表層類似度に基づく日本語テキスト含意認識

Surface-Similarity Based Textual Entailment Recognition for Japanese Text

服部 昇平
Shohei Hattori

名古屋大学大学院 工学研究科
Graduate School of Engineering, Nagoya University
syohei_h@nuee.nagoya-u.ac.jp

佐藤 理史
Satoshi Sato

(同上)
ssato@nuee.nagoya-u.ac.jp

駒谷 和範
Kazunori Komatani

大阪大学 産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University
komatani@sanken.osaka-u.ac.jp

keywords: RITE-2, recognizing textual entailment, surface similarity, overlap ratio

Summary

This paper proposes a surface-similarity based method for recognizing textual entailment (RTE) in Japanese. First, we experimentally show that there is a positive correlation between semantic similarity (textual entailment) and surface similarity between sentences. The most effective measure of surface similarity for RTE is the character overlap ratio, which achieves classification accuracy of 78.3%. Based on the result, we design a two-step RTE system for binary classification. The first step classifies a given text pair into positive or negative entailment based on the character overlap ratio. If the pair is classified into the positive class, the second step examines whether the assigned class should be flipped or not by using heuristic rules that detect the mismatch of named entities and numbers. In addition to the RTE system, we also implement the MC system that classifies a given text pair into one of four classes (forward entailment, bidirectional entailment, contradiction, and the others), by combining a contradiction detector and the RTE system.

In the RITE-2 formal run, the RTE system was ranked 7th among 42 systems at the RTE task, and the MC system was ranked first among 21 systems at the MC task. These results show that the surface-similarity based method achieves high performance in RTE.

1. はじめに

テキスト間の含意関係とは、次のような関係である [Bentivogli 11, Watanabe 13] .

人間があるテキスト t_1 を読んで、別のテキスト t_2 が正しいと推論可能であるとき、 t_1 は t_2 を含意するとみなす。

この説明に基づけば、例えば、以下の例のテキスト t_1 は、 t_2 を含意するとみなすのが妥当である。

t_1 : 川端康成は「雪国」などの作品でノーベル文学賞を受賞した t_2 : 川端康成は「雪国」の著者である
--

テキスト間に成立する含意関係の自動認識は、文書要約や質問応答といった自然言語処理アプリケーションを高度化する技術として期待されている。文書要約のタスクでは、含意認識によって、要約文章中で同一内容を含む文を検出し、削除することで、冗長性を排除できる [Tatar 08]。質問応答のタスクでは、テキスト間の含意関係を回

答生成に用いることにより、システム性能を向上させることができる [Harabagiu 06]。

テキスト含意認識に関する研究は、英語テキストに対しては、RTE (Recognizing Textual Entailment) [Dagan 05, Bar-Haim 06, Giampiccolo 07, Giampiccolo 08, Bentivogli 09, Bentivogli 10b, Bentivogli 11] という評価型ワークショップを中心に、活発に研究されている。日本語や中国語のテキストに対しては、2012 年に NTCIR-9 の RTE [Shima 11a] という評価型ワークショップが初めて実施された。この成功を受けて、2013 年には、NTCIR-10^{*1}において RITE-2 [Watanabe 13] が開催された。

テキスト含意認識には、言い換えを認識するための語彙知識や、論理的推論など様々な技術が必要とされている [永田 08]。確かに、意味関係を人間と同じように認識するためには、例えば、2 つのテキストが同じ事態を表すかどうかを判定することが不可欠であり、そのためには、述語項構造解析に基づく言い換え認識 [小谷 09] などが必要であろう。

*1 <http://research.nii.ac.jp/ntcir/ntcir-10/>

しかしながら、現時点では、それらの技術の成熟度はまだまだ低い。例えば、同じ意味内容を表す言い換えは、単語レベルだけでなく、より大きな句、節、文のレベルでも起こりうる。この問題の根本的な解決には、あらゆるレベルの言い換え表現を網羅するデータベースの構築が必要になるが、そのためには、膨大な時間と労力が必要と考えられる。

それでは、そのような意味解析技術が実現できなければ、テキスト間の含意関係は認識できないのであろうか。この疑問に答えるために、本研究では、テキスト間の表層類似度を用いた含意認識手法について検討する。このような方法を検討するのは、テキスト間の表層類似度は、テキスト間の意味内容の近さ(意味的類似度)と正の相関を持つと考えられるからである。もし、このような手法で、含意関係を70%以上の精度で正しく認識することができれば、残りの30%を意味解析技術によって解けばよいことになる。

日本語テキストの含意認識に対して、テキスト間の表層類似度が有効に働くことが、すでに文献[Pham 11]において指摘されている。しかし、どのような表層類似度が最も効果的であるか、そして、表層類似度のみで、どの程度の認識性能が得られるかは、明らかになっていない。

このような背景に基づき、本研究では、まず、どのような表層類似度が含意認識に最も効果的であるか、そして、表層類似度のみで、どの程度、含意認識が可能であるかを調査する。次に、その結果に基づき、含意認識システムを作成する。そして、作成したシステムをRITE-2タスクに適用し、テキスト間の表層類似度に基づく含意認識手法の有効性を示す。

本論文の構成は、次の通りである。2章では、関連研究について述べる。3章では、RITE-2について概説する。4章では、本研究の方針を説明する。5章では、含意認識に有効に働く、テキスト間の表層類似度を調査する。6章では、作成した含意認識システムについて説明し、7章では、矛盾の検出を加えたMCシステムについて説明する。8章では、システムの性能について述べ、9章では、言語現象毎の認識性能を考察する。最後に、10章でまとめる。

2. 関連研究

2.1 英語テキストの含意認識

英語テキストを対象とした含意認識の研究は、主にRTEワークショップを通じて活発に行われてきた。最新のRTEワークショップは、2011年に開催されたRTE-7[Bentivogli 11]である。

RTE-7において、最も良い成績をおさめたのは、含意認識スコアとフィルタリングを組み合わせた手法[Tsuchida 11]である。含意認識スコアは、各単語に重み付けを導入した、単語オーバーラップ率により計算される。この過

表1 RITE-2 スケジュール

日程	イベント
2012.7.5	開発用データリリース
2013.1.9-16	フォーマルラン
2013.6.18-21	NTCIR10 ワークショップ

程で、WordNet^{*2}などに基づく類義語や上位語が考慮される。フィルタリングで用いる2値分類器は、単語オーバーラップ率や述語項構造などを特徴とする機械学習によって作成される。この手法は、単語のマッチングに基づく手法の一つとして位置づけられるが、この他にも、同様の手法を採用したシステム[Yokote 11, Tan 11]が、RTE-7で上位の成績をおさめた。

2.2 日本語テキストの含意認識

2012年に開催されたRITE[Shima 11a]において、最も良い成績をおさめたのは、機械学習を用いたPhamらの手法[Pham 11]である。この手法は、日本語テキストだけでなく、機械翻訳によって得られた英語テキストも利用している。機械学習に用いる特徴には、テキスト間の編集距離といった表層的特徴や、単語のオーバーラップ率、係り受け構造のオーバーラップ率などが用いられている。用いた特徴の中では、テキスト間の表層類似度を表すものが最も有効であったと報告している。

RITEでは、この他に、述語項構造のマッチングに基づく手法[Shibata 11, Akiba 11, Shima 11b]や語彙機能文法(Lexical Functional Grammar)に基づく手法[Uemoto 11]が用いられた。

RITEに先立つ研究として、表層的特徴のみを用いる宇高らの研究[宇高 10]がある。この研究では、形態素、自立語、文節の n -gramの一致率を用い、そのしきい値を変化させることで、含意認識精度がどのように変化するか調査している。しかし、含意認識システムの実装には至っていない。さらに、調査に使用したデータ数も含意ペア、非含意ペアそれぞれ52個と少ない。

3. RITE-2

RITE-2[Watanabe 13]は、テキスト間の含意、換言、矛盾等の検出の技術開発を目的とした評価型ワークショップであり、表1に示すスケジュールで行われた。

評価型ワークショップでは、通常、まず、正解が付与された開発用データが配布され、参加者は、そのデータを用いてシステムを開発する。その後、システム評価(フォーマルラン)のための、正解が隠されたデータが配布される。参加者は、システムを実行してそのデータに対する答えを求め、ワークショップ主催者に提出する。最後に、提出した結果に基づくシステムの性能が、主催者によって公表される。

*2 <http://wordnet.princeton.edu/>

```

<pair id="1" label="Y">
  <t1>プロメテウスは人類に火を渡し、張り付けにされた。</t1>
  <t2>プロメテウスは人類に火を齎して罰を受けた。</t2>
</pair>
<pair id="2" label="Y">
  <t1>伊坂幸太郎は直木賞候補になった2003年の『重力ピエロ』で
  一般読者に広く認知されるようになった。</t1>
  <t2>『重力ピエロ』は伊坂幸太郎による小説で直木賞候補作品
  だった。</t2>
</pair>

```

図1 開発用データの一部

RITE-2 は、日本語テキストを対象とした5つのサブタスク、および、中国語(簡体字および繁体字)テキストを対象とした2つのサブタスクの計7つのサブタスクで構成されている。それぞれのサブタスクに対し、図1に示すような、開発用データが配布された。開発用データは、多数のテキストペアから構成されている。そのそれぞれは、テキストペアを構成する2つのテキスト t_1, t_2 と、その id, および、正解ラベルから構成されている。フォーマルラン用のデータもこれと同じ形式であるが、正解ラベルは含まれない。

7つのサブタスクのうち、本論文で主に扱うのは、日本語テキストを対象とした以下の2つのサブタスクである。

Binary Class サブタスク(以下、BC サブタスクと略記)は、与えられたテキストペア $\langle t_1, t_2 \rangle$ に対して、 t_1 が t_2 を含意するかどうか判定する2値分類問題である。分類ラベルは、以下の2種類である。

Y t_1 が t_2 を含意する

N t_1 は t_2 を含意しない

これらは、いずれも方向性を持った関係である。

一方、Multi Class サブタスク(以下、MC サブタスクと略記)は、含意の方向、矛盾の検出も含めた、次の4種類のラベルを出力する4値分類問題である。

F t_1 は t_2 を含意するが、 t_2 は t_1 を含意しない

B t_1 が t_2 を含意し、かつ、 t_2 が t_1 を含意する(換意)

C t_1 と t_2 が同時に起こり得ない(矛盾)

I 上記以外(無関係)

これらのうち、F のみが方向性を持った関係であり、他は双方向の関係である。

4. 方針

ここでは、概念的なレベルで、2つのテキスト間の意味的類似性と表層類似性の関係を考える。

一般に、2つのテキスト間の意味的類似性と表層類似性には、正の相関があると考えられる。含意認識にテキスト間の表層類似度が有効に働く [Pham 11] という事実は、このことの一つの傍証である。

ここで、意味的類似性と表層類似性を、それぞれ、高・低に分割すると、図2に示すような 2×2 の表を描くこと

		表層類似性	
		低い	高い
意味的 類似性	高い	Case2	Case1
	低い	Case3	Case4

図2 テキストの意味的類似性と表層類似性

ができる。2つの類似性には正の相関があるので、多くのテキストペアは、Case1 と Case3 に分類されることになる。

残りの2つの箱のうち、Case2 には、意味的類似性が高いが表層類似性は低いテキストペアが分類される。例えば、テキストペアが言い換えを含み、かつ、その言い換えの表層上の類似性が低い場合、そのようなテキストペアは、Case2 に分類されることになる。

逆に、Case4 には、表層類似性が高いが、意味的類似性が低いテキストペアが分類される。2つのテキスト間に見られる表層上の小さな差異が、意味的に大きな差異をもたらす場合、そのようなテキストペアは、Case4 に分類されることになる。

表層類似性の計算法を定めれば、Case1/4 と Case2/3 を分離することができる。テキストペアの表層類似性の計算法は、これまでに各種の方法が提案されているので、その中から最適な方法と最適なしきい値を定めれば良い。このような考え方に基づき、本研究では、まず、Case1/4 と Case2/3 の効果的な分類方法を調査する。

残された問題は、Case2 と Case4 の検出である。Case2 の検出のほとんどは、言い換え表現の認識となる。これを実現するには、大規模な言い換えのデータベースが必要となるので、今回は、Case2 の検出を断念する。

一方、Case4 の検出は、多くの場合、表層的な小さな差異が、意味的に大きな差異を持つかどうかの判定となる。その一部は、比較的簡単なヒューリスティックで実現できると考えられる。

このように、テキストの表層類似性に基づく大分類を最初に行えば、その後に解かなければいけない問題を限定することができる。

5. 含意認識とテキスト間の表層類似度

前章で述べた方針に従い、本章では、どのような表層類似度が、含意関係の認識に有効かを検討する。

5.1 オーバーラップ率

ここでは、表層類似度を定義するための基礎となる、オーバーラップ率を計算する一般関数を定義する。

ある集合 E が与えられたとき、2つのテキスト t_1 と t_2 において、集合 E の要素がいくつ共通して現れるかを、

次式によって定義する．

$$\text{overlap}(E; t_1, t_2) = \sum_{x \in E} \min(f(x, t_1), f(x, t_2)) \quad (1)$$

ここで, $f(x, t)$ は, 集合 E の要素 x が, t 中に出現する回数を表す．なお, 集合 E には, 文字の集合, 文字 bigram の集合, 単語の集合, 単語 bigram の集合などを想定する．

この式を用いて, 以下に示す 2 種類のオーバーラップ率を定義する．

$$\text{overlap_ratio}_B(E; t_1, t_2) = \frac{2 \text{overlap}(E; t_1, t_2)}{\sum_{x \in E} f(x, t_1) + \sum_{x \in E} f(x, t_2)} \quad (2)$$

$$\text{overlap_ratio}_D(E; t_1, t_2) = \frac{\text{overlap}(E; t_1, t_2)}{\sum_{x \in E} f(x, t_2)} \quad (3)$$

前者は, 式 (1) を t_1 と t_2 の両方の長さで正規化した値であるのに対し, 後者は, t_2 の長さのみで正規化した値である．

含意は, 方向性をもった関係である．そのため, t_1 から t_2 への含意関係の認識には, 後者の式 (3) を用いる．一方, 前者の式 (2) は, 方向性を持たない関係である矛盾の検出 (7.2 節) に用いる．

これらの式の集合 E を定めると, 特定の形式 (文字, 単語) のオーバーラップ率を定義する式となる．例えば, 集合 E として文字集合 (C^1) を採用することにより, 方向性を持った文字オーバーラップ率 $\text{cor}_D(t_1, t_2)$ を以下のように定義することができる．

$$\text{cor}_D(t_1, t_2) = \text{overlap_ratio}_D(C^1; t_1, t_2) \quad (4)$$

5.2 9 種類の表層類似度の比較

集合 E として以下の 9 種類の集合を採用し, それぞれの場合の含意認識の精度を比較した．

$$C^1, C^2, C^3, W^1, W^2, W^3, L^1, L^2, L^3$$

ここで, C^n は文字 n -gram の集合を表し, W^n は形態素出現形 n -gram の集合, L^n は形態素基本形 n -gram の集合を表す．テキスト中の句読点や記号は, 削除せず, それぞれの集合の要素として扱う．なお, テキスト t を形態素の列に変換する際には, MeCab^{*3} + UniDic^{*4} を用いた．

精度比較の調査には, RITE-2 の開発用データを用いた．使用したデータを表 2 に示す．全部で 1242 個 (含意関係が成立するもの 613 個, 成立しないもの 629 個) のテキストペアを使用した．なお, 表 2 の, 順方向 (f) は, $\langle t_1, t_2 \rangle$ のペアをそのまま使用したことを表し, 逆方向 (b) は, t_1 と t_2 を入れ替えて使用したことを意味する．

表 2 調査に使用したデータ

	タスク						total
	BC		MC				
ラベル	Y	N	B	F	C	I	
個数	240	371	83	207	65	193	
方向	f	f	f	b	f	f	f
含意あり	✓		✓	✓	✓		613
含意なし		✓			✓	✓	629

f: 順方向, b: 逆方向

表 3 9 種類の表層類似度の含意認識性能

E	th.	正解数		正解率
		Y	N	
C^1	0.69	499/613	473/629	78.3
L^1	0.66	468/613	471/629	75.6
W^1	0.64	452/613	477/629	74.8
C^2	0.45	484/613	473/629	77.1
L^2	0.35	413/613	487/629	72.5
W^2	0.29	443/613	447/629	71.7
C^3	0.36	410/613	490/629	72.5
L^3	0.19	365/613	472/629	67.4
W^3	0.18	355/613	480/629	67.2

調査の手順は, 次のとおりである．まず, それぞれのテキストペアに対して, 9 種類のオーバーラップ率を計算する．次に, それぞれのオーバーラップ率に対して, 最適なしきい値を決定する．具体的には, しきい値を 0.01 ずつ変化させていき, しきい値以上を Y (含意あり), しきい値未満を N (含意なし) と判定したときの正解率が最も高くなる値を求める．

調査結果を表 3 に示す．この表より, 1-gram, 2-gram では, いずれの集合 E に対しても 70% を超える判定精度が得られていることがわかる．9 種類の中では, 文字 1-gram のオーバーラップ率の判定精度 (78.3%) が最も高かった．

この場合 (C^1) の 0.01 刻みのヒストグラムを図 3 に示す．この図の上部は Y のテキストペアの個数を, 下部は N のテキストペアの個数を示している．この図より, 類似度の高いところでは Y が多く, 低いところでは N が多いことがわかる．

以上の結果より, 確かに, テキスト間の表層類似度が, 含意認識に有効であることがわかる．この調査結果を踏まえ, 本研究では, 表層類似度の計算に, 集合 C^1 を用いたオーバーラップ率, すなわち, 式 (4) で定義される文字オーバーラップ率 $\text{cor}_D(t_1, t_2)$ を使用する．

6. 含意認識システム

6.1 システムの概要

作成した含意認識システムの概要を図 4 に示す．RITE-2 の BC サブタスクには, このシステムを適用する．このシステムは, 以下の 2 段階で含意認識を行う．

*3 <http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>

*4 <http://www.tokuteicorpus.jp/dist/>

図3 $\text{overlap_ratio}_D(C^1; t_1, t_2)$ のヒストグラム

図4 含意認識システム概要

ステップ1 テキストペアの表層類似度を計算し、類似度が高い場合は Y (含意あり)、低い場合は N (含意なし) と判定する。

ステップ2 ステップ1で Y と判定した場合、そのテキストペアの差異を調査し、その結果を覆す必要があるかどうかを判定する。

以下では、これらのステップについて説明する。

6.2 ステップ1: 表層類似度に基づく判定

表層類似度に基づく判定では、以下の3つの条件のいずれかに該当する場合に、類似度が高いとみなす。

- (y1) $\text{cor}_D(t_1, t_2) \geq 0.73$
- (y2) $\text{cor}_D(t_1, t_2) > \text{cor}_D(t_1, t_2) \geq 0.69$
- (y3) $\begin{cases} 0.69 > \text{cor}_D(t_1, t_2) > 0.65 & \& \\ \text{cor}_D(t_1, t_2) > \text{cor}_D(t_1, t_2) + 0.1 \end{cases}$

これらのいずれの条件にも該当しない場合は、類似度が低いとみなす。

前章の結果に基づけば、類似度が高いとみなす条件は、 $\text{cor}_D(t_1, t_2) \geq 0.69$ である。しかしながら、しきい値付近では、Y と N が混在しており、これらをよりうまく分離するために、上記のように修正した。

まず、 $\text{cor}_D(t_1, t_2) \geq 0.73$ の場合は、無条件で Y と判定する。次に、 $0.73 > \text{cor}_D(t_1, t_2) \geq 0.69$ の場合は、追加の条件として、 $\text{cor}_D(t_1, t_2) > \text{cor}_D(t_1, t_2)$ を設定する。最後に、 $0.69 > \text{cor}_D(t_1, t_2) > 0.65$ の場合は、条件 $\text{cor}_D(t_1, t_2) > \text{cor}_D(t_1, t_2) + 0.1$ を満たす場合に、Y とする。

ここで用いた $\text{cor}_D(t_1, t_2)$ は、漢字およびカタカナの文字集合 K^1 におけるオーバーラップ率を表す。

$$\text{cor}_D(t_1, t_2) = \text{overlap_ratio}_D(K^1; t_1, t_2) \quad (5)$$

平仮名と比較した場合、漢字とカタカナは、内容語の一部として使用される割合が高い。そのため、 $\text{cor}_D(t_1, t_2)$ は、 $\text{cor}_D(t_1, t_2)$ よりも、より内容語を重視したオーバーラップ率となる。

なお、上記の条件に含まれるしきい値は、開発用データを用いて実験的に決定した。

6.3 ステップ2: 差異の調査

ステップ1で Y と判定した場合、そのテキストペアの差異を調査し、その結果を覆す必要があるかどうかを判定する。本システムでは、固有名詞と数字に着目する。

§1 固有名詞の不整合

含意関係にある2つのテキストでは、出現する固有名詞が共通している場合が多い[宇高12]。そのため、「 t_2 に出現する固有名詞が t_1 に出現しない場合、 t_1 から t_2 への含意関係は成立しない」と考えられる。このようなヒューリスティックを導入すれば、以下に示すテキストペアを、正しく N と判定できる。

t_1 : 比較宗教学の観点では世界宗教をエルサレムを聖地とするアブラハムの宗教、インド宗教、東アジア宗教の三つに分類し、他の分野も比較する。
 t_2 : 比較宗教学ではイスラム教、ユダヤ教、キリスト教をアブラハムの宗教に分類する。
 <dataset type = BC(dev.), id = 44, label = N>

このペアの文字オーバーラップ率 ($\text{cor}_D(t_1, t_2)$) は 0.76 と高いが、 t_2 に出現する「イスラム」という固有名詞は、 t_1 には出現しない。

上記のヒューリスティックを実現するためには、固有名詞の抽出と照合が必要となる。固有名詞の抽出は、Juman^{*5}を利用して t_2 を形態素解析した後、以下のいずれかの条件を満たす形態素を固有名詞として抽出する。

- 「自動獲得:Wikipedia」という素性が付与される形態素
- 「固有名詞」という品詞細分類が付与される名詞

固有名詞の照合、すなわち、抽出した固有名詞 w が t_1 に出現するかどうかの判定には、以下の3つの条件を用いる。この中のいずれかに該当すれば、出現するとみなす。

- (a) 固有名詞 w と同じ文字列が t_1 に存在する
- (b) 固有名詞 w から長音符号を削除した文字列が、 t_1 から長音符号を削除した文字列中に存在する
- (c) 固有名詞 w がカタカナ語のとき、 t_2 中で w に隣接するカタカナ文字列 w' が t_1 に存在する

ここで、(b) の条件は、カタカナ語の表記ゆれ(長音符号の有無)に対処するために導入した。(c) の条件は、主に

*5 <http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN>

人名を対処するために導入した (ファーストネームとラストネームのいずれかが存在すればよい) .

§2 数字の不整合

開発用データを分析した結果, t_2 に出現する数字が t_1 に出現しない場合, 含意関係が成立しない可能性が高いことが確認された. 数字に関して, 固有名詞と同様のヒューリスティックを導入すれば, 以下に示すテキストペアを, 正しく N と判定できる.

t_1 : 天京事変は, 天京で 1856 年に発生した太平天国の指導部の内紛で, 東王楊秀清・北王韋昌輝・燕王秦日綱が命を落とし, 2 万人余りが殺害された.
 t_2 : 1856 年 9 月には天京事変が発生, 楊秀清, 韋昌輝, 秦日綱が肅清された.
 <dataset type = BC(dev.), id = 321, label =N>

このペアの文字オーバーラップ率 ($\text{cor}_D(t_1, t_2)$) は 0.89 と高いが, t_2 に出現する「9」という数字は, t_1 には出現しない.

数字の抽出では, t_2 から, 出現するアラビア数字 (半角のみ) を抽出する. ただし, 直後に「つ」が存在する場合は抽出しない. 抽出した数字 n の照合には, 以下の条件を用いる. どちらかの条件に該当すれば, 出現するとみなす.

- (a) 数字 n と同じ数字が t_1 に出現する
- (b) 数字 n を漢数字に変換したものが t_1 に出現する

7. MC システム

本章では, RTE-2 の MC サブタスクで用いるシステムについて説明する. MC サブタスクでは, 含意だけでなく, 換言 (双方向含意) や矛盾の検出が必要となる. F (順方向含意) と B (双方向含意), I (その他) の検出は, 含意認識によって実現が可能である. そこで, 含意認識システムの性能確認を主眼に置き, このタスクにも参加した. 前章で述べた含意認識システムに, 新たに導入する矛盾検出モジュールを組み合わせた形で, MC システムを実装した.

7.1 システム概要

作成したシステムの概要を図 5 に示す. このシステムは, 以下の 3 段階のステップで意味関係の認識を行う.
 ステップ 1 矛盾検出モジュールにおいて, 4 つの条件のいずれかに当てはまるテキストペアを C (矛盾) と判定する.

ステップ 2 含意認識システムを用いて, t_1 が t_2 を含意するかどうか調べる. 含意関係が成立しない場合, I (無関係) と判定する.

ステップ 3 含意認識システムを用いて, t_2 が t_1 を含意するかどうか調べる. 含意関係が成立しない場合は, F (順方向含意) と判定し, 成立する場合は, B (双方向含意) と判定する.

図 5 MC システム概要

以下では, 新たに導入した矛盾検出モジュールについて説明する.

7.2 矛盾検出モジュール

MC サブタスクのポイントは, いかにして矛盾 (C) を検出するかにある. 後の 8 章で示すように, 矛盾の自動検出は, 順方向含意 (F) や双方向含意 (B) の自動認識に比べ, かなり難しい. そのため, 矛盾検出の条件には, 原則として, 順方向含意や双方向含意の認定条件と重ならないようにする方針をとった. その理由は, 矛盾検出の導入が, 含意の認識に悪影響を与えることを防ぐためである.

開発用データを分析した結果, 矛盾するテキストペアは, つぎのような性質を持つ傾向があることがわかった.

- (1) 表層類似度が高い. かつ, 双方向で表層類似度が高い場合が多い (矛盾は, 双方向で成立する関係である).
- (2) 表層の差異が, 意味的に大きな差異をもつことが多い.

これらの点を考慮し, 矛盾検出条件として次の 4 つの条件を設定した.

- (c1) $\begin{cases} \text{cor}_D(t_1, t_2) \geq 0.69 & \& \\ \text{overlap_ratio}_D(C^2; t_1, t_2) > 0.65 & \& \\ \text{Num_mismatch}(t_1, t_2) = 1 \text{ (mismatch)} \end{cases}$
- (c2) $\begin{cases} 0.69 > \text{cor}_D(t_1, t_2) > 0.59 & \& \\ \text{overlap_ratio}_B(C^1; t_1, t_2) > 0.5 \end{cases}$
- (c3) $\begin{cases} 0.69 > \text{cor}_D(t_1, t_2) > 0.5 & \& \\ \text{ht_ratio}(t_1, t_2) > 0.4 \end{cases}$
- (c4) $\begin{cases} \text{cor}_D(t_1, t_2) > 0.69 & \& \\ \text{ht_ratio}(t_1, t_2) > 0.75 \end{cases}$

これらの条件では, いずれも文字オーバーラップ率 (cor_D) を使用している.

最初の条件 (c1) では, 文字オーバーラップ率に, 6.3 節で説明した数字の不整合 (Num_mistach) と, 文字 bigram のオーバーラップ率 $\text{overlap_ratio}_D(C^2; t_1, t_2)$ を組み合わせている. 後者の値が高い場合, t_1 と t_2 は, 文字列としてかなり似ていることになる. それにも関わらず, 数字に不整合がある場合を, 矛盾とみなす. この条件により正しく矛盾と判定できる例を, 以下に示す.

t_1 : 地獄の天使は、1967 年 8 月 18 日に公開されたアメリカ合衆国の映画である。
 t_2 : 地獄の天使は 1930 年の映画である。
 <dataset type = MC(dev.) , id = 537 , label =C>

2 番目の条件 (c2) では、文字オーバーラップ率と、双方向の文字オーバーラップ率とを組み合わせている。この条件により正しく矛盾と判定できる例を、以下に示す。

t_1 : 留萌市は、北海道留萌管内にある市で、留萌振興局所在地で、留萌管内の中心都市である。
 t_2 : 留萌振興局は北海道の振興局のひとつで振興局所在地は新留萌市にある。
 <dataset type = MC(dev.) , id = 447 , label =C>

残り 2 つの条件では、 $ht_ratio(t_1, t_2)$ という関数を用いる。この関数は、次のように定義される。

まず、テキストペアに対し、共通する先頭部と末尾部を切り離し、それぞれを以下のように 3 分割する。

$$\begin{aligned}
 t_1: & [h_1 h_2 \dots h_i][d_1 d_2 \dots d_k][z_1 z_2 \dots z_j] \\
 t_2: & [h_1 h_2 \dots h_i][d'_1 d'_2 \dots d'_l][z_1 z_2 \dots z_j]
 \end{aligned}$$

ここで、 $[h_1 h_2 \dots h_i]$ は、 t_1 と t_2 で先頭から最も長く一致する文字列であり、 i が先頭から一致する最大の長さとなる。 $[z_1 z_2 \dots z_j]$ は、 t_1 と t_2 で末尾から最も長く一致する文字列であり、 j が末尾から一致する最大の長さとなる。

このような分割に基づき、 $ht_ratio(t_1, t_2)$ を以下の式で定義する。

$$ht_ratio(t_1, t_2) = \frac{2(i+j)}{(i+k+j) + (i+l+j)} \quad (6)$$

この値は、2 つのテキストの先頭および末尾でそれぞれ共通する部分が、テキスト全体のどれぐらいの割合を占めるか、ということを表す。式 (2)、(3) で示したオーバーラップ率が出現位置や順序を考慮しないのに対し、この値は、文字列の先頭と末尾の共通文字列にのみ着目した値である。

開発用データでは、この値が比較的大きな場合に矛盾となる場合が多かったため、(c3) と (c4) の条件を設定した。(c3) の条件により正しく矛盾と判定できる例を以下に示す。

t_1 : サイボウズ Live は、無料招待制でサービスをしている。
 t_2 : サイボウズ Live は招待制ではなく、だれでも自由に閲覧できる。
 <dataset type = MC(dev.) , id = 388 , label =C>

(c4) の条件により正しく矛盾と判定できる例を以下に示す。

表 4 システム性能

data	system	BC		MC	
		M-F1	rank	M-F1	rank
開発用	本システム	0.85	-	0.59	-
	本システム	0.79	7/42	0.60	1/21
評価用	Best	0.81	1/42	0.60	1/21
	Baseline	0.63	29/42	0.27	16/21
	(性能低下)	-0.06	-	+0.01	-

t_1 : 大きな政府とは、政府・行政の規模・権限を可能な限り小さくしようとする思想または政策である。
 t_2 : 大きな政府とは、政府・行政の規模・権限を拡大しようとする思想または政策である。
 <dataset type = MC(dev.) , id = 282 , label =C>

8. システム性能と考察

本章では、RITE-2 の開発用データと評価用 (フォーマルラン) データに対する本システムの性能を示し、それらの結果に対して考察する。

8.1 システムの性能

表 4 に、開発用データと評価用データのそれぞれに対する $MacroF1$ (M-F1) を示す。この表の rank は、フォーマルランにおける順位を表す^{*6}。Best は、それぞれのタスクで最も性能が良かったシステム、Baseline は、主催者が用意したベースラインシステム (機械学習を用いたシステム)[Watanabe 13] を意味する。この表に示すように、含意認識システムは 42 システム中 7 位、MC システムは 21 システム中 1 位の成績であった。

この表の最下段の値は、評価用データに対する性能が、開発用データに対する性能からどの程度低下したかを表す。開発用データに基づいてシステムを開発するため、評価用データに対する性能は、開発用データに対する性能より低いのが普通である。しかしながら、MC システムでは、逆に、評価用データに対する性能の方が高かった。

8.2 含意認識システムの検討

含意認識システムの各ステップの性能を、表 5 に示す。表 5 において、ステップ 1 における M-F1(開発用データ: 0.83/ 評価用データ: 0.78) と、ステップ 1 + ステップ 2 における M-F1(0.85/0.79) を比較すると、開発用データおよび評価用データのいずれに対しても、ステップ 2 を導入することにより、性能が向上していることがわかる。また、判定精度に関しても、性能向上が確認できる。向上の度合は、M-F1 で 0.02/0.01、精度で 2%/1%と、決して大きなものではない。これは、固有名詞と数字の情報を用いた判定しか行っていないためである。今後、よ

*6 RITE-2 では、システムの性能を測る指標として $MacroF1$ が用いられた [Watanabe 13]。

表 5 含意認識システムの各ステップの性能

		Y	N	total	recall	prec	M-F1
ステップ 1	N_1	28 / 41	298 / 260	326 / 301	80% / 73%	91% / 86%	-
ステップ 2	N_{21}	4 / 7	16 / 13	20 / 20	4% / 4%	80% / 65%	-
	N(固有名詞)	4 / 5	11 / 6	15 / 11	3% / 2%	73% / 55%	-
	N(数字)	0 / 2	5 / 7	5 / 9	1% / 2% +	100% / 78%	-
	Y_{22}	208 / 208	57 / 81	265 / 289	87% / 81%	78% / 72%	-
ステップ 1	$N (= N_1)$	28 / 41	298 / 260	326 / 301	80% / 73%	91% / 86%	0.86 / 0.79
	$Y (= N_{21} + Y_{22})$	212 / 215	73 / 94	285 / 309	88% / 84%	74% / 70%	0.81 / 0.76
	total	240 / 256	371 / 354	611 / 610	-	83% / 78%	0.83 / 0.78
ステップ 1 +	$N (= N_1 + N_{21})$	32 / 48	314 / 273	346 / 321	85% / 77%	91% / 85%	0.88 / 0.81
	$Y (= Y_{22})$	208 / 208	57 / 81	265 / 289	87% / 81%	78% / 72%	0.82 / 0.76
	total	240 / 256	371 / 354	611 / 610	-	85% / 79%	0.85 / 0.79

各欄は、「開発用データに対する性能 / 評価用データに対する性能」を表す
「+」は性能向上（評価用データに対する性能の方がよいこと）を意味する

り多くの情報を用いた判定を追加することにより、更なる性能向上を期待することができる。

先に述べたように、評価用データに対する含意認識システムの性能は、開発用データに対する性能よりも低い。表 5 より、その性能低下は、それぞれのステップの性能低下の積み重ねが原因となっていることがわかる。

ステップ 1 の表層類似度による判定の性能低下は、文字オーバーラップ率のしきい値、および、しきい値付近の分類精度を向上させるための拡張が、開発用データに最適な形に調整されていることによる。そのため、この性能低下はやむを得ない。

ステップ 2 のヒューリスティックには、いくつかの不備が見つかった。固有名詞のヒューリスティックでは、固有名詞の抽出および照合のいずれにおいても、失敗が存在した。固有名詞の抽出では、固有名詞ではない「古称」という名詞に「自動獲得:Wikipedia」の素性が付与されていたため、固有名詞とみなしてしまった。固有名詞の照合では、「祇園」と「祇園」の表記のゆれが原因で、照合に失敗する場合が存在した。

数字のヒューリスティックでは、「範囲」を表す表現への対処が必要であることが判明した。以下に例を示す。

t_1 : ジョン・ケネス・ガルブレイスは 1943 年から 1948 年にかけて「フォーチュン」誌の編集者を務め、1949 年にはハーヴァード大学の経済学教授に就任した。
 t_2 : フォーチュンは、1940 年代、ジョン・ケネス・ガルブレイスを編集員として起用した。
<dataset type = BC(form.), id = 158, label = Y>

この例では、 t_2 には「1940 年代」が、 t_1 には「1943 年から 1948 年にかけて」という表現が存在する。単純な数字の照合では、これらの表現が整合することを正しく判定できない。範囲を表す表現を含めた形で数字を抽出し、照合することが必要である。

8.3 MC システムの検討

MC システムのそれぞれのステップの性能 (recall と precision) を表 6 に示す。この表より、ステップ 2 以降の I, B, F の判定において、開発用データに対する性能より、評価用データに対する性能の方が高かったことがわかる。これらの判定は、6 章で述べた含意認識システムで行っている。つまり、MC サブタスクの好成績は、含意認識システムによるものである。

一方、C (矛盾) の判定性能は、recall と precision のいずれにおいても低下した。4 つの判定条件別にみると、開発用データにおいて比較的 recall が高かった (c2), (c1), (c4) の条件は、評価用データにおいては recall はいずれも低下し、precision も (c2) を除いて低下した。(c3) の条件は、100% の precision を保ったまま recall が向上したが、全体に対しては微々たる効果しかもたらさない。すなわち、全体として、C の判定に成功しているとは言いがたい。

B, F の判定性能と比較して、C の判定性能の低さは際だっている。今回の結果を見る限り、矛盾の検出は、表層類似度ヒューリスティックを組み合わせただけでは難しいと言わざるを得ない。

8.4 表層類似度に基づく手法の有効性

本研究では、テキスト間の表層類似度を用いて含意認識を行う方針 (4 章) を採用した。作成した含意認識システムは、この方針をそのまま実装した形となっている。

含意認識システムの BC サブタスクでの成績は、41 システム中 7 位であり、上位 20% に入ることはできたが、トップ 3 に入ることはできなかった。この原因の一つは、含意認識システムの開発に MC サブタスクの開発用データも利用したことにある。

表 7 に、文字オーバーラップ率の最適しきい値を決める際に用いるデータと、それぞれの場合に得られるしきい値、および、正解率を示す。ただし、この正解率は、文字オーバーラップ率のみを用いた場合 (ステップ 1) の正

表 6 MC システムの各ステップの性能

		B	F	C	I	total	recall	prec
1	C	10 / 8	9 / 11	20 / 11	20 / 13	59 / 43	31% / 18%	34% / 26%
	c2	7 / 6	7 / 9	9 / 6	19 / 10	42 / 31	14% / 10%	19% / 21% +
	c1	0 / 0	1 / 1	6 / 2	1 / 2	8 / 5	9% / 3%	75% / 40%
	c4	3 / 2	1 / 1	4 / 1	0 / 1	8 / 5	6% / 2%	50% / 20%
	c3	0 / 0	0 / 0	1 / 2	0 / 0	1 / 2	1% / 3% +	100% / 100%
2	I	11 / 9	40 / 31	21 / 31	147 / 170	219 / 241	76% / 80% +	67% / 71% +
3	B	48 / 44	16 / 7	3 / 2	7 / 8	74 / 61	58% / 63% +	65% / 72% +
	F	14 / 9	142 / 156	21 / 17	19 / 21	196 / 203	69% / 76% +	72% / 77% +

各欄は、「開発用データに対する性能 / 評価用データに対する性能」を表す
「+」は性能向上（評価用データに対する性能の方がよいこと）を意味する

表 7 データ毎の最適しきい値と BC サブタスクにおける正解率

しきい値を決める際 使用するデータ	BC	BC + MC
しきい値	0.73	0.69
BC 開発用データに 対する正解率 (%)	83.0	81.5
BC 評価用データに 対する正解率 (%)	76.6	75.1

解率である。この表より、BC の開発用データのみで最適しきい値を決定した方が性能が良かったことがわかる。我々は、より多くのデータを用いてしきい値を決定したが、オープンなデータに対して良い性能が得られると考え、MC サブタスクのデータも利用したが、結果的には、この選択は BC サブタスクに対しては最適ではなかった。

一方、含意認識システムは、MC システムの一部として威力を発揮した。我々の MC システムは、フォーマルランにおいて 1 位の成績を修めたが、その好成績は、先に述べたように、含意認識システムによってもたらされている。このことにより、今回採用した方針が、一定の成果を上げたと考えることができる。

9. ユニットテスト

本章では、RITE-2 のユニットテストを通じて、含意認識システムの言語現象毎の含意認識性能を調べる。この実験は、フォーマルランとは別に、後日実施した。

9.1 概要

ユニットテストは、RITE-2 の一部として実施された、言語現象毎の含意認識精度を評価するタスクである^{*7}。問題の形式は、BC サブタスクと同じである。

BC サブタスクでは、テキスト t_1 と t_2 の差異に、複数の言語現象が現れることがある。それに対し、ユニットテストデータでは、 t_1 と t_2 の差異は、例えば、同義語による置き換えなど、単一の言語現象に限られている。ユニットテストデータで用いられる言語現象のカテゴリは、

文献 [Bentivogli 10a, Sammons 10] を参考に、日本語に合わせたものが採用されている^{*8}。

9.2 結果と考察

言語現象毎の判定結果を表 8 に示す。この表の各行は、case_alternation (受け身や使役などの格交替) など、それぞれの言語現象における判定性能を示している。この表より、含意関係が成立する場合、inference (一般常識に基づく推論, 2 件) を除くすべての言語現象に対して、高い精度で判定ができていたことがわかる。

一方、含意関係が成立しないペアに対しては、Recall が全体的に低く、判定性能も低い。判定を誤った例を以下に示す。

t_1 : 新卒一括採用では、年長者が上司となる。 t_2 : 新卒一括採用では、年長者が部下となる。 <dataset type = Unittest(form.), id = 10-2, label =N, category=disagree:lex>	t_1 : 紅茶用のカップよりもコーヒー用のカップの口径の方が小さい傾向にある。 t_2 : 紅茶用のカップよりもコーヒー用のカップの方が小さい傾向にある。 <dataset type = Unittest(form.), id = 20-3, label =N, category=disagree:phrase>
--	--

上記の例のように、テキスト間で単語やフレーズが異なる場合、テキスト間の表層類似度は高いが、意味的な類似度が低くなり、含意関係が成立しない。このようなケースを検出するためには、固有名詞と数字の以外に、他の情報を利用したルールを追加する必要がある。

ユニットテスト全体に対する性能をまとめたものを、表 9 に示す。この表の Best は、ユニットテストのフォーマルランで最も良い成績をおさめたシステムを意味する。この表から、我々の含意認識システムは、フォーマルランで最も成績が良かったシステムよりも、高い性能を示すことがわかる。

*7 我々は、フォーマルランではこのタスクに参加しなかった。

*8 <http://www.cl.ecei.tohoku.ac.jp/rite2/>

表 8 言語現象毎の認識精度

Label	Category	Y:prec	Y:recall	Y:F1	N:prec	N:recall	N:F1
Y	case alternation	100 (7/7)	100 (7/7)	100	0 (0/0)	0 (0/0)	0
	inference	0 (0/0)	0 (0/2)	0	0 (0/2)	0 (0/0)	0
	spatial	100 (1/1)	100 (1/1)	100	0 (0/0)	0 (0/0)	0
	implicit relation	100 (17/17)	94 (17/18)	97	0 (0/1)	0 (0/0)	0
	list	100 (3/3)	100 (3/3)	100	0 (0/0)	0 (0/0)	0
	synonymy:phrase	100 (33/33)	94 (33/35)	97	0 (0/2)	0 (0/0)	0
	meronymy:lex	100 (1/1)	100 (1/1)	100	0 (0/0)	0 (0/0)	0
	apposition	100 (1/1)	100 (1/1)	100	0 (0/0)	0 (0/0)	0
	modifier	100 (42/42)	100 (42/42)	100	0 (0/0)	0 (0/0)	0
	transparent head	100 (1/1)	100 (1/1)	100	0 (0/0)	0 (0/0)	0
	synonymy:lex	100 (9/9)	90 (9/10)	94	0 (0/1)	0 (0/0)	0
	nominalization	100 (1/1)	100 (1/1)	100	0 (0/0)	0 (0/0)	0
	coreference	100 (4/4)	100 (4/4)	100	0 (0/0)	0 (0/0)	0
	temporal	100 (1/1)	100 (1/1)	100	0 (0/0)	0 (0/0)	0
	entailment:phrase	100 (43/43)	96 (43/45)	97	0 (0/2)	0 (0/0)	0
	hypernymy:lex	100 (3/3)	100 (3/3)	100	0 (0/0)	0 (0/0)	0
	scrambling	100 (15/15)	100 (15/15)	100	0 (0/0)	0 (0/0)	0
	clause	100 (14/14)	100 (14/14)	100	0 (0/0)	0 (0/0)	0
	relative clause	100 (8/8)	100 (8/8)	100	0 (0/0)	0 (0/0)	0
N	disagree:phrase	0 (0/7)	0 (0/0)	0	100 (18/18)	72 (18/25)	84
	disagree:temporal	0 (0/1)	0 (0/0)	0	0 (0/0)	0 (0/1)	0
	disagree:modality	0 (0/1)	0 (0/0)	0	0 (0/0)	0 (0/1)	0
	disagree:lex	0 (0/1)	0 (0/0)	0	100 (1/1)	50 (1/2)	67

表 9 ユニットテスト全体に対する性能

System	Y			N			Acc.	M-F1
	prec	recall	F1	prec	recall	F1		
本システム	95.3	96.2	95.8	70.4	65.5	67.9	92.5	81.8
Best	93.9	93.9	93.9	55.2	55.2	55.2	89.2	74.5

10. 結 論

本論文では、日本語テキスト含意認識に対する、テキスト間の表層類似度に基づく手法の有効性を検討した。まず最初に、テキスト間の表層類似度が含意認識に有効であることを実験的に示した。最も効果的な表層類似度は、テキスト間の文字オーバーラップ率であり、この値のみで 78.3% の判定精度が得られた。この結果を踏まえて作成した含意認識システムは、まず、文字オーバーラップ率を用いて、テキストペアを表層類似度が高い場合と低い場合に分類する。類似度が高い場合に対しては、固有名詞と数字の情報を用いた差異の調査を行い、最終的な判定を行う。さらに、この含意認識システムに矛盾検出モジュールを追加して、含意・換言・矛盾検出システム (MC システム) を実装した。

これらのシステムで参加した RITE-2 フォーマルランでは、含意認識タスク (BC サブタスク) で全 42 手法中 7 位、含意・換言・矛盾検出タスク (MC サブタスク) で全 21 手法中 1 位の成績をおさめた。この結果より、含意認識において、テキスト間の表層類似度を用いた我々の手法が、高い性能を示すことが実証された。

我々の含意認識システムは、2 段目の判定ルールを容易に追加できる構造となっている。現在は、固有名詞と数字に着目したルールのみを用いているが、今後、判定ルールを拡充することで、更なる性能向上を期待することができる。

今後の大きな課題は、表層類似度が低い含意関係が成立する場合の検出である。この実現に向けて、言い換え認識の問題に取り組む必要がある。

謝 辞

本研究では RITE-2 ワークショップにおいて配布されるデータを使用してシステムの開発、評価を行った。RITE-2 の運営に携った皆様に深謝します。

◇ 参 考 文 献 ◇

- [Akiba 11] Akiba, Y., Taira, H., Fujita, S., Kasahara, K., and Nagata, M.: NTTCS Textual Entailment Recognition System for NTCIR-9 RITE, in *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, pp. 330–334 (2011)
- [Bar-Haim 06] Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Gi-

- ampiccolo, D., Magnini, B., and Szepietor, I.: The Second PASCAL Recognising Textual Entailment Challenge, in *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, pp. 1–9 (2006)
- [Bentivogli 09] Bentivogli, L., Dagan, I., Dang, H. T., Giampiccolo, D., and Magnini, B.: The Fifth PASCAL Recognizing Textual Entailment Challenge, in *TAC 2009 Workshop Notebook* (2009)
- [Bentivogli 10a] Bentivogli, L., Cabrio, E., Dagan, I., Giampiccolo, D., Leggio, M. L., and Magnini, B.: Building Textual Entailment Specialized Data Sets: a Methodology for Isolating Linguistic Phenomena Relevant to Inference, in *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pp. 3542–3549 (2010)
- [Bentivogli 10b] Bentivogli, L., Clark, P., Dagan, I., and Giampiccolo, D.: The Sixth PASCAL Recognizing Textual Entailment Challenge, in *TAC 2010 Workshop Notebook* (2010)
- [Bentivogli 11] Bentivogli, L., Clark, P., Dagan, I., and Giampiccolo, D.: The Seventh PASCAL Recognizing Textual Entailment Challenge, in *TAC 2010 Workshop Notebook* (2011)
- [Dagan 05] Dagan, I., Glickman, O., and Magnini, B.: The PASCAL Recognising Textual Entailment Challenge, in *Quiñero-Candela et al.*, Vol. 3944, pp. 177–190, Springer-Verlag (2005)
- [Giampiccolo 07] Giampiccolo, D., Magnini, B., Dagan, I., and Dolan, B.: The Third PASCAL Recognizing Textual Entailment Challenge, in *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pp. 1–9 (2007)
- [Giampiccolo 08] Giampiccolo, D., Dang, H. T., Magnini, B., Dagan, I., Cabrio, E., and Dolan, B.: The Fourth PASCAL Recognizing Textual Entailment Challenge, in *TAC 2008 Proceedings* (2008)
- [Harabagiu 06] Harabagiu, S. and Hickl, A.: Methods for Using Textual Entailment in Open-Domain Question Answering, in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 905–912 (2006)
- [Pham 11] Pham, Q. N. M., Nguyen, L. M., and Shimazu, A.: A Machine Learning based Textual Entailment Recognition System of JAIST Team for NTCIR9 RITE, in *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, pp. 302–309 (2011)
- [Sammons 10] Sammons, M., Vydishwaran, V., and Roth, D.: “Ask not what Textual Entailment can do for You...”, in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1199–1208 (2010)
- [Shibata 11] Shibata, T. and Kurohashi, S.: Predicate-argument Structure based Textual Entailment Recognition System of KYOTO Team for NTCIR9 RITE, in *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, pp. 310–317 (2011)
- [Shima 11a] Shima, H., Kanayama, H., Lee, C.-W., Lin, C.-J., Mitamura, T., Miyao, Y., Shi, S., and Takeda, K.: Overview of NTCIR-9 RITE: Recognizing Inference in Text, in *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, pp. 291–301 (2011)
- [Shima 11b] Shima, H., Li, Y., Orii, N., and Mitamura, T.: LTI’s Textual Entailment Recognizer System at NTCIR-9 RITE, in *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, pp. 386–393 (2011)
- [Tan 11] Tan, Y., Zeng, J., Wang, X., and Hovy, E.: BUPTTeam Participation at TAC 2011 Recognizing Textual Entailment, in *TAC 2010 Workshop Notebook* (2011)
- [Tatar 08] Tatar, D., Tamaianu-Morita, E., Mihis, A., and Lupsa, D.: Summarization by Logic Segmentation and Text Entailment, *Advances in Natural Language Processing and Applications*, pp. 15–26 (2008)
- [Tsuchida 11] Tsuchida, M. and Ishikawa, K.: IKOMA at TAC2011: A Method for Recognizing Textual Entailment using Lexical-level and Sentence Structure-level features, in *TAC 2010 Workshop Notebook* (2011)
- [Uemoto 11] Uemoto, H. and Hattori, K.: Experiments of FX for NTCIR-9 RITE Japanese BC Subtask, in *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, pp. 412–417 (2011)
- [Watanabe 13] Watanabe, Y., Miyao, Y., Mizuno, J., Shibata, T., Kanayama, H., Lee, C.-W., Lin, C.-J., Shi, S., Mitamura, T., Kando, N., Shima, H., and Takeda, K.: Overview of the Recognizing Inference in Text (RITE-2) at the NTCIR-10 Workshop, in *Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies*, pp. 385–404 (2013)
- [Yokote 11] Yokote, ichi K., Tanaka, S., and Ishizuka, M.: Effects of Using Simple Semantic Similarity on Textual Entailment Recognition, in *TAC 2010 Workshop Notebook* (2011)
- [宇高 10] 宇高 邦弘, 山本 和英: n-gram を用いた日本語テキスト含意認識の検討, 言語処理学会第 16 回年次大会発表論文集, pp. 462–465 (2010)
- [宇高 12] 宇高 邦弘, 山本 和英: 含意要因となるテキスト中の表現と仮説の対を用いたテキスト含意認識, 言語処理学会第 18 回年次大会発表論文集, pp. 435–438 (2012)
- [永田 08] 永田 昌明, 藤田 早苗, 平 博順: 汎用的な意味解析技術への挑戦, NTT 技術ジャーナル, Vol. 20, No. 6, pp. 24–27 (2008)
- [小谷 09] 小谷 通隆, 柴田 知秀, 黒橋 貞夫: 言い換え表現の述語項構造への正規化とテキスト含意関係認識での利用, 言語処理学会第 15 回年次大会発表論文集, pp. 260–263 (2009)

〔担当委員: 鶴岡 慶雅〕

2014 年 1 月 21 日 受理

著者紹介

服部 昇平(学生会員)

2013 年名古屋大学工学部電気電子・情報工学科卒業。同年より、同大学院工学研究科電子情報システム専攻博士課程(前期課程)在学中。主に、自然言語処理の研究に従事。情報処理学会学生会員。

佐藤 理史(正会員)

1988 年京都大学大学院工学研究科博士後期課程電気工学第二専攻指導認定退学。京都大学工学部助手、北陸先端科学技術大学院大学助教授、京都大学大学院情報科学研究科助教授を経て、2005 年より名古屋大学大学院工学研究科教授。工学博士。自然言語処理、人工知能の研究に従事。

駒谷 和範(正会員)

1998 年 京都大学工学部情報工学科卒業。2000 年 同大学院情報学研究科知能情報学専攻修士課程修了。2002 年 同大学院博士後期課程修了。京都大学博士(情報学)。京都大学大学院情報学研究科助手、助教、名古屋大学大学院工学研究科准教授を経て、2014 年より 大阪大学産業科学研究所教授。2008 年から 2009 年まで米国カーネギーメロン大学客員研究員。2010 年から 2014 年まで JST さきがけ「情報環境と人」領域研究員兼務。情報処理学会平成 16 年度山下記念研究賞、FIT2002 ヤングリサーチ賞などを受賞。SIGDIAL Scientific Advisory Committee member。情報処理学会、電子情報通信学会、言語処理学会各会員。