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：14501
研究種目：若手研究（B）
研究期間：2009～2010
課題番号：21700169
研究課題名（和文）　言語資源からの知識の抽出・汎化と新仮説生成

研究課題名（英文）　Hypothesis Discovery via Knowledge Extraction and Generalization

研究代表者
関　和広（Seki Kazuhiro）
神戸大学・自然科学系先端融合研究環・助教
研究者番号：30444566

研究成果の概要（和文）：近年のコンピュータ関連技術の進歩により，我々は極めて容易に膨大な量のテキストを入手できるようになった．この研究では，テキストデータを基に従来人手で行ってきた遺伝子機能のアノテーションと仮説生成を計算機で高精度に実現するための研究を行った．前者については，カーネル法と呼ばれるパターン認識の手法を用いて，効率的かつ高精度なアノテーションを実現した．また後者については，イベント類似度という概念を定義することで，より妥当な仮説を発見する枠組を考案，評価した．

研究成果の概要（英文）：Owing to the recent advances in computer technologies, a large amount of texts have become available. This research attempted to automate highly intellectual tasks analyzing such texts, which usually require labor-intensive work and domain knowledge. Specifically, we investigated gene function annotation and hypothesis discovery. For the former, we proposed an approach based on semantic kernels and achieved both efficient and effective gene annotation as compared with existing works. For the latter, we defined event similarity to spot valid hypotheses. Evaluative experiments revealed that our proposed approach was more stable and effective than previous frequency-based approaches.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	2,100,000	630,000	2,730,000
2010 年度	1,300,000	390,000	1,690,000
年度			
年度			
年度			
総　計	3,400,000	1,020,000	4,420,000

研究分野：知能情報処理
科研費の分科・細目：情報学・知能情報学
キーワード：仮説生成，遺伝子機能アノテーション，イベント類似度

１．研究開始当初の背景

近年のコンピュータ関連技術の進歩により，我々は極めて容易に膨大な量のテキストを入手することが可能になった．しかしながら，自然言語で記述されたテキストに内包される情報は通常，人間がテキストを読むことによって初めて理解されるため，テキストが大量・豊富であるほど，そこに存在する情報を即座にかつ効果的に利用することが難しいというジレンマが生じる．爆発的に増大するテキストを我々が効果的に利用するために

は，情報検索や情報抽出，言語理解，要約といった知的な情報管理・処理技術の発展が重要である．

２．研究の目的

本提案研究では，これまで申請者が得た研究成果を基に，特に（a）テキストからの情報抽出に基づく遺伝子オントロジー（GO）自動アノテーション，および（b）仮説生成に焦点を当て，テキストに埋没した情報の利活用を図る．

３．研究の方法

GO アノテーションについては，カーネルを用いた機械学習の手法を採用する．生物医学文献の数は膨大であり，また遺伝子の種類も数多くある．そのため，各文献に記された遺伝子に対して遺伝子機能を手作業で付与（ラベル付け）するには，多大なコストが必要となる．その結果，機械学習を行うために必要な訓練データが，十分に集まらないことが多い．訓練データを必要としない手法として，従来，文字列一致が利用されてきたものの，この手法では，表記の揺れや未知語に対処できないという問題がある．そこで本研究では，付加的な情報を容易かつ効果的に取り込むことができ，計算量的にも優れた性質を持つカーネルを用いることで，これらの問題に対処する．また，マルチラベル分類による遺伝子機能付与を行う際に，各クラスごとに正則化を行うことで，ラベル付きデータの数が特定のクラスに偏っているデータ（不均衡データ）の問題にも対処する．より具体的には，以下のステップで GO アノテーションを行う．

情報抽出：文献中から，曖昧文字列一致を用いて遺伝子について言及した部分，および文献のタイトルとアブストラクトを抽出する．また，訓練事例に付与されている GO タームの定義文（たとえは，GO ターム「paclitaxel metabolic process (GO:0042616)」の定義文は，「The chemical reactions and pathways involving paclitaxel, an alkaloid compound used as an anticancer treatment.」）を擬似 的な正例として（学習時のみ）利用する．これによって，予測の際，もしテストデータの文献に GO ターム定義文の単語が存在すれば，それらの単語も GO タームの予測に考慮される．
前処理と単語集合表現：続いて，抽出したテキスト情報を単語集合表現へと変換する．形式的には，各事例を$x_i=(x_{i1}, ..., x_{iD}) \in R^D$という形で表現する．ここで，$x_{ij}$は，文書$i$にある$j$番目の語彙の出現回数であり，$D$は語彙数（単語の数）を表す．
学習：遺伝子機能のアノテーションはマルチラベル分類の問題として考えることができる．つまり，各文献を事例として考えた場合，対応する複数のラベルの組み合わせを予測することに相当する（対応するラベルがない場合もある）．提案手法では，マルチラベル分類を行うための方法として，各クラス（GO ターム）1 つにつき 1 つの二値分類器を構築するone-vs-allの手法を用いる．この手法は，モデルが簡潔であり，結果の解釈がし易く（どの特徴がGOタームの分類に役立ったかを確認できる），各分類器は独立であるために並列化が容易である．数式的には，文献x_iに対して，$y_i = (y_{i1},...,y_{iM}) \in \{-1, +1\}^M$を割り当てることに対応する．ここで，$M$はGO タームの数であり，$y_{ic}$=+1 は$c$番目のGOタームが文献$x_i$に付与されること，$y_{ic}$=−1 は付与されないことを表す．
予測と後処理：学習された分類器によって，GO タームの予測を行う．ただし，我々が使用する one-vs-all の手法では，二値分類器を逐次的に適用していくので，GO タームの付与中に矛盾する GO タームの組み合わせが生じる可能性がある．GO の構造である有向非巡回グラフの利点を生かした後処理を行えば，このような矛盾する GO タームの組み合わせを除去できるものと考えられる．そこで，後処理として，GO タームが祖先と子孫の関係にあるとき，それらの内で尤もらしい方のみを付与する．

仮説生成については，意味的に類似したイベントから生成された仮説はより妥当であると仮定し，仮説を構成するイベント間の意味的な類似度を用いて仮説の妥当性を定義する．そして，この妥当性に基づいて仮説を順位づけることで，真に重要な仮説を効率的に発見する．なお，イベントとは，生物医学要素の 2 項関係のことを指すものとする．

イベント間の類似度には，人手で体系化された MeSH と呼ばれるシソーラスを使用する．このイベントを特徴付けるために使用する MeSH 語は，MEDLINE の各レコート（論文）を索引付けするため，人手により通常十数個ほど付与されている．2009 年の時点では，25,186 個の定義語が存在し，意味的な構造を持つ 11 の階層に整理されている．この MeSH の階層構造の中では，上位にある語ほど一般的な意味を持ち，下位に行くほど厳密な意味を持つ語となる．なお，MeSH は文献を特徴付けるための索引語であり，必ずしも文献から抽出されたイベントを特徴付けるものではない．しかしながら本研究では，文献の内容を最も簡潔に表現したタイトルだけからイベントを抽出することで，文献に付与された MeSH 語をイベントの特徴と見なす．そして，イベントの特徴として MeSH 語間の類似

度を定義し，さらにこれをイベント間の類似度へと拡張する．

概念間の類似度を測る手法として，本研究では，Seco らが提案した手法を用いる．この手法の特徴は，シソーラスの構造を用いることで概念間の類似度を測ることにある．よって，頻度に依存することなく，概念間の類似度を計算することができる．

以下に示す類似度の定義は，イベントに対応する MeSH 語間の類似度をすべて求め，その平均をイベント間の類似度としたものである．このイベント間の類似度を当該イベントによって導出される仮説の妥当性（reasonability）と見なす．

$$\mathbf{R_{avg}}(e_i,e_j)=\frac{1}{|\mathbf{M}_i||\mathbf{M}_j|}\sum_{m_k\in\mathbf{M}_i}\sum_{m_l\in\mathbf{M}_j}\mathbf{sim}(m_k,m_l)$$

ここで，M_iとM_jはそれぞれイベントe_i，e_jに対応するMeSH語の集合を表している．MeSH語の集合Mは，イベントの抽出元である文献に付与されたMeSH語を重複しないように集めたものである．この妥当性 R_{avg}の定義の欠点は，似ていない概念同士の類似度か影響を持ち易いことである．次の妥当性 R_{max}はこの点を考慮し，最も類似する概念たけに着目する．イベントe_iに対応する各概念に対して，もう一方のイベントe_jにある概念と最も類似する概念を選び，互いの類似度の平均をもとめる．そして，イベントに対して対称になるように，e_iとe_jを入れ替えて同様の計算を行い，最終的なe_iとe_jの類似度をもとめる．

$$\mathbf{R_{max}}(e_i,e_j)=\frac{1}{|\mathbf{M}_i|}\sum_{m_k\in\mathbf{M}_i}\max_{m_l\in\mathbf{M}_j}\mathbf{sim}(m_k,m_l)+\frac{1}{|\mathbf{M}_j|}\sum_{m_l\in\mathbf{M}_j}\max_{m_k\in\mathbf{M}_i}\mathbf{sim}(m_k,m_l)$$

４．研究成果

GO アノテーションについては，TREC ゲノムトラックのデータによる実験を行い，提案手法によって，文字列一致（Stoica & Hearst）および異種間の情報（Seki et al.）を用いた従来手法よりも高い適合率と F1 スコアが得られることが示された（下表）．

手法	適合率	再現率	F_1 スコア
PROPOSED	0.38	0.24	0.29
PROPOSED (+O)	0.42	0.23	0.30
STOICA & HEARST	0.19	0.46	0.27
SEKI ET AL.	0.26	0.27	0.26

また，GO タームのアノテーションに起こりやすいラベル付きデータが不足する問題についても，潜在トピックをカーネルに取り込むことで 効果的に対処することが出来た（下表のK_{plsa} (+U)とK_{plsa} (+U+T)）．

カーネル	適合率	再現率	F_1 スコア
K_{linear}	0.36	0.20	0.26
K_{poly} $(d=2)$	0.35	0.19	0.25
K_{plsa}	0.38	0.20	0.26
K_{plsa} (+U)	0.39	0.22	0.28
K_{plsa} (+U + T)	0.38	0.24	0.29

加えて，GO ターム毎（クラス毎）に正則化を行うことで，貴重な訓練データを活用しつつ，不均衡データに対処する手法を提案した（下表の提案ヒューリスティクス）．

不均衡データの扱い	適合率	再現率	F_1 スコア
提案ヒューリスティック	0.36	0.20	0.26
ダウンサンプリング	0.26	0.18	0.21
不均衡データ考慮せず	0.14	0.09	0.11

仮説生成については，Swanson が発見したレイノー病と魚油の関係，および「偏頭痛とマグネシウムの関係を既知の仮説として，評価実験を行なった．そして，提案手法である意味的類似度に基づく妥当性の指標R_{max}とR_{avg}が，既存手法の頻度に基づく指標R_{tfidf}，R_{freq} に対して，どの程度妥当・非妥当な仮説を適切に順位付けできるのかを検証した．その結果，ほとんどの場合において，仮説を説明するイベントの頻度に関わらず，R_{max} は安定かつ適切な順位付けを行うことができた．一方，従来手法である頻度に基づく妥当性は，概念またはイベントの頻度に大きく左右されるため，不適切な結果を示す場合があった．

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 9 件）

①. Taiki Miyanishi, Kazuhiro Seki, and Kuniaki Uehara. Hypothesis Ranking Based on Semantic Event Similarities. IPSJ Transactions on Bioinformatics. (To appear)

②. Mathieu Blondel, Kazuhiro Seki, and Kuniaki Uehara. Tackling Class Imbalance and Data Scarcity in Literature-Based Gene Function Annotation.In Proceedings of the 31th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2011), July 2011.

③. Kazuhiro Seki and Kuniaki Uehara. Opinionated Document Retrieval Using Subjective Triggers. Journal of the American Society for Information Science and Technology (JASIST), Vol. 62, No. 5, pp. 861-876, 2011.
④. Mathieu Blondel, Kazuhiro Seki, and Kuniaki Uehara. Unsupervised Learning of Stroke Tagger for Online Kanji Handwriting Recognition. In Proceedings of the 20th International Conference on Pattern Recognition (ICPR 2010), pp. 1973-1976, August 2010.
⑤. Kazuhiro Seki, Huawei Qin, and Kuniaki Uehara. Impact and Prospect of Social Bookmarks for Bibliographic Information Retrieval. In Proceedings of the 10th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2010), pp. 357-360, June 2010.
⑥. Taiki Miyanishi, Kazuhiro Seki, and Kuniaki Uehara. Hypothesis Generation and Ranking Based on Event Similarities. In Proceedings of the 25th Annual ACM Symposium On Applied Computing (SAC 2010), pp. 1552-1558, March 2010.
⑦. 吉川幹人, 佐藤翔平, 関和広, 上原邦昭. リンク構造とコンテンツを複合的に用いた極少訓練事例によるスプログ検出. 情報処理学会論文誌：データベース, Vol. 3, No. 1, pp. 29-37, March 2010.
⑧. 関和広, 上原邦昭. 主観的トリガー言語モデルによる意見情報検索. 情報処理学会論文誌：数理モデル化と応用, Vol. 2, No. 3, pp. 27-38, December 2009.
⑨. Kazuhiro Seki, Yoshihiro Kino, and Kuniaki Uehara. Gene Functional Annotation with Dynamic Hierarchical Classification Guided by Orthologs. In Proceedings of the 12th International Conference on Discovery Science (DS 2009), pp. 425-432, October 2009.

〔学会発表〕（計 7 件）
①. 中菅 章浩, 関 和広, 上原邦昭. 救急医療トリアージノートを用いた症候群サーベイランス. 言語処理学会第 17 回年次大会発表論文集. 2011 年 3 月.
②. 宮西 大樹, 関 和広, 上原 邦昭. ネットワーク構造解析に基づく有望ノードの予測. 情報処理学会研究報告 2011-MPS-82. 2011 年 3 月.
③. Mathieu Blondel, 関 和広, 上原邦昭. 文献情報を用いたカーネル法による遺伝子機能アノテーション. 情報処理学会研究報告 2011-MPS-82. 2011 年 3 月.
④. 秦 華偉, 関 和広, 上原邦昭. 生物医学文献検索におけるソーシャルタグと統制語彙との比較. 電子情報通信学会技術研究報告, 言語理解とコミュニケーション研究会. 2011 年 1 月.
⑤. 関和広, 上原邦昭. 実空間検索メタデータとしてのソーシャルメディア. 電子情報通信学会技術研究報告, ライフインテリジェンスとオフィス情報システム研究会, pp. 1-6, 2010 年 5 月.
⑥. 萩村卓也, 関和広, 上原邦昭. 発想を支援するユーザエージェント. 電子情報通信学会技術研究報告, ライフインテリジェンスとオフィス情報システム研究会, pp. 99-103, 2010 年 5 月.
⑦. Shohei Sato, Mikito Yoshikawa, Kazuhiro Seki and Kuniaki Uehara. Splog Detection Exploiting Link Structure and Contents Based on Few Labeled Examples. Workshop on Search in Social Media (SSM 2009), co-located with ACM SIGIR 2009. July 2009.

６．研究組織
(1)研究代表者
関 和広（SEKI KAZUHIRO）
神戸大学・自然科学系先端融合研究環・助教
研究者番号：30444566

(2)研究分担者
なし
(3)連携研究者
なし