2016年2月18日
第4回NLP東京Dの会

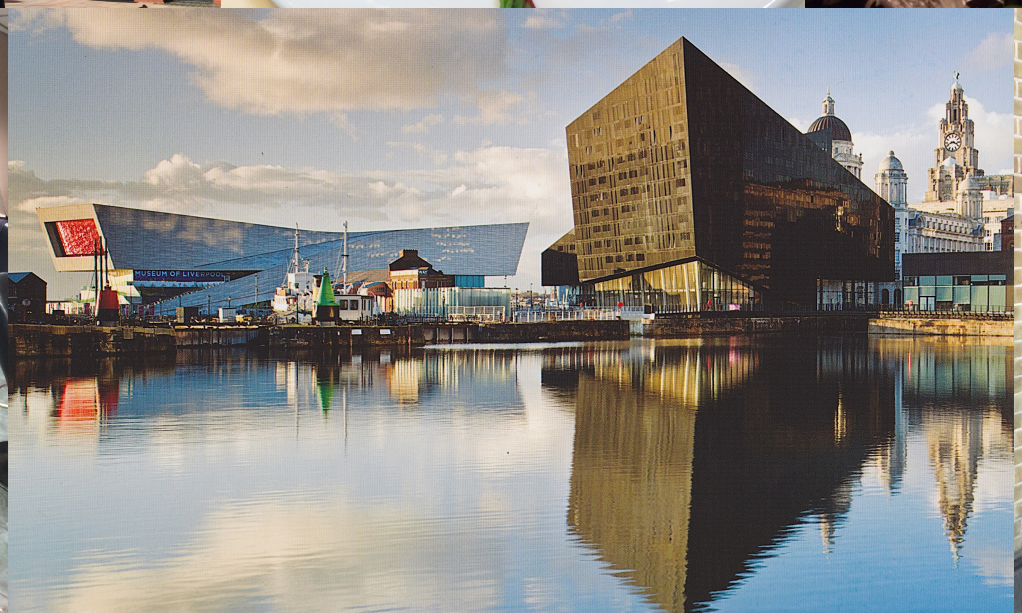
An Iterative Approach for the Global Estimation of Sentence Similarity

首都大小町研 D1　梶原 智之
https://sites.google.com/site/moguranosenshi/

自己紹介：梶原智之 @moguranosenshi

- 新居浜工業高等専門学校（2006年4月～2011年3月）
  - 電気情報工学科
  - 音楽情報処理（遺伝的アルゴリズムを用いた自動作曲）
- 長岡技術科学大学（2011年4月～2015年3月）
  - 修士課程：工学研究科 電気電子情報工学専攻
  - 自然言語処理（文章読解支援のための語彙平易化）
- 首都大学東京（2015年4月～）
  - 博士後期課程：システムデザイン研究科 情報通信システム学域
  - 自然言語処理（文章読解支援のためのテキスト平易化）
  - NLP若手の会プログラム委員
  - 2015年9月から12月までイギリスのリバプール大学

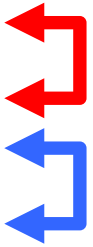
9月から12月まで英国リバプール大学で研究してきました。



An Iterative Approach for the Global Estimation of Sentence Similarity

- 情報検索や文書分類などの多くの自然言語処理課題で文間類似度の計算が重要である

- 特に短文間の類似度計算において「共通の単語が出現しないにも関わらず類似度が高い」という課題がある
  - I love dogs and cats.
  - I love dogs and hamsters.
  - My favorite pet is a cat.



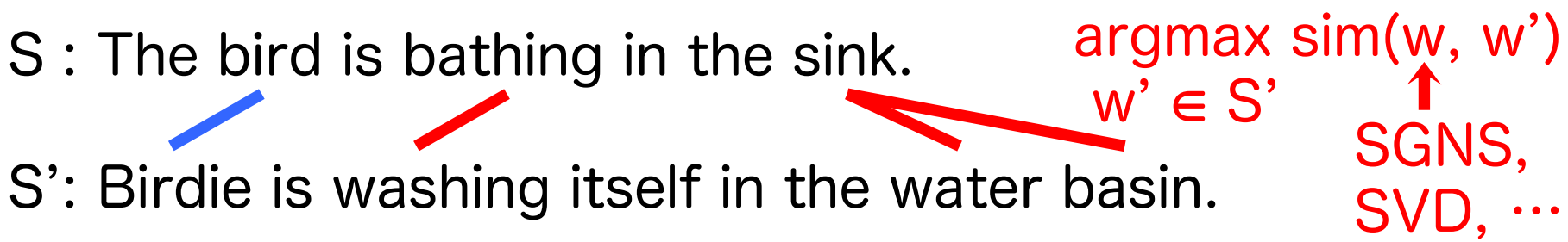
共通の単語が多く出現し類似している短文の組

共通の単語が出現しないが類似している短文の組

- 先行研究：2文の単語間類似度を平均して文間類似度を計算
- 本研究：類似度計算したいドメインのコーパスを利用し、そのドメインに特化した単語間類似度を計算してより正確に文間類似度を計算　（類似度計算の分野適応）

関連研究：Semantic Textual Similarity

- 評価型ワークショップ SemEval 2012-2016
  Semantic Textual Similarity タスク
  - 2つの文が与えられ、システムがその2文間の類似度を計算
  - 人間による正解の類似度との相関でシステムの性能を評価
- SemEval-2015 Task 2: STS English subtask
  - DLS@CU, ExBThemis, Samsungの3チームが上位を独占
  - いずれも単語アライメントをベースにしたアプローチ
  - これに加えて、WordNetやPPDBなどの外部資源を用いる



本研究の位置づけ

- 単語間類似度を改善する → 先行研究の性能改善
  - 対象のコーパスに特化した単語間類似度を計算する（類似度計算の分野適応）
  - 単語間類似度が良くなると、単語アライメントも最終的な文間類似度計算も良くなる
- この手法は単独でSTSタスクのstate-of-the-artを達成するというものではないです
  - が、多くの手法に適用できます
- 生コーパスだけあれば良い
  - 機械学習用のラベル付きデータは不要
  - シソーラスなどの外部資源も不要

提案手法：類似度計算の分野適応

- 2文間に共通の単語（類似度の高い単語）が多く出現する場合、残りの単語同士も類似している
- 2文間に類似度の高い単語が多く出現する場合、その2文は類似している
  - I love dogs and cats.
  - I love dogs and hamsters.
  - My favorite pet is a cat.

文の集合から、共通の単語を手がかりに残りの単語間の類似度を向上させていく

↕ 繰り返す

cat-hamsterと同様に、他の文のペアからI-Myやdog-petの類似度が高まる

$$\phi^{(t+1)}(x,y) = \phi^{(t)}(x,y) + \eta B_{\theta}^{(t)}(x,y)$$

更新後の単語間類似度

更新前の単語間類似度（一般的な類似度からスタート）

コーパスに特化した類似度

θ：「類似度が高い」の閾値

単語間類似度行列

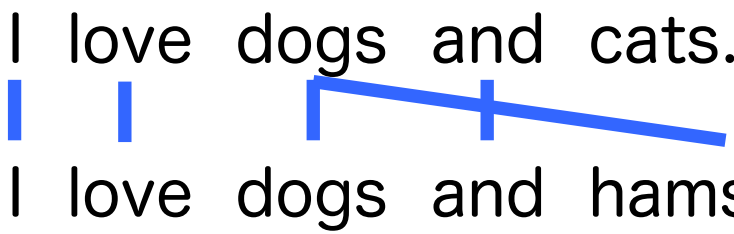
$$\phi^{(t+1)}(x,y) = \phi^{(t)}(x,y) + \eta B_\theta^{(t)}(x,y)$$

- $\Phi^{(0)}(x,y)$：一般的な単語間類似度
  - Skip-gram with Negative-sampling (SGNS)
- $B^{(t)}{}_\theta(x,y)$：コーパスに特化した単語間類似度
  - 単語アライメントに基づく正の自己相互情報量 (PPMI)
  - 単語アライメント行列 $A^{(t)}{}_\theta(x,y)$ を考える
    - 単語xが単語yにアラインされる頻度の行列
    - ただし単語間類似度 $\Phi^{(t)}(x,y)$ が θ 未満ならアラインしない

$$B_\theta^{(t)}(x,y) = \max(0, \log \frac{C_{ij}^{(t)} \sum_{ij} C_{ij}^{(t)}}{\sum_i C_{ij}^{(t)} \sum_j C_{ij}^{(t)}}) \qquad C_{ij}^{(t)} = \frac{1}{2}(A_{ij}^{(t)} + A_{ji}^{(t)})$$

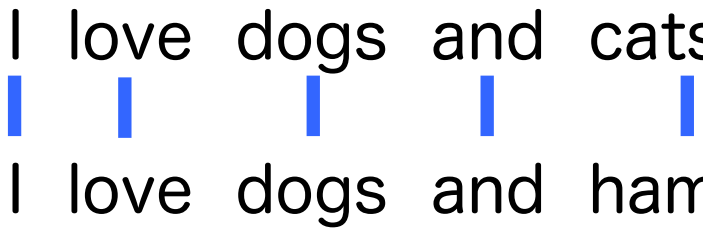
単語アライメント

- 1対多の単語アライメント (maximum: Amax)
  - 単語xに対して類似度最大の単語yをアラインする



※Φ(犬, 鼠) > Φ(猫, 鼠)

- 1対1の単語アライメント (hungarian: Ahun)
  - 2文間の単語アライメントを2部グラフの最大マッチング問題と考えてハンガリアン法で解く



単語間類似度 → 文間類似度

- 多対多の単語間類似度の平均 (average: Save)

$$S_A = \frac{1}{\|x\|\|y\|}\sum_i\sum_j \phi(x_i, y_j)$$

- 1対多の単語間類似度の平均 (maximum: Smax)

$$S_M = \frac{1}{\|x\|\|y\|}\sum_i \max_j \phi(x_i, y_j)$$

- 1対1の単語間類似度の平均 (hungarian: Shun)

$$S_H = \max_h \frac{1}{\|x\|\|y\|}\sum_i \phi(x_i, h(x_i))$$

評価①：SemEval-2015 Task 2

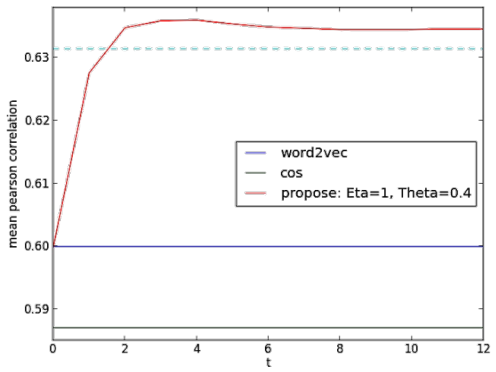
- SemEval-2015の Semantic Textual Similarity タスク
  - 英語、5ドメイン、合計 3,000 文対
  - 人手で付与された文間類似度とシステム出力との Pearson 相関

Method	Answers-Forums	Answers-Students	Belief	Headlines	Images	Mean
Amax+Save η=1.5, θ=0.5, t=12	0.4774	0.3747	0.2554	0.4568	0.4015	0.3998
Amax+Smax η=1, θ=0.4, t=4	**0.5012**	**0.6744**	**0.6710**	**0.6394**	**0.6439**	**0.6360**
Amax+Shun η=1, θ=0.4, t=4	0.4914	0.6746	0.6543	0.6269	0.6297	0.6260
Ahun+Save η=1, θ=0.4, t=5	0.4500	0.3616	0.2547	0.4430	0.3981	0.3888
Ahun+Smax η=1, θ=0.4, t=4	0.4989	0.6739	0.6706	0.6380	0.6424	0.6348
Ahun+Shun η=1, θ=0.4, t=4	0.4910	0.6742	0.6546	0.6260	0.6290	0.6255

評価①：SemEval-2015 Task 2

* Song and Roth, Unsupervised Sparse Vector Densification for Short Text Similarity, NAACL-2015.

Method	Answers-Forums	Answers-Students	Belief	Headlines	Images	Mean
COS類似度	0.4453	0.6642	0.6517	0.5312	0.6039	0.5870
Save*	0.0794	0.0434	0.1254	0.2943	0.3160	0.1890
Smax*	0.3912	0.6568	0.6366	0.6031	0.6260	0.5999
Shun*	0.3867	0.6820	0.6157	0.5901	0.6141	0.5969
Amax+Smax (PPMIのみ)	0.2320	0.6646	0.4960	0.4662	0.4480	0.4857
提案手法 Amax+Smax	0.5012	0.6744	0.6710	0.6394	0.6439	0.6360



COS類似度：弱いベースライン（完全一致）

Smax：強いベースライン（従来手法）
提案手法の左辺のみ（t=0）に相当

PPMI：提案手法の右辺のみ（t=0）に相当

結果：文間類似度推定タスクで性能を改善

知見：分野適応◎、繰り返しアプローチ◎

評価②：単語間類似度

- 6つの単語類似度ベンチマークデータセットで評価
  - 人手で付与された単語間類似度とシステム出力との Pearson 相関
- SemEvalのコーパスだけでは多くの単語が未知語なので UKWaC（イギリス英語のWebテキスト）約40万文も併用

Method	MC	MEN	RG	RW	SCWS	WS
SGNS	0.6330	0.5636	0.5347	0.3421	0.6111	0.5526
PPMI	0.4883	0.6307	0.4263	0.1206	0.5382	0.5297
提案手法 Amax	0.7065	0.5894	0.6258	0.1881	0.5428	0.4823

- 文間類似度を求める際に副産物として得られた単語間類似度も、SGNSやPPMIに比べて改善されている！

An Iterative Approach for the Global Estimation of Sentence Similarity

- 特定のコーパスから得られる単語間類似度を用いて
  一般的な単語間類似度を繰り返し更新し、文間類似度を
  正確に求める類似度計算のドメインアダプテーション
- 提案手法は一般的な単語間類似度のみを用いる
  ベースラインよりも統計的に有意に性能が向上した
- Semantic Textual Similarity タスクでは
  単語アライメントに基づくアプローチが主流で、
  外部知識や機械学習の手法を加えたものがSOTA
- 提案手法は単語アライメントに基づく多くの
  文間類似度計算手法の性能改善に寄与すると期待できる