様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 24 年 6 月 7 日現在

機関番号：12601
研究種目：研究活動スタート支援
研究期間：2010～2011
課題番号：22800010
研究課題名（和文）ウェブにおけるエンティティ間の関係検索に関する研究

研究課題名（英文）Searching for Semantic Relations between Entities on the Web

研究代表者
ボレガラ　ダヌシカ（BOLLEGALA DANUSHKA）
東京大学・大学院情報理工学系研究科・講師
研究者番号：10581712

研究成果の概要（和文）：本研究では２つのエンティティ間の意味的関係が検索できる潜在関係検索システムを作成しました．尚，関係の対称性に着目をし，同一関係を複数の対称性を使って予測することによって関係類似性計測の精度を向上させた．関係検索の候補のランキングに関するアルゴリズムの設計も行った．研究の成果を国際会議(AAAI 2011)で発表し，英文論文誌(Elsevier IPM)にまとめた．

研究成果の概要（英文）：In this research, I developed a search system that can search for the semantic relations between named entities. In particular, I focused on the symmetries that exist in proportional analogies to accurately estimate the relational similarity between two pairs of entities. Moreover, algorithms were developed to accurately rank the search results using relational similarity. The outcome of this project was presented at AAAI 2011 and published in Elsevier Information Processing and Management Journal.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,260,000	378,000	1,638,000
2011 年度	1,160,000	348,000	1,508,000
年度			
年度			
年度			
総　計	2,420,000	726,000	3,146,000

研究分野：工学
科研費の分科・細目：情報学・知能情報学
キーワード：関係検索，ウェブ工学，ウェブマイニング，情報抽出

１．研究開始当初の背景

既存のウェブ検索エンジンは殆どキーワードベースの検索エンジンであり，入力されたキーワードを含むページを結果として出力する仕組みとなっている．そこで，エンティティではなく，エンティティ間の関係を検索可能な方法を開発したいと考えた．

２．研究の目的

本研究計画ではエンティティ間の関係が検索できる関係検索エンジンの構築とその評価を目的とする。具体的には一つのエンティ

ティペア(A,B)ともう一方のエンティティ C が与えられた時に、A と B の間の関係 R を Web の情報から抽出し、その関係 R で C と結ばれているエンティティ D を探し、ランク付けて、ユーザーに提示する一連のシステムを構築する.

３．研究の方法

関係検索エンジンを実現させるため研究の方法は次のステップからなる.

(1) Web をクロールし、エンティティに関する情報を収集する.
(2) クロールした Web テキストからエンティティを抽出し、その同定を行う.
(3) エンティティ間の関係を表す語彙パターンを抽出し、抽出された語彙パターンを抽象化する.
(4) 同一関係を表す語彙パターンをクラスタリングする.
(5) エンティティと語彙パターンに固有 ID を割り当て、その対応関係を高速に検索できるようにそれぞれに対してインデックスを構築する.
(6) エンティティペアと語彙パターンの共起情報が高速に検索できるようにインデックスを構築する.
(7) 検索インタフェースを実装する。さらに、検索結果をランク付けて表示する.
(8) 検索エンジンの評価を行い、対外発表する.

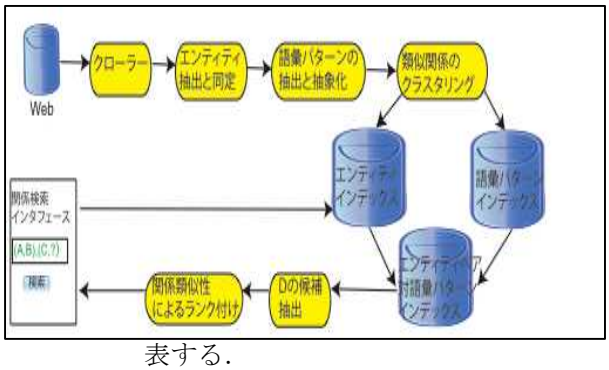

図１：研究方法の概念図

４．研究成果

本研究課題ではエンティティ間の関係抽出に関する研究を展開した．関係抽出を行う際に教師あり学習が教師なし学習に比べ、より良い精度を出しているが、ウェブのような多様な関係が膨大な数存在するドメインに関してはその全ての関係に関する学習データを人手で作成することは不可能であり、教師あり学習を使うには限界がある。そこで、本研究では対象とする関係に関するエンティティペアをいくつかのみ（シードという）を与えることで関係抽出を行う方法を考案した．提案手法により関係 A を抽出するために学習させた関係抽出器を別の関係 B を抽出するために適応できることが可能となった．提案手法ではまずエンティティペアに含まれる２つのエンティティ間の関係を語彙パターンを使って表現する．語彙パターンはその２つのエンティティが共起する文脈から部分シーケンスとして抽出する．次に、一つの関係についてのみ出現する語彙パターンと様々な関係について出現する語彙パターンをパターンのエントロピーを用いて分類する．パターンのエントロピーはあるパターンがどのようなエンティティペアと一緒に出現するかというパターンの出現頻度分布から計算できる．あるパターンが沢山のエンティティペアと一緒に出現すればその出現頻度分布が平らになり，エントロピーが高くなる．このことを利用し，語彙パターンを分類することができる．数多くのエンティティペアと共起する語彙パターンは様々な意味的関係をカバーできていると考えられるため，そのようなパターンをピボットとして使い，転移学習を行うことができる.

次に、同一エンティティペアについて抽出される異なる語彙パターンをエッジで繋げることによりパターンをノードとする２部グラフを構築する．２つの語彙パターン（グラフ上ではノード）はある同一のエンティティペアに対して共起していればそれらのノードをエッジで繋ぐことにする．尚，本提案手法ではエッジの重みはそのエッジが繋ぐ２つの語彙パターンを同時に満たす異なるエンティティペアの数にした．この２部グラフは関係に依存する語彙パターンと関係に依存しない語彙パターンの間の対応関係を示しているものだと考えられる．最後にこの２部グラフのグラフラプラシアンを計算することでどの関係に依存するパターンがどの関係に依存しないパターンに対応しているかを計算する．この対応関係が分かると例えばある関係 A を抽出するために学習させた学習器を別の関係 B を抽出するために使うことができる.

評価実験では２０種類の異なる関係について評価を行い、様々なベースライン手法と先行研究と比較した。本研究成果はウェブの分野の最高峰の国際会議である International World Wide Web や人工知能分野の最高峰の国際会議である International Joint Conference on Artificial Intelligence にて論文として採択されており国外でも高く評価された．尚，本研究成果は英文と和文論文誌に採択されている.

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　4 件）

① Danushka Bollegala, Tomokazu Goto, Nguyen Tuan Duc, and Mitsuru Ishizuka, Improving Relational Similarity Measurement using Symmetries in Proportional Word Analogies, Information Processing and Management, Elsevier, 査読有，採択済み，近日公開.

② Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka, Minimally Supervised Novel Relation Extraction using Latent Relational Mapping, IEEE Transactions on Knowledge and Data Engineering (TKDE), 査読有, Volume: PP, Issue 99 (Early Access), 2012.

③ Nguyen Tuan Duc, Danushka Bollegala, and Mitsuru Ishizuka, Cross-Language Latent Relational Search between Japanese and English Languages using a Web Corpus, ACM Transactions on Asian Language Information Processing (TALIP), 査読有, 採択済み, 近日公開.

④ 後藤友和, Nguyen Tuan Duc, Danushka Bollegala, and Mitsuru Ishizuka, 関係の対称性および予測語を用いた関係検索の性能向上法，人工知能学会論文誌，Vol. 26, no. 6, pp. 649-656, 2011.

〔学会発表〕（計 3 件）

① Nugyen Duc, Danushka Bollegala, and Mitsuru Ishizuka: Cross-Language Latent Relational Search: Mapping Knowledge across Languages, Proceedings of the 25th National Conference on Artificial Intelligence (AAAI 2011), 査読有, pp. 1237-1242, San Francisco, USA, 2011.

② Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka: Relation Adaptation: Learning to Extract Novel Relations with Minimum Supervision, Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011), 査読有，pp. 2205-2210, Barcelona, Spain, 2011.

③ Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka: From Actors, Politicians, to CEOs: Domain Adaptation of Relational Extractors using a Latent Relational Mapping, Proceedings of the 20th International World Wide Web Conference (WWW 2011) poster session, 査読有，pp. 13-14, Hyderabad, India, 2011.

〔図書〕（計　0 件）

〔産業財産権〕
○出願状況（計　0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
http://www.iba.t.u-tokyo.ac.jp/~danushka/
http://www.milresh.com/

６．研究組織
(1)研究代表者
ボレガラ　ダヌシカ（BOLLEGALA DANUSHKA）
東京大学・大学院情報理工学系研究科・講師
研究者番号：10581712

(2)研究分担者 なし
（　　）

研究者番号：

(3)連携研究者　なし

（　　）

研究者番号：