

第11回産業日本語研究会・シンポジウム

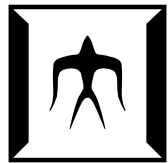
自然言語処理による ニュース記事執筆の現状と課題

岡崎 直観

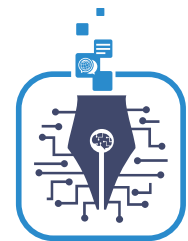
東京工業大学 情報理工学院

okazaki at c.titech.ac.jp

<https://www.nlp.c.titech.ac.jp/>

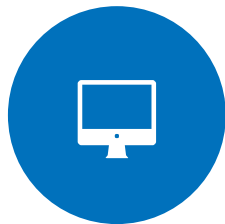


東京工業大学
Tokyo Institute of Technology



OKAZAKILAB

自然言語処理：人間のように言葉を理解・生成できるロボット



機械翻訳

ある言語の文章を別の言語に翻訳する



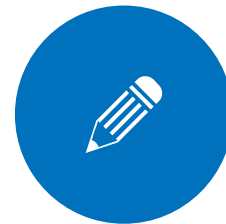
質問応答

自然言語で与えられた質問に答える



対話エージェント

コンピュータと人間の間で会話をする



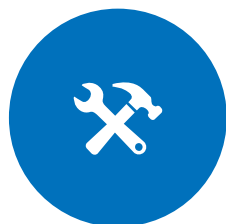
自動要約

情報を集約して文章を生成する



情報検索

大量の文書の中から必要なものを探す



情報抽出

文章から事実や知識を抽出する



意見・感情分析

文章から人間の主観的評価を抽出する



教育支援

人間が良い文章を書くように支援する

ニュース記事と自然言語処理

- ニュース記事をコーパスとして利用
 - 正しい日本語の「お手本」に統語解析・意味解析の「お手本」を付与
 - 例：京都大学テキストコーパス（1995年版毎日新聞）
- **データジャーナリズム**：テキストを分析してニュースの種を得る
 - 様々なテキストデータをマイニングして，社会を観測する
- **ロボットジャーナリズム**：自動もしくは半自動で記事を生成する
 - 記事の自動生成
 - 記事の自動校正
 - 見出しの自動生成

データジャーナリズム

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

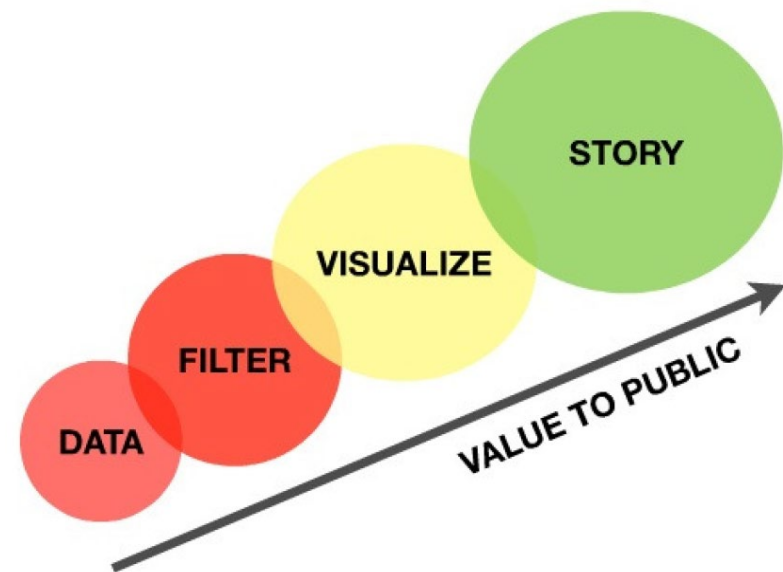
データを収集, 整形, 整理, 分析, 可視化, 公表することで, 新しいニュースを発見したり, ニュースを分かりやすく伝えようとする報道

データ処理による記事化

- データ: データを収集して構造化
- フィルタ: データから知識を抽出
- 可視化: グラフィックなどで表現
- ストーリー: 分析結果から記事のストーリーを作る

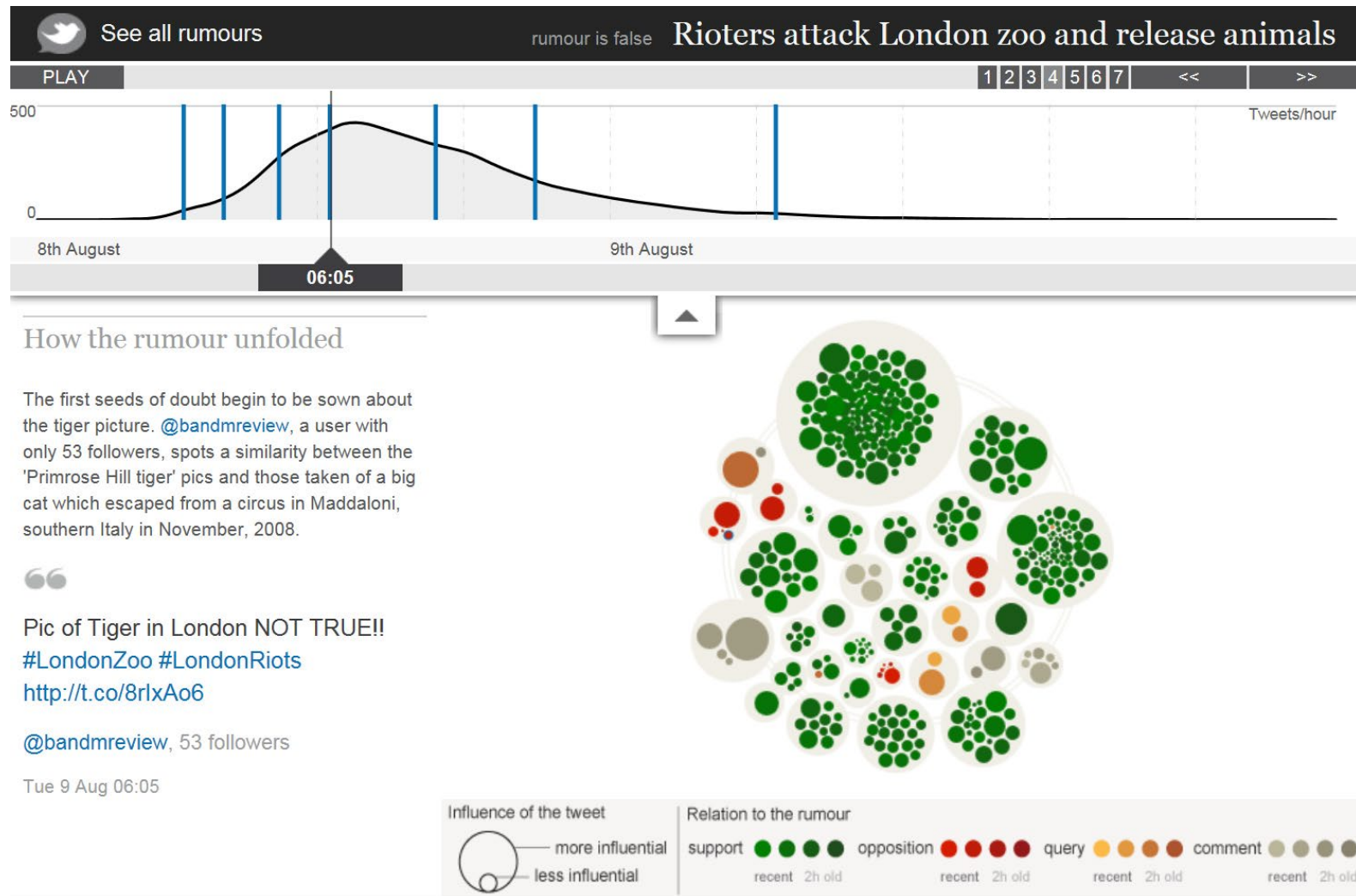
株価のデータを載せただけの記事はデータジャーナリズムとは呼ばない

DATA-DRIVEN JOURNALISM = A PROCESS



データジャーナリズム (Lorenz, 2010)

Reading the Riots (Guardian)



<https://www.theguardian.com/uk/interactive/2011/dec/07/london-riots-twitter>

Full-text visualization of the Iraqi War Logs (AP)

分析の狙い

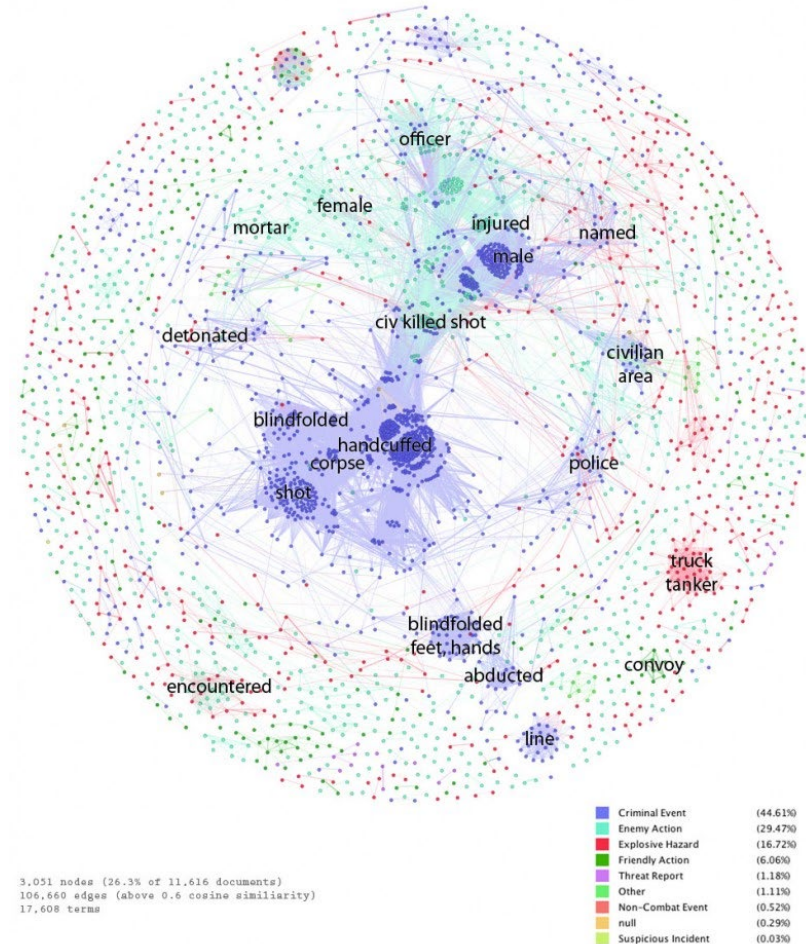
- WikiLeaksが公開したイラク戦争に関する機密文書をネットワーク可視化

分析方法

- 各文書の単語をTF*IDFで重み付け・ベクトル化
- 各文書を重みの高い3つの単語で表現
- 文書をノード、文書間のコサイン類似度でエッジを作成
- ノードの色は事件タイプ
- Gephiでネットワーク可視化



WikiLeaks Iraq SIGACTS (redacted) - Dec 2006



<http://jonathanstray.com/a-full-text-visualization-of-the-iraq-war-logs>

Message Machine (ProPublica)

7

People in our sample got this email

EMAIL 3

vs. EMAIL 1

++++ ----

About This Group

Gender

F

M

Age

17

74

Previous Donor?

YES

NO

Requested Donation

\$3

\$300

Subject Line

Emails

Janet just found out

6

You must be at least a little curious

1

The content of this message is similar to Email 1, but the phrasing is different. Some recipients were asked for \$3, others \$25, and one was asked for \$190.

EMAIL 1

EMAIL 2

EMAIL 3

EMAIL 4

EMAIL 5

EMAIL 6

~~"You're going to have dinner with the President."~~

~~There are only a handful of people who will~~ You must be at least a little curious.

probably don't know Janet from Accokeek, Maryland, Maryland.

But you two other supporters for dinner could be having with four supporters.

~~We're dinner with~~ President Obama, Janet, and Obama together sometime soon.

Janet was the first guest to be selected for the upcoming dinner the President's next Dinner with Barack. And we're counting down the hours until we pick the next winner. Will you be the next person who will hear that sentence?

~~Pitch~~ draw the next name.

The second seat could be yours -- you'll be automatically entered for the chance to join when you donate \$XX or whatever you can to help today.

Janet told us that when she donated \$XX last week she was just became trying to show her support this campaign today, and for the President.

No one of them.

~~She learned she's ever hear that sentence.~~ thinks they'll be picked until they are.

Take the chance. Chip in \$X-\$XX or more today to be automatically entered:

<https://donate.barackobama.com/The-Second-Guest>

Thanks,

Julianna

Julianna Smoot Deputy Campaign Manager Obama for America

東日本大震災ビッグデータワークショップ Project 311

概要
参加申し込み
提供データ
スケジュール
ディスカッション
FAQ
規約一覧
プロジェクト一覧
報道関係者向け資料
▼ イベント
▶ 報告会
経過報告

概要

2011年3月11日の東日本大震災発生時、ソーシャルメディアとマスメディアを通じて大量の情報が広がりました。正しい情報もあれば、様々な噂やデマも飛び交い、情報がいかに重要かを改めて知る良い機会でもありました。では、こうした情報はいったいどのように伝えられたのでしょうか？ また、本当に伝えなかった情報がなぜ伝えられなかったのでしょうか？ 当時をデータで振り返った時に、本当に必要なサービスは何だったのか、次の災害に備えるために、今我々はどんな準備をすればよいのでしょうか？

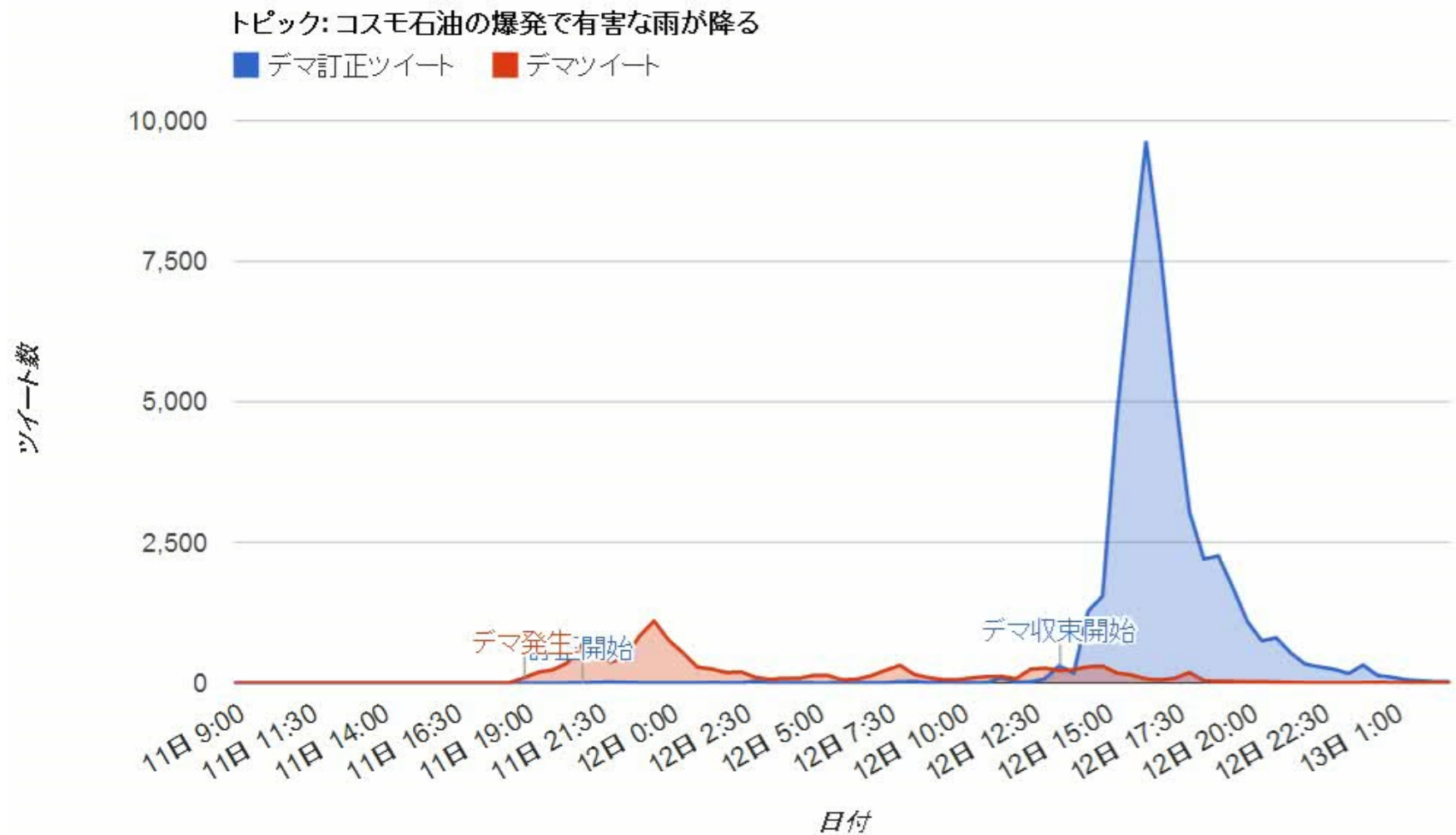
「東日本大震災ビッグデータワークショップ - Project 311 -」では震災発生から1週間の間に実際に発生したデータを参加者に提供いたします。参加者はそのデータを改めて分析することによって、今後起こりうる災害に備えて、どのようなことができるかを議論し、サービスを開発することができます。

提供データ

- 3月11日から1週間の朝日新聞記事 (提供: 株式会社朝日新聞社)
- Google Trends (提供: グーグル株式会社)
- 東日本大震災直後のテレビ放送テキスト要約データ (提供: JCC株式会社)
- 3月11日から1週間のツイート (提供: Twitter Japan 株式会社)
- NHK総合テレビ大震災発生直後から24時間の放送音声書き起こし及び頻出ワードランキング (提供: 日本放送協会)
- Honda インターナビ通行実績マップデータ (提供: 本田技研工業株式会社)
- レスキューナウの鉄道運行情報／緊急情報／被害状況のまとめ情報 (提供: 株式会社レスキューナウ)
- 混雑統計データ (提供: 株式会社ゼンリンデータコム)

<https://sites.google.com/site/prj311/>

東日本大震災時の誤情報の拡散・終息過程の可視化 (鍋島+, 2013)



鍋島 啓太, 渡邊 研斗, 水野 淳太, 岡崎 直観, 乾 健太郎. 2013. 訂正パターンに基づく誤情報の収集と拡散状況の分析. 自然言語処理, 20(3):461-484. (言語処理学会 2013年度論文賞)

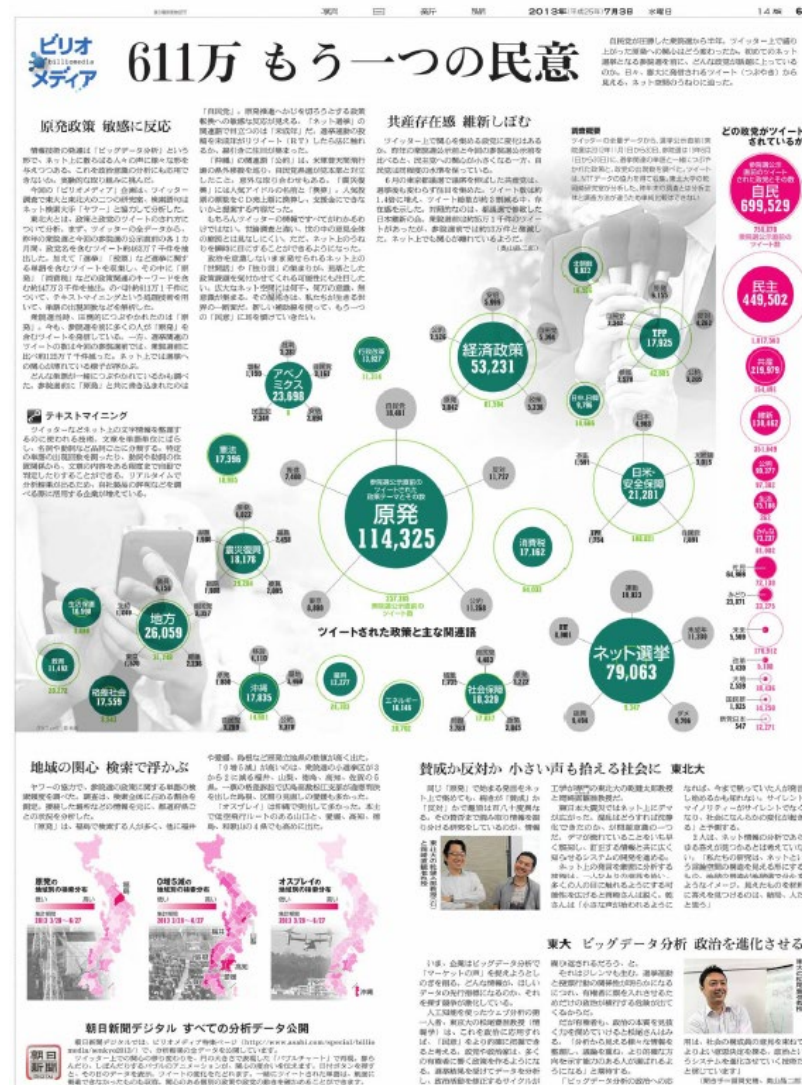
参院選公示日前のツイートの分析

分析の狙い

- 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

分析方法

- 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



2013年7月3日 朝日新聞朝刊（6面）「611万 もう一つの民意」

国会議事録からのキーワード抽出

発言分析（上位20／予測回数との差）

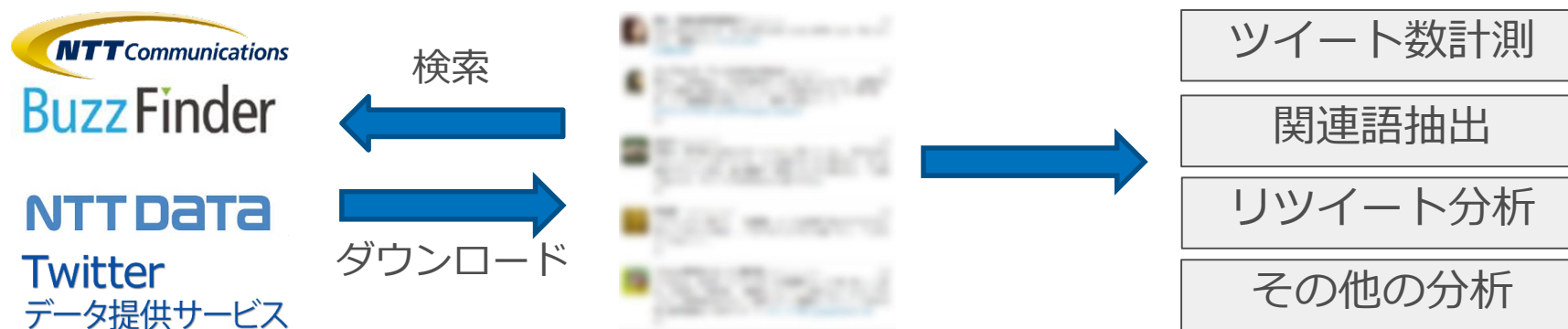
1	日本	(529.8)	11	世界	(179.4)
2	北朝鮮	(356.6)	12	問題	(169.7)
3	話	(345.2)	13	中	(163.2)
4	元気	(327.7)	14	ブラジル	(153.0)
5	辺	(313.0)	15	次	(132.3)
6	中国	(247.1)	16	質問	(126.3)
7	大統領	(207.5)	17	テレビ	(104.8)
8	アメリカ	(200.2)	18	国連	(95.2)
9	人	(195.3)	19	報道	(92.0)
10	ロシア	(183.6)	20	対話	(88.7)

アントニオ猪木氏のキーワード

2019年7月2日 朝日新聞デジタル公開

<https://www.asahi.com/senkyo/senkyo2019/special/politipedia/>

ツイート分析の流れ



ツイート分析の流れ

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

関連語抽出の例（名詞の頻度計測）

報道において大切なポイント

- 分析結果の正確性に責任を持つ
 - 自然言語処理の意味解析の精度は低い
 - 分析方法に対して100%の精度を保証する必要がある
- 分析手法の透明性を確保する
 - 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
- 可視化も重要
 - 一般読者の目を引くようなデータの見せ方を考える

ロボットジャーナリズム（自動化報道）

記事の自動生成（AP通信, Automated Insights社）

AP



Aug. 29, 2017



TORONTO (AP) — Bank of Nova Scotia on Tuesday reported fiscal third-quarter profit of \$2.1 billion Canadian.

The Toronto-based bank said it had earnings of \$1.66 per share Canadian.

The bank posted total revenue of \$6.89 billion in the period.

Bank of Nova Scotia shares have risen 11 percent since the beginning of the year. The stock has climbed 17 percent in the last 12 months.

Elements of this story were generated by Automated Insights (<http://automatedinsights.com/ap/>) using data from Zacks Investment Research. Access a Zacks stock report on BNS at <https://www.zacks.com/ap/BNS>

この記事の要素はZacks Investment Researchから提供されたデータを用い, **Automated Insights**により生成された

<https://www.apnews.com/32a904607c112a15524439fda93ede63/Bank-of-Nova-Scotia-reports-3Q-profit>

深層ニューラルネットワークによる記事生成: RotoWireコーパス

NBA バasketボールの試合のスタッツとそのダイジェストの組が収録されているコーパス（約4,900事例）．ニューラル文章生成モデルの性能はBLEUスコアで10～15くらいで，実応用には程遠い

TEAM	WIN	LOSS	PTS	FG_PCT	RB	AS ...
Heat	11	12	103	49	47	27
Hawks	7	15	95	43	33	20

PLAYER	AS	RB	PT	FG	FGA	CITY ...
Tyler Johnson	5	2	27	8	16	Miami
Dwight Howard	4	17	23	9	11	Atlanta
Paul Millsap	2	9	21	8	12	Atlanta
Goran Dragic	4	2	21	8	17	Miami
Wayne Ellington	2	3	19	7	15	Miami
Dennis Schroder	7	4	17	8	15	Atlanta
Rodney McGruder	5	5	11	3	8	Miami
Thabo Sefolosha	5	5	10	5	11	Atlanta
Kyle Korver	5	3	9	3	9	Atlanta
...						

The Atlanta Hawks defeated the Miami Heat , 103 - 95 , at Philips Arena on Wednesday . Atlanta was in desperate need of a win and they were able to take care of a shorthanded Miami team here . Defense was key for the Hawks , as they held the Heat to 42 percent shooting and forced them to commit 16 turnovers . Atlanta also dominated in the paint , winning the rebounding battle , 47 - 34 , and outscoring them in the paint 58 - 26. The Hawks shot 49 percent from the field and assisted on 27 of their 43 made baskets . This was a near wire - to - wire win for the Hawks , as Miami held just one lead in the first five minutes . Miami (7 - 15) are as beat - up as anyone right now and it 's taking a toll on the heavily used starters . Hassan Whiteside really struggled in this game , as he amassed eight points , 12 rebounds and one blocks on 4 - of - 12 shooting ...

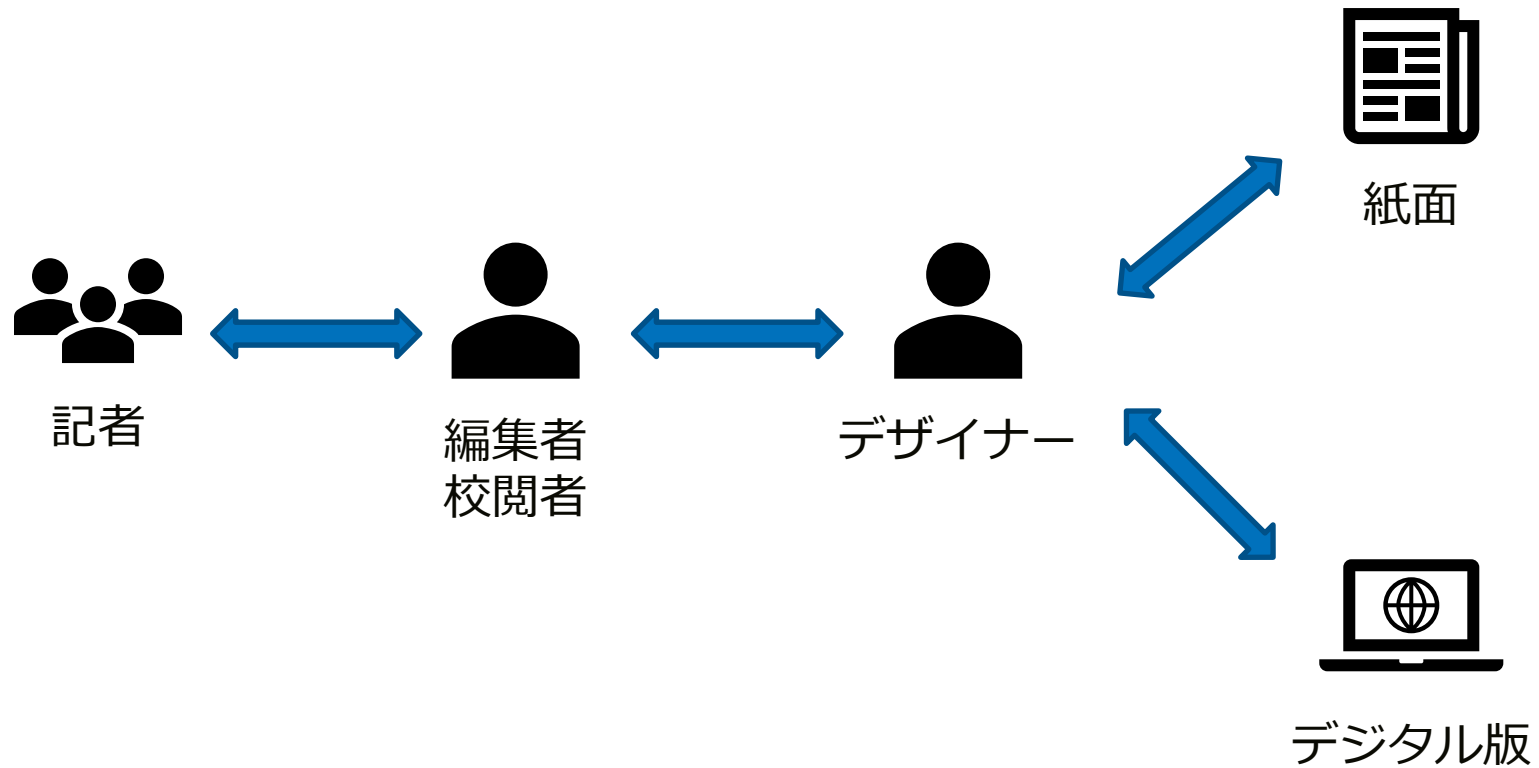
RotoWireデータセットのデータと記事の例 (Wiseman+ 2017)

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2017. Challenges in data-to-document generation. In Proc. of EMNLP, pp.2243-2253.

言語生成技術が報道の現場で使われるために (Leppänen+ 2017)

- **透明性**：報道機関の内部・外部の人が、報道のプロセスを監視、確認、批判、および干渉できる
- **正確性**：報道内容が客観的な事実に基づき、読者を誤解させない
- **修正可能性とドメイン適応性**：新しいドメインの記事を生成するときに、システムを容易に修正できる
- **流暢さ**：一貫性があって流暢な文章を生成できて、読者の満足度を損なわない
- **データの入手性**：構造化されたデータベースが入手できる
- **ニュースの話題性**：読者が興味を持つような話題のニュースを生成できる

記事の自動校正：新聞記事ができるまで



記事の自動校正：編集前後の記事の活用 (Tamori+ 2017)

- 記事の編集操作の履歴データは残っていない
 - 記事をどのような編集操作で書き換えたのか分からない
- 類似度を用い、編集前後の記事の文の対応関係を自動的に分析

A 裁判員経験者の意見を生かそうと宇都宮地裁は12日、裁判員経験者と裁判官、検察官、弁護士による意見交換会を開いた。**B** 今年1～8月にかけて裁判員として審理に参加した9人が出席し、審理日程や評議についての感想や改善点などを話した。**C** 参加した裁判員経験者は「社会の中で責任を果たす自覚を持った」、「貴重な経験ができた」などという前向きな感想が多かった一方、「弁護士の冒頭陳述がわかりにくかった」などの課題も指摘された。…

a 宇都宮地裁が裁判員経験者と裁判官、検察官、弁護士による意見交換会を開いた。**b** 裁判員経験者からは「社会の中で責任を果たす自覚を持った」、「貴重な経験ができた」などという前向きな声が出た一方、「弁護士の冒頭陳述がわかりにくかった」といった課題も指摘された。**c** 意見交換会は12日に開かれ、今年1～8月にかけて裁判員として審理に参加した9人が出席。**d** 審理日程や評議についての感想や改善点などを話した。…

c と d の 2 文に分割

編集・校閲前の記事

最終版の記事

記事の自動校正：記事の編集操作の推定 (Tamori+ 2017)

- 編集距離と同様のアルゴリズムで編集操作を推定
- 2015年10月から2016年9月までの新聞記事の編集前後の記事から
 - 文アライメントを求めることで2,197,739文対を獲得
 - 1,108,750文対は編集なし, 1,088,989文対は何らかの編集あり

~~裁判員経験者の意見を生かそうと~~宇都宮地裁はが12日、裁判員経験者と裁判官、検察官、弁護士による意見交換会を開いた。
(文 A から a への修正)

~~参加した裁判員経験者からは~~「社会の中で責任を果たす自覚を持った」、「貴重な経験ができた」などという前向きな声が出た~~感想が多かった~~一方、「弁護士の冒頭陳述がわかりにくかった」~~などの~~といった課題も指摘された。
(文 C から b への修正)

意見交換会は12日に開かれ、今年1～8月にかけて裁判員として審理に参加した9人が出席し、審理日程や評議についての感想や改善点などを話した。
(文 B から c,d への修正)

編集操作の推定

深層学習による記事の自動校正 (Hitomi+ 2017)

- 提案モデル

- エンコーダ: 双方向LSTM
- デコーダ: 文生成 (注意機構付きLSTM) と編集操作予測のマルチタスク

- 評価結果

評価対象	GLEU	WER	BLEU
すべての文対	70.68	23.90	74.48
編集された文対	68.63	35.55	65.31
変更されなかった文対	87.44	16.17	82.51

- 多くの個所は変更されずに残るので, BLEUスコアは高めの値となる
- 変更箇所に対する再現率は6.28%と低い

見出し生成デモ

ウェブサービスとして
一般公開中

国内2例目のコロナウイルス感染=武漢の外国人男性-厚労省



記事から自動生成

サンプル記事

記事を入力する

厚生労働省は24日、国内で2例目の新型コロナウイルスによる肺炎の患者が確認されたと発表した。同省などによると、発症したのは旅行で訪れた武漢市在住の40代の外国人男性。東京都内の病院に入院し、快方に向かっているという。男性は14日から発熱があり、渡航前の15、17両日に医療機関を受診。その際は肺炎と診断されなかった。

多言語音声翻訳コンテストのオプションAPIとして提供中

多言語
音声翻訳
コンテスト

TOP
トップ

IDEA
CONTEST
アイデアコンテスト

PoC
CONTEST
試作品コンテスト

NEWS
ニュース

FAQ
よくある質問

INQUIRY
お問い合わせ



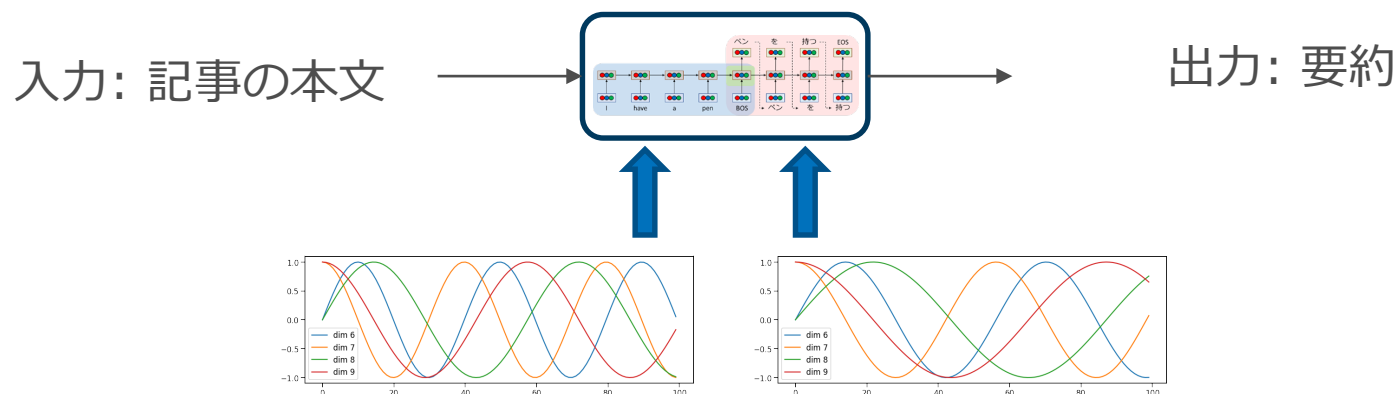
岡崎研究室 見出し生成API

ニュース記事を入力として受け取り、見出しを生成して返します。

<https://headline.okazakilab.org/>

出力長を制御できる見出し生成モデル (Takase+ 2019)

- エンコーダ・デコーダによる自動要約
 - 翻訳元を記事本文, 翻訳先を要約文とした「機械翻訳」
 - 日本語の新聞記事からの見出し生成の研究は少ない
- 機械翻訳と異なるのは出力長の制御
 - 機械翻訳では入力情報をすべて翻訳したら終了
 - エンコーダ・デコーダに出力長を制御する機構が必要



位置エンコーディングによる長さ制御 (左: 長さ40文字, 右: 長さ80文字)

出力長制御の効果 (Takase+ 2019)

提案手法は出力長を正確に制御できる

(出力長を指定して) 要約を生成した際の長さの分散

手法	10文字	13文字	26文字
ベースライン (Transformer)	8.989	5.581	2.510
先行研究 (Kikuchi+ 16)	0.005	0.002	0.002
提案手法	0.000	0.000	0.000

要約の質も向上する

生成された要約と人間の要約の類似度 (ROUGE-L)

手法	10文字	13文字	26文字
ベースライン (Transformer)	31.31	39.12	36.18
先行研究 (Kikuchi+ 16)	35.76	38.47	37.54
提案手法	40.83	43.30	41.21

記事から逸脱した見出しの削減

- 仮説:** 本文から逸脱した見出しが生成されるのはタスクや訓練データの問題？
- 記事が見出しを含意しない記事が学習データに含まれているのではないかな？

検証結果: 学習データにも**問題あり**

- 本文の一部だけを見出し生成器の入力として与えると、情報が不足する
- 入力文章に含まれない情報が見出しに現れる事例が全体の30-40%ある
- 含意関係認識器で学習データを精錬

記事が見出しを含意する割合 (%)

データセット	先頭 1文	先頭 3文	全文
Gigaword	70.3	—	92.8
JAMUL	—	61.4	94.2

- 実験結果:** 学習データから無理な事例を取り除くと**見出しの忠実性が改善**した
- ROUGEの向上は確認できないが、記事と見出しの含意率は向上（自動・人手評価）

学習データ	ROUGE-1	ROUGE-2	ROUGE-L	含意率	人手評価
無理な事例を除去しない	48.08	22.21	40.02	90.29%	89.91%
無理な事例を除去する	45.62	20.55	38.10	92.26%	92.66%

松丸 和樹, 高瀬 翔, 岡崎 直観. 含意関係に基づく見出し生成タスクの見直し. 第240回自然言語処理研究会, 2019年6月.
松丸 和樹, 高瀬 翔, 岡崎 直観. 見出し生成の忠実性の改善. 言語処理学会第26回年次大会 (NLP2020), 2020年3月.

自然言語処理によるニュース記事執筆の課題

- 日々生み出されている貴重なデータを保存していない
 - 記事の編集履歴を保存していない
 - 記事と短文要約との対応関係を保存していない
 -
- 人材不足
 - 新聞社の売上高は減少傾向が続いている
 - データがあっても分析・活用できる人材が社内にはいない
- 正確性に関する高いハードル
 - 新聞記事の生成を完全自動にすることは困難
 - 記者や編集者の作業を軽減できる性能・仕組みが必要

謝辞

- 東京工業大学：高瀬翔氏，松丸和樹氏
- 朝日新聞社：田森秀明氏，人見雄太氏，田口雄哉氏，浦川通氏
- 時事通信社：朝賀英裕氏，川上貴之氏

参考文献

- 岡崎 直観. データジャーナリズムとデータ科学. 電子情報通信学会誌, 99(4):339–346, 2016年.
- 岡崎 直観. ロボットジャーナリズムの現状と課題. 映像情報メディア学会誌, 72(2):70–75, 2018年.

解説

データジャーナリズムとデータ科学

Data Journalism and Data Science

岡崎直観

Topics 話題

ロボットジャーナリズムの現状と課題

関西支部
提案企画

岡崎直観