3.8 ビジョン・言語処理

人間同士や人間と知能機械のコミュニケーションを行うための表現メディアの解析・理解、その検索・変換・編集、これらを支える知識情報処理に関する研究開発分野をビジョン・言語処理と呼ぶ。

人の知的な振る舞いをコンピューター上で実現することを目指す、いわゆる人工知能の研究は 1950 年代から始まった。しかし、視聴覚から入力される知の源泉となる情報、すなわち、映像・画像や音声言語・文字言語などの処理(ビジョン・言語処理)が極めて難しいことから、その研究は容易ではなかった。そのため、初期の人工知能研究は、意味や知識処理のモデルに関する議論や、チェスに代表されるような単純で明確な規則を持つ問題を対象とした。

これに対して、近年、コンピューター環境の劇的進展、ウェブを始めとする大規模データ、いわゆるビッグデータの出現、さらに、データに正解を与えて解析器を学習する機械学習の発展があり、初期には非常に困難であった物体認識、音声認識、構文解析など、ビジョン・言語処理の基本処理の精度が劇的に改善した。そこで、このような基本処理の上で、いよいよ意味や知識の問題に取り組めるようになり、本格的な人工知能の研究が可能となってきた。また、人の知的活動を支援し、さらには社会を変革する可能性のある応用システムも生み出されつつある。最近では人工知能の発展に対する懸念の声が聞かれるようになってきたほどであるが、これは主にビジョン・言語処理の進展に対するものである。

このような研究開発分野は前回の俯瞰報告書では区分「知能/インタラクション」の一領域であったが、当該分野の急速な進展を背景として、今回は「ビジョン・言語処理」という区分をたて、その中でいくつかの領域について報告を行うこととした。

ビジョン・言語処理のミニ俯瞰図を図 3.8.1 に示す。ビジョン・言語処理の基本処理においても、厳しい実環境、例えば強い雑音下の音声認識やネット上のくだけた言い回しの解析などにはまだまだ改善の余地がある。しかし、本報告書では、今後の社会へのインパクトや政策的課題などを考慮し、比較的上位レイヤーの研究開発領域について俯瞰を行った。

まず、意味や知識の取り扱いに関しては、文字言語、すなわちテキストの処理からアプローチしやすい。コンピューターによって大規模テキストから知識を抽出し意味解析を行うことは、膨大なウェブ情報の利活用支援にもつながる。このことから、「大規模言語処理に基づく情報分析」領域をたてた。さらに、言語処理の応用として、近年の発展がめざましい「機械翻訳」と「音声対話」を項目とした。これらの応用システムは 2020 年の東京オリンピック・パラリンピックを契機に社会に大きなインパクトを与えるものと予想される。

ビジョン関係では、車の自動運転など既に社会実装レベルの応用システムも存在するが、本報告書ではいわゆる一般物体認識などをターゲットとする、より一般的・普遍的な「画像・映像の意味理解」を項目とした。さらに、言語の基本処理の高精度化と、一般物体認識の実現が視野に入りつつあることを受けて、「言語と映像の統合理解」の項目をたてた。この研究開発領域は現時点ではまだ萌芽的であるが、今後、このあたりをハブとして、認知科学、脳神経科学、ロボティクスなどの関連領域との本格的連携が生まれてくることが期待される。

ビジョン・言語処理は、いわば「地に足がついた」形で、データをきちっと集積して解析

するということをベースに発展してきており、わが国は国際的にも高い競争力を有している。今後もこの発展を持続・加速させ、わが国の競争力を維持・発展させるためには、各領域の報告でも共通して述べられているように、データの収集と研究利用を妨げない法律等の整備、さらに、HPC 基盤やデータ基盤の共有を含む研究連携や共同プロジェクトを推進する政策などが必要であると考えられる。

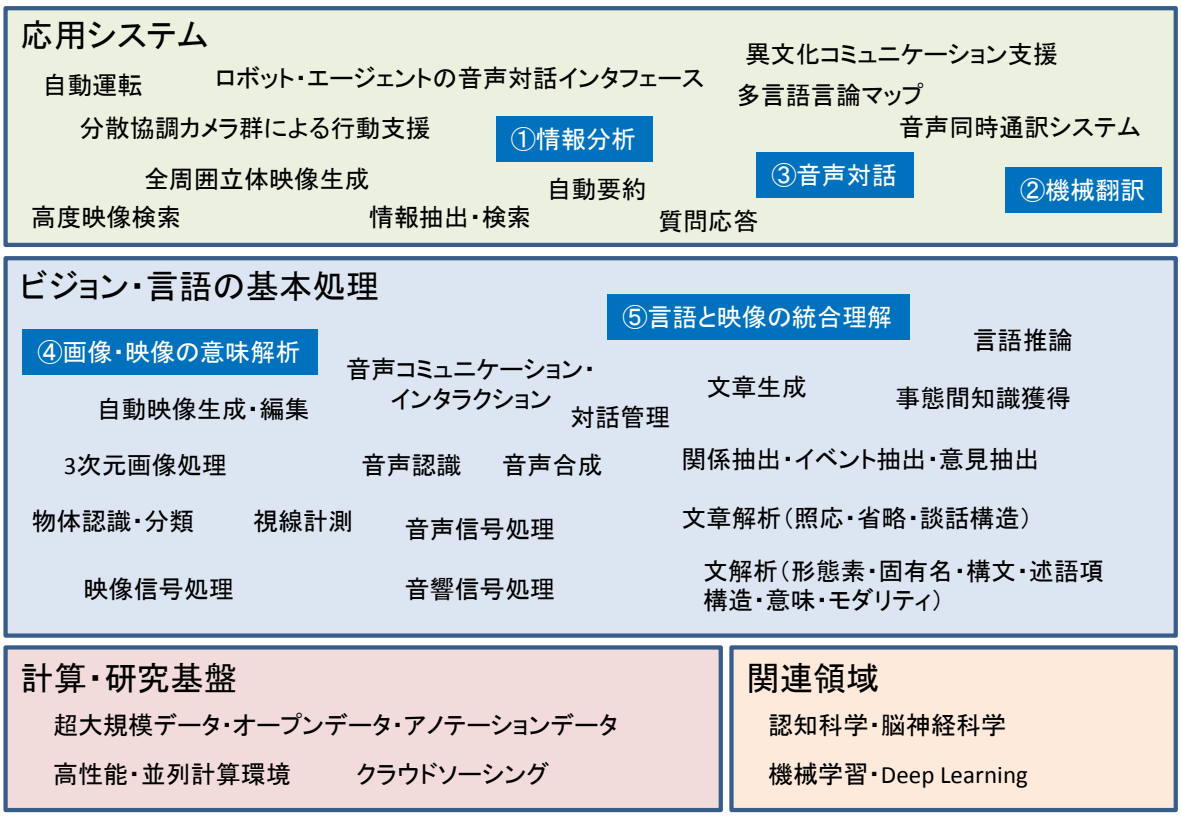

図 3.8.1 ビジョン言語処理の俯瞰

３．８．１　大規模言語処理に基づく情報分析

（１）研究開発領域名

大規模言語処理に基づく情報分析

（２）研究開発領域の簡潔な説明

Web、SNS、各種の論文アーカイブ等に存在する大量のテキストに言語処理技術を適用し、それらのテキストに書かれた膨大な情報や知識を抽出する。さらに、それらに現れる社会、コミュニティーの大域的な動向や、情報、知識の間の相関、関連、書かれた知識を組み合わせて作られる新たな仮説など、さまざまな価値ある分析結果をユーザーに提供できる情報分析技術を開発する。

（３）研究開発領域の詳細な説明と国内外の動向

大量のテキストは、仮に有用な情報を含むことが事前に分かっていたとしても、有効活用が困難である。これは人が読めるテキストの量には限りがあることによる。インターネットの出現以前にはそもそも人が入手できるテキストの量も少なく、この単純な事実は取り沙汰してもしょうがないことであった。ところがネットの出現により大量の電子可読のテキストが容易に入手可能になると状況は変わり、計算機がそうしたテキストを「自動的に読みこなして」有効活用につなげることが期待されるようになった。この有効活用を可能にする第一段階の技術が検索エンジンであるが、現在の課題の一つは、検索エンジンのように提示されたテキストの意味解釈をユーザーに任せるのではなく、計算機自身が意味解釈を行って、ユーザーの負担を可能な限り減らすことである。別の言い方をすれば、計算機にユーザーの知的活動をより広範に代行してもらい、ユーザーの介入を重要な最小限のポイントに限定することである。例えば、災害時に Twitter や各省庁、自治体の報告書等に膨大な災害関連テキストが出現するが、検索エンジンを利用したとしても、その有効活用は救援側のマンパワー、時間的制約の中では極めて困難である。また、分野をある程度限ったとしても、膨大な科学技術論文や Web 等に書かれた専門的知識をすべてカバーできる専門家、研究者はまずいない。この結果、被災状況や新たな科学的発見のチャンスを見落とすと言ったことが起こり得るし、また実際に明確に認識されることがないだけで至る所で起きているであろう。こうした状況において、システムが Web や SNS のものも含め膨大なテキストを「読みこなし」、重要なポイントだけを人が容易に理解可能な形、量で提示してくれるとすれば、社会全体の知的活動が多いに活性化し、より適切な意思決定が行われるようになるであろう。

本稿では「情報分析」という用語を、上述のような意味でのテキストの知的処理をターゲットとする自然言語処理の（機械翻訳を除いた）さまざまなタスク、システム、および、それらの組み合わせの総称として用いる。なお、情報分析を支える基盤として、構文解析などの基盤的な自然言語処理技術が必要であるが、これらに関してはいまだ改善の余地はあるものの、既により高次の情報分析技術の研究が可能な完成度が既に実現されている。

情報分析技術のはしりとしては、1980 年代後半から情報抽出というタスクの研究が開始された。これは、例えば特定のタイプの出来事（例：殺人事件）に関する記述を新聞記事等から自動的に発見し、その出来事の属性（例：犯人、被害者、犯行場所）をやはり自動的に抽出するといったものである。こうしたタスクをこなせるシステムは蓄積された膨大なテキ

ストの有効活用につながるとされ、米国では、DARPA などの支援により、多数の大学、企業が参加したコンテスト等が継続的に行われ、さまざまな基礎技術が開発された [1)]。また、その後、この技術の変種として、例えば「タバコ」と「肺がん」といった単語の組を因果関係に関する知識として大量のテキストから抽出するといった、知識獲得といったタスクも盛んに研究された。同様に、テレビのクイズショーで人間のチャンピオンを破り、大変有名になった IBM の Watson [2)]のように膨大なテキストをもとに自然言語で書かれた質問に回答する質問応答、あるトピックに関して書かれた評判を抽出して、肯定的評判／否定的評判等に分類する sentiment analysis、評判分析といった技術も開発が進んでいる。さらに、以上の技術でも必要となる、テキスト間の同義性や含意関係を認識する含意関係認識といった技術も研究開発されている。また、これらの技術についてもやはり DARPA の支援等により、性能に関するコンテストが実施されてきた [3)4)5)6)]。

ここで重要なポイントは、これらの技術も現在のところは入力となるテキスト中の単語、フレーズなどの短い断片を抜き出してそのまま出力としているものがほとんどであることである。IBM の Watson のような質問応答の成功例は、多くがテキストから抜き取った単語一語を回答とするタイプのものである。一方で、ある出来事の理由など、文章を回答とするタイプのより高次な質問応答で、一つの単語、一つの文の意味解釈だけで処理が終わらず、文章全体の解釈が必要なタスクにはいまだ課題が多い。本来、文章の解釈というのは非常に知的な作業であり、それを計算機が広範に代行するにはまだまだ課題が多いといっても良い。例えば、将来的には「大規模言語処理の研究開発動向は？」といった質問を受け取ると、大量の文章を分析し、その結果を論理的に整合性が取れるようにマージ、要約して、本稿のようなサーベイを出力するシステムも開発のターゲットとなろうが、現在の技術だけではこうした処理は到底不可能である。そうした意味で今後の発展の余地は非常に大きく、また、大きな社会的インパクトも予想される。

以上の技術はテキストに書かれた情報を抽出するものであるが、テキストに書かれていない知識を推論によって求める技術の開発も始まっており、生命科学において有力な仮説を提示して研究を加速するケース [7)]や、Web 上の情報から導かれた仮説が著名な科学ジャーナルの内容を先取りするケース [8)]なども一部で報告され始めている。よく「行間を読む」といった言い方がなされるが、これはまさに仮説を生成しながらテキストの解釈を行うといった状況を指すものと考えられ、上記の仮説生成技術も仮説生成自体を目的とするアプリケーションにとどまるのではなく、将来的にはテキスト解釈の必要不可欠な基礎的処理として、本来、仮説生成を陽に目的としないタイプの情報分析の高度化にも貢献する可能性がある。

なお、(5)注目動向の項で詳述するが、IBM の Watson、Google Knowledge Vault [9)]など、情報分析技術の実用化は急速に進展している。 また、こうした実用化の動きの中で、特に Web、SNS に関して問題となってきたのが情報の信ぴょう性である。東日本大震災ではネット上に大量のデマが出回ったことが記憶に新しい。これに関連し、信頼度の低い情報を検出する手法に関する研究も既に始まっている [10)11)12)]が、いまだ決め手となる手法等は出現しておらず、今後さらなる展開が待たれる。

（４）科学技術的・政策的課題

言語処理技術は個別言語に依存し、大規模な情報分析システムの研究開発にはターゲット

となる言語に精通した多くの研究者が必要である。世界的に見た場合、研究者が最も多いのは英語に関してであり、また、大手検索エンジンなども英語のサービスから開発、公開するのが通常である。こうした状況は日本語で書かれた情報が英語の情報に比べて有効活用されにくくなるという可能性を意味し、日本の国際競争力をそぐ可能性すらある。こうした点に鑑み、日本語に通じた多くの優秀な言語処理研究者を育成する必要がある。

一方で、研究実施に際しては、大規模な計算リソース上に、大量の文書を配置して処理を行う必要がある。米国の大手ネット企業ではそうした環境が身近に存在するが、日本の大学の研究室レベルではこうした計算リソースの確保は困難であり、大学、研究機関等でそうした計算リソースを共有する枠組みがまずは必要である。また、計算リソースが実際に利用可能であっても技術的なハードルがあり、Hadoop 等のツールも実際に使いこなせている言語処理研究者は少ない。今後 High Performance Computing と言語処理を橋渡しするさらに進化した枠組みが必要となろう。また、研究開発で利用できるデータに関していえば、米国においては大量の Web ページを収集してまとめた ClueWeb [13]といったデータセットが利用可能であるが、日本においては法的問題もあり、Web 文書を異なる組織間で共有することが難しく、研究開発の障害となっている。

以上は研究リソース、技術的課題についてであるが、研究者のスキルアップも課題である。例えば、戦略的に大規模実験のプランニングを行い限られた時間内で成果を挙げる、あるいは、大規模な情報分析のための柔軟なアーキテクチャを設計する、さらにはリーダーシップを取って大規模な情報分析プロジェクトをマネージメントするといったスキルを持つ人材は特に日本では極めて少ない。これは大学等の教育・研究が小規模なリソースで実施可能な要素技術に関するものに偏っているためと考えられ、そうした課題を視野に入れての人材育成が待たれる。また、大規模な情報分析システムを社会実装する上では著作権、プライバシー等、法律に関わる問題も無視できない。技術とそうした社会的、法的問題の両方に精通した人材の育成も急務である。

（５）注目動向（新たな知見や新技術の創出、大規模プロジェクトの動向など）

現在、最も脚光を浴びているのは、各種情報分析技術の実用化の試みである。まず、IBM の Watson による医療応用等が開始され、日本語で会話を行う家庭用ロボットへの組み込みの検討も開始されたとの報道もある [2)14)]。Apple の Siri [15)]、NTT ドコモの「しゃべってコンシェル」[16)]も一定規模のデータに対して情報分析を行っているシステムと考えられる。さらに、入力されたトピックに関する評判を sentiment analysis を用いて Web 等から抽出するサービスはさまざまな企業によって実用に供されており [17)]、大規模な知識獲得の結果をユーザーに提供する Google Knowledge Vault [9)]なども実用に供され始めている。

一方で、前述したように技術の潜在的可能性はいまだ膨大なものがあり、大学や公的研究機関での、より基礎研究指向の研究プロジェクトも多数存在する。例えば、米国 DARPA の支援によって DEFT [18)]、SMISC [19)]等のプロジェクトが実施されている。日本国内においては、NICT において数十億件規模の Web ページを用いて、仮説生成や質問応答を行う大規模システム WISDOM X の開発が進んでいる [20)21)8)]。また、SNS に関しては、NICT が Twitter からデータの提供を受け、災害時の情報に関して質問応答を行う対災害 SNS 情報分析システム DISAANA [22)23)24)]を開発している他、ソーシャルメディア等での社会的な動きを把握、

活用するための大規模な情報分析研究のため、Twitter が MIT に 1000 万ドルの資金援助をするといった動きも出ている[25]。

（６）キーワード

自然言語処理、情報分析、情報抽出、質問応答、知識獲得、関係抽出、sentiment analysis、評判分析、含意認識、矛盾認識、情報信ぴょう性、仮説生成、Big Data、Textual Big Data

（７）国際比較

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	○	→	言語処理学会[26]という学会を中心に基礎研究が進められており、年次大会には毎年３００件前後の発表がある。また、NICTが中心となって設立したALAGINフォーラム[27]においても言語資源等の共有が進められている。日本語の情報分析技術を中心に研究成果が出ており、英語に関する研究が圧倒的に多数であるトップカンファレンスACL[28]などでも一定数の発表がなされている。
	応用研究・開発	○	↗	JST CREST等において、具体的な応用を指向した情報分析の研究が進んでいる[29]。NICTにおいて、大規模Web情報分析システムWISDOM[10]が2010年に一般公開された。現在その後継として前述したWISDOM Xの開発が進んでおり、2014年度に一般公開予定である。また、やはり前述した災害関連情報を分析するシステムDISAANAも試用版が既に公開されている。NEC等の民間企業でも、例えば、含意関係認識技術に関して、ビジネス化を意識した開発が進んでいる[30]。
	産業化	○	→	前述したように、NTTドコモの「しゃべってコンシェル」がサービスインした他、企業等の評判をWeb等からマイニングするサービスもさまざまな企業から提供され、ビジネス化されている。また、2015年に発売予定のソフトバンクの家庭用ロボットもバックエンドは情報分析技術と呼んでよいと思われる[14]。
米国	基礎研究	◎	→	Stanford、CMU、MIT、ペンシルバニア大学、ワシントン大学等の有力大学のみならず、Google等の有力企業はそれぞれ強力な研究チームを持っており、基礎研究を遂行している。ACL、EMNLPなど有力な国際会議等での発表の多くは米国発である。
	応用研究・開発	◎	→	上述した基礎研究が比較的短期間に有力企業で応用研究・開発に回るサイクルが確立している。また、有力大学の優秀な大学院生が有力企業のインターンに行き、良い成果を挙げることも多い。
	産業化	◎	↗	もともとは大学発のベンチャーであったGoogleを始め、産業化の実例には枚挙に暇がない。最近ではIBMがWatsonの産業化に10億ドルを投じ、世界各国の企業とプロジェクトを開始している[14]。
欧州	基礎研究	○	→	欧州における言語処理研究は、欧州連合(EU)が多数の国から構成されていることもあって、機械翻訳に重点が置かれており、情報分析に関してはイスラエル等の例外を除き、突出した研究は少ない。
	応用研究・開発	○	→	European CommissionのFP7等のファンディングにより3年程度の期間で予算総額が300万EURO前後のプロジェクトが一定の数存在する。例えば、PHEME project[31]では、ソーシャルメディア上で広がっているうわさを検出・把握することを狙っており、また、NewsReaderプロジェクト[32]では複数言語のニュース記事等を用い、さまざまな事象の時系列的な展開を追跡し、未来予測まで行って意思決定の支援を行うことを目標としている。
	産業化	○	→	Google等グロバール企業の研究所が存在し、一定のアクテビティはあるが、米国の主導のもと産業化が行われている側面が強い。
中国	基礎研究	○	↗	北京大学、清華大学等の有力大学やMicrosoft Research Asia、Baidu等の民間企業の研究所を中心に基礎研究が進められている。また、ACL、WWW等のトップコンファレンスでは中国発の論文が多数発表されている。
	応用研究・開発	◎	↗	Microsoft Research Asiaでは、Webを用いる質問応答システムを開発している[33]。Baiduは、米国Silicon valleyに人工知能研究所を設立し、Deep Learning分野に3億ドルを投資すると発表した[34]。大規模な情報分析への応用が考慮されていると思われる。
	産業化	○	→	Microsoft、IBM等のグロバール企業におけるアジア地域の研究所が集中している。また、Baiduは検索エンジンの改良、音声処理技術の改良等を目指して大規模プロジェクトを進めている。

韓国	基礎研究	△	→	KAIST、ETRI、KISTI等の有力大学、国研を中心に基礎研究が進められている。BORA（言語資源銀行）[35]というコンソーシアムにおいて音声・言語資源等の共有が進められている。ただACL等のトップコンファレンスでの韓国発の発表は減少している。
	応用研究・開発	○	↗	ExoBrain[36]というプロジェクトでは、進化する知能を持つ質問応答システムに関し研究開発を進めている。また、InScite[37]という科学技術文献の情報分析プロジェクトも進行している。
	産業化	○	→	90年代末から2000年代初までに生まれたNaver、Daum等、情報分析技術が産業化された実例は多い。最近もカカオ社などベンチャー企業の創業が増加している。また、有力企業であるサムソン社、LG社等での製品化の事例もある。

（註 1）フェーズ
基礎研究フェーズ ：大学・国研などでの基礎研究のレベル
応用研究・開発フェーズ ：研究・技術開発（プロトタイプの開発含む）のレベル
産業化フェーズ ：量産技術・製品展開力のレベル
（註 2）現状
※わが国の現状を基準にした相対評価ではなく、絶対評価である。
◎：他国に比べて顕著な活動・成果が見えている、 ○：ある程度の活動・成果が見えている、
△：他国に比べて顕著な活動・成果が見えていない、×：特筆すべき活動・成果が見えていない
（註 3）トレンド
↗ ：上昇傾向、 →：現状維持、 ↘：下降傾向

（８）引用資料

1) TIPSTER Text Program,
NIST, http://www.itl.nist.gov/iaui/894.02/related_projects/tipster/
2) “Computer Wins on ‘Jeopardy!’: Trivial, It’s Not”, New York Times, Feb. 16, 2011.
http://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html?pagewanted=all&_r=0
3) Ellen M. Voorhees, “The TREC-8 Question Answering Track Report”, in the Proceedings of TREC 8, pp.77-82, 1999.
4) NIST Text Analysis Conference, http://www.nist.gov/tac/
5) PASCAL Textual Entailment Challenge, http://pascallin.ecs.soton.ac.uk/Challenges/RTE/
6) NTCIR、国立情報学研究所、http://research.nii.ac.jp/ntcir/index-ja.html
7) W. Scott Spangler, et al., “Automated hypothesis generation based on mining scientific literature”, In the Proceedings of KDD 2014, pp.1877-1886, 2014.
8) Chikara Hashimoto, et al., “Toward Future Scenario Generation: Extracting Event Causality Exploiting Semantic Relation, Context, and Association Features”, In the Proceedings of ACL 2014, pp.987-997, 2014.
9) Xin Luna Dong, et al., “Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion”, In the Proceedings of KDD’14, pp. 601-610, 2014.
10) 情報分析システム WISDOM -Web の健全な利活用を目指して
http://kc.nict.go.jp/project1/WISDOM_TR.pdf
11) Carlos Castillo, et al., “Information credibility on twitter”, In the Proceedings of WWW’11, pp.675-684, 2011.
12) Yusuke Yamamoto, et al., “Enhancing credibility judgment of web search results”, In the Proceedings of CHI’11, pp.1235-1244, 2011.
13) The ClueWeb12 Dataset, http://www.lemurproject.org/clueweb12.php/

14) “人工知能「ワトソン」、日本語版開発　ソフトバンクと”、日本経済新聞、2014 年 10 月 9 日、http://www.nikkei.com/article/DGXLASGM0900J_Z01C14A0EAF000/
15) “To Siri, with Love”, New York Times, Oct. 17, 2014, http://www.nytimes.com/2014/10/19/fashion/how-apples-siri-became-one-autistic-boys-bff.html?_r=0
16) しゃべってコンシェル、株式会社 NTT ドコモ、https://www.nttdocomo.co.jp/service/information/shabette_concier/
17) 口コミ係長、　株式会社ホットリンク、 http://www.hottolink.co.jp/service/kakaricho
18) DARPA DEFT, http://www.darpa.mil/Our_Work/I2O/Programs/Deep_Exploration_and_Filtering_of_Text_(DEFT).aspx
19) DARPA SMISC, http://www.darpa.mil/Our_Work/I2O/Programs/Social_Media_in_Strategic_Communication_(SMISC).aspx
20) Masahiro Tanaka, et al., “WISDOM2013: A Large-scale Web Information Analysis System”, In The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations, pp.45-48, 2013.
21) Jong-Hoon Oh, et al., “Why Question Answering using Intra- and Inter-Sentential Causal Relations”, In the Proceedings of ACL 2013, pp.1733-1743, 2013.
22) NICT 対災害 SNS 情報分析システム DISAANA 試用版、http://disaana.jp
23) “NICT が対災害 SNS 情報分析システムを Web 上で試験公開”、ITPro by 日経コンピューター、http://itpro.nikkeibp.co.jp/atcl/news/14/110501766/
24) Istvan Varga, et al., “Aid is Out There: Looking for Help from Tweets during a Large Scale Disaster”, In the Proceedings of ACL 2013, pp.1619-1629, 2013
25) “MIT Launches Laboratory for Social Machines with Major Twitter Investment”, MIT Media Lab, http://socialmachines.media.mit.edu/2014/10/01/mittwitter-launch-announcement/
26) 言語処理学会ホームページ http://www.anlp.jp
27) ALAGIN フォーラムホームページ http://alagin.jp
28) Association for Computational Linguistics (ACL) ホームページ http://www.aclweb.org
29) 知識に基づく構造的言語処理の確立と知識インフラの構築、JST、http://www.jst.go.jp/kisoken/crest/project/45/45_01.html
30) “分析エンジンの強みを生かす NEC のビッグデータクラウド”、ITPro、2012 年 11 月 8 日、ASCII.jp、http://ascii.jp/elem/000/000/742/742011/
31) PHEME Project, http://www.pheme.eu/
32) NewsReader Project, http://www.newsreader-project.eu
33) “Microsoft Research Asia: A Celebration of 15 Years of Innovation”, Microsoft Research Asia, http://www.msra.cn/zh-cn/um/events/innovationday/2013/english.html
34) A Chinese Internet Giant Starts to Dream, MIT Technology Review, Aug.14, 2014, http://www.technologyreview.com/featuredstory/530016/a-chinese-internet-giant-starts-to-dream/

35) BORA(言語資源銀行)、http://semanticweb.kaist.ac.kr/org/bora/
36) ExoBrain Project, http://exobrain.kr/
37) InScite、 KISTI、 http://inscite.kisti.re.kr/

３．８．２　言語情報処理応用（機械翻訳）

（１）研究開発領域名

言語情報処理応用（機械翻訳）

（２）研究開発領域の簡潔な説明

ある言語から別の言語への翻訳を自動的に行う技術の研究開発

（３）研究開発領域の詳細な説明と国内外の動向

［背景と意義］

機械翻訳は初期の計算機のアプリケーションとして 1940 年代から研究開発が行われてきており、異なる言語で書かれた文書からの情報収集、外国への特許出願やプレスリリースなどの情報発信、母国語が異なる人間同士の多言語による円滑なコミュニケーションを実現する上で非常に重要な技術である。初期の機械翻訳は構文解析、構造変換、生成といった過程でモデル化され、各モジュールを言語の専門家が知識を記述することで実現されていたが、コストが非常にかかるのが問題であった。これに対し、現在主流である統計的機械翻訳は、統計モデルに基づき二言語の対訳データから自動的に機械翻訳システムを構築する技術であり、新しい言語対や分野への適応を短期間かつ低コストで実現可能になった。

［これまでの取組み］

統計的機械翻訳は 80 年代後半からの IBM による単語アライメントモデルから始まり、句単位に翻訳を生成するモデルを実現、さらに構文知識を導入することで統語的に正しい翻訳が可能となった。同時に BLEU などの客観的かつ容易な翻訳評価手法およびその評価尺度を直接最適化する機械学習手法が開発された。

この機械翻訳の発展に寄与したのが DARPA によるファンディングであり、2001 年の TIDES に始まり、GALE [1)]では IBM および BBN、SRI の三つのチームによる競争的な研究開発が行われ、規模は縮小したものの現在の BOLT [2) 3)]につながっている。競争的な研究開発と同時に、世界各国からの参加が可能なオープンな評価型ワークショップを開催 [4)]することで機械翻訳の研究開発を促した。また、90 年代後半から、National Science Foundation (NSF)をスポンサーとして JHU Workshop [5)]が毎年開催され、機械翻訳を含む、音声認識および言語処理の先進的な研究課題に対し、学部生や院生、研究機関の研究者などさまざまなバックグラウンドを持つ人材でチームを組み、その問題解決に取り組んできた。

欧州では 2006 年から FP6 の EuroMatrix [6)]の下で、オープンソースの統計的機械翻訳ツール Moses の開発が始まった。Moses は研究のためのツールとして広く使われるだけでなく、実際のサービスに応用されている。また、定期的にハッカソン（MT Marathon）を行い [7)]、最先端の機械翻訳の研究成果を実装するだけでなく、新しい機能を提案して継続的に開発が進められている。

日本では、欧米と比較して大きなファンディングはなく、研究所あるいは大学の研究室レベルでしかなかった。また、特に日英翻訳では、対訳データが少ないという問題に加え、日本語および英語が文法的に非常に異なる言語対であるという難しさもあった。NTCIR-7 および8、9の特許翻訳タスク [8)]では、日英で同時に出願された特許から対訳データを作成し、

非常に大きな対訳データを自動的に構築可能となった。また、文法の違いを吸収するため、構文情報を利用し、あらかじめ原言語の入力文を目的言語の語順に並び替えることで性能向上を果たしている。

機械翻訳の応用として、音声認識と音声合成の技術を組み合わせた音声翻訳の研究開発は古くから行われ、C-STAR の枠組みにより日米欧共同で試みられてきた [9]。各国の研究機関で協力して多言語音声翻訳を実現する枠組みは、FP6 の TC-STAR [10]や A-STAR [11]、U-STAR [12]へと受け継がれている。同時に 2004 年より評価型ワークショップ IWSLT を開催し、特に音声翻訳技術への発展に貢献している [13]。

［今後必要となる取組み］

機械翻訳はある特定の分野で大量の対訳データがあれば精度の高い翻訳を実現可能であることが示されている。ところが他の言語や別の分野へと適用するたびに対訳データを必要とする。今後、機械翻訳システムの多言語化および多分野化を進めていく時、量による高精度な翻訳を追求すると同時に、構文解析や意味解析などの、より深い解析技術を組み合わせた、量によらない翻訳技術が必要となる。また、構文解析器や意味解析器が存在しない言語に対しても高精度な翻訳を実現するため、教師なし学習などの機械学習を取り入れ、翻訳に必要な情報を自動的に抽出する技術が求められる。

現在は文単位に翻訳を行っているが、文単位では意味の曖昧性を解消できない。今後は文章全体から参照関係を捉え、文章の意味を解析した上で、各文へと反映させることによる高精度な機械翻訳が求められる。また音声翻訳の延長線上では、同時通訳を可能とする逐次翻訳の技術開発が必要とされる。

（４）科学技術的・政策的課題

大規模対訳データから翻訳のための知識を抽出する手法が確立されてきており、日英など言語資源が豊富な言語対では精度の高い機械翻訳が実現されつつある。ところが多言語化あるいは多分野化を進めるとき、言語資源が乏しい言語、あるいは分野では高精度な翻訳が難しく、今後の課題である。

機械翻訳は構文解析などの自然言語処理の基礎技術だけでなく、機械学習やデータ構造など数多くの要素技術から構成されており、新たに機械翻訳の研究を始めるには非常に多くのことを勉強しないといけない。この複雑さのため全体像を把握して研究開発を進めることができる研究者は限られ、例えば大学の研究室レベルでの研究は非常に限定的とならざるをえない。このため、欧米にならい、強いファンディングおよびリーダーシップのもとで各大学の研究室や研究所等が何らかの形で連携を強める必要がある。

評価型ワークショップ等により最先端の技術をシステマティックに評価するだけでなく、オープンソース化などでその技術を幅広く普及する仕組みの充実が求められる。また、情報収集や情報発信、同時通訳など、機械翻訳システムが実際に使われる現場を想定し、翻訳結果の評価手法について吟味する必要がある。

（５）注目動向（新たな知見や新技術の創出、大規模プロジェクトの動向など）

［新たな技術動向］

日英など大幅に文法が異なる言語対に対して機械翻訳の性能は限定的であったが、原言語の入力文を構文解析し、あらかじめ目的言語の構造へと並び替える手法により大幅な性能向上を果たすことが可能になった。ただし、構文解析の精度に依存した手法であり、口語体など、文法を逸脱した文に対して効果は限定的であるため、今後の発展は限定的と思われる。

近年、ニューラルネットワークの技術を応用した深層学習が音声認識や画像認識などの人工知能の分野で大幅な性能向上に貢献している。深層学習の利点の一つとして自動的な特徴量の抽出にあり、ベクトルで表された単語の意味表現を大規模データから学習することで、曖昧性の解消などへと応用可能であり、今後、機械翻訳の分野でも大きく発展する余地があると思われる。

［注目すべきプロジェクト］

DARPA は古くから機械翻訳に対して大規模に投資を行ってきたが、現在実施されている BOLT [2]では、過去の TIDES や GALE と比較してその予算規模は縮小されている。機械翻訳プロジェクトを実施している NIST は、研究開発を進めるために、同時に機械翻訳システムのオープンな評価を行う評価型ワークショップ OpenMT [4]を主導してきた。この数年、実施されていないが、2015 年に久しぶりに実施されることになり、今後の発展が興味深い。

欧州の FP7 で実施されてきた EuroMatrixPlus [14]、およびその後継 MosesCore [7]は 2015 年に終了する。このプロジェクトではオープンソースの統計的機械翻訳ツール Moses の研究開発を行うと同時に、ヨーロッパ言語間の翻訳を中心とした評価型ワークショップ WMT (Workshop on Statistical Machine Translation) を運営し、研究開発を促している。Moses は実際のサービスにも応用されており、今後もプロジェクトは引き継がれると思われる。EU-BRIDGE は音声翻訳に重点をおいたプロジェクトであり、現在 IWSLT を主体的に運営している [15]。

中国では CWMT (China Workshop on MT) [16]を通して、特に中国語の翻訳を重点的に大学や企業での活発な研究開発を継続的に行ってきた。このような評価型ワークショップでは言語資源を集め、同じ条件で各機械翻訳システムを評価すると同時に、最新の技術を共有することでさらに基礎研究が進むと考えられる。

情報通信研究機構では「先進的音声翻訳研究開発推進センター」を立ち上げ、最先端の音声翻訳技術の研究開発を開始した [17]。2020 年の東京オリンピックに向けて多言語による円滑なコミュニケーションを目指す試みであり、今後の進展が注目される。

（６）キーワード

統計的機械翻訳、自然言語処理、機械学習、人工知能

（７）国際比較

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	◎	→	機械翻訳の基礎研究に関しては、NICTおよびNAISTが突出して強く、トップ国際会議や論文誌などに数多く採択されている。
	応用研究・開発	○	→	NICTはB2Bに特化し、特許庁など他の企業と共同で機械翻訳システムの研究開発を継続的に行っている。JSTは日本、中国の研究機関と協力し日中・中日機械翻訳実用化プロジェクトを推進している。NTTドコモは他の企業と共同出資した「みらい翻訳」によりさらに高精度な翻訳を目指して開発を行っている。
	産業化	◎	↗	NTTドコモによる携帯端末や、NICTによるVoiceTraやTexTraなどのスマートフォンのアプリで音声翻訳のサービスを実施中。
米国	基礎研究	○	→	BOLTで予算規模が縮小したものの、ISI (Information Sciences Institute)やCMU、UMDは継続的に機械翻訳の基礎研究を行っている。また、JHU Workshop (Frederick Jelinek Memorial Workshopへ名称を変更)により研究機関が連携し積極的な研究活動を行っている。
	応用研究・開発	○	→	BBN、IBM、SRIは継続的にDARPA向けの機械翻訳システムの開発行っている。また、SDLはB2Bでカスタマイズされた機械翻訳システムの提供をしている。
	産業化	◎	→	Googleは2007年より統計的機械翻訳のサービスを開始したパイオニアであり、アプリの提供も行っている。MicrosoftはBingだけでなくSkypeにも機械翻訳を提供している。
欧州	基礎研究	○	→	統計的機械翻訳ツールキットMosesを中心に継続的に基礎研究を行っている。
	応用研究・開発	○	→	MosesCore[7]によりMosesの宣伝および普及活動をするとともに、評価型ワークショップ WMTやハッカソン MT Marathonなどを主催し、積極的に応用研究を進めている。EU-BRIDGEではIWSLTを主催し、音声翻訳技術の開発を行っている。
	産業化	○	→	Mosesは商用のシステムとして実際のサービスに用いられている。
中国	基礎研究	◎	↗	中国科学院やMSRAなどの研究機関はトップ国際会議に数多く採択され、最先端の機械学習技術を取り入れると同時に、評価型ワークショップCWMTを中心に最先端の研究の普及活動をしている。
	応用研究・開発	○	→	MSRAは継続的に機械翻訳の研究開発を行ってきている。Baiduは2013年に北京、2014年にシリコンバレーに人工知能の研究所を立ち上げ、機械翻訳の研究開発を始めた。 各大学の研究室では応用研究を行っているものの、その成果は外からはわかりにくい。
	産業化	○	→	Baiduは機械翻訳のサービスを行っている。
韓国	基礎研究	△	→	KAISTなどが基礎研究を行っているが特に活発な研究発表はない。
	応用研究・開発	△	→	基礎研究同様、特に活発な応用研究は見られない。
	産業化	△	→	欧州で古くから機械翻訳の開発を行ってきたSystranを韓国のベンチャー企業CSLiが買収した。CSLiは携帯端末での機械翻訳アプリを提供しているが、Systranはアジアの言語に特に強いとはいえず、今後どのように発展するかは未知である。

（註 1）フェーズ
　　基礎研究フェーズ　：大学・国研などでの基礎研究のレベル
　　応用研究・開発フェーズ　：研究・技術開発（プロトタイプの開発含む）のレベル
　　産業化フェーズ　：量産技術・製品展開力のレベル
（註 2）現状
　※わが国の現状を基準にした相対評価ではなく、絶対評価である。
　　◎：他国に比べて顕著な活動・成果が見えている、　○：ある程度の活動・成果が見えている、
　　△：他国に比べて顕著な活動・成果が見えていない、×：特筆すべき活動・成果が見えていない
（註 3）トレンド
　　↗　：上昇傾向、　→：現状維持、　↘：下降傾向

（８）引用資料

1) http://www.itl.nist.gov/iad/mig/tests/gale/
2) http://www.darpa.mil/Our_Work/I2O/Programs/Broad_Operational_Language_Translation_(BOLT).aspx
3) http://www.nist.gov/itl/iad/mig/bolt.cfm
4) http://www.nist.gov/itl/iad/mig/openmt.cfm
5) http://www.clsp.jhu.edu/workshops/
6) http://www.euromatrix.net/
7) http://www.statmt.org/mosescore/
8) http://ntcir.nii.ac.jp
9) Takezawa, Toshiyuki, Morimoto, Tsuyoshi, Sagisaka, Yoshinori, Campbell, Nick, Iida, Hitoshi, Sugaya, Fumiaki, Yokoo, Akio and Yamamoto, Seiichi. 1998. "A Japanese-to-English speech translation system: ATR-MATRIX.." In the Proceedings of the ICSLP.
10) http://tcstar.org
11) http://www.mastar.jp/AStar/
12) http://www.ustar-consortium.com
13) http://www2.nict.go.jp/univ-com/multi_trans/WS/IWSLT2004/
14) http://www.euromatrixplus.net
15) http://www.eu-bridge.eu
16) Sitong Yang, Heng Yu, Hongmei Zhao, Qun Liu and Yajuan lv. 2014. “Review and Analysis of China Workshop on Machine Translation 2013 Evaluation”. In the Proceedings of AMTA2014.
17) http://www.nict.go.jp
18) 国際会議 ACL や NAACL、EMNLP あるいは論文誌 TACL (http://anthology.aclweb.org を参照)
19) http://www.nict.go.jp/press/2014/07/28-1.html
20) http://foresight.jst.go.jp/jazh_zhja_mt/
21) https://www.nttdocomo.co.jp/info/news_release/2014/09/29_01.html
22) http://www.darpa.mil/opencatalog/BOLT.html
23) http://www.sdl.com
24) http://ir.baidu.com/phoenix.zhtml?c=188488&p=irol-newsArticle&ID=1931950
25) http://www.systransoft.com/download/press-releases/systran-pr-csli-acquisition-20140425.pdf

３．８．３　言語情報処理応用（音声対話）

（１）研究開発領域名

言語情報処理応用（音声対話）

（２）研究開発領域の簡潔な説明

人間同士の音声コミュニケーションの分析や人間とコンピューターの音声による自然なやり取りを実現するための研究開発

（３）研究開発領域の詳細な説明と国内外の動向

人間のコミュニケーション手段で最も古くから用いられているものは音声によるやり取り、すなわち、音声対話である。情報、意思、感情の伝達など、人間は音声対話をさまざまな用途に用いている。音声対話をコンピューターで処理可能にすることは、人間同士の膨大なやり取りの情報にコンピューターがアクセスできるようになることのみならず、人間と共同で知的活動を行うシステムの構築につながる。

音声対話を行うコンピューターを実現するためには、まず、相手の音声を音声認識によってコンピューターが理解可能なシンボル列（例えば，認識文字列）に変換する必要がある。そして、そのシンボル列がどのような意味を持つかを解釈しなくてはならない。発話は単独で意味を持つ場合もあるが、一般には文脈によって解釈が変わる。そのため、発話は文脈に基づいて解釈する必要がある。相手に応答するためには、コンピューターは自身の意図に基づき発話を生成して、相手に音声で伝える必要がある。このとき、意図が的確に反映されるように、声の大きさや抑揚も調整する必要がある。音声対話分野は、人工知能の主要分野である信号処理、音声処理、自然言語処理だけでなく、音声知覚や感情理解に関わる脳科学や人間科学、物理的なインターフェースを実現するためのロボット工学や人間工学、社会において人間と円滑にやり取りを行うための社会学・心理学・認知心理学など複数の分野の融合領域である。

研究の歴史をひもとくと、音声認識技術が一定のレベルに達する 1990 年頃まで、研究の中心は対話理論の構築であった [1)]。例えば、対話で起こる現象を議論する上で必要な基本概念や人間同士の対話のモデル・理論が提案された。発話を物理的な行為とみなす発話行為論、対話を意図のレベルで構造化する談話構造理論などが有名である。1990 年代に音声認識の性能が改善されると、国内外で音声対話システムが構築され始めた。最も大きなインパクトを与えたのは、米国の DARPA 主導による、ATIS や Communicator といった巨大プロジェクトである。この成果により、フライト情報案内などの所定のタスクを遂行するタスク指向型音声対話システムの技術が確立された [2)3)]。同時期に、音声対話技術の IVR（音声自動応答）への適用も開始された [4)]。

2000 年代に入り、教師データからコンピューターに判断を学習させる機械学習が流行し、データさえあれば研究者でなくてもタスク指向型の音声対話システムが構築できる環境が整った。2010 年代には、スマートフォンの普及とクラウド型の音声認識により、商用の音声エージェントサービスが次々とリリースされるようになった [5)6)]。

現在、パーソナルロボット、車載、医療などの分野において音声対話への期待は高く、国内外で産業界から巨額の資金が投入されている。利用者増によるデータの集積は精度改善に

重要であるため、シェア争いが今後ますます激化すると予想される。大量の音声対話データを対象とした分析や情報抽出の研究も進められている。一般話者の自由発話（コンピューターに向かって話すようなコマンド発話ではなく人間同士の自由な発話）に対する音声認識精度はいまだ低く、実用にはまだ遠いと考えられているが、家庭内の会話や、通話、ミーティングの分析など幅広い用途が考えられ、この分野はこれから重要になると予想される。

（４）科学技術的・政策的課題

音声対話は人間にとって自然なコミュニケーション手段であるため、コンピューターにとっての難しさが理解されにくいが、人工知能の困難な課題を扱う分野である。しかし、ひとたび実用化のめどが立てば、人間同士のやり取りのほぼすべてに関わる分野であり、その適用範囲は非常に大きい。2020 年の東京オリンピックにおける外国人観光客の対応だけでなく、産業用ロボットの入出力、教育支援、介護ロボット、認知症予防、独居世帯のコミュニケーション支援など、少子高齢化が進む日本において音声対話システムの役割はますます大きくなると予想される。音声対話はわれわれの生活を根底から変える可能性のある分野であると認識して投資を行う必要がある。

音声対話サービスは乱立しているが、キラーアプリはまだ手探りの段階であり、日本が世界での競争に勝てる可能性は十分ある。ただ、音声対話システムは非常に多くの部分からなる複雑なシステムであり、商用レベルのシステムを構築し運用するためには、大量の学習データ、計算機資源、コンテンツ、高度な学習ソフトウエアが必要である。ブレークスルーの可能性を高めるためには、国内の多くの研究者がこれらのリソースを安価に使えるための施策が必要である。

音声対話システムがパターン認識の研究と大きく異なるのはその評価の難しさである。例えば、音声認識であれば、音声データとその書き起こしからなる評価データを用いて音声認識精度を計算できる。しかし、音声対話システムはやり取りを扱うため、評価をしようと思うと「やり取り」とその「良さ」を対応づけた評価データが必要となる。ここで、やり取りはシステムとユーザーの複数の発話であるため、やり取りが長くなればなるほど組み合わせが爆発し、適切な評価データが作れないという問題が起こる。これはビッグデータを用いても解決することが難しい。また、対話の良さについても、主観の入る要素が大きいという問題がある。現状、音声対話システムを評価するためには、ユーザーに実際に使ってもらい主観評価を積み重ねていくしかない。実ユーザーによる評価実験を行うのは多大なコストがかかる。実サービスを行う機関でなくても実証実験が可能となるような仕組み作りが必要である。なお、音声対話データは個人の声であることや話されている内容がプライベートであることが多いことから、インターネットのクロールデータよりも扱いが困難である。実サービスや実証実験によって得られるデータを複数の機関で利用しやすくするための法整備も必要である。

音声対話を用いたシステムが既に商用でサービスされていることから、音声対話の多くの問題が既に解決したと考えがちであるが、それは誤りである。コールセンターやミーティングにおける音声認識率が低いことからも分かるとおり、人間同士の自由発話はまだまだ処理が難しい [7]。また、現状の音声対話システムは決まり切ったタスクしか実行できないタスク指向型対話システムであり用途には限界がある。人間同士の会話を、文脈を踏まえて適切に

理解したり、人間の知的活動を支援したりできる音声エージェント実現のためには、依然、基礎研究から応用研究まで幅広い支援が必要である。なお、音声対話の研究は分野横断的であるが、複数の分野の研究者が共同で研究を行う営みは少ない。複数の分野の研究者を共同研究させる政策が必要である。

（５）注目動向（新たな知見や新技術の創出、大規模プロジェクトの動向など）

音声対話ビジネスが急速に勢いを増している。法人向けでは、Nuance を主要ベンダとして、車載、医療、IVR、CRM（顧客関係管理）、speech analytics (音声データの分析)の分野が活発である。IBM の人工知能である Watson もコールセンターの会話分析に用いられ始めた [8]。個人向けでは、Apple の Siri、Google の Google Now、Microsoft の Cortana、NTT ドコモの「しゃべってコンシェル」などの音声エージェントが一定の支持を得ている。Amazon は Amazon Dash や Amazon Echo という、日常生活を会話によってサポートしたり、商品名を話すだけで注文ができたりするデバイスを販売している。家庭用ロボットの商用化も始まっている。国内ではシャープが会話できる掃除機である COCOROBO を販売している。また、ソフトバンク社がパーソナルロボット Pepper を 2015 年に発売する。 MIT 発のベンチャーは 2015 年末に家庭用アシスタントロボット Jibo [9]を発売予定である。音声対話技術が日常に入り込みつつある。

研究面では、産業界の動きを反映して、実用的な音声対話システムに向けた動きが多い。Dialogue State Tracking Challenge [10] と呼ばれる共通タスクが Microsoft やケンブリッジ大の研究者を中心に開催され、タスク指向型対話システムにおけるロバストな意図理解の手法が競われている。また、Real Challenge [11]では、実ユーザーに実際に利用される音声対話システムのアイデアが競われている。GPS、スマートフォン、Google Glass、Kinect、iWatch などのデバイスから得られるセンサー情報を用いてユーザーの状況を適切に理解し、高度な対話処理を実現する研究も増加している。この分野は Situated Dialogue と呼ばれ、国際ワークショップ [12]や国際会議の特別セッション [13]も開かれている。扱われるアプリケーションとしては、カーナビゲーションや歩行者の案内システムが多い。

国内では、ヒト型ロボットの研究が盛んなこともあり、コンピューターとの自由な会話を実現するシステムの研究が増加している。Project Next NLP[14]と呼ばれる自然言語処理のプロジェクトの中では、産学合わせた 15 の機関が人間とコンピューターの雑談の分析を共同で行っている [15]。また、NII 主催のグランドチャレンジとして「ロボットは井戸端会議に入れるか」プロジェクトが立ち上がっている [16]。CREST や ERATO が支援する対話関連プロジェクトにおいても雑談の実現が目的に含まれている。なお、2014 年 6 月、チューリングテストに Eugene と呼ばれるチャットボットが始めて合格したと大きく報じられたが [17]、このシステムは数分間の対話を人手による膨大なルールで実現するものであり、本質的なブレークスルーではない。海外では、雑談よりも実際的な、交渉や議論といったやり取りが可能な音声対話システムを実現しようとする動きが強まっている。米国では軍がこの取組みを支援している [18]。

（６）キーワード

音声対話システム、音声エージェント、音声認識、自然言語処理、人工知能、パーソナルロボット、雑談、Situated Dialogue

（７）国際比較

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	○	↗	・大学・公的機関における基礎研究のレベルは高い。主要な国際会議では、米国とは差があるものの、英国、ドイツと並ぶ論文数が採択されている。NTT、NICT、HRI-JP（ホンダ・リサーチ・インスティチュート・ジャパン）、京大、阪大、奈良先端大が有力機関である。 ・対話システムシンポジウム、人工知能学会における特別セッション、「知的対話システム」の論文特集が企画されるなど、本分野の学会活動は活性化している。 ・CRESTのuDialogue、ERATOの石黒共生ヒューマンロボットインタラクションプロジェクト、Project Next NLPにおける対話タスク、NIIのグランドチャレンジの「井戸ロボ」などの対話関連プロジェクトが立ち上がっている。
	応用研究・開発	○	↗	・トヨタ・ホンダなどの自動車メーカー、シャープや東芝などの家電メーカー、NTTドコモやソフトバンクなどの通信会社が音声関連サービスの研究開発を進めている。CEATEC Japan 2014でも関連展示が多く見られた。 ・2020年に向けて、NICTが多言語観光案内音声対話システム開発のための取組みを始めている[19]。
	産業化	○	↗	・NTTドコモの「しゃべってコンシェル」、Yahoo!の「音声アシスト」など、音声エージェントが一般に広く利用されるようになっている。特に、しゃべってコンシェルでは数百のキャラクターとの音声対話が可能となっており若年層からも人気を集めている。 ・シャープのCOCOROBOやソフトバンクのPepperなどのパーソナルロボットが一般に販売されている。
米国	基礎研究	◎	→	・主要な国際会議における論文数では米国はトップである。 トップカンファレンスであるSIGDIAL2014[20]では論文の約半数が米国の研究機関によるものであった。カーネギーメロン大学、南カリフォルニア大学、Microsoftなどが有力機関である。 ・Dialogue state tracking challenge や Real challenge などの共通タスクは米国が牽引しており研究トレンドを生み出している。
	応用研究・開発	◎	↗	・医療分野では、DARPAが兵士やその家族の精神的サポートを行うSimSensei [18]と呼ばれるシステムを巨額なファンドで支援しており、実ユーザーを用いた実証実験が行われている。また、病院において看護師を支援するシステムの実証実験も進められている[21]。 ・Google、Microsoftなどの検索大手は自社サービスに向けた音声認識・音声理解技術の研究開発を進めている。Microsoftは Skype通話の解析も進めており、通話の自動翻訳サービスを予定している[22]。
	産業化	◎	↗	・Amazon によるAmazon DashやAmazon Echo、Nuanceによる車載・医療・ウェアラブル・IVR向けサービス、家庭用アシスタントロボットJiboなど多くの商品が発売されている。 ・音声ベンダのエキスポであるSpeechTEK 2014では400名を超える関係者が集まった。Nuanceの売り上げは17億ドルにも上り[23]、音声ビジネスを牽引している。 AT&T、SpeechProなどの多くの音声SIベンダが活発に活動している。 ・AppleのSiri、GoogleのGoogle Now、MicrosoftのCortanaなどの音声エージェントが出そろい、シェアの獲得競争を行っている。

欧州	基礎研究	◎	↗	・SIGDIAL 2014では欧州が米国に次いで発表件数が多い。特に、英国とドイツの発表が多く、英国ではケンブリッジ大学、ヘリオットワット大学、ドイツではビーレフェルト大学、ウルム大学が有力研究機関である。 ・欧州は、対話のモデル化、ドメイン適応、強化学習を用いた統計的対話制御などの基礎理論に関する研究が強い。近年の統計的対話制御に関する文献の大部分はケンブリッジ大のSteve Young教授、ヘリオットワット大のOliver Lemon教授の研究室から発表されている。 ・FP7が基礎研究を強く支援しており、ここ数年で10件近いプロジェクトが進行している。European Research Counselも先駆的な音声対話システムのプロジェクトを支援している[24]。2015年からはHORIZON2020の枠組みで同様の支援が継続される見込みである。
	応用研究・開発	○	→	・FP7のプロジェクトは企業も多く参画する。大規模な対話ストリームの分析を行うSENSEI[25]ではエリクソンやSAPを始め多くの企業が参画している。FP7ではシステム開発に資するプロジェクトも一定数あり、Port dial[26]は多言語化、DIRHA[27]は家庭内音声対話、PARLANCE[28]は実用性のある音声対話システムの構築が目標である。 ・ドイツでは、車載関連やユーザビリティ評価についての研究が多い。ダイムラーはウルム大と共同で運転中でも安全に利用できる音声エージェントの開発を行っている[29]。また、TU-Berlinは音声サービスの評価に関わる研究を積極的に進めている[30]。
	産業化	△	→	・欧州では産業界が米国ほど活発ではなく、産業化について目立った動きはない。しかし、基礎研究が強く、また、公的なファンドが安定して供給されていることから、今後産業化の動きが活発化する可能性がある。
中国	基礎研究	△	→	・公的ファンドとしてNational Natural Science Foundation of China (NSFC) のExcellent Young Researcher Project が音声対話研究を支援している。 ・国際会議でのプレゼンスは低いが、海外に留学している研究者も多く、基礎研究のレベルが低いとはいえない。 ・主要な研究機関としては、上海交通大学のKai Yu教授の研究室とChinese Academy of Sciencesがある。
	応用研究・開発	△	→	・BaiduやiFlyTekが音声エージェントの研究開発を進めている。Microsoft Research Asiaも対話システムの検討を進めている[31]。
	産業化	○	↗	・Baidu Voice Assistant とiFlyTekのYuDianが一般に利用されている。iFlyTekの中国語の音声認識性能は高く、YuDianの評判は非常に良い。
韓国	基礎研究	△	↗	・浦項工科大学校と西江大学校が主要研究機関である。浦項工科大学校のLee教授は韓国の音声対話を牽引する人物で国際的な知名度も高い。 ・公的ファンドによるプロジェクトDeveloping Personal assistant Software Platform with Multi-domain Dialogue Interfaceが進行中である。西江大学校のSeo教授が研究代表者である。 ・国際会議においては、韓国の発表件数は多くないが一定のプレゼンスは保っている。
	応用研究・開発	○	↗	・Samsung、LG、SK telecom、KT、NAVER、Daum がそれぞれの業務分野において音声を用いたアプリケーションの研究開発を活発に進めている。
	産業化	○	↗	・音声対話が可能なデバイスとして、SamsungはSmart TVやS-Voice をサービスしている。LGはSmart Watchを発売している。

（註1）フェーズ
基礎研究フェーズ ：大学・国研などでの基礎研究のレベル
応用研究・開発フェーズ ：研究・技術開発（プロトタイプの開発含む）のレベル
産業化フェーズ ：量産技術・製品展開力のレベル
（註2）現状

※わが国の現状を基準にした相対評価ではなく、絶対評価である。
◎：他国に比べて顕著な活動・成果が見えている、　○：ある程度の活動・成果が見えている、
△：他国に比べて顕著な活動・成果が見えていない、×：特筆すべき活動・成果が見えていない
（註 3）トレンド
↗ ：上昇傾向、　→：現状維持、　↘：下降傾向

（８）引用資料

1) 石崎雅人、伝康晴：談話と対話、東京大学出版会、2001。
2) 河原達也、荒木雅弘：音声対話システム、オーム社、2006。
3) 河原達也：音声対話システムの進化と淘汰-歴史と最近の技術動向-、人工知能学会誌、vol.28、No.1、pp.45-51，2013。
4) http://www.corp.att.com/attlabs/reputation/timeline/01hmihy.html
5) https://www.apple.com/jp/ios/siri/
6) https://www.nttdocomo.co.jp/service/information/shabette_concier/
7) 島津明、堂坂浩二、川森雅仁、中野幹生：話し言葉対話の計算モデル、電子情報通信学会、2014。
8) http://www-06.ibm.com/jp/press/2014/11/0602.html
9) http://www.myjibo.com/
10) http://research.microsoft.com/en-us/events/dstc/
11) https://dialrc.org/realchallenge/
12) http://www.uni-ulm.de/in/iwsds2014.html
13) http://www.interspeech2014.org/public.php?page=special_sessions.html#open-domain-situated
14) https://sites.google.com/site/projectnextnlp/
15) 東中竜一郎、船越孝太郎：Project Next NLP 対話タスクにおける雑談対話データの収集と対話破綻アノテーション、人工知能学会研究会資料 SIG-SLUD-72，2014。
16) http://research.nii.ac.jp/~bono/ja/aboutus/theme03.html
17) http://www.reading.ac.uk/news-and-events/releases/PR583836.aspx
18) http://ict.usc.edu/prototypes/simsensei/
19) 水上悦雄、岡本拓真、堀智織：多言語音声翻訳・対話システム構築ツールの公開に向けて、人工知能学会研究会資料 SIG-SLUD-72，2014。
20) http://www.sigdial.org/workshops/conference15/
21) http://relationalagents.com/projects/2.html
22) http://research.microsoft.com/en-us/about/speech-to-speech-milestones.aspx
23) http://www.nuance.com/company/company-overview/fast-facts/index.htm
24) http://www.irit.fr/STAC/
25) http://www.sensei-conversation.eu/
26) https://sites.google.com/site/portdial2/
27) http://dirha.fbk.eu/
28) https://sites.google.com/site/parlanceprojectofficial/
29) Hansjörg Hofmann, Mario Hermanutz, Vanessa Tobisch, Ute Ehrlich, André Berton, Wolfgang Minker: Evaluation of In-Car SDS Notification Concepts for Incoming Proactive Events,

Proc. IWSDS, 2014.

30) http://www.qu.tu-berlin.de/menue/qu/parameter/en/

31) http://www.msxiaoice.com/

３．８．４　画像・映像の意味解析

（１）研究開発領域名

画像・映像の意味解析

（２）研究開発領域の簡潔な説明

画像・映像等の視覚情報を対象とした、人間と知能機械のコミュニケーションを行うための表現メディアの解析・意味処理とその検索・変換・編集に関する研究開発。マルチメディアコンテンツのメタデータ付与、検索、利活用等を含む。

（３）研究開発領域の詳細な説明と国内外の動向

画像・映像処理とは、画像・映像をデータとして計算機で処理する技術のことを指し、特にその意味解析とは、画像・映像の意味内容を計算機により推定する等、画像・映像の意味内容に基づいた計算機処理を指す。そもそも画像・映像は、「百聞は一見にしかず」というように、人間にとっては、極めて情報豊富で、包括的な情報源であって、その意味内容も容易に理解可能であり、コミュニケーションにおける表現メディアとしても極めて有効である。加えて、画像・映像処理、特に意味解析への要望は強く、顔による入退室管理、郵便番号を含む紙ドキュメントの文字認識、Web 上の画像・映像検索、e-コマースや画像・映像提供サービスにおける検索や推薦、放送映像やオンデマンド番組提供における検索、市場調査における特定商品の放送映像・ネット（ブログやツイッターを含む)における画像としての露出調査、監視カメラ映像解析、農場における作物生育状況の自動監視等、その応用は極めて多岐にわたる。計算機処理により高速化・大規模化が可能となり、画像・映像の意味解析技術への期待が高まっている。

しかしながら、数値やテキストデータに比べ、画像・映像の計算機処理は一般に困難で、特に意味解析は極めて困難であり、人間が行うような柔軟な画像・映像の意味理解には通常遠く及ばない。現状の技術では、特定の対象に対する特定のタスクについて研究開発を行い、成果を挙げている。特に、視線方向の変動による透視投影変換歪み、照明条件の変動、物体の運動や変形等による視覚情報の不定性に対応した視覚特徴量の開発と、機械学習技術の応用により、当該対象・タスクに対応した大規模学習データに基づいた戦略的な技術開発が、いくつかの対象において功を奏している。

［これまでの取組み］

まずは、社会的要請の高さから、画像・映像中の顔の検出・照合・認識、ならびに（印刷と手書き両方を含む）文字画像の認識に関する研究が突出して進んでいる。入退室管理から、iPhoto や Picasa のような個人用写真管理ツールでも顔認識技術は実用化されており、文字認識も古くは郵便番号の自動読み取りシステムから、名刺管理システムや Evernote のような個人用情報管理ツールでも文字認識技術が利用されており、実用レベルにある。わが国の技術レベルも世界トップクラスである。顔認識については、そもそも画像入力から認識まで首尾一貫して計算機で実現した世界最初のシステムは CMU の金出武雄による京大博士論文の研究（1973 年）であるといわれている。現在も、米国標準技術局(NIST)による顔認証技術ベンチマーク(Face Recognition Vender Test: FRVT [8])において、2013 年に NEC が世界

第１位の認識性能を達成している。オムロンは１万人の顔をさまざまな表情・方向・照明条件等で撮影した大量の画像データベースをもとに顔認識関連技術 OKAO Vision を開発し、トップレベルの性能ならびに世界トップレベルのシェアを達成している[9]。

顔・文字の解析における研究黎明期には NIST 等による研究用データの整備（顔データベース FERET[10]、文字データベース MNIST[11]等）、大学も含む研究の進展が見られ、特に日本においても政府によるプロジェクト推進（1971 年より 10 年間、通産省工技院を中心としたパターン情報処理大型プロジェクト）、電子技術総合研究所（現、産業技術総合研究所）等によるデータ整備（ETL 文字データベース[12]）等により、特段の技術の進展があった。やがて、産業としての価値が認められると、各企業による独自のデータ整備と研究が進展するようになり、大学等は追従しにくくなる。日本でも、顔・文字については、このようにして企業による技術の進展が見られる。

一般の画像・映像を対象とした意味解析では、映っている事物の認識、シーンの分類、動作の認識等が目的であり、インターネット上の画像・映像の検索、e-コマースにおける物品画像の検索・推薦等を始め、さまざまな応用がある。こうした研究開発のため、Caltech(米)[13]、PASCAL VOC(欧)[6]、TRECVID(米)[4]、ImageNet(米)[1]等、大学や公的資金援助による大規模データセットが構築され、大学も含み研究が進展してきている。日本でも PASCAL や TRECVID において東大、東工大、国立情報学研究所、NTT 研究所等により世界的にも競争力のある技術が開発されている。日本においては、一般画像の意味解析を利用する産業応用(画像検索等)の素地ができておらず、企業の参加が鈍い。海外においては、Microsoft (Bing)、Yahoo!、アリババ・グループ、Baidu 等においてこうしたビジネスが立ち上がっており、画像検索のクリックログ等の独占利用データができつつあり、大学等こうしたデータを持たない研究機関による追従が難しくなりつつある。一般画像・映像の意味解析でも deep neural network (DNN)により高性能が達成されてきているが、DNN の先導的な研究者もこうした企業による囲い込みが進んでいる(DNN 生みの親の Geoffrey E. Hinton 教授は Google に、Yann LeCun 教授は Facebook に所属)。欧州においては、米国等による技術の独占先行を阻むために重要であるとの考えから、FP6/FP7 において画像・映像等のマルチメディア内容解析技術の特段の推進が図られ、多くの研究機関において特段の技術的な躍進が図られた。日本においては、この点でも弱い。

［今後必要となる取組み］

今やインターネットにおける通信量も大部分が映像であり、いわゆるビッグデータにおいても、現状では数値ならびにテキストデータの解析がほとんどであって、画像・映像のビッグデータ的な解析は未開拓の領域といえる。また、画像・映像は人間には理解しやすいが計算機には「理解」できないので、感知されないように情報をやり取りするメディアとしては最適であり、米国では画像・映像意味解析をホームランドセキュリティーのための重要な技術と位置付けていると考えられる。画像・映像の意味解析技術は、ビッグデータ、Web 検索、監視カメラ映像解析等、安心安全のための IT、農場自動調査等グリーンイノベーション、ライフサイエンス等、ありとあらゆる分野で必要とされる基盤技術である。今後、特に顔・文字に限らない一般の画像・映像の意味解析技術の研究開発の推進が重要と考えられるが、わが国においては、こうした技術を主として扱う産業の素地が脆弱であり、加えてそれ

を補完するような、画像・映像の意味解析を中心に据えた国プロ等の施策が少ない。当該技術の推進のためには、こうした点に対する対策が必要である。

（４）科学技術的・政策的課題

・大規模データの収集: 多様性を有する画像・映像解析の研究開発には、実際の対象と同等の特性を持ち、実際の多様性を反映する十分大規模な研究開発用データの収集が重要となる。昨今は、特に、インターネットから容易に収集可能なデータも多く、技術的には研究開発用の大規模データの収集は容易になってきているが、こうして構築したデータを研究コミュニティーに公開するのは、著作権・肖像権・個人情報保護法等の問題により難しい。米国ではフェアユース規定によりこうした活動が広く行われているが、日本においては一般により困難である。その一方、検索エンジンや e-コマース、SNS 等を有する企業による独自のデータを用いた開発も進んできている。例えば、DNN を用いた顔認識技術 deepface が Facebook の研究者らにより開発され、人間が行うのと同等の認識性能が報告された[14]が、その学習には Facebook が保有するデータから得た 440 万顔画像が利用されている。

・政策的な研究開発の底上げ: IT メディア分野の産業においては、特に研究開発から展開までの速さ、世界的な展開の容易さ等から、一つないしはごく少数の企業のみが「勝ち残る」ケースが多い。こうした「勝者」が米国に集中している状況では、米国以外では当該分野の研究開発が、自然発生的に世界的にも存在感を示すほど進展することは考えにくい。これを打破するためには、欧州における FP7 のマルチメディア内容解析技術への取組み、中国・韓国における国家的な研究推進戦略の一定の成功を見るに、日本でも政策的な研究開発の底上げが重要と思われる。

（５）注目動向（新たな知見や新技術の創出、大規模プロジェクトの動向など）

・ImageNet [1]は、米国スタンフォード大 Li Fei-Fei 教授らが構築している画像意味解析用のデータセットであり、1 万オーダーの。規模、品質ともに群を抜いており、当該研究分野の研究ならびに評価のデファクトスタンダードとなっている。

・フランス国立視聴覚研究所 (INA) [2]は、1974 年制定のラジオやテレビ放送を含む視聴覚資料の保存に関する法律に基づき、1975 年に設立された機関であり、現在までフランス国内のラジオ・テレビ放送を原則すべて蓄積し続けている。コンテンツは商用・研究用に提供されている。オランダにも視聴覚研究所(Netherlands Institute for Sound and Vision) [3]があり、オランダ国内のテレビ放送を含む視聴覚情報を蓄積し、研究用の提供も行っている。

・米国標準技術局(NIST)では、TRECVID [4]プロジェクトにより、映像解析・検索プロジェクトの推進を図っており、大規模データセットの構築・配布、研究マイルストーンの策定を行い、世界中の研究グループを主導している。顔認識に関しての取組みでは、FERET [10]は NIST により構築・配布された顔認識評価用のデータセットであり、FRVT [8]は NIST により主催されている主として産業界を対象とした顔認証技術ベンチマークである。

・Labeled Faces in the Wild(LFW) [15]は、より「ワイルド」な状況、すなわち、顔の向き、

表情、照明条件等に制限が一切ない顔画像データセットであり、インターネットから収集された 1680 人分 1 万 3000 枚以上の画像からなる。FERET よりも挑戦的であり、顔認識技術の評価のための新たなデファクトスタンダードとなっている。

・米国 IARPA ALADDIN プロジェクト[5]では、2011 年から 5 年計画で、インターネット上の映像を含む任意の映像中の動作等のイベント情報を極めて高い精度で検出する技術の実現に取り組んでいる。IARPA 主導のプロジェクトであることから、米国において、この挑戦的な研究開発が国家安全上重要な課題として認識されていると考えられる。

・欧州では、FP6/FP7 等を通じて、Web、検索、マルチメディア解析等を含むさまざまなプロジェクトを支援している。PASCAL VOC (Visual Object Classes) [6]は、FP6 の援助による PASCAL プロジェクトの一部であり、ImageNet と比較して小規模ながら極めて高品質の画像意味解析向けデータセットならびに研究マイルストーンを構築し、研究コミュニティーを主導している。Chorus+1 [6]は、FP7 の援助による、多数のプロジェクトを携えた巨大プロジェクトであり、マルチメディア検索エンジンの構築を主目標としている。Theseus や Quaero も含んでおり、Petamedia プロジェクトではマルチメディア解析のベンチマークにより研究コミュニティーを主導する MediaEval プロジェクト [7]を支援している。

・タスクが決定された状態における日本の技術水準は世界的に見ても高い。顔認証技術については、NEC の技術が 2013 年の FRVT [8]において世界第 1 位の精度を達成している。画像・映像意味解析・検索技術については、ImageNet [1]データを利用した PASCAL VOC [6]における大規模画像意味解析コンペティションにおいて、東京大学原田達也教授のチームが 2012 年に世界一位の認識精度を達成している。また、TRECVID [4]の SIN(映像意味分類)タスクにおいて東京工業大学篠田浩一教授のチームが 2011、2012 年に世界一位の認識精度を達成している。また、同 INS(物体検索)タスクにおいて、国立情報学研究所佐藤真一教授のチームが 2011、2013、2014 年に世界一位の検索精度を達成している。

（6）キーワード

大規模学習データ、マルチメディア、検索、内容解析、意味解析

（７）国際比較

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	○	→	・データセット構築・公開については△ ・わが国においては、著作権・肖像権・個人情報等の問題により、大量に収集され、研究用途に広く公開されている画像・映像データセットはまれである。 ・研究については○ ・顔・文字について、技術水準は世界的に見ても高く、基礎技術におけるブレークスルーについては落ち着いている感があるが、まだ広く研究が進められている。 ・一般の画像・映像の意味解析・検索等についても、大学・国研等において盛んに研究が行われている。
	応用研究・開発	○	→	・顔・文字については広く応用研究・開発が進んでいる。 ・一般の画像・映像の意味解析・検索等の応用についても研究・開発が進められている。
	産業化	○	→	・顔・文字については産業化も進んでおり、世界的にも存在感を示している。 ・一般の画像・映像の意味解析・検索に関する産業化については、特にインターネットにおける検索エンジンやe-コマース等において、こうした技術を核とした産業化が遅れている。
米国	基礎研究	◎	→	・研究を戦略的に振興するため、研究用データセットの構築ならびに配布、研究すべきタスクの設定ならびに研究コミュニティーへの発信等、積極的に世界でも中心的な役割を果たしてきている。NIST等がその中核的な機関となっている。 ・NSF等により、上記のような機関とも呼応して、米国内での基礎研究を特段にサポートしており、顕著な研究成果を挙げている。実際の研究コミュニティーにおいても、米国の大学等の研究者が主導的な役割を担っている。 ・IARPA ALADDINプロジェクトにおけるインターネット上の映像のイベント解析等、より挑戦的な画像・映像意味解析研究への政府主導の動きが見られる。 ・ディープラーニング等新たなブレークスルーについても、主導している。
	応用研究・開発	◎	→	・Google、Facebook、Microsoft等の米国内の巨大ITメディア企業でも研究部門を有し盛んに応用研究・開発を行っている他、大学とこうした企業との連携も盛んであり、加えて大学において、起業を目指した応用研究・開発も極めて盛んである。
	産業化	◎	→	・今後の研究開発において重要となる、画像・映像意味解析におけるビッグデータを実際に保有する企業は、Google、Facebook、Microsoftなど、ほとんど米国に集中している。基礎研究、応用研究・開発の層の厚さとも相まって、今後とも画像・映像意味解析研究の産業化においても米国は主導的な役割を果たすと思われる。
欧州	基礎研究	◎	→	・EUを母体とし、欧州内の複数の国家にまたがる共同研究が非常に盛んであることから、大規模データセットを共有したオープンな研究について非常に積極的である。 ・マルチメディアを対象とした研究にもこうしたアプローチをとっており、MUSCLE、Chorus、Quaero等のプロジェクトがある。 ・Image CLEF、PASCAL VOC、MediaEvalにおいては、実際の研究用マルチメディアデータセットを構築し、研究コミュニティーに公開している。
	応用研究・開発	○	→	・FP6/FP7ならびにHorizon2020において、産官学の連携によるプロジェクトの推進も振興されており、基礎研究の成果を応用するプロトタイプの開発なども盛んに行われている。
	産業化	○	→	・産学官連携コンソーシアムとしてのQuaeroにおける参画企業の積極的な活動等、産業化にも積極的であり、実際にExaleadの起業等の例もある。しかし、世界的に顕著な存在感を示すまでには至っていない。

中国	基礎研究	○	↗	・データセットの構築・配布、タスク設定による行うべき研究の方向性の提示等の例はほとんどない。 ・研究は極めて盛んに行われている。特にマルチメディア分野においては、中国の大学は、特に近年顕著に能力を上げてきており、トップ会議にも採択論文が散見されるようになってきている。 ・中国の大学において、米国や欧州の有力研究者との連携を図り、研究能力の飛躍的発展を実現できている例が多く見られる。
	応用研究・開発	○	↗	・Microsoft Research Asiaや、Baidu、アリババ・グループ等ならびにこうした企業と連携した大学等において、画像・映像の検索等の応用研究が盛んに行われている。
	産業化	○	↗	・Baiduやアリババ・グループ等、大量のデータも保有し世界的な存在感を急速に上げている企業が出てきている。
韓国	基礎研究	○	→	・データセットの構築・配布、タスク設定による行うべき研究の方向性の提示等の例はほとんどない。 ・研究は盛んに行われている。画像・映像圧縮、コンピュータービジョン等の分野では顕著な成果が見えている。一方、画像・映像の検索や応用を主としたマルチメディア分野の研究はあまり盛んではない。
	応用研究・開発	△	→	・基礎研究から応用に結びついた例はまだ多くは見られないが、基礎研究の進展を見るに、今後は応用研究・開発の進展も見込まれる。
	産業化	○	→	・サムスングループ等において研究開発が行われている。

（註 1）フェーズ
基礎研究フェーズ ：大学・国研などでの基礎研究のレベル
応用研究・開発フェーズ ：研究・技術開発（プロトタイプの開発含む）のレベル
産業化フェーズ ：量産技術・製品展開力のレベル
（註 2）現状
※わが国の現状を基準にした相対評価ではなく、絶対評価である。
◎：他国に比べて顕著な活動・成果が見えている、 ○：ある程度の活動・成果が見えている、
△：他国に比べて顕著な活動・成果が見えていない、×：特筆すべき活動・成果が見えていない
（註 3）トレンド
↗ ：上昇傾向、 →：現状維持、 ↘：下降傾向

（８）引用資料

1) ImageNet
http://www.image-net.org/
2) フランス国立視聴覚研究所(INA)
http://www.ina.fr/
3) オランダ視聴覚研究所(Netherlands Institute for Sound and Vision)
http://www.beeldengeluid.nl
4) TRECVID
http://trecvid.nist.gov/
5) IARPA ALADDIN プロジェクト
http://www.iarpa.gov/index.php/research-programs/aladdin-video
6) PASCAL VOC
http://pascallin.ecs.soton.ac.uk/challenges/VOC/
7) MediaEval
http://www.multimediaeval.org/
8) Face Recognition Vendor Test (FRVT)
http://www.nist.gov/itl/iad/ig/frvt-home.cfm

9) OMRON OKAO Vision
http://plus-sensing.omron.co.jp/technology/
10) FERET
http://www.itl.nist.gov/iad/humanid/feret/feret_master.html
11) MNIST
http://yann.lecun.com/exdb/mnist/
12) ETL 文字データベース
http://etlcdb.db.aist.go.jp/?lang=ja
13) Caltech 画像データベース
http://www.vision.caltech.edu/Image_Datasets/Caltech101/ http://www.vision.caltech.edu/Image_Datasets/Caltech256/
14) Yaniv Taigman, Ming Yang Marc'Aurelio Ranzato, Lior Wolf, DeepFace: Closing the Gap to Human-Level Performance in Face Verification, Proc. of Computer Vision and Pattern Recognition, 2014.
15) Labeled Faces in the Wild (LFW)
http://vis-www.cs.umass.edu/lfw/
16) Chorus+
http://avmediasearch.eu/

３．８．５　言語と映像の統合理解

（１）研究開発領域名

言語と映像の統合理解

（２）研究開発領域の簡潔な説明

文字・音声等の言語情報と画像・映像等の視覚情報の融合的な処理による相互の対応の認識・学習、またそれに基づく言語・視覚情報の統合的な理解・編集に関する研究開発

（３）研究開発領域の詳細な説明と国内外の動向

言語情報および画像・映像情報の意味解析技術がそれぞれ長足の進歩をとげつつある今日、これら異なるメディアの情報の間の意味的な対応関係を認識・学習し、両メディアの情報を統合的に解析・理解する次世代の研究への関心が次第に高まってきている。画像・映像データの意味内容を言語でより精緻に記述できるようになれば、例えば画像・映像データの検索や分類において、物体のカテゴリや名称だけでなく、物体間の関係や事象・情景といった高次の意味的情報が利用できるようになる。画像・映像データに対する要約やマイニング等の自動編集にも現実性が出てくる。言語処理の観点からも、画像・映像付きの言語データの意味内容が画像・映像と対応づく形で解析できるようになれば、言語だけでは伝わりにくい物理的位置関係や動作、物体の質感等の情報を含めたデータ解析につながる。モバイル端末からの情報検索、質問応答等の言語情報サービスにおいても、ユーザーの周囲の視覚情報を質問のコンテキストとして利用することによって、高度に文脈適応的な情報提供が可能になると期待できる。また、ロボットとの自然言語対話、言語による制御においても言語・視覚情報の統合理解は必須の技術である。

言語と画像の統合理解はもともと 1970 年代の初期の人工知能研究から想定されてきた目標であるが、わずか数年前まではまだ遠い目標と考えられていた。統合理解のための基本条件と考えられる画像・映像の意味解析、すなわち画像・映像中のどこに何があり、どのように動いているかの一般的な解析が言語の意味解析に比べてはるかに困難であったことが主な理由である。しかし、そうした見方が今大きく変わりつつある。第一に、「3.8.4 画像・映像の意味解析」領域の報告にもあるように、大規模学習データに基づく戦略的な技術開発、深層学習等の機械学習技術の発展などにより、一般の画像・映像を対象とする意味解析が徐々に現実味を帯びてきている。第 2 の変化は、キャプションなどの言語情報が付与された膨大な量の画像・映像データが Web から入手できるようになったことである。従来の画像認識研究では研究用に統制された学習データを用いるのが一般的であった。こうしたデータに比べ、Web から入手できるキャプション付きの画像・映像データは規模の面でも多様性の面でもはるかに大きく、事象・行為・情景などの高次の視覚情報と言語による記述の対応を学習できる研究環境が出現しつつある。こうした状況の変化を受けて、「言語による画像・映像の意味記述（キャプション生成）」、「実世界にグラウンドされた言語学習（grounded language learning）」などの切り口で言語と画像・映像を統合する萌芽的研究が一つの学際領域を形成し始めている。

米国では、画像認識、言語解析、機械学習をリードするいくつかの大学の研究グループに加え、Google、Microsoft、Baidu 等の企業研究所がこの分野の基礎研究を牽引する勢力に

なりつつある。欧州では、英国の研究助成機関 EPSRC に支援された研究コミュニティー（V&L Net [1)]）が同分野に関する国際会議を 2011 年から 3 回開催するなど、関心が高まっているが、学術的に強いインパクトを持つ動きはまだあまり見られない。日本は、言語解析、画像認識それぞれの要素技術において高い国際競争力を有しており、統合理解に向けた取組みでも東大、国立情報学研究所等がキャプション生成で先駆的な研究成果を挙げるなど、潜在的な競争力は十分にある。

重要な点は、いずれの地域においてもまだ萌芽的な基礎研究の段階にとどまっており、当該技術の応用開発・産業化によって将来大きな市場の開拓が見込めるにもかかわらず、どの地域もそこに踏み出せていないことである。わが国が戦略的にこの分野の技術開発を行えば、学術界・産業界の両方において強力な主導権を確保できる可能性が高い。

（4）科学技術的・政策的課題

・言語・画像解析の研究分野ではこれまで、共通の問題設定と研究開発用データセットをもとに技術評価型の会議を組織し、戦略的な技術開発振興を成功させてきた面がある。しかし、言語と画像の統合理解についてはまだ極めて萌芽的段階にあり、研究者が共有できる問題設定もデータセットも存在しないのが現状である。画像・映像解析の技術評価で中心的な役割を果たしている米国の TREC/TRECVID [2)]においても、キャプション生成を含め、言語・視覚情報の統合に踏み込んだ課題は見られない。今後、日本の研究機関がこの分野の標準となり得る問題設定と大規模なデータセットを早期に構築し、公開することができれば、わが国が同分野の主導権を握れる可能性もある。それには、国立情報学研究所が主体になって開催している評価型会議 NTCIR [3)]など、わが国の言語・画像解析技術の強みを支える既存の枠組みを活用する他、言語処理・画像処理・機械学習等の第一線の研究者の領域横断的な連携を促進するプロジェクト型研究振興施策が有効であると考えられる。

・画像・映像の意味内容を言語で記述するという課題は、単なる物体や動きの検出を大きく超えて、究極的には例えば登場人物がどのような状況で何のために何をしようとしているかといった理解を目指すものである。しかし、現在のキャプション生成の代表的な方法は、入力画像の個々の断片から単語を生成し、それらを統計的な言語モデルで並べ替えているにすぎない。一方、言語処理の側では、大量の言語データから多様な世界知識・領域知識を獲得する研究が発展してきており、そこで得られる知識を用いてどのように言語の深い意味理解（事象間の因果関係や意図の構造の解析など）を実現するかが最先端の課題になっている。京大、東北大、奈良先端大、NICT など、わが国の言語処理研究チームはこの点で先進的な研究を展開しつつあり、こうした言語理解の研究をどのように画像・映像の深い意味理解につなげるかが今後の大きな課題になると予想される。

・静止画の意味解析についてはディープ・ニューラルネット等の近年の技術開発で大きな発展をとげた。一方、映像の意味解析は静止画の解析よりも困難であり、映像をディープ・ニューラルネットでうまく解析できたという例も現時点では報告がない。しかし、人間の認知機構から見ても時系列情報の処理は極めて重要であり、この問題でのブレークスルーが言語・視覚情報の統合理解への重要な鍵になると考えられる。

・言語・視覚情報の統合理解はヒューマノイドを指向するロボット研究の文脈でも重要なテーマの一つである。ロボットを人間と同じ環境に置き、実世界や周囲の人間とのインタラ

クションから言語と画像を統合的に学習するといった実験がヒューマノイドロボット等の技術革新により可能になってきている。こうした試みはまだ実験室レベルを出ないが、「実世界にグラウンドされた言語学習」の実現をテクノロジーの立場から目指す研究としても、またロボットを使ったシミュレーションによって人間の知能のモデル化を目指すサイエンスの研究としても重要な意味を持っており、今後の発展が期待される。

（５）注目動向（新たな知見や新技術の創出、大規模プロジェクトの動向など）

・欧州の V&L Net が過去 3 回開催した Workshop on Language and Vision の他、言語処理分野 NAACL-2012、画像認識分野 CVPR-2013、人工知能分野 AAAI-2011、AAAI-2012、機械学習分野 NIPS-2011 などの国際会議でそれぞれ言語情報と視覚情報の統合に関するワークショップが開催され、また本会議で同分野の論文の受賞が相次ぐなど、新しい学際領域を形成しつつある。
・当該領域の研究開発推進に焦点を当てた大規模プロジェクトや大型の研究助成プログラムなどは国際的にも現時点ではまだ出てきていない。

（６）キーワード

画像・映像からのキャプション生成、実世界にグラウンドされた言語学習

（７）国際比較

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	◎	↗	・キャプション生成等で先駆的で影響力のある研究を行っている。 ・言語処理、画像処理のおのおのの要素技術で強い国際競争力を持っている。 ・課題設定・データの共有など研究推進の共通基盤は未整備。NTCIR等でも言語と画像の統合までは踏み込んでいない。
	応用研究・開発	△	→	・特筆すべき動きはまだ見えておらず、企業の動きも弱い。
	産業化	×	→	・特筆すべき動きはまだ見えていない。
米国	基礎研究	◎	↗	・数カ所の大学の研究グループ、およびGoogle、 Microsoft等の研究所で研究に着手。大学と企業の連携も見られる。 ・課題設定・データの共有など研究推進の共通基盤は未整備。TRECVID等も言語と画像の統合までは踏み込んでない。
	応用研究・開発	△	→	・特筆すべき動きはまだ見えていない。
	産業化	×	→	・特筆すべき動きはまだ見えていない。
欧州	基礎研究	○	↗	・英国EPSRCの支援で研究コミュニティーV&L Netを構築。ワークショップを開催。 ・課題設定・データの共有など研究推進の共通基盤は未整備。Horizon2020等の大型研究助成の枠組みでもカバーしていない。
	応用研究・開発	△	→	・特筆すべき動きはまだ見えていない。
	産業化	×	→	・特筆すべき動きはまだ見えていない。
中国	基礎研究	△	→	・画像・映像の意味解析の研究は活発になってきているが、言語との統合処理については目立った動きは見られない
	応用研究・開発	×	→	・特筆すべき動きは見られない
	産業化	×	→	・特筆すべき動きは見られない。
韓国	基礎研究	△	→	・画像・映像の意味解析の研究は活発になってきているが、言語との統合処理については目立った動きは見られない
	応用研究・開発	×	→	・特筆すべき動きは見られない
	産業化	×	→	・特筆すべき動きは見られない。

（註１）フェーズ
基礎研究フェーズ ：大学・国研などでの基礎研究のレベル
応用研究・開発フェーズ ：研究・技術開発（プロトタイプの開発含む）のレベル
産業化フェーズ ：量産技術・製品展開力のレベル
（註２）現状
※わが国の現状を基準にした相対評価ではなく、絶対評価である。
◎：他国に比べて顕著な活動・成果が見えている、 ○：ある程度の活動・成果が見えている、
△：他国に比べて顕著な活動・成果が見えていない、×：特筆すべき活動・成果が見えていない
（註３）トレンド
↗ ：上昇傾向、 →：現状維持、 ↘：下降傾向

（８）引用資料

1) The EPSRC Network on Vision and Language
http://www.vlnet.org.uk/
2) TRECVID
http://trecvid.nist.gov/
3) NTCIR
http://research.nii.ac.jp/ntcir/index-en.htm