未来イベント予測のための将来言及文における特徴語の調査

中島 陽子 [1]

Investigation of Future Reference Expressions in Newspaper Articles

Yoko NAKAJIMA

Abstract:I have worked on extracting future reference sentences to support future prediction and strategy from a text corpus in Japanese. I apply the Morpho-semantic method which combines the morphological structure and the semantic role of sentences so far and automatically extracts the future reference sentences. Since the Morpho-semantic method does not depend on kinds of words, versatility can be secured, and classification accuracy is about 76%. In this research, I focus on the surface representation of sentences and investigate the characteristics of future reference sentences in order to improve the accuracy of the method of automatically extracting future reference sentences from a text corpus. I aim to improve classification accuracy of future reference sentences by using it in conjunction with word-independent previous methods.

Key words: Information Extraction, NLP, Future Reference Sentences

1 はじめに

近年, 各種データベースや web 上に大量に蓄積されたデータから大規模データ獲得や分析の要求が高まっており, これらのテキストデータを用い統計的手法や自然言語処理技術を用いた機械学習や人工知能の手法により, 各分野における分析や知識獲得手法が提案されている. それらの多くはテキストデータの対象を過去のデータを蓄積した情報を用いて行っている. 例えば, 未来の事象に関する予測問題において, 過去の事象から因果関係を見つけ出し, キーワードに対し次に起き得る事象を予測している [7]. また, インフルエンザに関する SNS の現在の情報からどのくらいの規模で流行しているかまた, 過去のどの程度広がるかなどを予測する研究も行われている [1]. これらの研究は, 線形的な予測を行うためには有効な手法であり, 近年は特に非線形的な事象が発生する傾向にあり, これまでの予測手法では難しい局面が生じている [11].

将来情報ニーズに目を向けると, 潜在的な技術, 変化の予測, 将来の開発と応用, 個人的な計画と将来の動向に関連する傾向がある. Joho [2] らは, 顧客との対話の中で,「将来の関連情報は統計的予測や学術論文から得られる可能性が高いと感じたが, 理解するのは容易ではなかった. そのため, 将来の情報を見つけるためには, 私の理解を深めるための支援メカニズムが必要であると感じた」と述べており, 将来に関連する情報を見つけるのは困難であると言及している. 将来の関連情報の信頼性や信憑性の問題が挙げられており, その解決策として, トピックに関する専門家を見つけ, その出所（e.g. 解説, ブログ, レポートなど）から将来の見解を見つけることがあげられる. また, 未来関連のキーワードからアイディアを得ることも将来イベントへの戦略を立てる一助になることも言及している.

文章に書かれたイベントがいつ実施されるか特定したいという需要は古くからあり, 文章が過去, 現在, 未来のうち, いつ起きるイベントであるか具体的な日時タグ付けする研究 [3], [6] が行われており, 英文の時間タグを特定する Stanford Temporal Tagger:SUTime[1] や TimeML[2] などがある. それらは, 文章中に時間情報（年月日, 来月, 1 年後など）がある場合には有効である. しかし, 実際の将来言及している文には時間情報を持たない文も多くあることが Nakajima ら [4] の調査で明らかにしている.

未来時間表現（e.g. 来年, 明日, 2020 年 etc.）を含まず, 文法的には未来形は用いられていないが, 未来の出来事への言及している文章の例を以下に示す.

1. 彼は, 大統領が対テロ戦争の軸足をアフガニスタンに移す考えを改めて強調したことで, 米軍の早期イラク撤退が現実味を増したと喜んだ.
2. そのうち全体の 4 割に当たる 40 万人が失業すると試算した.
3. 中央銀行の協調姿勢が一層強まり, 円安に動く.
4. C 社が家庭用電源からも充電できるプラグイン・ハイブリッド車を発売する.
5. 大統領は, エジプトへの公式訪問を行う意向である.

文章中に時間表現がある場合は, その記述が書かれた時点を基準に時間表現からイベントが実施される日にちがわかれば, それが未来のことなのかそうでないのかは Main ら [3] や Strötgen ら [6] の手法により判定することができる. しかし, 時間情報の手がかりがなく,

[1]釧路工業高等専門学校　創造工学科

[1]https://nlp.stanford.edu/software/sutime.shtml
[2]http://www.timeml.org/

かつ, 文法的にも未来表現が使われていない場合は判別することは難しい.

日本語文法において, 発話時の時間関係（過去, 現在, 未来）を表す文法カテゴリ:テンスの観点からみると, “ル”で発音が終わるル形動詞を述語とし, かつ動的述語 (e.g. 起こる, 食べる, etc.) の場合それは未来の事象を表している [9] . しかし, 動的述語であっても一般的な出来事を述べる場合 (e.g. 太陽は西に沈む) も含んでおりその区別がつかない場合がある.

1–5 のような文も分類可能にするために, 述語項構造に基づく意味役割と形態論を用いた手法が Nakajima ら [5] によって提案されており, 分類精度 (F 値) は 76%としており, さらなる分類精度の向上が課題とされている.

前述したように, 時間表現や未来へ言及する表層的な語彙はスパースであるが将来イベントを言及する特徴として有用性が確認できれば, 意味役割と形態論と複合的に考慮することで分類精度向上を目指せると考える.

そこで, 本研究では, テキストコーパスから将来へ言及する文の分類精度向上を目指すために, 将来へ言及している文における 3 種類の特徴を見出し, それらの文に用いられる表層的な語彙について検討を行った.

2 章で将来言及文の調査について, 3 章で調査結果と考察, 最後に 4 章ではまとめと展望について述べる.

2 将来言及表現調査

本章では, 将来へ言及している文とそれ以外の文の特徴を比較し, 将来へ言及している文の特徴を定義するための調査について述べる.

将来へ言及している文（以降, 将来言及文）は, 大きく分類すると未来時間情報を含む文, 含まない文に分類できる. 例えば, 以下に示す 1 のように未来時間情報が明記されている文, 2,3,4 のように未来時間情報を含まないが文中で言及しているイベントが現在より先の動向を示唆する語「計画だ」「発売する」などを含む文がある. 4 の文のように文末が過去形になっている場合もある.

1. 2020 年の新車販売台数のうち 50 %はエコカーが占めるようになる.
2. A 社は, 太陽電池を製造する新しい設備を増やす計画だ.
3. C 社が家庭用電源からも充電できるプラグイン・ハイブリッド車を発売する.
4. A 銀行も投票前最後の政策決定会合で, 離脱の意向をあらためて強調した.

将来言及文の表層的にどのような特徴があるか, 新聞コーパス [3]1 年分から 5 分野（経済, 国際, 科学技術, 社会, スポーツ）を対象に 1 文ごとに含まれている語彙について調査を行う. 調査に用いるデータは将来言及文とその他の文の正例と負例になる 2 極を準備する. 新聞コーパスから分野ごとに無作為に抽出し, 将来言及文とその他の文の分類は, 執筆者が複数回の目視によって行った.

コーパス： 毎日新聞 2012 1 年分 (日本語)
分野：経済, 国際, 社会, 科学技術, スポーツ
将来言及文：467 文
その他の文：492 文

新聞コーパスから抽出した各分野の例文（将来へ言及している文とその他）を表 1 に示す.

調査方法は, まず, 各文ごとに未来を表す時間表現を取り出し, 時間表現の種類について調査を行う. 次に名詞と動詞 について形態素解析を行いそれらの出現頻度を求め, 将来言及文とその他の文との差を求める. その差が偶然の差ではないこと ($p < 0.05$) を確認し. それらがどのように使われているか語彙構成や語彙変化, 述語部における用法などに注目しながら検討を行う.

各文の形態素解析を行い, 名詞 (N) と動詞 (V) に該当する語と品詞（語 → 品詞）で表した例を以下に示す. 形態素解析は品詞の単位で分割され品詞種類が解析結果として出力されるが, ここでは名詞および動詞を確認するため, 名詞と動詞以外の品詞 (記号, 助詞など) は除くこととする.

形態素解析の例

文：“政府・与党はそのための基金（ 3 年間）を創設し, 財源を捻出する方針だ. ”

形態素解析：政府 →N / 与党 → N / ため →N / 基金 →N / 3→N / 年間 →N / 創設 →N / し →V / 財源 →N / 捻出 →N / する →V / 方針 →N

調査データ全てに形態素解析を行い, 将来言及文とその他の文における語彙の出現頻度の差が十分である語彙を求め将来言及文にある特徴の検討を行う.

3 調査結果と考察

未来時間表現, 名詞, 動詞について前章 2 で述べた方法で調査を行った. 未来時間表現は, 将来言及文に出現する表現を全て拾い, その他の文に含まれているか確認を行った. 名詞と動詞の頻度を求めた結果については, 未来を示唆する語彙（名詞, 動詞含む）と述語部に出現する動詞に特徴があると仮定し, 未来を示唆する語彙と動詞の種類の観点から検討を行った.

未来時間表現について述べる.

未来時間時間表現の種類は多様で出現頻度が少ないものが多いが将来言及文の約 46%に含まれており, その他の文にはわずか, 0.4%しか含まれていない. 従って, 将来を言及する文の特徴として有用であることは明確である. 未来時間表現の例を以下に示す. x は任意の数を表す.

[3]CD–毎日新聞データ集 2012.1.1–2012.12.31（毎日新聞社）

Table 1: 毎日新聞データ集 2012 から抽出した調査に用いる将来言及文とその他の文例

将来言及文 (例)
政府・与党はそのための基金（３年間）を創設し, 財源を捻出する方針だ.
メキシコと国境を接する州以外にも患者の所在地が広がり, 感染拡大への懸念が高まっている.
敷地を所有する千葉県と協議した上で着工し, 今季開幕前の完成を目指す.
１２月には気候変動枠組み条約第１５回締約国会議（ＣＯＰ１５）が開かれ, 京都議定書後の新たな枠組みが決まる.
航空業界では, 運航費軽減にも貢献するＣＯ削減対策に本気で取り組み始めた.
その他 (例)
２４時間発着がある関空では, 深夜や早朝に手軽に使える宿泊施設を増やすことが課題だった.
１０日の東京株式市場で, 日経平均株価は３営業日ぶりに値下がりして取引を終えた.
英国では立候補段階で党支部の党員投票により候補者が選ばれる.
練習環境が充実した高校だから, 雨の日は室内練習場でピッチングする.
就職希望者はその間に職を探すが, 高齢者や障害者の場合, 仕事も住まいも見つからないケースが多い.

x 年までに/に/から/まで, x 年 x 月までに/に/から/まで, x 年以内に, 今年の, 今月末, x 年度以降, 来年度, 今後も/の, 来年前半, 残る x 年, 今季開幕前の, 引き続き, オリンピック, etc.

具体的な数値を含んだ表現, 時間を表す修飾句, 数値と修飾句の混合, 名詞, 副詞や概知のイベント名が 1 種類または 2 種類以上の組み合わせで構成されている．例えば, “今月末”を形態素解析すると [今月 (名詞-副詞可能)→ 末 (名詞-接尾-副詞可能)] の品詞である．ここで, 品詞情報を表層的な品詞名さらに分類した情報を考慮すると (品詞-品詞細分類 1-品詞細分類 2) と表す.

同様に, [来年度 (名詞-副詞可能)→ 以降 (名詞-副詞可能)], [来年 (名詞-副詞可能)→ から (助詞-格助詞)] や [２ (名詞-数)→ ０ (名詞-数)→ ２ (名詞-数)→ ０ (名詞-数)→ 年 (名詞-接尾-助数詞)] のような品詞で構成されている．“名詞-副詞可能” は，曜日，月など時間を表す副詞的な用法を持つ名詞，“名詞-接尾-助数詞” は, 数に接続して名詞を形成する接尾を意味する．例に見たように, 品詞情報の組合せは未来時間表現を正規化する場合に有用な情報であるかもしれない.

その他の文に出現した未来時間表現の例を以下に示す.

1. “エルバラダイ事務局長の任期は今月末までだが, 週末を挟み, 30 日は IAEA の公休日のため, 27 日の定例理事会が最後の公務となった. ”

2. “将来は, まとまって利益が出た時に保険会社の個人年金に入ろうと思う. ”

1 の文は, “今月末”と “週末”, “30 日”は未来時間であり, “27 日”は過去時間と判断できる．このような場合は主節, 従属節それぞれから情報が得られるので, それぞれで考慮したい．2 の文は, “将来は”が未来時間の可能性はあるが, 述語部で “と思う”の希望を表す文であり, 時間表現だけでは判断できない文である．時間表現は表層的には未来の可能性もあるが, 他の表現と組合わせることが必要であることを示唆していると考える.

次に, 未来を示唆する語彙に特徴があるかを調べるために, 名詞と動詞について出現頻度を求め, 各語彙毎に将来言及文とその他の文を比較した語彙例

Table 2: 名詞と動詞の頻度を求め，各語彙毎に将来言及文とその他の文を比較した語彙例

語彙	頻度（回）		**語彙**	頻度（回）	
	将来言及文	その他		将来言及文	その他
可能性	26	0	期待	10	2
予定	25	1	示し	9	2
目指す	23	2	動き	7	1
方針	17	0	方向	7	1
見込む (め)	12	0	開始	8	2
交渉	18	3	向け (./て/た)	6	1
進め/ん	11	0	懸念	6	1
開（か/き/く）	17	3	必至	4	0
影響	15	2	恐れ	4	0
拡大	9	0	想定	4	0
そうだ	9	0	導入	4	0
見通し	13	2	模様 (./だ)	4	0
計画	8	0	予想	7	2
支援	12	2	広が	7	2
始 (め/ま)	17	6	高ま	7	2

に将来言及文とその他の文を比較を行なった．比較は将来言及文とその他の文の頻度の有意確率 p 値 < 0.05 の語彙について検討を行った．その結果の例を表 2 に示す.

これらの語彙は将来言及文とその他の文との差が十分あることから将来言及文の特徴と判断し, 将来を示唆する**未**来言及示唆語として 30 語を定義する.

日本語文法において，動詞（基本形）のうち，ル形 (“ル”の音で終わる動詞）は未来を表す [10]．そこで, このル形に着目し執筆者が一文ずつ目視を行い, 将来言及文とその他の文で使われれている述語部の動詞 (基本形)（e.g. 作る, 進める, 増えるなど）および名詞＋動詞 (基本形)（e.g. 投入する, 出席する, 販売するなど）について出現頻度の調査を行なった．将来言及文に出現している動詞の例を以下に示す.

求める, 進める, 加える, 図る, なる, 見送る, 増える, 支持する, 再開する, 出席する, 発足する, 販売する, 投入する, 維持する, 獲得する... etc.

その他の文に出現している動詞の例を示す.

Table 3: 将来言及文とその他の文にイベント分野ごとに比較した未来時間表現, 未来言及示唆語, 動詞（基本形）を含む割合 [%].

	経済	国際	スポーツ	科学技術	社会	5 分野
将来言及文						
未来時間表現	51.0	34.9	58.1	36.7	48.0	46.0
未来言及示唆語	48.0	38.4	28.0	25.6	43.9	37.0
動詞（基本形）	88.0	75.6	68.8	77.8	79.6	88.4
その他の文						
未来時間表現	0.0	1.0	0.0	0.0	1.0	0.4
未来言及示唆語	6.3	10.1	0.0	3.1	2.0	4.3
動詞（基本形）	24.0	17.2	8.0	39.8	20.2	21.7

考える, 感じる, 勝る, 伝わる, 由来する, 説明する, こぎつける, ... etc.

動詞の基本形が未来を表すことから，基本形（特にル形）に着目したが，形態素解析の情報のみでは分類は難しい．ル形のテンス（時制）は動的述語の場合は未来を表し静的述語は現在を表すことから，動的述語であるか静的述語であるかを検討する必要があると考える [8].

以上の 3 種類の特徴を将来言及文を表層的に表す特徴と仮定し, 将来言及文とその他の文間で未来時間表現, 未来言及示唆語, 動詞（基本形）の頻度比較を行なった.

未来時間表現, 未来を示唆する語彙, 動詞（基本動詞）を含む割合は将来言及文において, 未来時間表現は約 46%, 未来を示唆する語彙は 37%, 動詞（基本形）は 88%, 一方, その他の文においては, 同順に 0.4%, 4.3%, 20.7%である [表 3]. また, 分野別にもわずかでかるが違いがあり割合が大きいほど 3 種類の特徴で分類しやすいことを示唆しており, 3 種類の特徴の有効性を考慮できる. 未来時間表現の割合を見ると, いずれの分野においても将来言及文の特徴として有用と言える．また, 未来を示唆する語彙においても同様である.

動詞（基本形）については将来言及文に含む割合は多いものの, その他の文にも 8%–24%含んでおり, 動詞の基本形だけの特徴では将来言及していない文も獲得する可能性がある. 動詞（基本形）に条件を追加するなどの工夫が必要と考える．しかし, 動詞（基本形）は将来言及文に約 88%含まれていることは, 重要な手がかりになると考える.

さらに, 将来言及文とその他の文の未来時間表現, 未来言及示唆語, 動詞（基本形）が 1 種類または 2 種類以上の組合せで（混合）含む割合 [%] の比較を行う．表 4 において, 3 種類の特徴が混合で含まれている場合が 62.7%と 1 種類の場合の 31.4%よりも多いことがわかる．また, 将来言及文とその他の文を比較しても混合することで有用性が高まることを示唆している．前述した動詞（基本形）がその他の文に他の特徴と共起する場合は将来言及文と分類される可能性もあるが, この調査ではわずか 1.2%にとどまっている．全体では, 将来言及文の特徴 3 種類で 94.2%を網羅できているが, その他の文でも 20.7%が将来言及文の特徴を含んでいる.

3 種類の特徴が混合している場合の内訳をみると, 将来言及文において動詞 (基本形) が共起している場合は

Table 4: 将来言及文とその他の文の未来時間表現, 未来言及示唆語, 動詞（基本形）が 1 種類または 2 種類以上の組合せで（混合）含む割合 [%] の比較.

			将来言及文	その他の文
A:未来時間表現のみ			2.1	0.4
B:未来言及示唆語のみ			3.6	2.0
C:動詞（基本形）			25.7	19.1
混合			62.7	1.2
	内訳	(A+B+C	10.9	0.0)
		(A+B	1.7	0.0)
		(A+C	29.8	0.0)
		(B+C	20.3	1.2)
全体			94.2	20.7

61%であり無視できない特徴と考えられる. 動詞（基本形）についてさらに分析を進める必要があると考える.

4 まとめと展望

新聞記事 1 年分から将来言及文 467 文とその他の文 492 文を集め, 将来言及文の特徴の調査を行った.

将来言及文において, 表層的な特徴を未来時間表現, 未来言及示唆語, 動詞（基本形）の 3 種類の特徴により将来言及文 467 文の 94.2%を網羅できることを見出した. また, 将来言及文に特徴的に含まれる語彙である 30 語を未来言及示唆語として新たに定義を行った. 3 種類の特徴を用いると将来言及文の 94.2%を網羅できる一方, その他の文が約 20%が将来言及文の特徴が含んでいる．3 種類の特徴は 1 種類よりも 2 種類以上の組み合わせで含まれている場合, 将来言及文とその他の文を分類するためには有用であることも明らかになった.

今後の課題として, 動詞（基本形）が他の特徴と共起している場合で将来言及文の 61%であることから, 将来言及文の述語部分に含まれる動詞について分析を行う予定である.

5 謝辞

本論文は, 在外研究員（独立行政法人国立高等専門学校機構在外研究員制度）として New York University Computer Science Department (平成 28 年 4 月 1 日–9 月 30 日) にて行った研究報告である. 本研究を進めるにあたり, 素晴らしい研究環境を与えて頂くとともに, 多大なるご指導・ご助言を賜りました Prof. S.Sekine に誠意を表わすとともに厚く御礼申し上げます. また, 貴重なご意見を頂戴しました The Proteus Project の教授の皆様, メンバーの皆様に深く感謝申し上げます.

References

[1] Eiji Aramaki, Sachiko Maskawa, Mizuki Morita. 2011. Twitter catches the flu: Detecting influenza epidemics using twitter. *P*roceedings of the Conference on Empirical Methods in Natural Language Processing, pp.1568–1576.

[2] Hideo Joho, Adam Jatowt and Roi Blanco. 2015. Temporal Information Searching Behaviour and Tactics. *I*nformation Processing and Management Journal (IPM), Elsevier, Vol.51(6), pp. 834-850.

[3] Inderjeet Mani and George Wilson. 2000. Processing of News. *P*roceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL2000). pp.69–76.

[4] Yoko Nakajima, Michal Ptaszynski, Hirotoshi Honma and Fumito Masui. 2014. Investigation of Future Reference Expressions in Trend Information. *A*AAI Spring Symposium, SS-14-01, pp.3–38.

[5] Yoko Nakajima, Michal Ptaszynski, Fumito Masui and Hirotoshi Honma. 2016. A Method for Extraction of Future Reference Sentences Based on Semantic Role Labeling. *I*EICE Transactions on Information and Systems, E99-D(2), pp.514–524.

[6] Jannik Str ötgen and Michael Gertz. 2010. HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions. *P*roceedings of the 5th International Workshop on Semantic Evaluation(ACL2010). pp.321–324.

[7] Kira Radinsky, Sagie Davidovich and Shaul Markovitch. 2012. Learning causality for news events prediction. *P*roceedings of the 21st International Conference on World Wide Web, ACM.

[8] 庵功雄, 松岡弘, 中西久美子 その他. 2000. 初級を教える人のための日本語文法ハンドブック. スリーエーネットワーク.

[9] 庵功雄. 2001. 新しい日本語学入門 ことばのしくみを考える. スリーエーネットワーク.

[10] 寺村秀夫.1984. 日本語のシンタクスと意味第 *II* 巻. くろしお出版.

[11] 鷲田祐一. 2016. 未来洞察のための思考法: シナリオによる問題解決 (KDDI 総研叢書). 草書房.