2J3-NFC2-1

言語横断共訓練によるWikipediaからの上位下位関係の獲得

Bilingual Co-Training for Hyponymy Relation Acquisition from Wikipedia

呉　鍾勲　　山田　一郎　　内元　清貴　　鳥澤　健太郎　　橋本　力
Jong-Hoon Oh　Ichiro Yamada　Kiyotaka Uchimoto　Kentaro Torisawa　Chikara Hashimoto

情報通信研究機構
National Institute of Information and Communications Technology

本論文では大規模かつ高精度な知識獲得のため，言語横断共訓練（bilingual co-training）という新たな枠組みを提案する．言語横断共訓練においては，二つの言語のための二つの知識獲得プロセスが，対訳辞書などの対訳資源によって繋がれ，各プロセスが協調して処理を行い，両言語の知識獲得の性能を向上させる．実験では，知識獲得のタスクの一つである Wikipedia からの上位下位関係獲得を日本語と英語に対して行い，言語横断共訓練を適用することにより，F 値が約 3.6〜10.3%改善できることを示した．さらに，二言語の初期学習データと同量の学習データを用いた単言語の上位下位関係獲得処理との比較実験を行い，二言語を用いる言語横断共訓練が効果的であることを確認した．

1. はじめに

機械翻訳や質問応答などの高度な自然言語処理応用技術を開発する上で，シソーラスのような意味知識を獲得し蓄積することは極めて重要な課題である．しかし，大量かつ高品質な意味知識を獲得するのは難しく，その獲得技術はまだ発展途上と言える．本論文では単言語の意味知識，特に上位下位関係[*1]といった単語間の意味的関係を精度良く獲得するための新しい枠組みを提案する．以降では，この枠組みを**言語横断共訓練**（bilingual co-training）と呼ぶ．

単語間の意味的関係の獲得は，任意の語のペアに対し，ある特定の意味的関係があるか否かを二値分類するタスクとして扱われることが多い [Girju 07]．この二値分類のタスクには教師有り学習の方法が多用され，効果を上げている．しかし，教師有り学習では，一般に高い性能を得るために大量の学習データが必要であり，学習データの準備に高いコストがかかるという問題がある．

そこで，言語横断共訓練の枠組みは次のような考え方に基づいて，従来からの二値分類における問題に対処する．

- ある言語の学習データを別の言語に翻訳し，翻訳された言語での同一のタスクの学習データに加えることができれば，あまりコストをかけずに対象とする翻訳先の言語の学習データを拡張できる．
- ある言語の自動分類結果のうち信頼度の高いものをさらに別の言語に翻訳し，翻訳先の言語の学習データに加えることで学習データをさらに拡張できる．
- 追加される学習データは後に説明する理由により最終的な分類性能を向上させる上で効果的である．

一般に，言語が異なれば，同一，もしくは同じタイプのタスクであっても，素性集合，素性値，コーパスなど学習時の入力が異なる．そのため，ある言語では自動分類された事例の分類結果が高い信頼度を持つ場合でも，別の言語では，対応する事例の分類結果の信頼度が低いことがある．このような場合，ある言語における信頼度の高い自動分類結果を別の言語でも同様に信頼できるとみなして，学習データに加えることで，全体的な精度を改善することができる．また，このプロセスは図 1 のように言語を入れ替えつつ，学習データの追加，再学習，学習結果をうけての再分類，その再分類をうけての学習データのさらなる追加と，何度でも繰り返すことが出来る．これは，いわゆる共訓練（co-training）[Blum 98] の言語横断版と言える．

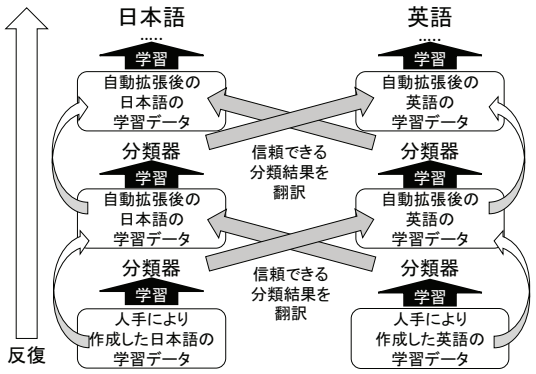

図 1: 言語横断共訓練の概念

本論文では，隅田ら [隅田 09] によって提案された Wikipedia からの上位下位関係獲得をタスクとして取り上げる．隅田らの手法は教師有り学習のアプローチに準拠しており，日本語を対象にした実験ではその性能が F 値で約 80%であることが示されている．このタスクを英語と日本語の Wikipedia から上位下位関係を同時に獲得するタスクへと拡張し言語横断共訓練を適用することにより，言語横断共訓練の有効性を定量的に示す．隅田らの手法では，与えられた単語対が上位下位関係にあるかを判定する処理において，上位語と下位語に共通する文字列が大きな手がかりとして用いられている．例えば，「**酵素**」とその下位語「加水分解**酵素**」の単語対は，一方の単語「酸素」が他方の単語の最終形態素として使われているため，上位下位関係の判定は比較的容易であると考えられるが，それぞれの英訳である *enzyme* と *hydrolase* は，共有する文字列がないため，上

連絡先: 呉鍾勲，情報通信研究機構，〒 619-0289 京都府相楽郡精華町光台 3-5，rovellia@nict.go.jp

*1 本論文では，上位下位関係を，「A は B の一種です」，もしくは「A は B の一例です」のいずれかを満たす A と B の関係と定義する．前者の条件は，A と B がともに概念である場合で，例えば「犬」と「哺乳類」がこの関係に該当する．後者は A がインスタンスで B が概念である場合で，例えば「清水寺」と「お寺」がこの関係に該当する．

位下位関係の判定はより難しいと考えられる．つまり，日本語の分類器が高い信頼度で上位下位関係にあると推定したペアの英訳を，英語の分類器はそれほど高い信頼度で上位下位関係とは推定できないと推測される．この場合，日本語において高い信頼度で上位下位関係にあると推測された単語対を英語に翻訳し，対応する英語の単語対を学習データに加えることにより，英語の分類器の弱い部分を補うことができる．さらに，英語と日本語を入れ替えて同様のプロセスを繰り返すことにより，各々の分類器の弱い部分が，より改善される可能性がある．また，このようにして自動生成された信頼度の高い事例数はプロセスを繰り返すことにより増加し，最終的には膨大な量の学習データを構築できる．このデータは，学習器の弱い部分を補強するために選ばれたものであり，同様の効果は，人手コストをかけて学習データを無作為に増やしただけでは得られないと考えられる．実験では，言語横断共訓練に基づく提案手法は，日英各言語の上位下位関係獲得を独立に行った場合と比べて，F 値で約 3.6〜10.3%改善できることを示した．さらに，単言語の学習データを二言語の初期学習データの総量に合わせて，単言語における上位下位関係獲得実験を行い，二言語を用いる言語横断共訓練が効果的であることを確認した．

2. 言語横断共訓練

二つの異なる言語をそれぞれ S，T とし，学習／分類の結果であるクラスラベルの集合を CL とする．クラスラベルは「yes」または「no」とする（$CL = \{yes, no\}$）．言語 S，T で獲得される知識の候補となるインスタンス（上位下位関係であれば，単語対）の集合を X_S，X_T とし，$X = X_S \cup X_T$ とする．分類器 c はインスタンス x に対してクラスラベル cl と信頼度 r を与える．つまり，$c(x) = (x, cl, r)$ とする．ただし，$x \in X$，$cl \in CL$，$r \in R^+$ である．実験では，分類器 c として SVM を採用し，サンプルと超平面との距離の絶対値を信頼度 r として用いる．学習データは $L \subset X \times CL$，学習プロセスは関数 $LEARN$ と表わし，学習データ L により分類器 c を学習する場合，$c = LEARN(L)$ と表現する．学習データのうち，特に，人手により作成した S と T の学習データをそれぞれ，L_S，L_T と表わす．

対訳インスタンス辞書 D_{BI} は言語 S と T のインスタンスのうち対訳関係があるものと定義する．すなわち，$D_{BI} = \{(x_s, x_t)\} \subset X_S \times X_T$ とする．ただし，この定義は簡略化したものであり，実際に上位下位関係獲得で用いる場合，x_s と x_t はそれぞれ上位下位関係候補の単語対となり，$D_{BI} = \{(x_s, x_t)\} \subset {X_S}^2 \times {X_T}^2$ と表現される．例えば，上位下位関係獲得において，対訳インスタンスペア (x_s, x_t) を (x_s=(*enzyme*, *hydrolase*), x_t=(酵素, 加水分解酵素)) のように表現でき，*enzyme* と酵素，*hydrolase* と加水分解酵素が対訳辞書によって翻訳できると考える．

対訳辞書は Wikipedia の言語横断リンク（cross-language link）に基づいて作成する．Wikipedia では，言語横断リンクにより，複数の言語間の記事が繋がれている．このリンクを持つ記事間には翻訳関係がある場合が多い．そこで，言語横断リンクで繋がれた記事のタイトルペアの集合を対訳辞書として使用できる [Erdmann 08]．本論文では，Wikipedia の英日言語横断リンクと日英言語横断リンクを利用し約 20 万の記事タイトル間の対訳ペアを取り出した（以後，*Wikipedia* **対訳辞書**と呼ぶ．）．

言語横断共訓練のアルゴリズムの疑似コードを図 2 に示す．まず，i 回目用の学習データである L_S^i および L_T^i を用いて，各言語の学習器 c_S^i および c_T^i を学習する．ただし，0 回目の

1: $i = 0$
2: $L_S^0 = L_S$; $L_T^0 = L_T$
3: **repeat**
4: $c_S^i := LEARN(L_S^i)$
5: $c_T^i := LEARN(L_T^i)$
6: $CR_S^i := \{c_S^i(x_S) | x_S \in X_S, \forall cl\ (x_S, cl) \notin L_S^i, \exists x_T\ (x_S, x_T) \in D_{BI}\}$
7: $CR_T^i := \{c_T^i(x_T) | x_T \in X_T, \forall cl\ (x_T, cl) \notin L_T^i, \exists x_S\ (x_S, x_T) \in D_{BI}\}$
8: $L_S^{(i+1)} := L_S^i$
9: $L_T^{(i+1)} := L_T^i$
10: **for** each $(x_S, cl_S, r_S) \in TopN(CR_S^i)$ **do**
11: **for** each x_T such that $(x_S, x_T) \in D_{BI}$ and $(x_T, cl_T, r_T) \in CR_T^i$ **do**
12: **if** ($r_S > \theta$ and $r_T < \theta$) or ($r_S > \theta$ and $cl_S = cl_T$) **then**
13: $L_T^{(i+1)} := L_T^{(i+1)} \cup \{(x_T, cl_S)\}$
14: **end if**
15: **end for**
16: **end for**
17: **for** each $(x_T, cl_T, r_T) \in TopN(CR_T^i)$ **do**
18: **for** each x_S such that $(x_S, x_T) \in D_{BI}$ and $(x_S, cl_S, r_S) \in CR_S^i$ **do**
19: **if** ($r_T > \theta$ and $r_S < \theta$) or ($r_T > \theta$ and $cl_S = cl_T$) **then**
20: $L_S^{(i+1)} := L_S^{(i+1)} \cup \{(x_S, cl_T)\}$
21: **end if**
22: **end for**
23: **end for**
24: $i = i + 1$
25: **until** a fixed number of iterations is reached

図 2: 言語横断共訓練の疑似コード

繰り返しでは学習データは人手で用意したもの，すなわち L_S，L_T を用いて学習する（2〜5 行目）．次に，学習した分類器 c_S^i および c_T^i によって X_S および X_T に含まれるインスタンスを分類する（6〜7 行目）．10〜16 行目は分類されたインスタンスの集合 CR_S^i のうち，高い信頼度を持つ分類結果から，言語 T の新たな学習データに加えるインスタンスを選択する方法を示している．$TopN(CR_S^i)$ は CR_S^i のうち信頼度 r_S が上位 N 位となるような分類結果 $c_S^i(x)$ の集合である．実験では $N = 900$ とした（4. 章を参照）．この選択の過程で，c_S^i は教師，c_T^i は生徒のように振る舞う．教師は分類結果 cl_S の信頼度がある一定レベル以上の場合（$r_S > \theta$），かつ，$r_T < \theta$ あるいは $cl_S = cl_T$ を満たす場合に限り，生徒に対し x_T のクラスラベルが cl_S であると教示する．17〜23 行目では，教師と生徒の役割は逆転し，c_T^i が教師，c_S^i が生徒となる．最終的に，一定回数の繰り返しが終了した段階で処理を終了する．

3. 上位下位関係の獲得

本節では，言語横断共訓練の枠組みを適用する隅田ら [隅田 09] の上位下位関係の獲得手法について述べる．隅田らは，Wikipedia 記事の定義文（記事の第 1 文），カテゴリ，階層構造から上位下位関係を獲得する手法を提案している．本論文では隅田らの階層構造を利用した手法に従い，単言語の上位下位関係候補の抽出を行う．Wikipedia 記事の階層構造を用い

る手法は Wikipedia 記事の定義文やカテゴリを用いる手法と比べて下記の利点がある．

- Wikipedia 記事の階層構造からは定義文やカテゴリに比べてより多くの上位下位関係が獲得できる．隅田ら [隅田 09] は Wikipedia 記事の階層構造から 150 万の日本語上位下位関係を獲得できると報告している．
- Wikipedia 記事の階層構造は言語依存性が低いため，英語と日本語の両言語に対してある程度言語独立な上位下位関係候補の抽出手法が適用できる．

この手法は「上位下位関係候補の抽出」,「上位下位関係候補の分類」の二つの処理で構成される．上位下位関係候補の抽出処理では，Wikipedia 記事から記事のタイトル，節のタイトル，リスト項目などのノードを取り出し，親子関係*2となっている二つのノードを，その親ノードを上位語，子ノードを下位語とした上位下位関係候補として扱う．例えば，節タイトル「Tiger」とその下位にあるリスト項目「Siberian Tiger」から，上位下位関係候補（Tiger, Siberian tiger）が抽出できる．この手法により，英語 Wikipedia 記事の階層構造から 3,900 万の英語上位下位関係候補を，日本語 Wikipedia 記事の構造からは 1,000 万の日本語上位下位関係候補を抽出することができた．以後，上位語候補を **hyper**，下位語候補を **hypo** と呼ぶ．

次に，抽出した **hyper** と **hypo** が，上位下位関係を持つか否かを分類する処理を行う．この処理では，SVM を分類器として用いる．隅田らの手法では，**hyper** や **hypo** に含まれる形態素や形態素の品詞などの「語彙素性」と，**hyper** と **hypo** を取り出した階層構造項目の種類（記事タイトル，節タイトルなど）や階層構造上の距離などの「構造素性」を用いて，分類器の学習処理を行っている．

提案手法では隅田らの素性に加え，以下の構造素性と「Infobox 素性」を利用する．

- 木構造ノードの種類（例：ルートノード，リーフノード）
- **hypo** の親ノードの語彙素性
- **hyper** の子ノードの語彙素性
- Wikipedia の Infobox とその属性の情報（Infobox 素性）

最後の項目の Infobox 素性は，Wikipedia の Infobox から取り出した Infobox 名，属性名，そして，属性値に関する特徴を表す．Wikipedia の Infobox には Wikipedia 記事タイトルに対して，その Infobox 名と属性名，属性値が記述されている．そこで，この（Infobox 名，属性名）を属性値の素性として考える [Auer 07]．例えば，Wikipedia 記事「クリスティアーノ・ロナウド」の Infobox には，Infobox 名として「サッカー選手」，属性名と属性値として「所属チーム名=レアル・マドリード」が存在する*3．この情報から，(サッカー選手，所属チーム) を，「レアル・マドリード」の Infobox 素性とすることができる．なお，Infobox の素性は，上位下位関係候補の **hyper**，もしくは **hypo** がいずれかの記事の Infobox から獲得した（Infobox 名，属性名，属性値）の属性値に該当する場合のみに付与されることに注意されたい．

*2 直接の親子関係だけでなく,「祖父－孫」や「曾祖父－ひ孫」など間接的な親子関係も含む．

*3 英語 Wikipedia から約 440 万の属性値に対する 590 万の（記事タイトルの上位語，属性名，属性値）の三つ組を，日本語 Wikipedia からは約 120 万の属性値に対する 160 万の三つ組を取り出し，Infobox 素性を生成した．

4. 実験

本節では，Wikipedia からの上位下位関係獲得タスクに，言語横断共訓練を適用した実験について報告する．実験では，2008 年 5 月版の英語 Wikipedia と 2008 年 6 月版の日本語 Wikipedia を対象とした．両言語の上位下位関係候補からランダムに各 24,000 を選んで実験データとした．

この実験データに対して，人手により上位下位関係の有無のタグ付けを行い*4，ここからランダムに 20,000 を学習データとして選び，残り 4,000 を等分して開発用データとテストデータとした．学習データは言語横断共訓練の初期分類器の学習で用い，開発用データによって言語横断共訓練のパラメータを最適化し，テストデータは提案手法の性能を評価するために使う．

分類器としては TinySVM*5の二次多項式カーネルを使う．言語横断共訓練の最大繰り返し回数は 100 に設定し，開発用データに対して最適な性能を示した $\theta = 1$ と $TopN$=900 を言語横断共訓練のパラメータとして実験を行った．

実験結果の評価では，上位下位関係と判定されたペアが正しい割合を示す適合率（P），テストデータ中の全上位下位関係に対して正しく抽出できた割合を示す再現率（R），そして，適合率と再現率の調和平均である F 値（F）を利用する．

4.1 言語横断共訓練の効果

言語横断共訓練の効果を評価するために，下記の四つの手法に対して実験を行った．

- SYT：隅田ら [隅田 09] の手法．
- INIT：言語横断共訓練の初期分類器（c^0）のみを使用する手法．
- TRAN：言語横断共訓練の初期分類器（c^0）のみを使用する手法．ただし，INIT と異なり，対訳インスタンス辞書で初期学習データを英語から日本語，あるいは日本語から英語に翻訳し，その翻訳結果を対象言語の初期学習データに追加して新たな学習データを生成する．この処理における，英語の新たに追加された学習データ数は 729，日本語は 486 であった．
- BICO：言語横断共訓練に基づく手法（提案手法）．

各手法の評価結果を表 1 に示す．表 1 において SYT は隅田ら [隅田 09] に報告された性能より低い結果となった．これは，学習データとテストデータの差，特にその量の差が原因と考えられる．本実験では 20,000 の学習データと 2,000 のテストデータを使用しているが，隅田ら [隅田 09] では 29,900 の学習データと 1,000 のテストデータを使用した．

INIT と SYT の比較は本論文で新たに提案した素性の影響を示している．両システムの F_1 値の差は 0.5〜1.8%で小さいが，INIT は一貫して SYT より高い F_1 値を示しており，これらの素性が有効であると考えられる．

BICO は，言語を問わず全ての比較手法に対して大幅な性能向上を示している（F_1 値で約 3.6〜10.3%）．TRAN と BICO の比較からは，単なる既存の学習データの翻訳では得られない性能向上が言語横断共訓練で得られていることが分かる．

言語横断共訓練によって得られた分類器を全ての英語と日本語の上位下位関係候補に適用することによって，約 540 万の英語の上位下位関係と約 241 万の日本語の上位下位関係を獲得することができた．

*4 人手での判定は一つの言語について約 2-3ヶ月を要した．各言語の実験データのうち，約 8,000 が上位下位関係として判定され，正例と負例の比率はおよそ 1:2（8,000:16,000）であった．

*5 http://chasen.org/~taku/software/TinySVM

表 1: 各手法の評価結果（%）

	英語			日本語		
	P	R	F_1	P	R	F_1
SYT	78.5	63.8	70.4	75.0	77.4	76.1
INIT	77.9	67.4	72.2	74.5	78.5	76.6
TRAN	76.8	70.3	73.4	76.7	79.3	78.0
BICO	78.0	83.7	80.7	78.3	85.2	81.6

表 2: 初期学習データの量と F_1 値（%）の関係 : 言語横断共訓練を適用した場合と適用しなかった場合

	単言語で $2n$		二つの言語で各 n	
n	INIT-E	INIT-J	BICO-E	BICO-J
2500	67.3	72.3	70.5	73.0
5000	69.2	74.3	74.6	76.9
10000	72.2	76.6	76.9	78.6

4.2 初期学習データを同数とした実験

前節の実験では，提案手法（BICO）の初期学習データ量は，二言語分を合わせた時に，他の手法と比較して多い．そこで，同じ初期学習データ量に対する BICO の効果を明確にするため，「単言語に対する $2n$ の学習データと，二言語に対する各 n（合計 $2n$）の学習データでは，どちらが効果的か？」という質問に対する回答を得る実験を行った．表 2 に，その結果を示す．

INIT-E と INIT-J は $2n$ の単言語の学習データで学習した英語と日本語の分類器の性能を示す．BICO-E と BICO-J は各言語で n の初期学習データ（二つ言語で合計 $2n$ の学習データ）で各言語の初期分類器を学習し，言語横断共訓練を適用した結果を示す．BICO の場合，INIT に使われた初期学習データの半分を単言語の初期学習データとして使用しているが，最終的に言語に限らず INIT より高い性能を示している．言語横断共訓練は，同じ量の初期学習データで学習した一つの単言語分類器より高性能であり，さらに同じコストで二言語の分類器を生成できる．つまり，最初の質問に対する回答は，「二言語に対する各 n（合計 $2n$）の学習データのほうが効果的」となる．

5. 関連研究

Li ら [Li 02] は訳語曖昧性解消のため二言語ブートストラップ（bilingual bootstrapping）という手法を提案した．この手法では言語横断共訓練と同様に二言語に対する各分類器が対訳資源を介して協調をする．しかし，二言語ブートストラップでは，対訳資源によって一方の言語の単語から他方の言語の単語へ対応付けが出来ない場合は，その単語を処理対象とすることができない．二言語ブートストラップの処理は対訳資源に依存すると考えられる．一方，言語横断共訓練では，学習データを獲得する処理のみで対訳資源を利用し，各言語における分類処理では対訳資源を必要としないため，対訳資源に含まれない単語に対しても分類処理を行える．

また，二言語の言語資源を利用した手法が，動詞の分類や名詞句の意味解析などのために提案されている [Merlo 02]．しかし，この手法は提案手法と違い，単言語の教師あり学習における二言語に関わる素性の生成のために二言語資源を使っている．

近年，文書からの意味的関係獲得の研究が注目されている．SemEval-07 [Girju 07] では，名詞句間の意味的関係分類のための様々な手法が提案された．それらの手法と本提案手法はとの違いは，本提案手法が単言語の意味的関係獲得（上位下位関係獲得）に複数言語の情報を利用している点にある．

6. まとめ

本論文では言語横断共訓練を提案し，Wikipedia から上位下位関係を獲得するタスクに適用した．実験で，言語横断共訓練を用いた手法は，日英各言語の上位下位関係獲得を独立に行った場合と比べて F 値で約 3.6〜10.3%改善できることを示した．また，単言語の初期学習データのみで学習した単言語の分類器より，それと同じ量の二言語の学習データを用いた言語横断共訓練の方が高性能であることを示した．今後は，日本語と英語以外の言語に対して，さらには，上位下位関係獲得以外のタスクに対して言語横断共訓練を適用し，その有効性を検証する予定である．

参考文献

[Auer 07] Auer, S. and Lehmann, J.: What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content, in *Proc. of the 4th European Semantic Web Conference (ESWC 2007)*, pp. 503–517, Springer (2007)

[Blum 98] Blum, A. and Mitchell, T.: Combining labeled and unlabeled data with co-training, in *COLT' 98: Proceedings of the eleventh annual conference on Computational learning theory*, pp. 92–100 (1998)

[Erdmann 08] Erdmann, M., Nakayama, K., Hara, T., and Nishio, S.: A Bilingual Dictionary Extracted from the Wikipedia Link Structure., in *Proc. of DASFAA*, pp. 686–689 (2008)

[Girju 07] Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., and Yuret, D.: SemEval-2007 Task 04: Classification of Semantic Relations between Nominals, in *Proc. of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 13–18 (2007)

[Li 02] Li, C. and Li, H.: Word Translation Disambiguation Using bilingual bootstrapping, in *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 343–351 (2002)

[Merlo 02] Merlo, P., Stevenson, S., Tsang, V., and Allaria, G.: A multilingual paradigm for automatic verb classification, in *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 207–214 (2002)

[隅田 09] 隅田 飛鳥, 吉永 直樹, 鳥澤 健太郎：Wikipedia の記事構造からの上位下位関係抽出, 自然言語処理, Vol. 16, No. 3, pp. 3–24 (2009)