様式 C-19

科学研究費補助金研究成果報告書

平成２３年　５月　２７日現在

機関番号：１３５０１
研究種目：基盤研究（C）
研究期間：2008～2010
課題番号：２０５００１２７
研究課題名（和文）　コーパスを利用した科学技術分野でのシソーラス構築と関連文書抽出への利用
研究課題名（英文）　Thesaurus construction using corpus of science and technology and its application to document retrieval
研究代表者
鈴木良弥（SUZUKI YOSHIMI）
山梨大学・大学院医学工学総合研究部・准教授
研究者番号：２０２０６５５１

研究成果の概要（和文）：特許公報文コーパスを利用した科学技術分野でのシソーラス構築と，構築したシソーラスを利用して関連文書抽出を行った．具体的には日本語特許公報文に焦点をあて，１．関連単語対の抽出，２．関連単語対の細分類，３．シソーラスの構築，４．シソーラスを利用した関連文書抽出，を行った．本研究で構築したシソーラスの情報を特許公報文の関連文書抽出に利用し，抽出精度を向上させることができた．

研究成果の概要（英文）：We studied thesaurus expansion using patent documents and document retrieval using the expanded thesaurus. In particular, using patent documents we investigated the following 4 subtasks: (1) extraction of similar word pairs, (2) classification of the extracted similar word pairs, (3) thesaurus expansion using the extracted similar word pairs, and (4) document retrieval using the expanded thesaurus.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	１，５００，０００	４５０，０００	１，９５０，０００
２００９年度	１，２００，０００	３６０，０００	１，５６０，０００
２０１０年度	７００，０００	２１０，０００	９１０，０００
年度			
年度			
総　計	３，４００，０００	１，０２０，０００	４，４２０，０００

研究分野：総合領域
科研費の分科・細目：情報学・知識情報学
キーワード：自然言語処理，情報検索

１．研究開始当初の背景

文書の電子化とインターネットの普及により，大量のテキストデータを個人が簡単に閲覧できるようになった．このような状況において利用者が膨大なテキストデータから本当に必要なデータだけを探し出すことは重要であると同時に非常に困難な作業となっている．特に学術分野では，自分の研究に関連する研究を大量の文献から探し出したり，特許出願に際して関連特許がすでに特許申請されているかどうかを検索することは非常に重要である．

現在，複数のキーワードを利用して特許や論文を検索するサービス（特許庁の特許検索サービス，Google Scholar など）は存在するが，検索キーワードと検索対象との文字列照合だけで検索を行っているため，関連文書内で，検索キーワードに使われた用語は含まず検索キーワードの類義語や上位概念語が使われていた場合，目的の関連文書を抽出することが出来ない．例えば，検索キーワードに「ワイン」を入力した場合，検索対象に「アルコール飲料」に関する文書があってもその文書に「ワイン」という単語が出現しなければその文書を関連文書として抽出することは出来ない．そのため，特許や学術論文の分類などに利用出来る辞書の構築が課題となっていた．

２．研究の目的

本研究は特許公報文コーパスを利用した科学技術分野でのシソーラス構築と，構築したシソーラスを利用して関連文書抽出を行うことを目的とする．具体的には日本語特許公報文に焦点をあて，（１）関連単語対の抽出，（２）関連単語対の細分類，（３）シソーラスの構築，（４）シソーラスを利用した関連文書抽出する．本研究の特色・意義は，（１）システムでも利用者でも利用可能なシソーラスの構築手法の提案，（２）他分野でも利用可能なシソーラス構築手法の提案，（３）構築したシソーラスの情報を利用した関連文書抽出の３点に集約できる．

３．研究の方法

本研究では特許申請文を知識源とし，抽出した類語を同義語，上位語-下位語，姉妹語，反意語などに自動的に分類し，シソーラスを構築する．構築したシソーラスを利用し，関連文書検索を行い，構築したシソーラスの効果を確かめる．以下に具体的な研究方法を示す．

(1)関連単語対の抽出

関連単語対の抽出のために Lin の手法を日本語に適用して単語間の類似度を計算している．Lin の手法は相互情報量の計算をもとにしており，構文情報と近隣情報を活用しているが，構文情報のより効果的な利用と文章の中の役割などの情報を利用することで，より正確な単語間類似度を計算している．

(2)関連単語対の細分類

関連単語抽出の実験時に抽出した関連単語対を詳しく調べた結果，３種類（上位/下位語，類義語，姉妹関係語）に細分類が可能であることがわかった．

関連単語の細分類を行うために，

・同一文書中では同じ意味を表す場合は同一単語を使用する場合が多い．

・固有表現は上位下位関係の終端ノードである．

・すでにあるシソーラス（JST 科学技術シソーラス，EDR の種々の辞書，WordNet（英語のシソーラス））の利用

・関連単語対の類似度の利用

・抽出した関連単語対を利用した関連単語グループの作成

などを利用している．

(3)シソーラスの構築

関連単語対の細分類で得られた情報（類義語，上位／下位概念，姉妹語）などを利用して見出し語毎に部分木を作製する．見出し語間で不整合が起きる場合があるが，単語対の類似度や多数決処理を行って，不整合を解消する．「２．関連単語対の細分類」で得られた細分類の情報をもとに類語辞典（シソーラス）を作成した．

(4)シソーラスを利用した関連文書検索

シソーラスを用いて検索キーワードを拡張する場合，拡張した単語の重みを決めなければならない．拡張された単語がキーワードの言い換えの場合には重みを大きくし，上位／下位語の場合には重みを小さくする．その重みと検索キーワードとの類似度を利用して，最終的な重みを決定しなければならない．まず，コンテキストの違い毎にどの分類（類義語，上位／下位語など）の重みを大きくすればいいかを調べ，その後，機械学習によって最適な重みをつけ文書検索を行う． 関連単語の細分類が実際のアプリケーションの精度を向上させるかどうかを確認するために文書分類実験を行った．

４．研究成果

本研究は４つの柱からなる．それぞれについて研究成果を説明する．

(1)関連単語対の抽出

大量のコーパス（特許公報文）を用いて単語毎の類似度を計算し，関連単語対の抽出を行った．Lin の手法を基にした手法を提案し，関連単語対の抽出実験を行い，結果を論文４で報告した．

(2)関連単語対の細分類
新しい単語をシソーラスに追加するためには類似度を計算するだけではなく，各単語との関係（上位/下位語，類義語，姉妹語）などを調べる必要がある．類似度や手がかり語などの情報を利用して関連単語対の細分類を行う手法を提案し，実験結果と共に論文⑪で報告した．
(3)シソーラスの構築
EDR の専門用語辞書をコアシソーラスとし，特許公報文に現れる専門用語をシソーラスに統合する方法を提案し，実験結果と共に論文（④，⑧）で報告した．
(4)シソーラスを利用した関連文書検索
構築したシソーラスを利用した特許公報文の分類システムを作成し，シソーラス利用により分類性能が向上したことを確認した．国際会議に投稿中である．本研究で構築したシソーラスの情報を特許公報文の関連文書抽出に利用し，抽出精度を向上させることができたことは学術的意義があると考えられる．本研究では，なるべく汎用的な手法で汎用的な用途に利用出来るシソーラスの作成を目指したが，研究を通して，より効果的にシソーラスを利用するためには利用する用途に応じて作成した汎用的なシソーラスの構造を修正する必要があるという知見を得た．

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　14 件）

① F. Fukumoto and Y. Suzuki and K. Yamashita, “Polysemous Verb Classification using Subcategorization Acquisition and Graph-based Clustering”, Human Language Technologies, Lecture Notes in Computer Science, 査読有, 2011, (To Appear)
② F. Fukumoto and Y. Suzuki, “Identifying Domain-Specific Senses and its Application to Text Classification”, Proc. of 2nd International Conference on Knowledge Engineering and Ontology Development, 査読有, 2010, pp. 59-64
③ F. Fukumoto, A. Sakai and Y. Suzuki, “Eliminating Redundancy by Spectral Relaxation for Multi-Documents Summarization”, Proc. of 2010 Workshop on Graph-based Methods for Natural Language Processing, 査読有, 2010, pp. 98-102,
④ Y. Suzuki and F. Fukumoto, “Integrating Technical Terms into Thesaurus using Extracted Related Words”, Proceedings of the International Symposium on Matching and Meaning, 査読有, 2010, pp.34-38.
⑤ F. fukumoto and K. Yamashita and Y. Suzuki, “Classifying Polysemies using a Graph-based Clustering”, Proc. of the 4th Language and Technology Conference, 査読有, 2009, pp. 210-214.
⑥ F. Fukumoto and Y. Suzuki, “Effect of Overt Pronoun Resolution in Topic Tracking”, Proc, of the 4th Language and Technology Conference, 査読有, 2009, pp. 350-354.
⑦ F. Fukumoto and Y. Suzuki, “Using Graph-based Indexing to Identify Subject-Shift in Topic Tracking”, Human Language Technologies, Lecture Notes in Computer Science, Vol. 5603, 査読有, 2009, pp. 392-404.
⑧ Y. Suzuki and F. Fukumoto, “Detecting Word Senses of Technical Terms in Patent Documents using Hierarchical Semantic Features”, Proceedings of 4th Language & Technology Conference, 査読有, 2009, pp.195-199.
⑨ Y. Suzuki and F. Fukumoto, “Classifying Japanese Polysemous Verbs based on Fuzzy C-means Clustering”, Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4), 査読有, 2009, pp.32-40.
⑩ F. Fukumoto and N. A. H. B. Zahri and Y. Suzuki, “Example-assignment to WordNet Thesaurus based on Distributional Similarity of Words”, Proc. of Workshop on Matching and Meaning, 査読有, 2009, pp. 1-5.
⑪ Y. Suzuki and F. Fukumoto, “Detecting unknown word senses using concept dictionary”, Proceedings of the Symposium on Matching and Meaning, 査読有, 2009, pp.49-51.
⑫ F. Fukumoto and Y. Suzuki, “The Effect of Combining Similarity Measures for Onomatopieia Clustering”, Proc. of Empirical Method in Asian Natural Language Processing, 査読有, 2008, pp. 37-51.
⑬ F. Fukumoto and Y. Suzuki, “Retrieving Bilingual Verb-Noun Collocations by Integrating Cross-Language Category Hierarchies”, Proc. of the 22nd International Conference on Computational Linguistics, 査読有, 2008, pp. 233-240.
⑭ F. Fukumoto and Y. Suzuki, “Integrating Cross-Language Hierarchies and its Application to Retrieving Relevant Documents”, ACM Trans. of Asian Language

Information Processing, 7(3), 査読有, 2008, pp. 1-22.

〔学会発表〕(計　１件)
① 森田一匡, 鈴木良弥　“Web 上のオノマトペの用例を共起単語で絞り込む用例抽出法”, 言語処理学会 第１６回年次大会講演論文集, 2010 年 3 月 8 日, pp.924-927, 東京大学.

６．研究組織
(1)研究代表者
鈴木　良弥（SUZUKI YOSHIMI）
山梨大学・大学院医学工学総合研究部・准教授
研究者番号：２０２０６５５１

(2)研究分担者
なし

(3)連携研究者
なし