様式 C-19

科学研究費補助金研究成果報告書

平成２３年２月２８日現在

機関番号：14301
研究種目：特定領域研究
研究期間：2007～2010
課題番号：19024040
研究課題名（和文）　構造的言語処理による情報検索基盤技術の構築

研究課題名（英文）　Construction of Information Retrieval Infrastructure Based on Structural Natural Language Processing

研究代表者
黒橋　禎夫（SADAO　KUROHASHI）
京都大学・大学院情報学研究科・教授
研究者番号：50263108

研究成果の概要（和文）：情報検索の本来の目的は、表面的なテキストではなく、その中の情報・知識を得ることであり、そのためには計算機によるテキストの理解、言語の理解が本質的に必要となる。構造的言語処理によって、語を単位とするのではなく述語項構造を単位とし、言語表現の多様性を吸収し、検索結果クラスタリングに基づく鳥瞰図的把握を提供する，次世代情報検索の基盤技術を構築した。

研究成果の概要（英文）：The essential purpose of Information Retrieval is not to get relevant documents, but to obtain relevant information and knowledge. In order to achieve this, we believe that text understanding by machine, or Natural Language Processing is the most important aspect. This research project constructed IR infrastructure based on structural NLP, analyzing predicate argument structures in texts, handling expressive diversity in natural language, and providing a bird's-eye view towards a given topic by organizing and relating information.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	8,700,000	0	8,700,000
2008 年度	8,700,000	0	8,700,000
2009 年度	9,000,000	0	9,000,000
2010 年度	9,000,000	0	9,000,000
総　計	35,400,000	0	35,400,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：自然言語処理、情報検索、述語項構造、柔軟マッチング、クラスタリング

１．研究開始当初の背景

ウェブをはじめとする電子テキスト情報が爆発的に増加し、我々の社会、生活の基盤となっている。この情報を整理し、人間が高度に利活用するためには、計算機によるテキスト・言語の高度な理解が必要である。しかし、従来の情報検索の研究は文書を単語の集合として近似することを基本とする研究や、種々の制限のもとで商用検索サイトの API を利用して行う研究が一般的であった。また、構造的言語処理を検索に使うプロジェクトとしては Powerset などがあるが、Wikipedia のテキストだけを検索対象とした小規模なものであった。

２．研究の目的

本研究課題では、言語の構造や意味を高度に解析し、これに基づく情報検索基盤技術を構築することを目指し、次のことを研究目的とした。

(1) 語が意味をあらわす基本単位と考えられ

るが、そこには、同義語・多義語、さらには俗語・専門用語・新語などの多様な問題が存在する。これらについて、ウェブスケールの自動獲得、自動解析を実現するとともに、これを情報検索において利用する枠組みをデザインする。

(2) テキストの意味は単語集合として近似するだけでは不十分であり、「誰が何をどうした」という述語項構造によってはじめて捉えることができる。様々な分野・タイプのテキストが存在するウェブに対して、文をまたぐゼロ照応現象を含めた頑健で高精度な述語項構造解析を実現する。

(3) 情報の整理・組織化のために、与えられた課題(クエリ)に対する重要概念を検出し、検索結果を組織化し、さらに、インタラクティブなやりとりによってユーザが検索結果を再組織化する方法を実現する。

３．研究の方法

(1) 計算機が利用できる形での知識の自動獲得を行い、これによって言語解析システムの高度化を実現する。これを可能とするのは、まさに爆発しているウェブ上の超大規模コーパス（電子テキスト集合）である。超大規模コーパスの自動解析結果から確からしい情報を抽出・集約して利用すること、また、Wikipedia などの比較的高品質な知識源を活用することでこれらを実現する。

(2) ウェブスケール、すなわち億の規模のテキスト処理、知識獲得、情報検索を実現するために、研究コミュニティとしての共通研究基盤を構築する。まず、ウェブテキストを言語資源として利用するためにページのメタ情報、文区切り、言語解析結果などを標準的なフォーマットのもとに管理する。また、新たな検索機能の追加、上位の情報分析システム開発の基盤となるオープンな検索エンジンを構築する。

(3) 上記の(1)、(2)を実現するために超大規模計算環境を利用する。具体的には、本特定研究の A02 班によって提供される多拠点分散計算環境(InTrigger)、東京工業大学 TSUBAME をはじめとする計算機センター、Microsoft 社のクラウド Windows Azure 等を活用する。

４．研究成果

(1) 開放型検索エンジン基盤 TSUBAKI の構築

日本語ウェブ文書 1.2 億ページを対象とする開放型検索エンジン基盤 TSUBAKI の開発・運用を行った。

TSUBAKI の目的の１つは、知識獲得のための言語資源としてウェブ文書の整備を行い、これをコーパスとして提供することである。ウェブテキストを対象として深い言語処理を指向する研究を行うには、研究に至るまでに直面する面倒な処理が多い。たとえば、ウェブテキストでは句点だけでなく HTML タグや顔文字が文の境界を意味することもあり、文区切りが不明瞭な場合が多い。また、単語分割などの基本的な処理を各研究者がそれぞれに行うことも効率的ではない。そこで、ウェブテキストに対して、タイトル、URL などのメタ情報や、文区切り、文の基本解析結果などを管理する XML 形式のウェブ標準フォーマットをデザインし、1.2 億ページをこの形式に変換し、16 億文の日本語コーパスとして整備した。この言語解析データは TSUBAKI の API を介して誰でも自由にアクセス可能であり、本特定領域の中でも鳥式改や言論マップなど多くの研究プロジェクトで利活用されている。

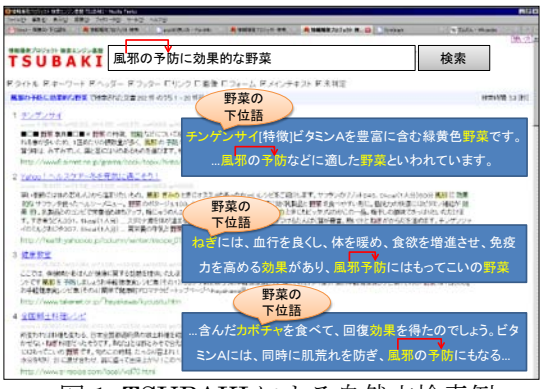

図 1. TSUBAKI による自然文検索例

TSUBAKI のもう１つの目的は、深い言語処理に基づくインデキシングのプラットフォームとなり、さらに、上位のより高度な情報分析システムの基盤となることである。既存の検索エンジンの多くが単語だけをインデックスに登録しているのに対し、インデキシング・プラットフォームとしての TSUBAKI は後述する同義関係、上位下位関係、述語項構造関係などをインデックスする枠組みを提供し、これらによって、クエリと文書に書かれている内容の意味的な一致をより柔軟・正確にとらえることを可能とし、ユーザの情報要求を直接的に表現する自然文検索を強力にサポートしている。その結果「風邪の予防に効果的な野菜」というクエリに対して、「効果＝効く」「野菜←ねぎ（上位下位関係）」などを考慮した検索を実現している（図１）。さらに、特定領域の中で研究された「トヨタ≒ト○タ」のような隠語のインデキシングも行っている。

TSUBAKI は強力な自然文検索を実現する基盤として様々な展開を行っており、京都大学附属病院の診断画像所見検索、英文構文解析器 Enju との統合による MEDLINE 検索な

どへの適用を進めている。また、情報分析システムの基盤として、本特定領域の種々の研究成果を統合した情報処理学会論文検索システムの基盤、根拠サーチ(東北大)の基盤、情報通信研究機構の情報分析システム WISDOM 等の基盤として活用されている。

(2) ウェブからの未知語自動獲得

大規模ウェブテキストを処理する際、これまでに人手で登録した語彙だけでは不十分であり、未知語（例えば「ググる」）に起因する解析誤りが問題となる。そこで、計算機によってウェブテキストから未知語を自動獲得し、自動的な辞書更新を行うことによりこの問題を解決する手法を考案した。

ここでのポイントは、テキストから未知語を獲得するためのメタ知識を、すでに人手により辞書登録された形態素（既知語）を利用して収集することである。また、人手により付与されるのと同等の情報を計算機が一気に獲得するのは困難なため、段階的に知識獲得を行う。すなわち、未知語検出、未知語同定、名詞の意味分類というサブタスクを設定し、それぞれに対して解法を提案した。

未知語検出: まず、テキスト中の未知語候補を検出するために、未知語を 2 種類に分類した。一つは、既知語では解釈し得ない未知語である。例えば未知語「グーグル」は「グー」から始まる既知語を持たないため、既知語とは解釈できず、自明に検出できる。一方、既知語の組み合わせで解釈できる未知語も存在する。例えば未知語「うざい」は既知語「う」と「ざい」の組み合わせと解釈でき、検出が自明でない。こうした過分割は、人間が見ると意味的に不自然であり、計算機にもその不自然さを気づかせる必要がある。そのために既知語に付与された表記ゆれの知識を用いる。「う」に対応する異表記は「卯」「雨」などであり、もしこの分割が正しいなら、「卯ざい」や「雨ざい」といった表記がテキスト中に出現するはずであるが、実際には出現しない。このように表記ゆれ知識を用いることにより過分割された未知語の検出を可能にした。

未知語同定: テキスト中の未知語の境界を同定し、文法役割を表す品詞を付与するタスクを未知語同定と呼ぶ。例えば未知の動詞「ググる」の境界を同定し、ラ行動詞という品詞を付与する。同定に用いるメタ知識として、形態変化の規則性に着目し、これを既知語のテキスト中での振る舞いから獲得する。例えば、既知のラ行動詞「走る」は、「走らない」、「走るとき」、「走れ」などの形をとることを知っておけば、未知語候補がテキスト中で「ググらない」、「ググるとき」、「ググれ」などの形で出現したとき、これがラ行動詞「ググる」であると同定できる。

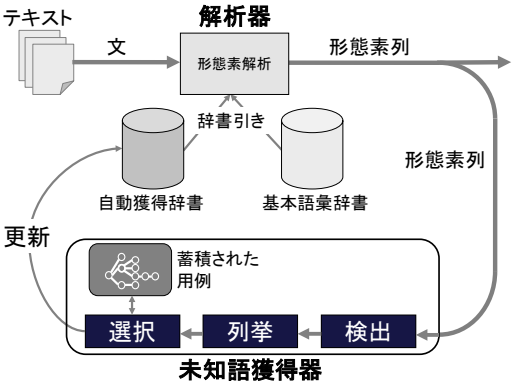

図 2. テキストからの未知語自動獲得

名詞の意味分類: 未知語同定により獲得された形態素（獲得語）のうち、名詞に対して「人名」、「地名」、「人」などの意味ラベルを付与する。こうした意味分類間には明確な文法的差異がないため、特定の語彙との結びつき（語彙的選好）を分類に用いる。例えば、テキスト中に「X 駅」とあれば X は地名らしい。また、「X を通り越す」とあれば、X は「地名」または「場所」と推定できる。こうした分類規則を複数組み合わせることにより、X の意味分類が特定できる。ただし、分類規則を人手で書き下すのは現実的でない。そこで、既知語のテキスト中の振る舞いからこうした規則を自動獲得し、獲得された知識を獲得語に適用する。

上記の未知語検出と未知語同定の提案手法を用いて、図２のようなオンライン未知語獲得システムを実現した。未知語獲得の従来手法が、テキスト全体を一度に処理するのに対して、オンライン未知語獲得システムは、逐次的に入力されるテキストから未知語を獲得する機能を持つ。この機能を用いると、例えば、人気ウェブサービス Twitter 上での人々のつぶやきからリアルタイムに新語を獲得できる。

未知語自動獲得に関する研究成果は次のようにまとめられる。まず、表記ゆれ知識を利用した未知語検出により、過分割されたひらがな未知語の検出率が大幅に向上した(34.5%から 72.0%)。ひらがな未知語は数は多くはないが、その同定は実用上重要である。ひらがな未知語の解析誤りは、応用処理に深刻な副作用をもたらすからである。

未知語同定については、比較的少数の例(4-7)から高精度(97.3-98.5%)で未知語が獲得できることが示された。従来研究が、頻度 10 未満の候補を獲得対象から外していたことを考慮すると、非常に効率が良い。また、獲得にともなう辞書拡張により、形態素解析の精度も向上することを示した。また、自動獲得した名詞の意味分類は F 値で 85.4%という実用的な精度を得た。

さらに、こうした研究成果を応用して、大規模ウェブコーパスから比較的高頻度の未知語を獲得した。得られた語彙は、日本語形態素解析システム JUMAN6.0 に同梱して配布している。

(3) 同義・上位下位関係の自動獲得と情報検索での利用

情報検索において、検索クエリとテキスト間の表現のずれが大きな問題となる。そこで、国語辞典・Wikipedia の定義文や大規模コーパスから語の同義・上位関係を自動獲得し、それらの知識を利用することにより、検索クエリとテキスト間の柔軟なマッチングを実現した。

まず、国語辞典・Wikipedia の定義文から同義語または上位語を抽出した。定義文の主辞を上位語とみなすことができ、また、定義文が短い場合は全体を同義語とみなすことができる。例えば、「我が国」の定義文が「自分の国。日本のこと。」であるので、「我が国」の上位語が「国」であり、「我が国」の同義語が「日本」であるという知識を獲得することができる。また、「EU」の定義文は「欧州連合」であるので、「EU」の同義語として「欧州連合」という知識を獲得することができる。このような処理を国語辞典と Wikipedia の定義文に対して行うことにより、語の同義・上位下位関係を網羅的に獲得することができた。

さらに、大規模テキストから、「(景気が) 冷え込む」と「(景気が) 悪化する」のように、述語単体では同義でないが文脈に依存して同義関係となる述語ペアを自動獲得した。同義関係の獲得には、例えば「(景気が) 冷え込む」、「(景気が) 悪化する」ともに、修飾する述語には「低迷」「増税」など、修飾される述語には「下落」「下げた」などが出現し、このような出現分布が類似していることを利用する。まず、格要素と述語を組とした単位に対して係り受け関係にある述語を要素とした素性ベクトルを構築した。そして、格要素が同一で述語が異なる任意の組に対して分布類似度を計算し、類似度の高いペアを同義表現として獲得した。

自動獲得した同義・上位下位関係を効率的に扱うために SynGraph というデータ構造をデザインした。SynGraph データ構造は、入力文の構文木を基に、基本句をノード、係り受け関係をエッジとしたグラフ構造であり、さらに、自動獲得した同義・上位下位関係がノードに付与されたものである。

図３の例では、「日本」に対して上位語「国」が付与され、また、「景気が冷え込む」という複数ノードに対して同義句「景気が悪化する」が付与される。また、「EU」には「欧州連合」以外にも「ユウロピウム」「愛媛大学」

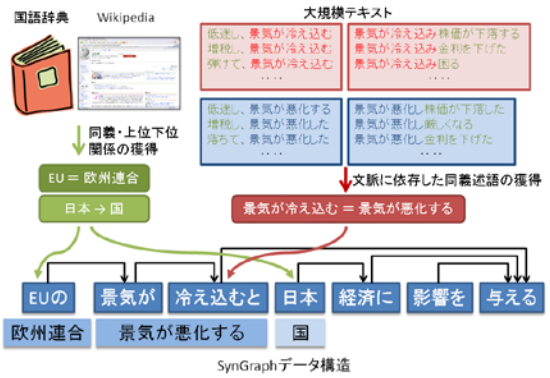

図 3. 同義・上位下位関係の自動獲得と SynGraph 構造

などのいくつかの意味があるが、多義性解消を行なうことにより、文脈での意味を同定し、この例の場合、同義語「欧州連合」のみが付与される。

このようにして獲得された同義・上位下位関係を検索エンジン基盤 TSUBAKI に導入した。クエリ・テキストともに SynGraph データ構造に変換して扱うことにより、クエリ・テキスト間の柔軟なマッチングを実現することができた。以下に利用された知識と、マッチングが行なわれたクエリ・テキストの例を示す。

クエリ: 高齢化社会で発展が見込める市場
テキスト: 高齢化社会の到来で、国内の医薬品需要のさらなる高まりを見込んだ、外資系メーカーの参入も活発化している。… 成長市場である欧米などでの...

クエリ: 風邪の予防に効果的な野菜
テキスト: ...カボチャを食べて、回復効果を得たのでしょう。ビタミンAには、同時に肌荒れを防ぎ、風邪の予防... (※上位下位関係)

クエリ: 大学を出るまでにいくらかかるか
テキスト: ...大学を卒業して一人前になるまでの 22 年間に、いったい、どれくらいの金額が必要なのでしょうか。...

(4) 述語項構造による文の深い理解の実現と検索の高度化

検索クエリに含まれる複数の語が出現するテキストであっても、それらの語と語の関係が検索クエリにおける関係と異なれば、検索意図に合うテキストではないと考えられる。より高度な検索を実現するためにはこのような関係を考慮する必要があり、そのためには、述語とその項の関係をはじめとしたテキスト中の語と語の関係を正しく認識することが必要となる。

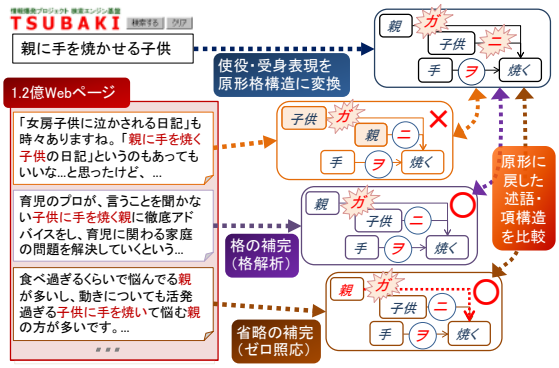

図 4. 述語項構造に基づく検索の高度化

日本語において述語項関係は格助詞の種類によってある程度判断できる。しかし実際のテキストにおいては、項が省略されていたり（これをゼロ照応とよぶ）、格助詞が明記されていなかったり、また、使役・受身表現の場合は格助詞が変化するなど、表層的な記述からは述語と項の関係が判断できない場合が多い。そこで、16 億文コーパスから自動獲得した述語と項の関係に関する知識（格フレーム）を基に、省略された項、および、格助詞を補完し、原形の格構造に直した述語項構造を認識するシステムを構築し、述語項構造レベルでの正確なマッチングに基づく検索を実現した。

この中で、160 万文から 16 億文までの異なるサイズのコーパスから格フレームを自動獲得し、格フレームに基づく構文・格解析モデル、および、照応解析モデルに適用した結果、より大規模なコーパスを用いることで格解析・照応解析の精度が向上すること、16 億文を利用した場合でもその精度は飽和していないことを示した（図５）。

ゼロ照応解析については、省略された項・格助詞を補完するために、語彙的な手掛かりについては大規模な格フレームから、構文的・談話的な手掛かりについては比較的小規模な人手で作成した述語項関係タグ付きコーパスから獲得・利用するモデルを構築した。また、使役・受身表現の原形格構造の認識に関しては、事前に使役・受身形の格フレームと対応する原形の格フレームの類似度を計算し、それらの対応を取ることで、使役・受身形を原形に変換するシステムを構築した。

人手で正しい照応関係を付与した Web テキスト 186 記事を用いた実験の結果、ゼロ照応解析において 0.40 程度の F 値で認識を行うことに成功した。さらに、検索において重要になると考えられる動作主の省略については F 値 0.55、これを同一文内の出現に限定する場合には F 値 0.70 を達成した。

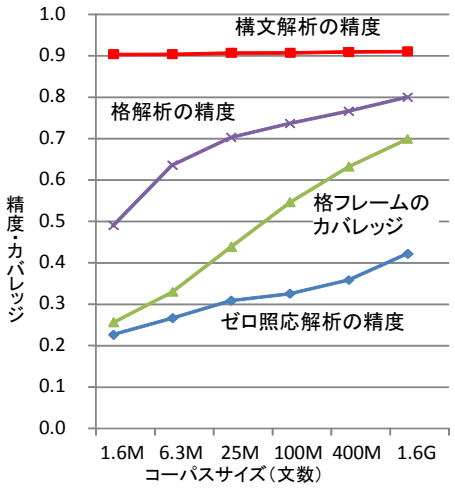

図 5. コーパスサイズと解析精度

(5) 検索エンジン基盤上でのクラスタリングシステムの構築

日本語 1.2 億ページの検索エンジン TSUBAKI を基盤として、クエリに対する重要関連表現を検索結果文章中から自動抽出し、各表現を含む文書を一つのクラスタと考えるラベルベースのクラスタリングシステムを構築した。

処理対象とするテキスト量について、従来の研究では既存検索エンジンの API 等で得られるスニペット（ページ要約文）100 文程度が対象であったのに対して、TSUBAKI の利用によって数万文の言語解析結果を高速に利用することができ、これによってクエリに対して重要関連語を網羅的に取得することを可能とした。さらに、関連表現抽出においては、自動獲得した同義関係知識を利用し、キーワードの表記ゆれ、同義語、包含関係などを徐々に集約していくキーワード蒸留という手法を考案し、これによって高精度に関連語を抽出することに成功した。抽出した関連語を固有名詞のタイプ、複合語の語構成などによって整理することにより、クエリの関連項目を鳥瞰図的に眺めることができるシステムを構築した。さらに、ユーザが関連語を選択すると、その関連語を含む文書を収集し、関連語に関する詳細な情報を提示する機能を実現した。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計８件）

① Ryohei Sasano, Daisuke Kawahara and Sadao Kurohashi: The Effect of Corpus Size on Case Frame Acquisition for Predicate-Argument Structure Analysis,

IEICE TRANSACTIONS on Information and Systems, Vol.E93-D, No.6, pp.1361-1368 (2010.6),査読有

② 村脇有吾,黒橋禎夫: 形態論的制約を用いたオンライン未知語獲得, 自然言語処理, Vol.17, No.1, pp.55-75 (2010.1), 査読有

③ 馬場康夫, 新里圭司, 柴田知秀, 黒橋禎夫: キーワード蒸留型クラスタリングによる大規模ウェブ情報の俯瞰, 情報処理学会論文誌, Vol.50, No.4, pp.1399-1409 (2009.4),査読有

〔学会発表〕(計２０件)

① Daisuke Kawahara and Sadao Kurohashi: Acquiring Reliable Predicate-argument Structures from Raw Corpora for Case Frame Compilation, In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC10), pp.1389-1393, Malta (2010.5.20).

② Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto and Sadao Kurohashi: TSUBAKI: An Open Search Engine Infrastructure for Developing New Information Access Methodology, In Proceedings of Third International Joint Conference on Natural Language Processing (IJCNLP2008), pp.189-196, Hyderabad, India (2008.1.8).

③ Tomohide Shibata, Michitaka Odani, Jun Harashima, Takashi Oonishi and Sadao Kurohashi: SYNGRAPH: A Flexible Matching Method based on Synonymous Expression Extraction from an Ordinary Dictionary and a Web Corpus, In Proceedings of Third International Joint Conference on Natural Language Processing (IJCNLP2008, poster), pp.787-792, Hyderabad, India (2008.1.9).

〔その他〕
報道関連

① 2010年8月12日 NHK教育 ITホワイトボックス2「第18回 検索の未来はどうなるの？」

② 2009年8月号 (2009年6月26日発売) Newton 「検索は「キーワード」から「文章」へ」

③ 2008年7月5日 朝日新聞b3面 「ポスト・キーワード検索」

④ 2007年8月21日 日経産業新聞10面 「情報爆発に立ち向かう(上) 疑問に答える検索技術」

URL: 検索エンジン基盤TSUBAKI
http://tsubaki.ixnlp.nii.ac.jp/

６．研究組織

(1)研究代表者
黒橋　禎夫 (KUROHASHI　SADAO)
京都大学・大学院情報学研究科・教授
研究者番号：50263108

(2)研究分担者
河原　大輔 (KAWAHARA　DAISUKE)
京都大学・大学院情報学研究科・准教授
研究者番号：10450694
柴田　知秀 (SHIBATA　TOMOHIDE)
京都大学・大学院情報学研究科・助教
研究者番号：70452315

(3)連携研究者
なし

(4)研究協力者
新里　圭司 (SHINZATO　KEIJI)
京都大学・大学院情報学研究科・特定研究員
笹野　遼平 (SASANO　RYOHEI)
東京工業大学・精密工学研究所・助教
研究者番号：70603918