E-008

構文解析を用いた常識的感覚データベース構築のための知識獲得
Syntactic Analysis-based Method for Automatic Creation of Sensory-related Common Sense Knowledge Base

三橋 奎太† ジェプカ・ラファウ† 荒木 健治†
Keita Mitsuhashi Rafal Rzepka Kenji Araki

1. はじめに

ロボットは常識的知識が不足しているため, 人間のように正しく言葉の意味を理解することができない. このような問題を解決するために常識的知識データベースを構築する研究が様々行われている. 常識の中でも最も基本となるのは, 人間が物体を認識することによって得られる知識であると考えられる. センサーによってロボットに物体を認識させることは可能であるが, 人間の言葉の意味を理解させるには, センサーによって得られた数値データを言葉と結びつける必要がある. そのため人間の言語理解には, 文書や音声データを扱う必要がある.

本研究では, 常識的感覚を収集し, 常識的感覚データベースを構築することを目的としている. 常識的感覚データベースとは, 物体が人間の五感に与える刺激情報を感覚ごとに分類したデータベースである.

常識的知識データベースとして有名なものに ConceptNet[1]があるが, 常識的感覚に関する知識は少なく, 感覚ごとに分類されていないため, 曖昧性が存在する. また, 渡部ら[2]は概念ベースと呼ばれる独自のデータベースを使用することで常識的感覚データベースを構築したが, 渡部らの手法では, 事前に手動で知識を登録する必要がありコストがかかる. ジェプカら[3]はブログデータからルールベースで知識を抽出し, 289 種の物理名詞に関して常識的感覚を付与し, 0.92 ポイントと高い精度を出したが, 再現率は 0.08 ポイントと低い値になった.

本稿では, ブログデータからルールと構文解析を用いることで常識的感覚データベースを構築する手法を提案する.

2. システム概要

YACIS コーパス[4]にルールと構文解析を用いて, 名詞(“一般”, “固有名詞”)と感覚語のペアを抽出し, 常識的感覚データベースを構築する. 感覚語とは, “赤い”, “暖かい” など五感に関する語であり, 感覚語のカテゴリ(“sight (視覚)”, “touch (触覚)”, “hearing (聴覚)”, “taste (味覚)”, “smell (嗅覚)”)に応じて知識を分類し, データベースに登録する. データベースの例を表 1 に示す. 具体的なデータベースの構築方法について以降説明していく.

表 1 常識的感覚データベースの例

	sight	touch	hearing	taste	smell
林檎	赤い	硬い		甘い	
犬	茶色い	温かい	うるさい		
ゴミ	汚い				くさい

† 北海道大学大学院 情報科学研究科, Graduate School of Information Science and Technology, Hokkaido University

2.1 抽出対象の感覚語の選定と感覚語カテゴリの付与

感覚語およびそのカテゴリには渡部ら[2]の作成したものを参考に決定した. 渡部らが作成した 98 種の感覚語のうち, 複数のカテゴリに属するもの 11 種は, 曖昧性があるため除外した. また, “痛い”, “痒い”, “多い”, “少ない” の 4 種の感覚語は物体の性質を示すものではないとして除外した. その結果合計 83 種の感覚語と感覚語カテゴリのペアが得られた.

2.2 名詞と感覚語のペア抽出

YACIS コーパスから以下の図 1 のルールにあてはまる名詞と感覚語のペアを抽出した. “—” は係り受け関係にあることを示す. 構文解析器には, CaboCha[6]を使用した.

ルール 1:
感覚語 + 名詞 + (“は” or “が” or “を” or “に” or “で”)
ルール 2:
文節末尾 { 感覚語 + (“の” or “な” or “null”) }
— 文節先頭 { 名詞 + (“は” or “が” or “を” or “に” or “で”) }
ルール 3:
名詞 + (“は” or “が” or “も”) + 感覚語
ルール 4:
文節末尾 { 名詞 + (“は” or “が” or “も”) }
— 文節先頭 { 感覚語 }

図 1 名詞と感覚語のペア抽出ルール

ルール 1 のみ (ジェプカら[3]の手法) 用いたものをベースラインとし, ルール 2 のみ用いたものを提案手法 1, ルール 1 とルール 3 を組み合わせたものを提案手法 2, ルール 2 とルール 4 を組み合わせたものを提案手法 3 とする.

2.3 名詞フィルタの作成

2.2 のルールのみであると, “気持ち” や “可能性”, “腰” など抽象名詞まで抽出してしまうため, それらを除外する名詞フィルタを作成した. 分類語彙表[5]の部門番号 1.2:主体, 1.3:活動, 1.4:生産物, 1.5:自然のいずれかに属する 368 種の分類項目について, 20 代男性 2 名(社会人, 理系大学生), 20 代女性 1 名(社会人)の被験者 3 名に対し, 実体の存在する集合か判定してもらい, 3 名全員が実体ありと判定した分類項目 89 種に属する名詞 7,449 件を名詞フィルタとした. このとき複数の分類項目に属する名詞 2,113 件は曖昧性があるため除外している.

2.4 データベースの構築

2.2 で得られたペアの名詞に対し, 2.3 で作成した名詞フィルタを用いてフィルタリングを行い, 残ったペアを感覚

語に付与されているカテゴリに応じてデータベースを構築した.

3. 正解データ

日常生活でよく使用される物理名詞について正解データを作成するため, Google N-gram[6]のコーパス中で感覚語と共起しやすい名詞 (フィルタリングしたもの) 上位 150 件を抽出した. これら 150 種の名詞と 83 種の感覚語とのペアを作成し, 20 代男性 3 名(社会人, 理系大学生, 理系大学院生)に “○:常識である”, “△:非常識ではない”, “×:非常識である” の 3 つの評価値で評価を行なってもらった. そして, 3 名全員が “○:常識である” と答えたものを “○”, 3 名全員が “×:非常識である” と答えたものを “×”, それ以外のものを “△” として正解データを作成した. このとき 1 名でも名詞の意味がわからないと回答したデータは正解データから省いた. “○” 91 件, “△” 6,304 件, “×” 4,227 件の合計 10,622 件の正解データが得られた. 正解データの一部を図 2 に示す.

○: (トイレ sight 白い), (桜 sight 綺麗), (馬 sight 茶色い), (うどん touch 柔らかい), (スープ taste おいしい), etc.
△: (風船 sight 緑), (バッグ sight 高級), (クッキー taste 苦い), (トマト sight 新しい), (うさぎ sight 大きい), etc.
×: (時計 taste おいしい), (チーズ sight 若い), (水晶 sight △), (歯 hearing 賑やか), (洋服 touch 硬い), etc.

図 2 正解データの一部

4. 評価実験

正解データのうち “○” あるいは “△” のデータをシステムが抽出できていた場合に正解として評価実験を行なった. 実験結果を表 1 に示す.

$$(\text{適合率}) = \frac{(DB\text{に含まれる "○" } or \text{ "△" の数})}{(DB\text{に含まれる "○" } or \text{ "△" } or \text{ "×" の数})}$$

$$(\text{再現率}) = \frac{(DB\text{に含まれる "○" } or \text{ "△" の数})}{(\text{正解データに含まれる "○" } or \text{ "△" の数})}$$

$$(F\text{値}) = \frac{2 * (\text{適合率}) * (\text{再現率})}{(\text{適合率}) + (\text{再現率})}$$

表 1　実験結果

	知識数	適合率	再現率	F 値
ベースライン	13,753	**0.793**	0.292	0.427
提案手法 1	15,132	0.779	0.309	0.443
提案手法 2	21,300	0.743	0.402	0.522
提案手法 3	**22,718**	0.735	**0.408**	**0.525**

注) 太字は最高値を示す.

5. 考察

ベースラインで用いた図 1 のルール 1 では, 適合率は高いが再現率が 0.292 ポイントと低い値になった. ここに構文解析を適用することで F 値が 0.016 ポイント上昇した(提案手法 1). また従来の手法に新たなルールを追加したことで再現率が 0.1 ポイント以上上昇し(提案手法 2), さらにそこに構文解析を適用することで獲得知識数と F 値が最大となった.

今回の抽出手法では感覚語の活用形について考慮していなかったため, 新たに感覚語の活用形を考慮した抽出ルールを加えることで, さらなる再現率の向上が期待できる.

提案手法 3 で得られた知識のうち非常識であると判定されたものの中には, “ワイン sight 高い”, “洋服 sight 高い” など感覚語ではない意味を持つ語によるものがあった. これは, “高いワインを買ってしまった” という文章や “高い洋服を着る” という文章が存在するためである. また, “コーヒー touch 暖かい” など感覚語の間違いによるものもあった. これは漢字の間違いによるものであり, 本来は “コーヒー touch 温かい” として抽出されなければならない.

感覚語以外の意味を持つ語は, “価格” や “値段” などの関連語を用いることで除外できると考えられる. また, 漢字の表記間違いは, 物理名詞と同音異義の感覚語の共起頻度をとることで改善できると考えられる.

6. まとめと今後の課題

ブログデータから, 自動的に常識的感覚データベースを構築する手法を提案した. 新たな抽出ルールと構文解析を組み合わせることにより従来手法よりも知識数を約 9,000 件, 再現率を約 0.1 ポイントと大きく上昇させることができた.

今回の実験では, 第一著者が一人で感覚語の選択を行なったため, “寒い” や “暖かい” など物体の持つ特徴として不適切な感覚語も残ってしまった. これは人数を増やして, 再度決定する必要がある. また正解データの作成の際に被験者に単語を提示し, 常識判定をしてもらったが, 常識を判定することは難しいため, 文章を被験者に提示し, その中から得られた知識が, 物体が人間の五感に与える刺激情報か判定してもらうなど正解データの作成方法を見直す必要がある.

今後の課題としては, 単語をベクトル化する Word2vec[8] を用いて, 知識の拡張を行うことや, 機械学習を用いて常識的感覚データベースを構築すること, 抽出した知識の信頼度を示すスコアを導入することなどが挙げられる.

参考文献

[1] Robert Speer, Catherine Havasi, “Representing General Relational Knowledge in ConceptNet 5 ”, In LREC, pages 3679-3686 (2012).
[2] 渡部 広一, 堀口 敦史, 河岡 司, “常識的感覚判断システムにおける名詞からの感覚想起手法”, 人工知能学会論文誌, Vol.19, pp. 73-82 (2004).
[3] Rafal Rzepka, Kenji Araki, “Web-Based Five Senses Input Simulation – – Ten Years Later”, ことば工学研究会：人工知能学会第 2 種研究会ことば工学研究会資料, Vol.43, pp.25-33 (2013).
[4] Michal Ptaszynski, Pawel Dybala, Rafal Rzepka, Kenji Araki, Yoshio Momouchi, “YACIS: A Five-Billion-Word Corpus of Japanese Blogs Fully Annotated with Syntactic and Affective Information”, 2nd Symposium on LaCATODA, pp.40-49 (2012)
[5] 分類語彙表 https://www.ninjal.ac.jp/publication/catalogue/goihyo/
[6] CaboCha: YetAnotherJapaneseDependencyStructureAnalyzer http://chasen.org/taku/software/cabocha
[7] Google N-gram, http://www.gsk.or.jp/catalog/gsk2007-c/
[8] T Mikolov, I Sutskever, K Chen, G Corrado, and J Dean. Distributed representations of words and phrases and their compositionality. NIPS pp.3111-31119 (2013).