# 大規模常識知識ベース構築のための常識表現の自動獲得

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ことを目指す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，名詞が格助詞付きで係る用言 (動詞，形容詞，サ変名詞) の集合をその名詞の持つ常識であると仮定し，これらの常識を大規模な Web テキストから自動的に獲得する手法を提案する．たとえば，名詞 “犬” と文中で係る “が-ほえる”，“と-散歩” といった用言の集合は，名詞 “犬” が持つ常識である．

更に，常識が類似した名詞は類似した常識集合を持つと仮定し，獲得した常識を利用して名詞同士の類似度計算を行う．名詞同士の類似度を測ることで，同時に名詞同士の関係性を推定することが可能となる（例えば，名詞 “犬” と名詞 “猫” は類似した常識を持つ類似した概念であり，双方の名詞に共通する常識集合は上位概念である “動物” が持つ常識集合に類似する）．各名詞に対し常識を定義し，更にそれらの名詞を常識の類似度に基づいた関係ネットワークで結ぶことで，最終的に大規模常識知識ベースの構築を目指す．

## 2　関連研究

人工知能研究の分野では，常識知識ベースは上位オントロジーと呼ばれることもある．上位オントロジーとは，大量の一般的概念を定義したオントロジーである．Ahrens et al.[1] は上位オントロジーの一種である SUMO[6] 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 and Pease[7] や，Hanett and Felbaum[4] は，上位オントロジー中の概念と既存の語彙資源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が多く，自然言語処理のタスクで扱いづら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，単語や短い文の単位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．しかし各概念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 [8]．

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，大量の常識を自動的に収集する手法を提案する．

## 3　処理対象となる名詞および用言

本研究では，日本語語彙大系【1】中で “名詞-具体” でラベル付けされている名詞 12,042 語を常識付与の対象として扱う．具体名詞のみを選別することで，常識付与の対象として相応しくない名詞（原因，理由といった名詞同士の関係を定義する名詞 等）を除外する．

次に，Web 日本語 N-gram【2】の 7-gram データを CaboCha【3】を利用して構文解析し，対象名詞に対し格助詞付きで係る用言をその名詞に対する常識として定義する．最終的に，1,631,209 個の名詞-用言対を獲得した．

## 4　常識として適切な用言の自動選定

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．

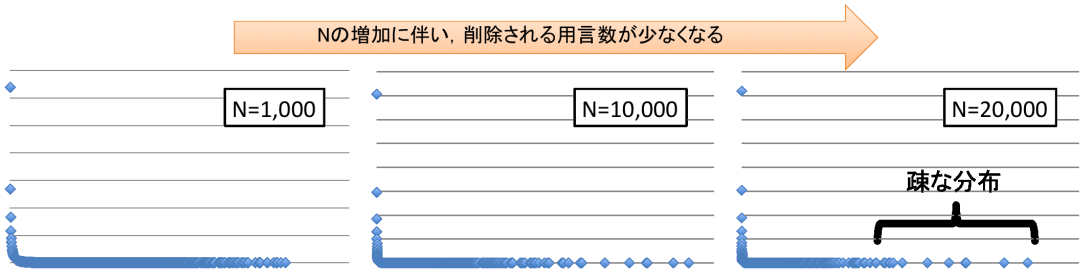


図 1: 係り先の用言数が多い上位 N 件の名詞について，N の値を変化させた場合の削除用言数の変化 (横軸：用言の出現名詞数，縦軸：用言の異なり数)

(1) 名詞 $n$ に対し高頻度で係る用言 $p$ 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，どのような名詞にも係る用言は常識として不適切である．

(3) 用言 $p$ が名詞 $n$ の常識として適切か否かは，その名詞の係り先の用言の異なり数に依存する．多くの名詞に係る用言でも，係り先の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 “が-走る” は，係り元の名詞 “ひと” を特徴づけないが，係り先の用言数が少ない名詞 “ランナー” を特徴づける）．

仮説 (1) および仮説 (2) から，特定の名詞に高い頻度で係る用言はその名詞にとって常識である可能性が高いが，その用言が多くの名詞に係る汎用的な用言であった場合は，常識として不適切となる．つまり，多くの名詞に対し係る用言を，常識として不適切であるとして除外する必要がある．

まず，係り先の用言の異なり数が多い (=多くの用言の係り元となる) 順に名詞を並べ替え，上位 N 件の名詞に対する用言の出現分布を調査した (図 1)．横軸は用言の出現名詞数 (e.g. 用言 “が-走る” が 500 種類の名詞の係り先となった場合，出現名詞数は 500 となる)，縦軸は用言の異なり数を示している (e.g. 出現名詞数が 500 の用言が 10 語あった場合，用言の異なり数は 10 となる)．結果から，出現名詞数の増大に伴い値の出現が疎となっている事が分かる．これは，一部の汎用的な用言 (e.g. ある, 行う) が極端に多くの名詞の係り先となっていることが原因である．本研究ではこの点に着目し，用言の異なり数が出現名詞数に対し疎となる範囲に属する用言を削除対象の用言 (=常識として不適切な用言) であるとした．

仮説 (3) に基づき，各名詞に対する削除用言を決定するため，N の値を 1,000 から 70,000 まで変化させた場合の，削除される用言数の変化を調査した．この結果を図 2 に示す．N=1,000〜20,000 の間においては削除用言数は累乗的に変化し，N が 20,000 を超えた段階でほとんど変化しなくなっていることが分かる．この結果から，各名詞における削除用言数を決定する．N を変化させた場合の削除用言数の変化から近似曲線を求め，N=1,000〜20,000 の間においては近似曲線の式から削除用言を決定する．削除用言数を求める式を以下に示す．

$$y = 13135.0 \times x^{-0.583} \quad (1)$$

N が 1,000 よりも小さい場合は N=1,000 で削除される 234 語を削除し，N が 20,000 を超える場合は，N=20,000 の場合に削除される 40 語を削除した．

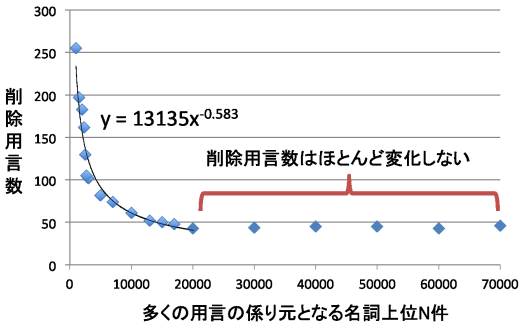


図 2: 多くの用言の係り元となる名詞上位 N 件における削除用言数の変化

以上の処理により選定した用言を，それぞれの名詞に対する常識として扱う．それぞれの用言が名詞に係って出現する頻度が高いほど，その名詞の常識として適切な用言であると考えられる．

## 5 各概念同士の類似度計算

常識知識ベースを構築するために，獲得した常識を用いて名詞間の意味的関係を調査する．我々は名詞間に現れる性質として以下の 2 つの仮説を立てた．

1. 名詞対に付与される常識集合が類似していた場合，その名詞対は類似した概念を持つ．

2. 名詞 $a$ と名詞 $b$ が類似しおり，かつ名詞 $b$ と名詞 $c$ も類似している場合は，名詞 $a$ と名詞 $c$ もまた類似した名詞である．

仮説 (1) より，名詞同士の類似度はその名詞同士の持つ用言集合の類似度で測れることになる．更に，係り受け解析誤り等によって付与された低頻度の用言による影響を抑えるために，解析対象となる名詞同士が持つ共通した用言集合のうち，それぞれにおいて最も頻度の低い用言以下の用言集合を削除した．図 3 に，名詞 “犬” と名詞 “猫” の例を示す．

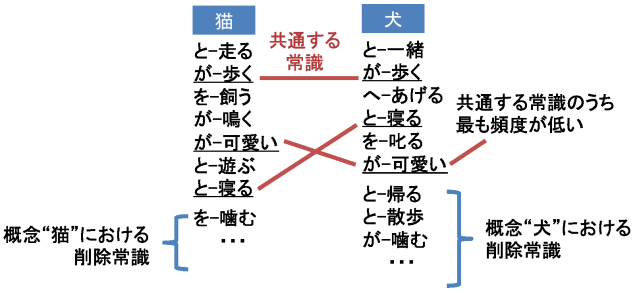


図 3: 削除対象となる低頻度用言の決定

次に実際に名詞同士の類似度を計算する．名詞 $w_i$ と名詞 $w_j$ が類似しており，更に名詞 $w_i$ と名詞 $w_a$ も類似していれば，仮説 (2) から，名詞 $w_j$ と名詞 $w_a$ もまた類似している可能性が高い．つまり，名詞 $w_i$ と名詞 $w_j$ の類似度が高い場合，名詞 $w_i$ とその他の名詞との類似度集合 $SIM_i$ と，名詞 $w_j$ とその他の名詞との類似度集合 $SIM_j$ の類似度も高くなる．

以上の考え方から，比較する 2 つの名詞とその他の名詞集合との類似度を計算し，両者の類似度集合の相関を求め，この相関係数を両者の類似度とする．

# 6 評価

作成した名詞の常識知識ベースについて，名詞に対し正しい常識が付与され，名詞同士の関係を正しく計算できているか評価する．本研究では，評価セットとして日本語語彙大系中で “名詞-具体” にラベル付けされ，更に日本語 N-gram 中で出現頻度の上位 90 %を占める 1,617 個の名詞を用いて，評価セットと正解セットにおける各名詞間の類似度集合の相関係数を求める．正解セットとして，日本語語彙大系中における名詞間の距離を計算した．概念 $x \in X$ を持つ名詞 $w_i$ と，概念 $y \in Y$ を持つ名詞 $w_j$ の類似度は以下の式で計算される．

$$ave\_sim(w_i, w_j) = \frac{1}{|XY|} \sum_{x \in X, y \in Y} \frac{2d(w_{i,x}, w_{j,y})}{d(w_{i,x})d(w_{j,y})} \quad (2)$$

$$max\_sim(w_i, w_j) = \max \left( \frac{2d(w_{i,x}, w_{j,y})}{d(w_{i,x})d(w_{j,y})} \right) \quad (3)$$

ここで $d(w_i)$ とは，根から $w_i$ までの深さ，$d(w_i, w_j)$ とは，根から名詞 $w_i$ と名詞 $w_j$ が共有する上位概念までの深さを表している．名詞 $w_i$ と名詞 $w_j$ が類似した概念を持っている場合，両 $sim$ 関数の値は高くなる．

## 6.1 比較手法

提案手法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（ベースライン 1）．

(2) 自己相互情報量（PMI）のスコアが $\beta$ 以下の用言を，類似度計算に悪影響を及ぼす用言であると考え削除する手法 (ベースライン 2)．本手法は相澤 [9] によって定義された名詞同士の類似度計算手法のうち，実験中で最も精度の高かった手法である．

本実験では文献 [9] と同様の手法で相関係数の変化を調査し，定数 $\beta = 0$ と設定した．

## 6.2 評価結果

以下に示す式を用いて，概念 $x \in X$ を持つ名詞 $w_i$ と，概念 $y \in Y$ を持つ名詞 $w_j$ の類似度を計算する ($Jac$: Jaccard 係数，$Simp$: Simpson 係数，$WJac$: 重み付き Jaccard 係数，$f(w_i, p)$: 名詞 $w_i$ に係る用言 $p$ の出現頻度)．

$$Jac(w_i, w_j) = \frac{|X \cup Y|}{|X \cap Y|} \quad (4)$$

$$Simp(w_i, w_j) = \frac{|X \cup Y|}{\min(|X|, |Y|)} \quad (5)$$

$$WJac(w_i, w_j) = \frac{\sum_p \min(f(w_i, p), f(w_j, p))}{\sum_p \max(f(w_i, p), f(w_j, p))} \quad (6)$$

ベースラインと提案手法に付与される用言のトップ 10 の例を，表 1 に示す．提案手法では，すべての用言がそれぞれの名詞に対する常識となっている．これに対し，どちらのベースラインも出現頻度は高くても多くの名詞に係る汎用的な用言 (e.g. “に-なる”) が上位にきてしまっている．これらの名詞は常識として相応しくなく，提案手法ではこうした不適切な用言の削除に成功している．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 [5, 2] と比較して，名詞同士の関係性を測れるだけではなく，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．

表 2 に，それぞれに付与した用言を利用して名詞同士の類似度を計算した結果を示す．提案手法はベース

表 1: 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：世の中 | | | 名詞：道路 | | |
|---|---|---|---|---|---|
| ベースライン 1 | ベースライン 2 | 提案手法 | ベースライン 1 | ベースライン 2 | 提案手法 |
| に-なる | に-なる | を-生き抜く | が-分断 | が-分断 | が-分断 |
| に-ある | に-ある | で-起こる | に-関連 | に-関連 | を-走る |
| を-生き抜く | を-生き抜く | に-存在 | を-走る | を-走る | に-面す |
| を-変える | を-変える | に-広める | に-面す | に-面す | を-挟む |
| に-いる | に-いる | に-必要 | を-使う | を-使う | を-直進 |
| で-起こる | で-起こる | に-送り出す | を-挟む | を-挟む | から-出入り |
| に-存在 | に-存在 | の-役に立つ | を-直進 | を-直進 | に-接す |
| に-広める | に-広める | に-役立つ | を-利用 | を-利用 | を-横断 |
| に-出る | に-出る | に-貢献 | から-出入り | から-出入り | を-渡る |
| に-必要 | に-必要 | を-動かす | に-ある | に-接す | が-整備 |

ラインと比較して高い精度を取ることが示された．特に Jaccard 係数を用いた場合に最高の精度を出している．このことから，提案手法の方が名詞に対してより常識として相応しい用言を付与出来ていることが確認できた．この結果は，本手法は名詞に対する常識集合を集められるだけではなく，類似度計算手法としても有用であることを示している．

表 2: 名詞同士の類似度の評価結果

| | | Jac | Simp | WJac |
|---|---|---|---|---|
| ベースライン 1 | ave | 0.443 | 0.326 | 0.378 |
| | max | 0.451 | 0.335 | 0.376 |
| ベースライン 2 | ave | 0.480 | 0.442 | 0.371 |
| | max | 0.481 | 0.446 | 0.364 |
| 提案手法 | ave | **0.607** | 0.499 | 0.582 |
| | max | **0.591** | 0.461 | 0.558 |

## 7 おわりに

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た．名詞が係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，名詞に対して不適切な用言を自動的に削除することに成功した．

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 2 つの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た．また 2 つのベースラインと比較して名詞同士の類似度計算の精度が高かったことから，本手法が常識の付与だけにとどまらず，名詞同士の類似度計算にも有用であることが分かった．さらにこの結果から，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り，またその用言が常識として適切か否かは常識の付与対象である名詞に依存するという，本研究における常識に関する仮説が立証された．

## 使用した言語資源及びツール


【1】白井 諭，大山 芳史，池原 悟，宮崎 正弘，横尾 昭男，“日本語語彙大系について”，情報処理研究報告.IM，vol.98, no.106，pp.47-52，1998.

【2】工藤 拓，賀沢 秀人，“Web 日本語 N グラム 第一版”，言語資源協会.

【3】工藤 拓，松本 裕治，“チャンキングの段階適用による日本語係り受け解析”，vol.43，no.6，pp.1834–1842，2002.


## 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] M. Hagiwara, Y. Ogawa, and K. Toyama. A comparative study on effective context selection for distributional similarity. *Journal of Natural Language Processing*, Vol. 5, No. 5, pp. 119–150, 2008.

[3] C. Havasi, R. Speer, and J. Alonso. Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *In Recent Advances in NLP*, 2007.

[4] 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.

[5] 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.

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

[7] 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.

[8] Rafal Rzepka, Koichi Muramoto, and Kenji Araki. Generality evaluation of automatically generated knowledge for the japanese conceptnet. In *Proceedings of 24th Australasian Joint Conference*, pp. 648–657, 2012.

[9] 相澤彰子. 大規模テキストコーパスを用いた語の類似度計算に関する考察. 情報処理学会論文誌, Vol. 49, No. 3, pp. 1426–1436, 2008.