2R-01

評価表現を用いた道徳判断システム

山本眞大† 萩原将文†
†慶應義塾大学

1. はじめに

人間の知能を計算機上でシミュレートするためには，常識の処理は極めて重要な要素であり[1],知識獲得に関する研究は多く行われてきた[2,3]．中でも特に，道徳的常識の獲得は，人工知能作成の際に重要であると考えられ，数多くの研究がなされてきた．

Wallach ら[4] は，演繹的モデルと帰納的モデルのハイブリッドモデルを提案している．まず，演繹的に基本的な指針を定め，次に細部を帰納的に補強する．しかしこの研究は，提案のみに留まっており，その有効性は示されていない．

一方，Anderson ら[5]は，MedEthEx というシステムを提案している．彼らは，医療分野において善悪判断を行うシステムをロボット NAO に組み込んだ．しかし，専門家による大量の推論事例が必要であるという点，及び特定の分野のみにしか適応していないという点には課題が残っていると考えられる．

本論文では，評価表現を用いた自動的な道徳判断システムを提案する．ここで道徳判断とは，「ゴミをポイ捨てするのはよくない.」などの善悪判断のことであり，提案システムではこの尤もらしさの推定を行なう．

以降本稿では，第 2 章で提案システムを説明し，第 3 章で評価実験，第 4 章でまとめを述べる．

2. 道徳判断システム

提案システムは，学習フェーズと道徳判断フェーズに分けることができる．まず，学習フェーズにおいては，道徳判断の際に効果を発揮する評価表現を取得することを目的とする.次に，道徳判断フェーズにおいては，自由文形式の１文を入力とする．そして，学習フェーズで獲得した評価表現との共起頻度を基にした道徳判断が行われる．出力形式は,1.00～5.00 の範囲のアナログ値であり，1.00 に近いほど悪い行為，5.00 に近いほど良い行為である.

Moral Judgment System using Evaluation Expressions
Masahiro Yamamoto † and Masafumi Hagiwara †
†Department of Information and Computer Science
Keio University Yokohama, 223-8522 Japan
{yamamoto, hagiwara}@soft. ics. keio. ac. jp

2.1 学習フェーズ

入力として，文部科学省の小学校の道徳の指導要領に書かれている文，及びその極性を用いる．学習フェーズにおいては，これらの文を 2 文節もしくは 3 文節の短い文に区切って用いる．

入力文は，形態素解析エンジンの Mecab により形態素解析され，名詞（一般，固有名詞，サ変接続，形容動詞語幹），動詞（自立），形容詞（自立）を抽出する．

次に，これらの抽出単語群と極性辞書中の単語が共起しているかどうかを調べる．その際，言語資源として，Web 本語 Nグラムを用いる．

そして，その共起情報を元に，評価表現辞書[6,7]中の単語にスコアリングを行う．具体的には，以下の式により positive 単語のスコアおよび negative 単語のスコアを算出する．

$$Score_{pw_i} = N_{pw_i,p} - aN_{pw_i,n} \quad (1)$$

$$Score_{nw_i} = N_{nw_i,n} - bN_{nw_i,p} \quad (2)$$

ここで$N_{pw_i,p}$は，極性辞書中の Positive な単語pw_iと共起した極性が Positive である抽出単語群の数であり，$N_{pw_i,n}$は，pw_iと共起した極性が Negative である抽出単語群の数である．同様に，$N_{nw_i,n}$は，極性辞書中の Negative な単語nw_iと共起した極性が Negative である抽出単語群の数であり，$N_{pw_i,n}$は，nw_iと共起した極性が Negative である抽出単語群の数である．

最終的には，$Score_{pw_i}$が大きい Positive 単語群および$Score_{nw_i}$が大きい Negative 単語群を出力する．

2.1 道徳判断フェーズ

道徳判断フェーズでは，学習フェーズで抽出された評価表現を元に，ユーザが入力した文の道徳判断をすることを目的とする．

自由文形式の１文が入力されると，名詞（一般，固有名詞，サ変接続，形容動詞語幹），動詞（自立），形容詞（自立）の抽出が行なわれる．

次に，抽出された単語群と学習フェーズで取得

Copyright ©2015 Information Processing Society of Japan.
All Rights Reserved.

された評価表現との共起頻度を求める．ここで，抽出単語群が多い場合，共起頻度の取得が困難となることがある．その際には，以下に説明する単語の簡約化を行い，重要度が低いと推定される単語を1つ取り除く．単語の簡約化は以下の2点を考慮して行われる．

・1文中に同じ単語が複数存在する場合，その単語に大きな重みを与える．
・言語資源中に頻出する単語が存在する場合，その単語に小さな重みを与える．

抽出単語群の簡約化が終わり，重要度の低い単語を取り除いた後，もう一度，学習フェーズで取得した評価表現との共起頻度を取得する．以降，共起頻度が取得できるまで，単語の簡約化を繰り返す．

共起頻度が取得できたら，スコアリングが行われる．スコアリングの式は以下の通りである．

$$\mathrm{Score} = \frac{P_{score} - N_{score}}{P_{score} + N_{score}} \times 2.00 + 3.00 \qquad (3)$$

ここで，P_{score}およびN_{score}は，それぞれ，入力文のpositive度合い，Negative度合いを表し，以下の式によって求められる．

$$P_{score} = \sum_{i=1} \frac{F_{pw_i,ew}}{F_{pw_i}} \qquad (4)$$

$$N_{score} = \sum_{i=1} \frac{F_{nw_i,ew}}{F_{nw_i}} \qquad (5)$$

ここで，F_{pw_i}は，あるPositive単語pw_iが言語資源中に現れる頻度を表し，$F_{pw_i,ew}$は，pw_iと抽出単語群(ew)の共起頻度を表す．同様に，F_{nw_i}は，あるNegative単語nw_iが言語資源中に現れる頻度を表し，$F_{nw_i,ew}$は，nw_iと抽出単語群(ew)の共起頻度を表す．

3. 評価実験

道徳判断の精度を確認するために実験を行った．まず，道徳判断が可能な文を20代の学生10名から116文収集した．そして，20代学生23名の被験者に，1〜5の5段階で評価を行ってもらった．なお，評価の際は，116文をランダムで表示した．

図1に人間の評価の平均とシステムの出力の関係を示す．システムの出力と人間の道徳判断の相関係数は0.65であった．

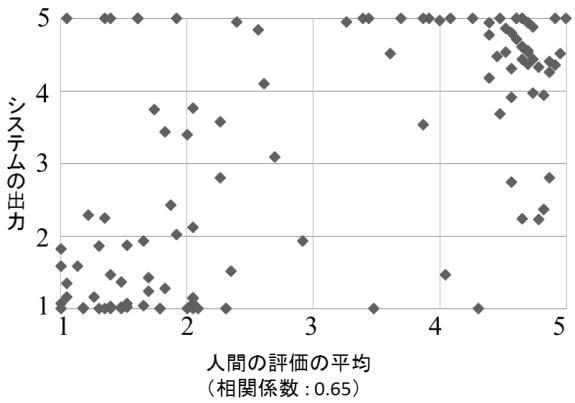


図1：人間の評価の平均とシステムの出力の関係

4. まとめ

本論文では，評価表現との共起を元に，スコアリングを行う道徳判断システムを提案した．提案システムは，学習フェーズと道徳判断フェーズから成る．学習フェーズでは，道徳判断を行う際に効果を発揮する評価表現が極性辞書から抽出される．道徳判断フェーズでは，入力文がどの程度善いことなのか，あるいは悪いことなのかが判断される．

評価実験においては，人間による道徳判断と，提案システムによる道徳判断の比較が行われた．

今後は，統計的な処理だけではなく，さらに意味内容まで考慮することにより精度向上を目指す．

参考文献

[1]M.Minsky, “ミンスキー博士の脳の探索 -常識・感情・自己とは-.”, 共立出版, 2009.
[2] D. Lenat, M. Prakash, M. Shepherd, “CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks.” AI Magazine, 6(4):65-85, 1986.
[3] P. Singh, “The open mind common sense project.” KurzweilAI. net, 2002.
[4] W. Wallach and C. Allen, “Moral Machines: Teaching Robots Right from Wrong.” Oxford University Press, 2009.
[5] M. Anderson, S. Anderson, and C. Aren, “An approach to computing ethics.” Intelligent Systems, IEEE, vol. 21, no. 4, pp. 56–63, 2006.
[6] 佐野大樹, “『日本語アプレイザル評価表現辞書（態度表現編）』の構築”, 言語処理学会第17年次大会, pp. 115-118, 2011.
[7] 東山昌彦, 乾健太郎, 松本祐治, “述語の選択選好性に着目した名詞評価極性の獲得”, 言語処理学会第 14 回年次大会論文集, pp. 584-587, 2008.

Copyright ©2015 Information Processing Society of Japan.
All Rights Reserved.