# 第７回産業日本語研究会・シンポジウム
# 予稿集

平成２８年２月２９日

於　東京・丸ビルホール（丸の内ビルディング７階）

高度言語情報融合フォーラム

一般社団法人
言 語 処 理 学 会

一般財団法人
日本特許情報機構

# 第７回産業日本語研究会・シンポジウム開催のご案内

平成２８年２月

## 人工知能と産業日本語の出会い

## ～先進的グローバル・ビジネスへの展開と躍進～

市場のグローバル化を背景に、我が国企業の海外展開が加速化しており、グローバル・ビジネスの場面においても産業・技術文書（特許明細書、技術論文、技術説明書など）の役割は、重要性を増しています。特に、情報発信力の円滑化、知的生産力の効率化、情報抽出・分析力の高度化を実現するために､産業･技術文書の正確な記述は欠かすことができません。

産業日本語研究会では、以上のような背景のもと、産業分野・科学技術分野における情報発信力や知的生産性の飛躍に貢献するとともに、我が国産業界全体の国際競争力の強化に資する「産業日本語」をテーマに、各種研究や提言に継続的に取り組んでいます。

また、平成２２年より毎年シンポジウムを開催し、特許、法令工学、翻訳、テクニカルコミュニケーション、データマイニング、システム開発など多分野の取組を紹介し、産業日本語の必要性につき議論を重ねてまいりました。

本年の「第７回産業日本語研究会・シンポジウム」では、産業日本語が人工知能と出会うことで、更なる展開と躍進を目指します。我が国企業によるグローバル・ビジネス領域での言語処理技術や人工知能技術の先進的な適用事例を紹介し、産業日本語に期待される重要な役割につき議論できる場といたします。

産業日本語研究会世話人会
顧問：長尾　眞　（京都大学名誉教授）
代表：井佐原 均　（豊橋技術科学大学）
辻井　潤一　（産業技術総合研究所）
橋田　浩一　（東京大学）
隅田　英一郎（情報通信研究機構）
山口　昌也　（国立国語研究所）
横井　俊夫　（日本特許情報機構特許情報研究所）
潮田　　明　（元奈良先端科学技術大学院大学）
河合　弘明　（日本特許情報機構）

■主催：高度言語情報融合フォーラム（ALAGIN）、言語処理学会、日本特許情報機構（Japio）
■後援：総務省、経済産業省、特許庁、国立国語研究所、情報通信研究機構、
工業所有権情報・研修館、情報処理学会、人工知能学会、
日本知的財産協会、アジア太平洋機械翻訳協会、大学技術移転協議会、
フジサンケイビジネスアイ
■日時：2015 年 2 月 29 日（月） 13：00-18：00
■場所：東京・丸ビルホール（丸の内ビルディング７階）
http://www.marunouchi-hc.jp/hc-marubiru/access.html
■テーマ：人工知能と産業日本語の出会い
～先進的グローバル・ビジネスへの展開と躍進～
■参加費：無料 （事前登録制、懇談会・意見交換会は要会費）

■プログラム：

【第一部】オープニング (13:00－14:05)

(1) 開会挨拶
長尾 眞 公益財団法人 国際高等研究所 所長
／国立大学法人 京都大学 名誉教授

(2) 基調講演：日本語の全体像を知るために---------- 1
―国立国語研究所による言語資源整備―
前川 喜久雄 大学共同利用機関法人 人間文化研究機構 国立国語研究所
副所長／言語資源研究系長 教授
(3) 特別講演：言語理解と人工知能---------- 7
辻井 潤一 国立研究開発法人 産業技術総合研究所
人工知能研究センター 研究センター長
／国立大学法人 東京大学 名誉教授

【第二部】産業界と産業日本語の交わり (14:05－15:10)

(4) 産学連携における学術用語の一般化----------31
山本 貴史 株式会社東京大学ＴＬＯ 代表取締役社長
(5) 日本企業の一層のグローバル化と産業日本語----------35
井上 二三夫 シスメックス株式会社 研究開発企画本部 副本部長 兼
知的財産部長／一般社団法人 日本知的財産協会 副理事長
(6) 研究会報告：「日本人のための日本語マニュアル」のご紹介----------47
横井 俊夫 一般財団法人 日本特許情報機構 特許情報研究所 顧問
／学校法人片柳学園 東京工科大学 名誉教授

（休憩 10 分）

【第三部】人工知能研究と言語処理研究の最前線 (15:20－17:00)

(7) 特別講演：質問応答タスクと自然言語処理----------63
武田 浩一 日本アイ・ビー・エム株式会社 東京基礎研究所 技術理事
(8) リクルート AI 研究所から見る言語処理研究への期待----------67
石山 洸 株式会社リクルートホールディングス
R&D 本部 Recruit Institute of Technology 室長

(9) ニュースアプリケーションにおける自然言語処理----------75
徳永 拓之　　スマートニュース株式会社 エンジニア
(10) 1000 社が採用 「見える化エンジン」が挑戦し続ける----------87
テキストマイニングのビジネス利用最前線
坂 慎弥　　株式会社プラスアルファ・コンサルティング
見える化エンジン事業部 テキストマイニング研究 G
グループマネージャー

**（休憩　5 分）**

**【第四部】人工知能と産業日本語の出会い**　(17:05－18:00)

(11) パネル討論
＜モデレータ＞
渡部 俊也　　国立大学法人 東京大学 政策ビジョン研究センター 教授
＜パネリスト＞
前川 喜久雄　大学共同利用機関法人 人間文化研究機構 国立国語研究所
副所長／言語資源研究系長 教授
武田 浩一　　日本アイ・ビー・エム株式会社 東京基礎研究所 技術理事
石山 洸　　株式会社リクルートホールディングス
R&D 本部 Recruit Institute of Technology 室長
徳永 拓之　　スマートニュース株式会社 エンジニア
坂 慎弥　　株式会社プラスアルファ・コンサルティング
見える化エンジン事業部 テキストマイニング
研究 G グループマネージャー
(12) 閉会挨拶
井佐原 均　　産業日本語研究会 代表世話人
／国立大学法人 豊橋技術科学大学
情報メディア基盤センター センター長 教授

**【第五部】ポスターセッション（ホワイエにて開催）**　(12:30－18:15)

(13) 「特許版・産業日本語委員会」活動報告
(14) 「特許ライティングマニュアル＜改訂版＞」（日本特許情報機構）
(15) 「日本語マニュアルの会」活動報告（日本語マニュアルの会）
(16) 特別展示：請求項パターンと統計に基づく特許の高精度自動翻訳（情報通信研究機構）
(17) 特別展示：1000 社が採用「見える化エンジン」™
（プラスアルファ・コンサルティング）

※懇親会・意見交換会
シンポジウム終了後、丸ビル内レストランにて懇親会・意見交換会（定員：先着 50 名、会費：3,000 円）の開催を予定しております。

# 第一部
# オープニング

# 基調講演
# 日本語の全体像を知るために
# ―国立国語研究所による言語資源整備―

産業日本語のような試みが必要とされる背景には、日本語の多様性の問題がある。言語の多様性というと世界諸言語の多様性に注目が集まりやすいが、実は日本語の内部にも一般に想像されるよりも大きな多様性がある。多様性をもたらす要因としては言語の歴史的変化と方言の分化があることが古くから知られているが、これらが要因のすべてではない。現代語の書き言葉だけに着目してもかなり顕著な文法のゆれが観察されるし、話し言葉の多様性についてはそれをどう把握するかについて試行錯誤が続いている段階である。日本語を含めた自然言語にはいわば内在的かつ本質的に多様性が存在しているように思われる。したがって日本語の科学的研究の重要な目的のひとつは多様性をはらんだ日本語の全体像を客観的に把握することであると考えられる。そのためには各種コーパスに代表される言語資源の組織的な整備が必須であるし、コーパスアノテーション技術や言語構造の自動解析技術などの基礎研究も重要である。この講演では国立国語研究所が近年推進してきたKOTONOHA計画による言語資源整備の進捗状況を紹介し、今後の整備計画を紹介する。産業日本語のような試みも日本語の多様性の把握に立脚して進められてこそ有効に機能するであろう。

前川 喜久雄

大学共同利用機関法人 人間文化研究機構

国立国語研究所副所長／言語資源研究系長 教授

# 日本語の全体像を知るために

―国立国語研究所による言語資源整備―

大学共同利用機関法人 人間文化研究機構
国立国語研究所副所長／言語資源研究系長 教授
前川喜久雄

## １．はじめに

産業日本語は日本語に対する領域限定的な規制の試みである。言語の規制に対する言語研究者の立場は、これを強く要請する者から全く不可とするものまで様々であるが、そもそも言語の規制というアイデアが浮かぶ原因が言語の多様性にあるという点については大方の一致が見られるものと思う。

言語の多様性というと世界に何千種類の言語があるという類の、言語間の多様性が注目されがちであるが、一言語の内部にも多様性がある。規制の問題と結びつくのは、この言語内多様性の問題である。

言語内多様性をもたらす要因には様々なものがある。そのうち言語の歴史的変化と地理的変化につては、古くから研究が行われてきており、データも不十分ながら整備されてきている。しかし、多様性をこのふたつの要因で説明しきることは到底不可能である。生きた言語を仔細に観察すると、歴史的変化とも地理的変化とも関係しない、あるいは双方と関係するように見える変異、いわば純粋に確率的な変異とでも呼ぶべき言語変異現象があり、変異の大部分はむしろそのような性質のものであると思われるからである。

そのような変異の研究は標準語を対象として行われることが多いが、現代の標準語は話し言葉としても書き言葉としてもデータサイズが膨大であるために、一個人の力でその全体像を把握することは極めて困難である。

国立国語研究所ではこの問題の解決に寄与するために、1990 年代末から一連のコーパスを構築し公開してきた。また 2016 年の 4 月から始まる次期中期計画期間(2021 年度までの 6 年間)においても、コーパスに代表される言語資源の整備が研究所の重要な活動目標のひとつとなっている。以下では言語資源の整備に関わる国語研のこれまでの成果と今後の活動計画を紹介する。

## ２．KOTONOHA 計画

KOTONOHA 計画とは国語研による言語資源整備計画の総称である。実際にこの名称を使い始めたのは 2006 年からだが、ここでは 1999 年に計画がスタートしたものとして説明する。最初に KOTONOHA 計画開始直前の状況を説明しておくと、現代語の書き言葉データとして、まとまった量を誰でも利用できたのは新聞各社の記事テキストデータベースのみであった。他に著作権の消滅した文芸作品を中心とする「青空文庫」のデータも利用でき

たがこれは現代というよりむしろ近代語のデータである。

### ２．１　『日本語話し言葉コーパス』（Corpus of Spontaneous Japanese: CSJ）

1999 年に構築をはじめ、2004 年に公開した話し言葉コーパスである。開発費は科技庁の科学技術振興調整費であり、東工大、情報通信研究機構との共同開発である。学会での研究発表や少人数の聴衆を前にしたスピーチなどのモノローグを中心に 662 時間 (750 万語) の音声を精密に転記し、短単位と長単位による二重形態素解析、節境界情報、印象評定情報などのアノテーションを施した。コアと呼ばれる 50 万語分のサブセットには X-JToBI による精密な分節音・韻律アノテーションも施した。CSJ はもともと自発音声の音声認識研究用に設計したコーパスであり、音声認識研究に大きく貢献したが、コアは音声学の領域でもよく利用されている。Google Scholar で検索すると引用件数は 800 件を超えている。

### ２．２　『現代日本語書き言葉均衡コーパス』（Balanced Corpus of Contemporary Written Japanese: BCCWJ）

2006 年に構築をはじめ、2010 年に公開した日本語初の均衡コーパスである。開発費は文科省科研費（特定領域研究）であった。規模は British National Corpus に倣って 1 億語であり、書籍、雑誌、新聞、白書、ブログ、ネット掲示板、法律、広報誌、韻文など 11 種のレジスターにわたる書き言葉のサンプルを可能な限りランダムサンプリングによって収集している。従来利用が困難であった各種書籍からのサンプルを大量に含んでいる点と、全サンプルに著作権処理が施されていて安心して利用できる点に特長がある。CSJ と同様、テキストには二重の形態素解析が施されている。

BCCWJ は DVD で全データが公開されているが、ウェブ上の検索アプリ（『中納言』）で形態論情報を検索することもできる。日本語学の広い領域で活発に利用されており、現時点での引用件数は 600 件を超えている。

### ２．３　『国語研日本語ウェブコーパス』(NINJAL Web Japanese Corpus: NWJC)

BCCWJ では量的に不足する研究がある。そのために母集団をウェブ上の日本語に限って 200 億語規模のコーパスを構築したのが『国語研日本語ウェブコーパス』である。開発費は運営費交付金（特別経費）である。2010 年に設計を開始して 2015 年に構築を終えた。現在は 2016 年の公開にむけて準備を進めている。短単位相当の形態論情報に加えて、文節係り受け構造の情報も提供する。

### ２．４　『日本語歴史コーパス』(Corpus of Historical Japanese: CHJ)

国立国語研究所は 1948 年に現代日本語を研究する国立試験研究機関として設置されたが、2009 年に大学共同利用機関法人に移管されるに際してこの制約が外された。そこで学界の要望に応えて過去の日本語を対象とするコーパスの構築を開始した。これが CHJ である。

現在は古典文学作品を対象として作業を進めており、本文テキストは㈱小学館のご厚意で同社の『新編日本古典日本文学全集』を利用させてもらっている。現時点で「平安時代編」約 73 万語と「室町時代編 I 狂言」約 24 万語が公開されている。テキストには短単位で形態素解析が施されており、BCCWJ と同じ『中納言』での検索が可能である。

CHJ とは別に近代語（明治・大正期の日本語）のコーパスも構築している。総合雑誌のテキストを対象としており、『太陽コーパス』『近代女性雑誌コーパス』『明六雑誌コーパス』『国民之友コーパス』を公開しているが、将来的には CHJ と統合する予定である。

## ３．今後の開発計画

以上のコーパス以外にも国立国語研究所内で構築作業が進められているコーパスがいくつかある。そのうち三つを紹介する。

第一のコーパスは、日本語学習者による日本語のサンプルを集めた『多言語母語の日本語学習者横断コーパス』(International corpus of Japanese As a Second language: I-JAS)である。日本を含む 20 の国と地域で、異なる 12 言語を母語とする日本語学習者 1000 人の話し言葉と書き言葉を収集することを目標としている。データの一部は今春公開の予定であるが、最終的な公開は次期中期計画期間年度末（2021 年度）を予定している。

第二のコーパスは「方言コーパス」（仮称）である。これは 1977～85 年に文化庁が実施した「各地方言収集緊急調査」で収録された方言談話資料（全国 224 地点×30 時間の録音）を活用したコーパスであり、方言談話中の語形を標準語で検索可能とし、転記テキストと方言音声を聴取可能とする予定である。

第三のコーパスは「大規模日常会話コーパス」(仮称)である。CSJ では対象から外れた日常の話し言葉を対象とするコーパスであり、首都圏の日常会話を個人の生活に密着する手法と特定の場面を想定する手法とで合計 1000 時間分収録する予定である。「方言コーパス」「大規模日常会話コーパス」ともに、やはり 2021 年度の公開を目標としている。

ここまで紹介してきたコーパスは、それぞれ独自の経緯をもって開発されたものであり、検索系はそれぞれに独立している。しかし、日本語全体を研究する立場からすれば、各コーパスをバラバラに利用するのではなく、一括して検索できることが望ましい。これは理論的にも技術的にもいくつかの難しい問題を孕んだ要請であるのだが、コーパス利用の利便性に大きくかかわる問題であるから、包括的な利用を可能にする検索環境を実現して公開したいと考えている。

国立国語研究所が開発した言語資源でコーパス以外に広く活用されているものに形態素解析用辞書 UniDic がある。これは BCCWJ のために開発された短単位辞書で、2005 年以降更新されていない。UniDic の拡張も次期中期計画期間の目標のひとつに掲げている。

## ４．現代標準語の言語内多様性

すでに構築が完了している CSJ、BCCWJ、NWJC などを検索することで、現代標準語の

言語内多様性を客観的に把握することが可能になってきた。本稿では紙幅の関係で省略に従うが、講演時には以下のような実例を示す予定でいる。①「来れる」などのラ抜き言葉の使用に関する意識と実態の乖離、②文法書では常に「～ている」の形で使われるとされる動詞類（「そびえる」等）の例外的用法、③「形容詞＋です」「動詞＋です」の用例、④「書けれる」のように可能動詞に可能の助動詞が付加されたと解釈できる二重可能の用例、⑤「～するべきでない」の意味で「～しないべき」を用いる例などである。

## ５．おわりに

最後に指摘しておきたいことがある。それは、少なくとも日本語の場合、言語コーパスや音声コーパスの構築に最初に取り組んだのは人文系の言語研究者ではなく、音声認識や機械翻訳などを研究する情報処理領域の研究者たちであったという事実である。

本稿では国立国語研究所の活動ばかりに触れたが、音声コーパスであれば、音声認識研究用に構築された種々の朗読音声コーパスが CSJ 以前に構築され広く利用されていた。また書き言葉についても、新聞記事を素材とする『京都大学テキストコーパス』が 90 年代に開発されている。

KOTONOHA 計画で開発した各種コーパスはこのような先行研究の成果に立脚して設計されたものであるが、わけても自動形態素解析に代表されるアノテーションの自動化技術は日本語コーパスを質量ともに一変させる効果があった。国立国語研究所におけるコーパス開発は今後とも情報科学の成果に多くを負う形で進められるであろう。反対に言語研究者が情報科学にできる貢献としては、計画的に設計され、品質管理されたコーパスの開発が重要であり、現に CSJ や BCCWJ は情報科学領域でも活発に利用されている。

今後、人文系の言語研究における諸問題、特に言語内多様性の分析にコーパスを活用するにあたり、階層ベイズモデルなどによる複雑かつ柔軟な統計モデリングが非常に重要な役割を果たすものと予想している。そのような研究において、言語研究者と情報科学研究者による相方向的な共同研究が実現できれば、言語の本質解明に到る新しい可能性が拓けるにちがいないと期待している。産業日本語も、そのような分析の成果に立脚して設計を進めれば、よりよい成果を得られるのではなかろうか。

## ６．本稿で紹介した言語資源の URL

BCCWJ: http://pj.ninjal.ac.jp/corpus_center/bccwj/
CHJ: http://pj.ninjal.ac.jp/corpus_center/chj/
CSJ: http://pj.ninjal.ac.jp/corpus_center/csj/
I-JAS: http://ninjal-sakoda.sakura.ne.jp/lsaj/
NWJC: http://pj.ninjal.ac.jp/corpus_center/nwjc/
UniDic: https://osdn.jp/projects/unidic/
各種音声研究用コーパス: http://research.nii.ac.jp/src/list.html
各種近代語コーパス: http://pj.ninjal.ac.jp/corpus_center/cmj/
京都大学テキストコーパス: http://nlp.ist.i.kyoto-u.ac.jp

# 特別講演
# 言語理解と人工知能

言語処理が言語の表層を操作する分野を指すのに対して、言語理解は、言語とそれ以外のものとを対応つける技術分野を指す。後者は、人工知能の中核分野の一つであるが、その長い歴史は必ずしも成功の歴史ではない。その原因はいろいろ考えられるが、言語を対応づける相手に関する技術が未発達であったこと、対応づける相手を「意味」という抽象的な存在だと考えたことで、経験主義的な健全な研究ができなかったことが、大きな阻害要因であったと思う。現在の人工知能ブームは、この２つの阻害要因を取り除くことで、言語理解の研究にもある種の飛躍をもたらすことが期待される。このような観点から、研究を組み立てることを考えて見たい。

辻井 潤一

国立研究開発法人 産業技術総合研究所

人工知能研究センター 研究センター長

／国立大学法人 東京大学 名誉教授

# 人工知能と言語理解
## ー人間との協調を目指してー

辻井潤一
研究センター長
産業技術総合研究所
人工知能研究センター

1

# 話の流れ

- 人工知能研究センターの紹介
- 人工知能と科学研究 （AI for Big Sciences）
- Big Mechanismと言語理解
- おわりに

2

# 自然知能と親和性の高い人工知能

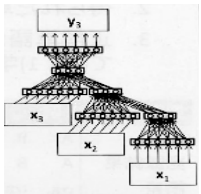
AIクラウド
脳型AI
データ・知識融合型AI

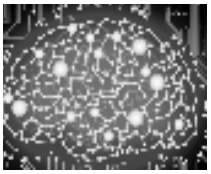
3

# 人工知能

## 人間に迫る人工知能

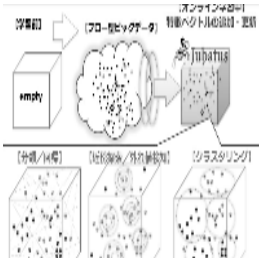
- **IBM ワトソン**: 言語理解、テキストと構造化された知識（事実）、検索と質問応答
- **コンピュータ将棋**: 大規模な探索空間, 機械学習
- **東大入試ロボット**: 言語理解、問題解決、知識に基づく推論
- **会話ロボット**: 身体性をもった知能, 特定の文脈下での言語理解
- **深層学習**: 脳からのヒント、計算原理の変革、自律性をもった機械学習
- **脳科学**: 人間知能の解明

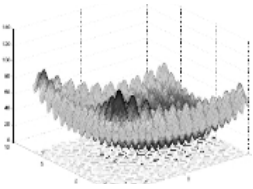
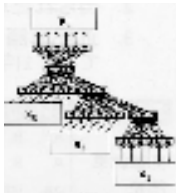



4

機械学習　大規模グラフ　GPU・スパコン　最適化技術　深層学習



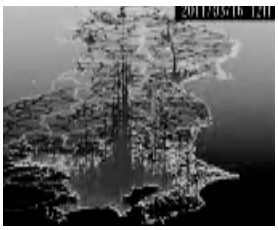


# もう一つの人工知能

ビッグデータ、データサイエンスからの人工知能
人間を超える人工知能





人間に迫る　人間を超える

# ２つの流れの統合

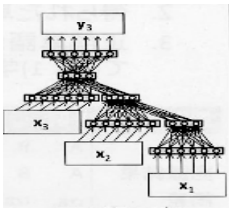
6

# AlphaGo（2016）
## シミュレーションと機械学習

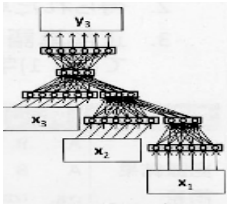
A game of perfect information

機械学習器

*v(s)*

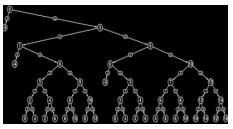


*p(a|s)*



学習データ

プロ棋士の対戦
棋譜データベース



シミュレーション
ゲーム



# AlphaGo by DeepMind (2016)

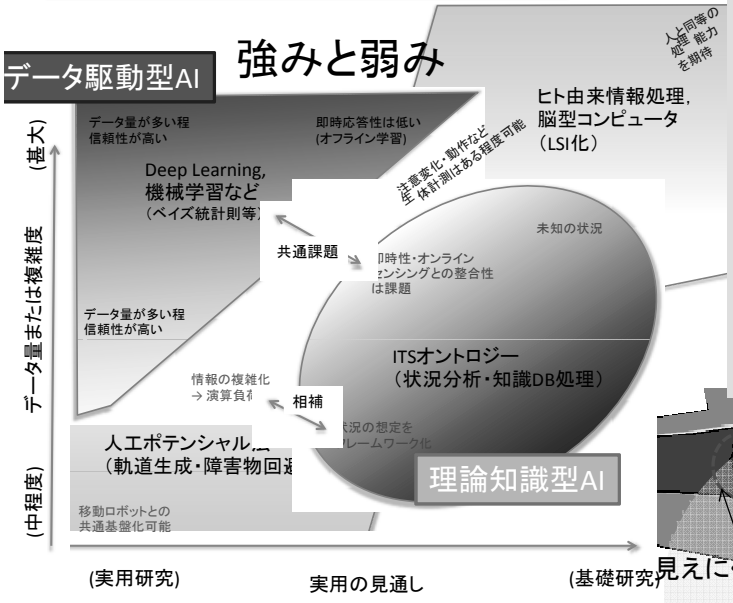
- *v(s)*:局面の評価
  - 深層学習（CDNN)
  - ステップ３:ランダムに選択した局面（30 million）から、*p(a|s)*を使ったシミュレーション・ゲームで最終局面まで
- *p(a|s)*: 指し手選択
  - 深層学習（CDNN)
  - ステップ１:教師付き学習（過去の棋譜データベース）
  - ステップ２:強化学習（シミュレーション・ゲームで最終局面）
- アルゴリズム（MCTS）、計算能力（GPU、CPU）
  - CPU（48, 1,202):シミュレーション
  - GPU(8, 176):深層学習

# 言語使用の特定状況へのGrounding
# 意図と行動計画（稲邑招聘研究員、NII）


多数のユーザがアバターとして
ロボットと会話

9

## 我妻、市瀬招聘研究員ら
九工大、NII、早大



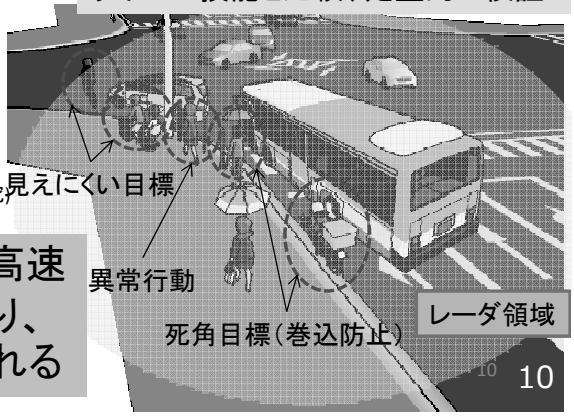

**目的**：自動運転における危険予測・回避行動判断

**提案方法**：二種のAI組合せの最適化からハイブリッドAI設計法を得る

**検証法**：AI実用に必須となる車メーカの製造者責任担保を可能にする。

・多種の実車計測データを活用し、ハード化により、ADAS支援の基準10ms-500ms実装を可能にする

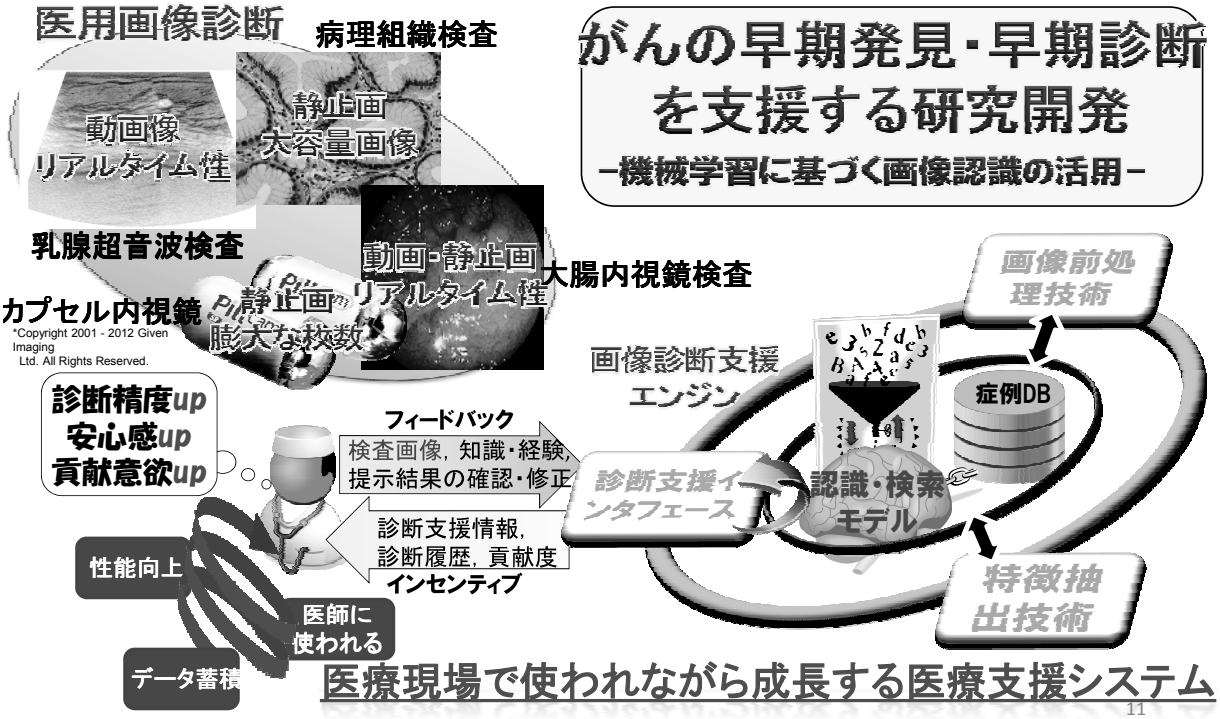
・オントロジー分枝構造から得られた「状況複雑度指標」を一般・熟練者ドライバー技能と比較、定量的に検証

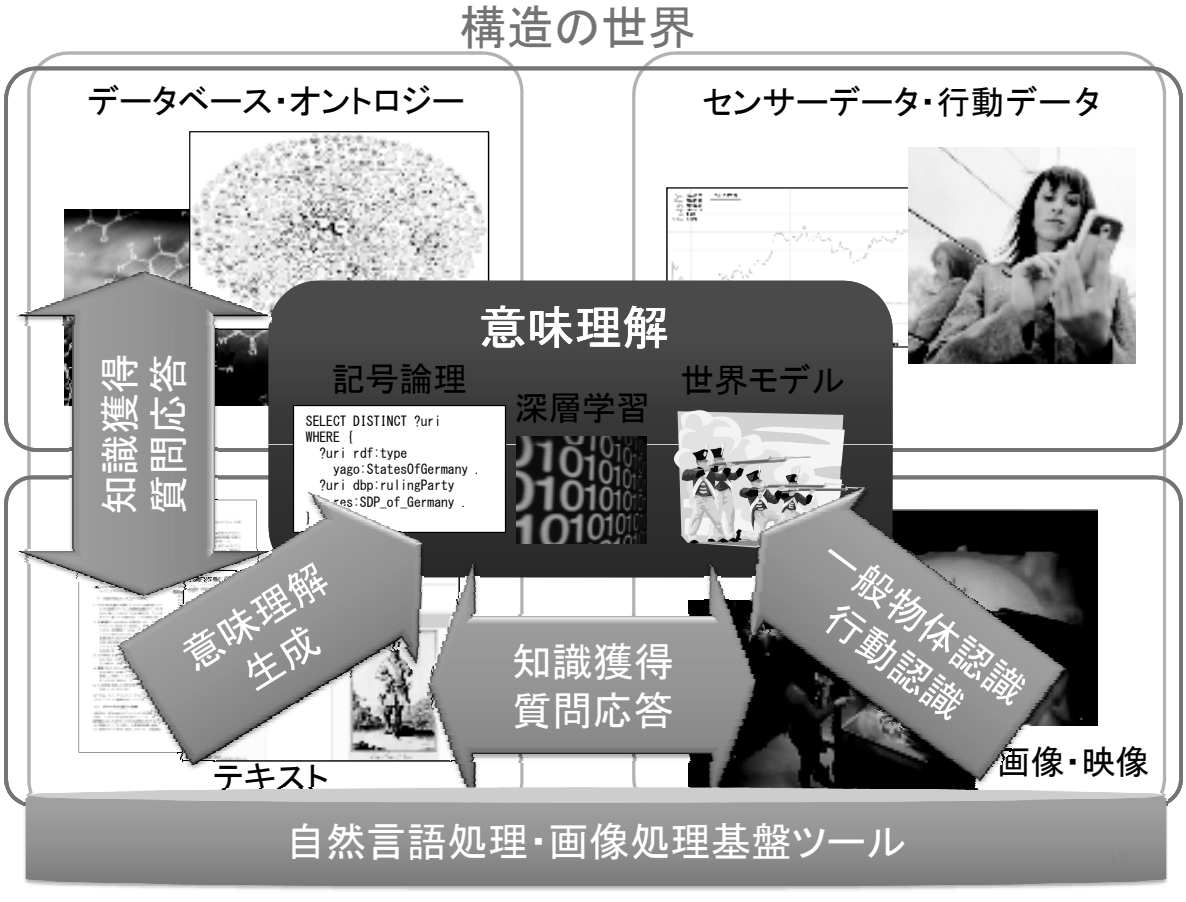



データ駆動型AIと理論知識型AIで、高速かつ「推論の説明責任」が可能になり、AIの自動運転分野での実用化が図れる

10

# 人工知能応用研究チーム: 医用画像診断支援



# 話の流れ

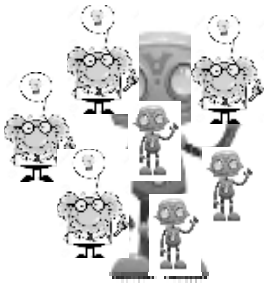
- 人工知能研究センターの紹介
- 人工知能と科学研究 （AI for Big Sciences）
- Big　Mechanismと言語理解
- おわりに

13

計算

**Open Access Journal/Open Data/Pathway Commons/Unique ID/Linked Data**

実験・観察



モデル・理論

データ解析

シミュレーション

**Machine Learning/Large Scale Search**

人工知能

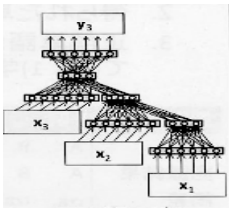
機械学習 ⇒ 計算モデル
探索 ⇒ 仮説の構築
ロボット ⇒ 仮説の実証

14

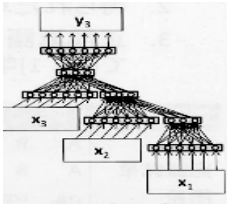
# AlphaGo
## シミュレーションと機械学習

A game of perfect information

機械学習器



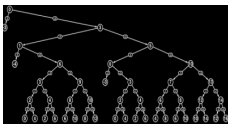
*v(s)*



*p(a|s)*

学習データ

プロ棋士の対戦
棋譜データベース



シミュレーション
ゲーム

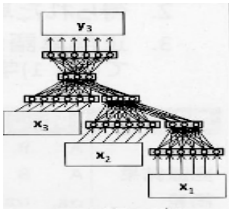


# 計算科学と人工知能
## シミュレーションと機械学習

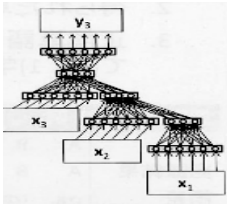
A game of perfect information

Unknowns

機械学習器

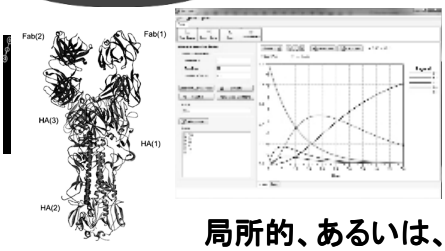


*v(s)*



*p(a|s)*

コンテキストの
異なる観察データ
プロ棋士の対戦
棋譜データベース
多様なオミクスデータ



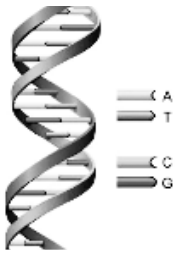
局所的、あるいは、
不完全なシミュレーション

# 話の流れ

- 人工知能研究センターの紹介
- 人工知能と科学研究 （AI for Big Sciences）
- Big Mechanismと言語理解
- おわりに

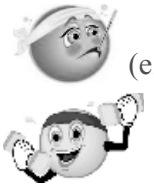
17

## Genome-Wide Association Studies (GWAS)



... ATTCGGATATTTAAGGC ...

... ATTCGGGTATTTAAGCC ...



Disease
(e.g., Alzheimer, Cancer)

Healthy

2000

Francis Collins (NIH)

"Genetic diagnosis of diseases would be accomplished in 10 years and that treatments would start to roll out perhaps five years after that."

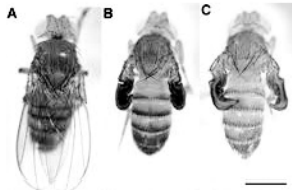
2010

"A Decade Later, Genetic Maps Yield Few New Cures" New York Times, June 2010.

***Hoifung Poon (MSR)***

18

# Traditional Biology




Targeted Experiments

Discovery

One hypothesis

***Hoifung Poon (MSR)***

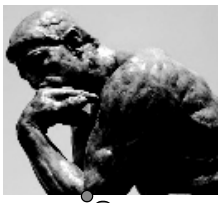
19

# Genomics

… ATTCGGATATTTAAGGC …
… ATTCGGGTATTTAAGCC …
… ATTCGGATATTTAAGGC …
… ATTCGGGTATTTAAGCC …
… ATTCGGATATTTAAGGC …
… ATTCGGGTATTTAAGCC …

High-Throughput Experiments

……



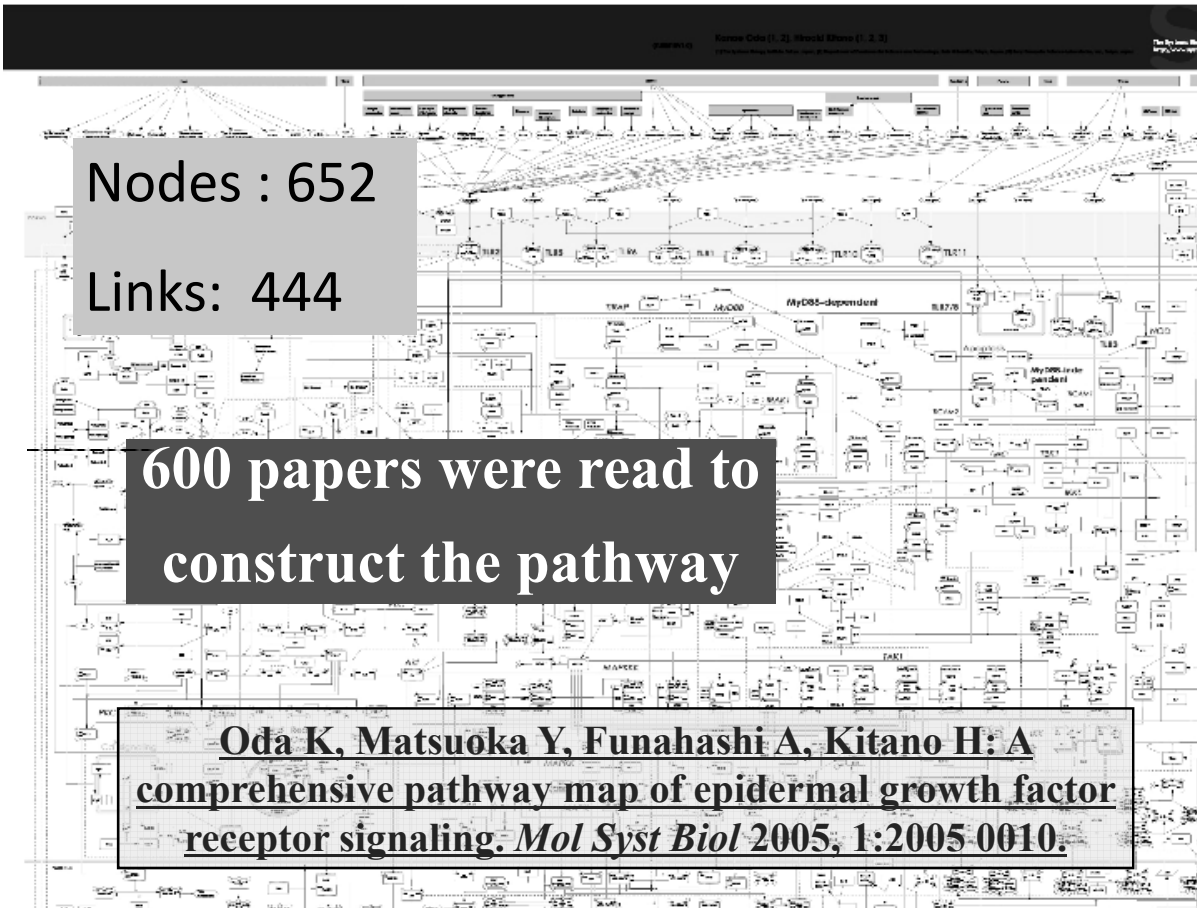
Discovery

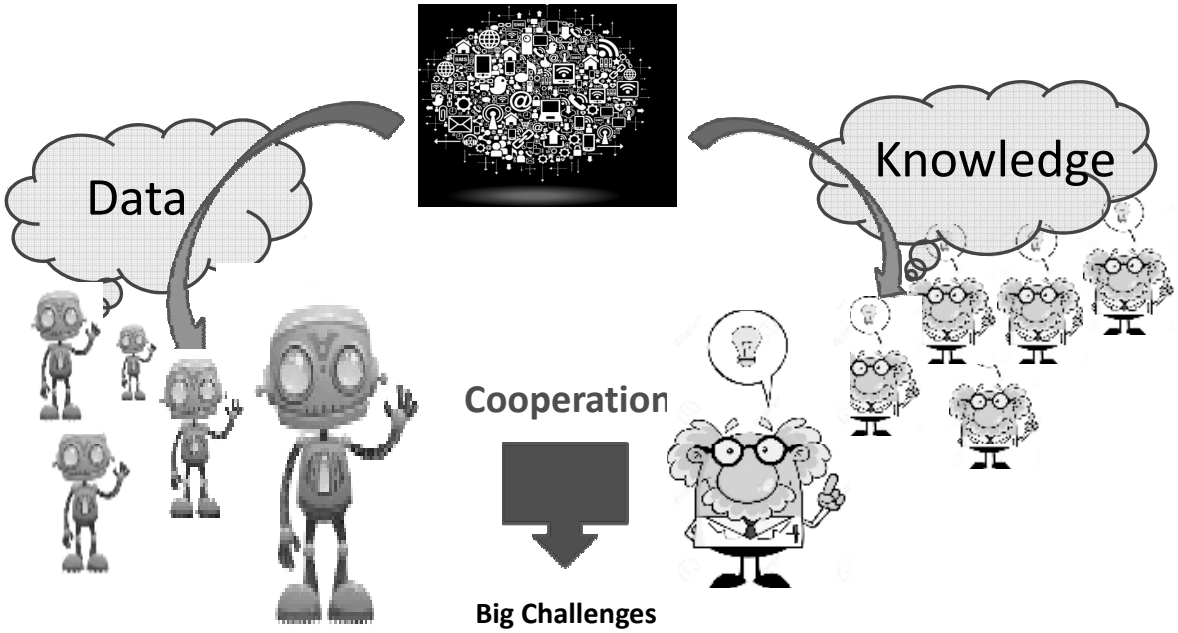
Too many hypotheses

Big Data

***Hoifung Poon (MSR)***

20



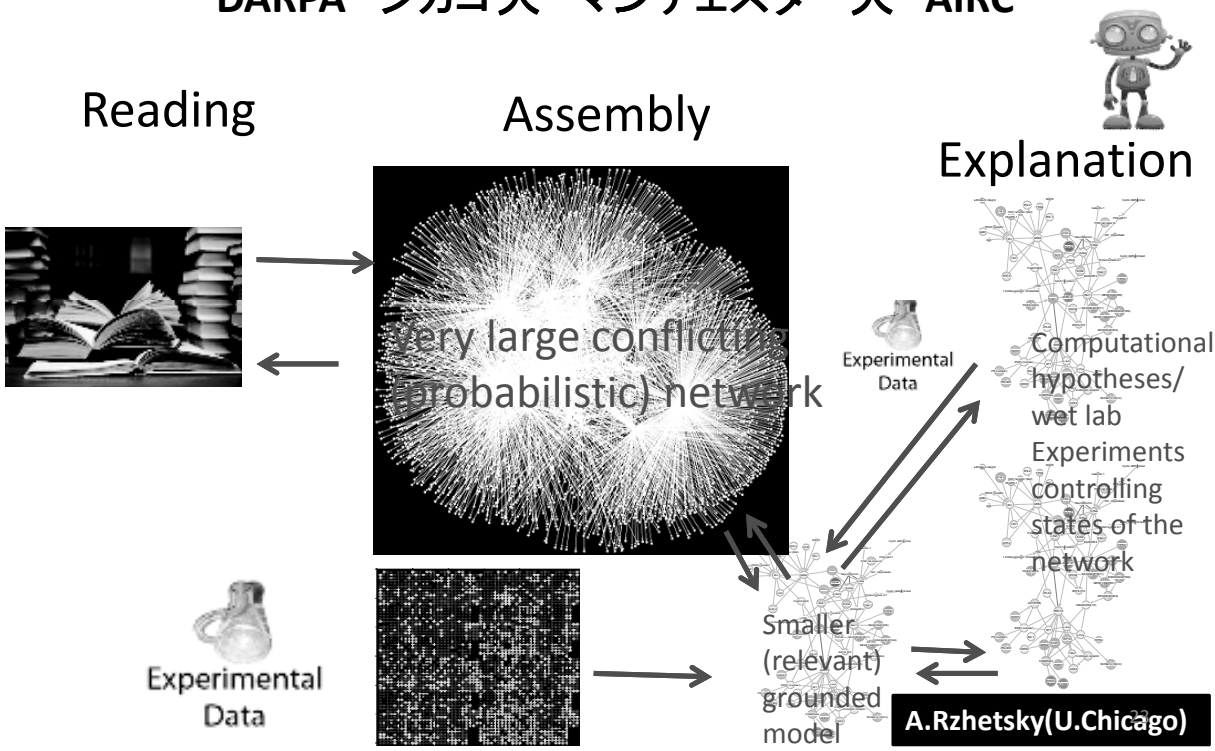

Artificial Intelligence

Natural Intelligence

22

## Big Mechanism: ロボット・サイエンティスト

**DARPA　シカゴ大　マンチェスター大　AIRC**




# Big Mechanism

- Project supported by DARPA
- Some of the systems that matter most to the Defense Department are very complicated. Ecosystems, brains and economic and social systems have many parts and processes, but they are studied piecewise, and their literatures and data are fragmented, distributed and inconsistent. It is difficult to build complete, explanatory models of complicated systems, and so effects in these systems that are brought about by many interacting factors are poorly understood.
- Big mechanisms are large, explanatory models of complicated systems in which interactions have important causal effects. The collection of big data is increasingly automated, but the creation of big mechanisms remains a human endeavor made increasingly difficult by the fragmentation and distribution of knowledge. To the extent that the construction of big mechanisms can be automated, it could change how science is done.

24

# The Need for Text Mining

**Types of documents**

- Full papers
- Abstracts
- Reports, discharge summaries
- EMR
- Textbooks, monographs
- Grey content, online discussion forums

**MEDLINE**

- 2005: ~14M
- 2009: ~18M
- 2013: ~22M
- 2015: ~26M

Overwhelming information in textual, **unstructured** format

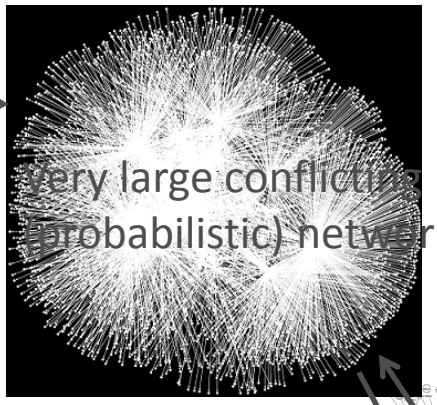
**S.Ananiadou (U.Manchester)**

# Big Mechanism: Reading-Assembly-Explanation

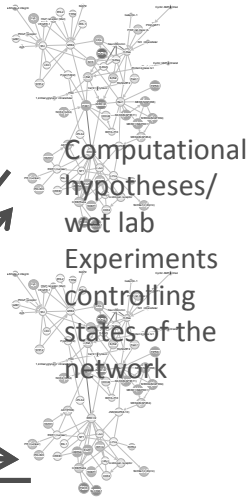
Reading

Assembly

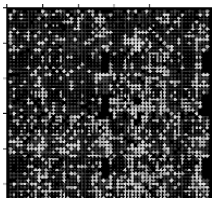
Explanation











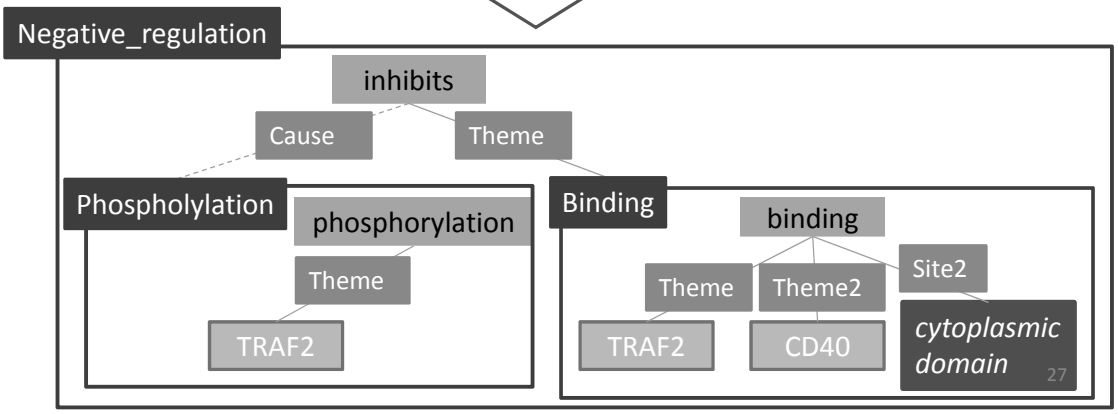

**A.Rzhetsky(U.Chicago)**

# Event E

**http://www.nactem.ac.uk/EventMine/**

Finding events ( trigger mentions , event types , and typed arguments including locations ) involving genes or gene products

*… In this study we hypothesized that the phosphorylation of TRAF2 inhibits binding to the CD40 cytoplasmic domain. …*




# Finding Evidence -EuropePubMed Central

- Currently: runs on 2,550, 328 full texts
- 82,198,474 facts in 38,411,661 sentences
- Full parsing used a version of Enju (Mogura)
- Parsing pipeline run on 60 machines at EBI ~30 days

http://labs.europepmc.org/evf

Europe PMC

**S.Ananiadou (U.Manchester)**

# Deep Reading: Reading with a Model

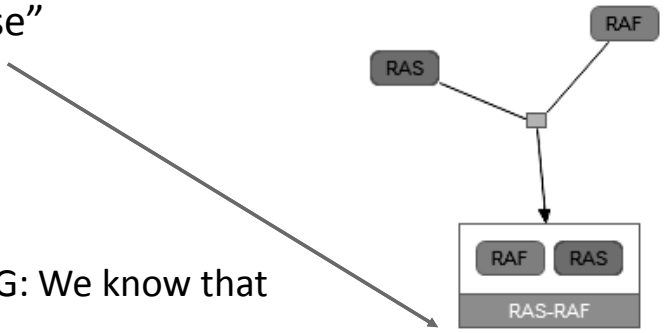
- Goal: evaluate how TM systems process text in relation to what is known about a pathway
- Performers asked to produce
  - Relationship/proposed change to the model (new/corroborating/conflicting information)
  - A model fragment describing the change
  - The source text supporting the change

L.Hirschman

29

# Reading against a Model (1)

"monoubiquitination of Ras enhances association with the downstream effectors Raf and PI3-Kinase"



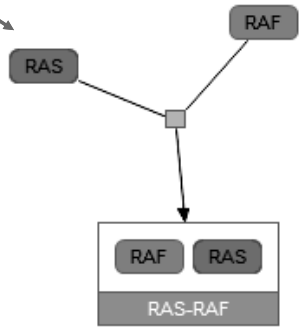

CORROBORATING: We know that *Ras binds Raf*

L.Hirschman

30

# Reading against a Model (2)

"monoubiquitination of Ras enhances association with the downstream effectors Raf and PI3-Kinase"




NEW MECHANISM: Ras binds PI3-Kinase.

BEL: complex(p(PFH:"Ras family"), p("PI3K"))

L.Hirschman 31

# Epistemic knowledge

- Enriches event-based search systems
  - Discovery of new knowledge
  - Negation, uncertainty, speculative claims in literature

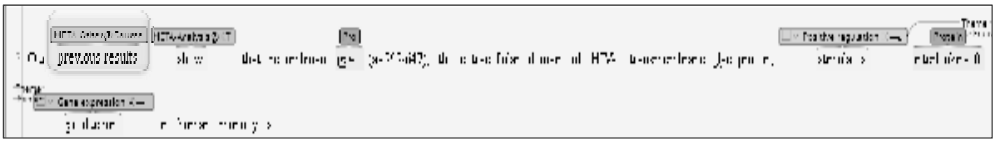
*… In this study we hypothesized that the phosphorylation of TRAF2 inhibits binding to the CD40 cytoplasmic domain. …*

Miwa, Thompson, McNaught, Kell, Ananiadou (2012). Extracting semantically enriched events from biomedical literature. **BMC Bioinformatics** 13, 108

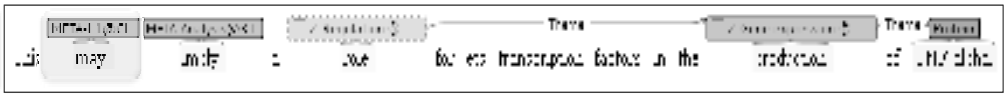
32

# Extracting epistemic knowledge

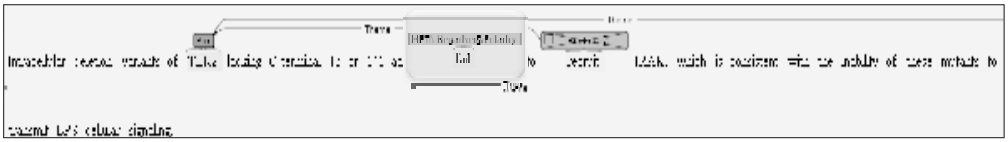
Source



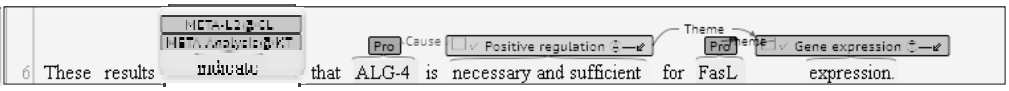
Uncertainty




Negation



Analysis




33

# Named entity recognition

- Summary of evaluation results (NaCTeM)

| Type | Precision | Recall | F-score |
|---|---|---|---|
| CellLine | 0.94 | 0.64 | 0.76 |
| ChemicalOrDrug | 0.82 | 1.00 | 0.90 |
| Complex | NA | 0.00 | NA |
| DrugClass | 1.00 | 0.29 | 0.44 |
| GeneOrProtein | 0.81 | 0.50 | 0.62 |
| Pathway | 1.00 | 0.86 | 0.92 |
| ProteinFamily | 0.73 | 0.88 | 0.80 |
| SubcellularLocation | 1.00 | 1.00 | 1.00 |
| **OVERALL** | **0.84** | **0.66** | **0.74** |

Top ranking team

34

# Event extraction

- Events scored for recall (NaCTeM)

| Counted element | Count |
|---|---|
| Required events retrieved | 106 |
| Required events not retrieved | 48 |
| Required events partially retrieved | 0 |
| Scored events (required + optional events retrieved) | 148 |

Recall = 106/148 = **0.72**

Top ranking team

35

# Event extraction

- Events scored for accuracy (NaCTeM)

| Counted element | Count |
|---|---|
| Correct events matched, including arguments | 95 |
| Events with wrong arguments | 5 |
| Partial events | 15 |
| Duplicate events | 24 |
| Out-of-scope events | 14 |

Precision = (95 + (0.5*15)) / 115 = **0.89**

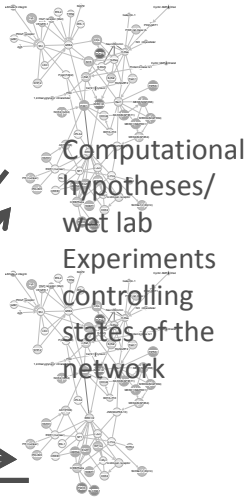
Top ranking team

36

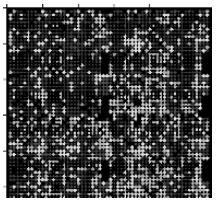
## Big Mechanism: Reading-Assembly-Explanation

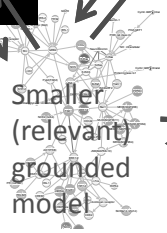
Reading Assembly Explanation












A. Rzhetsky 37

## 話の流れ

- 人工知能研究センターの紹介
- 人工知能と科学研究 （AI for Big Sciences）
- Big Mechanismと言語理解
- おわりに

38

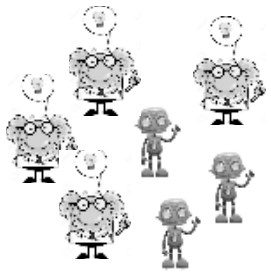
# オープンサイエンス

39

# 計算

**Open Access Journal/Open Data/Pathway Commons/Unique ID/Linked Data**

実験・観察



モデル・理論

データ解析

シミュレーション

**Machine Learning/Large Scale Search**

人工知能

機械学習 ⇒ 計算モデル
探索 ⇒ 仮説の構築
ロボット ⇒ 仮説の実証

40

# 第二部
# 産業界と産業日本語の交わり

# 産学連携における学術用語の一般化

大学から産業界への技術移転を行う際、アカデミアにおける学術用語を技術用語に変換しただけでは、円滑なコミュニケーションは行えない。大学の技術を事業化するのは産業界であり、その意思決定を行うのは、必ずしもその技術に精通した人であるとは限らないからである。また、昨今は海外企業への産学連携も増加傾向に有り、日本語のみのコミュニケーションでは十分とは言えない。これを外国語に翻訳し、誰もが理解できるフォーマットの開発も求められる。本講演では、産学連携の現場で求められる言語について言及したい。

山本 貴史

株式会社東京大学ＴＬＯ 代表取締役社長

# 産学連携における学術用語の一般化

東京大学 TLO　代表取締役社長　山本　貴史

大学における発明を知的財産権にし、産業界への技術移転（ライセンス）を行うのが TLO や知財本部の役割である。各大学において、産学連携の機運は高まっているが、各大学が独自のフォーマットを用い産業界への技術移転活動を行っているので、大学によってフォーマットも違うし技術の記述も異なる。そもそも産学連携活動を行うには、学術用語を技術用語に翻訳しただけでは不十分であり、技術のバックグランドを持たない経営者にも分かり易く伝えることが重要である。もし、大学における技術移転活動でフォーマットが統一され、用いる用語がある程度統一されれば自動翻訳等を用いて海外産学連携活動もより活性化されるであろう。本講演では、上記のように産学連携活動における学術用語の一般化と今後の可能性について言及する。

# 日本企業の一層のグローバル化と産業日本語

世界の GNI は 1990 年比で 3.4 倍に成長しているが、日本の成長は 1.5 倍程にとどまっている。また、1990 年代前半の日本の GNI の世界 GNI に占める割合は 17%前後であったが、現在はその 1/3 の約 6%に低下している。日本の GNI 成長には、日本企業の一層のグローバル化が必要となるが、そこには多くの課題が存在する。課題解決に当たり、まずは言語の障壁を取り除く必要があるが、産業日本語の浸透や機械翻訳の普及により多くの課題が解決し、日本企業のグローバル化が加速するものと考える。

井上 二三夫

シスメックス株式会社

研究開発企画本部 副本部長兼知的財産部長

／一般社団法人 日本知的財産協会 副理事長

# 第7回産業日本語研究会シンポジウム

## ～日本企業の一層のグローバル化と産業日本語～

2016年2月29日　東京・丸ビルホール

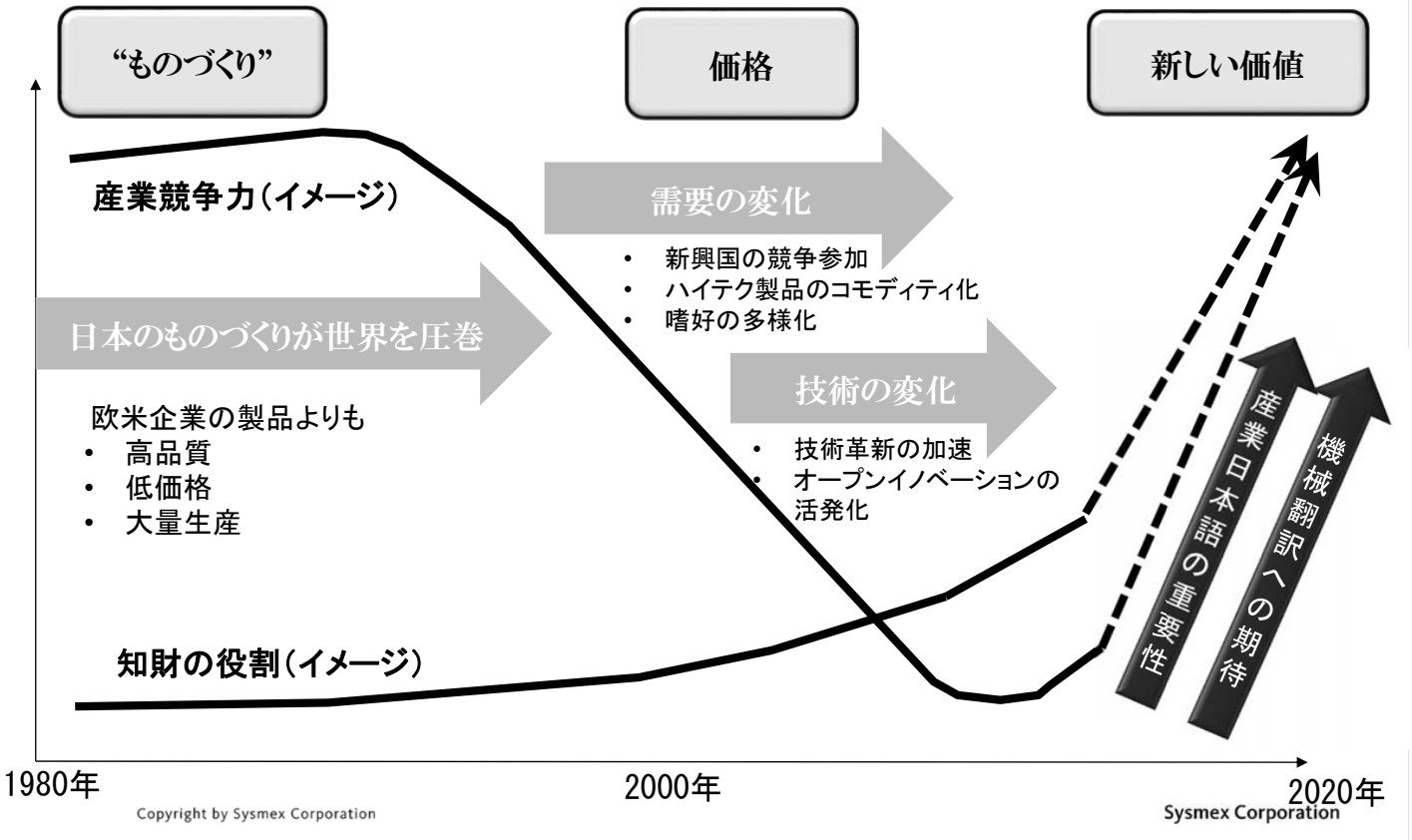

シスメックス株式会社
研究開発企画本部
副本部長兼知的財産部長

一般社団法人
日本知的財産協会
副理事長

井上二三夫


Sysmex Corporation

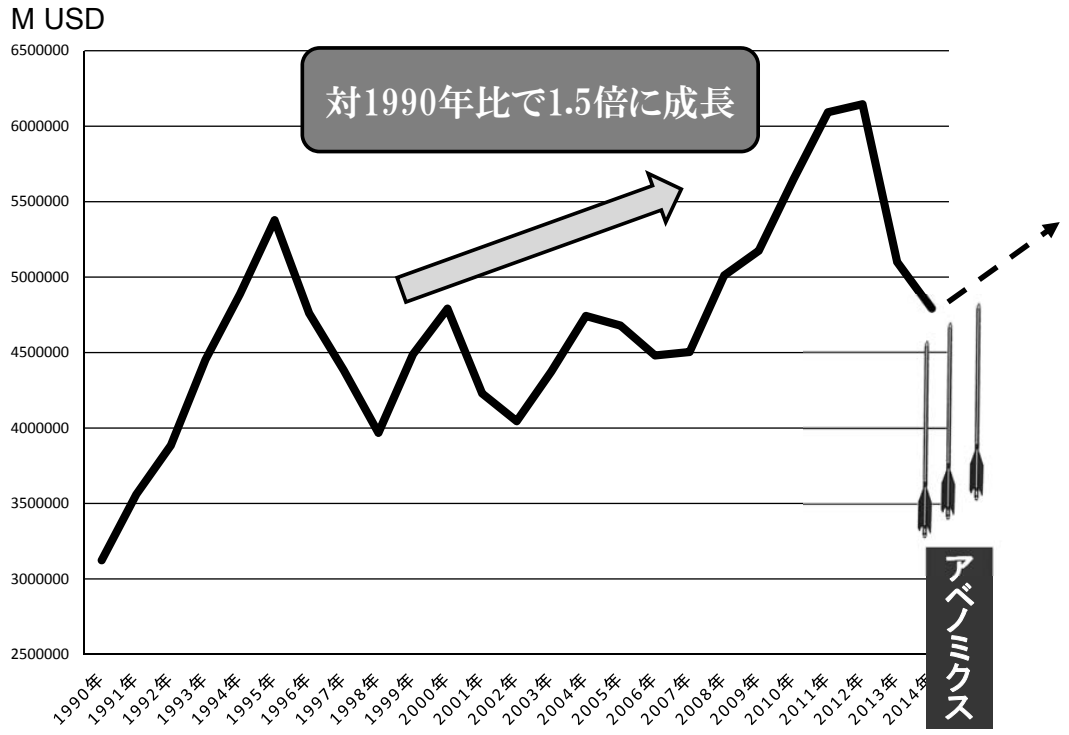
# 競争環境の変化






Sysmex Corporation

# 日本の名目GDPの推移

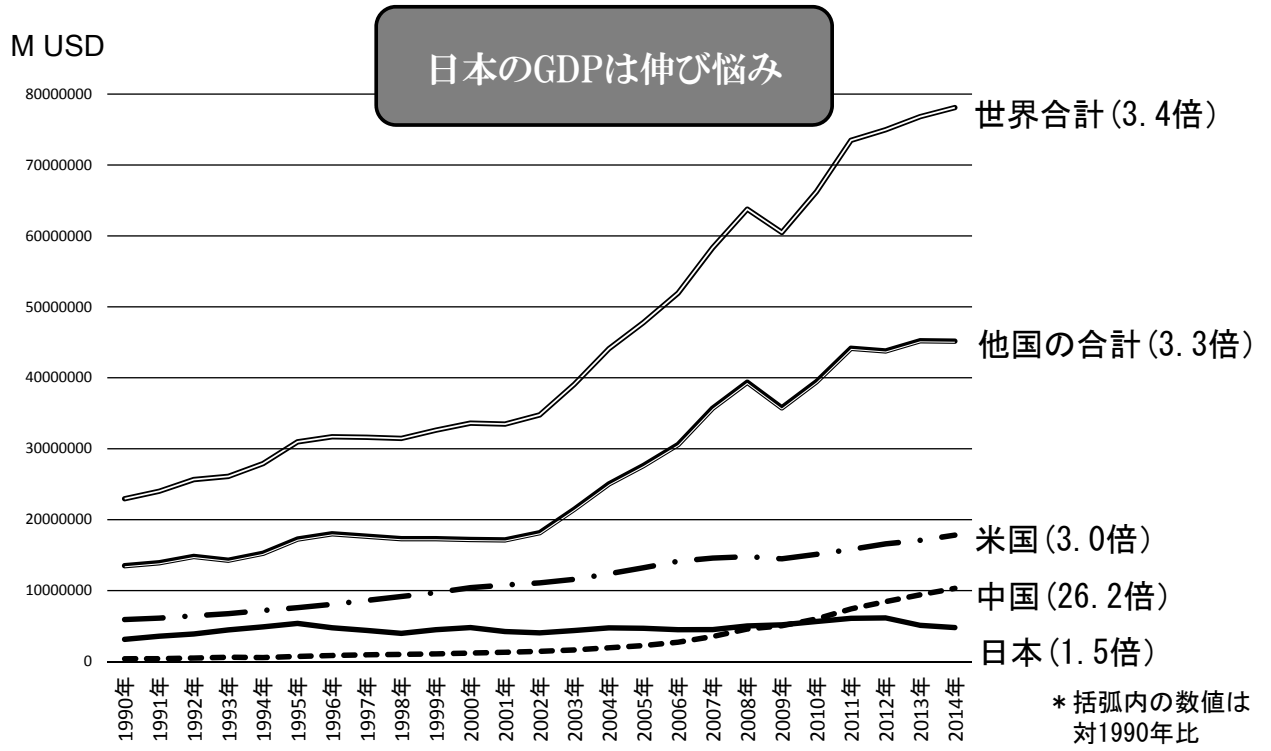



＊出典:国連のデータに基づいて作成



Sysmex Corporation

# 各国の名目GDPの推移

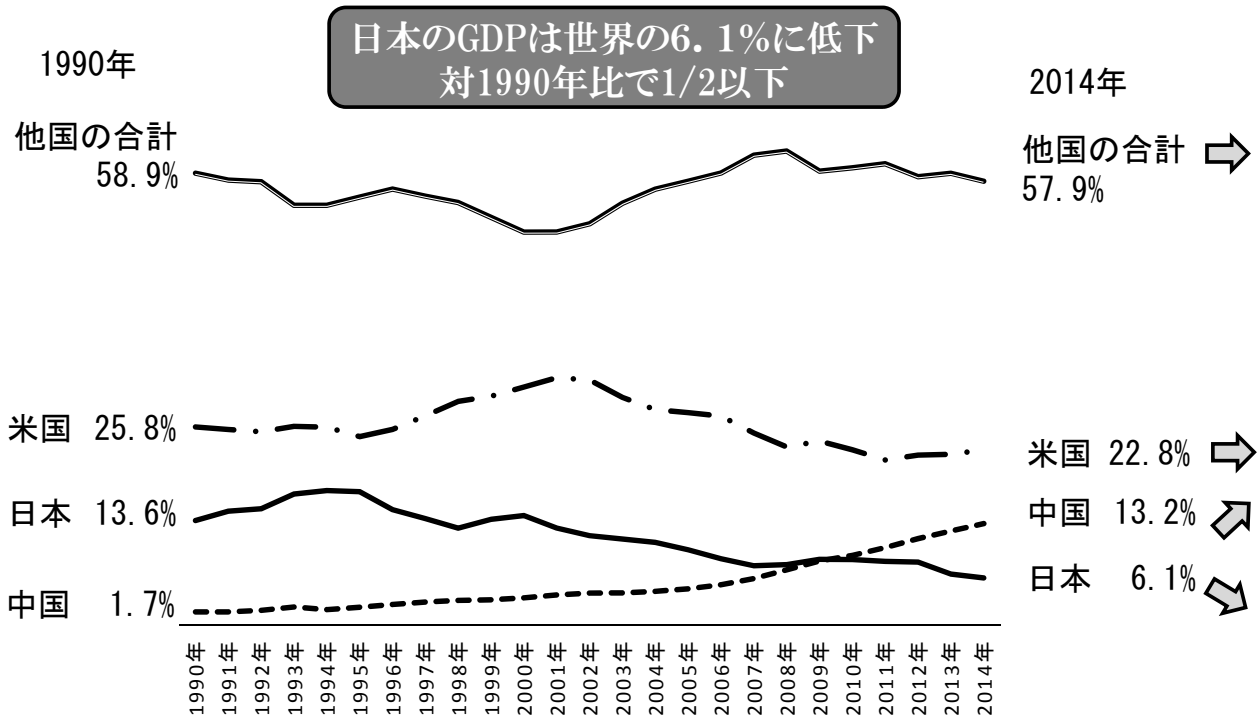



＊出典:国連のデータに基づいて作成



Sysmex Corporation

# 各国の名目GDPの推移(相対値)




＊出典:国連のデータに基づいて作成



Sysmex Corporation

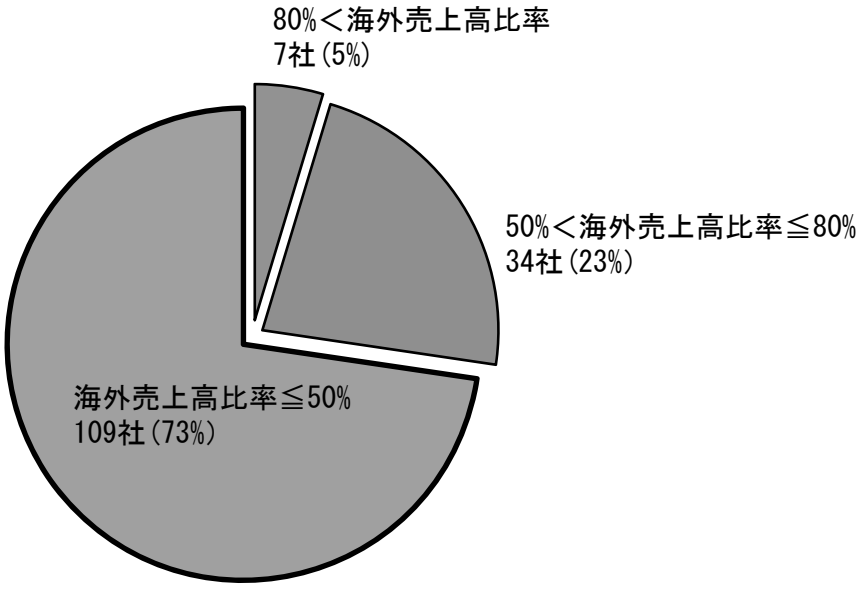
# 日本企業の海外での売上

**直近の<u>連結売上が1兆円</u>を超える日本企業(150社)の海外売上高比率**(注1)(注2)

大半の企業が国内依存

↓

**海外展開が課題**




(注1)東証一部上場企業の直近の連結売上げ (出典Yahooファイナンス)
(注2)海外売上比率は、Yahooファイナンスに記載の数値を使用



Sysmex Corporation

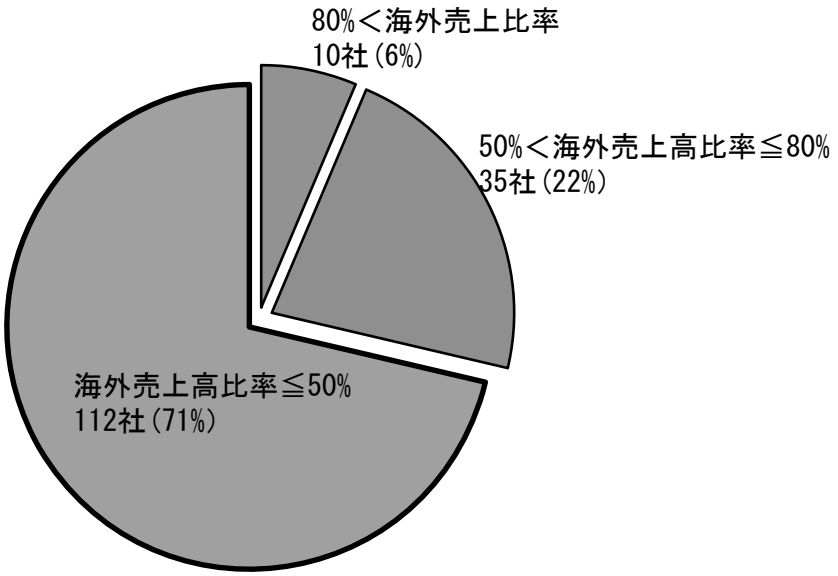
# 日本企業の海外での売上

**株式時価総額が1兆円を超える日本企業(122社)の海外売上比率(注1)(注2)**

大半の企業が国内依存

↓

**海外展開が課題**




(注1)株式時価総額は2016年1月4日東京株式市場の終値により算出 (出典Yahooファイナンス)
(注2)海外売上比率は、Yahooファイナンスに記載の数値を使用



Sysmex Corporation




Sysmex Corporation

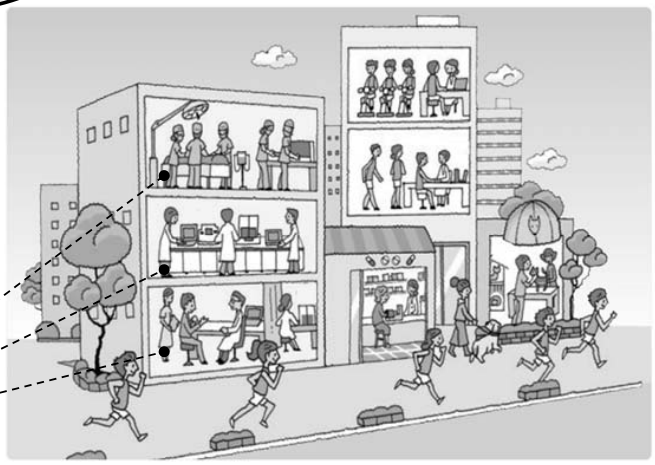
# グローバル化事例（シスメックスの例）

## 事業概要

| 400mL ・ 成分献血者 | | | | |
|---|---|---|---|---|
| 検査項目 | | 標準値 | 今回の値 | 前回の値 |
| 血球計数検査 | 赤血球数 RBC | 男性 435~575<br>女性 375~505 ×10⁴/μL | 426 | 468 |
| | ヘモグロビン量 Hb | 男性13.7~17.6<br>女性11.2~15.2 | | |
| | ヘマトクリット値 Ht | 男性38.9~49.0 %<br>女性34.3~44.8 % | | |
| | 平均赤血球容積 MCV | 80.0~100.0 fL | 93.4 | 91.8 |
| | 平均赤血球ヘモグロビン量 MCH | 26.0~34.0 pg | 30.5 | 29.6 |
| | 平均赤血球ヘモグロビン濃度 MCHC | 32.0~36.0 % | 32.7 | 32.3 |
| | 白血球数 WBC | 35~100×10²/μL | 57 | 90 |
| | 血小板数 PLT | 14.0~38.0×10⁴/μL | 21.9 | 22.9 |

見覚えありませんか？
血液検査の結果表です

**健康診断や病院などで行う**
**血液検査、尿検査に必要な機器や試薬、**
**ソフトウェアを開発し、製造･販売しています。**



検査機器

専用試薬

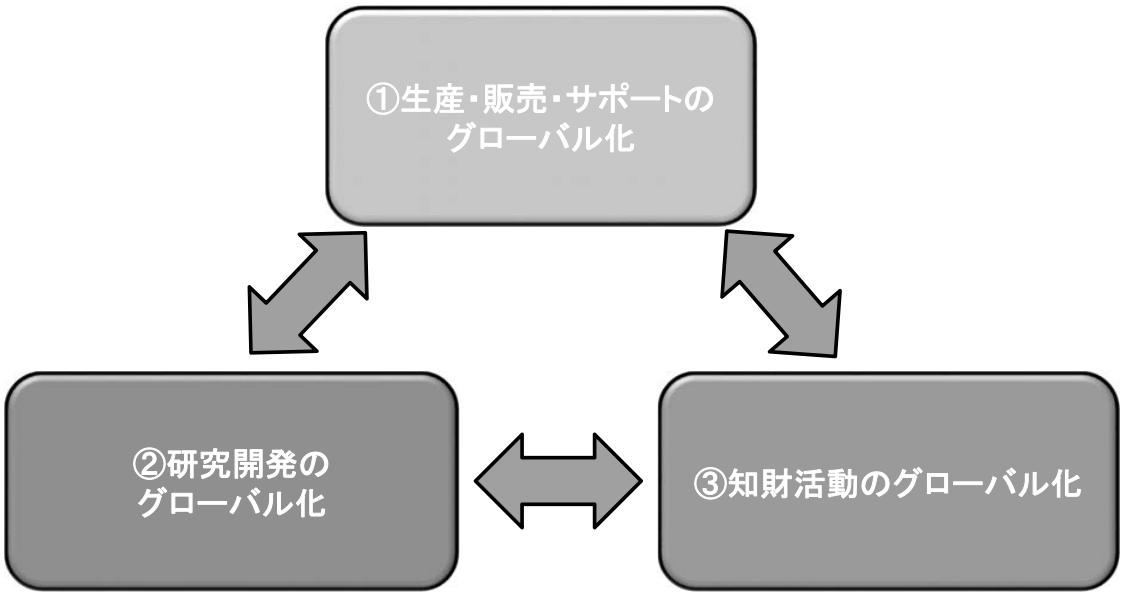
「検査」を通じてみなさまの健康をサポートしています



Sysmex Corporation

## 2. グローバル化（海外展開）推進施策

**3つの施策を同時並行で推進**






Sysmex Corporation

## ①生産・販売・サポートのグローバル化

**生産・販売・サポートのグローバル化**

- 海外企業とのアライアンス強化
- 海外での直販化推進
- 海外での生産力増強

**英語によるコミュニケーションが増大**

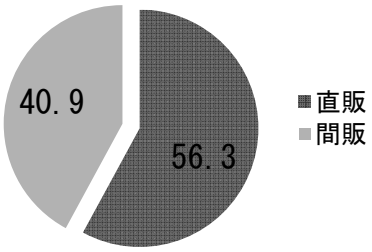
→ Win-Winの関係構築
自社の弱みをパートナーの強みで補完
パートナーの弱みを自社の強みで補完

→ 現地顧客との直接コミュニケーション

→ 材料の現地調達・人材の現地採用

**190カ国以上で販売**

2015年度　直販・間販比率








**190カ国以上で販売**

→ 30～40言語での
取扱説明書（300ページ400ページ）を用意
→ 膨大な翻訳費用

→ 製品は少量多品種
→ 1冊あたりの翻訳コスト大

## ②研究開発のグローバル化

**研究開発のグローバル化**

- 本社地区開発拠点のグローバル化 → 全世界から有能な人材を採用
- 海外R&D拠点強化 → 現地の有能な人材を現地採用
- グローバルなオープンイノベーション推進 → 世界トップレベルの研究者・研究機関とのコラボ

**日・米・独・仏・中 ニュージーランドに開発拠点**

本社によるグローバルR&D活動の統制
技術戦略の共有

**英語によるコミュニケーションが増大**





- 技術戦略書
- 技術企画書
- 商品戦略書
- 商品企画書
- 進捗レポート
- レビュー書類
- 各種提案書類
- ・
- ・
- ・

英語を
共通言語
（基本言語）

日常的に
大量の翻訳作業が発生

## ③知財活動のグローバル化

**知財活動のグローバル化**

- グローバルFTO → グローバルな知財調査・分析・鑑定
  190カ国以上のお客様に安心してお使いいただくため
- グローバル出願 → グローバルな競争優位性担保のため
- グローバルIP契約・紛争処理 → グローバルな事業活動の推進
  ユーザー・事業の保護のため

**187カ国に出願**

**グローバルIPネットワークの構築**

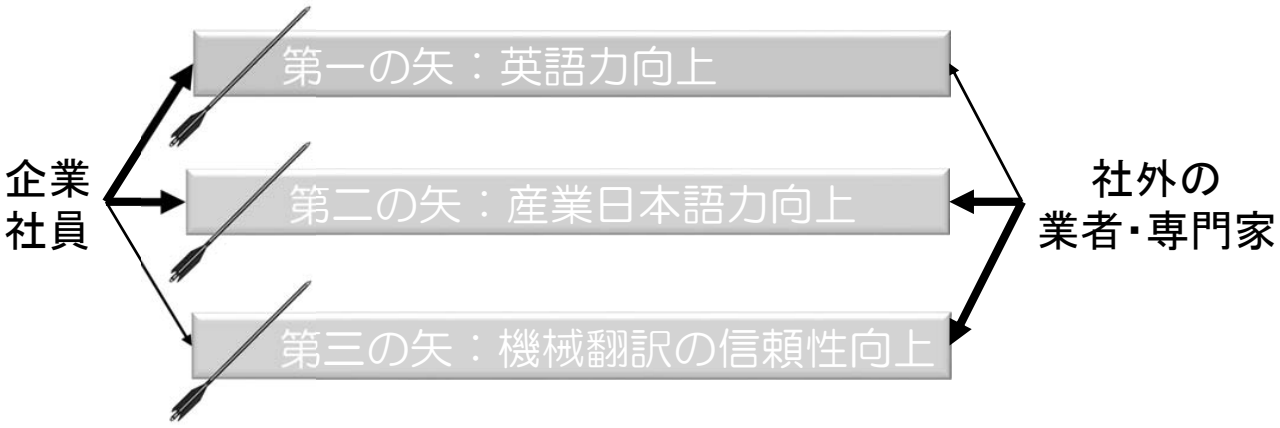
**英語によるコミュニケーションが増大**



Sysmex Corporation

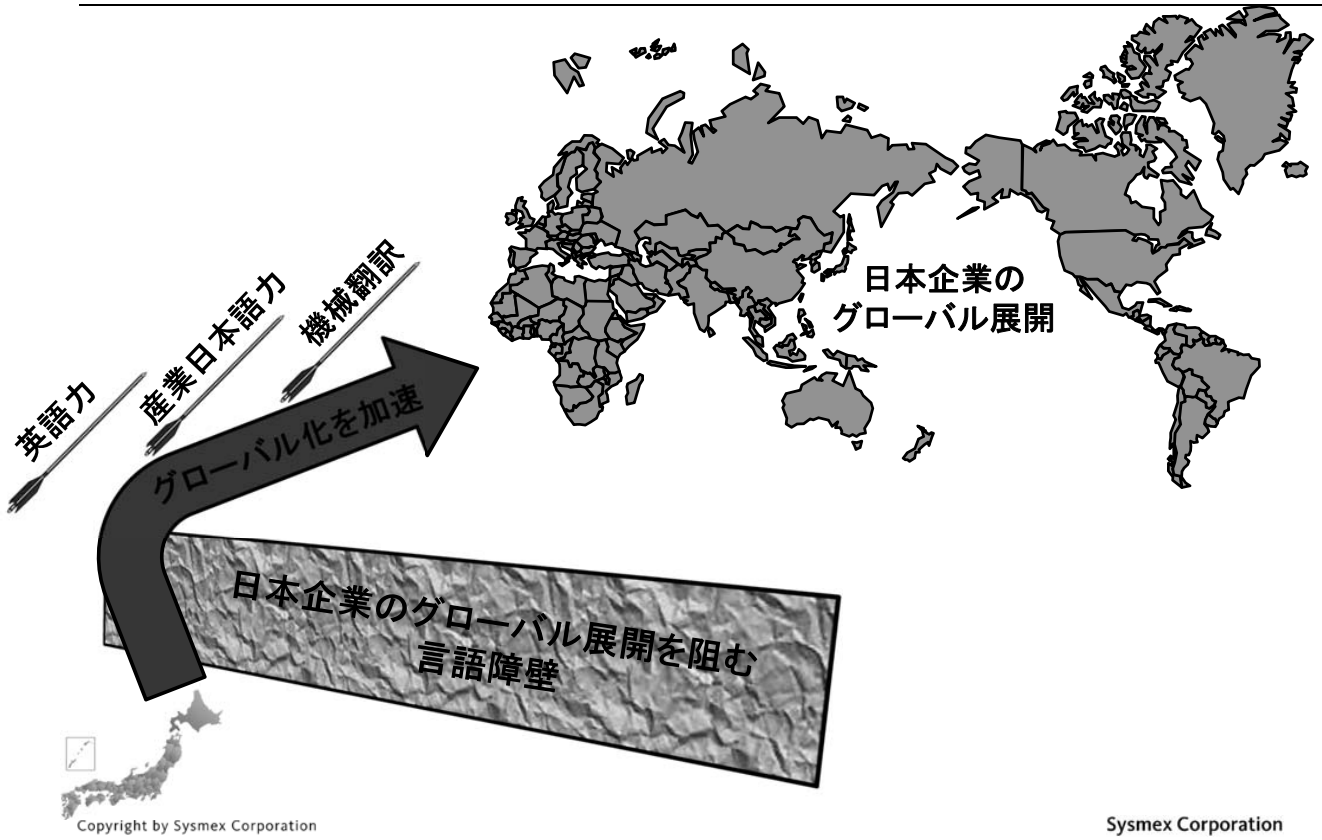
## 産業日本語・機械翻訳への期待

**日本企業の一層のグローバル展開に「英語」は必須**






Sysmex Corporation




ご清聴、ありがとうございました。

Thank you for your attention.
感谢您的关注。
感謝您的關注。
경청 해 주셔서 감사 했습니다.
Vielen Dank für Ihre Aufmerksamkeit.
Σας ευχαριστώ για την προσοχή σας .
Gracias por su atención .
Nous vous remercions de votre attention .
Grazie per l' attenzione .
Obrigado por sua atenção.
Tack för er uppmärksamhet .
Takk for oppmerksomheten .
Tak for din opmærksomhed.
İlginiz için teşekkür ederiz.
Dziękuję za uwagę.
Благодарим Вас за внимание .
Köszönöm a figyelmet .
Děkuji vám za pozornost .
Благодаря ви за вниманието .
Tänan teid tähelepanu eest .
Hvala vam na pažnji .
Dėkojame už jūsų dėmesį.
Paldies par jūsu uzmanību.
Ďakujem vám za pozornosť .
Vă mulțumesc pentru atenție .
Хвала на пажњи .
Hvala za vašo pozornost .
Kiitos huomiota .
Dank u voor uw aandacht .
شكرا لكم على اهتمامكم .
با تشکر از توجه شما.
شکریہ . آپ کی توجہ کے لئے آپ کا
Terima kasih atas perhatian Anda .

**井上二三夫**
Inoue.Fumio@sysmex.co.jp

# 研究会報告
# 「日本人のための日本語マニュアル」のご紹介

産業日本語研究会の活動として「日本語マニュアルの会」が制作した「日本人のための日本語マニュアル」を紹介する。本マニュアルは、日本人ビジネスマンを対象にしたビジネス文章ライティングマニュアルである。詳細は【第五部】ポスターセッションへお運びいただきたい。日本語へのアプローチとしては、日本語マニュアルが次世代知的テキスト処理とビックデータとを結ぶものとなることが期待される。

横井　俊夫

一般財団法人 日本特許情報機構 特許情報研究所　顧問

／学校法人片柳学園東京工科大学 名誉教授

# 「日本人のための日本語マニュアル」の紹介

横井俊夫
一般財団法人日本特許情報機構特許情報研究所顧問
東京工科大学名誉教授

目次：

1. 日本語マニュアルの制作
2. 日本語マニュアルの特徴
3. 日本語マニュアルの構成
4. 日本語マニュアルの内容
   4.1 文書・文章ライティングのモデルプロセス
   4.2 共通例文
   4.3 「伝える日本語」、「訳せる日本語」、「機械が訳せる日本語」が実現してくれること
5. 日本語マニュアルと人工知能・次世代言語処理

---

## 1. 日本語マニュアルの制作

『日本人ビジネスマンのためのビジネス文章ライティング日本語マニュアル
- 言葉の仕組を学び、外国語との対照を通じて日本語スキルを磨く -』

制作グループ：日本語マニュアルの会
横井俊夫（Japio 特許情報研究所顧問、東京工科大学名誉教授）
石崎 俊（慶応大学名誉教授、一般財団法人 SFC フォーラム理事）
佐野 洋（東京外国語大学教授）
石黒 圭（国立国語研究所准教授、一橋大学連携教授）
猪野真理枝（翻訳家、語学教材制作家）
烏日哲（一橋大学国際教育センター非常勤講師）

## 2. 日本語マニュアルの特徴

本マニュアルは、次の諸点において従来からの数多ある文章作成術関連書とは大きく異なる。

① ライティング全体のモデルプロセスに沿っている
② ライティングルールが言葉の仕組に裏付けられている
③ 他言語への翻訳プロセスが組み込まれている
④ 校正支援や機械翻訳などのテキスト処理ソフトの利用が組み込まれている。

## 3. 日本語マニュアルの構成

はじめに
1 章　文書・文章ライティングのモデルプロセスを学ぶ
　1.1　言葉の役割
　1.2　ビジネス文書とビジネス文章
　1.3　文章特性がライティングを特徴付ける
　1.4　ライティングのモデルプロセス
　1.5　マニュアルの対象範囲と利用手順
2 章　情報を表わし伝える言葉の仕組みを学ぶ - 日本語と外国語とを照らし合わす -
　2.1　文章技術のための言葉の仕組
　2.2　事例から学ぶ
　2.3　仕組を学ぶ
　2.4　言葉の仕組とライティングルール
3 章　「表す日本語」で書き、「伝える日本語」へと言い換える
　3.1　「表わす日本語」と「伝える日本語」の役割
　3.2　「表わす日本語」で書く - 日本語パラグラフライティング
　3.3　「伝える日本語」への言い換えルール
4 章　「訳せる日本語」へ言い換える
　4.1　「訳せる日本語」の役割
　4.2　「訳せる日本語」への言い換えルール
5 章　コンピュータの支援機能を活用する - 文章校正ソフトと機械翻訳ソフト
　5.1「伝える日本語」への言い換えを支援する文章校正ソフト
　5.2「訳せる日本語」の翻訳を支援する機械翻訳ソフト
おわりに

## 4. 日本語マニュアルの内容

## 4.1 文書・文章ライティングのモデルプロセス

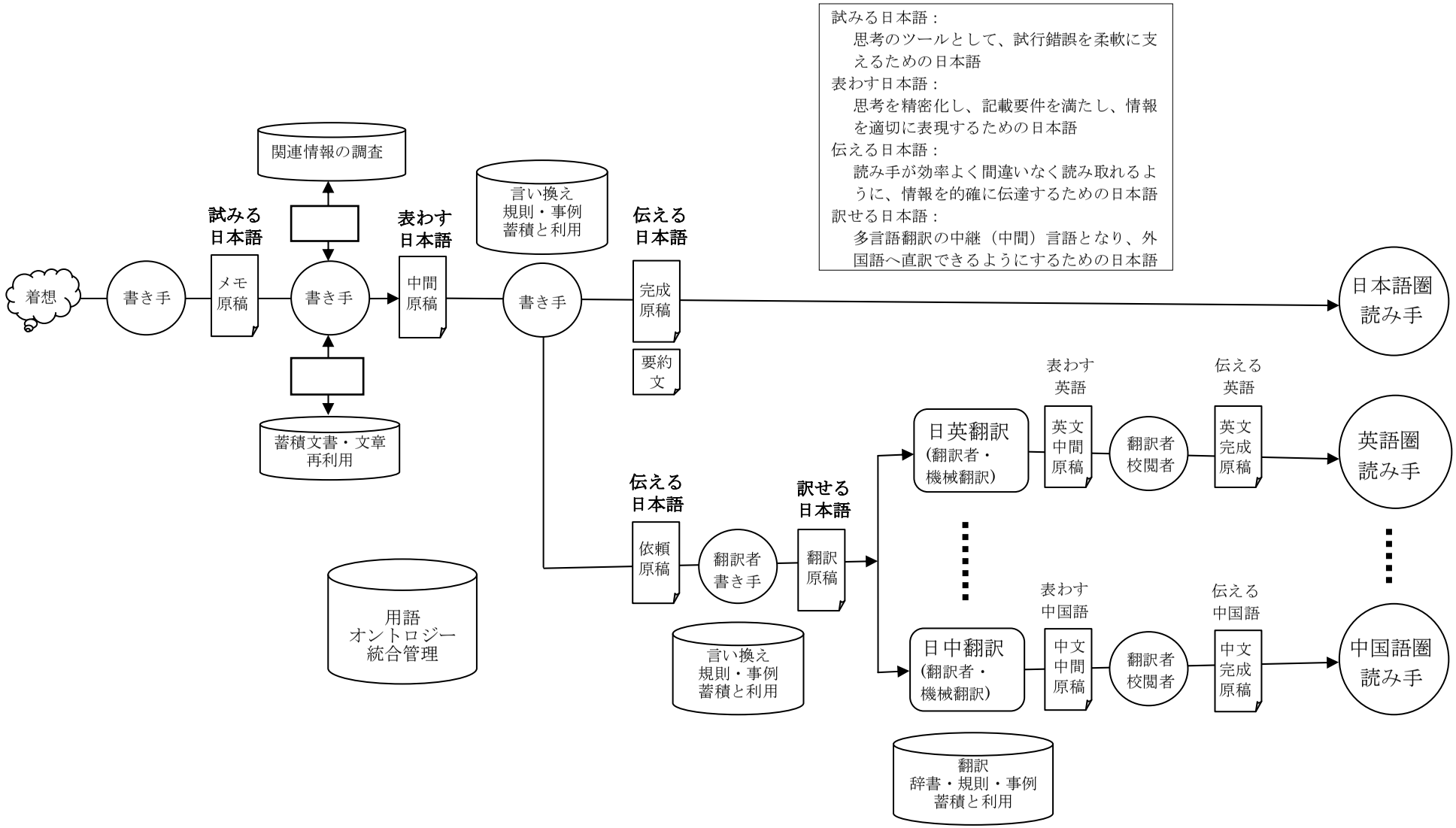



図 1.3　文書・文章ライティングのモデルプロセス

## 4.2 共通例文

ビジネス文章の実態を把握することは難しい。ビジネス文書の多くは、各企業・各組織の機密情報に属する。本マニュアルがどのようなビジネス文章を対象にしているのか、その想定を分かり易く示すのが共通例文である。

### ［共通例文 1］単純な「事柄」を表現する

【日本語】――

昨日電子メールで友人が新聞社に原稿を送った。

――――――――――

【英語】――

My friend sent the newspaper publisher a manuscript by e-mail yesterday.

――――――――――

【中国語】――

朋友昨天用电子邮件寄了原稿給报社。（朋友昨天用電子郵件寄了原稿給報社。）

注）カッコ内は繁体字

――――――――――

### ［共通例文 2］複雑な「事柄」を表現する

【日本語】――

日本人は、外国語が苦手なので、昨日の新聞に掲載された記事のように、その苦手の克服のために努力しなければならないことを議論してきた。

――――――――――

【英語】――

Since Japanese are weak at foreign languages, they have been debating what they have to try in order to overcome this weakness, as in an article that was published in yesterday's newspaper.

――――――――――

【中国語】――

因为日本人不擅长外語, 所以像在昨天的报紙上刊登的报道一样, 他们一直在讨论为了克服这个弱项不得不努力的事。

――――――――――

### ［共通例文 3］「小状況」を伝える

【日本語】――

日本人は、外国語が苦手である。その苦手を克服するために努力しなければならないこと

を様々な観点から議論してきた。議論の中には、日本語が世界の言語の中でも際立って特殊な言語であり、本来、日本人には、外国語は向かないという悲観論を説くものもあった。確かに、その悲観論も、欧米の諸言語の有様を調べてみると頷けないことはない。しかし、本稿は、すべての国々の人々にとって、自国語以外の言語を習得する困難さには、さしたる違いはないという観点に立つ。その観点から、苦手とする外国語を習得するために日本人が心得なければならない要点を 3 点挙げる。

――――――――

【英語】――

――――――――

【中国語】――

――――――――

## ［共通例文 4］「状況」を構成する

【日本語】――

日本人は、外国語が苦手である。その苦手を克服するために努力しなければならないことを様々な観点から議論してきた。議論の中には、日本語が世界の言語の中でも際立って特殊な言語であり、本来、日本人には、外国語は向かないという悲観論を説くものもあった。確かに、その悲観論も、欧米の諸言語の有様を調べてみると、頷けないことはない。しかし、本稿は、すべての国々の人々にとって、自国語以外の言語を習得する困難さには、さしたる違いはないという観点に立つ。その観点から、苦手とする外国語を習得するために日本人が心得なければならない要点を 3 点挙げる。

第 1 に、外国語を使うことが強く求められる環境に身を置くべきである。そのような環境が得られるなら、積極的にその機会を生かし、最大限に活用する努力をすべきである。必死になって努力せざるを得ない環境こそが、最良の教師となる。

第 2 に、思考言語としての日本語の能力を鍛えるべきである。どんなに外国語が上達しても、日本人にとっては、日本語が思考言語であり、外国語はコミュニケーション言語に止まる。日本語も外国語も同じように使いこなせるようになる能力も機会も、多くの日本人には無縁である。言語能力の土台となる日本語による思考能力を鍛えることが重要である。

第 3 に、自信を持って外国語で発信できるしっかりとした自分独自の考えや主張を持つべきである。自分の考えに自信が持てるなら、たとえ稚拙な外国語能力でも臆することなく主張することができる。一方、自信が持てない考えを言葉巧みに説明し相手を納得させるには、高度な外国語能力が必要になる。

以上の 3 点は、外国語によって表現したい内容が相当に高度な場合の要点である。日常生活や簡単な情報交換における軽いコミュニケーションのための外国語能力に関しては、もう少し異なった心得が必要になる。

――――――――

【英語】――
――――――――
【中国語】――
――――――――

## 4.3 「伝える日本語」「訳せる日本語」「機械が訳せる日本語」が実現すること

ライティングのモデルプロセスの中の外国語への翻訳プロセス部分に焦点を当ててみましょう。この翻訳プロセス部分は、この日本語マニュアルのアピールポイントとなります。「伝える日本語」に「訳せる日本語」と「機械が訳せる日本語」の仕組を加える翻訳プロセスによって、翻訳という作業に新たな機能を加えることができるようになります。その新たな機能を、市販されている機械翻訳ソフトの新たな活用方法という側面から取り上げましょう。

新たな活用方法とは、機械翻訳ソフトに次の 2 つの機能を付加できるようする試みです。

（1） 安定した翻訳結果が得られるようにする

（2） 文脈処理を含むより高度な翻訳に対応できるようにする

以下に、例文を用いて、その有様を詳しく説明します。使用する例文は、共通例文３の先頭の 2 文です。市販されている機械翻訳ソフトとしては、東芝ソリューション社の The 翻訳プロフェショナル V15（日英）を使いました。今回の利用体験を通じて、適切な利用方法を設定すれば、日本の機械翻訳ソフトが十分に実用性の高い翻訳ツールに成長してきていることを実感いたしました。

### ［文 1］

**伝える日本語：**

「日本人は、外国語が苦手である。」

The 翻訳：

“Japanese people are weak at a foreign language.”

**訳せる日本語：**

「日本人が外国語に苦手である。」

:「伝える日本語」に（訳せる文 - 1 - 3）を適用し、二重主語文を主語ひとつ文に言い換える。

翻訳者：

“Japanese people are weak at foreign languages.”

The 翻訳：

“<想定外の翻訳結果>”

注)「苦手だ」に二格が共起するとは解釈しないことと、ハガ構文の処理を優先するがための副作用が原因と思われる。興味深い現象であるが、詳細は割愛する。

**機械が訳せる日本語：**

「日本人が外国語に苦手である。」

:「訳せる日本語」のまま

———————

［文 1］は、「～は、～が苦手である」という二重主語文です。よく見かける簡単な文パターンです。したがって、「伝える日本語」から「訳せる日本語」への言い換えが、翻訳者にとっては当然ですが、The 翻訳にとっても不要であることがみてとれます。当然のことながら、「訳せる日本語」から「機械が訳せる日本語」へ言い換える必要もありません。

「訳せる日本語」では、外国語（英語）に直訳できない日本語特有の表現を直訳できるように言い換えます。実は、機械翻訳ソフトも日本語特有の表現をそのまま翻訳できる機能を備えています。わざわざ、「訳せる日本語」や「機械が訳せる日本語」に言い換える必要はありません。

The 翻訳も、「日本人は、外国語が苦手である。」は、翻訳できました。それでは、「外国語は、日本人が苦手である。」を翻訳させてみましょう。翻訳結果は、“A foreign language are weak at Japanese people.”となります。これは、おかしいですね。<「日本人」が主体的に行為を行うもの>であり、<「外国語」は行為の対象となるもの>であるという意味的な処理を行わず、単に、「A は B が苦手である」を“A is weak at B”に対応付けているのではないかと思われます。

機械翻訳ソフトが「訳せる日本語」への言い換えルールを完全にカバーしている、あるいは、カバーしている範囲が明解なら、その言い換えルールの適用をその範囲内でスキップしてもかまいません。ただし、そうでないのなら、「訳せる日本語」への言い換えを画一的に適用した方が、安定した翻訳結果が得られることになります。「外国語は、日本人が苦手である。」を「訳せる日本語」に言い換えると、「外国語に関しては、日本人がそれに苦手である。」あるいは、「外国語は、日本人の苦手に含まれる。」等となります。

「訳せる日本語」と「機械が訳せる日本語」の役割は、現在の機械翻訳ソフトに次の 2 つの機能をプラスできるようにすることです。

（1） 安定した翻訳結果が得られるようにする

（2） 文脈処理を含むより高度な翻訳に対応できるようにする

（1）についての説明は終えたとして、（2）について説明しましょう。簡単な例を使っ

て説明しましょう。以下の例文です。

「象は、鼻が長い。」

この例文は、日本語特有の構文である二重主語文やハガ構文の例文として、言語学でもよく取上げられる例文です。意外に、ビジネス文章にもよく使用される構文です。例えば、以下のような例もあります。

「金属球は、表面がきれいに磨かれている。」

「我が社は、今年の目標が新興国における増産である。」

「象は、鼻が長い。」という文の意味は、<象の鼻は長い>という簡単な「事柄」とその「事柄」を伝える際に<「象」に焦点を当て前提とする>との２つの事項から成り立っています。「象は、鼻が長い。」と「鼻は、象が長い。」の 2 つの文は、「事柄」に関しては、<象の鼻は長い>という同じことを表現していますが、伝える際に「象」を前提とするのか、「鼻」を前提とするのかが異なっています。

「象は、鼻が長い。」と「鼻は、象が長い。」を The 翻訳で翻訳してみましょう。いずれも、“An elephant’s trunk is long.”と翻訳されます。伝える際の前提の違いは、反映されません。機械翻訳ソフトは、文を一文ずつ、文を単独の文として翻訳します。文脈を考慮することはしません。伝える際の前提の違いは、多分に文脈にも関わる処理となります。The 翻訳の翻訳結果は、出来るだけ文脈にニュートラルな翻訳を心掛けた結果です。

しかしながら、翻訳の品質を上げるためには、文脈を考慮した処理が必要になります。以下の文例 A と B、そして、A’と B’をみて下さい。いずれも、4 つの文から成り立っています。

A :「ここには、多くの動物がいる。動物について特徴的な身体部位を挙げてみる。麒麟は、首が長い。象は、鼻が長い。」

B :「ここには、多くの動物がいる。身体部位について特徴的な動物を挙げてみる。首は、麒麟が長い。鼻は、象が長い。」

A’:「ここには、多くの動物がいる。動物について特徴的な身体部位を挙げてみる。麒麟の首が長い。象の鼻が長い。」

B’:「ここには、多くの動物がいる。身体部位について特徴的な動物を挙げてみる。麒麟の首が長い。象の鼻が長い。」

A と B は、2 番目の文の違いに対応して、3 番目と 4 番目の文が前提とするものが違うことを反映しています。一方、A’と B’は、2 番目の文の違いに対応せず、3 番目と 4 番目の文が文脈にニュートラルなものとなっています。明らかに、4 つの文の連鎖として納まりが良いのは、A と B でしょう。

A、B を The 翻訳で翻訳すると、A’、B’を英訳したのと同じ結果が得られるだけです。

内容的に間違っているわけではありませんが、訳文として上質であるとはいえません。「訳せる日本語」、「機械が訳せる日本語」と経由する翻訳プロセスによって、The 翻訳が A、B に対して文脈を考慮した翻訳ができるようになることを示しましょう。

A、B は、「伝える日本語」で書かれています。まず、「訳せる日本語」に言い換えます。以下の A-訳、B-訳です。

A-訳「ここには、多くの動物がいる。動物について、我々は特徴的な身体部位を挙げてみる。麒麟が長い首を持つ。象が長い鼻を持つ。」

B-訳「ここには、多くの動物がいる。身体部位について、我々は特徴的な動物を挙げてみる。首ついては、麒麟の首が長い。鼻については、象の鼻が長い。」

「訳せる日本語」への言い換え操作は以下のようになります。

① 2 番目の文が主語なし文であるため、主語あり文に言い換える
（訳せる文 - 1 - 1）を適用

② 3 番目と 4 番目の文が二重主語文であるため、主語ひとつ文に言い換える
（訳せる文 - 1 - 3）を適用

主語ひとつ文への言い換え操作は、ここでは、大きく、次の 2 つのいずれかとなります。

ⓐ「X は、Y が長い」を「X が長い Y を持つ」と言い換える。日本語では、主題（「X は」）と題述（「Y が長い」）で表現されているものを、英語に対応できる主語（「X が」）と述部（「長い Y を持つ」）に言い換える。

ⓑ「X は、Y が長い」を「X に関しては、Y のそれが長い」と言い換える。日本語では、主題（「X は」）と題述（「Y が長い」）で表現されているものを、英語に対応できるテーマ（「X に関しては」）とコメント（「Y のそれが長い」）に言い換える。

注）
伝える際の前提を表現する方法は、日本語では主題、英語では主語です。いずれも、文の先頭部分に置かれます。また、いずれの言語においても、場合によって、テーマ化する表現が使われます。

A-訳、B-訳を The 翻訳で翻訳しますと、以下のようになります。「訳せる日本語」と「機械が訳せる日本語」という仕組を利用することによって、The 翻訳は、文脈に渡る処理を行う上質な翻訳を行えるようになります。

A-訳 “Here, there are many animals. About an animal, we mention a characteristic body part. A giraffe has a long neck. An elephant has a long trunk.”

B-訳 “Here, there are many animals. About a body part, we mention a characteristic animal. About a neck, a giraffe’s neck is long. About a trunk, an elephant’s

trunk is long.”

なお、もとの A、B をそのまま The 翻訳で翻訳しますと、以下のようになります。ただし、2 番目の文に主語を補う言い換えだけは適用してあります。

A “Here, there are many animals. About an animal, we mention a characteristic body part. A giraffe's neck is long. An elephant's trunk is long.”

B “Here, there are many animals. About a body part, we mention a characteristic animal. A giraffe's neck is long. An elephant's trunk is long. ”

## [文 2]

**伝える日本語：**

「その苦手を克服するために努力しなければならないことを様々な観点から議論してきた。」

The 翻訳：

“It has debated about that it must try hard in order to overcome the weakness from various viewpoints.”

**訳せる日本語：**

「日本人は、その苦手を克服するために努力しなければならないことを様々な観点から議論してきた。」

:「伝える日本語」に（訳せる文 - 1 - 1）を適用し、主語なし文を主語あり文に言い換える。

翻訳者：

“They have been debating what they have to try in order to overcome this weakness from various standpoints.”

The 翻訳：

“Japanese people have debated about that it must try hard in order to overcome the weakness from various viewpoints.”

**機械が訳せる日本語：**

「日本人は、その苦手を克服するために日本人が努力しなければならないことを様々な観点から議論してきた。」

:「訳せる日本語」に（機械が訳せる日本語 - 2）を適用し、省略要素を明示化するよう言い換える。

The 翻訳：

“Japanese people have debated about that Japanese people must try hard in order to overcome the weakness from various viewpoints.”

さらに、

「日本人は、事柄を様々な観点から議論してきた。その事柄とは、その苦手を克服するために日本人が努力しなければならないことである。」

：さらに、（機械が訳せる日本語 - 1）を適用し、短文化を進め、構文構造を簡明に把握できるようにする。

“Japanese people have debated about a matter from various viewpoints. The matter is that Japanese people must try hard in order to overcome the weakness.”

———————

「伝える日本語」での［文 2］の The 翻訳の翻訳結果の検討から始めましょう。「伝える日本語」の段階での［文 2］には、主語がありません。私たちが日常書く文には、主語のない文が普通です。文脈から容易に判断できる成分は、省略するのが、ごく当たり前の日本語の用法だからです。The 翻訳は、ゼロ代名詞を“it”（三人称モノ代名詞）で置き換えるというディフォルトのルールに従って、主語なし文の翻訳を行っています。

次に「訳せる日本語」に言い換えます。ルール（訳せる文 - 1 - 1）を適用し、主語なし文を主語あり文に言い換えます。この言い換えルールは、人である翻訳者のためですので、主語を主題成分として文頭に置くことによって、主語あり文へと言い換えています。この「訳せる日本語」に対する翻訳者の翻訳と The 翻訳の翻訳とを比べてみましょう。次の 3 点です。

（1） 主語の翻訳

翻訳者は、文脈からの情報を利用して、主語を代名詞（“they”）に翻訳しています。一般に機械翻訳ソフトは、文脈を考慮せずに、文それぞれを独立のものとして翻訳します。The 翻訳も、主語を対訳辞書の訳語どおりに翻訳しています。ただし、他に問題点があります。主語は、主節の主語となる位置に追加されています。主題成分化されていますので、翻訳者であれば、この主語が従属節の主語ともなることは、容易に推論できます。しかし、The 翻訳には、その推論能力はありません。そこで、従属節の主語に関しては、ディフォルトルールに従って、“it”で対応しています。

（2） 「こと」の翻訳

「～努力しなければならないこと」という従属節を、翻訳者は、what 節に、The 翻訳は、that 節に翻訳しています。ここは、what 節に翻訳すべきでしょう。「～ならないということ」であれば、that 節となります。

（3） 「てきた」の翻訳

「議論してきた」のアスペクト属性を翻訳者は、現在完了進行形に、The 翻訳は現在完了形に翻訳しています。この場合は、現在完了進行形の方が適切です。しかし、多くの場合、「てくる＋た」は、現在完了形や過去形に対応付けられます。

以上の比較検討の中で、翻訳者であれば容易に推論できることにも、機械翻訳ソフトに対しては、手厚い配慮が必要であることが明らかになりました。その配慮を「機械が訳せる日本語」として取り上げます。この例文の場合は、従属節の省略主語の明示化です。厳密に言いますと、この例文の従属節は、二段構えです。「日本人が努力しなければならないこと」という従属節に、「(日本人が) その苦手を克服するために」という従属節が従属する形式となっています。多くの場合、末端の従属節は、準動詞句、この例の場合は、不定詞句に翻訳することができますので、省略主語の明示化は不要になります。

The 翻訳の翻訳結果では、補充された主語の訳は、対訳辞書の訳語で置き換えた形式となっています。このような場合、英語では、代名詞に置き換えるのが通常の翻訳となります。「機械が訳せる日本語」の段階で、主語を「日本人」ではなく「彼ら」に言い換えることによって対応することになります。ただし、このような文脈に関わる事項は、文章全体を見通してから判断できることです。そこで、文章全体を「機械が訳せる日本語」に言い換え、「機械が訳せる日本語」文章として、文脈の一貫性や整合性をチェックすることになります。このようにして、「訳せる日本語」と「機械が訳せる日本語」という仕組を経ることによって、The 翻訳は、文脈処理にも対応できるようになります。

この［文 2］は、対格（を格）成分が長く複雑です。長く複雑な成分を持つ文の解析は、機械翻訳ソフトは苦手です。機械翻訳ソフトの構文解析をより確実なものにするために、文を短文化し、構文構造を簡明化する言い換えを「機械が訳せる日本語」として行うこともできます。

## 5. 日本語マニュアルと人工知能・次世代言語処理

（1） 日本語マニュアルは、人工知能技術や次世代言語処理技術を駆使したソフトウェアシステムとして実現されるべきである。

（2） 日本語マニュアルによって、ブラックボックス AI をインタラクティブ AI へと進化させることができる。

（3） 日本語マニュアルによって、言語関連ビックデータをより良質なビックデータへと改善し、ビックデータの資産価値をたかめることができる。

# 第三部

## 人工知能研究と言語処理研究の最前線

# 特別講演
# 質問応答タスクと自然言語処理

TV クイズ番組に挑戦する質問応答システム Watson の開発において自然言語処理は極めて重要な役割を果たした。これに解候補や根拠の検索と、解のスコアリングのための機械学習を組み合わせたタスク設計が効果的であった。一方で、主に言語資源からなる情報源と学習データを確保しつつ人手によるオントロジー等の作成を回避できたことも大きな要因であった。本講演では特にこのコンテンツ面に注目したプロジェクトへの寄与について検討する。

武田 浩一

日本アイ・ビー・エム株式会社

東京基礎研究所 技術理事

# 質問応答タスクと自然言語処理

武田浩一

日本アイ・ビー・エム株式会社　東京基礎研究所

**1. 質問応答システム Watson の処理方式**

2011 年 2 月の米国 TV クイズ番組 Jeopardy!への挑戦において、いわゆる質問応答技術[1]が open-domain のトリビア質問に対して極めて高い解答精度を達成して以来、同技術の実用化が注目を集めている。IBM 研究者たちが設計開発した質問応答システム Watson[2, 3]は、図 1 に示すように 4 つの大きなステップで与えられた質問の解答を計算する方式を採用した。

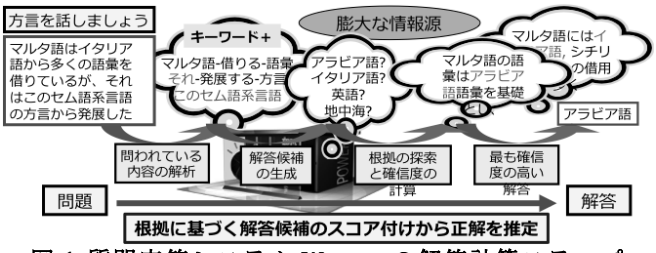



**図 1. 質問応答システム Watson の解答計算ステップ**

ステップ 1 では質問文分析による内容語および述語項構造の(これらを「手掛かり表現」と呼ぶ)抽出を行い、ステップ 2 では手掛かり表現を検索式に変換して解答候補を取得する。ステップ 3 では得られた各解答候補を手掛かり表現中の解答指示部分に代入し、個別の仮説を生成し、これらの仮説を再度検索式に変換して「根拠」と呼ばれるパッセージ集合を検索する。最後にステップ 4 では各解答候補ごとに仮説と根拠との整合性等を数値化して、確信度と呼ばれる解答候補のスコアを計算する。事前に設定された閾値以上の確信度を持つ解答候補が見つかれば、そのうちで最大の確信度の解答候補を回答する。これらの 4 つのステップはいわゆるランタイムの質問応答処理部分に相当し、それ以外に学習データとなる過去の質問および正解・不正解のペアから解答候補をスコア付けするロジスティック回帰分析の重みの学習や、情報源の前処理による情報抽出および索引付け等を行うためのバッチ処理が適用される。

**2. 既存コンテンツの活用**

Watson の実装において既存コンテンツの活用は極めて重要な成功要因であった。情報源として WordNet の辞書資源、Wikipedia, World Book Encyclopedia, Gutenberg プロジェクトの電子化された書籍など多様なデジタルテキスト、および DBPedia, Yago 等の構造化データベースを採用し、学習データとしては Jeopardy!の過去問約 25,000 問を正解・不正解の解答候補に対応づけた約 570 万対のデータが使われた。IBM 自身が大規模なコーパスや言語資源を全く構築することなく、これら既存コンテンツのみを用いて対戦時に約 88%の正答率を実現できたことは注目に価する。研究者を多大な時間と労力を必要とするコンテンツ開発からシステム構築に集約することで開発期間を短縮化できただけではなく、医療その他の分野に適合させる際の大きな知見を獲得することができたといえる。1999 年の TREC-8[4]でタスク定義されて以来多数の質問応答システムが提案されてきたが、今後は大規模オープンデータなどの言語資源を機械学習により最大限活用したデータ中心システムが実用化への最有力候補となるであろう。

一方でほぼすべての既存コンテンツは質問応答タスクを想定して作成されたものではないため、必ずしもコ

ンテンツの表現形式やその構造が質問応答タスクに適合せず、回答精度向上に直結しないという問題が発生しうる。たとえば Watson の実装においては、以下のような基準で情報源の取捨選択や変形を行い、コンテンツの可用性や再利用性を高めた。これらはすべてヒューリスティックな処理ではあるが、人手による介在をできるだけ避けたスケーラブルな作業であり、成熟した自然言語処理なしでは実現不可能であった。

- **Source Acquisition:** ランダムに選択した質問から主題の広がり(範囲)や解答に必要な情報の詳細度を検討し、エラー分析を通して、追加･削除すべき情報源を決定した。
- **Source Transformation:** 文学作品、聖書、歌詞などを Wikipedia のような見出し語つき文書形式に変換(作者＋有名な引用など)する。相互参照される(同義的な)見出しの文書はマージする。見出し語が省略された辞書形式の定義文には見出し語を補って完全な文形式に変換した。
- **Source Expansion:** 認知度の高い(被参照リンク数や出現頻度)固有表現の記述を Web 情報で補足し、質問と定義文の表現のゆらぎによる不一致の可能性を減少させた。各見出し語に対し約 100Web ページを取得し、関連度の高い表現部分を文書に追加した。

**3. 産業日本語への展開**

Watson プロジェクトの経験から産業日本語への貢献としては、質問応答に代表される Watson 要素技術の日本語対応と、利用可能な日本語言語資源およびその加工のための知見が含まれる。

前者については、機械翻訳の時代から蓄積された国内の自然言語処理技術のレベルが十分に高いため、日本語質問文や情報源の前処理および情報抽出に必要な自然言語処理は既に実現していると考えてよい。分野に応じた解答候補の評価のための属性(次元)やそのスコア計算はその都度設計開発する必要があるが、実装上の技術的な問題はないといえる。

後者についてはたとえば英語および日本語の Wikipedia の差異、WordNet や MEDLINE 文献抄録のように極めて利便性の高いコンテンツの有無が問題になっている。

産業日本語の大きな応用分野として、機械翻訳や言い換えの技術が注目されてきたが、Watson の商用化に関わってきた 5 年間の経験からいえば、質問応答や意思決定支援、商品推薦、医薬品や新製品開発における探索的知見獲得など、多くの産業界で高付加価値ソリューションへの需要が目立つ。データ中心のシステム実現には大規模な言語資源およびデータベースが不可欠なため、これらの応用のなかでも特許情報などの共用性の高い日本語データが存在する分野から具体的なソリューションが誕生するものと期待している。コールセンターやインターネット・サービス事業者のもとでも大規模な専有データが存在するが、業界全体で共有することは難しい。画像投稿と同様の気軽さで、テキストデータが新規に発生し、様々なメタデータが付与され再利用可能となるサービスを創出できれば、こういった技術を予想以上にタイムリに商業化できるのではないかと考えている。

**参考文献**

[1] 磯崎秀樹、東中竜一郎、永田昌明、加藤恒昭，(監修：奥村学): 質問応答システム， コロナ社 (2009).

[2] 金山 博、武田浩一：Watson: クイズ番組に挑戦する質問応答システム、情報処理 、Vol.52, No.7, pp. 840-849, 2011 年 7 月

[3] “This is Watson”, IBM Journal of Research and Development, Vol.56, Issue 3-4, 2012

[4] Voorhees, Ellen M.: The TREC-8 Question Answering Track Report, Proc. of TREC-8, 1999

# リクルートＡＩ研究所から見る
# 言語処理研究への期待

人材領域のビジネスを言語処理的な枠組みから表現すると、レジュメと呼ばれる非構造の自然言語で記述された求職者サイドのデータと、職務記述書と呼ばれる同じく非構造の自然言語で記述された企業によって提供される雇用者サイドのデータをマッチングするビジネスと表現できる。本講演では、リクルート AI 研究所が上記のような実問題を解くにあたって人工知能と言語処理に期待する役割について紹介する。

石山 洸

株式会社リクルートホールディングス

R&D 本部 Recruit Institute of Technology 室長

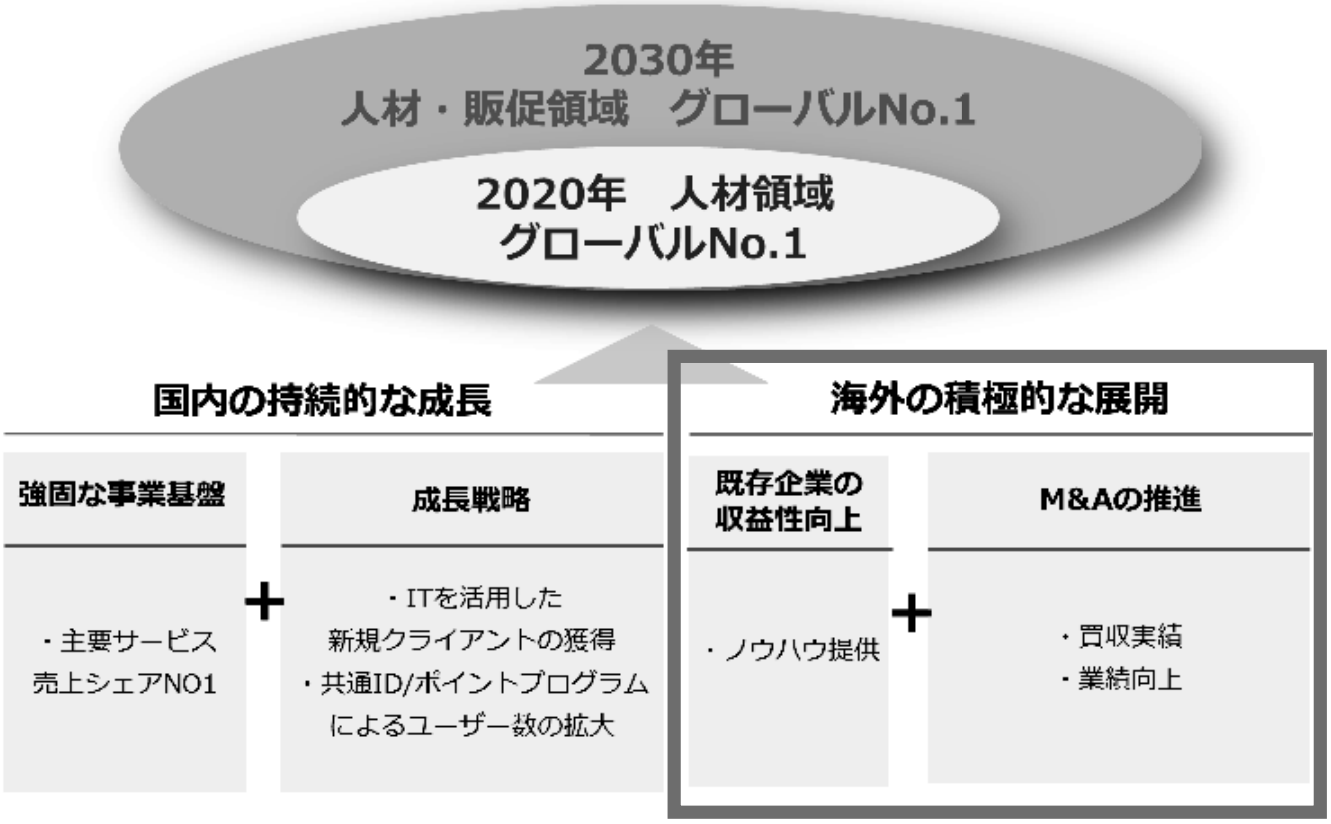
# リクルートAI研究所から見る
# 言語処理研究への期待

株式会社リクルートホールディングス
Recruit Institute of Technology 推進室 室長
石山 洸

## はじめに
## リクルートホールディングスのご紹介

| | |
|---|---|
| 社名 | 株式会社リクルートホールディングス |
| 創業 | 1960年3月31日（設立 1963年8月26日）<br>2012年10月1日「株式会社リクルート」より商号変更 |
| 本社所在地 | 〒100-6640　東京都千代田区丸の内1-9-2 |
| グループ従業員数 | 30,016名（2014年7月31日時点） |
| グループ会社数 | 119社（連結対象子会社、2014年7月31日時点） |
| 資本金 | 100億円（2014年10月15日より） |
| 連結売上高 | 11,915億67百万円（2013年4月1日～2014年3月31日） |
| 連結経常利益 | 1,220億50百万円（2013年4月1日～2014年3月31日） |

# リクルートの目指す姿




# 2020年に人材領域でグローバルNo.1になるための破壊的技術としてのAI開発

| | 人材紹介事業 | 人材派遣事業 | オンラインHR事業 |
|---|---|---|---|
| ブランド | RGF HR Agent<br>WORK IN JAPAN<br>伯乐 Bó Lè Associates, Ltd. | The CS Companies<br>Staffmark<br>people bank<br>CHANDLER MACLEOD<br>Advantage Resourcing | indeed<br>and search all jobs. |
| エリア | **アジアを中心に**<br>**11の国と地域28の都市**<br>日本、中国、香港、台湾、シンガポール、インド、ベトナム、インドネシア、フィリピン、マレーシア、タイ | **先進国を中心に**<br>**7か国約450拠点**<br>アメリカ・イギリス・オーストラリア・香港・シンガポール・インドネシア・マレーシア | **先進国を中心に**<br>**世界55カ国以上、28言語** |
| サービス | 1．日系現地法人向けの人材紹介<br>2、日本企業向けの外国人新卒の紹介<br>3．海外現地・外資系企業向けの人材紹介 | 海外現地での人材派遣 | アグリゲート型の求人専門検索エンジンの運営 |
| 特徴 | **人材紹介業**<br>**日系アジアNo.1** | **世界　5位** | **世界最大級のユーザー数** |

# AI研究所の設立背景

グローバルトップレベルの技術水準

Disruptiveな新規ビジネスへの活用

# 「グローバル x 新規ビジネス」を実現する研究チーム

Recruit Institute of Technology
エグゼクティブ・ディレクター

ALON HALVEY
h-index 94、バイアウト２回

アドバイザー・リサーチャー派遣先

OREN ETZIONI

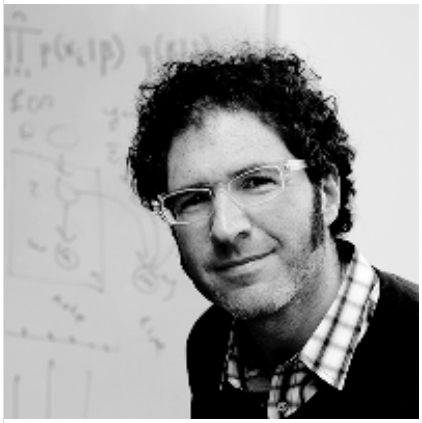
CHRISTOPHER D. MANNING

TOM M. MITCHELL

DAVID M. BLEI

ALEX‘SANDY’PENTLAND

## 言語処理チーム: David M. Blei教授



David M. Blei教授
(米コロンビア大学)

機械学習の理論研究者であり、膨大なデータの中からパターンを見つけるための代表的な機械学習手法であるトピックモデルの第一人者。幅広い分野で応用されているトピックモデル手法Latent Dirichlet Allocation (LDA)の考案者のひとり。ACM Infosys-Foundation Award受賞(2014)

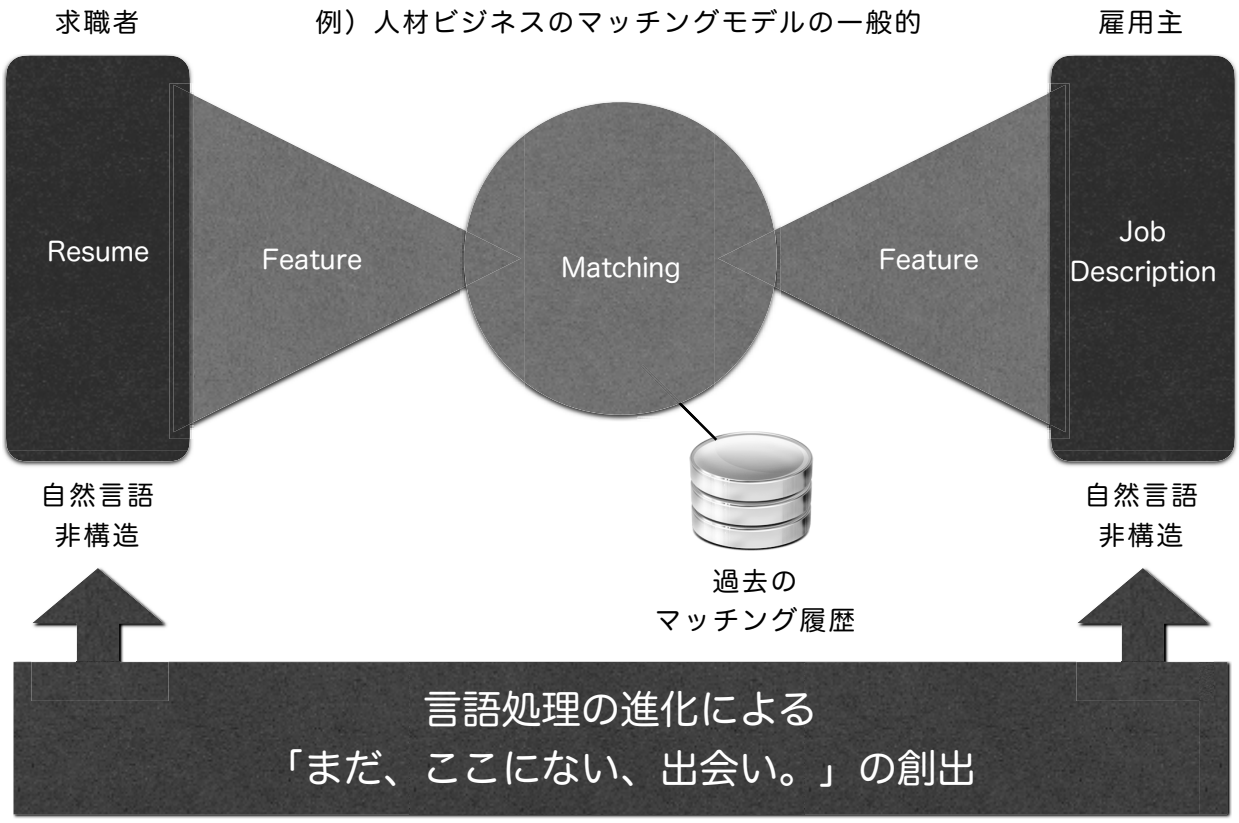


## 言語処理チーム: Christopher D. Manning教授

Christopher D. Manning教授
(米スタンフォード大学)

自然言語処理、情報検索分野における著名な研究者。自然言語処理の代表的な教科書である“Foundations of Statistical Natural Language Processing”、情報検索分野の代表的な教科書である“Introduction to Information Retrieval”の著者。ACM、AAAI、ACLフェロー。

# 言語処理への期待


求職者
例）人材ビジネスのマッチングモデルの一般的
雇用主
Resume
Feature
Matching
Feature
Job Description
自然言語
非構造
過去の
マッチング履歴
自然言語
非構造
言語処理の進化による
「まだ、ここにない、出会い。」の創出

# ニュースアプリケーションにおける自然言語処理

「スマートニュース」はスマートフォン専用のニュースアグリゲータである。アプリケーションを起動するとニュース記事などが表示されるだけであるが、その裏側ではインターネット上から膨大なニュースを収集、分析、選定し、アプリケーションへと配信している。本講演では、このスマートニュースの裏側の記事の分析作業に使われる自然言語処理技術について紹介する。

徳永 拓之

スマートニュース株式会社 エンジニア

第７回産業日本語研究会・シンポジウム

# ニュースアプリケーションにおける自然言語処理

スマートニュース株式会社
徳永拓之

SmartNews

ニュース、サクサク。

App Store からダウンロード

App Store でダウンロード

for iPhone/iPad

ANDROID アプリ Google play

for Android



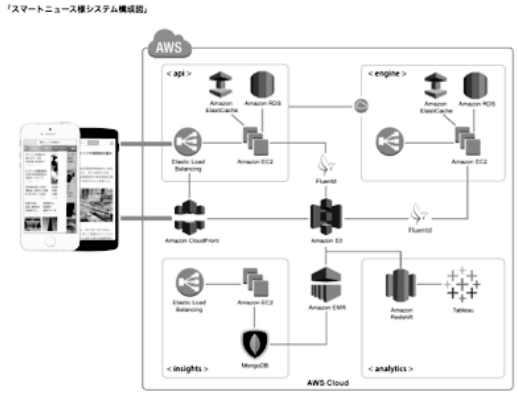
# スマートニュースのミッション


# アプリを開くと
# 記事一覧が表示される

- 見た目は非常にシンプル
- しかしその裏側では...


# その裏側では...

- 多くのサーバーが記事の解析・配信のために
  稼働している




http://aws.amazon.com/jp/solutions/case-studies/smartnews/

# スマートニュースの難しさ

- **どこ**に**どの記事**を出すのか？

- 記事を複数の媒体社様から頂いている
  - 人手で全部チェックするのは不可能
  - ソフトウェアによる自動的な解析が重要

# 記事解析において重要な技術

- 自然言語処理
- 画像処理
- ログ解析

# NLP以外の技術も活用してます

- 画像処理
  - サムネイル画像のクロッピング
  - セクシー画像の自動判定
- ログ解析
  - 記事クオリティの判定
  - アプリの改善

# 自然言語処理をどこで使うか？

- 記事カテゴリの分類
- NG記事の分類
- 関連記事の推薦
- 記事主題の抽出

# 記事カテゴリの分類

- 典型的な文書分類
  - 本文、タイトルから記事カテゴリを当てる
  - 機械学習 + ルールベース
- Paragraph Vector (Le et al., 2014) を使用

# NG記事分類

- 掲載基準を満たさない記事を分類する
- こちらも典型的な文書分類

# 関連記事の推薦

- ある記事を読んだユーザーに対して別の記事をおすすめする
- 現在はタイトルと本文の情報が類似する記事を推薦している

# 記事主題の抽出

- 記事の中心となるトピック（を表現するキーワード）を抽出する
- 曖昧性の問題が難しい
  - 「羽生」は「羽生善治」か「羽生結弦」か

# 記事解析と記事配信の関係

- 記事解析の結果は重要なシグナルとなり得る
  - 多くの記事が言及しているトピックは重要である可能性が高い
  - 他にも用途は考えられる
- どのカテゴリに配信するかも記事解析の結果で決まる

# 機械学習を使うことについて

- 記事解析システムは下記の性質が重要である
  - 高速に動作する
  - 挙動の予想がつきやすい
  - トラブル時に原因がわかりやすい

- 機械学習を使うと下の2つが満たされない事が多いが、NLPでは機械学習は重要である

# 機械学習とのつきあい方

- シンプルなケースは予めルールで処理する
  - ややこしいものだけ機械学習
- 機械学習の結果を上書きできる仕組みを用意する
- 機械学習の性質について周知が有効な場合は、周知を怠らない

# まとめ

- スマートニュースは、ニュース記事を自動的に解析し配信するプラットフォームである
- 記事解析を行う上で最も重要なパーツの1つが自然言語処理である
- 記事カテゴリ分類をはじめとして多くの部分に自然言語処理、機械学習が使われている

# 1000 社が採用
# 「見える化エンジン」が挑戦し続ける
# テキストマイニングのビジネス利用最前線

企業活動でのテキストであるアンケート自由回答やコールセンターログ、SNS データは、商品開発や品質向上の現場では VOC(Voice of Customer, 顧客の声）と呼ばれ、PDCA サイクルの重要な情報源として扱われている。このサイクルの中で、クラウド型のテキストマイニング分析ツール「見える化エンジン」は利用されている。しかしながら、顧客の声であるテキストは、アンケート回答のような書き言葉だけでなく、ツイッターで用いられる話し言葉、通話データやチャットのような対話など複雑化している。また、テキスト分析には、同じ言葉であっても業界や業種ごとに適した意味で示さなければならないという課題が常にある。経営判断の素材となるには、これらの課題の克服が必要である。本講演では、ビジネスの現場で扱われる多様性あるテキストへの、先進技術による日々の取り組みを紹介する。

坂 慎弥

株式会社プラスアルファ・コンサルティング

見える化エンジン事業部 テキストマイニング研究 G

グループマネージャー

産業日本語研究会
Technical Japanese Association

プラスアルファ コンサルティング

# 1000社が採用
# 「見える化エンジン」が挑戦し続ける
# テキストマイニングのビジネス利用最前線

見える化エンジン™

2016年2月29日

株式会社プラスアルファ・コンサルティング
坂　慎弥　saka@pa-consul.co.jp

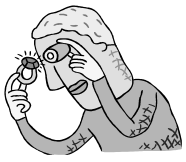


## プラスアルファは何の会社か？

ＩＴ技術×コンサルティングノウハウ

# 「お客様の声活用専門会社」


データマイニング



データの山の中から
ビジネスのヒントを発見！

# プラスアルファの特徴

テキストマイニング
国内シェアNo.1
導入：1000社
4年連続業界トップ！





強力な
言語解析技術部隊

80
業界トップを支える
テキストマイニング
専門集団




# さまざまな業界で1,000社以上の実績

消費財・家電・自動車・アパレル

Asahi Glico meiji MORINAGA LOTTE ハウス食品
SAPPORO NISSIN 日清食品 pokka sapporo DyDo エバラ 亀田製菓
MEGMILK キユーピー CALPIS カラダにピース 宝酒造 SARAYA
HONDA The Power of Dreams mizkan ブルボン マルコメ コイケヤ
HITACHI Inspire the Next mazda HINO TOTO brother 三井製糖
Combi Aprica FOODLIER
TAKARA TOMY TORAY MoonStar Wacoal
clarion hoyu KAI
MIZUNO asics adidas

金融・通信・IT・サービス・メディア・エンタメ・公共

docomo NTT東日本 NTT西日本 経済産業省 JAL
UQ WiMAX TOKAI 大阪ガス 全農 ALSOK
東武鉄道 GDO 日本郵政 Fidelity OZmall
三井住友信託銀行 新生フィナンシャル トーキョーPay Roland
大和ネクスト銀行 伊予銀行 ほけんの窓口 ZENRIN Antenna SHOFU 関西テレビ
FuRyu TIPNESS 湯快リゾート 東宝
Gakken 河合塾 Benesse
KLab 水晶院
東京大学 慶應義塾大学

ビッグデータ活用 × プレミアム会員

SUNTORY Eat Well, Live Well. SHISEIDO Nestlé SoftBank JR
AJINOMOTO KIRIN NISSAN FamilyMart KDDI
Panasonic TOSHIBA Leading Innovation
BELLE MAISON CO-OP DUSKIN NEXCO
MIZUHO BANDAI NAMCO Games SHOP CHANNEL ASKUL J:COM
TOKIO MARINE NICHIDO 損保ジャパン日本興亜 小林製薬 クリナップ YAMAHA BANDAI
MS&AD 三井住友海上 RECRUIT 大正富山医薬品株式会社
MUFG 三菱UFJニコス T&G Canon SUNSTAR ベルーナ 弥生

流通・小売・外食・通販・化粧品・製薬関係

SEVEN & i HOLDINGS イオン atré
MUJI 無印良品 CAINZ HOME scroll ディノス・セシール
ユニー KOKUBU 三菱食品 CO-OP 生活協同組合コープこうべ CO-OP
コープさっぽろ ココカラファイン シオノギ製薬 Santen
kao 第一三共株式会社 塩野香料株式会社
ENEOS ono 大日本住友製薬 Eisai KOSE
L'ORÉAL PARIS 牛乳石鹸 ROHTO Kracie janssen SANOFI

調査・マーケティング・広告代理店

PRAP JAPAN Inc. J.D.POWER AND ASSOCIATES DAIKO Kurt Salmon M3, Inc.
Tokyu Agency Inc. YOMIKO NIKKEI R KANTAR Ravigote
MACROMILL intage
dentsu HAKUHODO jeki
ADK McCANN ERICKSON

ソリューションベンダー・テレマ会社

NTT Solco DNP D&I Partners
SCSK KDDI Evolva TMJ IMJ GROUP Zation


※社名公開のご許可を頂いている企業様を一部抜粋させて頂いております。

# 今日お伝えしたいこと

## テキストマイニングは、
## ビジネスの “武器” となる

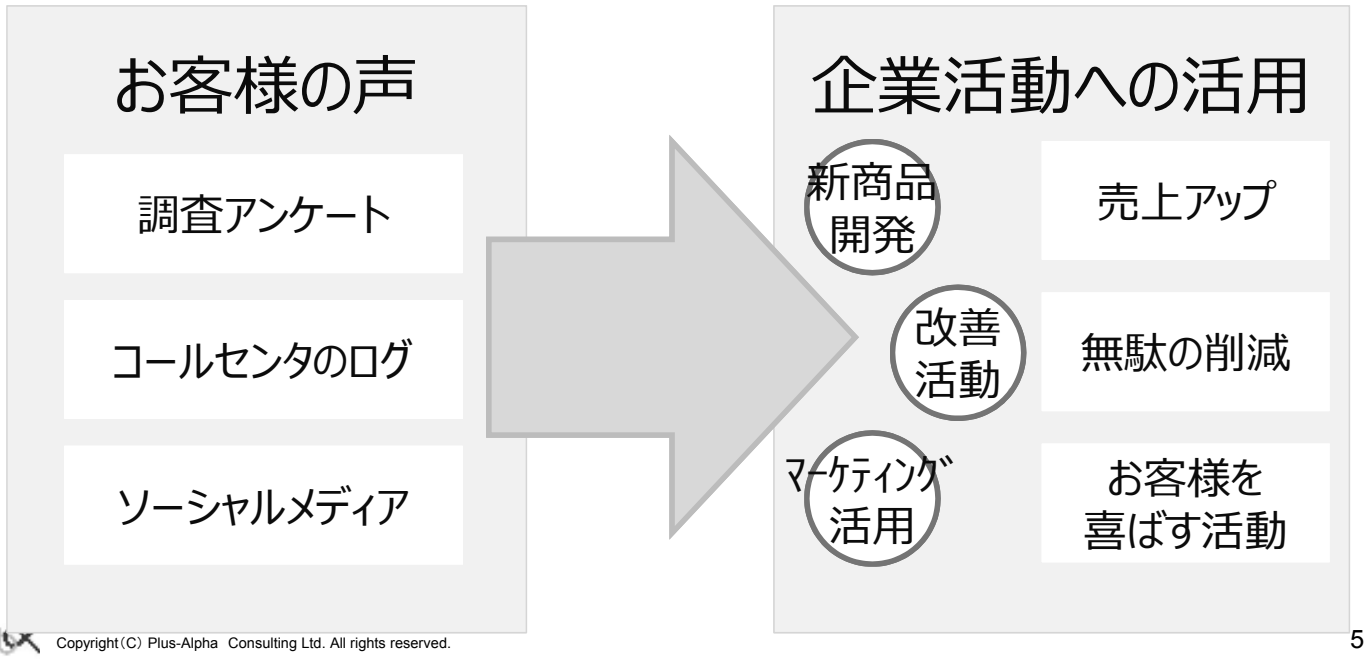
## お客様の声は、
## 多様化 し、進化している




【テキストマイニングはビジネスの武器となる】

# ビジネスにおいて進む お客様の声活用

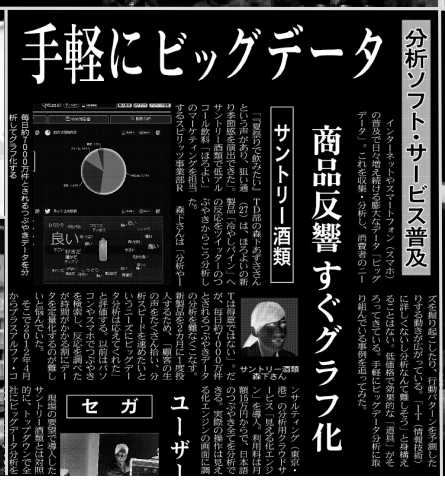
**ビジネスでは、テキストのことを**
**“お客様の声”と呼び、企業活動に活用をしている**



【テキストマイニングはビジネスの武器となる】

# お客様の声活用事例は年々増加している




【テキストマイニングはビジネスの武器となる】

# お客様の声と企業活動との“橋渡し”

## 言語処理解析 × テキストマイニング

### お客様の声

- 調査 アンケート
- コールセンタ ログ
- ソーシャル メディア

声を“解析”して、“表現”する

見える化エンジン™

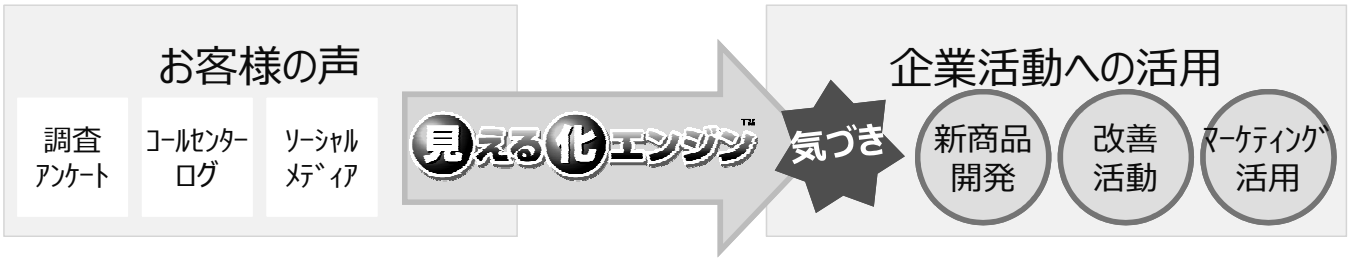
### 企業活動への活用

| | |
|---|---|
| 新商品開発 | 売上アップ |
| 改善活動 | 無駄の削減 |
| マーケティング活用 | お客様を喜ばす活動 |

【テキストマイニングはビジネスの武器となる】

# お客様の声と企業活動との“橋渡し”

## 見える化エンジン = 「気づき発見」ツール




思考のスピードを止めることのない、高速レスポンスと 直感的な表現を実現

| ①全体化から詳細へ | ②担当者から全社へ | ③希少意見の発見 |
|---|---|---|
| グラフを**クリック**すれば、必ず**原文テキスト**にたどり着く仕組み | 面白い分析結果はお気に入りとして**保存**<br>動的なレポートとして、他メンバーへの**共有** | **言語解析**技術を使った**希少**な価値ある声を抽出する仕組み |



【テキストマイニングはビジネスの武器となる】

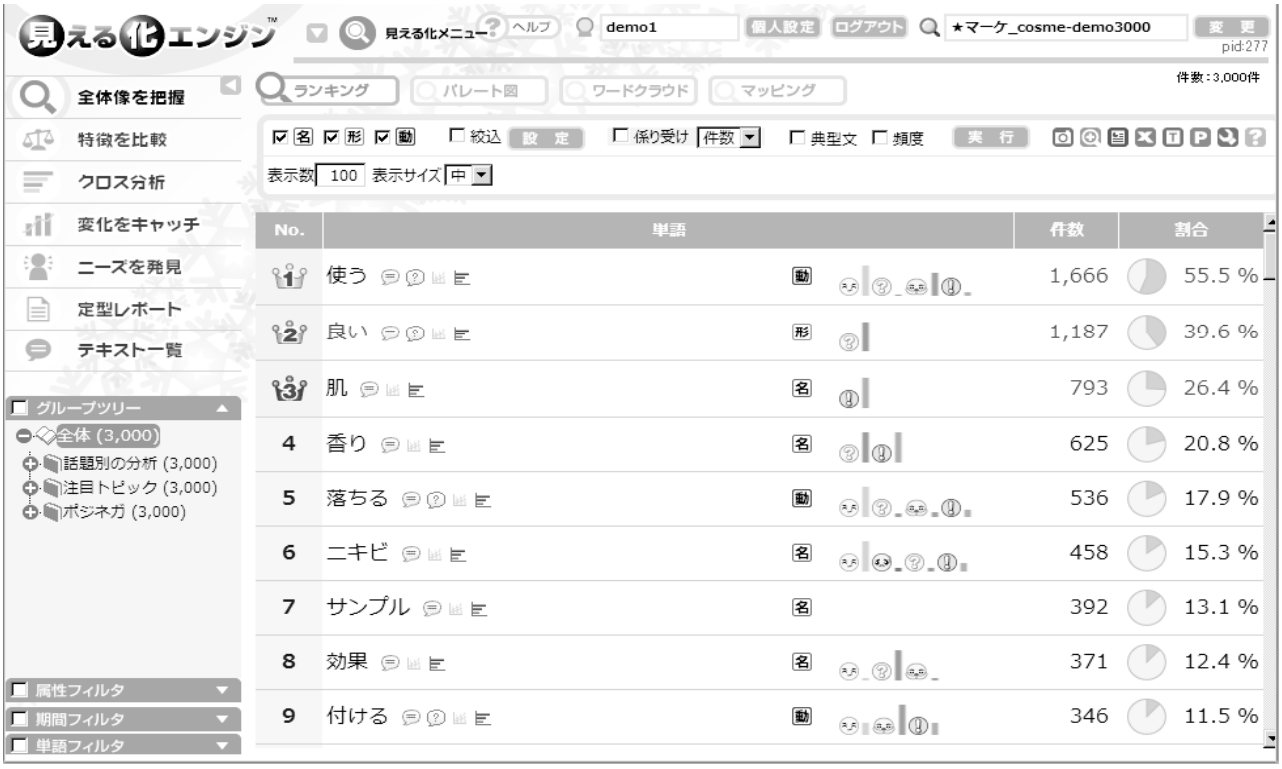
# 特徴① 全体から詳細へ

## クイズ：この化粧品の問題点はどこでしょう？

| No | | 年代 | 肌質 | 都道府県 |
|---|---|---|---|---|
| 1 | メイクや毛穴の汚れはかなり取れます。私は香り(オリーブの？)がダメで友達にあげました。ニキビには特には効果なし。 | 20代前半 | 普通肌 | 山梨県 |
| 2 | このクレンジングを使っているときは、毛穴は広がるし、ニキビが治らなかったんですが使うのをやめたら、毛穴が目立たなくなってきました。ニキビは少しはおさまったかな？雑誌で「メイクが良く落ちる」と載っていたので購入したのですが、確かにメイクは良く落ちますが顎のところが乾燥しちゃうし、においがダメ。わたしの肌には全然合いませんでした。 | 20代後半 | 混合肌 | 東京都 |
| 3 | においはなんだかかわいくってけっこう好きなんですよね。でもダメ！なんかひりひりするんです。私が敏感だからっていうのもあるかもしれないけど。。。3回くらい試しましたが、ひりひりするばっかりでした。 | 20代後半 | 脂性肌 | 大阪府 |
| 4 | 雑誌で評判がいいのと、ニキビにも効果あり、というので、ニキビジプシーしてた頃使ったんだけど。。。。まず感触がだめ。ごわごわしてるって書かれてた方がいたけど、その通り。すすいでも油が残っているような気もしました。コメドがぽろぽろ、ていうのも？？？そして困ってしまったのは、ニキビが悪化しちゃったこと。一月ちょい使って、捨てました。 | 30代前半 | 脂性肌 | 神奈川県 |
| 5 | 敏感肌でもないけどヒリヒリしてダメでした。にきびに効果があったのかよく分からなかったし。今は使っていません。 | 20代未満 | 普通肌 | 大阪府 |
| 6 | ニキビ用に、20代前半まで使ってました。若い頃はこの刺激が「キクーー！！」って感じだったのだけど、今はピリピリ刺激が強すぎてもーダメ･･･アルコール臭がきつすぎるのもどうかと… | 20代後半 | 混合肌 | 宮城県 |
| 7 | まず、匂いがダメなんです。それからひきしめとか、潤いとか･･･なんの効果も感じられないんです。だから、他の物はリピートしても、これはもういいです･･･。 | 30代前半 | 乾燥肌 | 東京都 |
| 8 | う〜ん、わたしは苦手です。まず臭いがダメ。あまいんだか、苦いんだか、なんとも形容しがたい臭いです。原液というから期待してたんですけど、効果も目に見えるものがありませんでした。肌に合えば、安いしお買い得なんですけどね。 | 20代後半 | 脂性肌 | 京都府 |
| 9 | メイクは良く落ちました。ただ、私にはあの独特な匂いがどうもダメで、１本使いきるのにいっぱいいっぱいでした。小鼻の黒ズミとかが改善されるわけでもなく、リピートはしていません。でも、肌がトラブルを起こしたわけでもなく、無事に1本使いきれたから、まぁいっかと言ったところです。 | 20代未満 | 混合肌 | 神奈川県 |
| 10 | 新商品という事で買ってみました。混合肌で吹き出物が出来やすタイプの肌なので、良いかも！っと思ったのですが、べたつきが気になって、ダメでした。2.3回使って乾燥肌のお友達にあげました。サッパリ感を期待してたので･････ | 30代前半 | 混合肌 | 大分県 |
| | ノンオイリーなので期待してたのですが、私にはちょっと潤いすぎました。なんかベタベタしてダメでしたね。まだサンプルが2つ | | | |

【テキストマイニングはビジネスの武器となる】

# 特徴①　全体から詳細へ

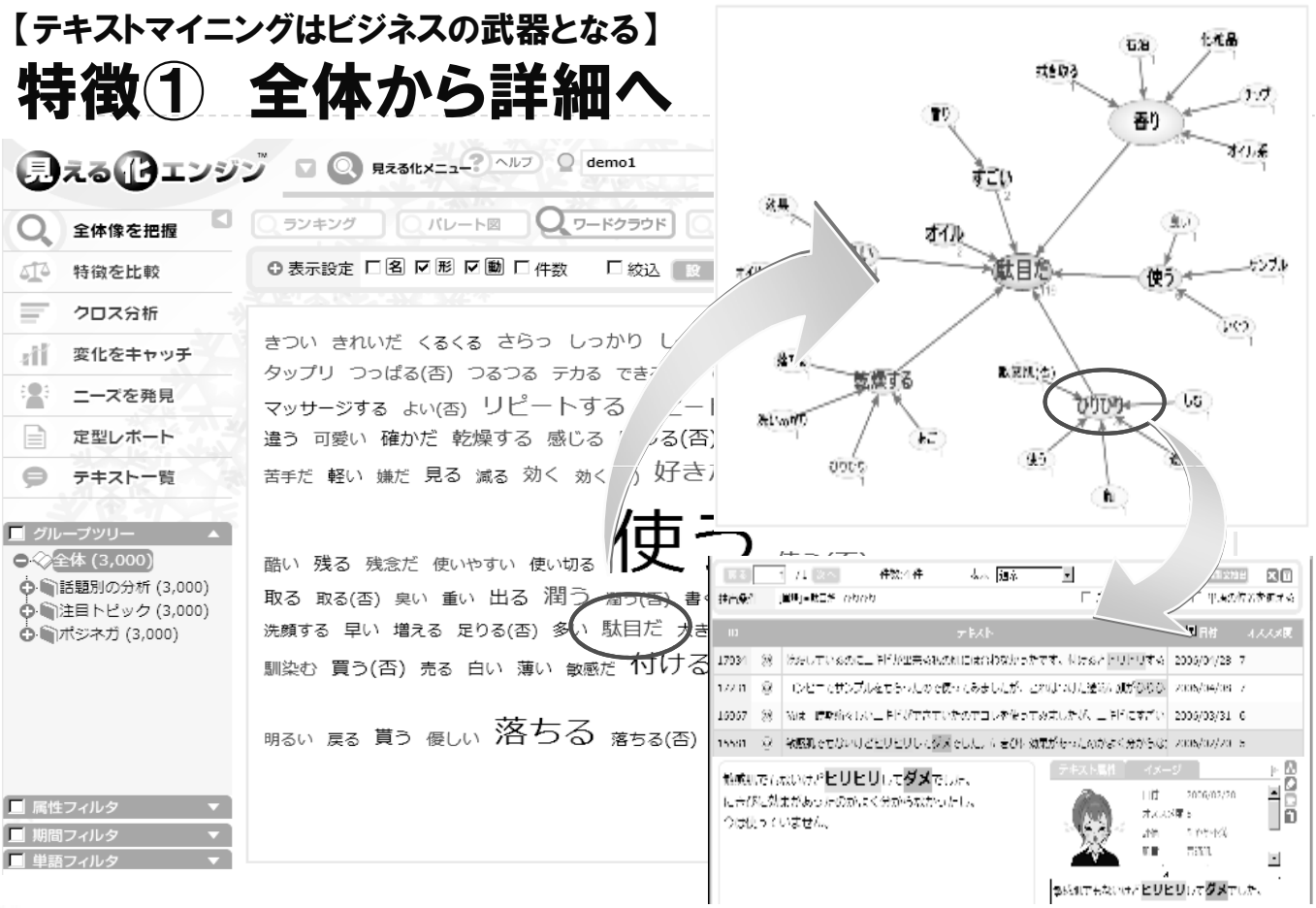







【テキストマイニングはビジネスの武器となる】

# 特徴①　全体から詳細へ



【テキストマイニングはビジネスの武器となる】

# 特徴② 担当者から 全社へ

閲覧者ごとの深掘可能な
**動的なレポート**

分析結果と“レポート形式”として共有


毎朝 スマホやPCに届く、
**レポートメール配信**





【テキストマイニングはビジネスの武器となる】

# 特徴③　希少意見の発見

言語解析技術をつかって、お客様の気持ちが強い箇所を洗い出す

## 感情の表現抽出

（要望、不可能、不可能など）

**要望**


5.93%

| | | |
|---|---|---|
| 1位 | 使いたい | 64件 |
| 2位 | リピートしたい | 17件 |
| 3位 | 試したい | 17件 |
| 4位 | 洗いたい | 6件 |

**不可能**


9.17%

| | | |
|---|---|---|
| 1位 | 使えない | 57件 |
| 2位 | 取ることができない | 26件 |
| 3位 | 言えない | 17件 |
| 4位 | 勧めることができない | 12件 |

**困難**


2.63%

| | | |
|---|---|---|
| 1位 | 使いにくい | 32件 |
| 2位 | 出しにくい | 5件 |
| 3位 | 泡立ちにくい | 4件 |
| 4位 | 泡立てにくい | 3件 |

## 比較の表現抽出

（○よりも△のほうが□だ）

【テキストマイニングはビジネスの武器となる】

# 特徴③　希少意見の発見

言語解析技術をつかって、お客様の気持ちが強い箇所を洗い出す

**比較の表現抽出**
（○よりも△のほうが□だ）

■ **比較表現 勝ち負け表**

| | クリアアリビのほうが | | 髪とホゾノのほうが | | プレミアムモルツのほうが | | スーパードライのほうが | | マービスのほうが | |
|---|---|---|---|---|---|---|---|---|---|---|
| クリアアサヒより | | | 冷たい | 88 | 苦い | 48 | 苦い | 22 | マイルド | 7 |
| | | | 冷たい | 82 | マイルド | 42 | 軽快さ | 22 | 飲みやすい | 7 |
| | | | 炭酸感 | 7 | 香り | 37 | 薄い | 7 | 美しい | 5 |
| | | | 上品さ | 6 | コク | 36 | スッキリ | 19 | 華やか | 5 |
| | | | 軽快さ | 6 | 安い | 8 | 上品さ | 7 | 安い | 4 |
| 髪とホゾノより | 安い | 31 | | | コク | 126 | 薄い | 27 | 華やか | 8 |
| | 飲みやすい | 21 | | | 濃い | 52 | 飲みやすい | 20 | 濃い | 7 |
| | 美しい | 8 | | | 華やか | 41 | スッキリ | 20 | 安い | 5 |
| | 苦い | 7 | | | すっきり | 7 | 軽快さ | 7 | マイルド | 5 |
| | 軽快さ | 6 | | | 冷たい | 6 | すっきり | 3 | いい香り | 5 |
| プレミアムモルツより | 薄い | 8 | 炭酸感 | 7 | | | 安い | 7 | コク | 4 |
| | 上品さ | 8 | 薄い | 15 | | | 華やか | 6 | 美しい | 4 |
| | 後味の軽い | 5 | 冷たい | 6 | | | 苦い | 6 | すっきり | 3 |
| | 軽快さ | 4 | 軽快さ | 3 | | | 軽快さ | 3 | 後味の軽い | 3 |
| | すっきり | 4 | 苦い | 4 | | | 冷たい | 4 | 上品さ | 4 |
| スーパードライより | ない | 56 | 苦い | 17 | うまみ | 26 | | | ない | 5 |
| | すっきり | 35 | うまみ | 8 | 冷たい | 23 | | | うまみ | 5 |
| | 後味の軽い | 23 | 後味さわやか | 6 | 華やかな | 21 | | | 上品さ | 5 |
| | 美しい | 5 | 濃い | 5 | コク | 18 | | | 華やか | 5 |
| | 濃い | 5 | 小麦の | 5 | 濃い | 7 | | | 軽快さ | 5 |





【テキストマイニングはビジネスの武器となる】

# ここまでのまとめ

## テキストマイニングは、ビジネスの “武器” となる


気づき発見ツール

思考のスピードを止めることのない、高速レスポンスと 直感的な表現を実現

**①全体化から詳細へ**　**②担当者から全社へ**　**③希少意見の発見**

# 今日お伝えしたいこと

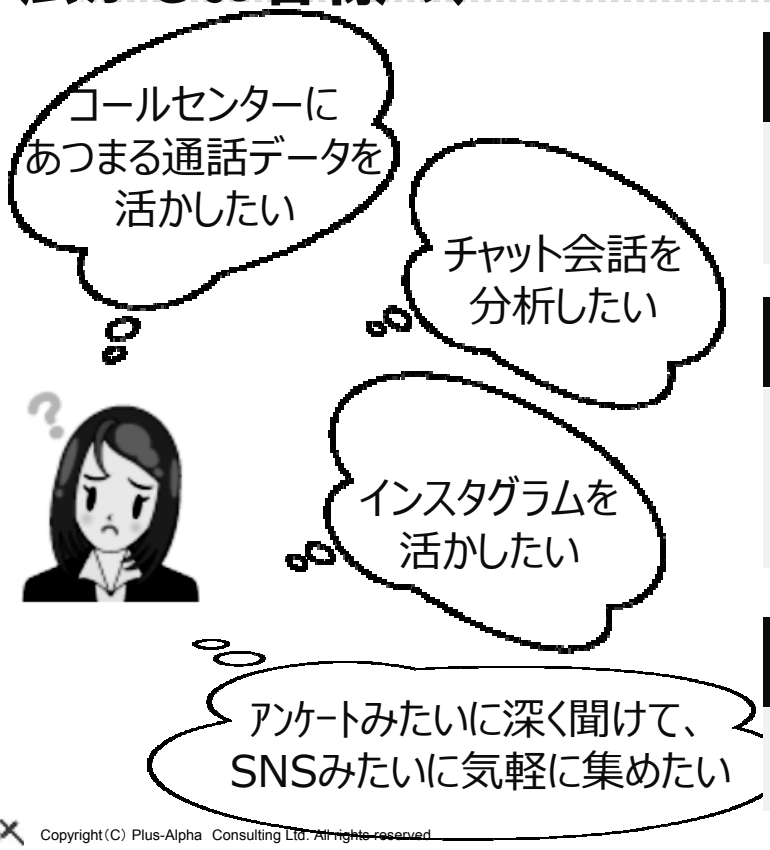
## テキストマイニングは、
## ビジネスの “武器” となる

## お客様の声は、
## 多様化し、進化している




【お客様の声は、多様化し、進化している】

# 広がるお客様の声




### 言語解析技術 の 向上

■日々 進化している言葉への対応

### テキストマイニング の 向上

■対話の自動要約
■会話分析（1対1、多対多）
■テキストマイニング×画像分析

### データ収集 の 向上

■瞬速リサーチ機能 (2h 100コメント)

【お客様の声は、多様化し、進化している】

# 言語解析技術 の向上

## 「生きている言葉」への言語解析技術の課題

新しい言葉や、使われ方こそ、まだまだ**少数派** 少数

文法ルールの基本があるものの**例外的**な使われ方が存在 例外

同じ言葉でも言回しや意味が**業界や年代で異なる** 固有

**ハイブリットな言語解析ｱﾌﾟﾛｰﾁ**

**ルールベース型**
人による知識と経験から得た**"理由"**による解析が可能

**機械学習型**
データとロジックから算出された**"傾向"**による解析が可能





【お客様の声は、多様化し、進化している】

# 言語解析技術 の向上 ①

# 「しかない」問題

# 「彼はプレモルしか飲まない」


○ 飲む （肯定）

× 飲まない（否定）

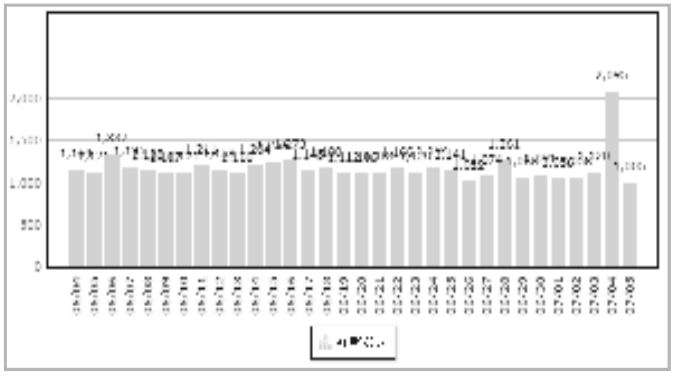
【お客様の声は、多様化し、進化している】

# 言語解析技術 の向上 ②

# 「好きくない」

## 日本語として、間違えている

「好きくない」をTwitterで検索すると1か月で
37,541件



**新しい言い方への対応できる進化が重要**

※「好き」は形容詞では　なく形容動詞（好きだ）





【お客様の声は、多様化し、進化している】

# 言語解析技術 の向上 ③

## ネット上では、日々、
## 新しい日本語が誕生。

・昨日町中で友達を見つけて思わずニヤついた(ヽ´ω`)

・先週彼女にフラレたショックから立ち直れません・・・

・やっばいくらいたくさんお菓子貰ったんだけど！(*´艸｀*)

・安モノの傘を買ったら1日で真っ二つに折れました

【お客様の声は、多様化し、進化している】

# 言語解析技術 の向上 ④

## 係り受けの難しさへの対応

やさしい お母さんの カレー

おいしい お母さんの カレー




【お客様の声は、多様化し、進化している】

# 言語解析技術 の向上 ⑤

## 業界ごとの固有表現への対応

「やばい」の同義語は？

10代では、○○○
60代では、○○○

「飲む」の同義語は？

飲料業界では、○○○
医薬品業界では、○○○

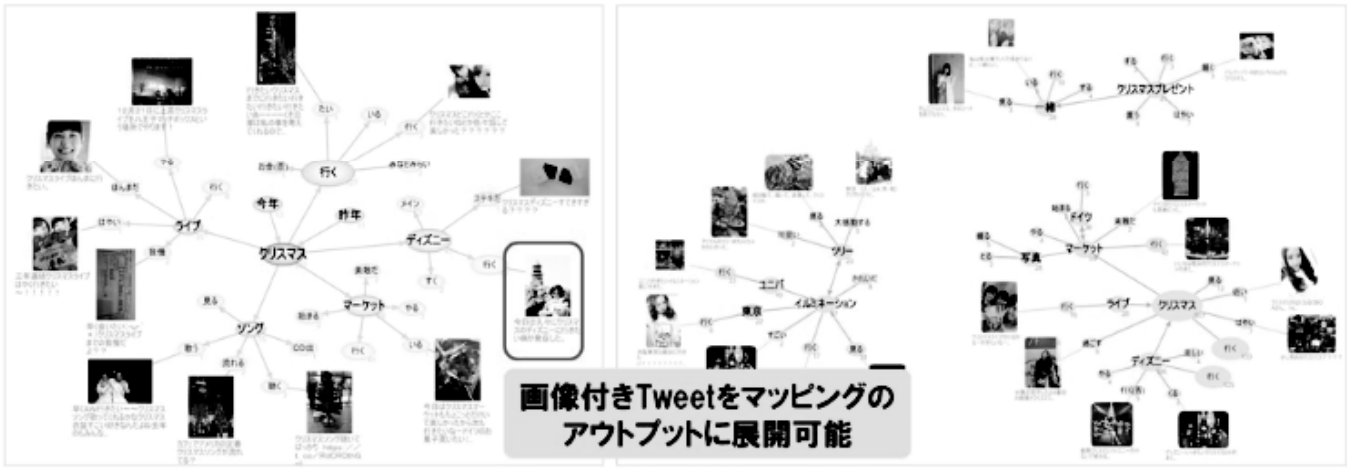
【お客様の声は、多様化し、進化している】

# テキストマイニング の 向上

## テキストマイニングでのマップ表示と 画像との融合

■ 「単語マッピング」話題内容の把握とともに画像で分かりやすく

■ 「全体マッピング」画像表示で、全体の話題構成を可視化




・プロモーションや新商品の反響分析

・未充足ニーズの発見



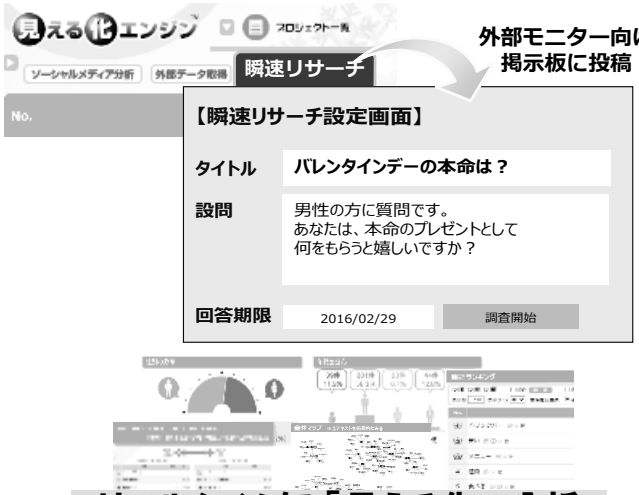


【お客様の声は、多様化し、進化している】

# データ収集 の 向上

## “いつでも”9万人に対して調査をすることが可能に

●アンケートのように、ピンポイントのテーマ収集
●新サービスへのテストマーケティング
●画像などを使った視覚調査も可能






リアルタイムに「見える化」分析

【テキストマイニングはビジネスの武器となる】

## ここまでのまとめ

**言語解析技術 の 向上**
生きている日本語への対応

**テキストマイニング の 向上**
テキストマイニング×画像

**データ収集 の 向上**
リアルタイムな深い声の収集

# お客様の声は、
# 多様化し、進化している

新しい言葉や、使われ方こそ、
まだまだ**少数派** → 少数

文法ルールの基本があるものの
**例外的**な使われ方が存在 → 例外

同じ言葉でも言回しや意味が
**業界や年代で異なる** → 固有

**ハイブリットな言語解析ｱﾌﾟﾛｰﾁ**

**ルールベース型**
人による知識と経験から
得た**"理由"**による
解析が可能

**機械学習型**
データとロジックから
算出された**"傾向"**による
解析が可能





## 今日お伝えしたこと

# テキストマイニングは、
# ビジネスの “武器” となる

# お客様の声は、
# 多様化し、進化している

# 「見える化」で担う社会的意義

## よりよい社会のための生活者と企業の橋渡し

**生活者**

**要望・意見**

**企業**


**よりよい商品**
**サービス**






**ビジネスの現場で、**
**言語処理技術、人工知能技術を磨きたい方、**
**ぜひご連絡ください**

**見える化エンジン™**

主催者　高度言語情報融合フォーラム（ALAGIN）
一般社団法人 言語処理学会
一般財団法人 日本特許情報機構（Japio）
発行日　平成２８年２月２９日
発行者　高度言語融合フォーラム（ALAGIN）内
産業日本語研究会・シンポジウム事務局
email: info@alagin.jp
TEL: 03-3351-8155