様式 C-19

科学研究費補助金研究成果報告書

平成２１年　５月２８日現在

研究種目：若手研究（Ｂ）
研究期間：2006～2008
課題番号：１８７００１５２
研究課題名（和文）　共起情報格納型のトライ構造を利用した単語概念知識獲得と意味理解への応用
研究課題名（英文）　Knowledge acquisition of word concepts using a trie structure of storing co-occurrence information, and application to semantic understanding

研究代表者
森田　和宏（MORITA KAZUHIRO）
徳島大学・大学院ソシオテクノサイエンス研究部・講師
研究者番号：２０３２５２５２

研究成果の概要：シソーラス（単語を意味で分類し体系化した語彙集）は，コンピュータに自然言語（日本語など人間が日常的に用いる言語）を理解させる上で重要であるが，人手による作成の労力や，作成者による分類の視点の揺れが問題であった．本研究は，シソーラスに代表される概念知識をコンピュータにより機械的に獲得し，体系を構築する技術確立を目的に実施され，成果として，概念知識をコンピュータが扱いやすい辞書形式で，少容量で保存する技術と，自動獲得，体系化を行う技術を考案した．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００６年度	600,000	0	600,000
２００７年度	500,000	0	500,000
２００８年度	400,000	120,000	520,000
年度			
年度			
総　計	1,500,000	120,000	1,620,000

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：共起情報，トライ構造，概念知識

１．研究開始当初の背景

単語を意味（語義）で分類し，体系化したシソーラス（統制語辞書）は，人間にとって理解しやすく，また機械へのインプリメントも容易であり，自然言語処理において多義語の曖昧性解消問題等に用いられる．また，大量の文書の意味内容による分類の手がかりとなるものとしても重要であり，利用価値は高い．シソーラスは従来，人手により作成されたものを用いることが多かったが，人手による作成は労力やコストがかかること，作成者による概念化の視点が揺れるなどの問題点がある．

また，これらの体系化された知識は，通常辞書に格納され，自然言語解析時に参照されるが，単語には新語や流行語などの未知語が存在するため，これらに対応するためには常に辞書を整備しておく必要がある．知識構築の実用システムが確立されれば，辞書の整備コストも無視できないものになるため，この問題の解決を念頭に置いた知識構築システムを考案する必要がある．

２．研究の目的

本研究の目的は，共起情報からの意味的類似性に着目し，係り語に対する概念知識の獲得，体系化技術の確立と，獲得知識の辞書構築の効率化，また構築した辞書を用いた意味理解への応用であり，次の点を明らかにする．

(1)トライ構造を用いた，共起情報とともに単語の概念情報を格納できる新しい辞書構造を提案する．この手法により，数十万語彙から無限に生成される共起情報を，数十 MB 程度の辞書に格納でき，係り語の持つ受け語群や，同一概念を持つ単語が相互検索できるようになる．また，探索時間計算量もO(k)(k：語彙長)となる．

(2)(1)の手法を用いて，共起情報から係り語を意味的に分類するアルゴリズムの提案，作成を行う．具体的には，係り語の持つ受け語群の頻度などから，特徴ベクトルを生成し，他の係り語との類似度から分類する．また，この分類は一語単位で行うため，頻繁な更新にも十分に対応できる．

(3)(2)で分類された，意味的に近い係り語群を，1 つのグループとしてまとめ上げ，グループ間の上位下位関係の生成と概念名を付与する手法を提案する．

(4)以上の技術を統合した概念知識構築エンジンを開発し，実用化されたシステムを作成，評価する．

３．研究の方法

(1)共起情報とともに単語の概念情報を格納できる辞書構造の提案と開発

過去の研究成果をもとに，トライ構造を用いて共起情報・概念情報を効率よく格納できる新しい辞書構造について研究する．また，辞書エンジンのプロトタイプを開発する．

(2)共起情報を格納し，係り語を意味的に分類するアルゴリズムの提案と開発

係り語の持つ受け語群の頻度などから，係り語を意味的に分類する手法を研究する．また，開発された辞書エンジンのプロトタイプと連動して，分類された係り語群をグループとして格納する辞書構造を開発する．

(3)概念知識構築アルゴリズムの考案とシステムの開発

係り語を意味的に分類するアルゴリズムによって分類されたグループ間の上位下位関係を生成する手法を研究する．また，グループに対してラベル(概念名等，そのグループに属する語の特徴を表す表現)を付与する手法を研究し，これらの情報を効率よく格納する辞書構造の開発と，各手法を結合した概念知識構築エンジンを開発する．

(4)エンジンの評価

開発した概念知識構築エンジンを使用して，新聞・Web などから収集した文書から共起情報を抽出し，概念知識を構築させる．抽出する共起情報は多岐かつ大量に必要だったが，Web 等からの文書収集では時間的効率に欠けるため，Web 日本語 N グラムデータを利用することとした．

構築した概念知識を，従来のクラスタリング手法や，シソーラスとの比較実験により，有効性を評価する．

４．研究成果

(1)トライ構造を用いた新しい辞書構造に関する成果

過去の研究成果であるリンクトライ構造は，2 項間の関係を格納できる構造であったので，これを 3 項以上の関係についても格納可能な辞書構造を考案した．これにより，共起関係と，それに関連する概念情報など，複数の項の関連情報を効率よく格納することに成功した．また応用として，通常の単語を分割して格納することにより記憶量を圧縮する手法を考案し，学会発表を行っている．

考案した辞書構造はトライ構造を基にしているため，その実装には通常ダブル配列を用いる．そこで，ダブル配列の実装方法を改善することにより，記憶量を削減する手法の考案と，ダブル配列の弱点であった動的更新速度を改善する手法を考案し，論文発表を行っている．

(2)概念知識の獲得に関する成果

①(1)の辞書構造を用いて，共起情報から係り語を意味的に分類し，分類したグループ間の上位下位関係を決定する手法を考案した．具体的には，係り語の持つ受け語群の頻度などから，特徴ベクトルを生成し，他の係り語との類似度を判定することによって分類を行う．この特徴ベクトルは，分類したグループごとにも生成され，グループ間の距離尺度を元に上位下位関係を決定する．この処理は，係り語を意味的に分類する処理と同時に行う．つまり，類似度判定によって係り語を分類する際に，どのグループにも属さなかった係り語は独自のグループを形成し，グループ間の距離尺度による上位下位の決定が行われる．

また，この分類処理は一単語単位で行うことができるため，頻繁な更新にも十分に対応でき，分類結果も同一の辞書構造内に格納できる．

この成果により得られた技術は，分野連想語辞書への新語の分類や，共起情報を利用した感性表現知識の獲得，コーパスからの話題語抽出の研究に応用し，その研究成果は学会・論文で発表を行っている．

②①の概念知識構築アルゴリズムを実現するシステムの開発を行った．このシステムは，図1のような共起情報とその頻度の集合を入力として与えると，分類結果を辞書に格納する．入力に与える共起情報に制限はなく，例えば係り語と受け語の関係を与えれば単語の概念分類に，文書とキーワードの関係を与えれば文書分類に利用できると考えられ，応用性の高いシステムとなっている．また，分類結果を視覚化して出力することも可能であり，図2は分類結果の出力例である．

ハードロック	の代名詞	85
ほうれん草	の焼き	62
拒食症	に苦しむ	33
スキャナ	が読み取っ	34
ハードロック	が聴ける	23
自動車	を使う	2126
ハクチョウ	の飛来	615
ほうれん草	をゆでる	511
サッカー	に携われ	28
	：	

図1　入力データの例

(3)国内外の位置付け

コーパスや共起情報から知識を獲得する国内外の研究は，様々なものが行われているが，特定の単語や分野に絞って知識獲得を行う場合が多い．またコーパスからの知識獲得であっても，一括して静的に処理が行われるので，頻繁に更新する場合は処理コストが無視できないと考えられる．

また，獲得された単語知識は辞書として使用されるが，単語には新語や流行語などの未知語が常に存在するため，獲得された知識が動的に，かつ頻繁に更新できる必要があると考えられる．

これに対して本研究成果は，これらの問題を解決することが可能であり，また同時に自然言語辞書で最も利用されるトライ構造を用いた辞書に構築された知識が蓄えられるので，構築された知識の応用性が確保されている．

また，分類はクラスタリングの観点から行われるが，一般的なクラスタリング手法は人手で作成された体系に対して分類を行うのに対し，本研究成果では，知識体系が全くない状態から構築するので，人の主観が反映されないという特徴がある．

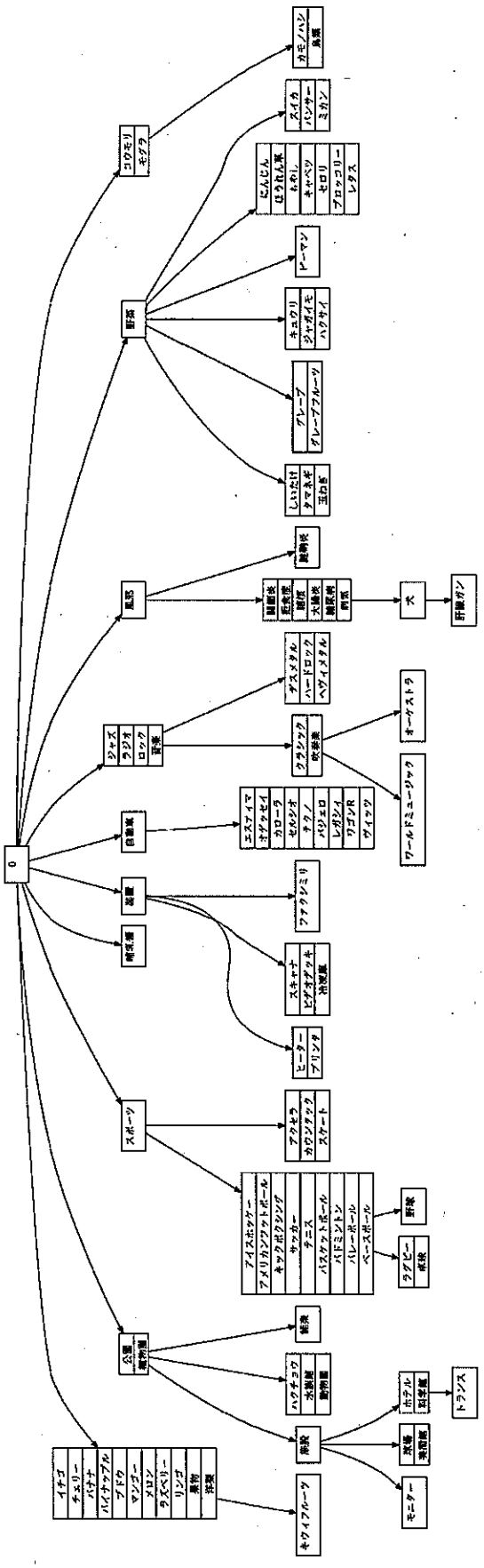

図2　分類結果の出力例

(4)今後の展望

直近においては，分類されたグループに対してラベルを付与する手法の研究を行ってきたが，現状では結果が芳しくないため，今後も継続する予定である．また，概念知識構築アルゴリズムについては，成果を纏め，学会・論文等での発表を行う予定である．

さらに，本研究により得られた成果を利用して，意味理解への応用研究へ発展させる．具体的には自然言語の持つ曖昧な表現(比喩など)を理解させるため，表現の関連知識獲得に本成果を利用する予定である．

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５件)

① Tomoko Yoshinari, Atlam EL-Sayed, Kazuhiro Morita, Kumiko Kiyoi and Jun-ichi Aoe, Automatic acquisition for sensibility knowledge using co-occurrence relation, International Journal of Computer Applications in Technology, Vol.33, No.2/3, pp.218-225, 2008, 査読有.

② Uddin Md. Sharif, Elmarhomy Ghada, El-Sayed Atlam, Masao Fuketa, Kazuhiro Morita and Jun-ichi Aoe, Improvement of building field association term dictionary using passage retrieval, Information Processing & Management, Vol.43, No.6, pp.1793-1807, 2007, 査読有.

③ Susumu Yata, Masaki Oono, Kazuhiro Morita, Masao Fuketa and Jun-ichi Aoe, An Efficient Deletion Method for an MP Double-Array, Software Practice & Experience, Vol.37, No.5, pp. 523-534, 2007, 査読有.

④ Susumu Yata, Masaki Oono, Kazuhiro Morita, Masao Fuketa, Toru Sumitomo and Jun-ichi Aoe, A Compact Static Double-Array Keeping Character Codes, Information Processing & Management, Vol.43, No.1, pp.237-247, 2007, 査読有.

⑤ El-Sayed Atlam, Ghada Elmarhomy, Kazuhiro Morita, Masao Fuketa and Jun-ichi Aoe, Automatic Building of New Field Association Word Candidates Using Search Engine, Information Processing & Management, Vol.42, No.4, pp.951-962, 2006, 査読有.

〔学会発表〕(計４件)

① Kazuhiro Morita, Yuya Iwabu, Atlam EL-Sayed, Masao Fuketa and Jun-ichi Aoe, A Method of Extracting Word Tendencies to Understand Popular Subjects, 5th International Conference on Innovations in Information Technology (Innovations'08), 2008/12/18, Al Ain, UAE.

② Tomoko Yoshinari, Atlam EL-Sayed, Kazuhiro Morita, Kumiko Kiyoi and Jun-ichi Aoe, Automatically Combination with Expressions using Collocation Relation, International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE 2007), 2007/9/1, Beijing, China.

③ Kazuhiro Morita, Atlam EL-Sayed, Elmarhomy Ghada, Masao Fuketa and Jun-ichi Aoe, A New Approach for Improving Field Association Term Dictionary Using Passage Retrieval, 10th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES 2006), 2006/10/11, Bournemouth, UK.

④ Masaki Oono, Masao Fuketa, Kazuhiro Morita, Yutaka Inada, Yo Murakami and Jun-ichi Aoe, A Compression Method Using Link-Trie Structure for Natural Language Dictionaries, International Conference on Computing and Informatics (ICOCI 2006), 2006/6/7, Kuala Lumpur, Malaysia.

6．研究組織

(1)研究代表者

森田　和宏 (MORITA KAZUHIRO)

徳島大学・大学院ソシオテクノサイエンス研究部・講師

研究者番号：２０３２５２５２

(2)研究分担者

(　　)

研究者番号：

(3)連携研究者

(　　)

研究者番号：