Wikipedia を利用した日本語 WordNet への用語追加の検討

山田一郎　　呉鍾勳　　鳥澤健太郎　　黒田航　　風間淳一　　村田真樹
情報通信研究機構 MASTAR プロジェクト 言語基盤グループ
E-mail: iyamada@nict.go.jp

1. はじめに

2009 年 2 月に日本語 WordNet [1]が公開され、今後、日本語 WordNet を利用した様々なアプリケーションの実現が期待される。しかし現状では、日本語 WordNet に登録されている名詞の数は 6 万語程度であり、特に固有名の登録は少ない。情報検索などの用途でシソーラスを利用する場合、多くの語が登録されていることが望まれる。そこで本稿では、Wikipedia に出現する用語を日本語 WordNet へ追加する手法を提案する。Wikipedia には大量の用語に関する知識が含まれるため、Wikipedia 中の用語が日本語 WordNet に登録され、2 つのリソース間のリンクが生成されることにより、検索などのアプリケーションで利用する際に相乗的な効果が期待できる。

従来、Snow らは単語が上位語と共起する際のパターンと、単語間の類似性を利用することにより、WordNet に新たな単語を追加する手法を提案している[2]。また、市瀬らは、2 つの階層的な構造を持つ辞書を、各ノードが持つインスタンスの共通項を手掛かりとして統合する手法を提案している[3]。提案手法では、これらの従来手法とは異なるアプローチをとり、Wikipedia におけるタグ情報などを利用して確度高く推定した上位語を利用する。上位語が推定された Wikipedia 中の用語に対して、WordNet の意味概念を表す synset を推定する。対象語と上位語のペアを処理対象とすることにより、対象語に多義性がある場合でも、その上位語により対象語の語義をそれぞれ特定することができ、同一語に対して異なる synset を推定することが可能となる。対象語の synset を推定する処理では、ALAGIN フォーラム[1]から公開されている文脈類似語リストと上位語階層データを利用する。実験では、対象語の上位語が属する WordNet の synset 数により分類した各処理について報告する。現状のリソースを利用することにより、約 51 万個の上位下位関係に対して 84.5%以上の精度で下位語の synset を推定できることを確認した。

2. 使用するリソース

本章では、用語の追加対象となる日本語 WordNet、用語追加のための情報源となる Wikipedia から抽出した上位下位関係[4]、上位語階層データ[5]、文脈類似語データベース[6]について説明する。

2.1 日本語 WordNet

日本語 WordNet は、Princeton WordNet 3.0 をベースとして構築されている。一つの synset が一つの意味概念に対応し、各語は最低一つの synset に属している。また、各 synset は上位下位関係などの多様な関係で結ばれている。処理対象としたリリース版の Wn-Ja0.92 では、synset は 49,655 概念、87,133 語（名詞は 59,439 語）、語と synset のペアは 146,811 語義が収録されている。本稿では、新たな名詞に対して、属する synset を推定し、日本語 WordNet の語彙数を増加させることを目的とする。

2.2 Wikipedia から抽出した上位下位関係

本稿では、隅田らの手法[4]により Wikipedia から抽出される上位下位関係を利用する。この手法では、Wikipedia の各記事に含まれる定義文、カテゴリー情報、階層構造を手掛りとして、SVM により上位下位関係か否かを判定することにより、90%の抽出精度で上位下位関係を得ることができる。我々は、この手法により Wikipedia から用語の上位下位関係を抽出するツールを公開している[2]。このツールによって 2007 年 3 月バージョンの Wikipedia を解析した結果、約 182 万組の上位下位関係(下位語の異なり約 87 万)が抽出された。Wikipedia から自動抽出した上位下位関係の一部を表 1 に示す。

Wikipedia から自動抽出した上位下位関係は、従来のシソーラスとは異なり、下位語に固有名が多く、ニュース記事などの実データを解析する際に有用なデータと考えられる。しかし、下位語の多くが固有名であるため、さらなる下位語を持つことが少ない。自動抽出した上位下位関係における共通する用語を繋げて、木構造を生成すると、木構造の木の深さは平均 2.89 と、浅い構造となってしまう。Wikipedia から自動獲得した上位下位関係だけで WordNet のような精錬された階層構造を持つシソーラスを構築することは難しい。

2.3 上位語階層データ

上位語階層データ[5]は、隅田らの手法[4]によって 2007 年 3 月バージョンの Wikipedia を解析し抽出した上位下位関係の上位語の約 69,000 語を対象としている。上位語を構成する形態素を語の先頭部分から削除した語（以後、縮退語と呼ぶ）を順次生成し、縮退語を元の語の上位語とした階層構造を生成する。この処理において、縮退語が縮退前

表 1. Wikipedia から自動獲得した上位下位関係例

上位語	下位語
抗ヘルペスウイルス薬	ガンシクロビル
太平洋の島	ケルマディック諸島
戦国武将	松前盛広
解熱鎮痛剤	ナロンＳ
近畿地方の鉄道路線	神戸電鉄粟生線
海上保安庁の巡視船	てしお

[1] http://www.alagin.jp/

[2] http://nlpwww.nict.go.jp/hyponymy/index.html

の語の上位語として相応しくない場合がある。そこで、縮退語を人手により評価を行い、評価値を付与している。例えば、「解熱鎮痛剤」という上位語は、下記のように縮退語が生成される。

剤 ＞ 鎮痛剤 ＞ 解熱鎮痛剤

「鎮痛剤」は「解熱鎮痛剤」の上位語として相応しいが、「剤」は「鎮痛剤」の上位語としては問題となる。この場合、「鎮痛剤」にはGood、「剤」にはDubiousという評価値が付与されている。上位語階層データでは、上位語としての相応しさから 5 種類(Good, Less Good, Dubious, Bad, Connector)の評価値を使用している。

本稿では、用語が属するsynsetを推定する際に、Wikipediaから抽出した上位語を利用する。しかし、上位語は複数の文節から構成される語であることが多く、そのような語はWordNetにおける登録が少ない。表1の上位語は、どれもWordNetの登録語ではない。そこで、上位語がWordNetに登録されていない場合は上位語階層データを利用し、WordNetに登録され、Good もしくはLess Goodの評価値を持ち、上位語に最も近い縮退語を、上位語の代わりに利用する。

2.4 文脈類似語データベース

文脈類似語データベース[6]では、100 万語の名詞に対して、約1億ページのWeb文書上での文脈が類似している名詞が、類似度とともに順に最大500個列挙されている。

類似性は、名詞があるクラスに属する確率分布間のJensen-Shannon divergenceにより評価している。助詞を介して動詞を修飾する係り受け構造を一つのクラスとみなした手法と、EM アルゴリズムにより名詞と助詞を介して動詞を修飾する係り受け構造をクラスタリングした結果を利用する手法による2種類のデータを公開している。

本稿では、クラスタリングを用いた手法による文脈類似語データを利用し、WordNet に登録された単語と、synsetを推定する対象単語やその上位語との類似性を評価することにより、最適なsynset推定を行う。

3. 対象語が属するsynset推定

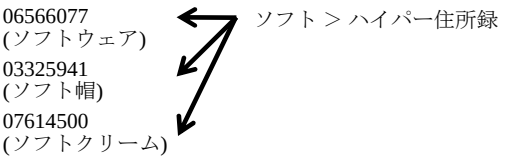
Wikipedia から抽出した上位下位関係の下位語を処理対象とし、対象語が属するWordNet中のsynsetを推定する。対象語の上位語、または、上位語の縮退語が属するsynset数によって、図1のようにデータ5つに分類して処理を行う。Wikipediaから抽出した1,884,513個の上位下位関係中、下位語が日本語 WordNet に既に登録されているものは58,504個であった。残りの1,826,009個の上位下位関係の下位語が、WordNetへの追加対象語となる。各対象に対するsynset推定処理について、以下の節で説明する。

3.1 上位語が一つのsynsetを持つ場合(Data 1)

対象語の上位語がWordNetに登録されており、さらに、その上位語が持つsynsetが一つしか無い場合、対象語が属するsynsetを、上位語の属するsynsetとする(手法 1)。WordNetにおける登録単語数の問題から、上位語に多義性があるにも関わらずWordNetにおけるsynsetが一つとなっている場合もある。このような場合には、対象語が属するsynsetは誤って特定される場合がある。また、Wikipediaから抽出した上位下位関係の精度は90%であるため、10%程度の誤りが含まれる。誤った上位語が抽出された対象語も、誤ったsynsetとなってしまう。上位語が一つのsynsetを持つ147,280個から200個を任意抽出し、各対象語(上位下位関係の下位語)に対して特定された synset が相応しいかを人手により評価した。結果を表2のData 1欄に示す。

3.2 上位語が複数のsynsetを持つ場合(Data 2)

対象語の上位語がWordNetに登録されており、さらに、その上位語が持つsynsetが複数存在する場合、対象語がどのsynsetに属するかを推定する必要がある。例えば、下記の例では、対象語「ハイパー住所録」の上位語「ソフト」には、3つの属するsynsetがあり、「ハイパー住所録」に最適な“06566077(ソフトウェア)”以外の synset を選択しては問題となる。



そこで、対象語に相応しいsynsetの推定を行う。我々はこれまでに、文脈類似語データベースの類似度をベースと

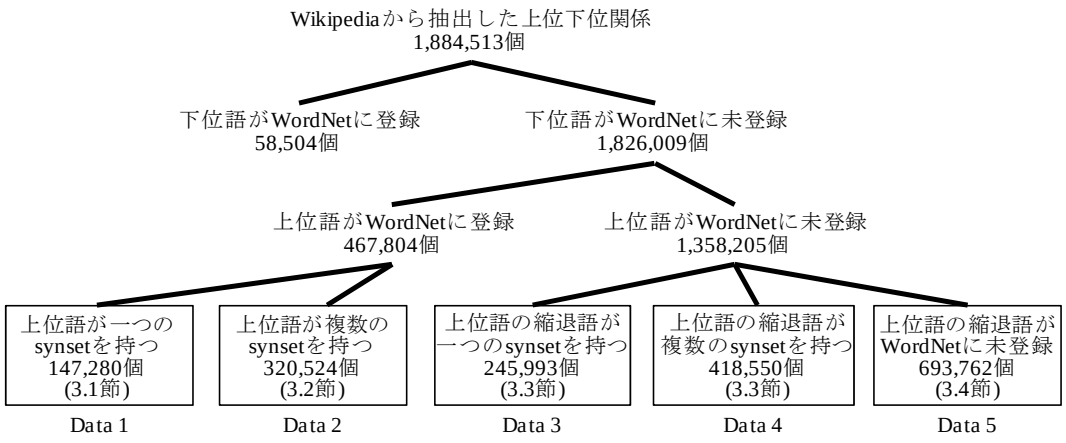


図 1. Wikipedia から抽出した上位下位関係の分類

表 2. 各データにおける synset 推定精度と処理可能な語数

対象データ	手法	精度(synset 推定正解数/サンプルした上位下位関係数)	処理可能な対象語(下位語)数
Data 1	(手法 1) 上位語の synset に特定	84.5% (169/200)	147,280 語
Data 2	(手法 2-1) 対象語からの *score* を利用	90.0% (180/200)	44,576 語
	(手法 2-2) ランダムに選択	72.0% (144/200)	320,524 語
Data 3	(手法 3) 上位語の縮退語の synset に特定	84.5% (169/200)*	245,993 語
Data 4	(手法 4-1) 上位語階層データを利用	89.0% (178/200)	75,135 語
	(手法 4-2) 上位語階層データを利用しない	56.0% (112/200)	102,923 語
Data 5	(手法 5-1) 上位語の synset 推定結果を利用	68.0% (136/200)**	94,296 語
	(手法 5-2) 上位語と兄弟語の synset 推定結果を利用	70.0% (140/200)**	94,296 語

*Data 1(手法 1)の精度，**Data 4 でサンプルした上位下位関係を対象

して任意の用語の上位語を推定する手法[7]を提案している。synset 推定処理では、この手法を利用し、対象語 n_{trg} が属する synset らしさを示す $score(n_{trg}, n_{syn})$を(1)式で評価する。

$$score(n_{trg}, n_{syn}) = \sum_{sn_{hypo}} d^{r} \times (1 + sim(n_{trg}, sn_{hypo})) \quad (1)$$

sn_{hypo} は n_{syn} の下位に属し、かつ、対象語 n_{trg} と類似する上位 m 個に含まれる用語を示す。r は n_{syn} と sn_{hypo} の木構造中の階層の差を示す。この階層の差に対するペナルティの値を d とする。$sim(n_{trg}, n_{hypo})$は、文脈類似語データベース中の n_{trg} と n_{hypo} の類似度の値を示す。この値を利用し、対象語 n_{trg} が属する synset を(2)式により推定する(手法 2-1)。

$$\hat{n}_{syn}(n_{trg}) = \underset{n_{syn} \in syn(hyper(n_{trg}))}{\arg\max} (score(n_{trg}, n_{syn})) \quad (2)$$

ここで $hyper(n_{trg})$は n_{trg} の上位語を示し、$syn(n)$は用語 n が属する synset を示す。対象語「ハイパー住所録」の例では、「ハイパー住所録」の上位語「ソフト」が属する 3 つの synset の中で、$score(n_{trg}, n_{syn})$の値が最大である“06566077(ソフトウェア)”が選択される。手法 2-1 では、上位語が属する全ての synset にスコアを与えられない場合もある。例えば、Wikipedia から自動抽出した上位下位関係で誤って抽出された上位語は、その synset には値が与えられないことが多い。そのため、上位語が持つ全ての synset に値が与えられなかった対象語は、処理対象から除いている。

実際に、上位語が複数の synset を持つものの中で、文脈類似語データベースの対象となっている 65,932 個を解析することにより、44,576 個の対象語に対して最低一つの synset にスコアが与えられ、属する synset を推定することができた。この中から 200 個を任意に抽出して人手による評価を行った。比較手法として、上位語が属する複数の synset からランダムに一つを選ぶ実験も行った(手法 2-2)。評価結果を表 2 の Data 2 欄に示す。評価結果から、対象語からの *score* を用いて synset を推定する手法の有効性がわかる。しかし、ランダムに synset を選んでも 72.0%は適当な synset が選択されている。例えば「学校」という単語は、教育機関として意味を持つ“08276720”と、校舎の意味を持つ“04146050”の 2 つの synset を持つ。「学校」の下位語である「西小学校」という語は、どちらの synset が適当か判断できないため、この場合ではともに正解と判定している。このように、上位下位関係の情報だけでは判断できないものが存在するため、ランダムに synset を選択しても一定の精度が得られている。

3.3 上位語の縮退語が WordNet に登録されている場合(Data 3, Data 4)

対象とする語の上位語が WordNet に登録されていない場合でも、上位語階層データによって上位語を WordNet に登録されている語に縮退することができれば、対象語が属する synset を推定することができる。上位語の縮退語が WordNet に登録されている 664,443 個の上位下位関係中、縮退語の synset が一つであるものが 245,993 個あった。これらの対象語(下位語)は、Data 1 における手法 1 と同様に、縮退語の synset に属すると推定する(手法 3)。残りの 418,550 個は、縮退語の synset が複数存在していた。この中で文脈類似語データベースの対象となっている 102,923 語を処理対象とし、(2)式の対象より synset 推定処理を行った(手法 4-1)。その結果、75,135 個の対象語に対して最低一つの synset にスコアを与えられ、対象語が属する synset を推定できた。また比較手法として、上位語階層データを用いない実験も行った。この手法では、全ての synset に対して(1)式により *score* の値を計算し、その値が最大である synset を抽出している(手法 4-2)。任意の 200 個の上位下位関係に対して、下位語の synset 推定処理を人手により評価した結果を表 2 の Data 4 欄に示す。

この結果から、上位語階層データを synset の推定処理で利用する手法の有効性が分かる。また、Data2 の結果と比較すると、上位語階層データを利用した手法は、上位語が WordNet に登録されている語とほぼ同程度の精度で synset を推定できることがわかる。

3.4 上位語、上位語の縮退語がともに WordNet に登録されてない場合(Data 5)

上位語が WordNet に登録されておらず、その上位語の縮退語も WordNet に登録されていない場合は、synset を限定することができない。3.3 節で言及した上位語階層データを利用しない手法(手法 4-2)を適用することができるが、精度は 56.0%と低い結果になっている。この手法では、Wikipedia から抽出した上位語の情報は利用できていない。そこで、対象語 n_{trg} の上位語 $n_{hyper(trg)}$に対しても(1)式を利用

して、各 synset に対するスコア $score'(n_{hyper(trg)}, n_{syn})$を計算する。(1)式では、$n_{syn}$ の下位の語のみを加算対象したが、ここでは n_{syn} の上位の語も加算対象とする。この値を利用し、対象語が属する synset を(3)式により推定する(手法 5-1)。

$$\hat{n}_{syn}(n_{trg}) = \arg\max_{n_{syn} \in S} \sum_{sn_{syn}} score(n_{trg}, n_{syn}) \times score'(n_{hyper(trg)}, n_{syn}) \quad (3)$$

ここで S は、WordNet の全ての synset 集合を示す。3.3 節の実験と同じデータを対象として、(3)式により対象語が属する synset を推定し、人手により評価した結果、その精度は 67.5%(135/200)に向上した(表 2)。

また、Wikipedia から抽出した上位下位関係では、同じ上位語を共有する兄弟語が数多く存在する。対象語の兄弟語も、対象語と同じ synset に属すると考えられるため、兄弟語 $n_{sibling(trg)}$に対しても(1)式により、属する synset らしさスコア $score(n_{sibling(trg)}, n_{syn})$を計算する。この値も利用し、対象語が属する synset を(4)式により推定する(手法 5-2)。

$$\hat{n}_{syn}(n_{trg}) = \arg\max_{n_{syn} \in S} \sum_{sn_{syn}} \{score(n_{trg}, n_{syn}) \times (score'(n_{hyper(trg)}, n_{syn}) + \sum_{n_{sibling(trg)}} \frac{score(n_{sibling(trg)}, n_{syn})}{N(n_{sibling(trg)})})\} \quad (4)$$

ここで、$N(n_{sibling(trg)})$は、兄弟語 $n_{sibling(trg)}$の数を示す。3.3 節の実験と同じデータを対象として、(4)式により対象語が属する synset を推定し、人手により評価した結果、その精度は 70.0%(140/200)であった(表 2)。精度は上位語のみを使う手法と比較して向上しているが、上位語に多義性がある場合は、兄弟語でも異なる synset が相応しいケースも存在するため、精度の向上は 2.5%に留まっている。

3.5 考察

表 2 から、Data 1 に属する 147,280 語は精度 84.5%で対象語の synset を特定できることが分かる。Data 3 に属する 245,993 語に対しても、同程度の精度が期待できる。また、Data 2 の 44,576 語、Data 3 の 75,135 語に対して、それぞれ精度 90.0%、89.0%で synset が推定できることを確認した。これらを合わせると、約 51 万個の上位下位関係(下位語の異なり数 283,938 語)に対して、84.5%以上の精度で下位語の synset が推定できることがわかる。synset 推定結果例を表 3 に示す。この処理結果を調査したところ、対象語の 42,233 語には、複数の synset が付与されていた。例えば、表 3 に示す「イトウ」という語には 02512053(魚)と 10200781(人物)という 2 つの synset が付与されている。多義性のある単語に対しても複数の synset を与えられていることがわかる。

Data 2、Data 4、Data 5 では、文脈類似語データベースに含まれた語のみを対象としているため、処理可能な語数が図 1 で示した数より少なくなっている。文脈類似語データベースは、今後、その対象語を増加する予定であり、本手法による日本語 WordNet への追加対象語も増加できると考えられる。

表 3. synset 推定結果例

対象語	synset 推定結果(括弧内は synset に属する代表的な用語)
ER 流体	14939900(流動体)
華為技術	080058098(会社)
アマンタジン	03740161(薬)
イトウ	02512053(魚)、10200781(人物)
スキタイ	07967982(民族)、08544813(国)
スーパージャンプ	05638063(技)、06596364(漫画雑誌)
スタンリー	00007846(人)、063438383(地名)、03684823(機関車)、09587565(キャラクター)

4. まとめ

本稿では、上位語階層データと文脈類似語データベースを利用することにより、Wikipedia に出現する用語を日本語 WordNet へ追加する手法を提案した。実験では、対象語の上位語とその縮退語を利用することにより、対象語の synset の推定を 89.0%～90.0%の精度で実現できることを示した。上位語とその縮退語が WordNet に登録されていない場合でも、上位語や兄弟語の synset 推定処理を行うことにより、対象語の synset 推定の精度を 70.0%まで向上できることも確認した。本手法は、多義性のある語に対しても適切に複数の synset を付与することができる。

現状のリソースを利用することにより、約 51 万個の上位下位関係に対して 84.5%以上の精度で下位語の synset の推定が可能となり、名詞の登録数が 6 万語程度の WordNet に対する有益な情報になると考えられる。今後、日本語 WordNet へ追加できる用語を拡張し、より効果的なリソースの公開を目指す。

【 参考文献 】

[1] F. Bond, H. Isahara, K. Uchimoto, T. Kuribayashi, K. Kanzaki, "Extending the Japanese WordNet," 言語処理学会第15回年次大会発表論文集, C1-4, pp.80-83. (2009)

[2] R. Snow, D. Jurafsky, A. Y. Ng, "Semantic Taxonomy Induction from Heterogenous Evidence," In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp.801–808. (2006)

[3] 市瀬, 武田, 本位田, "階層的知識間の調整規則の学習," 人工知能学会誌, Vol. 17, No. 3, pp. 230-238. (2002)

[4] A. Sumida, N. Yoshinaga and K.Torisawa, "Boosting Precision and Recall of Hyponymy Relation Acquisition from Hierarchical Layouts in Wikipedia," In *Proceedings of the Sixth International Language Resources and Evaluation*. (2008)

[5] 黒田, 李, 野澤, 村田, 鳥澤, "鳥式改の上位語データの人手クリーニング," 言語処理学会第第15回年次大会発表論文集, C1-3, pp.76-79. (2009)

[6] 風間, De Saeger, 鳥澤, 村田, "係り受けの確率的クラスタリングを用いた大規模類似語リストの作成," 言語処理学会第第 15 回年次大会発表論文集, C1-6, pp.84-87. (2009)

[7] I. Yamada, K. Torisawa, J. Kazama, K. Kuroda, M. Murata, S. De Saeger, F. Bond and A. Sumida, "Hypernym Discovery Based on Distributional Similarity and Hierarchical Structures," In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp.929-937. (2009)