

自然言語処理の医療応用

次世代人工知能社会実装WG

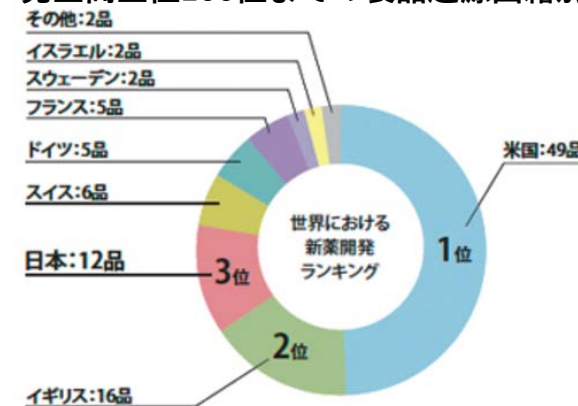


NAIST

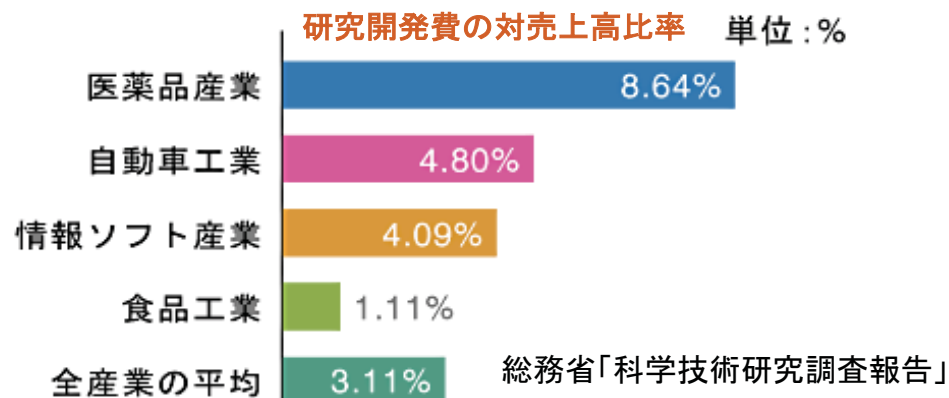
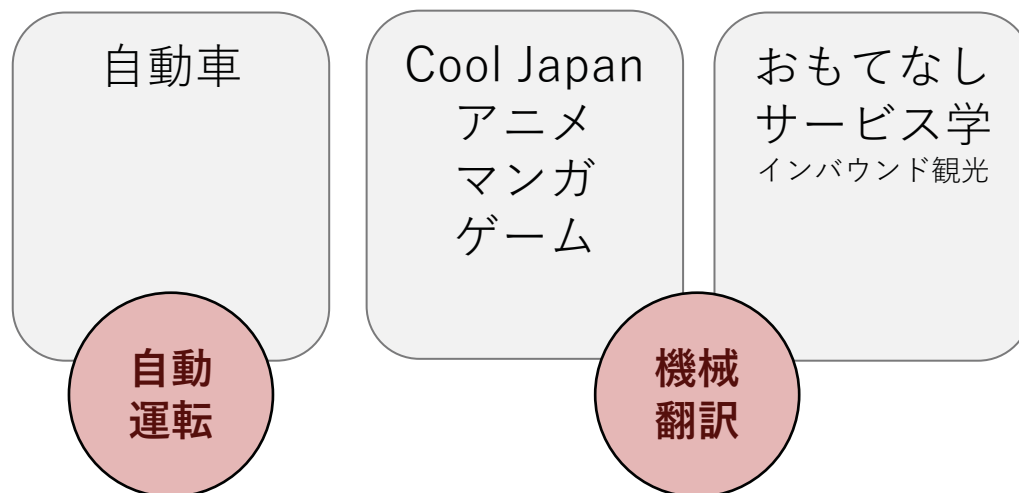
奈良先端科学技術大学院大学
ソーシャル・コンピューティング研究室
荒牧英治

日本を牽引する産業

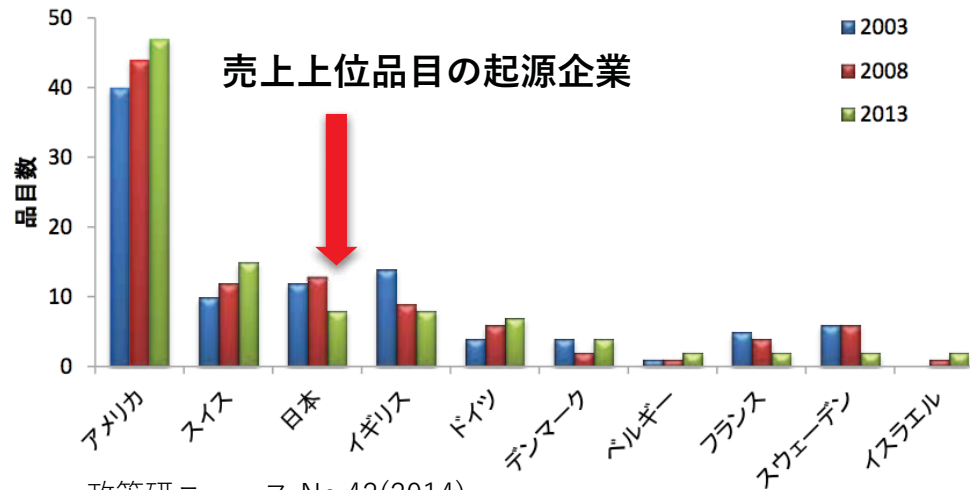
売上高上位100位までの製品起源国籍別



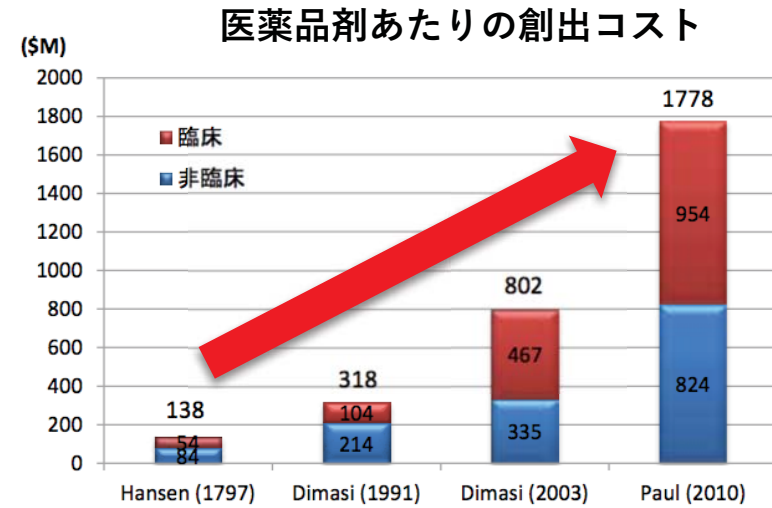
IMS LifecycleおよびPharmaprojectに基づき医薬産業政策研究所が再集計



医療（製薬）を取り巻く状況

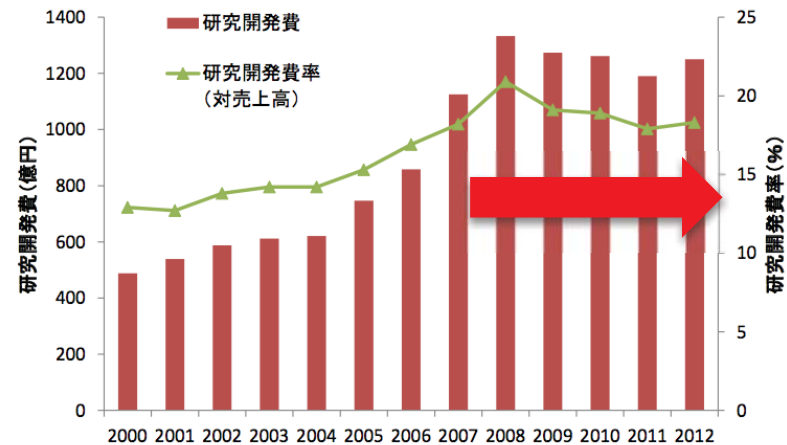


政策研ニュース No.42(2014)

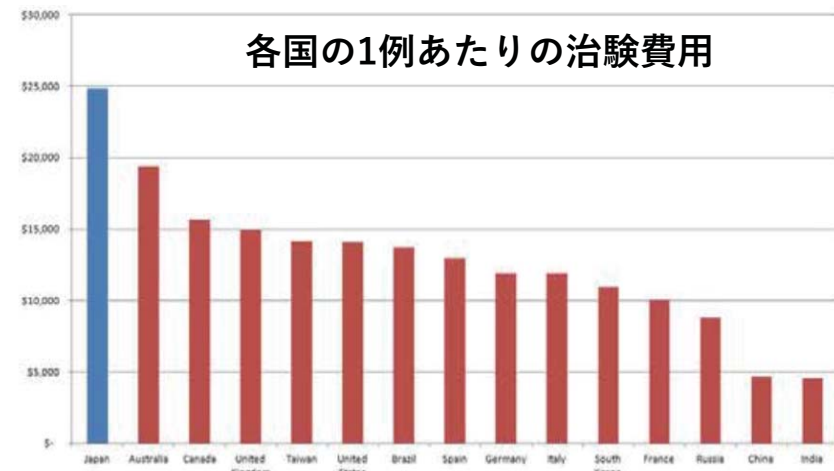


Paul, S.M. et al. (2010) Nat. Rev. Drug Discov. 9, 203-214

日本の製薬企業の研究開発費の推移(売上上位平均)



製薬協DATA BOOK 2014



2014IMSHealth、GrantPlan Data: Average Cost Per Patient in US Dollars

国際共同治験

実施試験数

Evaluate Pharma
(Clinical.Trial.gov)

順位	実施国	実施 試験数	Ph IIとPh IIIの割合		順位	実施国	実施 試験数	Ph IIとPh IIIの割合	
			Ph II (%)	Ph III (%)				Ph II (%)	Ph III (%)
1	米国	1,630	44.7	55.3	16	メキシコ	489	21.5	78.5
2	ドイツ	1,209	41.1	58.9	17	スウェーデン	442	30.5	69.5
3	カナダ	1,003	39.5	60.5	18	ルーマニア	440	26.6	73.4
4	スペイン	906	39.1	60.9	19	アルゼンチン	435	23.7	76.3
5	フランス	893	40.2	59.8	20	台湾	414	25.8	74.2
6	イタリア	835	37.8	62.2	21	ブラジル	401	22.7	77.3
7	イギリス	811	38.1	61.9	22	オーストリア	397	28.5	71.5
8	ポーランド	800	35.1	64.9	23	南アフリカ	371	23.2	76.8
9	ロシア	697	26.4	73.6	24	日本	358	28.2	71.8
10	ベルギー	668	36.2	63.8	24	インド	358	19.0	81.0
11	オーストラリア	599	34.2	65.8	26	デンマーク	324	32.7	67.3
12	ハンガリー	557	33.8	66.2	27	ウクライナ	320	27.2	72.8
13	韓国	553	32.7	67.3	28	スロバキア	305	26.9	73.1
14	オランダ	545	35.6	64.4	29	イスラエル	299	21.4	78.6
15	チェコ	533	30.6	69.4	30	ブルガリア	291	32.6	67.4

製薬

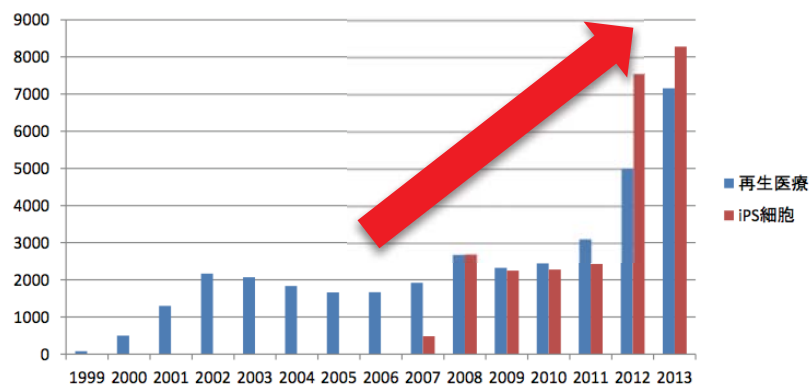
ゲノム医療

先制医療 先制医療が対象とする主な疾患

AI・BD

再生医療

再生医療・細胞に関する記事数



日経テレコン2014

アルツハイマー病
糖尿病
動脈硬化性疾患（冠動脈硬化症、脳動脈疾患など）
骨粗しょう症
閉塞性肺疾患
骨関節疾患
加齢黄斑変性症
発達障害（統合失調症、自閉症、ADHD など）
乳がん

実験医学 Vol. 30, 99 (2012)

AI・BD・IoT

医療情報の展望

～外へ 内へ～



DialBetics

Smartphone-Based Self-Management for Type 2 Diabetes Patients on Insulin Injections



LEGACY

病院から

+SOCIAL

外へ

+BRAIN

内へ

(年 月 日)

<医師の記録>
<<脳神経外科>>
<<19>>
OPが○日AM9:30分からに決定。

患者、家族に手術について説明。
手術説明書・同意書を渡す。

<看護師の記録>
<<脳神経外科>>
<<20>>
OPが○日AM9:30分からに決定。
以下の看護目標、プランを立案
看護目標:
1. 手術後の感染症を予防し、治療を継続できるように維持する。

看護計画: 手術前
1. 口腔ケア感染予防のために術前か
含嗽を行い、口腔内の細菌を除去する。
2. 口呼吸・含嗽の練習: 口呼吸の練
、鼻をつまんでの飲水、含嗽、食事の
習を行う。

看護計画: 手術後
1. 以下の症状の有無と程度の観察
①髄液鼻漏: 水分鼻汁の有無
②口渴、多尿、多飲水
③頭痛、吐きけ、嘔吐

2. 髄液鼻漏の予防:
①手術直後より鼻腔タンポン抜去時ま
仰臥位を保つ。
②術後鼻漏の発生を予防する。
③鼻腔や咽頭の流出感があっても、鼻を
かんだり、すすったりしないように説明
する。

7 / 25

NTCIR E

Upload tex

Choose File

Or input te

神奈川県在

12月12日

嘔吐と下痢

血液検査の

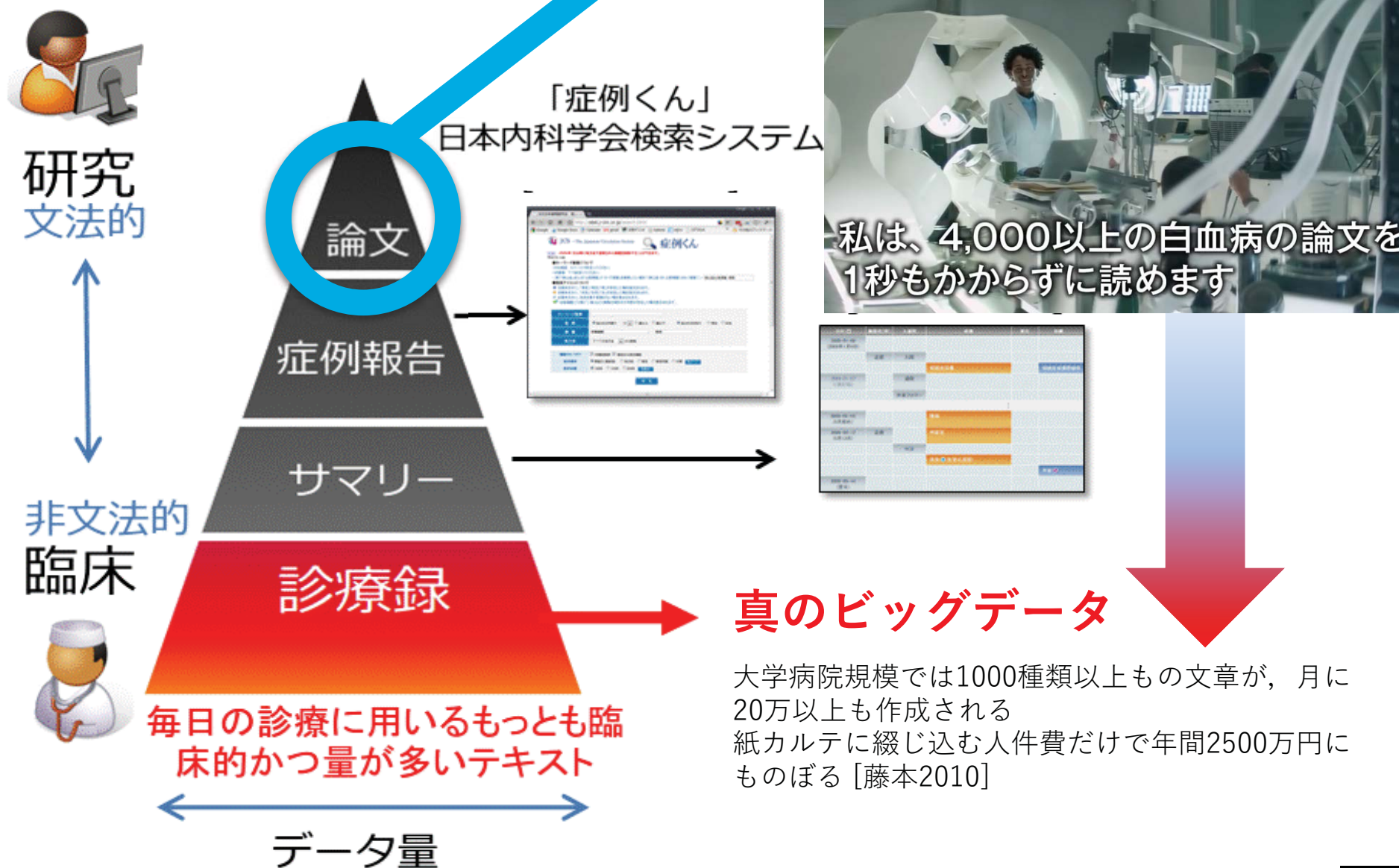
感染性胃腸

Output for

08:04:17



現在の人工知能 (IBM Watson)

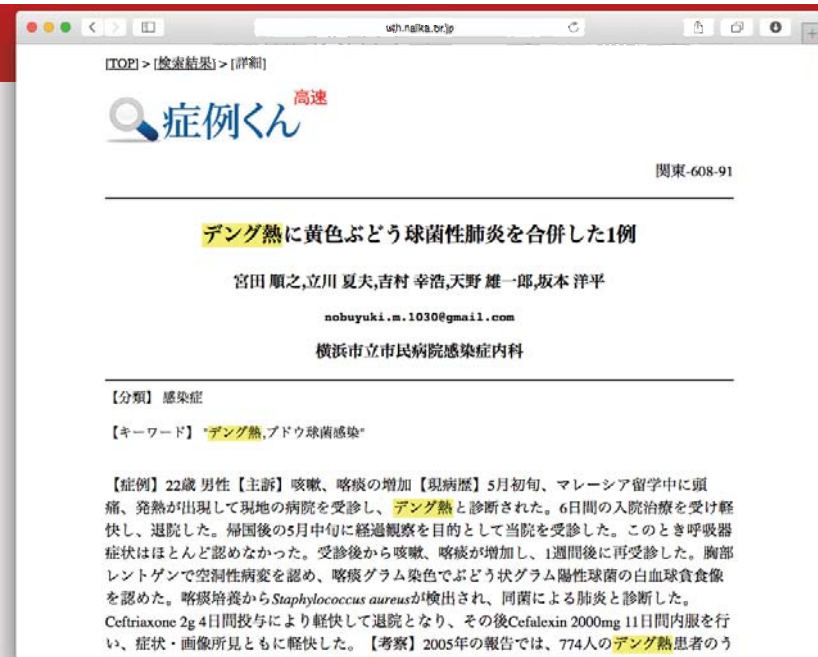


NAISTに集まる大規模データ

- 日本内科学会
 - 45,000症例
- 日本循環器学会
 - 25,000症例
- 東京大学医学部附属院 (大江和彦教授)
 - 300,000症例
- 自治医科大学 (永井良三学長)
 - 100,000-症例前向き収集開始 (H28補正予算)
- 思春期コホート (笠井登人教授)
 - 4,800例 (新学術) *
- 京都大学医学部附属病院 (木下彩栄教授)
 - 60症例 **
- 東京大学生産技術研究所 (熊谷晋一郎准教授)
- 福島県立医科大学
- 東北メディカルメガバンク



* インタビュー映像データ+MRI画像付
** インタビュー音声データ付



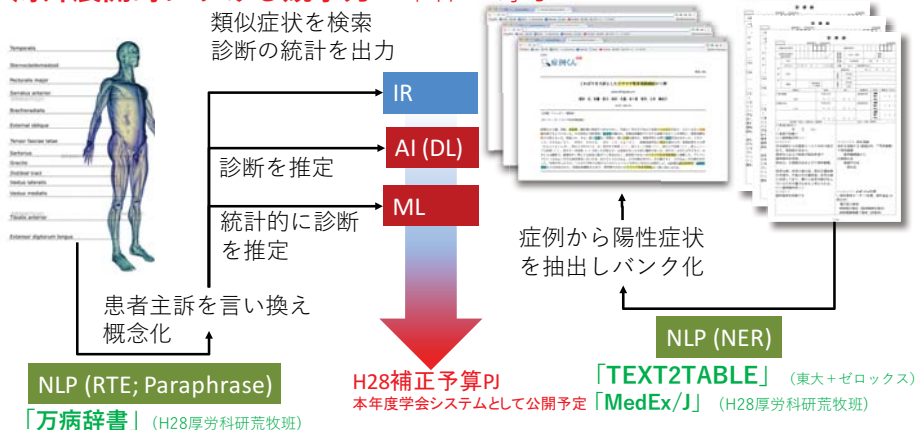
- 症例報告は日本の学会が長年行ってきた我が国独自の活動
- 立ち遅れた日本の医療AIを挽回する格好のデータベース

症例検索から診断支援のためのデータへ

症状から症例を検索

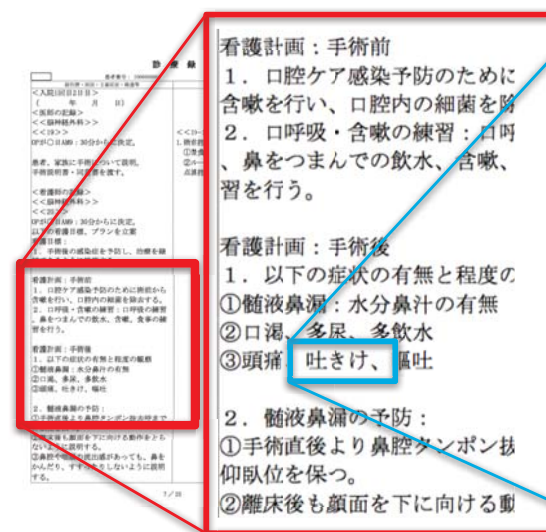
症状から診断名を推測する

- 稀な症例であっても重要な疾患集であり見落としの防止に貢献
- 海外展開時における競争力の確保に寄与



医療テキストから医療
表現 (病名, 症状)
の抽出

医療表現の標準化
(言い換え; 正規化)



ICD=R11
悪心及び嘔吐
悪心増悪
胃のむかつき
悪心
食後悪心
嘔気
吐き気 (け)
悪阻

現状の言語処理の精度

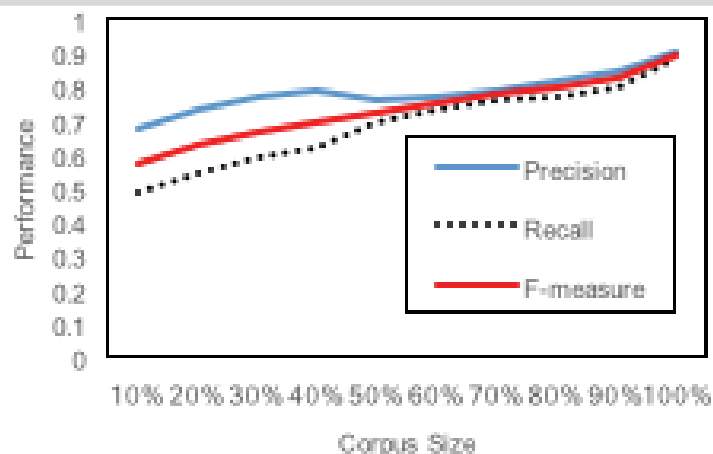
奈良先端科学技術大学院大学は従来から我が国の言語処理ツール開発において中心的存在

自然言語処理

NAISTで開発されている言語処理ツール群

1. 症例報告検索基盤「症例くん」
2. テキスト自動構造化ツール「TEXT2TABLE」
3. 自動病名標準化ツール「MedEx/J」（開発中）

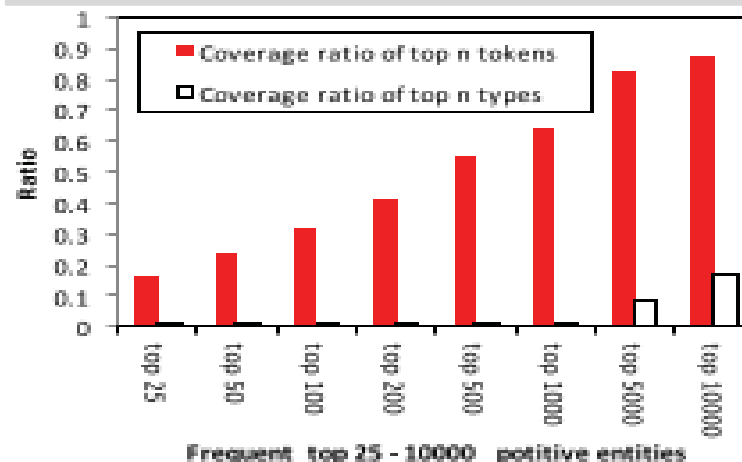
項目1: 症状特定 (F値80% → 95%見込み)



症状抽出精度と教師データ量 (1,800症例を用いた実験)

病名特定においては研究途中 (3万症例) で精度 (F値) 93%を達成, 研究終了時 (7万症例) では精度**95%が達成される見込み**

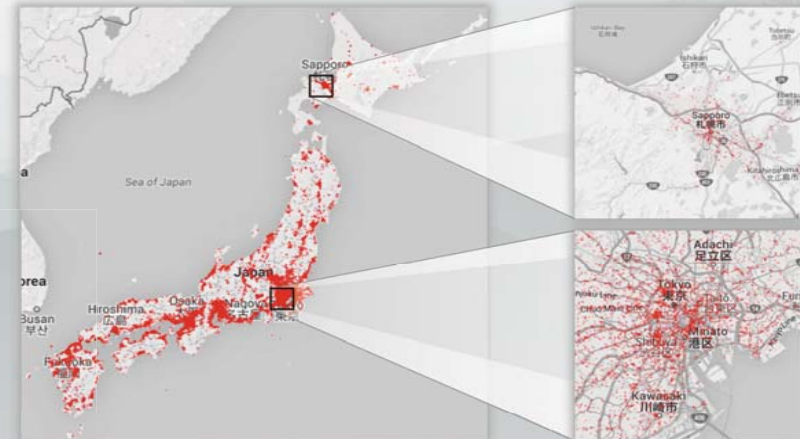
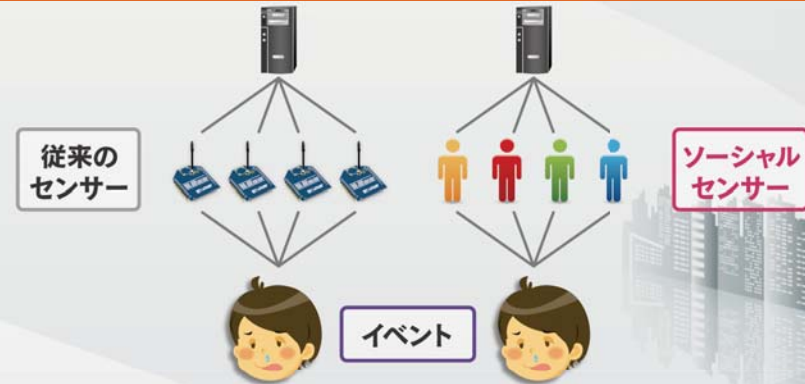
項目2: 症状標準化 (被覆率90%)



症例報告における頻出症状が全症状をカバーする割合

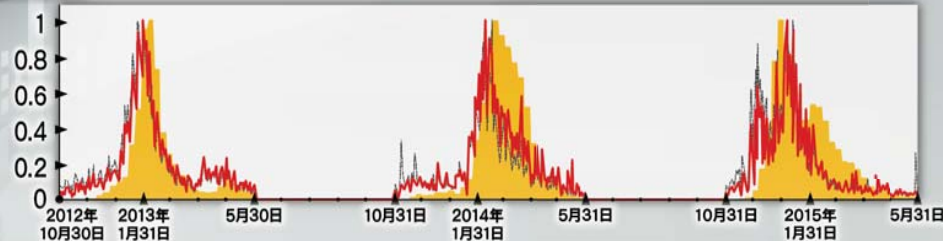
頻出する上位1万症状表現について知識ベース (ICD10やLiLak) へのマッピング作業を行う。試算では**90%の被覆率**で症状を標準化可能となる

汎用機 (core i7 6core 3.4GHz, Memory 32GB ; 70万相当) にて, **120秒で3,000症例報告**を処理可能. 本研究終了後には実用化に耐えうる見込み

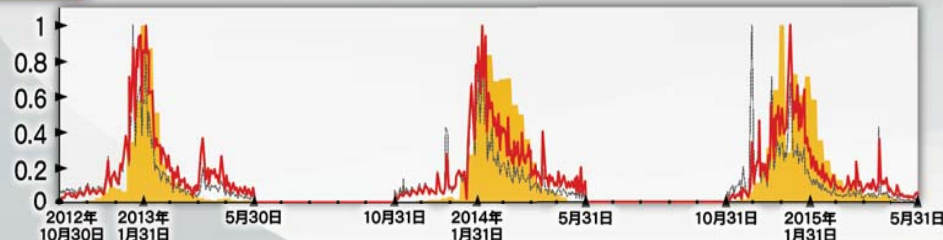


出展はこちらに入ります

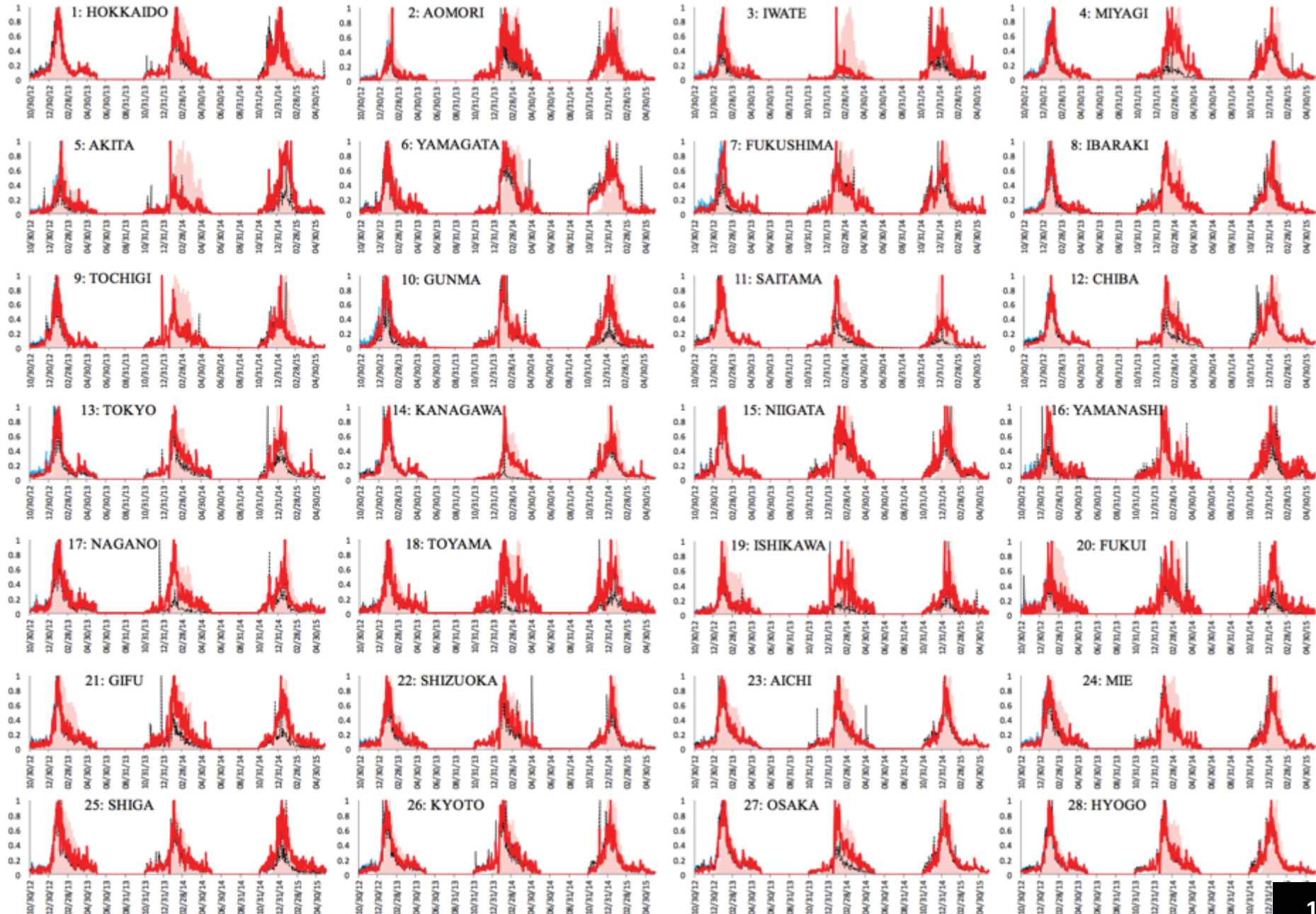
北海道



東京

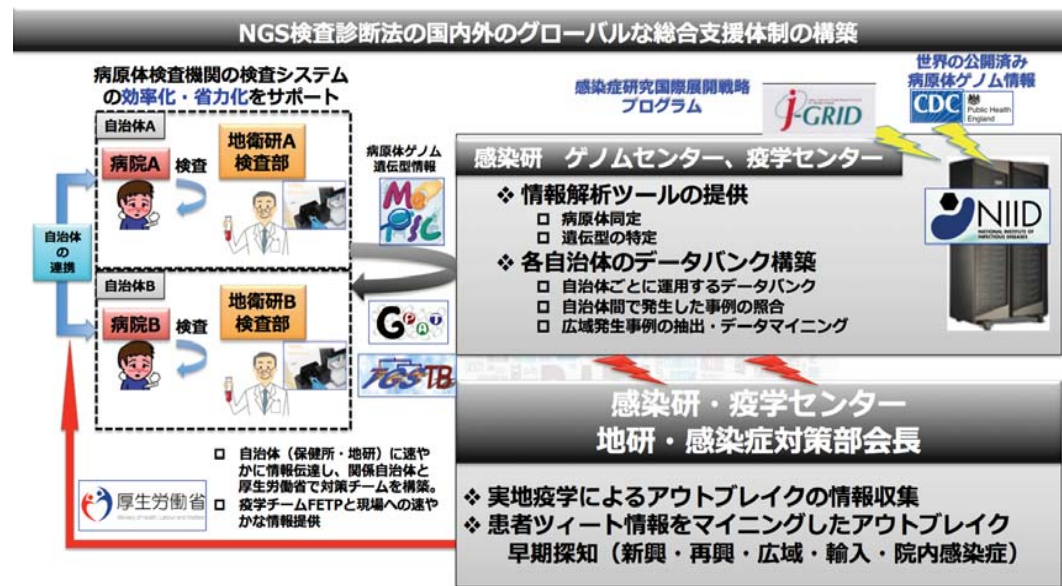
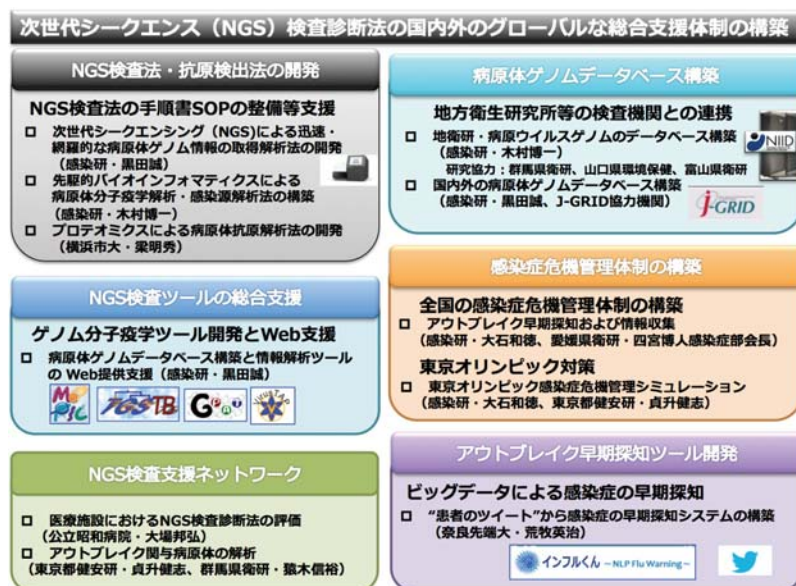


赤字=つぶやきから推定された患者数 ピンク=感染症情報センター報告の患者数





迅速・網羅的病原体ゲノム解析法の開発及び感染症危機管理体制の構築に資する研究



感染症危機管理体制の構築に資する網羅的な診断系を確立
Twitterの“つぶやき”から自然言語処理にて感染症発生を推定するツールを開発し、そのアラートが実際の現場（病院・地研・感染研）で正しい事象を捕捉しているのか否かを検証する。
海外来訪者の増大と東京オリンピック対策を兼ねて、感染研・東京都健康安全研究所との連携を蜜にし、実地疫学とNGS検査法を軸にネットワーク構築する。

黒田誠（国立感染症研究所）
 木村博一（国立感染症研究所）
 梁明秀（横浜市立大学大学院）
 大石和徳（国立感染症研究所）
 四宮博人（愛媛県環境衛生研究所）
 貞升健志（東京都健康安全研究所）
 大場邦弘（公立昭和病院）
 荒牧英治（奈良先端技術大学院大学）

群馬県における 感染性胃腸炎

感染性胃腸炎の患者は、年間を通して報告されています。特に例年、11月頃から患者数が増加し始め、その後急増して12月頃にピークとなる傾向があります。また、冬期にはノロウイルスによる感染症や食中毒が多く発生しています



疾病名	第3週の状況		第4週の状況	
	報告数	定点当たり	報告数	定点当たり
インフルエンザ	2,751	28.96	3,101	32.64
RSウイルス感染症	22	0.38	7	0.12
咽頭結膜熱	14	0.24	27	0.47
A群溶血性レンサ球菌咽頭炎	129	2.22	144	2.48
感染性胃腸炎	303	5.22	312	5.38
水痘	18	0.31	15	0.26
手足口病	26	0.45	23	0.40
伝染性紅斑	10	0.17	6	0.10
突発性発しん	23	0.40	27	0.47
百日咳	0	0.00	0	0.00
ヘルパンギーナ	1	0.02	3	0.05
流行性耳下腺炎	23	0.40	17	0.29
急性出血性結膜炎	0	0.00	0	0.00
流行性角結膜炎	11	0.79	9	0.64
細菌性髄膜炎	0	0.00	0	0.00
無菌性髄膜炎	0	0.00	0	0.00
マイコプラズマ肺炎	10	1.25	13	1.63
クラミジア肺炎	0	0.00	0	0.00
感染性胃腸炎（ロタウイルス）	4	0.50	2	0.25
インフルエンザ（入院患者）	33	4.13	42	5.25

つぶやきくん IG POWERED BY NAISTARS

◆「つぶやきくん IG」とは

マイクロプログラム「ツイッター (Twitter)」に投稿されている発言から「感染性胃腸炎 (IG: Infectious Gastroenteritis)」に関連のあるつぶやきを集め、都道府県ごとに推定した患者数を表示するシステムです。

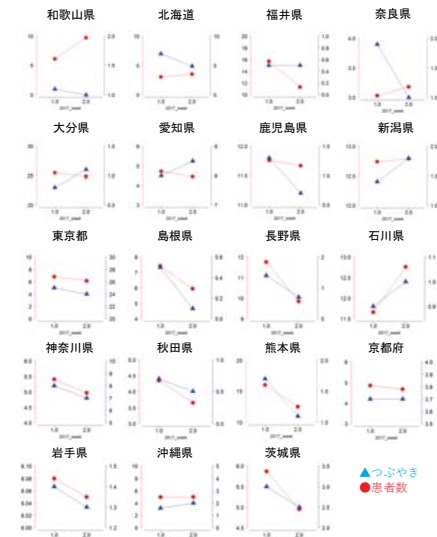


利用者の位置の特定: 「ツイッター (Twitter)」での発言時に GPS により付与された位置情報や、プロフィールで公開されている場所名などから、利用者の位置を都道府県単位で特定しています。

患者数の推定: 奈良先端科学技術大学院大学 ソーシャル・コンピューティング研究室が開発している自然言語処理技術を用い、実際に「感染性胃腸炎」と関係のある症状で苦しんでいる人のつぶやき (以降「関連性のあるつぶやき」とも記述しています) のみを抽出し、集計しています。

奈良先端科学技術大学院大学 ソーシャル・コンピューティング研究室 (<http://socialcom.jp>)
E-mail: socialcomputing-office@is.naist.jp

つぶやきくん IG Infectious Gastroenteritis



ソーシャルセンサーの問題

- インフラの不安定性の問題:** ラインやツイッターなどのソーシャルメディアは、電話のような公共性の高い通信インフラでなく、私企業が利益のために営んでいるものである。このため、いつ仕様の変更があるかは分からず、また、場合によっては停止することさえ考えられる。人命さえ関わることがあり得る重要性の高いサーベイランスをこのような不安定なインフラ上で実現してよいのかという問題がある。
- ユーザ・バイアスの問題:** ソーシャルメディアを使うのは都会に住む人々が多い。東京や大阪などの大都市に相対して、北海道など通常の都市は発言数が少ない。当然、ここから得られる情報も都会中心の偏ったサンプルとなる。このような地理的バイアスの他にも、ユーザの年齢層、性別、職業、教育レベルといった様々なバイアスが存在し、集計結果を湾曲させる恐れがある。
- 正確性の問題:** 発言者によっては虚偽の情報を流す場合もある、または、自分では正しいと思っていても、後から誤りと判る場合もある。仮に、正しい情報であっても、自然言語処理技術が、そこから、正しく情報を抽出できるかどうかという問題もある。
- 責任の不明確さの問題:** 上記の諸問題が行き着くところとして、もし、誤りがあった場合、だれがどのように責任をとるのか、という問題がある。個人情報問題とも関連する。

言語計測を用いた臨床応用

	Disease	Sample size	NLP measure	Classification Method
Roark (Roark, Mitchell et al. 2007)	Mild Cognitive Impairment	55 (MCI : 18)	Lexical Features, Syntactic Features	N/A
Pakhomov (Pakhomov, Smith et al. 2010)	Frontotemporal Lobar Degeneration	38	Acoustic Feature Lexical Features	ANOVA
Fraser (Fraser, Rudzicz et al. 2013)	Primary Progressive Aphasia	24	Acoustic Features Lexical Features	Naïve Bayes, Support Vector Machine, Random Forest
Rouhizadeh (Rouhizadeh, Prud'hommeaux et al. 2014)	ASD	60 (ASD : 21)	Lexical Features	Similarity (Word overlap)
Orimaye (Orimaye, Wong et al. 2014)	Alzheimer's Disease	484 (AD or probable AD :242)	Syntactic Features, Lexical Features	Support Vector Machine, Naïve Bayes, Decision Tree, Neural Networks, Bayes Nets
Lamers (Lamers, Truong et al. 2014)	Depression	3	Acoustic Feature	N/A
Yangfeng (Ji, Hong et al. 2014)	Asperger's and Autism	3278	LDA (Topic mining)	N/A
Schwartz (Schwartz, Eichstaedt et al. 2014)	Depression	28749	Lexical features	Regression
Aramaki (Aramaki, et al. 2015)	ASD	16	Lexical Features (JEL)	N/A



語彙量

語彙量推定

構文能力

構文解析

冗長性

文章要約

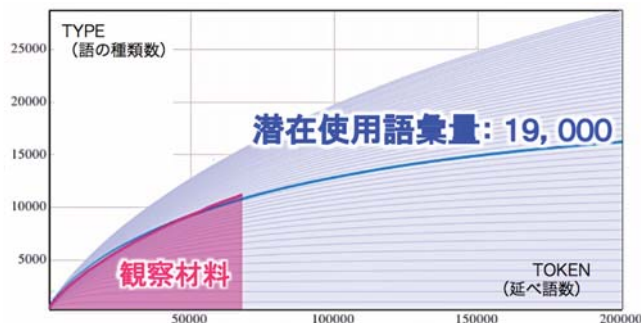
論理性

談話構造解析

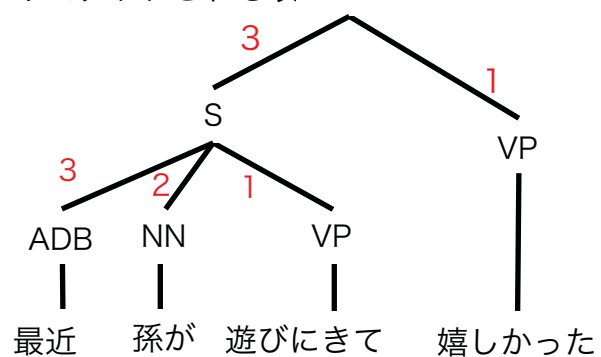
具体性

固有表現抽出

今後の課題



無限時間サンプリングした際の語種類数

[Aramaki2011] $PVS = \lim_{token \rightarrow \infty} (\# \text{ of types})$ 構文解析時に必要となる負担
≡ スタックされる項

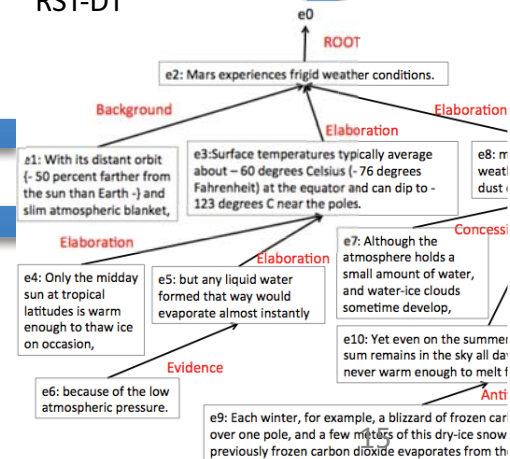
Modified Yngve Score (1960)

それは、あの一、我が家で…孫が、**自分の目指す大学に、無事、入学したこと**であります。えー、我が家に、えー、おいては、私の父の代から、あまり教育は、施されて来なかったんですけども、ようやく、私の…系図の中で…、孫が、**自分の目指す大学に、無事、入学**できたのは、本当に、えー、嬉しいことでございます。

孫が無事に希望大学に入学した。我が家は父の時代からあまり教育されてこなかったが、ようやく孫が大学に入学できて嬉しい。

要約した際の圧縮率

RST-DT



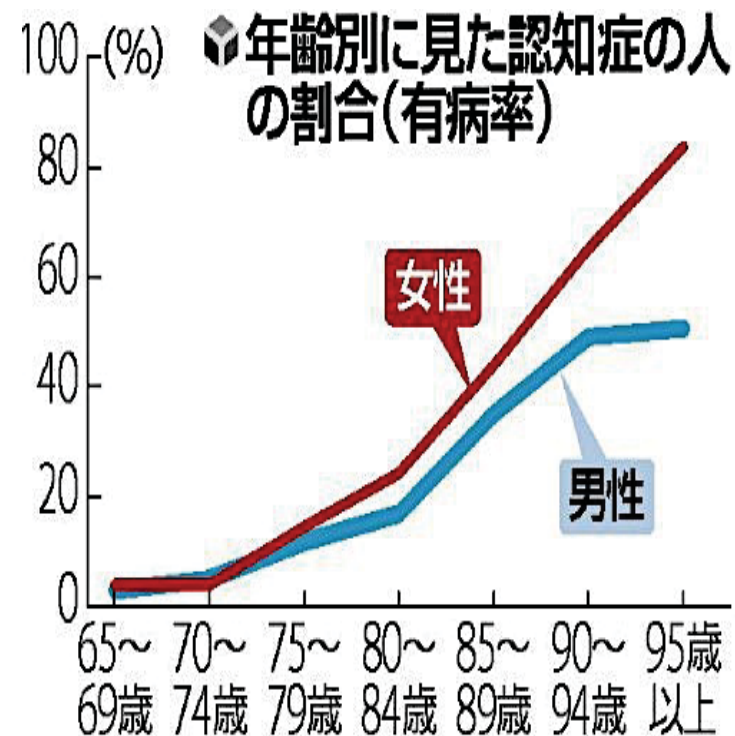
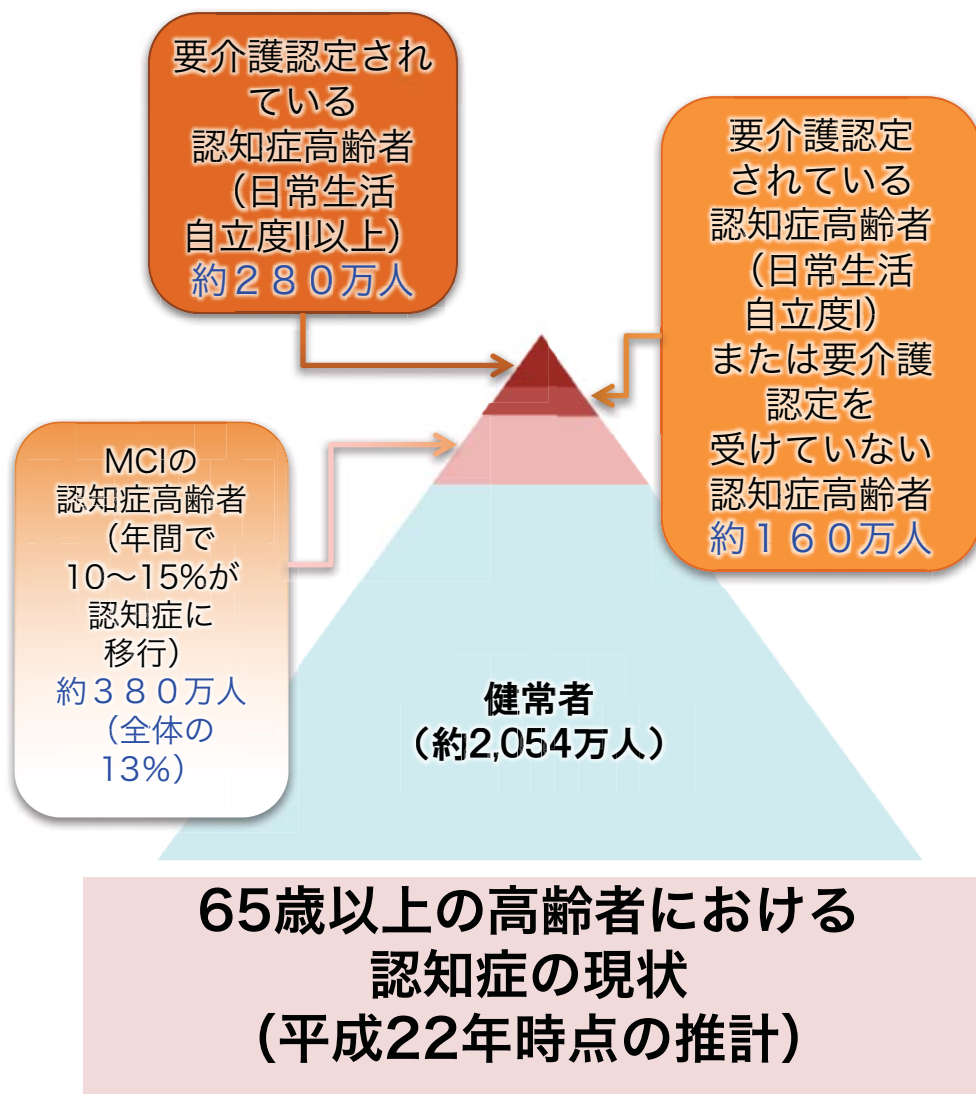
我が家で…孫が、自分の目指す大学に、無事、入学したことあります

孫の名前は？

学校←大学←○○大学

入学（時間格）（場所格）

認知症発症数と医療費



医療費：数千億円

介護費：6兆～7兆円

*既に算出している英国並みと仮定
すると：全体で10兆円を超す見込

参考：厚生労働省

<http://www.gov-online.go.jp/useful/article/201308/1.html> (2014/2/4)

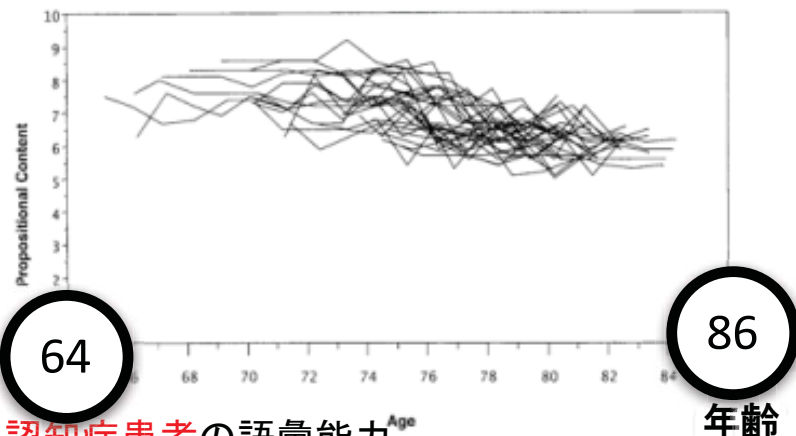
2014年1月31日 読売新聞

+ BRAIN

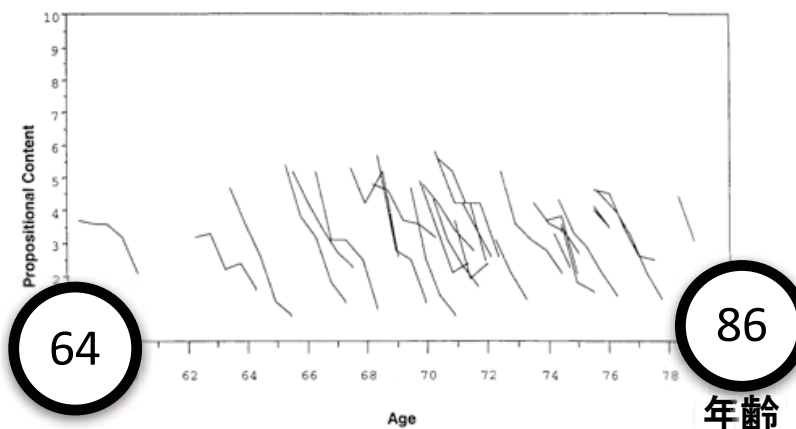
個人発話データの利活用

Snowdonや Kemperは英語言語にて認知症者の語彙力が発症の30年前からも予測できる可能性を示唆

健常者の語彙能力

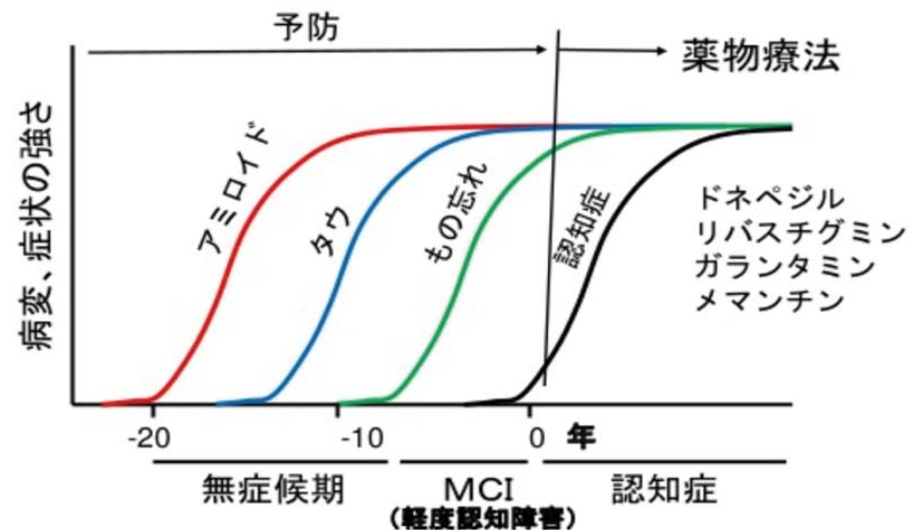


認知症患者の語彙能力



Snowdon, D. A., Kemper, S., Mortimer, J. A., Greiner, L. H., Wekstein, D. R., & Markesbery, W. R. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the Nun Study. *Journal of the American Medical Association*, 275, 528-532.

病理的妥当性



Brain

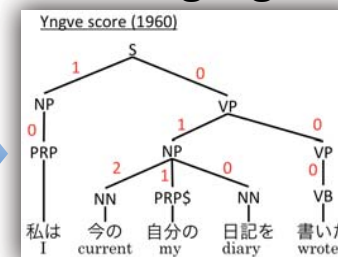


Imaging 測定



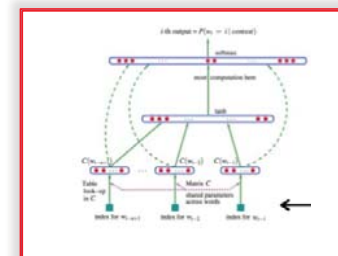
Biomarker

Language



産出

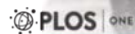
NLP 測定
Natural Language Processing



LangMarker
言語マーカー

言語冗長性
[H27特願]
潜在語彙量
[Aramaki+2015]
:

Small Vocabulary in MCI



RESEARCH ARTICLE

Vocabulary Size in Speech May Be an Early Indicator of Cognitive Impairment

Eiji Aramaki^{1*}, Shuko Shikata¹, Mai Miyabe², Ayae Kinoshita³

¹ Nara Institute of Science and Technology (NAIST), 8916-5 Takayama, Ikoma City, 630-0192, Japan, ² Wakayama University, Sakakidani 930, Wakayama City, 640-8510, Japan, ³ Kyoto University Graduate School of Medicine, 54 Kawahara-cho, Syogoin, Sakyo-ku, Kyoto City, 606-8507, Japan

©Eiji Aramaki, Shuko Shikata, Mai Miyabe, Ayae Kinoshita: Vocabulary Size in Speech may be an Early Indicator of Cognitive Impairment, PLoS ONE, Volume 11 (5), e0155195, 2016. (IF=3.2)

Table 6. Speech

	Healthy	MCI	p
TTR	0.43	0.53	n.s.
PVS	5.1	26.6	n.s.
VL	0.41	0.40	n.s.
DepD	2.03	2.05	n.s.
Yngve	1.60	1.52	n.s.

TTR, Type-Token Ratio; PVS, Potential Vocabulary Size; VL, Vocabulary Level; DepD, Dependency Distance.

Underlined values indicate statistically significant results at $p \leq .05$

doi:10.1371/journal.pone.0155195.t006

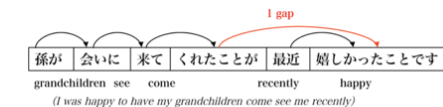


Fig 2. Dependency distance.

doi:10.1371/journal.pone.0155195.g002



Fig 3. Yngve score.

doi:10.1371/journal.pone.0155195.g003

ASD utilizes easy words
Autism Spectrum Disorders

Original Articles

Understanding the Relationship between Social Cognition and Word Difficulty

A Language Based Analysis of Individuals with Autism Spectrum Disorder

E. Aramaki¹, S. Shikata¹, M. Miyabe², Y. Usuda³, K. Asada³, S. Ayaya³, S. Kumagaya³

¹ Kyoto University, Kyoto, Japan;

² Wakayama University, Wakayama, Japan;

³ The University of Tokyo, Tokyo, Japan

Aramaki E, Shikata S, Miyabe M, Usuda Y, Asada K, Ayaya S, Kumagaya S. Understanding the Relationship between Social Cognition and Word Difficulty. A Language Based Analysis of Individuals with Autism Spectrum Disorder. Methods Inf Med. 2015;54(6) (IF=1.8)

Indicator	Indication	Part of Speech	Description
FPU	Peculiarity	Noun	(Average of FPU scores)
JEL	Difficulty	Noun	(Number of JEL high-level words)/(Number of JEL defined words)
TTR	Quantity	Noun	(Number of types)/(Number of tokens)
NER	Particularity	Named Entity	(Number of NEs)/

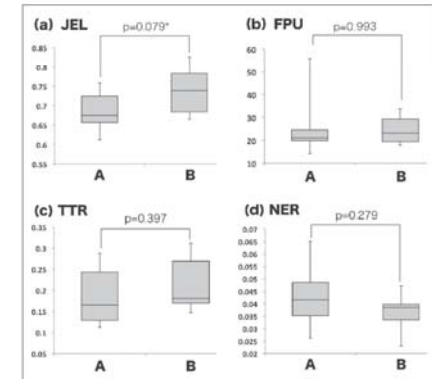
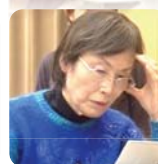


Figure 1. Differences in language scores between Group A (SRS™-2 score: ≥ 76) and Group B (SRS™-2 score: < 76); * $p < .10$

注目される言語+医療

(最近2ヶ月の報道)



H29.2.2 日刊工業新聞
Twitterでの位置推定



H29.1.22 Japan Times
Twitterで花粉症把握



H29.1.27 日経産業新聞
Twitterでインフル把握

H28.12.24 日経新聞
会話が認知症スクリーニング