Vol.2010-NL-199 No.1
2010/11/18

係り受け木を利用した単語類似度計算方法とそのシソーラス拡張への応用

鈴木　郁美[†1]　　原　　一夫[†1]
新保　　仁[†1]　　松本　裕治[†1]

本研究では並行ランダムウォークによりグラフ上の節点の類似度を測る方法を，自然言語処理のタスクの一つである，コーパスからの単語類似度測定に応用する．標準的な手法は，各々の単語に対し，コーパスから周辺単語を抽出する．その上で抽出した周辺単語から特徴ベクトルを作成し，比較したい単語対それぞれの特徴ベクトルが成す角度のコサイン (コサイン類似度) などを用いて単語対の類似度を測る．本研究では並行ランダムウォークにより単語類似度を測ることで, 周辺単語の情報と係り受け木の構造の両方を考慮に入れて類似度の改良を試みる．具体的には，注目する 2 つの単語をそれぞれ含む 2 つの係り受け木において，注目する単語をそれぞれ出発点とするランダムウォークを並行して行い，2 つの単語の類似度をウォークの類似度の重み付き和として定義する．MeSH シソーラスと GENIA コーパスを用いた実験で，並行ランダムウォークを用いた手法はコサイン類似度による手法を上回る結果を得た．

A New Word Similarity Measure Capturing Dependency Tree Structure and Its Application to Thesaurus Expansion

Ikumi Suzuki,[†1] Kazuo Hara,[†1]
Masashi Shimbo[†1] and Yuji Matsumoto[†1]

A new word similarity measure is presented. Generally, bag-of-words model is applied to construct feature vectors. And consine similarity is widely used to measure word similarity in various natural language processing applications. In this paper, word similarity is measured not only by bag-of-words model but also by considering dependency tree structures. In the proposed method, similarity of two words is obtained by random walk in the dependency tree structures. As starting the corresponding nodes of the words, the similarity is calculated as the sum of weighted walk-paths in the dependency trees. As a result, the proposed similarity measure outperformed conventional consine similarity in thesaurus expansion task.

1. はじめに

コーパスを用いて計算される単語類似度は，自然言語処理の多くのタスクにおいて解析精度向上のための手がかりの一つとして利用されている．たとえば，語義曖昧性解消タスクにおいては，周辺単語との共起の分布をもとに計算する単語類似度が有効な手がかりとなることが以前から知られている[1)]．また，シソーラス構築ではコーパスに現れる文脈の類似度をもとに兄弟（あるいは従兄弟）関係を捉えようとする研究があり[8)]，構文解析や照応解析においても大規模コーパスから獲得した単語の類似度（選択選好）を用いて確率モデルを構築する研究がなされている[6),7)]．

コーパスにおける出現文脈の類似性（分布類似度）を利用して，単語類似度を計算する標準的な手法は，各々の単語の特徴として周辺単語を用いる方法である．すなわち，各々の単語に対し，あるウインドウサイズで周辺単語を bag-of-words として抽出し，特徴ベクトルを作成する．その上で，2 つの単語の類似度として，たとえばそれぞれの単語の特徴ベクトルが成す角度のコサインと定める（コサイン類似度）．

単語の特徴として周辺単語を用いる類似度計算方法は，計算コストが少ない反面，単語が出現する文脈の文法的，意味的な構造を類似度計算に反映していない点に改善の余地があると考えられる．たとえば，2 つの単語が類似する係り受け木の構造の中で出現している，あるいは，類似する述語項構造のなかで同じ意味役割を付与されている，といった情報を類似度計算に利用することができない．

そこで本研究では，周辺単語の情報と係り受け木の構造情報の両方を考慮して，単語類似度を計算する手法を示す．先行研究として，グラフ上の節点間の類似度を計算する Desrosiers らによる手法[2)] がある．Desrosiers らは，鹿島らによるグラフの類似度を測る手法[4)] をヒントとして，2 つの節点の類似度を測るためにそれぞれの節点を出発点とするランダムウォークをグラフ上で並行して行い，ウォーク経路の類似度の重み付き和（期待値）を計算することで節点間の類似度とした．本研究における単語類似度計算法は，Desrosiers らによる手法の係り受け木のグラフに対する適用である．すなわち，2 つの単語の類似度を，それぞれが出現する文の係り受け木のグラフにおいてその単語を出発点とするランダムウォークを並行

†1 奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

Vol.2010-NL-199 No.1
2010/11/18

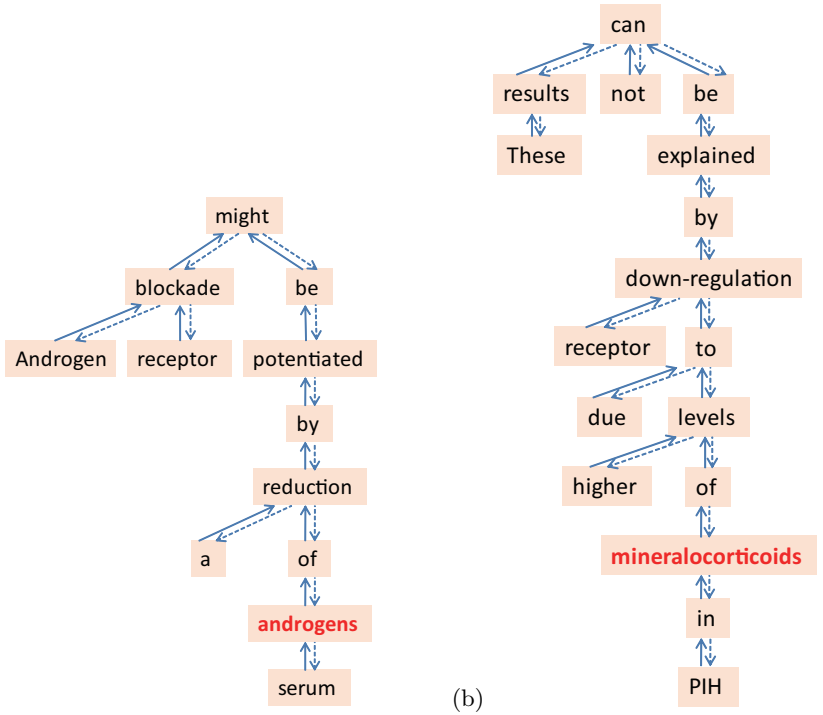

図 1 2 節で用いる例 (a) 文 S, (b) 文 S' に対する係り受け木のグラフ．実線は forward ラベル，破線は backward ラベルの枝である．

して行い，ウォーク経路の類似度の重み付き和として求める．

本論文の構成は以下の通りである．並行ランダムウォークによる単語類似度計算方法について 2 節で述べた後，評価タスクとして行うシソーラス拡張の説明を 3 節で行う．MeSH シソーラスと GENIA コーパスを用いて行った評価実験の結果を 4 節で報告し，5 節で関連研究を紹介する．最後に 6 節で今後の課題について述べるとともに本論文をまとめる．

2. 並行ランダムウォークによる単語類似度計算方法

2.1 係り受け木における並行ランダムウォーク

類似度を測る対象となる 2 つの単語を v_i, $v'_{i'}$ とし，それら単語が出現する文をそれぞれ

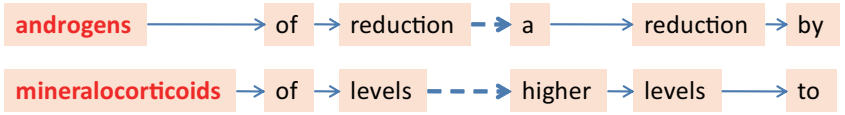

図 2 図 1 の係り受け木のグラフにおける長さ 6 の並行ウォークの例

S, S'，それら文に対する係り受け木のグラフをそれぞれ T, T' とする．係り受け木のグラフ T, T' は，文 S, S' の単語集合 W, W' を節点集合としてそれぞれ持ち，係り受け関係にある 2 つの単語 v_i, v_j（または $v'_{i'}$, $v'_{j'}$）に対応する節点間に双方向の枝 $e_{i,j}$, $e_{j,i}$（または $e'_{i',j'}$, $e'_{j',i'}$）を持つとする（ここで，節点と枝の下付き添字は，係り受け木のグラフにおけるインデックスを表す）．そして，係る方向と順向きの枝には forward ラベル，逆向きの枝には backward ラベルを付与する．次の 2 つの文：

文 S Androgen receptor blockade might be potentiated by a reduction of serum androgens.

文 S' These results can not be explained by receptor down-regulation due to higher levels of mineralocorticoids in PIH.

に対する係り受け木のグラフを図 1 に例示する．ここで，androgens および mineralocorticoids を，類似度を測る対象の単語対とする．

2 つの単語 v_i, $v'_{i'}$ の類似度を計算するために，係り受け木のグラフ T, T' において，それぞれ v_i, $v'_{i'}$ を出発節点とするランダムウォークを並行して行う．ランダムウォークは，停止確率を γ，節点 v_i からの遷移確率を $\frac{1-\gamma}{d_i}$ として行う．ただし，d_i は節点 v_i の出次数である．図 1 の係り受け木のグラフにおける長さ 6 の並行ウォークの例を図 2 に示す．

2.2 ウォークの類似度

係り受け木のグラフ T, T' における長さ L のウォークをそれぞれ

$$walk = v^1 e^{1,2} v^2 e^{2,3} \cdots v^L$$
$$walk' = v'^1 e'^{1,2} v'^2 e'^{2,3} \cdots v'^L$$

と書き（節点と枝の上付き添字は，ウォークにおけるインデックスを表す），ウォークの類似度 $Sim^L(walk, walk')$ を次の式で計算する．

$$Sim^L(walk, walk') = K_v(v^1, v'^1)K_e(e^{1,2}, e'^{1,2})K_v(v^2, v'^2)\dots K_v(v^L, v'^L)$$
$$= \prod_{l=1}^{L} K_v(v^l, v'^l) \prod_{l=1}^{L-1} K_e(e^{l,l+1}, e'^{l,l+1})$$

ここで，$K_v(v^l, v'^l)$ は 2 つの節点 v^l, v'^l に対応する単語対の類似度であり，$K_e(e^{l,l+1}, e'^{l,l+1})$ は 2 つの枝 $e^{l,l+1}$, $e'^{l,l+1}$ のラベル対の類似度である．単語対の類似度 $K_v(v^l, v'^l)$ の与え方については，4 節で述べる．ラベル対の類似度は，ラベル（forward, backward）が一致するとき 1 ，一致しないとき 0 として与える．なお，長さが異なるウォーク対の類似度は 0 とする．

2.3 単語類似度計算アルゴリズム

本研究では，2 つの単語 v_i, $v'_{i'}$ の類似度 $Sim^L(v_i, v'_{i'})$ を，それら単語をそれぞれ出発節点とする長さ L 以下のウォーク対の類似度の重み付き和（期待値）として計算する．並行ランダムウォークによる単語類似度計算のアルゴリズムは，Desrosiers らによる手法[2] に則って行う．その計算アルゴリズムを Algorithm 1 に示す．

Algorithm 1 の 4 行目で行われる再帰計算の解釈を説明する．まず右辺の第 1 項について説明すると，類似度を計算したい単語ペア $v_i, v'_{i'}$ について，$v_i, v'_{i'}$ を出発節点として，長さゼロ ($l = 0$) のウォークの類似度，$Sim^0(v_i, v'_{i'})$，を求める．これは，出発節点に止まっていることに対応し，出発節点の節点類似度に，その節点で止まる確率で重みつけたものになる ($K_v(v_i, v'_{i'}) \times \gamma^2$)．右辺の第 2 項では，出発節点 $v_i, v'_{i'}$ から 1 歩以上 l 歩以下のウォークの類似度を求めている．$Sim^{l-1}(v_j, v'_{j'})$ は，$v_j, v'_{j'}$ を出発節点として長さ $l-1$ 以下のウォークの類似度である．$v_i, v'_{i'}$ に隣接するすべての節点 $v_j, v'_{j'}$ に関して，$v_i, v'_{i'}$ から 1 歩進む類似度を求め，$Sim^{l-1}(v_j, v'_{j'})$ と掛け合わせることで，$v_i, v'_{i'}$ を出発節点とする 1 歩以上 l 歩以下の類似度を計算できる．$l = 1$ から再帰的に $Sim^l(v_i, v'_{i'})$ を求めることで，最終的に目的の長さ L 以下のウォーク対の類似度を計算できる．

2.4 周辺単語を bag-of-words と見る手法との違い

並行ランダムウォークにより類似度を獲得する手法は，周辺単語を bag-of-words と見る手法と異なり，周辺単語の情報と係り受け木の構造情報の両方を考慮して，単語類似度を計算することができる．それについて 2.1 節で示した例文を用いて説明する．例文中の androgens(アンドロゲン) と mineralocorticoids(鉱質コルチコイド) はホルモン（生理活性物質）の一種であり，分泌量が増える／減る，あるいは，体液中の濃度が高い／低いことを記述する文脈で用いられることが多い．実際，例文では “a reduction of androgens”, “higher levels of mineralocorticoids” という句を構成して出現している．こうした句の構造を，周辺単語を bag-of-words として扱う手法では捉えることができない．しかし，係り受け木の上で並行ランダムウォークを行う手法では，図 2 に示すようなウォークを用いることで，周辺単語の構造を捉えてそれらの類似度を測ることができる．

Algorithm 1 Calculate $Sim^L(v_i, v'_{i'})$

input:
類似度を測る対象となる単語対 $(v_i, v'_{i'}) \in T \times T'$
ランダムウォークの停止確率 γ
類似度計算に用いるウォークの長さの上限 L
単語の類似度 $K_v(v_i, v'_{i'})$, $\forall (v_i, v'_{i'}) \in T \times T'$
エッジの類似度 $K_e(e_{ij}, e'_{i'j'}) = \begin{cases} 1 & \text{if label of } e_{ij} = e'_{i'j'} \\ 0 & \text{otherwise} \end{cases}$
節点 $v_i, v'_{i'}$ の出次率 $d_i, d'_{i'}$
output: 長さ L 以下のランダムウォークによる単語対の類似度 $Sim^L(v_i, v'_{i'})$

1: $Sim^0(v_i, v'_{i'}) \leftarrow K_v(v_i, v'_{i'})$
2: **for** $l \leftarrow 1 \dots L$ **do**
3: **for all** $(v_i, v'_{i'}) \in T \times T'$ **do**
4: $Sim^l(v_i, v'_{i'}) = K_v(v_i, v'_{i'}) \times \gamma^2$
$+K_v(v_i, v'_{i'}) \times \frac{1-\gamma}{d_i} \times \frac{1-\gamma}{d'_{i'}} \times \sum_{v_j, v'_{j'}} Sim^{l-1}(v_j, v'_{j'}) K_e(e_{ij}, e'_{i'j'})$
5: **end for**
6: **end for**
7: **return** $Sim^L(v_i, v'_{i'})$

Vol.2010-NL-199 No.1
2010/11/18

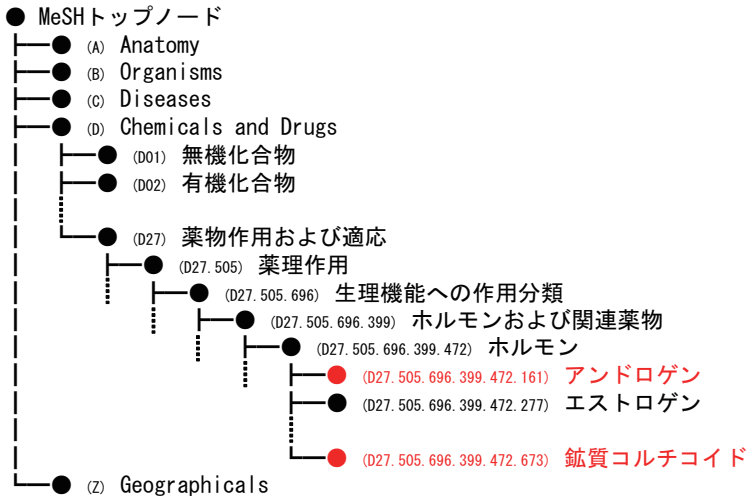

図 3 MeSH シソーラスの一部を示す．括弧内はシソーラス木のアドレスである．

3. シソーラス拡張

本論文ではライフサイエンス分野の専門用語シソーラスである MeSH ⋆1 を用い，前節で説明した単語類似度計算法を，シソーラス拡張タスクで評価する．

シソーラスは単語の上位下位関係を木構造で表した辞書である．たとえば，アンドロゲン (androgen) と鉱質コルチコイド (mineralocorticoids) はホルモンと上位下位関係にあり，互いに兄弟関係にある（図 3 参照）．

シソーラス拡張タスクは，シソーラス辞書の編集者が新しい単語を追加登録するのを支援することを目的の一つとする．すなわち，新しい単語と類似度の高い（シソーラス既登録の）単語を編集者に提示し，新しい単語をその近傍に配置することを推薦する．このとき，新しい単語とシソーラス既登録の単語の正確な類似度計算が必要になる．

⋆1 http://www.nlm.nih.gov/mesh/MBrowser.html

4. 評 価 実 験

4.1 実験データ

類似度計算の対象とする単語の集合として，GENIA treebank コーパス（1999 MEDLINE アブストラクト）⋆2 に出現し，かつ，MeSH に登録されている専門用語を用いる．ただし，語義曖昧性解消とは切り離した評価を行うために，複数の語義，すなわち，複数の MeSH アドレスを持たない専門用語に限定する．さらに，4.2 節で説明する評価方法に則り，MeSH の木において親子または兄弟の関係にある専門用語が少なくとも一つ GENIA コーパスに出現する専門用語に限定する．

また，使用する専門用語の頻度を統一した．なぜなら，専門用語により GENIA コーパスに出現する頻度が異なるが，出現頻度の違いにより，計算結果の精度に影響がでるのを避けるためである．今回の実験では，GENIA コーパスに 5 回以上出現する専門用語に限定する．さらに，各々の専門用語に対してランダムに選んだ 5 回の出現のみを実験に使用する．以上の条件を考慮にいれ，合計 234 専門用語（$234 \times 5 = 1170$ 出現）を，今回の実験で使用する専門用語集合とする．

なお，並行ランダムウォークにより単語類似度を測る方法は，各々の専門用語の出現に対して，それを含む文の係り受け木を必要とする．今回の実験では，GENIA treebank コーパスの句構造木を，係り受け木に変換するツール⋆3を用いて係り受け木を作成した．

最後に，前節で説明した並行ランダムウォークによる単語類似度計算アルゴリズムにおける，単語対の類似度 K_v の与え方について述べる．係り受け木のグラフの節点となりうるすべての単語タイプの対に対して K_v を計算するが，まず，GENIA コーパスに出現する各々の単語タイプについて特徴ベクトルを作成する．具体的には，単語タイプの出現（トークン）の各々に対して，GENIA コーパスが与える品詞，および，一定のウィンドウ範囲⋆4に現れる周辺単語を抽出し，それぞれの頻度を合計しておく．次に，それら頻度の合計を tfidf により重み付けし，各々の単語タイプの特徴ベクトルの要素とする．そして，単語対の類似度 K_v を，それら特徴ベクトルをもとに計算するコサイン類似度として与える．

⋆2 version 1.0: http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=GENIA+Treebank
⋆3 The LTH Constituent-to-Dependency Conversion Tool for Penn-style Treebanks:
http://nlp.cs.lth.se/software/treebank_converter
⋆4 今回の実験で設定したウィンドウは前後 5 単語である．

Vol.2010-NL-199 No.1
2010/11/18

4.2 評価方法

上述の専門用語集合から 1 つの専門用語をクエリとして選択し，残り（233 専門用語）をランキングすることを順次（234 回）行う．ランキングは，シソーラス木においてクエリの近傍に位置する専門用語が上位にランクされていることが望まれる．今回の実験では，クエリの最近傍の専門用語，つまり，クエリと親子または兄弟の関係にある専門用語のランクに注目する．そして，それらのうちの最上位ランクを評価に用いる．というのは，シソーラス辞書の編集者は，ランキング結果の上位の単語から順番にチェックしていくことが想定され，最近傍の専門用語のうちいずれかが上位に入っていれば，シソーラス辞書の編集作業を支援できると考えられるからである．

4.3 結　　果

ランダムウォークの長さ L の上限を 1〜10，停止確率 γ を 0.001〜0.5 の範囲で設定して実験を行い，クエリと親子または兄弟の関係にある最上位専門用語のランクの平均値を表 1 に示す．$L = 1$ は，係り受け木の構造情報を用いないコサイン類似度による手法に相当する．並行ランダムウォークにより単語類似度を測る手法は $L \leq 8$，$\gamma = 0.001$ のときにベストの平均ランク 37.21 位を得て，コサイン類似度による 42.48 位を約 5 ポイント上回る結果となった．

表 1 シソーラス拡張タスクによる評価実験の結果．
クエリと親子または兄弟の関係にある最上位専門用語のランクの平均値を示す．L はランダムウォークの長さ，γ はランダムウォークの停止確率である．$L = 1$ は，係り受け木の構造情報を用いないコサイン類似度による手法の結果に相当する．

γ	0.001	0.005	0.01	0.05	0.1	0.2	0.3	0.4	0.5
L＝1	42.48	42.48	42.48	42.48	42.48	42.48	42.48	42.48	42.48
L≦2	40.43	40.43	40.43	40.31	39.70	38.99	39.54	40.68	41.29
L≦3	38.49	38.5	38.50	38.21	38.85	40.72	41.36	41.82	42.06
L≦4	39.17	39.12	38.88	38.34	39.92	41.17	41.68	41.91	42.07
L≦5	38.08	37.79	37.81	39.95	40.92	41.47	41.75	41.91	42.06
L≦6	37.94	37.56	38.18	40.40	41.14	41.51	41.76	41.91	42.06
L≦7	37.67	38.3	39.31	40.92	41.24	41.53	41.76	41.91	42.06
L≦8	37.21	38.73	39.94	41.03	41.26	41.53	41.76	41.91	42.06
L≦9	38.26	39.7	40.17	41.16	41.30	41.53	41.76	41.91	42.06
L≦10	38.9	40.1	40.45	41.19	41.30	41.53	41.76	41.91	42.06

5. 関連研究

係り受け関係を手がかりに単語類似度を計算する関連研究には，萩原らの研究がある[3),9)]．彼らは類義語獲得を行うために，周辺単語および係り受け枝を辿って得られる単語を用いて単語の特徴ベクトルを作り，コサイン類似度により単語類似度を計算した．類義語獲得の実験で，係り受けの枝を辿る回数は 2 回以下が最適で，3 回の効果はほとんどない，と報告している[9)]．本研究のシソーラス拡張タスクの実験では，係り受けの枝を辿る回数は 8 回以下のときに最も良い結果を得たが，この顕著な相違はタスクによって最適な回数が異なる，あるいは，単語類似度の計算方法の違いに因る，と考えられる．また，文の類似度を測るために，文の係り受け構造を利用する研究もある[5)]．類似度を測りたい 2 つの文の係り受け木において，共通する経路の数をカウントし，それを 2 つの文の類似度としている．

6. まとめと今後の課題

本研究では，Desrosiers らによるグラフ上の節点間の類似度をランダムウォークにより獲得する方法[2)] を，自然言語処理において，コーパスから単語類似度を測る方法に適用し，その有効性を確かめた．周辺単語を bag-of-words として抽出して作成する特徴ベクトルによるコサイン類似度と異なり，並行ランダムウォークによる手法は文の係り受け木の上でランダムウォークを行い，周辺単語の構造を捉えてそれらの類似度を測ることができる．MeSH シソーラスと GENIA コーパスを用いたシソーラス拡張タスクの実験で，並行ランダムウォークを用いて単語類似度を測る手法はコサイン類似度による手法を上回る結果を得ることに成功した．

今後の課題は以下の通りである．まず，forward と backward のみとしている枝ラベルの情報量を増やすことによる効果を調べる．たとえば，句構造規則を枝のラベルとして用いることが考えられる．逆に，枝ラベルの種類を減らし，forward と backward を区別しない，あるいは forward ラベルの枝のみを用いることもできる．さらに，ウォークの類似度計算を，ウォークのスキップを許して行うことが考えられる．これは，and などの機能語をスキップすることに対応する．また，係り受け木以外の構造，たとえば日本語 HPSG パーザ Enju [⋆1] が出力する述語項構造活用することも考える．加えて，シソーラス拡張以外のタスクに，並行ランダムウォークによる単語類似度計算を適用することも念頭にいれる．

⋆1 http://www-tsujii.is.s.u-tokyo.ac.jp/enju/index.ja.html

Vol.2010-NL-199 No.1
2010/11/18

謝辞 本研究は，文部科学省統合データベースプロジェクト「ライフサイエンス分野の統合データベース整備事業」の支援を得て行われた．

参 考 文 献

1) Dagan, I., Lee, L. and Pereira, F.: Similarity-based methods for word sense disambiguation, *ACL-35: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, Morristown, NJ, USA, Association for Computational Linguistics, pp.56–63 (1997).
2) Desrosiers, C. and Karypis, G.: Within-Network Classification Using Local Structure Similarity, *ECML/PKDD*, pp.260–275 (2009).
3) Hagiwara, M., Ogawa, Y. and Toyama, K.: Selection of effective contextual information for automatic synonym acquisition, *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Morristown, NJ, USA, Association for Computational Linguistics, pp.353–360 (2006).
4) Kashima, H., Tsuda, K. and Inokuchi, A.: Marginalized Kernels Between Labeled Graphs, *ICML*, pp.321–328 (2003).
5) Kate, R.J.: A dependency-based word subsequence kernel, *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Morristown, NJ, USA, Association for Computational Linguistics, pp.400–409 (2008).
6) Kawahara, D. and Kurohashi, S.: A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis, *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, Morristown, NJ, USA, Association for Computational Linguistics, pp.176–183 (2006).
7) Sasano, R., Kawahara, D. and Kurohashi, S.: A Fully-Lexicalized Probabilistic Model for Japanese Zero Anaphora Resolution, *COLING*, pp.769–776 (2008).
8) Snow, R., Jurafsky, D. and Ng, A.Y.: Semantic Taxonomy Induction from Heterogenous Evidence, *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, Association for Computational Linguistics, pp.801–808 (2006).
9) 萩原正人，小川泰弘，外山勝彦：類義語自動獲得における間接依存関係の有効性，言語処理学会第 13 回年次大会ワークショップ「言語的オントロジーの構築・連携・利用」論文集，pp.43–46 (2007).