学会記事

A Method for Building a Commonsense Inference Dataset based on Basic Events

大村 和正†

まえがき

本稿では，EMNLP2020 に採択された表題の論文 "A Method for Building a Commonsense Inference Dataset based on Basic Events" (Omura, Kawahara, and Kurohashi) について解説する。採択論文は，拡張性・言語依存性のある常識推論データセットの構築手法を提案したものである。

1 はじめに

計算機による文章読解の実現は，自然言語理解の分野における大きな目標の一つである。この目標に向けて，計算機の文章読解能力を評価する問題設定を考え，そのデータを構築する研究が盛んに行われている (Wang et al. 2019a, 2019b)。本研究も例外でなく，「文章読解に向けた言語資源を構築する」という目標を設定したところから始まった。

具体的な問題設定を決めるにあたり，（公財）日本語能力検定協会が提供している文章読解・作成能力検定（文章検）の問題を参考に，文章読解に必要な知識の分析と問題のデザインの検討を行なった。この際に我々は，四級の文章検で出題される表 1 のような問題に目を付けた。この問題は次のような特徴があるために，自動生成を試みるにふさわしいと考えた。

(1) 基本的なイベント間の蓋然的関係1を問い，解答に常識を必要とする。
(2) 問題のベースとなるイベントは，読者側の知識が自動生成で抽出できる。

一点目の特徴は，計算機の常識推論能力の訓練／評価に利用できることを意味する。近年，深層学習はめざましい発展を遂げている。特に大規模テキストで事前学習した汎用言語モデル BERT (Devlin et al. 2019) などにより，言葉の認識や構文解析など，基礎的な文章読解能力を評価するタスクで人間に匹敵する精度が達成されている。しかし，これらの高性能なモデルでも，常識推論は依然として難しいことが示されている (Talmor et al. 2019)。そのため，計

†京都大学大学院情報学研究科
1ある事柄がある程度起こりうる／真であるという関係。因果関係より広範な原因-結果の関係に着目するという意図で，この用語を用いる。
2 提案手法

2.1 提案手法の鍵

1節で述べたように、テキストから抽出したイベントペアをベースに常識推論問題を作ること、そのイベントペアが常識的であることを保証する必要がある。そこで、本研究では基本イベントと蓋然的基イベントペアなるものを定義し、これらに着目する。基本イベントと蓋然的基イベントペアは以下のように定義される。

基本イベント テキストから抽出した述語對構造をクラスタリングし、その中的高頻度なものを核とする表現

蓋然的基イベントペア 蓋然的関係を持ち、前件・後件が基本イベントであるイベントペア

提案手法の鍵は、ある蓋然的基イベントペアをベースとしその他のペアの後件を誤り選択肢として、常識推論問題を自動生成できるというものである（表2）。蓋然的基イベントペアは、基本イベントを核とし、クラウドワークにによる確認を経るため、常識としての知識の一般性をある程度保証している。また、問題は蓋然的基イベントペアをもとに作られるため、常識推論能力を問うものになっている。

提案手法はテキストからの抽出をベースとするため、拡張性があり、ドメインが限定されることもない。クラウドソーシングについても、確認・フィルタリングを行うだけなので、低コストかつ作文によるannotation artifacts (Gururangan et al. 2018) の問題もない。また、人手

表1 四級の文章検例出題される問題例（実際の問題例より引用）。太字は正解選択肢である。

算機の常識推論能力を議論するために、表1のような問題を作ることは意義があると考えた。

また、二点目の特徴により、深層学習モデルの訓練データとして求められる規模のデータ構築が可能である。ただし、テキストから自動抽出したイベントペアは(1)の性質を持つ（常識的である）とは限らない。本研究では、基本的なイベント表現を含むことを条件付け、確認のためのクラウドソーシングを組み込むことで、この問題に対処している。

以上が、本研究の着想に至る経緯である。
A Method for Building a Commonsense Inference Dataset based on Basic Events

お腹が空いたので，ご飯を食べる
ご飯を食べたら，すごく疲れる
疲れて，コーヒーを飲む
激しい運動をすると，汗をかく

お腹が空いたので
a. コーヒーを飲む
b. ご飯を食べる
c. 汗をかく
d. すごく疲れる

表 2 難然的基本イベントベアの例（左）と，これらをもとにした常識推論問題の作成例（右）．太字は正解選択肢である．

図 1 提案手法の概要図

で構築されたデータやクラウドソーシングに強く依存しないため，談話理解を手掛かりにする言語においては同様の手法を適用することができる．

2.2 提案手法の全体像

提案手法は以下の 4 ステップから成る（図 1）．
(1) 格フレームから高頻度な述語項構造を基本イベントとして獲得する．
(2) テキストに係り受け解析・談話関係解析を適用し，その結果，難然的関係を持つと認識され，かつ基本イベントを核とするイベントベアを抽出する．
(3) 抽出したイベントベアが難然的関係を持つかどうかをクラウドソーシングで確認し，難然的基本イベントベアを得る．

2 本研究では，京都大学格フレーム (https://www.gsk.or.jp/catalog/gsk2018-b) を利用した．
表 3 常識推論問題の生成例。太字は正解選択肢である。

| 今週はずいぶん気温が下がったので | 南西の方角にエリの花を飾ると |
|---------------------------------|--------------------------------|
| a. ふるえることで体を温めようとします | a. 遠近法の効果で部屋が広く見えます |
| b. 問題なく行けるだろう       | b. 効果は別として、可愛い      |
| c. カゼひかないようにする      | c. ちょっと期待が高まる        |
| d. 早目に寝る                  | d. 多少運気が上がる          |

(4) 馥然的基本イベントペアから文脈と正解選択肢を作り、その他の程度に類似するイベントペアの後に到達選択肢を選択することによって、常識推論問題を生成する。各ステップの詳細については採択論文の 3・4 章を参照されたい。

2.3 提案手法の適用

7.1 億文を含む日本語ウェブコーパスに提案手法を適用した結果、常識推論問題が 10.4 万問生成された。表 3 に問題の生成例を示す。これらの常識推論問題について、計算機と人間による解答精度を検証した結果、BERT LARGE モデルの解答精度は 76.0%、多数決で集約したクラウドワークの解答精度は 88.9%であった（詳細は採択論文の 4・5 章を参照）。常識推論能力において、計算機と人間の間に聞きがあることを示している。

3 おわりに

採択論文では、テキストからの自動抽出とクラウドソーシングを組み合わせた、拡張性・言語依存性のある常識推論データセットの構築手法を提案した。各問題は、基本的なイベント間の関係性を問う多段選択式問題である。提案手法を日本語ウェブコーパスに適用し、大規模な常識推論データセットを構築した。計算機による解答実験の結果、高性能な転移学習モデルと人間の間に 10%程度の聞きがある、すなわち常識推論能力の評価に有効なデータセットであることを示した。構築したデータセットは、http://nlp.ist.i.kyoto-u.ac.jp/KUCI にて公開している。

今後は、構築したデータセットを利用して、常識推論向けモデルやアルゴリズムの改善に取り組む予定である。また、このデータセットで常識を学習し、省略・照応解析や会話関係解析など他のタスクに応用することを検討する。

提案手法は、実際の検定問題から着想を得たこともあり、文章題の作成支援にも利用できると考えられる。計算機だけでなく、人間の学習にも役立てられることを期待する。
A Method for Building a Commonsense Inference Dataset based on Basic Events

参考文献

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. (2018). “Annotation Artifacts in Natural Language Inference Data.” In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 107–112.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. (2019). “CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge.” In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4149–4158.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019a). “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems.” In Advances in Neural Information Processing Systems 32, pp. 3266–3280.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019b). “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding.” In International Conference on Learning Representations.

略歴

大村 和正: 2019年京都大学工学部電気電子工学科卒業。現在、同大学大学院情報学研究科修士課程に在籍中。