言語処理がつなぐビッグデータと社会

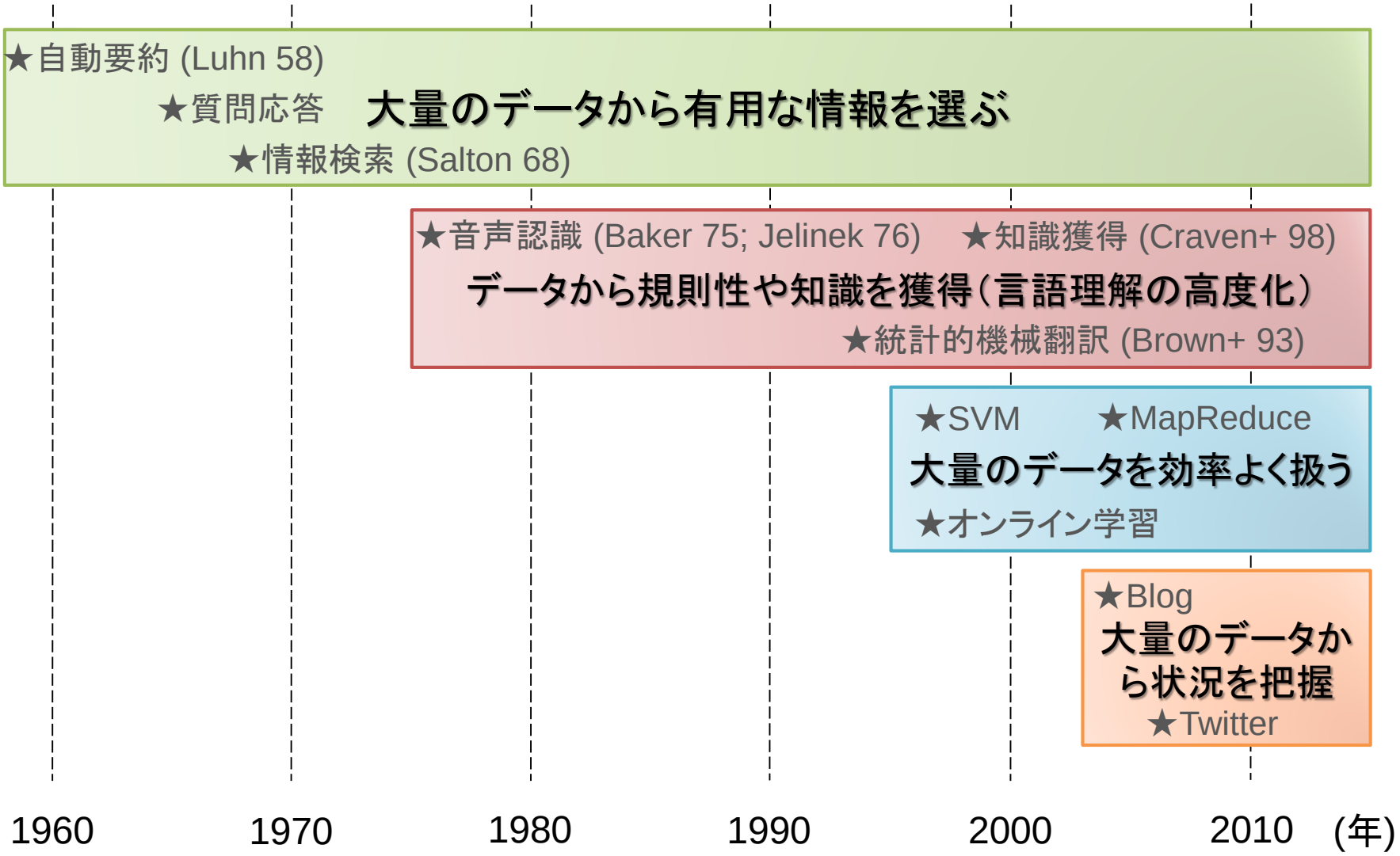
東北大学大学院情報科学研究科
JST 戦略的創造研究推進事業「さきがけ」・「情報環境と人」
岡崎　直観（okazaki@ecei.tohoku.ac.jp）

http://www.chokkan.org/
@chokkanorg

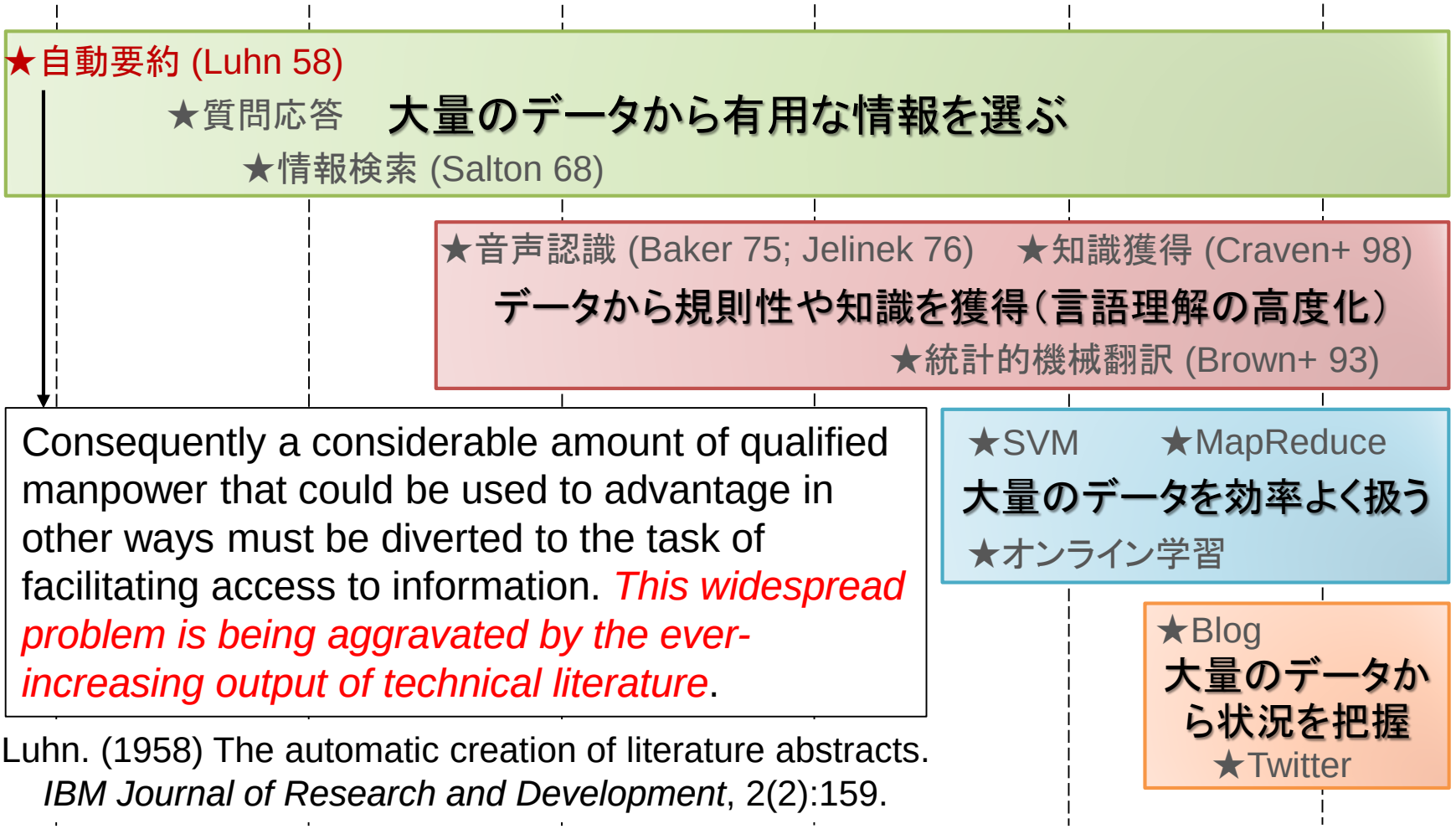
言語処理がつなぐビッグデータと社会？

- ビッグデータそのものを研究している訳ではない
- 「ビッグデータ」という言葉をタイトルに入れた
  - 大量のデータを扱う
  - 大量のデータが次々に到着する（今回は話さない）
  - データの種類や情報源がバラエティに富む
  - 分野外から見ればビッグデータ
    - 「この言語データの解析を自然言語処理で出来ませんか？」
- 自然言語処理と「大量のデータ」の密接な関係

大量のデータと言語処理

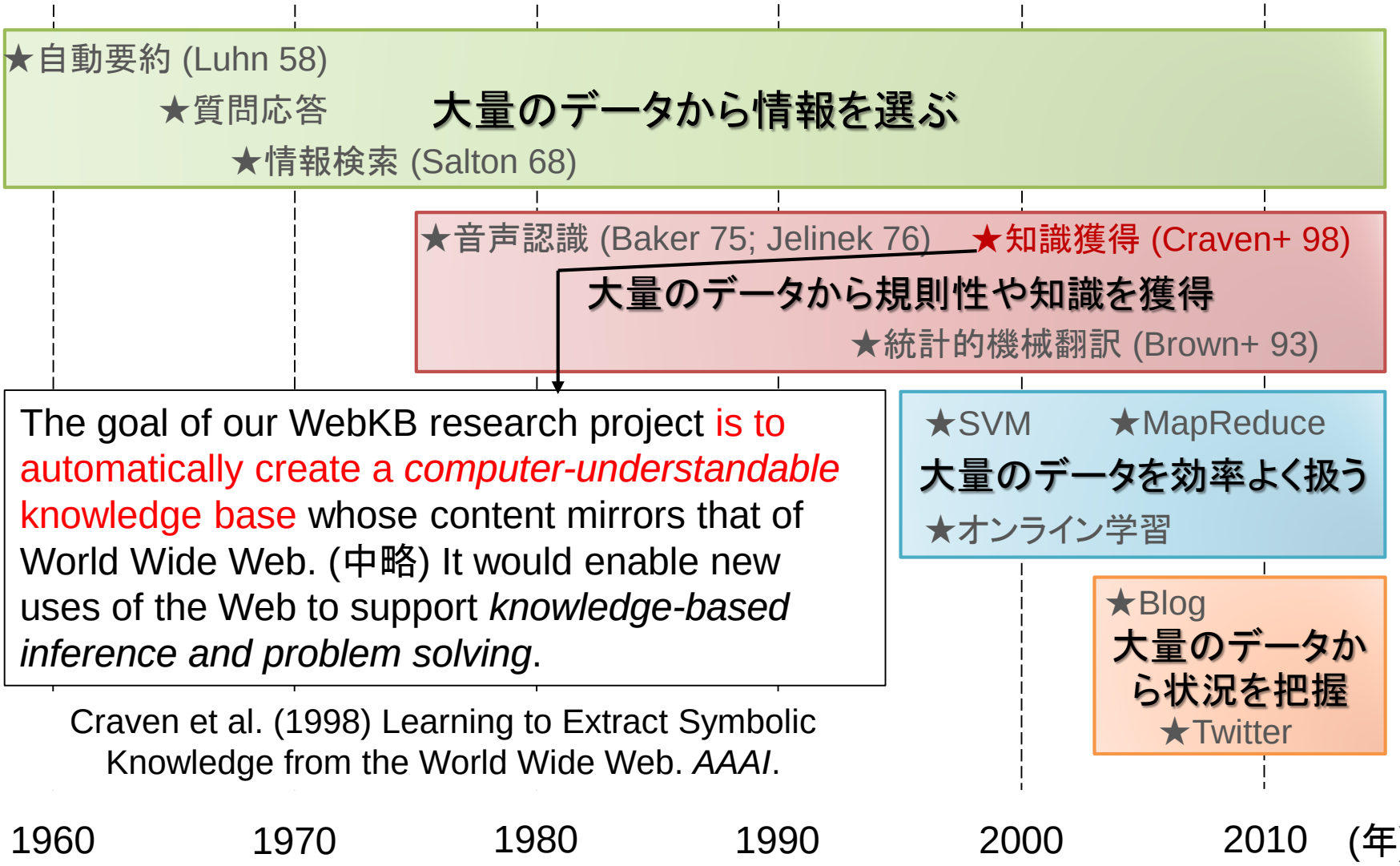


大量のデータと言語処理

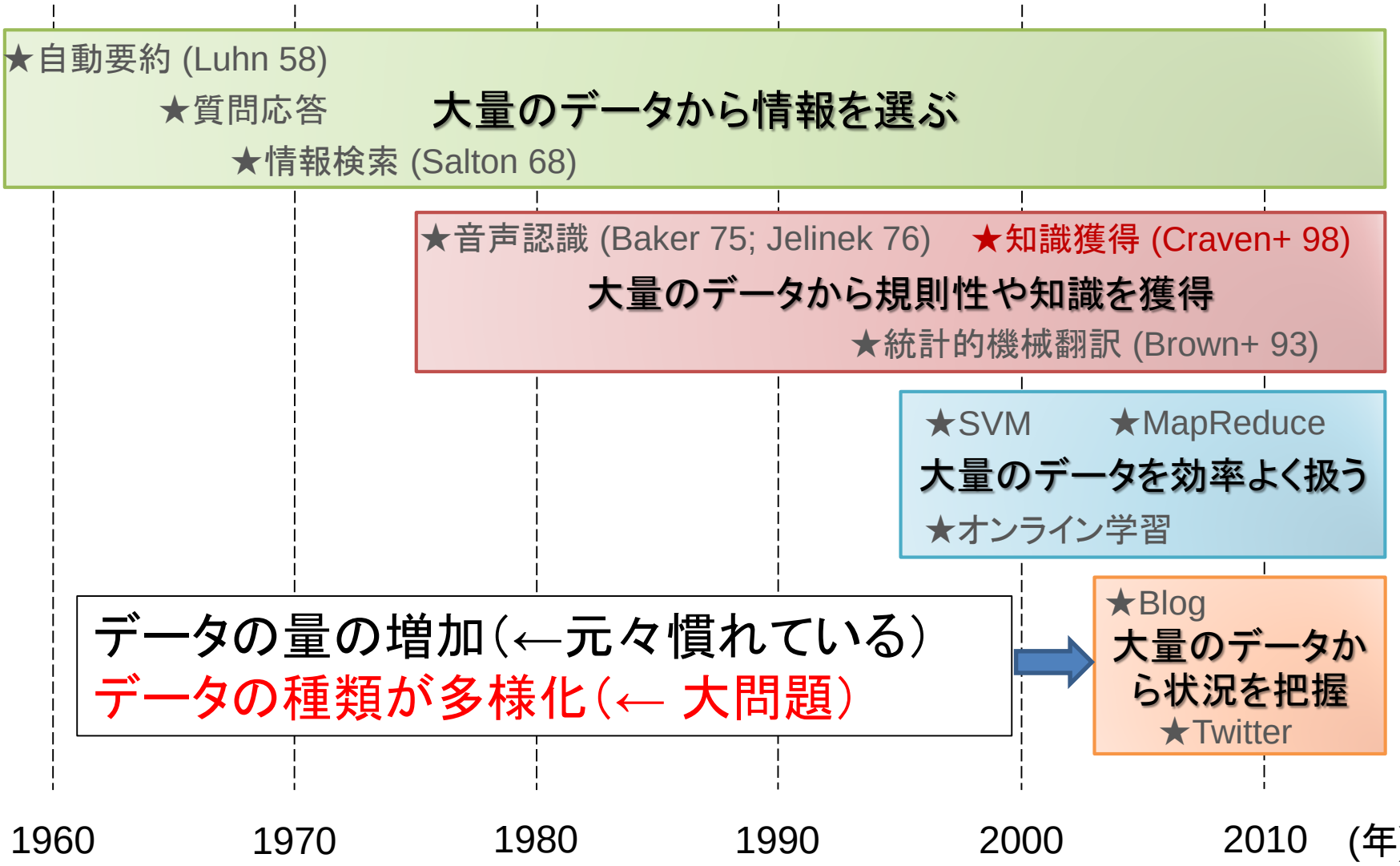


Luhn. (1958) The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159.

大量のデータと言語処理



大量のデータと言語処理



押し寄せる言語処理への期待

- 古典的な分析手法と言語処理による分析を対比したい
  - 災害時の人々の寄付行動を分析したい
  - 言語データおよび言語処理のポテンシャルを再考したい
- 玉石混交の情報の信憑性を判断したい
- 大量の情報を構造化・整理したい
- 他の研究分野と言語を繋ぎたい
  - 画像から説明文を生成する，曲から歌詞を生成する
  - 画像や動画から状況を認識する
- ビッグデータから世の中の動き・関心を探りたい
  - 選挙や株価市場のトレンドを予測する
  - データジャーナリズム

本講演の内容

- 「ビッグデータ」を言語処理で分析し，データから世の中の動向を探り，報道で活用された2事例を紹介
  - 世の中の興味・関心の抽出
  - 福島の桃に関する風評の分析
- これらの事例紹介の中で，言語処理でビッグデータと実社会を繋ごうとした時の問題点を説明
- 社会に浸透する言語処理に向けて

世の中の興味・関心の抽出

データジャーナリズム
(data-driven journalism, computational journalism)

- 「データジャーナリズムはプロセス」（Lorenz, 2010）
  - データ: データを収集し，クレンジング，構造化
  - マイニング: データの中から知識を取り出す
  - 可視化: 取り出した情報をグラフィックなどで表現
  - ストーリー: 記事化（ストーリーを作らなければただの統計資料）
- 「ストーリー化する前にデータを処理」（van Ess, 2013）

- 参考資料
  - Sarah Cohen, James T. Hamilton, Fred Turner. 2011. Computational Journalism. *Communications of the ACM*, Vol. 54 No. 10, Pages 66-71.
  - Jonathan Gray, Liliana Bounegru, and Lucy Chamber. 2012. *The Data Journalism Handbook*. Oreilly & Associates Inc.
  - 米国ジョージア工科大学の講義「Computation + Journalism」

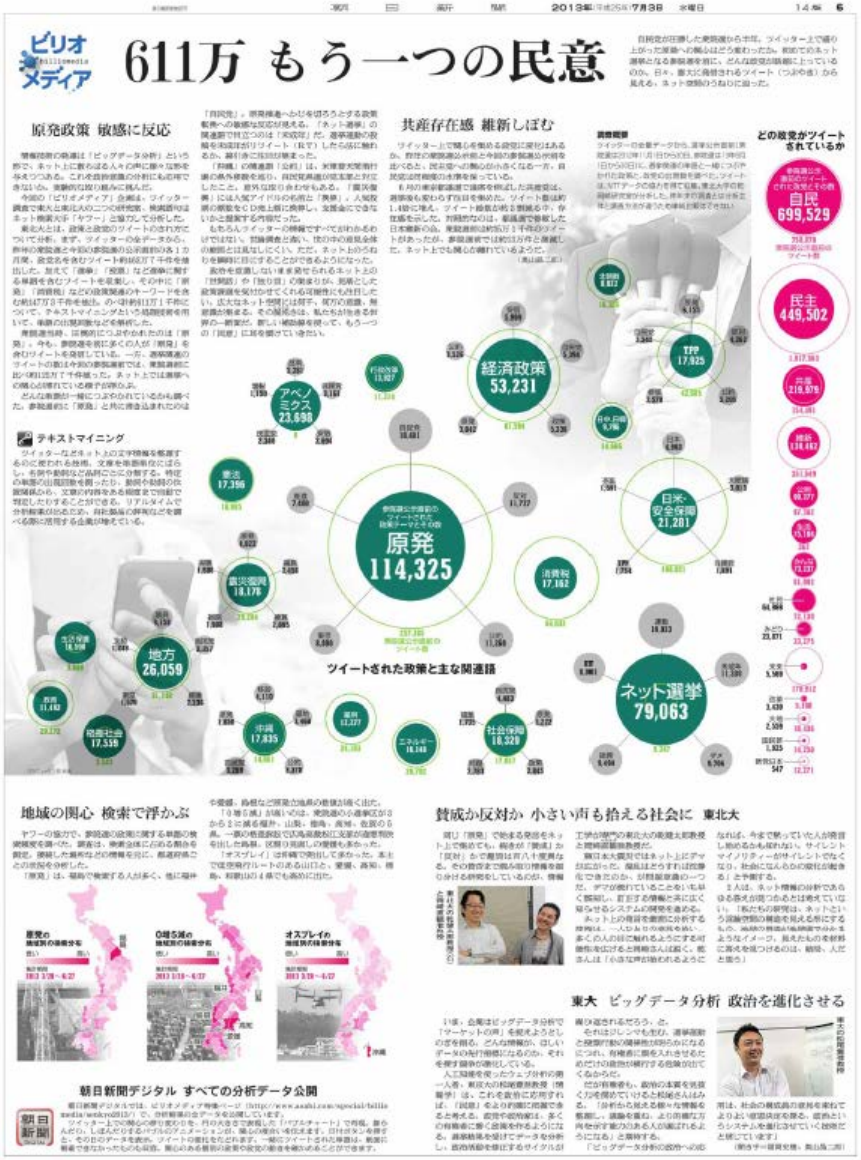
参院選公示日前のツイートの分析
（2013年7月3日 朝日新聞朝刊･6面掲載）

分析内容

政策･政党名でツイートを検索し，日ごとのツイート数，関連語を分析

知見

- ツイート総数が衆院選前と比べ3割減（参院選への関心低い）
- 共産党に関するツイートが増加
- 維新の会のツイートは激減
- ネット選挙に関連して，未成年のリツイート問題が話題に
- 原発を含むツイートが依然多い


ビリオメディア

611万 もう一つの民意

原発政策 敏感に反応

共産存在感 維新しぼむ

どの党がツイートされているか

テキストマイニング

ツイートされた政策と主な関連語

原発 114,325

経済政策 53,231

ネット選挙 79,063

地方 26,059

地域の関心 検索で浮かぶ

賛成か反対か 小さい声も拾える社会に 東北大

東大 ビッグデータ分析 政治を進化させる

朝日新聞デジタル すべての分析データ公開

ツイート分析の流れ

- 関連語抽出の例（名詞の頻度計測）

社民党Facebookの「萌えみずほたん」が似てなすぎると話題に
まさかの萌え化：**社民党**の福島瑞穂党首が萌え化してFacebookに降臨
社民党の福島瑞穂が萌えキャラ化！！お前らのみずほたんだろ

形態素解析で名詞を認識

各名詞の出現頻度を計測

関連語	頻度
萌え	4
みずほたん	2
Facebook	2
福島瑞穂	2

- スパム除去や形態素解析辞書の修正等，表には出ない泥臭い作業に追われる

見過ごしやすい検索クエリの部分文字列

- 「民主」でツイートを検索
  - ○「**民主**党の海江田万里代表は17日，・・・」
  - ×「秘密保護法案は国**民主**権と**民主**主義を・・・な法律である」
- 「民主党」でツイートを検索
  - ○「**民主党**の海江田万里代表は17日，・・・」
  - ×「自由**民主党**公認・参議院議員の○○○○をよろしくお願いします」
- 「共産党」でツイートを検索
  - ○「**共産党**はゆるキャラを使ったPRサイト・・・をオープンした」
  - ×「中国**共産党**指導部が地方当局者の・・・」

ツイートの目視確認は欠かせない

なぜ無難な分析手法になるのか

- ツイートや単語の頻度計測だけではつまらない
  - 研究の世界ではベースライン手法
- 分析結果の正確性に責任を持つ必要がある
  - 自然言語処理の解析は100%正しい訳ではない
- 分析手法の透明性を確保する必要がある
  - 一般読者を想定し，分かりやすい分析・結果が望まれる
  - 「頻度」よりも性能のよい統計尺度（例えばTF*IDF）が使えない
  - 特定の団体・政党への利益・不利益にならないように，中立性を保つ
  - 報道だけでなく，様々な局面で手法の透明性の確保は重要

紙面作りに貢献できたが・・・

- クエリに基づく仮説検証型の分析
  - 記者さんの目（検索クエリ）に依存した分析の切り口
  - SNSユーザの偏りに関する批判
  - 元々テキストマイニングは仮説検証に向かない

- （制約の多い中で）言語処理らしい分析を売り込む
  - 調べたい関心に合わせて最適な解決策を設計すべき
  - データから世論を探る（仮説発見）にはどうすればよいか？
  - 現状の自然言語処理で精度100%を目指してみる

世の中の関心を自動的に掘り起こす

（2013年7月26日 朝日新聞朝刊・9面掲載）

分析内容

記者のフィルターを介さずに，ツイートから社会の論点・関心を抽出

結果と課題

- 「児童ポルノ禁止法改正案」「Jリーグの2ステージ制」など，新聞が取り上げていない話題を抽出
- 分析結果がそのまま新聞記事に掲載された
- 賛否を分離して数を出すことは出来なかった

9 12版▲ 2013年(平成25年)7月26日 金曜日

つながる力 次こそ真価

ビリオメディア

賛成、または明確な反対ではないツイート

憲法改正 1,579

自民党 448

TPP 4,738

TPPに反対し、食の安全・農業・医療・中小企業を守る

TPPに入れば脱原発ができない

原発 3,617

大事故おきても誰も責任とらず

原発廃炉 770

外国人参政権 919

Jリーグの2ステージ制 816

閣僚決議案 461

東京五輪招致 378

消費税の増税 1,679

日本の政治権害する一方的行動 371

反対意見のツイート

「議論」ツイート上位20（6/21～7/20）

表現規制 1,382

児童ポルノ禁止法改正案 1,686

渡辺美樹ワタミ会長の自民党公認 850

戦争 827

憲法96条改正 1,290

憲法9条改正 372

オスプレイ 275

原発再稼働 205

選挙の陰で熱い議論

争点にならなくとも

児童ポルノ改正法案には反対です。立体起動装置をマスターした部下によって、賛成の政治家を駆逐させます。

6月3日 0:22 RT数:1212 返信 リツイート ★お気に入り

（朝日新聞社様にライセンス提供）

世の中の関心を自動的に掘り起こす
（渡邉ら，2013）

TPPやACTAに絶対に反対する
これは嫌なので**命を無駄にする徴兵制**には反対です
なぜアメリカ人が**銃規制**に反対するのか

前処理

1.文分割
2.賛成反対文抽出
3.ノイズ処理
4.文節解析

ルール適用

1.賛否が受動態なら「意見なし」にする
2.同じ節に賛否が複数ある場合，前の賛否を全て削除
3.賛否よりも前にストップワードが有る場合，それ以前を削除
4.賛成/反対の直前に接頭詞,形容動詞語幹,副詞可能ならその単語を消す
5.この/あの/そのなどの指示代名詞があると「意見なし」にする
6.賛成/反対の直前に来ると意味が変わるもの(上下反対)の除外
7.賛成/反対の直後に来ると意味が変わるもの(反対車線)の除外
8.「～的に」を含む節の除外
9.賛成/反対を含む節の前節に「形容動詞語幹or副詞可能+に」の時その節を除外
10.主格(も/は格)の消去
11.一番最初の単語が「助詞，接頭詞，副詞，フィラー，連体詞」の場合その節を削除

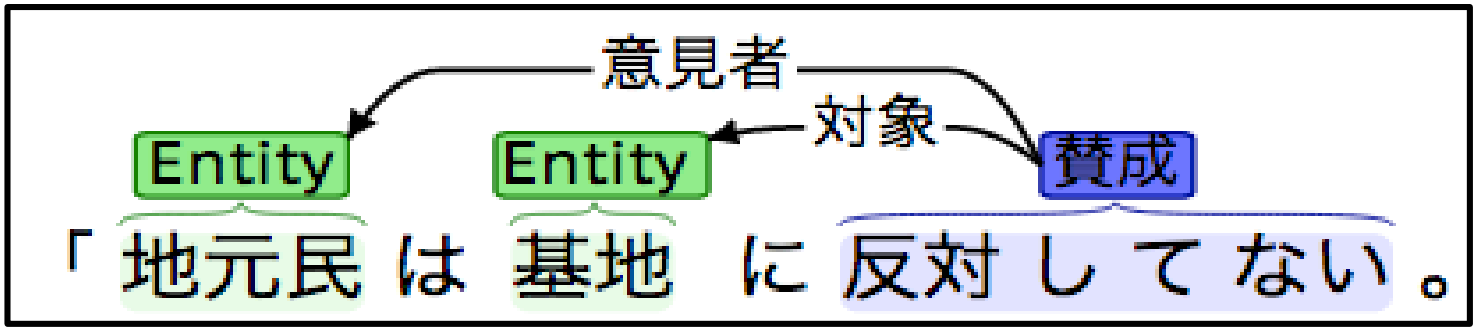
二格の抽出

1.二格の直前が人物の場合「意見なし」
2.二格「によって」の場合「意見なし」
3.二格の直前の名詞の「の→こと」に変更
4.二格が「意見/件」なら「意見なし」
5.途中に「並立助詞」があったら分割
6.先頭に「記号」があったら記号を削除

賛否の主体・極性の認識

紙面掲載後の取り組み（渡邉ら，2013）

- 社会の論点を探るために必要な要素
  - 対象: 賛成・反対のターゲットになっている事柄
  - 意見者: 誰が賛成・反対を主張しているのか
  - 極性: その意見者は賛成しているのか，反対しているのか
- 述語項構造解析の一種として取り組む



  - 意見者の認識精度は41.9 F1スコア（CRFで学習）
  - 賛否極性の認識精度は99.02%（「つつじ」等を素性に用いSVMで学習）

タスク設定の限界

- 賛否の対象として「憲法9条改正」と「憲法96条改正」は同じ？
  - 文字列類似度や文脈類似度は非常に高い

- 自然言語処理が扱う「意味の理解」の研究と，報道に耐えうる「世論の理解」の間にはギャップがある
  - 米著名投資家「アベノミクスには反対」
    意見者　対象　反対
    → 主体と日本の世論との関係
  - 東京五輪に反対する人間は日本人じゃない
    対象　反対　意見者
    → 意味表現と世論との関係
  - 多くの人が東京五輪に反対しているけど，
    意見者　対象　反対
    → 目的語の省略
    俺はやったらよいと思うよ.
    → 同一文内での意見の対比

福島の桃に関する風評の分析

震災ビッグデータⅡ
(2013年9月8日放送 NHKスペシャル)

研究の目的

- ツイッターから「福島産の桃」に関する風評を探る
  - 情報交換・議論はどのように行われているのか
  - 意見を変える人，変えない人の特徴はどこにあるのか
  - 考えを変えるためには，どのような情報がカギとなるのか

- 福島県産の「桃」に焦点をあてる
  - 原発や食品に関するツイート数はよく変動する（→語られる）
  - 食品の中では「米」に続き，「桃」がよく語られている
  - 嗜好品のため「買う」「買わない」の議論が起こりやすい
  - 福島県のPR戦略の中心に位置づけられている

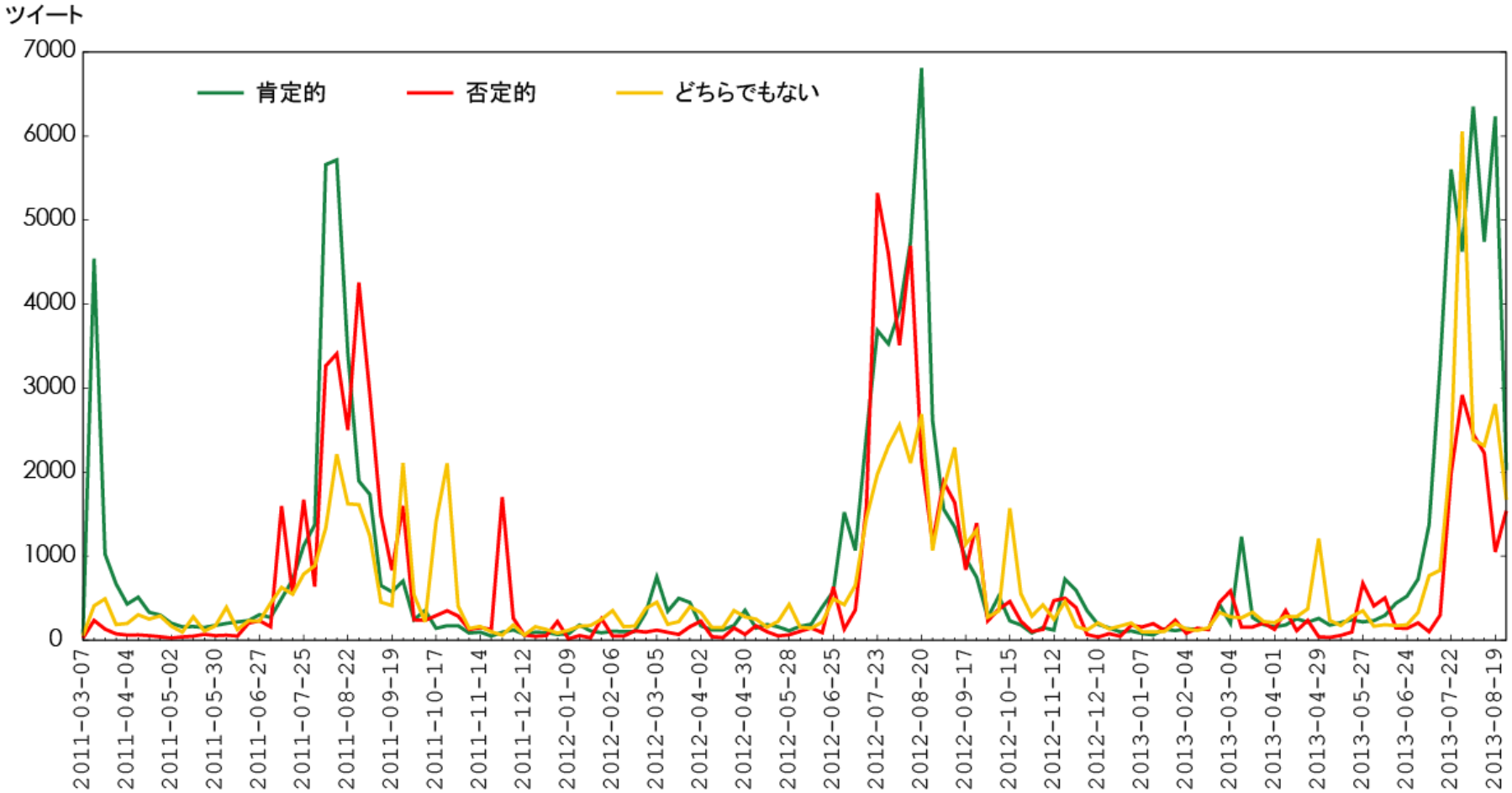
- 「福島の桃」を通じて風評の実態と対策を探る

福島の桃に対して肯定的か否定的か

- 各ツイートを以下の3つのカテゴリに自動分類
  - 肯定的: 福島の桃を消費するという行為を報告しているものや，その行為を推奨するもの，福島の桃を応援するツイート
  - 否定的: 福島の桃を消費するという行為を忌避したり，行為に懸念を示すようなツイート
  - その他: 福島の桃を消費することに中立的，肯定とも否定とも判断できない場合，単に事実を述べているだけのツイート

- 4,824件の訓練事例を用意し，分類器を学習
  - BOW，否定表現，評価表現辞書など，ひと通りの素性を導入
  - 自動分類の精度は74.27%（10分割交差検定による）

ツイート数の推移（肯定・否定別）

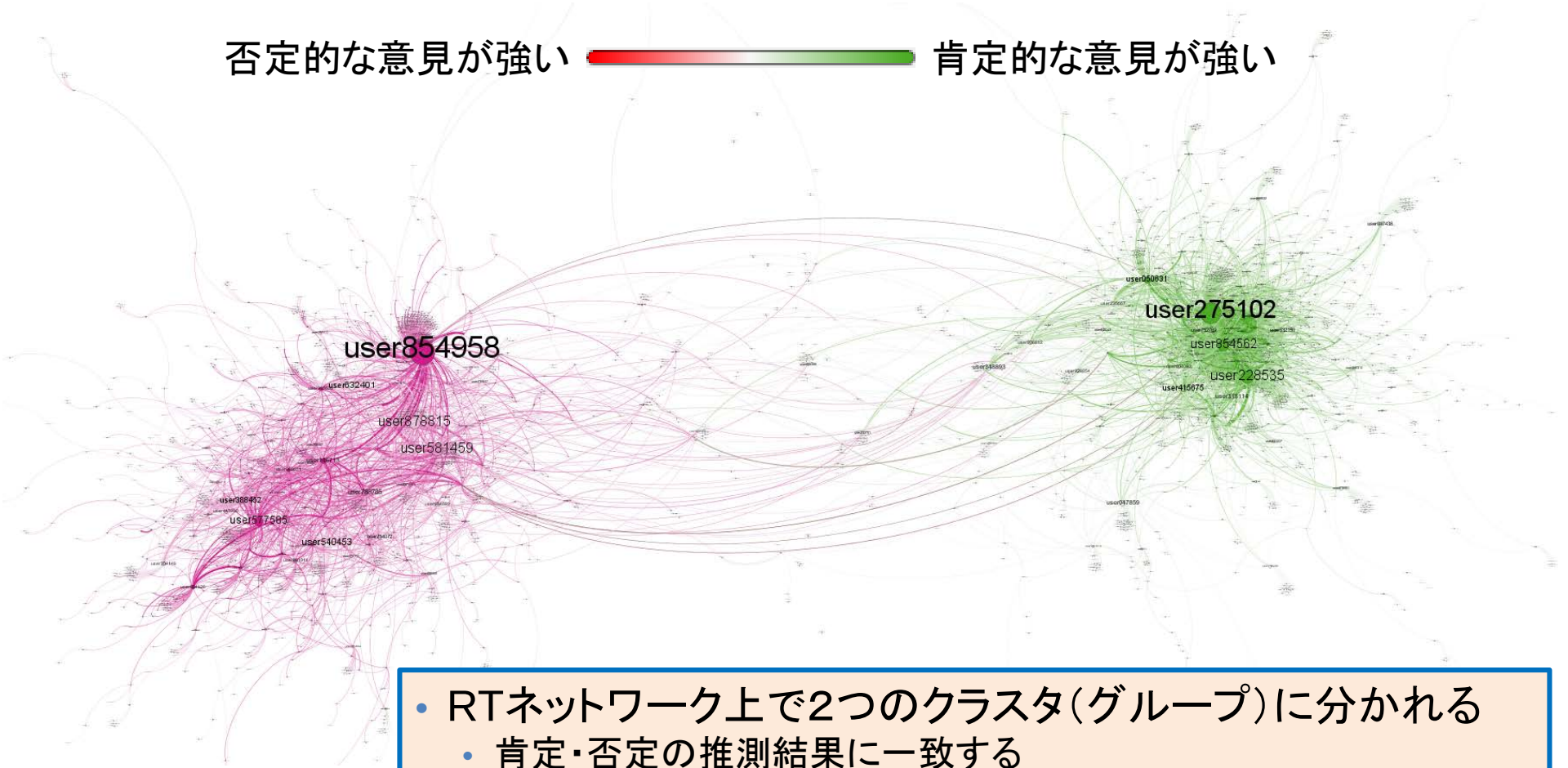


2011年～2013年夏の主な話題

年	肯定側	否定側
2011	おじいちゃんおばあちゃんの愛， 農家に向けられる冷たい視線に対する悲しみ， 福島の桃の当たり年	福島の桃の値崩れ→加工食品・産地隠蔽の懸念， 市長の不適切発言， 福島県発表で不検出も市民団体調査でセシウム検出続出
2012	福島の人の偽らざる声→普通， 桃農家の除染の努力による不検出， 福島のモモ900個タイで完売	産地偽装の噂・疑い， ピーチプロジェクトに対する批判， 数ベクレルの検出
2013	「桃の涙」， 福島の桃の取り寄せ， 福島の桃「おいしい」と両陛下	首相のパフォーマンスと出荷規制， 桃園のお土産と付近の空間線量， 秋田県でのヨウ素の誤検出， 知事の大阪でのキャンペーン

可視化したリツイート・ネットワーク

否定的な意見が強い　　肯定的な意見が強い



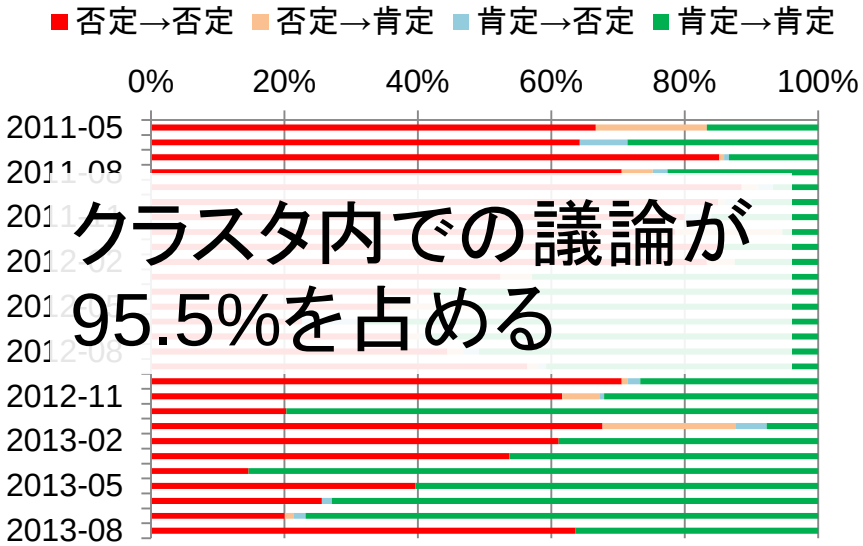
- RTネットワーク上で2つのクラスタ(グループ)に分かれる
  - 肯定・否定の推測結果に一致する
- 反対の立場のツイートはRTされない
  - 反対側の立場のツイートは拡散させたくない

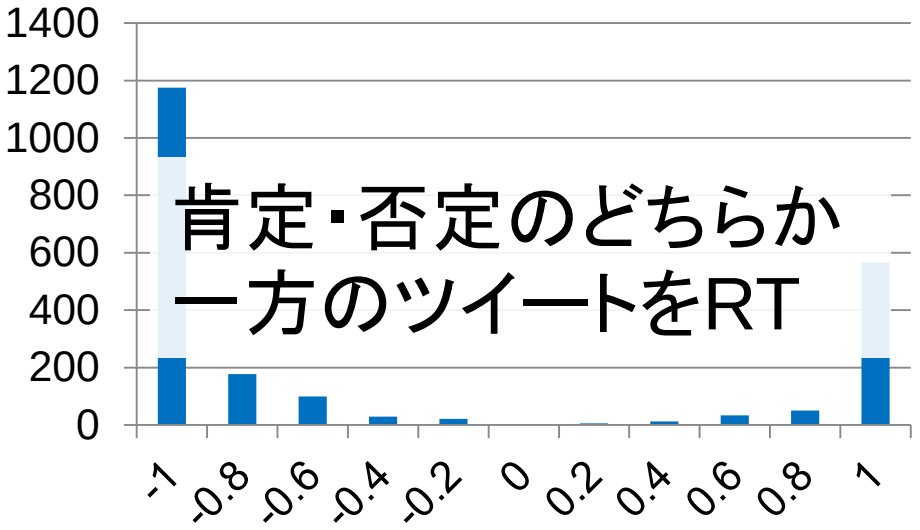
user854958さんの意見（2013年夏）

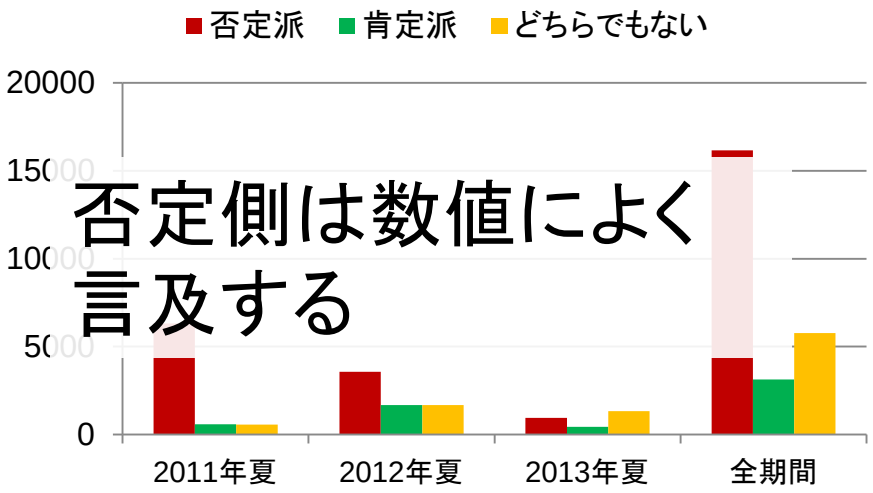
- セシウムの濃度は出来るだけ下げるべき
- 自分で情報を精査すべき
- 努力している福島の農家のことも知るべき
- 福島の桃を食べる
- デマ・危険情報に飛びつく人々への批判

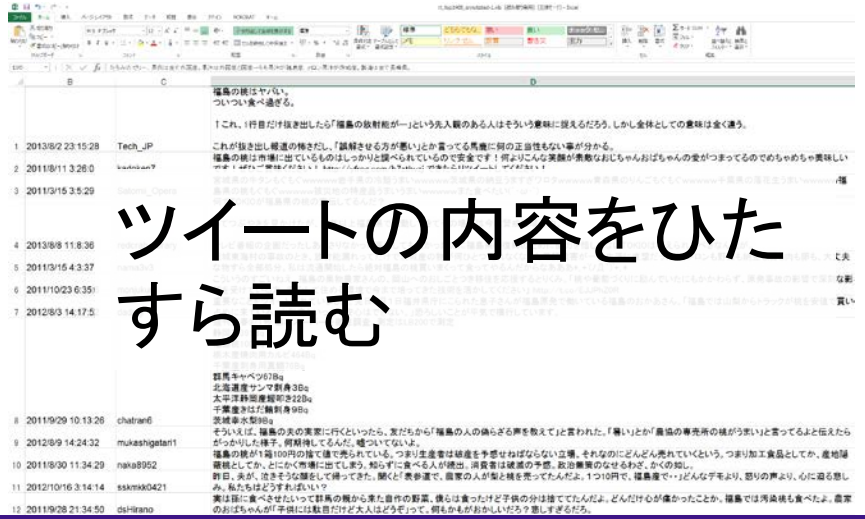
放射性物質に対して万全の対策を行い，正しい情報を届ければ，肯定的に判断する人も出る

別の分析手法で裏付けをとる









リスクコミュニケーション研究としての結論

- ツイッター上では福島の桃に肯定的／否定的な立場に分断る
- 風評に対する対応の難しさが改めて浮き彫りに
  - 考えを途中で変えた人は見つからなかった
  - 立場を横断した健全な議論も見つからなかった
- 産地ではなく精密な情報を使って自ら判断したいという意見を時系列ネットワーク分析で発見
- 検査結果に関する数値データを積極的に開示し，検査体制・検査結果の透明性のアピールしていくことで，両派の溝を解消する可能性が見えた

- 2013年日本リスク研究学会年次大会で優秀発表論文賞

苦戦する言語処理

- 評判分析の精度向上は容易ではない
  - 要知識: 極性を安定して判定するには知識が必要

    例:「今年の福島のももは安値で大阪でバンバン売れたって…」
  - 発言者同士の関係: 同一クラスタのユーザは似ている意見

    例:「県知事は，何故汚染水対応じゃなくて，大阪に桃を配りに来ているの？」

    「ホントだよ．ニュース見るたび泣いてるんだよ！ RT @xxxxx: 県知事は」
    - ネットワーク分析と組み合わせた評判分析
  - まんじゅうこわい（← 絶対に無理）

    例:「福島の桃が怖いです。すごく怖いです。怖くてたまりません。ひと箱でもふた箱でも怖いです。できれば、さくらんぼとかも怖いです」

- user854958さんを発見したのはネットワーク分析
  - 時間とともに評価極性が変化するユーザを探したが，ノイズに埋もれる

社会に浸透する
言語処理に向けて

1. フィールドに飛び込む

- 基礎研究と応用研究の両方に取り組む
  - 言語処理がどのように必要とされるのか分かる
  - 現状の言語処理では考慮していなかったテーマに気づく
  - 言語処理の健全な普及へ
  - 情報処理分野の全体を考えて（by 喜連川先生）

- 飛び込む方法は2つ
  - 出る: NLPのプロとして応用分野での問題にアプローチする
  - 組む: 応用分野の研究者と共同で研究を進める

  ※ 最初は我々が「出る」方がやりやすいと思う

- よい相手と組むと我々も勉強になる・楽しい

2. ツールやデータを整備する

- とは言え，我々の時間は限られている
- 個別のタスクを解くには人手による調整が不可欠

- 自分の分身としてツールやデータを使ってもらおう
  - 我々の研究成果の正しい使い方を布教する
  - 分野外の人でも使えるツールを（nltk, OpenCV, pylearn）
  - ツールのポテンシャルを説明することも大切（word2vec）
  - 再学習が容易な解析ツールを整備したい (by 宮尾さん)

3. 多様なデータへの統一的なアプローチ → グラウンディング再考？

- すべての対象ドメインに取り組むのは無謀
- ビッグデータを統一的に扱う方法論を確立すべき
- 「データと実社会を繋ぐ」＝「データをグラウンディング」
- ビッグデータ時代のグラウンディング問題を考える
  - 画像，音声，動画，センサーデータと言語を接続する
  - 人や場所に関する知識獲得と曖昧性解消
  - 実社会のモノ・コトに関する知識を自動獲得

飛び込もう

- 感謝いたします
  - 貴重な機会を与えて頂いた記念シンポジウムの実行委員会の先生方
  - 言語処理学会の発展にご尽力頂いた先生方
  - 「飛び込む」きっかけを与えて頂いたProject 311の関係者の皆様