2005－NL－170（2）
2005／11／21

イベントの生起時間帯判定

野呂太一[†]　乾孝司[††]　高村大也[†††]　奥村学[†††]

[†]東京工業大学大学院 総合理工学研究科
〒226-8503 横浜市緑区長津田町 4295
[††]日本学術振興会 特別研究員
[†††]東京工業大学 精密工学研究所
代表連絡先：norot@lr.pi.titech.ac.jp

ブログテキスト中のイベントの生起時間帯判定を行う，機械学習を利用した手法を提案する．従来の時間情報解析の研究は主な対象がニュース記事であり，明示的な時間表現の解析を軸としたイベントの時系列化などが主な目的であった．しかし，ニュース記事以外のブログなどで明示的に時間表現が記されることは稀なため，従来手法では限界がある．そこで本研究では，時間帯を連想させる表現を手がかりに，朝・昼・夕・夜の粒度でイベントの時間帯判定を行うことを目的とする．具体的にはナイーブベイズ分類器を EM アルゴリズムで補強する semi-supervised な手法を SVM と組み合わせ，時間帯情報分類器を作成する手法を提案する．
キーワード：時間情報解析，イベント，ブログ，機械学習，semi-supervised

Temporal Processing of Events in Text

Taichi Noro[†]　Takashi Inui[††]　Hiroya Takamura[†††]　Manabu Okumura[†††]

[†]Graduate School of Science and Engineering, Tokyo Institute of Technology
4259 Nagatsuta Midori-ku Yokohama, JAPAN, 226-8503
[††]Japan Society for the Promotion of Science
[†††]Precision and Intelligence Laboratory, Tokyo Institute of Technology
norot@lr.pi.titech.ac.jp

We propose a machine learning-based method for identifying when an event in blog text occurs: ,morning, daytime, evening, night. Earlier study analyzed explicit temporal expressions of events and mapped them on time-line in newswire texts. However, other texts such as weblogs contain few explicit temporal expressions. We therefore use various implicit temporal expressions extracted automatically. Specifically, we use naïve bayes classifiers backed up with the EM algorithm, and also use support vector machines.
Keywords：temporal information, event, blog, machine learning, semi-supervised

1. はじめに

近年，Web の発達とともに電子化されたテキストの量は増加を続けている．そしてそれらの中には，ある出来事，イベントについて記述されたものも少なくない．その代表としてニュース記事が挙げられる．

このようなテキスト群を“時間”に注目して整理する研究がある．その目的は，実際のイベントの発生とは異なった順序で記述されるニュース記事などを，実際の発生順に時系列に並べることで事実の理解を深めること，あるいは，ある一連のイベントについて複数回記述されている場合，時間情報をもとに整理することで要約の助けとすることである．

上記のような従来研究では，その多くがニュース記事のみを対象としていた．それは，ニュース記事が“3 日午後 3 時ごろ”や“今週金曜日”のように，明示的に時間情報を表記していることが多く，研究対象として扱いやすかったためである．しかし，イベントを記述しているテキストの全てがニュース記事のように明示的に時間情報を記

述しているわけではない．例えば Web 日記，ブログなど（以下まとめてブログと呼ぶ）もイベントを記したテキストと言える．ただ，個人が日常のイベントを記すブログでは，ニュース記事とは性質が異なり，イベントの記述の際に，その生起時間を明示的に示すことは稀である．

しかし，イベントの生起時間が明示されていないにも関わらず，人間がブログを読んだときには，その内容からおおよその生起時間が特定できる場合は少なくない．さらに，ブログの内容からイベントの生起時間の特定が可能になれば，検索における新たな軸としての時間情報の利用や，時間帯ごとの人々の行動統計（例えば“人は朝何を食べているのか”）などを把握することが可能になると考えられる．

そこで本研究では，ブログテキストを対象にし，テキスト中に記述されたイベントの生起時間帯を判定することを目的とする．もう少し具体的には，イベントを朝，昼，夕，夜の粒度に分類する．

明示的な時間表現がない場合に対応するために，機械学習手法を用いて自動的に時間帯を連想させる表現（以下，時間帯連想語）の情報を取り入れて，イベントの時間帯を特定する．

既存研究では，イベント情報の処理単位として，文書レベルから単語（動詞）レベルまでさまざまなものがあるが，本研究では文を処理単位とし，文毎に時間帯を特定する．ただし，ブログのような日記形式のテキストでも，全ての文がイベントを表した文というわけではない．そして当然のことながら，イベントを表していないその他の文に対して，イベントの生起時間帯判定を行っても意味がない．

これを踏まえ本研究は，

Step1: テキストからイベントについて記述された文を抽出する“イベント文抽出”，

Step2: 抽出された文のイベントの生起時間帯を判定する“イベント文の時間帯判定”，

の２つの課題を逐次的に行うことによって，文内で表現されたイベントの生起時間帯を判定する．

2. 関連研究

Setzer ら[1]や Mani ら[2]はニュース記事中の時間情報を解析するための取り組みとして，イベント，および時間情報へのアノテーションを研究目的としている．これによりイベント発生の絶対時間の決定を可能にするとともに，時間情報・イベント同士の相対的な順序関係に着目し，イベントの整列を行うことも目指している．小倉ら[3]は，ニュース記事を対象とし，一文章中のイベントの時系列化を目指した．特に，イベント同士の前後関係を求める時間推論に焦点を置き，データを絶対時間を持つものと，相対情報を持つものに分け，イベント群の時系列化を行っている．これらの研究はニュース記事を対象としたもので，明示的な時間情報がある程度含まれることが前提となっており，本研究とは方向性が異なるものである．

本研究と類似した目的を持つものに，土屋ら[4]の研究がある．これは，あらかじめ用意した時間判断知識のデータベースをもとに，未知語（時間判断データベースに存在しないもの）から連想される時間を導き出すものである．辞書の見出し語と説明文の関係を利用し，既知語と未知語の関連度を計算して，未知語から連想される時間情報を取得している．本研究では，辞書にあたるものを利用せず，人々の行動のデータ（ブログ）から時間情報の取得を目指している．これにより，辞書には記載されないような，日常生活に基づいた時間連想情報を得られると考えている．

3. コーパス

1 節でも述べたように，本研究ではブログを解析の対象とする．本節では，ブログエントリのテキストから作成するコーパスについて説明する．このコーパスは，4 節で説明する機械学習手法において，訓練・評価データとして使用する．

南野ら[5]が収集したデータを，1 文毎に自動的に切り出したものを使用する．ただし，ブログのデータは文末に正確に句点が記述されないことも多いので，正確に文に分割することは難しい．よって，ここでの 1 文とは，句点や HTML タグの情報によって単純にテキストを分割したものであり，不適切に分割されたもの*も含まれている．

3.1. タグ

各文に“event”，“time slot”の 2 種類のタグを付与した．それぞれについて説明する．

* 文分割エラーはおよそ 5%程度である．

3.1.1. event タグ

これは文がイベントを表しているか否かを“1”，“0”の 2 値で表したものである．文がイベントを表しているときは 1 を付与し，表していないときは 0 を付与する．イベントを表した文とは，発生した出来事について記述されたものである．イベントを表していない文とは，何らかの説明をしている，主張・感想を述べているものなどである．event=1 である文の例を例 1 に示す．

例 1

a. 福岡に行くために，羽田に行きました。

b. ひどく痛むので近くの整形外科に。

続いて event=0 である文の例を例 2 に示す．

例 2

a. 生姜紅茶を，一日一杯は飲んでます。(習慣)

b. ほんとにかわっちゃったの？ (台詞)

c. ご無沙汰しております。 (挨拶)

3.1.2. time slot タグ

これはイベントが生起した時間帯を“朝”，“昼”，“夕”，“夜”，“情報無し（時間帯不明）”の 5 値で表したものであり，event=1 の文にのみ付与される．各時間帯の目安として設定した定義を以下に示す．

朝：04:00~10:59，早朝から午前中，朝食

昼：11:00~15:59，昼から夕方の前，昼食

夕：16:00~17:59，夕方から日没前

夜：18:00~03:59，日没後から夜明け，夕食

文にそれぞれの値をつける判断は，文内の比較的明示的な表現によって付与可能になるものと，文内には明示的な表現がないが前後の文脈情報を見ることによって付与可能になるものがある．前者の例を例 3 に示す．

例 3

a. 朝から自転車で郵便局へ行く。(朝)

b. 昼は，定食屋で豚丼を食べた。(昼)

c. １６時過ぎには帰路につく。(夕)

d. 鍋を作り，その日の夕食とした。(夜)

後者の例を例 4 に示す．

例 4

文１．朝から自転車で郵便局へ行く。(朝)

文２．郵便局の帰りに某ショップへ。(朝)

ここで，文 1，2 は連続してブログに出現したものとする．この場合，文 1 を朝と判定し，それに続いて出現した文 2 も話の流れから朝だと判断できる．

例 5 のように，一文で複数のイベントを記述している文も存在する．

例５

今日は朝学校に行って，昼には弁当を食べ，夕方帰った。

このような場合は，文の末尾側に記述されているイベントのみに注目して，そのイベントの時間帯で判断することとした．つまり，例 5 では time slot=夕となる．

3.2. コーパス統計

コーパスは人手で作成している．ブログエントリの数は 6,158 であり、そこから分割した文の総数は 56,644 文である。ちなみに著者数は 238 人である。各タグの内訳を表 1，2 に示す．

表 1：event タグ内訳

event=1	10,903
event=0	45,741
計	56,644

表 2：time slot タグ内訳

time slot=朝	594
time slot=昼	491
time slot=夕	156
time slot=夜	906
time slot=情報無し	8,756
計	10,903

4. 提案手法

提案手法では，まずテキストからイベントにつ

いて記述された文を抽出する“イベント文抽出”を行い，次に，抽出された文のイベントの生起時間帯を判定する“イベント文の時間帯判定”を行う．本節ではこれらについて順に述べる．

4.1. イベント文抽出

文がイベントを表すか否かを，機械学習を用いて判定する．具体的には前述のコーパスを利用し，event=1 の文を正例クラス，event=0 の文を負例クラスとし，SVM によって分類器を作成する．

分類に利用する素性情報について説明する。文がイベントを表すか否かの判定には，文末に現れるモダリティ情報などが有効であると考えられる．これを次の 2 つの事例を用いて説明する．

例 6

a. 朝，トーストを食べた。(event=1)

b. 朝，トーストを食べる。(event=0)

例 6-a はイベントを表すが，例 6-b は習慣の説明をした文である．この場合，それぞれの文を構成する単語にはほとんど差異がないが，文末の表現によってイベント文か否かが判定される．このように，文末表現のタイプや，品詞の種類などがイベント文判定には有効だと考えた．

これを踏まえ，単語（名詞，動詞）に加え，表 3 に示すものを素性として使用した．

表 3：イベント文判定に使用した素性

最終文節に係る文節内の情報
“助詞-格助詞”の種類
“助詞-係助詞”の種類
末尾が副詞か
末尾が“助詞-連体化”か

最終文節内の情報
“助詞-格助詞”の種類
名詞・記号のみで構成されているか
動詞の有無
末尾の記号の種類
末尾が副詞か
末尾が“名詞-サ変接続”か
文末表現タイプ（横山[6]）

文節位置に関係のない情報
挨拶表現（失礼します等）が存在するか
“助詞-格助詞”の種類

ここでの文末表現タイプには横山[6]が使用したものを用いた．品詞体系は Chasen[7]に従う．

4.2. イベント文の時間帯判定

「朝食」という単語が，それを含む文を“朝”と判定する強い手掛かり，つまり時間帯連想語であることが分かっているとする．これによって，例えば「朝食にトーストを食べた」という文が“朝”であることが分かり，さらにこの文から，「トースト」が“朝”の連想語である可能性が高いことが分かる．このような考え方を繰り返すことにより，ブートストラップ的に時間帯連想語が獲得でき，同時に文を正しく分類できるようになると考えられる．

この考えを実現するためには，時間帯のタグが付けられたイベントコーパスを種として，時間帯のタグが付いていない大量のイベントコーパスを併せて利用する，半教師付学習を用いればよい．そこで，教師付き学習手法のナイーブベイズ分類器を Expectation Maximization(以下 EM)アルゴリズム[8]で補強する semi-supervised な方法を適用する．ナイーブベイズ分類器を用いたのは，EM アルゴリズムと組み合わせることにより，文書分類で高い性能を発揮することが Nigam ら[9]によって示されているからである．

4.2.1. ナイーブベイズ分類器による時間帯分類

ここではまずナイーブベイズ分類器（Naïve Bayes classifiers）の一種である多項モデルについて説明する．

多項モデルでは，カテゴリ c が与えられたときに，事例 x が生起する確率は，

$$P(x \mid c,\theta) = P(|x|)|x|!\prod_{w}\frac{P(w \mid c)^{N(w,x)}}{N(w,x)} \qquad (1)$$

となる．ここで，$P(|x|)$は長さ$|x|$の文が生起する確率であり，$N(w,x)$は文 x 中での素性 w の出現頻度である．文の生起は，全語彙の中から単語を一つ選び出す試行の繰り返しとして，モデル化される．

ナイーブベイズ分類器を文の時間帯分類に適用した場合，各文が事例 x に相当し，カテゴリ c は，朝，昼，夕，夜，情報無しのいずれかの値をとる．使用される素性は，文に出現する単語などである．

4.2.2. ナイーブベイズ分類器と EM アルゴリズムの組み合わせ

EM アルゴリズムはいくつかの変数（隠れ変数と呼ばれている）が観測できない状況で，モデルを最尤推定もしくは事後確率最大化推定する手法である．Nigam らはナイーブベイズ分類器と EM アルゴリズムを組み合わせることを提案している．

ナイーブベイズ・モデルの式において，関係ない要素を無視すると，次の式を得る：

$$P(x|c,\theta) \propto \prod_w P(w|c)^{N(w,x)}, \qquad (2)$$

$$P(x|\theta) \propto \sum_c P(c) \prod_w P(w|c)^{N(w,x)}. \qquad (3)$$

以降, モデルのパラメータ群をまとめて θ と表す．

c を隠れ変数とし，ディリクレ分布をパラメータの事前分布とすると，対数尤度の隠れ変数に関する期待値（Q 関数）は次のように定義できる：

$$Q(\theta|\bar{\theta}) = \log(P(\theta)) + \sum_{x\in D}\sum_c P(x|c,\bar{\theta}) \times \log\left(P(c)\prod_w P(w|c)^{N(w,x)} \right). \qquad (4)$$

ここで，$P(\theta) \propto \prod_c \left(P(c)^{\alpha-1} \prod_w \left(P(w|c)^{\alpha-1} \right) \right)$ であり，また，α はハイパーパラメータ，D はモデルの推定に用いられる事例の集合である．

この Q 関数より，次の EM 計算式が得られる：

E-ステップ：

$$P(c|x,\bar{\theta}) = \frac{P(c|\bar{\theta})P(x|c,\bar{\theta})}{\sum_c P(c|\bar{\theta})P(x|c,\bar{\theta})}, \qquad (5)$$

M-ステップ：

$$P(c) = \frac{(\alpha-1) + \sum_{x\in D} P(c|x,\bar{\theta})}{(\alpha-1)|C| + |D|}, \qquad (6)$$

$$P(w|c) = \frac{(\alpha-1) + \sum_{x\in D} P(c|x,\bar{\theta}) N(w,x)}{(\alpha-1)|W| + \sum_w \sum_{x\in D} P(c|x,\bar{\theta}) N(w,x)}. \qquad (7)$$

ここで $|C|$ はカテゴリ数，$|W|$ は素性の種類数を表す．ラベル付き事例については，式(5)は使用されない．その代わりに，c が事例 x のカテゴリならば $P(c|x,\bar{\theta})$ は 1 とし，そうでなければ 0 とする．

EM アルゴリズムの変種に tempered EM がある．この変種では，モデルの複雑さを調整することが出来る．tempered EM は，E-ステップで式(5)の代わりに次式を使用することで実現できる：

$$P(c|x,\bar{\theta}) = \frac{\left\{P(c|\bar{\theta})P(x|c,\bar{\theta})\right\}^{\beta}}{\sum_c \left\{P(c|\bar{\theta})P(x|c,\bar{\theta})\right\}^{\beta}}. \qquad (8)$$

ここで，β はモデルの複雑さを決めるハイパーパラメータで，値が大きいほどモデルは複雑になる．

ラベルなしデータに対してラベル有りデータが極端に少ないと，学習を繰り返していくうちにラベル無しデータの影響が強くなりすぎて，結果が悪くなってしまうことがある．そのため $\lambda(0 \le \lambda \le 1)$ を用いて，ラベル無しデータの影響が小さくなるように式(4)の右辺の第 2 項を次式と入れ換える：

$$\sum_{x\in D^l}\sum_c P(c|x,\bar{\theta}) \log\left(P(c)\prod_w P(w|c)^{N(w,x)} \right) + \lambda \sum_{x\in D^u}\sum_c P(c|x,\bar{\theta}) \log\left(P(c)\prod_w P(w|c)^{N(w,x)} \right).$$

ここで，D^l はラベル付きデータ，D^u はラベル無しデータである．この式が示すように，λ の値が小さいほどラベル無しデータの影響が小さくなる．

この新たな Q 関数を用いて導出したアルゴリズムを使用した．Q 関数の値の変化が十分に小さくなることを終了条件とした．

4.2.3. “time slot=情報無し”の文の問題点

ここで, 具体的な手法に入る前に, time slot の値に情報無しが付与された文の 2 つの問題点について述べる．

1 つ目は，時間帯を連想させる表現が存在しないという性質的な特徴である．他の値（朝～夜）が付与された文には，解析の焦点となる時間情報が含まれているが，この文にはそれが含まれていない可能性が高い．これにより，情報なしが付与された文では，他の値が付与された文と比べて素性の分布の特徴が著しく異なり，提案手法の計算に悪影響を与えることが予想される．

2 つ目は，他の値が付与された文と比べて，量が非常に多いという特徴である．表 2 を見ても分かるように，他のものと比べて 10 倍前後の差がついている．この差が，生起確率の計算に影響を

及ぼすことが予想できる．

以上のように time slot=情報無しの文は，分類器の学習において悪影響を及ぼす可能性が高いため，この問題点を考慮した分類器の作成を行う．

4.2.4. 時間帯分類手法

前述した問題点を考慮した，2 段階で分類器を作成する手法（以下，手法 A）について説明する．

1 段階目の分類器（以下，時間情報有無分類器）は，time slot の値が情報無しの文と，それ以外の文を分類する．この学習には SVM を使用し，使用した素性は品詞が名詞，動詞となる単語である．

そして，時間情報有無分類器によって time slot の値が時間帯情報有り（朝～夜）だと判定された文を，2 段階目の分類器（以下時間帯 4 値分類器）で朝，昼，夕，夜に分類する．学習には，前述したナイーブベイズ分類器と EM アルゴリズムを組み合わせたものを使用する．使用した素性は品詞が名詞，動詞となる単語である．

また，実験の比較対象として 4.2.3 節の問題点を考慮しない手法（以下，手法 B）を試す．これは，time slot の値が朝，昼，夕，夜，情報無しの文を分類する 5 値分類器（以下，時間帯 5 値分類器）を作成する手法である．学習には，前述したナイーブベイズ分類器と EM アルゴリズムを組み合わせたものを使用する．使用した素性は品詞が名詞，動詞となる単語である．

5. 実験と考察

5.1. イベント文抽出の結果

event=1 のデータを正例クラスとして 10864 文，event=0 のデータを負例クラスとして 45783 文使用して，SVM による分類実験を 10 分割交差検定で行った．結果を表 4 に示す．ソフトマージンパラメータは 0.2 である．なお，SVM の学習には TinySVM[10]を使用した．

表 4：イベント文分類結果

正解率	0.852
精度	0.660
再現率	0.477
F 値	0.551

単語以外の素性では，文末表現タイプが有効であったが，全体的にイベント文抽出はあまり良い結果を出すことが出来なかった．

イベント文判定は，人手でのコーパス作成の際にも，判断が困難な（曖昧な）文が多数含まれていたため，分類結果も悪くなったと思われる．

5.2. イベント文の時間帯判定の結果

まず，時間帯分類手法 A について説明する．時間情報有無分類器は time slot の値が時間帯情報有り（朝～夜）のデータを正例クラスとして 2147 文，情報無しのデータを負例として 8756 文使用して，SVM による分類実験を 10 分割交差検定で行った．結果を表 5 に示す．ソフトマージンパラメータは 0.5 である．

表 5：時間情報有無分類器結果

正解率	0.863
精度	0.717
再現率	0.475
F 値	0.562

時間帯 4 値分類器は時間情報有無分類器により，time slot が時間帯情報有り（朝～夜）だと判断したデータを用いてナイーブベイズ分類器＋EM アルゴリズムによる分類実験を 10 分割交差検定で行った．結果を表 6 に示す．ラベル無しデータには未知のデータ 58645 文を使用した．β の値は 0.01 である．

ベースラインは全ての文の time slot が夜だと判断した場合の正解率である．EM アルゴリズムの適用が成功していることが分かり，正解率でベースラインを 20%上回った．表 6 の上限値とは，仮に完璧な時間情報有無分類器が存在したとして，時間帯 4 値分類器を作成すると結果がどうなるかを実験したものである．つまり，理想的な環境での上限値を求める実験である．ラベル無しデータには同じく未知のデータ 58645 文を使用し，β の値は 0.01 である．

表 6：時間帯 4 値分類器結果

手法	正解率	上限値
ベースライン	0.422	0.422
ナイーブベイズのみ	0.504	0.587
ナイーブベイズ＋EM(λ=0.1)	0.581	0.646
ナイーブベイズ＋EM(λ=1.0)	0.624	0.668

次に，時間帯分類手法 B について説明する．time slot の値が朝，昼，夕，夜，情報無しのデータをそれぞれ，594 文，491 文，156 文，906 文，8756 文用いて，ナイーブベイズ分類器＋EM ア

ルゴリズムによる分類実験を 10 分割交差検定で行った．結果を表 7 に示す．ラベル無しデータには未知の（タグの付けられてない）データ 58645 文を使用した．β の値は 0.01 である．

表 7：時間帯 5 値分類器結果

手法	正解率
ベースライン	0.803
ナイーブベイズのみ	0.807
ナイーブベイズ＋EM(λ=0.1)	0.802
ナイーブベイズ＋EM(λ=1.0)	0.751

ベースラインは全ての文の time slot が情報無しと判断した場合の正解率である．EM アルゴリズムによって正解率は低下してしまった．また，ナイーブベイズのみでもベースラインをわずかに上回るのみである．

さらに 2 段階の分類器によって得られた最終的な手法 A の分類の正解率を，手法 B と比較して表 8 に示す．

表 8：手法比較

手法	最終正解率
手法 A	0.852
手法 B	0.807

時間帯分類手法 A と B の比較では，手法 A の方が良い結果（正解率で 4.5%上回った）を示した．これにより，4.2.3 節で述べた time slot=情報なしの文の問題点が分類器の学習に悪影響を与えていることが分かり，2 段階に分類器を作成する提案手法が有効であることが示せた．

5.3. ラベル無しデータの比較

EM アルゴリズムに影響を与える，ラベル無しデータとして最適なデータを調べる比較実験を行った．ラベル無しデータとしては，様々な種類の事例が含まれているもの，ラベル有りデータにできるだけ類似したものの適用が考えられる．

様々な種類の事例が含まれているものとして，未知データが考えられる．未知データにはイベント文ではない説明文などが多量に含まれるため，一見ラベル無しデータとしては不適切であるように思われる．しかし，イベント文以外にも時間情報を持つ文，例えば習慣の説明（“私は毎朝，トーストを食べます。”）などが含まれるため，正解率の向上に有効なデータである可能性がある．ラベル有りデータと類似したものは，データの素性の分布が類似しているので，正解率の向上が期待される．これには，時間情報有無分類器によって“time slot=情報無し”以外（朝，昼，夕，夜）だと判断されたデータが適当だと考えられる．そして，以上の 2 つの中間に位置するものとして event=1 であるデータもラベル無しデータとして有効である可能性がある．そこで，以下の 3 種類のデータで実験を行った．

データ 1：未知データ

データ 2：未知データからイベント分類器を使用して抽出した event=1 のデータ

データ 3：データ 2 から時間情報有無分類器によって“time slot=情報無し”以外だと判断されたデータ

訓練・評価データは 5.3 節の時間帯 4 値分類器で使用したものと同じであり，β の値は 0.01 である．結果を表 9 に示す．

データ 2 は λ の値をあげるにつれて正解率は上昇したものの，表 6 の結果と比べて，効果的とは言えない結果となった．データ 3 はラベル無しデータに影響する，λ の値をあげるにつれて，正解率が悪くなってしまった．データ 1 が 3 種類の実験の中で一番良い結果を示した．

表 9：ラベル無しデータ比較実験結果

λ	正解率		
	データ 1	データ 2	データ 3
0.1	0.562	0.490	0.544
1.0	0.596	0.528	0.521

データそれぞれに正解率の向上を期待したが，結果的にイベント文とそれ以外の文が混合されたデータが一番ふさわしいことが分かった．

5.4. 素性比較

本節では，手法 A の時間帯 4 値分類器の学習において，いくつかの素性を追加した実験を行う．追加した素性は，次の 3 種類である．

素性セット 1：係り受け関係にある名詞-動詞，名詞-形容詞のペア

素性セット 2：文内での位置情報（最終文節，最終文節にかかる節）

素性セット 3：前後の文の情報

β=0.01 で実験した結果を表 10 に示す．結果はどれも名詞，動詞のみの正解率 0.624 を下回ってしまった．

素性セット１は，“会社に行く”のような名詞-動詞の組み合わせから，“朝”を連想するような場合を想定したのだが，元の正解率を超えることは出来なかった．素性セット２は，例５のような場合を考え，最終文節内の情報が分類に有効だと考えたが，同じく元の正解率を下回った．3.1.2 節において説明したように, time slot タグには前後の文脈を見ることによって時間帯の判定が可能になった文が, 多数存在する. 素性セット３は，これを踏まえて追加した素性だったが，良い結果は出なかった．前後の文の情報を入れたことによって，必要な情報以上にノイズとなる情報を増やしてしまったと思われる．

表 10：素性比較実験結果

素性タイプ	正解率	最終正解率
素性セット１（係り受け）	0.617	0.851
素性セット２（文内位置）	0.604	0.849
素性セット３（前後文）	0.606	0.849

5.5. 取得連想語例

時間帯分類手法 A で得られた時間帯連想語の例を表 11 に示す．

表 11：取得連想語リスト

順位	朝	昼	夕	夜
1	圧雪	昼休み	夕方	花火
2	朝食	ちょうちょ	夕日	昨夜
3	今朝	授乳	松ぼっくり	更かす
4	パレード	お昼	乗り上げる	あっし
5	化す	昼飯	砂浜	知之
6	出港	湯麺	扇風機	弓
7	荷役	オムツ	道案内	夕食
8	午前	昼過ぎ	住職	ビーチ
9	朝	ランチ	夕暮れ	散らかす
10	ホイール	昼間	試飲	残業
11	開会	七夕	カジ	昨晩
12	通勤	午後	大森	夜
13	早朝	中華	主	夕飯
14	埋葬	ユキコ	集金	閉店
15	約定	昼食	下見	冷蔵庫
16	靴擦れ	ハヤシライス	受話器	夜中
17	不意打ち	天王寺	すべる	晩
18	まなこ	クレープ	帰路	毎晩
19	新神戸	祥	雲	ナポリ
20	成行	扇ぐ	坊	詩人

これは，素性 w が与えられたときにカテゴリ c のどこに出現しやすいかを表す，$P(c \mid w)$ の値で単語をカテゴリごとに降順に並べたものである．次に，ラベル無しデータのみに出現した時間帯連想語の例を示すと，“寝癖(朝:1432 位)”, “通学(朝:1703 位)”, “生ビール(夜:2013 位)”, “閉館(夜:2078 位)”などである．なお，訓練・評価・ラベル無しデータの出現単語総数はおよそ 22600 語であった．

6. おわりに

本研究では，テキストからイベント文を抽出し，そのイベントの生起時間帯を判定することを行った．連想語によって時間帯を判定するという考えを，機械学習の手法によって実現し，正解率で 85.2%という結果を出すことが出来た．

今後は，素性の改善などのほかに，文脈上での時間の流れも考慮するために，データ系列を学習し，品詞タグ付けなどで用いられる Conditional Random Fields を適用することも考えている．

[1] Andrea Setzer, Robert Gaizauskas. A Pilot Study on Annotating Temporal Relations in Text. In *Proceedings of the ACL-2001 Workshop on Temporal and Spatial Information Processing*, Toulose, France, July, pp.88-95, 2001.

[2] Inderjeet Mani, George Wilson. Robust Temporal Processing of News. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pp.69-76, 2000.

[3] 小倉牧人，田村直良．文間の時間制約モデルと事象の時系列化への応用に関する研究．情報処理学会研究報告「自然言語処理」，No.140-16, pp.111-118,2000.

[4] 土屋誠司，渡部広一，河岡司．連想メカニズムを用いた時間判断手法の有効性の検証．情報処理学会研究報告，2005-NL-168, pp.113-118, 2005.

[5] 南野朋之，鈴木泰裕，藤木稔明，奥村学．blog の自動収集と監視．人工知能学会論文誌，Vol.19, No.6, pp.511-520, 2004.

[6] 横山憲司，難波英嗣，奥村学．Support Vector Machine を用いた談話構造解析．情報処理学会自然言語処理研究会 NL-155, pp.193-200, 2003.

[7] http://chasen.naist.jp/hiki/ChaSen.

[8] Arthur P. Dempster, Nan M. laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, Vol. 39, No. 1, pp.1-38, 1977.

[9] Kamal Nigam, Andrew Mccallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, Vol. 39, No.2/3, pp.103-134, 2000.

[10] http://www.chasen.org/~taku/software/TinySVM.