# 第三部

# 言語処理分野の最前線１
# 産業日本語としての医療カルテ文章

荒牧　英治

京都大学デザイン学ユニット　特定准教授

／独立行政法人科学技術振興機構　　さきがけ研究員

# 産業日本語としての医療カルテ文章

荒牧英治　京都大学

## 自己紹介

- 学部: 京大 総合人間学部
- 修士: 〃　　言語メディア研
- 博士: 東大 情報理工

自然言語処理

---

- 東大 医学部附属病院
  - 助教 (2006-2008)
- 東大 知の構造化センター
  - 講師 (2008-2013) 研究主導者(PI)
- 京大 デザイン学ユニット
  - 准教授 (2013-) 研究主導者(PI)

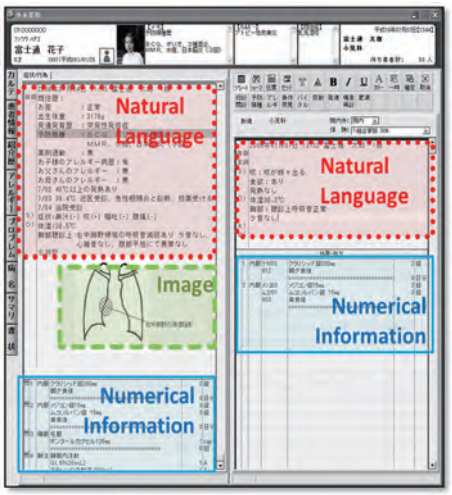
医療分野での言語処理研究に従事

# 医療における
# 自然言語処理とは

# 電子化 ≠ 標準化

電子カルテは Natural Language（自然言語）を含んでいる

病院ごとに，各診療科ごとに，（場合によっては）医師ごとに表現が異なる




様々な表現にバリエーションを扱うために言語処理 （Natural Language Processing; NLP) が必要

Marchesani Syndrome

マルケサーニ症候群
マルケサニー症候群
マルケサニ症候群
マルケザーニ症候群
マルケザニー症候群
マルケザニ症候群

ICD 10 = Q871

**Transliteration (翻字)**

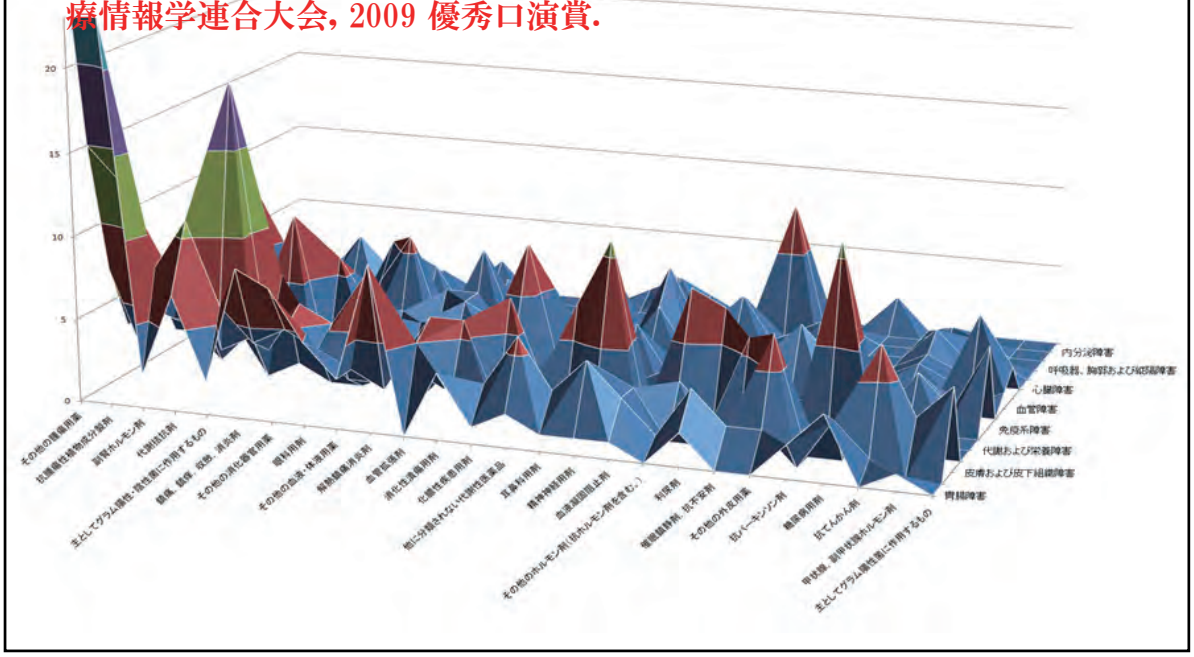
悪心増悪
胃のむかつき
悪心 (sickness)
食後悪心
嘔気 (vomiting)
吐き気
悪阻 (nausea)

ICD 10 = R16
悪心及び嘔吐
(sickness & vomiting)

Paraphrase（言い換え）

# 東大病院での副作用情報抽出

外池昌嗣, 大熊智子, 荒牧英治, 三浦康秀, 増市博, 大江和彦: 自然言語表現の現病歴情報を時系列表形式で表示するシステムとその評価, 第29回医療情報学連合大会, 2009 優秀口演賞.



# 医療言語処理システムの精度

工場に勤めている64歳の男性。2025年月8月2日（来院5日前）頃から腹痛が生じるとともに，食欲不振，嘔気・嘔吐出現した。体幹は温かいが，末梢は湿潤冷汗でショック状態。明らかな運動麻痺はみられず。翌日，意識障害出現し，腎機能障害の増悪を認めて徐々に尿量低下し，8月9日18時10分に心肺停止。8月9日21時44分死亡確認。

**症状表現の抽出**

22システム中1位
（2-stage CRF+辞書）

**時間表現の抽出**

15システム中3位
（2-stage CRF）

86%

| Composition | Precision | Recall | F |
|---|---|---|---|
| BASELINE | 87.87% | 81.43% | 84.53 |
| SYMPDIS | 87.46% | 84.18% | 85.79 |
| MEDDIC | 88.57% | 83.45% | 85.94 |
| FULL | 88.39% | 84.76% | 86.54 |

87%

| Composition | Tag | Precision | R | F |
|---|---|---|---|---|
| BASELINE | a | 86.67% | 69.94% | 77.23 |
| | h | 98.51% | 88.00% | 92.96 |
| | t | 90.42% | 85.07% | 87.66 |
| | overall | 91.26% | 83.74% | 87.34 |

NTCIR-MedNLP1 (2012)

[荒牧，大江2009]

# 症例報告の有効活用に関する研究

- 二つの学会に同義／表記ゆれ検索機能をもった症例検索システムを提供

社団法人 日本内科学会 The Japanese Society of Internal Medicine　社団法人 日本循環器学会

高速 症例くん

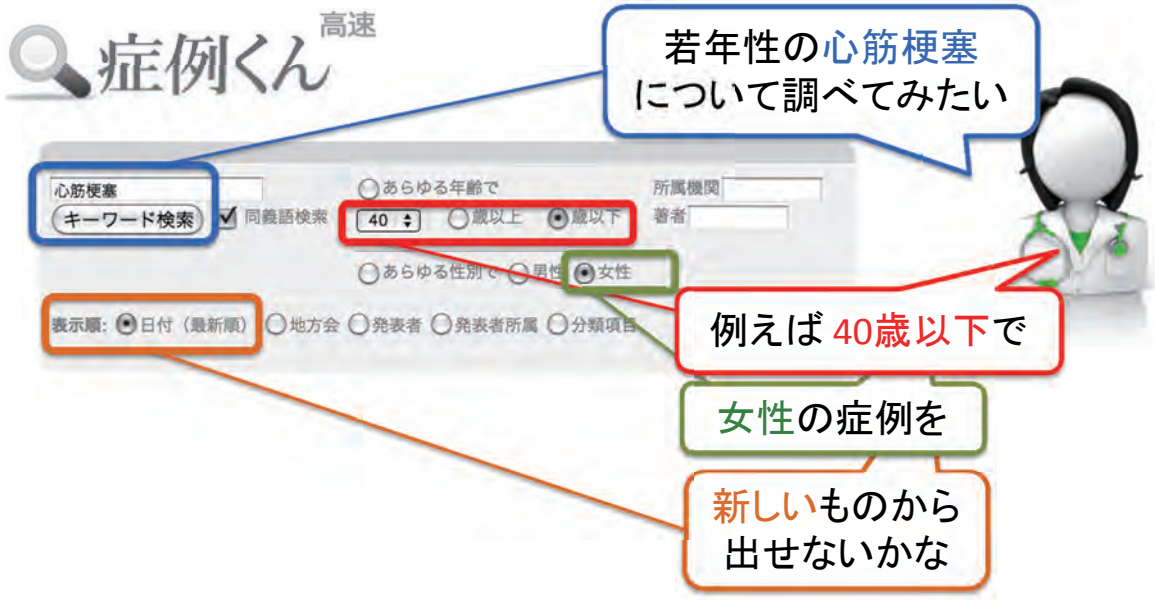
現在データベースには、すでに約25,000件の演題が登録されており、毎年、新たに約3,100件の演題が追加登録


http://member.naika.or.jp/member/content/ninsho1/search.html

# 「症例くん」の検索画面


http://uth.naika.or.jp

# 特徴1) さまざまな要求に耐えうる症例検索




http://uth.naika.or.jp

# 特徴2) 同義語拡張

高速 症例くん

同義語検索：「狭心症」以外に「虚血性胸痛」でも検索します.

バージャー病 ➡ 閉塞性血栓血管炎
BUERGER病
ビュルガー病
血栓閉塞性動脈炎

約20,000の同義語辞書による検索クエリ展開

狭心症 ➡ 虚血性共通

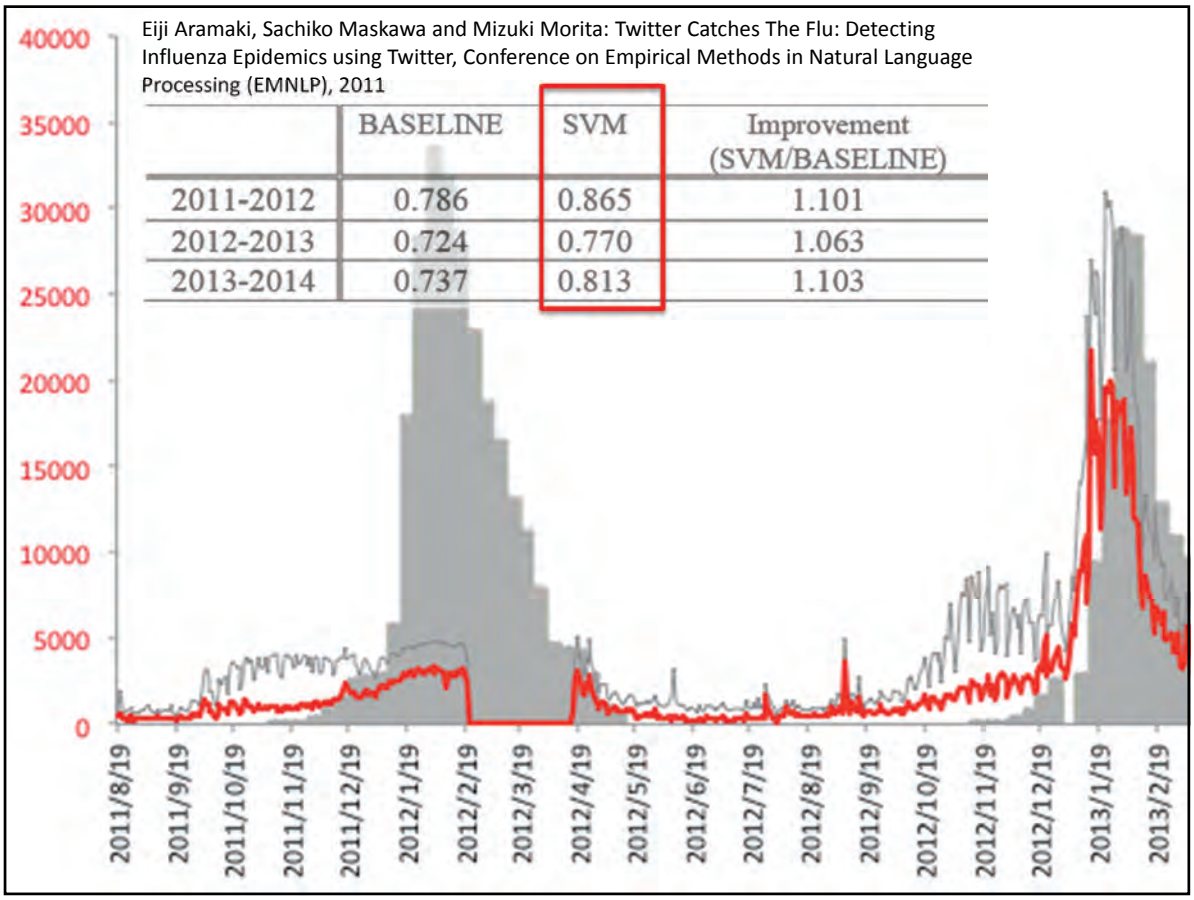
NASH ➡ 非アルコール性脂肪性肝炎
非アルコール性脂肪肝炎

# 日々蓄積される大量の医学知識（上）と臨床知識（下）

# ビッグデータで病気を防ぐ

Eiji Aramaki, Sachiko Maskawa and Mizuki Morita: Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter, Conference on Empirical Methods in Natural Language Processing (EMNLP), 2011

| | BASELINE | SVM | Improvement (SVM/BASELINE) |
|---|---|---|---|
| 2011-2012 | 0.786 | 0.865 | 1.101 |
| 2012-2013 | 0.724 | 0.770 | 1.063 |
| 2013-2014 | 0.737 | 0.813 | 1.103 |



# もう１つの医療記録
# 闘病記・闘病ブログ


悪心増悪
胃のむかつき
悪心 (sickness)
食後悪心
嘔気 (vomiting)
吐き気
悪阻 (nausea)

ICD 10 = R16
悪心及び嘔吐
(sickness & vomiting)

Paraphrase（言い換え）

吐き気
ムカつき
ムカムカ
気持ち悪さ
オエーとなる
リバースする

ICD 10 = R16
悪心及び嘔吐
(sickness & vomiting)

# 認知症者のブログ

発症 → 4年後

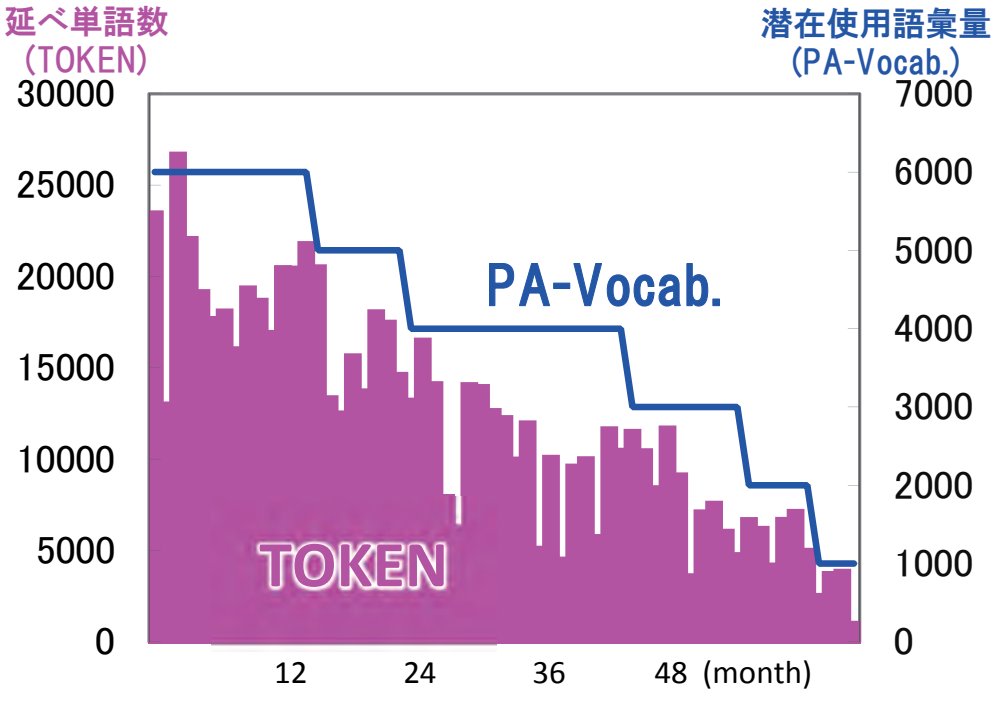
聞き取りをした後長谷川式簡易知能評価スケール（後に本で知る）のテストを受けた。
例貴方は何歳ですか？今日は何年何月何日ですか？私たちが今居るところはどこですか？これから言う言葉を言ってみてください。
桜　　猫　　電車　次に身長・体重・尿・血液検査をして一日目は終わった。
緊張から開放されると大きなため息がでた。
なんで桜や猫なんだ……と帰りの車の中でいらいらした。

2006 日記開始直後

パソコンで文字がかけていたわたしですがなぜか文字の変換が出来ない今日です。
言葉をかくにはそれなりの言葉をさがし読んでいただける言葉にすれなはとはよういではなしのです。
とみにそのかいすが多くなった。
なんで・・・言葉が変換さかと私は私の頭をコツコツと叩いてみましがそんなことでは反応がなかったです。

2010 日記停止1ヶ月前 （原文ママ）

# TOKENとPA-Vocab.の推移



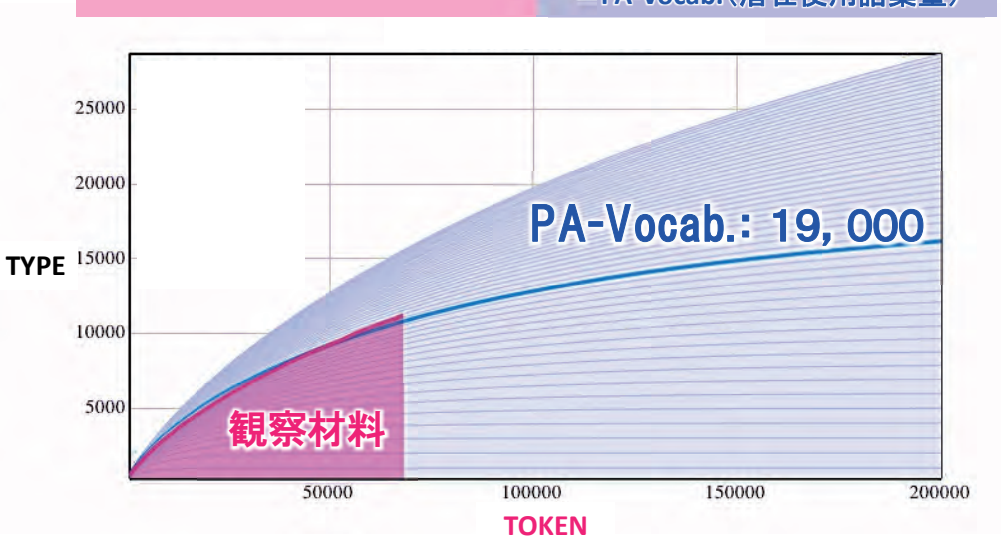

24

# 潜在使用語彙量：PA-Vocab.

Aramaki + , 2012

一定TOKEN（延べ単語数） → 一定TYPE（使用語彙量）

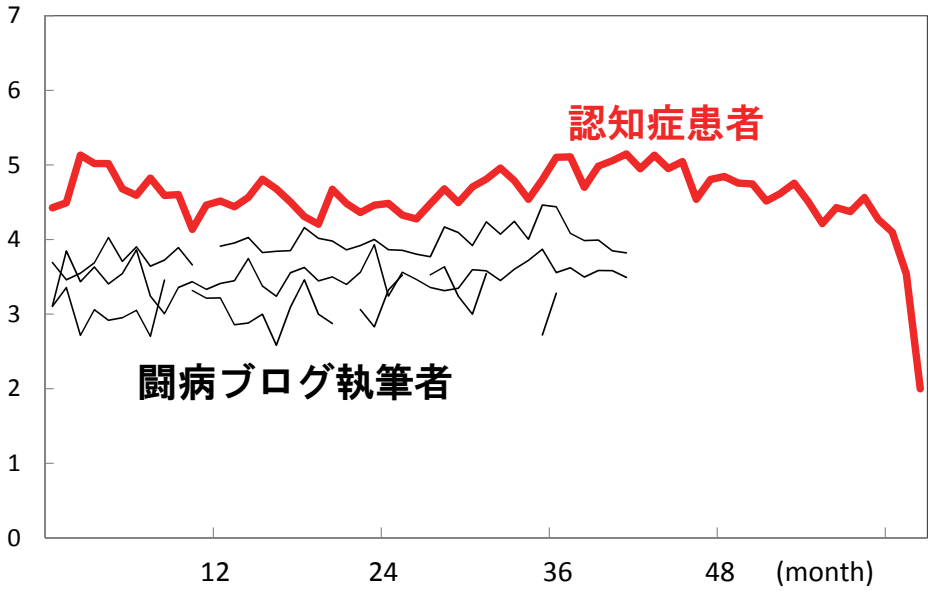
∞ TOKEN → TYPEの収斂 ＝PA-Vocab.（潜在使用語彙量）




25

# 闘病ブログとの対比 ②

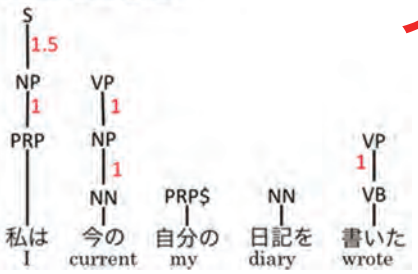
MT-Depth 値の推移



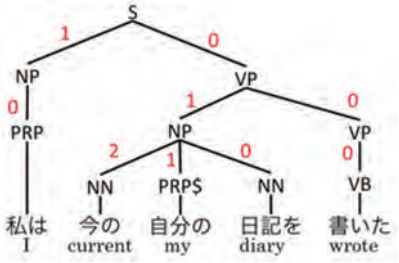

26

# 文長にあまり依存しない構文の複雑さの提案

Frazier score (1985)



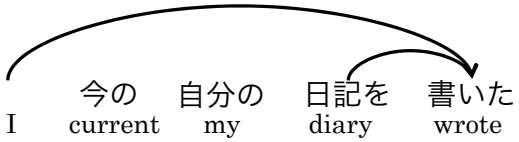

Yngve score (1960)




Dependency distance (1995)


Case Structure Probability (2006)

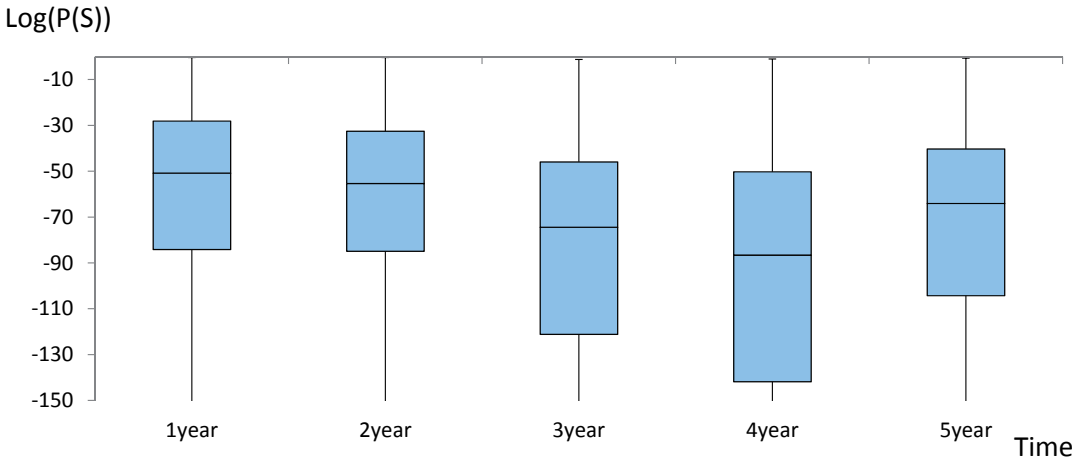
$P(I_{sub}, diary_{obj} | wrote) = e^{-8}$　　P(wrote)=0.002




文確率 ≒ 構文解析確率
= P(書く) Psub(私|書く) Pobj(日記|書く)

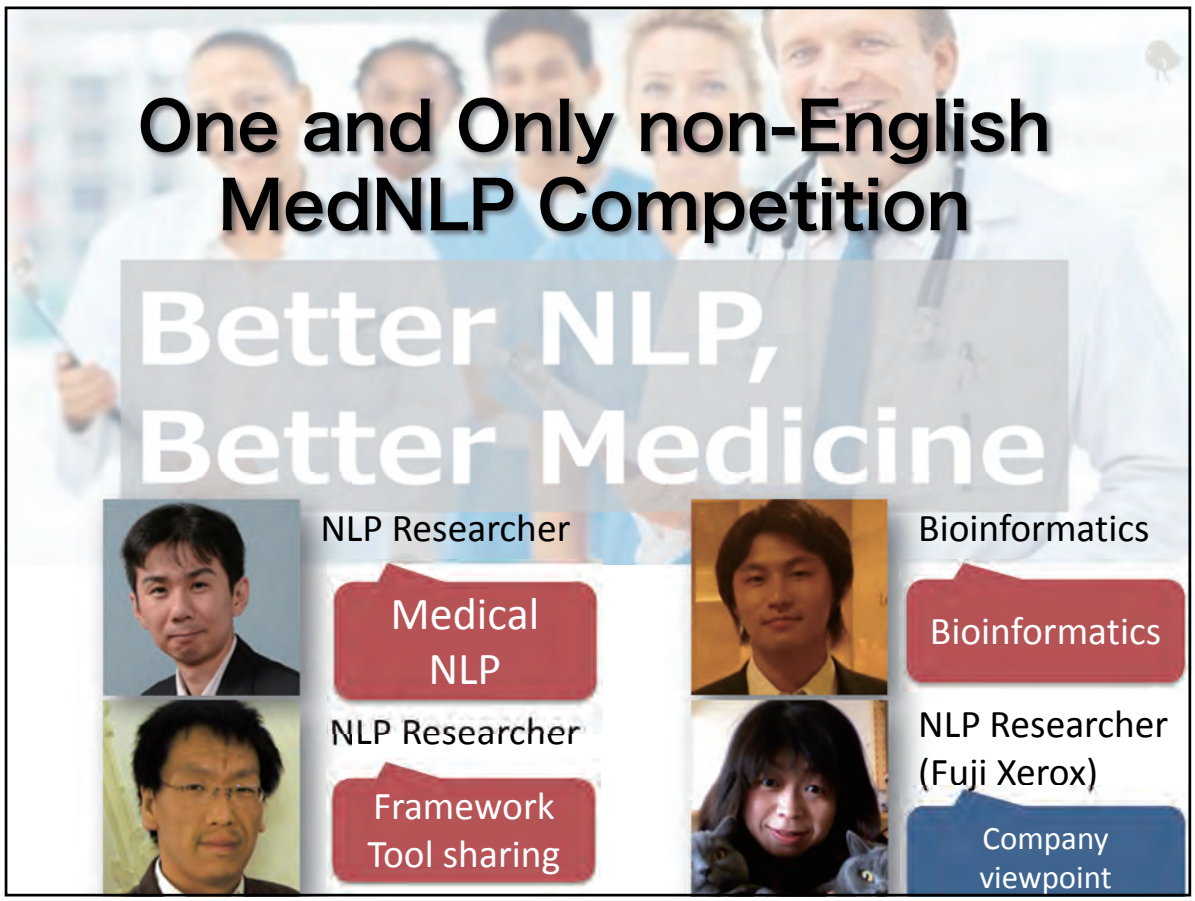
- 構文解析　$ARGMAX_{t \in T} Pr(t)$
  - あらゆる解析結果Tの中から，もっとも解析確率の高い解釈を選ぶ
- 文の生成確率 ≒ 構文の自然さ　$MAX_{t \in T} Pr(t)$
  - その時の確率を文の自然さとみなす

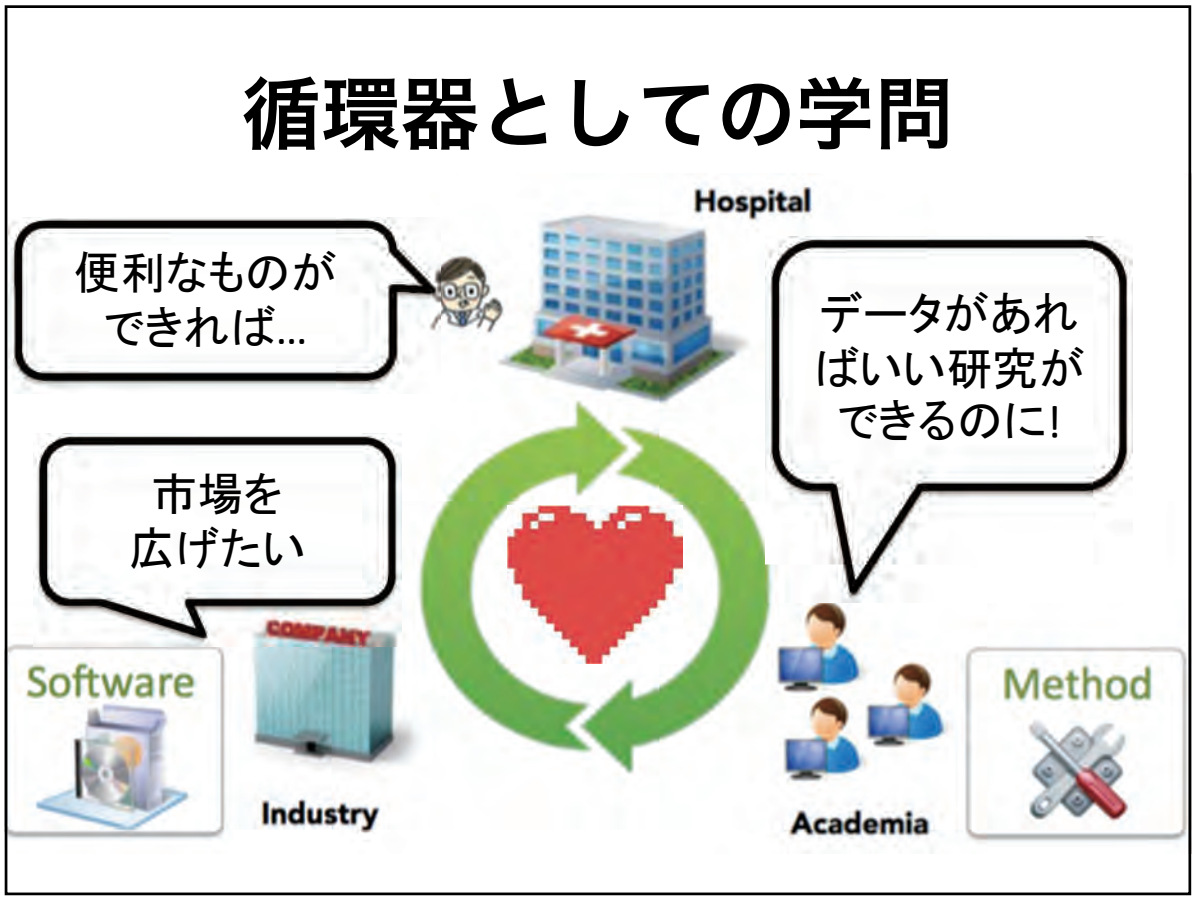
# 構文能力も変化する！




言 秤
～コトバカリ～
言秤とは、話したことばを、自動でリアルタイムに解析して、言語能力を測定することができる装置です。
何が測れるの？
言語能力が測れます。
おしゃべりをするだけで、言語能力のひとつ、語彙能力（単語や話の内容）を測定することができます。
語彙能力にはいろいろな種類があります
言秤～コトバカリ～では、６つの特徴が測定できます。
1. 使う語彙の難しさ
2. 使う語彙の量
3. 使う語彙の具体性
4. 使う語彙の特殊性
5. 使う文体の難しさ
6. 使う文体の丁寧さ
言語能力で何がわかるの？
海外の研究では認知症は言語能力と関係があると言われています。
中でも、語彙能力が高い人が認知症になる割合は低い人にくらべて大幅に低いと言われており、更にバイリンガルの人が認知症になる割合も、そうでない人とくらべて大幅に低いと言われています。
今日本では、言秤～コトバカリ～で、日々おしゃべりしながら言語能力を測る習慣をつけることで、認知症予防に役立てようという研究が始まっています。
ご意見・お問合せ先：京都大学 学際融合教育研究推進センター デザイン学ユニット MedNLP
http://mednlp.jp E-MAIL: mednlp.office@gmail.com
修徳学区認知症を地域で見守るネットワーク主催
第3回
認知症あんしん相談会
参加無料
とき 平成27年3月15日（日） 午後2時～4時（開場午後1時半）
ところ 修徳ふれあい福祉会館4階せんだんホール
京都大学木下先生と京都府立医科大学占部先生による第3回目の認知症相談会です。
どなたでも参加していただける、認知症予防のイベントもたくさん！ どうぞお申込みください。
事前お申込み
先着10名様
認知症の個別相談
京都大学 木下彩栄先生
京都府立医科大学 占部美恵先生
ご自身やご家族のことについて、専用ブースで個別相談に応じます。病気やお薬、対応についてなど、専門的なアドバイスが受けられます。
大好評！セルフチェック出来る！
タブレットによる、もの忘れ診断
当日午後2時～受付順
京都大学大学院生 野田泰葉さん
簡単なタブレット操作で、最新のもの忘れ診断ができます。今から予防を！
京都発お目見え！
言語能力測定「言秤」
当日午後2時～受付順
京都大学 四方朱子さん
京都大学 宮部真衣さん
おしゃべりするだけで、あなたの言語能力をリアルタイムで自動分析。語彙能力をあげることで、認知症の予防にもつながります！
～おはなしとエクササイズ～
転倒予防で認知症も恐くない！
―今日からできる体操とお部屋の工夫―
当日午後2時15分～
京都大学 正木光裕
春になりお出かけの機会も増えますね。転倒しないよう身体もお部屋もちょっと動かしてみましょう！
「個別相談」のお申込み、またイベントについてのお問い合わせは
高齢サポート・修徳
（京都市修徳地域包括支援センター）
☎075-351-2153
京都市下京区富永町110-1 総合福祉施設 修徳 1F
修徳


さいごに


One and Only non-English MedNLP Competition
Better NLP, Better Medicine
NLP Researcher
Medical NLP
NLP Researcher
Framework Tool sharing
Bioinformatics
Bioinformatics
NLP Researcher (Fuji Xerox)
Company viewpoint


循環器としての学問
Hospital
便利なものができれば…
データがあればいい研究ができるのに!
市場を広げたい
Software
COMPANY
Industry
Academia
Method

# 仮想患者のカルテを配布

- (I) 模擬病歴報告 (70症例)
  - 仮想の患者を想定して，医師が記述した病歴報告
- (II) 授業用教材のデータ（50症例）

| | (I) | (II) |
|---|---|---|
| 消化管･腹壁･腹膜疾患 | 4 | 11 |
| 肝･胆･膵疾患 | 2 | 10 |
| 心臓･脈管疾患 | 12 | 10 |
| 内分泌･代謝･栄養疾患 | 5 | 9 |
| 腎･泌尿器疾患 | 4 | 9 |
| 免疫･アレルギー性疾患･膠原病 | 5 | 6 |
| 血液･造血器疾患 | 1 | 7 |
| 感染症 | 6 | 9 |
| 呼吸器･胸壁･縦隔疾患 | 11 | 11 |


# MedNLP 参加者

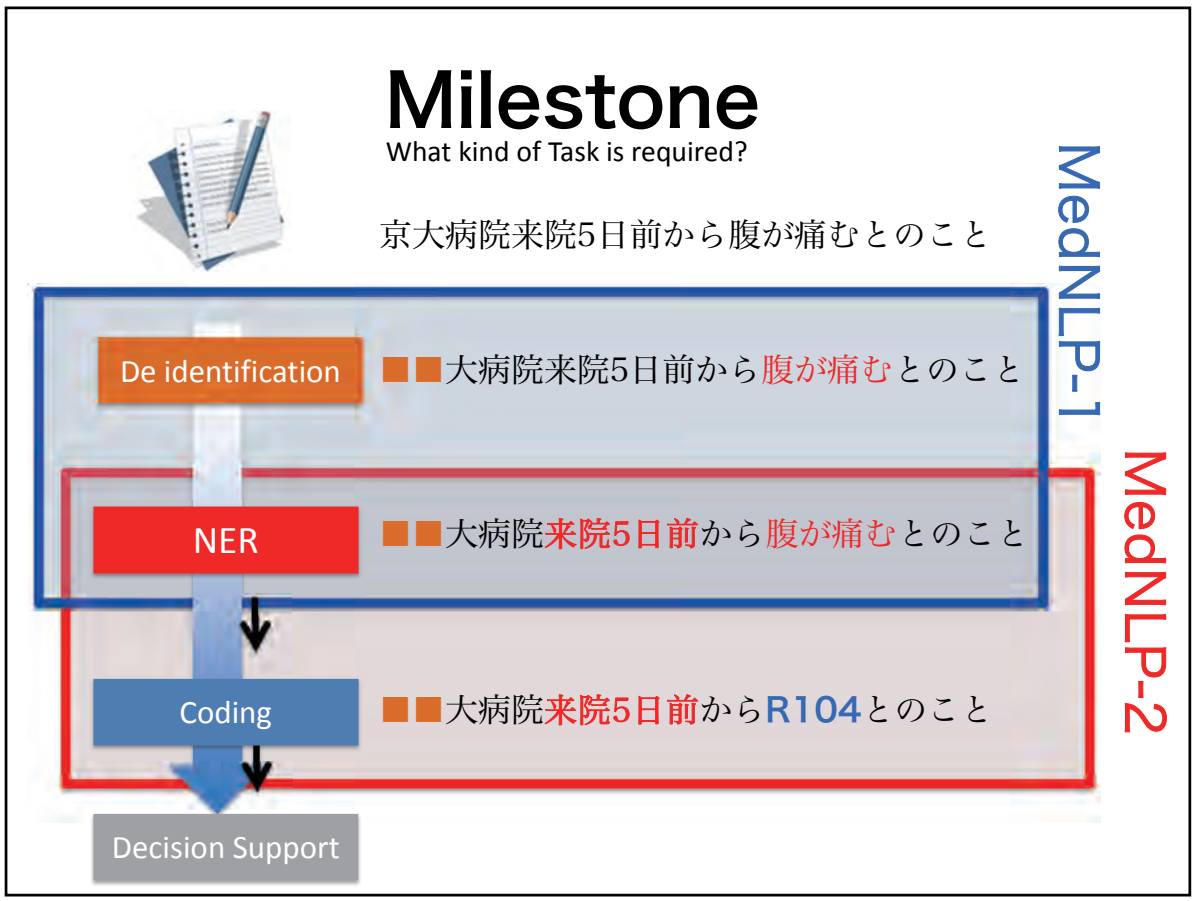
国立保健医療科学院
National Institute of Public Health
トーマツ
Deloitte Touche Tohmatsu

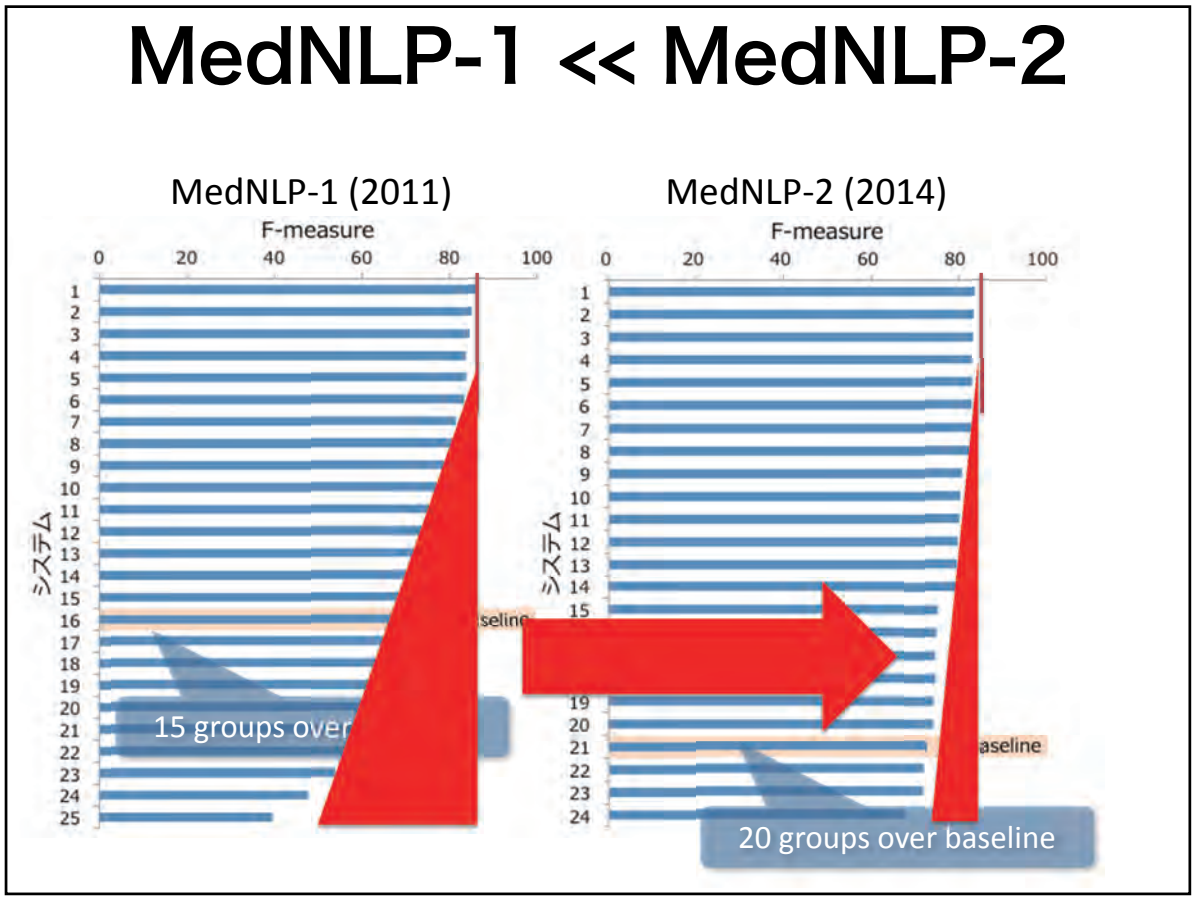
北陸先端科学技術大学院大学
JAIST
北海道大学
Hokkaido University
京都大学
Kyoto University
岡山大学
Okayama Prefectural University
岡山県立大学
Okayama Prefectural University
東京大学
The University of Tokyo
奈良先端科学技術大学院大学
Nara Institute of Science and Technology
安田女子大学
Yasuda Women's College

国立中央大学（台湾）
National Central University
朝陽科技大学（台湾）
Chaoyang University of Technology
南京大学 （中国）
Nanjing University
中央研究院 （台湾）
Academia Sinica
ダブリン大学（英国）
Dublin City University
NEC アメリカ（米国）
NEC USA

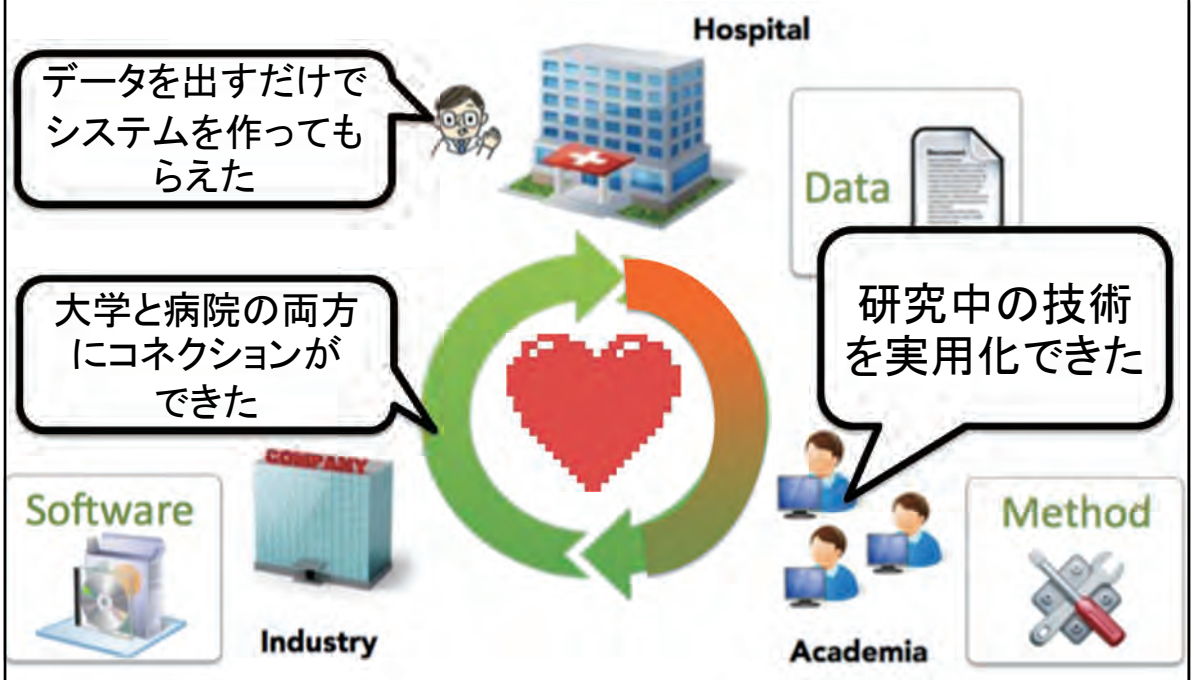
日本ユニシス
Nihon Unisys, Ltd
日立中央研究所
Hitachi, Ltd.
NTT研究所
NTT Science and Core Technology Laboratory Group
数理システム
MSI Knowledge
富士ゼロックス
Fuji Xerox
日立中央研究所
Hitachi, Ltd.
NTTデータ
NTT Data

| Task | MedNLP-1 | MedNLP-2 |
|---|---|---|
| De-identification | 6 groups (11 systems) | - |
| NER | 11 groups (15 systems) | 10 groups (24 systems) |
| ICD-coding | - | 9 groups (19 systems) |
| Free | 1 groups (1 systems) | 2 groups (2 systems) |


Milestone
What kind of Task is required?
京大病院来院5日前から腹が痛むとのこと
MedNLP-1
MedNLP-2
De identification
■■大病院来院5日前から腹が痛むとのこと
NER
■■大病院来院5日前から腹が痛むとのこと
Coding
■■大病院来院5日前からR104とのこと
Decision Support


MedNLP-1 << MedNLP-2
MedNLP-1 (2011)
MedNLP-2 (2014)
F-measure
システム
15 groups over
20 groups over baseline

# 循環器としての学問




# MedNLP: 医療言語処理プロジェクト

| | 原著論文 | 国際会議 | 受賞 |
|---|---|---|---|
| 2014年度 | 5編 | 4編 | 7賞 |
| 2013年度 | 4編 | 11編 | 4賞 |
| 2012年度 | 3編 | 6編 | 6賞 |
| 2011年度 | 2編 | 2編 | 4賞 |
| 累計 | 25編 | 48編 | 29賞 |

## 外部獲得資金（５年以内）

NTCIR MedNLP-2, NII共同研究費, 2013年度-2014年度, 研究代表者：荒牧英治, 2,000千円.
挑戦的萌芽研究, 2013年度～2015年度,「テキストの安全な匿名化に関する研究」研究代表者：荒牧英治, 2,800千円.
基盤研究（A）, 2011年度～2014年度,「確率関係モデルによる医療臨床データの高度活用に関する研究」, 研究分担者：荒牧英治（代表者：麻生英樹）, 47,970千円.
若手研究（A）, 2011年度～2014年度,「表記ゆれ及びそれに類する現象の包括的言語処理に関する研究」, 研究代表者（個人型研究）, 17,940千円.
挑戦的萌芽, 2010年度～2012年度,「ダミー診療録の構築および自動構造化に関する研究, 研究代表者：荒牧英治, 3,3500千円.
特定研究, 2009年度～2011年度,「コミュニティ型コンテンツのコンテンツホール検索の研究」, 研究分担者：荒牧英治（研究代表者：灘本明代）, 5,000千円.
若手研究（A）, 2008年度～2011年度,「非文法的かつ断片化したテキストからの情報抽出に関する研究」, 研究代表者：荒牧英治, 9,980千円.

奨学寄付金, 2012年度,（株）カレン, 2012年度,「ソーシャルメディア上の発言の分類器の構築に関する研究」, 研究代表者：荒牧英治.
奨学寄付金, 2012年度,（株）UTIX,「気象データを用いた風邪流行予測サイト・カゼミルの高度化」, 研究代表者, 研究代表者：荒牧英治.
奨学寄付金, 2012年度,（株）ブルワーク,「生命科学DB横断検索」, 研究代表者：荒牧英治.
奨学寄付金, 2012年度, Microsoft Research Asia（マイクロソフトアジア研究所）, CORE8,「患者モデルを用いた疾患観測モデルの構築に関する研究」研究代表者：荒牧英治.
奨学寄付金, 2010年度, Microsoft Research Asia（マイクロソフトアジア研究所）, Microsoft Research Asia eHealth Theme Program, 研究代表者.
奨学寄付金, 2010年度,（株）UTIX,「ウェブからの疾病情報の大規模かつ即時的な抽出手法」, 研究代表者, 研究代表者：荒牧英治.

# プロジェクトの受賞（5年以内）

2014, グループウェアとネットワークサービスワークショップ2014, ベストプレゼンテーション (9%=2件/22件) （研究主導者(PI)として）.
2014, グループウェアとネットワークサービスワークショップ2014, ベストペーパー賞 (9%=2件/22件) （研究主導者(PI)として）.
2014, 第13回情報科学技術フォーラム (FIT2014), FIT奨励賞 (15%=1件/約7件) （研究主導者(PI)として）.
2014, 京都大学 学際研究着想コンテスト, 優秀賞 (10%=3件/32件) (研究主導者(PI)として).
2014, 日本災害食学会, （カゴメ賞（企業賞））（共著者として）.
2014, 第33回医療情報学連合大会 (第14回日本医療情報学会学術大会), 優秀論文賞 (3%=数件/約200件) （研究主導者(PI)として）.
2014, データ工学と情報マネジメントに関するフォーラム(DEIM), 最優秀インタラクティブ賞 (0.6%=1件/約160件) （研究主導者(PI)として）.
2013, 京都大学 学際研究着想コンテスト, 奨励賞 （研究主導者(PI)として）.
2013, 第12回情報科学技術フォーラム (FIT2013), FIT奨励賞 (12%=1件/約8件) （研究主導者(PI)として）.
2013, 第32回医療情報学連合大会, 優秀口演賞 (3%=数件/200件) (共著者として).
2013, 第31回社会言語科学会, 研究大会発表賞 (3%=1件/30件) (研究主導者(PI)として).
2012, WebDBフォーラム, 企業賞（チームラボ賞）（共著者として）.
2012, 言語処理学会 第18回年次大会, 優秀賞 (2-3%=数件/約340件)（研究主導者(PI)として）.
2012, 言語処理学会 第18回年次大会, 若手奨励賞 (2-3%=数件/約340件) （共著者として）.
2012, テキストアノテーションワークショップ 2012, 奨励賞 (16%=3件/18件) (研究主導者(PI)として).
2012, マルチメディア、分散、協調とモバイルシンポジウム(DICOMO), ヤングリサーチャ賞 (10%=30件/300件) (研究主導者(PI)として).
2012, 第31回医療情報学連合大会, 優秀賞 (1.5%=5件/300件) (共著者として).
2011, CLIO Healthcare Awards 2011, Gold Awards (最高賞) （技術提供サイト「カゼミル」による）（世界最高峰の広告コンクール）
2011, Spikes Asia Advertising festival (スパイクス アジア 広告祭) 2011, デジタル部門 Gold Awards (第二位)（技術提供サイト「カゼミル」による）
2011, 第81回グループウェアとネットワークサービス研究会, 優秀発表賞 (研究主導者(PI)として)（10%=2件/19件）.
2011, Emerald Literati Network 2011 Awards for Excellence, Outstanding Paper Award Winner (共著者として).
2010, 言語処理学会 第16回年次大会, 優秀発表賞 (2%=5件/270件).
2010, 第29回医療情報学連合大会, 研究奨励賞 (2%=数件/300件) (共著者として).
2010, 第29回医療情報学連合大会, 優秀口演賞 (2%=数件/300件) (共著者として).
2010, データ工学と情報マネジメントに関するフォーラム (DEIM), 最優秀論文賞 (共著者として) .

患者文書プロジェクト