

数量表現を伴う文における含意関係認識の課題分析

成澤 克麻 渡邊 陽太郎 水野 淳太 岡崎直観 乾 健太郎 (東北大学)

概要

- ▶数量表現を伴う文の含意関係認識が重要
- ▶日本語含意関係コーパス (RITE、[小谷 08]) を対象に、数量表現が問題となる事例を抽出、分析。解決のため必要となる処理を明確にした

- ▶結論：半分の事例については基本的な問題を同定。残りの問題は複雑な推論を要するAI的な問題であり部分問題には分けにくい
- ▶基盤技術と考えられる、数量表現の規格化システムを作成・公開した

背景

- ▶テキストtが仮説hを含意：tからhを推論可能である

t: 「ネット広告は15%伸びたが、テレビの広告は3%しか伸びなかった。」
h: 「ネット広告はテレビよりも伸びている。」

- ▶このような数量表現を伴う文の含意関係認識は有益
- ▶しかし、現状の研究は十分な対応ができていない。
→既存研究における問題の分析不足が原因の1つ

既存研究

- ▶ RTE-6で、数に関する簡単な処理を行っているシステムは全18システム中5つ → **難しい処理は手つかず**
- ▶ [Sammons 2010] RTE-5のコーパスとシステムを分析。**数の推論**の問題が解決できていないと指摘
- ▶ [LoBue 2010] Sammonsらと同様にRTE-5を分析。四則演算のような処理、そのための知識が必要と指摘
→ **どちらも指摘のみ。詳しく分析されていない**

よって、本研究では課題の分析を行う

問題の分類、分析

数量表現を伴う文における含意関係認識の問題

文節レベルの問題 (58事例、49%)

数量表現間の含意関係

「宇宙の年齢は130億歳だ」
「宇宙の年齢は100億歳以上だ」

- 単位の統一、数の包含関係の認識が必要 ← 規格化により解決
(数量表現以外の) 量を表す表現との含意関係
・客観的な量を表す表現

「日本人のおよそ5割がそれを信じている」
「日本人のおよそ半数がそれを信じている」

- 対応する数量表現についての知識が必要 ← 表現の数は限られている
・主観的な量を表す表現

「AKBのライブには毎回1万人が押し寄せる」
「AKBのライブには毎回大勢の人々が押し寄せる」

- 量の大小に関する知識が必要 ← 対象が無限にあるので、単純な知識獲得では難しい

2つの項をもつ数量表現による含意関係

「人間の遺伝子は予測を含めて3万2615個で、ショウジョウバエの遺伝子は約1万5千個である」
「人間の遺伝子は予測を含めて3万2615個で、ショウジョウバエの遺伝子はその半分である」

- 「何の」割合なのかを認識、処理する必要 ← 照応のような処理

並列関係にある表現との含意関係

「北京の展覧会には日本、中国、韓国の漆芸作家の作品が並ぶ」
「北京の展覧会には三国の漆芸作家の作品が並ぶ」

- 並列表現の認識、それを「数える」表現の認識が必要

数量表現の規格化

- ▶規格化を行う事で、数量表現間の含意関係は認識できる
▶例：3トン → [kg, 3000 ~ 3000]
2000kg以上 → [kg, 2000 ~ ∞]
単位が等しく、数の範囲の包含関係も正しいので、含意

評価実験

- ▶対象：NAISTテキストコーパス (新聞記事)
- ▶文中の数量表現・時間表現を認識・規格化できたかを人手で評価
- ▶結果：Precision 0.95 recall 0.91

文構造レベルの問題 (19事例、16%)

被限定名詞の同定

「韓国では22.3%の女性が整形経験者である」
「韓国では整形経験者の女性が22.3%いる」

- 数量表現が限定する名詞を同定する必要 ← 述語項構造が分かれば可能

意味レベルの問題 (61事例、52%)

テンプレートに基づく立式と評価 (8事例)

「五羽の子ウサギが産まれて、三羽が死んでしまった」
「二羽の子ウサギが生きている」

- テンプレート (例：「変化前+変化量=変化後」) を用いた情報抽出により、このような問題は解決できる可能性

その他 (複雑な推論) (53事例)

「21世紀半ばには最悪の場合、全人口の7割以上にあたる70億人が水不足に直面する」
「近い将来、世界は深刻な水不足になると懸念されている」

- 様々な要素の複合問題 (「全人口の7割以上が水不足」ならば「深刻な水不足」)

今後の課題

- リソースの整備が必要
- ▶分析した事例数が少ない。更なる分析のためのデータセットが必要
- ▶解決システムを評価するための評価セットが必要