B2TB2036

卒業論文

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

2016 年 3 月 31 日

東北大学
工学部 情報知能システム総合学科

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

内容梗概

自然言語文を理解するためにはモノに関する知識が必要である．モノに関する知識は主にモノの利用や効果に関する情報が多く，イベント知識で表されることが多い．しかしイベント知識を獲得している既存研究はエンティティ同士の簡単な関係にのみ着目しているものが多い．また，現在利用可能な大規模知識ベースでもモノの利用・効果に関する情報は少なく，十分な知識獲得がなされているとは言えない．本研究では様々な表現で表されるモノの利用に関する知識の獲得を目指し，テキスト中に現れる利用・効果に関する表現のための意味関係ラベルを設計しベンチマークデータを作成・CRF で自動認識を行う．

キーワード

keyword1, keyword2

*東北大学 工学部 情報知能システム総合学科 卒業論文, B2TB2036, 2016 年 3 月 31 日.

Contents

1 序論

自然言語文を理解するためにはモノに関する知識が必要である．以下の Winograd Schema Challenge[1] と Choice of plausible Alternatives[2] の例を考える．

(1) a. The policeman finally caught the criminal, because he used pepper spray and handcuffs.

b.
- Promise: The man broke his toe. What was the CAUSE of this?
- Alternative 1: He got a hole in his sock.
- Alternative 2: He dropped a hummer on his foot.

人間は (1a) の代名詞 *he* は *policeman* を指しており，(1b) の Alternative 2 は Promise と強い因果関係があると判断できるが，人間のような常識的知識を持たない計算機には難しい．例えば (1a) では，*pepper spray*（催涙スプレー）や *handcuffs*（手錠）が逮捕するために使われるという知識が必要となる．また (1b) では，*hummer* が衝撃を与えたりものを壊す効果があるという知識があればより妥当な選択肢を選ぶことができる．こうしたモノに関する知識が文章理解に役立つと考えられる．

このようなモノの利用や効果についての情報を集めた資源は照応解析や含意関係認識など，さまざまなタスクに利用することができる．また，医学・薬学系の自然言語処理分野では，薬を利用したときの効果や副作用というドメインに特化した知識獲得が行われている．このようにモノの利用に関する知識は様々な分野で利用価値がある．

本研究ではこうしたイベント知識獲得のタスクを設計し，自動認識のためのベンチマークコーパスを作成した．ここではまず，例 (2) で表されるような利用や効果についての情報が多く記述される健康ドメインのモノに着目した．

(2) a. Fish-oils ... are known to reduce inflammation n the body, ... (*fish oil*)

b. Alcohol-based hand sanitizers are more effective at killing microoganisms than soaps... (*hand sanitizers*)

c. BB cream and CC cream are both tinted moisturizers ... (*CC cream*)

d. ... the American Dental Association reports that up to 80% of plaque can be eliminated with this method. (*dental floss*)

人間はこうした記述からこれらのモノを利用したときの効果を簡単に理解することができる．例えば *fish-oil* は炎症を抑え (2a)，*hand sanitizer* は雑菌などの微生物を殺す働きがある (2b)．*BB cream* は肌に着色したり，肌を保湿する効果があり (2c)，*dental floss* はプラークを除去することができる (2d)．こうした情報は動詞句 (2a)，動名詞 (2b)，名詞句 (2c)，節 (2d) など様々な形で表されるので，計算機による自動認識は難しいタスクである． 2 章では既存研究でテキストからの十分な知識獲得がなされていないことを示す．

このようなモノの利用に関する情報は主にイベントで表現される．近年大規模な知識ベースが多く利用できるようになってきているが，そのほとんどが人物や組織名などのエンティティと，エンティティ間の関係（*IsPresidentOf* など）に着目しているためイベントで表現される知識は十分に集められていない． 2 章では近年代表的な知識ベースに収録されている知識について詳しく述べる．

本研究では，(i) モノの利用に関する知識を表す意味関係ラベルを定義しタスクを設計した．また，(ii) 定義したラベルを用いて実券的にベンチマークコーパスを作成した．(iii) 作成したコーパスを用いて試験的に CRF (conditional random field) を用いて自動認識を行った．

2　関連研究

2.1　目的役割の獲得

Pustejovsky[3] は生成語彙論において語の意味を表す特質構造を定義した．4 つの特質構造のうち，主体役割 (agentive role) と目的役割 (telic role) は本研究が着目するモノの利用や効果の表現に関連している．主体役割は対象の発生や起源に関する事例を表す．(例えば，*book* の主体役割は *write, public* など．) 目的役割は，対象の持つ機能や目的を表す．(例えば，*book* の目的役割は *read*.) この理論に基づいて，これらの表現をテキストから自動抽出する研究がなされてきた []. また，この目的役割に関連して，準備（*preparation*）・用途（*urilization*）表現を獲得する研究がある．[4] 準備表現はあるモノを使うための手順を表し，用途表現は「モノを使う」の言い換えになる表現を指す．

これらの既存研究はモノの知識を獲得することに焦点を当てているが，本研究とはモノの効果に着目するという点で異なる．例えば，*book* を利用したときの効果は *write* や *read* よりも *learn* である．中には，例 (3) のように効果と用途表現の両方になる表現も存在する．これに対し，本研究ではモノの効果と使い方を分けて定義している（ 3.1 章）．このことから，本研究は既存研究に対して相補的な立場であると言える．

(3) a. It is used for hair and skin care. (*Egg oil*)

2.2 イベント知識

例 (2) で示したように，モノを使うことによる効果は主にイベントで表現される．WordNet[5] や FrameNet[] などの代表的な知識ベースは人手で整備されているため，カバレッジが広くない．さらに，近年大規模な知識ベースの需要が高まっているが，ほとんどがエンティティとそれらの間の関係に着目しているので，イベント知識は十分に整理されていない．我々は代表的な知識ベースである ConceptNet[6] と Freebase[7] におけるモノの利用に関する知識のカバレッジを調べた．

試験的なドメインとして，Wikipedia にエントリを持つ健康分野の利用物 100 個を選択した．次に，それぞれの知識ベースで設定されている関係のうち，利用物の効果を表す関係を人手で収集した．これらの関係を表??と表??の一番左の行に示す．それぞれの関係を持つインスタンスの数を数え，表に示した．

どちらの知識ベースも効果を表すインスタンスの数は少ないが，Freebase と ConceptNet に収録される情報には質的な差があった．Freebase に収録されている情報のほとんどは薬を使ったときの効果を表すものであった．一方で，ConceptNet は常識的知識を収集しているので，より一般的な事象についての記述が多い．これらの結果から，既存の知識ベースの利用物の効果に関する情報は量的・質的に不十分であると言える．

最近では，イベント中心の知識を獲得する研究が行われているが，これらの研究は時事的なイベント（経済危機，大統領選挙，FIFA ワールドカップなど）とそれに関連する参加者・日付・場所などの知識の獲得に焦点を置いているため，モノの利用や効果といったイベント知識の獲得はまだ不十分である．本研究は，より精緻なイベント知識獲得の足がかりとなることを目指す．

2.3 イベント抽出

イベント知識の獲得の研究は以前からさまざまな手法で行われてきた．近年の大規模な知識ベースに収録されてる関係は数が少なく限られているのに対して，TextRunner[] やReVerbを用いたOpenIE[8]システムは言語パターンを用いてWebから大量の関係を抽出している．それぞれの関係インスタンスは，$\langle Tramp,\ lost,\ the\ election \rangle$ のように関係を表すフレーズとその項で表される．こういったシステムは様々な自然言語処理のタスクに役立っているが，このような関係インスタンスはテキストに現れた表現を集めているにすぎず，構造化されていないものが多い．利用物の効果を表す表現を精緻に捉えるためには，文の表層に現れる表現だけでなく単語の意味を考える必要がある．例えば，$\langle BBcream, IsA, tintedmoisturizer \rangle$ という表現では，*tinted moisturizer* という名詞句に肌を着色し保湿するという効果が現れている．このような表現を獲得するためには，言語パターンのような表層的な特徴を手がかりとする手法はあまり有用ではない．

また，医学・薬学系の自然言語処理分野では，薬を使った時の効果や副作用など，ドメインに特化したイベント知識獲得が行われている．

3　コーパス作成

3.1　意味関係ラベルの設計

テキスト中に現れる利用物の効果とそれに関する情報を表す 12 個のラベルを定義した．表 1 に示す．以下の (4) はこれらのラベルを *hand sanitizer* についてのテキスト中のセグメントに対して適応した例である．

(4) a. Alcohok-based hand sanitizers are more effective at killing microorganisms than soaps... (*hand sanitizer*)

b. alcohol-based: Version

c. hand sanitiers: Target

d. more effective: Degree of Effect

e. killing microoorganisms: Effect

12 種類のラベルのうち，Effect と Means of Use は利用物に関する情報のうち最も重要である．それ以外のラベルは Effect の補助的な情報を表す．まず，Effect の補助的な情報を表すものとして，Null Effect，Degree of Effect，Certainty of Effect をそれぞれ定義した．Means of Use，Composed of，Part of，Location，Time，User は Effect が発生する為の条件を表す．例えば，Wikipedia の *lip stain*（口紅）に関する *it can dry the lips and is not recommended for winter* という文の中には，*dry the lips* という Effect とその時間的条件（Time）となる *for winter* という記述がある．

Table 1: 利用物の効果についての知識をとらえるための意味関係ラベル

ラベル	定義	例
TARGET	利用物を指示する．別名や代名詞も含む．	BB cream stands for blemish balm, blemish 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 の効果がないという情報を表す．	The myth of its effectiveness (Wikipedia: beer'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/信頼度を表す．	have not been proven to given 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	利用物を含むものを表す．	Cinnamon is a spice obtained 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．	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 gross)

3.2 データ規模

本研究では，健康に関する利用物 100 個について書かれている英語版の Wikipedia 記事を収集した．アノテーションにはそれぞれの記事の導入部のうち，最初から 5 文目までを用いた．Wikipedia 記事本文ではなく導入部を使用し分量を制限することで記事ごとの文章量や情報量のばらつきを抑えた．アノテーションに用いたテキストは 100 記事で 384 文となった．

3.3 アノテーション

定義した意味関係ラベルを用いて，アノテーションを行った．ラベルとアノテーション方法について説明した英語話者 2 名の作業者が，アノテーションツール

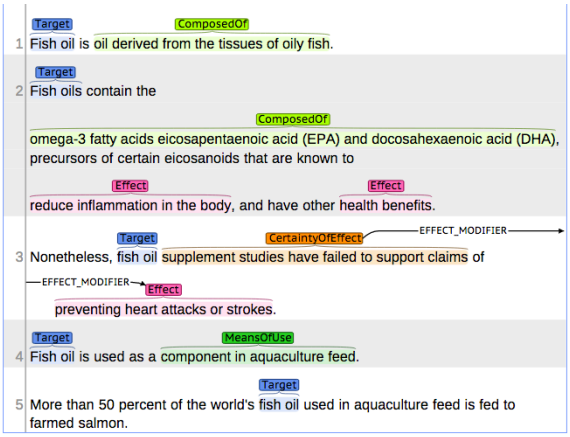

Figure 1: *fish oil* の記事の brat によるアノテーション

brat[] を用いて行った．図 1 にアノテーションの様子を示す．

3.3.1 作業者間一致

コーパスの品質を評価するため，作業者間での一致を測った．表 2 に結果を示す．より正確な評価を行うため，TARGET ラベルが付与されたセグメントを考慮に入れる場合とそうでない場合の 2 通りで一致率を測った．この結果より，このタスクは文章を深く理解しなければならない難しいタスクであるにもかかわらず，セグメント範囲に揺れはあるがラベルの種類とラベルを付与する場所については高い一致が得られた．範囲の揺れはほとんどが助動詞や冠詞などの機能語によるものである．

アノテーションにおいて観察された問題として，複数のラベルに該当する表現

Table 2: 作業者間一致

一致の度合い	TARGET を含む	TARGET を含まない
ラベルとセグメント範囲が一致	1248 (59.5%)	458 (36.1%)
ラベル一致，範囲部分一致	325 (15.5%)	288 (22.7%)
ラベル不一致，範囲部分一致	159 (7.6%)	159 (12.5%)
ラベル不一致，範囲一致	81 (3.9%)	81 (6.4%)
片方の作業者がラベルを付与しなかった	283 (13.5%)	283 (22.3%)
F (lenient agree)	77.2%	52.5%

での揺れが多数発生することがあげられる．例えば，以下の例 (5) では *hair and skin care* というセグメントにおいてラベルの不一致が発生した．

(5) It is used for topical applications such as hair and skin care. (*egg oil*)

片方の作業者はこのセグメントの Effect ラベルを付与したが，もう片方の作業者は Means of Use のラベルをつけた． **??**章では，Effect は利用物を使った時に起こる効果，Means of Use は「利用物を使う」の言い換えになる表現と定義したが，この例のようにどちらにもなる表現では同様のラベル不一致が起こった．

3.3.2 ベンチマークデータの統計

表 3 に，作業者 2 名によって作成したベンチマークデータの統計を示す．利用物を表す Target，Version は代名詞なども含むため数が多い．それ以外では，効果・利用を表す Effect と Means of Use が最も多く，利用物についての重要な記述であるこれらの情報がこのコーパスから獲得できていることが示された．一方で，Effect の補助的な情報を表す Null Effect，Degree of Effect，Certainty of Effect は Wikipedia の導入部にはあまり現れないため，数が極端に少なくなっていると考えられる．

Table 3: ラベルの統計

ラベル	作業者 A	作業者 B
Target	444	462
Effect	190	189
Certainty of Effect	32	19
Degree of Effect	13	19
Null Effect	0	0
Means of Use	124	59
Composed of	98	112
Part of	12	14
Location	12	26
Time	11	16
User	19	25
Version	100	103
Total	1060	1038

4 自動認識

3 章で作成したデータを用いて，条件付き確率場（Conditional Random Field, CRF）を用いて自動認識を行った．試験的に行うため，作成したデータのうち数が多かった Effect，Means of Use，Composed of，Version の 4 つのラベルが付与されたセグメントを認識の対象とした．

4.1 実験設定

この実験では，記事タイトル（利用物の名前）が与えられた状態で文中のセグメントに対して正しいラベルを付与する系列ラベリングを行なう．データは 3 章で作成した 384 文を用い，10 分割交差検定を行った．今回はセグメントの範囲が前後にずれていても一致とみなすため，トークンごとにラベルの一致を集計する評価指標を用いた．

$$Precision = \frac{\text{システムが正しくラベルを付与したトークン数}}{\text{システムがラベルを付与したトークン数}}$$

Table 4: 学習に用いた素性

素性名	定義	例
Token	今見ている単語	Perfume
Lower	単語の小文字化	perfume
POS	単語の品詞	NNP
Target	利用物を表す単語であるか	T
Disease	単語が病名リストに入っているか	F

$$Recall = \frac{\text{システムが正しくラベルを付与したトークン数}}{\text{正解データ中のトークン数}}$$

$$Fscore = \frac{2 * Precision * Recall}{Precision + Recall}$$

4.2 素性

学習に用いた素性は表 4 [1] に示した 5 種類の素性の組み合わせである．テキストのタイトルを利用するため，今注目している単語が利用物を表す単語出会った場合に発火する素性 Target を定義した．また，Effect ラベルや Means of Use ラベルが付与されたセグメントには病名が含まれることがある．これを利用するため，今見ている単語が病名リスト [2] に含まれていれば発火する素性 Disease を定義した．

文脈として前後 3 単語を考慮し，作られた素性は 10 分割交差検定の平均で約 250,000 個であった．

素性の組み合わせは以下の 5 種類である．

(6) a. Token のみ

b. Target + Lower

c. Target + POS

d. POS + Lower

e. Disease + POS

この素性を a から e まで順に増やしたときの精度を表 5 に示す．

[1] 例には今注目している単語が *Perfume* であったときの素性の値を示している

[2] Freebase に記載される病名をリストにしたもの

Table 5: 素性と精度

素性	F 値（マクロ平均）(%)	F 値（マイクロ平均）(%)
a	38.9	71.7
a+b	37.7	71.6
a+b+c	42.8	73.3
a+b+c+d	42.7	73.2
a+b+c+d+e	44.0	73.1

この結果より，素性を増やすと精度は向上した．また，ドメインに特化した素性である Disease(6e) を増やしたとき，マクロ平均で F 値が大きく改善したことから，今後ドメインを拡大して同様に自動認識を行なう際にもある程度このような素性が必要であると考えられる．

自動認識で正解した結果を分析すると，*is used for* など，効果や利用を表すフレーズが使われる表現では正しくラベルが付与されていた．また，(7) のように，名詞が列挙されるフレーズは VERSION である確率が高いということが学習されていた．

(7) Such preparations are available in the form of tablets, capsules, pastilles, powders, ... (*Multivitamin*)

一方で，(8a) のように，*is used for* のようなパターンに当てはまらず直接効果が記述される表現には正しくラベルを付与することができなかった．また，(8b) のように，パターンに当てはまらず，*herbal* の「薬用の」という意味を知らなければ EFFECT であると判断できないものも観察された．また，人間同士でラベルの不一致が起こった表現は，自動認識でも正しいラベルを付与することは難しい．

(8) a. Powder tones the face and gives an even apperarance. (*Face powder*)

b. Hibiscus tea is a herbal tea. (*Hibiscus tea*)

5　結論

イベントに関する密な知識をテキストから獲得するためには，単純な言語パターンに当てはまる表現以外の様々な表現を考慮した獲得モデルが必要である．本研究では，イベントに関する知識獲得をテキストからの関係抽出問題として定義し，タスク仕様の設計を行い意味関係ラベルを定義した．また，定義した意味関係ラベルを用いてベンチマークコーパスを作成し，その品質が十分であることを確認した．作成したコーパスデータを用いて試験的に CRF で自動認識を行った．今後は自動獲得にむけて，データを増やしコーパスの品質向上を試みる．また，自動認識モデルを改善し，大規模コーパスに適応することで知識獲得を行なう．

謝辞

本研究を進めるにあたり，ご指導頂いた乾健太郎教授，岡崎直観准教授に感謝いたします．研究活動，本論文の執筆全般にわたり，直接のご指導，適切な助言をくださった折田奈甫研究特任助教に感謝いたします．日常の議論や研究会で様々な知恵や示唆をくださった乾・岡崎研究室の皆様に感謝いたします．

References

[1] Altaf Rahman and Vincent Ng. Resolving complex cases of definite pronouns: the winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 777–789. Association for Computational Linguistics, 2012.

[2] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, 2011.

[3] James Pustejovsky. The generative lexicon. *Computational linguistics*, Vol. 17, No. 4, pp. 409–441, 1991.

[4] Kentaro Torisawa. Automatic acquisition of expressions representing preparation and utilization of an object. *Proc. 5th RANLP*, pp. 556–560, 2005.

[5] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, Vol. 38, No. 11, pp. 39–41, 1995.

[6] Robert Speer and Catherine Havasi. Representing general relational knowledge in ConceptNet 5. In *LREC*, pp. 3679–3686, 2012.

[7] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 1247–1250. ACM, 2008.

[8] Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam Mausam. Open Information Extraction: The second generation. In *IJCAI*, pp. 3–10, 2011.