

テキストデータの未来関連情報における予定変更情報の獲得に関する研究

栗原 俊明^{1,a)} 豊田 正史¹ 喜連川 優²

概要: 日々多量に生成されるニュース記事やブログ等のテキストデータには、未来に関する情報も多く含まれている。これらの Web 上の未来関連情報は、我々が未来を予測し適切な行動をとる上で役に立つと考えられる。しかし、未来の予定や計画は変更されることが多々あり、未来関連情報の中には誤った情報も存在する。そこで本研究では、未来の予定や計画が変更されたことを示す“予定変更情報”をテキストコーパスから獲得する手法を提案する。未来関連情報における変更を捉えることは、それらの情報を用いた未来予測を行う上で有用であると考えられる。

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

Knowledge Acquisition Method of Schedule Change Information in Future-timestamped Text Data

TOSHIKI KURIHARA^{1,a)} MASASHI TOYODA¹ MASARU KITSUREGAWA²

Abstract: Huge amount of text data such as news articles and blogs have been generated into the Web, and some of them are future-related information. Future-related information in text data can help us to predict the future and take appropriate action. Future schedule often change, however, and information about change of future schedule will play an important role in predicting the correct future. In this paper, we propose a method to acquire information about future schedule change from text corpus. Detecting the change information will be necessary to maximize the predicting power of future-related information in text data.

Keywords: Future Information Retrieval, Temporal Information Analysis, Knowledge Acquisition, Text Mining

1. はじめに

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が日々多量に生成され、Web 上に蓄積されている。これらの膨大なテキストデータを分析することで、多くの有用な情報を得ることが期待されている。

本研究では“未来関連情報”を、テキストが作成された時間点よりも未来の時間点を示す時間表現を含むセンテ

ンスと定義する。未来関連情報の例を以下に示す。

- 未来関連情報
 - 安倍晋三首相は 10 月 1 日、消費税率を“2014 年 4 月”に現行 5 % から 8 % に引き上げることを表明した。
 - フジテレビ系のバラエティ番組「笑っていいとも！」が“来年 3 月”で終了することが 22 日、分かった。

これらの未来関連情報を活用することは、未来を予測する上で有用である。テキストデータ内には多量の未来関連情報が存在しており、例えば本研究で使用しているブログデータ（約 23 億センテンス）のうち、1600 万センテンス以上が未来の時間点（2015 年、来年 1 月 etc.）を参照しているこ

¹ 東京大学

The University of Tokyo

² 東京大学生産技術研究所, 国立情報学研究所

The University of Tokyo, National Institute of Informatics

^{a)} kurihara@tkl.iis.u-tokyo.ac.jp

とがわかっている。また、近年では“Recorded Future”^{*1}という、Web上の未来関連情報活用に特化したサービスを提供する企業等も登場している。

本研究は、未来関連情報の中に存在する“予定変更情報”に着目した。予定変更情報とは、既存の未来関連情報における日時の変更や中止を表す情報である。以下に予定変更情報の例を示す。

- 予定変更情報

- － オバマ米大統領は、“来月”にモスクワで予定されていたロシアのプーチン大統領との首脳会談を中止することを決めた。
- － “今年12月18日”に北米公開予定だったジョージ・クルーニー監督作『ミケランジェロ・プロジェクト』が、“来年2月”に公開が延期になった。

予定変更情報の出現は、既存の未来関連情報の一部を誤った情報に転換させるという点で、その影響力は大きい。情報が変更になったことに気づかずに、誤った未来関連情報を用いれば、誤った未来予測をしてしまうことになる。

そこで本研究は、テキストデータからこのような予定変更情報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を正規化した。その後、予定変更情報によく用いられるパターンを手で収集し、そのパターンにマッチしたセンテンスに対して機械学習によるフィルタリングを行うことで、精度高くできるだけ多くの情報を獲得することを目指した。また、本研究で用いているデータは、毎日新聞コーパス（期間：1996年1月～2012年12月、全15,732,878センテンス、平均センテンス長：43.2文字）とブログコーパス（期間：2005年1月～2013年6月、全2,340,932,402センテンス、平均センテンス長：33.1文字）であるが、本稿では毎日新聞コーパスからの情報獲得に関する実験について紹介する。

本稿の構成は以下の通りである。まず第2章では関連研究について述べる。第3章では、未来関連情報を抽出するための、時間表現の正規化について述べ、毎日新聞コーパスとブログコーパスにおける時間表現の違いを検討する。第4章では、予定変更情報獲得の提案手法について述べ、毎日新聞コーパスを用いた実験を紹介する。最後に5章でまとめと今後の課題について述べる。

2. 関連研究

これまで、テキストデータ内の時間表現の活用を目指した研究は数多く行われている。また、その中でも未来の時間表現に着目し、未来予測に役立てようとした研究もいくつか行われてきた。しかし、本研究で獲得対象とする予定変更情報に着目した研究はこれまでに行われていない。

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]。以前はテキストデータの時間情報活用という、ドキュメント作成時間（DCT: Document Creation Time）のみの活用にとどまっていたが、これらのテキスト内の時間表現を用いることは、類似度の高いテキストをみつけること等、多様な用途で役に立つと考えられている。また、SemEvalのワークショップであるTempEval[2][3][4]では、文構造などに着目して、テキスト内のドキュメント作成時間、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[5]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Jatwot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[6]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[7]等を行っている。Diasら[8]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[9]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、Horiら[10]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、Kanhubuaら[11]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。

3. 未来関連情報の抽出

本研究では、未来の時間点を示す時間表現を含むセンテンスを未来関連情報として扱う。本節では、未来関連情報の抽出を目的として、テキストデータにおける時間表現の正規化に関する実験を紹介する。

本研究で用いているデータは、毎日新聞コーパス（期間：1996年1月～2012年12月、全15,732,878センテンス、平均センテンス長：43.2文字）とブログコーパス（期間：2005年1月～2013年6月、全2,340,932,402センテンス、平均センテンス長：33.1文字）の二種類である。

3.1 テキストデータにおける時間表現

Pustejovskyら[12]によると、文章内の時間表現の種類には大きく分けて4種類ある。「期間」、「セット」、「日付」と「時間」である。1つ目の「期間」は、例えば“they have been traveling through the U.S. for three years”の“for three years”のように、時間の長さを表現する。2つ目の「セット」は、“She goes to the gym twice a week.”の“

^{*1} <https://www.recordedfuture.com/>

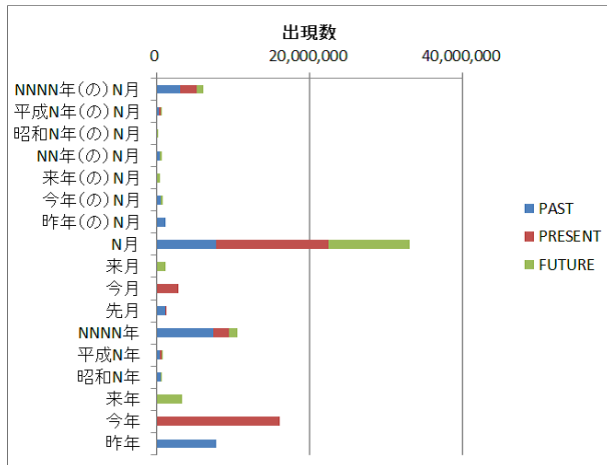


図 1 正規化表現と実験データ内における出現数 (ブログ)

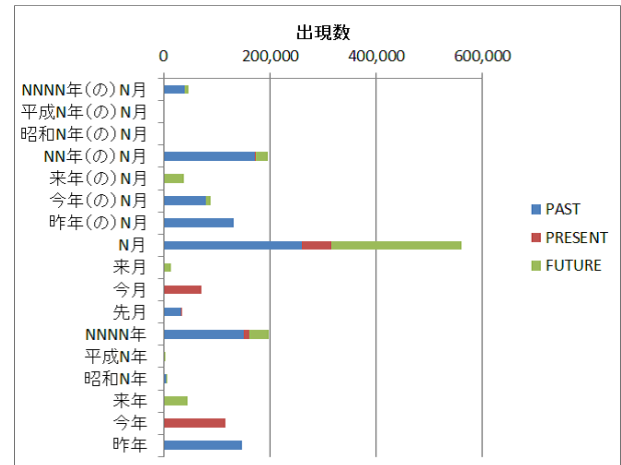


図 2 正規化表現と実験データ内における出現数 (毎日新聞)

twice a week”のように時間の周期性について表現する。そして「日付」と「時間」は，“3 p.m.”や“January 25, 2010”のように、具体的な時間点を指す。

また、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がある。

3.2 時間表現の正規化

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。正規化した表現と、各々の表現のデータ内での出現数を図1、図2に示す。ただし、“NNNN年”は西暦4桁の年度表現を表し、“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。

“N月”(1月, 2月 etc.)の時間表現の正規化にはセンテンスの自制を用いる等の手法も考えられるが、今回はドキュメント作成時間と近い時点へ正規化した。例えば、“12月”という表現について、1月に作成されたテキストデータの場合には前年の12月、11月に作成されたデータの場合は同年の12月へと正規化した。

また、各表現とも参照時間点のドキュメント作成時間との前後関係から、過去、現在、未来で分けてその出現数を示している。例えば年単位の粒度の表現(NNNN年等)では、参照時間点がドキュメント作成年と一致していれば“現在”，その翌年以降であれば“未来”としている。同様に月単位の表現では、参照時間点店がドキュメント作成月と一致していれば“現在”，翌月以降であれば“未来”としている。

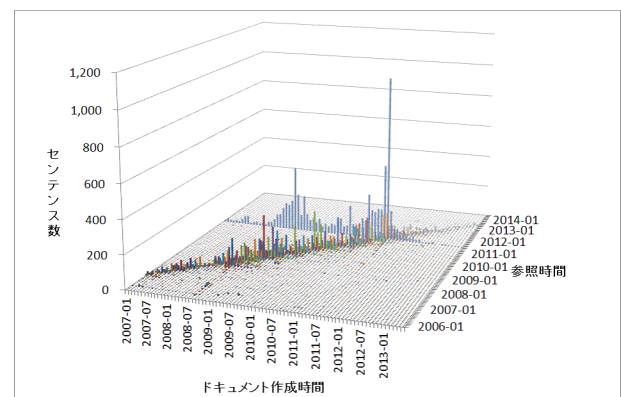


図 3 ドキュメント作成時間と参照時間の関係 (地デジ化)

3.3 ドキュメント作成時間と参照時間の関係性分析

テキストデータは、前節で正規化を行ったテキスト内の時間情報(参照時間: Reference Time)と、テキストが作成された時間に関する情報(ドキュメント作成時間: Document Creation Time)の主に二つの時間情報を持つ。前節の方法で正規化されたデータを用いて、2つの時間情報とそのセンテンス数を軸として3次元グラフにプロットする。

図3は、“地上デジタル放送への移行”に関する情報のドキュメント作成時間と参照時間の関係性をプロットしたものである。前節で正規化したもののうち、月単位の粒度のもののみをプロットしている。ただし、棒グラフの色は、同じ参照時間のものを見やすくするため、参照時間ごとに色分けしている。ドキュメント作成時間と参照時間が同じまたはその前後の情報(すなわちドキュメントが作成された月と同じ月に関する情報)は基本的によく参照されている。さらに、地上デジタル放送への完全移行月である“2011年7月”への参照も多く、ドキュメント作成時間の経過により、その参照数も上下している。特に予定などの延期もなかったため、2011年7月”は実際にその月を迎えるまで、参照され続けている。

一方、“スペースシャトルディスカバリー”に関する情報

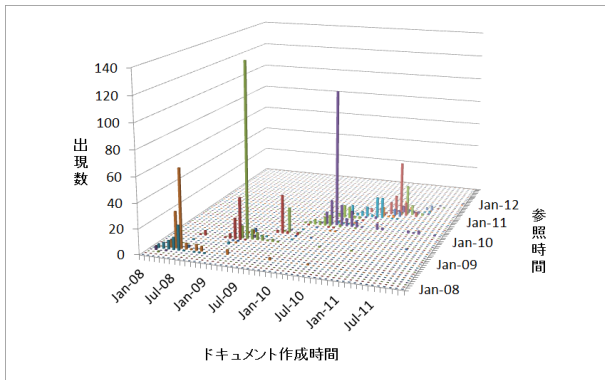


図 4 ドキュメント作成時間と参照時間の関係 (スペースシャトル ディスカバリー)

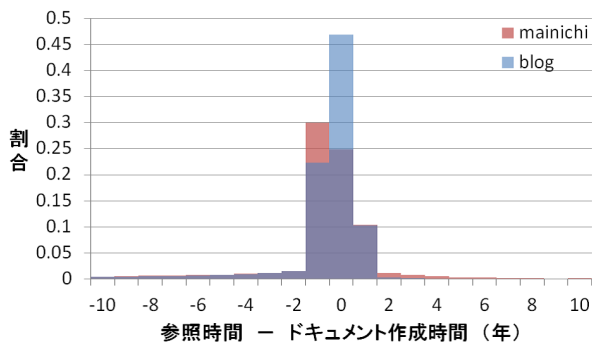


図 5 ドキュメント作成時間と参照時間の時間差 (年粒度)

は、頻繁に打ち上げ延期が発生したため、その参照パターンは異なる。図 4 に“スペースシャトル”と“ディスカバリー”を含むセンテンスのプロットを示す。2008 年 1 月から 2011 年 12 月の期間で、ディスカバリーは 5 回打ち上げられている。各塊は各々の打ち上げに関する情報であると考えられる。しかし、どの塊も 1 つの参照時間をずっと参照しているわけではなく、ドキュメント作成時間が進むと、別の参照時間にずれるポイントがある。例えば最も左の塊では水色からオレンジ色に 1 ヶ月後ろにずれている。これは“打ち上げの延期”によるものである。

3.4 毎日新聞データとブログデータ間での時間表現比較

図 5 と図 6 は正規化した時間表現とドキュメント作成時間との時間差について、その出現割合をプロットしたものである。図 5 は年粒度のもの、図 6 は月粒度のものをプロットしたものである。

これらの図から、毎日新聞データとブログデータにおいて出現する時間表現の違いがわかる。年粒度、月粒度ともブログデータは毎日新聞データよりも時間差 0 の割合が高い。これは、ブログデータは毎日新聞データに比べて、ドキュメントが作成された日時と同じ年月の情報が多いことを示唆している。一方、毎日新聞は過去から未来までブログデータよりも幅広い情報を提供していると考えられる。

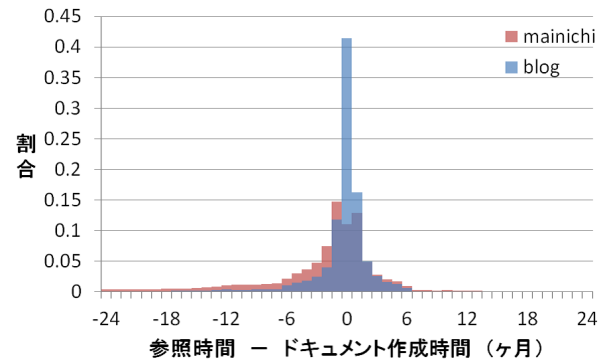


図 6 ドキュメント作成時間と参照時間の時間差 (月粒度)

4. 予定変更情報の獲得

4.1 概要

毎日新聞コーパスからの予定変更情報の獲得を目標とした実験を紹介する。前節で述べた手法により時間表現を正規化した毎日新聞コーパスのうち、ドキュメント作成時間よりも未来の時間表現を含む 387,835 センテンスを用いた。

本研究において予定変更情報とは、以下の性質をもつ情報のことである。

- 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)

実験の大まかな流れについて述べる。まず、コーパス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パターンを収集した。その後、それらのパターンにマッチしたセンテンス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、精度向上を目指した。

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^{*2} [13] を用いた。学習割合や繰り返し回数等のパラメータはデフォルト値を用い、カーネルは線形カーネルとした。オプションとしては、学習データのシャッフリング (-s) とパラメータの平均化 (-a) を設定した。

4.2 パターンマッチによる抽出

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 2,000 センテンスを手でチェックし、予定変更情報に特有のパターンを収集した。2,000 センテンスのうち 56 センテンス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パターンおよびそこから推測されるパターンを 40 個選択した。図 7 にその 40 パターンを示す。

次にコーパス全体から図の 40 パターンを含むセンテ

^{*2} <http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/opal/>

延期、延長、変更、中止、断念、保留、前倒し、見送る、先送り、ずれ込む、
ずれ込む、間に合わない、見合わせる、見合わせる、諦める、持ち越す、
もちこす、持ちこす、計画～白紙、計画～練り直す、

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、だったが、予定より、計画より、を待たずに、
と発表していた、に控えていた、表明していた、
まで続投する、まで～続けることを決めた、できないことが分かった、
以降も当然、以降になる見通し、日程～誤り

図 7 パターンマッチに用いた表現 (40 表現)

	total	positive	negative
train	200	91	109
test	200	108	92

図 8 学習データとテストデータのラベル付け結果

スを抽出した。その結果 17,946 センテンスが抽出された。
このセンテンス群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 センテンスずつランダムに選択し、
ラベル付した。図 8 にこの結果を示す。positive は予定変
更情報,negative は予定変更情報でないものである。

このパターンマッチによる抽出で、どれだけの予定変
更情報がとれているかを概算で検討する。全コーパスは
387,835 センテンスで,2,000 センテンス中 56 センテンスが
予定変更情報であったため、概算で 10,859 センテンス存在
することになる。一方、パターンマッチによる抽出後は、全
体で 17,946 センテンスあり,400 センテンス中 199 センテ
ンスが positive 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 センテンス (17. 8%) である。

4.3 機械学習を用いたフィルタリング

前節で抽出したセンテンス群に機械学習でフィルタリ
ングを行い、精度向上を目指す。学習データ、テストデータは
前節でラベル付けしたものを用いる。

本手順で除外したいセンテンスの例としては、以下のよ
うなものがある。

- 不確定
 - ー 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。
- 否定
 - ー 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。

本研究では、上記の表現を除外するには、前節でマッチし
たパターンの直後の表現が大きな手がかりになることに着
目した。例えばマッチパターンが“延期”の場合に、直後に
“可能性”“～しない”などの表現があれば、除外すべき情

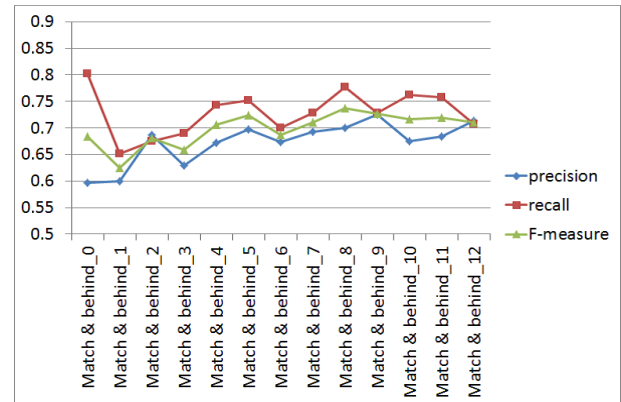


図 9 素性とする形態素数による分類性能 (学習データにおける交差
検定)

	Precision	Recall	F-measure	知識獲得量 (概算)
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

図 10 各手法の分類性能

報である可能性は高い。そこで、センテンスを MeCab^{*3}で
形態素解析及び見出し語化し、マッチパターンおよびその
直後の数形態素をそのセンテンスの特徴ベクトルとすること
にした。

マッチパターン後のいくつか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 9 にその結果を示す。“Match &
behind.N” は、マッチパターンとその直後 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-measure を示した
ので、これを採用する。

4.4 評価

提案手法の評価を行う。図 10 にベースラインおよび提
案手法の各評価値を示す。Recall は、パターンマッチによ
る抽出後の予定変更情報全体をもとに算出している。ベー
スラインとしては、全てを positive 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-measure におい
て,All-positive より 0. 029,BOW より 0. 063 高い結果と

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

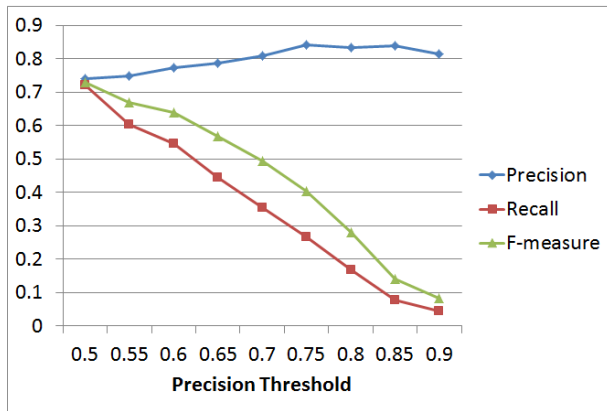


図 11 設定する分類確率閾値による各評価値の変化

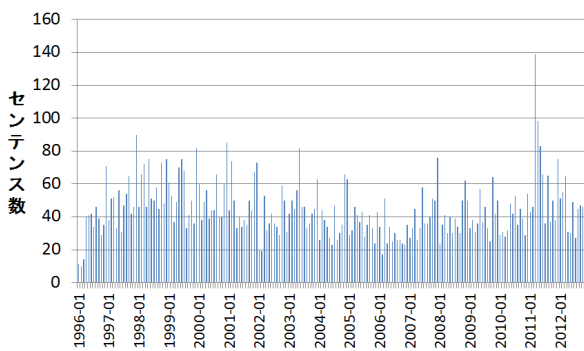


図 12 ドキュメント作成月別の獲得センテンス数

なった。

また、機械学習の分類において、ある閾値以上の分類確率によって positive に分類されているものを予定変更情報として獲得する場合、各閾値による評価値を図 11 に示す。上記の場合と同様に、評価値は 5 回の分類の平均値を用いている。閾値を上げることで、閾値 0.75 とした場合には Precision は 0.842 まで上昇した。

4.5 獲得した予定変更情報の分析

前節までで述べた提案手法を用いて、コーパス全体から予定変更情報を獲得し、分析を行った。図 12 は獲得した予定変更情報を、ドキュメント作成月別に分け、そのセンテンス数をプロットしたものである。この図から、2011 年 3 月から数ヶ月間の間、他の月よりもセンテンス数が多いことがわかる。これは東日本大震災という事故が発生したことで、多くの予定や計画が日程変更や中止されたことによると考えられる。

また、図 13 は 1 月から 12 月の各月でのセンテンス数の 17 年間の平均をとったものである。この図から、12 月は他の月に比べて予定変更情報が多いことがわかる。これは、12 月は例年、特別国会や臨時国会の会期末になること、次年度の予算編成などが行われること、予定などの期限として設定されやすいことなどが理由として考えられる。

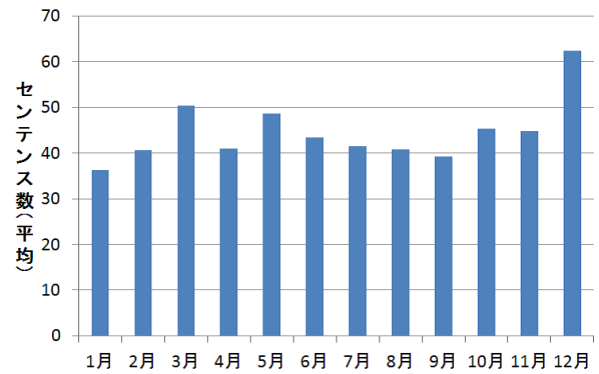


図 13 月別の獲得センテンス数 (17 年平均)

5. おわりに

本研究では、テキストデータの中でも未来の時間表現を含む未来関連情報から、日程の変更や中止を表す予定変更情報を獲得する手法を提案した。まず、ブログデータと毎日新聞データ内の時間表現を正規化し、出現する時間表現について比較を行った。次に、予定変更情報に特有のパターンを収集し、それらのパターンを含むものを抽出した。そして、そこから不要な情報をフィルタリングして精度を高めるために、マッチパターンとその直後の数形態素を特徴ベクトルとする手法を提案した。その結果、機械学習を用いないケースや全品詞を Bag-of-words で分類するケースと比べて F-measure は向上し、その有効性が示された。

今後の課題として、素性選択をより工夫することで分類性能を向上できると考えられる。提案手法では、マッチ表現の直後に着目したが、その他にも素性として有効なものと考えられる。例えば、マッチ表現が「当初」などの場合は、文末に重要な情報がある場合がある。また、「雨天の場合は延期する」など、マッチ語より前の「条件」を表す表現は、情報が不確定であることの手がかりとなると考えられる。

参考文献

- [1] O.Alonso, J.Strotgen, R.Baeza-Yates and M.Gertz: *Temporal Information Retrieval: Challenges and Opportunities*, TAWA'11.
- [2] M.Verhagen, R.Gaizauskas, F.Schilder, M.Hepple, G.Katz and J.Pustejovsky: *Semeval-2007 task15: Temporal relation identification*, In proceedings of the Four Int. Workshop on Semantic Evaluations (2007).
- [3] M.Verhagen, R.Gaizauskas, F.Schilder, M.Hepple, J.Moszkowicz and J.Pustejovsky: *The tempeval challenge: identifying temporal relations in text*, (2009).
- [4] 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).
- [5] R.Baeza-Yates: *Searching the Future*, In Proceedings of ACM SIGIR workshop MF/IR 2005.
- [6] A.Jatwot, K.Kanazawa, S.Oyama and K.Tanaka: *Sup-*

- porting Analysis of Future-Related Information in News Archives and the Web*, In Proceedings of JCDL 2009.
- [7] A.Jatwot and C.A.Yeung : *Extracting Collective Expectations about the Future from Large Text Collections*, CIKM'11.
 - [8] G.Dias, R.Campos and A.Jorge : *Future Retrieval: What Does the Future Talk About?*, SIGIR 2011 Workshop on Enriching Information Retrieval.
 - [9] H.Kawai, A.Jatowt, K.Tanaka, K.Kunieda and K.Yamada : *ChronoSeeker: Search Engine for Future and Past Events*, ICUIMC 2010 SKKU.
 - [10] 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.
 - [11] N.Kanhabua, R.Blanco and M.Matthews : *Ranking Related News Predictions*, SIGIR'11.
 - [12] 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).
 - [13] N.Yoshinaga and M.Kitsuregawa: *Kernel Slicing: Scalable Online Training with Conjunctive Features*, In proceedings of COLING (2010).