Wikipedia からの意外な恩恵事例の抽出

尾崎 立一 ‡　橋本 力 *　村脇 有吾 ‡　黒橋 禎夫 ‡　颯々野 学 §
‡京都大学 大学院情報学研究科　*§ヤフー株式会社
{ozaki,murawaki,kuro}@nlp.ist.i.kyoto-u.ac.jp　msassano@yahoo-corp.jp

1　はじめに

現代社会は地球規模での相互依存性の深化とともに複雑さが増し続けている．それにともない，リーマンショックや新型コロナウィルスの世界的流行のような予測困難な事態が発生し，それが社会に重大な影響を与えることがある．その一方で，そのようなネガティブな事態によって，意外なポジティブな事態が生まれることもある．新型コロナウィルスが旧来のシステムの問題点を明らかにし，デジタル化・オンライン化推進の契機になるような例である．このような意外な恩恵を含めて俯瞰的に問題を捉えることは重要であり，千載一遇のチャンスを掴む可能性を高めることにもなる．

そこで本研究は，意外な恩恵の事例を大規模テキストから自動抽出することを目的とする．大規模テキストとして Wikipedia 記事群を用いる．Wikipedia は百科事典であり，様々な事物を広範に扱っていること，また多くの人々が編集に参加し内容を精査しているため，情報の信憑性も高いと考えられることがその理由である．対象言語は英語である．以下に本研究で対象とする，意外な恩恵の例を挙げる．

Smoking reduces the risk of Parkinson's disease.

健康にとって有害だと広く知られている喫煙に，実はパーキンソン病のリスクを減らしうること[1)]が述べられている．この例が示す通り，本研究における意外な恩恵は，専門家を含む一部の人々には既知でもその他大勢の人々にとっては意外と感じられる恩恵のことである．

提案手法の基本的なアイデアは次の通りである: **ネガティブなものと広く知られているエンティティを何らかの恩恵をもたらすものとして述べている文は，その意外な恩恵を述べている可能性が高い．** 上述の *Smoking* の例は，健康にとって有害である点でネガティブといえる喫煙を，パーキンソン病リスク低減の恩恵をもたらすものとして述べている．*reduces the risk of Parkinson's disease* のような恩恵を述べている箇所をポジティブ説明，ネガティブなものと広く知られているエンティティをネガティブエンティティと呼ぶことにすると，上記アイデアを次のように言い換えることができる: **ネガティブエンティティについてのポジティブ説明を述べている文は，その意外な恩恵を述べている可能性が高い．** このアイデアを実現するには，ネガティブエンティティとポジティブ説明を自動認識する方法が必要である．2 節で詳述する．

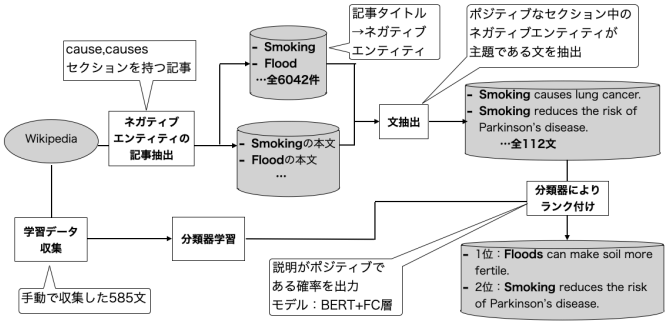

図 1　提案手法の全体像

英語 Wikipedia 全体からランダムに抽出した 300 文を調べたところ意外な恩恵を述べている文は 1 文も無かった．これは本タスクの難易度が非常に高いことを示している．一方，提案手法は英語 Wikipedia 全体から意外な恩恵の事例を 40 件自動抽出できた．改善の余地は大きいが，意外な恩恵の割合がそもそも非常に低いため，全数チェック等のナイーブな手法と比べ，はるかに効率的に意外な恩恵の事例を抽出できる．実験については 4 節で述べる．

2　提案手法

図 1 に示す通り，提案手法は，Wikipedia からネガティブエンティティを抽出し，次に，それが主題となっている文を抽出する．最後に，分類器によって，ネガティブエンティティのポジティブ説明を述べているものだけを残す．

2.1　ネガティブエンティティ抽出

本研究におけるネガティブエンティティは，病気や災害，事故，不況，テロ，戦争等の不運，不幸な

* 現在は楽天株式会社在籍．
1)　https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.10277

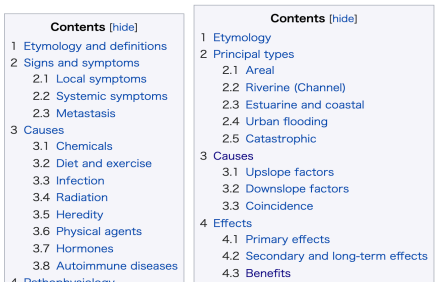

図 2　*Cancer*（左）と *Flood* のセクション群

事物，またはそれらの原因となりうる事物を指す．

英語の Wikipedia 記事として取り上げられている（記事タイトルとなっている）エンティティのうち，記事中に *Cause* または *Causes* というセクションが設けられているものは，*Cancer*，*Flood*，*Traffic collision*，*Great Depression* 等[2] ネガティブエンティティである場合が多い．これは，何らかの不運，不幸につながりうる，または実際そうなった事物について百科事典に記述する際は，その原因が重要な記載項目と見なされるからだと推測される．図 2 に *Cancer* と *Flood* の記事のセクション群を示す．

以上の観察結果に基づき，提案手法は *Cause*，*Causes* セクションを持つ記事のタイトルをネガティブエンティティとして抽出する．

2.2　文抽出

抽出したネガティブエンティティが主題の文を意外な恩恵事例の候補（2.3 節で述べる分類器への入力）として Wikipedia から抽出する．ネガティブエンティティが主題である文に限定するのは，単に当該エンティティを含む文だと，その説明がポジティブであれネガティブであれ，そもそも極性が当該エンティティに関するものではない可能性があるからである．

ここで問題になるのはいかにして文の主題を認識するかである．提案手法は，Wikipedia 記事中の当該記事のタイトルを含む文はそのタイトルを主題とした文である可能性が高い，という仮説のもと，ネガティブエンティティがタイトルである記事から，そのタイトルを含む文を抽出する．例えば，タイトルが *Flood* の記事から文字列 *Flood*/*flood* を含む文を抽出する．

加えて，意外な恩恵事例抽出の精度を上げるべく，文の抽出元を，*Benefit*，*Advantage*，*Positive* のいずれかの文字列をセクション名に含むセクションのみに限定する．つまり，提案手法は，ネガティブエンティティがタイトルである記事の中の，*Benefits* や *Advantages*，*Positive Effects* といったポジティブな

2）　2020 年 11 月 25 日現在

図 3　*Flood* の *Benefits* セクション

内容が書かれている可能性が高いセクションからのみ文を抽出する．ネガティブエンティティに関する記事の一部でありながらポジティブな内容が書かれているセクションは少数ながら存在する．例えば図 2 右にあるように，*Flood* はネガティブエンティティでありながら，その記事には *Benefits* セクションが含まれている．図 3 に示す通り，そうしたセクションにはネガティブエンティティのポジティブな側面が書かれる傾向がある．一方で，そうしたセクションにある文全てが意外な恩恵を述べているわけではない．そこで 2.3 節で述べる分類器により，抽出した文を意外な恩恵を述べている可能性の高いものとそれ以外に分類する．

2.3　分類器

提案手法の分類器は，抽出された 1 文を入力として受け取り，そのうちのネガティブエンティティ以外の箇所，つまりネガティブエンティティの説明がポジティブである可能性を示すスコアを出力する．具体的には，ネガティブエンティティを<ENT>という記号に置換した文を受け取り，0 から 1 の間のスコアを出力する．

分類器は，BERT[3] と，1 層の feed-forward network から成る．入力文はトークン列に分割され，先頭に [CLS] トークン，末尾に [SEP] トークンが連結される．BERT の最終層の [CLS] トークンに対応する embedding を feed-forward network に入力しスコアを得る．

3　分類器の学習

分類器の学習データは次の通りに構築した．まず，英語 Wikipedia（2019 年 1 月 7 日版）全体から，粗いフィルタリングとして，セクション名に Advantage,Benefit などのポジティブなキーワードを含むセクションの文のみを残す．次に，それらの中から，何らかの事柄のポジティブな面について述べている文を人手で見つけ，ポジティブ説明とする．非ポジティブ説明については，上記以外の

3）　https://huggingface.co/bert-base-cased

病気	*Prosopamnesia*	*Hypopnea*
災害	*Frank Slide*	*Cataract*
事故	*Barnes rail crash*	*JAL Cargo Flight 8054*
経済	*Stagflation*	*2010s oil glut*
戦争	*Battle of Rayy*	*Acre War*

図 4　提案手法で抽出したネガティブエンティティの例

セクション中から，何らかの事柄のネガティブな面，またはポジティブでもネガティブでもない面について述べた文を人手で見つけた．最終的に 585 文からなる学習データができた．そのうち，266 文がポジティブ説明である．

上記データを用いて分類器を学習した．Epoch 数は 500, Loss 関数は Cross Entropy, 最適化には AdamW を用いた．

4　評価実験

提案手法により意外な恩恵候補文抽出し，分類器によりランキングした．その評価結果と抽出できた事例について述べる．

4.1　ネガティブエンティティ・意外な恩恵候補文の抽出

提案手法により，英語 Wikipedia から 6042 件のネガティブエンティティを抽出した．50 件をランダムに抽出して調査したところ 42 件が実際にネガティブエンティティと言えるものだった．図 4 に抽出したネガティブエンティティを例示する．病気や災害，事故，経済的困難，戦争等幅広いネガティブエンティティが抽出できた．

次に．提案手法を用いて実際に英語 Wikipedia から意外な恩恵事例の候補文を抽出した．その結果，Wikipedia 全体から 112 文を抽出した．

4.2　ベースライン手法

ベースラインとして，Wikipedia 全体からランダムに 300 文を抽出する方法を選んだ．これが最もシンプルな手法であること，意外な恩恵事例抽出の先行研究が著者らの知る限り存在しないこと，Wikipedia 全体に占める意外な恩恵事例の割合を推定したかったことがその理由である[4]．

4.3　評価方法

抽出した事例の適合率と抽出数を評価する．正解判定は著者のうち 2 名が別々に行う．判定作業

4）　他にベースラインとして，感情分析を応用した次の手法が考えられる: 主題にあたる句の sentiment がネガティブで，それ以外の箇所がポジティブである文を抽出する．紙面の都合上詳細は割愛するが，こうしたベースラインの精度は提案手法を大幅に下回るものだった．これは，感情分析におけるポジティブ/ネガティブが，映画や商品のレビューに代表されるように，主観的で個人的であるのに対し，本研究におけるポジティブ/ネガティブはより客観的，普遍的であり，両者の性質が大きく異なるためであると考えられる．

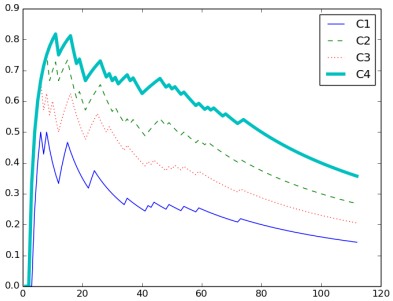

図 5　提案手法のランクごとの適合率

では，Yes（抽出した文は意外な恩恵事例である），No（そうした事例ではない），？（判断に迷う）の 3 つのラベルのいずれかを付与する形で行う．次の基準のいずれかで両者の判定をとりまとめ，最終的な判定とする．

1. 両者とも Yes としたものだけを正解とする．
2. 一人でも Yes としたものを正解とする．
3. 両者とも Yes か？としたものを正解とする．
4. 一人でも Yes か？としたものを正解とする．

公正さを担保するため，判定作業に先立ち，提案手法とベースラインの出力を 1 つにまとめ，シャッフルし，どの文がどの手法の出力かわからないようにする．

両者の判定の Cohen's kappa の値は 0.65 (substantial agreement) だった．

4.4　評価結果

実験の結果，ベースラインの抽出数と適合率はともに 0，つまりその出力 300 件はいずれも意外な恩恵事例とは言えないものだった．そのため，Wikipedia 全体に占める意外な恩恵事例の割合は推定できなかった．この点は今後の課題としたい．

提案手法の抽出数（スコア上位 k 件）とその適合率をプロットしたグラフを図 5 に挙げる．凡例の C1 から C4 は 4.3 節の基準 1 から 4 に対応する．条件 C4 では適合率 80%以上で 16 件の意外な恩恵事例を抽出した．Wikipedia 全体から抽出した事例候補全 112 件中 40 件が意外な恩恵事例と言えるものだった．

表 1 に提案手法の抽出結果の例を挙げる．誤りの要因は (1) ネガティブエンティティ抽出の誤り，つまり *Cause*/*Causes* セクションのある記事のタイトルだがネガティブとは言えないエンティティを誤って抽出した場合（表 1 の Rank 79 位），(2) 文抽出の誤り，つまり，*Gram positive bacteremia* 等のように，*Benefit*，*Advantage*，*Positive* のいずれかを含むセクションタイトルのセクションではあるが，当該エンティティに関するポジティブな内容ではないセクションから文抽出した場合（Rank 85 位），(3) 分類器の誤り，つまり当該エンティティの説明を誤ってポジティブと

Rank	判定	Negative <ENT>ity	抽出された文
13	Y/Y	*Flood*	*Bird populations may also profit from the boost in food production caused by <ENT>ing.*
14	Y/Y	*Invasive species*	*<ENT> have the potential to provide a suitable habitat or food source for other organisms.*
25	Y/?	*Population decline*	*And fertility moderately below replacement and <ENT> would maximize standards of living when the cost of providing capital for a growing labor force is taken into account.*
26	Y/N	*Grandiose delusions*	<ENT> frequently serve a very positive function for the person by sustaining or increasing their self-esteem.
49	Y/Y	*Short stature*	*<ENT> decreases the risk of venous insufficiency.*
79	N/N	*Home advantage*	*However, if the initial match is drawn (tied), <ENT> for the replay is given to the other team.*
85	N/N	*Bacteremia*	*Enterococci are an important cause of healthcare-associated <ENT>.*
87	N/N	*Population decline*	*More recently (2009–2017) Japan has experienced a higher growth of GDP per capita than the United States, even though its <ENT>d over that period.*

表 1 提案手法の抽出結果の例: ランク，判定者 2 名の判定，ネガティブエンティティ，抽出された文

判定した場合（87 位）の大きく 3 つに分類できる．

(1) ネガティブエンティティ抽出は，*Cause*/*Causes* セクションの有無という非常にナイーブな手がかりのみに基づいているが，そうした手がかりを利用して学習データを効率的に収集し，教師あり学習を用いることで精度向上が期待できる．(2) 文抽出も同様に，ポジティブな内容が書かれているセクションのタイトルかどうかの判定に教師あり学習を導入することで改善が期待できる．(3) 現状の分類器は 585 件のわずかなデータで学習されたものであり，学習データの拡充による精度向上の余地は大きいと考えられる．

5 関連研究

本研究は，不確実な未来に備えるための，あまり知られていない因果関係知識の獲得の研究と見なすことができ，未来シナリオ予測 [2, 4]，因果関係知識獲得 [3, 1]，トリビア知識獲得 [5] の 3 つが関連の深い領域となる．

未来シナリオ予測は，将来起こりうる重大な事態をその可能性がいかに低くても事前に把握する，という目的のもと，起こりうる未来のシナリオを予測，生成する研究である．因果関係知識獲得は，原因と結果を表すテキストを獲得する研究であり，その獲得結果は未来シナリオを自動生成する際に用いられることがある．こうした過去の研究と本研究の大きな違いは，意外な結末にフォーカスし，その自動抽出技術を開発したことにある．従来の未来シナリオ予測でも意外な結末を対象にしてはいたが，それは人手によるところが大きい．一方従来の因果関係知識獲得は常識的知識を対象としており，意外なものをシステマティックに獲得することは困難だった．トリビア知識獲得は，オバマ元大統領のグラミー賞受賞等，あまり知られていない意外な事実を自動獲得する研究である．しかし，意外な結末に関する事実を獲得する研究はこれまでなかった．

これまでになかった意外な恩恵事例抽出を可能にしたのは，1 節で述べた，**ネガティブエンティティのポジティブ説明を述べている文は，その意外な恩恵を述べている可能性が高い**，というアイデアである．

6 おわりに

本研究では意外な恩恵の事例を自動抽出する手法を提案した．精度，抽出数ともに改善の余地は大きいが，4.4 節で述べた通り，機械学習の導入や学習データの拡充で精度向上が大いに期待できる．

今後の方向性として意外な落とし穴事例抽出が挙げられる．つまり，善意の施策やサービス等が何らかの不運，不幸な結末につながる事例の抽出であり，例えば，食糧増産のための政策が逆に大飢饉につながった毛沢東による大躍進政策が挙げられる．

また，本研究では対象を Wikipedia に限定したが，Wikipedia 以外への拡張も今後の課題である．具体的には，作成したネガティブエンティティのリストと説明分類器を新聞記事，ブログ，SNS 等に適応し，意外な恩恵の文を抽出する．その際，提案手法では Wkipedia の記事タイトルとセクション名を利用したが，そのような Wikipedia 特有の構造が利用できない点が問題となる．

参考文献

[1] Chikara Hashimoto. Weakly supervised multilingual causality extraction from Wikipedia. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2988–2999, 2019.

[2] Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 987–997, 2014.

[3] Canasai Kruengkrai, Kentaro Torisawa, Chikara Hashimoto, Julien Kloetzer, Jong-Hoon Oh, and Masahiro Tanaka. Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3466–3473, 2017.

[4] Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. Learning causality for news events prediction. In *Proceed-*

ings of the 2012 World Wide Web Conference (WWW), pp. 909–918, 2012.

[5] David Tsurel, Dan Pelleg, Ido Guy, and Dafna Shahaf. Fun facts: Automatic trivia fact extraction from wikipedia. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 345–354, 2017.