

関係事象の組織化によるニュース記事の理解支援

堀江 伸太郎^{1,a)} 切通 恵介^{1,b)} 馬 強^{1,c)}

概要：本研究では、具体的な要素を含まず簡潔にまとめられ記述されたニュース記事に対応する、具体的なイベント（事象）を抽出し関係事象の組織化を行うことで理解支援を行う手法を提案する。ニュース報道では出来事その由来や背景知識について、まとめて簡潔にしか記述しない場合が多く、ユーザーにとって理解が困難な場合がある。我々は、関連記事から事象を表す SVO タブルの抽出、また時間や場所の表現、及び動詞の性質を分析して簡潔に記述された事象に対応する具体事象の抽出を行う。また、得られた具体事象の主体と行動に基づいて分類し、その関連によって記述の根拠を示す。これらを共に提示することでユーザーのニュース記事への理解を支援することを目的とする。

1. はじめに

近年、日々報道される膨大なニュース記事から、それぞれの関連性や連続性に基づいて有用な情報を抽出する研究が盛んに行われている。その中でも特に、ユーザーへ情報を提示することで理解支援を行う方法については多くの関心が寄せられている。しかし、多くの研究は記事の報じられ方や記事間の関連性、分類等に着眼しており、ユーザーの理解困難な部分について補足する部分を明示的に示す研究は少ない [1][2][3][4]。

ニュース記事には、その由来や背景知識について多くが語られず、簡潔な要約のみ記述されることが多く、その事象について詳細の知識を持たないユーザーにとって理解が困難な場合がある。そのような場合、ユーザーはこれまでにかきた事象を検索し、時系列順にユーザー自身でひとつずつ遡ってニュース記事を読んでいかなければならない。日々膨大な数のニュースが報じられる中、目的の記事を見つけ出すことは容易ではなく、多くの記事を読む必要があり、多大な労力を伴う。

そこで我々は、そのような事象に対して過去に実際どのような事があったか、どのような行動かを示すことで利用者の理解支援を行う手法を提案する。

過去に報じられた事象を簡潔にまとめながら言及している例を以下に示す。

“The Park Geun-hye administration is drawing flak for its poor response to the Middle East respiratory syn-

drome outbreak, despite rising public concerns of the surging number of confirmed or suspected patients.”^{*1}

この記事は、2015年6月3日に韓国政府の MERS への対応が不十分だったことについて非難が殺到していることに関する記事である。‘poor response’ といった記述からは、パク大統領が何らかの MERS への対策を行ったが、それが有効ではなかったということが推測されるが、具体的にパク大統領がどのような対策を行ったかという情報は省略されている。このように、過去に報道された事象に対し、誰が行動したのか、ということが書いてあった場合においても、具体的にいつ、何をしたのかについて言及されることは少ない。また、‘poor’ といった評価の根拠も記述中には示されていないため、記事を初めて読む読者にとっては何故そのように評されているのか理解出来ないと考えられる。本研究では、パク大統領、MERS のような固有エンティティ（人、組織、地域など）に着目して事象を組織化すると共に、抽象的にまとめられた事象に対応する具体事象を発見して提示することによって、簡潔に記述されたニュースの理解を支援する手法を提案する。

我々はまず、事象の構造化のために石井らの SVO モデル [1] を利用し、ニュース記事に記述されている事象を SVO タブルを用いて表現することで、記事から抽出された事象を行動とその関係する主体（人物、組織、地域、etc.）を考慮して分類して組織化を行う。また、分類された事象集合間の関係について分析し、ニュースにおける主体とその行動の関係を明らかにすることによってユーザーのニュース理解を支援する。

同時に、事象の SVO タブルを抽出する際、その周辺テ

¹ 京都大学大学院情報学研究科

a) horie@db.soc.i.kyoto-u.ac.jp

b) kiritoshi@db.soc.i.kyoto-u.ac.jp

c) qiang@i.kyoto-u.ac.jp

*1 KoreaHerald:Blue House blasted for MERS response
<http://www.koreaherald.com/view.php?ud=20150603001128>

キスト中の時間や場所などの表現の有無に注目し、時間や場所を伴う事象がより具体的であると判定する。さらに、VerbNet^{*2} などの大規模コーパスを用いて動詞 (SVO タブルの V に相当する部分) の性質を分析し、具体的な動作を伴う動詞であるかどうかを推測して事象の具体的な記述を発見する。発見した具体的な事象を時系列順にユーザに提示してその理解の支援を試みる。

以下、第 2 節にて関連するニュース記事の理解支援に関する研究を紹介し、3 節では提案するシステムの概要、4 節にて提案手法について述べ、5 節にて実験結果及び考察、6 節では結論と今後の課題について述べる。

2. 関連研究

ニュース記事の理解支援に関する研究は多く行われている。バイアスを排除することで理解支援を行う研究として、NewsCube[2] では、ユーザーはニュースには多様な側面があるということ認識しなければならず、それを認識することが理解支援につながるとし、Aspect-level を定義、これに従い分類を行うことでメディアのバイアスを最小化し理解支援を行っている。

灘本ら [3] は、閲覧しているニュース記事が掲載されている Web サイトとは異なるニュースサイトに掲載されている、類似する文章を含む記事を検索、抽出し並べて表示することで比較を行える The Comparative Web Browser(CWB) を提案している。馬ら [4] は、様々なメディアから得られたニュース記事を、記事の視点や話題といったトピックに基づいて偏りや多様性を分析し分類、ユーザーに提示することで理解支援を行う TVBank の提案を行っている。ニュース記事を直接提示する方法以外にも、NewsStand[5] では報じられている事象がいつ起きたのか、どこで起きたのかに注目し地図上に表示するための web インターフェースを提供している。

また、記事中に含まれる固有エンティティに関する記述を分析する研究として、小川らのステークホルダーマイニングが挙げられる [7]。これは、ユーザーが閲覧している記事中に登場する固有エンティティに対する記述の極性を分析することで、固有エンティティ間の利害関係をグラフ化しユーザーの理解支援を行っている。

web 文書から因果関係を抽出することで事象間の関係性を整理する研究として、青野ら [6] は、手がかり表現をもとに web 検索を行い、因果関係ネットワークの可視化を行うことでユーザーの理解支援を行っている。因果関係ネットワークの構築のためにニュース記事の事象を構造化する研究として、石井らは文章中に含まれる「主語・述語・目的語」の構造に着目した SVO モデルの提案を行っている [1]。彼らは、ニュース記事から得られた因果関係ネッ

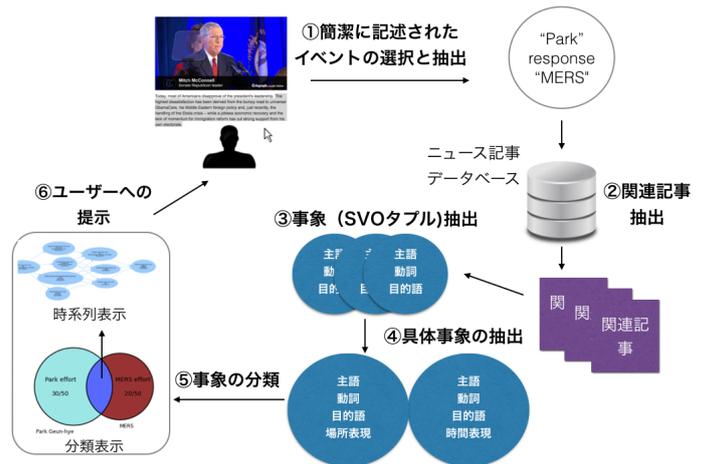


図 1 システムの概要

トワークを構築するノードのマージを行うために、事象を (主語, 述語, 目的語) の三属性に分け構造化することで、キーワードによる類似性の一致によって発生する諸問題への解決を図っている。

また、このように文章における述語構造に着目した研究として、石田らの記述の主観性を考慮したニュース発信者の特徴分析が挙げられる [9]。この研究では、文章が著者の主観に基づいたものなのか、あるいは客観的なものであるかを判別するために、固有エンティティを伴った SVO の構造に着目し、この構造によって表現されるものが客観的事実を指すものであると定義している。

我々の研究では、ユーザーが記述のどの部分を理解したいか選択することが起点となり具体事象を提示するという点でこれらの研究と異なっている。

3. システムの概要

本節では、本システムの想定する利用シナリオに基づき、手法の概要について述べる。システムの概要を図 1 に示す。

本システムではまずユーザーが簡潔に記述されているため理解が困難であると判断した記述を選択することを起点とする。次に、SVO タプルとして選択された事象の抽出を行い、ニュース記事を保存しているデータベースから関連記事の抽出を行う。関連記事から SVO タブルの抽出を行った後、それらが具体事象であるかを判定し、具体事象を抽出する。最後に、得られた具体事象の行動主体及び行動に基いて分類することで関連の分析、また理解が困難であると指定された事象に関する具体事象を時系列に表示した結果をユーザーに表示することで理解支援を行う。

4. 提案手法

4.1 関連記事の抽出

具体事象の提示のためには、具体化の対象となる事象と関連する事象が記述されている記事を網羅する必要がある。

*2 <http://verbs.colorado.edu/verb-index/>

関連する事象とは、前節で述べたように、‘poor response’ と評されるに至った両行動主体それぞれに関するものである。そのため、ここではではまず大規模な記事集合から対象とする事象に関連する記事の抽出方法について述べる。

我々は関連する記事の抽出のために、共通した話題が記事で扱われる際には共通の人物、機関、団体等が登場するという仮定をし、具体化を対象とする記事の固有エンティティに着目する。

具体的に上述した例で言えば、Park Geun-hye と、Middle East Respiratory Syndrome が固有エンティティの例であり、パク大統領の行った行動を含む記事集合には必ずパク大統領が含まれ、同様のことが MERS についても言える。このような固有エンティティをキーワードとして検索を行う手法として、数原らのブログ記事から固有名詞間の動作関係抽出を行ったものがあげられる [8]。彼らは、対象となるブログ記事の検索クエリとして関係の抽出を目的とする固有エンティティを AND で結び検索を行っている。本研究では、GoogleNews^{*3} から得られた約 50 万件の記事集合から固有エンティティを本文に含む記事の抽出を行い関連記事集合とした。

4.2 SVO タブルの抽出

上記で述べた記事集合中の文章から、事象を抽出する方法について述べる。ここでいう事象とは、ある固有エンティティによって表される主体が行った行動を含む文章を指す。このような事象を SVO モデルを利用することで表現する。

SVO タブルの抽出のためには、主語および述語構造の解析が必要となる。そのため、我々は Stanford Core NLP^{*4} を用いて、係り受け構造の解析を行い、主語、動詞、目的語それぞれの抽出を行った。また、he や she などの代名詞で記述されることもあるため、共参照・照応解析を行うことでこれを補完することとした。

4.3 具体事象の抽出

SVO タブルとして抽出したものについて、具体事象であるかの分類を行う。1 節で述べた例を用いると、‘The Park Geun-hye administration is drawing flak for its poor response’ とあったときに考えられる具体事象として、以下のようなものが考えられる。

- South Korean President Park Geun-Hye scolded health officials today over their ‘insufficient’ response to an outbreak of the MERS virus
- Infections with the Middle East Respiratory syndrome (MERS) in South Korea increased to 138
- President Park Geun-hye postpones U.S. Trip

^{*3} <https://news.google.com>

^{*4} <http://nlp.stanford.edu/software/corenlp.shtml>

パク大統領を行動主体とした、MERS への対応行動の他、‘poor response’ と評されるに至った MERS が感染を拡大しているという事実も具体事象に含まれると考えられる。ここで重要なのは、MERS への対応といった記述への具体的な事象は、必ずしも直接的に MERS への行動として現れるわけではないということである。

このような事象の抽出方法として考えられるのは、ある文脈上で語られる固有エンティティの行動を具体事象であると捉え、関連する記事を網羅的に収集し、注目する固有エンティティを主語とする事象を抽出することである。ここでいう文脈とは、それぞれの固有エンティティを含む記事の事を指している。例えば、MERS の文脈でのパク大統領の具体的な行動であれば、MERS を本文に含む関連記事集合中における、パク大統領を主語とする記述のことを指す。

我々は、具体的な行動を伴う固有エンティティを含む記述の判定のために、具体的な行動には特定の日付もしくは行動を行った場所が伴うという仮定を置き、固有エンティティ、述語、目的語、時間もしくは場所からなるタブルを具体事象と定義した。述語が他動詞である場合に目的語を伴わないため、ここでは目的語が存在しない場合も具体事象と判断することとした。

ニュース記事集合 \mathcal{D} 、与えられた固有エンティティ e_s, e_o があつたとき、これを本文に含むニュース記事 d の集合 D_s, D_o はそれぞれ以下のように表すことが出来る

$$D_s = \{d | e_s \in w_d, d \in \mathcal{D}\}$$

$$D_o = \{d | e_o \in w_d, d \in \mathcal{D}\}$$

ここで、 e_s, e_o はそれぞれ主語、目的語を表す固有エンティティであり、 w_d はニュース記事 d に含まれる単語である。

ニュース記事 d_i から抽出される j 番目の SVO タブルを $\tau_{d_i j}$ とおくと、固有エンティティ e_s を主語とする具体事象は以下のように表される。

$$\tau_{d_i j}(e_s, v_i, o_i, t_i, l_i) \text{ where } d_s \in D_s \quad (1)$$

$$\tau_{d_o j}(e_s, v_i, o_i, t_i, l_i) \text{ where } d_o \in D_o \quad (2)$$

ここで v, o, t, l はそれぞれ動詞、目的語、時刻、場所を表したエンティティである。ただし、 t_i, l_i はどちらかの値は少なくとも NULL ではないことが条件である。

(1) の式は固有エンティティ e_s を含む文書 d_s に含まれる、主語 e_s の具体事象を表している。これに対して、(2) の式は、 e_o を含む文書 d_o に含まれる e_s を主語とする具体事象を指している。

4.4 具体事象の判定

ニュース記事の構造として、第 1 パラグラフでニュース

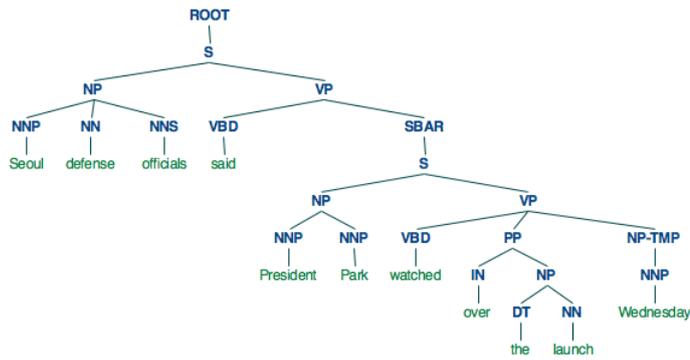


図 2 構文解析木

の概要を語り、第 2 パラグラフで詳細を述べるといった形がとられることが多い。我々は、ニュース記事で述べられるある人物に対する日付や場所を伴った具体的な行動は、主題にて述べられていると仮定し、SVO 抽出対象を第 2 パラグラフまでとした。

次に、得られた SVO タプルから具体事象の抽出のため、目的語それぞれに係る時間表現もしくは場所表現の探索を行った。‘Seoul defense officials said President Park Geun-hye personally watched over the successful launch Wednesday’の文章を例に、これを StanfordCoreNLP にて解析した構文解析木で表すと、図 2 のようになる。探索方法として、S もしくは ROOT ノードの一つ手前の節まで親を辿り、その節の子となる節について S でなければ子を辿っていく深さ優先探索によって葉の探索を行っている。Wednesday の係る先は President Park watched であり、Seoul defence officials には係らない。このとき、said という動詞に係る時間もしくは場所表現があるか探索を行う場合、まず親である VP まで辿り、その後子節を辿る。この例の場合、直後に S ノードが現れるため、探索は終了となる。watched に係る表現の探索を行った場合には、一つ親の VP までさかのぼった後、PP ノード、NP-TMP ノードの子節の探索となる。これを再帰的に行うことで葉ノードである Wednesday に辿りつき、watched は時間表現 Wednesday が係ることがわかる。したがってこの例では、(President Park, watched) が具体事象となる。

具体事象の判定として、VerbNet を用いて動詞の性質を考慮する方法について述べる。VerbNet では動詞の格フレームから、述語の意味的役割を確認することが出来る。そのため、得られた動詞と意味的な役割が類似したものを SVO タプルから抽出することでより正確な具体事象を得る事ができると考えられる。

4.5 具体事象の分類

具体事象が得られた後、具体事象を主体と行動に基いて分類を行うことで、関連の分析を行う。関連の分析の目的は、どのような種類の行動を行ったのか、その期間と頻度

表 1 固有エンティティを含む記事と具体事象数

固有エンティティ	記事数	具体事象	Park Geun-hye	Middle East Respiratory Syndrome
Park Geun-hye	701	382	19	0
Middle East Respiratory Syndrome	1700	1040	13	0

について整理することで、記述の根拠を示すことである。

図 1 で示した分類表示では、円の左がパク大統領の行動を表す事象数、右が MERS の行動を表す事象数を示しており、積集合部分がパク大統領の MERS 関連の行動及び MERS のパク大統領関連の具体事象を表している。

ユーザーが指定した事象に対する具体的事象として、この積集合部分から得られた具体的事象をノードとした時系列でのグラフ表示を行うことでパク大統領、MERS、それぞれの行動が可視化できる。

5. 予備実験

5.1 実験データの準備

実験対象のニュース記事として、GoogleNews を対象に 2015 年 4 月 22 日から、World, Business, Technology, Entertainment, Sports, Science, Health, Politics の 8 つのカテゴリを対象に 50 万件のニュース記事の収集を行った。本稿では、予備実験として 1 節に示したパク大統領の MERS への対応の例 (Park Geun-hye, Response, Middle East Respiratory Syndrome) を用いて、関連記事を収集し、具体事象の抽出を行い、結果の評価を行った。

5.2 抽出結果

Park Geun-hye または Middle East Respiratory Syndrome を含む記事数、また記事から抽出した時間表現または場所表現を伴う SVO タプルである具体事象の数、及びタプルを構成する主語の内それぞれを主語とするタプルの数を 1 の表に示す。例えば固有エンティティ Park Geun-hye を例にとると、これを含む記事数が 701 件あり、そのうち場所もしくは時間表現を伴った SVO タプル数が 382 件で、更に Park Geun-hye を主語とするものが 19 件、及び Middle East Respiratory Syndrome を主語とするものが 0 件であったことを示す。

また、得られた具体事象を時系列に表示したものを図 3 に示す。これはパク大統領が含まれる記事集合の内、主語に Park Geun-hye を含む具体事象、すなわちパク大統領の行動が示されている。ノードは上から Subject, Verb, Object となっており、最下段に記事の発行日が示されている。ここでノードに記す時間表現を発行日としたのは、時間表現が具体事象に付随しないことがあるからである。辺で結ばれたものは時間の連続を示しているが、今回時間表

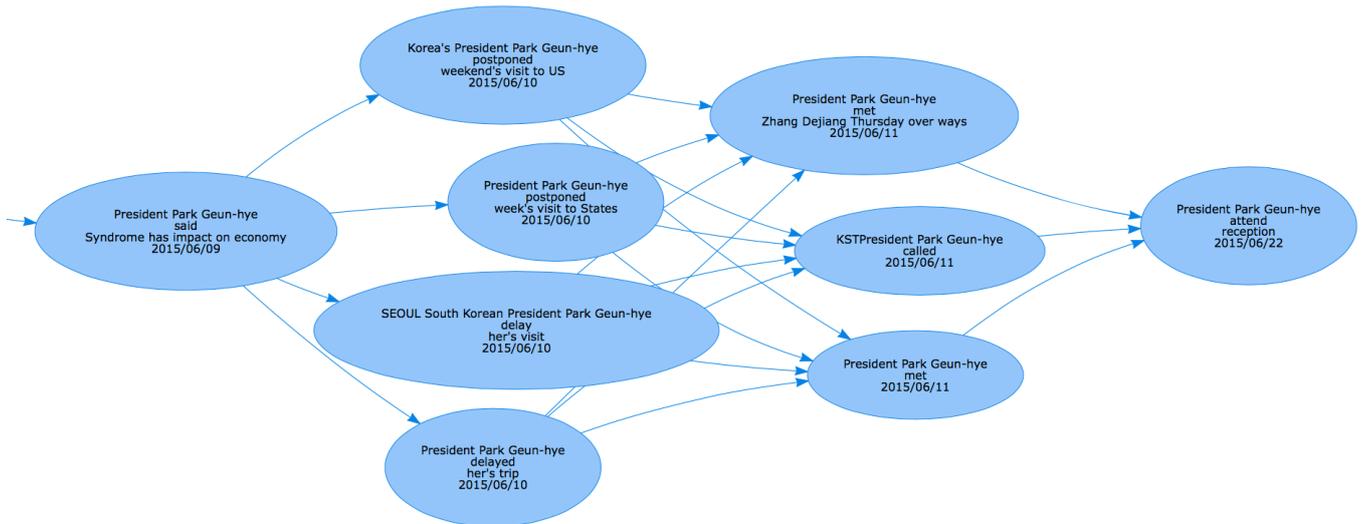


図 3 具体事象を時系列順に表示

現の粒度を日までとしたため、前後する日付のノード間には全て辺が張られている。

表 1 に示した SVO タプル抽出の結果から、*Middle East Respiratory Syndrome* を主語とする SVO タプルが抽出出来ないことがわかった。これは、例えば、*Middle East Respiratory Syndrome* は、記事中では MERS という略称で述べられることが多いように、表現揺れが存在する場合に結果が得られないことに起因している。

この問題を解決するため、周辺単語から意味的類似度によって類似語の抽出を行うことを目的とし、Skip-gram モデルを利用した Word2Vec [10] を用いてベクトル表現への変換を行った。意味的に類似する語は、文脈あるいはトピックに依存するため、関連記事集合をトレーニングセットとして用いて単語間の類似度を算出した。また、固有エンティティは複数の語からなっていることが多いため、bigram 表現を含めた学習を行っている。表 1 に示す ‘Middle East Respiratory Syndrome’ を例にすれば、‘Middle_East’, ‘East_Respiratory’, ‘Respiratory_Syndrome’ といったような語を与え、*Park Geun-hye* を含む記事、及び *Middle East Respiratory Syndrome* を含む記事それぞれについて類似語の抽出を行った。

得られた類似語の内、類似度のしきい値として 0.50 以上であるものを類似語として採用し、これを主語とした具体事象の抽出を行ったそれぞれの抽出結果を表 2 に示す。表では、類似語 ‘president’ を主語とする具体事象の記事件数が、*Park Geun-hye* を含む記事集合中では 29 件、*Middle East Respiratory Syndrome* を含む記事集合中で 21 件抽出されたことを示している。

5.3 具体事象抽出の評価

類似語を用いずに抽出した具体事象及び、類似語を用い

表 2 意味的に類似する語を用いた場合の抽出結果

類似語	Park Geun-hye を含む記事中の具体事象数	Middle East Respiratory Syndrome を含む記事中の具体事象数
president	29	21
plans_visit	0	0
Narendra_Modi	11	0
WASHINGTON_AP	0	0
hosted	0	0
New_Delhi	1	0
First_Lady	0	0
healing	0	0
PTI	1	0
TOKYO_AP	0	0
country	0	2
Respiratory_SyndromeMERS	0	0
masks_prevent	1	8
Syndrome	0	0
precaution_Middle	2	2
contracting_Middle	0	0
stand_guard	0	0
Tourists_wearing	0	0
mask_prevent	0	0
walk	0	0
wears_mask	0	0
walk_past	0	0
Middle_East	0	0
country	0	3
virus	0	2
	0	3

て抽出した具体事象が適切であるかの評価を行った結果を表 3 に示す。記述の簡略化のため、*Park Geun-hye* を Park、*Middle East Respiratory Syndrome* を MERS と記述して

表 3 具体事象の正解率

対象	記事集合	Park	MERS
類義語なし	Park	0.684	0.000
	MERS	0.538	0.000
類似語あり	Park	0.413	0.333
	MERS	0.476	0.350

いる。表では、具体事象の主語が類似語を用いたものであるか、またどの固有エンティティから得られた記事集合であるか、固有エンティティを主語とする具体事象の正解率を示している。

5.4 考察

実験の結果、*Middle East Respiratory Syndrome* の類似語を用いた場合には、パク大統領に関する記事集合では33%程度の正解率であり、MERSに関する記事集合内では35%程の結果となった。類似語を用いない場合よりも正解率の30%程度の向上が見られたのは、‘Middle_Easte’の類似語としてMERSが得られたからである。

類似語を用いた場合、パク大統領を主語とした具体事象の正解率についてはどちらの記事集合についても低下した。表2に示す意味的に類似する語を見ると、パク大統領のかわりに主語となり得るのは‘president’のみであることがわかる。このように、誤った語を類似語として抽出したこと、また、‘President Obama’のような他国の大統領を主語とした事象の抽出も行われており、正解率を下げる要因となっている。

記事本文の抽出精度が低く、記号等が本文中に含まれることでNLPツールによるパース精度が著しく低下していることも大きく正解率を押し下げる要因となっていたため、本文抽出方法についても検討が必要である。

正解率の向上には、上記にあげられる原因の対策のほか、今回使用していない動詞の性質を用いることも必要である。方法としては、動詞と属する格フレームが一致する動詞を含むタプルを抽出することが考えられるが、この方法の場合、‘react’や‘reply’といった語が対象となり、今回抽出されるべきであった、‘infects’や‘scolded’などといった動詞を抽出することは出来ない。これについては、対象動詞前後の単語を利用し文脈を考慮することで対応する方法も考えられる。

6. おわりに

本稿では、関連記事から簡潔に記述された事象に対応する具体事象の抽出を行い、得られた具体事象の主体と行動に基いて分類しその関連を示すことで、記述の根拠とともにそれらを時系列に示す手法の提案を行った。評価実験では、固有エンティティをそのまま利用しSVOタプルの抽出を行う際、表記揺れがある場合に正解率が著しく低く

なる結果が得られた。そのため、抽出した記事集合を訓練データとすることでドメインを限定したWord2Vecによる単語の意味の類似度を評価することによってこのような言い換え表現への対応を行ったところ、若干の精度向上が見られた。

今回、手法の適用範囲がパク大統領とMERS間という、限定されたトピックを扱ったが今後は多数のトピックに対して評価実験を行うことが今後の課題としてあげられる。また、MERSに関連する行動がパク大統領に比べ極端に少ない理由として、MERSが人ではないために主語として語られることが少ないことがあげられる。例えば、‘South Korea has confirmed the 11th death from MERS’といった文章はMERSによる被害でありMERSの行動、と捉えられるが文書の主語は韓国でありこのような記述を我々の方法で抽出することは出来ないため対応が必要になる。

今後、具体的事実の抽出後についても、図3に示した具体事象の可視化方法について検討する必要がある、ノード間の辺をどのように設定すべきか、各日付ごとにクラスタリングしたほうが理解支援に効果的であるか、といった項目について検討を行いたい。

謝辞 本研究の一部は、科研費(課題番号25700033)とSCAT研究費助成による。

参考文献

- [1] Hiroshi Ishii, Qiang Ma, Masatoshi Yoshikawa: Incremental Construction of Causal Network from News Articles, JIP 20(1): 207-215 (2012)
- [2] Souneil Park, Seungwoo Kang, Sangyoung Chung, June-hwa Song: Delivering multiple aspects of news to mitigate media bias, CHI: 443-452(2009)
- [3] Akiyo Nadamoto, Katsumi Tanaka: A Comparative Web Browser (CWB) for Browsing and Comparing Web Pages, WWW 2003, 727- 735
- [4] Qiang Ma, Masatoshi Yoshikawa: Topic and Viewpoint Extraction for Diversity and Bias Analysis of News Contents, APWeb/WAIM:150-161(2009).
- [5] Benjamin E. Teitler, Michael D. Lieberman, Daniele Panozzo, Jagan Sankaranarayanan, Hanan Samet, Jon Sperlberg: NewsStand: A New View on News, GIS: No.18(2008).
- [6] 青野 壮志, 太田 学: 要因検索による因果関係ネットワークの構築と因果知識の獲得, DEIM: F9-1(2010).
- [7] Tatsuya Ogawa, Qiang Ma, Masatoshi Yoshikawa: News Bias Analysis Based on Stakeholder Mining, IEICE Transactions 94-D(3):578-586 (2011).
- [8] 数原 良彦, 戸田 浩之, 櫻井 彰人: ブログ記事を用いた複数話題語間の動作関係抽出手法, 電子情報通信学会論文誌 D Vol. J91-D No.3:619-627 (2008).
- [9] 石田 晋, 馬 強, 吉川 正俊: 記述の主観性を考慮したニュース発信者の特徴分析とその応用, DEIM Forum: B10-3(2010).
- [10] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean: Efficient Estimation of Word Representations in Vector Space, ICLR(2013).