意味的な情報検索のための関係パタンの獲得

兼岩研究室　鈴木諒

RDF

- RDFとはリソースを記述するためのフォーマット
- 主語、述語、目的語の3ペアで表現する

例: 電気通信大学のRDF(一部)

電気通信大学, locationCity, 東京都
電気通信大学, ウェブサイト, <http://www.uec.ac.jp/>
電気通信大学, 創立年, 1918
電気通信大学, キャンパス, 調布が丘(東京都調布市)
電気通信大学, 学校種別, 国立
⋮

http://ja.dbpedia.org/page/電気通信大学

目的

- コンピュータは次の2つの文を同じ意味の文章だと理解できない

緋色の研究を執筆した人はコナンドイルである
緋色の研究を書いた人はコナンドイルである

- 人間は執筆した人と書いた人が同じ意味だと理解できる
- Web上のテキストから単語の意味や単語の意味的関係に関する知識をRDFとして構築することでコンピュータが人間と同じレベルで文を理解することを可能にしたい
- 単語とRDF上に既に定義されているリソースと同じであるとしてRDF上では扱いたい

関係パタン

- 例えば以下のような関係パタンが抽出できる

XはYを書く

- XとYにはそれぞれ適切な名詞が当てはまる
- RDFの主語、述語、目的語の形とほぼ同じなので、これを中間表現としてテキストからRDFを構築できる

私は本を書く

夏目漱石はこゝろを書いた

コナンドイルは緋色の研究を書いた

⇩

XはYを書く(関係パタン)

{{X,Y}∈{{私,本},{夏目漱石,こゝろ},
{コナンドイル,緋色の研究}}}

提案手法

- 下図のように生成したベクトルによる名詞の分類の改善の提案
  - 係り受け関係を考慮したベクトルを生成
  - 係り受け関係を考慮した名詞のクラスタリングや機械学習が可能

- word2vecを用いて生成した特徴ベクトルと、提案手法のベクトルの次元数を減らして生成した2種類の特徴ベクトルを用いる

・コナンドイルはシャーロックホームズシリーズを執筆した
・コナンドイルはシャーロックホームズシリーズを書いた
・コナンドイルによるバスカヴィル家の犬
・コナンドイルは緋色の研究を書いた
・コナンドイルは緋色の研究を書いた
・コナンドイルは緋色の研究を書いた

⇩

コナンドイル ＝

(Xは,執筆する)	(Xは,書く)	(XによるY)	…
1	3	1	…

- このままでは次元数が多すぎるので、ベクトルのうち、有用なベクトルのみを抽出して使用する
- 特徴抽出の仕方は以下の3通りを考える
- この3通りのそれぞれの組み合わせで検討する

  - W2Vベクトル(既存手法)
    word2vecを用いて生成した特徴ベクトル
  - FREQベクトル
    提案手法のベクトルのうち、
    出現頻度の高い上位500次元の特徴ベクトル
  - IGベクトル
    提案手法のベクトルのうち、
    情報利得率の高い上位500次元の特徴ベクトル

現状と今後

- 現状
  - 分類の精度が不十分
  - 計算に非常に時間がかかる

- 今後
  - 他の特徴抽出の仕方を検討する
  - 分類器の最適なパラメーターを調査する
  - プログラムを修正して計算時間を削減を図る