様式 C-7-2

自己評価報告書

平成２２年５月１１日現在

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

研究分野：自然言語処理
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　クラスタリング　述語項構造　柔軟マッチング

１．研究計画の概要

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，究極的にはテキスト・言語の計算機による理解が必要である．この問題意識のもとに，本計画研究では構造的言語処理による情報検索基盤技術に関する研究として，格フレームに基づく省略照応解析，同義異表記の知識獲得と利用，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を行う．

２．研究の進捗状況

(1) テキストで述べられていることは「誰が何をどうした」という述語項構造を基本単位として考えることができる．このうち文中の明示的な構造は構文解析によって 90%程度の精度で解析が可能である．一方，文中で省略・代名詞化されている項については，文脈の中でそれらを特定する必要があるが，ウェブの多様なテキストを対象とした場合にはその精度は平均的には 20%程度であり，構造的言語処理の最大のボトルネックであった．この問題に対して，例えば「会社が商品を発表する」などの述語項構造のパターン（格フレーム）を 16 億文の大規模コーパスから自動学習し，格フレームとの対応付けの整合性を文書全体で最適化することによって省略照応解析の精度を 41%に向上させた．さらに，格フレームの学習コーパス量と格フレームのカバレージ，省略照応解析の精度の関係を明らかにし，学習コーパスの増加によって一層の解析精度向上がみこめることを示した．

（2）同義語や言い換え表現などの同義異表記に対して，一般用語レベル，専門用語レベルそれぞれにおける網羅的自動獲得を行った．一般用語については，国語辞典に示されている同義語および短い語釈文を抽出し，約 6 千の同義関係を獲得した．たとえば「最寄り」の語釈文「いちばん近い所。近所。」から，「最寄り」と「近所」が同義語であり，「最寄り」が「いちばん近い（所）」と言い換え表現であることを抽出した．専門用語については，文書中の「メタボリックシンドローム（内臓脂肪症候群）」「内臓脂肪症候群（メタボリックシンドローム）」のような双方向の括弧表現，および Wikipedia から約 1 万の同義関係を獲得した．

さらに，これらの同義異表記を構文木の各語／句に付与した Syngraph とよぶデータ構造で表現し，これによって「最寄り＝いちばん近い＝もっとも近い」などの同義異表記の組み合わせを扱うことを可能とした．また，この結果を検索のインデックスとして利用し，1 億ウェブページを対象として網羅的に同義異表記を処理する検索を可能とした．

（3）申請者らが本領域支援班で構築している日本語 1 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，キーワードの表記揺れ，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

３．現在までの達成度
①当初の計画以上に進展している．
（理由）
省略照応解析については当初の予定どおり進展している．同義異表記問題については，文脈に依存する句の同義表現獲得も実現し，さらに同義性と多義性の統合処理に進展させている．クラスタリングシステムについても，キーワードによるシステムはすでに構築し，クラスターの要約文を生成する課題に挑戦している．

４．今後の研究の推進方策
（１）これまでに得られた知見から，格フレームの学習コーパス量の増加によって省略照応解析の精度向上がみこまれる．現在は 16 億文コーパスで精度 41%であるが，60 億文規模のコーパスを整備し，これによって精度 60%程度が達成されることを確認する．一方，情報検索・情報組織化への貢献を考えた場合，省略照応解析について少なくとも 80%程度の精度が要求される．しかし，この精度はウェブ全体を対象とする必要はなく，比較的高品質で情報価値の高い部分に対して達成できればよい．そこでテキストの種々の特徴量から情報価値の高い部分を自動選択し，その部分について 80%の省略照応解析が行えるよう，エラー分析に基づき解析手法を高度化する．

（２）同義性と多義性は深く関係している．情報検索の高度化のためにはこの問題の統合的解決が必要となる．すなわち，多義語についてその文脈での意味を特定し，その意味での同義語とのマッチングを許す．多義語についてどのような意味があるかは国語辞典から基本語彙 3 万語，ウェブから専門語彙 2 万語についてすでに自動学習を行っているので，それぞれの意味について文脈中の共起語を学習し，多義性解消を行うシステムを構築し，この結果を情報検索で利用する．

（３）これまでに，与えられたクエリに対して重要関連語を検出し，これを固有名詞のタイプ，語構成の情報から整理するシステムを構築している．クエリ関連話題の把握をさらに促進するために，たとえば「グリンピース」「商業捕鯨」を独立に示すのではなく，関連語を含む述語項構造を検索結果テキストから抽出することにより「グリンピースが商業捕鯨に反対している」という関連語間の主要な関係を提示するシステムを構築する．

５． 代表的な研究成果

〔雑誌論文〕（計 7 件）

① 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，査読有，Vol.17，No.1，pp.55-75，2010.
② 柴田知秀，姜ナウン，黒橋禎夫，同一文抽出に基づく類似ページの検出と分類，人工知能学会論文誌，査読有，Vol.25，No.1，pp.224-232，2010.
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，査読有，Vol.50，No.4，pp.1399-1409，2009.

〔学会発表〕（計 13 件）

① Ryohei Sasano, Daisuke Kawahara and Sadao Kurohashi, The Effect of Corpus Size on Case Frame Acquisition for Discourse Analysis, North American Chapter of the Association for Computational Linguistics - Human Language Technologies, pp.521-529, Boulder, Colorado (2009.6.3).
② 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, Third International Joint Conference on Natural Language Processing, pp.787-792, Hyderabad, India (2008.1.9).

〔その他〕
報道関連
・検索は「キーワード」から「文章」へ，日経産業新聞（2007 年 8 月 21 日 10 面）
・「情報大爆発」どうさばく，朝日新聞 be（2008 年 7 月 5 日 b3 面）
・情報爆発に立ち向かう，Newton 2009 年 8 月号（2009 年 6 月 26 日発売）
ホームページ情報
・検索エンジン基盤 TSUBAKI URL
http://tsubaki.ixnlp.nii.ac.jp/