利用物に関する知識のコーパスアノテーション

上村 明衣　　折田 奈甫　　岡崎 直観　　乾 健太郎
東北大学 工学部／大学院情報科学研究科
{mei.uemura, naho, okazaki, inui}@ecei.tohoku.ac.jp

1 はじめに

昨今様々な大規模知識ベースが利用可能となっている．従来の知識ベースはエンティティに関する情報とエンティティ間の関係を中心に構築されているが，近年，事象（出来事/コト/*event*）中心の知識ベース構築の重要性が指摘されている [6, 3].

既存の知識ベースにおいても事象的知識を含むものはあるが，構造化されていないか単純な関係に限定されているなど小規模である．例えば，常識的知識を主に人手で集めている ConceptNet[7] では，単語や述語を含む短い句の間の関係知識を収集しているが，常識的知識が中心のため，専門的な物の知識に乏しい．Freebase[*1]などのエンティティ中心の知識ベースでは，事象知識は構造化されておらず，Wikipedia から引用されたテキストが未整理のまま記述されているのみである．

事象に関する知識獲得には，いわゆるスクリプト知識の獲得 [1] と，項と項の関係インスタンスを獲得する OpenIE[2] のような研究がある．スクリプト知識の獲得では，文書内の複数の事象の系列から一般的な事象の繋がりパターンを帰納する．この一般化のためには同程度の粒度の記述が複数必要だが獲得が困難なため，密な知識はこれまでに獲得できていない．一方で，⟨*X Cause Y*⟩のような言語パターンを用いて，関係インスタンスにより事象間の関係を獲得する方法がある [2]．このような獲得手法では，関係インスタンスは一つの文の中での項と項の関係に限定される典型的なパターンであり，多くの場合名詞と名詞の関係である．しかし実際のテキストを観察すると，事象間の関係は，文を越えて，典型的な言語パターンに当てはまらない名詞以外の多様な表現の関係として存在している．

例えば，(1a) では *facial prosthetic* の効果である *altering the curve of a cheek or nose* が，文を越えて記述されている．(1b) は表面的には *is-A* 関係であるが，*moisturizer* という語の意味から，*CC cream* を使用すれば保湿されるという因果関係があると人は解釈する．同様に，(1c) の *add depth and dimension to one's eyes* も，*eye shadow* を用いた時に起こる因果関係的な事象であるが，関係インスタンスでは ⟨*Eye shadow Add depth ...*⟩ と表現され，事象間の関係を意味的に十分に捉えていない．また，(1d) の *in portable coolers* のような事象の発生に関する条件を表すものや（*ice pack* を *portable coolers* に入れることによって *keep food cool* という事象が生じる），(1e) の *more effective* のように事象の効果の程度を表すもの（*alcohol-based hand sanitizers* は *killing microorganisms* という事象において *more effective* である），(1f) の *have not been proven* のように事象の信頼度を表すような表現もある．

(1) a. A facial prosthetic or facial prosthesis is ... Effects can be as subtle as altering the curve of a cheek or nose, (Wikipedia: facial prosthetic)

b. BB cream and CC cream are both tinted moisturizers... (Wikipedia: CC cream)

c. Eye shadow can add depth and dimension to one's eyes. (Wikipedia: eye shadow)

d. These packs (ice packs) are commonly used to keep food cool in portable coolers, ... (Wikipedia: ice pack)

e. Alcohol-based hand sanitizers are more effective at killing microorganisms... (Wikipedia: hand sanitizers)

f. anti-aging products ... have not been proven to give lasting or major positive effects (Wikipedia: anti-aging cream)

[*1] https://www.freebase.com

事象に関する密な知識を得るためには，このような情報を獲得する必要がある．このためには，文を越えた典型的なパターンには当てはまらない，多様な表現による関係を獲得するモデルの構築が必要である．本研究は，事象に関する知識獲得の問題を，テキストからの関係抽出問題と捉え直し，このためのタスク仕様の設計とベンチマークデータ開発を試みた．

本稿では，(1) のような事象に関する情報に対し意味関係タグを定義しアノテーションを行う．アノテーションには短く簡潔に記述されている英語の Wikipedia の定義文を用いる．Wikipedia の記事は，健康・美容分野において利用される物（以降「利用物」）に限定し，物を利用した時の効果・影響という事象知識と，これに関連する条件，度合い，確信度などの知識に焦点を当てる[*2]．物を利用した時にどういう効果・影響があるかという事象間関係知識に限定することにより複雑さを軽減し，事象に付随する知識の検討を容易にした．

物を利用した時の効果・影響やそれに関する知識は，実用面においても必要性がある．医学薬学分野におけるテキストマイニングでは，薬を投与した時の副作用などを文書から自動で獲得する研究がされている [4]．本研究はより一般的な健康・美容ドメインにおける物の効果・影響に着目しているため，このような知識は自然言語処理技術において有用であるだけでなく，様々な分野での実用が期待できる．

2 アノテーションの仕様

2.1 意味関係ラベル

著者らは英語 Wikipedia の利用物についての記事を 25 件ほど観察し，最も多く分布している情報は利用物の効果についてであると仮定した．この仮定に基づき，利用物の効果とそれに関連する情報についてパイロット的にアノテーションを行い，意味ラベルを暫定的に定義した．

アノテーションには，英語 Wikipedia の健康・美容分野の利用物に関する記事の冒頭の定義文を用いた．定義文は記事全体の簡潔な要約となっているため，少ないデータからある程度の質の良い知識が獲得できると期待される．25 記事にパイロットアノテーションを行い，利用物と定義文中に記述されている情報との関係を表す 12 種類のラベルを暫定的に定義した（表 1 参照）．

ほとんどの利用物は，使われ方，使われる場所，時間，利用者によって異なる効果を持つ．このため，Means of Use, Location, Time, User などの利用物の効果が生じるための条件にラベルを付与する．例えば，*bolster*（長枕）についての Wikipedia の定義文には，“*a bolster is usually placed at the head of one's bed and functions as head or lower back support*” という記述があるが，*placed at the head of one's bed* は *head or lower back support* という効果を得るための条件であり，Means of use というラベルを付与する[*3]．

2.2 アノテーション実験

アノテーション仕様作成時に使用した Wikipedia の記事とは別の 100 記事の定義文に対し，アノテーションガイドラインの説明を受けた英語話者 2 名がアノテーションを行った．

Wikipedia には利用物というカテゴリは存在しないため，amazon の健康・美容に関するカテゴリリスト[*4]などを参考にした．記事の内容は確認せずに利用物であるかを人手で選択し，文量のばらつきを抑えるために各記事 5 文目までを対象テキストとした．アノテーションツール brat[8] によるアノテーション例を図 1 に示す．

3 分析

作業者間の一致について，数と全体に占める割合を表 2 に示した．2 名の作業者がつけたラベルの数に大差はなかったが，ラベルごとのインスタンス数に偏りがあった．特に Target が全インスタンスのうち 40~50 %を占めており，他のラベルに比べて完全一致が多く，全体としての完全一致率を上げるため，他のラベルの一致度の分析が困難になる．そのため，表 3 には Target を考慮しない作業者間一致の結果も示している．Target を考慮しない場合， 完全に一致するものが 31.5 %，部分的に一致するものを含めると 60.6 %となった．以下，部分一致と完全不一致についての分析を示す．

3.1 部分一致の分析

部分一致はラベルが異なる場合とセグメントの範囲が異なる場合の 2 種類に分類することができる．

[*2] 物の利用という観点では，利用表現の獲得研究がある [5, 9, 10]．これらの研究は，物の利用，用途，準備表現などの獲得が中心となっており，物を利用した時の効果・影響とそれに関連する知識に焦点を当てている本研究とは相補的な関係にある．

[*3] 本研究では利用物の使い方と効果を区別する．Effect は「使った結果起こること」を表し，Means of use は「Effect を起こすための使い方」を表す．

[*4] https://www.amazon.com/gp/help/customer/display.html

表 1: 利用物とその効果に関する意味ラベル

ラベル	定義	例文
TARGET	利用物を指示する．別名や代名詞も含む．	BB cream stands for blemish balm, blemsh base (Wikipedia: BB cream)
EFFECT	利用物の効果を表す．期待されない効果も含む．	to decorate and protect the nail plates (Wikipedia: nail polish)
NULL EFFECT	ある特定の EFFECT について効果がないという情報を表す．	The myth of its effectiveness (Wikipedia: bear's grease)
DEGREE OF EFFECT	ある特定の EFFECT の程度を表す．	a poor substitute for protective clothing (Wikipedia: barrier cream)
CERTAINTY OF EFFECT	ある特定の EFFECT の確信度/信頼性を表す．	a have not been proven to give lasting or major positive effects (Wikipedia: anti-aging cream)
MEANS OF USE	利用物の使い方を表す．	is applied around the contours of the eye(s) (Wikipedia: eye liner)
COMPOSED OF	利用物を構成している要素を表す．	consisting mainly of triglycerides (Wikipedia: egg oil)
PART OF	利用物を含む/構成素とするものを表す．	Cinnamon is a spice obrained from the inner bark (Wikipedia: cinnamon)
LOCATION	利用物が使われる場所を表す．	It is often used ... where sunlight can impair seeing (Wikipedia: eye black)
TIME	利用物を使用する時間を表す．	soon after birth (Wikipedia: kohl(cosmetics))
USER	利用物を使う人/効果を受ける人を表す．	mothers would apply kohl to their infants' eyes (Wikipedia: kohl(cosmetics))
VERSION	利用物の別のバージョンを表す．	It is distributed as a liquid or a soft solid (Wikipedia: lip gloss)

表 2: 作業者間一致

	Target あり	Target なし
完全一致	1116 (55.9%)	387 (31.5%)
部分的に重なる	10 (0.5%)	10 (0.8%)
片方を真に含む	57 (2.9%)	54 (4.4%)
ラベルのみ違う	63 (3.2%)	64 (5.1%)
範囲の最初だけ違う	92 (4.6%)	83 (6.8%)
範囲の最後だけ違う	174 (8.7%)	147 (12.0%)
完全不一致	484 (24.3%)	484 (39.4%)

ラベルが異なる例は全体の 5.1 %で，EFFECT と MEANS OF USE の間での揺れが多い．例えば (2) の *hair and skin care* や *medicine* は，効果を表しているように見えるが，使った結果何が起こるか明記されていない．作業者が，薬を使うと良い効果があるという前提知識を持っているため EFFECT ラベルを付ける傾向がある．

(2) a. It is used for topical applications such as hair and skin care. (Wikipedia: egg oil)

b. These herbal distillates have uses as flavorings, medicine, and cosmetics. (Wikipedia: herbal distillate)

範囲が異なる例の合計は 23.9 %となった．これらは助動詞，前置詞，冠詞などの機能語を含むかどうかの揺れが多い．

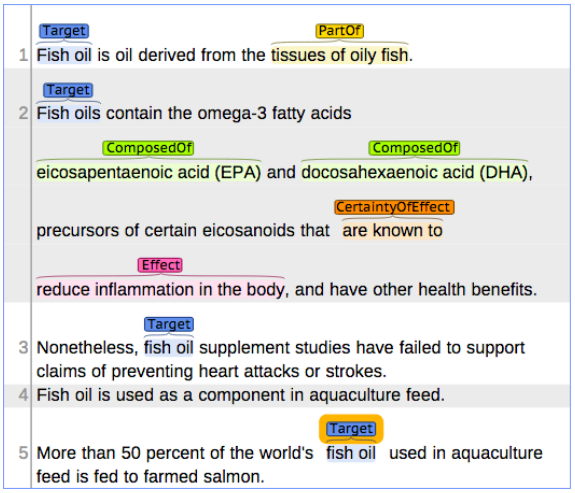

図 1: アノテーション例（Wikipedia: Fish oil）

3.2 完全不一致の分析

完全不一致は 39.4 %で，両方がラベルをつけたが選んだラベルとセグメント範囲がそれぞれ違ったものと，片方がラベルをつけなかったものの 2 種類に分けられる．後者の例を以下に示す．

例 (3) の *health benefits* は *green tea* の EFFECT としてラベル付けされた例だが，*health benefits* は使って何が起こるかを具体的に表していないため，一方の作業者は

表 3: 付与されたラベルの数

ラベル	作業者 A	作業者 B
Target	430	466
Effect	152	189
Certainty of Effect	4	19
Degree of Effect	10	13
Null Effect	5	0
Means of Use	79	59
Composed of	127	112
Part of	27	14
Location	27	26
Time	12	16
User	22	25
Version	56	105

ラベルを付与しなかった．

(3) Numerous claims have been made for the health benefits of green tea. (Wikipedia: green tea)

このように，作業者間での前提知識や明示されていない取り決めの解釈の違いによって不一致が生じることがあるものの，現行の意味関係ラベルの仕様でも作業者間で一定の共有が可能であることが確かめられた．

3.3 付与されたラベルの分布の分析

対象とする利用物そのものを指示する Target を除けば，Effect ラベルの数がどちらの作業者においても最も多く，アノテーション前の想定と一致した．これは効果に関する情報が一般的に必要とされている事を示唆しており，物の利用と効果という事象関係に付随する情報を構造化するという本研究の妥当性が示された．

また，Effect や Means of Use が付与されたセグメントには，*moisturizer*，*tenderizer*，*cold compress* などの事象が名詞化された表現が散見された．これらの典型的な言語パターンに当てはまらない事象表現をどう扱うかは，今後の課題である．

Certainty of Effect, Degree of Effect, Null Effect の 3 つのラベルは付与されたインスタンス数が少ない．これは，アノテーション対象を Wikipedia の定義文に限定したためであると考える．今後は定義文以外の記述についてもアノテーションを検討する．

ラベルが全く付与されなかった箇所にも傾向が観察された．利用物の起源や歴史に関する記述など効果に関係のないものや，利用物の製作過程に関する情報が多い．利用物の製作過程については，効果との関連性を考慮し意味関係ラベルに含めるかを検討したい．

4 結論

事象に関する密な知識をテキストから獲得するためには，文を越えた，典型的なパターンには当てはまらない多様な表現による関係を獲得するモデルが必要である．本研究では，事象に関する知識獲得の問題を，テキストからの関係抽出問題と捉え直し，このためのタスク仕様の設計とベンチマークデータ開発を試みた．

今回のアノテーションでは，同一文書内のラベル間の関係を扱っていないが，何がどの効果の条件となっているのかという依存構造があるため，現在ラベル間の関係アノテーションをしている．今後は，作業者間の相違を解消したゴールドデータを作成・公開し，アノテーションの規模を拡大する．また，作成したデータを基にこのような知識の自動獲得手法の研究に繋げたい．

謝辞 本研究は，JST 戦略的創造研究推進事業 CREST の一環として行われた．また文部科学省科研費（15H05318）から部分的な支援を受けて行われた．

参考文献

[1] Nathanael Chambers and Daniel Jurafsky. Unsupervised learning of narrative event chains. In *ACL*, pp. 789–797, 2008.
[2] Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam Mausam. Open Information Extraction: The second generation. In *IJCAI*, pp. 3–10, 2011.
[3] Michael Färber, Basil Ell, Carsten Menne, and Achim Rettinger. A comparative survey of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web*, Vol. 1, pp. 1–26, 2015.
[4] Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, Vol. 45, No. 5, pp. 885–892, 2012.
[5] James Pustejovsky. The generative lexicon. *Computational linguistics*, Vol. 17, No. 4, pp. 409–441, 1991.
[6] Roxane Segers, Piek Vossen, Marco Rospocher, Luciano Serafini, Egoitz Laparra, and German Rigau. ESO: a frame based Ontology for Events and implied Situations. *Proceedings of Maplex2015*, 2015.
[7] Robert Speer and Catherine Havasi. Representing general relational knowledge in ConceptNet 5. In *LREC*, pp. 3679–3686, 2012.
[8] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. Brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th EACL*, pp. 102–107, 2012.
[9] Kentaro Torisawa. Automatic acquisition of expressions representing preparation and utilization of an object. *Proc. 5th RANLP*, pp. 556–560, 2005.
[10] 常吉高弘, 小町守, 松本裕治. 生成語彙論に基づく日本語の特質構造のランキング学習による自動獲得. 言語処理学会第 18 回年次大会, 2012.