意味辞書WordNetの不整合抽出

平尾, 拓也
九州大学大学院システム情報科学府

鈴木, 孝彦
九州大学情報基盤研究開発センター

宮田, 光樹
九州大学システム情報科学府

Miyata, Koki

他

出版情報：情報処理学会研究報告, pp.1-5, 2012. 情報処理学会
バージョン：

意味辞書 WordNet の不整合抽出

平尾拓也[†1]　鈴木孝彦[†2]　宮田光樹[†1]
廣川佐千男[†2]

意味辞書である日本語 WordNet は，自然言語処理に有用である．しかしながら，日本語 WordNet 中には 5%ほどのミスが存在している(公式発表)．本稿では，日本語 WordNet のミスを分類し，それらを機械的に抽出する方法について論ずる．

Detection of Inconsistency in Japanese WordNet

TAKUYA HIRAO[†1]　TAKAHIKO SUZUKI[†2]　KOUKI MIYATA[†1]
SACHIO HIROKAWA[†3]

Conceptual dictionary Japanese WordNet is a useful tool in natural language processing. However, it is officially announced that the Japanese WordNet contains 5% errors. In this paper, we classify errors in the Japanese WordNet and discuss automatic detection methods of the errors.

1. はじめに

日本語 WordNet[1,2]は Prinston WordNet[3] を基に作られた意味辞書である．日本語 WordNet は，自然言語処理研究において有用なツールでありさまざまな研究で利用されている．一般利用者を対象とした Web 上の類義語辞書サービスとして[4]も日本語 WordNet は提供されている．
しかしながら，現在の日本語 WordNet (Wn-Ja 1.1 版)には 5%程度の不整合が含まれていることが公式にアナウンス[2]されており，研究および一般の利用において問題となる可能性がある．

本論文では，日本語 WordNet 上の不整合を分析し，機械的に抽出する手法についての研究状況を述べる．不整合を機械的に抽出することができれば，その修正が容易になり，より信頼性の高い意味辞書作成が可能になる．
本論文において，不整合とは，「語と概念のつながりが成り立たないものでありながら，つながりが存在すると辞書上に誤って登録されているもの」と定義する．

不整合の抽出のために，文書集合（青空文庫の文学小説部）を用い，対象となる語の用例の前後の語（3-gram）の発生頻度を測定する．その分布を，対象となる語の同義語の用例に対する分布と比較することによって，対象の語が不整合であるかどうかを検出する．対象とする語の品詞としては動詞を選択した．

予備実験として，日本語 WordNet の不整合を手動で検出し，不整合の例を蓄積するとともに，不整合の原因を分類した．

関連する研究としては次のようなものがある．多国語間で概念と同義語のつながりを機械的に検出し，WordNet を拡張しようとする試みに Universal Wordnet[5]および BabelNet[6] がある．

秋良ら[7]は，日本語の同義語を自動的に抽出するために，単語を構成する漢字を利用する手法を提案している．鈴木ら[8]は，格助詞で繋がれた単語集合の出現頻度から，ドメイン固有の同義語を抽出する手法を提案している．

WordNet とコーパスを用いて，用例中の単語の意味を確定する試みには例えば[9]がある．

以下，2 節では日本語 WordNet の構造，3 節で不整合の抽出のための予備調査に関して述べる．4 節では不整合抽出実験の構成とその結果，5 節では追加分析，6 節は今後の方針について述べる．

2. 日本語 WordNet および関連事項

日本語 WordNet は，独立行政法人情報通信研究機構（NICT）により作成された日本語の意味辞書である．大規模かつ誰でも使用可能な日本語意味辞書の作成を目指し，2006 年から開発が行われている．基本的な構造は Princeton WordNet に準拠している．しかし，日本語と英語では語彙体系が異なるため，日本語 WordNet ではオリジナルの WordNet の概念に加え，日本語独自の概念を追加している[1,2].

日本語 WordNet (Wn-Ja 1.1 版)に収録された Synset 数, 語数, および語義数は次のようになっている．

・57,238 概念 (Synset 数)
・93,834 words 語

†1 九州大学システム情報科学府
1TE10173M@kyushu-u.jp (平尾)
2ie13089s@s.kyushu-u.ac.jp (宮田)
†2 九州大学情報基盤研究開発センター
{suzuki, hirokawa}@cc.kyushu-u.ac.jp

・158,058 語義 (Synset と語のペア)

日本語 WordNet では，概念と概念の間に上位下位の関係が存在し，階層構造をなす形になっている．また，概念（Synset）はそれぞれ一意の Synset-id が与えられている．Synset には，概念の簡単な定義(概要：gloss) およびその概念を表す同義語集合が付随している．複数の意味を持つ語は，複数の Synset に付随している（図 1）．

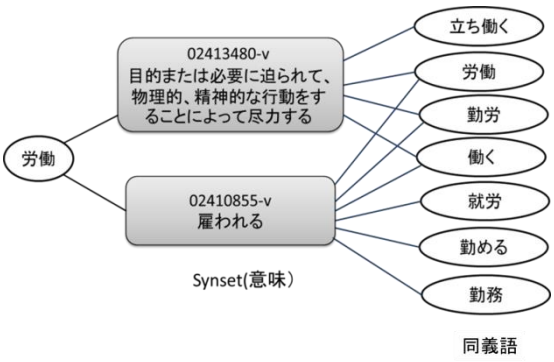

図 1 日本語 WordNet の Synset と語
Figure 1 Synsets and words in Japanese WordNet.

3. 予備調査

語と概念の間に存在する不整合が，どのような傾向を持つかを調べるため，手作業で語と概念のつながりを調査した．

今回は，日本語能力試験[10]に過去出題された動詞のうち，無作為に抽出した 900 語を含む Synset を調査した．

まず，900 語について 1 名の代表検査者がスクリーニングを行い，不整合候補と思われる Synset を抽出した．続いて，4 名の検査者が抽出された不整合候補の Synset について再判定し，全員が不整合と判定したものを最終的に不整合とした．

3.1 結果と傾向

約 9%の 82 語(Synset)において不整合が抽出された．不整合の原因の傾向としては，元の WordNet を和訳する際に，単語を直訳してしまったせいで，関連性の無い概念にもつながりが存在していると判断してしまったものが多いと思われる（翻訳時の誤り）．また，一部ではあるが，不整合の原因がまったく推測できないものも存在した．

3.2 不整合の分類

翻訳時の誤りを検出するためには，原文と翻訳文を人間が比較する必要があり，機械的抽出の判断基準とするには不適切である．そのため，不整合が存在する概念を中心として，その傾向を場合わけした．その結果，大きく分けて 4 種類のパターンが検出された．

タイプ 1 は，概念と日本語の単語（和語）のつながりのうち，整合と不整合の両方が混在しているものである（一部型）．おそらく，語の和訳においてミスをしたためにおきた形態であろう．

タイプ 2 は，概念と和語のつながりが，すべて不整合であるというパターンである（全滅型）．これは，概念の説明文である gloss（概要）の和訳を失敗したことによって起こったものと考えられる．

タイプ 3 は，概念と語のつながりが一通りしかなく，かつそのつながりが不整合であるというパターンである（単独型）．このタイプの発生原因は特定できない．

最後にタイプ 4 は，日本語として見た場合 WordNet の概念と語のつながりにおいて元から不整合が存在していると疑われるパターンである（根源型）．

82 例の不整合のうち，タイプ 1 の一部型が 26 例　タイプ 2 の全滅型が 18 例，タイプ 3 の単独型が 28 例，タイプ 4 が 9 例であった．

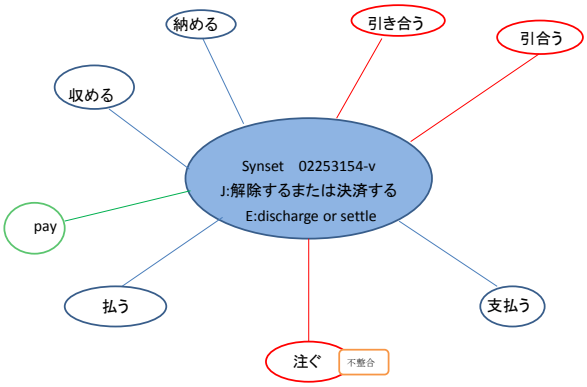

図 2.1 タイプ 1　一部型の例
Figure 2.1 Type-1 Partly-wrong

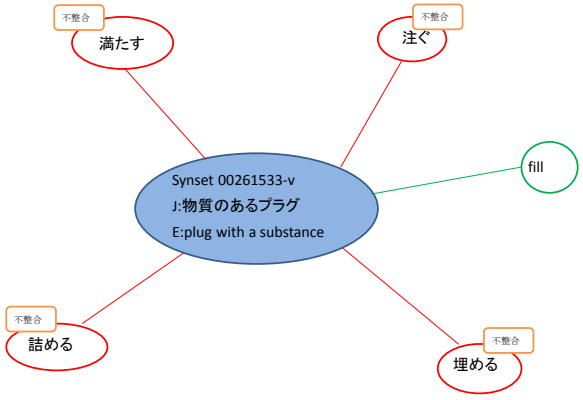

図 2.2 タイプ 2　全滅型の例
Figure 2.2 Type-2 All-wrong

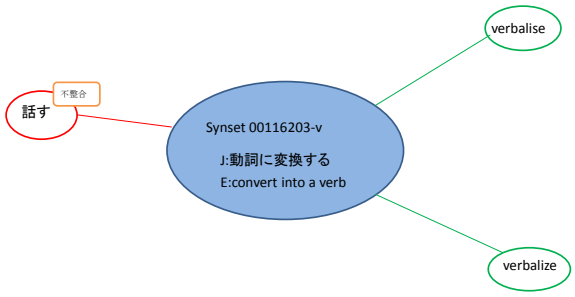

図 2.3 タイプ 3　単独型の例
Figure 2.3 Type-3 One-to-one-wrong

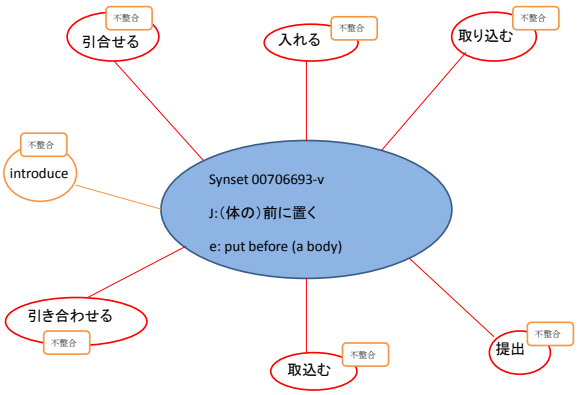

図 2.4 タイプ 4　根源型の例
Figure 2.4 Type-4 Fundamentally-wrong

以上の 4 つのパターンの発生原因を確定することは困難であるが，機械的にパターンを検出し，不整合を抽出することは可能である．

以下の実験では 4 パターンのうち,「パターン 1　一部型」と呼称している不整合のパターンに注目し，このパターンの検出を機械的に行おうとしている．その他のパターンについては，WordNet の階層構造や gloss の情報を利用することで類似した手法が適用可能である．

4. 不整合抽出実験

予備実験で確認した不整合のうち，タイプ 1 (一部型) 26 例について機械的に判定することを試みた．青空文庫[11]を形態素解析したコーパス「青空文庫　形態素解析データ集」(64,247,761 語 4.5Gb)[12] の新字新仮名全データを使用した．

対象となる語が使われている文例において，語の前後 2 語 (3-gram)の形態素の分布を比較する．また分布を比較するにあたり，平仮名一文字の形態素を助詞と判断し取り除いている．対象の分布の近似度を測る尺度として，ナイーブベイズによる判定を試みた[13].

4.1 実験の手順

1) ある Synset S に繋がっている同義語の集合 Ws ={w_1 ,...,w_n} (n >=2)　から対象となる 1 つの語 w_i を取り除く．
2) Ws-{w_i} のそれぞれの語を独立した n-1 個のカテゴリとする．
3) w_k ∈　Ws-{w_i} それぞれについてコーパス中のすべての用例 3-gram に出現する語の分布をカテゴリのベクトルとしてナイーブベイズ分類機を構成する．
4) w_i について同様にベクトルを定義し，n-1 個のどれかのカテゴリに属する確率を計算する．

w_i が他の語 w_k と同義であるならば，そのカテゴリに属する確率が高く，そうでなければ，属する確率が低くなる．w_i がどのカテゴリに属す確率も低い場合に注目すれば不整合が抽出できると考えた．

4.2 実験結果

26 例の一部型不整合において，不整合である語に対して上記実験を行ったところ，すべて，他の同義語のカテゴリに高い確率で分類されるという結果になった．

また，予備調査で不整合が無いと判定された Synset 7 例について，総当り的に同義語を指定して上記の実験を行ったところ，すべて他の同義語のカテゴリに分類される，という結果になった．

5. 追加分析

実験が不調に終わった原因について，(a) 対象となる語の用例 3-gram を使ったベクトル化そのもの　(b) ナイーブベイズによる分類器の性能，どちらに問題があるかを調べるために次のような追加実験を行った．

5.1 追加分析の手順

1) Synset S に繋がっている同義語の集合 Ws ={w_1 ,...,w_n} (n >=2)　に属するそれぞれの語について，用例の 3-gram 中に出現する語の頻度 (%) を計算する．
2) 対象とする語を w_i　とした場合，w_i 以外の Ws の要素 w_k にそれぞれについて，3-gram に出現する上位 10 語まで頻度順に並べ, ヒストグラムを描く．
3) w_i の 3-gram 中の頻度を，w_k の頻度グラフに重ね合わせて描く．

一部型に分類した，Synset 00047317-v　同義語＝{試す，為る，見る，試みる，*try, try on*} (為るが不整合),

gloss =「サイズが合ってよく見えるかどうか衣類を着てみる」について，結果を図 3.1,3.2 に示す．なお,「試みる」についてはコーパス中の全用例が 100 例未満であるため結果から省略している．「為る」の用例の 3-gram 中に,「こと」が高頻度で出現するのに対し,「見る」「試す」に

ついて「こと」の出現頻度には大きな差があるように見える．

比較のため，不整合が無いと思われる Synset 01224744-v={動かす，操作，操縦，操る，コントロール，operate, control} gloss =「手で触って機能させる」について，２つの結果を図 4.1，4.2 に示す．図 3.1 および 3.2 のような見かけ上の差を，ナイーブベイズ分類器は上手く検出できていない．以上から，4 節における実験がうまくいかなかった一つの原因は，ナイーブベイズ分類器の性能にあると推測できる．

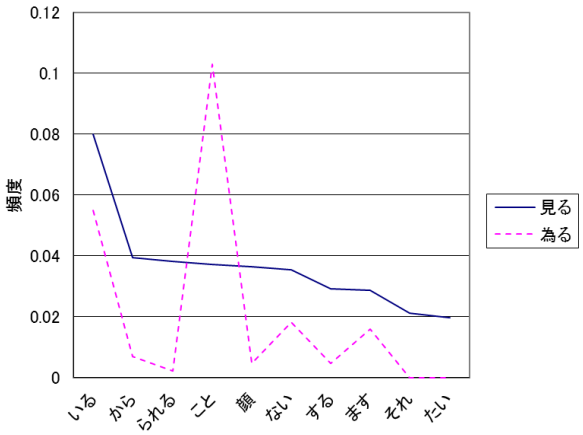

図 3.1 「為る」と「見る」の 3-gram 出現頻度比較
Figure 3.1 Comparison of frequency of words which appear in 3-gram of 為る and 見る

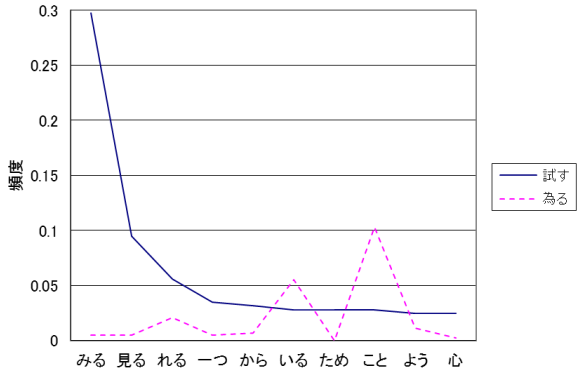

図 3.2 「為る」と「試す」の 3-gram 出現頻度比較
Figure 3.2 Comparison of frequency of words which appear in 3-gram of 為る and 試す

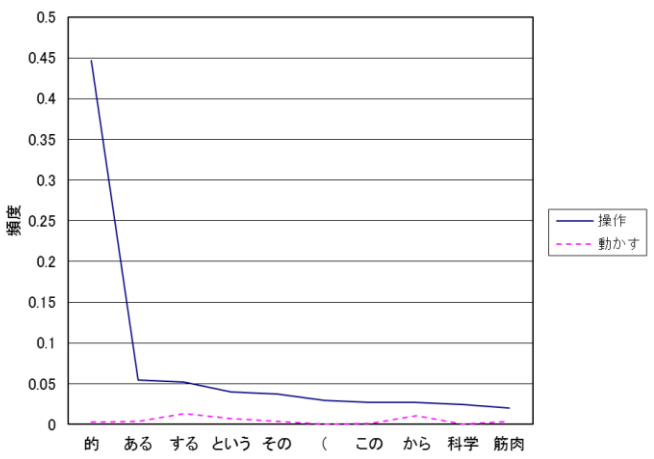

図 4.1 「操作」と「動かす」の 3-gram 出現頻度比較
Figure 4.1 Comparison of frequency of words which appear in 3-gram of 操作 and 動かす

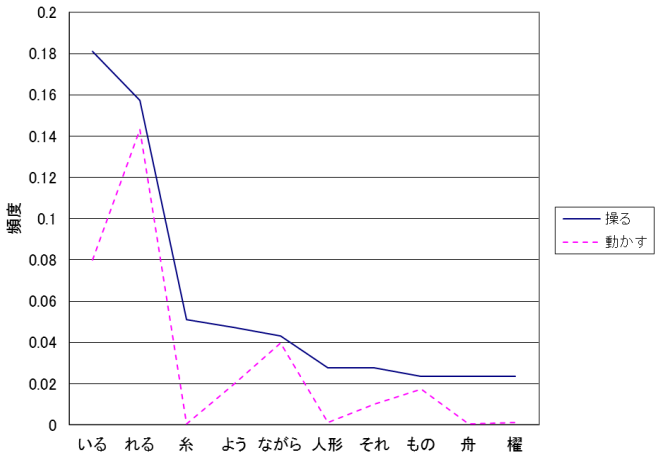

図 4.2 「操る」と「動かす」の 3-gram 出現頻度比較
Figure 4.2 Comparison of frequency of words which appear in 3-gram of 操る and 動かす

6. まとめと今後の方針

本論文の 4 節で行った実験では，日本語 WordNet の不整合を機械的に抽出することはできなかった．5 節の追加分析で示したように，抽出失敗の一因は，使用したナイーブベイズ分類器が，本研究の目的に合致していなかったことにある．

5 節で示した例，および我々が行ったその他の分析例において，不整合がある場合，用例 3-gram 中の語の出現頻度に偏りがある可能性は無視できないものと思われる．

今後の方針として２つをあげる．一つは，分類器の高性能化である．ナイーブベイズ以外に，ノイズに強いとされる SVM[14]を用いた検出器を使うことが考えられる．また，同義語によって用例数が異なるため，それを正規化することも考えに入れる．

もう一つの方針は，他の資源の活用である．EDR[15]等

の他の意味辞書と比較することによって，不整合の検出を効率化できる．しかしながら、この方法を取った場合、参考となる意味辞書が整備されていない言語には応用できない。どのような資源を利用すべきか研究を進めたい。

参考文献

1) Fransis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, Enhancing the Japanese WordNet, ALR7 Proceedings of the 7th Workshop on Asian Language Resources Pages 1-8,Association for Computational Linguistics Stroudsburg, PA, USA.
2) NICT Information Analysis Laboratory, National Institute of Information and Communications Technology, Japanese WordNet, http://nlpwww.nict.go.jp/wn-ja/index.en.html
3) Princeton University "About WordNet." WordNet. Princeton University. 2010, http://wordnet.princeton.edu
4) Weblio 類語辞典, http://thesaurus.weblio.jp/content/
5) Gerard de Melo, Gerhard Weikum, Towards a Universal Wordnet by Learning from Combined Evidence, CIKM '09 Proceedings of the 18th ACM conference on Information and knowledge management Pages
6) Navigli, R., Ponzetto, S.P. , BabelNet: Building a Very Large Multilingual Semantic Network, ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, pp. 216-225
7) 秋良直人, 森本康嗣, 共起漢字ベクトルを用いた同義語抽出方式, 情報科学技術フォーラム一般講演論文集 3(2), 189-190, 2004-08-20
8) 鈴木良弥, 既存の汎用辞書を利用したオンデマンド概念辞書の構築と質問応答システムへの利用, 電気通信普及財団 研究調査報告書 No.28 2013, pp. 493-504
9) George A. Miller, Martin Chodorow Hunter, Using corpus statistics and WordNet relations for sense identification, Computational Linguistics - Special issue on word sense disambiguation archive Volume 24 Issue 1, March 1998 pp. 147-165
10) Japan Foundation, Japan Educational Exchanges and Services, Japanese Language Proficiency Test, http://www.jlpt.jp/e/index.html
11) 青空文庫, http://www.aozora.gr.jp/
12) 青空文庫形態素解析データ集, http://aozora-word.digiweb.jp/index.html
13) 持橋大地, 隅田英一郎, 階層 Pitman-Yor 過程に基づく可変長 n-gram 言語モデル, 情報処理学会論文誌 48(12), 4023-4032, 2007-12-15
14) Brendan Flanagan, Chengjiu Yin, Kiyota Hashimoto, Sachio Hirokawa, Clustering English Writing Errors based on Error Category Prediction, Proceedings of the 3rd International Symposium on Engineering, Energy and Environments, pp. 733-739, 2013
15) EDR 電子化辞書 http://www2.nict.go.jp/out-promotion/techtransfer/EDR/JPN/TG/Doc/EDR_J01a.pdf