過去の出来事に関する短文の分類

澄川　靖信† Jatowt Adam††

† 東京理科大学理工学部情報科学科　〒 278–8510 千葉県野田市山崎 2641
†† 京都大学大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
E-mail: †yas@cs.is.noda.tus.ac.jp, ††adam@dl.kuis.kyoto-u.ac.jp

あらまし　過去に生じた出来事は，歴史の教科書，新聞記事，年表といったあらゆる文章で参照される．このような形式で参照される出来事の記述は簡潔にまとめられている傾向がある．これまでに行われてきた出来事の分類に関する研究は，新聞記事が主題として報じているときのような，十分な記述量がある文章を対象としていた．本研究では，少数の文で記述された文章でも高い精度を得られるように，各文章に対して 9 種類の特徴ベクトルを定義し，それらを基に分類器を訓練することによって先行研究よりも高い精度で分類できることを示す．

キーワード　歴史，出来事，文章分類，短文，特徴抽出

1. はじめに

過去に起きた出来事は簡潔な記述によって文章中で言及されていることがある．例えば，地震が起きたことを報じる新聞記事が，これから取るべき対策について過去に生じた地震の結果を踏まえて議論することがある．他にも，歴史上の都市に関する文章中では，過去にその場所でどのような出来事が起きたのかを述べることがある．なお，このような過去の出来事は，現上記のような長文の中で言及されるだけでなく，Wikipedia の Current Portal(注1) のような時系列に列挙されることや，関連する出来事をまとめてリスト化されている場合もある．表 1 に Wikipedia の Current Portal に記述されている出来事の例を示す．

本研究では，出来事に関する短文を分類するための効果的な特徴抽出を提案する．これまでに行われた出来事分類の研究は，新聞記事のような十分な記述量のある文章を対象としている [4]．しかし，上述したように，副次的に述べられる場合や時系列に列挙される出来事はわずかな分量で簡潔に記述されている傾向があるので，曖昧性解析や意味解析のみで良い結果を得ることが困難である．また，出来事は必ずしも名前が付与されているとは限らないので(注2)，固有表現抽出（named entity recognition, NER）が有効に働く状況は限定される．

以上の問題を解決するために，本研究ではトピック解析や潜在意味解析，Wikipedia や VerbNet(注3) といった外部の情報源を用いて特徴量を増やす．ここで，本研究では列挙された文章も対象としているので，外部リンクや出来事に関する記述の前後にある段落や文章といった文脈情報は利用できないものと仮定する．すなわち，本手法は出来事に関する記述のみから上記の字句解析や意味解析を適用して得られた特徴ベクトルを用いて分類器を訓練する．

本手法の効果を確認するために，Wikipedia の Current Portal に記録されている出来事の中でカテゴリが付与されている 32,618 個のデータだけを用いたところ，各特徴ベクトルだけで訓練した SVM と短文に特化した先行研究よりも，本手法の方が高い精度が得られ，F 値は約 80%であった．さらに，Current Portal の出来事よりも前に起きた過去の出来事に対して本手法の効果を確かめるために，上記のカテゴリが付与されている 1500 年から 1999 年までの出来事を収集し，こちらのデータセットにおいても提案手法を用いて訓練した SVM が本稿で用いた分類器の中で最も良く，F 値が約 85%であることを確認した．

出来事の短文を高い精度で分類できることによって，以下のような研究を発展させることができると考えられる．

（1） アジアで起きた災害や事故のリストなどの歴史的な出来事の潜在的なカテゴリに基づいた時系列リストの作成や構造化といった分類研究．

（2） 出来事の重要性とその潜在的なカテゴリに基づいた未来予測の研究．

（3） どのような種類の出来事が新聞記事中でよく参照され

表 1　全てのクラスの出来事例．各クラスの名称は以下のように省略している．Armed Conflicts & Attacks (AA), Arts & Culture (AC), Business & Economy (BE), Disasters & Accidents (DA), Health & Environment (HE), Law & Crime (LC), Politics & Elections (PE), Science & Technology (ST), Sport (S).

クラス	例文
AA	Bombs across Iraq detonate, killing 18 people.
AC	The Beatles release their back catalogue on iTunes.
BE	Brazil's economy falls into recession.
DA	A bus crashes into a ravine in Tibet, killing at least 44 people.
HE	The number of Zika virus infected in Singapore rises above 40.
LC	The Constitutional Council of France upholds a ban on fracking.
PE	Voters in Costa Rica go to the polls for a general election.
ST	Iran successfully puts the Fajr satellite in orbit using a Safir-B1 rocket.
S	The Winter Olympics in Sochi, Russia officially concludes.

(注1)：https://en.wikipedia.org/wiki/Portal:Current_events
(注2)：一般的に，有名なイベントや歴史的に重要なイベントだけが固有の名前を付与されている．
(注3)：https://verbs.colorado.edu/~mpalmer/projects/verbnet.html

るのかの分析研究.

（4） 出来事の種類を節とし関連する出来事間を辺で表すグラフを用いた，関連する出来事の集合を基にした出来事の類似性検出に関する研究.

本稿の貢献をまとめると，次の通りである.

（1） 歴史的な出来事に関する短文を分類するために有効な特徴を明らかにする.

（2） 複数の特徴ベクトルを組み合わせて分類器を訓練することが分類精度を高めることを明らかにする.

（3） カテゴリが付与されていない出来事への自動的な分類や手動で行う際の支援として本手法を用いることができる.

2. 関 連 研 究

これまでに行われた短文分類に関する研究は，文脈情報を用いるものと外部の情報源を用いるものの 2 種類に分けられる.

前者の研究としては，ツイートの分類のために Twitter ユーザのプロフィールやツイートに含まれる URL とハッシュタグを用いた手法 [10]，Q&A 文章のために画像や動画を利用した単純ベイズ分類器を利用したもの [8]，Web 検索におけるユーザの目的を明らかにするために，各ユーザがクリックした情報を利用したクエリ分類 [6] といった研究が行われている.

一方，外部の情報源を用いた後者の研究として，Wikipedia 上でモデルを構築した LSA を適用した特徴ベクトルを基に分類器を訓練した手法 [12]，LSA [2] や LDA [1] といった潜在意味解析を用いた分類器の訓練法を一般化した枠組みの提案 [9] が行われている．また，明示的意味解析（Explicit Semantic Analysis, ESA）を用いて短文に対応する Wikipedia 記事を取得し，その文章を基に分類器を訓練する手法も提案されている [11].

本研究では，過去の出来事に関する短文を分類するために，固有表現や動作といった特徴の中でどれが本目的に対して有効なのかを調査する.

3. データの収集と分析

3.1 ク ラ ス

本稿で用いるクラスは, [4] でも利用されている，Wikipedia 編集者とガイドラインによって定義された 9 個のクラス `Armed Conflicts & Attacks` (`AA`), `Arts & Culture` (`AC`), `Business & Economy` (`BE`), `Disasters & Accidents` (`DA`), `Health & Environment` (`HE`), `Law & Crime` (`LC`), `Politics & Elections` (`PE`), `Science & Technology` (`ST`), `Sport` (`S`) を用いる．なお，[4] ではこれらのクラスの出来事に分類されている新聞記事から TF-IDF ベクトルを作成し，SVM の精度を調査している.

3.2 データセット

本稿では短期間データセットと長期間データセットの 2 種類のデータセットを集めた．短期間データセットは Wikipedia の Current Events ポータルに記録されている，2010/1/1 ~ 2016/12/17 の間に生じた 32,618 個のラベル付きデータを収集して作成した．このデータセットでは，各文章は平均的に約 25 単語で記述され，短いものでは 10 単語だけで記述されている．各クラスでの平均単語長を表 2 に示す.

表 2 短期間・長期間データセットに含まれる各クラスの出来事の平均的な単語数と出来事数

	短期間データセット		長期間データセット	
クラス	出来事数	単語数	出来事数	単語数
`AA`	8,886	23.6	-	-
`AC`	1,800	22.9	683	20.0
`BE`	2,517	23.6	606	38.8
`DA`	4,961	23.1	-	-
`HE`	487	28.7	204	25.9
`LC`	4,984	27.5	1,623	28.7
`PE`	5,517	25.2	319	20.2
`ST`	1,066	24.6	7,253	16.3
`S`	2,400	23.3	9,654	19.0
Total	32,618		20,342	

長期間データセットは，短期間データセットと同じクラスの出来事が別々の記事としてまとめられている(注4)．本稿では 1500 年から 1999 年までの出来事 20,342 個をラベル付きデータとして収集した．収集した各文章は短期間データセットと同様に 1~2 文で各出来事を記述している．なお，長期間データセットには `AA` と `DA` のクラスの出来事が含まれていない．これらのクラスの出来事を収集できるように，本手法を各年の記事(注5)に記録されている出来事に対して本手法を適用して，ラベル付きデータ数を増やすことが今後の重要な課題である．さらに，短期間データセットの出来事は約 7 年分，長期間データセットには 500 年分の出来事が含まれるが，それぞれのデータセットに含まれる出来事数は，約 3 万個と約 2 万個であった．前述した Wikipedia の各年で生じた出来事をまとめた記事には AD 1 年から 1999 年までの 2000 年分の 38,625 個の出来事が記録されているので，これらの出来事に対してラベルを付与することも重要な今後の課題である.

4. 特徴ベクトルの作成

4.1 特 徴 抽 出

本節で，本研究が用いる特徴について述べる.

4.1.1 語彙に基づく特徴

本研究では，まず，単語に基づいた解析を行うために全ての文章に対して TF-IDF ベクトル (F_1) を作成する.

4.1.2 潜在的な意味に基づく特徴

次に，潜在的な意味を解析するために，Doc2Vec [5] (F_2) と LSA (F_3) を適用した特徴ベクトルを作成する.

4.1.3 動詞に基づく特徴

出来事の記述において，動詞はその出来事で何が起きたのかを表す重要な役割を持つ．VerbNet は， `destroy` クラスは `demolish` や `ruin` といった動詞を含むように，意味が類似している動詞をまとめたクラスを定義しているので，本手法は各文章に含まれる動詞を Stanford POS タガーを用いて取得し，

(注4)：以下に収集した記事のタイトルを列挙する．なお，`{year}` は各年の数字を表す．`{year}_in_art`, `{year}_world_oil_market_chronology`, `{year}_in_the_environment`, `{year}_in_organized_crime`, `{year}_in_politics`, `{year}_in_science`, `{year}_in_sports`

(注5)：例えば `https://en.wikipedia.org/wiki/1900`

それら全てに対して VerbNet の動詞クラスを取得した後，各クラスの出現数に基づいた特徴ベクトルを作成する (F_4)．なお，VerbNet には 429 個のクラスが定義されている．

4.1.4 固有表現に基づく特徴

多くの出来事の記述では，人や組織，場所といった，誰がその出来事を起こしたのか，どこで生じたのか，といったことを示すために固有表現を含む．このような固有表現の種類は，会社名は BE の出来事に出現しやすい，のように，出来事のクラスに強く関連していると考えられる．同様に，ある特定の種類の固有表現が無いことも出来事クラスとの関連が無い傾向があることを示す．例えば，いかなる場所も明記されていない場合，その出来事は DA では無いと考えることができる．さらに，異なる種類の固有表現の組み合わせは異なる出来事クラスの組み合わせを示唆しうる．そこで，本研究では，全ての固有表現をそれらの種類で一般化し，その数を基にした特徴ベクトル (F_5) を作成する．本稿では，固有表現とその種類の取得のために Yodie [3] を用いた．

4.1.5 先頭の固有表現と動詞に基づく特徴

短文で出来事を記述する際，その出来事を生じた人や組織の名前が，その文の最初に出現する傾向がある．そこで，本研究では，最初に出現した固有表現の種類だけを用いて特徴ベクトル (F_6) を作成する．また，同様に，最初に出現する動詞もその出来事において特に重要なことが多く，先頭の固有表現の潜在的な表すものとみなして，その動詞に対する VerbNet のクラスを用いた特徴ベクトル (F_7) を作成する．

4.1.6 概念に基づく特徴

本研究では Wikipedia を用いて類似する出来事と潜在的な概念を検出する．まず，各出来事の文章に対して ESA [7] を適用して対応する Wikipedia の記事をランキング形式で取得する．なお，出力された Wikipedia 記事の多くは，入力された出来事に類似する過去の出来事に関するものである．次に，出力された記事の上位 10 件からその記事に付与されている Wikipedia カテゴリの単語の特徴ベクトル (F_8) とそのタイトルの単語の特徴ベクトル (F_9) を作成する．これら 2 つの特徴ベクトルの値は TF-IDF によって重みづけしている．

4.2 特徴選択

以上のすべての特徴ベクトルを 1 つのベクトルとして結合し，特徴選択を適用して次元削除を行う．本稿では短期間データセットに対してはランダムフォレスト(注6)に基づく特徴選択を適用し，重要度の高い上位 2,000(注7)個の特徴を採用した．一方，長期間データセットに対しては L1 ベースの特徴選択を適用した．

5. 実　　験

5.1 準　　備

データコレクション．本稿で構築したデータセットのために収集した固有表現，Wikipedia 概念，Wikipedia カテゴリの数を表 3 に示す．

表 3　短期間・長期間データセットそれぞれで収集した固有表現，Wikipedia 概念，Wikipedia カテゴリの数と，それぞれの一つの出来事に付与されている平均数．

	固有表現	平均	Wiki. 概念	平均	Wiki. カテゴリ	平均
短期間	17,503	2.5	72,540	10	116,809	22.5
長期間	118,798	5.8	20,342	10	464,257	22.8

パラメータ．LSA と Doc2Vec の次元数はそれぞれ 300 とした．この数字は，100 から 10,000 までの 100 刻みでそれぞれの特徴ベクトルを作成し，それらを用いて SVM を訓練して最も良い F 値のものであった．

比較手法．本手法の精度を確認するために，以下の手法の精度を評価する．

（1） TF-IDF+SVM: TF-IDF で重みづけされた BOW ベクトル上で訓練した SVM [4]

（2） MaxEnt: 短文分類器として広く利用されている，潜在トピック解析を用いた特徴ベクトルを用いて訓練した MaxEnt 分類器 [9]

（3） All+NB: 第 4. 節の特徴ベクトルを用いて訓練した単純ベイズ分類器（NB）

（4） All+RFs: 第 4. 節の特徴ベクトルを用いて訓練したランダムフォレスト（RFs）

（5） All+SVM-rbf: 第 4. 節の特徴ベクトルと RBF カーネルを用いて訓練した SVM

（6） All+SVM-linear: 第 4. 節の特徴ベクトルと線形カーネルを用いて訓練した SVM

さらに，本稿で作成した各特徴ベクトルだけを用いて SVM を訓練し，すべての特徴ベクトルを組み合わせて特徴選択を適用することが精度を向上することを示す．これらの手法の精度は，分割数を 10 とした交差検定と One-vs-All による多項分類で確認した．

5.2 短期間データセットの結果

表 4 に全ての分類器の F 値を示す．全ての特徴ベクトルを組み合わせて特徴選択を適用して得られたベクトル上で訓練した SVM が，AC と HE を除いたクラスと全クラスの中で最も良い結果であった．特に RBF カーネルで訓練した SVM は 5 つのクラス（BE, DA, PE, ST, S）と全てのクラスの F 値の平均で最も良い結果であった．図 1 に 2 つの先行研究の手法と，本稿で作成した特徴ベクトルで訓練した単純ベイズ分類器・RFs・SVM のマイクロ平均の ROC 曲線を示す．この結果においても SVM-rbf の結果が最も良いことを示すので，以降は SVM-rbf の結果だけ詳細に分析する．

次に，各特徴ベクトルだけで SVM-rbf を訓練したときの F 値（表 4）について分析する．すべてのクラスにおいて，各特徴ベクトルの結果よりも全てを組み合わせたものの方が良い結果であった．特に AC, BE, ST の 3 つのクラスでは最も良い結果であった単一のものよりも 10%以上も向上している．

次に提案手法の各クラスでの適合率，再現率，F 値の結果を図 2 に示す．全体的に 80%程度の高い精度が得られた．しかし

(注6)：http://scikit-learn.org/stable/index.html

(注7)：この値は本稿のデータセットから少数のデータを使用して幾つかの値の中から最も良い結果であった

表 4　短期間データセットの各特徴ベクトルだけで訓練した SVM の F 値，先行研究の F 値，全特徴ベクトルを利用して訓練した SVM，NB，RF の F 値.

	各特徴ベクトルで訓練した SVM										提案手法			
クラス	F_1 ([4])	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	[9]	All+NB	All+RFs	All+SVM-rbf	All+SVM-linear
AA	68.2%	10.1%	83.8%	52.5%	28.5%	23.7%	0.0%	69.6%	28.6%	52.3%	50.0%	46.4%	85.3%	**85.8%**
AC	21.6%	9.1%	41.4%	13.0%	6.2%	0.0%	0.0%	44.8%	7.3%	**79.9%**	70.0%	68.4%	59.7%	65.1%
BE	36.5%	3.7%	66.2%	21.8%	7.6%	0.0%	0.0%	59.8%	1.4%	73.3%	61.9%	58.2%	**75.5%**	74.4%
DA	65.5%	22.0%	83.8%	37.5%	1.8%	30.3%	0.0%	68.3%	9.1%	84.4%	64.7%	60.1%	**88.4%**	86.5%
HE	42.9%	4.4%	54.5%	8.2%	2.2%	3.8%	0.0%	25.9%	3.3%	**89.3%**	72.2%	66.5%	54.0%	64.2%
LC	42.0%	11.4%	68.3%	33.2%	14.5%	0.0%	0.0%	46.8%	10.2%	65.2%	49.0%	44.4%	72.0%	**72.6%**
PE	52.7%	15.0%	72.6%	39.0%	20.3%	0.0%	0.0%	61.9%	9.7%	65.5%	58.6%	50.0%	**77.6%**	75.1 %
ST	31.3%	6.8%	58.6%	8.9%	5.0%	0.0%	0.0%	63.6%	0.0%	8.3%	43.0%	44.9%	**71.8%**	66.7%
S	66.7%	8.2%	85.0%	36.5%	14.4%	0.0%	13.7%	81.8%	10.5%	57.3%	50.3%	52.0%	**89.3%**	88.9%
Total	*54.5%*	*27.2%*	*74.7%*	*40.9%*	*28.5%*	*11.6%*	*1.0%*	*62.6%*	*46.0%*	*64.0%*	*58.3%*	*54.6%*	***79.7%***	75.5 %

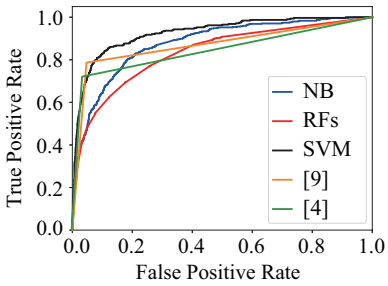

図 1　短期間データセット ROC 曲線

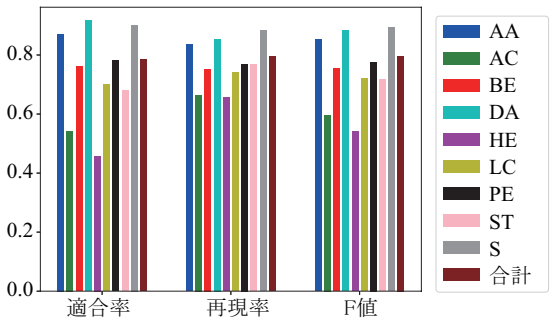

図 2　短期間データセットで訓練した提案手法の SVM-rbf の適合率，再現率，F 値.

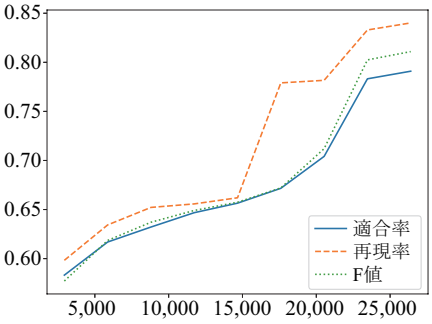

図 3　短期間データセットの訓練データ数を変更したときの提案手法の SVM-rbf の適合率，再現率，F 値.

HE と AC の 2 つのクラスでは比較的精度が低い傾向がある．これは，表 1 に示したように，これらのクラスのデータ数が比較的少ないことが原因だと考えられる．図 3 は訓練データ数を変更した際の適合率，再現率，F 値の推移を示す．全ての結果はデータ数の増加と共に精度が向上している．特に，再現率に関

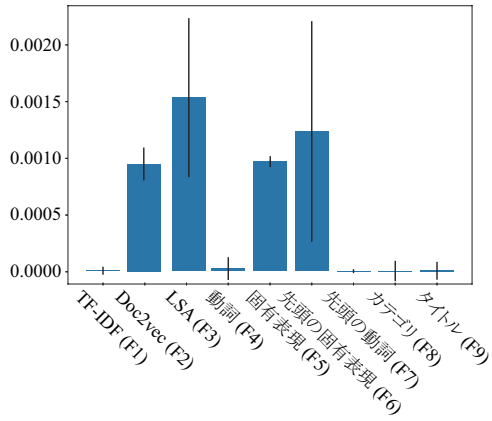

図 4　短期間データセットの各特徴の重要度.

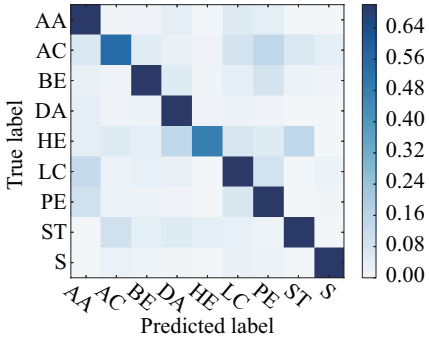

図 5　短期間データセットからランダムにサンプリングした 3,260 個のデータに関するコンヒュージョンマトリックス.

してはデータ数が 15,000~18,000 のとき，適合率と F 値に関してはデータ数が 21,000~23,000 のときに，75~80%に達することがわかる.

本稿のデータセットの各特徴の重要度を図 4 に示す．青線は各重要度の値の平均値を示し，黒線は標準偏差を示す．この結果から，LSA と固有表現（特に先頭のもの）が短文分類器にとって重要であることがわかる．一方，TF-IDF，動詞，Wikipedia 記事のタイトルとカテゴリの文章はあまり重要ではないことがわかる.

最後に，図 5 に分類器が誤分類した結果を示す．AC は PE に誤分類される傾向があり，一方 LC と PE は AA と誤って分類される傾向がある．これらの誤りの理由として，これらのカテゴリでは，出来事を起こしたものが国や国籍と共に言及される人

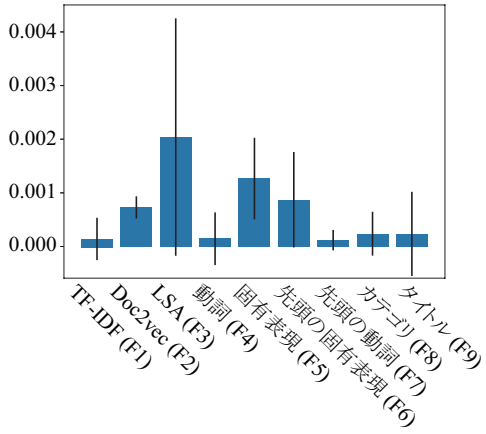

図 6　長期間データセットの各特徴の重要度.

表 5　先行研究の F 値，全特徴ベクトルを利用して訓練した SVM，NB，RF の F 値.

クラス	[9]	All+NB	All+RFs	All+SVM-rbf	All+SVM-linear
AC	93.0%	**97.4%**	97.2%	96.2%	80.0%
BE	8.7%	56.9%	55.7%	59.5%	**97.8%**
HE	23.5%	94.9%	93.2%	94.0%	59.3%
LC	4.9%	63.6%	47.1%	50.0%	**96.3%**
PE	96.2%	**97.3%**	95.9%	96.4%	72.0%
ST	87.3%	93.3%	94.3%	88.7%	**94.9%**
S	81.5%	48.0%	70.4%	85.3%	**94.3%**
Total	*56.4%*	*78.8%*	*79.1%*	*81.4%*	***84.9%***

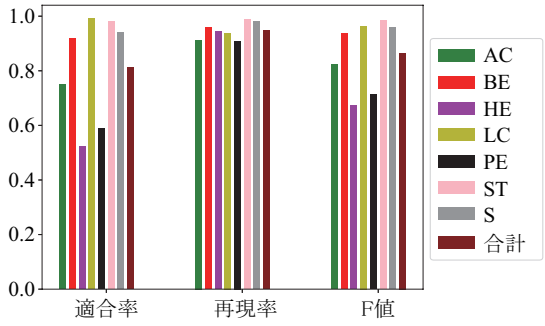

図 7　長期間データセットの訓練データ数を変更したときの提案手法の SVM の適合率，再現率，F 値.

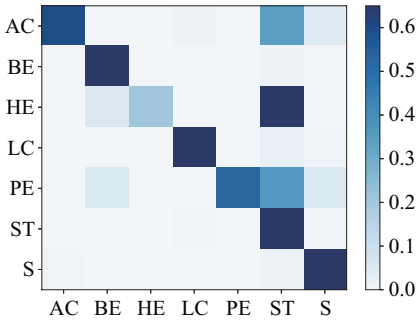

図 8　長期間データセットからランダムにサンプリングした 3,260 個のデータに関するコンヒュージョンマトリックス.

（例えば，日本人の科学者）であることが多いことが考えられる．HE の出来事は ST に誤分類されることがある．これはこれらのカテゴリでは共に薬や生物といった単語が出現する傾向があるのが理由だと考えられる．また，HE と DA の出来事は同一の他の出来事がきっかけとなって生じることがある．例えば，2016 年に生じたジカウィルスは，多くの死者を生じ（HE の出来事），また蜂の数を減少させた（DA の出来事）.

5.3　長期間データセットの結果

次に長期間データセットに関する議論を行う．まず，図 6 に各特徴の重要度を示す．この結果は短期間データセットでの結果（図 4）と同様の傾向が確認できる．すなわち，LSA と固有表現が特に重要である．標準偏差に着目すると，短期間データセットでも値が大きい傾向があった LSA (F_3) と先頭の固有表現 (F_7) は長期間データセットでも同様の傾向がある．さらに，TF-IDF (F_1), 動詞 (F_4), 固有表現 (F_5), カテゴリ (F_8), タイトル (F_9) に関しては長期間データセットでは高い値が確認できる．これは，長期間データセットには 1500 年から 1999 年までの 500 年間文の出来事が記録されていることから，多様な種類の単語が含まれているのが理由として考えられる.

表 5 に，短文向け分類器の先行研究と提案手法の F 値を示す．長期間データセットにおいても SVM の結果が AC と PE を除いた全クラスで良い結果であった．特に，線形カーネルを用いた SVM は 4 つのクラス（BE，LC，ST，S）と全てのクラスの平均値で最も良い結果であった．したがって，本節の以降の議論では線形カーネルを用いた SVM の結果だけを示す.

図 7 に提案手法の各クラスの適合率，再現率，F 値を示す．適合率は HE と PE を除いた 6 クラスが約 80%程度の結果であった．同様に，これら 2 つのクラスを除いた 6 クラスの F 値は約 80%程度である．一方，再現率は 8 クラスすべてにおいて約 90%程度であった．図 8 に分類器が誤分類した結果を示す．HE クラスの出来事は，短期間データセットと同様に ST クラスに誤分類される傾向があり，約 74%の HE 出来事が ST と誤分類されている．AC クラスと PE クラスの出来事も共に約 35%が ST クラスに誤分類されている．誤分類が特に多い HE クラスは，本稿で用いたデータ数が約 200 個であり，一つの出来事を平均して約 26 単語で記述されていることが適合率の結果と誤分類の理由として考えられる．長期間データセットでは，PE クラスでは過去に生じた争いの際に開発された道具に関するものが記録されている(注8)．このような出来事は ST クラスでも存在するので，このような PE クラスの出来事が誤分類されたと考えられる.

6.　ま　と　め

過去に起きた出来事を簡潔な記述と共に参照することはよく行われる．このような出来事をカテゴリ化することによって，歴史と現代の類似する出来事や固有表現の接続，歴史的類推を促すためのモデルの構築，Wikipedia 記事のような歴史的な出来事に関する記事の構造化，などの発展的な研究の基礎となると考えられる．本研究では 2~3 文程度で記述された歴史的な出来事を約 6 万 3 千個収集し，それらを効果的に分類するための特徴の選択とその精度について報告した．上記の発展的な研究

（注8）：例えば 1968 年の June 30？The Lockheed C-5 Galaxy heavy military transport aircraft first flies in the U.S. This model will still be in service 40 years later.

に加えて，今後の課題として，長期間データセットをより細かい粒度に分解し，それぞれのサブセットでの分析を行い，クラスごとの出来事分布の分析，同一クラス内で分類精度が向上する年の閾値の検討，固有表現とクラスの年代別の相関値の分析，などが考えられる．

謝辞 本研究の一部は科研 (17H01828, 17K12792) および戦略的情報通信研究開発推進事業（171507010）の助成を受けたものである．

文　　献

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[2] S. Deerwester, S. T. Dumais, G. W. Furnas, L. Thomas K., and R. Harshman. Indexing by latent semantic analysis. *J. Amer. Soc. Inform. Sci.*, 41(6):391–407, 1990.

[3] G. Gorrell, J. Petrak, and K. Bontcheva. Using @twitter conventions to improve #lod-based named entity disambiguation. pages 171–186, Springer-Verlag New York, Inc., New York, NY, USA, 2015. ESWC'15.

[4] A. Košmerlj, E. Belyaeva, G. Leban, M. Grobelnik, and B. Fortuna. Towards a complete event type taxonomy. pages 899–902, ACM, New York, NY, USA, 2015. WWW '15 Companion.

[5] Q. Le and T. Mikolov. Distributed representations of sentences and documents. volume 32, pages 1188–1196, Bejing, China, 22–24 Jun 2014. ICML'14.

[6] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. WWW '05, pages 391–400, New York, NY, USA, 2005. ACM.

[7] M.-W. M. Chang, L. L. Ratinov, D. Roth, and V. Srikumar. Importance of semantic representation: Dataless classification. In *AAAI*, 7 2008.

[8] L. Nie, M. Wang, Z. Zha, G. Li, and T.-S. Chua. Multimedia answering: Enriching text qa with media information. SIGIR '11, pages 695–704, New York, NY, USA, 2011. ACM.

[9] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. WWW '08, pages 91–100, New York, NY, USA, 2008. ACM.

[10] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. Short text classification in twitter to improve information filtering. SIGIR '10, pages 841–842, New York, NY, USA, 2010. ACM.

[11] X. Sun, H. Wang, and Y. Yu. Towards effective short text deep classification. SIGIR '11, pages 1143–1144, New York, NY, USA, 2011. ACM.

[12] S. Zelikovitz and F. Marquez. Transductive learning for short-text classification problems using latent semantic indexing. *International Journal of Pattern Recognition and Artificial Intelligence*, 19(2):146–163, 2005.