常識表現となり得る用言の自動選定の検討

真嘉比 愛　　山本 和英
長岡技術科学大学　電気系
{makabi, yamamoto}@jnlp.org

1　はじめに

言葉の意味を理解する計算機を実現するためには，言語の文法的理解とともに，大量の常識 (世界知識) が必要となる．そのため，それらの常識を集め，自然言語処理で利用可能な常識知識ベースを構築する研究が注目されている．

本研究では，自然言語処理の意味解析に利用できる常識知識ベースを構築するために，文中で名詞と共起する用言 (動詞，形容詞，サ変名詞) の集合はその名詞の持つ常識であると仮定し，これらの常識を自動的に獲得する手法を提案することを目的とする．たとえば，名詞 “いぬ” と文中で共起する “ほえる”，“散歩”，といった用言の集合は，名詞 “いぬ” が持つ常識である．

本稿では，常識知識ベースの構築にあたり，常識として適切な用言の選定方法について述べる．

2　関連研究

Ahrens et al.[1] は既存の上位オントロジー (=常識オントロジー) である SUMO (Suggested Upper Merged Ontology) [4] をベースとして，SUMO 中で定義される概念をテキストコーパス中の語へマッピングする手法を提案している．また Niles et al.[5] や，Jan et al.[2] は，SUMO と既存の語彙資源 (WordNet[1], FrameNet[2]) を組み合わせることで，常識を組み込んだ汎用的な知識ベースの構築を試みている．しかし上位オントロジーを使ったこれらの研究は，厳密に定義された常識を利用できる反面，上位オントロジー上で定義される常識表現と実際の語彙表現との対応が取れないことも多く，常識表現を適用可能な語彙の数が少ないという問題がある．

これに対し，MIT メディアラボが構築している常識知識ベース ConceptNet[3]は，各 concept に対し様々な関係 (e.g. IsA,CapableOf,RelatedTo) で結ばれる語や文を付与することで，各々の concept が持つ常識を定義している．ConceptNet は自然言語で常識定義を行なっているため，上位オントロジーと比較して自然言語処理のタスクに適応しやすいというメリットがある．しかし各 concept が持つ常識の大半が人手で集められたものであり，常識の網羅性が低いという問題がある．ConceptNet を自動的に拡張しようとする研究もあるが，十分な拡張には至っていない [6].

そこで本研究では，自然言語処理で利用可能な常識知識ベースを構築するために，自然言語を利用して常識定義を行うとともに，自動的に常識知識ベースを構築する手法の提案を目指す．

3　常識表現の自動選定

本研究では，名詞を特徴付ける用言をその名詞の持つ常識と定義し，常識の持つ性質として以下の仮説を立てた．

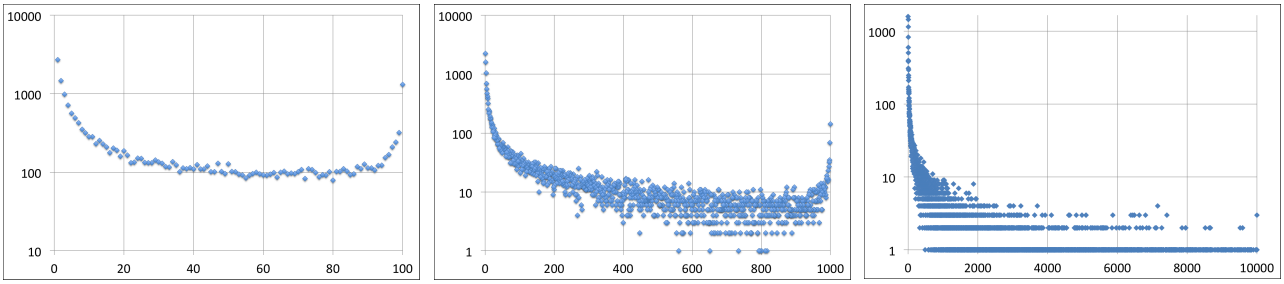
(1) 名詞 n に対して多く共起するほど，用言 a は名詞 n の常識である可能性が高い．

(2) 名詞 n は常識の集合によって特徴づけられるはずなので，どのような名詞とも共起する用言は常識として不適切である．

(3) 用言 a が名詞 n の常識として適切か否かは，その名詞と共起する用言の異なり数に依存する．多くの名詞と共起する用言でも，共起する用言数が少ない名詞に対しては常識となる場合がある (e.g. 用言 “はしる” は，共起する用言が多い名詞 “ひと” を特徴づけない (=常識としては不適切) が，共起する用言数が少ない名詞 “ランナー” を特徴づける)．

この仮説をもとに，名詞と共起する用言の中から常識として不適切な語を除外することで，適切な用言の

[1] http://wordnet.princeton.edu/
[2] https://framenet.icsi.berkeley.edu/fndrupal/
[3] http://conceptnet5.media.mit.edu/



(1) N=100　　(2) N=1000　　(3) N=10000

図 1: 名詞と共起する用言の異なり数上位 N 件における用言の出現分布 (横軸：用言の出現名詞数，縦軸：用言の異なり数 (対数))

選定を行う．解析対象として，Web 日本語 N グラム第 1 版【1】中の 7 グラム内で共起する名詞と用言の対を用いた．Web 日本語 N グラムには，200 億文から出現頻度が 20 回以上の文が格納されており，今回利用する 7 グラムの総数は 570,204,252 個である．

まず MeCab【2】による形態素解析を行い，名詞と共起する用言との対を抽出する．この際，表記統合辞書【3】を用いて各語の表記ゆれを吸収している．抽出した名詞と用言は 605,363,630 対で，異なり数は 29,434,191 対であった．このうち，名詞の異なり数は 655,038 語，用言の異なり数は 26,455 語である．

次に，共起する用言の異なり数が多い名詞順に名詞を並び替え，上位 N 件における用言の出現分布を調査する．N=100，1,000，10,000 と変化させた場合の用言の出現分布を図 1 に示す．横軸は用言の出現名詞数 (e.g. 用言 “走る” が，共起する用言の異なり数が多い名詞上位 1,000 件中において 500 種類の名詞と共起した場合，出現名詞数=500 となる)，縦軸は用言の異なり数を対数表示している (e.g. 出現名詞数=500 の用言が 10 語あった場合，用言の異なり数=10)．用言の出現分布より，出現名詞数が極端に少ない場合と極端に多い場合について，用言の異なり数が加速度的に増加していることが分かる．また N の値が増加するごとに出現名詞数が少ない用言が増加し，出現名詞数が多い用言が減少している．

事前に仮定した常識の持つ性質 (2) に従い，多くの名詞と共起する用言は常識として不適切とみなし除外する．本研究では，出現名詞数が多い場合に用言の異なり数が加速度的に増加している点に着目し，その範囲内に属する用言を除外の対象とした．範囲の決定には図 2 に示すような累乗近似曲線を用いて，近似曲線上に連なる点 (図中の赤い点) までを削除範囲とした．

N の値を 100 から 4500 まで 100 刻みに変化させた場合の，削除される用言数の変化を図 3 に示す．

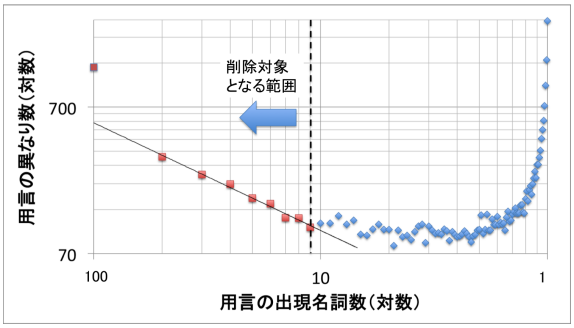


図 2: 図 1(1) における用言の出現分布と累乗近似曲線

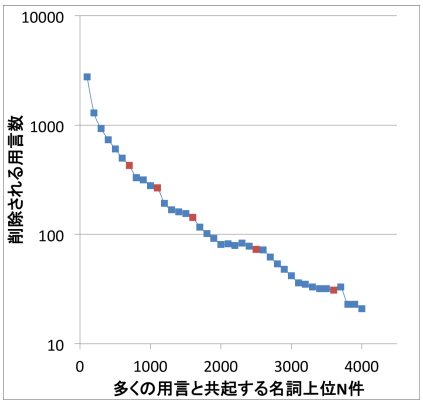


図 3: N の変化に伴う削除用言数の変化

削除される用言数は N の値を増加させる毎に階段状に減少しており，N=700，1,100，1,600，2,500，3,600 の際に特異点を持つ．この結果を踏まえ，常識の性質 (3) に従い各名詞に対する削除用言数を決定する．共起する用言の異なり数が多い名詞ほど汎用的な用言では特徴付けられない (=常識として不適切) ため，削除される用言数が多くなる．各名詞における削除用言数を表 1 に示す．

表 2: 名詞に対して付与される用言の違い (スコア順上位 10 件)

名詞：いぬ				名詞：小学校			
baseline1	baseline2	baseline3	提案手法	baseline1	baseline2	baseline3	提案手法
かう	かう	喰わない	散歩	入学	就学	離任	入学
なる	一緒	飼わない	しつける	教育	入学	訓導	教育
いる	生活	かみころす	病気	ある	付属	めざめない	卒業
ある	販売	吠えない	つれる	なる	参観	さかしい	授業
生活	たのしい	薬殺	くらす	卒業	給食	加減乗除	受験
みる	やすい	繋留	訓練	授業	受験	そばだつ	かよう
ない	わかる	訓練	ほえる	受験	授業	歌わす	学習
いう	登録	やせこける	かわいい	かよう	担任	やり直さない	指導
一緒	大きい	かまない	介護	学習	卒業	のびゆく	依頼
できる	かんがえる	代参	飼育	指導	かよう	実験	就学

表 1: 各名詞に対する削除用言数 (N=共起する用言の異なり数)

対象となる名詞の範囲	削除数
N≤700	427
700<N≤1,100	267
1,100<N≤1,600	143
1,600<N≤2,500	73
それ以外	33

例えば，N=1,000 の名詞については 227 個の用言が削除対象となる．ただし N=3,600 の際に削除される 33 個の用言は，対象となる名詞を選ばない用言が多く含まれていたため，そもそも常識としては不適切であると判断し，全ての名詞に対する削除用言とした．図 4 に，すべての名詞に対する削除用言 (= 常識として不適切) を示す．

わかる，もつ，みる，なる，ない，とる，できる，つく，しる，くる，おもう，おおい，いる，いう，ある，良い，入る，でる，つくる，つかう，きく，かく，おこなう，紹介，よい，ゆく，たつ，たかい，おる，いい，関係，やる，かける

図 4: すべての名詞に対する削除用言 (共起する名詞の異なり数が多い順)

常識として選定した用言を用いて，名詞に対する常識の付与を行う．名詞と共起する用言に対して Harman 正規化した TF を用いて重み付けを行い，スコアが高い用言ほどその名詞に対する常識として適切であるとした．名詞 n と共起する用言 a について，Harman 正規化した TF の計算式は以下の通りになる．ここで $n_{a,n}$ は，用言 a の名詞 n における出現回数を表す．

$$TF(a,n) = \frac{\log_2(n_{a,n}+1)}{\log_2(\sum_k n_{k,n})} \quad (1)$$

4 評価

4.1 評価方法

付与される常識集合を以下のベースラインと比較する．

(1) 用言の削除は行わず，Harman 正規化した TF で重み付けした用言を用いた場合 (baseline1).

(2) 用言の削除は行わず，用言を TF-IDF に則った式 (2) を用いて重み付けした場合 (baseline2).

(3) 表 1 中で示される N≤700 の場合に削除される用言 427 個をすべての名詞中から削除し，Harman 正規化した TF で重み付けした場合 (baseline3).

(1)，(2) と比較することで，頻出用言を削除することの有用性を確認する．また (3) と比較することで，名詞ごとに削除する用言を変化させることの効果を確認する．

baseline2 で用いた，名詞 n に対する用言 a の重み $wgt(a,n)$ を以下の式で定義する．$|N|$ は名詞の総数，$|N_a|$ は用言 a と共起する名詞の総数をそれぞれ表す．

$$wgt(a,n) = TF(a,n) \times \left\{ \log_2 \frac{|N|}{|N_a|} + 1 \right\} \quad (2)$$

4.2 評価結果

2 つの名詞を例に，付与される用言 (=常識) のスコア上位 10 件を，表 2 に示す．提案手法はほとんどの名詞において常識として適切な用言が付与されており，それは baseline1，baseline2 と比較しても明らかである．この結果から，多くの名詞と共起する用言を削除する本手法の有効性を確認した．本研究は動詞を利用して名詞同士のシソーラスを自動構築する従来の研究

[3][7] と比較して，不適切な用言を削除してしまうことで，名詞に付与する用言レベルでの細かい比較が可能となっている．

また baseline3 の結果では，名詞の常識を表現する語まで削除されてしまった影響で，常識として不適切な用言が含まれてしまっている．このことから，名詞ごとに削除する用言の数を変更する本手法は適切であったといえる．

4.3 適切な用言の付与失敗例について

適切な用言が付与できなかった例を表 3 に示す．付与失敗の原因として，文中で名詞と共起はしても，実際にはほとんど関係のない用言が多いことがあげられる．この解決策として，文中で名詞に係っている用言のみを常識候補として利用することなどが考えられる．また数は少ないが，名詞 “月” のように接尾辞的な使われ方 (e.g. 6 の月に入籍する，月ごとに決算する) をする名詞に対しては，適切な用言が付与できなかった．“月” に関しては，天体の月なのか月日の月なのかといった曖昧性の問題も存在する．この問題は，例えば名詞 “太陽” とともに使われる場合には天体の意味合いである可能性が高いなど，文中で共起する他名詞との関係性を考慮することで緩和できる．更に名詞 “理由” や “原因” などは，名詞同士の関係を定義する際に使われるものであり，そもそも常識を付与する対象として適切かどうかを議論する必要があり，今後は常識を付与する対象である名詞をどう制限していくのかについても考慮する必要がある．

表 3: 常識の付与がうまくいかなかった例

名詞：月	名詞：理由	名詞：原因
必着	やむをえない	救命
決算	返品	老化
施行	稼げない	つきとめる
公布	拒絶	故障
利上げ	解雇	ひきおこす
ずれこむ	削除	病気
入籍	志望	食中毒
ぞくする	却下	出火
連結	ことわる	肥満
着工	上告	くすむ

5 おわりに

本稿は，常識知識ベースの構築にあたり，常識として適切な用言の選定方法について述べた．共起する用言の異なり数順に名詞をソートし，上位 N 件の名詞と用言の出現頻度の関係について調査した結果，N=700, 1,100，1,600，2,500，3,600 の際に削除される用言を，指定範囲の名詞に対する除外用言と決定した．

各名詞に対して付与される常識集合を評価したところ，提案手法はベースラインと比較して適切な用言が常識として付与されていることが確認でき，本提案手法の有効性を確認した．

今後は，今回は扱わなかった，共起する用言の少ない名詞に対する常識表現の付与についても検討していく予定である．

使用した言語資源及びツール

【1】Web 日本語 N グラム第 1 版, http://www.gsk.or.jp/catalog/GSK2007-C/catalog.html.

【2】MeCab 0.993，http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html.

【3】表記統合辞書, http://www2.ninjal.ac.jp/lrc/index.php?%A1%D8%C9%BD%B5%AD%C5%FD%B9%E7%BC%AD%BD%F1%A1%D9，国立国語研究所.

参考文献

[1] K. Ahrens, S.F. Chung, and C.R. Huang. Conceptual metaphors: Ontology-based representation and corpora driven mapping principles. In *Proceedings of the ACL 2003 workshop on Lexicon and figurative language*, Vol. 14, pp. 36–42. Association for Computational Linguistics, 2003.

[2] H. Hennett and C. Fellbaum. Linking framenet to the suggested upper merged ontology. In *Formal Ontology in Information Systems: Proceedings of the Fourth International Conference (Fois 2006)*, Vol. 150, p. 289. Ios PressInc, 2006.

[3] D. Hindle. Noun classification from predicate-argument structures. In *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, pp. 268–275. Association for Computational Linguistics, 1990.

[4] I. Niles and A. Pease. Towards a standard upper ontology. In *In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, pp. 17–19, 2001.

[5] I. Niles and A. Pease. Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology. In *Proceedings of the IEEE International Conference on Information and Knowledge Engineering*, pp. 412–416, 2003.

[6] 村本晃一, ジェプカラファウ, 荒木健治. 自動生成された常識的知識を表現する文の自然性判定. 第 10 回情報科学技術フォーラム講演論文集, pp. 217–220, 2011.

[7] 萩原正人, 小川泰弘, 外山勝彦. シソーラス自動構築における PLSI の利用 (シソーラス・辞書). 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2005, No. 22, pp. 71–78, 2005.