情報処理学会論文誌　Vol. 52　No. 12　3558–3573 (Dec. 2011)

言語横断共訓練による単語間の上位下位関係の獲得

呉　鍾勲[†1]　山田一郎[†2,†1]　内元清貴[†1]
鳥澤健太郎[†1]　橋本力[†1]

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ける各言語の知識獲得プロセスは，対訳辞書などの対訳資源によってつながれ，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 1 つである Wikipedia からの上位下位関係獲得を日本語と英語を対象として行い，言語横断共訓練を適用することにより，F_1 値が約 3.6〜10.3%改善できることを示した．さらに，2 言語で用意した学習データは，その総量が同じ単言語における学習データと比較して，上位下位関係獲得処理においてより効果的であることを示した．

A Bilingual Co-training Algorithm for Hyponymy Relation Acquisition

Jong-Hoon Oh,[†1] Ichiro Yamada,[†2,†1]
Kiyotaka Uchimoto,[†1] Kentaro Torisawa[†1]
and Chikara Hashimoto[†1]

This paper proposes a novel framework called *bilingual co-training* for a large-scale, accurate acquisition method for *monolingual* semantic knowledge. In this framework, we combine the independent processes of monolingual semantic-knowledge acquisition for two languages using bilingual resources to boost performance. We apply this framework to large-scale hyponymy-relation acquisition from Wikipedia. Experimental results show that our approach improved the F-measure by 3.6–10.3%. We also show that bilingual co-training enables us to build classifiers for two languages in tandem with the same combined amount of data as required for training a single classifier in isolation while achieving superior performance.

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うえで，シソーラスのような意味知識を獲得し蓄積することはきわ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いえる．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[10]．また，二値分類のタスクには教師あり学習の方法が多用され，効果をあげている．しかし，教師あり学習では，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．これは英語だけでなく，日本語や他の言語における意味的関係獲得においても直面する問題である．

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いる．

- 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，そ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．
- 学習データの拡張された部分は最終的な性能を向上させるうえで効果的である．

一般に，言語が異なれば，素性集合，素性値，学習データ，コーパスなど学習時の設定が異なるため，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

†1 情報通信研究機構
National Institute of Information and Communications Technology
†2 NHK 放送技術研究所
Science & Technical Research Laboratories, Japan Broadcasting Corporation

⋆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，たとえば「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，たとえば「清水寺」と「お寺」がこの関係に該当する．

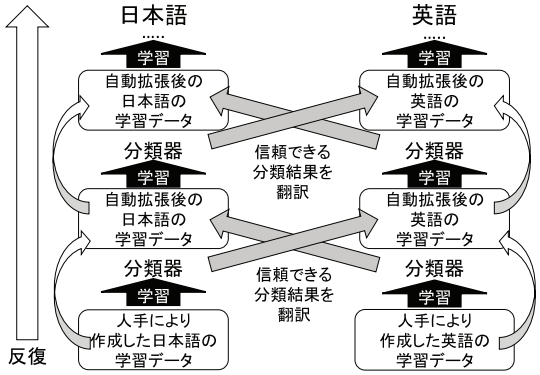


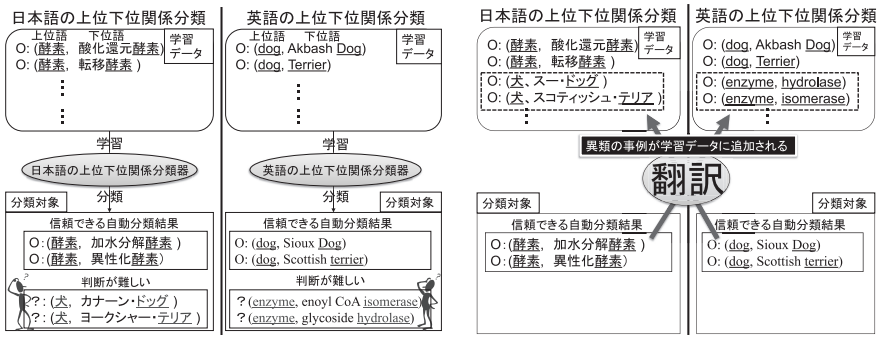
図 1　言語横断共訓練の概念
Fig. 1　Concept of bilingual co-training.

言語における信頼度の高い自動分類結果は別の言語でも同様に信頼できると見なすことで，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できる．これは，いわゆる共訓練 (co-training)[4] の言語横断版といえる．

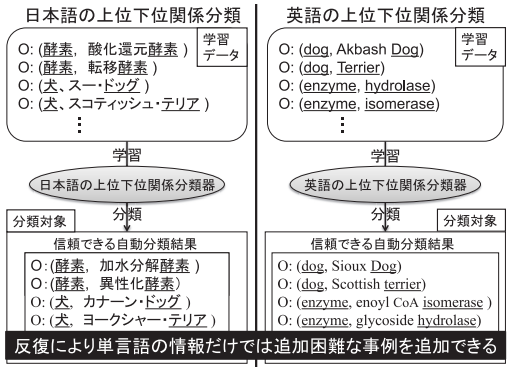
図 2 に言語横断共訓練による上位下位関係獲得の例とその期待される効果を示す．各言語の上位下位関係の分類器は異なる分類結果を生成し，信頼度の高い分類結果と信頼度の低い分類結果が異なる（図 2 (a)）．そこで，一方の信頼できる分類結果を翻訳して他方の学習データに追加し（図 2 (b)），新たな学習データで分類器を再学習する．この結果，新たな分類器は判定が難しかったインスタンスに対しても信頼できる分類結果を生成する（図 2 (c)）．

本論文では，Sumida ら[22] 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，我々の提案する言語横断共訓練の枠組みの有効性を検証する．Sumida らの手法は教師あ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_1 値で約 80%であることが示された．以下では，このタスクを英語，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，上述の言語横断共訓練を適用することを考え，そのメリットについて具体的な例をあげて説明する．

Sumida らの手法では，与えられた 2 語が上位下位関係にあるかの判定において，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．たとえば，「**酵素**」と



(a) 初期学習データで学習した分類器による分類結果　(b) 翻訳による新たな学習データの追加



(c) 拡張した学習データで学習した分類器による分類結果

図 2　言語横断共訓練の例：上位下位関係の獲得
Fig. 2　Examples of bilingual co-training: hyponymy relation acquisition.

「加水分解**酵素**」の 2 語は，末尾に「**酵素**」という文字列を共有しているため，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にあるとは推定しないと推測される．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 2 語を英語に翻訳し，対応する英語の 2 語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，分類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通常行われるようにランダムサンプリングに基づいて単純に増やしただけでは得られないと考えられる．

本論文では，英語，日本語の Wikipedia からの上位下位関係の獲得において言語横断共訓練が実際に有効であることを定量的に示す．実験では，言語横断共訓練に基づく提案手法により，言語横断共訓練を用いない場合，つまり，日英各言語の上位下位関係獲得を独立に行った場合に比べて，性能を F_1 値で約 3.6〜10.3%改善できることが分かった．さらに，2 言語から得た学習データは，同じ量の単言語のみから作成した学習データと比較して，上位下位関係獲得処理において効果的であることを確認した．さらに，言語横断共訓練の枠組みは，対象とする 2 つの言語の橋渡しをする対訳辞書，訳語選択処理の質に依存しないことも示す．

本論文の構成は次のとおりである．まず，2 章で言語横断共訓練の枠組みについて述べ，3 章では，この枠組みに基づく上位下位関係獲得システムについて詳細に述べる．そして，5 章で実験結果をあげ，その結果について考察する．7 章では，関連研究について述べ，最後に，8 章で結論と今後の課題について述べる

2. 言語横断共訓練

2 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 のインスタンス，つまり最終的にクラスラベルを振りたい言語表現の集合，それぞれを X_S，X_T とし，$X = X_S \cup X_T$ とする．なお，上位下位関係獲得タスクでは上位下位関係を持つ可能性のある単語ペアをインスタンスとする．分類器 c はインスタンス x に対してクラスラベル cl（$\in CL$）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，$D_{BI} = \{(x_s, x_t)\} \subset {W_S}^2 \times {W_T}^2$ と表現される．ここで，W_S と W_T は言語 S の単語と言語 T の単語を表す．たとえば，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．なお，後述するように本論文では Wikipedia から自動的に構築した対訳辞書で対訳インスタンス辞書 D_{BI} を作成する．

言語横断共訓練のアルゴリズムの疑似コードを**図 3** に示す．まず，i 回目用の学習データ

```
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 ∈ X_S, ∀cl (x_S, cl) ∉ L_S^i, ∃x_T (x_S, x_T) ∈ D_BI}
7:    CR_T^i := {c_T^i(x_T)|x_T ∈ X_T, ∀cl (x_T, cl) ∉ L_T^i, ∃x_S (x_S, x_T) ∈ 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) ∈ TopN(CR_S^i) do
11:      for each x_T such that (x_S, x_T) ∈ D_BI and (x_T, cl_T, r_T) ∈ CR_T^i  do
12:         if (r_S > θ and r_T < θ) or (r_S > θ and cl_S = cl_T) then
13:            L_T^(i+1) := L_T^(i+1) ∪ {(x_T, cl_S)}
14:         end if
15:      end for
16:   end for
17:   for each (x_T, cl_T, r_T) ∈ TopN(CR_T^i) do
18:      for each x_S such that (x_S, x_T) ∈ D_BI and (x_S, cl_S, r_S) ∈ CR_S^i do
19:         if (r_T > θ and r_S < θ) or (r_T > θ and cl_S = cl_T)  then
20:            L_S^(i+1) := L_S^(i+1) ∪ {(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 3　言語横断共訓練の疑似コード
Fig. 3　Pseudo-code of *bilingual co-training*.

である L_S^i および L_T^i を用いて，各言語の分類器 c_S^i および c_T^i を学習する．ただし，0 回目の繰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）．インスタンスの集合 X_S のうち，分類器の学習に使われた学習データ L_S^i に含まれず，かつ，対訳辞書 D_{BI} に含まれるようなインスタンスに対し，分類器 c_S^i を適用して分類した結果の集合を CR_S^i とする．これは翻訳されて言語 T の学習データに加えられるインスタンスの候補となる．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（5 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，つまり，$r_S > \theta$ であり，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．ここで，x_T は実際には対訳辞書 D_{BI} による x_S の翻訳である．$cl_S = cl_T$ という条件は，付与するべきクラスラベルについて，生徒も自らの意見に対してある一定レベルの信頼度を持っており，かつ，教師の意見と異なる場合，つまり，$r_T > \theta$ かつ $cl_S \neq cl_T$ となる場合に，クラスラベルを教師に合わせて換えてしまうのを回避するためのものである．この場合，教師は何もせず，そのインスタンスを無視する．一方，$r_T < \theta$ の場合は，生徒の信頼度が低いと考えられるため，生徒と教師との間で意見の相違があったとしても，教師の意見を優先する．すべての条件が満たされる場合には，(x_T, cl_S) を言語 T の学習データ $L_T^{(i+1)}$ に追加する．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．

ここで，オリジナルの共訓練（co-training）[4] との関係についてまとめる．まず，共訓練では，2 つの分類器は**同じタスク**に携わる．つまり，両分類器の分類対象と分類対象に与える分類クラスの集合が同じである．しかし，言語横断共訓練では，同じタスクというよりも**同じタイプのタスク**に 2 つの分類器が携わる．つまり，両分類器が分類対象に与える分類クラスのタイプは同じだが，分類対象は言語によって異なる．別のいい方をすれば，共訓練と言語横断共訓練の大きな違いはインスタンスの集合である．共訓練ではインスタンスは同じで着目する素性が異なるが，言語横断共訓練ではインスタンスそのものが言語により異なるものの，インスタンスのいくつかは対訳辞書でつながっているため，あたかも同じ集合のインスタンスを対象にしているかのように扱われる．結果として，言語横断共訓練では，同じタイプの素性が 2 つの言語の分類器で使われることはあるが，その具体的な出現，あるいは分類結果に対する貢献は大きく異なることとなる．この素性の働き方の差によって，分類器の得意とするインスタンス，不得意とするインスタンスが言語ごとに変わり，結果として，異なる素性セットを使う 2 つの分類器によって効果をあげる共訓練と同様の効果を期待できる．

3. 上位下位関係の獲得

本章では，言語横断共訓練の枠組みを適用する Sumida ら[21),22)] の上位下位関係の獲得手法について述べる．この手法は Wikipedia 文書からの上位下位関係獲得手法であり，「上位下位関係候補の抽出」，「上位下位関係候補の分類」の 2 つの処理で構成される．以下に，各処理について説明する．

3.1 上位下位関係候補の抽出

Sumida らの手法では，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．本論文では Sumida 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より多くの上位下位関係が獲得できる．Wikipedia 記事の定義文とカテゴリから上位下位関係を獲得する手法では，獲得できる上位下位関係の下位語は記事のタイトルに限定されるため，その数は少ない．Sumida ら[22)] は Wikipedia 記事の定義文からは 17 万，Wikipedia 記事のカテゴリからは 42 万，Wikipedia 記事の階層構造からは 150 万の日本語上位下位関係を獲得できたと報告している．
- 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．

上位下位関係候補の抽出処理では，まず，**図 4** のような Wikipedia 記事の階層構造を**図 5** のように木構造へ変換する．記事のタイトル，節のタイトル，リスト項目などが木構造のノードになる．この木構造のあるノードと，そのノードのすべての下位ノードとの単語ペアを上位下位関係候補として扱う．たとえば，図 5 のノード「Tiger」とその下位ノード「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した．抽出した候補は 3.2 節に説明する分類器によって上位下位関係か否かを判定する．

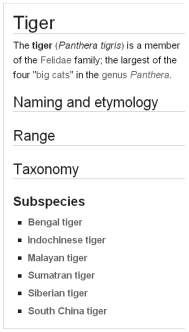


図 4　Wikipedia 記事，TIGER の階層構造
Fig. 4　A layout structure of Wikipedia article TIGER.

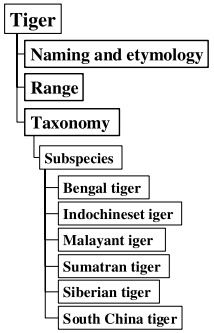


図 5　図 4 の階層構造を木構造に変換した結果
Fig. 5　A tree structure of the layout structure in Fig.4.

3.2　上位下位関係候補の分類

抽出した候補が上位下位関係を持つか否かを分類するために，SVM [14] を分類器として用いる．分類器の学習では，語彙素性（LF_1〜LF_5），構造素性（SF_1〜SF_5），Infobox 素性（IF）を使う．それぞれの説明を**表 1** に示す．表 1 では，**hyper** を上位語候補，**hypo** を **hyper** の下位語候補，(**hyper,hypo**) を上位下位関係候補と表記しており，以降，この表記を使用する．語彙素性（LF_1〜LF_5）と構造素性の SF_1，SF_2 は Sumida ら[22] によって提案されたものである．一方，構造素性の SF_3〜SF_5 と Infobox 素性は本論文で新たに提案する素性である．

以下に，使用した素性を簡単に説明する．語彙素性（LF_1〜LF_5）は **hyper** および **hypo** の形態素，単語，品詞などの語彙的な情報として上位下位関係の判定に使われる．LF_1 と LF_2 では，**hyper** と **hypo** の「主要部名詞」（主要部名詞は「名詞句内の主要部名詞」である．日本語では，末尾の形態素が主要部名詞となる場合が多い）と「主要部名詞以外」を区別し，その形態素の表記と品詞情報のそれぞれを素性として登録した．また，**hyper** と **hypo** の主要部名詞の一致の有無も素性として登録した．たとえば，**hyper**=酵素 と **hypo**=加水分解 酵素 の「主要部名詞」は「酵素」であり，**hypo** の「加水」と「分解」は「主要部名詞以外」の形態素であり，それぞれが独立に素性として登録される．そして，上位下位関係候補の **hyper** と **hypo** がその末尾の主要部名詞を共有するため，上位下位関係となる可能性が高い．LF_1〜LF_3 はそのような特徴を表すために使われる[⋆1]．LF_4 は Wikipedia 記事のタイトルと節のタイトルによく現れる特徴的な語彙統語パターンを表す[⋆2]．たとえば，Wikipedia 記事タイトルや節のタイトルには「X の一覧」や「有名な X」のような表現がよく現れる．しかし，「X の一覧」や「有名な X」などの表現は上位語としてふさわしくないため，「の一覧」や「有名な」などの表現を削除する必要がある．さらに，「X の一覧」や「有名な X」で表現される記事のタイトルや節タイトルは，その階層構造における下位には X のインスタンスとなる単語がよく現れ，X とそのインスタンスのペアは上位下位関係になる場合が多い．語彙統語パターンは上記の 2 つの目的で使われ，「上位下位関係候補が持つ **hyper** あるいは **hypo** が語彙統語パターンに合致するか否か」と「その一致した語彙統合パターンの表記」のそれぞれを LF_4 の素性として登録する．なお，語彙統合パターンに合致する **hyper** と **hypo** は語彙統合パターンによって「の一覧」や「有名な」などの表現が削除される．たとえば，（サッカー選手の一覧，クリスティアーノ・ロナウド）という上

表 1　使用した素性の種類
Table 1　Feature type.

素性の種類	説明
LF_1	**hyper**，**hypo** の主要部名詞の形態素の表記と， **hyper**，**hypo** の主要部名詞以外の形態素の表記． そして，**hyper**，**hypo** の主要部名詞の一致の有無
LF_2	**hyper**，**hypo** の主要部名詞の形態素の品詞と， **hyper**，**hypo** の主要部名詞以外の形態素の品詞
LF_3	**hyper**，**hypo** の表記
LF_4	節タイトルなどに頻出する語彙統語パターンとの一致の有無，そして，その一致した語彙統合パターンの表記（例：「主な X」，「X のリスト」など）
LF_5	節タイトルなどに頻出する単語との一致の有無，そして，その一致した単語の表記（例：「参照」，「分類」など）
SF_1	**hyper** と **hypo** の Wikipedia 階層構造上の距離
SF_2	階層構造項目の種類（例：節タイトル，リスト項目）
SF_3	木構造ノードの種類（例：ルートノート，リーフノード）
SF_4	**hypo** の親ノードの LF_1 と LF_3 素性
SF_5	**hyper** の子ノードの LF_1 と LF_3 素性
IF	Wikipedia infobox における属性情報

⋆1 語彙素性は形態素解析の結果を基に生成し，英語の形態素解析器 TagChunk [6] と日本語の形態素解析器 MeCab (http://mecab.sourceforge.net/) を使用した．
⋆2 日本語の語彙統語パターンは Sumida ら[22] と同じものを使い，英語の語彙統語パターンは日本語の語彙統語パターンを人手で翻訳して作成した．

位下位関係候補に語彙統語パターン「X の一覧」が適用され「サッカー選手の一覧」の「の一覧」が削除される．この結果，候補が（サッカー選手，クリスティアーノ・ロナウド）になる．また，「歴史」，「分類」のように Wikipedia 記事の節タイトルでよく使われる単語が上位語候補になる場合，その上位語候補を持つ上位下位関係候補は上位下位関係になる可能性が低い．LF_5 はこのような特徴を表す．

構造素性は，上位下位関係候補を抽出した記事の階層構造や木構造に関する素性である．SF_1 は階層構造における **hyper** と **hypo** の距離を表す．SF_2 は **hyper** と **hypo** が現れた階層構造項目の種類（記事タイトル，節タイトル，リスト項目など）を表す[22]．

SF_3〜SF_5 は本論文で新たに提案した構造素性である．SF_3 は **hyper** と **hypo** が現れたノードの種類を表す．上位下位関係候補 (**hyper,hypo**) の **hyper** と **hypo** が木構造のルートノード（および Wikipedia 記事のタイトル）とルートノードの子ノード（節のタイトル）から抽出される場合，その上位下位関係候補は上位下位関係になる可能性が低い（たとえば，図 4 の記事タイトルと節タイトルから抽出できる上位下位関係の候補（Tiger，Range）と（Tiger，Taxonomy）は上位下位関係ではない）．SF_3 はそのような特徴を表現するために使われる．SF_4 と SF_5 はそれぞれ，**hyper** の子ノードの語彙素性と **hypo** の親ノードの語彙素性を示す．親ノードや子ノードを共有する単語は，類似していると考えられるため，このような類似性を表す特徴として SF_4 と SF_5 を利用している．

Infobox 素性 IF 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，その上位語[⋆1]と属性名，属性値が記述されている．そこで，この（記事タイトルの上位語，属性名）を属性値の素性として考える[3),25)]．たとえば，Wikipedia 記事「クリスティアーノ・ロナウド」の Infobox「サッカー選手」には「所属チーム名=レアル・マドリード」という属性名と属性値のペアが存在する[⋆2]．この情報から，(サッカー選手，所属チーム) を，**hyper** または **hypo** として出現している「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 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．

⋆1 Infobox 名は記事タイトルの上位語と考えられる．たとえば，Wikipedia 記事「Microsoft」の Infobox 名は「会社」であり，これは「Microsoft」の上位語と考えられる．

⋆2 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 3 つ組を，日本語 Wikipedia からは約 120 万の属性値に対する 160 万の 3 つ組を取り出し，Infobox 素性を生成した．

4. Wikipedia からの対訳辞書の獲得

言語横断共訓練に使われる対訳インスタンス辞書 D_{BI} を構築するため，Wikipedia から対訳辞書を獲得する．対訳辞書の獲得は Wikipedia の言語横断リンク（cross-language link）に基づく．Wikipedia では，言語横断リンクにより，複数の言語間の記事がつながれている．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，言語横断リンクでつな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[7]．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）．

日本語の上位下位関係候補 $(hyper_J, hypo_J)$ と英語の上位下位関係候補 $(hyper_E, hypo_E)$ に対して，下記の条件を満たす「英日上位下位関係候補ペアの集合」を対訳インスタンス辞書 D_{BI} として使う．

- $(hyper_J, hyper_E)$ と $(hypo_J, hypo_E)$ の両方が対訳辞書に存在する．

約 20 万の対訳ペアを持つ Wikipedia 対訳辞書を用いて上記の条件を満たす約 18 万ペアの対訳インスタンスを取り出し，D_{BI} とした．

5. 実　　験

本章では，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[⋆3]，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 を開発用データとテストデータ

⋆3 人手での判定は 3 人で行い，1 つの言語について約 2–3 カ月を要した．3 人の被験者の一致率を示す Kappa 値は英語で 0.752，日本語で 0.798 と，十分な一致率であると考えられる．また，2 人以上が一致した判定結果に基づいて「上位下位関係である」,「上位下位関係ではない」と判定した．各言語の実験データのうち，約 8,000 が上位下位関係であると判定され，正例と負例の比率は 1 : 2（8,000 : 16,000）であった．「頻繁に現れる節タイトル」や「Wikipedia 記事のタイトル」は階層構造の上位にあるため，上位下位関係候補の上位語候補が「頻繁に現れる節タイトル」や「Wikipedia 記事のタイトル」で構成される場合がより多い．しかし，このような上位語候補を持つ上位下位関係の候補は上位下位関係ではない場合が多いため，評価データには負例が正例より多かった．

表 2　各言語における実験データ量
Table 2　Statistics on test set.

種類	日本語	英語
学習データ	20,000	20,000
開発用データ	2,000	2,000
テストデータ	2,000	2,000

とした（**表 2**）．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[⋆1] の 2 次多項式カーネルを使う．言語横断共訓練の最大繰返し回数は 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（$TopN$ と θ 値に関わる実験は 5.1 節を参照）．

本論文では，言語横断共訓練の効果，初期学習データ量の影響，対訳辞書の影響，対訳辞書サイズの影響を評価する 4 つの実験を行う．対訳辞書の影響を評価する実験では，Wikipedia 対訳辞書を含め 4 つの対訳辞書を使って実験を行い，残り 3 つの実験では Wikipedia 対訳辞書のみを使った．

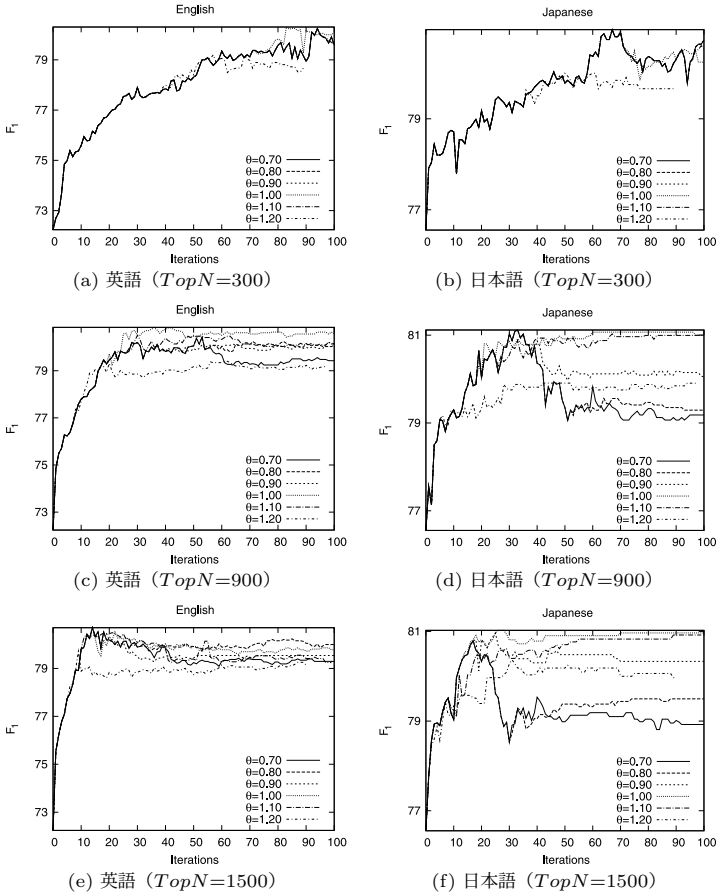
性能は式 (1) で定義された適合率（P），再現率（R），F_1 値（F_1）で測定する．式 (1) で Rel はテストデータのうち人手で上位下位関係と判定された候補の数，$HRbyS$ はテストデータのうち各手法で上位下位関係と判定された候補の数を表す．

$$P = |Rel \cap HRbyS|/|HRbyS| \quad (1)$$
$$R = |Rel \cap HRbyS|/|Rel|$$
$$F_1 = 2 \times (P \times R)/(P + R)$$

5.1 予 備 実 験

言語横断共訓練におけるパラメータ，θ と $TopN$ の値を決定するため，そして，θ と $TopN$ に対して提案手法がどの程度の頑健性を有するかを考察するため，予備実験を行った．予備実験では異なる 6 つの θ（$\theta = \{0.7, 0, 8, 0, 9, 1, 1.1, 1.2\}$）と 3 つの $TopN$（$TopN = \{300, 900, 1500\}$）の組合せを開発用データに適用した．

図 6 に示す実験結果から，θ が 1 以上の場合は繰返しが進むにつれ F_1 値が改善されていることが分かる．一方，θ が 1 未満の場合は，ある繰返し段階から F_1 値が下がっている．言語横断共訓練において θ は分類結果を「信頼できる分類結果」と「信頼できない分類結果」に分割するしきい値であり，この θ の値が高いほど，「信頼できる分類結果」における誤りが少なくなる．θ の値が低くなると「信頼できる分類結果」の数は多くなるが，その分，



(a) 英語（$TopN$=300）　(b) 日本語（$TopN$=300）
(c) 英語（$TopN$=900）　(d) 日本語（$TopN$=900）
(e) 英語（$TopN$=1500）　(f) 日本語（$TopN$=1500）

図 6　$TopN$，θ と F_1 値（%）の関係
Fig. 6　F_1 curves by different θ and $TopN$ values.

⋆1 http://chasen.org/˜taku/software/TinySVM

言語横断共訓練のノイズとなる分類結果の誤りを含む可能性も高くなり，このようなノイズが言語横断共訓練において悪影響を与えてしまう．初期の繰返しの段階では，$TopN$ のすべての分類結果は高い信頼度を持つため，θ による影響は少なく F_1 の差が小さい．しかし，繰返しが進むにつれ，低い θ の値（たとえば 0.7）により判断された「信頼できる分類結果」に言語横断共訓練のノイズとなる分類結果の誤りが出現するようになり F_1 の値に差が見られるようになるものと思われる．

$TopN$ は言語横断共訓練の 1 回の繰返しによって追加できる新たな学習データの数の最大値を示し，分類器における精度改善に必要な繰返し回数と最終的な精度に影響する．$TopN$=900 と比べ $TopN$=300 は各繰返し処理における F_1 値の改善が少なく，最終的に得られる F1 値も低い．また，$TopN$=900 と $TopN$=1500 は繰返し処理における F_1 値の改善の割合は同程度だが，最終的に得られる F_1 値は $TopN$=900 がより高い．

予備実験の結果から，θ と $TopN$ の値によらず提案手法による F_1 値の向上が見られ，$\theta = 1$ と $TopN$=900 の組合せにより最も高い F_1 値が得られることが分かった．この結果を受けて本実験では θ=1，$TopN$=900 という設定を用いた．

5.2　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 4 つの手法に対して実験を行った．

- SYT：Sumida ら[22] 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 3** に示す．表 3 において SYT は Sumida ら[22] 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，Sumida ら[22] 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 SF_3〜SF_5 と IF 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．

表 3　各手法の評価結果 (%)
Table 3　Performance of different systems (%).

	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

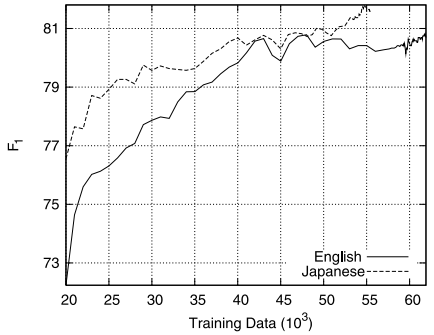


図 7　言語横断共訓練によって追加された学習データの量と F_1 値（%）の関係
Fig. 7　F_1 curves by the training data size.

BICO は，言語を問わずすべ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．

図 7 は言語横断共訓練によって追加された学習データによる F_1 値の変化を示している．言語横断共訓練の初期分類器の学習に使われた初期学習データは両言語でともに 20,000 のため，図 7 のグラフは学習データの量が 20,000 から始まる．英語の場合，言語横断共訓練により増えた学習データ量が約 42,000 のため，学習終了時の学習データの総計は約 62,000 である．日本語の場合，言語横断共訓練により増えた学習データ量が約 35,000 のため，学習終了時の学習データの総計は約 56,000 である．図 7 では学習データ量が増加することにより F_1 値も上がっているため，言語横断共訓練が両言語の分類器の性能向上に有効であると解釈できる．

表 3 の BICO の結果では，日本語，英語ともに再現率の向上が顕著に見られる．BICO

表 4　INIT と BICO の学習データにおける正例と負例の統計値
Table 4　Statistics for positive and negative samples in training data for INIT and BICO.

	英語		日本語	
	INIT	BICO	INIT	BICO
正例の数	6,655	19,231	6,805	17,456
正例の異なる上位語	3,650	4,949	3,238	3,954
負例の数	13,345	42,681	13,195	38,078
負例の異なる上位語候補	9,781	20,408	9,944	20,706

においては，**表 4** に示すように INIT に比べて，正例，負例，ともに学習データが 3 倍程度増大している．再現率の向上はこうした学習データの増大によるものと考えられる．また，学習データ中における上位語の分布の違いも再現率の向上の要因の 1 つと考えられる．表 4 によれば，INIT の学習データは，6,655 個の英語の正例を，6,805 個の日本語の正例を持つ．それぞれの正例が持つ上位語の異なり数は英語が 3,650，日本語が 3,238 であるため，1 つの上位語は平均 2.0 個の正例に含まれる．BICO の学習データは，19,231 個の英語の正例を，17,456 個の日本語の正例を持ち，正例が持つ上位語の異なり数は英語が 4,949，日本語が 3,954 であったため，1 つの上位語のは平均 4.1 個の正例に含まれる．つまり，言語横断共訓練によって同じ上位語を持つ正例が増えていることが分かる．一方，INIT と BICO が持っている負例は正例より多様な上位語候補を持っている．同じ上位語を持つ複数の正例で学習した分類器は，その上位語を持つ上位下位関係候補を「上位下位関係」と分類する傾向がある（特に学習データにこのような上位語を持つ負例が含まれていない場合，あるいはこのような負例が少ない場合）．その結果，このような上位語を持つ「上位下位関係」に対して正しく「上位下位関係」と判定する場合が多くなり，分類結果の再現率が向上すると考えられる．

なお，言語横断共訓練によって得られた分類器をすべ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 210 万の下位語に対する約 540 万の英語の上位下位関係と約 70 万の下位語に対する約 241 万の日本語の上位下位関係を獲得した．獲得した上位下位関係は，約 20 万単語が登録された WordNet [5),8)] と比べて，英語では約 10 倍以上の単語カバレージを，日本語では 3 倍以上の単語カバレージを持っている．そして，獲得した上位下位関係には，WordNet [5),8)] などの既存のシソーラスに含まれていない固有名に関わる上位下位関係が多く含まれているため，既存のシソーラスの拡張や固有名の知識源として有効であると考えられる．

表 5　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
Table 5　F_1 based on training data size: with/without bilingual co-training.

n	単言語で $2n$		2 つの言語で各 n	
	INIT-E	INIT-J	BICO-E	BICO-J
2,500	67.3	72.3	70.5	73.0
5,000	69.2	74.3	74.6	76.9
10,000	72.2	76.6	76.9	78.6

5.3　初期学習データ量の影響

言語横断共訓練に使われる初期学習データ量の影響を評価するため，2 つの実験を行った．1 つ目は，「学習データの作成コストが 2 つの言語で同じ場合，単言語に対する $2n$ の学習データと，2 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である．**表 5** はその結果を示す．

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（2 つの言語で合計 $2n$ 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．つまり，言語横断共訓練によって同じ量の初期学習データで学習した 1 つの単言語分類器より高性能の分類器を構築でき，さらに同じコストで 2 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，「2 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．

2 つ目は，「言語横断共訓練に携わる 2 つの分類器のうちの一方の性能が低い場合にも両言語で言語横断共訓練によってその性能を向上させることができるか？」という質問に対する実験を行った．この質問の答えが “yes” であれば，言語横断共訓練により新たな言語（たとえば，フランス語，中国語など）の上位下位関係獲得に必要な学習データの作成コストを減らすことができる．たとえば，英語の分類器あるいは日本語の分類器を高性能な分類器と想定し，学習データが少ない新たな言語の分類器を性能の低いものと想定する．上記の質問の答えが “yes” なら，その新たな言語の分類器に対しては少量の学習データのみで高い性能を得ることができる．以後，高性能な分類器を「強い分類器」，性能の低い分類器を「弱い分類器」と呼ぶ．強い分類器の学習には初期学習データのすべてを，弱い分類器の学習には初期学習データの一部（データ数 1,000，5,000，10,000，15,000 のいずれか）を用いる．初期学習データの一部のみを用いて弱い分類器を学習するのは，学習データ数が少な

表 6　初期学習データの量と F_1 値（%）の関係：英語の分類器が強い場合.
Table 6　F_1 for each training data size: when the English classifier is strong one.

英語の初期学習データの量	INIT-E	BICO-E	日本語の初期学習データの量	INIT-J	BICO-J
20,000	72.2	79.6	1,000	64.0	72.7
20,000	72.2	79.6	5,000	73.1	75.3
20,000	72.2	79.8	10,000	74.3	79.0
20,000	72.2	80.4	15,000	77.0	80.1

表 7　初期学習データの量と F_1 値（%）の関係：日本語の分類器が強い場合.
Table 7　F_1 for each training data size: when the Japanese classifier is strong one.

英語の初期学習データの量	INIT-E	BICO-E	日本語の初期学習データの量	INIT-J	BICO-J
1,000	60.3	69.7	20,000	76.6	79.3
5,000	67.3	74.6	20,000	76.6	79.6
10,000	69.2	77.7	20,000	76.6	80.1
15,000	71.0	79.3	20,000	76.6	80.6

いと分類器の性能が低くなる傾向にあることが容易に予想されるからである.

表 6 と**表 7** は 2 つ目の実験の結果を示す．INIT は各言語で初期学習データを使って学習した分類器の結果を，BICO は言語横断共訓練を適用した結果を表す．すべての設定で BICO が INIT より高性能を示すことが分かった．つまり，弱い分類器側の言語や初期学習データ数に関係なく，強い分類器は常に弱い分類器の性能向上に寄与し，弱い分類器も強い分類器の性能向上に寄与する．この実験の結果から，2 つ目の質問の答えは “yes” であり，一方の分類器の性能が低い場合でも言語横断共訓練は効果的に動作することが分かる.

5.4　対訳辞書の影響

対訳辞書の影響を評価するため 5 種類の対訳辞書を言語横断共訓練へ適用した．対訳辞書として，EDR と EDICT,「科学技術振興機構辞書」（専門用語辞書），Wikipedia 対訳辞書を用いた．以後それぞれを D1，D2，D3，D4 と呼ぶ．そして，D1〜D4 をマージして得られた一対訳辞書を D5 と呼ぶ．これらの辞書は一般用語と専門用語の割合や固有名詞の数などにおいて大きく異なる．いい換えれば，長所短所が互いに大きく異なる．**表 8** に，異なる対訳辞書の選択が性能（F_1）にもたらす影響をまとめた．表 8 において，Entry は対訳辞書の対訳ペアの数を，E2J と J2E は対訳辞書の 1 つの英語の見出し語（E2J の場合），もしくは 1 つの日本語の見出し語（J2E の場合）に対する対訳ペア数の平均値を，D_{BI} は各対訳辞書に基づいて得られた対訳インスタンス数を表す．たとえば，1 つの英語の見出し

表 8　言語横断共訓練への対訳辞書の影響
Table 8　Effect of different bilingual dictionaries.

対訳辞書		F_1（%）		対訳辞書の情報			D_{BI}
		英語	日本語	Entry	E2J	J2E	
D1	α=5	74.7	77.7	135K	2.01	2.23	27K
	All	73.8	76.6	425K	8.29	4.99	113K
D2	α=5	76.5	78.4	588K	1.80	1.77	83K
	All	75.0	77.2	990K	7.17	2.52	132K
D3	α=5	76.9	78.5	667K	1.89	1.55	65K
	All	77.0	77.9	750K	3.05	1.71	79K
D4	α=5	80.7	81.6	197K	1.03	1.02	177K
	All	80.7	81.6	197K	1.03	1.02	177K
D5	α=5	77.8	80.0	1,233K	1.35	1.49	212K
	All	79.0	80.7	2,056K	4.14	2.43	458K

語が 3 つの日本語への訳語を持つ場合，その見出し語に関して 3 つの対訳ペアができる．言語横断共訓練で，この訳語の曖昧性の影響を示すため，2 つの条件（α=5 と All）で言語横断共訓練による分類器の性能を評価した．All はすべての対訳ペアを使うという条件で，α=5 は 1 つの見出し語が 5 つ以下の対訳ペアを持つ場合のみその対訳ペアを使うという条件である．$\alpha = 5$ の条件における E2J と J2E の値は，ALL より小さい．つまり，α=5 の条件では訳語の曖昧性が低い対訳ペアのみを用いて対訳インスタンス辞書（D_{BI}）を作成される．この条件によって D_{BI} の対訳のカバレージが下がるが，訳語曖昧性のよる対訳の誤りが減ると予想できる.

表 8 では，D4，つまり，Wikipedia 対訳辞書を利用した場合が最良の性能を示し，D1，D2，D3，D5 ではそれより低い性能を示すことが分かる．訳語の曖昧性に関する 2 つの条件 α=5 と All に対する評価結果では，α=5 の場合のほうが高い F_1 値を得られていることが多いが，その差は小さい．上位下位関係候補と Wikipedia 対訳辞書は Wikipedia という同じソースから取り出されたため，上位下位関係候補に対する Wikipedia 対訳辞書のカバレージが他の辞書より高いと考えられる．このため，D4 の対訳ペアの数は D1〜D3 の対訳ペアの数より少ないが，対訳辞書から構築される D_{BI} の対訳インスタンス数は D1〜D3 より D4 が多い．D4 とその他の間の性能差は対訳辞書のカバレージの違いに起因すると考えられる．D5 は D1〜D4 をマージして得られた一対訳辞書であるため，D5 による D_{BI} の対訳インスタンスの数は最も多いが，F_1 値は D4 より低い値であった．D5 は，1 語あたりの訳の数が D4 より多く，曖昧性が高い対訳辞書であることが原因と考えられる.

表 3 に示されているように言語横断共訓練を使わない INIT の F_1 値は英語で 72.2，日本語で 76.6 であるが，これらの値は表 8 の D1〜D5 の F_1 値より低い．つまり，言語横断共訓練は，使用する対訳辞書によらず両言語の分類器の性能向上に有効であると解釈できる．

5.5　対訳辞書サイズの影響

対訳辞書サイズが言語横断共訓練の性能に及ぼす影響を調べるため，対訳辞書のサイズを変化させて上位下位関係獲得実験を行った．**図 8** にその結果を示す．この実験では，Wikipedia 対訳辞書の一部を用いて異なるサイズの対訳辞書を作成し[⋆1]，各対訳辞書を利用した言語横断共訓練の効果を調べた．図 8 で 0%は言語横断共訓練を適用しなかった場合の性能（表 3 の INIT に対応）である．実験の結果，言語横断共訓練を適用しなかった場合（0%）と比べて，小さい対訳辞書（たとえば，10%と 20%の場合) でも言語横断共訓練による分類器の性能の向上が見られ，これらの結果から，辞書カバレージがそれほど大きくなくても性能が向上することが分かり，辞書のカバレージに対して言語横断共訓練がある程度の頑健性を持つと考えられる．一方，60%程度の対訳辞書サイズから F_1 値の向上が少なくなり，対訳辞書が一定量を超えると大きな性能向上にはつながらないことが分かる．

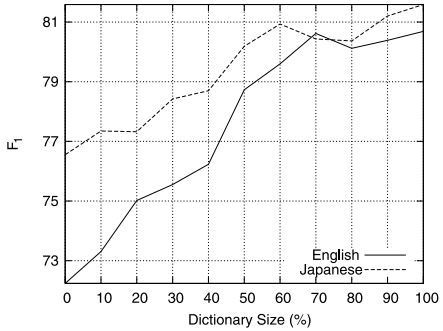


図 8　対訳辞書サイズと F_1 値（%）の関係
Fig. 8　F_1 by different size of a bilingual dictionary.

⋆1 対訳ペアの 10%，20%，30%，40%，50%，60%，70%，80%，90%をランダムに選択した 9 種類の対訳辞書を作成した．表 8 に記述したとおり Wikipedia 対訳辞書は約 20 万の対訳ペアを持っているため，Wikipedia 対訳辞書の 10%で作った対訳辞書は約 2 万の対訳ペアを持つ．

6. 考　　察

本章では，言語横断共訓練において性能改善に寄与した素性について考察する．また，2 種類の上位下位関係（「A は B の一例です」と「A は B の一種です」）に対する言語横断共訓練の効果についても検討する．

6.1　素性の種類による評価結果の解析

「言語横断共訓練によって追加された学習データ中のどのような素性が上位下位関係獲得の性能改善に寄与したか？」を明らかにするため，語彙素性（LF），構造素性（SF），Infobox 素性（IF）のさまざまな組合せで分類器を学習し，評価実験を行った．**表 9** に結果を示す．この結果では，いずれの素性のタイプ，それらの組合せにおいても，また，英語，日本語のいずれにおいても，程度の差こそあれ，言語横断共訓練によって性能が向上している．これらの結果はいずれのタイプの素性も言語横断共訓練による性能向上に貢献している可能性が高いことを示唆している．

また，これらの性能向上がどのような理由で生じたのかを知るために，言語横断共訓練を行う前（すなわち，INIT の学習データ）と後（すなわち，BICO の学習データ）での素性の異なり数の変化を素性のタイプごとに計算した．その結果を**表 10** に示す．これによれば，いずれのタイプの素性においても，言語横断共訓練を行うことで大幅に素性の異なり数

表 9　素性ごとの評価結果（%）
Table 9　Performance by different features (%).

	素性	INIT			BICO		
		P	R	F_1	P	R	F_1
英語	LF	80.6	58.9	68.1	78.8	79.7	79.2
	SF	63.7	56.7	60.0	66.9	69.2	68.1
	IF	70.0	21.3	32.7	66.9	25.5	36.9
	$LF+SF$	76.9	65.8	70.9	77.4	83.1	80.1
	$LF+IF$	79.9	65.2	71.8	79.2	80.4	79.8
	$SF+IF$	68.5	60.1	64.0	70.9	76.1	73.4
	$LF+SF+IF$	77.9	67.4	72.2	77.9	83.7	80.7
日本語	LF	73.7	75.4	74.5	77.8	80.7	79.2
	SF	61.9	70.0	65.7	68.8	75.6	72.0
	IF	70.1	27.0	38.9	74.6	29.5	42.3
	$LF+SF$	73.3	77.4	75.3	78.0	84.7	81.2
	$LF+IF$	72.7	78.5	75.5	77.9	81.2	79.5
	$SF+IF$	64.2	71.9	67.9	71.4	78.3	74.7
	$LF+SF+IF$	74.5	78.5	76.6	78.3	85.2	81.6

表 10　言語，学習データ，素性の種類による素性の分布：素性の数は素性の異なり数で計算
Table 10　Distribution of features according to feature type, language and training data: the figures are based on the number of unique features.

言語	素性の種類	INIT の学習データ	BICO の学習データ
英語	LF	93,838（29.6%）	167,174（39.0%）
	SF	216,592（68.3%）	252,744（58.9%）
	IF	6,687（2.1%）	9,140（2.1%）
	素性の総計	317,117（100%）	42,9058（100%）
英語の学習データの量		20,000	61,912
日本語	LF	73,217（46.6%）	131,533（53.6%）
	SF	80,853（51.5%）	109,734（44.7%）
	IF	2,918（1.9%）	4,142（1.7%）
	素性の総計	156,988（100%）	245,409（100%）
日本語の学習データの量		20,000	55,534

が増加していることが分かる．このことだけから，確定的な結論を導くのは難しいが，こうした素性の異なり数の増加が性能向上に貢献している可能性は高いものと思われる．

このような素性の増加がどのように性能向上に貢献しうるかを具体例に即して述べると，上位下位関係の判定においては，上位下位関係候補の形態素に関わる「上位語と下位語が共有する主要名詞部」と「上位下位関係によく現れる上位語と下位語が持つ主要名詞部の形態素対（あるいは主要名詞部の単語対）」が有力な手がかりとなる．たとえば，日本語の上位下位関係（酵素，加水分解酵素）は，上位語「酵素」と下位語「加水分解酵素」が「酵素」という主要名詞部を共有するため，上位下位関係の判定は比較的容易であり，分類器による分類結果の信頼度も高くなると考えられる．一方，（酵素，加水分解酵素）の英語の対訳である（enzyme, hydrolase）に対して，主要名詞部の共有以外に有効な手がかりとないとすると，上位下位関係の判定はより難しいといえる．ここで，言語横断共訓練により，英語の（enzyme, hydrolase）が上位下位関係であることが，日本語の（酵素，加水分解酵素）の信頼できる分類結果の翻訳により判明すると，英語の分類器は上位下位関係候補の分類に有効な新たな手がかり（「上位語の主要名詞部は enzyme であり，下位語の主要名詞部は hydrolase であれば，上位下位関係と判定できる」といった手がかり）を学習できる．実際にこうした手がかりは，新たな上位下位関係候補（enzyme, glucoside hydrolase）が適切な上位下位関係であることを判定するうえでの有力な手がかりとなる．本実験が示すところは，こうした事例が以上のような語彙素性だけでなく，他の素性，すなわち構造素性，Infobox 素性のいずれにおいても類似の現象が生じ，性能向上に貢献している可能性を示している．

表 11　「英語の初期学習データを 1,000 個，日本語の初期学習データは 20,000 個にした場合」と「英語の初期学習データを 20,000 個，日本語の初期学習データは 1,000 個にした場合」の素性ごとの評価結果（F_1 値：%）
Table 11　F_1 by different features: when the number of the initial training data for English is 1,000 and that for Japanese is 20,000 and when the number of the initial training data for English is 20,000 and that for Japanese is 1,000.

素性	$INIT^{1000}$		$BICO^{E1000}$		$BICO^{J1000}$	
	英語	日本語	英語	日本語	英語	日本語
LF	58.3	61.5	68.9	77.6	77.8	69.7
SF	10.7	61.0	67.6	73.8	70.8	68.8
IF	26.7	22.2	31.5	42.8	32.5	36.7
$LF+SF$	58.0	64.3	69.4	78.6	78.5	71.9
$LF+IF$	57.9	62.4	68.0	78.5	76.9	70.0
$SF+IF$	54.9	61.8	66.6	75.1	71.3	69.3
$LF+SF+IF$	60.3	65.0	69.7	79.3	79.6	72.7

表 12　「英語と日本語の初期学習データを 2,500 個にした場合の」素性ことの評価結果（F_1 値：%）
Table 12　F_1 by different features: when the number of the initial training data for English and Japanese is 2,500.

素性	$INIT^{2500}$		$BICO^{EJ2500}$	
	英語	日本語	英語	日本語
LF	62.4	63.0	70.6	71.1
SF	19.2	61.8	67.6	69.2
IF	28.2	25.2	32.8	35.3
$LF+SF$	62.6	66.5	72.7	74.5
$LF+IF$	63.4	63.1	69.1	72.3
$SF+IF$	56.3	62.2	66.4	70.8
$LF+SF+IF$	64.0	68.4	76.6	78.3

また，より少量の学習データを用いて言語横断共訓練を行った場合に関しても同様の評価実験を行った．より具体的には「英語と日本語の両方の初期学習データを 2,500 個にした場合」（$INIT^{2500}$ と $BICO^{EJ2500}$），「英語の初期学習データを 1,000 個，日本語の初期学習データは 20,000 個にした場合」（$INIT^{1000}$ と $BICO^{E1000}$），そして「英語の初期学習データを 20,000 個，日本語の初期学習データは 1,000 個にした場合」（$INIT^{1000}$ と $BICO^{J1000}$）という設定で評価実験を行った．**表 11** と**表 12** に結果を示す．いずれの場合においてもすべてのタイプの素性で性能向上が得られたことが分かる．つまり，少量の学習データを用いた実験においても，すべての素性に関して言語横断共訓練が有効であるとい

表 13　評価データにおける二種類の上位下位関係候補の統計値
Table 13　Statistics for two types of hyponymy relation instances in evaluation data.

	Is-A 関係候補					
	候補数	正例数	一般辞書		Wikipedia 辞書	
			カバー率	平均の訳語数	カバー率	平均の訳語数
英語	737	197	33.4%	11.20	22.1%	1.00
日本語	619	106	45.9%	5.20	24.1%	1.05
	クラス・インスタンス関係候補					
	候補数	正例数	一般辞書		Wikipedia 辞書	
			カバー率	平均の訳語数	カバー率	平均の訳語数
英語	1,263	502	17.5%	4.14	28.8%	1.00
日本語	1,381	521	25.3%	1.67	37.4%	1.03

える．

6.2　上位下位関係の種類による評価結果の解析

本論文における「上位下位関係」は「上位語と下位語がともに概念であり，前者の概念が後者の概念を包含する関係」と「下位語は上位語が表す概念のインスタンスである関係」に分けられるが，以下では前者をは「Is-A 関係」と呼び，後者を「クラス・インスタンス関係」と呼ぶ[⋆1]．以下では，このそれぞれの関係に対する上位下位関係獲得精度と，使用する対訳辞書の影響を調べるため，獲得した上位下位関係を 2 つの関係に分類したうえで分析を行った．まず，「下位語が固有名か否か」に従い，評価データに含まれた上位下位関係候補を「Is-A 関係候補」（下位語候補が固有名ではない場合）と「クラスとインスタンス関係候補」（下位語が固有名である場合）に人手で分類した．人手での判定は 3 人のアノテータで行ったが，その一致率を示す Kappa 値は英語で 0.773，日本語で 0.880 であった．そして，2 人以上が一致した判定結果に基づいて「Is-A 関係候補」と「クラスとインスタンス関係候補」に分類した．

表 13 に両言語の評価データにおける Is-A 関係候補の数とクラス・インスタンス関係候補の数，候補に含まれた正例（上位下位関係）の数，2 種類の上位下位関係候補における対訳辞書の下位語候補のカバー率，そして，対訳辞書により訳語を持つ下位語候補の平均の訳語数を示す．対訳辞書による下位語候補のカバー率は，Is-A 関係候補，あるいはクラス・インスタンス関係候補の全上位下位関係候補に対して上位下位関係候補の下位語候補が対訳辞書に登録された割合を示す．評価データの Is-A 関係候補における正例の割合は，英語が約 27%，日本語が約 17%であり，クラス・インスタンス関係候補における正例の割合は，英語が約 40%，日本語が約 38%であった．クラス・インスタンス関係候補は全体の評価データの 6 割を占める．Wikipedia 対訳辞書による下位語候補のカバー率は，Is-A 関係候補に対して英語が約 22%，日本語が約 24%，クラス・インスタンス関係候補に対して英語が約 29%，日本語が約 37%であった．まだ，Wikipedia 以外の辞書すなわち，EDR，EDICT，「科学技術振興機構辞書」をマージして得られた一般辞書による下位語候補のカバー率は，Is-A 関係候補に対して英語が約 33%，日本語が約 46%，クラス・インスタンス関係候補に対して英語が約 18%，日本語が約 25%であった．クラス・インスタンス関係候補の下位語候補における対訳辞書のカバー率は，Wikipedia 対訳辞書が一般辞書より 10%以上高い．一方，Is-A 関係候補の下位語候補における対訳辞書のカバー率は一般辞書と比べて 10%以上高い．

表 14　上位下位関係候補の種類と F_1 値の関係（%）
Table 14　F_1 for each type of hyponymy relation instances.

	英語			日本語		
	INIT	BICO^G	BICO^W	INIT	BICO^G	BICO^W
Is-A 関係候補	59.0	71.2	73.2	58.8	62.4	63.1
クラス・インスタンス関係候補	75.4	79.3	83.6	79.8	82.4	85.3

表 14 は 2 種類の上位下位関係候補においての INIT，BICO^G（一般辞書を言語横断共訓練へ適用した場合の結果），BICO^W（Wikipedia 対訳辞書を言語横断共訓練へ適用した場合の結果）を示している．実験の結果から，対訳辞書の下位語候補のカバー率によらず，Is-A 関係候補，クラス・インスタンス関係候補のいずれにおいても，BICO^W，すなわち Wikipedia 対訳辞書が最良の精度を示すことが分かった．クラス・インスタンス関係候補に対して Wikipedia 対訳辞書がより高い精度を示すことはその下位語候補のカバー率の差から容易に予想されるが，Is-A 関係候補に関してこのような結果が得られるというのは予想外であった．この理由を厳密に特定するのは非常に難しいが，「一般辞書における非固有名詞，すなわち，一般概念の訳語の多義性」が原因の 1 つと考えられる．つまり，一般辞書における Is-A 関係候補の下位語の訳語は，Wikipedia に比べてより多様なものがあり[⋆2]，これが言語横断共訓練に悪影響を与えたと考えられる．より詳細な分析は今後の課題としたい．

⋆1 上位下位関係といった場合には Is-A 関係のみを指す場合もあるが，固有名詞に関するクラス・インスタンス関係の有用性に鑑み，本論文ではこのような立場をとった．

⋆2 Is-A 関係候補の下位語候補に対して一般辞書による平均の訳語数は，日本語で 5.20，英語で 11.20 であり，Wikipedia 対訳辞書による平均の訳語数の約 5〜11 倍であった．

7. 関連研究

Li ら[16] は訳語曖昧性解消のため 2 言語ブートストラップ（bilingual bootstrapping）という手法を提案した．2 言語ブートストラップ手法では言語横断共訓練と同様に 2 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，2 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できない場合は，その単語を処理対象とすることができない．つまり，2 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，言語横断共訓練でも上述の条件が必要条件となるが，各言語のすべての分類対象がその条件に満たす必要はない．つまり，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，2 言語の言語資源を利用した手法が，動詞の分類，評価文書の自動分類，名詞句の意味解析などのために提案されている[9),17),23),24)]．しかし，この手法は提案手法と違い，教師あり学習における 2 言語に関わる素性の生成のために 2 言語資源を使っている．特に Wan[23] と Wei ら[24] の手法では，英語の学習データを中国語に翻訳し，あるいは評価文書の自動分類に主要な手がかりとして使われる英語と中国語の単語間の関係を学習し，学習データが存在しない中国語の評価文書の自動分類を行った．このような手法は言語横断共訓練において補完的に使うことができる（たとえば，言語横断共訓練におけるある言語の初期学習データが存在しない場合）．

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．意味的関係獲得の手法として，人手で作ったルールに基づく手法や機械学習を Wikipedia に適用した手法などが提案されてきた[12),15),18)–20),22)]．また，SemEval-07[10] では，名詞句間の意味的関係分類のためのさまざま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．

Adar ら[1] は多言語の Wikipedia の Infobox 情報を統合することにより，単言語の Infobox 情報を補完，生成する手法を提案した．このような Infobox 情報は提案手法で用いる Infobox 素性に対応するため，Adar らの手法を扱うことによって提案手法の Infobox 素性がより豊かになると期待できる．

語彙統合パータンを用いてテキストから上位下位関係を獲得する手法[2),11),13)] が提案されている．したし，これらの手法は単言語の上位下位関係獲得に複数言語の情報を利用している提案手法と違い，単言語のテキストからの上位下位関係のみを獲得対象にする．このようなテキストからの上位下位関係獲得に言語横断共訓練を適用することにより性能向上が期待されると思われる．

8. 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．

実験では言語横断共訓練が 2 言語の上位下位関係獲得の性能向上に効果的であることを示した．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 2 言語の初期学習データを用いて言語横断共訓練に基づいて学習した単言語の分類器の方が高性能であることを示した．

さらに，言語横断共訓練は一方が強い分類器で，他方が弱い分類器である場合にも両分類器の性能向上に役に立つことを示した．この結果は，学習データ作成のコストを抑えつつ，新たな言語の上位下位関係獲得ができるという可能性を示唆している．また，言語横断共訓練が辞書のクオリティ，カバレージに対しても一定の頑健性を持っていることも実験により確認した．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

1) Adar, E., Skinner, M. and Weld, D.S.: Information arbitrage across multi-lingual Wikipedia, *Proc. 2nd ACM International Conference on Web Search and Data Mining*, pp.94–103 (2009).
2) Ando, M., Sekine, S. and Ishiza, S.: Automatic Extraction of Hyponyms from Japanese Newspaper Using Lexico-syntactic Patterns, *LREC'04: Proc. 4th International Conference on Language Resources and Evaluation* (2004).
3) Auer, S. and Lehmann, J.: What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content, *Proc. 4th European Semantic Web Conference* (*ESWC 2007*), Springer, pp.503–517 (2007).
4) Blum, A. and Mitchell, T.: Combining labeled and unlabeled data with co-training, *COLT'98: Proc. 11th Annual Conference on Computational Learning Theory*, pp.92–100 (1998).
5) Bond, F., Isahara, H., Kanzaki, K. and Uchimoto, K.: Boot-Strapping a WordNet Using Multiple Existing WordNets, *LREC'08: Proc. 6th International Language Resources and Evaluation* (2008).
6) Daumé III, H., Langford, J. and Marcu, D.: Search-Based Structured Prediction

as Classification, *Proc. NIPS Workshop on Advances in Structured Learning for Text and Speech Processing*, Whistler, Canada (2005).
7) Erdmann, M., Nakayama, K., Hara, T. and Nishio, S.: A Bilingual Dictionary Extracted from the Wikipedia Link Structure., *Proc. DASFAA*, pp.686–689 (2008).
8) Fellbaum, C.: *WordNet: An Electronical Lexical Database*, The MIT Press (1998).
9) Girju, R.: Out-of-context noun phrase semantic interpretation with cross-linguistic evidence, *CIKM'06: Proc. 15th ACM International Conference on Information and Knowledge Management*, pp.268–276 (2006).
10) Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P. and Yuret, D.: SemEval-2007 Task 04: Classification of Semantic Relations between Nominals, *Proc. 4th International Workshop on Semantic Evaluations* (*SemEval-2007*), pp.13–18 (2007).
11) Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora, *COLING'92: Proc. 14th Conference on Computational Linguistics*, pp.539–545 (1992).
12) Herbelot, A. and Copestake, A.: Acquiring ontological relationships from Wikipedia using RMRS, *Proc. ISWC 2006 Workshop on Web Content Mining with Human Language Technologies* (2006).
13) Hovy, E., Kozareva, Z. and Riloff, E.: Toward completeness in concept extraction and classification, *Proc. 2009 Conference on Empirical Methods in Natural Language Processing*, Vol.2, pp.948–957 (2009).
14) Joachims, T.: *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*, Kluwer Academic Publishers (2002).
15) Kazama, J. and Torisawa, K.: Exploiting Wikipedia as External Knowledge for Named Entity Recognition, *Proc. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp.698–707 (2007).
16) Li, C. and Li, H.: Word Translation Disambiguation Using bilingual bootstrapping, *Proc. 40th Annual Meeting of the Association for Computational Linguistics*, pp.343–351 (2002).
17) Merlo, P., Stevenson, S., Tsang, V. and Allaria, G.: A multilingual paradigm for automatic verb classification, *Proc. 40th Annual Meeting of the Association for Computational Linguistics*, pp.207–214 (2002).
18) Nastase, V. and Strube, M.: Decoding Wikipedia Categories for Knowledge Acquisition, *Proc. AAAI'08*, pp.1219–1224 (2008).
19) Ruiz-casado, M., Alfonseca, E. and Castells, P.: Automatic extraction of semantic relationships for Wordnet by means of pattern learning from Wikipedia, *Proc. 10th International Conference on Applications of Natural Language to Information Systems*, pp.67–79, Springer Verlag (2005).
20) Suchanek, F.M., Kasneci, G. and Weikum, G.: Yago: A Core of Semantic Knowledge, *Proc. 16th International Conference on World Wide Web*, pp.697–706 (2007).
21) Sumida, A. and Torisawa, K.: Hacking Wikipedia for Hyponymy Relation Acquisition, *Proc. 3rd International Joint Conference on Natural Language Processing* (*IJCNLP*), pp.883–888 (2008).
22) Sumida, A., Yoshinaga, N. and Torisawa, K.: Boosting Precision and Recall of Hyponymy Relation Acquisition from Hierarchical Layouts in Wikipedia, *Proc. 6th International Conference on Language Resources and Evaluation* (2008).
23) Wan, X.: Co-training for cross-lingual sentiment classification, *Proc. Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp.235–243 (2009).
24) Wei, B. and Pal, C.: Cross lingual adaptation: An experiment on sentiment classifications, *Proc. ACL-10* (*Short Papers*), pp.258–262 (2010).
25) Wu, F. and Weld, D.S.: Autonomously semantifying Wikipedia, *CIKM '07: Proc. 16th ACM International Conference on Information and Knowledge Management*, pp.41–50 (2007).

(平成 23 年 4 月 11 日受付)
(平成 23 年 9 月 12 日採録)

呉　　鍾勲

独立行政法人情報通信研究機構ユニバーサルコミュニケーション研究所情報分析研究室専攻研究員．1998 年韓国成均館大学校情報工学科卒業．2005 年 KAIST（韓国科学技術院）電子電算学科電算学専攻博士課程修了．同年 KAIST 研究員を経て，情報通信研究機構に専攻研究員として着任．博士（工学）．自然言語処理の研究に従事．

山田　一郎（正会員）

1991 年名古屋大学工学部情報工学科卒業．1993 年同大学大学院修士課程修了．博士（情報科学）．同年 NHK 入局．1996 年より NHK 放送技術研究所にて自然言語処理を利用した情報抽出，メタデータ生成の研究に従事．2003〜2004 年スタンフォード大客員研究員．2008〜2011 年情報通信研究機構専門研究員．現在，NHK 放送技術研究所主任研究員．映像情報メディア学会，言語処理学会各会員．

内元　清貴（正会員）

1994 年京都大学工学部電気工学第二学科卒業．1996 年同大学大学院修士課程修了．同年郵政省通信総合研究所入所．2009 年内閣府出向．2011 年独立行政法人情報通信研究機構研究マネージャー．現在に至る．博士（情報学）．自然言語処理の研究，研究成果の社会還元活動に従事．言語処理学会，ACL 各会員．

鳥澤健太郎（正会員）

1992 年東京大学理学部卒業．1994 年同大学大学院修士課程修了．1995 年同大学院博士課程中退．同年同大学院助手．1998 年科学技術振興事業団さきがけ研究 21 研究員兼任（2002 年まで）．北陸先端科学技術大学院大学助教授を経て，2008 年より独立行政法人情報通信研究機構言語基盤グループ，グループリーダー．2011 年より同機構情報分析研究室室長，現在に至る．博士（理学）．自然言語処理の研究に従事．日本学術振興会賞等受賞．言語処理学会，人工知能学会，ACL 各会員．

橋本　　力（正会員）

1999 年福島大学教育学部卒業．2001 年北陸先端科学技術大学院大学博士前期課程修了．2005 年神戸松蔭女子学院大学大学院博士後期課程修了．京都大学大学院情報学研究科産学官連携研究員を経て，2007 年山形大学大学院理工学研究科助教，2009 年より独立行政法人情報通信研究機構専攻研究員．2011 年京都大学大学院情報学研究科博士後期課程修了．現在に至る．自然言語処理の研究に従事．博士（言語科学，情報学）．言語処理学会，ACL 各会員．