時系列テキストを用いた恒久性と一意性に基づく関係の分類*

高久　陽平[†a)]　吉永　直樹[††b)]　鍛治　伸裕[††c)]　豊田　正史[††d)]
喜連川　優[††e)]

Identifying Constant and Unique Relations by Using Time-Series Texts*

Yohei TAKAKU[†a)], Naoki YOSHINAGA[††b)], Nobuhiro KAJI[††c)],
Masashi TOYODA[††d)], and Masaru KITSUREGAWA[††e)]

あらまし　ウェブに存在する膨大な量のテキストを知識源として，固有表現間の関係知識を獲得する研究が盛んに行われている．しかしながら，現実世界は刻一刻と変化しているため，テキストに含まれる関係知識の中には，現在では既に成り立たない（現在成立していても将来的に成り立たなくなる）関係が存在する．そのため，テキストから得られた関係を整合性のとれた知識として集積する際に問題が生じる．このような問題を解決するために，本研究では恒久性と一意性に基づく関係の分類を提案する．我々は，大規模時系列テキストから得られる時系列頻度情報と言語情報に基づく素性を導出し，機械学習の分類問題として各分類問題を定式化する．提案手法を用いて実験を行った結果，時系列頻度情報が恒久性の分類において再現率，一意性の分類において精度の向上にそれぞれ有効であることを確認した．

キーワード　関係抽出，情報抽出，自然言語処理

1. まえがき

大規模化するウェブテキストを知識源とし，固有表現間の関係知識を獲得する研究が近年盛んに行われている [1]～[7]．こうした研究の成果によって，膨大な数の関係知識が得られるようになり，質疑応答システムや含意関係判定 [8] など実世界知識を必要とする高度な自然言語処理技術が実現されつつある．

従来の関係知識獲得に関する研究では，テキストから得られた関係知識を全てそのまま蓄積し，利用することが暗黙的に仮定されている．しかしながら，テキストから得られる関係は，そのテキストが書かれた時点においては成立していても，時間の経過に伴い成り立たなくなってしまう可能性があるため，上記のような仮定は明らかに不適当である．例えば，以下のような文から獲得される関係について考えてみよう．

(1) a. 1Q84 は 村上春樹 により書かれた．
　　b. モーゼル川 は ドイツ を流れている．
　　c. 米国 の大統領は ジョージ・ブッシュ である．
　　d. ペンタックス は K-5 を販売している．

ここで，下線部は固有表現，太文字はそれらの間の関係を表している．1 a と 1 b に記述されている関係は，時間によらず常に成立すると考えられる．そのため，これらの文からは，単純に関係を抽出して集積しても問題はない．しかしながら，1 c と 1 d に記述されている関係は時間的に変化し得る．まず，1 c から抽出される関係は，米国の大統領が交代すれば成り立たなくなってしまう．そのため，別の文から米国大統領に関する新しい関係知識（「米国の大統領はバラク・オバマである」など）が獲得された場合には，1 c から得られた古い関係を上書きする必要がある．一方，

† 東京大学大学院情報理工学系研究科，東京都
Graduate School of Information Science and Technology, The University of Tokyo, 7–3–1 Hongo, Bunkyo-ku, Tokyo, 113–0133 Japan
†† 東京大学生産技術研究所，東京都
Institute of Industrial Science, The University of Tokyo, 4–6–1 Komaba, Meguro-ku, Tokyo, 153–8504 Japan
a) E-mail: takaku@tkl.iis.u-tokyo.ac.jp
b) E-mail: ynaga@tkl.iis.u-tokyo.ac.jp
c) E-mail: kaji@tkl.iis.u-tokyo.ac.jp
d) E-mail: toyoda@tkl.iis.u-tokyo.ac.jp
e) E-mail: kitsure@tkl.iis.u-tokyo.ac.jp
* 本論文は学生論文特集秀逸論文である．

1 d の場合，新しい関係が獲得されても，必ずしも古い関係を削除する必要はない．

上記のような関係知識の管理を可能にするため，本論文では関係を恒久性と一意性(注1)に基づいて分類することを提案する．例えば，1 a と 1 b は恒久性がある関係である．一方，1 c と 1 d は共に恒久性がない関係であるが，1 c は一意性があるのに対して，1 d は一意性がない．

提案手法では，与えられた関係が恒久性を有するか，あるいは一意性を有するかを教師あり学習における分類問題としてそれぞれ定式化する．我々は，大規模時系列テキストを用いて，分類に有効な素性を導出する．具体的には，時間窓を用いた頻度情報と言語情報を用いることで，分類を可能とした．

実験では，約 6 年間の日本語ブログアーカイブ（約 23 億文）から獲得された関係知識から 1000 関係を選択し，提案手法の分類精度を評価した．その結果，時系列テキストを用いた素性により，分類精度が顕著に向上することを確認した．

本研究の貢献をまとめると以下のようになる．

- 我々は恒久性に着目した関係の分類という新たなタスクを提案した．既存研究では，Weikum ら [11] も述べているように，関係知識が時間的に不変であることを暗黙に仮定しているため，テキストから獲得した関係を整合性をとって管理することが困難になっている．我々の提案は，このような問題の解決に寄与するものである．
- 我々は，関係獲得における時系列テキスト情報の新しい活用方法を提案して，その効果を検証する実験を行った．実験の結果，時系列テキストから得られた統計情報は，恒久性の分類のみならず，一意性の分類においても有用であることが確認された．

本論文の構成は以下のようになる．**2.** では関係の恒久性と一意性について述べ，本研究で取り組む関係分類タスクの設定について述べる．**3.** 及び **4.** では，恒久性と一意性の分類タスクのために，時系列テキストから獲得する素性について述べる．**5.** では，評価実験について述べる．**6.** では，関連研究について述べ，最後に **7.** で本研究のまとめと今後の課題について述べる．

2. 恒久性と一意性に基づく関係の分類

2.1 恒久性と一意性

まずはじめに，関係の恒久性及び一意性という二つの性質について議論をする．以下，本論文では ⟨**arg1** の大統領は **arg2**⟩ のような形式の関係知識を議論する．**arg1** と **arg2** は具体的な固有表現が入るスロットである．

表 1 恒久性がある，ない関係の例
Table 1 Examples of constant and non-constant relations.

恒久性あり	恒久性なし
arg1 の出身地である **arg2**	**arg1** の首相の **arg2**
arg1 の父親の **arg2**	**arg1** が所属する **arg2**
arg1 の著者の **arg2**	**arg1** が住む **arg2**

表 2 一意性がある，ない関係の例
Table 2 Examples of unique and non-unique relations.

一意性あり	一意性なし
arg1 の出身地である **arg2**	**arg1** の創設者の **arg2**
arg1 の本社がある **arg2**	**arg1** の要素である **arg2**
arg1 の首相の **arg2**	**arg1** と国境を接する **arg2**

恒久性がある関係とは，**arg1** にある語が代入されたときに，**arg2** に入る語が時間的に変化しない関係と定義する．例えば，⟨**arg1** の出身地は **arg2**⟩ は，個人の出身地は決して変わらないので恒久性がある関係となる．一方，⟨**arg1** の首相である **arg2**⟩ は恒久性がない関係である．なぜならば，例えばアメリカの大統領は 2012 年 5 月の時点ではバラク・オバマであるが，以前はジョージ・ブッシュやビル・クリントンであったように，時間によって変化し得るからである．表 1 に恒久性がある関係，及び恒久性がない関係の例をそれぞれ示す．

一意性がある関係とは，**arg1** にある語が代入されたときに，**arg2** に代入できる語が任意の時点において唯一に定まる関係と定義する．例えば，⟨**arg1** の出身地の **arg2**⟩ は，出身地は常に一意に決まるため一意性のある関係である．また，⟨**arg1** の本社がある **arg2**⟩ も一意性がある関係であるが，本社の住所は時間によって変化し得るため恒久性はないという点で前に述べた関係と異なる．一方，⟨**arg1** の創設者の **arg2**⟩ は，ある会社の創設者は二人以上存在し得るため，一意性がない関係である．表 2 に一意性がある関係，及び一意性がない関係の例をそれぞれ示す．

2.2 考　　察

恒久性と一意性の判定を行う際，我々は多少の例外を許容するものとする．例えば ⟨**arg1** の大統領の **arg2**⟩ という関係を考える．我々はこれを恒久性がな

(注1)：関係の時間的変化は考慮していないが，Ritter ら [9] や Lin ら [10] は functional relation と呼んでいる．

く，一意性のある関係と捉えるが，以下のような例外的な場合を想定することもできる．例えば，紛争状態にある国では，同時に複数の大統領が存在することがある．また，独裁国家においては，大統領が変わらないことも考えられる．しかしながら，これらはいずれも特殊な場合であると考えられるため，本研究では⟨**arg1** の大統領の **arg2**⟩は恒久性のある関係や一意性のない関係としては考えない．

上記の考察から，関係の恒久性と一意性は客観的に決定することが難しいことが分かる．しかしながら，関係が恒久性及び一意性という性質をもっているという考え方は直感的に受け入れられるものであり，それらの性質は複数の人間がある程度の一貫性をもって判定可能であると考えている．評価実験においても，被験者の間においてある程度の一致を確認することができた（**5.** を参照）．

2.3 問題設定とアプローチ

本研究では，与えられた各関係を恒久性と一意性に基づいて分類することを問題として設定した．恒久性と一意性は，二つの独立した分類問題として定式化し，教師あり学習を用いてこれらの分類問題を解く．

正確な分類を行うためには，どのような素性を学習に用いるかが問題となる．**3.** では恒久性の分類に用いる素性，**4.** では一意性の分類に用いる素性について述べる．いずれの素性も，時系列頻度情報に基づく素性と，言語情報に基づく素性に分けることができる．

3. 恒久性の分類に用いる素性

3.1 時系列頻度情報

恒久性の分類に用いる素性として，異なる期間において **arg2** に出現する語の頻度分布変化を用いることが考えられる．

[時系列テキスト]

時系列テキストとしては，2006 年 2 月から 2011 年 9 月までの期間で蓄積した日本語のブログアーカイブを用いた．このテキストデータは，全部で約 23 億文から成り立っている．各ブログ記事には収集された時間情報が付加されており，月別にまとめられている．そのため，以降では，1 か月を単位時間として議論をする．

[アプローチ]

恒久性がある関係（例：⟨**arg1** の出身地は **arg2**⟩）では，**arg1** にある値（例：“モーツァルト”）が与えられたとき，相異なる二つの期間において **arg2** は似たような値をとると考えられる．

一方で，恒久性のない関係（例：⟨**arg1** が所属する **arg2**⟩）では，**arg1** のある値に対して，**arg2** が取り得る値は時間によって異なると考えられる．例えば，図 1 は ⟨**arg1** の所属する **arg2**⟩ において，**arg1** の値としてプロサッカー選手である “本田圭佑” が与えられたときの **arg2** の値の時系列頻度分布を表している．本田圭祐は 2008 年から 2010 年の間に，VVV フェンロから CSKA モスクワに移籍している．そのため，2008 年と 2010 年の時点では，**arg2** に出現する語が大きく異なることが確認できる．以下では，このような時系列頻度の情報を，どのようにして恒久性分類の素性として利用するのかを具体的に述べていく．

[時系列頻度情報に基づく素性]

我々は，**arg2** に出現する語の頻度分布の類似度をコサイン類似度を用いて求める．**arg1** に出現するすべての語について，**arg2** の分布のコサイン類似度を求めて平均すると以下のようになる．

$$\frac{1}{N}\sum_{e\in\mathcal{E}_N(r)}\cos(F_{w_1}(r,e),F_{w_2}(r,e))$$

ここで，r は関係（例：⟨**arg1** の大統領の **arg2**⟩），e は **arg1** に出現する語（例：“アメリカ”），$F_w(r,e)$ は r において **arg1** の値として e が与えられたときの **arg2** の頻度分布である．w_1 と w_2 は，頻度分布の情報源となる時間窓を表している．$\mathcal{E}_N(r)$ は **arg1** における頻度上位 N 語の集合である．$\mathcal{E}_N(r)$ にあたっては，時系列テキスト全期間を用いて頻度の順序を決定する．

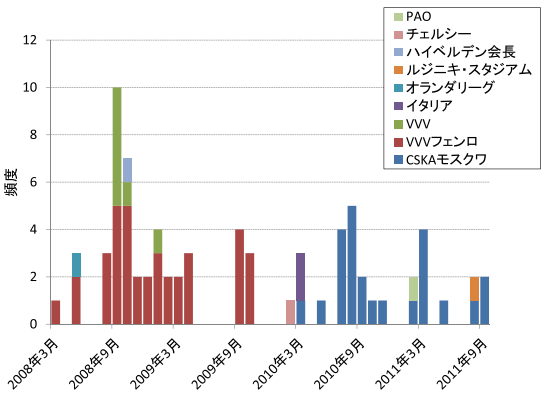


図 1 関係 ⟨**arg1** が所属する **arg2**⟩ の時系列頻度分布
Fig. 1 Time-series frequency distribution of ⟨**arg1** *belongs to* **arg2**⟩ when **arg1** takes Keisuke Honda.

ここで，w_1 や w_2 という二つの時間窓をどのように選択するかという問題が生じる．ある関係が恒久性をもたないことを把握するためには，二つの時間窓において **arg2** の頻度分布が異なることが必要である．しかしながら，事前にそのような時間窓のペアを知ることは不可能である．

そこで，提案手法では全ての時間窓のペアに対する類似度の平均値，最小値，最大値を素性として学習に用いる．

$$\frac{1}{N}\sum_{e\in\mathcal{E}_N(r)}\operatorname*{ave}_{\substack{w_1,w_2\in\mathcal{W}_T\\ w_1\neq w_2}}\cos(F_{w_1}(r,e),F_{w_2}(r,e))$$

$$\frac{1}{N}\sum_{e\in\mathcal{E}_N(r)}\max_{\substack{w_1,w_2\in\mathcal{W}_T\\ w_1\neq w_2}}\cos(F_{w_1}(r,e),F_{w_2}(r,e))$$

$$\frac{1}{N}\sum_{e\in\mathcal{E}_N(r)}\min_{\substack{w_1,w_2\in\mathcal{W}_T\\ w_1\neq w_2}}\cos(F_{w_1}(r,e),F_{w_2}(r,e))$$

ここで，$\mathcal{W}_T$ は幅が T となる全ての時間窓の集合である．実際に我々が用いるブログアーカイブは 2006 年 2 月から 2011 年 9 月までの 68 か月期間から収集されているため（詳しくは **5.1** で述べる），例えば T=3（か月）と設定した場合，$\mathcal{W}_T$ は 66 個の時間窓からなる集合となる．固有表現の数 N と時間窓の期間 T はハイパパラメータである．本研究では，N=100 とした．また，時間窓の幅 T として，1，3，6，12 か月の四期間を設定し，それぞれ前述のコサイン類似度の平均値，最小値，最大値を計算することで得られた計 12 個の素性を用いる．

3.2 言語情報

本節では，恒久性の分類に用いる言語情報に基づく素性について 2 種類述べる．

[接頭辞]

恒久性がない関係では，関係を表す語彙統語パターンの中で以下のような接頭辞が用いられることが多い．

(2) a. 米国の**前**大統領のジョージ・ブッシュ
　　b. 米国の**初**大統領のリンカーン

接頭辞 “前” や “初” は大統領が時間によって変化することを暗に示しているため，⟨**arg1** の大統領の **arg2**⟩ は恒久性がない関係であると判断する素性として使うことができる．

表 3 に挙げた時間変化を示唆する接頭辞を用いて，以下の手順で素性を設定した．まず，日本語形態素解析用辞書である NAIST-jdic(注2) から接頭辞のリストを作成する．次に，分類対象の関係から名詞を抽出し，その名詞の直前に各接頭辞が出現した頻度を計算する．この頻度の計算にあたり，我々は **3.1** で述べた日本語ブログ全記事を用いた．この場合，時間情報は重要ではないため，頻度は全てのテキストから計算したものを用いる．最後に，各接頭辞の頻度がしきい値 θ_1(注3) を超えた場合 1 を，それ以外の場合は 0 を素性の値として学習に与える．もし，関係に名詞が含まれない場合は 0 として与える．

表 3　恒久性の分類に用いる接頭辞の例
Table 3　Japanese prefixes and adjectives indicating non-constant relations.

前，新，元，現，旧，初，次など

[時制と相]

時制と相は，恒久性がない関係を認識するための手掛りとして使うことができる．ここで，以下の文章について考える．

(3) a. 米国の大統領はビル・クリントンであった．

3 a のように，テキストにおいて過去形で記述されている関係があれば，その関係は恒久性を有さない可能性が高いと考えられる．

提案手法では，時制として “た”，相として “ている”，“てる” をキーワードとして用いる．“た” は過去の時制，“ている” と “てる” は文脈のよって持続や進行形を意味する．

関係に含まれる動詞の直後に，各キーワードが出現する頻度に基づいて素性を定義する．この素性は，もし頻度がしきい値 θ_2(注4) を超えたならば 1 を，それ以外は 0 を値とする．頻度は接頭辞に基づく素性と同様にして計算し，もし関係に動詞が含まれない場合は値を 0 とする．

4. 一意性の分類に用いる素性

本章では一意性の分類に用いる素性について述べる．これらの素性も恒久性の分類と同様に，時系列頻度情報に基づくものと言語情報に基づくものに分けること

(注2)：http://sourceforge.jp/projects/naist-jdic/ (version mecab-naist-jdic-0.6.0-20090616)
(注3)：評価実験では θ_1=10 とした．ここではその言語表現が使われうるかを素性としているため，誤解析による誤検出の影響を除くことができるしきい値を，実際にデータを見て恣意的に決定した（以後のしきい値も同様）．
(注4)：評価実験では θ_2=3000 とした．

ができる．

4.1 時系列頻度情報

本節では，一意性の分類に用いる時系列頻度情報に基づく素性について 2 種類述べる．

[固有表現の種類数]

一意性の分類に用いる素性として，**arg2** に出現する語の種類数を計算することが挙げられる．一意性がある関係では，理想的にはある期間において **arg2** に出現する語の種類数が 1 種類になるはずである．書かれた時点の事実と異なる **arg2** がノイズとして観測されることもあるかもしれないが，いずれにせよ **arg2** の種類数が少ないほど一意性を有する可能性が高いことには違いない．

しかしながら，そのような単純なアプローチには関係の恒久性を考慮していないという欠点が存在する．例えば，恒久性がなく，一意性がある関係として ⟨**arg1** の本社がある **arg2**⟩ を考える．もし，頻度情報を多く得るために時間窓の幅を広く設定すると，この関係は恒久性がないために過去や未来に成り立つ語も頻度分布に現れてしまい，まるで一意性がない関係に見えてしまう．では，時間窓の期間を狭くすればよいことになる．しかしながら，時間窓の期間を狭くすると今度はデータスパースネスの問題が生じ，頻度情報が十分に得られないという問題が生じる．

このようなトレードオフは，時間窓の適切な幅を決定することが困難であるという問題を提起している．提案手法では，このような問題に対処するために **3.1** と同様に，時間窓の幅 T として四つの値を設定した素性を用いる．

$$\frac{1}{N}\sum_{e\in\mathcal{E}_N(r)} \underset{w\in\mathcal{W}_T}{\mathrm{ave}}\ \#\mathrm{type}(F_w(r,e))$$

$$\frac{1}{N}\sum_{e\in\mathcal{E}_N(r)} \max_{w\in\mathcal{W}_T}\ \#\mathrm{type}(F_w(r,e))$$

$$\frac{1}{N}\sum_{e\in\mathcal{E}_N(r)} \min_{w\in\mathcal{W}_T}\ \#\mathrm{type}(F_w(r,e))$$

なお，$\#type(\cdot)$ は **arg2** に出現する語の種類数を表している．

[頻度上位二語間の頻度割合]

提案手法では更に **arg2** に出現する頻度上位二語間の頻度割合を素性に用いる．ここで，e_{1st} と e_{2nd} を，それぞれ **arg2** に出現する最も頻度が高い語と 2 番目に頻度の高い語とする．もし，e_{1st} の頻度が e_{2nd} の頻度より十分大きい場合，一意性を有する可能性が高い．

そこで，提案手法では全ての時間窓に対する頻度割合の平均値，最小値，最大値を素性として学習に用いる．

$$\frac{1}{N}\sum_{e\in\mathcal{E}_N(r)} \underset{w\in\mathcal{W}_T}{\mathrm{ave}}\ \frac{f_w(e,r,e_{1st})}{f_w(e,r,e_{2nd})}$$

$$\frac{1}{N}\sum_{e\in\mathcal{E}_N(r)} \max_{w\in\mathcal{W}_T}\ \frac{f_w(e,r,e_{1st})}{f_w(e,r,e_{2nd})}$$

$$\frac{1}{N}\sum_{e\in\mathcal{E}_N(r)} \min_{w\in\mathcal{W}_T}\ \frac{f_w(e,r,e_{1st})}{f_w(e,r,e_{2nd})}$$

ここで，$f_w(e,r,e')$ は関係 r における，**arg1** と **arg2** がそれぞれ e と e' のときの頻度である．また，w は時間窓を表す．

4.2 言語情報

文章の並列構造や列挙を示唆する助詞・接尾辞は，一意性がない関係を認識するための手掛りとして使うことができる．以下の文章を考える．

(4) a. フランスと国境を接するイタリア**と**スペイン

b. フランスと国境を接するイタリア**など**

4 a の助詞 “と” による並立構造は，フランスと国境を接する国が複数ある（ここではスペイン）ことを意味しているため，⟨**arg1** と国境を接する **arg2**⟩ は一意性がないと判別できる．そこで提案手法では，並立構造に用いられる各助詞（表 4 を参照）が，**arg2** の直後に出現する頻度に基づいて素性を定義する．もし，頻度がしきい値 θ_3[(注5)] を超えた場合は 1 を，それ以外は 0 を素性の値とする．

また，4 b の助詞 “など” も，フランスと国境を接する国が複数あることを意味しているため，⟨**arg1** と国境を接する **arg2**⟩ は一意性がないと判別できる．そこで提案手法では，“など” に加え四つの接尾辞[(注6)]が，**arg2** の直後に出現する頻度に基づいた素性も定義する．もし，しきい値 θ_4[(注7)] を超えた場合 1 を，それ以

表 4 並立構造の特定に用いた助詞
Table 4 List of Japanese particles that are used to form coordination structures.

と，とか，や，やら，だの，なり，か

(注5)：評価実験では θ_3=10 とした．
(注6)：“ら”，“等”，“たち”，“達” である．
(注7)：評価実験では θ_4=10 とした．

表 5　恒久性・一意性の分類に用いる素性の概要
Table 5　Summary of features used in the classifications of constancy and uniqueness.

分類	時系列頻度情報	言語情報
恒久性	$\frac{1}{N}\sum_{e\in\mathcal{E}_N(r)}\{\mathrm{ave},\max,\min\}_{\substack{w_1,w_2\in\mathcal{W}_T\\ w_1\neq w_2}}\cos(F_{w_1}(r,e),F_{w_2}(r,e))$	$f_i=\begin{cases}1 & frequency(w_i^j)>\theta_j\\ 0 & otherwise\end{cases}$ $w_i^{\text{接頭}}=\{$ 前，新，元，現，旧，初，次 $etc\}$ $w_i^{\text{時制，相}}=\{$～た，～ている，～てる $\}$
一意性	$\frac{1}{N}\sum_{e\in\mathcal{E}_N(r)}\{\mathrm{ave},\max,\min\}_{w\in\mathcal{W}_T}\#\mathrm{type}(F_w(r,e))$ $\frac{1}{N}\sum_{e\in\mathcal{E}_N(r)}\{\mathrm{ave},\max,\min\}_{w\in\mathcal{W}_T}\frac{f_w(e,r,e_{1st})}{f_w(e,r,e_{2nd})}$	$f_i=\begin{cases}1 & frequency(w_i^j)>\theta_j\\ 0 & otherwise\end{cases}$ $w_i^{\text{並列}}=\{$ や，と，とか，やら $etc\}$ $w_i^{\text{接尾}}=\{$ など，ら，等，たち，達 $\}$

外は 0 を素性の値とする．

このようにして，言語情報に基づく素性として合計 12 個の素性を用いる．

3. と **4.** では，それぞれ恒久性，一意性の分類に用いる素性について述べた．ここで，これまで述べた素性を表 5 にまとめる．次章では，これらの素性を用いた評価実験について述べる．

5. 評価実験

本章では，正解データを人手によって作成し，提案手法を用いた恒久性と一意性の分類を評価する．実験では，パラメータである固有表現数 N 及び時間窓の期間 T が分類の性能に及ぼす影響を調査し，誤分類の原因について分析を行った．

5.1 実験データ

実験データの作成にあたり，**3.1** で述べた時系列テキストから関係知識を獲得し，そのうち約 1000 関係に人手で各分類タスクにおける正解ラベルを付与した．具体的な手続きについては以下に述べる．

まず，時系列テキストを係り受け解析し，二つの固有表現とそれらの間の最短係り受けパス(注8)を関係知識として獲得する．例えば，文章として「昨日，本田圭佑が所属する CSKA モスクワが快勝した.」が与えられた場合，図 2 のように関係知識が獲得される．係り受け解析には，Yoshinaga と Kitsuregawa による高速分類手法 [12] を用いた構文解析器 J.DepP(注9) を用いた．また，類義関係や上位下位関係に起因する表記の多様性を吸収するために，本研究では日本語 Wordnet [13] を用いた．具体的には，同一語義 (synset) に属する語，及び，上位下位関係にある synset に所属する語を同一のものとして扱った．

我々は，獲得された二つの固有表現の組合せ数上位 1000 関係を選択し，各関係について恒久性があるか否

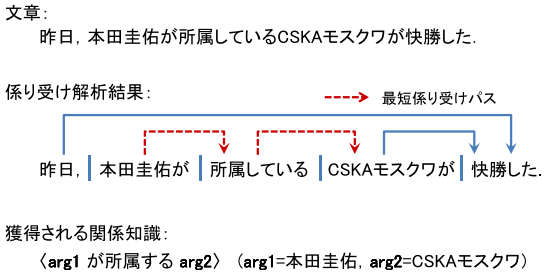


図 2　係り受け構造を用いた関係知識の獲得
Fig. 2　Example of extracting a relation using dependency path.

か，一意性があるか否かを被験者に依頼してラベル付けしてもらった．各関係当り三人の被験者を割り当て，多数決により最終的な正解ラベルを決定した．kappa 値 [14] は恒久性が 0.346，一意性が 0.428 となっており，Landis ら [15] の基準によるとそれぞれ妥当な一致と相当な一致となる．

被験者のラベル付けが一致しなかった主な要因として，各被験者が知っている **arg1** の語に知識差があることが挙げられる．例えば，〈**arg1** が **arg2** と戦う〉を考えよう．この場合，**arg1** としては，サッカーチームや野球球団，またはボクサーなどを考えると対戦相手は一意に決まるため，この関係は一意性を有するように見える．しかし，陸上競技や水泳など複数のチームや選手が同時に戦う競技もあるため一意性は有さない．このような要因による不一致は恒久性についても生じる．正確なラベル付けには網羅的な世界知識が必要であり，被験者間でそのような知識にばらつきがあると不一致が生じやすくなる．以下では，上記の他に被験者のラベル付けが一致しなかった代表的な要因を

(注8)：活用語は原型に正規化する．
(注9)：http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/jdepp/

恒久性と一意性のそれぞれについて述べる．

[arg1 の語の粒度の違い]

恒久性のラベル付けで生じた不一致の代表的な要因として，固有表現の粒度の違いが挙げられる．例えば，⟨**arg1** が **arg2** で開幕する⟩を考えよう．このとき，**arg1** の語として“オリンピック”を考えると，4 年ごとに開催地は変わるため恒久性を有さない．しかし，**arg1** の語として“ロンドン オリンピック”を考えると，開催地は“ロンドン”で変化しないため恒久性を有する．このように，恒久性を判別するとき，被験者が想定する **arg1** の語の粒度が異なると違った結果となってしまう．

[関係の意味的曖昧性]

一意性のラベル付けで生じた不一致の代表的な要因として，関係が複数の意味に解釈され得ることが挙げられる．例えば，⟨**arg1** が **arg2** で受賞する⟩を考えよう．この関係では，制作者と受賞した祭典名（例：“アカデミー賞”）との関係とも考えられるし，“英国王のスピーチ”といった受賞した作品名との関係とも考えられる．前者の場合は一意性を有するが，後者の場合は同時に複数の作品で受賞することも考えられるため一意性を有さないと判断することもできる．このように，関係を係り受けパスだけで表現すると複数の意味として捉えられる場合があり，このような場合には一意性を判別することができない．

Lin ら [10] の研究で用いた実験データでは，一意性の判別において二人の専門家間で 95.5%の一致が得られたと報告されている．これに対し，我々の実験データでは最も一致した二人の被験者間で，恒久性が 91.6%，一意性が 80.1%であり，彼らの数値と比べると低い一致率となった．しかしながらこれは，彼らの実験データでは，関係の恒久性を考慮せず，時制を含んだ関係を扱っているという点で我々のデータセットと異なっているためだと考えられる．時制が過去である関係（例：⟨**arg1** が訪れた **arg2**⟩）は，一意性がない関係になることが多く，被験者のラベル付けの一致が高くなることが予想される．本研究では時制や相は取り除いて関係を扱っており，一意性の判別がより難しいデータセットであったことから，このような kappa 値となったといえる．

5.2 実験結果

上記の実験データを用いて，恒久性，一意性のそれぞれの分類に対し五分割交差検定を行った．分類器としては，多くの自然言語処理タスクでその有効性が示されており [16], [17]，実装が容易で高速な Passive Aggressive アルゴリズム [18] を用いた．

[恒久性の分類]

図 3 に恒久性の分類における再現率・精度曲線を示す．関係の恒久性の分類に関する既存研究は存在しないため，ベースライン手法には以下の式で表せられるコサイン類似度を用いた簡易な手法を用いた．

$$\frac{1}{N}\sum_{e\in\mathcal{E}_N(r)}\cos(F_{w_1}(r,e),F_{w_2}(r,e))$$

なお，時間窓 w_1 は固有表現 e が最初に観測される月，w_2 は固有表現 e が最後に観測される月で決定される．上記の類似度があるしきい値以上である関係 r を恒久性があると分類する．実験結果の再現率・精度曲線は，このしきい値を変化させることで描かれている．

図 3 から，提案手法により再現率・精度共にベースライン手法に比べて大きく上回る分類結果が得られた．ベースライン手法の再現率・精度が極端に低いのは，最初と最後の月の頻度分布のみに注目するだけでは，**5.5** で述べるような誤分類の原因の影響を大きく受けてしまうからである．

恒久性の分類では，言語情報に基づく素性は時系列頻度情報に基づく素性よりも有効であった．また提案手法では，特に再現率 0.69 以上において言語情報に基づく素性のみを用いた場合より精度が高くなっており，言語情報に基づく素性と時系列頻度情報に基づく素性が相補的な関係となっていることが分かる．

[一意性の分類]

図 4 に一意性の分類における再現率・精度曲線を示す．ベースライン手法としては Lin ら [10] の手法

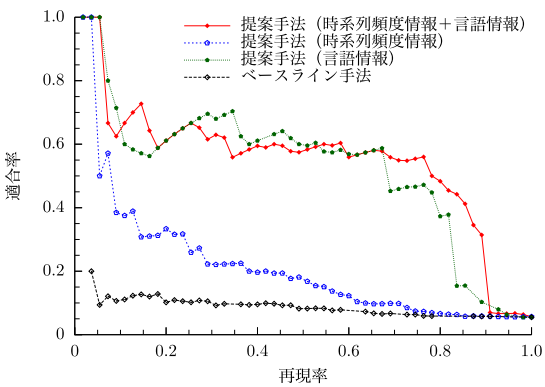


図 3　再現率・精度曲線（恒久性の分類）
Fig. 3　Recall-precision curve. (constancy classification)

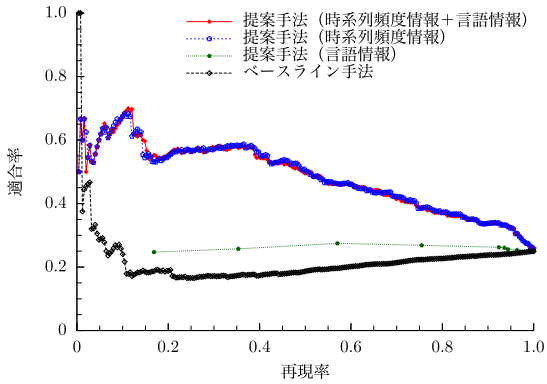


図 4　再現率・精度曲線（一意性の分類）
Fig. 4　Recall-precision curve. (uniqueness classification)

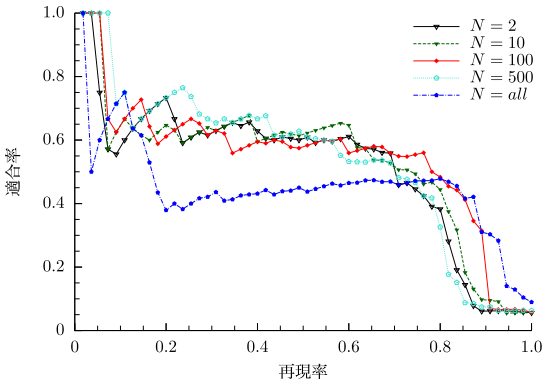


図 5　N の値による分類精度（恒久性の分類）
Fig. 5　Comparison with the methods varying a value of N for constancy classification.

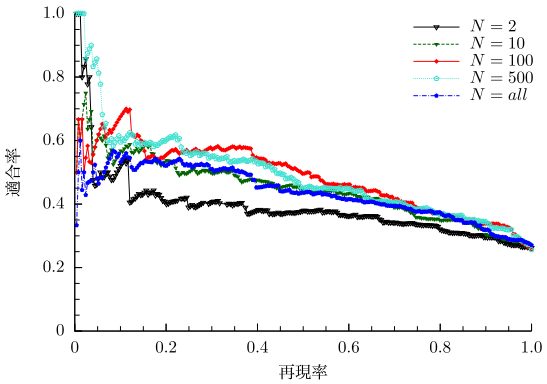


図 6　N の値による分類精度（一意性の分類）
Fig. 6　Comparison with the methods varying a value of N for uniqueness classification.

（KLFUNC，KLDIFF 及びそれらの平均値）を実装し，特に彼らの研究において最も性能が良かった KLFUNC と KLDIFF の平均値を用いた手法をベースライン手法とした．

図 4 から，提案手法により再現率・精度共にベースライン手法に比べて大きく上回る分類結果が得られた．Lin らの手法は我々の提案手法に類似した手がかりを用いているが，時系列情報を考慮していないという点で異なる．よって，時系列情報を用いることが恒久性のみならず一意性の分類において有用であることが示されたといえる．

一意性の分類では時系列頻度情報に基づく素性が分類において中心的な役割を果たし，言語情報に基づく素性はあまり有効に働いていないことも分かった．この要因として考えられるのは，並立助詞が使われているからといって必ずしもその関係が一意性を有さないことにはならないということが挙げられる．例えば「イチローが所属するマリナーズとヤンキースが対戦した.」といった文章における並立助詞 “と” は，イチローがマリナーズとヤンキースの 2 球団に同時に所属していることを表しているわけではない．これらを正確に識別するにはより高度な解析が必要である．

5.3　固有表現数 N の分類精度への影響

3.1 で述べたように，提案手法では **arg1** の固有表現数 N の値を 100 に設定した．この設定の有用性を確かめるために，固有表現数 N の値をそれぞれ 2，10，20，100，500 及び全ての固有表現(注10)に設定したときの分類精度を評価した．

図 5 から，恒久性の分類において，N=2〜500 では値が大きいほど分類の性能が良くなっているように見えるが，差は無視できる程度にとどまっている．また，N を全固有表現にした場合は分類精度が大きく低下した．これは，下位の固有表現は出現頻度が少ないためノイズの影響が無視できなくなるためと考えられる．

一方で，一意性の分類においては，N の値の違いにより大きな分類精度の変化が見られた（図 6 を参照）．一意性の判別においては，より良い分類結果を得るために N の値を調整する必要があるといえる．

5.4　時間窓の期間 T の分類精度への影響

提案手法では，**3.1** や **4.1** で述べたように，素性として時間窓の期間を複数設定したものを用いた．この手法の有用性を確かめるために，単一の T に基づく手法と提案手法の分類精度を比較した（図 7，図 8）．

(注10)：実験データでは最大で 30655 語

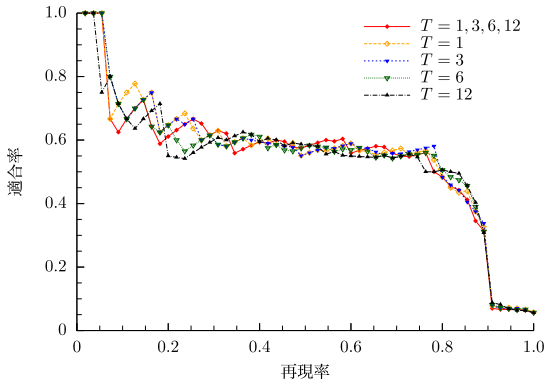


図 7　T の値による分類性能（恒久性の分類）
Fig. 7　Comparison with the methods using only a single value of T for constancy classification.

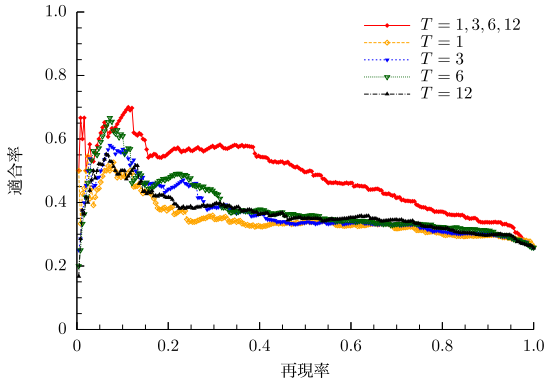


図 8　T の値による分類性能（一意性の分類）
Fig. 8　Comparison with the methods using only a single value of T for uniqueness classification.

図 8 から，一意性の分類において，提案手法は単一の T に基づく手法よりも分類精度が向上することが確認できる．特に，再現率 0.345 において提案手法による精度の向上が顕著である．このとき，それぞれ $T = 1$，$T = 12$ として生成した素性のみを用いて正しく分類できた関係 S_1，S_{12} を比較したところ，一致した割合 $|S_1 \cap S_{12}|/|S_1|(= |S_{12}|)$ は 0.558 にとどまった．例えば，⟨**arg1** の新曲の **arg2**⟩ という一意性がある関係では，一般的に新曲は 1 年に数曲でることが多いため，$T = 1$ では正しく分類されているものの，$T = 12$ ではそれらを同時に扱ってしまうため正しく分類されなかった．

このことから，$T = 1$，$T = 12$ において生成される素性が，それぞれ異なる関係を正しく分類するのに有効となっているといえる．提案手法ではこれらの素性を含め，様々な T で生成された素性を同時に用いているため，相補的に精度が向上したと考えられる．

表 6　誤分類の原因
Table 6　Major factors of misclassifications.

原因	恒久性の誤分類数	一意性の誤分類数	合計
類義語・上位下位概念語の不特定	14	48	62
話題性による記述の偏り	2	44	46
短期的な時間変化	14	39	53
過去の事実の記述	9	11	20
誤情報の記述	3	10	13
長期的な時間変化	6	0	6

以上から，時間窓の期間を複数設定した素性を用いることが有効であることが示された．

一方で，恒久性の分類においては，時間窓の期間を複数設定する提案手法の有用性は見られなかった（図 7 を参照）．

5.5　誤分類の原因分析

誤分類した関係のうち両タスクから合計 200 関係を無作為に選び，その原因について調査した結果，表 6 のようになった．

[類義語・上位下位概念語の不特定]

本研究では類義語・上位下位概念語の特定のため，日本語 Wordnet を用いた．しかしながら，それでも認識できない異表記が存在し，誤分類の原因となっていた．例えば，**arg2** がとる固有表現に “グーグル” や “Google” といった翻字に起因する表記揺れが存在するために，時間的に変化しているように見えてしまい，恒久性がない関係に誤分類してしまう場合がある．また，**arg1** の固有表現として “松井” があったときに，“松井秀喜”，“松井稼頭央”，“松井大輔” などの複数の対象を指し得る略称が用いられると，**arg2** の固有表現としては，その略称が指す全ての固有表現 **arg1** に対応した固有表現 **arg2** をとることとなり，一意性がある関係を一意性がない関係と誤分類してしまう．こうした異表記を扱うには，翻字認識や名寄せに関する研究成果を取り込んでいく必要があると考えられる．

[話題性による記述の偏り]

ブログ記事において記述される話題には偏りがあり，誤分類の大きな原因となる．例えば，ハリウッド俳優であるウィル・スミスの息子のジョイデン・スミスは，時系列テキストにおいて頻繁に記述されている．これは，ジョイデン・スミスが父と映画で共演したからであり，他の子供についてはブログにはほとんど記述されない．このため，例えば ⟨**arg1** の息子の **arg2**⟩ といった一意性がない関係を，一意性があると誤分類し

てしまう．

[短期的な時間変化]

我々が用いた時系列テキストは 1 か月を単位時間としている．そのため，それよりも短い間隔で変化する関係については，時系列頻度情報に基づく素性が分類精度向上に寄与しにくくなる．例えば，⟨**arg1** が **arg2** を下す⟩ の **arg1** に海外のサッカーチームである “レアル・マドリード” を代入したときを考えよう．サッカーの試合は 1 か月に複数回行われるので，1 か月を単位時間とすると複数の相手チーム（例えば，“バルセロナ” や “バレンシア”）と同時に試合をしたように見えてしまう．このため，一意性がある関係を一意性がないと誤分類してしまう．しかしながら，このような誤分類が起こる関係の多くは，⟨**arg1** が **arg2** に勝利する⟩ や ⟨**arg1** が **arg2** を訪れる⟩ といったように瞬間的な動作を表す動詞を伴うものがほとんどであり，動詞そのものの性質を別に特定することによって判断可能だと考えている．

また，表 6 の原因にある「長期的な時間変化」では，ブログ記事の収集期間内では変化しなかった恒久性がない関係を恒久性があると誤分類してしまう．今回頻度情報の抽出に用いた時系列テキストは 6 年分なので，⟨**arg1** のライバルの **arg2**⟩ のようなそれより長い期間で変化するような関係は恒久性を有すると判断されていた．これらに対処するためには，**6.** で述べるような文中に記述されている時間表現（“1980 年” や “平成 18 年” など）を用いる必要があると考えている．

[過去の事実の記述]

本研究では，各ブログ記事が収集されたときのタイムスタンプを時間情報として用いている．そのため，ある記事に過去の出来事について記述されていた場合，提案手法では記事が収集されたときの事実として扱われてしまうという問題がある．しかしながら，ブログ記事には比較的最近の出来事について書かれることが多く，また分類の素性においても時系列情報だけでなく，そのような影響を受けにくい言語情報も用いているため，誤分類の大きな原因とはならなかった．

[誤情報の記述]

ウェブテキストには，誤った情報や憶測に基づく記述があり，それらの記述がノイズとなって誤分類の原因となってしまう．例えば，⟨**arg1** が **arg2** に買収される⟩ の **arg1** に “Yahoo!” を代入すると，**arg2** として “Microsoft” や “Google” といった固有表現が獲得された．これは，Microsoft や Google が Yahoo!の買収に乗り出しているのではないかという憶測によるものである．事実ではないこのような記述のために，まるで複数の企業が同時に買収したように見えてしまい，一意性がある関係を一意性がないと誤分類してしまう．

6. 関連研究

大規模テキスト，特にウェブテキスト（**1.** を参照）から関係知識を獲得する研究は，近年盛んに研究されている．しかしながら，これまでの研究の多くは単にテキストから関係を獲得することに着目していた．そのため，整理・管理を目的として関係知識を分類する研究は少ない．

関係の恒久性を分類する研究はこれまで行われていない．関係の恒久性を考慮した関係獲得研究としては，Temporal Information Extracton タスク [19], [20] が挙げられるであろう．このタスクでは，テキスト（例：“オセロはシェイクスピアによって 1602 年に書かれた．”）からイベントとそれが起きている期間の情報を獲得する．テキストから直接獲得される時間情報は関係の恒久性を知る上で確かに有益であるが，それだけで関係の恒久性を判定することは難しい．

一方で，関係の一意性に関しては研究は幾つかの研究が行われている．Ritter ら [9] は，関係の一意性が様々な自然言語処理タスク（矛盾検出や数量詞の範囲特定，類義語の特定など）を解く上で有用であることを指摘した．彼らは EM アルゴリズムに基づく手法により，関係の一意性をスコア付けする手法を提案した．Lin らは，一意性を判別するために，三つのアルゴリズムを提案した．これらの研究では，本研究と同じ一意性の分類問題について扱っているが，一意性の判別の際に恒久性の観点を考慮していない（**4. 1** を参照）という点で提案手法とは異なる．

7. むすび

本研究では，実世界テキストから獲得した固有表現間の関係知識を集積する際に，知識の整合性の問題が生じることを指摘した．更に，その問題を解くために関係知識の恒久性・一意性の概念が有用であることを議論し，獲得された関係知識を関係の恒久性と一意性という観点から分類する手法を提案した．具体的には，時系列テキストを用いることで，各分類タスクに対して考案した言語情報に基づく素性と，時系列頻度情報に基づく素性を導出し，これを線形分類器の素性として用いた．評価実験の結果から，時系列頻度情報に基

づく素性が両関係分類タスクにおいて分類精度の向上に大きく寄与することを確認した．

今後の課題としては，得られた結果を用いて整合性のとれた大規模関係知識を集積することが挙げられる．

文　献

[1] P. Pantel and M. Pennacchiotti, “Espresso: Leveraging generic patterns for automatically harvesting semantic relations,” Proc. ACL, pp.113–120, 2006.

[2] M. Banko, M.J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, “Open information extraction from the web,” Proc. IJCAI, pp.2670–2676, 2007.

[3] F.M. Suchanek, G. Kasneci, and G. Weikum, “YAGO: A core of semantic knowledge unifying WordNet and Wikipedia,” Proc. WWW, pp.697–706, 2007.

[4] F. Wu, R. Hoffmann, and D.S. Weld, “Information extraction from Wikipedia: Moving down the long tail,” Proc. KDD, pp.731–739, 2008.

[5] J. Zhu, Z. Nie, X. Liu, B. Zhang, and J.-R. Wen, “StatSnowball: a statistical approach to extracting entity relationships,” Proc. WWW, pp.101–110, 2009.

[6] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, “Distant supervision for relation extraction without labeled data,” Proc. ACL-IJCNLP, pp.1003–1011, 2009.

[7] F. Wu and D.S. Weld, “Open information extraction using Wikipedia,” Proc. ACL, pp.118–127, 2010.

[8] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A.A. Kalyanpur, A. Lally, J.W. Murdock, E. Nyberg, J. Prager, N. Schlaefer, and C. Welty, “Building watson: An overview of the deepqa project,” AI Magazine, vol.31, no.3, pp.59–79, 2010.

[9] A. Ritter, D. Downey, S. Soderland, and O. Etzioni, “It’s a contradiction—no, it’s not: A case study using functional relations,” Proc. EMNLP, pp.11–20, 2008.

[10] T. Lin, Mausam, and O. Etzioni, “Identifying functional relation in web text,” Proc. EMNLP, pp.1266–1276, 2010.

[11] G. Weikum, S. Bedathur, and R. Schenkel, “Temporal knowledge for timely intelligence,” Proc. BIRTE, pp.1–6, 2011.

[12] N. Yoshinaga and M. Kitsuregawa, “Polynomial to linear: Efficient classification with conjunctive features,” Proc. EMNLP, pp.1542–1551, 2009.

[13] H. Isahara, F. Bond, K. Uchimoto, M. Utiyama, and K. Kanzaki, “Development of japanese wordnet,” Proc. LREC, pp.2420–2423, 2008.

[14] J.L. Fleiss, “Measuring nominal scale agreement among many raters,” Psychological Bulletin, vol.76, no.5, pp.378–382, 1971.

[15] R.J. Landis and G.G. Koch, “The measurement of observer agreement for categorical data,” Biometrics, vol.1, no.33, pp.159–174, 1977.

[16] C. Koby, D. Mark, and K. Alex, “Multi-class confidence weighted algorithms,” Proc. EMNLP, pp.496–504, 2009.

[17] Y. Naoki and K. Masaru, “Kernel slicing: Scalable online training with conjunctive features,” Proc. COLING, pp.1245–1253, 2010.

[18] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shawartz, and Y. Singer, “Online passive-aggressive algorithms,” J. Machine Learning Research, vol.7, pp.551–583, 2006.

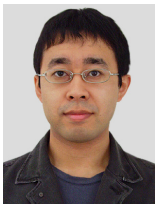
[19] X. Ling and D.S. Weld, “Temporal information extraction,” Proc. AAAI, pp.1385–1390, 2010.

[20] Y. Wang, M. Zhu, L. Qu, M. Spaniol, and G. Weikum, “Timely YAGO: Harvesting, querying, and visualizing temporal knowledge from Wikipedia,” Proc. EDBT, pp.697–700, 2010.

（平成 24 年 6 月 5 日受付，10 月 3 日再受付）

高久　陽平

2010 東工大・工・情報工学卒．2012 東京大学大学院情報理工学系研究科修士課程了．



吉永　直樹

2000 東大・理・情報科学卒．2002 同大大学院理学系研究科修士課程了．2005 同大大学院情報理工学系研究科博士課程了．博士（情報理工学）．2002 より 2008 まで日本学術振興会特別研究員 (DC1, PD)．2008 東大・生産技術研究所特任研究員，特任助教を経て 現在，同大学生産技術研究所特任准教授．計算言語学・機械学習の研究に従事．

鍛治　伸裕

2005 東京大学大学院情報理工学系研究科博士後期課程了．情報理工学博士．2007 東京大学生産技術研究所特任助教を経て現在，同大学生産技術研究所特任准教授．自然言語処理の研究に従事．

豊田　正史

1994 東工大・理・情報科学卒．1996 同大大学院情報理工学研究科修士課程了．1999 同大学院情報理工学研究科博士後期課程了．博士（理学）．同年，科学技術振興事業団計算科学技術研究員．2001 東大・生産技術研究所学術研究支援員，同大学同研究所産学官連携研究員，同大学生産技術研究所特任助教授，助教授を経て現在，同大学生産技術研究所准教授．ウェブマイニング，ユーザインタフェース，ビジュアルプログラミングに興味をもつ．ACM，IEEE CS，情報処理学会，日本ソフトウェア科学会各会員．

喜連川　優　（正員：フェロー）

1978 東大・工・電子卒．1983 同大大学院工学系研究科情報工学専攻博士課程了．工博．同年同大学生産技術研究所講師．現在，同教授．2003 同所戦略情報融合国際研究センター長．データベース工学，並列処理，Web マイニングに関する研究に従事．2009 ACM SIGMOD Edgar F. Codd Innovations Award 受賞．現在，本会副会長，日本データベース学会理事，情報処理学会フェロー，SNIA-Japan 顧問，本会データ工学研究専門委員会委員長 (1997～1998)，ACM SIGMOD Japan Chapter Chair (1999～2002) 歴任．VLDB Trustee (1997～2002)，IEEE ICDE，PAKDD，WAIM 等ステアリング委員，IEEE ICDE Program Co-chair (1999 年)，General Co-chair (2005)．