

大規模言語モデルを利用したコーパスの意味分析入門

文を超えた意味の計算

横井 祥

国立国語研究所・東北大学・理化学研究所・JST創発

2025-06-14

英語コーパス学会 2025年度春季研究会・総会
@青山学院大学

自己紹介



横井 祥 (よこい しょう)

- <http://www.cl.ecei.tohoku.ac.jp/~yokoi/>

🕒 略歴

- **B** … 京大 計算機 (機械学習, 指導教員: 鹿島久嗣先生)
- **M/D** … 東北大 情報 (自然言語処理, 指導教員: 乾健太郎先生)
- **PD** … 東北大 情報, 理研 AIP
- **現** … 国語研 **E3Pセンター**, 東北大 言語AIセンター, 理研 AIP

🎓 **専門** … 自然言語処理・表現学習

👁️ 興味

- **表現学習と意味の形**
- 人間の高次の知的処理もだいたいパターンマッチなのでは?
 - [slide] 確率的なオウムにできること, またそれがなぜできるのかについて
- 経験主義的な言語(哲)学の実験科学・数理科学としてのリバイバル
 - [slide] コーパスを丸呑みしたモデルから言語の何がわかるか

自己紹介



横井 祥 (よこい しょう)

– <http://www.c>

これまでは**計算機科学**・**情報科学**のコミュニティにいました

🕒 略歴

- B … 京大 **計算機** (機械学習, 指導教員: 鹿島久嗣先生)
- M/D … 東北大 **情報** (自然言語処理, 指導教員: 乾健太郎先生)
- PD … 東北大 **情報**, 理研 **AIP**
- 現 … 国語研 **E3Pセンター**, 東北大 **言語AIセンター**, 理研 **AIP**

🎓 専門 … 自然言語処理と機械学習

言語 × **認知** × **数理**

👁️ 興味

について考える未曾有のチャンス
だと思っています

- 表現学習
- 人間の高次の知的処理もだいたいパターンマッチなのでは?
 - [slide] 確率的なオウムにできること, またそれがなぜできるのかについて
- 経験主義的な言語(哲)学の実験科学・数理科学としてのリバイバル
 - [slide] コーパスを丸呑みしたモデルから言語の何がわかるか

自己紹介



横井 祥 (よこい しょう)

- <http://www.cl.ecei.tohoku.ac.jp/~yokoi/>

🕒 略歴

- B … 京大 計算機 (機械学習, 指導教員: 鹿島久嗣先生)
- M/D … 東北大 情報 (自然言語処理, 指導教員: 乾健太郎先生)
- PD … 東北大 情報, 理研 AIP
- 現 … 国語研 E3Pセンター, 東北大 言語AIセンター, 理研 AIP

🎓 専門 … 自然言語処理・表現学習

👁️ 興味

- 表現学習と意味の形

- 人間の高次の知的処理もだいたいパターンマッチなのでは?
 - [slide] 確率的なオウムにできること, またそれがなぜできるのかについて
- 経験主義的な言語(哲)学の実験科学・数理科学としてのリバイバル
 - [slide] コーパスを丸呑みしたモデルから言語の何がわかるか

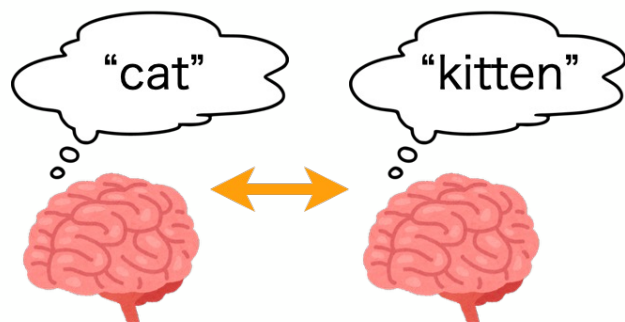
今日はこちらの話をします
直感的な説明をできればと思っています
数理的な話が気になる人向けに少しだけ追加説明もします

導入：意味の形

言語のモデルのご利益：意味の形がわかる

工学的な便利さを超えた

“意味”の問題、
たとえばふたつの単語がどの程度“似ている”かを、
一般的に定義し計算するのは難しい



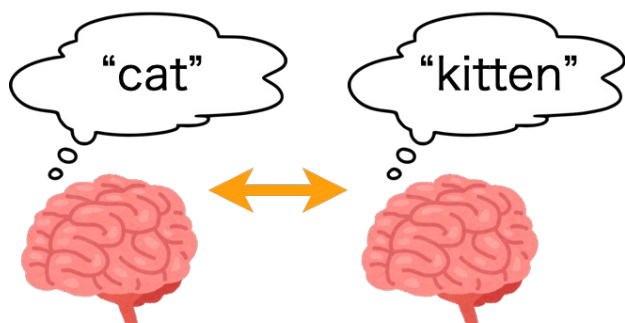
言語のモデルのご利益：意味の形がわかる

工学的な便利さを超えた

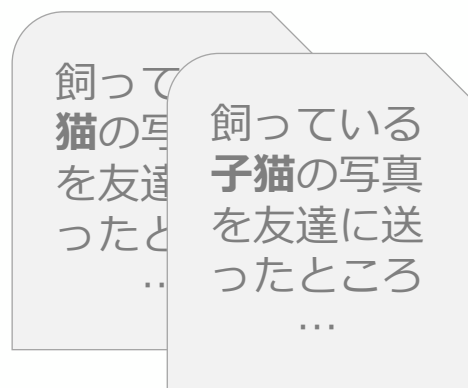
最近の言語モデルの魅力のひとつ

→ **意味**が**データ上の統計情報**を介して**言語モデル**に転写される

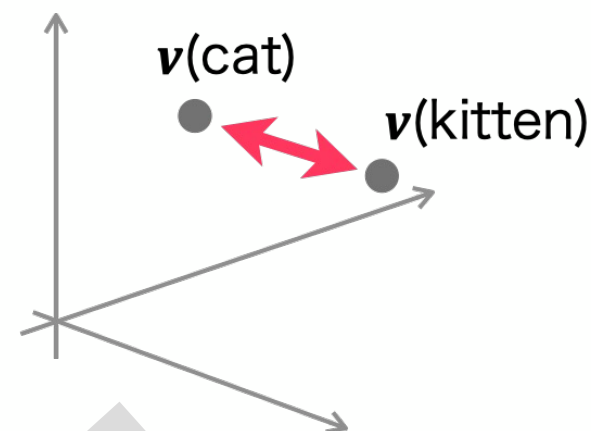
似た意味



似た分布



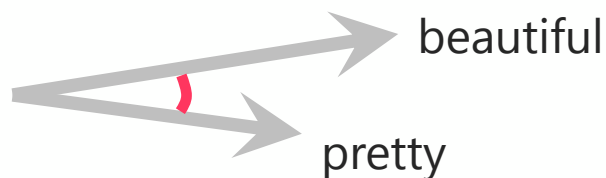
似た形



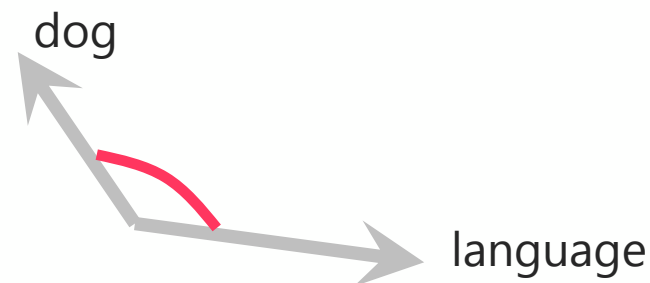
→ **言葉や意味**について、直感的で手触りのある**形**の情報を通して計算・理解・制御できて嬉しい（私が）

単語埋め込み空間における「意味 as 形」の例

- 「単語ペアに対して人間が感じる**意味の類似度**」と「単語ベクトルペアの**なす角**」が高い相関



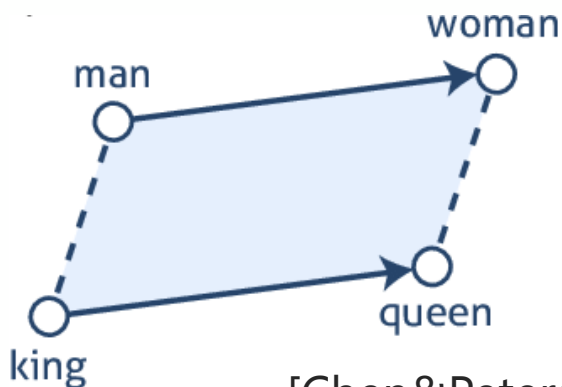
人間が似ていると思う単語ペアはベクトル同士が同じような向き



人間が似ていないと思う単語ペアはベクトル同士が関係ない向き

単語埋め込み空間における「意味 as 形」の例

- アナロジーの計算



[Chen&Peterson'17]

- 意味的な関係

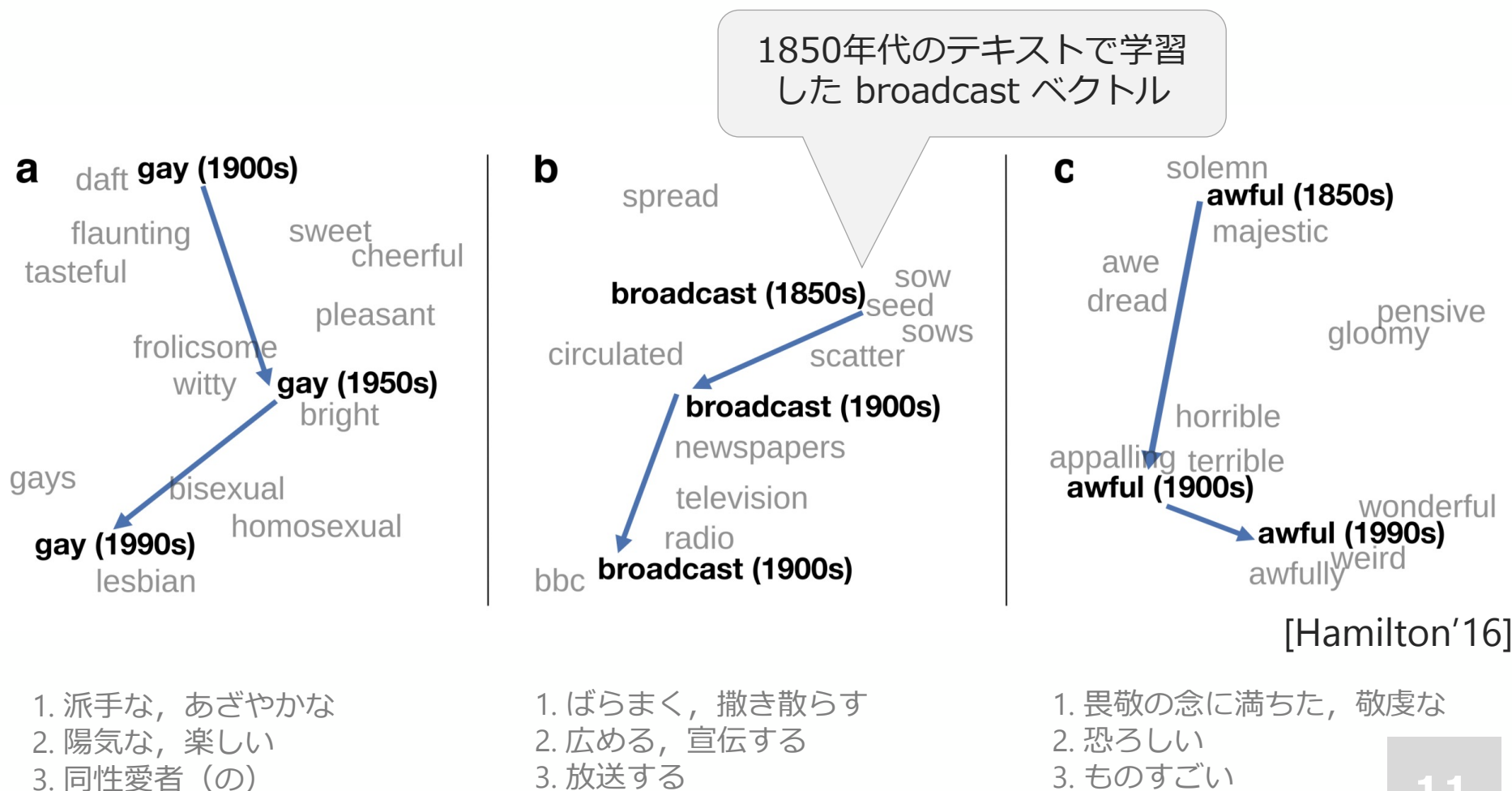
- 「king と queen の意味の異なり」
≈ 「man と woman の意味の異なり」

- ベクトル同士の関係

- 「king ベクトルと queen ベクトルの違い」
≈ 「man ベクトルと woman ベクトルの違い」

単語埋め込み空間における「意味 as 形」の例

- 単語の意味の変化をベクトル表現の軌跡でよく捉えられる



個人的な興味のひとつ： 「意味の形」の数学的な背景をはっきりさせる

- 普段おこなっている、意味と形の関係を調べる研究
 - 語の意味の強さ $\leftrightarrow$ 周辺分布の偏り $\leftrightarrow$ 単語ベクトルの長さ
[EMNLP'20, ICLR'21, EMNLP'23, EMNLP Findings'23, NeurIPS'24]
 - 語義への文脈の影響 $\leftrightarrow$? $\leftrightarrow$ 埋込空間での“混ざり具合”
[EMNLP'20, EMNLP'21, ICLR'24]
 - 文の意味的類似度 $\leftrightarrow$ 単語単位の集合類似度 $\leftrightarrow$ 点群間の幾何的類似度
[EMNLP'20, ACL'23, EMNLP Findings'23, NAACL'24]
 - 語義の排他性 $\leftrightarrow$ 周辺分布の均一さ $\leftrightarrow$ 埋込空間における局所性
[ACL Findings'25]
 - etc.
- ご興味を持たれたかたがおられたら、な動画
 - 🗣️ “言葉の形を教えてくれる自然言語処理” [IPSJ-ONE'22]
 -  [動画](#) (5min), [スライド](#)
 - 🗣️ “最適輸送と自然言語処理” [NLP'22]
 -  [動画](#) (90min), [スライド](#)
- 共同研究しましょう！

このあといくつか紹介します

今日の話題：「意味の形」に関する 色々な研究・計算ツール・考え方をご紹介します

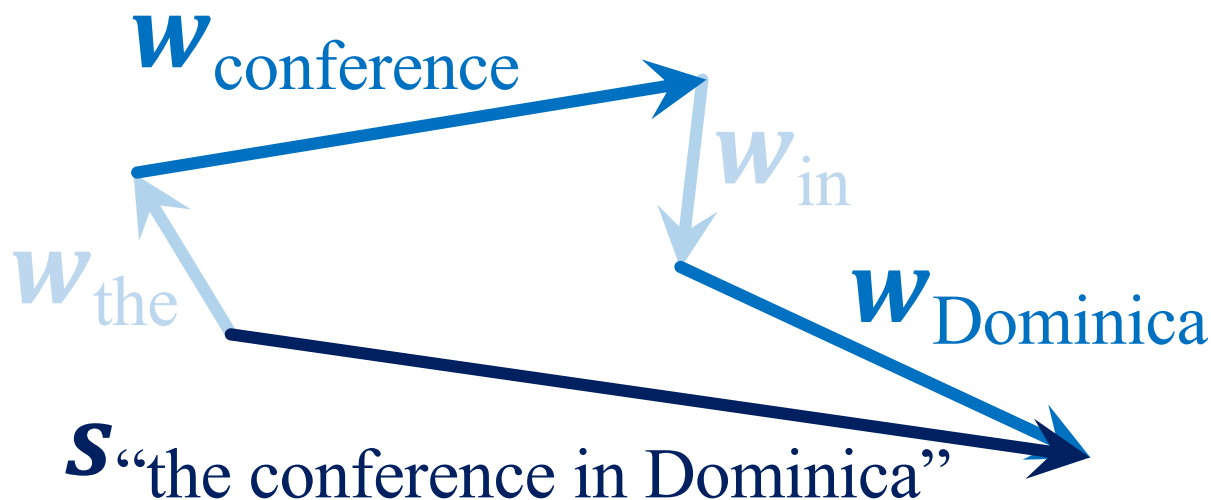
1. 準備：文のトピックを決める力が強い単語はベクトルが長い
2. 大きな言語ユニットの意味計算を
単語ベクトルを使って柔らかく実現する
 - 文の類似度や意味の違いを
「輸送のしやすさ」や「輸送できないこと」で柔らかく測る
 - 単語（ないし概念）集合間の「かつ」や「または」を
単語ベクトル空間で柔らかく計算する
 - 言語間の対訳辞書を最適輸送で柔らかく見つける
 - コーパス探索のための柔らかいコンコーダンスを作る

途中で QA の時間を入れます

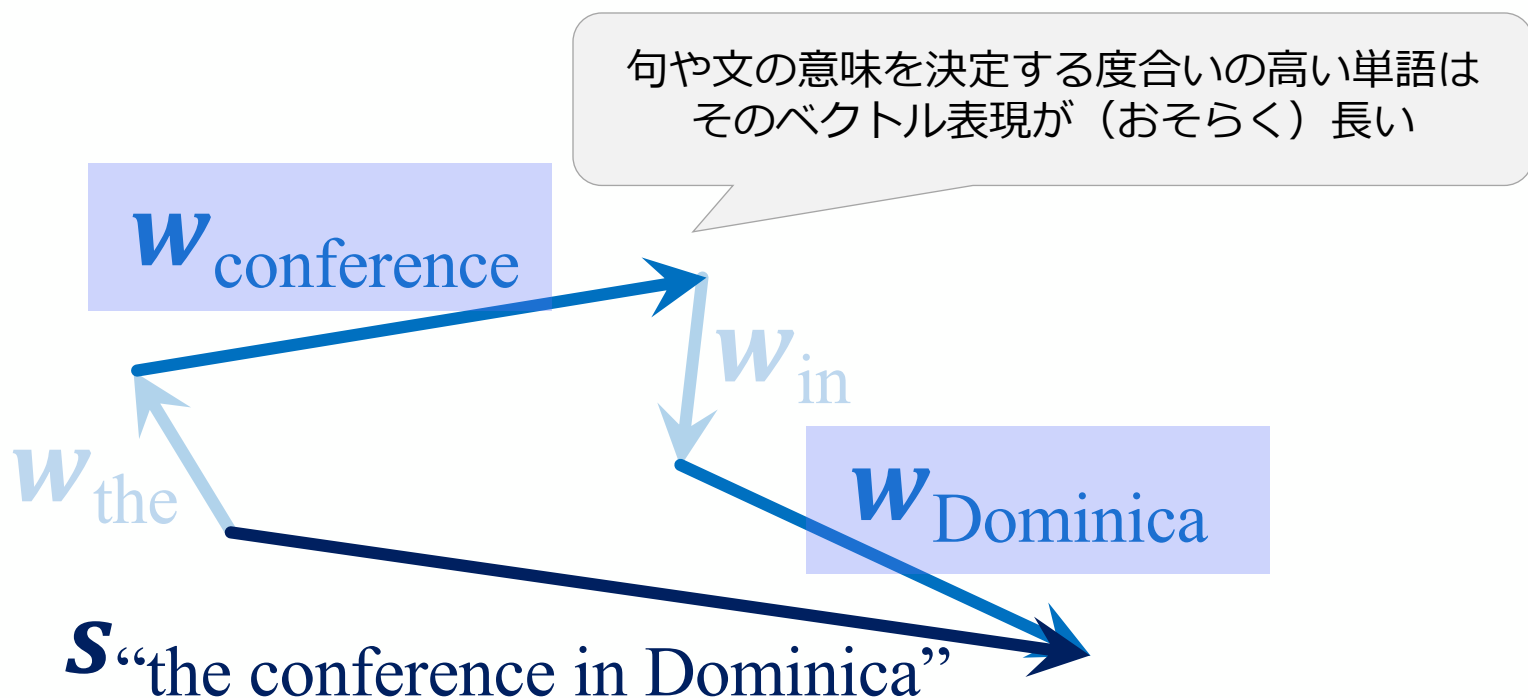
文のトピックを決める力が強い単語は そのベクトル表現が長い

- 「意味の形」の一例として
 - このあとの話の準備を兼ねて
-

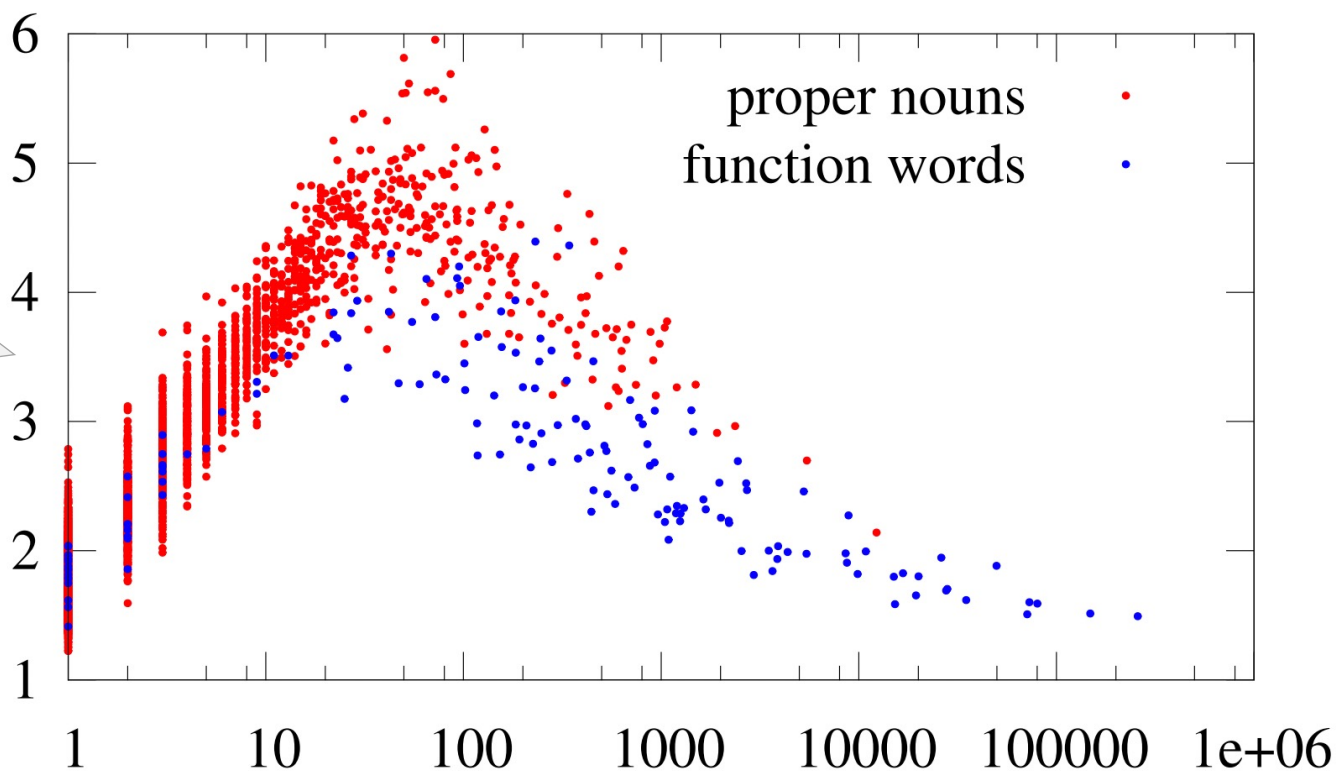
単語ベクトルを足すと, 良い (= 各種タスク
をうまく解ける) 句・文のベクトルができる



単語ベクトルを足すと、良い（＝各種タスクをうまく解ける）句・文のベクトルができる



固有名詞の単語ベクトルは、 機能語の単語ベクトルよりも長い [Schakel&Wilson'15]



長さ

コーパス頻度

単語が文のトピックを決める度合い ↔ 長さ

[Oyama, Y, Shimodaira, EMNLP'23; Y, Bao, Kurita, Shimodaira, NeurIPS'24]

トピックを決定する度合いを
情報理論の言葉で定式化

トピック
決める度

ベクトルの長さ

長さも真面目に測る
※ 計量という考え方を使う

gdi 国内総所得

10.13

78.2



phosphide リン化物

10.10

82.2



tachycardia 頻脈

10.00

82.1



...

of

0.09

1.2



and

0.08

0.9



the

0.08

1.0



 Oyama+, Norm of Word Embedding Encodes Information Gain (EMNLP 2023)

 Yokoi+, Zipfian Whitening (NeurIPS 2024)

ものすごく綺麗に対応する

長さの話は、学習方法からの必然だと言える

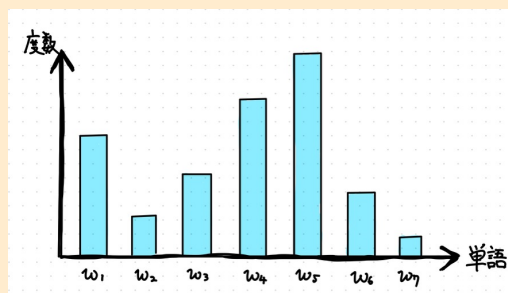
[Oyama, Y, Shimodaira, EMNLP'23; Y, Bao, Kurita, Shimodaira, NeurIPS'24]

単語がどの程度文の
トピックを狭めるか

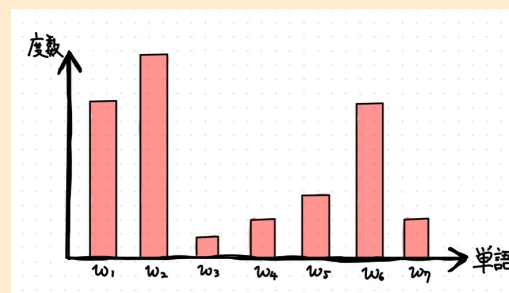
長さ

$$2 \text{KL}(p(\cdot) \parallel p(\cdot | w)) \approx \|w\|^2$$

コーパス全体の単語頻度分布

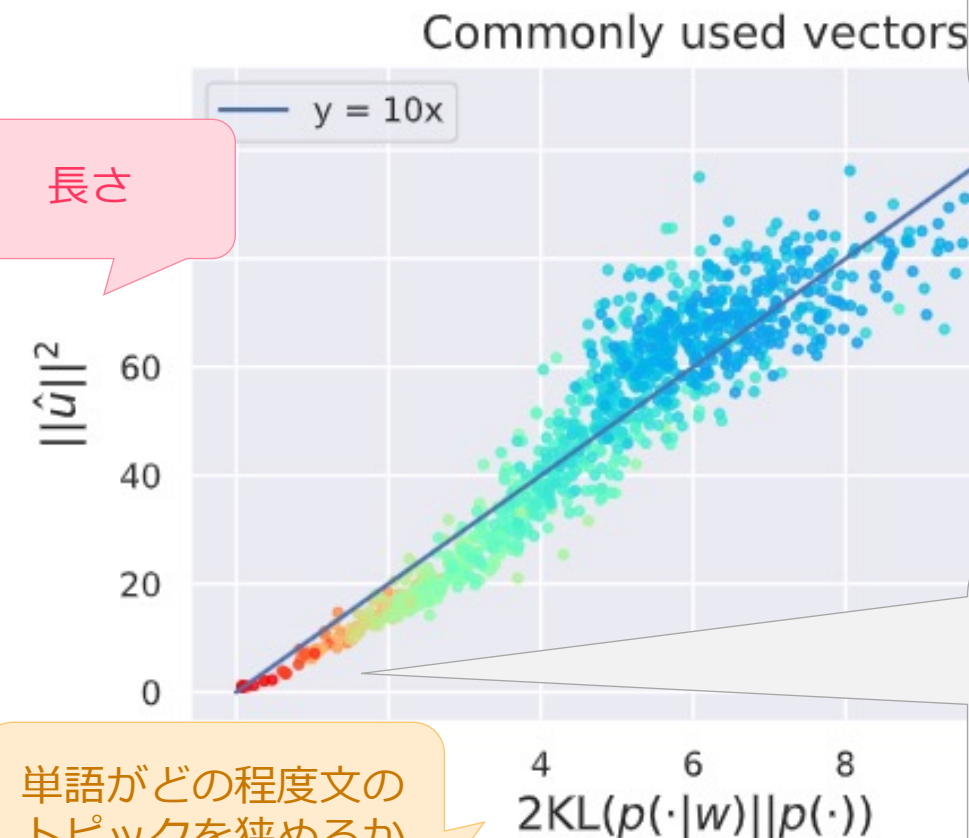


単語 w の周辺単語頻度分布



- 「ある単語 w が存在することで周辺単語の分布がどれくらいズレる？」 「文の意味が w の存在によってどれくらい決まる？」
 - 「ある文に “the” が入っている」 ← 周辺単語についてはほとんど何もわからない
 - 「ある文に “agglutinative” が入っている」 ← 周辺単語の候補が一気に狭まる

実際の単語ベクトルも 理論通りの挙動



Top 3 (w.r.t. KL)

単語	2 * KL	ノルム^2
tachycardia	9.997	82.119
phosphide	10.101	82.283
gdi	10.135	78.180

ノルムの長い側には
トピックを狭く決める語が集中

1. 《医》頻拍、頻脈
2. リン化物
3. 国内総所得

実際のモデルでも
綺麗に比例

Bottom 3 (w.r.t. KL)

単語	2 * KL	ノルム^2
the	0.076	1.036
and	0.083	0.904
of	0.094	1.190

ノルムの短い側には
トピックを決める力の弱い単語が集中

文の類似度や意味の違いを 「輸送のしやすさ」や 「輸送できないこと」で柔らかく測る

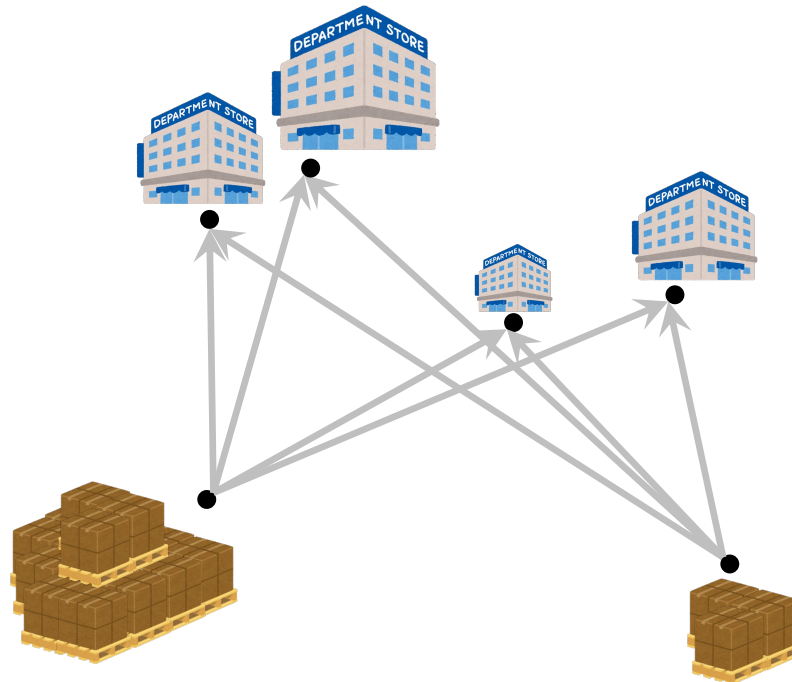
単語よりも大きな単位の意味の計算 (1)

最適輸送と自然言語処理

超簡略版

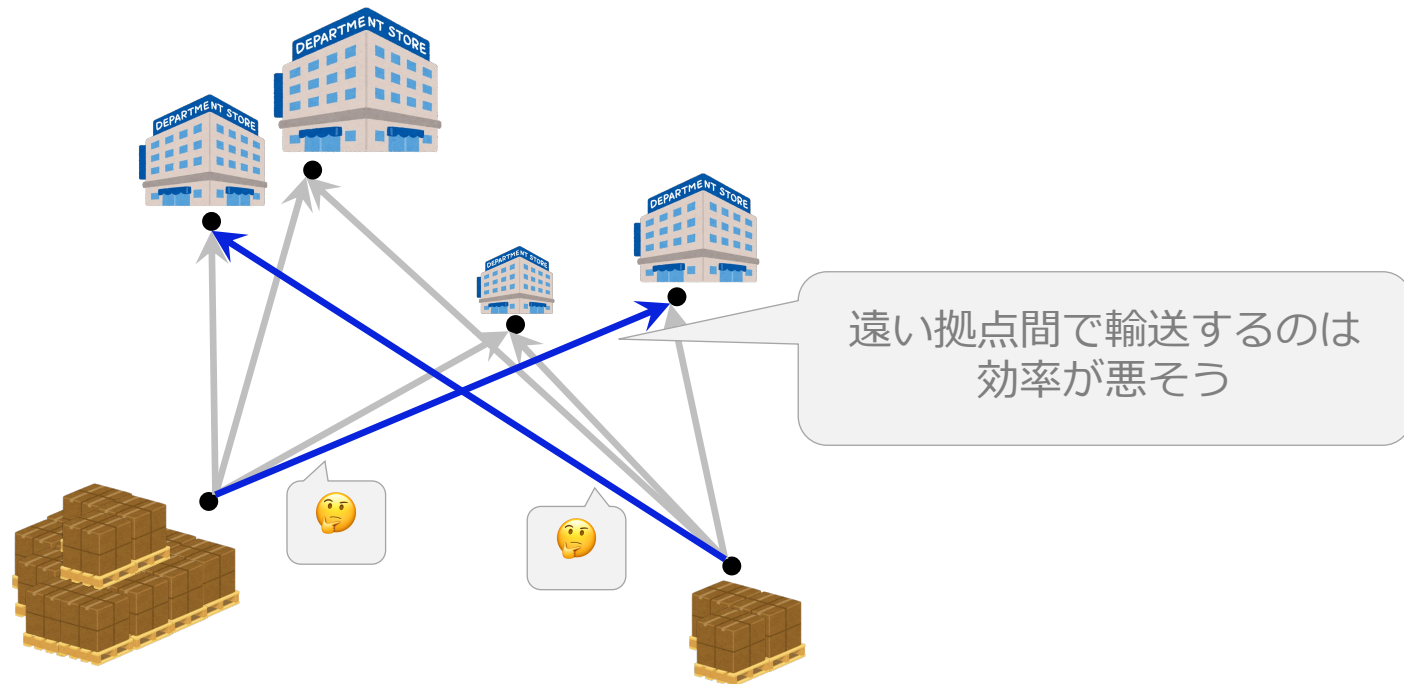
最適輸送

- 工場で作った特産品を市内のデパートに届けたい。
もっとも効率的な届け方は？



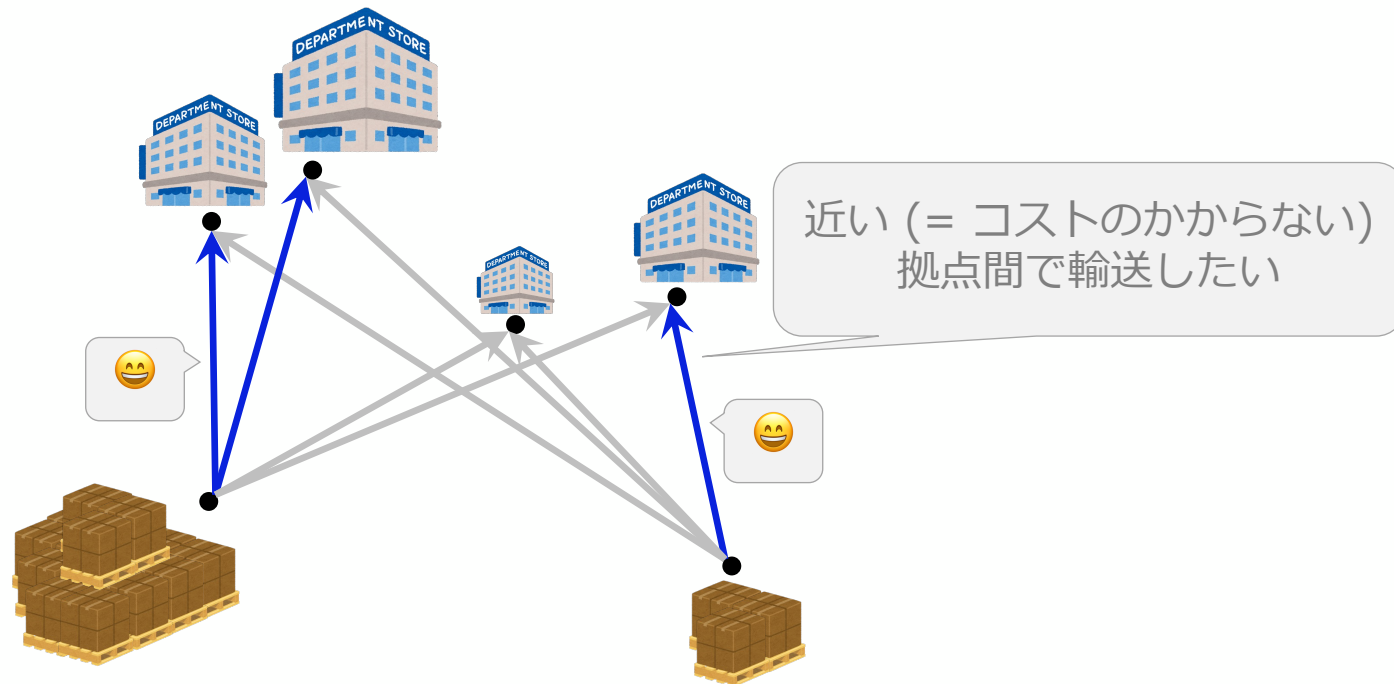
最適輸送

- 工場で作った特産品を市内のデパートに届けたい。
もっとも効率的な届け方は？



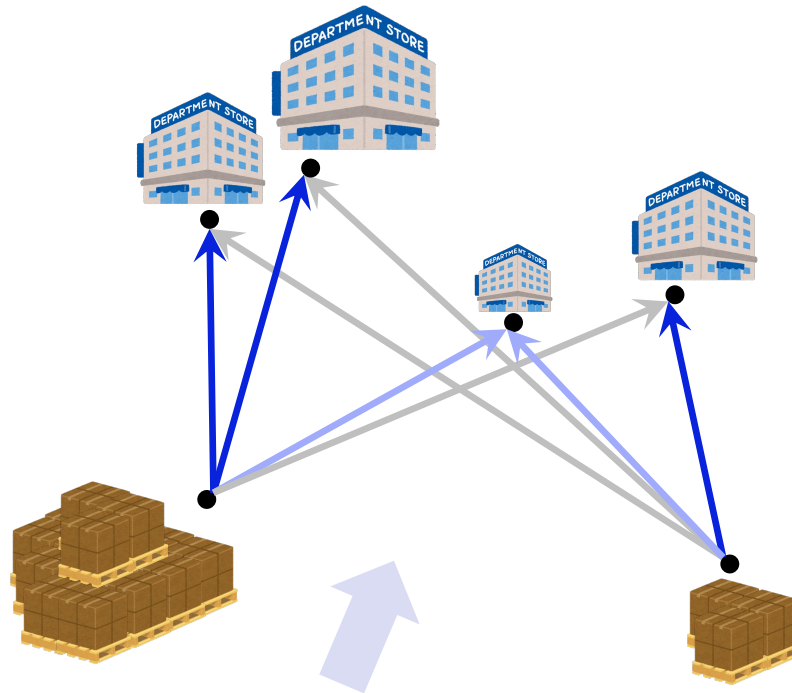
最適輸送

- 工場で作った特産品を市内のデパートに届けたい。
もっとも効率的な届け方は？



最適輸送

- 工場で作った特産品を市内のデパートに届けたい。
もっとも効率的な届け方は？



- 最適輸送** = 最良の輸送計画 と総輸送コストを求める道具

横井, 最適輸送と自然言語処理 (NLP 2022)

横井, 構造を持った言語データと最適輸送—二種類の「アラインメントに基づく類似度」 (NAIST DSC NLP Seminar 2022)

Semantic Textual Similarity (STS)

- ふたつの**テキスト**の**類似度**の測定

入力

出力

s: “Two boys on a couch are playing video games.”

s': “Two boys are playing a video game.”

4/5 mostly equivalent

s: “The woman is playing the violin.”

s': “The young lady enjoys listening to the guitar.”

1/5 not equivalent, but
are on the same topic

テキストの類似度測定

- ふたつの**テキスト**の**類似度**の測定
 - 自然言語処理で極めて頻繁に必要とされるサブルーチン
 - **例1** テキスト生成：**システムが生成した文** ↔ **人間の書いた見本の文**
 - システムの出力文は理想的な出力とどれくらいギャップがある？
 - 損失関数, 自動評価尺度
 - **例2** 関連テキスト検索：**文** ↔ **文**, **文書** ↔ **文書**, **クエリ** ↔ **文書**
 - 検索機能で拡張した (retrieval-augmented; RAG) 言語モデル, 類似性に基づく説明, etc.

チャットボットの訓練等に
使われる

「手元のテキストを参考にし
ながら答えてよ」

テキストの類似度測定

- ふたつの**テキストの類似度**の測定
 - 基本指針のひとつ：**要素単語の重なる度合い**を測る

Robin Warren was awarded a Nobel Prize .

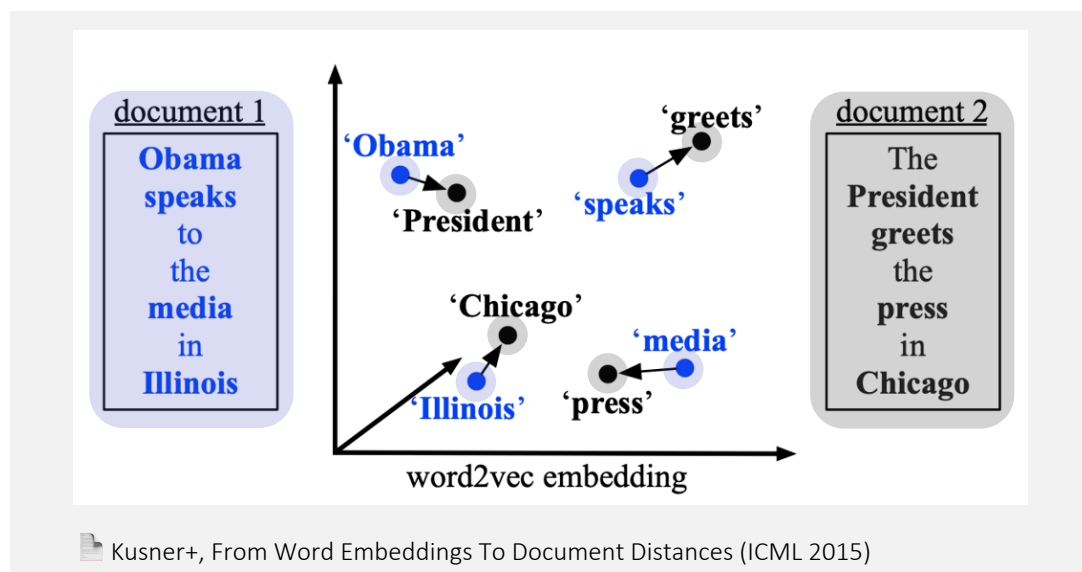
... Australian doctors Robin Warren and Barry Marshall have received the 2005 Nobel Prize in ...

 Sultan+, DLS\$@\$CU: Sentence Similarity from Word Alignment and Semantic Vector Composition (SemEval 2015)

Word Mover's Distance [Kusner+'15]

文類似度 $\leftrightarrow$ 最適輸送コスト

- 単語ベクトル空間上で点群間の最適輸送コストを考える
 - 文1：単語埋込空間で重みが乗っている位置 (before)
 - 文2：単語埋込空間の重みが乗っている位置 (after)



Word Mover's Distance [Kusner+'15]

文類似度 $\leftrightarrow$ 最適輸送コスト

- アイデア：単語ベクトル空間を地図だと思ってみる
 - 文1：単語埋込空間の工場の位置
 - 文2：単語埋込空間のデパートの位置

意味が近い単語同士（工場とデパート）は輸送コストが低い



Kusner+, From Word Embeddings To Document Distances (ICML 2015)

- 意味の近さの問題が輸送コストの大きさの問題に置き換わる
 - 文1, 文2: テキストの意味が似ている $\leftrightarrow$ 低コストで輸送しきれる

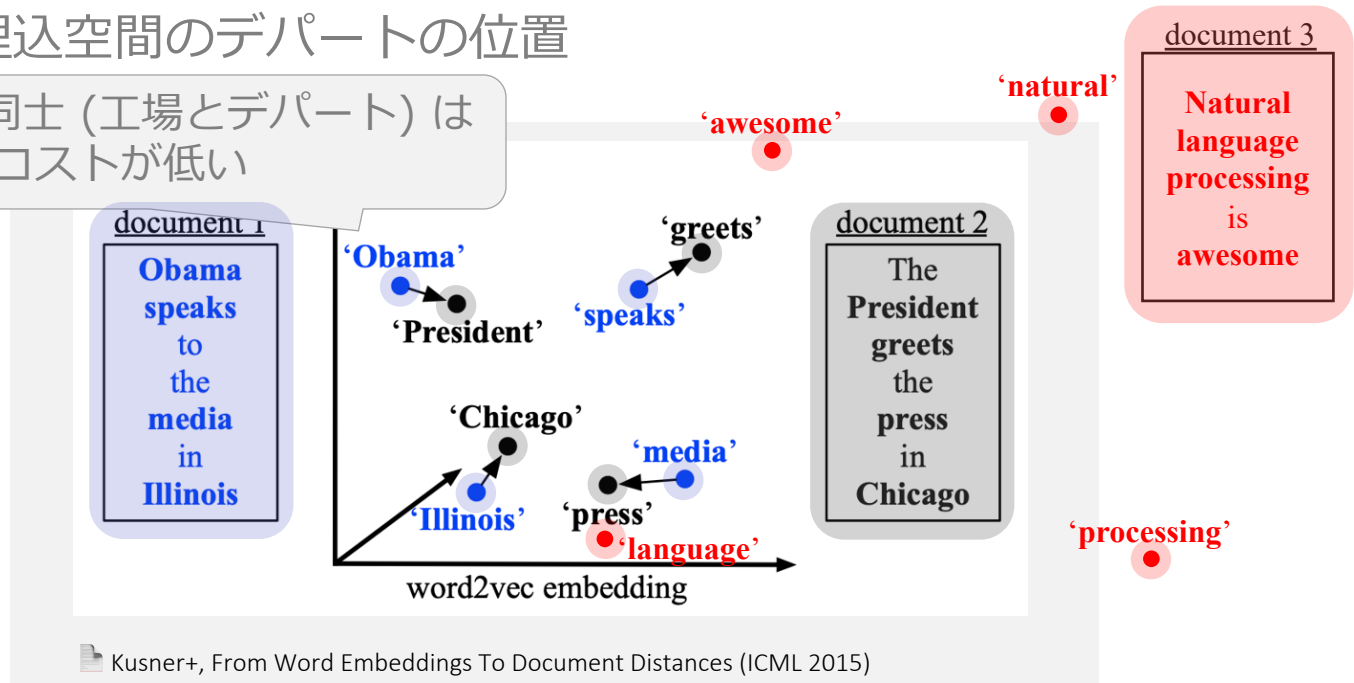
Word Mover's Distance [Kusner+'15]

文類似度 $\leftrightarrow$ 最適輸送コスト

- アイデア：単語ベクトル空間を地図だと思ってみる

- 文1：単語埋込空間の工場の位置
- 文2：単語埋込空間のデパートの位置

意味が近い単語同士（工場とデパート）は
輸送コストが低い



- 意味の近さの問題が輸送コストの大きさの問題に置き換わる

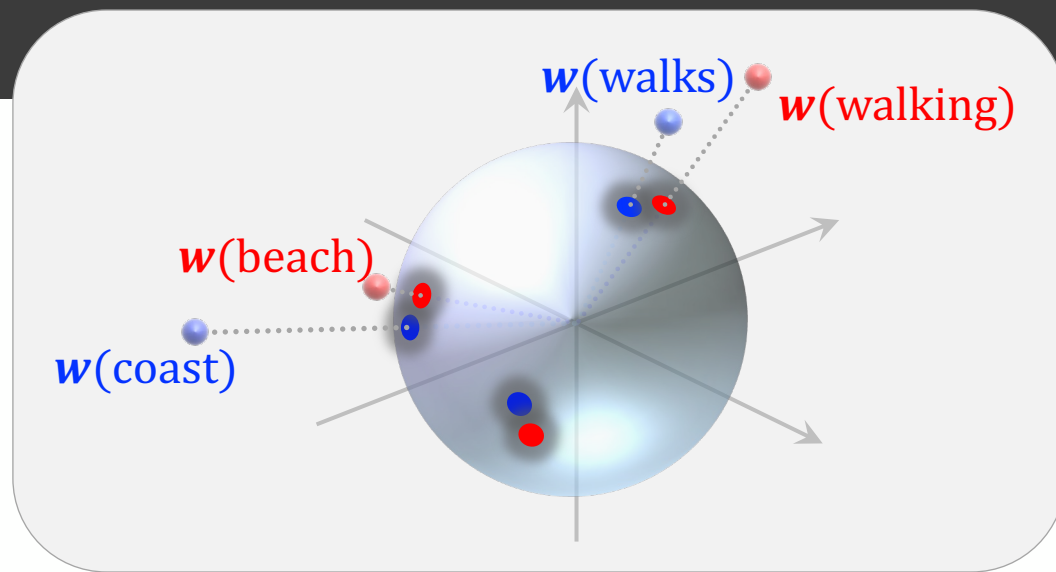
- 文1, 文2: テキストの意味が似ている $\leftrightarrow$ 低コストで輸送しきれる
- 文1, 文3: テキストの意味が似ていない $\leftrightarrow$ 高い輸送コストがかかる

影同士を輸送する

Word Rotator's Distance [Yokoi+'20]

文類似度 $\leftrightarrow$ 最適輸送コスト

- 提案：球面の上にベクトルの影を落として、球面上で影を輸送する



- 実験：文類似度推定の性能が跳ね上がる

	STS'15	STS-B	Twitter
GloVe – Additive Composition			
GloVe [†]	56.08	45.57	29.35
GloVe + WR [†] (Arora et al., 2017)	67.74	62.85	40.03
GloVe + SUP [†] (Ethayarajh, 2018)	74.38	71.03	50.24
GloVe – Considering Word Alignment			
WMD [†] (Kusner et al., 2015)	67.11	52.19	45.04
WRD	68.80	54.03	43.86
WRD + VC(SWC)	<u>77.63</u>	<u>73.14</u>	<u>56.81</u>

お気持ち

重みと輸送コストをどう与えると“自然”か

- **Word Mover's Distance** [Kusner+'15]

自然言語処理の言葉に直すと……
「文の意味を考える際の
各単語の重要度は同じ」

- 確率分布 : $\mathbf{a} = (1/n, \dots, 1/n)$, 確率分布 : $\mathbf{b} = (1/m, \dots, 1/m)$

- 輸送コスト : $\mathbf{C}_{ij} = \|\mathbf{w}_i - \mathbf{w}'_j\|_2$

- 文の非類似度のモデル : $W_1(\mathbf{a}, \mathbf{b}; \mathbf{C})$

自然言語処理の言葉に直すと……
「単語の類似度は単語ベクトルの
間の距離で測る」

おそらく数理科学の人が
「ベクトルの集まりの違いを最適輸送で
測りましょう」となったときの
もっとも自然な初手

Distance [Yokoi+'20]

$\mathbf{a} \propto (\|\mathbf{w}_1\|_2, \dots, \|\mathbf{w}_n\|_2)$, $\mathbf{b} \propto (\|\mathbf{w}'_1\|_2, \dots, \|\mathbf{w}'_m\|_2)$

- 輸送コスト : $\mathbf{C}_{ij} = 1 - \cos(\mathbf{w}_i, \mathbf{w}'_j)$

- 文の非類似度のモデル : $OT(\mathbf{a}, \mathbf{b}; \mathbf{C})$

📄 Kusner+, From Word Embeddings To Document Distances (ICML 2015)

📄 Yokoi+, Word Rotator's Distance (EMNLP 2020)

📄 Oyama+, Norm of Word Embedding Encodes Information Gain (EMNLP 2023)

📄 Yokoi+, Zipfian Whitening (NeurIPS 2024)

重みと輸送コストをどう与えると“自然”か

• Word Mover's Distance [Kusner+'15]

「言語の意味がベクトル空間にどう転写されているか」を考えると
こちらの方がはるかに自然

確率分布 : $b = (1/m, \dots, 1/m)$

$b; C$

単語ベクトルの長さには
単語の重要度が近似的に埋め込まれている
[Oyama+'23], [Yokoi+'24]

• 変更例 : Word Rotator's Distance [Yokoi+'20]

- 確率分布 : $a \propto (\|w_1\|_2, \dots, \|w_n\|_2)$, $b \propto (\|w'_1\|_2, \dots, \|w'_m\|_2)$
- 輸送コスト : $C_{ij} = 1 - \cos(w_i, w'_j)$
- 文の非類似度のモデル : $OT(a, b; C)$

なす角 が単語の非類似度として有用

Kusner+, From Word Embeddings To Document Distances (ICML 2015)

Yokoi+, Word Rotator's Distance (EMNLP 2020)

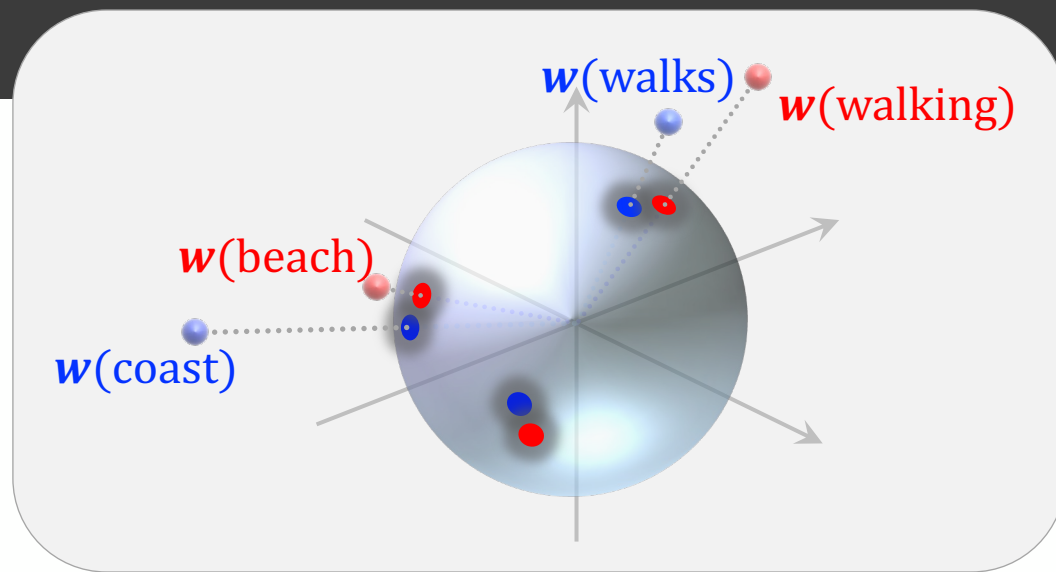
Oyama+, Norm of Word Embedding Encodes Information Gain (EMNLP 2023)

Yokoi+, Zipfian Whitening (NeurIPS 2024)

Word Rotator's Distance [Yokoi+'20]

文類似度 $\leftrightarrow$ 最適輸送コスト

- 提案：球面の上にベクトルの影を落として、球面上で影を輸送する



- 実験：文類似度推定の性能が跳ね上がる

	STS'15	STS-B	Twitter
GloVe – Additive Composition			
GloVe [†]	56.08	45.57	29.35
GloVe + WR [†] (Arora et al., 2017)	67.74	62.85	40.03
GloVe + SUP [†] (Ethayarajh, 2018)	74.38	71.03	50.24
GloVe – Considering Word Alignment			
WMD [†] (Kusner et al., 2015)	67.11	52.19	45.04
WRD	68.80	54.03	43.86
WRD + VC(SWC)	<u>77.63</u>	<u>73.14</u>	<u>56.81</u>

「対応する単語がない」 に対応する

テキストの類似度測定

- ふたつの**テキストの類似度**の測定
 - 基本指針のひとつ：**要素単語の重なるの度合い**を測る

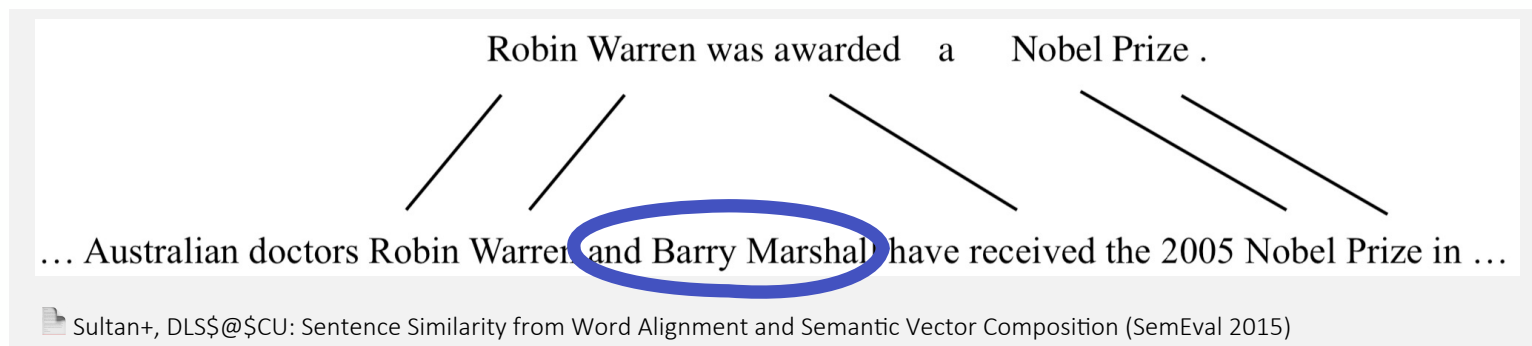
Robin Warren was awarded a Nobel Prize .

... Australian doctors Robin Warren and Barry Marshall have received the 2005 Nobel Prize in ...

 Sultan+, DLS\$@\$CU: Sentence Similarity from Word Alignment and Semantic Vector Composition (SemEval 2015)

テキストの類似度測定

- ふたつの**テキストの類似度**の測定
 - 基本指針のひとつ：**要素単語の重なり**の度合いを測る



「輸送してはいけない」単語もありそう

モチベーション:

一部だけマッチングさせたい

- 長さの大きく異なる文を比較したい [Wang+'20]

文B と 文C がともに否定的な意味であることを計算したければ

“awful” と “sad” だけマッチングできれば十分



Wang+, Robust Document Distance with Wasserstein-Fisher-Rao metric (ACML 2020), Figure 1, 改変

文C 内のほかの単語たち (“key”, “restaurant”, ...) は無視したい
「運びませんでした, おしまい」にしたい

モチベーション: 一部だけマッチングさせたい

モチベーションとなる例
をもうひとつ

- 関連文書を当てたい [Swanson+'20]

これら2つの質問投稿が
関連した内容だとわかるのは

Can I find duplicate songs with different names?

I have so many duplicate songs but they have different names.

◆ Is there an application I can use to find and delete the duplicates?

◆ How to find (and delete) duplicate files?

I have a largish music collection and there are some duplicates in

ere. Is there any way to find duplicate files. At a minimum by doing
hash and seeing if two files have the same hash. ... I'm happy using
command line if that is the easiest way.

緑の文同士 が対応して
いて 青の文同士 も対応
しているから

Swanson+, Rationalizing Text Matching: Learning Sparse Alignments via Optimal Transport (ACL 2020)

他の文たちについては
相手方の質問投稿内に対応する文
が存在していなくても構わない

「運ばない荷物があっても良い」を考慮した最適輸送 不均衡最適輸送

- (詳細略)

- Optimal Partial Transport [Caffarelli&McCann'10][Figalli'10] :
全体の輸送量の合計に上限を与える
- (エントロピー正則化つき) 不均衡最適輸送 [Chizat+'18 etc.] :
荷物の輸送に過不足が生じたらその分だけペナルティを加える
- [Figalli&Gigli'06] :
領域の境界に, **無限の荷物を飲み込めて・生成できるブラックホール**
を作っておく
- これらの一般化 [Liero+'18][Chizat+'18][Sato+'20] :
 - たとえば Optimal Partial Transport は過不足のコストが L1 (total variation) で入っているとみなせる

 Caffarelli&McCann, Free boundaries in optimal transport and Monge-Ampère obstacle problems (Ann. Math. 2010)

 Figalli, The Optimal Partial Transport Problem (Arch. Ration. Mech. Anal. 2010)

 Chizat+, Scaling algorithms for unbalanced transport problems (Math. Comput. 2018)

 Figalli&Gigli, A new transportation distance between non-negative measures, with applications to gradients flows with Dirichlet boundary conditions (J. Math. Pures Appl. 2006)

 Liero+, Optimal Entropy-Transport problems and a new Hellinger–Kantorovich distance between positive measures (Invent. Math. 2018)

 Chizat+, Unbalanced Optimal Transport: Dynamic and Kantorovich Formulation (J. Funct. Anal. 2018)

 Sato+, Fast Unbalanced Optimal Transport on a Tree (NeurIPS 2020)

参考：最適輸送は，数式を追わなくても使える 便利なツールが出ています

- 利用しやすく安心して使える**ライブラリ (ソルバ)**
 - Python Optimal Transport (POT) (2020)  [GitHub](#)
 - Optimal Transport Tools with JAX (OTT-JAX) (2022)  [GitHub](#)
- **計算技法・理論に関する勉強しやすい教科書・資料**
 - 英：Peyré&Cuturi, Computational Optimal Transport (2019)
 - 和：佐藤，最適輸送の理論とアルゴリズム (2022)

arXiv にも最新版が
上がっています

不均衡最適輸送 for 単語アラインメント

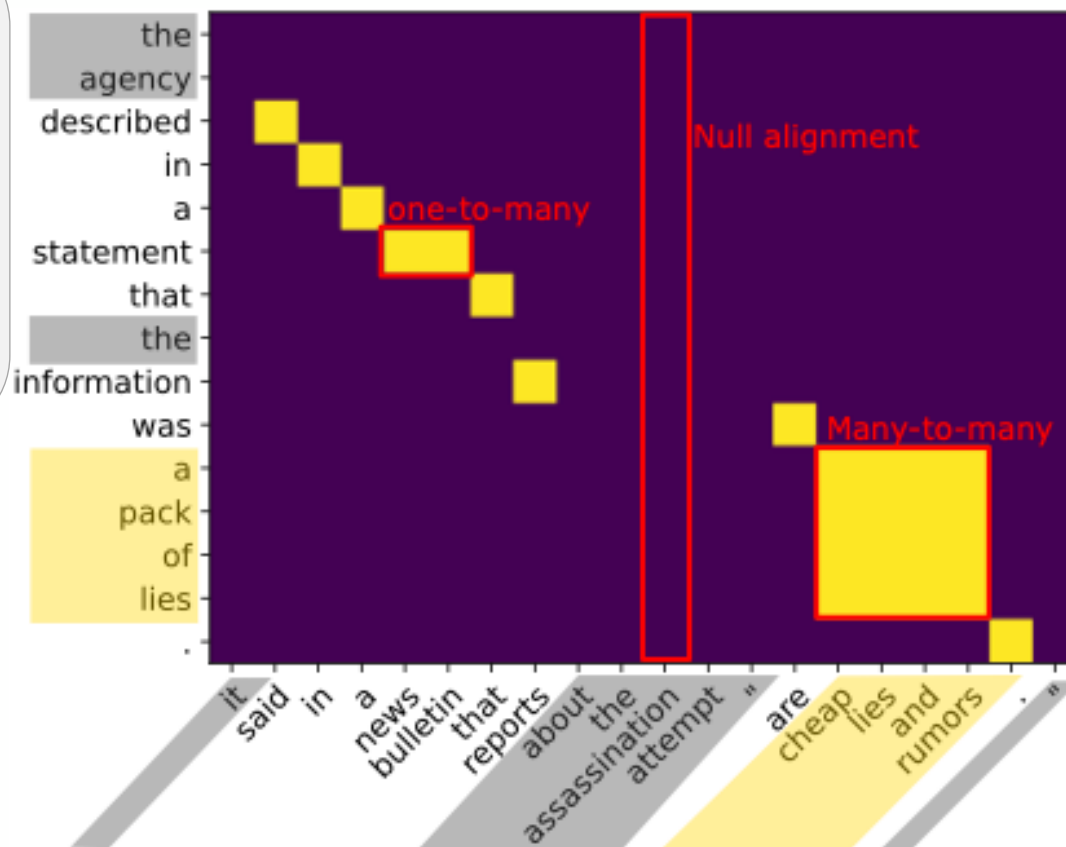
[Arase, Bao, Y, ACL'23]

言い換え表現間の単語アラインメント

簡単な単語で言い直してほしい,
おしゃれな英語にしてほしい, etc.

Q. ChatGPT 氏は正確に
抜け漏れなく言い換えてる…?

アラインメントは
(正確性が気になる一部の) ユーザ
にとって嬉しい説明



不均衡最適輸送 for 単語アラインメント

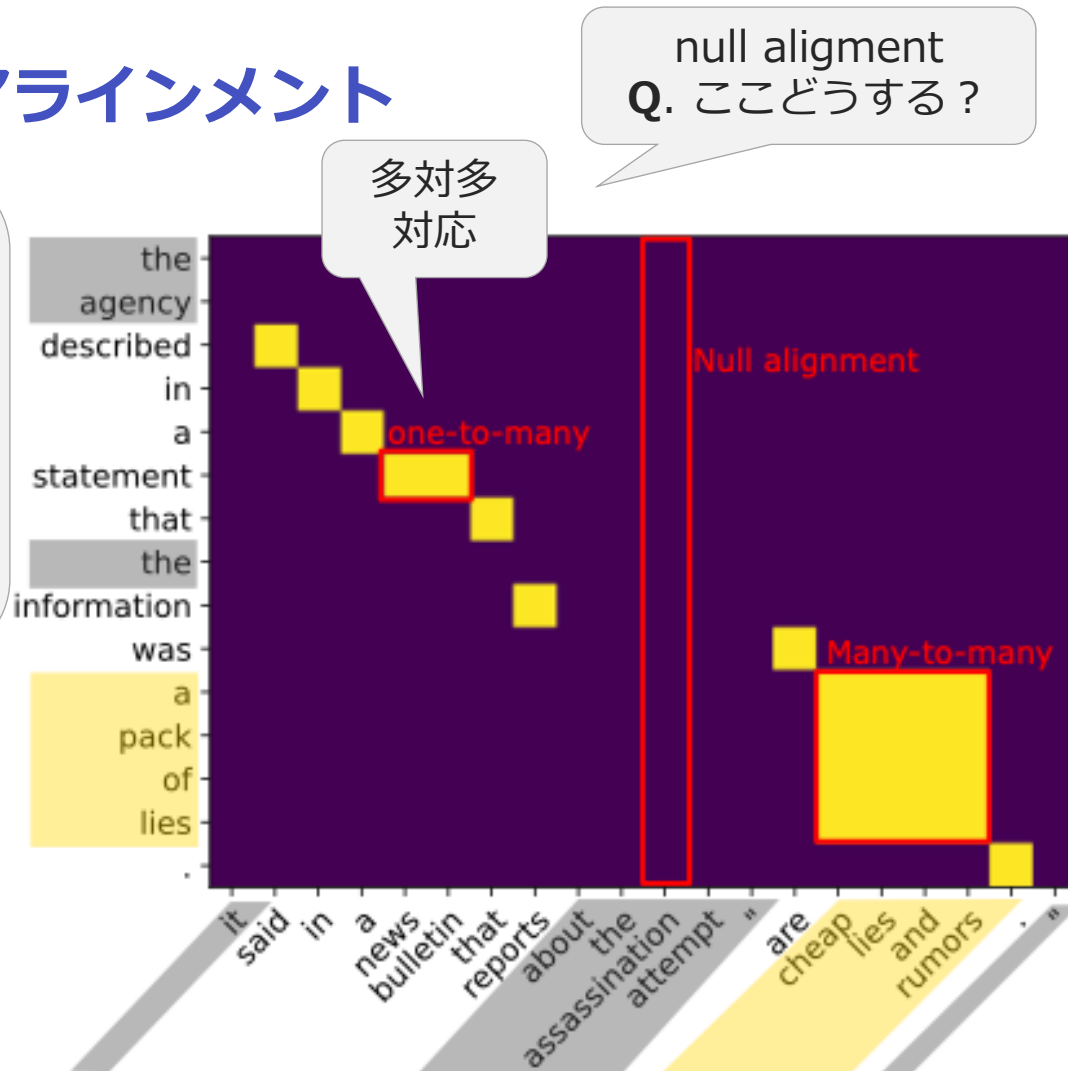
[Arase, Bao, Y, ACL'23]

言い換え表現間の単語アラインメント

簡単な単語で言い直してほしい,
おしゃれな英語にしてほしい, etc.

Q. ChatGPT 氏は正確に
抜け漏れなく言い換えてる…?

アラインメントは
(正確性が気になる一部の) ユーザ
にとって嬉しい説明

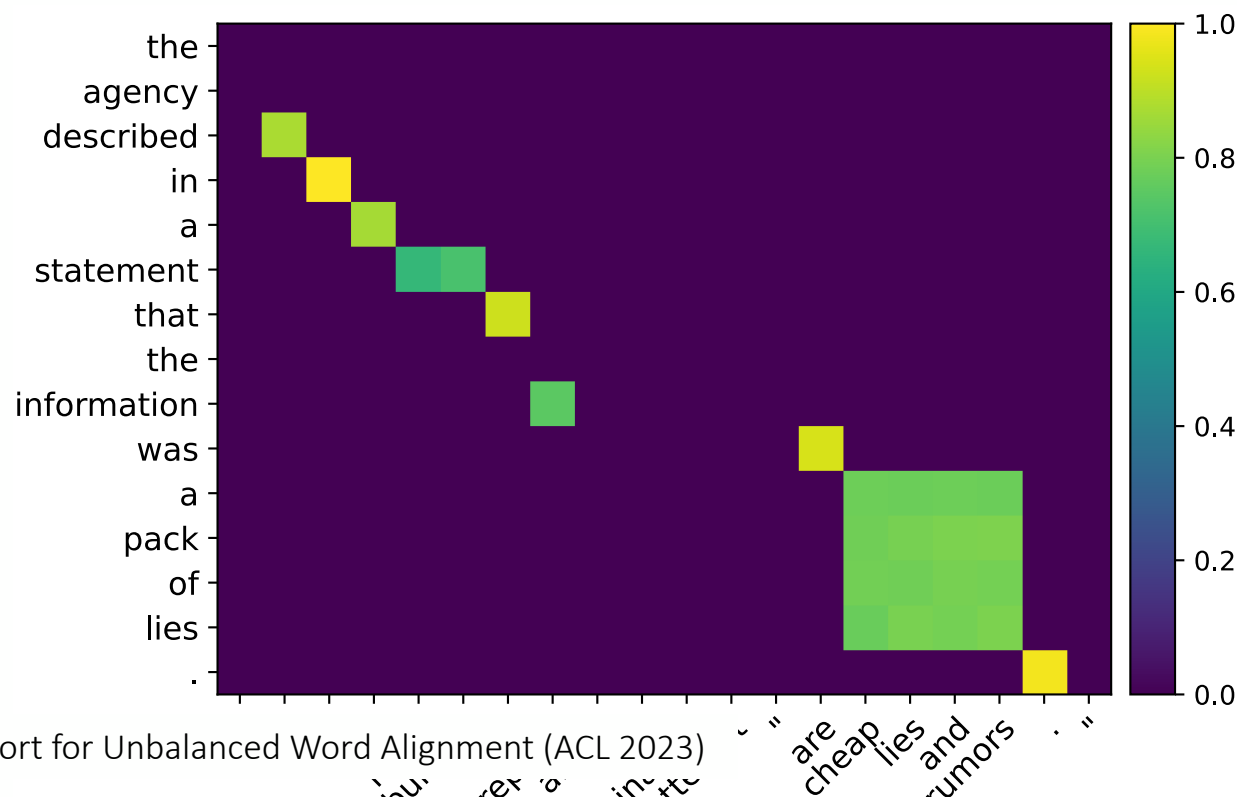


不均衡最適輸送 for 単語アラインメント

[Arase, Bao, Y, ACL'23]

• 言い換え表現間の単語アラインメント

- 1. null alignment があるので**不均衡最適輸送**を採用
- 2. 最適輸送の**最適値**（＝必要な合計コスト＝全体での意味的な違い）ではなく、**最適解**（＝見出される対応関係）を最適化
- → 綺麗に動く



QA

単語 (ないし概念) 集合間の 「かつ」や「または」を 単語ベクトル空間で柔軟かく計算する

単語よりも大きな単位の意味の計算 (2)

概念の共通部分を空間の重なりで計算

[Ishibashi, Y, Sudoh, Nakamura, NAACL'24]

- 言語を使って表される「概念」と, その共通部分
 - Color = {red, orange, ……}, Fruit = {apple, orange, ……}
 - Color $\cap$ Fruit = {orange, ……}

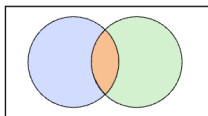
共通部分

- 概念を表す空間の重なりで計算できる

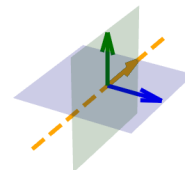
(固い) 概念の世界

(柔らかい) 形の世界

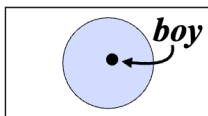
Intersection
 $Color \cap Fruit$



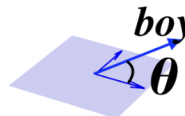
Intersection
 $\mathbb{S}_{Color} \cap \mathbb{S}_{Fruit}$



Set membership
 $\mathbb{1}_{\text{set}}[boy \in Male]$



Subspace indicator function
 $\mathbb{1}_{\text{subspace}}(\mathbf{v}_{boy}, \mathbb{S}_{Male})$



概念の共通部分を空間の重なりで計算

[Ishibashi,Y,Sudoh,Nakamura,NAACL'24]

- 言語を使って表される「概念」とその共通部分

– Color =	S_9	tour	open	golf	...
– Color ∩	$\mathbf{D}^{\text{Intersect}}(100) S_{72}$	poker	casino	gambling	...
	$S_9 \cap S_{72}$	money	won	player	...

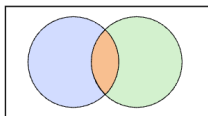
やわらかく「だいたいこのあたり」と計算することができる

- 概念を表す空間の重なりで計算できる

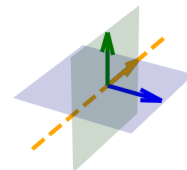
(固い) 概念の世界

(柔らかい) 形の世界

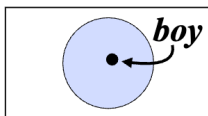
Intersection
 $Color \cap Fruit$



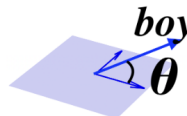
Intersection
 $\mathbb{S}_{Color} \cap \mathbb{S}_{Fruit}$



Set membership
 $\mathbb{1}_{\text{set}}[boy \in Male]$



Subspace indicator function
 $\mathbb{1}_{\text{subspace}}(\mathbf{v}_{boy}, \mathbb{S}_{Male})$



概念の共通部分を空間の重なりで計算

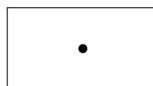
[Ishibashi,Y,Sudoh,Nakamura,NAACL'24]

- やったこと：集合（ものの集まり）とその上の演算を，単語ベクトル空間で柔軟に計算できるようにした

（固い）概念の世界

（柔らかい）形の世界

Element
 $king$

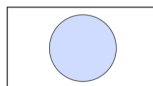


Vector
 $\mathbf{v}_{king}$



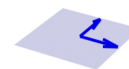
Set

$Male = \{king, man, \dots\}$

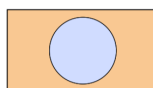


Subspace

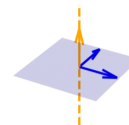
$\mathbb{S}_{Male} = \text{span}(\mathbf{v}_{king}, \mathbf{v}_{man}, \dots)$



Complement
 $\overline{Male}$



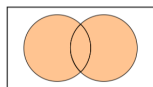
Orthogonal complement
 $(\mathbb{S}_{Male})^\perp$



ふたつの集合を
組み合わせる

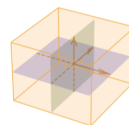
Union

$Male \cup Female$

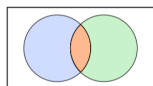


Sum space

$\mathbb{S}_{Male} + \mathbb{S}_{Female}$

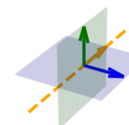


Intersection
 $Color \cap Fruit$



Intersection

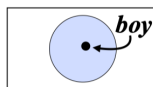
$\mathbb{S}_{Color} \cap \mathbb{S}_{Fruit}$



集合に要素が
含まれているかを
チェックする

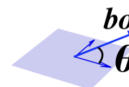
Set membership

$\mathbb{1}_{\text{set}}[boy \in Male]$



Subspace indicator function

$\mathbb{1}_{\text{subspace}}(\mathbf{v}_{boy}, \mathbb{S}_{Male})$



概念の共通部分を空間の重なりで計算

[Ishibashi, Y, Sudoh, Nakamura, NAACL'24]

- やったこと：集合（ものの集まり）とその上の演算を，単語ベクトル空間で柔軟に計算できるようにした

(固い) 概念の世界

(柔軟に) 形の世界

Element

Vector

※ 余談

「集合同士の演算」は**ブール束**という数理構造を持つ

↓

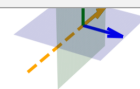
提案法 (= **線形部分集合間の演算**) は，**オーソモジューラー束**というブール束とそっくりの数理構造を持つ

= 単語ベクトル空間の中で
やりたかった集合演算が数学的にも（ほとんど）正確に実現できる

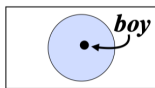
Intersection
 $Color \cap Fruit$



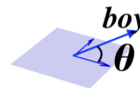
Intersection
 $\mathbb{S}_{Color} \cap \mathbb{S}_{Fruit}$



Set membership
 $\mathbb{1}_{\text{set}}[boy \in Male]$



Subspace indicator function
 $\mathbb{1}_{\text{subspace}}(\mathbf{v}_{boy}, \mathbb{S}_{Male})$



概念の共通部分を空間の重なりで計算

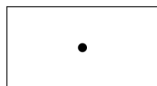
[Ishibashi, Y, Sudoh, Nakamura, NAACL'24]

- やったこと：集合（ものの集まり）とその上の演算を，単語ベクトル空間で柔軟に計算できるようにした

（固い）概念の世界

（柔軟に）形の世界

Element
king

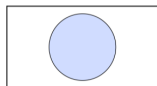


Vector
 v_{king}



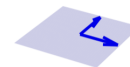
Set

$Male = \{king, man, \dots\}$



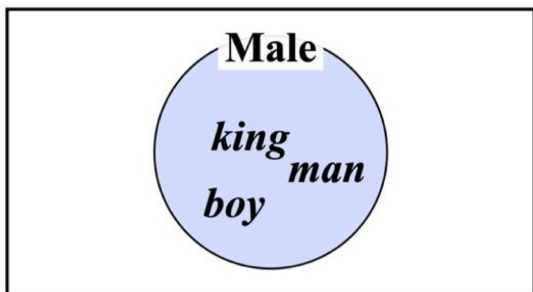
Subspace

$S_{Male} = \text{span}(v_{king}, v_{man}, \dots)$

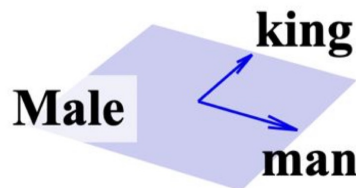


複数の要素を集めて
集合を作る

$Male = \{king, man, \dots\}$



$S_{Male} = \text{span}\{v_{king}, v_{man}, \dots\}$



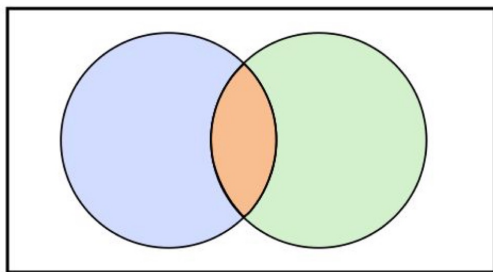
“張る空間”
を作る

概念の共通部分を空間の重なりで計算

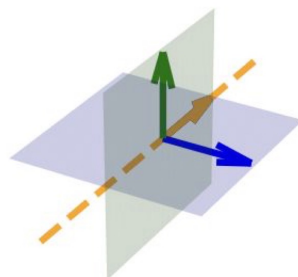
[Ishibashi,Y,Sudoh,Nakamura,NAACL'24]

- やったこと：集合（ものの集まり）とその上の演算を、
た

Color $\cap$ Fruit



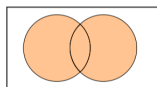
$$\mathbb{S}_{\text{Color}} \cap \mathbb{S}_{\text{Fruit}} \\ = \{ \mathbf{v} \mid \mathbf{v} \in \mathbb{S}_{\text{Color}}, \mathbf{v} \in \mathbb{S}_{\text{Fruit}} \}$$



「かつ」

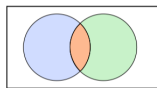
Union

$Male \cup Female$



Intersection

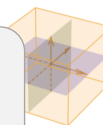
$Color \cap Fruit$



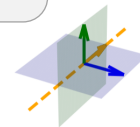
ふたつの空間の
重なり

Intersection

$\mathbb{S}_{\text{Color}} \cap \mathbb{S}_{\text{Fruit}}$

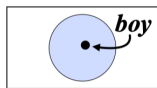


ふたつの集合を
組み合わせる



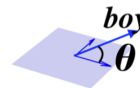
Set membership

$\mathbb{1}_{\text{set}}[boy \in Male]$



Subspace indicator function

$\mathbb{1}_{\text{subspace}}(\mathbf{v}_{\text{boy}}, \mathbb{S}_{\text{Male}})$



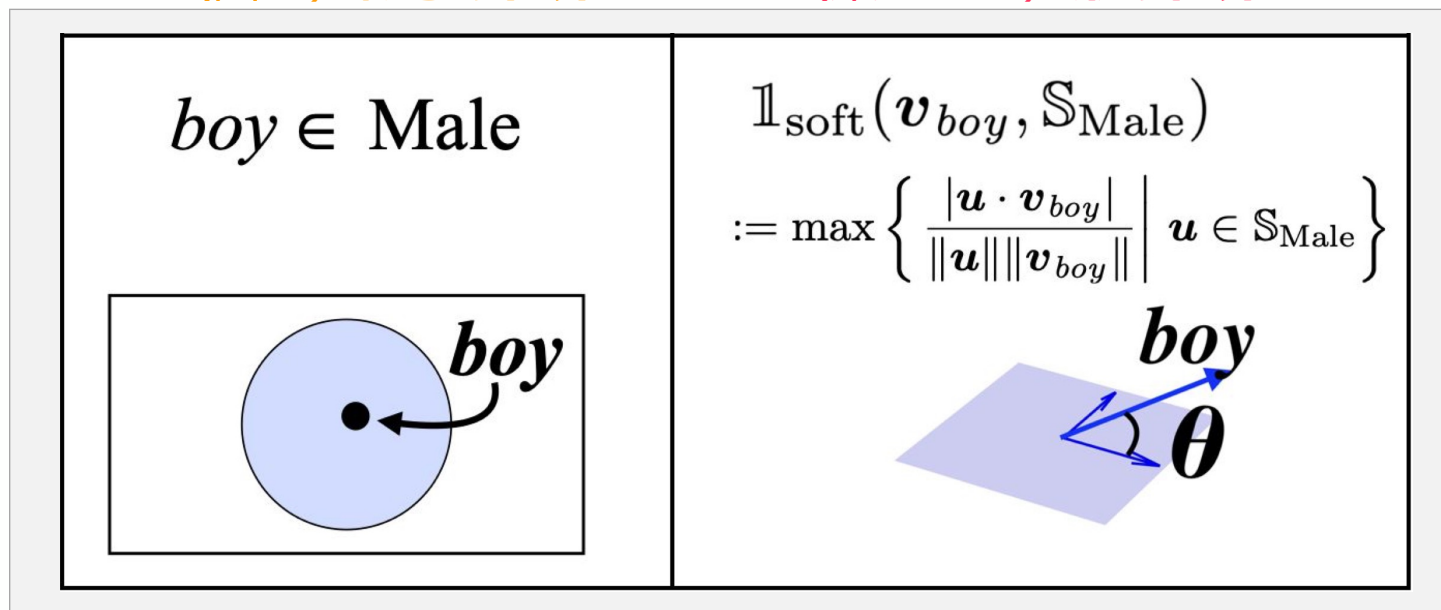
概念の共通部分を空間の重なりで計算

[Ishibashi,Y,Sudoh,Nakamura,NAACL'24]

- やったこと：集合（ものの集まり）とその上の演算を，単語ベクトル空間で柔軟に計算できるようにした

（固い）概念の世界

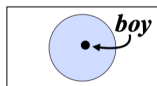
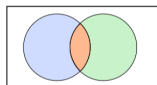
（柔らかい）形の世界



Intersection

「含まれている」

Set membership
 $\mathbb{1}_{\text{set}}[boy \in \text{Male}]$



Intersection

「単語ベクトルと，概念集合を表す空間が，近い」

$\mathbb{1}_{\text{subspace}}(\mathbf{v}_{boy}, \mathbb{S}_{\text{Male}})$



集合に要素が含まれているかを
チェックする

概念の共通部分を空間の重なりで計算

[Ishibashi,Y,Sudoh,Nakamura,NAACL'24]

- 言語を使って表される「概念」とその共通部分

Color =	S_9	tour	open	golf	...
Color ∩	$\mathbf{D}^{\text{Intersect}}(100) S_{72}$	poker	casino	gambling	...
	$S_9 \cap S_{72}$	money	won	player	...

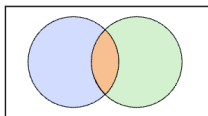
やわらかく「だいたいこのあたり」と計算することができる

- 概念を表す空間の重なりで計算できる

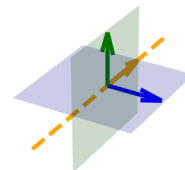
(固い) 概念の世界

(柔らかい) 形の世界

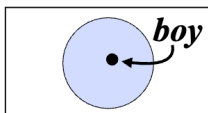
Intersection
 $Color \cap Fruit$



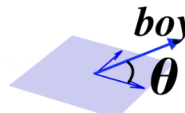
Intersection
 $\mathbb{S}_{Color} \cap \mathbb{S}_{Fruit}$



Set membership
 $\mathbb{1}_{\text{set}}[boy \in Male]$



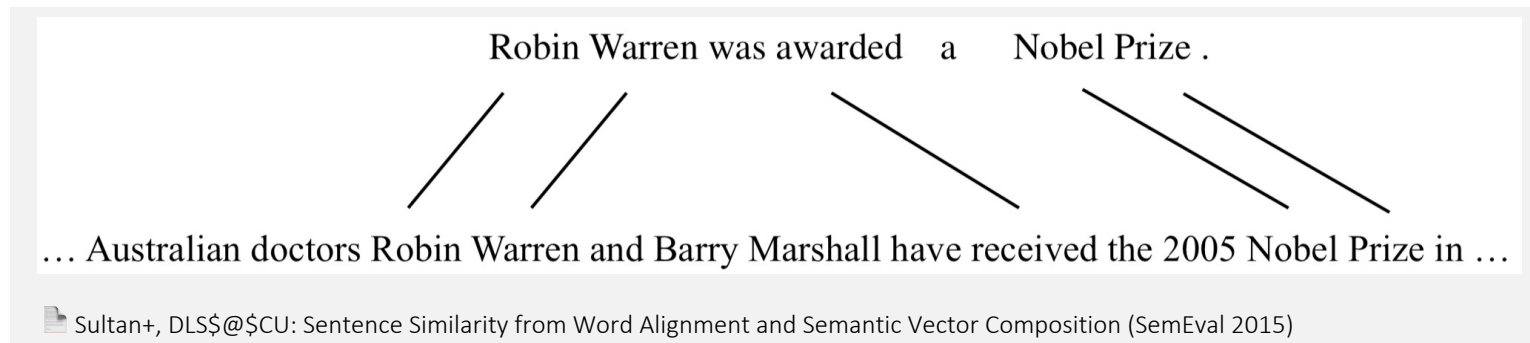
Subspace indicator function
 $\mathbb{1}_{\text{subspace}}(\mathbf{v}_{boy}, \mathbb{S}_{Male})$



概念の共通部分を空間の重なりで計算

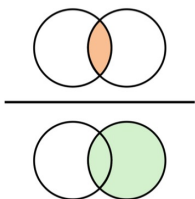
[Ishibashi,Y,Sudoh,Nakamura,NAACL'24]

- さらに： **文の意味的な類似度**の粗い計算にも使える
 - 基本指針のひとつ： **要素単語の重なり**の**度合い**を測る

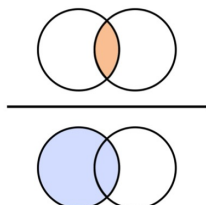


- 集合の類似度だと思うと…… 「F値」を測れば良さそう

Precision



Recall



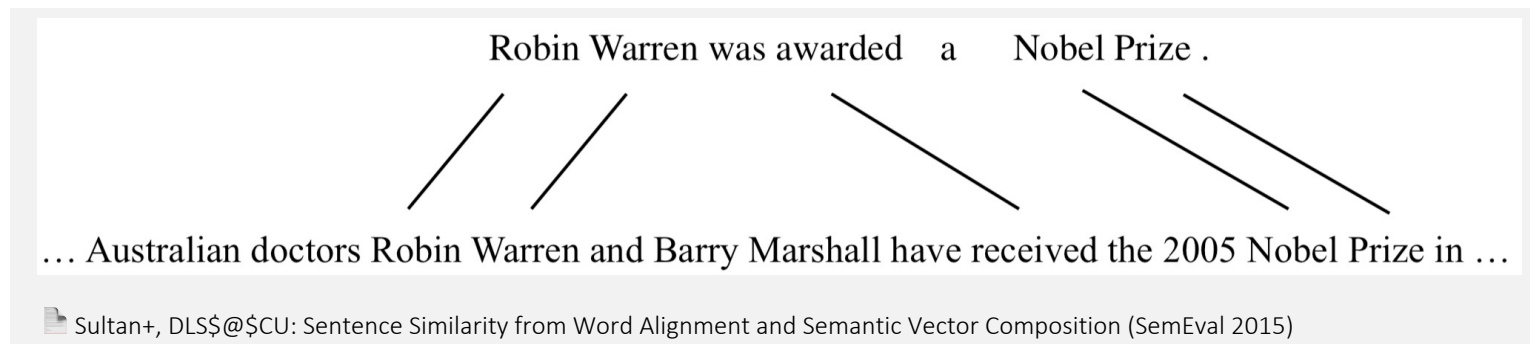
F-score

$$2 \frac{\frac{\text{orange intersection}}{\text{orange area}} \times \frac{\text{orange intersection}}{\text{blue area}}}{\frac{\text{orange intersection}}{\text{orange area}} + \frac{\text{orange intersection}}{\text{blue area}}}$$

概念の共通部分を空間の重なりで計算

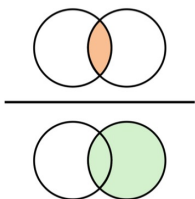
[Ishibashi,Y,Sudoh,Nakamura,NAACL'24]

- さらに：**文の意味的な類似度**の粗い計算にも使える
 - 基本指針のひとつ：**要素単語の重なりの度合い**を測る

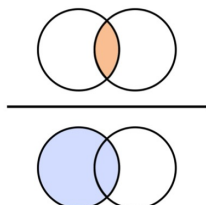


- 集合の類似度だと思ったら…… 「F値」を測れば良さそう

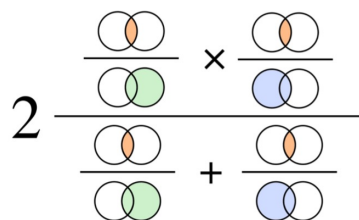
Precision



Recall



F-score



提案法を使うと、
すべて**単語ベクトル空間**
の言葉に翻訳できる

=**柔らかい集合類似度**
を計算できる

概念の共通部分を空間の重なりで計算

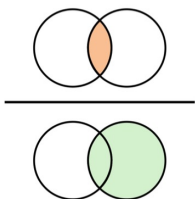
[Ishibashi,Y,Sudoh,Nakamura,NAACL'24]

Method	Metric	Weighting	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
CLS-cos	-	-	.215	.321	.213	.379	.442	.203	.424	.314
Avg-cos	-	-	.309	.599★	.477					
WMD	-	-	.238	.443	.389					
WRD	-	-	.241	.502	.410					
DynaMax	-	-	.322	.518	.432					
BERTScore	<i>F</i>	-	.312	.546	.450	.602	.636	.446	.553	.506
	<i>P</i>	-	.261	.532	.462	.576	.622	.443	.559	.494
	<i>R</i>	-	.350	.527	.416	.602	.623	.430	.522	.496
SubspaceBERTScore	<i>F</i>	-	.335	.573	.476	.610	.650	.479	.562	.526
	<i>P</i>	-	.282	.550	.488	.580	.630	.475	.568	.511
	<i>R</i>	-	.369★	.552	.436	.611	.639	.462	.530	.514

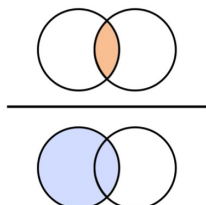
提案法を使うと、
深層学習に基づく「文類似度推定器」
の性能が上がる

- 集合の類似度だと思ったら…… 「F値」を測れば良さそう

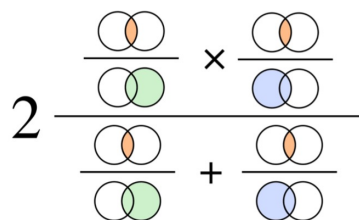
Precision



Recall



F-score



提案法を使うと、
すべて**単語ベクトル空間**
の言葉に翻訳できる

=**柔らかい集合類似度**
を計算できる

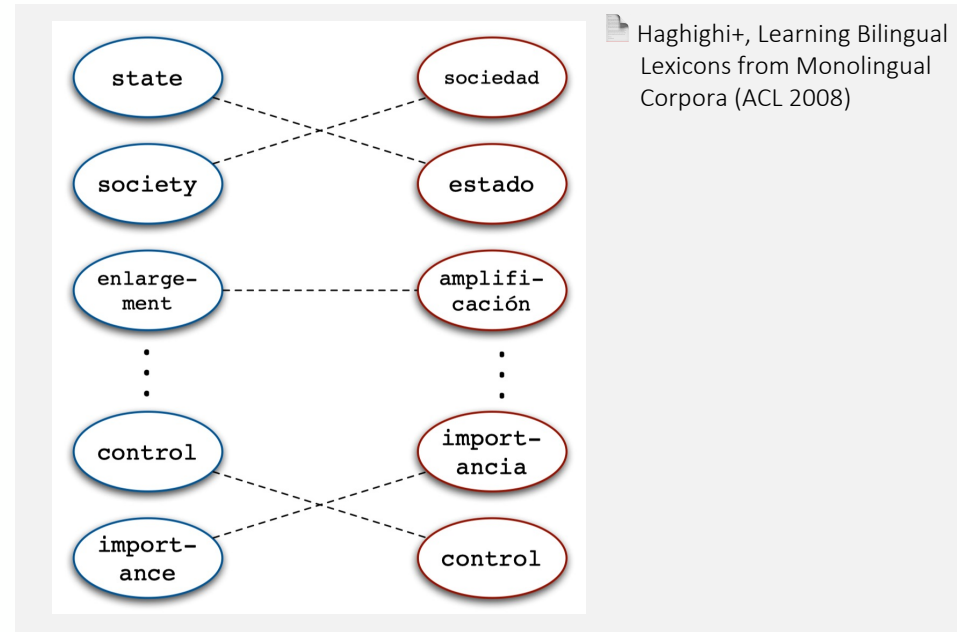
言語間の対訳辞書を 最適輸送で柔らかく見つける

単語よりも大きな単位の意味の計算 (3)

単語ベクトルを使うと対訳辞書を自動生成できる

[Alvarez-Melis&Jaakkola'18]

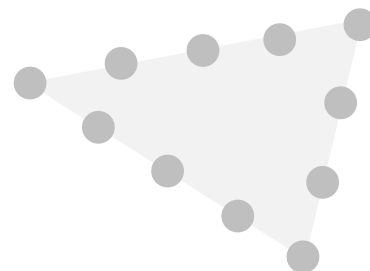
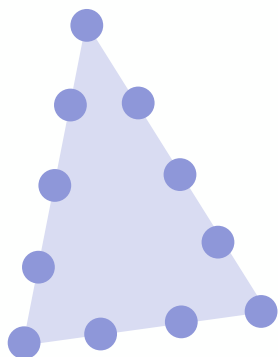
- 問題：教師なし辞書構築
 - “猫” $\leftrightarrow$ “cat”
 - “科学” $\leftrightarrow$ “science”



- 二つの言語の単語ベクトル空間を最適輸送の一種を使って自動で重ねることができる
(= 自動で対訳辞書を作ることができる)

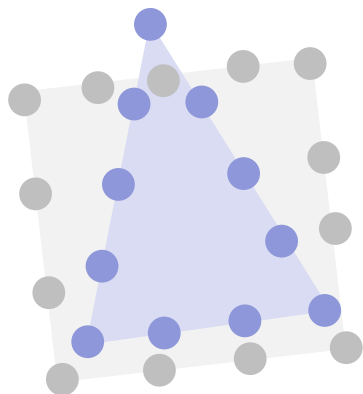
全体の類似性 = 形の類似性 という考え方

- 全体が似ている = **形**が似ている という考えかた

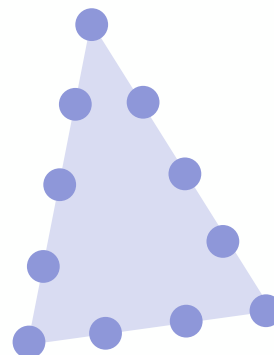


- 素直に**最適輸送**を使うと……

🤔 輸送コスト：小

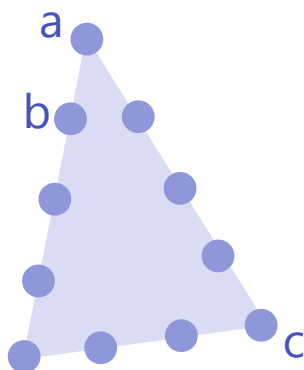


🤔 輸送コスト：大



全体の類似性 = 形の類似性

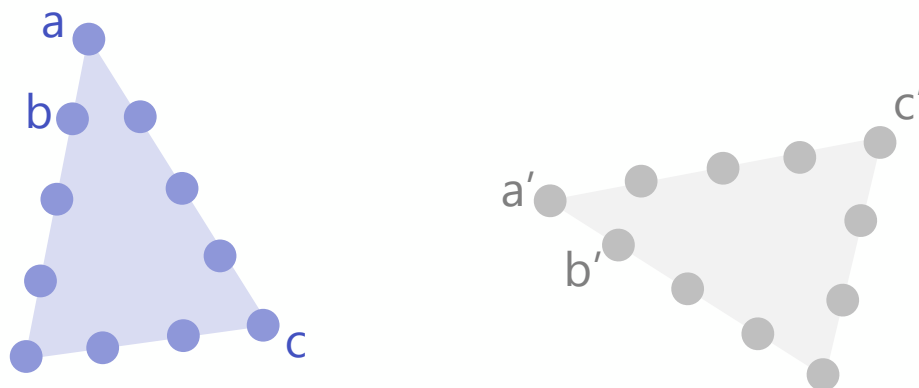
- なぜこれらの点群が「似ている」と思っているのか



- 全体の形状を保存する良いアラインメントが存在する

全体の類似性 = 形の類似性

- なぜこれらの点群が「似ている」と思っているのか



- 全体の形状を保存する良いアラインメントが存在する
 - 近い点对は近い点对に対応 ... $d(a,b) \approx d(a',b')$
 - 遠い点对は遠い点对に対応 ... $d(a,c) \approx d(a',c')$
- ※合同 = 等長変換 (合同変換) が存在

アイデア

グロモフ=ワッサーズタイン距離 [Mémoli'11]

内部構造に着目した最適輸送コスト

• グロモフ=ワッサーズタイン距離 (詳細略)

– 入力

– 荷物の量分布 × 2 (工場, デパート) : $\mathbf{a} \in \mathbb{R}_+^n$, $\mathbf{b} \in \mathbb{R}_+^m$

– コスト行列 × 2 (**工場同士**, **デパート同士**の位置関係) : $\mathbf{D} \in \mathbb{R}_+^{n \times n}$, $\mathbf{D}' \in \mathbb{R}_+^{m \times m}$

cf. 最適輸送で使うのは
工場とデパート の距離

– 最適化

良いアラインメント
を探す

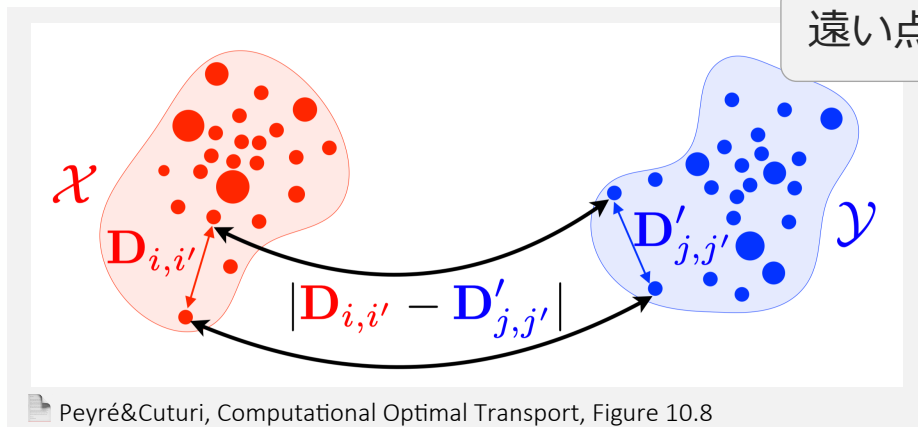
近い点ペアは

移動後も近い

$$\text{GW}((\mathbf{a}, \mathbf{D}), (\mathbf{b}, \mathbf{D}'))^2 \stackrel{\text{def.}}{=} \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \sum_{i,j,i',j'} |\mathbf{D}_{i,i'} - \mathbf{D}'_{j,j'}|^2 \mathbf{P}_{i,j} \mathbf{P}_{i',j'}$$

遠い点ペアは

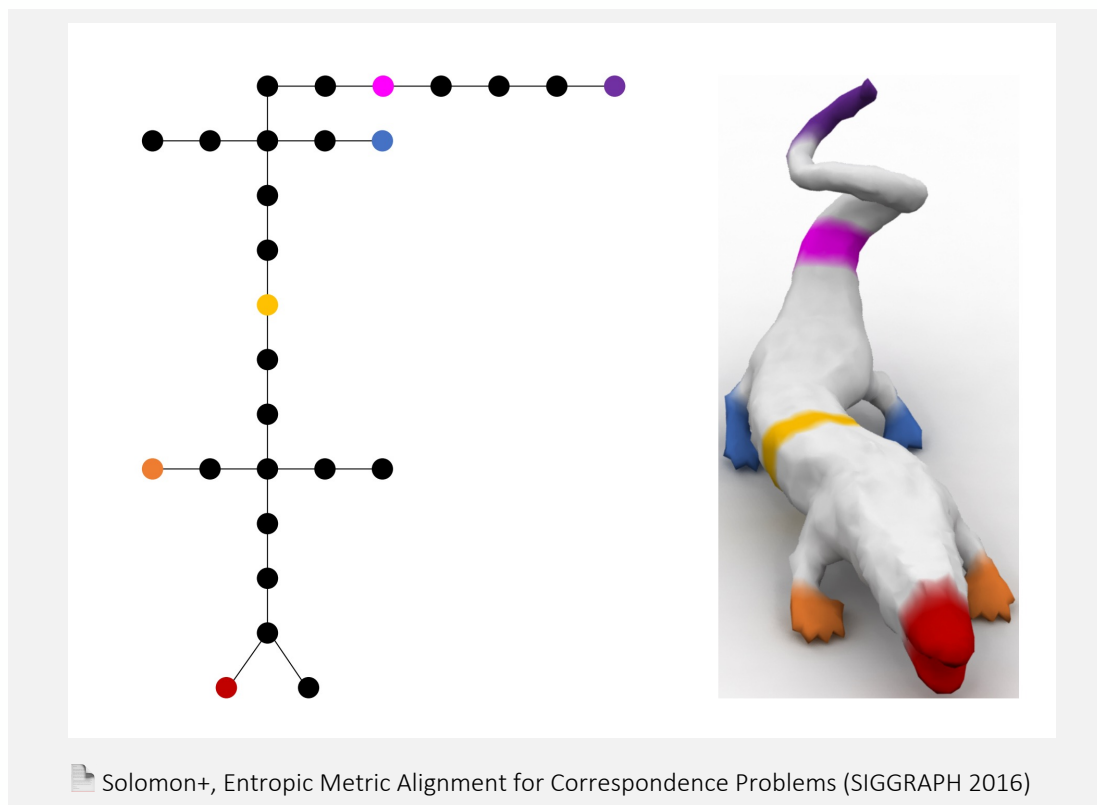
移動後も遠い



グロモフ=ワッサーズタイン距離 [Mémoli'11]

内部構造に着目した最適輸送コスト

- 異なる空間に置かれた点群のマッチングができる



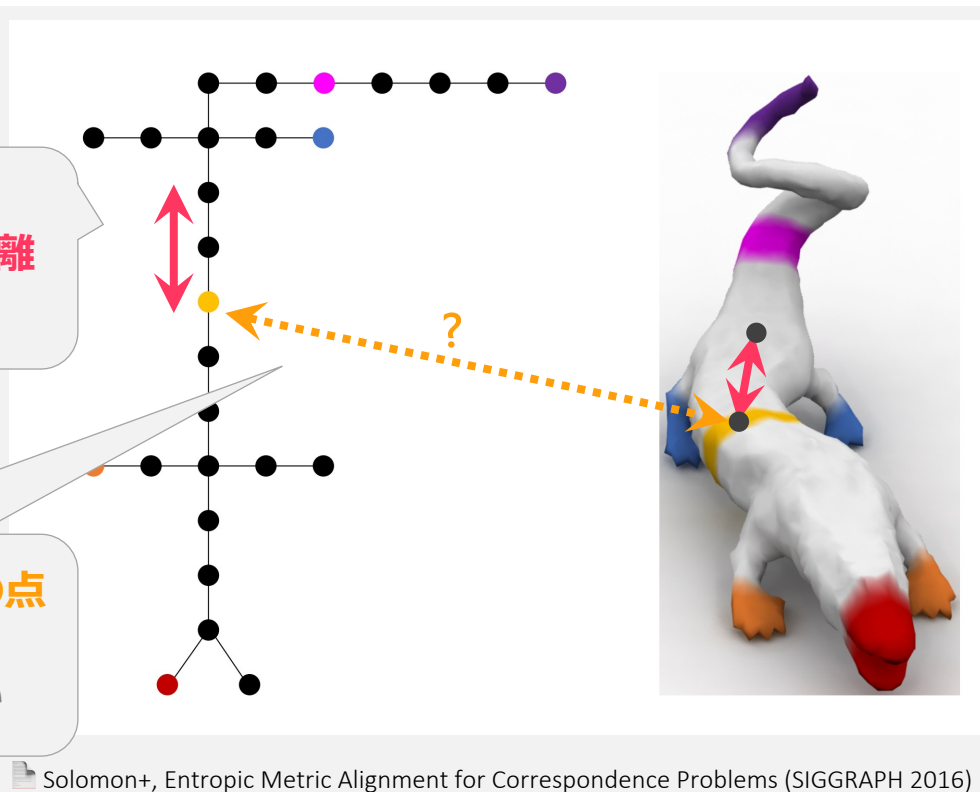
グロモフ=ワッサーズタイン距離 [Mémoli'11]

内部構造に着目した最適輸送コスト

- 異なる空間に置かれた点群のマッチングができる

グラフ, 物体表面
それぞれで**内部の点間距離**
が定まっていれば良い

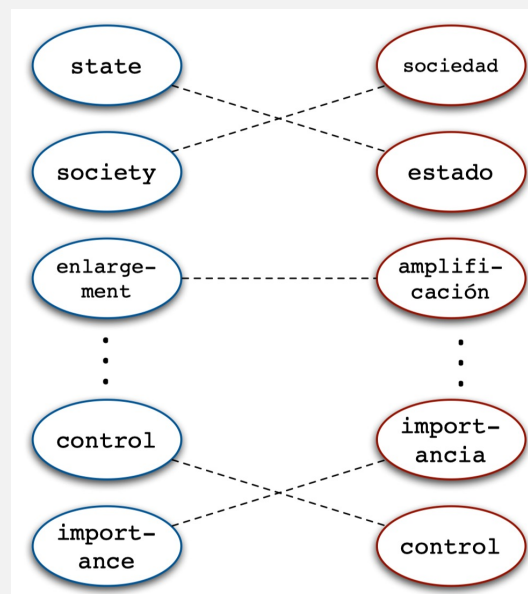
グラフノードと物体表面の点
の距離 (とは...) は
定まっていなくても良い



グロモフ=ワッサーズタイン距離 for 自然言語処理

[Alvarez-Melis&Jaakkola'18]

- 問題：教師なし辞書構築
 - “猫” $\leftrightarrow$ “cat”
 - “科学” $\leftrightarrow$ “science”



Haghighi+, Learning Bilingual Lexicons from Monolingual Corpora (ACL 2008)

グロモフ=ワッサーズタイン距離 for 自然言語処理

[Alvarez-Melis&Jaakkola'18]

- 異なる言語の埋込集合は、回転（直交変換）でよく重なり合うことが知られている [Xing+'15]

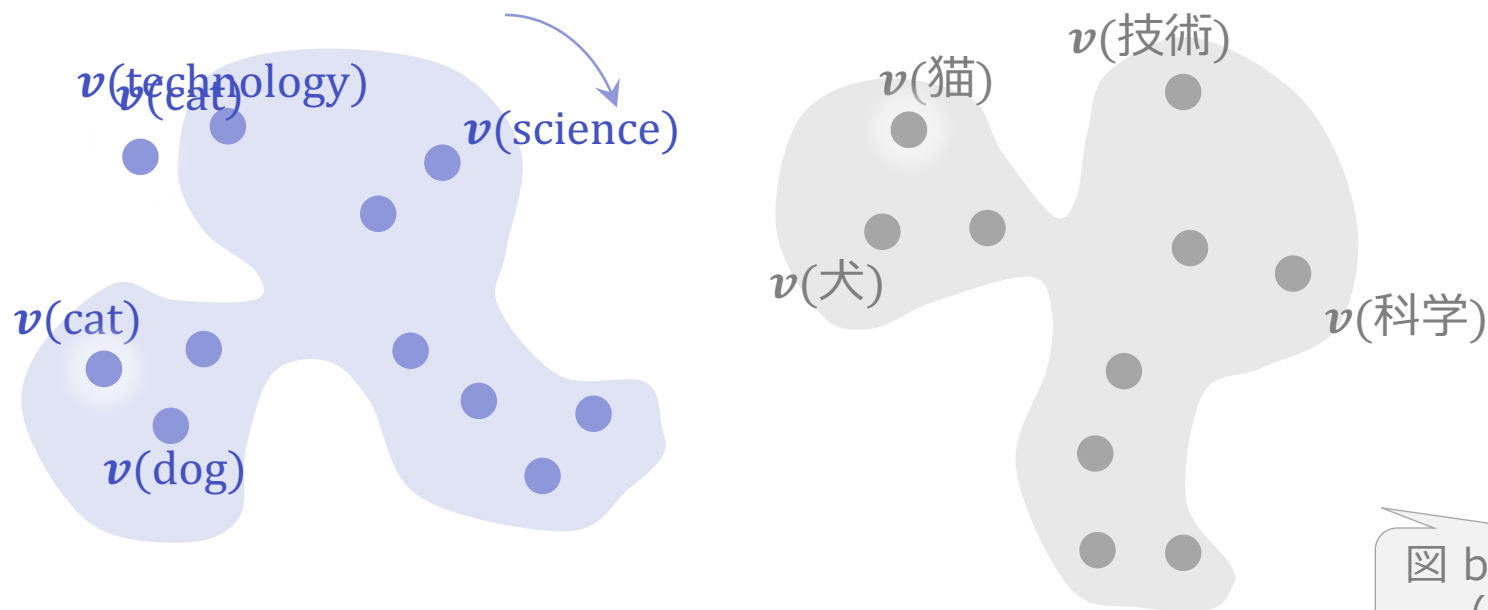


図 by 赤間さん
(東北大)

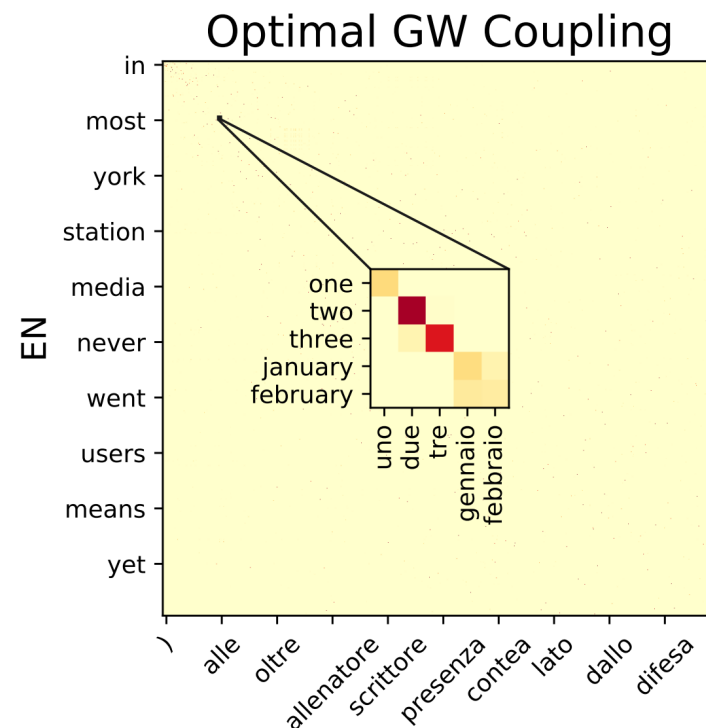
- $d(V_{\text{cat}}, V_{\text{dog}}) \approx d(V_{\text{猫}}, V_{\text{犬}})$, $d(V_{\text{cat}}, V_{\text{science}}) \approx d(V_{\text{猫}}, V_{\text{科学}})$
- 位置関係を保存するようにアラインメント… アレの出番では…**

グロモフ=ワッサーズタイン距離 for 自然言語処理

[Alvarez-Melis&Jaakkola'18]

- アイデア：2つの言語 (単語ベクトル集合) をグロモフ=ワッサーズタイン距離でマッチング
- 結構できてしまう

二つの言語の単語ベクトル空間を
最適輸送の一種を使って
自動で重ねることができる
(= 対訳辞書を作ることができる)

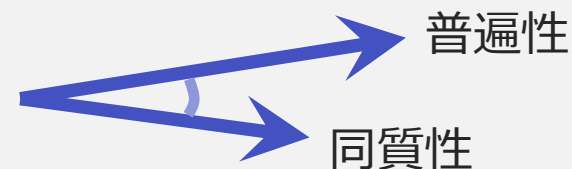


コーパス探索のための 柔軟いコンコーダンスを作る

単語よりも大きな単位の意味の計算 (4)

似ているフレーズを矢印の似かたで検索

[Deguchi+ICLR'25]



SoftMatcha 

言語の普遍性

×

Search

Results

Hits: 8

Search Time: 0.089 sec

リンゼイ・J・ウェイリー(2006).言語類型論入門— ^{1.00 1.00 1.00 1.00}言語の普遍性と多様性, 岩波書店(原書an Introduction to typologyは1996年)

たものの、かつてから宗教によって分断されてきたそれまでの民族規定に ^{1.00 1.00 0.55 1.00}変えて、言語の同質性によって南スラヴの統一を主張したものであるとした。しかしながら、wachtelの見解には次のような異議がもたれ得る。

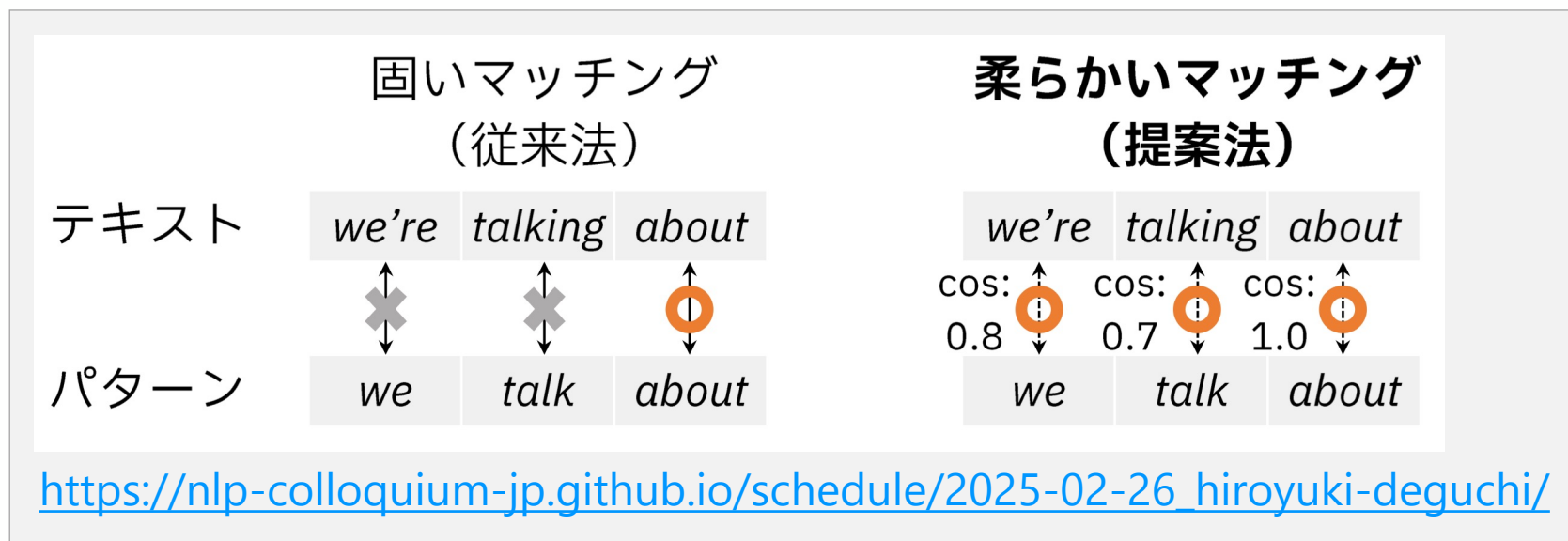
^{0.54 1.00 0.58 1.00}英語の多様性 "diversity"の語源は、ラテン語ではdiverstiasに求められ、この言葉は、最初には、一致可能なものに反すること、矛盾、対立、不一致、といった消極的な意味を有したが、第二義的に、相違、多様、

10億語レベルの巨大言語データから、1秒以内に「ちょっと似ているもの」を全列挙できる

似ているフレーズを矢印の似かたで検索

[Deguchi+ICLR'25]

- やりたいこと：単語ベクトルを使って柔らかくマッチ



- Q. これをパターンマッチ（クエリテキストを巨大な文書群で照合する形式）でどう実現？
- Q. 十分高速に動かすには？

SoftMatcha [Deguchi+, ICLR'25]

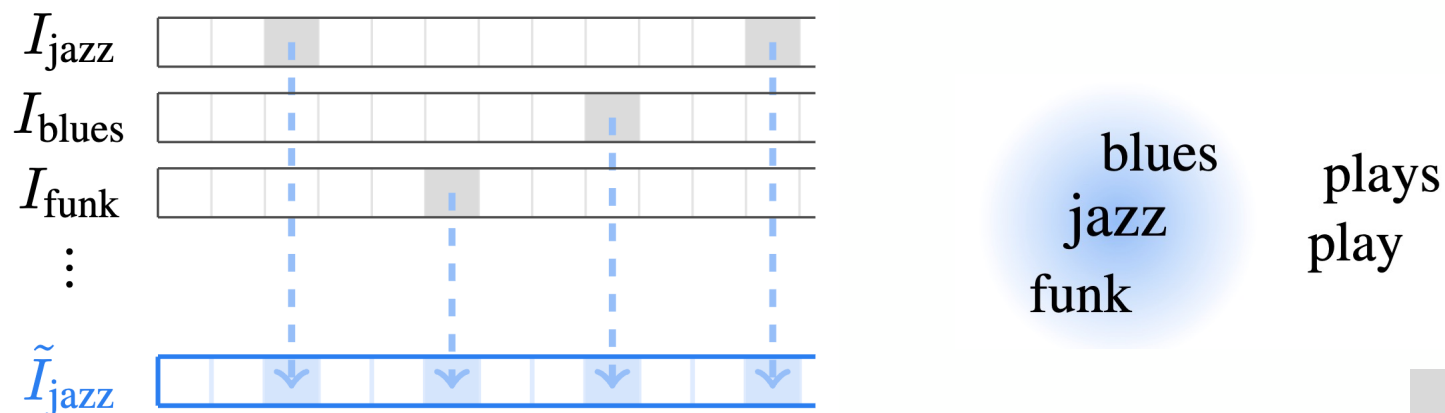
柔らかい grep/kwic

「a」はコーパス全体の2単語目, 8単語目, 51単語目, ……,
2039987911番目 に存在」

1. 各単語の登場位置の一覧を保存



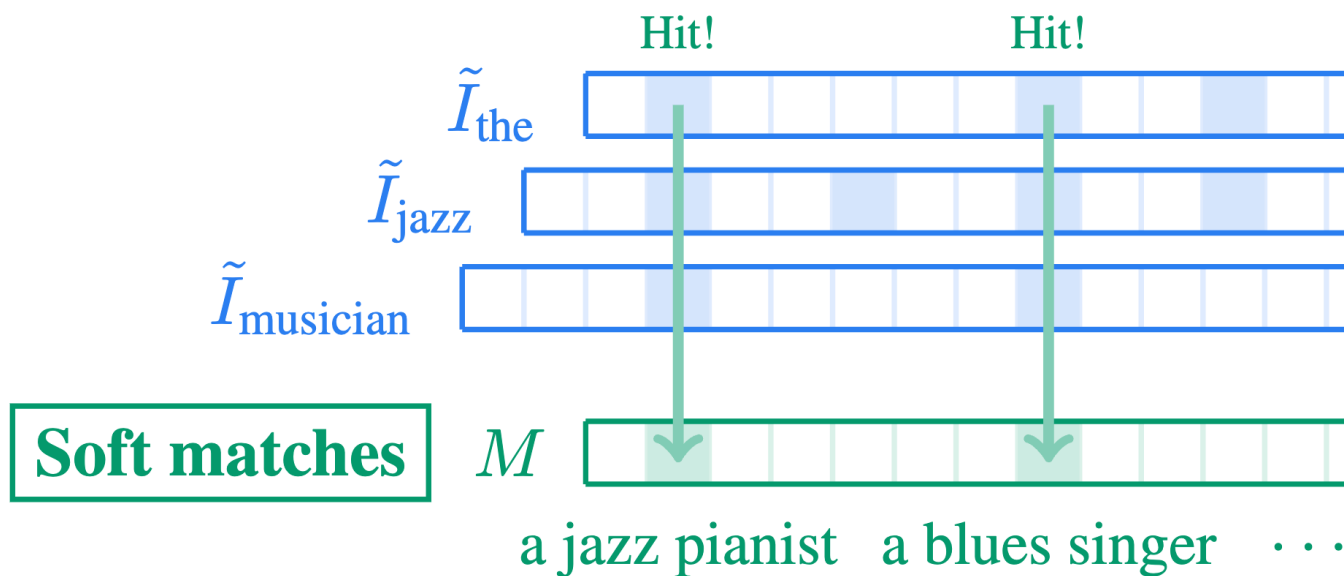
2. このリストを単語ベクトルの近さを使って膨らませる



SoftMatcha [Deguchi+, ICLR'25]

柔らかい grep/kwic

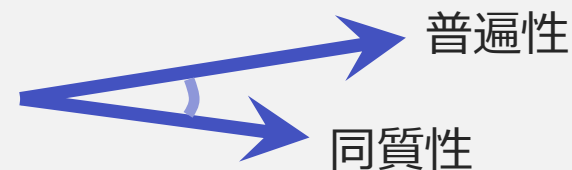
3. クエリ（例：“the jazz musician”）の単語順に、
リストを少しずつずらして重ねる
→ ヒット位置が見つかる



重なる = “the jazz musician” がその順番で連続で出てくる
※ しかも**似ている単語**も含めて検索できる

似ているフレーズを矢印の似かたで検索

[Deguchi+ICLR'25]



SoftMatcha 

言語の普遍性

×

Search

Results

Hits: 8

Search Time: 0.089 sec

リンゼイ・J・ウェイリー(2006).言語類型論入門— ^{1.00 1.00 1.00 1.00}言語の普遍性と多様性, 岩波書店(原書an Introduction to typologyは1996年)

たものの、かつてから宗教によって分断されてきたそれまでの民族規定に ^{1.00 1.00 0.55 1.00}変えて、言語の同質性によって南スラヴの統一を主張したものであるとした。しかしながら、wachtelの見解には次のような異議がもたれ得る。

^{0.54 1.00 0.58 1.00}英語の多様性 "diversity"の語源は、ラテン語ではdiverstiasに求められ、この言葉は、最初には、一致可能なものに反すること、矛盾、対立、不一致、といった消極的な意味を有したが、第二義的に、相違、多様、

10億語レベルの巨大言語データから、1秒以内に「ちょっと似ているもの」を全列挙できる

まとめ

まとめ

1. 準備：文のトピックを決める力が強い単語はベクトルが長い
2. 大きな言語ユニットの**意味計算**を
単語ベクトルを使って柔軟に実現する
 - 文の**類似度や意味の違い**を
「輸送のしやすさ」や「輸送できないこと」で**柔軟に測る**
 - 単語（ないし概念）集合間の「**かつ**」や「**または**」を
単語ベクトル空間で**柔軟に計算**する
 - 言語間の**対訳**辞書を最適輸送で**柔軟に見つける**
 - コーパス探索のための**柔軟な**コンコーダンスを作る

自然言語処理もしくは経験的な言語科学・ 言語工学にご興味がある/生えたみなさんへ

NLP コロキウム <https://nlp-colloquium-jp.github.io/>

新着の面白論文の著者をお呼びしてトーク(20分) + QA(30分)
日本時間 水曜正午より, 月1~2回開催

- | | | | |
|----|--|--|---|
| 80 | 2025/07/16 (Wed)
12:00 (JST) | 藤井一喜 / Kazuki Fujii
(東京科学大学) | 論文では語られないLLM開発において重要なこと - Swallowプロジェクトを通して (仮)
[概要] |
| 79 | 2025/06/25 (Wed)
12:00-13:30 (JST) | 陣内佑 / Yuu Jinnai
(株式会社サイバーエージェント) | Minimum Bayes Riskデコーディングのイントロダクション
[概要] [論文] (ACL 2025) |
| 78 | 2025/05/21 (Wed)
12:00 (JST) | Juno Kim
(東京大学, UC Berkeley) | Transformers Provably Solve Parity Efficiently with Chain of Thought
[概要] [スライド] [論文] (ICLR 2025) |
| 77 | 2025/05/07 (Wed)
12:00 (JST) | 前川政司 / Seiji Maekawa
(Megagon Labs) | Holistic Reasoning with Long-Context LMs: A Benchmark for Database Operations on Massive Textual Data
[概要] [スライド] [論文] (ICLR 2025) |

「大規模言語モデルをこじ開ける」という視点で最近チュートリアルをしました

- 言語モデルの内部機序（内部表象・計算メカニズム）について既存のどんな資料よりも詳しく分かりやすく説明しています
[[slides](#)][[video](#)]

数学的な視点というよりは、
まず as is で、言語モデルの内側で一体何が
起きていそうかを調べる研究群のまとめです

言語モデルの 内部機序

analysis interpretation
解析 と 解釈

Benjamin HEINZERLING¹², 横井 祥³²¹, 小林 悟郎²¹

¹理研, ²東北大, ³国語研

ラボを立ち上げました

- **ラボを立ち上げました [2025-04-]**

- 情報系出身の自分は「**大きなホワイトボードとおいしい珈琲と元気な若者を居心地の良い部屋に入れておくと (?) 面白い議論が生える**」メソッドの信奉者です
- 現在：博士課程学生 × 2, 外部研究員 × 2, バックオフィス × 3

- **主に扱う予定のトピック**

- 言葉の意味の統計的・幾何的特徴付け
- 経験主義的な言語(哲)学の実験科学・数理科学としてのリバイバル
- 研究というプロセス自体の自動化もおおいにおこなう予定です

- **ご興味がある皆さんぜひご連絡ください！**

- “この手の” 研究トピック・研究スタイルにご興味のあるみなさん、ぜひぜひ気軽にご連絡ください
- 各種バックオフィス業務にご興味のある皆さんもぜひご連絡ください
- yokoi@ninjal.ac.jp