



ソーシャルメディア分析サービスにおけるNLP技術の活用と近年における使用技術の変化

株式会社ホットリンク 榎 剛史

自己紹介・会社の概要

2019/7/7

軽く自己紹介

榎 剛史 (株)ホットリンク R&D部部长
 東京大学 客員研究員
 @tksakaki



Earthquake shakes Twitter users
<https://dl.acm.org/citation.cfm?id=1772777>
 T Sakaki 著 - 2010 - 被引用数: 3821 - 関連記事
 2010/04/26 - Twitter, a popular microblogging service, has received a lot of attention recently. For example, when an earthquake occurs, people make many Twitter tweets, which enables detection of earthquake occurrence.

- 興味領域
 - Artificial Intelligence
 - Computational Social Science
 - Natural Language Processing
 - Machine Learning
- 経歴
 - 2006年：修士号（電子情報学）取得
 - 2006~2009年：東京電力にて勤務
 - 2009年10月：博士課程入学（松尾研究室）
 - 2013年12月：博士号（技術経営学）取得
 - 2014年~2015年：東京大学 特任研究員
 - 2015年~現在：現職

第22回SIG-AM研究会 2019/7/7

会社概要

社 名 本 社 設 立 資 本 金 株 式 市 場 代 表 事 業 内 容 連 結 子 会 社	株式会社ホットリンク 東京都千代田区富士見1-3-11 富士見デュープレックスビル5階 2000年6月26日 2,357百万円(2018年7月末時点) 東京証券取引所マザーズ 代表取締役社長 内山 幸樹 ソーシャル・ビッグデータの分析・販売事業 クラウドサービス事業 インバウンドプロモーション支援事業など 株式会社トレンドExpress(100%子会社) EFFYIS, inc. (100%子会社) 流行特急(100%中国小会社)
--	---

第22回SIG-AM研究会 2019/7/7

事業コンセプト

Big Data

ソーシャル・ビッグデータを活用し、
「データとAIで意思決定をサポートする」ことを目指し、
マーケティングに関わる事業を運営・提供しています。

AI

第22回SIG-AM研究会 2019/7/7

会社概要

ソーシャル・ビッグデータを活用し、
「データとAIで意思決定をサポートする」ことを目指し、
マーケティングに関わる事業を運営・提供しています。

ソーシャル・ビッグデータ 解析ツール事業 	クチコミの マーケティング活用 	SNSアカウント活用 の効率化
ソーシャル・ビッグデータ 流通・販売事業 	24種類のロコミデータ 流通・販売 	クロスバウンド・ マーケティング 支援事業

第22回SIG-AM研究会 2019/7/7

ソーシャルメディア分析ツール

クチコミ係長
kuchikomi@kakaricho

<https://service.hottolink.co.jp/service/kakaricho/>

世界観

ネット世界

リアル世界

俯瞰

投影

第22回SIG-AM研究会 2019/7/7

ソーシャルメディアアカウント運用ツール

Buzz Spreader

<https://service.hottolink.co.jp/service/buzzspreader/>
<https://hashtag-ai.buzzspreader.com>

世界観

企業アカウント

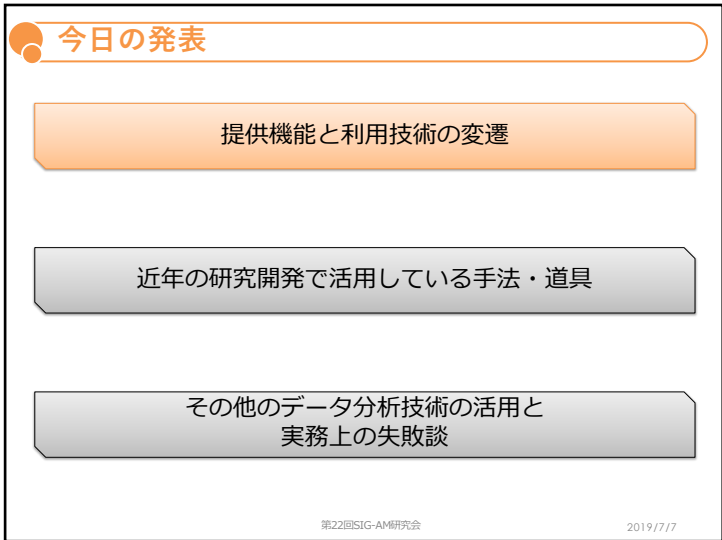
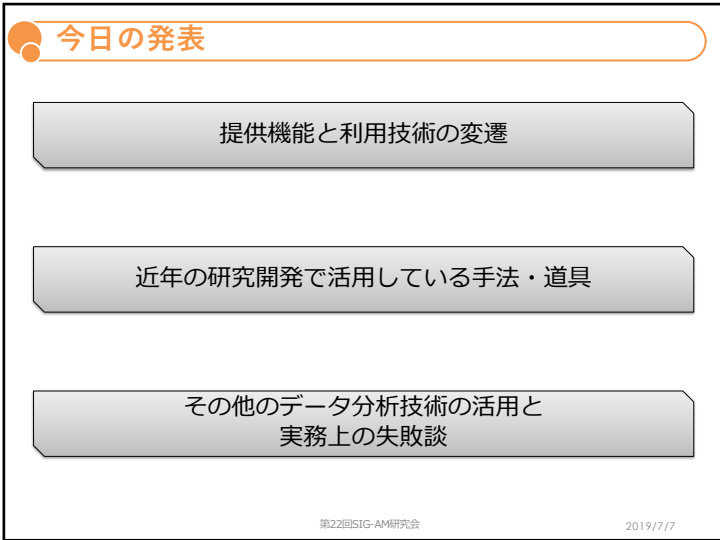
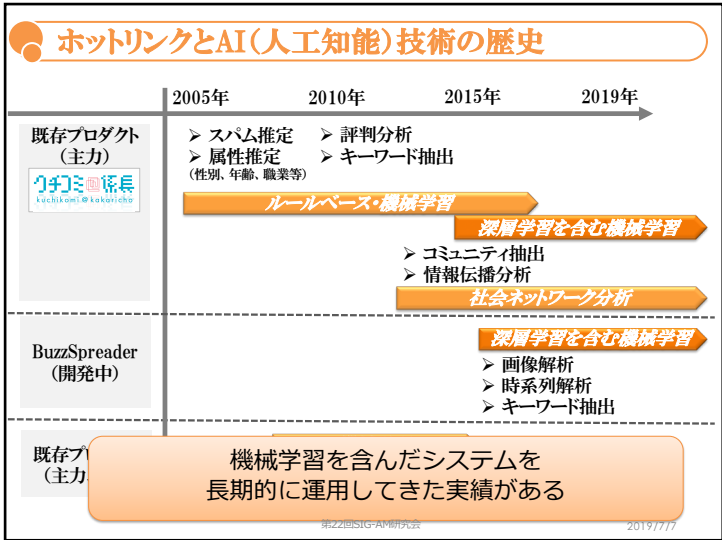
企業

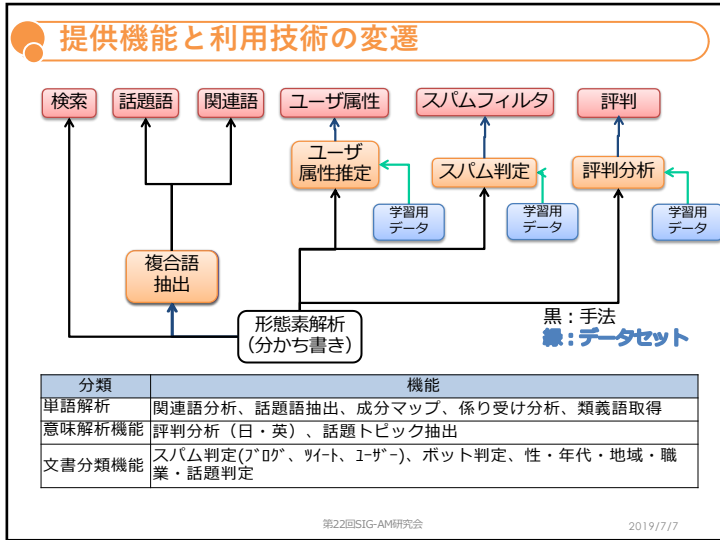
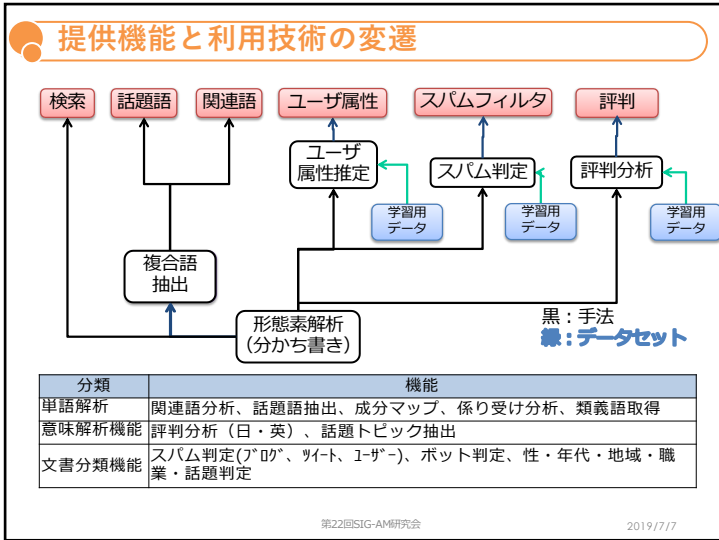
UGC

これまでの情報発信

SNSを活用した情報発信・拡散

第22回SIG-AM研究会 2019/7/7





提供機能と利用技術の変遷

複合語抽出

目的

- SNS投稿について、適切な単語単位で分かち書きを行う

機能

- 入力：SNS投稿（もしくはスニペット）
- 出力：複合語を含む分かち書き

サンプル

- 昨日の君の名はのついたー実況はめっちゃ盛り上がった
 - 通常
 - 昨日 | の | 君 | の | 名 | は | の | つ | い | っ | た |ー | 実 | 況 | は | め | っ | ち | ゃ | 盛 | り | 上 | が | っ | た |
 - 複合語抽出あり
 - 昨日 | の | **君の名は** | の | **ついたー** | 実況 | は | め | っ | ち | ゃ | 盛 | り | 上 | が | っ | た |

第22回SIG-AM研究会 2019/7/7

提供機能と利用技術の変遷

評判分析

目的

- SNS投稿について、指定された対象語に対するポジ/ネガ/ニュートラルを出力する
- SNS投稿について、文全体のポジ/ネガ/ニュートラルを出力する

機能

- 入力：SNS投稿, 対象語 / SNS投稿
- 出力：ポジ/ネガ/ニュートラルの3値及びスコア

サンプル

- 今年のお皿は2枚セットだよ!あと、**サンリオ**くじが素敵すぎて.....(*´v´)
- 全体：Positive サンリオ：Positive
- 子持ちシヤモは食べれる！**イクラ**は無理でした。。
- 全体：Neutral イクラ：Negative

第22回SIG-AM研究会 2019/7/7

提供機能と利用技術の変遷

スパム投稿判定

目的

- 入力された文書について、スパム投稿かどうかを二値で判定する

機能

- 入力：投稿
- 出力：スパム/非スパムの2値

サンプル

- シャツ メンズM ディープブルー ¥13000 送料無料
- [アダルトワード]興味あります。どMさんいない？
- 友人が台湾でスマホ盗まれた結果www

第22回SIG-AM研究会 2019/7/7

提供機能と利用技術の変遷

ユーザ属性推定

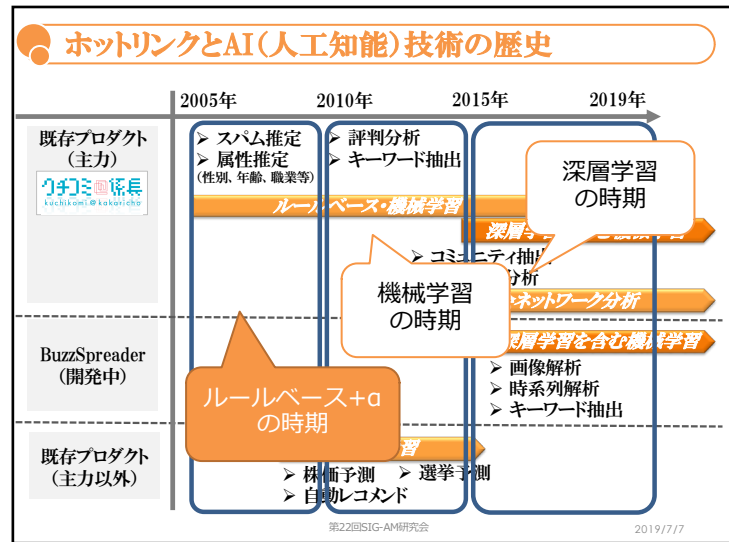
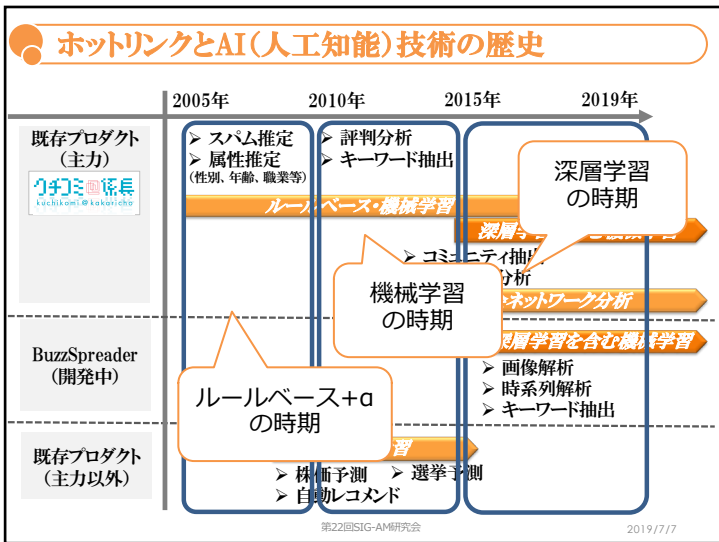
目的

- 特定のユーザによる投稿群から、ユーザの属性を推定する

機能

- 入力：投稿群
- 出力：各属性のクラス（性別/年代）
 - 性別：男性/女性
 - 年代：10代/20代/30代/40代/50代/60代

第22回SIG-AM研究会 2019/7/7



ルールベース+ α の時期

- 時期：2006年～2010年
- 分析対象
 - ブログ/2ch中心
- 道具立て

<ul style="list-style-type: none"> - 形態素解析 <ul style="list-style-type: none"> • MeCab：2006年9月にInitial release <ul style="list-style-type: none"> - Chasen/JUMANは速度の面から利用せず • 辞書：IPADIC <ul style="list-style-type: none"> - Unidic(2006~2007)/Naistdic(2007) - 係り受け解析 <ul style="list-style-type: none"> • CaboCha：商用利用不可 <ul style="list-style-type: none"> - 京都大学テキストコーパス Ver4.0 2002年 	<ul style="list-style-type: none"> - 分類器 <ul style="list-style-type: none"> • Support Vector Machine：自前で実装 <ul style="list-style-type: none"> - TinySVM(2002)/Liblinear(2008)/classias(2009)
---	---

第22回SIG-AM研究会

2019/7/7

ルールベース+ α の時期

複合語抽出

機能

- 入力：SNS投稿（もしくはスニペット）
- 出力：複合語を含む分かち書き

アプローチ

- 品詞の接続情報に対するルールベースアプローチ
- 品詞/品詞詳細情報ごとに、前の語との結合を規定
 - 「名詞,一般」は前の語と結合する
 - 「接頭詞」は前の語と結合しない

第22回SIG-AM研究会

2019/7/7

ルールベース+ α の時期

スパム投稿判定

機能

- 入力：投稿（ブログ）
- 出力：スパム/非スパムの2値

アプローチ

- コンテンツ/タイトルについて、ルール・辞書ベースアプローチ
 - タイトル：スパムワードによる判定
 - コンテンツ：スパムワード/URLの出現パターン/文書内の改行・空白の出現パターンから判定

第22回SIG-AM研究会

2019/7/7

ルールベース+ α の時期

評判分析

機能

- 入力：文書，対象語 / 文書（2つのタイプの対応）
- 出力：ポジ/ネガ/ニュートラルの3値及びスコア

アプローチ

- 対象語と評価語辞書の距離を用いた辞書・ルールベースアプローチ
 - 入力文書から対象語の位置を抽出
 - 入評価語辞書に含まれる単語のうち、入力文書に含まれる全ての単語とその位置を抽出
 - 文書に含まれた評価語について、対象語からの距離を重みとして、ポジティブ/ネガティブスコアを算出する
 - 全ての評価語について、評判スコアを合計し、最終的なポジティブ・ネガティブ・ニュートラルを判定する

第22回SIG-AM研究会

2019/7/7

ルールベース+ α の時期

ユーザ属性推定

機能

- 入力：投稿群
- 出力：各属性のクラス
 - ※ブログ著者，性別のみ

アプローチ

- Support Vector Machineによる文書分類
 - 特徴量：Bag of Words
 - 人手による学習・テスト用データ整備

第22回SIG-AM研究会 2019/7/7

ホットリンクとAI(人工知能)技術の歴史

時期	技術/製品	機能
2005年	既存プロダクト (主力)	スパム推定 属性推定 (性別, 年齢, 職業等)
2010年	既存プロダクト (主力以外)	株価予測 選挙予測 自動レコメンド
2010年	BuzzSpreader (開発中)	画像解析 時系列解析 キーワード抽出
2015年	既存プロダクト (主力以外)	深層学習を含む機械学習
2019年	既存プロダクト (主力以外)	深層学習の時期

第22回SIG-AM研究会 2019/7/7

機械学習の時代

- 時期：2010年～2015年
- 分析対象
 - ブログ/Twitter中心
- 道具立て

<ul style="list-style-type: none"> - 形態素解析 <ul style="list-style-type: none"> • MeCab • 辞書：IPADIC <ul style="list-style-type: none"> - Ipadic-neologd(2015年3月) - 係り受け解析 <ul style="list-style-type: none"> • CaboCha/JdepP：商用利用可能 <ul style="list-style-type: none"> - KNBコーパス：2009年9月 	<ul style="list-style-type: none"> - 分類器 <ul style="list-style-type: none"> • Support Vector Machine <ul style="list-style-type: none"> - liblinear/Classias - 分散表現 <ul style="list-style-type: none"> • word2vec
---	---

第22回SIG-AM研究会 2019/7/7

機械学習の時代

複合語抽出

機能

- 入力：SNS投稿（もしくはスニペット）
- 出力：複合語を含む分かち書き

アプローチ

- 品詞の接続情報に対するルールベースアプローチ
- マルコフ連鎖(k=2)を想定した実装
 - 「名詞,一般」と「名詞,一般」は結合する
 - 「名詞,一般」と「接頭詞,名詞接続」は結合しない

第22回SIG-AM研究会 2019/7/7

提供機能と利用技術の変遷

評判分析

機能

- 入力：文書，対象語 / 文書（2つのタイプの対応）
- 出力：ポジ/ネガ/ニュートラルの3値及びスコア

アプローチ

- 係り受け解析結果と評判語辞書に基づくルールベース
 - ・ 入力文書を分かち書き/係り受け
 - ・ 対象語と係り受け関係にあるのうち，評判語辞書にあるものを抽出
 - ・ 抽出した評価語において，ポジ/ネガのスコアを算出。ただし，係り受けのパターンによりスコアの反転などを実施
- 係り受け器：JDepP

第22回SIG-AM研究会

2019/7/7

提供機能と利用技術の変遷

スパム投稿判定

機能

- 入力：投稿（ツイート）
- 出力：スパム/非スパムの2値

アプローチ

- スパムツイートの内容に基づくNaive Bayesアプローチ
 - ・ 特徴量：Bag of Words, Bag of Word-bigrams
 - ・ Naive Bayesを用いてスパム判定モデルを学習
- 手動アノテーションによる学習・テスト用データ整備

第22回SIG-AM研究会

2019/7/7

ルールベース+ α の時期

ユーザ属性推定

機能

- 入力：投稿群
- 出力：各属性のクラス
 - ・ ※Twitterユーザ，性別・年代

アプローチ

- Support Vector Machineによる文書分類
 - ・ 特徴量：Bag of Words
- 手動アノテーションによる学習・テスト用データ整備

第22回SIG-AM研究会

2019/7/7

SNSアカウントのプロフィール推定

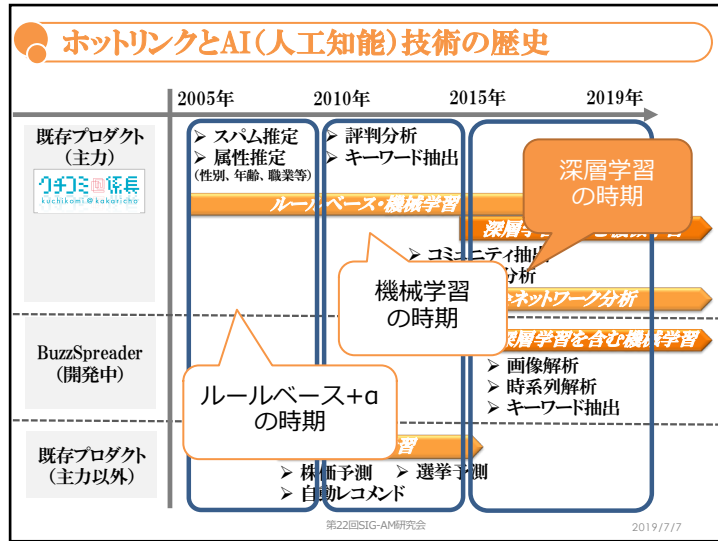
各ユーザ属性の分類に有効な特徴量

男性	女性
自己紹介：男子	自己紹介：女の子
自己紹介：愛しています	自己紹介：女性
発言：腐女子	自己紹介：キスマイ
発言：女子高生	発言：腐男子

大学生	それ以外	50代	それ以外
自己紹介：回生	自己紹介：元気	自己紹介：おじさん	自己紹介：女子
自己紹介：女子大	自己紹介：海外旅行	自己紹介：おばさん	自己紹介：ゲーム
発言：レポート	自己紹介：キロ	自己紹介：読書	自己紹介：在住
発言：サークル	発言：遠足	自己紹介：蕎麦	自己紹介：社会人
発言：履修登録	発言：職場	発言：膝	発言：(t,t)

ソーシャルメディア分析におけるAI技術活用とその失敗談

2019/05/25



深層学習の時代

- 時期：2015年～ 現在
- 分析対象
 - Twitter中心 / ブログが Optional
- 道具立て
 - 形態素解析
 - MeCab :
 - 辞書：IPADIC-neologd
 - 分散表現
 - 単語分散表現：word2vec
 - 文分散表現：BERT
 - 分類器
 - fasttext
 - Neural Network(Attention Mechanism)
 - データセット構築
 - Yahoo!クラウドソーシング
 - 自動アノテーション

第22回SIG-AM研究会 2019/7/7

深層学習の時代

複合語抽出

機能

- 入力：SNS投稿（もしくはスニペット）
- 出力：複合語を含む分かち書き

アプローチ

- 辞書の拡充
 - 名詞：mecab-ipadic-neologd に依存
 - 用言（形容詞・動詞）：SNSに頻出する表現を抽出
 - エモい/ググる/ジワる
- 品詞の接続情報に対するルールベースアプローチ
 - ルールの整理
 - 「基礎日本語文法に基づき言語現象を整理/対応
 - 有限状態オートマトンで実装

第22回SIG-AM研究会 2019/7/7

深層学習の時代

評判分析

機能

- 入力：文書（タイプ2の対応）
- 出力：ポジ/ネガ/ニュートラルの3値及びスコア

アプローチ

- BERT + fine-tuning
 - 1年分のツイートによる文分散表現を学習
- クラウドソーシングによる学習・テストセットデータ構築

第22回SIG-AM研究会 2019/7/7

深層学習の時代

スパム投稿判定

機能

- 入力：投稿（ツイート）
- 出力：スパム/非スパムの2値

アプローチ

- Fasttextによる文書分類
- データセット構築
 - 学習用データ：いくつかの仮説に基づく半自動アノテーション
 - テストデータ：手動アノテーション

第22回SIG-AM研究会 2019/7/7

深層学習の時代

ユーザ属性推定

機能

- 入力：投稿群
- 出力：各属性のクラス

アプローチ

- Fasttextによる文書分類
- NN-basedアプローチによる4種類属性の同時推定
 - 性別/地域（都道府県）/年代/職業
 - 特徴量：word2vec
- 自動アノテーションによる学習・テストセットデータ構築
- モデル自動更新の仕組みを導入

第22回SIG-AM研究会 2019/7/7

提供機能と利用技術の変遷

■ 利用技術の変遷 まとめ

時期	複合語抽出	スパム判定	ユーザ属性推定	評判分析
ルールベース +a	ルールベース	ルールベース	Support Vector Machine	分かち書き+ルールベース
機械学習	ルールベース	Naïve Bayes 手動アノテーション	Support Vector Machine 手動アノテーション	係り受け+ルールベース
深層学習	ルールベース	Fasttext 自動アノテーション	fasttext NN+ word2vec クラウドソーシング	BERT + fine-tuning クラウドソーシング

第22回SIG-AM研究会 2019/7/7

提供機能と利用技術の変遷

■ 評判分析を事例とした性能比較

評判分析 性能比較

指標	分かち書き+ルール	係り受け+ルール	BERT+Fine-tuning
Precision	0.48	0.48	0.72
Recall	0.32	0.52	0.72
F-value	0.38	0.50	0.72

Accuracy比較 (深層学習)

モデル	Accuracy
分かち書き+ルール	0.52
Conv1DClassifier (2017)	0.65
BiLSTM (2018-2017)	0.68
BiSVM+smi-supervised	0.68
BiSVM+seed (smi-supervised)	0.70

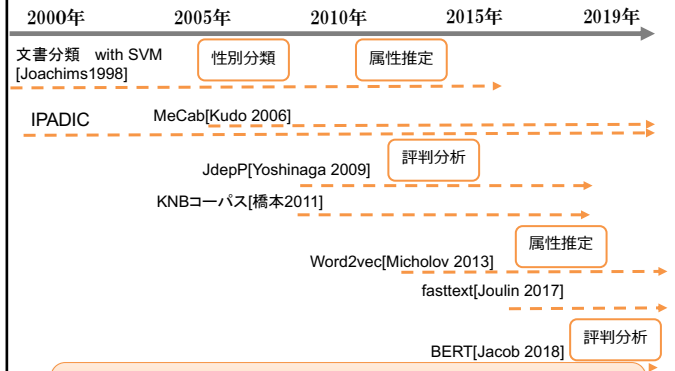
第22回SIG-AM研究会 2019/7/7

提供機能と利用技術の変遷

- ルールベース + α の時代
 - 分析対象データの収集
 - ルールと辞書の整備
 - ルール整備 => テスト => 誤り分析 => ルール整備・・・の繰り返し
- 機械学習の時代
 - 学習・テストデータの収集
 - 学習・テストデータへのラベル付け
 - モデルの選定/パラメータチューニング
- 深層学習の時代
 - 学習・テストデータの収集

徐々に泥臭い作業は減りつつある
 ルールベースの知見が活用できている
 事前学習済みモデルの活用が進みつつある

ホットリンクとAI(人工知能)技術の歴史



技術の実用化までの時間が短い
 技術の寿命が縮んでいる

今日の発表

- 提供機能と利用技術の変遷
- 近年の研究開発で活用している手法・道具
- その他のデータ分析技術の活用と実務上の失敗談

近年の研究開発で活用している手法・道具

- クラウドソーシングによるアノテーション
 - 小規模・大量のタスクをクラウドワーカーに依頼する
- メリット
 - 大規模なアノテーションデータが得られる
 - 支払い金額を高くすれば、作業時間を短縮できる
- デメリット
 - アノテーターの質・信頼性が均一でない
 - アノテーション結果をそのまま採用することはできない
 - 複雑なタスクを依頼することはできない

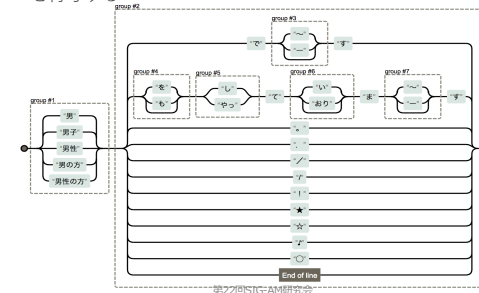
近年の研究開発で活用している手法・道具

- クラウドソーシングによるアノテーション
 - 小規模・大量のタスクをクラウドワーカーに依頼する
- デメリットの解消策
 - アノテーターの質を担保する
 - チェック質問に回答できたワーカーの結果のみを採用
 - 同じ質問を複数ワーカーに尋ねて、一致率が高い結果のみを採用
 - 1つ辺りのタスクを設問数を多くしすぎない
 - アノテーション作業を簡単な部分問題に分解する
 - 例：評判分析におけるラベルの付け直し
 - あるデータセットにおいて、不適切と思われるラベルのみ修正する
 - 付与されたラベルが適切かどうかを尋ねるタスク
 - 新たにラベルを付与するタスク

第22回SIG-AM研究会 2019/7/7

近年の研究開発で活用している手法・道具

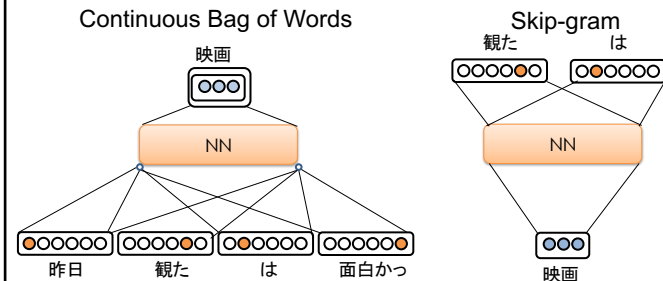
- 自動アノテーション
 - いくつかの手掛かりを元に自動で正解ラベルを付与する
- 例：ユーザのプロフィール推定
 - 人手で整備したルールに適合するプロフィールを持つユーザに正解ラベルを付与する



第22回SIG-AM研究会 2019/7/7

近年の研究開発で活用している手法・道具

- 単語分散表現(word2vec)
 - 言わずとした今の分散表現ブームの火付け役
 - 設定された予測タスクを解くことでNNを分散表現を学習する
 - 分散表現はいくつかでているが、実用上はword2vecで十分



第22回SIG-AM研究会 2019/7/7

近年の研究開発で活用している手法・道具

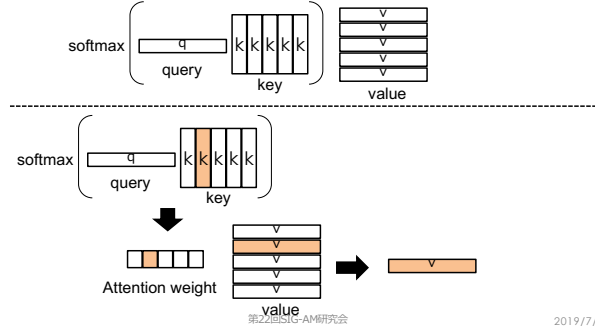
- fasttext (文書分類, 単語分散表現)
 - FAIRによる単語分散表現/文書分類モデル構築のためのツール
 - 単語分散表現はsub-word informationの利用に対応

第22回SIG-AM研究会 2019/7/7

近年の研究開発で活用している手法・道具

■ Attention Mechanism

- Encoder-decoderモデルにおいて、入力ベクトルのうち、ピンポイントで参照すべき情報に注目するための仕組み
- queryに一致するkeyを呼び出し、対応するvalueを取り出す仕組み

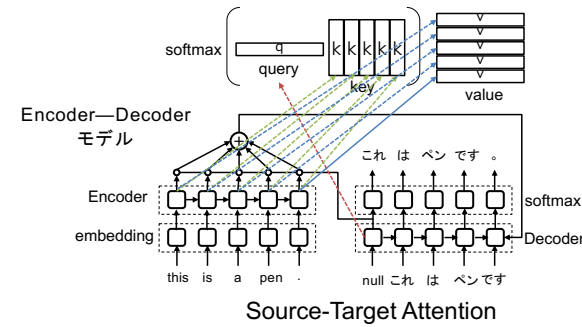


2019/7/7

近年の研究開発で活用している手法・道具

■ Attention Mechanism

- Encoder-decoderモデルにおいて、入力ベクトルのうち、ピンポイントで参照すべき情報に注目するための仕組み
- queryに一致するkeyを呼び出し、対応するvalueを取り出す仕組み

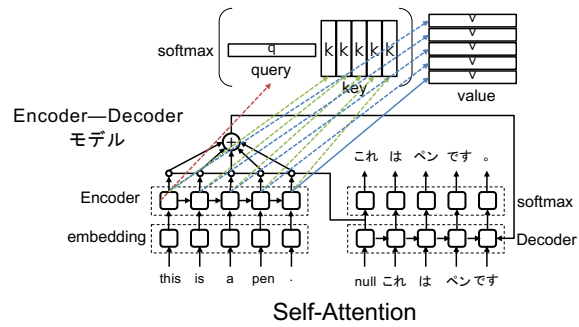


2019/7/7

近年の研究開発で活用している手法・道具

■ Attention Mechanism

- Encoder-decoderモデルにおいて、入力ベクトルのうち、ピンポイントで参照すべき情報に注目するための仕組み
- queryに一致するkeyを呼び出し、対応するvalueを取り出す仕組み

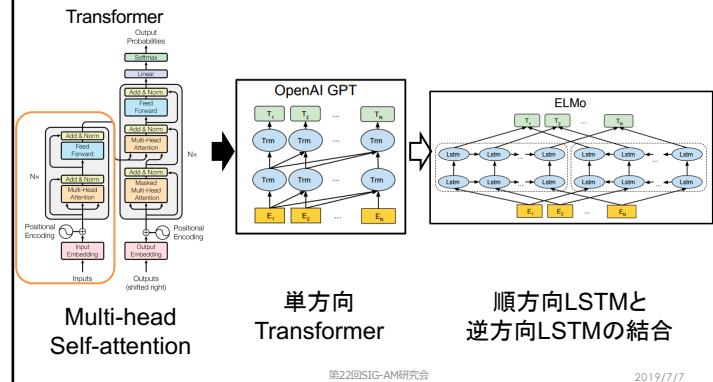


2019/7/7

近年の研究開発で活用している手法・道具

■ BERT (文分散表現)

- 文分散表現の発展



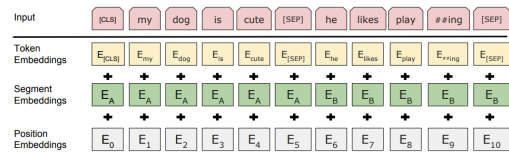
第22回SIG-AM研究会

2019/7/7

近年の研究開発で活用している手法・道具

■ BERT (文分散表現)

- 既存手法
 - Open AI: 単方向Transformer
 - 次の単語を予測するタスクを解くため、先の単語を予測できない
 - ELMO: 順方向LSTMと逆方向LSTMの連結による双方向学習を実現
 - 順方向LSTMと逆方向LSTMを同時に学習することができない
- 双方向Transformer
 - Randomにマスクされた単語を周辺情報から予測する



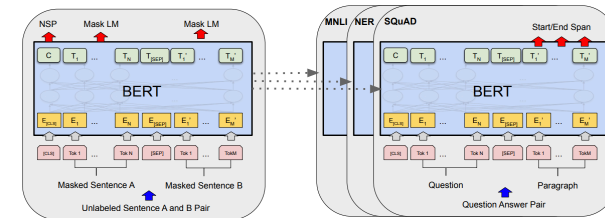
第22回SIG-AM研究会

2019/7/7

近年の研究開発で活用している手法・道具

■ BERT (文分散表現)

- Fine-Tuning
 - 大規模データから文分散表現を事前学習
 - 事前学習したモデルを用いて、各タスクに合わせてモデルをFine-Tuning
 - 評判分析であれば、出力層をポジ/ネガ/ニュートラルの3出力



第22回SIG-AM研究会

2019/7/7

今日の発表

提供機能と利用技術の変遷

近年の研究開発で活用している手法・道具

その他のデータ分析技術の活用と
実務上の失敗談

第22回SIG-AM研究会

2019/7/7

インタラクションに基づくユーザ属性推定

目的

- ソーシャルメディア上のインタラクションに基づくユーザ属性を作成したい

方法論

- Twitter上でのインタラクションからユーザのネットワークを構築する
- ユーザネットワークにコミュニティ抽出の手法を適用してコミュニティを抽出したのち、プロフィール文で特徴付けし、それをユーザ属性の一つとして扱う

貢献

- Twitterユーザ特有の興味・関心を反映したラベルをユーザに付与することができる

第22回SIG-AM研究会

2019/7/7

インタラクションに基づくユーザ属性推定

仮定
相互メンションしているユーザは
同じコミュニティ

第22回SIG-AM研究会 2019/7/7

インタラクションに基づくユーザ属性推定

ソーシャルメディア上のソーシャルキャピタルを分析することで、
定性的に理解可能なコミュニティを構成することができる

1. Twitter上の相互コミュニケーションからユーザネットワークを構築
2. ユーザネットワークからコミュニティを抽出 (Louvain法)
3. コミュニティごとにユーザプロフィール文を取得し、コミュニティ文書を構築
4. 各コミュニティ文書から特徴語群を抽出 (TF-IDF)
5. 特徴語を用いて、各コミュニティにWikipediaタイトルによるラベルを付与

第22回SIG-AM研究会 2019/7/7

インタラクションに基づくユーザ属性推定

インタラクションをベースとしたネットワークを用いることで、
定性的に理解可能なコミュニティを構成することができる

種類	人手ラベル	自動ラベル	特徴語
地域	新潟	新潟市	野球 北越 長岡 向陽 新津 niigata
地域	福島	福島市	郡山 野球 明成 白河 安積 白河
趣味	野球	日本のプロ野球選手一覧	ファン 応援 選手 阪神 カーブ 観戦
趣味	ポケモン	ポケットモンスターのゲーム用語一覧	スマ ブラ パズ ドラ レート アニメ

第22回SIG-AM研究会 2019/7/7

インタラクションに基づくユーザ属性推定

インタラクションをベースとしたネットワークを用いることで、
定性的に理解可能なコミュニティを構成することができる

種類	人手ラベル	自動ラベル	特徴語
職業	エンジニア	Python	haskell python エンジン vim microsoft engineer ruby
職業	トレーダー	投資信託	投資 トレーダー fx 株式 トレード 相場 先物
政治・思想	ネット右翼	自由民主党 (日本)	安倍 原発 反日 支持 日本 保守 政権
政治・思想	左翼	原子力発電	原発 反対 nukes _m racism tpp 戦争 被曝

第22回SIG-AM研究会 2019/7/7

インタラクションに基づくユーザ属性推定

インタラクションをベースとしたネットワークを用いることで、定性的に理解可能なコミュニティを構成することができる

属性種類	属性ラベル
興味・関心	サッカー、野球、アニメ（女性）、アニメ（男性）、ゲーム、テーマパーク、創作（小説、絵、歌）
ファン	アイドル、ジャニーズ、女性声優、男性声優、ミュージシャン（J-POP、K-POP）
政治思想	自民党支持、民進党支持
職業	研究者、トレーダー、エンジニア
地域高校	静岡県、栃木県、大阪府、沖縄県
地域大学	東京都、九州、中部、近畿

インタラクションに基づくユーザ属性推定

- Louvain法
 - Modularityという指標を最大化するクラスタリング手法
 - 既存のModularity-basedクラスタリング手法よりも、高速に処理することが可能

$$- \text{Modularity} : Q = \sum_{i \in C} \left\{ \frac{e_{ii}}{2m} - \left(\frac{a_i}{2m} \right)^2 \right\}$$

実際にクラスタ i に含まれるエッジ数の割合

ランダムグラフにおいてクラスタ i に含まれるエッジ数の割合

インタラクションに基づくユーザ属性推定

- TF-IDF法
 - ある文書集合において、各文書に特徴的な単語に高いスコアを与える
 - $tfidf(t, d, D) = tf(t, d) \times idf(t, D)$
 - $tf(t, d) = \frac{n_{t,d}}{\sum_k n_{k,d}} = \frac{\text{文書}d\text{内の単語}t\text{の出現頻度}}{\text{文書}d\text{内の全単語出現頻度}}$
 - 単語 t の文書 d における出現頻度に関する重み
 - 単語 t の出現回数が多いほど大きくなる
 - 文書 d が長い文書であるほど小さくなる
 - $idf(t, D) = \frac{|D|}{|\{d: t \in d\}|} = \frac{\text{単語}t\text{が含まれる文書数}}{\text{全文書集合}D\text{に含まれる文書数}}$
 - $idf(t, D) = \log \frac{1}{df(t, D)}$
 - 単語 t が含まれる文書数に関する重み
 - 単語 t が含まれる文書数が少ないほど値が大きくなる

ハッシュタグ推薦（画像）

- 目的**
 - 入力した画像からInstagramらしいハッシュタグを取得したい
- 方法論**
- 貢献**
 - 社外秘
(スライド公開時はマスク予定)
 - 特定のユーザ群をターゲットとする場合のコンテンツ作成やクリエイティブ作成の手掛かりとなる

ハッシュタグ推薦 (画像)

第22回SIG-AM研究会 2019/7/7

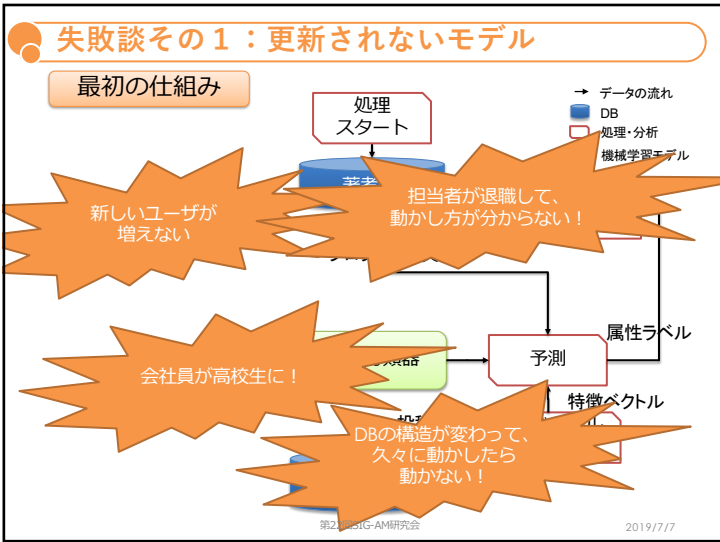
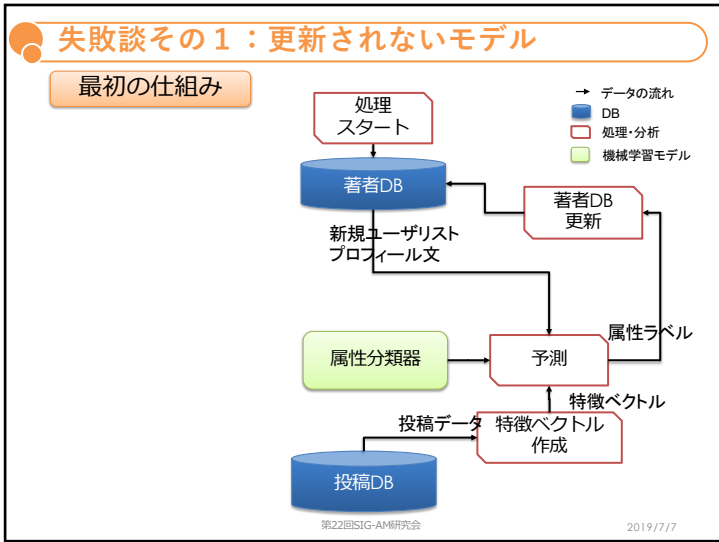
失敗談その1：更新されないモデル

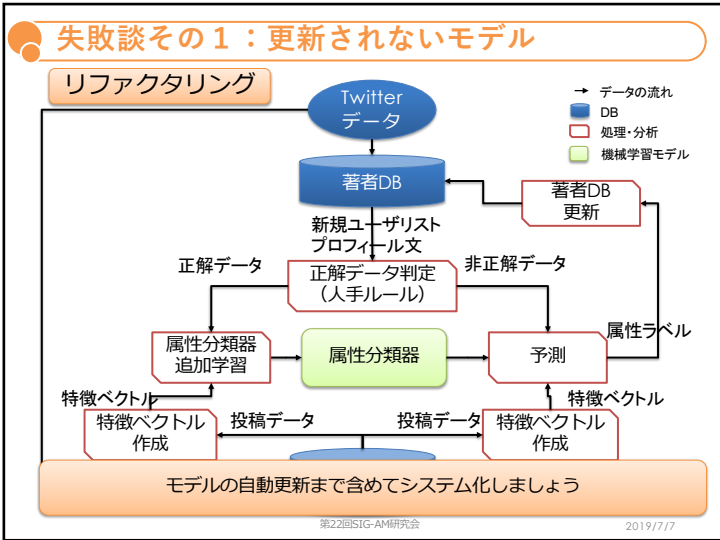
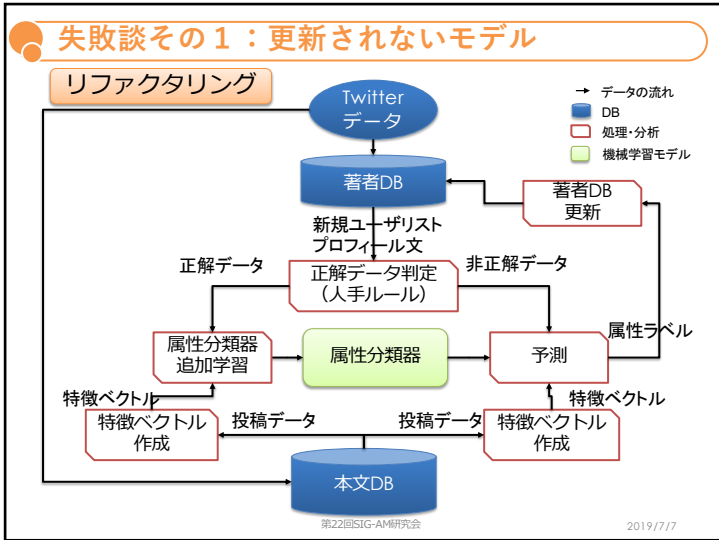
自社製品/ブランドについて、どこの誰が興味を持っているかがわからない

マーケティング担当者

ユーザの属性推定

第22回SIG-AM研究会 2019/7/7





失敗談その2：死蔵累々

- 数々の死蔵品・・・

社外秘
(スライド公開時はマスク予定)

第22回SIG-AM研究会 2019/7/7

失敗談その2：死蔵累々

- 数々の死蔵品・・・

弊社の顧客は日本企業ばかりなので中国語読めない！

問題は、むしろコミュニケーションだった！

予測できてもアクションがうてない！

欲しかったのはもっと中長期の予測

その 이슈が、本当に必要とされているかを精査しましょう

第22回SIG-AM研究会 2019/7/7

失敗談その3：新しすぎる技術の本格活用

- 新たなテキスト分析処理基盤開発プロジェクト
 - 2012年当時

elasticsearch
全文検索エンジン
Ver1.x

hazelcast
分散処理
(インメモリーデータグリッド)

共起語分析
係り受け分析
評判分析

第22回SIG-AM研究会 2019/7/7

失敗談その3：新しすぎる技術の本格活用

- 新たなテキスト分析処理基盤開発プロジェクト

srs / elasticsearch-hazelcast

Latest commit 1d86550 on 3 Oct 2012

README.md	first commit	7 years ago
build.gradle	Added more files	7 years ago
settings.gradle	Added initial gradle files	7 years ago

第22回SIG-AM研究会 2019/7/7

失敗談その3：新しすぎる技術の本格活用

- 新たなテキスト分析処理基盤開発プロジェクト
 - 2012年当時

分析機能を使うとすぐ高負荷に!

Hazelcastとの互換性維持のためにバージョンアップできない!

マッチしていないテキストが分析結果に含まれる!

elasticsearch
全文検索エンジン
Ver1.x

hazelcast
分散処理
(インメモリーデータグリッド)

共起語分析
係り受け分析
評判分析

大規模システムの本番開発には、
ある程度枯れた技術/技術の組み合わせを使いましょう

第22回SIG-AM研究会 2019/7/7

おわりに

おわりに

第22回SIG-AM研究会 2019/7/7

おわりに

- ホットリンクでの利用技術の変遷を紹介
 - 徐々にモデル構築のみに注力できるようになってきた
 - 最先端技術が実用化されるまでの期間が年々短くなってきている
 - ・ つまり、技術の賞味期限も短くなってきている
- 研究開発で活用している手法・道具の紹介
 - アノテーションの労力を減らす手段が増えてきている
 - 事前学習モデルの活用により、精度向上が容易に
- その他のデータ分析技術の活用と実務上の失敗談
 - 自然言語処理と他の技術の組み合わせも有用
 - 実務では、長期に使われることも想定する必要がある

第22回SIG-AM研究会

2019/7/7

参考文献

- [Joachims1998] Text categorization with support vector machines: Learning with many relevant features, Thorsten Joachims, ECML, 1998
- [Kudo 2006] MeCab, Taku Kudo, <https://taku910.github.io/mecab/>
- [Yoshinaga 2009] JDepP, Naoki Yoshinaga, <http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/jdepp/>
- [橋本2011] 構文・照応・評判情報つきブログコーパスの構築, 橋本力, 黒橋禎夫, 河原大輔, 新里圭司, 永田昌明. (2011). 自然言語処理 Volume 18, Number 2, pp.175-201.
- [Micholov 2013] Distributed Representations of Words and Phrases and their Compositionality, Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean, CoRR
- [Joulin 2017] Bag of Tricks for Efficient Text Classification, Joulin, Armand and Grave, Edouard and Bojanowski, Piotr and Mikolov, Tomas, EACL, 2017
- [Jacob 2018] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova

第22回SIG-AM研究会

2019/7/7

CM：学習済みモデルの配布

- 単語分散表現：hottoSNS-w2v
 - <https://github.com/hottolink/hottoSNS-w2v>
- 文分散表現：hottoSNS-bert
 - <https://github.com/hottolink/hottoSNS-bert>

モデル	相関係数
日本語大規模SNS+Webコーパス	0.548
Wikipedia (ホットリンク)	0.478
Wikipedia (東北大)	0.472

モデル名	分かち書き	学習言語	学習ドメイン
BERT Multi	WordPiece	多言語	Wikipedia
BERT JP	SentencePiece	日本語	Wikipedia
hottoSNS-BERT	SentencePiece	日本語	Twitter

第22回SIG-AM研究会

2019/7/7

- 以下、おまけスライド

第22回SIG-AM研究会

2019/7/7