情報処理学会論文誌　Vol. 52　No. 12　3338–3348 (Dec. 2011)

Webテキストを対象とした語義曖昧性解消のための言語資源の半自動構築

村本英明[†1]　鍜治伸裕[†2]
吉永直樹[†2]　喜連川優[†2]

近年のWeb上のCGM拡大により，社会分析やマーケティングの対象としてWebテキストに大きな注目が集まっている．そうしたテキストから有用な情報を抽出するためには，多義語の意味を正確に区別する処理（語義曖昧性解消）が非常に重要となる．しかし，訓練事例や辞書といった，語義曖昧性解消に必要な言語資源の構築には大きな作業コストが発生することから，高い精度で語義曖昧性解消を行うことは依然として実現困難となっている．特に，Webテキストのように多様な話題を含んだテキストを対象とする場合，この問題はいっそう深刻なものとなる．この問題の解決を図るため，本論文では，既存のWeb資源を活用することによって，語義曖昧性解消に必要な言語資源を半自動的に構築するための方法を提案する．実験においては，Wikipediaと Webテキストに対して提案手法を適用することによって，実際に大規模な言語知識が構築可能であることを確認した．また，それらの言語資源をもとに語義曖昧性解消システムを構築し，その性能についても調査を行った．

Semi-automatically Building Linguistic Resources for Word Sense Disambiguation of Web Text

Hideaki Muramoto,[†1] Nobuhiro Kaji,[†2]
Naoki Yoshinaga[†2] and Masaru Kitsuregawa[†2]

With the recent advent of consumer generated media (CGM) on the Web, the textual data on the Web has been given much attention as a target of social analysis or marketing. To extract useful information from such texts, it is crucial to precisely distinguish meanings of polysemous words (i.e., word sense disambiguation or WSD). However, due to the tremendous labor required to build a large amount of linguistic resources for WSD (e.g., training examples or dictionaries), it is still hard to perform WSD with enough accuracy. This is especially problematic in dealing with Web texts, which contains much more diverse topics than conventional news articles. To overcome this, we present a semi-automatic approach to building those linguistic resources from existing Web data. Our experiments confirmed that the proposed method is indeed able to build much larger linguistic resources than existenting ones. We also investigated the performance of WSD systems learned from those linguistic resources.

1. はじめに

近年のWebの爆発的な普及にともない，人々はブログなどのCGM（Consumer Generated Media）を通じて自由に情報発信を行うことが可能となった．これにより，Web上には人々の意見や感情が表出したテキストが多数流通することとなった．このようなテキストデータは社会分析やマーケティングなどの情報源として高い潜在的価値を有することから，これを解析するための自然言語処理技術に大きな期待が集まっている[8),10),20),21)]．

そうしたテキストから情報を自動抽出する際には，自然言語が持つ多義性，すなわち1つの語句が複数の意味を持ちうることが大きな問題となる．たとえば「ライオン」という語は，少なくとも会社と動物の2つの意味を持つ．そのため，Webテキストから会社の「ライオン」に関する言及を抽出したい場合には，会社の「ライオン」と動物の「ライオン」を何らかの方法で区別する必要がある．

多義語の意味を機械的に区別するためのテキスト処理技術は語義曖昧性解消（word sense disambiguation）と呼ばれ，自然言語処理の分野において古くから研究が行われている．この問題を解くために様々なタスク設定が提案されているが，基本的には分類タスクとして定式化される．すなわち，テキストに出現する語句に対して，あらかじめ定義された意味カテゴリの集合の中から，その文脈に最も適したカテゴリを割り当てるタスクとして扱われることが一般的である[11),16)]．

語義曖昧性解消の処理を実現するためには様々な言語資源が重要な役割を果たす．たとえば，教師あり学習の手法を用いるためには，意味カテゴリの情報が付与された訓練事例が大量に必要となる．また，各語句（e.g., ライオン）に対して，それに割り当てられる可能性がある意味カテゴリ（e.g.,《会社》と《動物》[⋆1]）を列挙した辞書（**意味カテゴリ辞書**と呼

†1 東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, the University of Tokyo
†2 東京大学生産技術研究所
Institute of Industrial Science, the University of Tokyo
⋆1 本論文では意味カテゴリ名は《》で囲って表記する．我々が実際に使用した意味カテゴリ集合については2章を参照されたい．

ぶ）も分類精度を向上させるために有用であると考えられる．

Web テキストを対象として語義曖昧性解消を行う際の問題点の 1 つとして，こうした言語資源の構築が難しいことがあげられる．Web は多様な話題が混在するきわめて大規模なテキストデータである．そのため，そこに出現する言語表現を十分カバーできるような言語資源を構築し，さらに，それを継続的にメンテナンスしていくことは，きわめて負荷の大きな作業であるといわざるをえない．

こうしたことをふまえると，言語資源の構築を完全に手作業で行うことはコスト面から考えて望ましくなく，これを自動化もしくは半自動化しておくことは実用上非常に有用であると考えられる．そこで本論文では，既存の Web 資源を活用することによって，語義曖昧性解消に必要な言語資源を半自動的に構築するための方法を提案する．また，構築された言語資源を用いて語義曖昧性解消の実験を行うことによって提案手法の有効性の検証を行う．

本論文の構成は以下のとおりである．

- まず 2 章では，本論文で議論する語義曖昧性解消のタスク設定を説明する．そして，そのタスクを教師あり学習を用いて解く方法について述べる．
- 3 章では，Wikipedia の記事間リンクを利用することによって，人手で作成した少数の規則から大量の訓練事例を半自動生成する方法を述べる．実験では，従来入手可能であった訓練事例[19),24)] と比較して，10 倍以上の大規模なデータを生成することに成功した（5 章を参照）．
- 4 章では，意味カテゴリ辞書を Web テキストから構築する方法について述べる．提案する意味カテゴリ辞書の構築方法は，古典的な語彙統語パターン[7)] に基づくものであり，技術的に高い新規性があるものではない．しかし，従来の語義曖昧性解消の研究は，人手で作成された辞書を主に使用しており，Web などのテキストから構築された大規模な辞書の有効性は十分な検証が行われておらず，これを行った点が本研究における学術的貢献の 1 つである．
- 5 章では，上記の 2 つの手法を用いて言語資源（訓練事例と意味カテゴリ辞書）を構築した結果を示す．また，これらの言語資源をもとに語義曖昧性解消システムを構築し，その分類精度について報告を行う．
- 6 章では関連研究の紹介を行い，最後に 7 章で本論文のまとめを行う．

2. 語義曖昧性解消手法の概要

本論文で議論するタスクは，ある語句（**分類対象語**と呼ぶ）とその出現文脈（分類対象語が出現する文を考える）が与えられたときに，あらかじめ定義された意味カテゴリのいずれか 1 つを分類対象語に割り当てるという分類タスクである．以下本章では，まず我々のタスクで用いる意味カテゴリ集合について説明したのち（2.1 節），関連研究で用いられている意味カテゴリ集合との比較について述べる（2.2 節）．そして，本タスクを教師あり学習を用いて解く方法を説明する（2.3 節）．

表 1　47 の意味カテゴリ（「ID：カテゴリ名」で表記している）．各カテゴリの詳細については関根の拡張固有表現階層における定義[17)] を参照されたい

Table 1　47 semantic categories, represented in the form of 'ID: category name'. For detailed explanations of each category, interested readers may refer to the definitions of Sekine's extended named entity hierarchy.

1：《人》	2：《神》	3：《国際組織》	4：《公演組織》
5：《家系》	6：《民族》	7：《競技組織》	8：《法人》
9：《政治的組織》	10：《温泉》	11：《GPE》	12：《地域》
13：《地形》	14：《天体》	15：《遺跡》	16：《GOE》
17：《路線》	18：《製品その他》	19：《材料》	20：《衣類》
21：《貨幣》	22：《医薬品》	23：《武器》	24：《賞》
25：《勲章》	26：《罪》	27：《キャラクター》	28：《乗り物》
29：《食べ物》	30：《芸術作品》	31：《出版物》	32：《主義方式》
33：《規則》	34：《称号》	35：《言語》	36：《単位》
37：《催し物》	38：《事故事件》	39：《自然災害》	40：《元素》
41：《化合物》	42：《鉱物》	43：《生物》	44：《生物部位》
45：《動物病気》	46：《自然色》	47：《その他》	

2.1 意味カテゴリ集合

Web テキストのマーケティングおよび社会分析への活用を念頭に置いた場合，会社や政党や製品など，いわゆる固有表現が解析対象として重要となることが予想される．たとえば，Web テキストから評判分析を行う場合には，何らかの商品や，それを製造している会社などに対する言及の抽出が必要になることが考えられる．

そこで，こうした固有表現を詳細に分類したオントロジである関根の拡張固有表現階層[17)] を利用して，意味カテゴリ集合の設計を行った（**表 1**）．我々が用いた意味カテゴリ集合は，拡張固有表現階層の第 2 層に位置する 46 カテゴリ[⋆1]と，そのどれにも対応しない語句の分類先として用意した意味カテゴリ《その他》の合計 47 カテゴリである．本論文における議論および実験は，この意味カテゴリ集合に基づいたものとなっている．これ以外の意味カテ

⋆1《名前_その他》や数値表現や時間表現などを除いてある．

ゴリ集合に対する提案手法の適用可能性については 5.5 節で議論を行う．

2.2　他タスクとの比較

従来の語義曖昧性解消の研究においては，様々な種類の意味カテゴリ集合が語句の分類先として用いられている．最も代表的なのは，WordNet [13] の synset や国語辞典の定義文などを意味カテゴリとして用いるアプローチである[11),16)]．また一方で，そうした従来のアプローチは，実応用において必要とされる以上に詳細な意味カテゴリを導出する傾向があることから[9)]，近年ではより粗い意味カテゴリを用いる動きも見られる．たとえば，国際ワークショップ SemEval-2007 では，英語の語義曖昧性解消タスクの 1 つとして，粗粒度語義曖昧性解消（coarse-graind WSD）が提案されている[15)]．また，超語義タグ付け（supersense tagging）[4)] やクラスに基づく語義曖昧性解消（class-based WSD）[9)] も，これと同様の試みと見ることができる．表 1 の意味カテゴリ集合は，WordNet の synset や国語辞典の定義文と比較して粗いものとなっているため，本研究のタスク設定もこうした粗い粒度の語義設計に基づくものの 1 つと考えることができる．

語義曖昧性解消において粗い粒度の意味カテゴリを用いることは，固有表現認識[14)] と類似したタスクに相当するのではないかと考えることも可能である．しかし，従来の固有表現認識において用いられる意味カテゴリはせいぜい 10 個程度と数が少なく，多義語の意味を区別することを念頭において設計されているとはいい難い．たとえば，IREX [19)] で定義されている意味カテゴリに基づくと「洋服のワンピース」と「漫画のワンピース」はともに《人工物》に分類されてしまうため，これら 2 つの意味を区別することはできない．

固有表現認識の立場から見れば，本研究におけるタスク設定は，従来よりも詳細化された意味カテゴリ集合を用いたものとして位置付けることができる．ただし，これまでに詳細なカテゴリ設計を行っている研究が皆無であるというわけではなく，固有表現認識に関する研究全体から見れば少数ではあるが，いくつかの研究事例が報告されている．たとえば，Sekine らは 200 の意味カテゴリを固有表現認識に用いることを提案している[17),18)]．また Whitelaw らは，32 の意味カテゴリを定義し，それに基づく固有表現認識器の学習を行っている[22)]．

2.3　教師あり学習による解法

本研究では教師あり学習の手法を用いて上記のタスクを解く方法を考える．まず本節では，これを単純な 47 値分類として解く方法を述べる．そして，4 章では，意味カテゴリ辞書を用いて分類先の意味カテゴリを絞り込む方法について議論を行う．

本研究で議論するタスクは，基本的に既存の教師あり学習手法を直接適用することが可能であるが，意味カテゴリ《その他》の扱いには工夫が必要となる．残る 46 の意味カテゴリとは異なり，《その他》は何か具体的な概念を表すものではない．そのため，どのようなデータを学習の正例として使えばよいのか自明ではない．

これに対して我々は 1 対他法（one-versus-the-rest）に基づく単純な解決策を講じた．まず学習時には《その他》を除く 46 の意味カテゴリを対象として，1 対他法を用いて分類器の学習を行う．このとき，カテゴリ数が 46 と多いことから，負例数が正例数に比べて極端に多くなる．そこでサンプリングによって負例数を削減し，正例と負例の数を均一にしたのちに学習を行う[3)]．分類時には，通常の 1 対他法を用いた場合と同様の処理を行うが，すべての分類器が負例と判断した事例は《その他》に分類する．これにより《その他》に対する正例を明示的に与えることなく，47 値分類器を構築することが可能となる．

学習アルゴリズムには，高速なオンライン学習手法として知られる平均化パーセプトロン[6)] を用いる．素性は，分類対象語が出現する文の bag-of-words と，分類対象語の前後の単語の表層形 n-gram（n=1, 2, 3），および，分類対象語の係り先の動詞の原形を用いる．ただし，訓練事例中に 5 回未満しか出現しない低頻度語とストップワードは素性から除外する．

3.　Wikipedia を用いた訓練事例の半自動生成

高精度な分類器を学習するためには大量の訓練事例が必要不可欠となる．本章では，Wikipedia の記事間リンクを用いることによって訓練事例を半自動生成する手法について述べる．

3.1　Wikipedia

Wikipedia は，固有名詞を中心とする様々な事物に関する常識的知識を記述した大規模な百科辞典サイトである．我々の提案手法は，次に述べるような Wikipedia の特徴を利用したものとなっている．

- Wikipedia のデータは記事と呼ばれる単位から構成される．Wikipedia 記事においては，その見出し語の簡潔な説明文が先頭に置かれることが多い．たとえば，記事「トヨタ自動車」の先頭文は次のようになっている．

  > トヨタ自動車株式会社は，愛知県豊田市と東京都文京区に本社を置く日本最大の自動車メーカーであり日本最大の企業．

  このような文は見出し語の定義文と見なすことができるため，以下では Wikipedia 記事の先頭文のことを定義文と呼ぶ．

- Wikipedia 記事においては，別記事へのハイパーリンクをユーザが自由に作成することができる．こうした記事間リンクは Wikipedia のソーステキストにおいて
  [リンク先の記事の見出し | アンカーテキスト]
  という形式で記述される．以下に具体例を示す．

(1)　a. プリウスは [トヨタ自動車 | トヨタ] が発売したハイブリッドカーである．
　　b. プリウスは トヨタ が発売したハイブリッドカーである．

(1a) は Wikipedia のソーステキストの例である．これは，画面上では (1b) のように表示され，下線部「トヨタ」が記事「トヨタ自動車」へのハイパーリンクになっている．

3.2　記事間リンクに基づく意味カテゴリラベルの自動付与

提案手法は，記事と意味カテゴリとの対応関係をもとに，意味カテゴリラベルを Wikipedia 中のテキストに自動付与することによって，訓練事例を半自動的に生成する．たとえば (1) において，リンク先の記事「トヨタ自動車」が意味カテゴリ《法人》に対応するという情報が，事前に人手で与えられていたとする．そうすれば，以下に示すように，分類対象語「トヨタ」に意味カテゴリラベル《法人》を付与した訓練事例を生成することができる．

(2)　a. プリウスは トヨタ《法人》が発売したハイブリッドカーである．

このとき，全記事に対して意味カテゴリとの対応関係を手作業で記述するという方法は明らかに非効率的である．そこで，Wikipedia の定義文の末尾にはその記事[⋆1]の上位語が出現しやすいことに着目し，記事と意味カテゴリではなく，上位語と意味カテゴリの対応を人手で記述することによって訓練事例を効率的に生成する．

訓練事例の生成方法は具体的には以下のとおりである．まず，上位語と意味カテゴリの対応ルールを人手で記述する．次に，各記事の定義文から隅田らの手法[25] を用いて上位語を抽出し，対応ルールと照合することによって，記事と意味カテゴリの対応関係を得る．そして最後に，記事間リンクに基づいてアンカテキストに意味カテゴリタグを自動付与する．**表 2** に，我々が記述した対応ルールと，そのルールに基づいて意味カテゴリとの対応関係が得られる記事の例を示す．

以下では，対応ルールに含まれる上位語のことを**シード上位語**，対応ルールを介して意味カテゴリとの対応が得られた記事のことを**シード記事**と呼ぶ．

⋆1 ここでは正確には記事の見出し語のことであるが，文脈から明らかな場合は記事の見出し語のことを単に記事と呼ぶこととする．

表 2　上位語と意味カテゴリの対応ルール，および意味カテゴリとの対応関係が得られる記事
Table 2　Mapping rules between hypernyms and semantic categories, and Wikipedia articles that are associated with the semantic categories.

対応ルール	意味カテゴリへの対応付けが得られる記事
メーカー →《法人》	ライオン (企業)，ソニー，…
銀行 →《法人》	三菱東京 UFJ 銀行，三井住友銀行，…
哺乳類 →《生物》	ライオン (生物)，うさぎ，…
魚類 →《生物》	ウナギ，アナゴ，…

3.3　対応ルールの作成

訓練事例の生成に必要な作業コストを最小限におさえるという観点からは，できる限り少ない数の対応ルールで，大量の訓練事例を生成できることが望ましい．そのためには，場当たり的に対応ルールを記述するのではなく，作業を効率化させる何らかの指針があれば有用であると考えられる．

ある上位語 h の対応ルールを記述した場合，h を上位語とする全記事へのインリンク数 LinkNum(h) が，その対応ルールを記述した結果として得られる訓練事例数となる．たとえば，上位語「メーカー」と意味カテゴリ《法人》の対応ルールを記述すれば，上位語が「メーカー」である全記事（トヨタ自動車，ソニーなど）に対するインリンク数が，得られる訓練事例の数となる．

このとき，単語の頻度分布におけるジップ則と同様の傾向が，上位語 h と LinkNum(h) の間にも存在することが期待できる．すなわち，LinkNum(h) の値が大きい上位語 h の種類は少なく，一部のそうした上位語について対応ルールを記述すれば，すべての上位語に対して対応ルールを記述した場合と比べても遜色のない数の訓練事例が得られる可能性がある．

このような考えに基づき，LinkNum(h) の値が大きな上位語 h から順に対応ルールの記述を行い，十分な数の訓練事例が得られた時点でルールの記述を終了することとした．ただし，上位語が多義である場合には，単純な対応ルールを記述することが難しいため，多義語と判断した上位語はルール記述の対象から除外した．実際に記述した対応ルール数および得られた訓練事例数は 5 章において報告する．

4.　Web テキストからの意味カテゴリ辞書の構築

4.1　分類先の絞り込み

2 章で定義した意味カテゴリの総数は 47 個であるが，ある分類対象語に対して実際に考

表 3　意味カテゴリ辞書の例
Table 3　Examples of the semantic category dictionary.

分類対象語	意味カテゴリ
ライオン	《法人》《生物》
オレンジ	《食べ物》《自然色》《芸術作品》

慮すべき意味カテゴリは，その一部のみであると考えられる．たとえば，我々が調べた限り「ライオン」は《法人》《生物》《芸術作品》の 3 カテゴリだけを分類先として考えれば十分である．

こうしたタスク特性を利用するため，**表 3** のような意味カテゴリ辞書を利用することを考える．辞書に登録されている意味カテゴリだけを分類先として考慮すれば，単純に 47 値分類として解いた場合よりも分類精度が向上することが期待できる．ただし，意味カテゴリ辞書は必ずしも網羅的である必要はなく，分類対象語が意味カテゴリ辞書に登録されていなかった場合には，すべての意味カテゴリを分類先候補とする．

一般的な語義曖昧性解消においては，WordNet や国語辞典といった既存の意味カテゴリ辞書を利用して，分類先カテゴリの絞り込みが行われる．しかし，Web のように，新語や固有名などの未知語が豊富なテキストを処理とする場合には，既存の辞書資源による絞り込みには十分な効果を期待することが難しい．そこで我々は Web テキストから意味カテゴリ辞書を構築し，それを絞り込みに用いることを提案する．

4.2　辞書構築手法

意味カテゴリ辞書は以下の手順で構築を行う．まず，語彙統語パターン[7] を用いることによって，Web テキストから上位下位関係にある語句の組 (w, h) を抽出する．ただし，ここでは h が w の上位語であるとする．実験では，安藤ら[23] と隅田ら[26] の研究を参考にして，以下の語彙統語パターンを用いた．

w という h
w などの h
w 以外の h
h「w」

次に，たとえば意味カテゴリ《法人》に対する「企業」や「メーカー」のような，各意味カテゴリ c に特徴的な上位語 h_c の収集を行う．ここでは，以下のいずれかの条件を満たす語句を h_c とした．

条件 1　意味カテゴリ c のシード上位語（3 章）．

条件 2　上記の語彙統語パターンにおいて，意味カテゴリ c のシード記事（3 章）と共起しやすい上位語 h．具体的には，以下の確率 $p(c\,|\,h)$ が閾値 τ を超える h．

$$p(c\,|\,h) = \frac{freq(c,h)}{\sum_{c' \in \mathcal{C}} freq(c',h)}$$

ここで，$freq(c,h)$ は意味カテゴリ c のシード記事と上位語 h が，4 つの語彙統語パターンにおいて共起した回数の総和を表す．また $\mathcal{C}$ は全 47 カテゴリの集合である．

最後に，4 つの語彙統語パターンのいずれかにおいて，語句 w が上位語 h_c と 1 回でも共起すれば，w は意味カテゴリ c に分類される可能性があると考え，それらの組を意味カテゴリ辞書に登録する．

5. 評 価 実 験

本章ではまず，訓練事例の生成結果および，それを用いて学習した分類器の精度について報告する．そして次に，Web テキストから構築した意味カテゴリ辞書による絞り込み手法の効果の検証を行う．

5.1　実験の設定

訓練事例の半自動生成には，2010 年 3 月 17 日の Wikipedia データを用いた．Wikipedia の先頭文からの上位語の抽出には上位下位抽出ツール[⋆1]を利用し，形態素解析には MeCab[⋆2]，構文解析には J.DepP[⋆3]を利用した．意味カテゴリ辞書構築のための Web テキストとしては，2006 年から 2009 年の間に収集したブログ記事約 20 億文を用いた．

分類精度の評価に用いるデータは，訓練事例と同様に Wikipedia の記事間リンクを利用して作成した．しかし，評価事例を作成するさいには，3 章の半自動手法とは異なり，Wikipedia 記事と意味カテゴリの対応付けをすべて手作業によって行った．評価事例における分類対象語としては，一義語と多義語を別々に Wikipedia データから無作為に選択したものを用いた（**表 4** と**表 5**）．

実際に Web テキストに対して語義曖昧性解消処理を行う場合には，訓練事例に出現していない語が分類対象となる可能性がある．このことを考慮するため，評価事例における分類対象語が訓練事例にも出現していた場合には，それを訓練事例から除外して分類器の学習を行った．

⋆1 http://alaginrc.nict.go.jp/hyponymy/index.html
⋆2 http://mecab.sourceforge.net/
⋆3 http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/jdepp/

表 4　多義語データにおける分類対象語（カッコ内は評価事例中で付与されている意味カテゴリの ID）
Table 4　Target words in the polysemous data. The numbers in the parentheses are the indices of the semantic categories assigned in the data.

イルカ (1, 43), エセックス (11, 23, 28), オアシス (4, 12, 13, 47), オセロ (4, 18, 30), オレンジ (11, 29, 43, 46), カサブランカ (11, 30), クリーム (4, 29), サラトガ (11, 23, 28), サルサ (29, 30, 32, 47), シカゴ (4, 11), ジェネシス (4, 23, 28), ジャングル (12, 13, 30, 32, 47), スズキ (8, 43), セオドア・ルーズベルト (1, 23, 28), タンポポ (4, 43), ハイヒール (4, 20), ハンニバル (1, 30), ホーネット (18, 23, 28), ボール (18, 23, 47), レベル (8, 36, 47), レンジャー (23, 28, 34), レーダー (1, 2, 18, 23, 27), ヴァルナ (2, 11), 安全地帯 (4, 47), 少年 (31, 47), 忍者 (4, 34), 石 (36, 42, 47), 羞恥心 (4, 30, 47), 銀河 (14, 17, 23, 28), 長征 (28, 38)

表 5　一義語データにおける分類対象語（カッコ内は評価事例中で付与されている意味カテゴリの ID）
Table 5　Target words in the monosemous data. The numbers in the parentheses are the indices of the semantic categories assigned in the data.

つるの剛士 (1), もののけ姫 (30), アトランタ (11), アルストム (8), インド洋 (13), ウラジオストク (11), オペレッタ (30), セルビア人 (1), セーラー服 (20), ソルボンヌ大学 (16), ツァーリ (34), テイチクエンタテインメント (8), テニス (32), ヒスパニック (6), ボーイング (8), ヤッターマン (27), ロッテルダム (11), 京都 (11), 信越放送 (8), 名古屋市営地下鉄 (17), 国立劇場 (16), 執事 (34), 強制収容所 (16), 待合室 (16), 情熱大陸 (30), 戦闘機 (28), 最高経営責任者 (34), 朝日放送 (8), 漢 (6), 漫才師 (34), 特殊相対性理論 (32), 笑福亭鶴瓶 (1), 紀行番組 (30), 統一地方選挙 (37), 西条市 (11), 連合艦隊司令長官 (34), 重油 (18), 高射砲 (23), 高野山 (13), 魔法のプリンセスミンキーモモ (30)

5.2　訓練事例の半自動生成

実験では 600 個の対応ルールを記述することにより，約 540 万の訓練事例を生成することができた（**表 6**）．これは，仮に Wikipedia の全記事間リンクを使うことができたとして，その場合に得られる事例数の約 80%に相当する量であった．また，対応ルール作成に要した作業時間はおよそ 1 日であった．この結果から，提案手法が効率的に訓練事例を生成可能であることを確認することができた．

提案手法によって生成された訓練事例と，橋本ら[24] によって作成された「拡張固有表現タグ付きコーパス」の比較を行ったところ，我々が半自動生成した訓練事例は「拡張固有表現タグ付きコーパス」の 10 倍以上の規模であることが分かった（**表 7**）．拡張固有表現タグ付きコーパスは，拡張固有表現階層に準拠した意味カテゴリタグを付与した言語資源としては，現在のところ最大規模のものである．このことから，少ない作業量で大量の訓練事例を生成するという，本研究の目的は十分に達成されたと考えることができる．

表 6　意味カテゴリごとの訓練事例数
Table 6　Number of training examples in each semantic category.

意味カテゴリ	訓練事例数	意味カテゴリ	訓練事例数
《人》	859,435	《賞》	18,134
《神》	14,454	《勲章》	2,679
《国際組織》	4,077	《罪》	7,203
《公演組織》	34,195	《キャラクター》	16,621
《家系》	29,665	《乗り物》	73,036
《民族》	13,047	《食べ物》	42,667
《競技組織》	76,437	《芸術作品》	382,701
《法人》	395,991	《出版物》	58,750
《政治的組織》	183,641	《主義方式》	196,006
《温泉》	4,286	《規則》	54,937
《GPE》	1,294,979	《称号》	103,700
《地域》	10,614	《言語》	84,948
《地形》	209,546	《単位》	35,871
《天体》	15,173	《催し物》	85,566
《遺跡》	2,160	《事故事件》	94,170
《GOE》	459,867	《自然災害》	5,842
《路線》	160,917	《元素》	16,781
《製品その他》	172,486	《化合物》	50,605
《材料》	2,283	《鉱物》	5,137
《衣類》	4,742	《生物》	92,794
《貨幣》	3,871	《生物部位》	10,946
《医薬品》	3,040	《動物病気》	16,434
《武器》	22,484	《自然色》	5,719
		合計	5,438,637

表 7　既存の言語資源との比較
Table 7　Comparison with existing linguistic resources.

	橋本ら[24]	本研究
用いた言語資源	新聞，白書	Wikipedia
作成方法	人手	半自動
訓練事例の総数	326,966	5,438,637

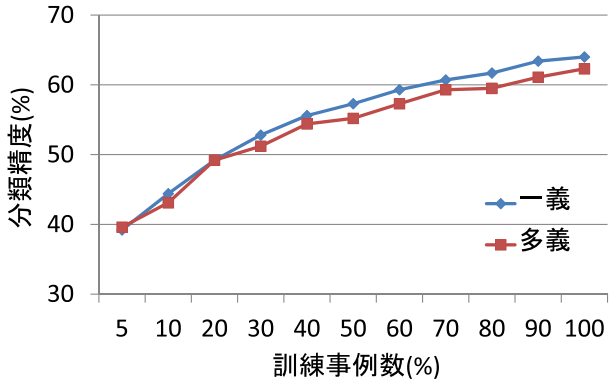


図 1　訓練事例数と分類精度の関係
Fig. 1　Relation between the number of training examples and classification accuracy.

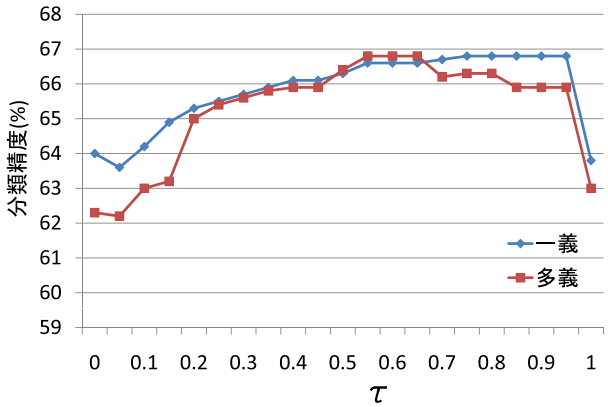


図 2　意味カテゴリ辞書による絞り込みの効果
Fig. 2　Effect of pruning by the semantic category dictionary.

この訓練事例を用いて分類器の学習を行い，評価事例に対する分類精度を調べた．このとき，訓練事例数と分類精度の関係もあわせて調査するために，ランダムサンプリングにより訓練事例数を変化させながら分類精度を計測した（**図 1**）．図の横軸は訓練事例の数であり，最も右側の点が，本実験で得られた全訓練事例を用いて学習した場合の結果に対応する．また縦軸はすべての分類対象語（表 4 および表 5）に対する平均分類精度であり，赤線と青線はそれぞれ多義語データと一義語データの結果を表している．

図 1 からは，一義語データおよび多義語データの両方において，訓練事例数の増加とともに，分類精度が単調に上昇していることが見て取れる．この結果から，提案手法が生成した訓練事例は分類器の学習に有用であることが分かった．これと同時に，訓練事例の規模拡大が精度向上に有効であることも確認することができた．これは，半自動的な手法によって，大量の訓練事例を取得しようとする本研究のアプローチの有効性を示す結果といえる．

5.3　意味カテゴリ辞書の効果検証

次に，構築した意味カテゴリ辞書を用いて絞り込みを行った場合の分類精度を調査した．辞書構築手法にはパラメータ τ が存在するため，実験ではこれを 0 から 1 までの間で変化させた（**図 2**）．τ=0 の場合は，すべての上位語が h_c に含まれるようになるため，絞り込みを行わない場合（すなわち図 1 において全訓練事例を用いた場合）とまったく同じ精度となることに注意されたい．また τ=1 の場合は，4.2 節の条件 1 のみを用いて，条件 2 を用いなかった場合にあたる．

この図から，多義語データ（赤線）と一義語データ（青線）のいずれにおいても，意味カテゴリ辞書を用いて絞り込みを行うことによって，絞り込みを行わなかった場合（τ=0）よりも分類精度が向上していることが分かる．精度向上の度合いは τ の値にもよるが，各データにおいておよそ 3%から 5%の精度向上が実現されており，我々が構築した意味カテゴリ辞書の有用性を示す結果が得られたといえる．最適な τ の値をどのように選択するかという技術的課題は残されているが，少なくとも本実験で試したほぼすべての τ において精度向上が確認できたため，これは実用上大きな問題にはならないと考えられる．

多義語データにおいては，τ の値がおよそ 0.5 から 0.7 の領域において最も高い精度が観察された．このような傾向が見られた理由として，次のようなことが考えられる．まず，τ の値が小さすぎる場合には，辞書には誤った意味カテゴリが大量に登録されることとなり，絞り込みの効果が小さくなっていると考えられる．一方，逆に τ を大きく設定しすぎると，辞書から正解カテゴリが抜け落ちてしまい，やはり分類精度が低下していると考えられる．

一方，一義語データにおいて同様の傾向は見られず，τ=1 の場合を除き，τ の値を上げるにつれて分類精度はほぼ単調に上昇した．このような結果が得られた詳細な理由について現時点で十分な結論を導くことは難しいが，多義語と一義語では有効な辞書構築手法が異なる可能性もあり，今後さらに調査を進めて手法の改善を行っていきたい．

以下に，文脈情報のみから正解意味カテゴリを判断することが困難であっても，意味カテ

ゴリ辞書を用いることによって適切に分類できた例を示す．

(3)　セオドア・ルーズベルト 大統領は全国レベルの行政改革を行った．

ここで「セオドア・ルーズベルト」を，たとえば「アメリカ」などの国名と置き換えても，文の意味は通じる．そのため，この「セオドア・ルーズベルト」が《国》であるのか《人》であるのかを文脈情報から判断することは困難であると考えられる．我々の実験では，意味カテゴリ辞書を用いることによって《国》が分類候補から除外されたため，この事例を正しく分類することができた．

自動生成した訓練事例を用いたときの分類精度自体は，決して高い値ではなく，我々の設定したタスクの難しさを示している．タスクが難化した原因の 1 つとしては，従来の語義曖昧性解消とは異なり，評価事例と訓練事例の間に同じ分類対象語が出現しない設定になっていることが考えられる．

5.4　他の分類手法との比較

以下の 4 つの比較手法の実装を行い，提案手法との精度比較を行った．まず，最も単純な比較手法として，評価事例全体において最も多く出現した意味カテゴリを，すべての単語に予測する手法を用いた．

2 つ目の比較手法として，橋本らの拡張固有表現タグ付きコーパス[24] を用いて，提案手法と同一の方法を用いて学習した 47 値分類器を用いた．この手法と提案手法を比較することによって，自動収集された訓練事例が，既存コーパスと比較して，どの程度精度向上に寄与できるのかを調べることができると考えられる．

次に，語義曖昧性解消におけるベースラインとして広く用いられている first sense heuristics [11),16)] を用いた．ただし，この手法は，多義語の語義曖昧性解消を前提として提案されたものであり，一義語に対してはつねに精度が 100%となってしまうため，一義語の評価に用いることは不適切である．そのため，この比較手法については，多義語データに対する精度のみを報告する．

残るもう 1 つの比較手法としては，評価事例を用いて分類対象単語ごとに 10 分割交差検定を行い，評価と学習において同一の分類対象語のみを用いた場合の分類精度を求めた．これは，個別の分類語に対して十分な量の訓練事例が入手可能であるという仮定のもとで，どの程度の分類精度が実現可能かを調べていることに相当する．なお，この比較手法も，first sense heuristics と同様に，一義語データに対しては精度がつねに 100%となるため意味をなさない．そのため，この手法に関しても，多義語データに対する精度のみを報告する．

表 8　他の分類手法との精度比較
Table 8　Accuracy comparison with other classification methods.

	一義語データ	多義語データ
提案手法	64.0	62.3
最頻意味カテゴリ	12.5	11.5
拡張固有表現タグ付きコーパス[24)]	18.4	21.3
First sense	Nan	75.7
評価データを用いた 10 分割交差検定	Nan	95.1

これら 4 つの比較手法のうち最後の 2 つに関しては，訓練事例と評価事例の間に同じ分類対象語が出現する設定で分類実験を行っていることになる．一方，提案手法においては，訓練事例と評価事例の間に同じ分類対象語が含まれていないため，提案手法とこれらの比較手法は直接結果を比較できるものではないことに注意をされたい[⋆1]．

提案手法と 4 つの比較手法の分類精度を**表 8** に示す．提案手法の精度は，最頻意味カテゴリをつねに選択する手法および，橋本らの拡張固有表現タグ付きコーパスを学習に用いた手法の両方を上回っている．この結果から，提案手法の有効性を確認することができる．

橋本らのコーパスを用いた手法が高い精度を達成することができなかった原因としては，コーパスが比較的小規模であり，多くの意味カテゴリにおいて十分な数の訓練事例が蓄積されていないことが考えられる．これと比較すると，提案手法は，大規模な訓練事例を利用することにより，精度を改善することに成功していることが分かる．

一方，first sense heuristics の精度は 75.7%，評価データを用いて交差検定を行った場合の精度は 95.1%と，いずれも提案手法を大きく上回る結果が得られた．このことから，訓練事例と評価事例に同じ分類対象語が含まれているというタスク設定においては，分類が容易になることが示唆される．こうした設定における精度との差をいかに埋めていくかということは，今後の重要な課題になると考えられる．

5.5　分類誤りの分析と今後の課題

以下に実験において分類を誤った例を示す．

(4)　a. 松竹芸能は，1990 年代は森脇健児，近年では オセロ が大ブレイクした
　　b. 日本プロ野球初となる オレンジ 色のホームユニフォーム

⋆1 特に Wikipedia テキストから作成した訓練事例と評価事例において，分類対象語が共通していた場合には，学習がきわめて容易になることが Bunescu ら[2)] によって指摘されている．詳細な議論を文献を参照されたい．

例文 (4a) の「オセロ」は，女性お笑いコンビの「オセロ」の意味であるため正解カテゴリは《公演組織》であるが，分類器は誤ってこれを《人》に分類した．また (4b) において「オレンジ」は《自然色》に分類されるべきであるが，野球チームなどに与えられる意味カテゴリ《競技組織》に分類された．

例文 (4a) においては，前後の文脈のみから「オセロ」が《人》であるのか《公演組織》であるのかを判断することは困難であると思われる．そのため，これを正しく分類するためには，意味カテゴリ辞書を用いた絞り込みにより，意味カテゴリ《人》を候補から除外しておくことが必要になると考えられる．しかし実験においては，以下のような Web テキストに語彙統語パターン（h=芸人，w=オセロ）がマッチし，さらに「芸人」は《人》に特徴的な上位語 h_c であると判定されていた．

(5)　芸人「オセロ」が司会のテレビ番組

そのため，意味カテゴリ辞書には「オセロ」の分類先候補の 1 つとして《人》が登録されており，意味カテゴリ《人》を分類先候補から除外することができなかった．(5) のようなテキストが記述される理由として，人間は《人》と《公演組織》というカテゴリを，必ずしも厳密に区別していないことが考えられる．そのため，今後は，意味カテゴリを再設計することも含めて，この問題に対処する方法を検討していきたい．

一方 (4b) は「オレンジ」の直後に位置する「色」を手がかりとすれば，適切な意味カテゴリが推定可能であるにもかかわらず，分類を誤っている．この理由として「プロ野球」や「ユニフォーム」など，より遠方に位置する語の影響を強く受けすぎていることが考えられる．この問題に対しては，たとえば Villeneuve ら[1] が遠方に出現する語の重みを減衰させるモデルを提案しており，こうした研究成果を参考にすることによって今後対応したい．

本論文では，表 1 にある 47 の意味カテゴリ集合を分類先として議論および実験を行ってきたが，意味カテゴリ集合の定義によっては，上位語と意味カテゴリの対応ルールを記述することができなくなり，提案手法を適用することが難しくなる可能性がある．たとえば，表 1 には《芸術作品》という意味カテゴリが含まれているが，これの代わりに《絵画》や《映画》や《テレビ番組》などの，より詳細な意味カテゴリを用いることを考える．我々の調べた限り《映画》や《絵画》カテゴリに所属する Wikipedia 記事の多くは，上位語が「作品」となっているため，上位語からそれが《映画》なのか《絵画》なのかを判別することは難しい．我々の提案する枠組みにおいて，Wikipedia の定義文に出現する上位語よりも詳細な意味カテゴリを扱うことは本質的に困難であり，今後はこれに対応可能な方法についても考えていくことが重要であろう．

一方，より粒度の粗い意味カテゴリを用いた場合であれば，そのような問題が発生する可能性は低く，提案手法が有効に働くことが期待できる．たとえば，IREX の意味カテゴリを用いた場合，単位ルールあたりの作業コストは今回の実験で用いた意味カテゴリと同程度であると仮定すると，およそ 1 日程度の作業時間で，600 個の対応ルールを作成することができると考えられる．また，一般的に，カテゴリ数が少なくなれば分類問題として解きやすくなると考えられる．そのため，その 600 個の対応ルールもとに分類器を構築して，IREX の意味カテゴリ分類を行った場合には，今回の実験で得られた精度と同等もしくはそれ以上の精度が期待できる．

6. 関 連 研 究

Wikipedia を利用して語義曖昧性解消を行うというアイデアは，Bunescu ら[2] や Cucerzan[5] によっても提案されている．しかし，彼らの手法においては，Wikipedia 記事が語句の分類先となっているため，Wikipedia に登録されていない語句には適切な分類先を設定できないことが問題となる．これに対して提案手法は，Wikipedia に未登録の語句であっても，表 1 の意味カテゴリに適切な分類先が存在すれば，扱うことができるところが利点である[⋆1]．

Mihalcea[12] は，本研究と同様に，意味カテゴリを分類先とした語義曖昧性解消システムを Wikipedia から学習している[⋆2]．しかし，Wikipedia 記事と意味カテゴリの対応付けを人手で行っているため，大規模な訓練事例を生成する場合には作業コストが問題となる．実際，Mihalcea の実験において用いられた訓練事例数は 9,489 と少ない．これに対して本論文は，頻出する上位語に着目することによって作業量が軽減可能であることを指摘し，現実的な作業時間（およそ 1 日）で大規模なコーパスが構築できることを示した．

Whitelaw ら[22] は，固有表現認識のための訓練事例を Web テキストから自動生成する方法を提案している．しかし，彼らの手法は英語が持つ言語特性や特定のツールに強く依存したものとなっており，英語以外の言語に対してどのように適用すればよいのか自明ではない．これに対して，提案手法において必要となるのは，Wikipedia および少数の語彙統語パ

⋆1 Bunescu らは Wikipedia に登録されていない語義を扱う方法についても言及しているが，見出し語自体が登録されていない場合や，複数の語義が未登録である場合に関する議論は見られない．
⋆2 Mihalcea は WordNet をもとに意味カテゴリを定義している．

ターンのみであり，英語以外の言語に対しても，比較的容易に適用できることが期待される．

一般的な語義曖昧性解消の研究においては，分類対象語が所属しうる意味カテゴリを過不足なく登録した辞書資源の存在が仮定されている[11),16)]．しかし，実際に語義曖昧性解消システムを使用するときに，そのような辞書が必ずしも入手可能であるとは限らない．こうした問題意識から，我々は意味カテゴリ辞書を自動構築し，それを用いた絞り込みの効果について検証を行った．同様のことは藤井ら[27)]によっても議論されているが，彼らは，意味カテゴリ辞書の構築と，評価事例の作成の両方において，Wikipedia のリンク先の記事という同一の情報源を利用している．そのため，彼らの実験結果自体は非常に有用な知見を含んでいるものの，実際に語義曖昧性解消処理を行う場合よりも，意味カテゴリの絞り込みが容易なタスク設定で実験が行われていると考えられる．これに対して我々は，意味カテゴリ辞書と評価事例の作成にはまったく異なる情報源を利用して実験を行っている．この点において，本論文は，藤井らが議論を行ったのと同様の問題に対し，より実応用に近い設定において実験を行ったものと位置付けることができる．

7. おわりに

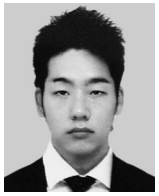
本論文では，Web テキストを対象とした語義曖昧性解消処理を実現するために必要な言語資源を半自動的に構築する手法について述べた．実験では，Wikipedia と Web テキストからそれぞれ訓練事例と意味カテゴリ辞書の構築を行い，大規模な言語知識が構築可能であることを確認した．また，構築された言語資源を用いて分類器の学習を行い，その有用性についても検証を行った．今後は，情報抽出などの実際の応用における語義曖昧性解消処理の有用性についても検証を行っていきたい．

参考文献

1) Brosseau-Villeneuve, B., Nie, J. and Kando, N.: Towards an optimal weighting of context words based on distance, *Proc. COLING*, pp.107–115 (2010).
2) Bunescu, R. and Pasca, M.: Using Encyclopedic Knowledge for Named entity Disambiguation, *Proc. EACL*, pp.9–16 (2006).
3) Chawla, N., Japkowicz, N. and Kotcz, A.: Editorial: Special Issue on Learning from Imbalanced Data Sets, *SIGKDD Explor. Newsl.*, Vol.6, pp.1–6 (2004).
4) Ciaramita, M. and Y., A.: Broad-coverage Sense Disambiguation and Information Extraction with a Supersense Sequence Tagger, *Proc. EMNLP*, pp.594–602 (2006).
5) Cucerzan, S.: Large-Scale Named Entity Disambiguation Based on Wikipedia Data, *Proc. EMNLP*, pp.708–716 (2007).
6) Freund, Y. and Schapire, R.E.: Large Margin Classification Using the Perceptron Algorithm, *Machine Learning*, Vol.37, No.3, pp.277–296 (1999).
7) Hearst, M.: Automatic Acquisition of Hyponyms from Large Text Corpora, *Proc. COLING*, pp.539–545 (1992).
8) Inui, K., Abe, S., Hara, K., Morita, H., Sato, C., Eguchi, M., Sumida, A. and Murakami, K.: Experience Mining: Building a large-scale database of personal experiences and opinions from Web documents, *Proc. WI-IAT*, pp.314–321 (2008).
9) Izquierdo, R., Suárez, A. and Rigau, G.: An Empirical Study on Class-based Word Sense Disambiguation, *Proc. EACL*, pp.389–397 (2009).
10) Kitsuregawa, M., Tamura, T., Toyoda, M. and Kaji, N.: Socio-Sence: A system for analysing the societal behavior from long term Web archive, *Proc. APWeb*, pp.1–8 (2008).
11) McCarthy, D.: Word Sense Disambiguation: An Overview, *Language nad Linguistics Compass*, Vol.3, pp.537–558 (2009).
12) Mihalcea, R.: Using Wikipedia for Automatic Word Sense Disambiguation, *Proc. HLT-NAACL*, pp.196–203 (2007).
13) Miller, G.A.: WordNet: A Lexical Database for English, *Comm. ACM*, Vol.38, pp.39–41 (1995).
14) Nadeau, D. and Sekine, S.: A Survey of Named Entity Recognition and Classification, *Lingvisticae Investigationes*, Vol.30, No.1, pp.3–26 (2007).
15) Navigli, R., Litkowski, K.C. and Hargraves, O.: SemEval-2007 Task 07: Coarse-Grained English All-Words Task, *Proc. SemEval-2007*, pp.30–35 (2007).
16) Navigli, R.: Word Sense Disambiguation: A Survey, *ACM Computing Surveys*, Vol.41, pp.1–69 (2009).
17) Sekine, S., available from ⟨http://sites.google.com/site/extendednamedentityhierarchy⟩.
18) Sekine, S., Sudo, K. and Nobata, C.: Extended Named Entity Hierarchy, *Proc. LREC*, pp.1818–1824 (2002).
19) Sekine, S. and Isahara, H.: IREX: IR and IE Evaluation Project in Japanese, *Proc. LREC* (2000).
20) Shinzato, K., Shibata, T., Kawahara, D., Hashimoto, C. and Kurohashi, S.: TSUBAKI: An open search engine infrastructure for developing new information access methodology, *Proc. IJCNLP*, pp.189–196 (2008).
21) Torisawa, K., De Saeger, S., Kakizawa, Y., Kazama, J., Muratam, M., Noguchi, D. and Sumida, A.: TORISHIKI-KAI: An Autogenerated Web search directory, *Proc. ISUC*, pp.179–186 (2008).
22) Whitelaw, C., Kehlenbeck, A., Petrovic, N. and Ungar, L.: Web-scale named entity recognition, *Proc. CIKM*, pp.123–132 (2008).

23) 安藤まや，関根　聡，石崎　俊：定型表現を利用した新聞記事からの下位概念単語の自動抽，情報処理学会研究報告，pp.77–82 (2003).
24) 橋本泰一，乾　孝司，村上浩司：拡張固有表現タグ付きコーパスの構築，情報処理学会研究報告，自然言語処理（NL-188-17），pp.113–120 (2008).
25) 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.1, No.3, pp.3–24 (2009).
26) 隅田飛鳥，鳥澤健太郎，新里圭司：全文検索エンジンを用いた単なる名詞列からの概念具体物関係の抽出，自然言語処理学会第 12 回年次大会，pp.504–507 (2006).
27) 藤井裕也，飯田　龍，徳永健伸：Wikipedia 記事を利用した曖昧性のある表現の固有表現クラス分類，言語処理学会第 16 回年次大会，pp.15–18 (2010).

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



村本　英明

2009 年東京大学工学部電子情報工学科卒業．2011 年同大学大学院情報理工学系研究科修士課程修了．情報理工学修士．自然言語処理の研究に従事．現在は株式会社三菱総合研究所に勤務．

鍛治　伸裕（正会員）

2005 年東京大学大学院情報理工学系研究科博士課程修了．情報理工学博士．現在，東京大学生産技術研究所特任助教．自然言語処理の研究に従事．

吉永　直樹（正会員）

2005 年東京大学大学院情報理工学系研究科博士課程修了．2002 年より 2008 年まで日本学術振興会特別研究員（DC1，PD）．2008 年 4 月より東京大学生産技術研究所特任助教．博士（情報理工学）．計算言語学と機械学習の研究に従事．

喜連川　優（フェロー）

東京大学大学院工学系研究科情報工学専攻博士課程修了（1983 年）．工学博士．東京大学生産技術研究所講師．助教授を経て，現在，同教授．東京大学地球観測データ統融合連携研究機構長．東京大学生産技術研究所戦略情報融合国際研究センター長．文部科学官．文部科学省「情報爆発」特定研究領域代表（2005〜2010 年），経済産業省「情報大航海プロジェクト」戦略会議委員長（2007〜2009 年），情報処理学会フェロー，副会長（2008〜2009 年），データベース工学の研究に従事．