

ショートメッセージの部分空間表現と 宛先推定

清水 彰人 酒井 智弥 小林 透

SNSに不慣れな人を支援したい

- SNSの多くは、主にテキストメッセージを流通させることが前提
- 端末の操作に不慣れな利用者や高齢者はSNSに不慣れ
- SNSに不慣れな人のメッセージ配信を支援するためには、メッセージの内容から宛先を推定する機能を実現するべき

課題：メッセージが短文、少量

- 迷惑メールフィルタや文書分類で成功している機械学習の既存手法は、大量のメッセージ例を必要とする（単純ベイズ分類器、tf-idfなど）
- 高齢者などからのメッセージ例を大量に用意することは非現実的
- 既存手法を流用して解決できない

word2vecを使った手法の提案

- 筆者は、出現頻度等で定量化した個々の単語の重要性ではなく、**単語の組合せが表す意味**に着目
- 例えば、親族宛メッセージの関連キーワードとして「正月」、「帰省」、「孫」等を想定していれば、「今度、新幹線で帰っておいで。」が親族宛であることは容易に推定できる
- **単語やその組合せが表す意味の類似性を計量**できれば、ショートメッセージから宛先を推定する手法を設計できる
- そこでword2vecです

word2vecとは

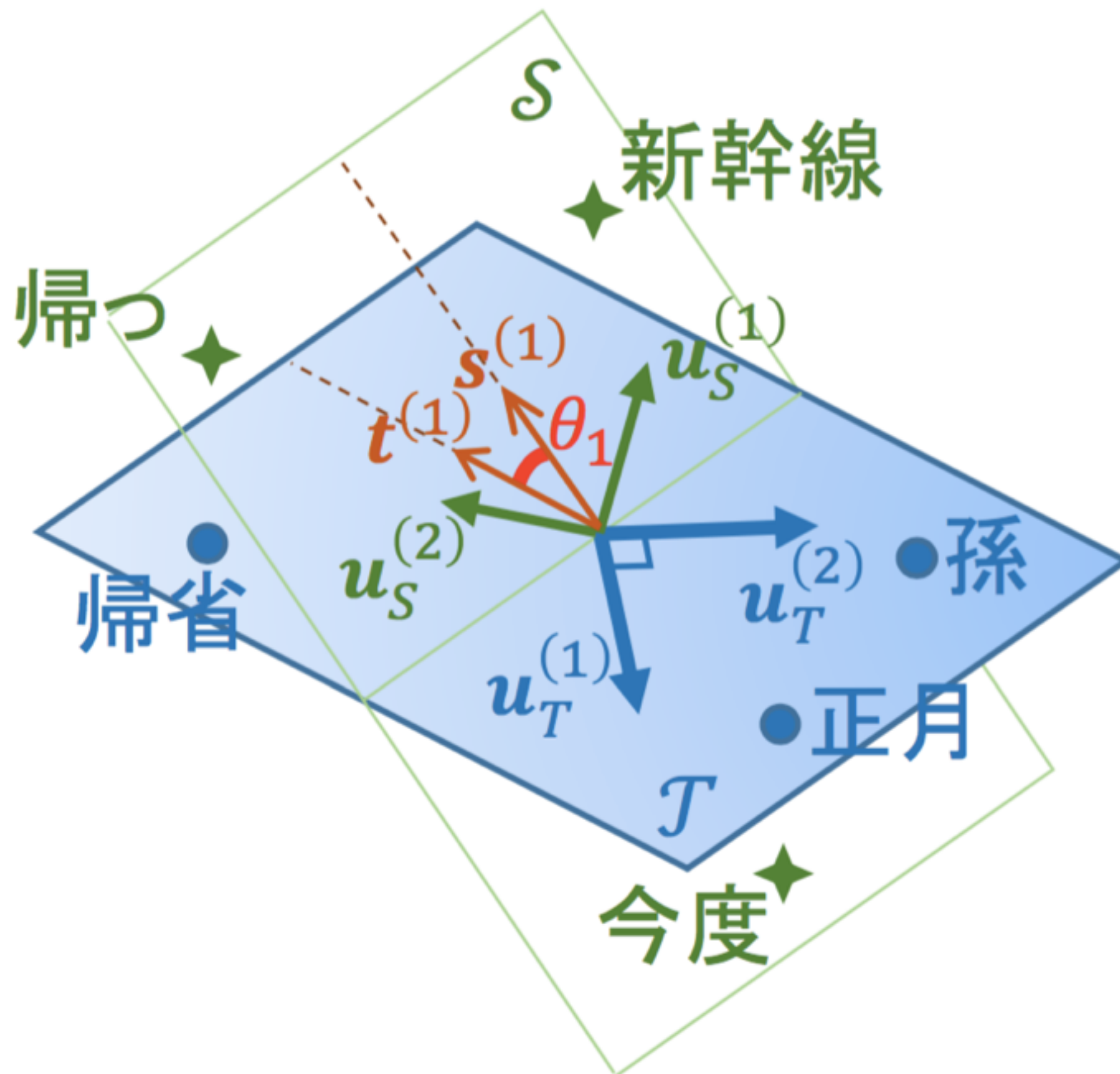
- 「似た文脈に出現する単語は互いに似た意味を持つ」という分布仮説に基づき、単語の分散表現を効率的に教師なし学習する手法
- **単語の意味のベクトル表現**が得られ、線形代数的な演算も可能
- $v(\text{"king"}) - v(\text{"man"}) + v(\text{"woman"}) = v(\text{"queen"})$ の近似ベクトル

提案手法のアイデア

- ショートメッセージを少数の単語の組合せと見なしてword2vecすると、ショートメッセージは線形ベクトル部分空間として表せる
- ショートメッセージの潜在的意味はこの部分空間に対応づけられ、意味が近いショートメッセージの部分空間は互いに類似していると考えられる
- 宛先についても同様に、少数のキーワードが与えられれば線形ベクトル部分空間で表現できる
- 2つの部分空間の類似度を計算できれば、宛先を推定できる

提案手法の手順

1. 大量の文書から単語ベクトルを作成しておく(一般的な単語の意味の関係を獲得できればよい)
2. 各宛先の関連キーワードを事前に集め、それらの単語ベクトルを列に持つデータ行列を**特異値分解**し、あらかじめ各宛先の基底行列を求めておく
3. メッセージが与えられたら、その単語ベクトルを列に持つデータ行列を**特異値分解**し、メッセージの基底行列を計算する
4. **相互部分空間法**を用いて類似度が最大となる宛先を探す



宛先推定の実験 (STEP 1/4)

大量の文書から単語ベクトルを作成しておくSTEP

1. 2015年11月2日時点の日本語Wikipediaの全記事の平文を取得
2. 形態素解析器MeCab (辞書はNAIST-jdic) で分かち書き
3. 各単語をword2vec

405,565語の単語ベクトルが得られた

宛先推定の実験(STEP 2/4)

各宛先の基底行列を求めておくSTEP

表 1 宛先とキーワード.

病院 T_1	医者, 病院, 薬, 処方箋, ケア, 通院, 患者, 救急車, 看護, 手術, 安静, 薬学, クリニック, 医療, 福祉, 介護, 内科, 診察, 白衣, 院長, ベッド, 通院
宅配 T_2	宅配, コンビニ, 出前, 寿司, 配達, 電話, ラーメン, 餃子, 中華, 料理, 野菜, 値段
親族 T_3	帰省, 母, 娘, 盆, 夏休み, 正月, 家族, 父, 祖父, 祖母, 渋滞, 孫, 墓, 跡継ぎ, 実家, 田舎, 息子

宛先推定の実験 (STEP 3/4)

メッセージの基底行列を計算するSTEP

表 2 各宛先へのメッセージ.

病院宛 S_1	先生 , 子供 が 熱 出 し た の で 診 て く だ さ い .
宅配宛 S_2	弁 当 の 注 文 を お ね が い し ま す .
親族宛 S_3	今 度 , 新 幹 線 で 帰 っ て お い で .

宛先推定の実験 (STEP 4/4)

類似度が最大となる宛先を探すSTEP

表 3 メッセージと宛先との類似度.

メッセージ \ 宛先	病院 T_1	宅配 T_2	親族 T_3
病院宛 S_1	0.84	0.55	0.77
宅配宛 S_2	0.49	0.85	0.57
親族宛 S_3	0.52	0.59	0.65

考察

- 親族宛メッセージ S_3 について、親族は他の宛先と比べて類似度が顕著に高くない
- S_3 「今度、/新幹線/で/帰っ/て/おい/で。」に含まれる単語「帰っ」は親族のキーワード「帰省」と意味が近い
- 一方、「新幹線」という単語は病院や宅配に挙げられているキーワード「救急車」、「配達」、「出前」に意味が近い(乗り物や移動に関連する)

まとめ

word2vecを利用して、短いメッセージと宛先候補をそれぞれ線形ベクトル部分空間で表すことで、相互部分空間法による類似性の定量化の手法を提案した

所感

- 「SNSに不慣れな人のメッセージ配信を支援するためには、メッセージの内容から宛先を推定する機能を実現すべき」← そう？
- 数式…
- word2vecはおもしろい