天気予報文作成支援のためのテキスト分析モデルに関する研究

1X15C100-1　松元 琢真
指導教員　後藤 正幸

1　研究背景と目的

新聞記事や TV 報道などに使用される天気予報文は，気象予報士により各メディアのフォーマットに応じて個々に作成される．しかし，気象予報士が 1 つの天気予報文の作成に使える時間は高々数分程度であることから，作業の効率化が重要な課題となっている．天気予報文は，全国の気象台から定刻に発表される数値予報や，各地に対する予報文 (以下，府県概況文) を入力情報として作成される．加えて，天気予報文には「今日の予報」と「明日の予報」が記述されており，当日に天気予報文を作成する際には，前日に作成した天気予報文が参照される．しかし，天気予報は時間経過とともに更新されるため，気象予報士は前日に作成した天気予報文から内容変更が必要と判断される場合には，数値予報や府県概況文から新たに天気予報文を作成する．その一方で，内容変更が必要ない場合には，前日に作成した天気予報文を引用することで当日の天気予報文を作成することができる．ここで引用とは，適当な言い換えを行いながら，同義の文章表現をすることを指す．このような内容変更の有無を自動判別する手法は，天気予報文作成の一助になると考えられる．

そこで本研究では，前日に作成した天気予報文から引用が可能か否かを判別する分類器を構築する．その際，府県概況文にも「今日の予報」と「明日の予報」が記述されていることから，前日に作成した天気予報文の「明日の予報」と，その後に気象庁から発表される府県概況文の「明日の予報」の内容変化を学習することで分類器を構築する．文書の内容変化を捉える際に，一般的な文書分類問題では単語の出現頻度が考慮される．しかし，本研究で扱う天気予報文では，天気の状態を表す際に多様な表現が用いられているため，各単語の出現頻度による分析では対応することができない．このような表現の多様性に対して有効な手法の 1 つにトピックモデル [1] がある．これにより，膨大な単語集合上の頻度分布ではなく，少数のトピック数からなるトピック分布という形で文書を表現することができるため，表現の多様性を考慮して文書の内容変化を捉えることが可能となる．

本研究では，府県概況文と天気予報文の「明日の予報」にトピックモデルを適用し，当日の天気予報文の「今日の予報」に引用可能か否かを判別するモデルを提案する．また，提案モデルの有効性を示すために，実際の府県概況文と天気予報文を用いて分類実験を行い，結果の評価と考察を行う．

2　準備

2.1　対象データ

本研究では，関東甲信地方 1 都 8 県の各気象台から発表される府県概況文，および一般財団法人日本気象協会により作成される天気予報文を分析対象とする．すなわち，天気予報文は，府県概況文等を基に作成される関東甲信地方をまとめた「今日」と「明日」の天気概況の解説文書である．ここで，府県概況文とは気象台ごとに発表される，各地の「今日」と「明日」の天気概況の解説文書である．天気予報文の作成時には，主に前日に作成した天気予報文と気象台から発表された府県概況文の内容や数値予報などの情報が活用される．

2.2　Latent Dirichlet Allocation (LDA)

トピックモデルの代表的な手法の 1 つに LDA がある．LDA は，1 つの文書には複数の潜在的なトピックが混在しており，かつ文書内の各単語は各トピックが持つ単語分布に従って生成されると仮定した確率モデルである．トピックごとに単語分布を仮定することで，似た意味を持つ単語のトピック分布が類似性を持つため，本研究で対象とする文書に出現するような表現の多様性に対応することが可能となる．

3　提案

3.1　概要

天気予報は時間経過とともに変化しているため，前日に作成した「明日の予報」の内容と，その数時間後に発表される府県概況文の「明日の予報」の内容は同じとは限らない．しかし，府県概況文での予報が変化していたとしても，その後作成する当日の天気予報文の「今日の予報」に，前日に作成した「明日の予報」が引用可能か否かは，気象状況により一概には定まらない．したがって，前日に作成した「明日の予報」と数時間後に発表された府県概況文の「明日の予報」の内容変化と，実際に天気予報文へ引用されたか否かの関係性を過去のデータから学習することで，新規に府県概況文が与えられた時に，前日の「明日の予報」の内容を当日の「今日の予報」に引用可能か否かを判別するモデルを構築する．提案モデルのイメージを以下の図 1 に示す．

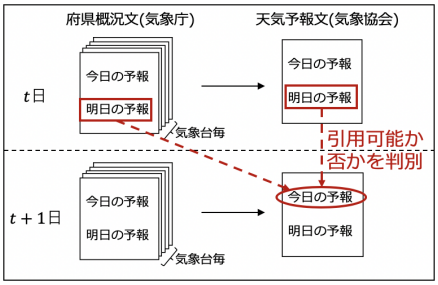

図 1: 提案の概要

このような文書分類を行う際に，一般的には全文書中に出現する単語数を次元とするベクトル空間が構成され，各文書の単語の出現頻度に基づいて文書がベクトル化される．しかし，本研究では府県概況文と天気予報文の性質の異なる 2 つの文書に出現する単語数を次元として各文書をベクトル化するため，文書量の違いから府県概況文と比べて，天気予報文の単語頻度ベクトルはスパースになってしまう．また，府県概況文と天気予報文では,「晴れ」を「さわやかな天気」や「青空が広がる」など，ある気象上の意味を持つ語を多様な表現を用いて言い換えることがある．そのため，単語の出現頻度のみを考慮する分類では，実際には同じ天気であっても出現単語の違いから異なる天気を示す文書と判別されてしまうと考えられる．そこで，府県概況文と天気予報文を用いて分類器を構築する際に LDA を適用することを考える．LDA は単語の共起関係により各文書を，膨大な単語数を次元とする単語頻度ベクトルではなく，少数のトピック数を次元とするトピック分布で表現することができる．そのため，LDA はスパース性や表現の多様性のある府県概況文と天気予報文の特徴ベクトル化に対して有効であると考えられる．

そこで本研究では，まず前日に作成する天気予報文の「明日の予報」と前日に発表される府県概況文の「明日の予報」に対して LDA を適用し，出現する単語の異なる文書同士をトピックを用いて表現する．その後，得られたトピック分布を入力として学習を行うことで，前日に作成した天気予報文の「明日の予報」が当日作成する天気予報文の「今日の予報」に引用可能か否かを自動判別するモデルを提案する．

3.2　提案アルゴリズム

提案する判別モデルの構築アルゴリズムを以下に示す．

STEP1) 学習データの対象期間すべての府県概況文と天気予報文から「明日の予報」に言及する箇所中の名詞・動詞・形容詞を抽出して辞書を作成し，それぞれを日ごとに単語頻度ベクトルに変換する．

STEP2) 府県概況文のトピック数を J_I，天気予報文のトピック数を J_O として，府県概況文と天気予報文それぞれの単語頻度ベクトルを LDA に適用し，トピック分布を推定する．

STEP3) 府県概況文と天気予報文それぞれのトピック分布を日ごとに連結する．

STEP4) 連結したトピック分布を入力として学習を行い，分類器を構築する．

4 評価実験

前日に作成した天気予報文の「明日の予報」と当日に作成する天気予報文の「今日の予報」の間で，内容が変更されていた日を正例，変更されていなかった日を負例として以下の表 1 の基準に基づき，人手によりラベル付けを行った．

表 1: 天気予報文の正解ラベルの付与基準

ラベル	内容
変更なし：負例	1 日を通じて天気予報が合致している
	一部の地域のみ予報が外れているが，1 日を通じて天気予報が合致している
	一部の地域や時間帯のみ記述されており，その内容が合致している
変更あり：正例	変更なし以外

4.1 実験条件

対象データは 2017 年と 2018 年の 1 月 1 日から 10 月 31 日の 608 日間であり，2017 年を学習用，2018 年をテスト用とした．ここで，天気予報文はメディアの要請により文字数が定められており，「今日の予報」に重きを置くために「明日の予報」に言及しない日も存在する．そのため，「明日の予報」に言及していない日を除外した 2017 年の 259 日間，2018 年の 299 日間のデータを用いて分析を行った．

単語頻度ベクトルの次元数は 684，LDA のトピック数を $J_I = 10$，$J_O = \{3, 5, 10\}$ とした．なお，J_O については「明日の予報」に言及した部分の府県概況文の文書長が天気予報文よりも 10 倍程度長いため，J_I と同等かそれ以下で設定した．分類器には Radial Basis Function(RBF) をカーネルとする Support Vector Machine (SVM) [2] を用いた．また，正例のデータ数が負例のデータ数に比べて著しく少ないため，事前にオーバーサンプリングによりデータ数を等しくしてから分類器の学習を行う．また，比較手法は府県概況文と天気予報文それぞれの単語頻度ベクトルを LDA に適用せず，そのまま入力とする SVM とする．さらに，評価指標は再現率，適合率，F 値 [3] と，Area Under the Curve (以下，AUC)[3] を用いた．ここで，AUC は Receiver Operatorating Characteristic curve (ROC 曲線) の曲線下面積により求められ，1 に近いほど高い分類性能を示す．

4.2 実験結果と考察

以下の表 2 に実験結果を示す．

表 2: LDA の有無における F 値，AUC

	LDA なし	LDA あり		
J_O	−	3	5	10
F 値 (最大値)	0.1879	**0.2629**	0.2093	0.2146
AUC	0.5368	**0.6265**	0.5162	0.5107

表 2 より，$J_O = 3$ の場合に，F 値，AUC ともに最大となった．一方で，$J_O = 5, 10$ の場合，LDA を適用しない場合よりも F 値は高くなったが，AUC は低下してしまった．これは，LDA のトピック数を大きくしたため学習データが不足し，適切な文書表現が得られなかったためと考えられる．また，$J_I = 10, J_O = 3$ のように，府県概況文と天気予報文のトピック数の差を大きくしたときに分類性能が高くなったことは，府県概況文の「明日の予報」は文章が長く単語数も多いことが理由として考えられる．そのため，天気予報文のトピック数が小さい方が，単語数の違いとその表現の多様さを適切に学習できていると推測される．実際にトピック内に出現する単語を分析すると，天気予報文のトピック数を大きくした場合に，ほとんどの単語の出現確率が同じとなってしまう現象や，同じトピックにおける単語分布で「晴れ」「雨」など異なる天候を表現する単語のトピック分布に違いが生じない事例も多く見られた．これらは，LDA が「複数文書で共起しやすい単語は同一のトピックに含まれやすい」という仮定を置いていることに起因していると考えられる．その一方で，天気予報文のトピック数を小さくした場合では「晴れ」や「雨」の状態を表すようなトピックが得られた．このため，トピック数を小さくすることで適切に分類が可能になったと考えられる．

ここで，$J_O = 3$ の場合の再現率，適合率，F 値を以下に示す．

表 3: $J_O = 3$ の場合の再現率，適合率，F 値

適合率	再現率	F 値
0.1538	0.9032	0.2629

表 3 より，適合率が約 15%，再現率は約 90%であることから，本研究の結果では多くの日で「変更あり」と出力していることがわかる．ここで，天気予報文作成においては，引用すべきか否かを正しく分類することよりも前日の天気予報文の内容を引用できない日をもれなく列挙することの方が重要である．すなわち，前日から内容を変更すべき日をできるだけ多く列挙できているかを示す指標である再現率を重視すべきであると考えられる．このことから，再現率が約 90%を示した提案モデルは有効であると言える．また，本研究において検出することのできなかった約 10%に関する対応は，今後の課題である．

5 まとめと今後の課題

本研究では，天気予報文作成支援を目的とし，前日の「明日の予報」が当日の天気予報の「今日の予報」に引用可能か否かを判別するモデルを提案した．この手法では，府県概況文と天気予報文に LDA を適用し，トピック分布を活用することにより，単語頻度のスパース性や表現の多様性を考慮した判別が可能となった．また，提案手法を実際の天気予報データに適用することで，提案手法の有効性を示した．

今後の課題として，辞書作成時に係り受け解析や単語 N-gram 頻度を用いることや，府県概況文と天気予報文の「今日の予報」，「明日の予報」に言及している文の自動抽出，また，「変更あり」と分類された場合に，天気予報文作成に有効となる候補文リストの自動抽出などが挙げられる．

参考文献

[1] Blei M.D., Ng Y.A., Jordan I.M.,“Latent Dirichlet Allocation,”*Journal of Machine Learning Research,* Vol3, pp.993–1022, 2003.

[2] 高村大也, 松本裕治,“SVM を用いた文書分類と構成的機能学習”, 情報処理学会論文誌, Vol. 44 No. SIG3, 2003.

[3] Tom Fawcett,“An introduction to ROC analysis,”*Pattern Recognition Letters,* Vol27, Issue 8, pp.861–874, 2006.