日本語WordNet類義語の誤り検出 : コーパス利用の試み

平尾, 拓也
九州大学大学院システム情報科学府

宮田, 光樹
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

廣川, 佐千男
九州大学情報基盤研究開発センター

出版情報：電子情報通信学会技術研究報告 : 信学技報. 114 (339), pp.13-18, 2014-11
バージョン：

一般社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

日本語 WordNet 類義語の誤り検出
ー コーパス利用の試みー

平尾　拓也† 　宮田　光樹† 　鈴木　孝彦‡ 　廣川　佐千男‡

†九州大学大学院システム情報科学府　〒819-0395　福岡市西区元岡 744 番地
‡九州大学情報基盤研究開発センター　〒812-8581 福岡県福岡市東区箱崎 6-10-1
E-mail: †{2ie14071e, 2ie13089s}@s.kyu-shu-u.ac.jp, ‡{hirokawa, suzuki}@cc.kyushu-u.ac.jp

あらまし　日本語 WordNet は自然言語処理において有用なツールであるが、5%の間違いが存在すると公式に認められている。結果として、言語処理のデータベースとして信頼性の面に疑問が残る。本論文では、日本語 WordNet 内の間違いを抽出するいくつかの手法を提示する。

前半では日本語 WordNet それ単体のみを使用した手法を提示し、その結果を記述する。後半では日本語 WordNet と、外部より準備したコーパスを使用し、構造上の間違いを抽出する手法を提示する。

キーワード　シソーラス,　WordNet,　日本語 WordNet,

Detection of Error Synonyms in Japanese WordNet
－A trial of using corpus－

Takuya HIRAO† 　Kouki MIYATA† 　Takahiko SUZUKI‡ 　Sachio HIROKAWA‡

†Kyushu University Graduate School of Information Science and Electrical Engineering　774, Motooka, Nishi-ku, Fukuoka, 819-0395 Japan
‡Kyushu University Research Institute for Information Technology 6-10-1, Hakozaki, Higashi-ku, fukuoka, 812-8581 Japan
E-mail: †{2ie14071e, 2ie13089s}@s.kyu-shu-u.ac.jp, ‡{hirokawa, suzuki}@cc.kyushu-u.ac.jp

Abstract　Lexical Database the Japanese WordNet is a useful tool in natural language processing. However, it is officially announced that Japanese WordNet contains 5% errors. In this paper, we discuss error detection methods in the Japanese WordNet.

キーワード　Thesaurus,　WordNet,　Japanese WordNet,

1. はじめに

日本語 WordNet[1,2]は Princeton 大学が開発した WordNet[3]を用いた言語データベースである。日本語 WordNet は自然言語処理において有用であり、様々な実験に使用されている[4][1]。フリーの Web シソーラスサービスにおいて、日本語 WordNet は一般的に使用されている。しかしながら、現行の日本語 WordNet は間違いを 5%ほど含んでいると作成者らが認めており[2]、それらの間違いが日本語 WordNet の使いやすさに影響を及ぼしている可能性がある。

本論文では、われわれが検証した日本語 WordNet の間違い探知手法において議論する。間違い探知は日本語 WordNet の間違い修正の第一段階である。この手法は、大規模言語データベースの作成に有用であると考える。我々は特に日本語 WordNet の似たような間違いの発見を主眼にしており、この間違いのことを「類義語の間違い」と呼んでいる。

英語でない WordNet や WordNet に似た言語データベースの作成という点において、複数のプロジェクトが行われている。日本語 WordNet や Chinese Open WordNet[5]は、ブートストラップの段階で、Princeton WordNet のマッピング手法を用いて半自動生成されている。

また、Universal WordNet[6]や Babel Net[7]、Open Multilingual WordNet[8]といった、WordNet の拡張による統合、多言語概念字句データベースの生成の試みもなされている。概念と語句、または複数の概念間の関係は、Wikipedia やタグ付けコーパスのような様々な資源から自動的に抽出することが可能である。それらに

[1] Weblio, http://ejje.weblio.jp

よって得られた統合データベースの品質は、生成者自身や、ネットワークコミュニティによって評価されてきた。

WordNet は、オントロジーのひとつとしてみなすことができる。多言語 WordNet を生成する場合には、言語数に応じたオントロジー間のマッピングをする必要がある。そのため、オントロジーの間違いの検出と修正、オントロジー間のマッピングに関する研究がなされてきた。これらの研究において、オントロジー内で分類が間違っているものや、冗長もしくは不適切である、または間違った関係性を生成されている箇所を修正する試みがなされてきた。

間違い検出の手法として、日本語 WordNet のみを使用した手法を動詞に適用した場合をベースラインとして提示する[9]。また、コーパスを用いた単語をベクトル化し、これらのコサイン類似度によって名詞の間違い検出の手法として使用できないかを議論する。

本論文では、第 2 節で WordNet と日本語 WordNet の説明、第 3 節で本論文のコンセプトと WordNet の構造における「同義語の間違い」の一例を紹介する。第 4 節では間違いの抽出法に関するわれわれの手法の説明、第 5 節では手法を用いた場合の結果の提示を行う。第 6 節では、本手法の Princeton WordNet における応用例と、関心を持っている別手法に関しての説明、第 7 節で word2vec を用いた単語のベクトル化とそれらを用いた間違い検出の実験、第 8 節に今後の展望と課題を述べる。

2. WordNet と日本語 WordNet

2.1. Princeton WordNet

Princeton WordNet は英語の大規模言語データベースである。名詞、動詞、形容詞、副詞といった品詞ごとに、明確なコンセプトを持った「Synset」という認知同義語のセットに纏められる。各 Synset は固有の ID によって管理されており、Gloss と呼ばれる、Synset の簡単な意味を説明するテキストがリンクされている。Synset は概念-意味関係もしくは字句トークン関係で相互リンクを持っている。単語が持つ意味を Synset によってグループ化することができるため、WordNet はシソーラスとして使用できる。多義である単語が存在するため、単語は複数の Synset に属することがある。

2.2. 日本語 WordNet

日本語 WordNet は Princeton WordNet を基にした、日本語の語彙データベースである。日本語 WordNet のプロジェクトの目的は、誰でも自由に使用可能な大規模日本語データベースを提供することである。このデータベースは 2006 年から開発されている。

日本語 WordNet の構造は、Princeton WordNet に準拠している[1]。しかし、日本語と英語という言語の違いが存在するため、日本語 WordNet は Princeton WordNet に含まれていないオリジナルの Synset を含んでいる。また、日本語 WordNet は、シソーラスとしての精度より多数の概念を包括することに主眼を置いている。

現行の日本語 WordNet の規模は以下のとおりである。

- 57,238 概念（Synset 数）
- 93,834 語（日本語）
- 158,058 語義（単語-synset ペア数）

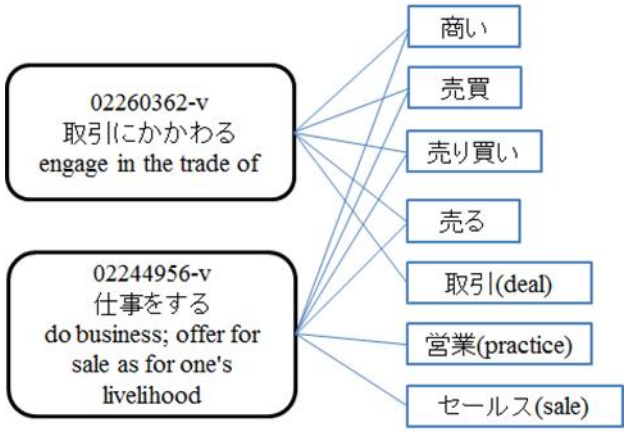

図 1 日本語 WordNet の Synset-同義語間リンク例

日本語とリンクを持つ Synset は日本語の gloss を持っている。日本語 WordNet のカバー範囲の拡張のために、SUMO や Wikipedia、GoiTaikei[10]といった他のリソースが使用されている。

2.3. 他言語の WordNet と WordNet の拡張

Princeton WordNet を基にした、様々な言語の言語データベース作成プロジェクトが存在する。一部のプロジェクトでは WordNet、Wikipedia[2]、Wiktionary[3]及びその他の言語資源を用いて、多言語のごくデータベースを作成しようと試みている。

既存の言語資源と新しいデータベース間のマッピングの正確さは、それによって出力されるデータベースの整合性の正しさを証明する指標になるので非常に重要である。新しい言語の WordNet を作成することは、他の言語からなる新しいオントロジーで表現されている、既存のオントロジーからマッピングで作成するとみなすことができる。

[2] Wikipedia, http://ja.wikipedia.org

[3] Wiktionary, http://ja.wiktionary.org

3. 日本語 WordNet の間違い

間違いの訂正は、新しく作成したオントロジーや、オントロジー間のマッピングの整合性の確認において重要である。日本語 WordNet の現行のバージョンでは、約 5%の間違いが含まれている。また、Chinese Open WordNet も、それに匹敵するエラー率である。本節の残りでは、同義語における間違いにおける、エラーの種類に焦点を当てる。

3.1. 同義語の間違い

WordNet の構造において、「同義語の間違い」を、語 w_{miss} が属している synset（S とする）の Gloss と合致しない語であると定義する。

図 2 では、Synset 02651424-v について図示している。この Synset は「泊める」、「収容」、「宿る」、「持ち込む」という 4 つの同義語を持っている。

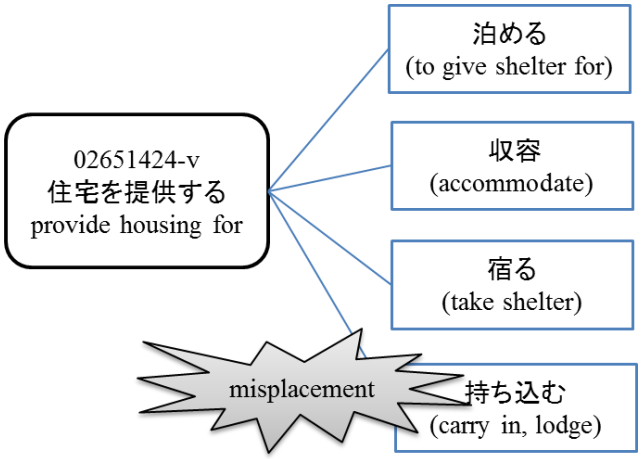

図 2 同義語の間違いの具体例

このうち、「持ち込む」という単語はこの Synset 02651424-v において同義語の間違いであるといえる。

3.2. 予備実験

我々は日本語 WordNet 内の間違いを手動でチェックした。今回対象としたのは動詞である。理由としては、WordNet の構造内での間違いは、名詞より動詞のほうが多く報告されていたためである。

以下は、今回の実験で間違いを確認した際の手順である。

1) 日本語能力検定(JLPT)に登場する単語のうち、WordNet に登録されているものを無作為に 900 語抽出する。
2) 抽出した単語の同義語を含む Synset を日本語 WordNet から抽出する。
3) スクリーニング担当者がすべての Synset と動詞とペア関係をチェックした。担当者はその後、間違いの確認用のリストを作成した。
4) スクリーニング担当とは別のチェック担当者たちが独立して確認を行った。チェック担当者全員が間違いと判断した Synset と単語のペアを最終的な間違いとしてマークした。

3.3. 間違いの種類

結果として 900 単語中、81 語（9%）が間違いだと判断された。それらは 3 つのエラーパターンに分類された。

Synset S 内のすべての日本語の同義語を Syn(S)と表現する。また、Synset S 内の間違いを mis(S)と表現する。

・Syn(S)内の同義語が一つでなく、Syn(S)=mis(S)である場合、S には全部型の間違いが存在していると呼称した。

・Syn(S) / mis(S) ≠ ϕ かつ mis(S) ≠ ϕ であるとき、この Synset S は一部型の間違いが存在していると呼称した。

・Syn(S)内の同義語が一つしかなく、Syn(S)=mis(S)である場合、S には単独型の間違いが存在していると呼称した。

全 81 個の間違いのうち、26 個が一部型、27 個が全部型、28 個が単独型の間違いであった。この種類を数える際、我々は 81 個の単語と Synset の間のリンクを対象とした。

4. 間違いの抽出方法

我々は日本語 WordNet 単体で間違いを抽出する方法を試した。われわれの手法は日本語 WordNet 以外の、限定的な情報しか持たない WordNet 構造のデータベースにも使用することができる。

4.1. Synset-同義語間リンクによる抽出

はじめに提示する抽出手法は、Synset と同義語のリンクのみを用いた手法である。以降の記述では Synset がリンクを持っている単語を w と定義する。

単語 w と Synset S においての、w の重複 Synset SC は以下の式で表現される。

$$SC(w) = \{Sk \mid w \in syn(Sk)\}$$

Synset 重複は、単語 w とリンクを持つすべての Synset-ID について定義される。図 3 に SC(売る)を例として図示する。

図 1 での単語である商う，売り買い，売買，売る は同じ Synset 002260362-v と 002244956-v の 2 つにリンクを持っている。つまり、SC(商う), SC(売り買い),

SC(売買) ,SC(売る) は 2 個あるいはそれ以上の要素を持っていることになる。

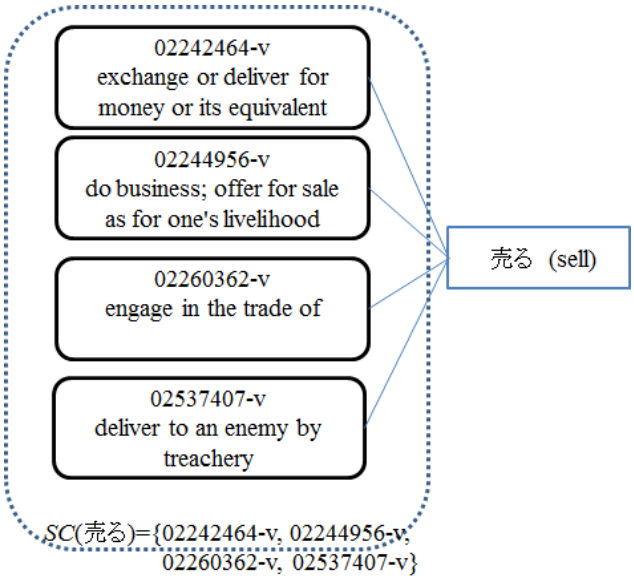

図 3 SC(売る)の図

4.2. Gloss-同義語リンクによる抽出

二つ目の手法は gloss に含まれている単語を使用する。

Synset S に含まれている gloss と文例に存在するすべての日本語の単語を glossw(S)={w1,w2,...}.とする。単語 w における gloss-coverage (GC) は以下の式で表される。

$$GC(w)= \{uk| uk \in glossw(Sj), Sj \in SC(w)\}$$

SC(w1) と SC(w2)が同じ要素を持つとき、 GC(w1)と GC(w2)も同様に同じ要素を持つことは自明である。

Gloss overlap は Synset overlap と似たような傾向を持つ。

4.3. 検証手法

1)から 4)という順序で、SC(w)を用いた間違いの抽出手法を記述していく。

1) 抽出手法を適用する対象の Synset をリスト化する。
2) リスト上の Synset Si について、各ペア wk, wj ∈Syn(Si)の Synset Overlap 、SOL(wj,wk)を計算していく。
   SOL(wj,wk)=#(SC(wj) ∩ SC(wk))/ #(SC(wj) ∪ SC(wk))
   (#(S) は S の基数を示す)
3) SOL(Si)の最小値をとる。
   mSOL (Si) = min (SOL (w_j,w_k)) (for all w_j,w_k ∈ Si)
4) mSOL(Si) < γ.である場合、間違いの可能性があるとタグ付けする。
   γ は閾値である。(γ < 1)

GC(w)を用いた手法は、SC(w)をそれぞれ SC(w), SOL(w_j,w_k), and mSOL(Si) から GC(w), GOL(w_j,w_k) (Gloss Overlap)、mGOL(S_i) (Synset S の gloss overlap 最小値) に置き換えたものとなる。

4.4. 手法の適用性

4.3 の手法を適用するには、Synset が一定の状態を満たしている必要がある。

状態 4.3

・Synset 内には 2 つ以上の単語が登録されている必要がある。

・われわれの手法は単独型の間違い抽出に使用することができない。

・仮説が絶対的なものではない。たとえば「切る」と「混ぜる」は Synset 01418667-v (ランダムな順序や配置になるように混ぜる)で同義語であるが、他の Synset で同時に出現することが起こっていない。

5. 日本語 WordNet 内での一部型の間違いにおける結果

日本語 WordNet には動詞を表す Synset が 10,324 個存在している。そのうち、検証手法の対象となるのは 3,031 個であった。2 名の検証者がそれらのデータの間違いを手動でチェックしていき、一部型であるか全部型であるかの確認も行っていった。結果として、125 個の全部型の間違いと 121 個の部分型の間違いが発見された。

図 7 は、一部型の mGOL(S)の Precision、Recall、F 値を示している。横軸は mGOL(S)の大きさを昇順に並べたものをとっている。

F 値は γ =0.0455 の時に最大となっている。表 1 は具体的な数値を示している。図 7 の縦線は F 値が最大となっているところをあらわしている。

表 1　F 値が最大値をとるときのデータ

	Synset rank	γ	Precision	Recall	F
mSOL	331	0.0526	0.195195	0.481481	0.277778
mGOL	313	0.0476	0.191693	0.495868	0.276498

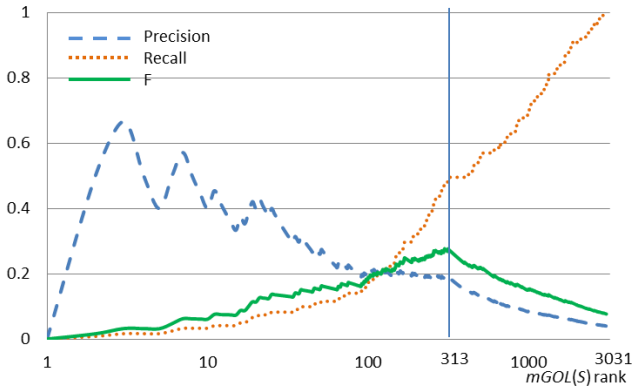

図 7　mGOL(S)の結果

6. 詳細結果

6.1. 全部型の間違い

mGOL(S)値の数値が高くなれば全部型の間違いは少なくなっていっているように見えるが、この手法では全部型の抽出をうまくできていない。

全部型の間違いの原因について、我々は 2 種類の分類をしている。

1) 以下の条件を同時に満たす場合

・英語の同義語が Synset 内に存在しており、それが一般的によく使われるものである(get など)。

・Synset 内に存在している英語の同義語の数が少ない。

・英語の同義語の gloss が一般的にあまり使われないものである。

・日本語の同義語が英語から取られているように推測される。

2) 日本語の gloss が誤訳されているため、日本語の同義語がマッチしない状態になっている。

われわれの手法では 2)の間違いは抽出することができない。理由としては日本語 gloss のみが間違っているため、同義語同士の不整合が見られないことが多いためである。

6.2. mSOL(mGOL)の低スコアに対する誤検知の原因推定

mSOL(S)や mGOL(S)のスコアが低くとも、間違いでない同義語が存在している。最小値が 500 位以内の Synset のうち、人間が間違いと判断しなかったものが 392 個ある。

一定の条件を満たすと、mGOL(S)や mSOL(S)が低くなるという現象を確認している。説明のために、mSOL(S)の例を挙げる。

条件 6.2

Synset 内の同義語の数が 2 つしかなく、同義語がそれぞれより多くの Synset とリンクを持っている場合。

この条件を満たした場合、mSOL(S)や mGOL(S)が低くなる現象が発生する。

mSOL(S)が 100 位以内の Synset 中に、同じ同義語が繰り返されているものがあった。表 2 は特に発生していた 4 単語を抜き出したものである。

表 2　頻出した単語の出現傾向

	Times of appearing	Partially wrong	#SC(w)
切る(cut)	14	1	31
掛ける(cover)	10	4	23
考える(think)	7	0	24
為る(become)	7	4	21

表 2 の単語は複数の意味を持つことは自明である。「切る」と「考える」は多くの低 mSOL(S)スコアを持っており、誤検知を引き起こしている。

6.3. 高 mSOL(mGOL)値を持つ間違い

mSOL(S)(mGOL(S))値が高い場合でも間違いが存在するパターンが発見されている。mSOL > 0.1 を満たす間違いは 29 個発見されており、うち 25 個では SC(w1) ∩ SC(w1) がひとつしか存在していなかった。

これは 6.3 と逆の現象であり、Synset 内に含まれている同義語の数が少ないことから発生したものと考えられる。

7. ベクトルを用いた間違い抽出法の検証

word2vec[11]を使い、コーパス中の用例について単語をベクトル化することができる。我々はこれを用いて日本語 WordNet の間違い抽出が可能ではないかと考えた。

青空文庫[4]の新字新仮名作品に対して形態素解析を行ったコーパス[5]が公開されており、これを用いて単語をベクトル化した。

その後、動詞の Synset に存在する単語のペアに対して、生成したベクトルを用いてコサイン類似度を求め、その数値を先述した mGOL のように指標として用いた。

動詞による検証結果はベースラインとした mGOL の結果よりもやや低い F 値の推移となっていた。

コーパスを調べたところ、サ行変格活用である動詞が名詞と混同されていることが判明したため、再度名詞のみを対象として検証を行った。

名詞による検証に使用した Synset 数は約 7000 個で、そのうち間違いを含んでいると判断した Synset は 161 個だった。

結果として、F 値は一部を除いて 0.1 を超えることはなく、このままでは間違い抽出への利用は不適切である。

なお、名詞における mGOL においても F 値はベクトルを用いた際と大きな差はなかった。

8. 課題と今後の展望

本論文では、WordNet の構造に存在する「同義語の間違い」の抽出を目標としてきた。結果として、mGOL は、日本語 WordNet で「一部型」と定義した種類の動詞の間違いの抽出に効果を示した。最小 Gloss Overlap を使用することで、313 個の Synset 中に 50%の一部型の間違いを集約することを可能にしている。また、特別に新しく情報資源を準備する必要がなく、日本語 WordNet それ単体で手法を実行することが可能である。

ベクトルを用いた間違い抽出に関しては、コサイン類似度を用いて類似性を判断した。今回の検証では間違いとそうでないものの区別がついておらず、課題が残る。

第一はコサイン類似度の大きさが類義語の正しさと相関関係を持っているのかを正しく判断しなければならない。Word2vec では、コサイン類似度が大きければ語は似たような意味を持つ[11]とあるが、すべての類義語がペアとなった際にコサイン類似度が高いとは限らない。類似性の判定には他の分類手法も試すべきであると考える。

第二に、語の多義性の問題がある。今回は単語に対してベクトルを抽出したが、単語の中には複数の意味を持つものが存在しており、ベクトルが意味関係の方向性を含んでいるならば、多義性を持つ単語であれば多義性を持つベクトルが出力されている可能性があると考える。

また、今回使用したコーパスである青空文庫は著作権切れの古い作品が多く、Synset 内の単語がコーパス内に存在しないという現象も起こっている。

これらのことから、ベクトルを用いた検証に関しては改良の余地が多いと考えている。

文　献

[1] F. Bond, H. Isahara, S. Fujita, K. Uchimoto,T. Kuribayashi, Enhancing the Japanese WordNet, *ALR7 Proc. the 7th Workshop on Asian Language Resources*, pp. 1-8 ,*Association for Computational Linguistics*. pp. 1-8, 2009

[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] K. Miyata, et al., Difficulty and Ambiguity of Verbs-Analysis based on Synsets in Japanese WordNet-, AIT2013, 2013

[5] S. Wang, F. Bond, Building the Chinese Open WordNet (COW): Starting from Core Synsets, *Proc. International Joint Conference on Natural Language*, pp. 10–18, 2013

[6] G. Melo, G. Weikum, Towards a Universal WordNet by Learning from Combined Evidence, *CIKM '09 Proc. 18th ACM conference on Information and knowledge management*, pp. 513-522, 2009

[7] R. Navigli, S.P. Ponzetto, BabelNet: Building a Very Large Multilingual Semantic Network, *ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics Proc.*, pp. 216-225, 201

[8] F. Bond, R. Foster, Linking and Extending an Open Multilingual WordNet, *in Proc. 51st Annual Meeting of the Association for Computational Linguistics*, pp. 1352-1362, 2013

[9] T. Hirao, T. Suzuki, K. Miyata, S. Hirokawa, Detection Methods for Misplacement of Synonyms in the Japanese WordNet, International Journal of Computer and Information Science, to appear

[10] S. Ikehara, et al., *"Nihongo GoiTaikei [Japanese Lexicon]"*, Iwanami Shoten, in Japanese , 1997

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

[4] 青空文庫, http://www.aozora.gr.jp

[5] 青空文庫形態素解析データ集, http://aozora-word.hahasoha.net