ARG WI2 No.13, 2018
WI2-2018-22

常識的知識の自動評価手法及び英日辞書を用いた ConceptNet の拡張

首藤 聖矢　　ジェプカ・ラファウ　　荒木 健治

北海道大学大学院情報科学研究科

{shudo,rzepka,araki}@ist.hokudai.ac.jp

概要　人工知能分野での常識的知識獲得は重要な問題の一つとして取り組まれている. このような知識を収集するオントロジー研究として ConceptNet があり多言語に対応しているが, 英語と非英語間の知識量の差は大きい. よって, 本研究では英語版の知識を日本語に対応させることで知識を拡充することを目的とする. 我々が行なった研究として, ブログコーパスと対象の知識の関係性を表す手がかり語を用いて日本語の常識的知識の一般性を自動評価する研究がある. 本稿では, 英日辞書と常識的知識の自動評価手法を用いて日本語の常識的知識を自動獲得する手法を提案する. 本手法に基づく実験システムを作成し, 獲得した知識を対象とした評価実験及び考察において, 提案手法の有効性を示す.

キーワード　常識的知識獲得, 自動評価, ConceptNet

1　はじめに

人間と同等の知的システムを構築する上で重要な問題の一つとして常識的知識の不足が挙げられる. このような知識を蓄積した代表的なデータベースとして ConceptNet[1] がある. ConceptNet には *MadeOf* (紙, 木) のような 2 つの概念とその概念間の関係性の 3 つ組のアサーション (以下 *R(C1, C2)* とする) とそれに対する自然文で格納されている. ConceptNet には OMCS[1]の知識の他, GWAP で獲得された知識 [2] や WordNet[2]のデータが格納されている. このデータベースは多言語に対応しており, 英語とそれ以外の言語の知識量に格差があるため, 英語版の知識を日本語に翻訳し知識を拡充する必要がある. 我々のこれまでの研究として日本語の常識的知識の一般性をブログデータにおける共起頻度を用いて自動評価する研究 [3] がある. そこで本稿では, 辞書と常識的知識の自動評価手法を用いて日本語の常識的知識を獲得する手法を提案する.

2　関連研究

日本語の常識的知識の獲得を対象としたクラウドソーシングを用いる手法として, なぞなぞのゲームを用いた中原らの研究 [2], 音声アシストを用いた連想ゲームとして大谷らの研究[3] がある. Wikipedia の dump データを用いる手法としては Krawczyk らの研究 [4] がある. さらに, 機械翻訳と知識データベースを用いて言語間の投影を行った大谷らの研究 [5] がある. これらの研究と比較して, 本研究はクラウドソーシングを用いず, ローデータを用いてより常識的な知識を対象としており, 概念同士の関係性を表す手がかり語を利用している点, 利用する辞書もオープンソースを利用しており, 低コストでシステム作成を行える点が異なる.

3　システム概要

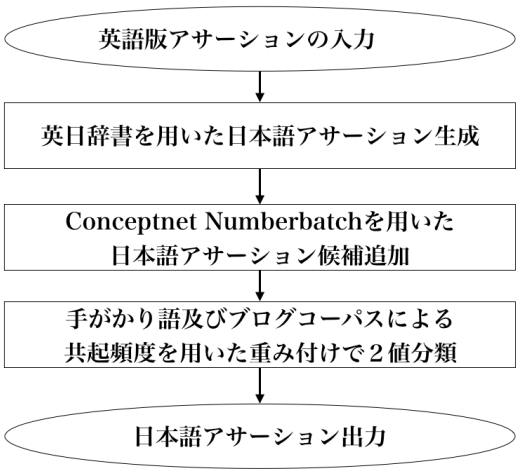

図 1 システム構成

本研究のシステム構成を図 1 に示す. 入力は CoceptNet の英語版のアサーションとして, 英日辞書 Edict[4]から各訳語を用いて日本語のアサーション候補を決定する. その後, 生成されたアサーションの *C1*, *C2* の類義語を ConceptNet の分散表現が含まれている Conceptnet Numberbatch [1] を用いて, 上位 10 件を新たな候補とした. 次に, 候補から正しいアサーション選定のために, ブログコーパス [6] での検索を行った. 検索の際は *C1* と *R* を表す手がかり語, また *C2* と *R* を表す手がかり語の組で形態素解析機

[1]https://github.com/commonsense/omcs
[2]https://wordnet.princeton.edu
[3]https://v-assist.yahoo.co.jp
[4]http://www.edrdg.org/jmdict/j_jmdict.html

ツール MeCab[5]を用いて検索結果のスニペットを抽出した. ここで, *R* を示す日本語の意味の手がかり語は人手で選定した. 例えば, *MadeOf*(紙, 木) に対しては「作る」という手がかり語を選定し, 以前の我々の研究 [3] と同じく, *C1-R* では「紙は作」, *C2-R* では,「木で作」等 4 件とした. さらに, 今回は *R-C1*, *R-C2* の組を 24 件追加した. 同じ例では,「作る紙」,「作っている紙」,「作られる紙」,「作られている紙」,「作った紙」,「作っていた紙」,「作られた紙」,「作られていた紙」等とした. 次に *C1-R*, *C2-R* に関する, 共起頻度による重み付けの算出方法のスコアをそれぞれ式 (1), (2) に示す.

$$Score1 = C1 - R + weight * R - C1 \quad (1)$$

$$Score2 = C2 - R + weight * R - C2 \quad (2)$$

上記の重み付けで *Score1*, *Score2* のスコアがともに正のものを正例, 0 または負のものを負例として判定した.

4 評価実験

正解データの作成はアンケート調査を用いて行った. 被験者は 20 代男子大学生 3 名, 20 代社会人男性 1 名, 30 代男子大学院生 1 名の計 5 名であった. 扱ったデータは英語版 ConceptNet の *MadeOf* アサーションからランダムに抽出した 100 件である. 被験者は英日辞書で翻訳された日本語のアサーションの候補 632 件, Conceptnet Numberbatch を用いて生成されたアサーションの候補 7430 件からそれぞれランダムに抽出された 100 件 (以下それぞれ Edict Dataset, Numberbatch Dataset 呼ぶ) の合計 200 件を大谷ら [5] に従い評価を行った. ここでは, 各アサーションをテンプレートの自然文に対して「1. 常に誤りまたは意味不明」「2. 誤りとは言えない」「3. 正しい場合がある」「4. 多くの場合正しい」「5. 常に正しい」から選択した. *MadeOf*(紙, 木) を例とすると,「紙は木から作られる」という自然文となる. 被験者の評価は中央値を正解データとし, 各評価に閾値を基準に 2 値に分類したものを用いた. 共起頻度による重みがない場合をベースラインとし, 提案手法との正解率での比較を行った. また, 提案手法の重みについては, Edict Dataset, Numberbatch Dataset に関してそれぞれ, 重み 11, 重み 3 とした. ここで, 各データセットに対する閾値 (1, 2, 3, 4) における実験結果をそれぞれ表 1, 表 2 に示す. 表 1 に関しては閾値が 2 以上, 表 2 に関しては全ての閾値において提案手法がベースラインを上回った.

実験結果より, 閾値を高くすることにより正解率が向上していることがわかる. Numberbatch Dataset では 90%以上の高い正解率であるため, Conceptnet Numberbatch で関連語を増やしアサーションを追加する方法は有効である. 問題点としては, 閾値が高いために獲得できる知識が存在しない場合である. 例えば, Numberbatch Dataset で閾値 3 の場合は獲得できた知識は *MadeOf* (書面, 書くこと) という知識で, 閾値 4 の場合は正例の知識を獲得していない. 不適切なものを除去できるが, 新たな知識を獲得するには改善の余地があると考えられる. 今回はある関係性の全てのアサーションを対象にしたため, 被験者が「5. 常に正しい」を選択しづらく, 正例の正解データが少ないことが原因と考えられる. 自然なコンセプトを予め選定することで, 獲得できる知識を増やすことが考えられる.

[5]http://taku910.github.io/mecab/

表 1 Edict Dataset

閾値	1	2	3	4
ベースライン (重みなし)	0.37	0.55	0.82	0.85
提案手法 (重み 11)	0.36	0.62	0.93	0.98

表 2 Numberbatch Dataset

閾値	1	2	3	4
ベースライン (重みなし)	0.72	0.82	0.85	0.82
提案手法 (重み 3)	0.73	0.85	0.88	0.91

5 おわりに

英日辞書及び自動評価手法を用いた常識的知識獲得手法の有効性を示した. 今後は評価対象のデータを選定し, より多くの正例の正解データを収集することで, 獲得知識を増やすことが可能であることが確認された. また, 対象の関係性を増やすことを検討する予定である.

参考文献

[1] Speer, R., Chin, J., and Havasi, C.: ConceptNet 5.5: An Open Multilingual Graph of General Knowledge, AAAI, pp. 4444-4451, 2017.

[2] 中原和洋, 山田茂雄: 日本でのコモンセンス知識獲得を目的とした Web ゲームの開発と評価, Unisys 技報 : Unisys technology review, Vol. 30, No. 4, pp. 295-305, 2011.

[3] Shudo, S., Rzepka, R., and Araki, K.: Automatic evaluation of commonsense knowledge for refining Japanese ConceptNet.: Proc. of the 12th Workshop on Asian Language Resources, pp. 105-112, 2016.

[4] Krawczyk, M., Rzepka, R., and Araki, K.: Extracting location and creator-related information from Wikipedia-based information-rich taxonomy for ConceptNet expansion, Knowledge-Based Systems 108, pp. 125-131, 2016

[5] Otani, N., Kiyomaru, H., Kawahara, D., et al.: Cross-lingual Knowledge Projection Using Machine Translation and Target-side Knowledge Base Completion, Proc. of the 27th International Conference on Computational Linguistics, pp. 1508-1520, 2018.

[6] Ptaszynski, M., Dybala, P., Rzepka, R., Araki, K.: Annotating affective information on 5.5 billion word corpus of Japanese blogs, In Proc. of The 18th Annual Meeting of The Association for Natural Language Processing, pp. 405-408, 2012