第2週B4 ゼミ

オントロジー構築

WikipediaとWebの情報を組み合わせたオントロジー構築の試み

出典

- 白川 真澄 中山 浩太郎 荒牧 英治　原 隆浩 西尾 章治郎
- WikipediaとWebの情報を組み合わせたオントロジー構築の試み
- 掲載誌名 電子情報通信学会論文誌. D, 情報・システム / 電子情報通信学会 編
- 掲載年 2011　3月
- 掲載巻 94
- 掲載号 3
- 掲載通号 471
- 掲載ページ 525〜539
- http://ci.nii.ac.jp/naid/110008460412/

概要・目的

- Wikipediaの概念の網羅性を利用した大規模オントロジー
  - 概念間の関係の網羅性が低い
- is-a関係、part-of関係のみのオントロジーが多い
- 情報検索などのアプリケーションは固有名詞、専門用語概念間の関係を定義したオントロジーが必要

- Webの情報を組み合わせることで
  - 固有名詞及び専門用語の網羅性
  - 概念間の関係の種類の豊富さ

- 以上の2点に着目した概念間の関係を抽出する手法の提案

関連研究

- Wikipediaを解析し大規模オントロジーを構築した例
- DBpedia
  - インフォボックスに注目
- YAGO
  - WorldNetのクラスにWikipediaのカテゴリーをマッピング
- 隅田らの研究
  - Wikipediaのカテゴリー構造、リスト構造からis-a関係を抽出

- is-a関係、part-of関係など定義、形式化が容易な関係のみを対象としたオントロジー

関連研究

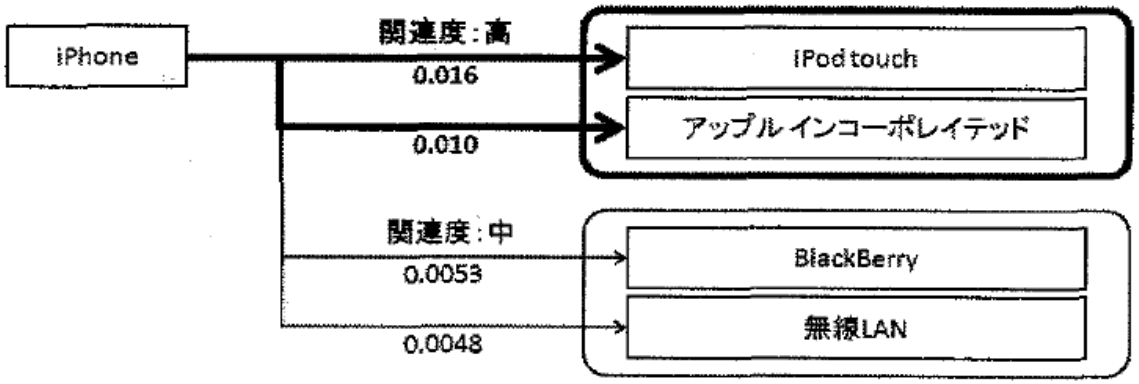
- 格フレーム辞書
  - 名詞のクラスタリング
    - 名詞間の関係を動詞によって表現
- 名詞を概念、動詞、格助詞を関係
  - 一種のオントロジー

- 固有名詞、専門用語は共起情報が少ないためあまり定義されていない

提案手法

関連のある概念ペアの抽出

- Wikipediaシソーラスを利用



- 固有名詞、専門用語についても関連度が定義されている

手法	Top10	Top20	Top30
語の共起性を用いた手法	46.2%	35.4%	30.7%
Chenらの手法	50.0%	50.9%	41.7%
Wikipediaシソーラス	93.2%	86.2%	83.1%

Web検索を用いた概念間の関係抽出

- 概念間の関係(動詞)を抽出
- 「孫正義」、「ソフトバンク」
  - 「孫正義がソフトバンクを創業する」
  - 「孫正義がソフトバンクを経営する」
- 概念ペアに付属する格助詞と動詞の組み合わせが得られる

- 頻度情報と同義語情報により重要度を算出する

格概念の同義語と信頼度の修得

- 表記の揺れ、省略形
  - 網羅性、精度が低下

- Wikipediaのリンクテキスト
- ある概念の記事に出現したリンクが設定された文字列
  - リンク先の概念を端的に表した表現が使用される
    - 同義語として使用可能
    - 例) ソフトバンク、日本ソフトバンク、Softbank

格概念の同義語と信頼度の修得

$$R_C(l_i) = \frac{\ln(M_C(l_i))}{\ln(\max(M_C(l_k)))}, l_k \in L_C$$

c：概念 l:リンクテキスト　M_c：使用回数　R_c：信頼度

i,k：変数　L_c：概念cへのリンクテキストの集合

リンクテキストは同義語として使用できるが、使用回数が多いほど同義語として適している

格助詞を付与したクエリによるWeb検索

- クエリ生成の時点で格助詞を付与し、Web検索を行う
- 数原らの研究　格助詞を付与しない場合
  - 箇条書きやリスト、題名がヒットし関係(動詞)が抽出できない場合がある
- 概念の関係は文章中に記述されている可能性が高い

関係抽出

- 各クエリによって得られたWebページ上位N件
  - 句点、ピリオド、改行などにより区切る
- この中から各概念が両方含まれる文を抽出
  - 概念の判定は同義語の情報を用いる

- CaboChaによる係り受け解析、MeCabによる形態素解析により概念ペアの両方が係っている動詞を抽出
  - 自立語の動詞とサ変接続の名詞

- 「する」は関係を表す動詞として意味をなさないので除外
- 得られた関係の出現回数を重要度として順位づけを行う

評価

- Wikipediaシソーラスからランダムに306の固有名詞同士の概念ペアについて格助詞を付与したクエリを生成
- 検索結果が多い格助詞の組み合わせ上位5つについてWebページ50件ずつ取得
- 人手で正解、部分正解、不正解に分類
  - 部分正解：補足情報が必要なもの、場合によって正しくないもの

	比較手法	提案手法
部分正解を含む	0.141	0.258
部分正解を除く	0.081	0.160

まとめ

- Wikipediaマイニングによって抽出した情報にWeb全体の情報を組み合わせることで、概念及び概念間の関係を網羅した関係抽出法を提案した

- 比較手法に比べ2倍程度の数の正しい関係を抽出した