論文構造解析に基づく専門分野知識の抽出

Domain-specific Knowledge Extraction based on Content Structure Analysis of Academic Papers

内山清子 Kiyoko UCHIYAMA　亀田尭宙 Akihiro KAMEDA　武田英明 Hideaki TAKEDA　相澤彰子 Akiko AIZAWA

何がわかる？

論文の記述内容・文脈を解析することにより、専門分野における必要な知識や関係を知ることができる。

(1) 専門分野で使われる用語の必須性や基礎性

(2) 論文を代表する語句の間の関係

どんな研究？

(1) 専門分野において最低限学習する必要がある用語、その関連連用語の分野基礎性を識別する研究

(2) 論文内で関連研究を記述した部分に着目し、用語や周辺単語を手がかりに解析することにより、論文間の関係とそのキーワードを抽出する研究

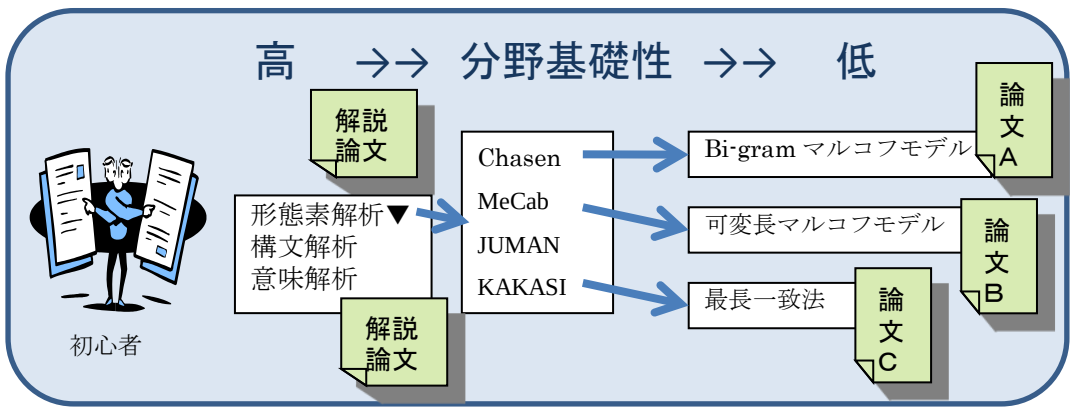
背景

学術論文を読む、書く、検索する時に

- 新しく研究する分野において最初に読むべき論文がわからない
- 分野の概要を効率的に把握したい
- 論文を書く時に参考となる論文リストを効率的に検索したい
- 分野における中心となる研究者や、話題の研究者を知りたい

新しい分野をこれから学ぶ初学者(学部生や他分野の研究者)にとっては、毎年大量に出版される論文の中から自分の知識レベルに適した論文を探すことが難しい。論文を理解するために必要な用語、用語間の関係などの知識をあらかじめ提示することにより、論文の内容を効率的に学習することができる。これらの知識を獲得するための手法を提案し、初学者が最初に読むべき論文リストを推薦するシステムに応用する。

専門用語の分野基礎性



分野基礎性とは

特定分野において、その用語の理解がなければ論文を読み進めることができない重要で基礎的な用語と定義する。分野基礎性が高い語とそれに関連する用語とのリンクがわかれば、分野を学習する際、優先度が明確で効率的である。また、用語と関連する論文の関係がわかれば、学習を進めて行く上で読むべき論文を簡単に選択できる。

分野基礎性が高い用語の出現傾向の一つとして、基となる用語の前後に単語が接続して出現する傾向、多くの複合語を生成する語構成性に着目

仮説:基礎性が高い語は、複合形式での出現が多い

例:日本語形態素解析、形態素解析システム、形態素解析処理

【対象データ】

情報処理学会自然言語処理研究会(NL)14年分1993年から2006年までの1421論文の書誌情報(タイトル、抄録、キーワード)

【正解データ】

事典の索引語と対象データの著者キーワードの中から専門家が分野基礎性が高いと判断した500語を4段階に分類したもの

(a)分野基礎性が高い語の抽出方法

対象複合語(CN)について左右に連接する頻度、コーパス中にCNが出現する頻度$n(CN)$、CNを含むより長い複合名詞の出現頻度$t(CN)$・異なり語数$c(CN)$、構成単名詞数$length(CN)$、CNの単独出現頻度$f(CN)$を属性として、MC/C-Value, FLR等を使って、上位100から500語の順位を求め、F値を算出して評価

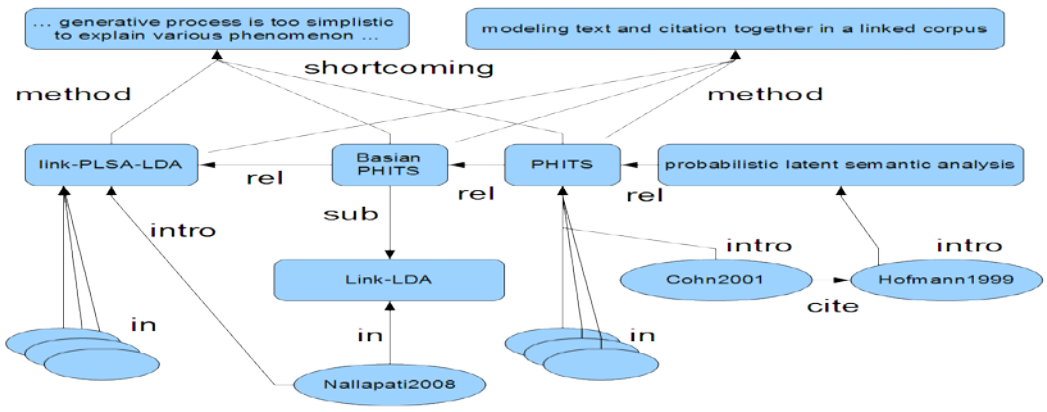
(b)分野基礎性の度合いを識別する方法

C-Valueによってランキングされた用語の各レベルにおける平均順位を算出

$$C\text{-}Value(CN) = (length(CN) - 1) \times (n(CN) - t(CN)/c(CN))$$

【結果と今後の課題】C-Valueによる用語スコア付けで分野基礎性が高い語が上位にランキングされた。分野基礎性の度合いは、4段階ではなく、基礎性が高い語を中心として関連用語とのリンク付けをする方が理解しやすい。

論文間の意味関係抽出



意味関係とは

本研究では、論文の中で手法(e.g. machine learning [機械学習]; clustering[クラスタリング])やテーマ(e.g. modeling text and citation[テキストと引用関係のモデリング]; trust information[信頼情報])といったものを表す概念を抽出し、さらにこれらの関係を数種類に分類して抽出する. これら概念とその関係を介して結び付けられる論文間の関係を、論文間の「意味関係」と呼んでいる。抽出する意味関係の例を以下にあげる。

【意味関係の例】

"… and was called PHITS (Cohn 2001). PHITS proposed a topical clustering of citations in a manner similar to the topical clustering of words proposed in PLSA (Hofmann 1999). …"

ここから (Cohn 2001)で提唱されているPHITSと(Hofmann 1999)で提唱されているPLSAに何らかの関連性があることが抽出でき、上の文章がその関連の説明文になっていることが研究者に提示できるため, 例えば, PLSAについて関連技術を調べている研究者を助けることができる。

【抽出手順】

1. 論文内の関連研究(Related Work) 章を取り出す
2. 論文内の引用文献標識(e.g. [1], [Kameda 2011])を抽出する
3. GeniaSSを用いて一文単位に切り取る
4. NLTKを用いてPOS タグの付与や, 名詞句節のタグの付与を行い, データを生成する
5. ルールを用いて概念と文型から抽出できる明示的な関係を抽出する
6. Conditional Random Fields (CRF) を用いて各文のレトリカルな役割を推定し, それを利用して, 文型からは抽出できない暗黙的な関係を抽出する
7. "A , B and C"など纏めて抽出した概念を分割する

【今後の課題】各ルールやCRFによる抽出の再現率と精度に付いて量的な分析を行うため, 現在100程度の論文のデータを用意し, 分析を行っている. 今後はその結果を踏まえ, ルールの選択を含め性能を高めるための工夫をしていきたい。

連絡先: 内山清子 (Kiyoko Uchiyama): 国立情報学研究所　学術コンテンツサービス研究開発センター　Email : kiyoko@nii.ac.jp
亀田尭宙 (Akihiro Kameda): 東京大学大学院　情報理工学系研究科　コンピュータ科学専攻　Email : kameda@nii.ac.jp