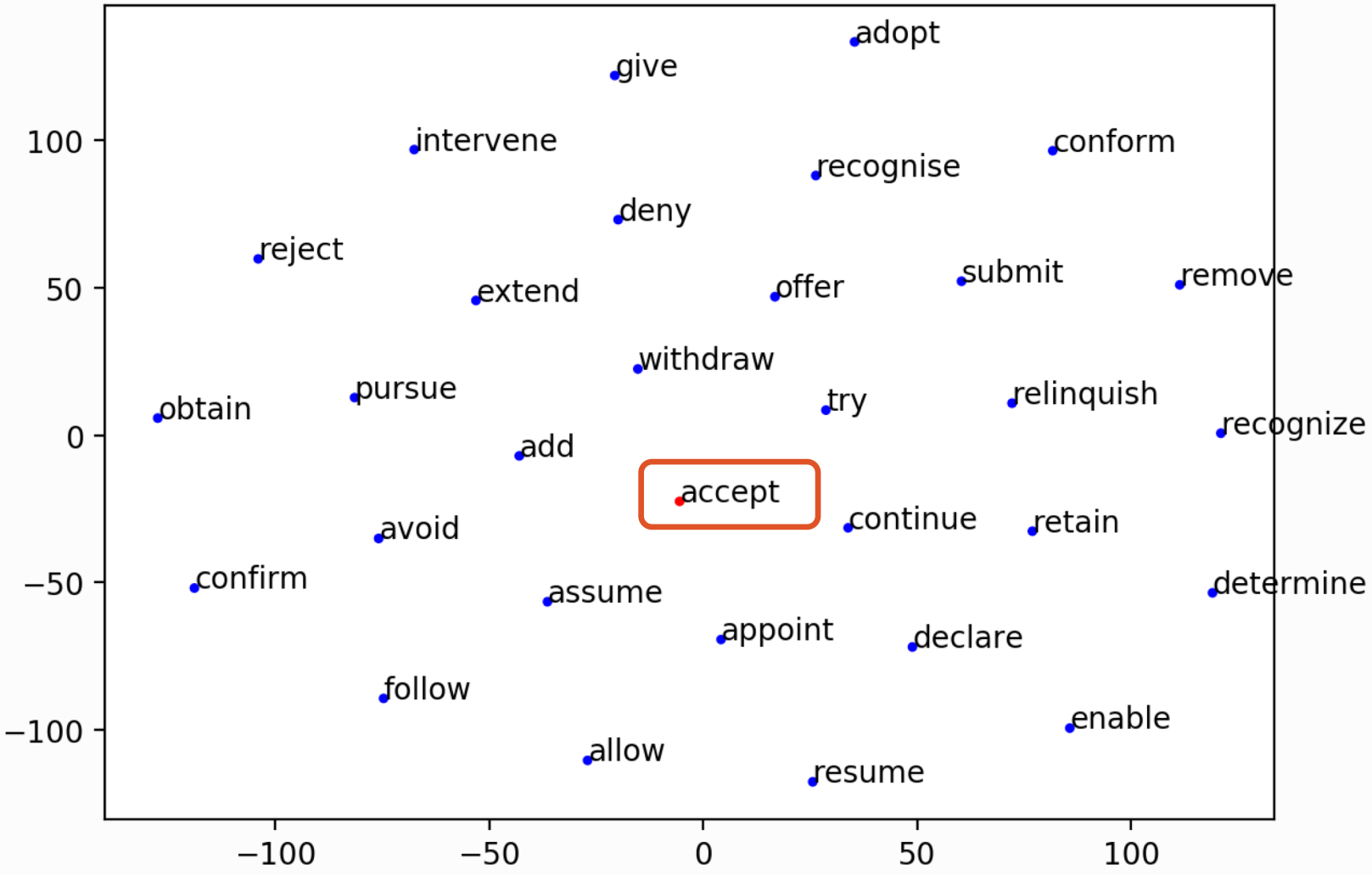
t-SNE（t-Distributed Stochastic Neighbor Embedding）により
300次元を2次元に圧縮して可視化

word2vecの使用例2



教育分野：学校関連の “accept”にまつわる事象を表す語が確認できる

word2vecの使用例2



政治分野：条約や法案を扱う厳格な手続きを思わせるような語が観察される

word2vecの使用例3

ビジネス サブコーパス

処理対象異なり語数： 39,338

コンピュータ・ソフトウェア サブコーパス

処理対象異なり語数： 31,932

```
model.wv.similar_by_word ('send')
```

ビジネス

```
[('give', 0.88),
 ('find', 0.88),
 ('allow', 0.87),
 ('deliver', 0.86),
 ('add', 0.86),
 ('keep', 0.85),
 ('overcome', 0.85),
 ('collect', 0.85),
 ('handle', 0.85),
 ('retain', 0.84)]
```

コンピュータ・ソフトウェア

```
[('receive', 0.87),
 ('upload', 0.85),
 ('listen', 0.82),
 ('transmit', 0.81),
 ('browse', 0.79),
 ('scan', 0.79),
 ('incoming', 0.79),
 ('synchronize', 0.78),
 ('securely', 0.78),
 ('retrieve', 0.78)]
```

word2vecの使用例3

word2vecでは語の意味を「計算」できる」とはどういうことか

定番の例

father + daughter – mother = son

france + tokyo - japan = paris

2つのサブコーパスを使って実験

```
model.wv.most_similar(positive = ['ios', 'google', 'mobile'],
                      negative = ['apple'])
```

ビジネス

```
[('android', 0.89),
 ('server', 0.87),
 ('sql', 0.85),
 ('linux', 0.85),
 ('windows', 0.84), ... ]
```

コンピュータ・ソフトウェア

```
[('android', 0.70),
 ('apps', 0.68),
 ('skype', 0.68),
 ('messenger', 0.64),
 ('bing', 0.64), ...]
```

word2vecの使用例3

領域が異なると語どうしの距離も異なることがある（仲間はずれ検知）

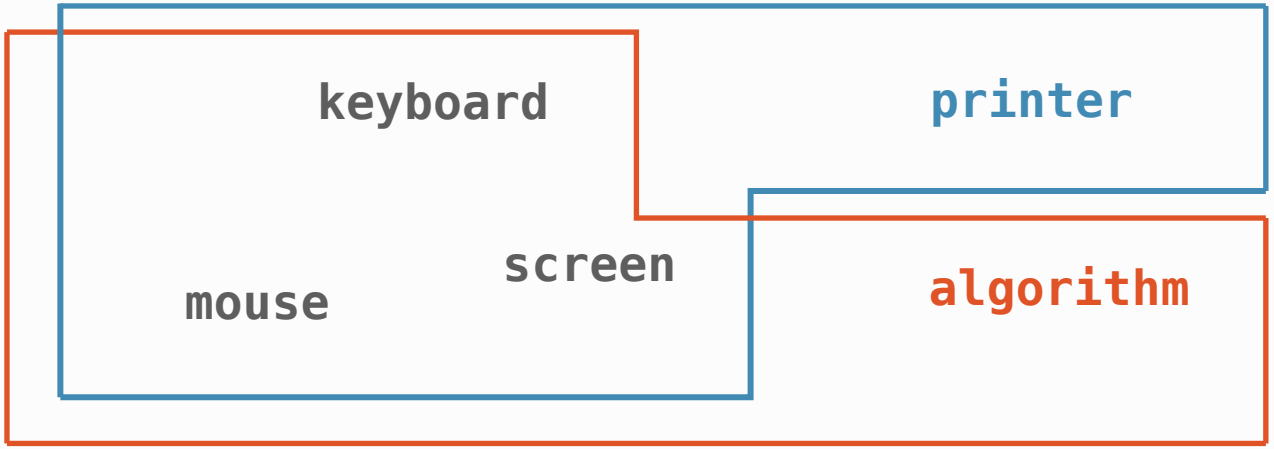
```
model.wv.doesnt_match(['keyboard',
                       'mouse',
                       'screen',
                       'printer',
                       'algorithm'])
```

ビジネス

```
'printer'
```

コンピュータ・ソフトウェア

```
'algorithm'
```



word2vecの制約・課題

- **多義語**の扱いが難しい
- **反義語**を区別させるのが難しい
- **連語**の扱いが難しい
- 語の意味の類似度に影響する**要因**をつかむことが難しい

Cf. Desagulier (2017)

まとめ

word2vec

- 語どうしの近接性を知る／語の意味を「計算」できる
- 大規模データから高速に語彙ネットワーク構築

Enwiki-Intro Corpus

- 特定の事物や概念を「説明」したテキストのコーパス
- 特定のカテゴリーに属するテキストでサブコーパスを構築

制約はあるが、上記の2つを組み合わせることは、特定領域の語彙の概念ネットワークを構築するための有望な方法と思われる。

サブコーパスを階層的に構築して、サブコーパス間にみられる差異や共通点に着目することで、将来的には**語の意味的粒度**（e.g. 抽象カテゴリー、基本レベルカテゴリー、細別カテゴリー）の違いを射程に入れた立体的な語彙ネットワークのモデル構築につながるかもしれない

言及文献

Desagulier, Guillaume. 2017. Can word vectors help corpus linguists? <halshs-01657591>

長谷部陽一郎. 2006. Wikipedia 日本語版をコーパスとして用いた言語研究の手法.『言語文化』9-2: 373-403.

Langacker, Ronald W. 1990. *Concept, Image, and Symbol: The Cognitive Basis of Grammar.* Mouton de Gruyter.

McCormick, Chris. 2016. Word2Vec Tutorial: The Skip-Gram Model. < http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/>

Mikolov, Tomas, et al. 2013. Efficient Estimation of Word Representations in Vector Space. <arXiv:1301.3781>

西尾泰和. 2014『word2vecによる自然言語処理』オライリー.