研究課題別評価

1. 研究課題名 :自然言語による知識の表現と利用

2. 研究者氏名 :黒橋　禎夫

3. 研究の狙い：

人間の知を理解する上で、また工学的にも、計算機に知的な振る舞いをさせることは大きな目標であり、そのためには計算機上での知識の取り扱いが重要な課題となる。人間は自然言語によって知識を操作するが、計算機にとっては自然言語は曖昧で扱いづらいものであった。そのため、計算機における知識操作には人工的な形式言語が必要であると考えられ、形式言語によって常識を人手で記述するということが試みられた。しかし、そのような方法はコストが膨大であり、また形式言語で書いた知識は保守・拡張が非常に困難であることが認識されてきた。このように、知識の取り扱いの難しさが人工知能研究の大きな障害であった。

これに対して、計算機環境の劇的な進歩と、辞書、テキストなどの大規模データが利用可能となったことで、自然言語処理の研究はこの 10 年間で大きく進展し、テキストの形態、素・構文解析、固有名詞抽出などの基本的な処理ではかなりの精度がえられるようになった。すなわち、自然言語が計算機にとって扱い可能な言語となり、自然言語によって知識を操作するということが少しずつ可能となってきた。

計算機が自然言語で記述された知識、すなわち自然言語テキストを「使いこなせる」ようになれば、web をはじめとする既存の膨大な自然言語テキストを知識源とすることができ、また、人間が日常行っているテキストの更新・追加という方法で知識の保守・拡張が可能となる。そして、なによりも、人間と同じメディアを使うことで、人間と計算機のコミュニケーションが格段に容易になる。本研究では、このようなことを実現するための基礎的研究を行った。

4. 研究結果：

自然言語テキストを計算機が知識として利用するためには、次の 2 つのことを行わなければならない。

1. テキストの構造化(テキスト中の様々な要素の関係付け)
2. 同義異表記の問題の解決

例えば、知識ベース中に「今大会の注目は室伏選手。昨年、世界記録を更新し. . .」というテキストがあり、これを用いて「ハンマー投げで世界記録を塗りかえたのは誰？」という質問に答える状況を考える。上記 1 は、知識ベーステキストの構文・格・省略解析などを行い、「昨年、室伏選手がハンマー投げで世界記録を更新した」という構造を理解することを意味する。上記 2 は、質問の「記録を塗りかえる」と知識ベースの「記録を更新する」との同義性を認識することである。この 2 つの問題を解決することにより、質問に対して「室伏選手」と答えることが可能となる。これは、人間との質問応答だけでなく、知識ベースの内部において推論などを行う場合にも同様である。

このような処理を実現するためには、各語の意味・使われ方に関する正確な情報が必要となり、さらに、言語が表現する世界において一般にどのようなことがおこるか、すなわち常識に関する情

報も必要となる。このようなことを人手で規則として与えることはもちろん不可能である。

そこで、本研究では、各語の基本的な意味に関して国語辞典を、語の使われ方、さらには常識に相当する知識に関して大規模コーパスを知識源とし、これらをうまく組み合わせて利用することにより、自動的に、徐々に高度な知識を獲得していくというアプローチをとった。具体的には、まず大規模コーパスから格フレームと呼ぶ知識構造を抽出し、これを用いてテキストの構造化を行い、さらに格フレームと国語辞典を用いて同義異表記を認識・言い換えるということを行った。これらの各研究項目の成果について以下で詳しく説明する。

4. 1. 格フレーム辞書の自動構築

格フレーム、すなわち各述語(動詞、形容詞など)がどのような項(主語、目的語など)をとるかという情報は、文・文章を構造的に解釈するための必須の情報である。格フレームは各述語に固有であり、さらに、一つの述語は複数の意味をもち、それぞれに異なる格フレームを持つ場合が多い。そのため、カバレージの高い格フレーム辞書を人手で記述することはほとんど不可能である。

そこで、大規模コーパスの構文解析を行い、その解析結果のうち信頼性の高い述語項構造だけを抽出し、その結果を各述語の各用法ごとにクラスタリングすることによって自動的に格フレーム辞書を構築する方法を考案した。

しかし、単純にコーパスから述語項構造を収集するだけでは、二重主語構文、連体修飾の外の関係などの複雑な構文を扱うことができない。そこで、まず大規模コーパスの構文解析結果から単純な格フレームを学習し、次にそれを用いて大規模コーパスの格解析を行い、その結果を収集することによって、より頑健な格フレーム辞書を構築する方法を考案した。

たとえば、「この車はエンジンが良い」という実例について、はじめに構文解析をする段階では「車は」は解釈することができないので、「エンジンが良い」という部分だけが収集される。このようにして収集した「エンジンが良い」の格フレームを用いてもう一度この実例を解析すると、格フレームにヲ格や他の格がないことから、「車は」はガ格であり、「エンジンが良い」は二重主語構文をとることがわかる。これに対して、例えば「その問題は彼が図書館で調べている」の「問題は」は、他の実例データから収集される「図書館で調べる」のヲ格の名詞集合と近いことから、2 回目の格解析では単にヲ格と解析されるだけである。このような原理で、2 回目の格解析の結果から二重主語構文の 2つめのガ格(外のガ格)を収集することができる。

まったく同じ原理により、一度目の解析では「業務を営む免許」の「免許」は扱われないが、2 回目の格解析ではこのような被連体修飾詞がガ格やヲ格であるかどうかが調べられ、そうでない場合には外の関係の名詞であると判断される。外の関係の名詞には、この「免許」のように特定の用言の場合に外の関係となるものと、「可能性」、「結果」、「見通し」などのように一般的に外の関係になりやすいものとがある。これらの区別も、外の関係と判断される名詞の(用言との)分布によって自動的に判断することができる。

この処理を新聞記事 20 年分の大規模コーパスに適用した結果、2 万 3 千の述語について、平均 14.5 個の格フレームを持つ、カバレージの広い実用的な辞書を構築することに成功した。この格フレーム辞書は、今後の自然言語処理の様々な場面で利用できる基礎的知識源となるもので、本研究においても、以下に述べる省略解析、言い換え規則の自動学習などで利用している。

4. 2. 関係」タグ付きコーパスの作成

文章中に存在する種々の関係性を計算機で正確に取り扱うためには、まずその関係のバリエーションを実際のデータ調査に基づいて整理し、それらの関係をタグ付けしたコーパスを整備することが必要である(これはコーパスベースの自然言語処理の基本的な考え方である)。省略・照応を含めた述語項構造、関係名詞などの種々の名詞間関係、共参照関係について、それらの関係を整理し、タグ付けコーパスを作成した。このコーパスは一般に公開する予定である。

4. 3. 省略解析

日本語の文章では格要素が頻繁に省略される(ゼロ代名詞となる)。文章の照合(検索)、要約などを正確に行うためには、文章中の述語項構造の認識、すなわち省略された格要素の復元が必要となる。

一般に、ゼロ代名詞の先行詞はゼロ代名詞から距離が近いところにある傾向がある。しかし、従来の研究では、先行詞から何単語離れているかというような、文の構造を考慮しない単純な距離尺度が用いられてきた。これに対して、我々は従属節、主節、埋め込み文など、構造的にゼロ代名詞と先行詞候補の関係をとらえ、その中でどの位置にあるものがどの程度先行詞となりやすいかを尺度とした。

この尺度の具体的値は、ガ格、ヲ格、ニ格それぞれのゼロ代名詞について。文章中の種々の関係を人手でタグ付けした関係コーパスを利用して計数した。すなわち、先行詞候補の位置 L に対して、(先行詞が L にある回数)/(L にある先行詞候補の数の和)の値を計算した。

入力文に対する省略解析を行う際には、ゼロ代名詞の格ごとに、上記の値が大きな位置にある先行詞候補から順に調べ、ある基準を満たす場合に先行詞とする。ある基準とは、自動学習した格フレームと比較して先行詞と述語の関係が妥当であるかどうか、および、種々の手がかりの機械学習結果から先行詞と判断されるかどうか(学習コーパスは先述の関係コーパス)の AND 条件とした。

関係コーパスの一部、100 記事に対して省略解析の評価実験を行ったところ、まず、格フレームと入力とのマッチングによって判断される、ゼロ代名詞があるかどうかの同定(ある述語に対して省略されている格要素があるかどうかの同定)は適合率 88.7%、再現率 74.7% であった。そして、ゼロ代名詞があると判断された場合の先行詞同定の精度は 63.0%であった。これらの掛算となる全体の適合率・再現率はそれぞれ 55.9%(503/900)、47.1%(503/1068)であった。この値は、この結果を直接利用するアプリケーションの立場からはまだ低いものであるが、省略解析についての学習・評価コーパス、基本的な扱い方、ベースラインとなる精度がえられたので、今後この評価結果を詳細に検討し、種々の言語的制約などを整備していくことで、解析精度を向上させていく予定である。

4. 4. 国語辞典と格フレームを用いた用言の言い換え規則の学習

同一、またはほぼ同一の意味内容に対して、多くの表現が存在するという同義異表記の問題を解決しなければ、計算機による質問応答や、計算機内での自然言語による推論を行うことはできない。この問題を解決するために、国語辞典の情報をもとに言語表現をより平易な方向に言い換え、それによって同じ意味のさまざまな表現を一つの表現に収束させることを考えた。

国語辞典に基づく言い換えの基本的なアイデアは、与えられた文中の各単語を、国語辞典から得られる上位語と置き換えるというものである。しかし、この方法で簡単な単文の用言を言い換えるだけでも、多くの問題を解決しなければならない。例えば「他社をしのぐ」を『しのぐ』の定義を用いて「他社より優れている」に言い換える場合、次のような問題がある。

1. 『しのぐ』の多義性解消、
2. 『しのぐ』と置き換えられる定義文中の部分表現の決定(たとえば「体得する」を言い換える場合、単に「つける」ではなく身につける」としなければならない)、
3. 表層格パターンの対応付け(「をしのぐ」が「より優れる」になる)。

これらの問題を、先に述べた格フレーム辞書を利用することによって解決する方法を考案した。例えば上記の例では、言い換えるべき用言(『しのぐ』)の各格フレームについて、最も類似する言い換え先用言(「耐え忍ぶ」「優れる」)の格フレームを見つけ出し、さらにその間で格要素間の対応付けを行うことにより、上記の 3 つの問題を一気に解決することができる。この方法は、国語辞典という人手で高度に整理された情報を、大規模コーパスを用いてさらに補足・強化したものと考えることができる。

この方法を実験文 220 文で評価したところ、77%の精度で適切な言い換え表現をえることができ、単純なベースラインの手法(辞書中の最初の語義を採用し、格助詞の言い換えなどを行わない方法)と比較して 11 ポイントの精度向上を得ることができた。ここで学習した言い換え規則は、今後、同義異表記の問題を扱う上での基本的なモジュールとして利用していく予定である。

4. 5. 迂言表現と重複表現の認識と言い換え

前節で扱ったような異なる表現ではなく、ある表現に余分なものが付加された、ある種の冗長性によって表現のずれが生じる場合がある。例えば、「パソコンを買う」のような「体言＋格助詞＋用言」という形の句を考えた場合、一般には体言と用言の意味は独立であり、その二つの意味の組合せとして句の意味が構成される。しかし、次の(1a)、(2a)の表現はこのような一般的な表現ではない。

(1) a. 改革を断行する / b. 思い切って改革する

(2) a. 貯金をためる / b. 貯金する / c. お金をためる

(1a)の例では、「断行」にいわゆる述語や項の意味はなく、副詞的な意味のみがある。ゆえに、これをより簡潔な副詞的表現に置き換えた(1b)の言い換えが考えられる。また、(2a)では「貯金」と「ためる」の間に意味の重複があり、この重複を取り除くと、(2b)、(2c)のより簡潔な表現が得られる。ここでは、(1a)を迂言表現、(2a)を重複表現とよぶことにする。

迂言表現と重複表現に関するずれを吸収するために、国語辞典の語の定義文を利用することにより、それらを認識し、言い換える方法を考案した。迂言表現は、付属要素の語の定義文の特徴から認識することができる。たとえば「断行」の定義文は「思いきってやること」、「強制」の定義文は「むりにさせること」となっている。このように、定義文が、「副詞的表現＋一般的な動作表現＋ヴォイスやアスペクトを表す付属語」であるものが迂言表現の付属要素と考えることができる。「一般的な動作表現」とは「物事をする」、「行う」、「やる」などで、辞書を調査したところ 20 程度の語句であることがわかった。付属要素がわかれば迂言表現の言い換えは比較的簡単で、(1a)に対する(1b)や「寄付を強制する」に対して「むりに寄付させる」を得ることができる。

一方、重複表現は、包含する方の(大きな意味の)語の定義文に、包含される方の語が含まれていることで認識できる。たとえば(2)の例であれば、「貯金」の定義文が「お金をためること」とあるので、この中に「ためる」があることで「貯金をためる」は重複表現とわかる。これが認識されれば、「ためる」を削除し、「貯金」を用言化して「貯金する」としたり、さらに「貯金」の定義文を用いて「お金をためる」と言い換えることができる。

このような手法の有効性を確かめるために、例解小学国語辞典を用い、新聞記事からランダムに取り出した600表現を対象に評価実験を行った。まず、人手でこれらの600表現を調べたところ、84の迂言・重複表現があった。これに対して、システムの解析結果は適合率65%、再現率58%であった。今後、この方法によって迂言・重複表現のずれを吸収することが情報検索、対話システムなどでどの程度有効であるかを検討する予定である。

5. 自己評価：

研究計画においては、1.言い換え関係の認識・生成による言語の冗長性の吸収、2.テキスト中の関連性の複合的関連付け、3.これらの機能を質問応答、対話システム等に組み込むこと、を目標とした。このうち、1.については国語辞典による用言の言い換えと、迂言、重複表現の言い換えを実現し、2.については格フレームの自動構築、省略解析システムの構築、学習用データとしての関係コーパスの構築などを実現した。また、3.についてもPC環境に関するヘルプシステムにおいて、言い換え、関係付けの成果の導入を行い、その有効性を検証した。このように、当初目標とした3項目それぞれについて、十分な成果をおさめることができた。

今後は、さきがけ研究で進めてきた方法論、すなわち、大規模コーパスと辞典を融合して利用する方法を発展させることにより、上位下位関係に代表される用語のオントロジーの自動獲得、名詞を中心とした関係性辞書の自動構築、さらに、述語項構造の連鎖関係、すなわち因果関係的な常識の自動獲得に発展させることを検討している。

また、PC環境に関するヘルプシステムはすでにwebを通して実際にサービスを行っているものであり、このシステムを自然言語理解技術のテストベッドとして、要素技術の統合、高度化を進める予定である。

6. 研究総括の見解：

計算機による自然言語理解は、情報化社会の諸問題を解決するキー・テクノロジーの一つとして期待されている。本研究では、大規模な電子テキストや電子化辞書をもとにして、自然言語理解のための知識を、自動的に、段階的に獲得する枠組みを考案した。特に、文理解の基本となる格フレーム辞書について、従来の人手で構築されてきた辞書を質、量ともに上回るものを自動処理によって構築したことは大きな成果である。さらに、この辞書を利用することによって、同義異表記関係のパターンを自動学習する方法を考案しており、知識獲得のブートストラップのプロセスを具体的に示した功績も大きい。研究提案時には、言語の基本語彙の問題に取り組むという項目も挙げられていたが、この問題に十分に取り組めなかった点は残念である。今後の研究の発展に期待したい。

7. 主な論文等：

学術雑誌および国際会議

1. 鍜治伸裕, 黒橋禎夫: 迂言表現と重複表現の認識と言い換え, 自然言語処理, Vol.11, No.1, (2004.1 採録予定)
2. 清田陽司,黒橋禎夫,木戸冬子: 大規模テキストl知識ベースに基づく自動質問応答 ーダイアログナビー, 自然言語処理, Vol.10, No.4, pp.145-175(2003.7)
3. 鍜治伸裕,河原大輔,黒橋禎夫,佐藤理史: 格フレームの対応付けに基づく用言の言い換え, 自然言語処理, Vol.10, No.4, pp.67-81 (2003.7)
4. Daisuke Kawahara and Sadao Kurohashi: Fertilization of Case Frame Dictionary for Robust Japanese Case Analysis, In Proceedings of 19th COLING (COLING02), pp.425-431 (2002.8).
5. Youji Kiyota, Sadao Kurohashi, and Fuyuko Kido: Dialog Navigator" A Question Answering System based on Large Text Knowledge Base, In Proceedings of 19th COLING (COLING02), pp.460-466 (2002.8).
6. Nobuhiro Kaji, Daisuke Kawahara, Sadao Kurohashi, and Satoshi Sato: Verb Paraphrase based on Case Frame Alignment, In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL02), pp.215-222 (2002.7).
7. Daisuke Kawahara, Sadao Kurohashi, and Koichi Hasida: Construction of a Japanese Relevance-tagged Corpus, In Proceedings of The Third International Conference on Language Resources & Evaluation, pp.2008-2013 (2002.5).
8. 河原大輔,黒橋禎夫: 用言と直前の格要素の組を単位とする格フレームの自動構築, 自然言語処理, Vol.9, No.1, pp.3-19 (2002.1).
9. Daisuke Kawahara and Sadao Kurohashi: Japanese Case Frame Construction by Coupling the Verb and its Closest Case Component, In Proceedings of First International Conference on Human Language Technology Research (HLT 2001), pp.204-210, San Diego, California, (2001.3.18-21).

解説記事、招待・依頼講演等

1. 黒橋禎夫: 自然言語処理の紹介, 土木学会第1回情報社会基盤小委員会 (2003.7.23).
2. 黒橋禎夫: コンピュータによる自然言語処理：入門と応用 質問に答えるコンピュータ」, 国際日本文化研究センターシンポジウム (2003.7.15).
3. 黒橋禎夫: 大規模テキストl知識ベースに基づく自動質問応答システム, 電子情報通信学会言語理解とコミュニケーション研究会 質問応答 (QA)技術最前線 - QA の現状と今後の可能性」講習会 (2003.1.27).
4. 黒橋禎夫,清田陽司,木戸冬子: 自動質問応答システム・ダイアログナビの現状と課題, 情報処理学会研究会 音声言語情報処理 43-4, pp.19-24 (2002.10.25).
5. 黒橋禎夫: 自然言語処理を支える文法, 月刊 言語」,Vol.31,No.4, pp.52-57 (2002.4).
6. 黒橋禎夫: 大規模テキストl知識ベースに基づく自動質問応答, 第3回音声言語シンポジウム, pp.37-42 (2001.12.21).
7. 黒橋禎夫: 言語処理・資源に関する世界の Initiative の動向, JEITA 自然言語処理技術シン

ポジウム (2001.12.12).
8. 黒橋禎夫: 計算機による言語の理解, 第 15 回人工知能学会全国大会 AI レクチャ (2001.5.24).