Webニュースを用いた未来情報年表の自動構築

吉田光男[†1] 乾 孝司[†1] 山本幹雄[†1]

将来にわたる戦略を立てるためには，起こりうるイベントを予測することが欠かせない．また，未来予測は人々の普遍的欲求である．本論文では，Webニュースステキストに含まれる明示的な未来情報を抽出し，特定の情報にフォーカスしない未来情報年表を自動的に構築する手法を提案する．さらに，年表を構築するために必要となる要約技術を新たな研究課題として提示する．

Automatic Generation of Future Timeline using Web News Corpus

MITSUO YOSHIDA,[†1] TAKASHI INUI[†1] and MIKIO YAMAMOTO[†1]

Future prediction is necessary for people and companies to develop strategies for the future. In addition, people basically enjoy making predictions for fun. In this paper, we propose methods to extract sentences including future events from a collection of Web news and generate "Future Timeline" automatically using the sentences to get a quick overview of the future.

1. はじめに

科学技術・文化の進歩，および近年のグローバル化により社会状況の複雑さが増す一方である．このような状況下で，将来にわたる戦略を立てるためには，起こりうるイベントを予測することが欠かせない．また，これまで人々は科学技術の進歩によって実現される未来像を常に描いてきた．例えば，報知新聞社は1901年に『二十世紀の豫言[⋆1]』と題した100年後を予測する記事を掲載し，朝日新聞社は1955年に『未来生活[⋆2]』と題した未来の生活像を予測する16回の連載記事を掲載した．現在も，Webなどの様々な媒体に同様の記事が掲載されている．未来予測は，人々の普遍的欲求であると考えられ，今後も需要が存在し続けると予想できる．

一方，Webには膨大なテキストデータが存在し，絶えず増加している．さらに，これらには未来に関する情報が書かれている場合もある．例えば，政府による数十年後の到達目標，アナリストによる将来予測，新製品発売に関する情報などは，すべて未来に関する情報である．Webニュースを対象とした予備実験では，全体の4%の記事に未来情報が含まれていた[⋆3]．

情報検索システムは，検索したい対象が曖昧であるが故に「どんなキーワードで検索すればよいか迷う」という問題を常に抱えている[⋆4]．一般的に，未来のイベントには現在において社会的に定着していない事柄を多く含み得る．そのため，検索する対象を未来の情報にすると，現在の情報よりも曖昧性が高まり，キーワードの選定は一般の情報検索システムより難しくなる．ユーザが気軽に未来情報にアクセスするためには，キーワードを入力せずに未来情報を検索できるシステムが求められる．未来情報には時間情報が含まれる傾向があるため，時間情報で整理し，年表形式で提示することによりキーワードの選定を不要にする．

本論文では，未来情報を気軽に知りたいという需要に対応できる未来情報年表を，膨大なテキストデータを用いて，自動的に構築する手法を提案する．さらに，年表を構築するために必要となる要約技術を，新たな研究課題として提示する．

2. 関連研究・サービス

本研究は，博報堂生活総合研究所が提供する「未来年表[⋆5]」から着想を得た.「未来年表」は，新聞記事などから未来予測関連の情報を厳選した，人手による未

†1 筑波大学大学院 システム情報工学研究科
Graduate School of Systems and Information Engineering, University of Tsukuba

⋆1 報知新聞 明治34年1月2日，3日

⋆2 朝日新聞 東京夕刊 昭和30年1月4日〜20日，19日，20日

⋆3 2010年9月に収集した日本語のWebニュース12万件を対象に，未来の西暦年を含む記事の数から推定した．

⋆4 2010年6月に実施したマイクロソフト社による「検索エンジン利用に関するアンケート」によれば，回答者の32.8%が「どんなキーワードで検索すればよいか迷う」と回答している．
http://keyword.jp.msn.com/bing/summary.htm
(accessed 2010-11-26)

⋆5 http://seikatsusoken.jp/futuretimeline/
(accessed 2010-11-26)

図 1 博報堂生活総合研究所が提供する「未来年表」
Fig. 1 "FutureTimeline" which "Hakuhodo Institute of Life and Living" provides.

来情報データベースであり，2010 年 11 月 26 日現在，17,553 件の未来情報が登録されている（**図 1**）．人手によって構築されているため，更新頻度が月 1 回に限られ，典拠となる情報ソースも少ないが，更新頻度を上げたり，典拠となる情報ソースを増やしたりすることは，人的なコストの上昇につながる．本研究では，このような未来情報データベースを自動的に構築したいと考えた．

金田[1] は，百科事典から年表を構築するための年代情報抽出法を提案している．既定の文字列パターンにマッチする文字列を抽出するという単純な手法であるにもかかわらず，非常に高い精度で年代情報を抽出できるという特徴がある．木村ら[2] は，Web から人物の経歴情報を自動収集する手法を提案している．これらの研究は，主に過去の情報に焦点をあてているが，本研究では，未来の情報に焦点をあてる．

未来情報検索の初期の研究として，Baeza-Yates[3] の研究が挙げられる．この研究では，未来情報検索のために，時間情報と起こりうるイベントの信頼度を加味した検索モデルを提案している．

金澤ら[4),5)] は，ユーザの入力したキーワードに関連する将来情報を集約し，グラフを用いて可視化する手法を提案している．彼らの手法の特徴は，Web 検索エンジンを用いることで将来情報が含まれる文書を効率的に収集する点，情報の信頼度にもとづいて将来情報を集約する点である．また，河合ら[6),7)] は，ユーザの入力したキーワードを拡張し，効率的に過去および未来の情報を検索する ChronoSeeker を提案している．彼らの手法の特徴は，ユーザの入力したキーワードに関連する過去および未来の情報を網羅的に収集する点，過去と未来の検索に対しそれぞれ異なるクエリ拡張を行う点である．これらの研究は，キーワード検索を拡張することにより未来情報の検索を行い，出力の際は，文書のタイトルが表示される．本研究では，ユーザは提示される未来情報に対する予備知識を持ち合わせていないものと仮定し，キーワードを入力せずに未来情報を閲覧できるシステムを目指した．

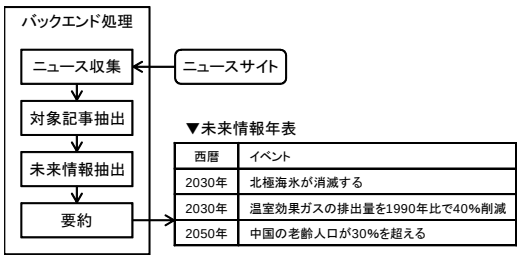

図 2 提案手法の概観
Fig. 2 The overview of the proposed method.

3. 提 案 手 法

提案手法の概観を**図 2** に示す．本システムは，時間情報（西暦）と項目情報（イベント）を対とする「未来情報年表」をユーザに提示する．このシステムを実現するため，Web ニュースを収集し，未来情報の抽出および要約を行う．

本システムの最大の特徴は，「未来情報年表」の提示までに，本システムを使用するユーザを含め，一切の人手を必要としない点である．このため，人的なコストをかけずに「未来情報年表」を構築できる．さらに，ユーザが本システムを使う際，キーワードの選定などの事前準備が不要である．

本システムの提案にあたり，年表を構築するために必要となる要約技術を，新たな研究課題として提示する．年表に提示する項目情報として，従来手法では，ニュース記事のタイトルや検索スニペットを用いていた．これらは，年表の項目情報として適しておらず，新たな要約技術が必要となる．詳細は，3.4 節，5 節で述べる．

以下の小節で，本システムを実現するのに必要な処理の詳細を述べる．

3.1 ニュースの収集

Web には膨大なテキストデータが存在するが，本論文では，未来情報に対する予備知識を持たないユーザを想定し，信頼性が高いと考えられる Web ニュースをもとに未来情報年表の構築を行うことにした．構築のもととなるテキストデータを絞ることにより，提示される未来情報の信頼性を上げる．

ニュース以外の情報源として，政府の文書，企業の株主向け文書が考えられる．これらの文書は，ニュー

スの一次情報になる傾向があり，ニュースの情報と重複する可能性が高い．ニュースの一次情報にならなかったそれらの文書，およびブログなどの情報の活用については，今後の課題とする．

筆者は，日本語のニュースサイトを 300 以上確認しており，これらのニュースサイトのうち，トップページに最新ニュースが表示されるなど，収集が容易である 80 のニュースサイトを監視し，記事を収集している[⋆1]．さらに，ニュース記事にはコンテンツ以外の不要な部分が多々存在するため，吉田ら[8] の手法により，あらかじめ不要な部分を除去している．以下，Web ニュースはコンテンツ抽出処理をした本文部分を指すものとする．

3.2 対象記事抽出

未来情報年表を構築するためには，収集した Web ニュースから，未来情報を含むニュースを抽出する必要がある．未来情報は，明示的な情報，暗示的な情報の 2 つに大分できるが[9]，本論文では，処理を簡略化するために明示的な情報のみを扱う．

明示的な未来情報の抽出は，4 桁の西暦に着目して行った．具体的には，現在よりも未来を示す「YYYY 年」を含む記事を抽出した後，さらにフィルタリングを行う．フィルタリングは簡易的なものであり，金融債権の契約満了日を示す「償還」を含むニュースを未来情報抽出の対象から外した．

3.3 未来情報抽出

未来情報を含むニュースであっても，ニュース全体としては現在のことを指す場合が大半である．以下の引用文は，AFPBB News が報じた『今夏の北極海氷面積、史上 3 番目の小ささ 米研究』と題したニュース[⋆2]のリード文である．記事のタイトルからもわかるが，今年の海氷面積がニュースの主題であり，含まれる未来情報は従属的である．

> この夏の猛暑の影響で、夏季終了時点での北極の海氷面積が観測史上 3 番目の小ささとなったとの研究報告が 15 日、相次いで発表された。地球温暖化によって 2030 年 9 月には北極海氷が消滅する恐れさえあるとしている。

上の例からもわかる通り，未来情報を含むニュースの中から，未来情報が書かれた部分を抽出する必要がある．句点（。），感嘆符（！），疑問符（？）で区切られる文字の連続を 1 文とし，文の中に未来を示す「YYYY 年」が含まれる場合，未来情報が書かれた部分（文）として抽出する．ただし，括弧の中で出現する，句点，感嘆符，疑問符は文の区切りとしない．上の引用文から抽出を行うと以下となる．

> 地球温暖化によって 2030 年 9 月には北極海氷が消滅する恐れさえあるとしている。

3.4 要　　約

未来情報を含む文のままでは，文中に未来の情報と現在の情報が混在したり，年表として表示するには長すぎたりするなど，未来情報年表の項目情報として適さない場合がある．さらに，従来手法では記事のタイトルを用いていたが，3.3 節で述べたように，含まれる未来情報が従属的であることから，未来情報年表の項目情報として適さない場合が多い．

要約処理では，混在する現在の情報を除去したり，適切な長さに圧縮する処理を行う．本論文では，混在する現在の情報を除去することにとどめ，さらに圧縮したり不足した情報を補完したりする処理は，今後の課題とする．

要約処理は，CaboCha[⋆3]を用いて文の係り受けを解析し，その係り受け構造をもとに行う．要約処理は，以下の手順で行う．

（1） 未来を示す「YYYY 年」を含む文節を探す
（2） (1) から係る動詞を含む最初の文節を探す
（3） 動詞にかかる文節を再帰的に探す
（4） (2)(3) を文節順に出力する

(1) の処理では，「YYYY 年」が CaboCha が示す固有表現「DATE」であるかどうかも加味する．固有表現「DATE」でない場合は，未来情報を含まないものとして終了する．(4) の処理の際は，(2) を含む文節は動詞までの出力とし，また，その動詞は基本形に変換して出力する．

3.4.1 要 約 の 例

本節では，以下に引用する中国国際放送局が報じた『中国、2050 年までにタバコによる死者 1 億人の恐れ』と題したニュース[⋆4]から抽出した文を例に，上の要約処理を説明する．引用文の係り受け構造は**図 3** の通りである．

> 中国の民間機構・新探健康発展研究センターの呉宜群副主任は、このほど「タバコの流行

⋆1 収集した記事は Ceek.jp News (http://news.ceek.jp/) を通じて検索することができる．
⋆2 http://afpbb.com/article/environment/2756394/6187958 (accessed 2010-11-26)
⋆3 Cabocha v.0.53, ChaSen v.2.4.4 http://chasen.org/~taku/software/cabocha/ (accessed 2010-11-26)
⋆4 http://japanese.cri.cn/881/2010/09/27/147s164437.htm (accessed 2010-11-26)

がこのまま進めば、中国では、タバコが原因で死亡する人の数は 2050 年までに年間 300 万人になり、合わせて 1 億人に上る恐れがある」と明らかにしました。

まず，(1) の処理により「2050 年までに」の文節がみつかる.「2050 年」は固有表現「DATE」であるため次の処理に進む．(2) の処理により「なり、」の文節がみつかる．そして，(3) の処理を行う．(3) の処理では，まず，(2) の文節に直接係る文節をみつける．直接係る文節は「300 万人」「2050 年までに」「数は」の 3 文節である．さらに，これら 3 文節に直接係る文節をみつけ，という処理を再帰的に繰り返し，(4) の処理に移る．最後の (4) の処理を行うと，以下のように出力される．

タバコが原因で死亡する人の数は 2050 年までに年間 300 万人になる

3.5 未来情報年表

図 2 に示すように,「YYYY 年」を西暦に，3.4 節で得た要約文をイベントに対応させ,「未来情報年表」を作成する．3.3 節，3.4 節で用いた例文を使って構築すると，**表 1** のようになる．

表 1 未来情報年表のサンプル
Table 1 The sample of the “Future Timeline”.

西暦	イベント
2030 年	地球温暖化によって 2030 年 9 月には北極海氷が消滅する恐れさえある
2050 年	タバコが原因で死亡する人の数は 2050 年までに年間 300 万人になる

4. 構築例・考察

本節では，提案手法による未来情報年表の構築例を示す．構築に使用したデータは，2010 年 8 月 1 日から 9 月 30 日にかけて収集した日本語 Web ニュース 248,984 件である．2020 年以降の未来情報年表生成を行ったところ，648 件のイベントが抽出できた．抽出したイベントから 12 件抜粋した結果が**表 3** である（10 件はランダム抽出）．なお，要約文中の角括弧は，未来の時間情報を含む文節を示す.

本論文では，未来情報の抽出に成功しているか，記事のタイトルは未来情報年表に使えるか，項目情報の要約文が適切であるか，要約文中の時間情報を除去しても構わないか，の 4 点を検討した.

まず，未来情報の抽出に成功しているかどうかであるが，いずれも未来情報であるといえる．しかし，(6)〜(8) は議論がわかれる可能性が高い．これらは，「訴える」「目指す」という行為が，現在行われているからである．時間情報が連体修飾語，または複合語の一部になっている場合は，未来情報年表から外すことを検討しており，現在，出現パターンを精査している．

次に，記事のタイトルが未来情報年表の項目情報（イベント）に適すかどうかであるが，抜粋した 12 件の中では (1) と (11) を除き，適さないことがわかる．(1) と (11) は時間情報を含み，未来のイベントを示しているが，これら以外は現在のイベントがタイトルとして現れており，未来情報年表に適していない．

そして，要約文が適切かどうかであるが，過不足のない完璧な要約文は抜粋の中に存在しない．(1)(2)(3)(11)(12) は，完璧に近いものの場所に関する情報が欠けている．これらについては，文中の固有表現「LOCATION」を併記したり，要約文に組み込む必要がある．また，(4)(5) は情報源または条件が付与されており，冗長な要約文になっている．(10) は,「恐れさえある」を除去すると単文として自然な文になる．上の検討とあわせ，未来情報年表を構築するためには，適切な要約文を生成することが重要であることがわかる．この議論は，5 節でも述べる．

最後に，要約文中の時間情報を除去して良いかどうか検討した．時間情報は未来情報年表の「西暦」に記載されるため，要約文にも出現すると冗長である．抜粋した結果をみる限り，時間情報が連体修飾語，または複合語の一部になっている場合を除けば，時間情報を除去しても文が不自然にならない.

5. 今後の課題

提案手法では，時間情報の粒度は年単位しか考慮していない．季節，月，日などより粒度の細かい時間情報を加味することにより，より正確な年表を構築できると考える．また，西暦 4 桁の時間情報のほかに，和暦，省略表記，相対表記などに対応することで，より多くの未来情報が抽出できる．その他，暗示的な未来情報に対する対応も課題として挙げられる．例えば「第 30 回夏季オリンピック」は明示的な時間情報を含まないものの，第 28 回が 2004 年に，第 29 回が 2008 年に実施されたという情報があれば，2012 年に実施されるであろうことが推定可能である.

未来の時間情報を含むものの，未来情報に該当しない情報もある．金融債権に関する情報や，映画のストーリーなどフィクションの情報である．提案手法では，文字列のパターンマッチで金融債権に関する記事のみを除外したが，教師情報を準備し，SVM を用いて判定する手法などが考えられる.

本論文では，Web ニュースを用いて未来情報年表を構築したため，情報源が限られている．構築した未来情報年表を拡張するには，2 節で述べたキーワード検索を拡張する手法を使うことができる．つまり，未来情報年表の項目情報に含まれる，特徴的なキーワードを用いて関連する未来情報を収集し，その項目情報と関連づけることで，より大きな未来情報年表を構築できると考えている．

大きな課題として，年表に適した要約技術の確立が挙げられる．年表に適した要約とは，十分に短い文であり，時間情報と関連付けられ，報知的（informative）な要約である．時間情報との関連付けは，目的に特化しているため focused な要約であるといえる．

要約研究の軸を**表 2** に示す．focused とは目的に特化した要約であり，generic とは特定の目的を想定しない要約である．また，指示的（indicative）とは原文を参照する前に用いられる要約であり，報知的（informative）とは原文の代わりとして用いられる要約である[10]．年表に適した要約は，表 2 の「年表」に該当し，かつ十分に短い文を出力する要約であるが，この条件に該当する要約研究は，未だに行われていない．この要約技術は,「時間情報」を任意のキーワードに置き換えれば，そのキーワードを端的に説明できるようになるなど，応用範囲が広いと考える．

現在，3.4 節の要約結果に対し，山本ら[11] による文末整形手法，平尾ら[12] による文短縮手法を適用し，改良することを検討している．さらに，時間情報に隣接する「○○は」という主題を残すことで，要約結果に不足した情報を補えないか検討している．

表 2 要約技術の軸
Table 2 The axis of the text summarization.

	focused	generic
指示的	検索スニペット	一般的な要約
報知的	年表	字幕

6. おわりに

本論文では，Web ニュースから未来情報を抽出し，未来情報年表を自動的に構築する手法を提案した．実際に未来情報年表を構築し，要約文を用いることの有用性を確認した．

提案手法は，未来情報年表を全自動で構築するため，非常に小さな労力で未来情報を検索するシステムを構築できる．また，キーワードの選定などを必要としないため，ユーザは気軽に未来情報にアクセスできると考えられる．

今後，本研究の要素の一つである，時間情報との関連を保ったまま要約を行う要約技術の改良を行う．また，未来情報を含むテキストデータの抽出性能を向上させ，より大規模な年表構築を行う．そして，本研究の成果を Web サイトで公開し，ユーザのフィードバックを受けながら実用化を進め，集合知を活用したりする未来予測の基盤システムとなることを目指す．

参 考 文 献

1) 金田　泰：百科事典から動的に年表を生成するテキスト検索法のための年代情報の抽出法と表現法, 情報処理学会研究報告, 1999-FI-055, Vol.1999, No.57, pp.81–88 (1999).
2) 木村　塁, 小山　聡, 田中克己：Web からの人物事典生成のための経歴情報の自動収集，日本データベース学会 Letters，Vol. 5, No. 2, pp. 29–32 (2006).
3) Baeza-Yates, R.: Searching the Future, *ACM SIGIR Workshop MF/IR 2005* (2005).
4) 金澤健介，Jatowt, A.，小山　聡，田中克己：Web 上の将来情報の発見と集約的提示，楽天研究開発シンポジウム 2009 論文集，pp. 57–60 (2009).
5) 金澤健介，Jatowt, A.，小山　聡，田中克己：Web 上の将来情報の集約的提示，Web とデータベースに関するフォーラム (WebDB Forum) 2009 (2009).
6) 河合英紀，Jatowt, A.，田中克己，國枝和雄，山田敬嗣：未来戦略立案のための情報検索，楽天研究開発シンポジウム 2009 論文集，pp. 17–20 (2009).
7) 河合英紀，Jatowt, A.，田中克己，國枝和雄，山田敬嗣：ChronoSeeker: Web からの過去・未来情報のオンデマンド検索エンジン，Web とデータベースに関するフォーラム (WebDB Forum) 2009 (2009).
8) 吉田光男，山本幹雄：教師情報を必要としないニュースページ群からのコンテンツ自動抽出，日本データベース学会論文誌，Vol.8, No.1, pp.29–34 (2009).
9) 金澤健介，Jatowt, A.，小山　聡，田中克己：Web からの明示的･暗示的な将来情報の抽出，第 2 回データ工学と情報マネジメントに関するフォーラム (DEIM 2010) (2010).
10) 奥村　学，難波英嗣：テキスト自動要約，オーム社 (2005).
11) 山本和英, 池田諭史, 大橋一輝：「新幹線要約」のための文末の整形，自然言語処理，Vol.12, No.6, pp.85–111 (2006).
12) 平尾　努，鈴木　潤，磯崎英樹：構文情報に依存しない文短縮手法，情報処理学会論文誌データベース（TOD），Vol.2, No.1, pp.1–9 (2009).

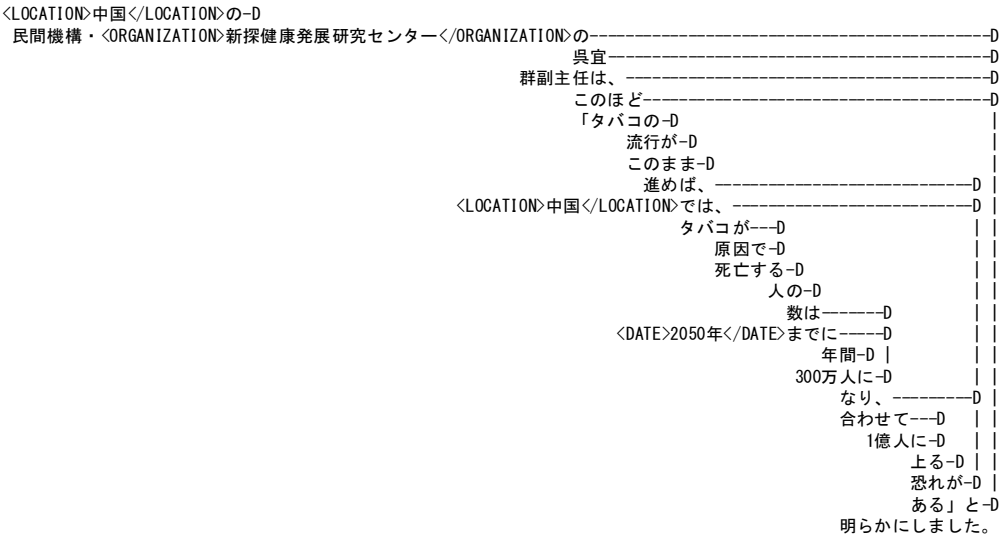

図 3 CaboCha による構文木解析の例.「2050 年に」は「なり、」に直接係り，他から直接係っていないことがわかる．また，提案する手法で要約すると「タバコが原因で死亡する人の数は 2050 年までに年間 300 万人になる」となる．
Fig. 3 An example of the syntax tree analysis by CaboCha.

表 3 自動生成された未来情報年表（12 件を抜粋）
Table 3 “Future Timeline” which was generated automatically.

	西暦	イベント（提案手法による提示） 要約文の基となった Web ニュースのタイトル（従来手法による提示）
(1)	2020 年	[2020 年までに] 世界の自動車市場で 15 %のシェアを確保する 現代自と起亜自、20 年までに世界市場の 15 %シェア確保狙う―中央日報
(2)	2020 年	[2020 年までに] 温室効果ガスを 1990 年比で 25 %削減する 核なき世界へ決意表明＝「最小不幸社会」実現促す―首相国連演説
(3)	2020 年	[2020 年までに] 30 テラワット／時間 (TW/h) の風力発電インフラを構築する スウェーデンがスマートグリッドの先陣を切る スマートメーターの導入率が 100 %を達成
(4)	2020 年	国務院の発表した『北京市都市全体計画（2004-2020）』によりますと、[2020 年までに、] 北京市の総人口は 1800 万人近くに抑える 北京、実際の常駐人口は 2000 万人に
(5)	2020 年	仮に日本の情報サービス業界が中国で 10 %のシェアを獲得できたとすれば、[2020 年には] 中国だけで 13 兆円の売り上げを得られる計算になる 中国ソフトウェア産業協会（CSIA）、情報サービス市場規模で
(6)	2020 年	[2020 年までの] 核廃絶を訴える 秋葉市長にマグサイサイ賞
(7)	2020 年	[2020 年夏季五輪の] 招致を目指す 広島市が五輪計画案公表、8 月 7 日に開会式など 年内に立候補判断
(8)	2022 年	サッカーのワールドカップ（W 杯）[2022 年大会開催を] 目指す エムボマ氏が招致大使に＝サッカー W 杯
(9)	2029 年	次回の接近は [2029 年に] なる 秋分と満月、19 年ぶり同じ夜に
(10)	2030 年	地球温暖化によって [2030 年 9 月には] 北極海氷が消滅する恐れさえある 今夏の北極海氷面積、史上 3 番目の小ささ 米研究
(11)	2050 年	タバコが原因で死亡する人の数は [2050 年までに] 年間 300 万人になる 中国、2050 年までにタバコによる死者 1 億人の恐れ
(12)	2050 年	[2050 年までに] 人口の 31 %が 60 歳以上となる 中国が米国を越えられない 4 つの理由―米メディア

※要約文中の角括弧は，未来の時間情報を含む文節を示す．また，(10)(11) は，3.5 節で示した未来情報の再掲である．