2008－NL－186（9）
2008/7/18

主要・対立表現の俯瞰的把握 － ウェブの情報信頼性分析に向けて

河原 大輔† 黒橋 禎夫†,†† 乾 健太郎†

† 情報通信研究機構 〒619–0289 京都府相楽郡精華町光台 3–5
†† 京都大学大学院情報学研究科 〒606–8501 京都府京都市吉田本町
E-mail: dk@nict.go.jp, kuro@i.kyoto-u.ac.jp, inui@is.naist.jp

あらまし ウェブ上の情報は玉石混淆であり、多種多様な報道、主張、意見などが存在する。人々は、これらの情報の信頼性・信憑性を判断することを日常的に行っているが、ウェブ上の情報が爆発的に増え続けている今日において、このような判断を支援するシステムが必要不可欠になりつつある。我々は、このような問題意識のもとに、情報内容、情報発信者、情報外観などの観点から情報信頼性を分析するシステム、WISDOM を開発している。本稿では、まずWISDOM について紹介し、次に情報内容の信頼性分析に向けて、述語項構造に基づき、あるトピックに関する主要・対立表現を俯瞰的に提示するための手法について述べる。さらに、得られた主要・対立表現の評価実験について報告し、今後の方向性について議論する。

キーワード ウェブ, 信頼性, 述語項構造, 対立

Grasping Major Topics and their Contradictions toward Information Credibility Analysis of Web Contents

Daisuke KAWAHARA†, Sadao KUROHASHI†,††, and Kentaro INUI†

† National Institute of Information and Communications Technology
3-5 Hikaridai Seika-cho, Soraku-gun, Kyoto, 619–0289 Japan
†† Graduate School of Informatics, Kyoto University
Yoshida Honmachi, Kyoto, 606–8501 Japan
E-mail: dk@nict.go.jp, kuro@i.kyoto-u.ac.jp, inui@is.naist.jp

Abstract The world wide web contains a large variety of news reports, arguments, opinions and so forth, which vary widely in quality. People judge the credibility of information on the web for decision making in daily life. While the quantity of information on the web is explosively increasing these days, it is of urgent necessity to develop a system that supports such judgment. We have been developing a information credibility analyzing system, WISDOM, from the viewpoint of information contents, information sender, and information appearance. This paper first introduces WISDOM, and describes a method for providing a bird's eye view of major linguistic expressions and their contradictions about a given topic. We evaluate the obtained expressions in our experiments, and report the experimental results. Furthermore, we discuss our future direction.

Key words Web, credibility, predicate-argument structure, contradiction

1. はじめに

今日、ウェブ上の情報は爆発的に増え続けており、さまざまな事柄について多種多様な報道、主張、意見などが存在する。しかし、情報の質、信頼性という観点からみると、日常生活に本当に役立つ情報と、何の根拠もない嘘やデマといった情報が玉石混淆の状態で存在している。

人々は、単純に知りたいことを調べるだけではなく、実世界における行動に対して意思決定するための情報を収集するために、ウェブを利用し始めている。そのために、一般には Yahoo! や Google などの検索エンジンを利用するのであるが、これらの検索結果から情報の信頼性や信憑性の判断を行うのは容易ではない。検索エンジンは、一つの尺度でページのランキングを行っているだけであり、各ページの信頼性の判断を行っているわけではないからである。それぞれのページの信頼性を判断するためには、利用者が各ページを実際に見て判断するしかない

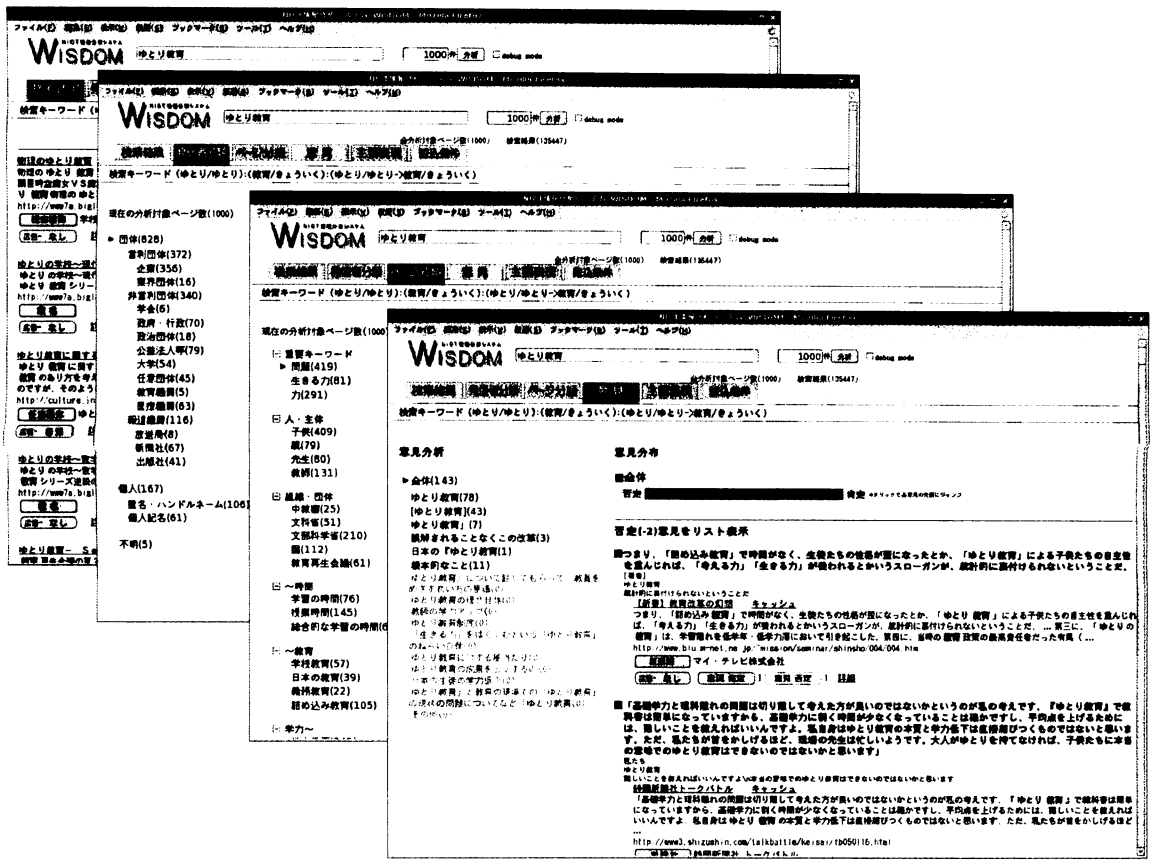

図 1 WISDOM のスクリーンショット
Fig. 1 Screenshots of WISDOM.

という現状である。

たとえば、アガリクスという健康食品について既存の検索エンジンを用いて調べてみると、健康に良いという宣伝をするページが上位に表示され、その他の情報は下位のどこかに埋もれてしまい、本当に健康に良いのかどうかを判断するのは極めて難しい。

信頼できる情報を見つけるためには、ウェブを俯瞰して全体的にどのような事実や意見が書かれているか、どんな人・組織が書いているか、連絡先や情報源は明示されているかといった俯瞰的かつ多角的な観点から情報を整理し利用者に提示する必要があり、既存の検索エンジンとは異なるシステムの構築が不可欠となる。

我々は、ウェブ上の情報信頼性を多角的に分析するシステム WISDOM を開発している [7], [8]。WISDOM を用いて、信頼性評価の各尺度で条件をさまざまに変化させて情報を閲覧することにより、利用者は自分の興味のあるトピックについて、信頼できる情報をより確実に見極めることができるようになる。

本稿では、WISDOM における情報内容の分析の一つとして、主要・対立表現を俯瞰的に提示する手法を提案する。本手法は、与えられたクエリ (トピック) に関するウェブページ集合を対象として、その中で高頻度に出現する言語表現を主要表現とし、それに対立する言語表現を対立表現として抽出、提示するものである。このような提示により、そのトピックに関してどのような事実や意見があるかを一覧として見ることができるようになる。また、その一覧には主要な言語表現だけではなく、それと対立する言語表現を出現頻度が少数であっても抽出できるため、全体把握がしやすいという特徴がある。

本稿の構成は以下の通りである。2 章で情報信頼性分析システム WISDOM の概要を述べる。3 章で主要・対立表現の俯瞰的把握を行うための手法について詳説し、4 章で主要・対立表現の評価実験について述べる。5 章で関連研究について述べ、最後に 6 章でまとめる。

2. 情報信頼性分析システム WISDOM

我々は、情報信頼性を多角的に評価することを目的としたシステム WISDOM を開発している。WISDOM においては、ウェブ上の情報の信頼性を、情報内容、情報発信者、情報外観といった基準でとらえることを提案しており、これらを述語項構造を単位とする自然言語処理によって論理的に分析・組織化することを目指している。

情報内容はウェブページの本文に書かれている内容に着目した観点であり、文内容の分類・要約、意見情報の抽出・分類と

いった処理が含まれる。情報発信者は、発信者の身元に着目した観点であり、所属の分類やその分野での専門性の有無に関する処理が含まれる。情報外観は、ウェブページの見た目に着目した観点であり、情報ソースや連絡先の明記、デザインや文体の適切さに関する処理が含まれる。

WISDOMは、ウェブページ数億件を対象として設計されており、現段階では日本語ウェブページ1億件を対象に動作している。各ウェブページに対しては、文抽出、形態素・構文解析、意見解析[10]、発信者分類[6]、外観に関するページ信頼度付与をあらかじめ行っている。これらの結果は、それぞれウェブ標準フォーマットなどのXMLデータとして格納されている。

利用者はウェブブラウザからWISDOMにアクセスし、分析したいトピックをクエリとして入力する(図1)。分析ボタンをクリックすることにより、そのトピックに関連したウェブページを、発信者、内容および意見で分類した一覧として表示することができ、各項目による全体的な動向や概要を把握できる(図1の発信者分類、ページ分類、意見)。また、特定のページの情報を表示させることで、そのページの発信者や外観の適正さ、意見内容と全体の中での位置づけなどを確認することができる。このように、利用者は、興味のあるトピックについてさまざまな観点から分析結果を閲覧することができ、信頼できる情報をより確実に見極めることができるようになる。

3. 主要・対立表現の俯瞰的把握

WISDOMにおける情報内容の分析の一つとして、与えられたトピックに関する主要・対立表現を抽出し、提示するということを行う。

主要表現とは、与えられたトピックに関するウェブページ集合において、高頻度に出現する言語表現のことである。それに対して、対立表現とは、主要表現に対立、矛盾する言語表現である。たとえば「ゆとり教育」というトピックに対しては、「学力が低下する」が主要表現であり、「学力が向上する」がその対立表現となる。

WISDOMにおいて主要・対立表現を表示したスクリーンショットを図2に示す。このような主要・対立表現の提示により、トピックに関してどのような事実や意見などの言語表現があるかを一覧として見ることができる。一覧には、主要な言語表現だけではなく、それと対立する言語表現が出現頻度が少なくても抽出される。このように、トピックに強く関連する言語表現をマイナーなものも含めて提示することができるので、トピックの全体的な把握を行いやすいという特徴がある。

主要・対立表現の単位としては、述語項構造を用いる。述語項構造とはテキスト文書中の「誰が何をどうした」といった文中の単語間の意味的関係であり、これを単位とした分類、要約、意味解析や、既存知識との比較、整合性検証といった論理的分析を行うことで、信頼できる情報を的確に利用者に提示することができるようになると考えている。

処理の手順としては、以下のように三つの処理に大別される。

(1) 述語項構造の抽出

(2) 述語項構造の集約

(3) 主要・対立述語項構造の同定

これらの処理は、与えられたトピックに関するウェブページ集合を対象として行う。このウェブページ集合は、トピックをクエリとして検索エンジンTSUBAKI[5]にアクセスすることによって取得する。取得するウェブページ数は1,000件とする。

以下では、上記の三つの処理について順に説明する。

3.1 述語項構造の抽出

述語項構造は、述語一つと、それに係る一つ以上の項からなる。述語と項については以下の条件で抽出する(注1)。

- 述語

句読点、括弧と一部の機能語(「ます」「いる」など)以外の形態素列を抽出し、最後の形態素だけ基本形にする。モダリティ、否定、目的、条件を表す表現を含んでいれば、それぞれのフラグを付加する。モダリティとしては、意志、命令、評価、蓋然性などの11種類を扱う。目的と条件は、「～ため」「～ば」などという表現に対して付加するフラグであり、その述語項構造が実際には実現していないことを区別するために導入する。

- 項

自立語列を抽出する。

たとえば、以下の文(1a)の下線部からは「ゆとりで学力が低下する」、文(1b)の下線部からは「学力が低下する〈否定〉」という述語項構造を抽出する。

(1) a. 保護者の83%が「ゆとりで学力が低下した。」と考えていることが明らかになった。
b. ゆとり教育を実践したから学力が低下しているのではなく、…

与えられたトピックに関するウェブページのそれぞれについて、以下の処理を行うことによって述語項構造を抽出する。

(1) ウェブページから重要文を抽出する。重要文としては、トピックを含む文の周辺とする。抽出する重要文の数は、各ページから15文とする。

(2) 形態素解析器JUMAN(注2)と構文・格解析器KNP(注3)を用いて、重要文に形態素・構文・格解析を適用し、述語項構造を抽出する。

(3) 述語項構造をフィルタリングすることによって、機能的な述語項構造を削除する。ウェブ全体と与えられたトピックにおける出現確率比を利用し、与えられたトピックに特徴的に出現する述語項構造以外を削除する。以下に、削除される述語項構造の例を挙げる。

(2) a. 気がつく
b. ことはない
c. ことになる
d. 場合がある

(注1): 本研究の抽出対象は、述語にモダリティなどの態度を付与しているので、「陳述」と言った方が正しいが、本論文では「述語項構造」と呼ぶことにする。
(注2): http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
(注3): http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

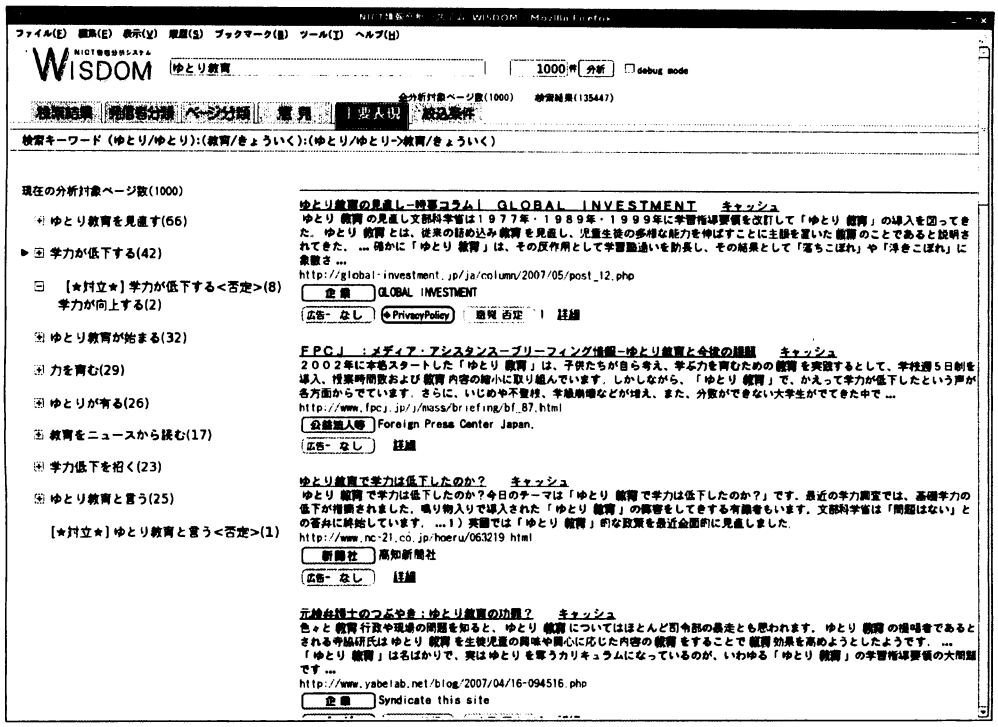

図 2　WISDOM における主要・対立表現表示

Fig. 2　A screenshot of major-contradiction display on WISDOM.

3.2　述語項構造の集約

同一または同義の述語項構造をマージし、頻度を計数することによって、述語項構造の集約を行う。基本的には、同一の述語項構造のマージを行うが、それ以外にも以下のマージを行う必要がある。

- 同義の述語項構造のマージ

述語の同義関係を利用して、述語項構造のマージを行う。述語の同義関係は、国語辞典から抽出された同義語辞書 [4] を用いる。この辞書は約 6,000 エントリからなる。

(3) a. 学力が 低下する
　　b. 学力が 下がる

- 述語項構造の包含関係のチェック

述語項構造の包含関係をチェックすることによって、述語項構造のマージを行う。

(4) a. 学力が低下する
　　b. ゆとりで学力が低下

一般には、このようなマージを行うと、意味が異なる述語項構造もマージしてしまう恐れがある。しかし、このタスクにおいては、与えられたトピックに関するウェブページ集合が対象になっており、文脈が狭められているので、このようなマージをしても多くの場合は問題がないと考えられる。

3.3　主要・対立述語項構造の同定

頻度上位の述語項構造を主要述語項構造として抽出する。また、主要述語項構造に対立する述語項構造があれば抽出する。対立述語項構造は、主要述語項構造に対して次の 2 種類の変換を行い、それを用いて述語項構造集合を検索することによって抽出する。

- 否定フラグの反転

主要述語項構造の述語に否定フラグがついていなければ、否定フラグを付加する。主要述語項構造の述語に否定フラグがついていれば、それを削除する。

(5)　学力が 低下する → 学力が 低下する〈否定〉

- 述語の反義語への置換

主要述語項構造の述語を反義語に置換する。反義語への置換は、国語辞典から抽出された反義語辞書 [4] を利用する。この辞書は約 2,000 エントリからなる。

(6)　学力が 低下する → 学力が 向上する

4.　評価実験

4.1　実験設定と結果

以下の 20 トピックについて、得られた述語項構造の評価を行った。

表 1 抽出された述語項構造の精度

Table 1 Accuracy of obtained predicate-argument structures.

	主要述語項構造	対立述語項構造
良い (○, △)	133/157 (84.7%)	32/42 (76.2%)
問題なし (○)	95/157 (60.5%)	27/42 (64.3%)
マージされるべき (△)	38/157 (24.2%)	5/42 (11.9%)
誤り (×)	18/157 (15.3%)	10/42 (23.8%)

マイナスイオン, レーシック手術, 介護保険制度, 全国学力テスト, 消費税, NHK 受信料, BSE 問題, バイオエタノール, ドラフト制度, 郵政民営化, 合成洗剤, 海洋深層水, カテキン, CO2, IP 電話, 赤ちゃんポスト, 道路特定財源, ゆとり教育, 捕鯨問題, 還元水

各トピックについて、頻度上位 8 個までの主要述語項構造を出力するという条件下で実験を行った。対立述語項構造は、それぞれの主要述語項構造に対して得られていれば出力した。その結果、主要述語項構造は 157 個得られ、対立述語項構造は 42 個得られた。つまり、約 27%(42/157) の主要述語項構造に対して、対立述語項構造が得られたことになる。抽出された主要・対立述語項構造ペアの例を表 2 に示す。

得られた述語項構造の評価として、次の三つへの分類を人手で行った。

○: 良い

△: 良いが、ほかの述語項構造にマージされるべき

×: 誤り

述語項構造が良いかどうかについては、以下の基準をすべて満たせば良いと判定した。

- トピックと関連がある
- 機能的な表現ではなく、意味内容をもつ
- 原文における意味と矛盾することなく抽出されているものが存在する (肯定・否定の極性が反転していないなど)

得られた述語項構造とその評価の例を表 3 に示す。トピック「ゆとり教育」において、「学力低下を招く」はトピックと関連しており良いと考えられるが、「学力が低下する」と同義でありマージされた方が良いと思われる。そのため、「学力低下を招く」は△と評価されている。また、「ゆとり教育と言う」はほとんど意味をなさないので、×と評価されている。

評価の結果を表 1 に示す。得られた述語項構造のうち約 80%は良い (○または△) ことがわかる。

4.2 考　　察

「ゆとり教育」に対しては、「ゆとり教育を見直す」「学力が低下する」などが主要述語項構造として抽出され、「学力が低下する」の対立述語項構造として「学力が向上する」が抽出されている。これによって、「ゆとり教育」について「学力が (低下 | 向上) する」ことが深く関連していることがわかる。また、「学力が低下する」ことが多くのウェブページで述べられている一方、少ないながらも「学力が向上する」ことも述べられているとわかる。実際に、「学力が向上する」が抽出されたウェブページを見てみたところ、ゆとり教育により学力が向上した事例が報告されており、通常の検索エンジンではなかなか見つけることができないウェブページを発見することができた。

一方で、△と評価された、マージされるべき述語項構造が約 20 は、同義語辞書になかったためにマージされなかったものが多い。

(7) a. 受信料を 支払う
　　b. 受信料を 払う

「支払う」の国語辞典の定義文は「相手にお金を払う」であり、この定義文全体が同義句と判定されており、「払う」とは同義語とみなされていない。これらを同義語とみなしマージするということは、述語項構造の包含関係のチェック (3.2 節) と同様に問題が少ないと思われるので、今後検討する予定である。

表 2 抽出された主要・対立述語項構造の例

Table 2 Examples of obtained major expressions and their contradictions.

・トピック: レーシック手術
手術を受ける ↔ 手術を受ける 〈 否定 〉
視力が回復する ↔ 視力が回復する 〈 否定 〉
・トピック: 合成洗剤
合成洗剤を使う ↔ 合成洗剤を使う 〈 否定 〉
環境に悪い ↔ 環境に良い
・トピック: 郵政民営化
郵政民営化に反対する ↔ 郵政民営化に賛成する

表 3 主要・対立述語項構造の評価の例

Table 3 Evaluation examples of obtained major expressions and their contradictions.

トピック: ゆとり教育	
ゆとり教育を見直す	○
学力が低下する	○
↔ 学力が向上する	○
ゆとり教育が始まる	○
力を育む	○
ゆとりが有る	○
教育をニュースから読む	○
学力低下を招く	△
ゆとり教育と言う	×
↔ ゆとり教育と言う 〈 否定 〉	×
トピック: NHK 受信料	
受信料を払う 〈 否定 〉	○
↔ 受信料を払う	○
受信料を支払う	△
↔ 受信料を支払う 〈 否定 〉	△
NHKを見る	○
↔ NHKを見る 〈 否定 〉	○
NHK受信料が利用料に含まれる 〈 否定 〉	○
受信設備を設置する	○
受信料を徴収する	○

(8) a. 消費税が 課税される
b. 消費税が かかる

「課税」の定義文は「国や都道府県などが、会社や個人に税金をわりあてて、はらわせること」となっており、この情報だけではマージを行うことはできない。格フレームの類似度や文脈の類似度を用いて、さらに同義表現を認識する必要がある。

×と判定された述語項構造は、大部分は機能的な表現であった。次に例を示す。

(9) a. 介護保険制度の概要
b. 始めに言及される
c. 上で意義深い

これらは、述語項構造のフィルタリングで削除されなかったものであり、フィルタリングの閾値を適切に設定することにより、削除されるようにする必要がある。

また、トピック「マイナスイオン」において「マイナスイオンを発生させる」に対する対立述語項構造として「マイナスイオンを発生させる〈否定〉」が抽出されていたが、これは誤抽出であった。この対立述語項構造は次の文から抽出されていた。

(10) マイナスイオン効果を得る為には、ただマイナスイオンを発生させるのではなく、プラスイオンを抑えてマイナスイオンを増やす事が重要です。

この文は局所的には「マイナスイオンを発生させる」ことを否定しているが、文の意味としては結局「マイナスイオンを発生させる」ことになる。このような現象を正しく扱うためには、「ただ〜ではなく」のようなパターンを用いて認識する必要がある。

5. 関連研究

本研究では、ウェブページ集合における主要な述語項構造を抽出しているが、複数の文書を対象とする意味では、複数文書要約の研究と関連がある。複数文書要約に関しては多くの研究があるが、特に本研究と関連があるものとして、Radev らの研究 [3] が挙げられる。Radev らは、文の同義・包含関係などの解析をベースにして、複数文書を要約する手法を提案している。

本研究において述語項構造のマージを行っているが、そこで重要となるのは述語の同義関係の認識である。同義・含意関係に関しては、近年、英語を対象として RTE(Recognizing Textual Entailment) と呼ばれる評価型のワークショップが開催されており、活発に研究が行われている [1]。2007 年の第 3 回ワークショップでは、オプショナルのタスクとして、矛盾関係の認識も行うことも始まっており、本研究と関連が深い。対立・矛盾関係の同定に焦点を合わせた研究として Harabagiu らによるものがある [2]。Harabagiu らは、否定表現、反義語および談話関係を用いて、対立・矛盾関係の同定を行っている。

6. おわりに

本稿では、情報内容の信頼性分析に向けて、述語項構造に基づき、あるトピックに関する主要・対立表現を俯瞰的に提示するための手法について述べた。さらに、得られた主要・対立表現の評価実験を行い、その有効性を示した。情報信頼性分析システム WISDOM における一つの観点として、主要・対立表現の一覧を見ることにより、与えられたトピックに関する情報をより多角的に俯瞰することができ、信頼できる情報の把握に役立つと考えられる。今後の課題として、主要・対立表現の抽出元ページの表示方法を改良することが挙げられる。現在はスニペット方式で表示しているが、KWIC(keyword in context) 方式で表示することもできれば、主要・対立表現の文脈付きの傾向を把握しやすくなると思われる。

提案手法は簡単なものであり、一定の精度は実現可能であるが、本当に精度の良い実用的なものにするためには、さまざまなレベルの同義表現の吸収や、より難しい対立・矛盾表現の同定を行う必要がある。これに関しては、今後、奈良先端科学技術大学院大学と共同で研究を進めていく予定である [9]。

文　献

[1] Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pp. 1–9, 2007.

[2] Sanda Harabagiu, Andrew Hickl, and Finley Lacatusu. Negation, contrast and contradiction in text processing. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*, 2006.

[3] Dragomir R. Radev, Hongyan Jing, Malgorzata Styś, and Daniel Tam. Centroid-based summarization of multiple documents. *Information Processing and Management*, Vol. 40, pp. 919–938, 2004.

[4] Tomohide Shibata, Michitaka Odani, Jun Harashima, Takashi Oonishi, and Sadao Kurohashi. SYNGRAPH: A flexible matching method based on synonymous expression extraction from an ordinary dictionary and a web corpus. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP2008)*, pp. 787–792, 2008.

[5] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto, and Sadao Kurohashi. TSUBAKI: An open search engine infrastructure for developing new information access methodology. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP2008)*, pp. 189–196, 2008.

[6] 加藤義清, 乾健太郎, 黒橋禎夫. Web ページの情報発信者の同定とその関係の抽出. 言語処理学会第 14 回年次大会, pp. 737–740, 2008.

[7] 赤峯亨, 宮森恒, 加藤義清, 中川哲治, 乾健太郎, 黒橋禎夫, 木俵豊. Web 情報の信頼性検証のための情報分析システム WISDOM. 言語処理学会第 14 回年次大会, pp. 721–724, 2008.

[8] 宮森恒, 赤峯亨, 加藤義清, 兼岩憲, 角薫, 乾健太郎, 黒橋禎夫. 情報の信頼性分析に向けた評価データおよびプロトタイプシステム WISDOM. 情報処理学会 自然言語処理研究会 2007-NL-180, pp. 103–108, 2007.

[9] 村上浩司, 松吉俊, 隅田飛鳥, 森田啓, 佐尾ちとせ, 増田祥子, 松本裕治, 乾健太郎. 言論マップ生成課題：言説間の類似・対立の構造を捉えるために. 情報処理学会 自然言語処理研究会 2008-NL-186, 2008.

[10] 中川哲治, 宮森恒, 赤峯亨, 乾健太郎, 黒橋禎夫. Web 上の客観的記述からの評価情報抽出に関する技術的検討. 言語処理学会第 14 回年次大会, pp. 344–347, 2008.