

意味談話解析勉強会 Measuring Semantic Similarity

2007年11月26日(月)
奈良先端大D1小町守

紹介する論文

- Peter D. Turney. Measuring Semantic Similarity by Latent Relational Analysis. IJCAI-05.

概要

- Latent Relational Analysis(LRA)を提案
 - Vector Space Model(VSM)の拡張
- コーパスからパターンを自動で獲得
- 頻度のスムージングのためにSingular Value Decomposition(SVD)を用いる
- 同義語を使う

例

- 猫:にゃー = 犬:わんわん
 - 「猫」と「にゃー」の類似度を測る
 - 「猫」と「にゃー」は「犬」と「わんわん」と同じくらい近い→LRAで表現

イントロダクション

- Relational similarity と Attributional similarity
 - Attributional similarity が高い=synonymの関係
 - Relational similarity が高い=analogousな関係

作りたいもの

- テキスト2つ与えたらそれらの類似度を返すような「ブラックボックス」
 - でもそのようなものはない
- “plant” の2つの語義
 - Lesk などのアルゴリズムでは「工場」と「植物」の語義との類似度を計算するが、関係性は unclear
 - “food for the plant” と “food at the plant” のように、間にある単語で関係が分かる

VSMのアプローチ

- 単語XとYの間の単語を素性とするベクトルを作って cosine similarity を計る
 - “X of Y”, “Y of X”, “X for Y”, “Y for X” など 128次元
 - SAT の analogy question (5択問題)で47%

LRAの使うリソース

- 大規模なコーパスを用いた検索エンジン
- 同義語が大量に入ったシソーラス
- SVDの効率的な実装
- ラベル付きデータは不要

LRAの手順

1. A:Bの original pair から alternate pair を作る(A:B から A':B と A:B' を作る)
2. Alternate pair のうち頻度の低いものを取り除く
3. 検索エンジンで original pair と alternate pair 両方検索し、フレーズ抽出 (←“food for the plant”のように間にある単語列が関係を規定する)

LRAの手順(2)

4. 間にある単語列でパターンを構築(表層パターンおよびワイルドカード)
5. 単語を類似度行列の行に対応づけ
6. パターンの類似度行列の列に対応づけ
7. 疎行列の作成
8. エントロピーの計算(各要素は頻度が入っているが、それをlog取ってエントロピーで重みづけしたもので置き換える)

LRAの手順(3)

9. SVDを走らせる($U\Sigma V^T$ の計算)
10. $U_k \Sigma_k$ の計算(cos 類似度を測るため)
11. Cos 類似度の計算(alternates の評価)
12. Relational similarityの計算(cos 類似度の平均)

SATのanalogy questionで評価

- 2つの単語を与えられて、それと同じ関係になっている単語対を5択で選択
 - 正解すれば1点、不正解だと0点
 - 分からない場合はスキップすれば0.2点
- LRAはスコア56.4%
- VSMは47.3%、アメリカの高校生の平均は約57%(LRAのスコアと統計的に有意ではない)

名詞修飾関係で評価

- 30個のfine-grainedな関係 (cause="flu virus", location="home town", part="printer tray", topic="weather report"など)と5個のcoarse-grainedな関係 (causal, temporal, spatial, participatory) で評価
- Nearest neighbour のLRAスコア

LRAとVSMの比較

- Fine-grained でのaccuracyはVSMで27.8%、LRAで39.8%
- Coarse-grained でのaccuracyはVSMで45.7%、LRAで58.0%
 - Cf. 先週のCJEで紹介したDisambiguating Noun Compoundsは20クラスの分類で42.6%

議論

- VSMよりLRAのほうがはるかによい性能
 - 実アプリではprecisionとrecallをいじるとよい
- LRAの速度は遅い
 - ほとんどの操作は並列化可能なので、アルゴリズムの変更で対応可能
- VSMはコーパスサイズの影響を受ける
 - LRAは小さいサイズのコーパスでもVSMを上回る性能

問題点

- どのパターンがタスクに有効か分からない
 - 類似度を測っているだけ
- →1年後の論文 Peter D. Turney. Expressing Implicit Semantic Relations without Supervision. COLING-ACL 2006. へ。