テキストデータにおける予定変更情報および影響の獲得

栗原　俊明†　　豊田　正史††　　喜連川　優††,†††

† 東京大学大学院情報理工学系研究科　〒 113-8656 東京都文京区本郷 7-3-1
†† 東京大学生産技術研究所　〒 153-8505 東京都目黒区駒場 4-6-1
††† 国立情報学研究所　〒 101-8430 東京都千代田区一ツ橋 2-1-2
E-mail: †{kurihara,toyoda,kitsure}@tkl.iis.u-tokyo.ac.jp

あらまし　我々人間が適切な意思決定を行う上で，未来の予定や計画について知っておくことは必要不可欠である．例えば，企業は未来の市場変化や競合他社の計画を知ることはビジネスの成功に大きく役に立つ．政府機関であれば国際情勢の変化，一般的の家族であれば週末のイベントなどを知りたいとも考えるだろう．しかし，未来の予定や計画は変更されることが多々ある．変更になったことに気づかずに，誤った未来を想定していれば，不適切な意思決定をしてしまうことにつながる．さらに，予定間には複雑な関連性があるため，一つの予定の変更はその他の予定の実施にも影響を及ぼす．そこで本研究では，テキストデータにおける予定変更が未来の予定や計画に与える影響を把握することを目的として，テキストデータから予定変更情報を獲得し，さらにその予定変更情報が与える影響を獲得する手法を提案する．

キーワード　未来情報検索, 時間情報分析, 知識獲得, テキストマイニング

1. はじめに

我々人間が適切な意思決定を行う上で，未来の予定や計画を知っておくことは必要不可欠である．例えば，企業は未来の市場変化や競合他社の計画を知ることはビジネスの成功に大きく役に立つ．政府機関であれば国際情勢の変化，一般市民であれば週末のイベントなどを知りたいとも考えるだろう．

近年多量にウェブ上に蓄積され続けているブログ，オンラインニュース記事，ツイッター等のテキストデータは未来の予定や計画を知る上での貴重な資源である．例えば，本研究で使用しているブログデータ（約 23 億文）のうち 1600 万文以上が未来の時間点を示す表現（”来年”，”2015 年 1 月”etc. ）を含んでいる．また，CIA や Google が出資していることでも話題となった“ Recorded Future ”(注1) という企業は，Web 上の未来関連情報の活用に特化したサービスを提供している．

しかし，予定や計画は変更されることが多々ある．予定の変更があった場合，我々がこれまで想定していた未来を塗り替える必要がある．変更になったことに気づかずに，誤った未来を想定していれば，不適切な意思決定をしてしまうことにつながる．そのため，我々はこれらの予定変更に関する情報を常に把握していることが望ましい．

さらに，予定間には複雑な関連性があるため，一つの予定の変更はその他の予定の実施にも影響を及ぼすことが想定される．例えば，東日本大震災による原発事故で“ 2030 年に向けたエネルギー基本計画 ”が白紙に戻されたが，これは日本のエネルギーに関する様々な予定に影響を与えた．例えば，原発の建設計画，再生可能エネルギーの普及，新技術開発，海外への原発輸出政策，地球温暖化対策としての CO2 削減目標等，多くの予定が影響を受けた．つまり，一つの予定が変更された際に，他に影響を受ける予定をも把握することができれば，正しい未来を思い描く上での助けとなる．

そこで本研究では，予定変更がテキストデータにおける未来関連情報に与える影響を把握することを目的として，テキストデータから予定変更情報を獲得し，さらにその予定変更情報が与える影響を獲得する手法を提案する．まずテキストデータ内の時間表現に正規化を行った．その後，手がかり表現と機械学習によるフィルタリングを用いて，未来関連情報から予定変更情報を獲得する手法を提案した．さらに，テキストコーパスにおける前提関係を表す表現を用いた教師なし学習により，予定変更情報が影響を与える未来関連情報を獲得する手法を提案した．

本論文の構成は以下の通りである．まず第 2 章で，これまでに行われた関連研究について順に俯瞰をした後，本論文の位置付けについて説明する．第 3 章では，テキストデータにおける時間表現を正規化し，未来関連情報を抽出した．第 4 章では，未来関連情報から予定の変更に関する情報を獲得する手法について説明する．毎日新聞データを用いた実験により，提案手法の有効性を示す．第 5 章では，予定変更情報により影響を受ける未来関連情報を獲得する手法について説明を行う．最後に，第 6 章で全体のまとめと今後の課題について述べる。

2. 関連研究

テキストデータ内の時間表現を活用する試みとしては，Temporal Information Retrieval なる分野が存在している [1] [2] ．これまでテキストデータの時間情報としては文書作成時間（DCT: Document Creation Time）のみの活用にとどまっていたが，本研究分野ではテキスト内の時間表現を，クエリに対してタイムライン形式で話題を俯瞰することを支援したり，類

(注1)：https://www.recordedfuture.com/

似度の高いテキストをみつけること等に活用することを目指している．

また，SemEval のワークショップである TempEval [4] [5] [6] では，文構造などに着目して，テキスト内のの文書作成時間，イベント，時間表現の関係性を特定する手法の研究等が行われている．

テキストデータにおける未来関連情報に関する研究を初めて行ったのは Baeza-Yates による研究 [7] であるとされており，テキストと時間情報からなるクエリに対して，ニュースアーカイブから未来関連情報を検索する future retrieval というタスクを定義した．その後 Jatowt らは，人名，地名，組織名，イベント名などのクエリに対して，Google News アーカイブから未来関連情報を検索し，それらを分類し可視化する手法の提案 [8] や，関連するイベントが発生する年度の確率分布を求める研究 [9] 等を行っている．Dias ら [10] は参照時間を用いて未来関連情報をトピックごとに分類する研究等を行った．また，Kawai らはクエリとテキスト内の未来時間の関連性の判定等の研究 [11] を行っている．

その他にも，未来関連情報の応用的な活用として，Ho ら [12] は未来関連情報のうち，地名を含むものに着目し，位置情報を利用した推薦システムに応用している．そして，Kanhabua ら [13] は，オンラインニュース記事のユーザーに対して，記事と関連性の高い未来関連情報を提供するためのランキング手法を提案した．

上記のように，テキストデータにおける未来関連情報に関する研究はいくつか行われている．しかし，これらの研究の中で予定変更情報に焦点を置いた研究は行われていない．

3. 時間表現の正規化および未来関連情報の抽出

本節では，テキストデータにおける時間表現を正規化し、未来関連情報を抽出する手法について説明する．本研究では，未来の時間点を示す時間表現を含む文を未来関連情報として定義する．未来関連情報の一例を以下に挙げる。

- 未来関連情報の例

– “安倍晋三首相は１日夕，官邸で記者会見を行い，平成２６年４月に消費税率を現行の５％から８％に引き上げることを正式に表明した．”

– “国際オリンピック委員会（ＩＯＣ）は７日（日本時間８日），2020 年の第 32 回夏季オリンピック競技大会（五輪）を東京で開催することをブエノスアイレス総会で決めた．”

本研究では，未来の時間点を示す時間表現を含む文を未来関連情報として定義する．本節では，テキストデータにおける時間表現を正規化し、未来関連情報を抽出する手法について説明する。

Omar ら [1] によると,「日付」と「時間」については，表現方法によって,「明示的表現」「相対的表現」「黙示的表現」に分けることができる.「明示的表現」は，“ January 25, 2010 ”のように，時間軸上の 1 つの点を指す.「相対的表現」は，“ today ”や“ next month ”のように，その文章の文脈を考慮することで，指す日付を特定できるものである．そして“ 黙示的表現 ”は，“ New Year’s Day 2002 ”等のように，黙示的に日付を指し示しているものである．

表 1　正規化を行った時間表現

	絶対的表現	相対的表現
月単位	NNNN 年（の）N 月 平成 N 年（の）N 月 昭和 N 年（の）N 月 NN 年（の）N 月	来年（の）N 月 今年（の）N 月 昨年（の）N 月 去年（の）N 月 N 月，来月，今月，先月
年単位	NNNN 年 平成 N 年，昭和 N 年	来年，今年，昨年，去年

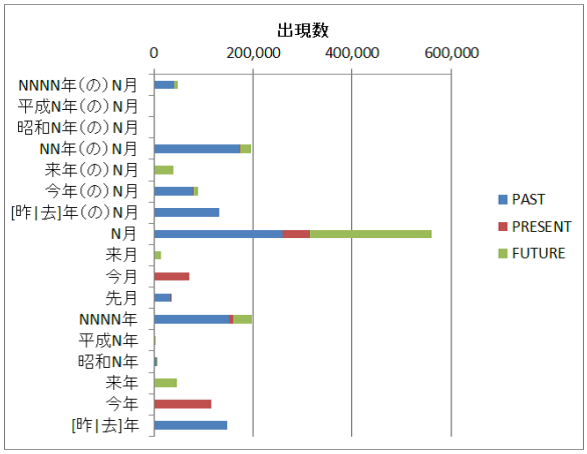

図 1　各時間表現の出現数

本研究では，特定の日付を示す時間表現のうち主な明示的表現と相対的表現を正規化した．表 1 に正規化した表現を示す．ただし，Nは任意の数字列，“ NNNN 年 ”は西暦 4 桁の年度表現，“ NN 年 N 月 ”は“ 98 年 1 月 ”等の西暦年を省略して２桁で表しているものを指す．

相対的表現については，文書作成時間を用いて正規化を行った．また，本研究では“ N 月 ”（1 月，２月 etc.）の時間表現の正規化については、文書作成時間と近い時点へ正規化した．例えば，1 月に作成された文書において“ 12 月 ”という表現がある場合，候補としては“前年の 12 月 ”と“同年の 12 月 ”が特に有力であると考えられるが，文書作成時間により近い“前年の 12 月 ”に正規化する等とした．

まず，本研究で用いる実験データについて説明する．本研究では，毎日新聞データを用いている．期間 1996 年 1 月から 2012 年 12 月で，全文数は 15,732,878 文である．

毎日新聞データにおいて時間表現を正規化した結果を，それぞれ図 1 に示す．“昨年 ”と“去年 ”については，同じ時間点を示すパターンであるため，まとめて表示した．また，青色は文書作成時間よりも“過去 ”，赤色は文書作成時間における“現在 ”，緑色は文書作成時間よりも“未来 ”の時間点を示しているものである．月粒度の表現では，文書作成時間と同じ月を示す表現の場合は“現在 ”，年粒度の場合は，文書作成時間と同じ年を示す表現の場合に“現在 ”というようにしている．

4. 予定変更情報の獲得

予定や計画は変更されることが多々ある．特に，災害や事故等の不測の事態が発生した際には，多くの予定が変更を余儀なくされる．予定の変更があった場合，我々はこれまでの予定情報を更新する必要がある．変更になったことに気づかずに，誤った未来を想定していれば，不適切な意思決定をしてしまうことにつながる．そのため，我々はこれらの予定変更に関する情報を常に把握していることが望ましい．

そこで，未来関連情報から予定変更情報を獲得する．本研究における予定変更情報は，以下の性質をもつ文として定義する．

- 予定変更情報の性質

– 未来の特定の年月日に設定されていた予定や目標を，別の日程に変更する，または，中止することを表現する．

– 情報として確定されたものである．(可能性を示唆するのみのものは含まない)

予定変更情報の一例を以下に挙げる。

- 予定変更情報の例

– “東日本大震災で大きな被害を受けた岩手，宮城，福島の 3 県について，7 月 24 日に予定されていた地上デジタル放送の完全移行を最大 1 年間延期すると総務省が発表した．”

本手順の流れは以下である．まず，未来関連情報全体からランダムに抽出したデータから，予定変更情報において特徴的な手がかり表現を収集した．その後それらの手がかり表現にマッチした文をコーパス全体から抽出し，機械学習を用いてフィルタリングを行い，獲得精度を向上させる．

4.1 提案手法

4.1.1 予定変更を表す手がかり表現による抽出

毎日新聞コーパスのうち，未来の時間表現を含むもののうち 2000 文を人手でチェックし，予定変更情報を表す手がかり表現を収集した．2000 文のうち 56 文が予定変更情報であった．

これらの予定変更情報内で発見された手がかり表現およびそこから推測される手がかり表現を 40 個選択した．これらの手がかり表現を選択する上で，日本語 WordNet (注2)等も利用した．表 2 にその 40 手がかり表現を示す．主な手がかり表現としては“延期”や“中止”等の予定の変更を表す名詞や動詞や，“～予定だった”や“当初は～”のように予定を過去化する表現などがあった．

4.1.2 機械学習によるフィルタリング

予定変更を表す手がかり表現で抽出された文集合に対して機械学習を用いてフィルタリングを行う．本手順で除外すべき文の例を以下に挙げる．

- 不確定

– 活動期間は来年 3 月までの予定だが，延長の可能性もある．

– 併せて、来月 31 日の控訴趣意書提出期限の延期を求めた．

– 米国の利上げが 11 月の大統領選挙後にずれ込むとの見方が、市場関係者の間で支配的になってきた．

表 2 予定変更を表す手がかり表現に用いた表現

予定変更を表す名詞・動詞
延期，延長，変更，中止，前倒し，断念，保留，見送る 先送り，諦める，間に合わない，ずれ込む，ずれこむ 見合わせる，見合せる，持ち越す，持ちこす，もちこす
予定を過去化する表現
当初，もともと，予定だった，計画だった，つもりだった はずだった，が予定されていた，と発表していた に控えていた，表明していた，だったが
その他
予定より，計画より，計画～白紙，計画～練り直す，を待たずに 日程～誤り，まで続投する，まで～続けることを決めた できないことがわかった，以降も当分，以降になる見通し

表 3 学習データとテストデータ

	全体	正例	負例
学習データ	200	91	109
テストデータ	200	108	92

- 否定

– 来年 5 月からの実施予定に変更がないとの立場を表明した．

– 今年 12 月としている工事の完了予定は変更しない．

本研究では，上記の表現を除外するには，前節でマッチした手がかり表現の直後の表現が大きな手がかりになることに着目した．例えばマッチした手がかり表現が“ 延期 ”の場合に, 直後に“ 可能性 ”や“ ～しない ”などの表現があれば, 除外すべき情報である可能性は高い．そこで，文を MeCab (注3) で形態素解析及び見出し語化し，マッチした手がかり表現およびその直後のN個の形態素をその文の特徴ベクトルとした．パラメーター N については，実験において学習データの交差検定を行い，そこで最も高い F 値を示したものとする．

4.2 実　　験

4.2.1 予定変更を表す手がかり表現による抽出

コーパス全体からの表 2 の 40 手がかり表現を含む文を抽出した．その結果，全体で 387,835 文あった未来関連情報のうち，17,946 文が抽出された．

この予定変更を表す手がかり表現による抽出で，どれだけの予定変更情報がとれているかを概算で評価する．全コーパスは 387,835 文で，2,000 文中 56 文が予定変更情報であったため，概算で 10，859 文存在することになる．一方，予定変更を表す手がかり表現による抽出後は，全体で 17,946 文あり，400 文中 199 文が正例 であった．よって概算で 8,928 存在することになる．つまり，この手順による予定変更情報の損失は概算で 1,931 文（17.8%）である．

4.2.2 機械学習を用いたフィルタリング

まず、前節で抽出した文集合から後の機械学習で用いる学習データとテストデータを各々200 文ずつランダムに選択し，ラベル付した．表 3 にこの結果を示す．正例は予定変更情報，負例は予定変更情報でないものである．

(注2)：http://nlpwww.nict.go.jp/wn-ja/

(注3)：http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

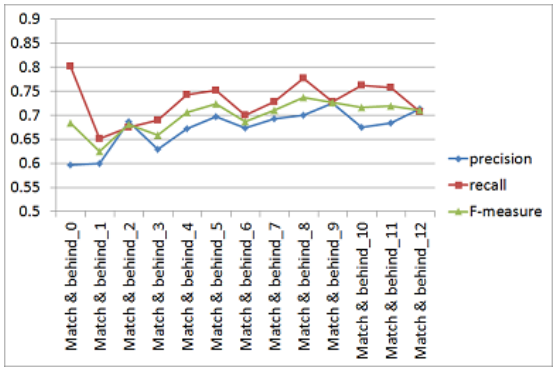

図 2　素性とする形態素数による分類性能（学習データにおける交差検定）

表 4　各手法の分類性能

	Precision	Recall	F-measure	概算獲得文数
All-positive	0.540	1.000	0.701	8928
BOW	0.706	0.631	0.667	5638
Match&behind_8	0.741	0.720	0.730	6432

前節で抽出した文集合に機械学習でフィルタリングを行い，精度向上を目指す．また，本研究の機械学習では機械学習を用いた手法では，パッシブアグレッシブアルゴリズムであるオンライン学習器 opal(注4) [14] を用いた．カーネルは線形カーネルを用いた．

マッチした手がかり表現の後のいくつの形態素を素性とするかについては，学習データの交差検定で最も良い評価値を示したものを採用した．図 2 にその結果を示す．“ Match & behind_N ”は，マッチした手がかり表現とその直後 N 形態素を用いたケースを示している．交差検定としては，ランダムな 2 分割交差検定を 5 回行い，その平均値を評価値とした．その結果,Match & behind_8 が最も高い F 値を示したので，これを採用する．

本手順の評価を行う．図 4 に比較手法および提案手法の各評価値を示す．Recall は，予定変更を表す手がかり表現による抽出後の予定変更情報全体をもとに算出している．比較手法としては，全てを正例 と評価する場合（All-positive）と，機械学習において全品詞の Bag-of-words で行った場合（BOW）を採用した．また，オンライン学習器においては学習データの学習順序によって結果が異なるため,5 回の分類評価値の平均によって評価している．その結果，提案手法（Match & behind_8）は F 値において,All-positive より 0.029,BOW より 0.063 高い結果となった．

最後に，提案手法により獲得された予定変更情報の例を以下に示す．

- 獲得された予定変更情報の例

– “菅直人首相は 3 月 31 日に，福島第 1 原発の事故を踏まえ，2030 年までに原発を現状より 14 基以上増やすとした政府のエネルギー基本計画を白紙にして見直す方針を表明している．”

– “鳩山首相は米軍普天間飛行場の移設問題の結論を来年に先送りし，アフガン支援策の策定を優先させる意向を表明した．”

(注4)：http://www.tkl.iis.u-tokyo.ac.jp/ ynaga/opal/

表 5　前提関係を表す表現

動詞の仮定形＋ば（すれば、ならば etc.）
に向けて，に向け，を目指し，に伴い

5.　予定変更による影響の獲得

本節では，ある予定が変更された際の影響をより詳細に把握することを目的として，予定変更情報によって影響を受ける予定を獲得する手法を提案する．具体的なタスクとしては，前節で獲得した予定変更情報を入力として，未来関連情報の中から影響を受けるものを出力とする．

具体的な例を用いて説明する．例えば，2011 年 05 月に以下のような予定変更情報が獲得されている．菅直人首相は 3 月 31 日に、福島第 1 原発の事故を踏まえ、2030 年までに原発を現状より 14 基以上増やすとした政府のエネルギー基本計画を白紙にして見直す方針を表明している。

この予定変更情報から 2030 年のエネルギー基本計画が変更になったことがわかるが，その他にも多くの予定が影響を受けたことが考えられる．例えば，原発の建設計画，再生可能エネルギーの普及，新技術開発，海外への原発輸出政策，地球温暖化対策としての CO2 削減目標等である．このような影響を受ける予定を獲得することが本手法の目的である．

本手法の提案における主な課題を二つ挙げる．まず一つ目は，出力の多様性を確保することである．本タスクにおいてまず考えられる手法は語句の類似度を尺度として，類似度の高い情報を獲得するというものだが，それでは予定変更情報と似た語句を持った情報以外は獲得することはできない．そこで，語句類似度以外の手法が必要となる．

二つ目は，影響の“方向”を考慮した手法設計が望まれるということである．例えば，2014 年 4 月の“消費税率引き上げ”に伴い，“鉄道の運賃値上げ”が行われる見通しである．この際，“消費税率の引き上げ”が中止されれば，“鉄道の運賃値上げ”も中止になることは十分に考えられる．一方で，“鉄道の運賃値上げ”が中止になったとしても，“消費税率引き上げ”が中止になるとは考えにくい．このように，二つのイベントの間の影響伝播関係の方向性も考える必要がある．

5.1　提 案 手 法

5.1.1　前提表現を含む文を用いた分類器の構築

本研究では，テキストコーパス内にある“前提関係を表す表現”に着目した．本研究では，表 5 の表現を“前提関係を表す表現”とした．以下ではこれらの表現を前提関係表現と呼ぶこととする．前提関係表現を含む文の例を以下に挙げる．

- 前提関係を含む文の例

– “消費税率の引き上げに伴い，鉄道の運賃を値上げする．”

– “オリンピックの開幕に向けて，スタジアムの建設が進む．”

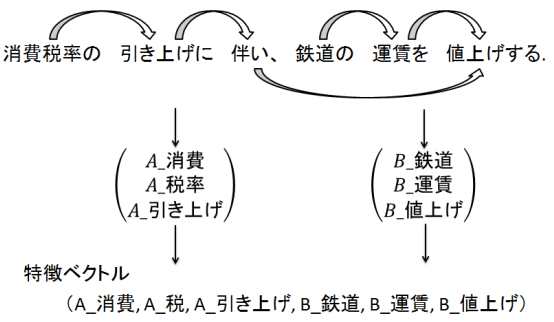

図 3　係り受け解析を用いた学習データの作成

– “ 不信任案が成立すれば，大統領は内閣総辞職か議会解散のどちらかを迫られる．”

– “ 原発が再稼働すれば，節電目標幅を縮小する．”

– “ 再生可能エネルギー拡大に向け，大規模太陽光発電や新エネ・省エネ設備の規制を緩和する．”

– “ 環境にやさしい再生可能エネルギー技術の開発・普及促進を目指し，学術的な意見交換を行う．”

– “ ソニーは放送設備分野でも世界的に圧倒的なシェアを持っており、これに放送事業が加われば、次世代テレビや衛星放送受信機などの標準規格競争で有利になる。”

乾らの研究 [15] や中島らの研究 [16] でもこのように手がかり表現に着目した研究が行われている．しかしこれらの論文では因果関係やイベント連鎖の獲得に焦点を当てており，さらに手がかりとしている表現も異なっている．Radinsky らの研究 [17] においても因果関係を表す表現に着目し，ニュースにおける未来のイベントを予測する研究が行われている．しかし，本研究で焦点を当てている予定変更による影響関係だけではなくより一般的なイベントの因果関係に着目している点や，テキストデータ内から獲得するのではなく，推論規則による予測を目的としている点で，本研究とは異なっている．

本手法では，これらの前提関係表現を含む文を用いた教師なし学習により分類器を構築する．分類器構築の手順を以下で説明する．

係り受け解析結果から各文の特徴量を作成する手順の概要を図 3 に示す．まず，前提関係表現を含む文の係り受け解析を行う．本研究の係り受け解析では吉永らが開発した J.DepP (注5) を用いた．前提関係表現に係っている文節のうち，前提関係表現から 2 つ以内の係り受け関係にあるものを抽出し，形態素解析器 MeCab を用いて形態素解析した上で，それらを “ 前提前 ” の特徴量とする．ただし，本手法では動詞と名詞のみを特徴量の対象とした．また，前提関係表現が係っている文節を見つけ，その文節に係っている文節のうち 2 つ以内の係り受け関係にあるものを含めて “ 前提後 ” の特徴量とする．そして，これらを前提前と前提後で管理した上で統合し，正例の特徴量とする．具体的には前提前の特徴量の前にはには “ A_ ”，前提後の特徴量の前には “ B_ ” とつけて管理した．

学習に用いる負例の作成方法については以下の手順で行う．二つの前提関係表現を含む文をランダムに選び出し，上記の手順で特徴量ベクトルを作成する．そして，一方の前提前の特徴量と他方の前提後の特徴量を統合し，それを負例の特徴量とした．これを負例が正例と同数になるまで繰り返した．

上記の手順で作成した正例と負例から，前節と同様オンライン学習器 opal を用いて分類器を構築した．本手法では，前提前と前提後の特徴量の組み合わせが大きな手がかりとなるため，２次の多項式カーネルを用いた．

また，より精度の高い前提関係を捉えため，本研究では一般的な単語（“ する ”，“ ある ” etc.）は分類器を構築する前に除外した．具体的には，各単語の idf 値を以下の式により算出し，idf 値が 5.5 以下のものを除外した．

$$idf_i = log(M/m_i) \qquad (1)$$

ただし，M はテキストコーパスの全文数，m_i は単語 i の出現する文数である．これにより，テキストコーパス全体にある約 13 万種類ある動詞と名詞のうち，最も一般的な約 400 種類の単語を除外した．

5.1.2　分類器による入力のスコア付け

前節で構築した分類器を用いて，予定変更情報による影響を獲得する．まず，入力として，各々の予定変更情報と全ての未来関連情報のペアを作成する．ただし，本研究では，予定変更情報がテキストデータ内に出現した時点で，それ以前に作成された未来関連情報から，予定の影響を獲得することを想定している．そこで，各予定変更情報について，予定変更情報が作成された月の前月までの未来関連情報のうち，予定変更情報が作成された月の翌月以降の参照時間をもつ未来関連情報を獲得対象とする．

上記条件で作られた予定変更情報と未来関連情報のペアを特徴量ベクトルに変換する．まず，予定変更情報を形態素に分割し，動詞と名詞を抽出する．そして，それらの前に “ A_ ” とつけて管理する．次に，同様に未来関連情報の動詞と名詞を抽出し，それらの前に “ B_ ” とつけて管理する．最後に両者を統合して一つの特徴量ベクトルとし，分類器の入力とする．ただし，前節と同様に一般的な単語については除外しておくこととする．

本研究では，各予定変更情報と未来関連情報を分類器にかけ，分類器によって算出されたマージンを各入力のスコアとした．マージンにしきい値を設け，しきい値以上のスコアをもつ入力を，変更情報により影響を受けるものとして出力する．

5.2　実　　験

2011 年 1 月から 2012 年 12 月における予定変更情報を用いて実験を行った．前章の手法により，本期間には 1587 文の予定変更情報が獲得されていた．

また，分類器を構築する際は，各予定変更情報の前月までのテキストコーパスを用いることとし，各月ごとに別々の分類器を用いた．これは，予定変更後の情報を使うことで本来なかったはずの前提条件を利用してしまうことで不公平性が生じないようにするためである．前提表現を含む文は，1996 年から 2012 年までの間に約 47 万文存在した．機械学習には前節同様オンライン学習器 opal を用い，多項式カーネルを用いた．

(注5)：http://www.tkl.iis.u-tokyo.ac.jp/ ynaga/jdepp/

表 6　提案手法の評価

	語句類似度	提案手法
精度	0.45	0.60

表 7　エネルギー基本計画が白紙になったことを表す予定変更情報から獲得された影響の例

世界のエネルギー需要が 2030 年に現在の 50%増となるとの予測も紹介したうえで、原発は二酸化炭素排出量が少ないことなどから、原子力は温暖化とエネルギー問題の「中核的解決手段の一つとなりうる」と位置づけた。
同法案は温室効果ガス削減中期目標を「2020 年までに 90 年比で 25%削減」と定め、国内排出量取引▽地球温暖化対策税▽再生可能エネルギーの全量固定価格買い取り制度を 3 本柱として施策の検討や実施を盛り込んでいた。
国内の原発関連メーカーは、老朽化した原発の建て替え需要が見込める 2030 年ごろまで輸出でつなぐしかない情勢で、経済産業省も総合資源エネルギー調査会の原子力部会で輸出政策を議論する見通しだ。
同委員会は、2020 年代を目標に行われる最終処分地の選定過程で、処分の実施主体となる新たな認可法人「原子力発電環境整備機構」が、関係地方自治体に事前に十分な情報を公開し自治体と意思疎通を図るなど選定作業の透明性と公正さの確保を求める付帯決議を行った。
二酸化炭素排出量の削減促進を目的に、2400 億円の増税となる地球温暖化対策税を来年 10 月から段階的に導入。

表 8　被災地における地上デジタル放送の完全移行を延期したことを表す予定変更情報から獲得された影響の例

原口一博総務相は 16 日、共同通信本社で開かれた放送協議会総会で講演し、約 1 年後に迫った地上デジタル放送への完全移行について「延期の選択肢は今はまったく考えていない」と、予定通り来年 7 月 24 日にアナログ放送を終了する考えを強調した
総務省や放送局、家電メーカーなどでつくる「地上デジタル推進全国会議」は 1 日、ケーブルテレビ事業者が地上デジタル放送をアナログ信号に変換して送信し、アナログテレビでも視聴できるようにする暫定措置「デジアナ変換」の期間を、15 年 3 月末と決めた。
地上テレビ放送のデジタル化に伴う周波数の大規模再編で、総務省は 14 日、2011 年 7 月に停波して空きが出るアナログ放送の周波数を、携帯電話などの通信サービスや ITS、携帯端末向け放送などに割り当てることを決めた。
総務省の情報通信審議会の情報通信政策部会は 23 日、2011 年 7 月 24 日に地上アナログテレビ放送が終了し地上デジタル放送に移行することに伴い、生活保護世帯に地デジ対応の専用チューナーを現物給付する答申をまとめた。
受信可能世帯は、当初は放送局からの電波が届く範囲に限られるが、中継局を順次増設し、11 年 7 月までには、東北全域で受信可能になる予定。
増田寛也総務相は 9 日、2011 年 7 月 24 日に現行の地上アナログ放送が停止し、地上デジタル放送に完全に移行するのに合わせ、難視聴対策や経済的弱者支援など総額約 2000 億円の総合対策を 09 年度以降約 3 カ年で行うと発表した。

提案手法を評価する上での比較手法は，語句類似度を用いた手法とする．本手法では，各文を MeCab で形態素に分割し，TF-IDF ベクトルに変換した後，そのコサイン類似度を各入力のスコアとした．TF-IDF の式を以下に示す．

$$tfidf_{i,j} = n_{i,j} * log(M/m_i) \qquad (2)$$

$n_{i,j}$ は単語 i の文 j における出現数，M は全文数，m_i は単語 i の出現する文数である．ただし，これまでと同様，品詞は名詞と動詞を用い，idf 値が 5.5 以下のものは除いた．

評価方法としては，提案手法と比較手法においてスコアの高いものから同数の結果を獲得対象として，その中からランダムに選んだ 100 個にラベル付をして，その精度を比較した．両者の獲得数は，3 万（一つの予定変更情報あたり約 20 文）とした．

提案手法、比較手法における精度を表 6 に示す．ただし，本手法の評価に焦点をおくため、正規化や予定変更情報獲得における誤りがあった場合はそれらを除き、合計 100 個になるまで評価を行った．これにより，提案手法の方が比較手法よりも獲得性能において向上していると考えられる．

提案手法によって獲得された影響について、具体的な例を示す．以下の予定変更情報から獲得された未来関連情報は 26 個あったが、そのうちの一部を表 7 に示す．

- 入力とした予定変更情報（1）

– 菅直人首相は 3 月 31 日に，福島第 1 原発の事故を踏まえ，2030 年までに原発を現状より 14 基以上増やすとした政府のエネルギー基本計画を白紙にして見直す方針を表明している．

同様に，以下の予定変更情報から獲得された未来関連情報の一部を表 8 に示す．

- 入力とした予定変更情報（2）

– 東日本大震災で大きな被害を受けた岩手，宮城，福島の 3 県について，7 月 24 日に予定されていた地上デジタル放送の完全移行を最大 1 年間延期すると総務省が発表した．

6. おわりに

本論文では，テキストデータから予定変更情報を獲得し，さらにその予定変更情報が与える影響を獲得する手法を提案した．

まず，テキストデータ内の時間表現を正規化し、未来関連情報を抽出した．次に未来関連情報から予定変更情報を獲得する手法を提案した．まず約 2000 文および毎日新聞データを人手で確認し，予定変更に特有の表現を 40 パターン収集した．実験により，この 40 パターンで約 82.2%の予定変更情報をカバーできることがわかった．その後，機械学習によるフィルタリングを行った．本研究では不確定な情報や否定の情報を除去するため，予定変更に特有の表現の直後を特徴量とする手法を提案した．実験により，提案手法により F 値が向上することを示し，手法の有効性を証明した．

また，予定変更情報によって影響を受ける未来関連情報を獲得する手法を提案した．語句類似度を用いる手法では，多様な情報の獲得，影響の方向性の考慮において問題があることから，本研究では，テキストコーパス内の前提条件を含む文に着目した．前提条件を含む文から教師なし学習を用いて分類器を構築し，各予定変更情報と未来関連情報のペアを入力として算出したマージンがしきい値以上のものを出力とした．実験の結果，

語句類似度を用いた手法よりも高い獲得性能を示し，提案手法の有効性を示した．

文　　献

[1] O.Alonso, J.Strotgen, R.Baeza-Yates and M.Gertz, Temporal Information Retrieval: Challenges and Opportunities, TWAW'11.

[2] J.Pustejovsky, J.M.Castano, R.Ingria, R.Sauri, R.J.Gaizauskas, A.Setzer, G.Katz and D.R.Radev, TimeML: Robust Specification of Event and Temporal Expressions in Text, In proceedings of the AAAI Spring Symposium on New Directions in Question Answering (2003).

[3] R.Kessler, X.Tannier, C.Hag'ege, Finding Salient Dates for Building Thematic Timelines, ACL(2012) .

[4] M.Verhagen, R.Gaizauskas, F.Schilder, M.Hepple, G.Katz and J.Pustejovsky, Semeval-2007 task15: Tempeval temporal relation identification,In proceedings of the Four Int. Workshop on Semantic Evaluations (2007).

[5] M.Verhagen, R.Gaizauskas, F.Schilder, M.Hepple, J.Moszkowicz and J.Pustejovsky, The tempeval challenge: identifying temporal relations in text (2009).

[6] J.Pustejovsky and M.Verhagen, SemEval-2010 Task 13: Evaluating Events, Time Expressions, and Temporal Relations,Proceedings of the NAACL HLT Workshop on Semantic Evaluations (2009).

[7] R.Baeza-Yates, Searching the Future,In Proceedings of ACM SIGIR workshop MF/IR 2005.

[8] A.Jatwot, K.Kanazawa, S.Oyama and K.Tanaka, Supporting Analysis of Future-Related Information in News Archives and the Web,In Proceedings of JCDL 2009.

[9] A.Jatwot and C.A.Yeung, Extracting Collective Expectations about the Future from Large Text Collections, CIKM'11.

[10] G.Dias, R.Campos and A.Jorge, Future Retrieval: What Does the Future Talk About, SIGIR 2011 Workshop on Enriching Information Retrieval.

[11] H.Kawai, A.Jatowt, K.Tanaka, K.Kunieda and K.Yamada, ChronoSeeker: Search Engine for Future and Past Events, ICUIMC 2010 SKKU.

[12] S.Ho, M.Lieberman, P.Wang and H.Samet, Mining Future Spatiotemporal Events and their Sentiment from Online News Articles for Location-Aware Recommendation System, ACM SIGSPATIAL MobiGIS'12.

[13] N.Kanhabua, R.Blanco and M.Matthews, Ranking Related News Predictions, SIGIR'11.

[14] N.Yoshinaga and M.Kitsuregawa, Kernel Slicing: Scalable Online Training with Conjunctive Features, In proceedings of COLING (2010).

[15] 乾孝司, 乾健太郎, 松本裕治, 接続標識「ため」に基づく文書集合からの因果関係知識の自動獲得, 情報処理学会論文誌 Vol. 45, No. 3, pp. 919-933, 2004.

[16] 中島直哉, 吉永直樹, 鍛冶伸裕, 豊田正史, 喜連川優, 時期依存性を有するイベント連鎖の獲得, 日本データベース学会論文誌 vol.12, No.1, pp.103-108, 2013.

[17] K.Radinsky, S.Davidovich, S.Markovitch, Learning Causality for News Events Prediction, WWW2012.