3　教育・学習支援環境の実現に向けて

3-1　言語機能研究の新たな展開に向けて

3-1 Toward the New Horizon of Computational Linguistics

井佐原　均
Hitoshi ISAHARA

要旨

我々は自然言語処理の研究、自然言語の基礎研究、実用システムの開発、外部機関との協力に注力し、研究活動を行っている。自然言語処理の研究においては、学習に基づく自然言語処理の研究を進めている。神経回路網モデルを用いた言語解析技術の一層の高精度化、実用化も行っている。自然言語の基礎研究においては、日本語連体修飾要素に関する語彙意味論的研究を行っている。日本語待遇表現の研究においては、被験者実験に基づく敬語表現運用の数値モデルの開発や、敬語表現の変遷に関する研究を行った。また、音楽等の感性情報を表現する形容詞表現の研究を被験者実験に基づいて行うとともに、三者間の対話モデルの構築の検討を行った。

This paper describes an overview of Computational Linguistics Group of Communications Research Laboratory of Japan. We focus on the following topics; research on natural language processing, fundamental research on natural language, development of practical systems and collaboration with other research organizations. As for the research on natural language processing, we are doing research on natural language processing using learning mechanism and neural network models. As for fundamental research on natural language, our research includes lexical semantics, representation of emotion and dialogue model with three participants.

[キーワード]
自然言語処理，知識処理，計算言語学
Natural Language Processing, Knowledge Engineering, Computational Linguistics

1　まえがき

けいはんな情報通信融合研究センターでは、人間の知的活動を支援する環境の開発を目標に、自然言語処理の研究、自然言語の基礎研究、実用システムの開発、外部機関との協力に注力し、研究活動を行ってきた。特に自然言語処理の研究においては、学習に基づく自然言語処理手法の開発を行い、成果を上げている。

自然言語処理の研究においては、学習に基づく自然言語処理の研究を進めている。最大エントロピー法を用い、形態素解析、係り受け解析、固有表現抽出、語順の決定に至る一連の処理を同じ枠組みで行う手法を開発した。また、学習に基づくコーパス修正、言い換え、テキストセグメンテーションに関する研究を進めている。神経回路網モデルを用いた言語解析技術の一層の高精度化、実用化も行っている。情報検索、情報抽出、要約に関して研究を進めるとともに、客観的な評価を行うコンテストに参加し優秀な成績を上げた。今後は要約の研究を更に進め、単なる重要文抽出ではなく、言い換え技術を用いた真の要約を生成する技術の開発を進める予定である。また、学習に基づくシステムに規則

を後処理として用いることにより、精度の向上を図っているが、これらを融合処理することにより、より人間の言語処理に近いモデルの構築が可能となろう。

自然言語の基礎研究においては、日本語連体修飾要素に関する語彙意味論的研究を行っている。特に神経回路網モデルを用いて日本語名詞の自己組織化意味マップを作成することにより、理論の実証を行った。今後は、意味マップを階層化することにより、文章の意味を表現する枠組みを開発する予定である。日本語待遇表現の研究においては、被験者実験に基づく敬語表現運用の数値モデルの開発や、敬語表現の変遷に関する研究を行った。また、音楽等の感性情報を表現する形容詞表現の研究を被験者実験に基づいて行うとともに、三者間の対話モデルの構築の検討を行った。

実用システムの開発においては、知的ニュースリーダの実運用と公開を行った。今後、実運用時の各機能の精度の検証を行う予定である。また英語学習支援システムの開発に向けて、学習者コーパスの作成とエラータグの検討を行った。学習者コーパスは日本人の英語学習者に対するインタビューを書き起こしたものであり、学習者の英語レベルの判定結果がついていることが特徴である。本研究開発は、通信・放送機構の技術移転プロジェクト「適合型コミュニケーション技術の研究開発」のもとで行っている。また、例文検索システム（KWIC システム）、情報検索システム、コーパス作成支援システムの開発と公開も進めている。

外部機関との協力においては、科学技術振興調整費開放的融合研究推進制度のもとで、国立国語研究所とともに、「言語的・パラ言語的情報に基づく『話し言葉工学』の構築」の研究開発を行っている。ここでは、大規模話し言葉コーパスの作成と、コーパスに基づく話し言葉工学の研究を行っている。ブレークスルー21の委託研究として、東京都神経科学総合研究所と失語症患者の語音認知に関する共同研究を行った。

これらの成果をもとに、けいはんな情報通信融合研究センターにおいて、引き続き自然言語に関する研究開発を推進する予定である。本稿では、自然言語グループの今後の研究活動計画を概説する。

2　自然言語処理の研究

2.1　話し言葉テキスト用解析システムの研究

これまでに我々は、書き言葉テキスト（新聞記事）に対応した学習に基づく解析システムを開発してきた。しかし、書き言葉テキスト用に開発されたシステムをそのまま話し言葉テキストに適用しても高精度は望めない。なぜなら、話し言葉と書き言葉との間には、両者では表現が異なる、話し言葉では文の単位が明確ではない、ポーズやアクセント、音の強弱や話す速さなどの韻律情報を含むなど、様々な違いがあるからである。現在、「言語的・パラ言語的情報に基づく『話し言葉工学』の構築」において、話し言葉処理の研究を推進しているところであり、今後は形態素解析システム、構文解析システムを対象に、話し言葉に特徴的な情報である韻律情報にも着目し、書き言葉用に開発された解析システムを話し言葉テキスト解析用に改良する予定である。実験により、韻律情報が解析システムの精度向上に有効利用できるかどうかを確かめたい。開発したシステムは近い将来、話し言葉テキストに形態的、構文的な付加情報を自動付与するのに用いる予定である。

2.2　翻訳メモリに基づく機械翻訳の研究

現在までに開発してきた学習に基づく日本語解析、生成システムはテキストと、文節の係り受け関係を表わす統語構造との間のマッピングを目的としていた。ここで更に日本語の統語構造を英語の統語構造にマッピングする技術が開発できれば、日本語テキストを英語テキストへ翻訳することが可能となる。このような方向性のもと、まず、統語構造を構成する最小単位である単語のマッピング、つまり、訳語選択のためのシステムを開発する予定である。訳語選択は容易ではない。なぜなら、例えば一つの日本語単語でもその英語訳は複数あることが多く、どの訳語を選ぶべきかどうかはその単語のまわりの文脈に依存しているからである。そこで、各単語ごとにその単語を含む句のバリエーションを対訳とともに用意しておき（これを翻訳メモ

リと呼ぶ)、どの対訳対を選択するかを類似性に基づいて決定することにより訳語を選ぶ手法を開発する。将来的には教師データを整備し、学習に基づくシステムを作成する予定である。

2.3 要約に関する研究

我々は現在、要約システムの開発に取り組んでいる。これまでに、日本語の文章を対象として、重要文抽出に基づく要約システムを作成した。今後の目標としては、要約の対象とする文章の範囲を広げること、より自然な要約を出力することの二つを目指している。

まず、日本語だけでなく英語の文章も対象とするようにシステムを拡張し、言語に依存する部分と依存しない部分とを切り分ける。これに基づいて、複数の言語に対して適用可能な要約システムを構築する。また、話し言葉コーパスを対象として要約を行い、書き言葉の性質に基づいた要約技術がどの程度話し言葉で利用可能か、話し言葉の要約のためにはどのような技術が新たに必要かを調べる。

さらに、より自然な要約を出力するために、文章から特定の情報を取り出す技術である情報抽出と自動要約とを組み合わせる。要約に情報抽出の技術を取り入れることによって、表面的な変形だけでなく、意味の理解にまで踏み込んだ要約へつながるものと考えている。一方で、情報抽出において問題となっている、対象分野にシステムが特化してしまうことへの解決策も、要約と情報抽出を統一して考えることで見えてくるであろう。

2.4 テキスト分割に関する研究

テキスト分割とは、複数トピックからなる文章を切り分けて、それぞれの切り分けた部分が一つのトピックになるようにすることを言う。

テキスト分割は、情報検索や要約などにおいて重要である。まず、情報検索においては、文書全体ではなく、ユーザの検索要求を満たす部分(トピック)だけを検索した方が効果的である。また、要約においては、長い文書をトピックに分ければ、それぞれのトピックごとに要約を作成することにより、文書全体の要約を作成できるし、重要なトピックだけを選んで要約を作成することもできる。

情報検索や要約等が対象とする文書は、分野を限定しない文書であるので、それらを分割する手法も分野を限定しないものである必要がある。我々の手法は、テキスト内の単語分布のみを利用してテキストを分割する。そのため、訓練データが存在する分野に限られることなく、どんな分野のテキストでも分割できる。

この手法は、テキストの分割確率が最大となるような分割を選択するというものである。このようなアプローチは、分野を限定しないテキスト分割としては、新しいアプローチである。

なお、従来の研究で、分野を限定しないテキスト分割の研究では、主に、語彙的な結束性を利用してテキストを分割している。その例としては、意味ネットワーク上での活性伝播に基づく結束性を利用するものや、単語分布の類似度(コサイン)を結束性としたものや、単語の繰り返し状況に基づいて結束性を計るものや、文間の類似度としてコサインを直接使うのではなくコサインの順位を結束性の指標とするものなどがある。

我々の手法は、従来手法と比べて、同等以上の精度でテキストを分割することができた。このことは、この手法がテキストの分割に有用であることを示している。今後、実際の応用でのテキスト分割の有効性を調べたい。

2.5 質問応答システムに関する研究

計算機が言語を理解したかどうかを判断する明確な基準はまだない。その一つの基準として、計算機に質問をし、それに対して適切に回答ができるかを調べるという方法が考えられる(これに類似した基準としてチューリングテストと呼ばれる方法がある。これは、被験者には計算機と会話しているか人と会話しているかをふせておいて、計算機と会話しているか人と会話しているかを被験者が判断できない場合、その計算機は会話に関して人と同等程度の能力があると判断するものである。)。例えば、計算機に「日本の首都はどこですか」と聞いて、「東京です」と答えることができれば、そのような答えを出すことができる程度の能力を持っていると判断できる。このようなシステムは質問応答システ

ムと呼ばれ、我々はこのシステムの初歩的なものを既に作成している。

我々の質問応答システムでは、百科事典や数年分の新聞記事などの大量の電子化テキストから、質問に対する答えが含まれていそうな文章を抜き出し、その文章と質問文を照合することで答えを抽出するということを行っている。例えば、「パーキンソン病の兆候は脳のどの部分にある細胞の死に関係していますか」という質問を入力すると、大量の電子化テキストから「パーキンソン病は、中脳の黒質にあるメラニン細胞が変性し、黒質細胞内で作られる神経伝達物質のドーパミンがなくなり発病する、とされている」といった文を探し出し、「黒質」と解答を出力する。

現在のところ、システムは質問の文と知識ベースの文を大雑把に照合し、疑問詞「どの部分」に対応する「黒質」を取り出すということを行っている。しかし、この方法では、大雑把に照合しているだけで解の確実性が乏しい。今後は、意味が等価な言い換えの表現を集めるということを行い、

“「細胞が変性する」＝「細胞が死ぬ」”

“「AはBして発病する」＝「Aの兆候はBすることに関係する」”

といった同義表現に関する知識により、質問文と知識ベースの文を言い換えていき、照合をより確実なものにすることを検討している。例えば、知識ベースの文を上の言い換え表現により「パーキンソン病の兆候は、中脳の黒質にあるメラニン細胞が死ぬことに関係する」に言い換え、さらに“「死ぬこと」＝「死」”、“「にある」＝「の」”といった同義表現に関する知識により、「パーキンソン病の兆候は、中脳の黒質のメラニン細胞の死に関係する」と言い換えてから、知識ベースの文と質問文を照合するとより確実に「黒質」という解答を得ることができるようになる。我々は現在、類似テキスト対を照合することにより、上記の言い換えの表現の獲得の研究も試みている。

3　自然言語の基礎研究

3.1　神経回路網モデルによる自然言語の研究

神経回路網モデルの研究は、脳の神経回路網構造と振舞いを模倣しようとして始められ、ブームが去った今、理論、脳・認知科学、そして、工学的応用という三つの方面で地道な研究が重ねられてきている。その工学的な応用としてパターン認識の分野で大きな威力を発揮している一方、言語獲得、知識表現、脱曖昧化など自然言語の研究にも大きく寄与している。海外では、コネクショニストモデルと称して、このような自然言語の研究が盛んに行われている。しかし、対照的に、日本ではこのような研究があまり重視されていないのが現状である。自然言語グループは、この研究領域の草分け的な存在として、1993年という早い時期から神経回路網モデルを用いた自然言語の研究に着手し、種々の機械学習手法と同等以上の性能を保ちながら、更に付加するべき知見を追求して連想記憶、知識表現、単語意味の可視表現、形態素解析、そして、大規模コーパスの誤り検出などの研究を神経回路網モデルという同一の枠組みで研究してきた。

今後は引き続きこれらの研究を行っていくとともに、それらの技術を更に構文解析や固有名詞同定などの自然言語処理の基本課題へ拡張していく予定である。そして、最終的にはこれらの技術と他の機械学習手法とをうまく融合して、実世界の様々な誤りを含む言語データをその誤りを高精度まで自動修正して学習することによって、言語処理の様々なタスクに必要とされる知識や情報を獲得することの可能な学習機械の開発に力を入れていきたい。

3.2　語彙意味論の研究

計算機による言語の理解・生成を可能とするためには、処理に必要となる言語知識を計算機上に適切に蓄えておかなくてはならない。そのような知識を用いて、例えば言語理解においては、どのようにして異なった構文構造から同じ意味表現を生成するか、また、どのようにして意味的に曖昧な文から、文脈に応じてそれぞれの曖昧性に対応する意味表現を生成するかを考える必要がある。本研究が目指すものは、静的

あるいは固定的に記述された語彙の意味辞書ではなく、動的に語彙の意味同士を結びつけた辞書の構築である。つまり、辞書にあらかじめコード化されている語彙項目から動的に語彙項目を生成したり、各語彙項目をシステマティックに関連づけたりするメカニズムを導入した意味辞書の構築を目指す。これを行うためには、単語が実際の運用の中で意味を実現するときの語彙的規則（Lexical Rule）を記述する必要がある。本研究では、言語学の知見と工学の知見及び技術を積極的に取り入れ、この両面から研究を進める。

Lexical Ruleを導入するためには、現実の言語現象において、どのように単語が機能しているかという「語の意味の実体」を正確に把握しなければならない。正確な意味記述ができなければ、辞書の精度は落ち、不正確な意味処理や、語彙の過剰生成に陥ったり、文の意味的な曖昧性が解消できなかったりする。そこで、まず言語学的な知見を利用して、正確な意味の実体の把握を目指す。

一方、言語学的観点からの分析作業と併行して、自己組織型神経回路網モデル（ニューラルネットワークモデル）の知見と技術を生かして分析作業の検証を行う。現在、このモデルを利用して、日本語名詞の意味マップを構築している。日本語名詞の意味マップとは、名詞の意味の親疎によって、名詞同士を近い位置に配置したり遠い位置に配置したりする「意味の地図」である。また、配置された名詞と密な関係にある連体修飾語句（形容詞など）との間にもリンクが張られ、全体として、日本語語彙のネットワークを作ることができる。

3.3　敬語表現の誤用に関する研究

近年、敬語表現の乱れ、あるいは変遷が指摘されている。敬語表現の誤用（ここでは、言語学的観点からは規範的でないと考えられる敬語表現をこう呼ぶことにする）としては、語形の単純な誤りや機能の誤解による運用の誤りなど、いろいろな種類がある。敬語表現の誤用が相手に与える違和感の強さは、誤用の種類に依存する可能性がある。そこで、敬語表現における誤りには、どのようなものがあるか、また、それらを人々がどの程度、不自然と感じているのかを調査し、統計的に分析する。

具体的には、誤用にあたる敬語表現を収集・生成した上で、一対比較法による心理実験を行い、誤りを含む敬語表現に対する不自然さ（自然さ）の心理的印象を数値化する。これまでに小規模実験により、自然さの程度は、誤用の種類や被験者属性への依存性があることを確かめているが、これらがより明らかになれば、敬語学習システムのような教育システムへの適用等が可能であると考えられる。

今後は、規模を拡大した実験の結果に関し、上記の依存性についての詳細な分析を行うとともに、正しい敬語表現に対する認知との比較により、性別や年齢による敬語習得の傾向や表現ごとの認知の度合いなども分析していく予定である。

3.4　コミュニケーション過程に関する研究

コミュニケーションの柔軟性と頑健性を可能にする認知機構の解明を目指して研究を行う。特に、言語行為の対話内での機能、会話進行のモデル（会話連鎖構造の規則性の生成過程、会話における関連性に基づく推論のメカニズム）、会話主体の認知モデル（対話主体のプラン形成・認識や信念の文脈感応性の研究）に焦点をおく。

現在、話し言葉コーパスへの談話レベルタグの付与の検討、対話コーパスの構築、談話行為・関連性タグの付与・見直し、会話の進行に応じて適切に振る舞う社会的エージェントの実現とその基礎となる三者対話データの収録及び三人対話モデルの構築、コミュニケーションを通じての教育・学習を実現するための基礎的枠組みの考察を進めている。

3.5　音に関する感性情報の研究

音楽と言語は両方ともコミュニケーションの形態であるが、それらは異なった目的を持っている。言語の第一目的は、考えを正確に伝えることであるのに対して、音楽の主要な目的の一つは、情緒を高めて美しく表現することにある。しかし、言語の中でも、考えが率直に伝わらない場合があり、例えば、詩や短歌などは、著者の思い描いた情景に読者が共感するかどうかは

分からない。正確な理解のためには、文字として記述されたもの以外の情報を詩や短歌に付与する必要があるであろう。さらに、話し言葉となると、時間の流れとともに言葉が消えていくため、聞き手が話の内容を本当に理解しているかどうかは、話し手の技量と聞き手の注意に依存する。話し手は、相手に理解してもらおうと、声の音量を大きくして強調したり、同じ言葉を繰り返したり、印象付けをするため、情緒豊かに話す。そして、聞き手は、話し手の呼吸や癖などを読みとる。それは一種の音楽的表現であると考えられる。演奏者は、情緒豊かに演奏し、感情を外に発散させる。それを受け取った聴取者は自然と演奏者と呼吸を同期させ、演奏者の演奏意図を感じとる。

話し言葉には、正確に伝える目的がありながらも、正確に伝える手段は人間の感情にある、といったあいまいな部分がある。このように、音楽と言葉に共通性がみられることから、これまで行ってきた、音楽情報を変数とした数々の実験やモデルが言語にも適用できるのではないかと考え、これからは、以下に示すような事柄を検討する。

話し言葉の個性について考えると、学会の講演では、冒頭部分や話題の転換部分、終了部分などに様々な接続詞が使われる。それは、人によって異なるが、各個人の中では講演の中で同じ接続詞が数回繰り返されることがある。まだまだ個性となるような箇所はあると思われるが、それらの個性のグループ化を目指す。その人の話の個性が発見できれば、その後の様々な処理が容易になると思われる。例えば、ある人は一文の初めに「それでは」を多用する“それでは型”であることが事前に分かっていれば、話題の転換部分等が容易に発見できると考えられる。

情緒豊かな表現に必要な形容詞の役割について考えると、形容詞は、音や音楽演奏を表現するための一つの手段といえる。現在は、音楽も文書と同じようにコンテンツとなってしまったので、音楽メディアにアクセスすることが必要となる場合がある。そのような場合、音楽や音楽演奏を適切に表現することが必要になる。その手助けができることを目指して、音楽学、システム開発の両面から研究を進める。

4 実用システムの開発

4.1 適合型コミュニケーション技術の研究開発

英語学習支援システムの開発に向けて、学習者コーパスの作成とエラータグの検討を行っている。学習者コーパスは日本人の英語学習者に対するインタビューを書き起こしたものであり、学習者の英語レベルの判定結果がついていることが特徴である。本研究開発は、通信・放送機構の技術移転プロジェクト「適合型コミュニケーション技術の研究開発」のもとで行っている。

高度情報社会の進展に伴い、人間同士、あるいは人間とコンピュータ間の情報交換を柔軟に行うコミュニケーションの必要性が高まっている。そのため、ちょっとした言葉遣いの誤りや言い間違いを、そのままに受け取ったり、誤りだと反応するのではなく、話し手が何を言いたかったかを推論して適切なコミュニケーションを継続するための技術である「適合型コミュニケーション技術」の確立が重要である。これまで開発してきた自然言語解析技術をもとに、定型的な誤りが多い日本人による英語発話を一例として、誤りを含む発話に強い言語情報付与技術を開発し、併せて誤りを含む発話の言語情報付きデータベースの作成と公開を行うことを目的としている。

適合型コミュニケーションに必要な技術としては、発話意図理解技術や要約文生成技術などがあるが、これらの中核となる技術は、人手を介さずに文章を単語に区切り、各単語の品詞や単語間の係り受け関係などの言語情報を文章に付与する「言語情報自動付与技術」である。従来の言語情報自動付与技術は、文法的に誤っていたり、表面的には話し手の意図を表していないような入力に対しては良好な性能が得られず、そのため適合型コミュニケーション技術へ応用することは難しかった。ここでは、上記の目的に向けた技術移転を実現するための研究開発として、

- 誤りを含む発話の言語情報付きデータベースの作成と公開。
- 誤りを含む発話に強い言語情報付与技術の開発。
- データベース及び言語情報付与技術の有効

性の実証。

を実施する。本研究開発では、誤りを含む発話の具体的対象として、日本人が発話した英語を用いる。英語の能力が低い話し手の発話においては、言外の意味といったものを意識して話すということが少なく、基本的な辞書と文法に関わる定型的な誤りが多いため、誤りのパターンの検出と規則化が容易であるため、早期の研究成果の実現及び技術移転に適した対象である。

今回の研究開発においては、誤りを含む発話の言語情報付きデータベースの作成・公開と、誤りを含む発話に強い言語情報付与技術の開発とを並行して行う。言語情報付与システムの開発において、一つ一つの誤りに対して文法や辞書を人手で修正していくということは現実的ではなく、システムがあらかじめ蓄えられた「誤って使われている言葉」のデータから、文法や辞書を自動獲得することが必要となる。我々が開発したシステムは、少量の言語データから効率よく情報を獲得し、高水準の言語情報付与精度を実現している。本研究開発では、この技術を拡張し、技術移転を推進することを目指している。

4.2　共生コミュニケーションにおける言語処理

IT技術の進展と普及に伴い、IT技術を使いこなせる人と使いこなせない人とのギャップが拡大しつつある。このギャップを埋め、誰もがIT技術を縦横に利用できるためには、印刷されたマニュアルを読むといったものとは対極にある人間的なコミュニケーションによる学習（支援）が必要となる。

ここでは、言語によるコミュニケーションだけではなく、視線・指さしといったnon-verbalな情報伝達、更には形象型のインタフェースによる実在感をも含めたコミュニケーション環境により、人間的な学習支援システムの実現を目指す。このシステムにおいて、自然言語処理が直接関わる部分は、対話処理部と知識処理部である。

ここに組み込まれる対話処理部では、non-verbalな情報と、言語情報とを組み合わせ、ユーザ側から入力された情報を記号化（テキスト化）する。また、知識処理部が作成した応答用情報をもとに、適切な応答をメディアごとに作成する。ここでは、従来からの談話解析研究に加えて、インタフェース部からのマルチモーダル情報を前提とした談話処理手法の研究開発を行うことが必要である。話者モデルに基づく適切な情報提示と、マルチモーダルの有効利用を実現したい。

知識処理部では、知識源として、教科書、マニュアル、新聞記事といったものを用意し、対話処理部の作成した、入力の記号化情報を検索のキーとして、応答するべき情報を抽出し、記号化する。ここでは、大規模文書からの高精度の情報抽出手法を開発する。また、マルチモーダル情報の付加されたテキスト情報の解析と生成手法を開発する。

5　むすび

ここでは、自然言語グループが行っている自然言語に関わる研究開発の一端を述べた。言語学的な基礎研究から、実際のシステム開発に至るまで、幅広い研究開発活動を行っている。今後とも、それぞれの方面に向けての研究開発を進めていく予定である。

井佐原　均（いさはら　ひとし）
情報通信部門　けいはんな情報通信融合研究センター自然言語グループリーダー　工学博士
自然言語処理